Section 4

Technical Description Summary

PIP Preliminary Down-Select Process and Results

Major Requirements, Functionality, and Approaches

Key Technologies

Extensibility to SKA2
The CSP Element provides essential real-time signal processing facilities for the SKA. Inputs to the CSP are sampled signals from receiver elements, and outputs are used by the SDP for imaging and non-imaging science.

As defined in the Baseline Design, there are 3 telescopes requiring CSP capabilities. These are the low frequency aperture array (LOW-AA), the mid-frequency dish array (MID-DISH), and the survey dish array (SUR-DISH). For imaging, each of these requires a correlator. For non-imaging, currently only the MID-DISH telescope requires a Central Beamformer (CB) and a Non-Image Processor (NIP) in the form of a Pulsar Search engine, and a Pulsar Timing engine.

The LOW-AA, and SUR-DISH sub-elements (correlators) are located at the MRO (Murcheson Radio Observatory) site in Western Australia where the ASKAP telescope is located. The MID-DISH correlator, CB, and NIP sub-elements are located in the Karoo desert in South Africa where the KAT-7 test array and eventually MeerKAT telescope is located. In both cases, eventually both ASKAP and MeerKAT processing (and by extension receiver elements) will be subsumed and included in the SKA Phase 1 (SKA1) CSP Element. All investigations within this Consortium have assumed this to be the case.

### 4.1 PIP Preliminary Down-Select Process and Results

Due to the competitive nature of approaches and technologies to implement sub-elements of the CSP, a preliminary proposal and down select process was implemented within this Consortium. This was in the form of the requirement that proponents for particular approaches and technologies write "white papers" that flesh out in more detail their solutions. A guide for writing these white papers was developed (see Appendix J), a deadline was established, and a quasi-democratic process (i.e., vote of significant contributing Consortium organizations) was executed to try to select to carry forward in Stage 1 (and if successful Stage 2), only those approaches that would likely ultimately be feasible to implement, recognizing that each "Physical Implementation Proposal" (PIP) requires significant effort and management/system engineering overhead.

In broad-brush terms "hardware-centric" and "software-centric" white papers were written and evaluated. Roughly speaking, hardware-centric approaches rely on custom hardware solutions and a programming method that requires more intimate knowledge of digital hardware techniques, and software-centric approaches rely on commercial off-the-shelf (COTS) equipment using programming methods that do not require such intimate knowledge of digital hardware and therefore in principle (and in practise as demonstrated in existing systems) easier and faster to implement.

To level the playing field, only those technologies that had a high probability of being ready for full production in late 2016 were allowed to be used in white papers and associated performance and cost analysis. The purpose of pre-construction activities required by this Consortium is to, by time of completion in late 2016, be ready to move into the construction phase. Such a late-2016 restriction virtually guarantees that this is the case, and that the solutions proposed, studied, and costed in further detail could be firmly relied on for the CSP, although likely at least one “technology refresh” should occur before full SKA1 production if cost-effective to save cost and power.

The result of the white papers and votes pretty clearly illustrate that hardware-centric solutions, for end of 2016 technologies, have significantly less power and therefore operating costs (in absolute and relative terms) than do software-centric solutions and with fairly identical/competitive design-build (i.e., NRE and capital) costs. However, technology is moving fast and new programming techniques for hardware are being developed, and so a compromise approach that we believe will be beneficial to the SKA project as a whole is developed and included in this Proposal, to effectively involve experts in software approaches and methods in the Consortium’s efforts and provide a greater link/conduit for effort and methods with the SDP Consortium. The “.SE-STG1.SMART” work package has been created to consolidate these activities, and additionally a full LOW-AA software correlator will be investigated in at least Stage 1 to fully flesh out feasibility.
In the hardware-centric solution space of correlators and central beamformers, the baseline technology is FPGAs with allowance for investigation of other technologies such as multi-core CPUs and ASICs (Application Specific Integrated Circuits) where found to be cost-effective and where they have the capability of meeting requirements where FPGA approaches alone could be unwieldy, impossible, or too power hungry. However, it is important to note that each PIP lead identifies and determines which approach is used within each particular PIP. A range of platforms (PCB assemblies and surrounding infrastructure) are identified for use, currently “PowerMX”, “Redback”, and “Uniboard”.

The NIP Pulsar Search white paper presented approaches and costs for what could be done now with the large body of algorithms and software available, on COTS platforms they currently run on, might likely be done within a reasonable cost and power budget, and likely to meet Baseline Design goals. It is likely that further investigation and “cross-pollination” with hardware-centric technologies in correlators and beamformers could cause further drops in capital and operating costs in this area.

The NIP-Pulsar Timing problem is relatively small in terms of cost and operating budget and there is likely no situation where its implementation will not meet science requirements contained in the Baseline Design.

“Clock and Timing” within the CSP Element is called out as a separate sub-element but there was no white paper and associated downselect process required. It may be the case that clock and timing could use at least one hardware technology being investigated for correlator/beamforming processing and therefore may largely be a protocol/interface definition and fault-tolerance/redundancy problem to solve, although it is of course up to the clock and timing leader to flesh out in more detail. This sub-element is somewhat paradoxical in that it is a relatively simple problem (in terms of cost and power), but it is critically important that it be very reliable and fault-tolerant since if clock and timing for the CSP is down, the entire telescope is down.

LMC is also called out as a separate sub-element within the CSP but never went through a white paper and downselect process as it is largely a software effort and must be provided independent of any particular technology proposed for any sub-element implementation. It is a separate work package and is the interface, set of activities, requirements, and processing to tie together low-level processing, configuration settings, and monitoring with the Telescope Manager. Each of the 5 major sub-elements will require such software, each with different configuration and monitoring requirements. There are multiple methods for LMC implementation and interface to sub-elements and therefore information gathering including experiences and effectiveness of various approaches must be examined to develop an approach to be used for the SKA.

In all cases, world-leading experts in each of the sub-elements are engaged within this Consortium’s activities, whether leading or participating. Refer to appropriate sections of this Proposal for further information on individual and institute qualifications.

A table summarizing the lead institutes, approaches, and technologies to be investigated for each major sub-element is as follows:

<table>
<thead>
<tr>
<th>Sub-Element</th>
<th>Lead Institute(s) / Person</th>
<th>Approach(es)</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW-AA correlator</td>
<td>Oxford/Zarb-Adami</td>
<td>“Hardware-based”. Custom platform: PowerMX, Uniboard-2, or Redback, or other TBD.</td>
<td>Primarily FPGAs, but will consider ASICs and GPUs, possibly a mixture of technologies.</td>
</tr>
<tr>
<td>LOW-AA correlator</td>
<td>Curtain/Steve Ord</td>
<td>“Software-based” programming methods. Results could impact Stage 2 investigations in all correlator sub-elements.</td>
<td>COTS NVIDIA GPUs as a starting point, using all COTS equipment.</td>
</tr>
</tbody>
</table>
It needs to be said that in all cases sub-element leads are, and will continue to be, responsive to new developments, approaches, methods, and technologies during investigations while moving forward for construction readiness.

4.2 Major Requirements, Functionality, and Approaches

The CSP Element contains 5 major sub-elements, some located in Western Australia and some located in South Africa. The base functionality of each sub-element is defined in the Baseline Design and further described in each white paper used in the downselect process as described above. A summary of each sub-element’s requirements and functionality and technical approach is contained within the following sub-sections; refer to the Baseline Design, CSP white papers, and associated detailed technical descriptions (in Appendix J) for further detailed information.

4.2.1 LOW-AA—Correlator for the Low Frequency Aperture Array in Australia

The major requirements affecting correlator capital and operating cost for LOW-AA, as currently understood are:

- ~911 stations, one dual-polarization beam, 250 MHz/polarization at an approximate centre frequency of ~108 MHz. The actual observing frequency does not affect the correlator too much, only in determining the maximum earth-rotation phase rate as a function of delay rate that must be handled, if any\(^2\).
- Maximum baseline ~100 km. Must correlate all ~414k baselines; for baselines < 6 km must be able to integrate for at least 10.6 seconds and need not dump data faster than that; for other baselines must be able to dump as fast as every 0.6 seconds and need not integrate longer than that.
- Correlator must provide for correlation of 262,144 channels per baseline per polarization product. There is a possibility that with post-correlation RFI excision, fewer channels may be transferred to the SDP, but the actual number is TBD.

\(^1\) This is likely generally the case for other sub-elements as well.
\(^2\) i.e., if there is no effective down-conversion, then the earth-rotation phase correction is zero.
• Maximum ~30 dB spectral leakage between adjacent channels, maximum -60 dB leakage between any non-adjacent channels, and integrated dynamic range of at least 60 dB\textsuperscript{3}.

• Maximum ~1.5\%\textsuperscript{4} coherence loss in any part of any channel. Maximum un-calibrated systematic amplitude and phase variations are TBD.

There are many other fine requirements of course, but these are the major ones that have a major impact on correlator capital and operating cost. Further detailed requirements definition occurs in Stage 1.

The LOW-AA correlator must perform the following steps (closely following Figure 8 in the Baseline Design):

• Receive optical signals from all ~911 AA stations. This is nominally 8-bit (complex or real, in either case the data rate is the same) data, sampled at 500 Ms/s. It may or may not already be in the form of coarse frequency channels. Currently the “Optical Distribution” box of Figure 8 is not costed or considered in correlator white paper studies, and therefore is not currently considered part of the CSP.

• For each station, write data and read data from a delay tracking memory buffer such that the error between the actual implemented delay and a model of the delay out of this buffer is +/-0.5 samples.

• For each station, channelize the data into fine frequency channels. On the output of the channeliser rotate the phase of each channel to effect very fine delay (<< +/-0.5 samples) and earth-rotation phase correction (if required, as previously noted).

• Re-arrange the data through a “corner turner” such that each correlator entity (X-engine part) receives “time bursts” of identical frequency channels sourcing from every station. This allows all baselines for each time burst to be calculated. The maximum time burst length is typically some sub-multiple of the minimum integration time.

• Integrate and transmit data to the SDP. Due to the large number of baselines (and associated memory capacity and bandwidth requirements), a reduction in the net number of channels integrated and transmitted to the SDP is desirable.

There are two approaches (“PIPs”) that will be investigated for LOW-AA within this Consortium summarized in the following sub-sections.

4.2.1.1 Oxford-Led Hardware-Centric LOW-AA Correlator PIP

The University of Oxford, led by Kristian Zarb-Adami, will be the lead of this investigation. The approach, as surmised in the downselect white paper is summarized as follows:

• Use of custom-built hardware (the PowerMX common platform) and FPGAs will be the primary approach that will be investigated. PowerMX technology development is in a separate .TECH.PMX work package and therefore this sub-element-level investigation will primarily place technology requirements on PowerMX and then use the results in a “COTS-like” manner to build/implement the correlator.

• Uniboard-2 and Redback custom platforms will be investigated by participating JIVE/Astron and CSIRO members (respectively) of this sub-element (probably) in separate .TECH work packages. Results/capabilities will be investigated and considered at this sub-element level as described above.

\textsuperscript{3} This is not believed to be explicitly in the Baseline Design, but is a reasonable assumption. It is an important parameter as it determines the number of bits used in all station-based mixers/multipliers in the correlator.

\textsuperscript{4} This is not believed to be explicitly in the Baseline Design, but is a reasonable assumption allowing for the possibility of few-bit processing in the correlator if need be to save cost and power.
• ASICs and GPUs will be investigated and used if found to be cost-effective compared to FPGAs. These may sit on PowerMX boards or a separate carrier although the former might be desirable to minimize additional development costs.

• In conjunction with SMART work package investigations it may be the case that more complex processing devices (GPUs or multi-core CPUs) could be integrated into the correlator either tightly as PowerMX mezzanine cards or loosely via separate COTS boards networked/in-between the CSP and the SDP.

4.2.1.2 Curtain-Led Software-Centric LOW-AA Correlator PIP

A sub-consortium consisting of a number of members (ICRAR, VUW, ASTRON, NCRA, Cambridge University, and KASI) and led by Steve Ord at Curtain University in Perth Australia, will lead this investigation. The approach, as surmised from the down-select white paper is summarized as follows:

• Assume that the compute platform is a NVIDIA GPU product line housed in servers, connected via 10 Gb Ethernet and Infiniband. A particular technology/approach is not necessarily championed, but this is a starting point.

• All hardware and associated installation is purchased and/or contracted in a COTS-like manner and therefore all physical issues (safety, reliability, cooling, mechanics) are already addressed by industry engineering.

• Existing or soon-to-be available compute clusters and time on the MWA (at the MRO site) will be used for developing and testing software algorithms.

• The primary focus of this activity will be to flesh out a detailed PIP to act as a vehicle to fully explore the feasibility of software correlators for the SKA, both in the immediate term, and the long term for SKA2. If feasible, carrying forward software correlator approaches into Stage 2 will not be restricted to LOW-AA and may impact approaches for other CSP sub-elements. Everyone should definitely keep their pencils sharp.

4.2.2 MID-DISH—Correlator and Central Beamformer for the Mid-Frequency Dish Array in South Africa

The major requirements affecting correlator capital and operating cost for MID-DISH, as currently understood are:

• 254 antennas, one dual-polarization beam, 2.5 GHz/polarization, in the range of 1-10 GHz. The actual observing frequency does not affect the correlator too much, only in determining the maximum earth-rotation phase rate as a function of delay rate that must be handled. The Baseline Design was quite confusing in terms of actual required processed bandwidth and so 2.5 GHz was used as a baseline, but could change as required.

• Maximum baseline ~200 km. Must correlate all ~32k baselines; for baselines < 10 km must be able to integrate for at least 1.6 seconds and need not dump data faster than that; for other baselines must be able to dump as fast as every 0.08 seconds and need not integrate longer than that.

• Correlator must provide for correlation of 262,144 channels per baseline per polarization product. There is a possibility that with post-correlation RFI excision, fewer channels can be transferred to the SDP, but the actual number is TBD.

• Maximum ~30 dB spectral leakage between adjacent channels, and maximum -60 dB leakage between any non-adjacent channels.
Technical Description Summary

- Maximum ~1.5% coherence loss in any part of any channel. Maximum un-calibrated systematic amplitude and phase variations are TBD.
- Provide for 50 pulsar phase bins, with bin width as narrow at 1 msec, but trading off number of spectral channels for phase bins. The Baseline Design was not specific in either of these cases, and we feel that these parameters provide a reasonable trade-off that actually could be achieved in practice.

The major requirements for the MID-DISH central beamformer are as follows. Note that there are two distinct beamformers, one whose output provides real-time data for the Pulsar Search engine, and one whose output provides real-time data for the Pulsar Timing engine.

First, the Pulsar Search engine central beamformer:
- 2100 combined polarization beams at 300 MHz per beam. Maximum array aperture is 900 m comprised of 125 antennas. Each polarization must be independently beamformed, but polarizations for the same beam may be (intensity/power) added together before output to the NIP Pulsar Search engine. Each output beam is on a different delay centre on the sky and is a function of all 125 antennas. It is assumed no aperture amplitude taper is required, but it could likely be accommodated if necessary.
- 20 kHz spectral resolution, providing ~15,000 channels per beam. The Pulsar Search white paper, however, refines this number and indicates that 4000 channels (75 kHz resolution) is sufficient to achieve science goals and actually be implemented. At either of these resolutions, and for an array aperture of 900 m, “phase-delay” beamforming can be comfortably utilized reducing costs tremendously and keeping coherence loss integrated across each channel at <5% (noting that the White Paper TOC (Appendix J), page 12 directive of +/-45 deg is a factor of 2 too high).
- Must re-arrange/format the data such that each output beam spigot contains all frequency channels for just that beam squared and summed/integrated in time to minimize the data rate out of the beamformer.
- It should be noted that slight systematic side-effects in the data are not a major concern here and this may be exploited in implementation to reduce cost, in close consultation with Pulsar Search experts.

Second, the Pulsar Timing engine central beamformer:
- 10 dual-polarization beams at 1.4 GHz per beam. Must be able to beamform all 254 antennas in the array with baselines up to 200 km. Each output beam is on a different delay centre on the sky. It is assumed no aperture amplitude taper is required, but it could likely be accommodated if necessary.
- Timing resolution of 100 nsec. Naively, this implies a spectral channel width of 10 MHz, however Pulsar Timing experts (van Straten, Bailes) indicate that 1/16th of this or ~625 kHz channels are possible, noting that the impulse response/pulse smearing effects of poly-phase filterbanks must be carefully taken into account so as not to limit timing resolution. The actual spectral channel width and beamforming operations are therefore TBD.
- Superior channel-to-channel isolation (requirements TBD), signal fidelity, and time-keeping are required so as to provide absolute accuracy in Pulsar Timing measurements. To this end, it is advisable to minimize the amount of processing that signals undergo before delivery to the Pulsar Timing engine.
- Must re-arrange/format data such that each output beam spigot contains all frequency channels for just that beam. In this case, the voltage signal must be delivered to the PST engine.

There are many other fine requirements of course, but these are believed to be the ones that have a major impact on correlator and central beamformer capital and operating cost. Further detailed requirements definition occurs in Stage 1.
The MID-DISH correlator and beamformer must perform the steps as defined in Figure 16 of the Baseline Design. These are similar to those described for LOW-AA, except with the addition of central beamforming and will not be redundantly described here. Refer to the “Feasibility White Paper, SKA CSP, 1.3.2.2 SKA.TEL.CSP.DBF-STG1.MID.PIP”, Carlson et. al, in Appendix J for details on signal processing steps.

There is one approach (PIP) that will be investigated for MID-DISH within this Consortium summarized in the following sub-section. The approach is flexible in that it may allow for concurrent or mutually exclusive operation of the correlator and beamformer(s), depending on science requirements and cost.

4.2.2.1 NRC-Can/SKA-SA-Led MID-DISH Correlator and Central Beamformer PIP

NRC Canada (National Research Council), in close cooperation and coordination with SKA South Africa (SKA-SA) will lead this investigation, with possible contributions from MDA Corporation (under contract to NRC-Can), and contributions from STFC-RAL UK, New Zealand, INAF-Italy, and CETC-38 (or “KLASSA” as it has recently be renamed) in China.

The approach, as contained within the downselect white paper is as follows:

• Use the PowerMX common platform as the baseline PCB assembly, using FPGAs. Mounting of PowerMX boards in blades and plugging into backplanes currently seems attractive and technologically feasible, although other connectivity approaches using fibre and passive or active signal routing will be considered.

• (As with LOW-AA) PowerMX technology development is in a separate .TECH.PMX work package and therefore this sub-element-level investigation will primarily place technology requirements on PowerMX and then use the results in a “COTS-like” manner to build/implement the correlator. What this means in practical terms is that the bulk of the sub-element level effort is performing device utilization and I/O studies to determine how many boards, blades, and interconnect signals are required, feeding those to PowerMX .TECH to determine feasibility and cost, and then rolling up the results to define/design the complete sub-element.

• The full flexibility of PowerMX is in the use of mezzanine cards to hold processing devices, however, “hardened versions” of PowerMX (i.e., “PowerHX” not using mezzanine cards) is the fallback position if the mezzanine approach proves infeasible.

• ASICs will be investigated for implementation of some parts of the correlator/beamformer if it is found that implementation in FPGAs is impossible, unwieldy, or that significant power savings can be had in terms of an effective return on investment. Fair warning is given to science teams that pulsar phase binning requirements could push implementation of the X-part of the correlator into an ASIC if FPGA external memory bandwidth and processing capacity is exceeded.

• In conjunction with .SMART work package investigations it may be the case that more complex processing devices (GPUs or multi-core CPUs) could be integrated into the correlator either tightly as PowerMX mezzanine cards or loosely via separate COTS boards networked/in-between the CSP and the SDP.

4.2.3 SUR-DISH—Correlator for Survey Dish Array in Australia

The major requirements affecting correlator capital and operating cost for MID-DISH, as currently understood are:

• 96 antennas (comprised of 36 existing ASKAP dishes and 60 SKA dishes), 36 dual-polarization beams, 500 MHz/polarization, in the range of ~700-1600 MHz. The actual observing frequency does not affect
the correlator too much, only in determining the maximum earth-rotation phase rate as a function of delay rate that must be handled, if any.

- Maximum baseline ~50 km. Must correlate all 4560 baselines for every beam, meaning that there are essentially 36, 4560 baseline correlators. For baselines < 10 km must be able to integrate for at least 1.4 seconds and need not dump data faster than that; for other baselines must be able to dump as fast as every 0.3 seconds and need not integrate longer than that.

- Correlator must provide for correlation of 262,144 channels per baseline per polarization product. There is a possibility that with post-correlation RFI excision, fewer channels can be transferred to the SDP, but the actual number is TBD.

- Maximum ~30 dB spectral leakage between adjacent channels, and maximum -60 dB leakage between any non-adjacent channels.

- Maximum ~1.5% coherence loss in any part of any channel. Maximum un-calibrated systematic amplitude and phase variations are TBD.

No central beamformer is explicitly called out for in the Baseline Design.

The SUR-DISH correlator must perform the steps as defined in Figure 26 of the Baseline Design. These are similar to those described for LOW-AA, except that there are 36 identical correlators and will not be redundantly described here. Refer to the “SKA1-Survey Physical Implementation Proposal Feasibility White Paper”, Ensor et. al, Section 2 in Appendix J for details on signal processing steps.

There is one (PIP) that will be investigated for SUR-DISH within this Consortium summarized in the following sub-sections. It is multi-facetted, will initially consider many technologies and platforms, and as quickly as possible will internally down-select and develop the full PIP package only for those believed to be competitive at SRR/PDR.

### 4.2.3.1 New Zealand-Led SUR-DISH Correlator PIP

The New Zealand Alliance, led by Andrew Ensor from the Auckland University of Technology, will be the lead of this investigation. The approach, as surmised in the down-select white paper is summarized as follows:

- The investigation starts with a broad approach looking at multiple technologies and platforms for implementation. These, along with preliminary investigations, are described in detail in the associated white paper. Platforms include Redback, PowerMX, and COTS CPUs/GPUs. FPGAs within Redback and PowerMX, as well as multi-core CPUs, ASICS, and/or GPUs on PowerMX mezzanine cards (or other cards) will be explored.

- As quickly as possible within Stage 1, down-select is performed so as to carry forward into more detailed investigations only those technologies that warrant further effort. The key metrics will of course be cost (capital) and operating power (operational cost).

Leadership of the New Zealand Alliance in the SUR-DISH correlator PIP, in conjunction with PowerMX technology development, and involvement in the .SMART work package brings a holistic multi-level approach that can only serve to provide the best end result to meet SUR-DISH SKA processing and imaging results provided that efforts and resources are effectively focussed towards the end goal of building the instrument.
4.2.4 NIP—Pulsar Search

The NIP Pulsar Search engine, located in South Africa at the SKA site, and taking as its input the ~2100 beams\(^5\), 300 MHz/beam MID-DISH central beamformer output is a challenging task simply because of the scale of the problem; the algorithms are very well established and tested, but the search parameter space is enormous.

During the preliminary down-select process mentioned previously, a white paper fleshing out the Pulsar Search engine, processing steps, technologies, costs, and power was written and can be found in Appendix J. From that white paper, the processing steps for each beam are summarized as follows:

- **Sum polarizations and magnitude.** After central beamforming and per-polarization residual delay and phase calibration/correction, both polarizations for a particular beam are summed and squared to form a short-term integration such that a time-series of spectral power values are further processed. Depending on interface definition, this operation could be performed within the beamformer or the Pulsar Search engine. The required time resolution is 53.2 µsec, indicated that each power point for each particular frequency is the sum of the squares of 4 contiguous samples, at 75 kHz channel resolution.

- **RFI mitigation.** This will likely occur within the beamformer and the Pulsar Search engine, as appropriate. The .SE-STG[1|2].RFI work package within the WBS will help define what steps are needed where.

- **Incoherent dedispersion.** Radio signals propagate through the interstellar medium and get dispersed in time as a function of frequency, with the lowest frequencies dispersed (arriving later) than the higher frequencies. When searching for pulsars, therefore, the first step is to search through a wide range of dispersion measures (DMs), estimated at 16000 to reach the maximum DM.

- **Acceleration processing.** Pulsars can be in binary orbits and have higher order (acceleration) terms that affect the regularity of pulses arriving at the observer. Thus, each DM must be searched in acceleration space and the corrections can be made in either time or frequency (Fourier-transformed time) space, the specifics of which will be investigated in Stage 1.

- **FFT & Harmonic Sum & Threshold detection.** A large swath of DM and acceleration-searched time series are Fourier Transformed, harmonic summed, and searched for peaks to detect possible pulsar candidates.

- **Candidate detection.** This is a “sifting algorithm” that, as best and automatically as possible tries to sort out real pulsar candidates from RFI or other sources of confusion. An important part of this step is cross-beam comparison since RFI will typically show up in multiple beams but an actual pulsar should show up in only one.

- **Candidate data.** Perhaps millions of candidate pulsars will be detected, and more sophisticated selection/learning will be required. This step may occur in the SDP.

Refer to the Baseline Design Figure 18, and the Pulsar Search white paper (“Draft White Paper on Non-imaging Processing Pulsar Search Sub-element”, Stappers et. al., V1.0, 2013-05-03) in Appendix J for further details on Pulsar Search processing.

4.2.4.1 “Time Domain Team” Led Pulsar Search PIP

The single Pulsar Search PIP will be led by the “Time Domain Team”, headed by Ben Stappers at the University of Manchester. From the white paper investigation, it is clear that the Pulsar Search engine will dominate CSP capital and operating costs, although it would seem that at significantly lower cost and power

---

\(^5\) From the Pulsar Search engine implementation perspective, the aperture and number of antennas used to form each beam are free parameters.
than estimated in the Baseline Design. This team is well aware of the challenges involved, in particular operating power, and is poised to employ a range of technologies suited to the task.

The general approach will be to use NVIDIA GPUs for complex processing tasks and FPGAs (and if cost-effective, ASICs) for accelerator tasks such as FFTs and acceleration trials as the starting point, noting that new technologies are emerging for programming FPGAs (in particular “System C” for Xilinx FPGAs, and OpenCL for Altera FPGAs) which could have an impact on the final result. Some research results from the SMART work package in this regard might find utility here.

In any case, the final implementation is likely to be some sort of hybrid of technologies, each one suited/optimized for the complexity and sheer brute force requirements of the task, to provide the maximum science capability within a reasonable/affordable budget. It seems likely that a significantly capable Pulsar Search engine will emerge from this effort.

### 4.2.5 NIP—Pulsar Timing

The NIP Pulsar Timing Sub-Element located at the South Africa SKA site, and taking as its input 10 beams/polarization and 1.4 GHz/polarization/beam is, in comparison to the Pulsar Search and other CSP correlator/beamformer problems, a relatively small problem in terms of cost and power. It is making measurements of known pulsars over an extended period of time to determine residuals/changes in such known pulsars’ spin and orbital parameters. The key issues are fidelity and absolute timing accuracy of results, which place requirements on other CSP sub-elements that are not onerous, but require special due care and attention.

The main requirements for the Pulsar Timing Sub-Element are as follows:

- 10 beams/polarization 1.4 GHz/polarization/beam. (All 254 antennas\(^6\) in the array are summed, noting that this does not affect Pulsar Timing implementation.) Polarizations must be kept separate and cannot be added as in the case of Pulsar Search.

- 100 nsec final timing resolution. This implies 10 MHz channel bandwidth, but this is not the case. The Pulsar Timing white paper located in Appendix J makes this point, and indicates channel bandwidths as low as \(~10\) MHz/16 ~= 625 kHz are possible; however, the pulse-smearing effects of long FIR poly-phase filters must be taken into account and so further investigation is required.

- From the Baseline Design, 3–12 minutes of integration on each pulsar to acquire stable integrated pulse profiles, not taking into account pulse jitter noise effects.

#### 4.2.5.1 Swinburne-Led Pulsar Timing PIP

The Pulsar Timing PIP will be led by Willem van Straten, Matthew Bailes, and Andrew Jameson at Swinburne University in Melbourne Australia. This is world-leading expert team. Based on their white paper (“Feasibility White Paper, Non-imaging Processor – Pulsar Timing, Physical Implementation Proposal” contained in Appendix J), it will be entirely implemented using COTS GPUs and can be done so for a cost and power budget that is essentially in the noise of the cost and power of the rest of the CSP Element. Refer to the white paper for further detailed information.

\(^6\) Fewer may be summed if the effect on sensitivity is negligible and beamforming complexity and cost is reduced appreciably.
4.2.6 .TECH System Engineering Work Package

The .TECH work package is a place for one or more technology-level developments that may be common across CSP sub-elements and even applicable to other SKA elements. “Technology-level developments” encompass anything that might be common across sub-elements, and require specialized infrastructure to develop and make available (i.e., generally not COTS).

Technology developments could occur here and be used by one or more sub-elements, or each sub-element lead could entirely contain such developments within its particular sub-element. At this level, technology requirements flow from sub-element work initially in a quasi-informal query-response/iteration manner, and then if feasible, in more formal requirements and work execution manner (noting that the currently known technology platforms/approaches listed below have a pretty good start on this process already).

Solutions developed in this work package then can be used by the progenitor sub-element and be made available and used by other sub-elements if so desired. This work package is concerned with technology solutions and is not concerned with the details of specific signal processing requirements, which is the domain of sub-element work.

Currently there are .TECH WBS sub-work packages for the PowerMX platform, the Redback platform, and the Uniboard-2 platform; more can be added as seen fit. CSP Lead’s ultimate vision is for these varying solutions to converge on a single ubiquitous common platform that can be employed with a shared development and use model across sub-elements. Selection of which platforms and technologies to carry forward to Stage 2 will occur at SRR/PDR. This is a vision, and not a requirement however, and room must be made for different approaches and solutions to be investigated and used.

Refer to the 1.2.28.1 SKA.TEL.CSP.SE-STG[1|2].TECH.PMX, 1.2.28.2 SKA.TEL.CSP.SE-STG[1|2].TECH.RDBK, and 1.2.28.3 SKA.TEL.CSP.SE-STG[1|2].TECH.UNI2 WBS descriptions and activities, as well as associated technical descriptions in Appendix J for further information on each platform.

4.2.7 .SMART System Engineering Work Package

The .SMART (Software Methods, Approaches, Research, and Technologies) work package is in place to deploy significant expertise in high-performance computing and software approaches, available within or associated with the Consortium, to a set of problems that can impact all CSP sub-elements, “cross-pollinate” processing and approaches/interfaces with the SDP, and all with the goal of providing a holistic approach to tackling the significant total signal/image processing problem for maximum scientific benefit for the SKA.

There will be a mirrored .SMART work package in the SDP to act as a conduit for sharing and use of results within each element. In past radio telescope systems it has been something of a cliché that “it will take years before image processing performance can keep up with correlator output”. It is fine to make this statement, but it does nothing to tackle the problem and therefore maximize quality of science output. .SMART aims to help facilitate taking on this challenge in the following ways:

- Investigate and develop software methods for programming a range of technologies (e.g., GPUs, CPUs, FPGAs, perhaps other DSP/compute devices), and inform CSP and SDP PIPs/sub-elements of these results. This is especially relevant as FPGAs manufacturers are now recognizing that it is to their advantage to provide software methods for programming their devices.
- Considering the whole signal/image processing problems, what are some things that might be done within the CSP to cost-effectively reduce the (already enormous) effort/compute problem in the SDP. For

---

7 Commonality is not a strict requirement; a particular sub-element could push its sub-element-specific technology development here if so desired.
example, is there some special post-correlation processing or networking or storage that might occur to optimize the entire image-processing pipeline? What is the optimized interface/data format for data flow from the CSP to the SDP?

- Facilitate cross-pollination of technologies within the CSP and SDP for use as accelerators or more complex/optimized processing platforms/approaches within each element. For example, some developments in .TECH, coupled with software programming methods might be employed in the SDP to reduce power and increase performance. In another example, some processing chips and techniques developed in the SDP/SMART might be incorporated into the CSP to provide better RFI mitigation or pre-SDP processing.

The results of this work package will be reports to PIPs and a final report deliverable, for each Stage, for SKAO consideration. It is envisioned and hoped that during Stage 1 and 2, significant iteration and communication between .SMART, .TECH, and sub-element PIP/DDs activities will occur to take advantage of the results of this effort.

### 4.3 Key Technologies

From the preliminary PIP down-select process and investigations contained within the white papers feeding into it, it is reasonably clear that the key baseline technology that will enable CSP correlator and beamformers to meet requirements within a reasonable capital and development cost, for quite a reasonable operating (power) cost, and for relatively low risk is FPGAs. In particular 20 nm and 14 nm FPGA technology is reasonably assured within the near future (by the end of pre-construction) and smaller feature sizes will only improve the situation if they become available. Other technologies will be investigated in each PIP at each PIP leads’ discretion (such as ASICs, specialized signal processing chips, multi-core devices etc.), however, there would have to be obvious advantages over FPGAs to fully carry forward such technologies to SRR/PDR as they may incur more risk and development cost. Some of these may indeed be needed depending on processing requirements. GPU/software-based correlator approaches seem in the white paper investigations to consume considerably more power and therefore significantly more operating cost, however one software-based PIP will be carried through to SRR/PDR to further flesh out feasibility and cost of such an approach as technology is constantly improving and we believe further investigation is warranted. Additionally this competition (S/W using COTS equipment vs. H/W often using custom solutions) has additional advantages in spurring innovation in both worlds.

Within the Pulsar Search Sub-Element, a mixture of COTS/GPU and custom or COTS FPGA technologies is likely. The Pulsar Timing Sub-Element appears to quite comfortably be implemented in COTS GPU technologies and is not a concern.

Other key technologies are likely to be:

- High-speed, high-capacity memory and memory modules. These are crucial to mate with FPGAs, ASICs, or any other chips for large memory capacity with significant memory I/O bandwidth. Requirements should be able to be met with existing DDR3 SDRAM technology, however, likely DDR4 will be mature soon and utilized in pre-construction design. In some cases ASIC implementations may use on-chip eDRAM (embedded DRAM) if science requirements exceed external memory bandwidth requirements.

- High-speed multi-mode fibre optic technology in the form of industry-standard COTS or MOTS modules. Currently, ~10 Gbps per fibre transmission capability is well-established and the CSP does not rely on anything faster (but could benefit of course if the cost and utility is clear).

- Active cross-connect (switch) or passive fibre mesh technology. It doesn’t seem to be the case that technologies like wavelength switching or other exotic capabilities would be required.
• There is something of an “arms race” in copper and fibre-based interconnects and drivers/receivers in the industry. It may be the case that copper interconnects within portions of particular CSP sub-elements might be feasible and not require fibre, potentially saving cost and interconnect complexity.

• High-performance PCB technology such as HDI (High Density Interconnect) microvias, controlled impedance, using the most advanced and yet affordable PCB technologies (e.g., Megatron-6) available. In all cases, only the technology needed to support the application should of course be employed. It doesn’t seem to be the case that requirements will exceed technologies in this regard.

• High-performance, high-efficiency low voltage DC power supplies (LVDC). Processing devices require low voltages at high currents and the power supply industry has responded with some spectacular capability.

• Thermal solutions. Air cooling is preferred as it is the simplest and most reliable. Liquid cooling may be employed at various levels where air cooling is not capable.

• COTS networking and CPUs/SoCs (System on a Chip) for monitor and control. There doesn’t seem to be any case where requirements exceed technologies here.

### 4.4 Extensibility to SKA2

The RfP package does not seem to place much emphasis on the importance of technology approaches’ extensibility to SKA2. Within the white paper preliminary downselect process, Consortium participants were asked to “make statements and cost estimates on viability of scalability to SKA2” (see Appendix J), somewhat hedging this uncertainty with at least asking proponents to address the issue.

All white paper authors responded and made statements in this regard and indicated that with ~4X silicon density improvement over late 2016 technology, construction of the CSP for SKA2 would not be out of the question, noting that correlators and central beamformers get much larger, but the NIP sub-elements are unaffected since the number of beams that need NIP processing does not change as long as (in the Pulsar Search case) the physical aperture of ~900 m that forms those beams does not change.

In one case, the PowerMX platform was specifically conceived with enough I/O capacity such that it would take several years if not a decade or more for silicon processing to catch up, virtually guaranteeing that it would be viable as a platform for SKA2 as designed. Other approaches would see generations of silicon advancements incorporated into the basic design, adding I/O in-concert with silicon advancements. Yet other approaches would rely on “big industry” keeping up with the technology curve, and purchasing COTS components for implementations.

In summary, it would seem that there is no fundamental and practical technology limitation to constructing the CSP for SKA2, either building out/on from existing developments, or developing new platforms using the technologies of the day.
Canada-led SKA CSP Consortium—Description of Constraints, Metrics, and TOC for Feasibility White Paper required for CSP Technical Proposal and Preliminary Downselect; including Downselect Method and Timeline

Author: B. Carlson
Version: 1.1
Date: April 22, 2013

Introduction and Purpose of White Papers

The SKAO RFP states that we must provide a “Technical description of the proposal” in our response (SKA-TEL.OFF.RFP-SKO-RFP-001, Revision 1, section 8.1) as well as, among others, a “Budgetary estimate”. As there are multiple Subsystems within the CSP Element and likely competing notions of the technology and approach to each of these, we require that the lead of each Subsystem “PIP” develop a short White Paper that establishes foundational feasibility for the purposes of inclusion in the response to the RFP and to determine if the proposed approach will be managed and carried forward within this (NRC/MDA-led) CSP consortium (i.e. a preliminary downselect). Additionally, White Papers for supporting baseline technologies and approaches, if not already established as COTS solutions, must be provided.

As we only have limited resources (in real paid FTE terms within the consortium and in general terms for the SKA project pre-construction budget as a whole) to manage and run the consortium we must limit ourselves to approaches that are ultimately feasible if carried forward to construction to give us the best chance of having our Proposal accepted by the SKAO.

White Paper TOC

The White Paper TOC (Table of Contents) is as follows. Feel free to add further clarification sections that help to support your position.

- Subsystem description and how it fits into the CSP Element to demonstrate understanding and scope of the problem. A top-level simplified CSP Element context diagram, key Element level requirements, and explanation is required as well.
- Signal processing definition/scope, again, to demonstrate understanding of the problem. Much of this is contained in simplified form in the SKA1 SYSTEM BASELINE DESIGN (SKA-TEL-SKO-DD-001 Revision 1), but should be re-stated and repeated here for completeness and clarity.
- Technical description in as much detail as possible to establish conceptual feasibility. Identify risks, tradeoffs, fallback positions, risk mitigation etc.
• Rough development schedule trying to keep within the schedule outlined in the RFP package (see section 3.1.3, page 9, of STATEMENT OF WORK FOR THE STUDY, PROTOTYPING, AND DESIGN OF AN SKA ELEMENT, SKA-TEL.OFF.SOW-SKO-SOW-001, Revision 1, 2013-03-11). This will help identify activities that must be started sooner (such as ASICs) or later (COTS solutions).

• A performance analysis against the performance established in relevant CSP sections of the SKA1 SYSTEM BASELINE DESIGN.

• NRE, construction, and operating costing in as much detail as possible using the applicable metrics and categories (or better) presented in the following Cost Constraints and Cost Metrics section. The better your cost estimates, the more credible will be your proposal.

• Statements and cost estimates on viability of scalability to SKA2. For antenna-based processing (including central beamforming), multiply all capital and operating cost numbers by 10. For baseline-based processing (correlation), multiply all capital and operating cost numbers by 100. Non-imaging processing requires no multiplier. Project technology use 4 years beyond the technology freeze date for SKA1 of end of 2016, for the same operating lifetime as SKA1. The importance/weight given to extensibility to SKA2 is incorporated in the downselect decision process.

Cost Metrics & Constraints

The SKAO’s RFP has left something of a hole in the information needed to perform this downselect, in particular the NRE, capital, and operating budget for the CSP Element. We issued a formal request to try to obtain the most accurate numbers possible on some of these from the SKAO (see page 7), but did not receive any quantitative reply. Therefore, a first cut at these constraints and metrics is as follows:

Cost Constraints

1. Total SKA1 construction (including pre-construction) budget ~$600 M Euros. Maximum total CSP pre- and construction budget ~<10%, ~$60 M Euros (max). Other telescopes (EVLA, ALMA) were in the 2-3% range for cost of the correlator (NRE and capital) compared to total telescope costs/value, although they never had ambitious central beamforming and pulsar searching requirements, so 10% might be about right.

2. For the purposes of total life cycle cost comparisons (which is the ultimate comparison), use 10 years of operating life (starting after delivery of the first installation of the particular Subsystem to the SKA site(s); your proposal could include multiple installations and in this case the full cost of each development and installation must be included) as the metric.

3. Use reasonable cost and performance estimates for technology nodes (and technology types) used in implementation backed by sufficient industry evidence; for construction build of the first installation use end of 2016 as the technology node freeze date (the technology node must be in full production at that time). Include NRE refresh and retest/qualification of the design prior to production to obtain use of the latest technology unless otherwise justified.
4. All technology components in concepts must be in advertised literature/data sheets at TRL6 or higher now or at the latest, the time of Stage 2 start (“mid-2014) and be ready for qualification in demonstrable prototypes in Stage 2 at TRL7 or higher. Refer to the SKA SYSTEM ENGINEERING MANAGEMENT PLAN, SKA.TEL.SE-SKAO-MP-001 Revision 1, page 39, for definition/explanation of TRLs. In practical terms this means that from a “catalog engineering/published specifications” standpoint, it must look feasible now. Forecasting to technology in production at the technology freeze date should be reasonable and have a high degree of confidence. If technology is not at these TRLs, special attention/flagging of risk and risk mitigation must be paid within the White Paper.

5. The boundary for this calculation is as follows. On the input side assume wideband sampled data from the antennas but do not include the cost of the (fiber) receivers or power in the calculation. On the output side (of the correlator), include output (transmitter) capital and power costs, but not any subsequent switching fabric or long-haul data transfer. Otherwise for all intra-CSP Element data transmission, include these costs as best as possible in your estimates.

6. Include software development costs to get the Subsystem’s components at a reasonable API (Application Programmer Interface) level for use by/plug into the LMC. State what this API is (examples: poke/peek via packets; xml configs to embedded processor).

**Cost Metrics**

1. Operating budget, including personnel, spares, repairs, power, for both sites, maximum 10% of construction, but likely 5% is the target and is therefore 3-6 (absolute max) M Euros/year.

2. As a baseline personnel cost, use 100,000 Euros per FTE year.

3. Grid power costs, not including power delivery infrastructure etc. is ~1 Euro per Watt per year. Off-grid power costs, not including power delivery infrastructure etc. is ~2.5-3X grid costs. The South African site is on grid power. The Australian site is off-grid.

4. Mains power delivery efficiency ~97%. E.g. if a piece of equipment plugged into mains power consumes 1 kW, the actual used mains power is ~1.03 kW.

5. Heat removal costs are roughly 50% of power consumption costs. E.g. if a piece of equipment consumes 1 kW, the total mains power is ~1 kW/0.97 * 1.5 ~ 1.55 kW. This includes the full scope of heat removal costs, including fans, liquid transfer, heat exchangers etc.

6. Power delivery infrastructure capital cost is ~1.5-2 Euros per W. E.g. if a piece of equipment uses 1 kW, it costs about 1500-2000 Euros to install the infrastructure to deliver that power. Similar numbers for capital costs for heat removal. Use 5% per year of power infrastructure capital costs for yearly maintenance costs.

7. For GPUs/CPUs, use 80% utilization as a baseline for performance, power, and capital costs unless you have evidence of more or less efficient use with quoted or actual benchmark tests. In any case state the basis of estimation and note in risk.

8. For FPGAs use 80% utilization (and internally, 70% of rated internal speed with one logic element LUT level between pipeline stages) as a baseline for performance, power, and capital
costs unless you have evidence of more or less efficient use with quoted or actual bench mark tests. In any case state the basis of estimation and note in risk.

9. For GPUs/CPUs and FPGAs use published manufacturer performance metrics (e.g. “TMACs/sec” where a MAC is a “large enough-ish” bit width for every problem tackled) for known technology nodes reasonably extrapolated to construction freeze date for the first construction+build installation. For guidance on these metrics refer to appropriate sections of the “SKA1 SYSTEM BASELINE DESIGN” SKA-TEL-SKO-DD-001, Revision 1.

10. For ASICs assume 4X power reduction, 4-8X density, and 2X performance compared to FPGAs¹ for the same processing (and technology node) but assuming bit widths roughly tuned to the task (e.g. if the FPGA’s “processing unit” is 16 bits wide, but really the task requires only 8 or 10 bits then that reduction in bit width is roughly contained within the 16X density/power factor), unless you have sufficient evidence to the contrary. E.g. if an FPGA can do 4 PPFBs, (conservatively) assume that 4X that same functionality (16 PPFBs) can be implemented in an ASIC, with optimized bit widths, at 1/4 the power, and at 2X the speed of the FPGA. For ASICs assume that one technology node behind the node available in production at the technology freeze date (end of 2016) is affordable (unless you can provide evidence to the contrary). We leave it up to you to provide evidence-backed costs for NRE and production, and to change these factors if you have evidence either way.

11. For custom solutions providers (PCB assemblies/modules) assume that the tested, delivered, packaged, turn-key costs, including bare PCB is 1.4X raw parts costs. E.g. if the raw parts costs are $5000, the delivered cost of that complete widget is ~$7000. In cases where there are a few components of high value (such as large FPGAs), this number could be reduced if evidence is provided to support such reduction. We leave it up to you to provide evidence-backed costs for NRE including any surrounding laboratory/infrastructure/tools costs that get charged to the SKA project (as credits).

12. Include reasonable estimates of maintenance + spares, as well as maintenance/operating personnel in your costs. As a baseline, assume minimum 5% (of total installed base) spares of all hardware on hand, 2 FTEs per Subsystem (maintenance + operating personnel and support infrastructure, which works out to ~0.4 FTEs 24/7) and 100,000 Euros/year per FTE².

13. For scale of processing required for each Subsystem (e.g. in TMACs/sec) refer also to appropriate sections in the SKA1 SYSTEM BASELINE DESIGN document unless you provide sufficient evidence to the contrary that they are substantially incorrect.

It is advisable that your costing breakdown/spreadsheet be parameterized as much as possible with these metrics such that when more accurately tuned numbers are available they can be fed in to obtain more accurate cost numbers. If you have better numbers (with supporting evidence/references) for any

² This assumes that site maintenance personnel pay costs are less than development personnel, which may be entirely incorrect!
of the above metrics, kindly email Brent.Carlson@nrc-cnrc.gc.ca so that they may be updated to provide the best numbers possible on an on-going basis (this document will be posted on a NRC/MDA-led CSP consortium Group Page on cyberska.org). If there is a conflict between numbers provided by consortium members and the SKAO, the SKAO numbers will take precedence until and unless they are refined accordingly by the SKAO.

It is not clear if power delivery and heat removal costs are to be contained within the CSP Element or not however, clearly these costs will have to be within the total SKA1 budget in any case and therefore to allow fair comparisons across technologies and techniques include these in your estimates. It also makes our total proposal more sound and competitive if we are not overtly or unknowingly hiding costs. For each cost estimate, provide an estimate of the uncertainty in percentage, such that ultimately there is a calculated min-to-max range in NRE+capital cost, operating cost, and total life-cycle cost\(^3\) for each proposed Subsystem.

Admittedly this is not an exhaustive treatment of all of the metrics and “what-have-yous” for complete and highly accurate downselect process, however, time is of the essence and it will have to form the foundation for preliminary feasibility analysis and downselection.

**Cost, Feasibility, Downselect Judgements, Decision Process, & Timelines**

As there are several distinct major Subsystems within the CSP Element and as there will likely be competing technologies, approaches, and costs to providing these, it would ideally be best to allocate a certain budget to each Subsystem and make the selection/downselect based on that Subsystem-specific budget. Unfortunately, it is not that easy and some sort of juggling of Subsystems and approaches will be used to come up with the best combination, providing the best science for the best cost, and a reasonable distribution among consortium members, at a judged acceptable risk. It may turn out that some science requirements are, for the foreseeable future, infeasible to meet within the cost cap or any reasonable cost cap, and so these would have to be de-scoped in any case. De-scoping of one Subsystem cannot go to the point where it doesn’t exist (e.g. each of the telescopes requires a correlator).

Therefore, there is no absolute cost cap on any particular Subsystem, but there is of course a cost cap on the entire CSP Element to work within. Therefore, one key decision we must make as a consortium is which PIPs we will include in our proposal for each sub-element for execution during Stage 1. This is an important decision because it will affect the solution space we explore, the scope of work we commit to, and therefore the team structure and management effort required.

The initial assessment by NRC/MDA of our consortium funding status is that we are marginal in terms of having sufficient resources to complete the required work. So, we likely can not simply accept all

---

\(^3\) Costs must be broken down like this so that budgeting judgements can be made regarding construction vs operating cost; e.g. some proposal might have the lowest total life cycle cost, but a NRE+construction cost that is infeasible or requires special “handling” to carry forward.

proposals without some cap. Furthermore, we only want to proceed with investigations that have a high likelihood of producing solutions that are affordable, available within the timeframes of SKA1, and will meet SKA requirements. Given that the consortium as a whole will be ultimately responsible for successfully delivering the CSP CDR package, this decision needs to be made by the group.

Therefore, a review panel will be created with one representative from each organization from an SKA-signatory nation which has pledged, by April 19, >= 1 FTE average throughout the pre-construction timeframe. The review panel and decision process is proposed to be as follows:

- Each review panel member will assess all PIP White Papers between Apr 29 and May 2.
- At the review telecon May 2, the review panel can ask questions of PIP proposers to seek clarification. Only voting panel members can ask questions at the telecom; other attendees only respond if questioned.
- At the conclusion of the telecon (limit to say 1.5 hr of questions), each review panel member will email to David Stevens (drs@mdacorporation.com) their assessment (vote) of each sub-element PIP as being either 1) highly likely to lead to a successful solution and therefore recommended for Stage 1 execution or 2) not, therefore reject.
- If multiple PIPs are entered for a particular sub-element, and each is likely to lead to a successful solution, they will be ranked as well by each review panel member (on a scale of 1 to 10; 10 being the highest rated).
- Majority rules and voting is on a per-sub-element PIP basis (i.e. each sub-element PIP is voted on). NRC/MDA gets a single vote but will use an additional vote as a tie breaker, if required.
- Once voting is complete, results will be disseminated to review panel members (only) for discussion on how to proceed. If more than one PIP is chosen for a given sub-element then a follow-up assessment will have to be conducted as to whether there are sufficient resources within the current NRC/MDA management team, and within the STFC-led system engineering team, to successfully execute the work.
- If it is determined that more resources are required, then voting organizations will be required to increase their contribution to cover the management gaps or the number of PIPs will have to be reduced. If there is still no consensus on how to reduce the number of PIPs to stay within management effort, then NRC/MDA and STFC will decide, together, on which PIPs will be carried forward for Stage 1 investigations.

This process will be discussed, further refined, and finalized at the face-to-face meeting in Vancouver, April 15-19. It was, and the voted conclusion was to proceed with this process as defined here.
Cost Constraints and Metrics Feedback from the SKAO

To get a sanity check on key cost constraints and metrics defined in this document, eight questions were “officially” sent to the SKAO on March 22 on behalf of the consortium. The eight questions, and resulting answers (received March 27), are provided below:

Questions:

1. For Section 8.7, we assume that operations costs, for say 10 years post-delivery, should also be a key component of the costing model and influence the proposed solutions?
2. Site operating power costs per Watt per year for each site.
3. Site power infrastructure costs per Watt of capacity, both delivery and cooling.
4. Rough order of magnitude CSP capital budget. Is 60 M Euro reasonable to assume?
5. Annual operating budget as percentage of capital budget. Is 5% reasonable to assume?
6. Spares requirements (% of installed) at delivery.
7. FTE cost per year for construction, installation, maintenance. Is 100,000 Euro reasonable to assume?
8. Is there an overall costing approach or strategy that should be followed? Should we bid in Euros using current currency exchange values from each country?

Answers:

We do not prescribe the metrics that you mention for the purposes of evaluation.

1. The proposed solution costing shall address operations costs.
2. This is an outcome of the design phase. You may parameterize your proposed cost for costs that are not provided.
3. This will be designed as part of the infrastructure. We will conclude on numbers during the first year of the project when the interfaces are defined.
4. This is an outcome of the design phase and proposal.
5. This is also an outcome.
6. It is typical that lifetime spares are provided with the supply. COTS components are not provided. If there are special considerations for the CSP consortium these will be negotiated.
7. See our answer about power costs. You may parameterize in terms of man years.
8. As long as the cost is structured the conversion can be done either way. Please indicate source and nature of supply with current cost estimate.

We trust that these answers are sufficient for you to continue your deliberations.
Interpretations of the Baseline Design

This section contains notes, interpretations, and questions discussed on Thursday April 18, 2013 at the face-to-face meeting at MDA Corp in Richmond B.C. In some cases there are action items for CSP Lead to ask the SKAO for clarification, however, it is unlikely these will be forthcoming within the timeframe required to generate White Papers. Therefore, comments are provided in following sections on what might be reasonable numbers to use where there is uncertainty.

Use these notes as appropriate to help to interpret the Baseline Design to inform your White Paper contents.

General guidelines for correlation in the presence of RFI, and regarding sensitivity loss:

From a sensitivity loss factor perspective, the minimum number of bits, for each of I and Q that shall be correlated is 4 (with a “1.5% sensitivity loss) noting that the sample rate for a complex signal, and therefore resulting number of complex multiplications is a factor of two lower than for a real signal.

If you use a 4-bit to < 8-bit correlation architecture, then include some statements regarding the impact of real-time temporal/spectral RFI excision and flagging (of excised channels) in your White Paper to prevent/deal with channel saturation. i.e. that you will do this in the F-part of the correlator before conversion to 4-bit to < 8-bit correlation, and an indication of how it will be done and flagging will be handled.

If you are planning ≥ 8-bit correlation, then it is sufficient to perform post-correlation RFI excision at the smallest dump time indicated for any baseline for a particular telescope (e.g. 0.6 seconds for SKA1-low) without worrying about dynamic channel saturation (i.e. set some level at the beginning of an observation good for all channels for the duration of the observation).

In both cases, channel averaging within the correlator after RFI excision may be used to achieve the indicated final number of channels to be delivered to the SDP, as recommended by Larry D’Addario in his “Cost Drivers and Tradeoffs” presentation at the face-to-face meeting on April 18, 2013 at MDA in Richmond, B.C. (see following sections for numbers). Note that channel averaging may only be performed post-correlation and not on the real-time signal and that in all cases the number of fine frequency channels prior to post-correlation RFI excision is as stated.

SKA1-Low:

The use of third party instrumentation (e.g. spigots) is vague – what has to be provided (meta-data, e.g.) etc? Assume no impact on CSP cost or configuration.

Is the Optical Distribution part of the cost of the CSP? (e.g. Figure 8 of the Correlator functional architecture.) Assumption: The cost of the Optical Distribution and Receiver modules are within the SADT element and is not to be included in the CSP cost.

The buffer (delay tracking buffer) has to be sized for +/- 50 km – to absorb the worst case wavefront travel time across the array plus differences in fiber delay from the stations to the correlator (at ~1/3 c). Assume the delay buffer must be able to absorb +/- (50/3e5 + 50/1e5) ~ = +/- 700 microseconds of
relative delay (+/-350,000 8-bit samples at 500 MHz) and that this buffer undergoes active +/-0.5 sample delay tracking. Assume there is a separate circular transient buffer with an externally supplied trigger able to capture 1 second worth of data (500 Msamples each polarization), but that it does not undergo active delay tracking, noting that the cost of this should be trivial as it can be relatively easily accommodated in external dynamic RAM without the complication of active delay tracking. The actual length of this transient buffer requires further science input.

Assume that baseline-based integration is required to minimize the data rate into the SDP, with minimum two settings required. For baselines $4 \leq 6 \text{ km}$, must be able to integrate for at least 10.6 seconds and need not dump to the SDP any faster than that. For baselines $> 6 \text{ km}$, must be able dump as fast as every 0.6 seconds, and need not integrate longer than that. Assume $\frac{1}{2}$ the baselines are within each category.

Assume a two way arrow between Local M&C and the Telescope Manager. Typical points that are monitored are environment (i.e. voltage, temperature, power), on-line error statistics (e.g. communications errors, internal errors), and configuration errors/warnings. Typical control is correlator configuration and delay models. These are for information only, and should not have any significant impact on White Paper metrics and cost analyses.

Design for test will add additional requirements to the design – these won’t be dealt with in any detail in the White Papers.

There is no requirement mentioned for aliasing due to leakage, pass-band ripple etc. For now, assume 30 dB maximum aliasing between adjacent channels, and 60 dB reject band attenuation from the center of the adjacent channel(s) on out; you need only use these if they affect your estimates compared to “load factors” in the Baseline Design.

The TMACS/s number in Table 4 in section 6.12 of the Baseline Design is calculated as follows:

$$O\left(\left[N_{\text{antennas}} \cdot \log_2(N_{\text{channels}}) \cdot 2 + 16 \cdot \frac{N_{\text{antennas}}^2}{2}\right] \cdot \frac{\text{sample rate}}{2}\right)$$

The factor of 2 in the first term is roughly that required for the poly-phase FIR in front of the FFT. The factor of 16 is 4 polarization products x 4 multiplies in a complex multiplier. The sample_rate/2 factor is noting that the sample rate for complex signals is one-half that for real signals (but the net data rate in both cases is the same). This equation assumes that additions are negligible cost and power impact compared to multiplies in the case where large multiplies dominate. The baseline-based $N^2/2$ factor dominates this equation.

Assume that 262,144 channels per polarization product, per baseline, are required but only for post-correlation RFI excision.

---

4 Which may include stations outside the core; it is the baseline length that is the key parameter here.

5 Recollection of Larry D’Addario’s statement at the Apr 15-19, 2013 f2f meeting.
Assume that each visibility (each frequency channel, each baseline) delivered to the SDP requires 64 bits at full resolution. Any meta-data that might be provided to the SDP is negligible for White Paper purposes.

Assume the final number of spectral channels per polarization product per baseline delivered to the SDP is 4096⁶ (ref: D’Addario “Cost Drivers” talk on April 18, 2013 at the CSP face-to-face meeting: http://www.cyberska.org/pg/file/read/27403/33-larry-daddario-options-and-tradeoff-cost-drivers-and-tradeoffs-in-pdf-format ).

Assume any switch fabric between the CSP and SDP is not included in White Paper costing.

**SKA1-Mid:**

For the purposes of the proposal, the MeerKat Interface box will not be costed.

The use of third party instrumentation (e.g. spigots) is vague – what has to be provided (meta-data, e.g.) etc? Assume no impact on CSP cost or configuration.

Is the Optical Distribution part of the cost of the CSP? (e.g. Figure 8 of the Correlator functional architecture.) Assumption: The cost of the Optical Distribution and Receiver modules are within the SADT element and is not to be included in the CSP cost.

Assume the SKA1-MID CBF must be for the combined array of 254 antennas, 2 8-bit data streams per antenna (two polarizations), at 5 Gsamples/sec (2.5 GHz) each. Use the equation in the SKA1-Low section to calculate the “correlator load factor” and use it instead of the Baseline Design metrics if in conflict.

The buffer (delay tracking buffer) has to be sized for +/- 100 km – to absorb the worst case wavefront travel time across the array plus differences in fiber delay from the stations to the correlator (at ~1/3⁷ c). Assume the delay buffer must be able to absorb +/- (100/3e5 + 100/1e5) ~ = +/- 1400 μseconds of relative delay (+/-7 million 8-bit samples at 5 GHz sample rate) and that this buffer undergoes active +/- 0.5 sample delay tracking. Assume there is a separate circular transient buffer with an externally supplied trigger able to capture 1 second worth of data (5 Gsamples each polarization), but that it does not undergo active delay tracking, noting that the cost of this should be relatively trivial as it can be easily accommodated in external dynamic RAM without the complication of active delay tracking. The actual length of this transient buffer requires further science input.

Assume that baseline-based integration is required to minimize the data rate into the SDP, with minimum two settings required. For baselines ⁷ ≤ 10 km, must be able to integrate for at least 1.6 seconds and need not dump to the SDP any faster than that. For baselines > 10 km, must be able dump

---

⁶ This number may be affected by the desired image FOV, and varies with baseline. Further work in Stage 1 REN-ADD investigations is required.

⁷ Which may include stations outside the core; it is the baseline length that is the key parameter here.
as fast as every 0.08 seconds to the SDP, and need not integrate longer than that. Assume ½ the baselines are within each category⁸.

Assume a two way arrow between Local M&C and the Telescope Manager. Typical points that are monitored are environment (i.e. voltage, temperature, power), on-line error statistics (e.g. communications errors, internal errors), and configuration errors/warnings. Typical control is correlator configuration and delay models. These are for information only, and should not have any significant impact on White Paper metrics and cost analyses.

Design for test will add additional requirements to the design – these won’t be dealt with in any detail in the White Papers.

There is no requirement mentioned for aliasing due to leakage, pass-band ripple etc. For now, assume 30 dB maximum aliasing between adjacent channels, and 60 dB reject band attenuation from the center of the adjacent channel(s) on out.

Assume that 262,144 channels per polarization product, per baseline, are required prior to post-correlation RFI excision.

Assume that each visibility (each frequency channel, each baseline) delivered to the SDP requires 64 bits at full resolution. Any meta-data that might be provided to the SDP is negligible for White Paper purposes.

Assume the final number of spectral channels per polarization product per baseline delivered to the SDP is 4096 (ref: D’Addario “Cost Drivers” talk on April 18, 2013 at the CSP face-to-face meeting: http://www.cyberska.org/pg/file/read/27403/33-larry-daddario-options-and-tradeoff-cost-drivers-and-tradeoffs-in-pdf-format).

Assume any switch fabric between the CSP and SDP is not included in White Paper costing.

For the central beamformer delivery of frequency channels to the NIP-PSS (pulsar search), assume 20 kHz spectral resolution (not lower, not higher). Assume 2100 beams, formed on the inner 900 m diameter core of 125 antennas, at 300 MHz bandwidth must be formed. Both polarization spigots can be added together⁹ to deliver a net 2100 spigots, 15,000 channels across 300 MHz, each. Note that if “channel averaging” is used here to obtain the required minimum 20 kHz resolution, it must be performed with proper bandwidth synthesis and not simple averaging. Assume the central beamformer delivers “voltages”.

For central beamformer delivery of frequency channels to the NIP-PST (pulsar timing), assume 10 beams per polarization, formed on all 254 antennas, 1.4 GHz bandwidth, resolution of 100 nsec. The minimum channel bandwidth is therefore 10 MHz and may be larger. Polarizations cannot be added together so a

---

⁸ Recollection of Larry D’Addario’s statement at the Apr 15-19, 2013 f2f meeting.
⁹ Assuming any phase and delay differences between polarizations have been compensated for to allow coherent voltage sum rather than incoherent power sum.
total of 20 spigots, 1.4 GHz and 140 spectral channels (max), must be delivered. Assume the central beamformer delivers “voltages”.

Note that each beam produced by the central beamformer for both NIP-PSS and NIP-PST must be on an independent delay center on the sky, within the +/- half-power points of the 15 m antenna primary beam. Thus, beam-dependent delay buffer sizes can be calculated, assuming the main correlator delay center is antenna primary beam boresight. Assume the absolute maximum pk-pk phase error within a channel due to beam-dependent fine phase-delay tracking is +/-45°. This results in maximum ~10% sensitivity loss at the channel edge, but averages out to about ½ that (ref: applicable sections of Thompson, Moran, and Swenson, Interferometry and Synthesis in Radio Astronomy, 1986, Wiley, New York [or more recent edition]).

The central beamformer must handle any re-formatting of beam outputs such that each beam spigot contains all beam-formed spectral channels in frequency and time sequence (in that order) for only that beam.

Pulsar gating flag is a gate to correlate / not correlate to maximize the signal to noise ratio of the correlated pulsar signal. It is simply a mechanism to turn on or off correlation for a particular duty cycle with a particular period, synchronized to a model of a pulsar ephemeris (spin model).

For pulsar binning, integration is synchronized to a model of the pulsar. Assume 1 millisecond minimum bin width (subject to later definition during Stage 1) that bins must be programatically optimally bunched/allocated across the pulsar period, and that it is acceptable to tradeoff number of fine or final frequency channels for spectral resolution in this case. This means that pulsar binning does not change the integration memory requirements of the correlator. Since pulsars are point sources on any SKA1 baselines, assume that integration times and final data rates to the SDP do not need to exceed, or be less than, that required without binning.

**Action Item:** NRC / MDA to seek clarification as to whether simultaneous imaging and pulsar search is required. Can the number of beams be reduced if simultaneous imaging / pulsar search is supported due to higher availability?

For now – the White Papers should assume the superset – simultaneous imaging / pulsar search and then describe what trade-offs (and resulting cost impacts) could be performed if this is not required. Further science definition and cost/power impact analysis required for final decision.

**SKA1-Survey:**

There were no particular notes taken regarding the CSP for this telescope. Refer to the Baseline Design, and previous sections for calculations and indications, noting:

- 500 MHz bandwidth, dual-polarization, 8-bit sampling per each of 36 telescope beams.
- 262,144 channels per polarization per baseline; post-RFI excision channel averaging to 4096 channels per polarization per baseline.
- 50 km maximum baselines.
• Required minimum integration time of 1.4 seconds on baselines ≤ 10 km, and no need to dump faster. Required minimum integration time of 0.3 seconds on baselines > 10 km, and no need to integrate longer.
• There are, effectively, 36 identical correlators.
## Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Changes/Notes</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAFT</td>
<td>March 18, 2013</td>
<td>First external release.</td>
<td>B. Carlson, J. Lim</td>
</tr>
<tr>
<td>V1.0</td>
<td>April 8, 2013</td>
<td>V1.0 release. PIP decision process described. Feedback from SKAO on key metrics noted. No feedback from consortium member on metrics, therefore current metrics stand for now.</td>
<td>B. Carlson, D. Stevens</td>
</tr>
<tr>
<td>V1.1</td>
<td>April 22, 2013</td>
<td>Add section ‘Interpretations of the Baseline Design’, from the Thursday April 18, 2013 ‘Baseline Design Summary’ session in the F2F meeting at MDA Corp in Richmond, B.C. according to Marie Goulding’s notes, email, also using Larry D’Addario’s cost drivers analysis, and some (hopefully) reasonable assumptions where not discussed but requiring clarification. Change technology freeze date to end of 2016, rather than end of 2018.</td>
<td>B. Carlson, M. Goulding</td>
</tr>
</tbody>
</table>
SKA1-LOW Hybrid PIP

Authors: Kristian Zarb-Adami, John Bunton, Brent Carlson, Gianni Comoretto, Larry D’Addario, Andrew Ensor, Andre Gunst, Jonathan Hargreaves, Arpad Szomoru

Abstract

This white paper addresses possible ‘custom-built’ hardware for the implementation of an SKA1-LOW Correlator. The main design philosophy that this PIP aims to achieve is an upgradable and scalable path on a common platform, combining the power of FPGA(s), ASIC(s) and GPU(s) in order to minimise the total cost of ownership in terms of Non-Recurrent Expenditure, Cost and Power. Throughout this work, we will assume the requirements set out in the baseline reference design document and derive a physical implementation that could deliver such requirements. To this effect, we derive four possible technology options based on FPGA(s) (Redback, Uniboard, Power-MX), GPU(s) and ASIC(s). Whilst the numbers themselves may be inaccurate, it looks like that an SKA-1 LOW correlator can be designed, built and operated for a total of 10 years with ≤€10M. Oxford and STFCs primary investigation will be PowerMX with FPGAs, in concert with other sub-element investigations in this regard, towards using a common platform. Astron/JIVE will investigate Uniboard-2, CSIRO will investigate Redback, JPL will investigate ASICs. One or more of these will be carried forward to full PIP documentation for SRR/PDR. It is possible that one or more will be downselected internally during Stage 1.

Executive Summary

Through stage 1 this hardware PIP will investigate the use of FPGA(s), ASIC(s) and GPU(s) as appropriate processing elements on a common hardware platform so that through an internal downselect process only one of the proposed solutions is taken forward to SRR/PDR.
Summary of Estimated Costs

Table 1 summarises the total cost of ownership of the implementations proposed in this White Paper. Each column represents a different platform, but it should be noted that various assumptions have gone into each costing, so whilst this table is meant to give an indication of the total cost of ownership, the reader should bear in mind that there is considerable uncertainty in the final numbers.

Assumptions made:

- NRE cost of ASIC(s) divided equally among the three CBF sub-elements
- The sample quantization at the correlator input is 4b+4b minimum
- NRE cost for all PCBs (Uniboard, Redback, Power-MX, and ASIC-based) is divided equally among the three CBF sub-elements
- Uniboard and Redback NRE and Firmware costed, but in truth this is already available through other funding streams
- The various FPGA-based options utilize the built-in multipliers in different ways. If they were all fully optimized, the cost results might be closer to each other
- Redback correlator is 7+7b rather than 4+4b
- FPGA(s) costed at $800 (as per volume quotation from Xilinx Virtex-7)
- FPGA technology available in 2016 is assumed.
- We will have FPGAs that interface to DDR4 at 2066MHz
- The NVIDIA VOLTA is assumed for the GPU, capable of 5TFMAC/s and 1TB/s, assuming 250W per GPU card at a cost €1500 each
- For sizing of the FPGA and GPU options, the total computational load was assumed to be 1700 TMAC/s. (The Baseline Design document gives 1660 TMAC/s for correlation, and we have added 40 TMAC/s to cover channelization.)
All cost numbers in Euros of 2013. All other numbers in the given units.

<table>
<thead>
<tr>
<th>Total Cost of Ownership</th>
<th>Uniboard</th>
<th>Redback</th>
<th>GPU(s)</th>
<th>Power MX</th>
<th>ASICs</th>
<th>FPGA&amp;ASIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NRE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main PCB (correlator)</td>
<td>33,000</td>
<td>33,000</td>
<td>-</td>
<td>33,000</td>
<td>38,402</td>
<td>38,402</td>
</tr>
<tr>
<td>Mezzanine PCB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control PCB</td>
<td></td>
<td>20,000</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter bank PCB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firmware</td>
<td>300,000</td>
<td>300,000</td>
<td>200,000</td>
<td>300,000</td>
<td>150,000</td>
<td>150,000</td>
</tr>
<tr>
<td>Correlator ASIC (incl prototypes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter bank ASIC (incl prototypes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototypes (10 Units)</td>
<td>109,000</td>
<td>83,000</td>
<td>50,000</td>
<td>210,000</td>
<td>512,821</td>
<td>(incl above)</td>
</tr>
<tr>
<td><strong>Total NRE</strong></td>
<td>442,000</td>
<td>435,000</td>
<td>250,000</td>
<td>580,000</td>
<td>996,154</td>
<td>444,872</td>
</tr>
</tbody>
</table>

**PRODUCTION**

<table>
<thead>
<tr>
<th>Processing board (filter bank)</th>
<th>(incl below)</th>
<th>(incl below)</th>
<th>(incl below)</th>
<th>(incl below)</th>
<th>2,308</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASICs per board</td>
<td>877</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost per filter bank board</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing board (corr. or combined)</td>
<td>4,000</td>
<td>3,000</td>
<td>4,000</td>
<td>2,308</td>
<td>2,308</td>
</tr>
<tr>
<td>Mezzanine board</td>
<td></td>
<td>1,300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control board (with FPGA)</td>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPGA or ASIC, cost of each</td>
<td></td>
<td>800</td>
<td>800</td>
<td>877</td>
<td>877</td>
</tr>
<tr>
<td>FRGAs or ASICs per board, each</td>
<td></td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>FPGA boards per unit</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Infrastructure per unit</td>
<td>500</td>
<td>500</td>
<td>(VOLTA)</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Cost per processing unit</td>
<td>18,500</td>
<td>8,300</td>
<td>1,300</td>
<td>21,000</td>
<td>16,838</td>
</tr>
</tbody>
</table>

**PROC. UNITS (filter bank)**

<table>
<thead>
<tr>
<th>(incl below)</th>
<th>(incl below)</th>
<th>(incl below)</th>
<th>(incl below)</th>
<th>57</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spares 10%</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Total Units</td>
<td></td>
<td></td>
<td></td>
<td>63</td>
<td>18</td>
</tr>
<tr>
<td>PROC. UNITS (corr. or combined)</td>
<td></td>
<td></td>
<td></td>
<td>57</td>
<td>12</td>
</tr>
<tr>
<td>Spares 10%</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Total Units</td>
<td></td>
<td></td>
<td></td>
<td>77</td>
<td>27</td>
</tr>
<tr>
<td>GPU hosts</td>
<td></td>
<td></td>
<td></td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>Spares 10%</td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total hosts</td>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Cost per host</td>
<td></td>
<td></td>
<td></td>
<td>700</td>
<td></td>
</tr>
<tr>
<td>Misc costs for ASIC-based system</td>
<td></td>
<td></td>
<td></td>
<td>269,231</td>
<td>269,231</td>
</tr>
</tbody>
</table>

**Processing unit hardware cost**

<table>
<thead>
<tr>
<th>Processing unit hardware cost</th>
<th>839,200</th>
<th>474,760</th>
<th>801,488</th>
<th>2,079,000</th>
<th>1,172,666</th>
<th>1,085,262</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infrastructure (racks/cables)</td>
<td>30,000</td>
<td>30,000</td>
<td>100,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
</tr>
<tr>
<td><strong>Total hardware cost</strong></td>
<td>869,200</td>
<td>504,760</td>
<td>901,488</td>
<td>2,109,000</td>
<td>1,172,666</td>
<td>1,115,262</td>
</tr>
</tbody>
</table>

**POWER**

<table>
<thead>
<tr>
<th>Max power per FPGA or GPU, W</th>
<th>37</th>
<th>37</th>
<th>250</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute load</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Compute per FPGA, W</td>
<td>25</td>
<td>29</td>
<td>200</td>
<td>29</td>
</tr>
<tr>
<td>V/I etc per FPGA, W</td>
<td>12</td>
<td>12</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>Power DRAM, Optic, Fans Etc, W</td>
<td>14</td>
<td>14</td>
<td>93</td>
<td>14</td>
</tr>
<tr>
<td>Power per unit, W</td>
<td>439</td>
<td>129</td>
<td>1,000</td>
<td>220</td>
</tr>
<tr>
<td>Power per ASIC filter bank board, W</td>
<td></td>
<td></td>
<td>18,176</td>
<td></td>
</tr>
<tr>
<td>Power per ASIC correlator board, W</td>
<td></td>
<td></td>
<td>39,375</td>
<td>39,375</td>
</tr>
<tr>
<td>Power for misc. ASIC system hv, W</td>
<td></td>
<td></td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>Line efficiency and cooling factor</td>
<td>1.54639</td>
<td>1.54639</td>
<td>1.54639</td>
<td>1.54639</td>
</tr>
<tr>
<td><strong>Total Power, W</strong></td>
<td>47,525</td>
<td>26,478</td>
<td>168,170</td>
<td>30,552</td>
</tr>
<tr>
<td>Power cost/W/Year (Euro)</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Maintenance labor, 10 years</td>
<td>2,000,000</td>
<td>2,000,000</td>
<td>2,000,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>Replacement parts, 5%/year, 10 years</td>
<td>484,650</td>
<td>252,180</td>
<td>450,744</td>
<td>1,054,500</td>
</tr>
<tr>
<td>Power cost over 10 Years</td>
<td>1,188,122</td>
<td>661,954</td>
<td>4,304,348</td>
<td>763,793</td>
</tr>
<tr>
<td>Power infrastructure</td>
<td>118,812</td>
<td>66,155</td>
<td>420,423</td>
<td>76,379</td>
</tr>
<tr>
<td>Operating power 10 Years</td>
<td>2,241,283</td>
<td>2,000,129</td>
<td>7,075,416</td>
<td>3,954,672</td>
</tr>
<tr>
<td><strong>Total Cost NRE+Cap+Op</strong></td>
<td>5,052,883</td>
<td>3,921,289</td>
<td>8,226,904</td>
<td>6,586,672</td>
</tr>
</tbody>
</table>

Notes:

[1] "ASICs" column uses two ASICs, one for filter banks and one for correlation, each on a custom PCB.
[2] "FPGA&ASIC" column uses FRGAs on Power-MX for filter banks and ASICs on custom PCBs for correlation.
Introduction to SKA1-LOW

SKA1-LOW will primarily address observations of the highly red-shifted 21 cm hyperfine line of neutral hydrogen from the Epoch of Reionization and earlier. It will also be well suited for conducting low radio frequency observations of pulsars, magnetized plasmas both in the Galaxy and intergalactic space, radio recombination lines, and potentially extrasolar planets.

The telescope receptors will consist of an array of ≈250,000 log-periodic dual-polarised antenna elements. Most of the elements will be arranged in a very compact configuration (the ‘core’) with a diameter of ~1 km, the rest of the elements will be arranged in stations about 35-m in diameter. The stations will be configured in three equally spaced spiral arms. The maximum radius of the configuration is ≈45 km. The antenna array will operate from 50MHz to ≈350MHz. Its sensitivity will be ≈1000 m²/K at frequencies above 110MHz at the zenith, assuming an instantaneous bandwidth of 250MHz. The core will have a brightness temperature sensitivity of ≈1mK over the same frequency range at the zenith.

The elements will be grouped into 911 35-m diameter stations, whose elements will be beam-formed to expose a field-of-view of ≈20 deg² in a single smooth beam. Possibilities exist for more elaborate beamformers in the core, if needed. Signals from the beamformers will be transported to a central signal processing building, where they will be channelized and cross-correlated with each other. Output data from the correlator will be transported to the science data processing centre in Perth. Additional signal processing equipment may be warranted but is not included at present. The required processing of the science data will be varied, and probably elaborate, and will likely include calibration, image-cube (i.e., spatial plus spectral) formation on various scales, and statistical analysis. The entire array can be built within the boundaries of Boolardy station, and operated concurrently with the SKA1-survey telescope on the same property.
Context

The context diagram for central signal processing associated with the SKA1 Low is show in Figure 1. The main inputs are the data for single beam from the 911 low stations and the main output are the visibilities generated by the correlator which are sent to Science Data Processing located in Perth. In normal operation the system also receives a time reference signal in addition to the timing data that is embedded in the PAF beam data. Optional features that will be included in the design are a single external transient trigger and the output of an array beam. No consideration is given to spigots for 3rd party equipment as function and form of the requirements for this equipment is currently undefined. It is assumed that the system is housed in racks that provide power and cooling.

The main input to the CSP processor is the single beam form 911 Low stations. Each beam is a dual polarisation beam with a bandwidth of 250 MHz. The data is 8 bit resolution at 500MS/s. This data must be processed by a second stage filterbank in CSP to bring the data to the final frequency resolution of ≈1 kHz. The input into CSP low from a single station is 500MS/s by 8 bits by two polarisations is 8Gb/s. Over all 911 stations this is 7.3 Tb/s. Data for a single antenna can be transported on a single 10Gb/s link or data for four stations on a single 40Gb/s link. As the beamformers for SKA1_low are likely to be aggregated in bunker, it is assumed here that 40Gb/s links will be used

- Input to CSP low is in the form of 40 Gb/s links each carrying data for 250 MHz of bandwidth for four stations.
The basic functional architecture for CSP Low is shown in Figure 2. Full Stokes correlation (I,Q,U,V) is provided by the SKA1-low correlator in the form of a set of integrated product-pairs of like signals from the array stations (u-v data-points or visibilities) for each frequency channel. Note that delay correction and some channelization is required before correlation can be performed on the data. Final channelization can also be carried after correlation in some designs. Figure 2 shows an FX implementation of the functional architecture which is commonly regarded as the most efficient. This readily maps to differing technology including ASIC, FPGA and Software implementations.

The processing load for a large correlator is dominated by the product, $N^2B$, where $N$ is the number of antennas inputs being correlated and $B$ is the total bandwidth. The output data rate from the correlator is a critical factor in the loading of downstream science data processing and in transmitting data to the science processing centre. This rate depends upon the integration time (dump time), which in turn is proportional to the radius of the processed field-of-view and the length of the particular baseline in wavelengths. There is typically some tolerance of coherence loss (smearing) at the edge of the field. For the purposes of design it is taken as 2% in this document, but it can be adjusted at observation time. Since the baseline length is variable, there are opportunities to integrate longer for many of the baselines in the array, thus reducing this rate. The correlator will be capable of generating more data than can be handled downstream. The maximum rate depends on the science program, particularly whether full-field imaging is required at the longest baselines at the lowest frequencies.

Hence the CSP-processor must receive the signal from a SKA-LOW station, each signal consisting of two-polarisations having 250MHz of bandwidth and sampled at 8-bit per sample, therefore requiring approximately 10Gbit/s per station, and following this

- apply geometric delay
- perform first stage polyphase filtering
– apply fringe rotation and fractional delay correction
– perform second stage polyphase filtering
– compute ACF spectrum
– detect and blank RFI on single channels
– requantize the output samples to 4+4 bit format
– perform a Corner Turner to transpose the data
– correlate every channel/antenna pair
– output visibilities at a minimum dump-time of 0.6s and a maximum dump-time of 10.6s

**Key Design Parameters**

The Baseline reference design provides the specifications for SKA1-LOW and are represented in Table 2 below.
<table>
<thead>
<tr>
<th>Aperture Array</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Frequency</td>
<td>50 MHz</td>
</tr>
<tr>
<td>Upper Frequency</td>
<td>300 MHz</td>
</tr>
<tr>
<td>Number of antennas per station</td>
<td>289</td>
</tr>
<tr>
<td>Total physical aperture</td>
<td>$8.0 \times 10^7$ m²</td>
</tr>
<tr>
<td>Area per antenna</td>
<td>2.15 m²</td>
</tr>
</tbody>
</table>

| Element filling factor in station | 0.7 | Areal filling factor |
| Dense/Sparse Transition | 111 MHz | $A_p$ per element is equal to packing density |

<table>
<thead>
<tr>
<th>Array Configuration</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Station Diameter</td>
<td>35 m</td>
</tr>
<tr>
<td>Number of stations</td>
<td>911 stations</td>
</tr>
<tr>
<td>Core (radius &lt;600 m)</td>
<td>~50% (~433 st/ns)</td>
</tr>
<tr>
<td>Core (radius &lt;1000 m)</td>
<td>~75% (650 st/ns)</td>
</tr>
<tr>
<td>Spiral Arms</td>
<td>~4% (45 stations)</td>
</tr>
<tr>
<td>Av’g St’n filling factor (radius &lt;220 m)</td>
<td>0.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station Beam Forming</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of beams</td>
<td>1</td>
</tr>
<tr>
<td>Instantaneous bandwidth per beam</td>
<td>250 MHz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Digital Outputs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample streams</td>
<td>2</td>
</tr>
<tr>
<td>bits per sample</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signal Transport System</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate per station</td>
<td>10 Gb/s *</td>
</tr>
<tr>
<td>Radius &lt; 3000 m</td>
<td>8.7 Tb/s</td>
</tr>
<tr>
<td>3 km &lt; Radius &lt; 50 km</td>
<td>450 Gb/s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signal Processing System</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine Frequency channels**</td>
<td>$2.5 \times 10^7$</td>
</tr>
<tr>
<td>Complex Correlations</td>
<td>$4.1 \times 10^{11}$</td>
</tr>
<tr>
<td>Complex Correlations: Spiral Arms</td>
<td>$0.4 \times 10^{11}$</td>
</tr>
<tr>
<td>Core (radius&lt;3 km) Dump Time</td>
<td>~10.6 s</td>
</tr>
<tr>
<td>Minimum Dump Time</td>
<td>~0.6 s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Science Computing System</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data rate (1 kHz channels)</td>
<td>$8.4 \times 10^7$</td>
</tr>
<tr>
<td>Input data rate (100 kHz channels)</td>
<td>$8.4 \times 10^{10}$</td>
</tr>
</tbody>
</table>

Table 2: Baseline Reference Specifications
Derived Specifications for the SKA1-LOW CSP

Table 4 in the Baseline reference design document specifies the following design specifications for the Central Signal Processing requirements for SKA1-LOW. This is reproduced in table 3 below.

Table 3: Derived specifications for SKA-1 LOW

<table>
<thead>
<tr>
<th>Correlator SKA1-Low Log Periodic Option</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine Frequency channels***</td>
<td>2.5 x 10^4</td>
</tr>
<tr>
<td>Actual channelisation</td>
<td>262,144</td>
</tr>
<tr>
<td>Channeliser stop band rejection</td>
<td>&gt;60dB</td>
</tr>
<tr>
<td>Complex Correlations</td>
<td>4.1 x 10^11</td>
</tr>
<tr>
<td>Complex Correlations: Spiral Arms</td>
<td>0.4 x 10^11</td>
</tr>
<tr>
<td>Corr. Load Factor for 250 MHz bandwidth</td>
<td>1660 TMACS</td>
</tr>
<tr>
<td>Core (radius&lt;3 km) Dump Time****</td>
<td>~10.6 s</td>
</tr>
<tr>
<td>Minimum Dump Time</td>
<td>~0.6 s</td>
</tr>
<tr>
<td>Channel Bandwidth = 1kHz</td>
<td></td>
</tr>
<tr>
<td>As power of 2 for FFT: 2^{18}</td>
<td></td>
</tr>
<tr>
<td>&gt; Telescope spectral dynamic range</td>
<td></td>
</tr>
<tr>
<td>911^{1/2} baselines x (1) bms x 4 pol’ns prod’x 2.5 x 10^5 chans</td>
<td></td>
</tr>
<tr>
<td>(911^{1/2}-866) / 2 baselines</td>
<td></td>
</tr>
<tr>
<td>Station diameter = 34 m; max baseline = 6 km</td>
<td></td>
</tr>
<tr>
<td>Station diameter = 34 m; max baseline = 100 km</td>
<td></td>
</tr>
</tbody>
</table>

For the channelization requirements we derive a total of:

911 Stations, 2 Polarizations, 2^{18}-FFT (could be split as 2 stages of 2^9) according to the formula:

\[ C_C = B \times N_{TOT}N_P (\log_2 (N_C) + N_{taps}) \]

Where \( C_C \) is the number of complex MAC(s), B is the bandwidth, \( N_{TOT} \) is the total number of antennas, \( N_P \) is the number of polarisations, \( N_C \) is the number of channels and \( N_{taps} \) is the number of taps in the polyphase filter.

For SKA1-LOW the number we derive from this formula is: **23 TMAC(s)**.

Adding the extra processing due to the FIR-filtering, fringe rotation and auto-correlation we obtain a total of **36-48 TMAC(s)**, depending on oversampling in the polyphase filter [from G. Comoretto]

The memory required in the channelizer is linear with the length of the FFT and can be derived from the following formula:

\[ C_{C-coarse} = N_{TOT}N_P N_{C-coarse} N_{taps} n_c \]

Where \( C_{C-coarse} \) is the total number of bytes, and \( n_c \) is the number of bytes required to store a complex-valued term in the transform. Assuming 2-bytes for each term we obtain 4Gbytes in RAM. This mean that external DRAM will be required in order to achieve this. Memory for delay compensation is a fraction of this value. For +/-50 km baseline and 500 MHz sampling rate, total memory requirement is less than 0.5 Gbytes. Both functions can be integrated in the same physical memory.

Correlation of 911 antennas, 2 polarizations require a total of 1660 k complex MAC functions. At 250 MHz total bandwidth, this correspond to 415 TMAC(s), or 1660 TMAC(s). In actual implementations, multipliers will be grouped in large arrays, leading to some inefficiencies. Correlation is performed with limited precision, and it is possible to implement a complex MAC using much less resources than for a full precision multiplication. A conservative estimate could be to have 1 4x4 bit CMAC using 1 full precision MAC.
THERE IS SOME CONFUSION HERE BECAUSE THE WHITE_PAPER_TOCS_v1.1 STATES THAT CHANNELISATION IS INCLUDED IN THE 1660TMAC/s WHEREAS WE DON'T THINK THAT THIS IS THE CASE

Table 4 shows the requirements of an SKA1-LOW PIP.

Table 4: Top-level requirements for an SKA1-LOW PIP

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Processing</td>
<td>1660+40 TMAC(s)</td>
<td>Correlation(4x4 bit) + Channelisation(8-18 bit)</td>
</tr>
<tr>
<td>Total I/O Bandwidth</td>
<td>911 * 2 * 500MSA/s * 8 = 8 Tbit/s</td>
<td>Assuming 8-bit sampling</td>
</tr>
<tr>
<td>Total Memory requirements</td>
<td>218 length FFT = 4GB</td>
<td>Assuming 2-bytes/complex term</td>
</tr>
</tbody>
</table>

Subsystem description

If we assume that the architecture of the SKA-1 LOW will be an F-X architecture then a possible generic implementation of such an architecture is described below.

F-engine

As shown in figure 3, data received from the station beamformer is first compensated for delay variations across a correlator dump time (up to 10 s). A range of ≈150 samples (+/- 75) is adequate to track the delay for the maximum baseline and 10 seconds. The major part of the delay is removed in DRAM, between the two FFTs, and is kept constant during the correlator integration.

The channelization is performed in two stages. A first interleaved FFT produces a set of spectral channels with a channel spacing of 0.5 MHz, and a data rate increased by an oversampling factor. Fractional sample delay and fringe rotation are performed in the frequency domain on the resulting spectra. Only frequency channels inside the array’s nominal band are processed. Channels affected by sampler aliasing or Nyquist analogue filter are discarded.

Filters overlap by the oversampling factor, and the polyphase filter is designed in order to provide adequate insulation (>60 dB) and flatness across the bandpass region. The required number of taps increases with decreasing overlap.

External memory bandwidth is a major design bottleneck. Memory size is not usually a significant cost factor, so it is advantageous to use the same memory for several functions in order to minimize memory transactions. In the design shown here the external memory stores a complete section of data for all input stations, all
frequencies and a complete dump time of 0.6s. In this way the memory is used for frequency-time corner turning, geometric delay compensation and a first stage of frequency-station corner turning, using only one write and one read operation on the data.

After the second stage of polyphase filtering, a final stage computes the total power (ACF) spectrum, requantizes the signal to 4+4 bit complex samples, and flags channels affected by RFI. A further stage of corner turn is performed in groups of 64 signals (32 stations). Final streams consist in blocks of 0.6s, 64 signals, and 16 fine frequency channels.

X-engine

Data from the ≈1000 antennas are processed, in groups of 16, by 64 station processor units (FPGAs, ASIC(s) or GPUs), that perform delay/fringe rate correction, channelization to the final 1KHz resolution, buffering over the 0.6 s minimum dump time, and a first phase of the corner turn operation. Two adjacent station processors merge together the signals from 32 stations, sending them on 32 separate links to a second array of corner turners, that reorder the signals in blocks of one dump time, all antennas-polarizations, and one frequency channel. These blocks are sent to the correlation processors along individual links as is shown in figure 4.

Each correlation processor computes all the correlation coefficients for one frequency channel at a time, by reusing a matrix of 64x64 complex MAC multiple times. Different frequency channels are processed sequentially, each one for the whole elementary integration time. In this example a total of 512 correlation processors analyse a total of 500 frequency channels (1/500 of the total bandwidth) each. The integrated correlation coefficients are accumulated in a long term accumulator unit.

Figure 4: A Possible Correlation architecture
Physical Implementation Proposals

UniBoard2

UniBoard, an EC-funded project that ended in 2012, produced a generic FPGA-based digital processing platform for radio astronomy in the form of a board with 8 Altera Stratix IV devices. UniBoard², an international collaboration again funded by the EC (total budget 1.4 Meuro), aims to create a similar board using the latest available technology, with a strong focus on compactness, flexibility and power efficiency. The budget covers the design and production of the board, as well as the development of a number of demanding firmware applications, such as a correlator, beam former and digital receiver. This project has started July 2012 and will finish end of 2015. At this time the intention is to create a prototype board using Altera 20nm Arria-10, and a first revision using the 14nm Stratix-10 devices. As details about these chips are not yet publicly available, calculations in this document are based on the use of Xilinx Virtex-7 chips, with a total of 2.8Tbit/s and 3.5TMAC/s. Naively this would suggest that we require ≈480FPGA(s) for the processing power (@ 0.7 programming efficiency).

Whilst Virtex-7 are used here to compare the same FPGA-technology with that used in the Power-MX, it is expected that the prototype of the UniBoard will use the Altera Arria and the first revision will use the Altera Stratix-10.

Figure 5: UniBoard² features as currently considered
8 UniBoards are required in order to get the signal from the antennas onto the first-stage processing. The total amount of processing required for the first stage channelization, delay filtering and second stage channelization is approximately 40TMAC/s, which comfortably fits within the 8-uniboards which have a total of 224TMAC/s of processing power. The front-end FPGA(s) are used for the first-stage of channelization and the back-end FPGA(s) are used for second stage channelization and few correlations. In fact each front and back-end FPGA in the channelization are only performing 2.88TMAC/s.

At the output of the first stage, the output data-rate is still equal to the input data-rate and this has now got to be fanned out to the X-part of the correlator. One way of doing this is shown in Figure 7.

The correlator part can be made of different types of technologies, including UniBoards. The total amount of processing required for the cross-correlation is approximately 1600TMAC/s which requires 60-Uniboards for processing.

The total number of UniBoards required for this would be ≈70 UniBoards giving a total processing power of 1960TMAC/s, which could accommodate the total 1660TMAC/s of processing required (this is assuming 70% programming efficiency of the FPGA(s).) This means that each FPGA processes 527kHz of bandwidth.

Each UniBoard consumes ≈440W (350W+ 150W overhead) which means that total power of the system = 47kW.

Both memory and bandwidth issues are resolved, due to fast DRAM available.
This fan-out can either occur on a switch or through a dedicated back-plane since the paths are deterministic. This could look something like figure 7. Note in this case, it is also conceivable the X-engine could be made up of different technologies such as ASIC(s) or GPU(s).

Figure 7: A possible fan-out of 8-UniBoards to 40 of them.
Development schedule

1. Design of Uniboard\textsuperscript{2} Starts \hspace{1cm} 2013
2. Uniboard\textsuperscript{2} prototype \hspace{0.5cm} (using Arria) \hspace{0.5cm} 2014
3. Uniboard\textsuperscript{2} first revision \hspace{0.5cm} (using Stratix-10) \hspace{0.5cm} 2015
4. Uniboard\textsuperscript{2} Production \hspace{0.5cm} \hspace{0.5cm} 2016

Performance analysis

This design conforms to the required specifications of SKA-1 Low.

Cost and Power

This is summarised in Table 1 at the beginning of the document.

Redback – John Bunton

Proposed Processing Module

The proposed processing module is based on the Redback-3 module used in the ASKAP beamformer. It is built on a 1U chassis with ATX power supply. At the rear are a set of fans to cool the hardware and all data, clock and timing connections are at the front of the module. Also at the front is the command and control board which also implements shelf control and data downloads. Behind the front panel is a fibre management module. The front panel is perforated at greater than 30\% to allow air flow from the rear to the front. The Redback board is mounted in the remaining space. Each FPGA has 4 DDR4 and two pairs of 12 multimode fibre optical RX and TX. For connection to the data from stations or antennas each FPGA has in addition connections to four QSFP+ cages. A possible layout of the module with all components in proportion is shown in Figure 8.

The QSFP+, DRAM and multimode fibre optical transmitter are all socketed and can be populated as needed either to match the performance of lower computation capacity FPGAs or systems with a lower I/O intensity. The I/O intensity is inversely proportional to the number of antennas. The I/O intensity of this board is designed to meet the needs of SKA1_Survey. SKA1_Mid needs about one quarter of the I/O on this board and SKA1_Low about one tenth.
The command and control has been separated out as a separate board. This board has an FPGA and multimode fibre transmitters and receivers for timing, clocks, command and control and data downloads. There are three SFP+ card cages for 1 or 10G Ethernet and two optical receivers for clock and timing. The SFP+ slots are populated as needed. In the simplest configuration it has a single 1GE link to provide command and control. More links can be added for higher capacity data product upload or downloads. The eight fibres for these interfaces are amalgamated onto a single 12 fibre ribbon. The only electrical connection is a JTAG input and output at the front of the board. This connection is used for testing purposes. All JTAG capable devices are accessible from these connectors on the front panel.

An additional function of command and control board is shelf control. It monitors temperatures in the 1U chassis and uses this to control the fan speeds. If the temperature increases to dangerous levels for the Redback board it deprograms the FPGA and if this is insufficient it commands the ATX power supply to power down the processor board. While Redback processor is powered down the command and control board can still monitor its temperature and report this to local monitor and control.

**Module performance**

It is estimated that a design can be implemented in 14nm devices. The estimated performance of a midrange FPGA is at this process step is 3.5TMAC/s at a cost of $800. Power dissipation is 10.5 W/TMAC/s and this includes serial I/O and DDR4 I/O. This gives the six FPGAs on a board a compute capacity of 21 TMAC/s or 5.25 TCMAC/s. At 80% usage this becomes 4.2 TCMAC/s.

I/O per FPGA will also increase from just over 30 to ≈100. For this proposal each FPGA has the following connections.
Table 5 FPGA SERDES connections

<table>
<thead>
<tr>
<th></th>
<th>No of SERDES per link</th>
<th>No of FPGAs connected to</th>
<th>Links per FPGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>For paired FPGA</td>
<td>20</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>For Triples</td>
<td>14</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>For Other FPGAs</td>
<td>7</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>For 40Gb/s link</td>
<td>16</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>For multimode Fibre</td>
<td>24</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>Total Links</td>
<td></td>
<td></td>
<td>102</td>
</tr>
</tbody>
</table>

If this exceeds the available SERDES then the fall back position is to provide only 7 links between every pair of FPGAs. This reduced the count to 75. Each link can run at up to 14Gb/s and possibly 28Gb/s but here 10Gb/s is assumed. This has already been proven on the Redback-3 system.

With these connection the FPGAs can implement a full cross connect simultaneously for the 40Gb/s input data and the data from the 12 multimode fibre receiver. The 40Gb/s link is WDM with four optical frequencies. Each frequency and each multimode fibre have been demonstrated to support 10Gb/s each. In later discussion this is assumed but operation at 14Gb/s and higher is possible this has not been demonstrated by CSIRO at this time.

Each FPGA has 4 DRAM. This will be DDR4 at the time of the hardware build. This is expected to operate at 2.066GHz giving a transfer rate of up to 132Gb/s per DRAM. If the DRAM achieves 80% usage then the total I/O is 2.54 Tb/s for the module.

The estimated performance of the Redback processing module is given in Table 6.

Table 6 Performance parameters for the Redback processing module.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute capacity (18-bit)</td>
<td>5.25TCMAC/s (4.2TCMAC/s at 80% usage)</td>
</tr>
<tr>
<td>DRAM (DDR4) I/O</td>
<td>3.173 Tb/s (2.54 Tb/s at 80%) unidirectional, halve for R/W</td>
</tr>
<tr>
<td>DRAM memory</td>
<td>24 DDR4 (196 Byte with 8GB DIMMs)</td>
</tr>
<tr>
<td>Multimode Fibre bidirectional I/O on Redback board</td>
<td>1440 Gb/s at 10Gb/s per fibre</td>
</tr>
<tr>
<td>Single mode Fibre bidirectional I/O on Redback board</td>
<td>960Gb/s with QSFP+ at 40 Gb/s</td>
</tr>
<tr>
<td>Data download/upload</td>
<td>Up to 3 links 1 or 10Gb/s each</td>
</tr>
</tbody>
</table>

**Correlation cell**

The above performance assumes 18-bit multiplication but the correlator requires fewer bits, nominally 8. This allows a number of approaches to improving the performance. For Altera FPGAs each pair of 18 bit multipliers can implement three 8-bit multiplies. This improves the performance of the Redback processing module to 7.88 TCMAC/s (6.3TCMAC/s at 80% usage).
The performance is doubled if 7-bit correlations are sufficient. A 7-bit correlator is 99.97% efficient when run with a signal with an rms of 17.5 quantisation levels. It has good linearity, within 0.1%, for rms signal levels of 2-28. For efficiency greater than 99.9% the level is limited to the range 9 to 22, a 7.8 dB range. Clipping only occurs on 3 sigma or greater excursions making them rare events.

The method used is to change to a sign magnitude representation and implement two 6-bit unsigned multiplies in the single 18-bit multiplier. The bit placement is illustrated below in figure 9.

![Figure 9 Bit ordering for two 6-bit unsigned multiplies. The ten bit product a.d in bits 11-22 and b.d in bits 0-11](image)

It is seen that the two 12 bit unsigned results are in the bottom 23 bits of the multiplier output and that the top bit of the bottom result b.d is added to the top result. This mixing can be unscrambled by noting that the bottom bit of a.d is easily calculated. If bit 11 is equal to this then the top bit of b.d was a zero completing the calculation of b.d and a.d is correct. If bit 11 is not equal to the expected result for bit 0 of a.d then the top bit of b.d is 1 and 1 needs to be subtracted from the result for a.d in bits 11 to 22 to get the correct value of a.d. This correction can be applied in the following adder.

This method allows two 18-bit multipliers form all four unsigned multiplies. After this the products are summed in an adder/subtractor with and adder used if the signs are the same and a subtracter if they are different. This gives a twos complement result whose sign may be incorrect. This corrected with a second adder/subtractor when the result is accumulated. Compared to a two complement implementation the only change is that adders have become adder/subtracters and there is a small amount of logic based on the sign bits of the input to determine the operation of the adder/subtracters.

For a Xilinx implementation a 25x18 bit multiplier is available and the same technique allows full 8 bit complex multiplies to be implemented in two 25x18 bit multipliers. This technique doubles the processing performance for correlation to 10.5 TCMAC/s (7 to 8 bit) for the Redback processing module or 8.4TCMAC/s at 80% usage.
Using a similar technique to the one above a four bit correlation can be implemented using single 18-bit multiplier. This has been implemented on the SKAMP correlator [Reference]. The output of the multiply has all the parts of the final result but further processing is required to get the final result. This results in a multiplier that is logic limited not multiplier limited. SKAMP managed to implement 128 4-bit complex multipliers in an FPGA that had 192 18-bit multipliers. This suggests that the technique can increase the performance of a correlator by a factor of 2.7. This is a 35% improvement over the previous scheme. But it does not improve the antenna based processing. Adding this processing leaves an overall improvement of ≈25%.

Problems with a 4-bit correlator are correlator efficiency and correlator gain variation with signal level. The correlator efficiency is 98.8% this is the equivalent of losing 11 antenna stations. This is sufficient to cancel out the cost improvement possible with a 4-bit FPGA correlator for SKA_Low. Hence in the following a 7 bit correlator is assumed with the following performance (table 7).

<table>
<thead>
<tr>
<th>Table 7 Performance parameters for the Redback processing module.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compute capacity (18-bit)</strong></td>
</tr>
<tr>
<td><strong>Compute capacity (7-8 bit)</strong></td>
</tr>
<tr>
<td><strong>DRAM (DDR4) I/O</strong></td>
</tr>
<tr>
<td><strong>DRAM memory</strong></td>
</tr>
<tr>
<td><strong>Multimode Fibre bidirectional I/O on Redback board</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Single mode Fibre bidirectional I/O on Redback board</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Data download/upload</strong></td>
</tr>
</tbody>
</table>

**Redback-4 implementation**

The total processing load is 10 TMAC/s for the channeliser and 415 TCMAC/s at 7 bits for the correlator. The correlator can implement two 7 bit correlations in the resources for an 18 bit CMAC this gives and equivalent compute load of 217.5 TMAC/s, 18 bit. Each Redback-4 module can process 4.2 TCMAC/s at 80% usage. Dividing this into the compute load gives a system which requires 52 modules to meet the compute load requirements.

The DRAM memory has two forms of I/O one is for the buffer which requires a rate twice that of the input data or 14.8 x10^{12} b/s and the correlator. The number of correlations is

\[
250,000 \text{ frequency channels x 4 Stokes parameters x } 911 \times 919/2 \text{ baselines} = 414 \times 10^9
\]

Assuming a correlator integration of 512 samples then the accumulation for a 2 kHz channels takes 0.5 seconds. Each accumulation requires a 64 bit read and write giving a data rate of 256 b/s per correlation. The correlator DRAM must support a data rate of 106 x10^{12} b/s. The total DRAM I/O required is 121 x 10^{12} b/s. Each module has DRAM I/O of 2.54 x 10^{13} b/s. To meet the DRAM I/O requirements the correlator needs 47 modules.

The DRAM must buffer the data for the 1024 samples at 2 kHz. This is at least a double buffer so the total storage required is for 1 second of data. This is 7.4 x 10^{12} bits or 0.84 Tbytes. The DRAM must also double buffer
the 412 x 10^9 correlations. With 8 bytes per correlation this is 6.6 Tbytes. Each module, with 8GB DRAMs has 0.196Tbytes of storage. To meet this requirement the system needs 34 modules.

To meet the requirements for compute capacity, DRAM I/O and DRAM capacity at least 52 modules are required.

The data rate from a single antenna is 250MHz x 8/7 oversampling x 16bit/Hz x2 pol = 9 x 10^9 b/s (8.5 Gb/s)

If 40 Gb/s links are used at 90% capacity then a link supports data from 4 stations. Total input is 228 links. This requires at least 10 modules and is not the limiting factor in the design. The minimum number is 52 which corresponds to 4 or 5 inputs per module.

The VLBI tied array beam can be implemented directly on the input data. Each FPGA has at most one 40Gb/s link attached. The data per FPGA for the VLBI array beam is about 32 Gb/s. The sample rate is 1.14GS/s (complex). The compute load is 1.4 GCMAC/s and 8.4 GCMAC/s if all FPGAs have the same code. This is 0.5% of the module compute capacity. It is suggest that the beamforming be implemented as a ring beamformer on the 1MHz input data. This could be daisy chained to the other modules so the final output is sum of data from all antennas. A single 10 G link is required between modules for this.

After the channeliser the data must be cross connected between the processor modules. The first cross connect is on the board. After this cross connect each FPGA has 42 MHz of data for at most 20 stations (5 input links). The cross connect between the 52 modules proceeds in two stages, a 4 then 13 way cross connect. The total data rate for 5 links after the channeliser is 142 Gb/s. The board keeps one quarter of this and transports 106 Gb/s to the other three boards in the 4-way cross connect. This uses two optic TX per FPGA capable of transmitting 120 Gb/s. Each of the other three boards receives four of these twelve fibres and ten after a cross connect between the FPGAs each has data for 10.5 MHz of data.

The next stage is a 13-way cross connect. Again there are 12 output fibres and each module keeps 1/13 of the data. After this passive optical cross connect and cross connect operation within the board each FPGA hold 808 kHz of data for all 911 stations and the correlations can now be processed.

The maximum rate from the correlator is 414 x 10^9 correlations every 0.6 second. Each correlation is 64 bit word which give a data rate of 44 x 10^12 b/s. This would give a dump rate of 141 x 10^9 b/s per FPGAs. This requires 15 10 Gb/s links per FPGA. There are also 2 per FPGA per cross connect giving a 19 per FPGA and 114 for the module. In addition there is one for VLBI.

Table 8 Redback processing module resource usage

<table>
<thead>
<tr>
<th></th>
<th>Available resource</th>
<th>Resource used</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 Gb/s link</td>
<td>24</td>
<td>At most 5</td>
<td>21%</td>
</tr>
<tr>
<td>Multimode fibre links</td>
<td>144</td>
<td>115</td>
<td>80%</td>
</tr>
<tr>
<td>DRAM I/O per FPGA</td>
<td>3.173Tb/s</td>
<td>2.32Tb/s</td>
<td>73%</td>
</tr>
<tr>
<td>DRAM</td>
<td>196 Gbytes</td>
<td>131 Gbytes</td>
<td>67%</td>
</tr>
<tr>
<td>Processing</td>
<td>5.25TCMAC/s</td>
<td>4.18 TCMACS (18-bit)</td>
<td>80%</td>
</tr>
</tbody>
</table>
Risks

The design is based on an existing design with all data links already proved to operate at 10 Gb/s. In terms of technology the only unproven items is the DDR4 DRAM. Without physical testing it is not known if this DRAM can operate at 2GHz. A fallback position is reducing the clock rate an increase the number of modules.

There is a risk that the next generation 14nm FPGAs will not be ready on time. The fall back position is to use 20 nm FPGAs which will have about half the performance. This will double the number modules required.

There is a risk that the FPGA dissipation will exceed the capacity of the module fans to remove the heat. Based on the current design this is a low probability. The fallback position is to use water cooling.

The performance of all other parts of the system is already proven.

Development Schedule and Manpower requirements

First installation use is required at the end of 2016. Time from layout to a production batch was 20 months for systems this design is based on. This was achieved with the effort of 1FTE who also was building a parallel system and coding the command and control firmware. So layout needs to start before May 2015. Layout can start as soon as details of the FPGA are available which is expected by the beginning of 2014. This leaves a window of more than a year for start of board layout.

The projected schedule could be

- First two quarters of 2015 develop and document design
- Third quarter of 2015 layout of boards, processing, and command and control
- Four quarter board testing and revision.
- First three quarters of 2016 design testing, revision and verification. Development of all procurement for production
- Last quarter 2016 production

Development of the firmware will be in parallel with the hardware. Prototype firmware exists for ASKAP and APERTIF. This will shorten the design cycle. With this background each firmware component needs ≈3FTE months to implement. Test-beds and documentation will add ≈6FTE months. There are six or seven modules so the total effort over 20 months 60 FTE months or 3 programmers. This will take place in parallel to with the hardware development. But it is noted that many of the modules are common across the three correlators. This is expected to halve the effort attributable to the SKA1 Survey correlator to 30 FTE months. Management adds ~20% so the effort needed is ≈36FTE months or 0.3MEuro. Ten prototypes will add ≈0.15MEuro

Performance analysis

The above design for meets the requirements of the baseline design. Table 3 shows the margin by which the resources meet the requirements.
Cost

The correlator and tied array beamformer consists of 52 Redback-4 modules. If there are 26 1U modules per cabinet then the full correlator minus control computers occupies 2 cabinets with space for LMC, timing hardware and switches to connect data download and LMC computers to the correlator.

The processor board, in volume, production is estimated from ASKAP to cost = $4000 excluding the cost of FPGAs, the control board $500, and the chassis and ATX power supply $600, sundry fans, electrical and optical cables $400. This gives a cost of = $5,500 for the system less FPGAs. Midrange FPGAs in volume are estimated to cost = $800 but there is considerable uncertainty. It is expected that for the SKA the number of pieces required will bring the cost within this range. With this proviso the capital cost of a module $16,300. This includes fabrications of all boards but not assembly. Assembly and testing adds another $500. With and exchange rate of 0.8 Euro to the dollar the capital cost is 9,600 Euro. Each the passive optical cross connects adds 100 Euro per module bringing the cost to 13.340 Euro.

For the 52 modules in the system the cost is 0.70 M Euro.

Operating costs are 5% per year this adds 0.35 M Euro over a 10 year period.

NRE is mainly labour estimated at 36 FTE months or 3 FTE years for firmware and 0.5 FTE for hardware design. At 100 k Euro per year this is 0.3 M Euro. Plus 10 systems for prototyping and revisions will add 0.15 M Euro. The total NRE is 0.5 M Euro.

The FPGA power consumption is estimate at 7.5 W/TCMAC/s or 30 W/TCMAC/s (18 bit). To account for I/O etc this increases to 42 W/TCMAC/s. The compute load is 217.5 TMAC/s, 18 bit. This gives a dissipation of 9.2 kW. In addition there is = 60 W for DRAM and multimode RX a TX (QSFP+ module cost belong in signal transport) and = 30 W of fan power and 10 W for the command and control. This adds 100 W per module or 5.2 kW for 52 modules. The total power is 14.4 kW. If the of grid cost of power is 2.5 Euro per watt per year then over a 10 year lifetime the cost is 0.36 M Euro.

The power infrastructure is estimated to be 1.5 Euro/W with 5% maintenance per year. Over 10 years the infrastructure cost is 2.25 M Euro or 0.03 M Euro.

Detailed costs are summarised in Table 1 at the beginning of the document.

Scalability to SKA2

SKA2 is required approximately two process generations after SKA1. This will increase FPGA performance by a factor of 4. There should also be a 25% increase in clock speed giving a total improvement of 5.

The antenna based processing increases by 10 and the correlation by 100. Antenna based processing is negligible. The increase in processing load is 100 and the design is compute limited. If FPGA were to increase in computational capacity by 5 then the system size increases by 20. The size increases from 52 1U modules to 1040 modules at a cost of = 14 M Euro for the capital cost and total operating cost over 10 years of = 12 M Euro.
The number of modules has scaled faster than the number of antennas. So there is no I/O problem or buffer storage problem. The problem to be solved is achieving the correlator dump rate. This is a requirements problem as the rate without any change goes from 44Tb/s to 4.4Pb/s.

The FPGA system is estimated to cost ≈26MEuro including operation for 10 years and occupy ≈30 cabinets making it a possibility for SKA2.

The cost of the correlator is such that an ASIC solution should be explored. The ASIC is one generation behind but is expected to have 4 times the computation capacity. This could possibly reduce the cost of the system by a factor of 4. Buffering requires single FPGA per processing board and the rest is for the ASIC. A significant reduction in the number of modules is possible. The problem to be solved is sufficient I/O to DRAM and to SDP. This may make this approach infeasible.

**Power-MX**

Power-MX is a platform being developed by Carlson et al. [PowerMX_TechnicalDescription_VDRAFT_Apr10-2013_Release.pdf]. A simplified diagram showing how data flows through the board is shown in the figure below. The main advantage of this system is that it can hold up to four processing sites and is agnostic of the type of processing element that is installed on-site. This makes it an extremely flexible motherboard which is likely to be ubiquitous throughout the SKA-processing systems.

![Simplified PowerMX block and data flow diagram. Each lane of each highway contains 12 hi-speed serial paths.](image-url)
Each of these 4 processing sites have high performance connectors to the Power-MX motherboard and I/O(s). Each mezzanine site (90x90 mm) can host a single large device (45x45 mm) and memory modules (4 shown for each), several smaller chips, or whatever circuitry that fits and is required for the task. If motherboard I/O is not required, the mezzanine card can be extended in one dimension (horizontal in the above figure) to provide additional capability and perhaps custom I/O.

Each site has associated I/O highways known as “M1-M4” motorway in figure 10; each motorway contains 4 lanes, and each lane contains 12 hi-speed data paths up to 28 Gbps each (and pending further study, possibly up to 40 Gbps each). Therefore, each highway contains 48 hi-speed data paths, with direction determined by the mezzanine card and I/O modules. In total, along each edge, there are 192 hi-speed paths.

Each lane at each edge of the board routes to a high-speed 12-signal connector, and the connector can host a desired pluggable transceiver module, or 2 or more connectors at each edge can be ganged to plug in a user-defined I/O module extending a user-defined distance away from the edge of the board. There are also 8 lanes between adjacent sites, with direction determined by the mezzanine modules.

There are also dual 10GBASE-T connections from each site to motherboard RJ-45 connectors providing 8x10G auxiliary I/O paths (not shown). For M&C, there are 1G or 10 G links to a central M&C SoC card (System on a Chip, also not shown), and 1G or 10G Ethernet on fiber or copper from that SoC to the outside world. Also not shown are clock I/Os with ADC-quality clock delivery to each mezzanine site for radio astronomy applications. The motherboard is ≈17” x 7”, and, depending on power and cooling, can fit within a 1U (1.75”) pizza box or other carrier. Further information can be found in the technical description [Carlson et al.]

**F-part**

Similar to what has been described in the previous sections, the F-part has to perform the coarse and fine channelization as well as fringe rotation.
Here we assume that the aggregation of beams has happened in the station processing. The delay buffer is implemented in off-chip memory. Having 32-antennas streaming into each FPGA, we would require 262kB words for the FFT and a further \((350kB \times 32 = 11MB)\) 11MB for delay (as stated in the TOC V1.1). This should **NOT** constitute a problem for doing the first and second stage processing.

Following the same method we used for Uniboard approach, if we assume that each FPGA gives us 3.5TMAC/s, then we require 10 Power-MX boards for the channelization, but for the bandwidth we require 30 of these POWER-MX boards.

**X-part**

In order to carry out the X-engine we propose to use the double POWER-MX board as shown in the figure 12. This is now almost exactly equivalent to the Uniboard (in terms of processing), but much more flexible and can also be re-arranged into a triple POWER-MX board. As before for we would require 60 POWER-MX boards for the X-part. This means that the POWER-MX system is very closely equivalent to the Uniboard approach albeit **MUCH MORE FLEXIBLE.**
Each chip correlates 1/128th of the BW (19.5 MHz/poln) all baselines.

Rear of 16-20 slot mesh backplane

Polyphase filter bank

Dual 10GBASE-T from each FPGA to SDP (or spare/auxiliary fibers)

Dual-PowerMX Blade

Pre-
There are no critical path long-lead time items in this proposal that need to be expedited beyond what the general SKA schedule mandates. For example it doesn’t seem, if the cost and power numbers in this analysis are correct, that ASICs for correlation or beamforming would be cost effective in this case.

The main technology item is development, test, and setting in motion volume manufacturing of the PowerMX board and mezzanine cards, as well as setting in place firmware and software IP management and practises that maximize design re-use across projects (sub-elements within the CSP, and if desired, other consortia that might use the same platform) to minimize development and re-development effort.

Development of PowerMX can occur concurrently with sub-element system engineering, provided base specifications (functionality and performance, form factor, mounting holes, connector pinouts, M&C definition) are in place. Much of this has been done at a draft level in the PowerMX technical description document. Further feasibility investigation and refinement leading to a “base standard” during Stage 1 will (if successful)
lead to a situation where now multiple parties can start building, testing, and deploying according to that standard.

*Performance analysis*

The above design meets the technical specifications of SKA-1 LOW, with sufficient margin (≈20%) to spare.

*Cost and Power*

This is detailed in Table 1 at the beginning of the document, but the basic methodology is hereby reproduced.

If we assume a single-blade costs $15,412 and a double-blade costs $30,124 then the total cost for the SKA-1 LOW processing becomes:

\[(30 \times 15,412) + (60 \times 30,124) = 1,958,860 = \$2,000,000\]

If we assume that each single Power-MX blade consumes 193W and each double Power-MX blade consumes 386W, this means that the total system requires:

\[10 \times 193W \text{ (Single Power-MX)} + 60 \times 386W \text{ (Double Power-MX)} = 25kW \times 2.5 = 63kW\]

(N.B. This is slightly lower and maybe due to the extra overhead added in the Uniboard control/data)

Running this for 10 years would cost $630,000

Adding 10% Spares gives: $15,412 + 90,372 = $105,784

Total Cost = $2,750,000

Miscellaneous costs:

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-20 slot mesh backplane slice (i.e. one lane) ea</td>
<td>$20,000</td>
<td>This should be highly conservative.</td>
</tr>
<tr>
<td>16-20 slot shelf, dual or triple blade depth</td>
<td>$8,000</td>
<td>This should be conservative.</td>
</tr>
<tr>
<td>Bare (custom) rack to hold dual back-to-back, up to 35” wide shelves</td>
<td>$10,000</td>
<td>Doesn’t include NRE. 4 posts, welded, sides and doors. Fan costs accounted for later.</td>
</tr>
</tbody>
</table>

Total Cost = $3,000,000 without maintenance costs.

Detailed costs are to be found in Table 1.
**GPU/PC**

Assuming the availability of Volta with STFMAC/s (SP) and 8Tbit/s per board, we find that we require 435GPU-cards to perform the processing requirements (again at 0.8x programming efficiency). Assuming 200W per card and multiplying by an overhead of 1.5x (difficult to know if this is correct) then the power consumption ≈130kW.

One important thing to consider is the hybrid approach where the F-part of the Correlator is done in the FPGA, where the processing is I/O bound while the X-engine is done in a GPU-cluster where the processing becomes compute bound.

A further advantage of the GPU architecture is that it may be possible to incorporate imaging and calibration systems within the same compute architecture.

**Development schedule**

Since this is based on Custom-off-The-Shelf (COTS) hardware, it can be bought and installed as the hardware becomes available commercially. The development of the code can be all done before the availability of the hardware unless the programming libraries significantly change, but this is not expected.

**Performance analysis**

There is enough compute and I/O to reach the required specifications for SKA-1 LOW, but there is not enough experience within the author list of this white paper to comment on the efficiency numbers of both the algorithms and their implementations.

**Cost and Power**

This is summarised in Table 1 at the beginning of the document.

**ASICS (Larry D’Addario)**

This section describes the use of ASICS to implement either channelization or correlation or both. It is shown that such choices result in lower power consumption by a factor of 50 to 500 compared with the alternatives previously described.

The description here is less complete than some of the preceding sections in that it does not cover all system requirements. In particular, delay tracking, fringe rotation, and long-term integration are not discussed. This is because only about 3 person-days were available to create this part of the white paper. It is nevertheless sufficient to demonstrate the power reduction and to provide a credible cost estimate in view of the fact that channelization and correlation constitute the vast majority of the processing.

**Correlator**

The design here is based on a correlator ASIC that is under development at JPL and whose RTL design is nearly complete. The existing design uses 2b+2b input samples, but the estimates here are based on scaling the design to 4b+4b input samples. Details are given in Appendix A.

The correlator subsystem for SKA1-low would consist of 32 custom PC boards, each containing 16 of the correlator ASICS and one Virtex 7 FPGA to handle board-level I/O. Each of the ASICS computes all cross
correlations among 2048 signals (more than the 1822 required) for 0.488 MHz of the bandwidth. The FPGA might also be able to handle long-term integration, but that feature has not been included in the estimates here. The board-level input data rate is 64 Gb/s; with no long-term integration, the output rate is 78 Gb/s. These rates are relatively easy to handle with ~12.5 Gb/s lanes.

Filter Banks
For the SKA-1 LOW, we require to channelize 1822 signals using and FIR-FFT polyphase filterbank. This can be done using ASICs having the architecture described in Appendix B. This is an architecture-only description and no work on the actual ASIC design has been done, so the estimates are less reliable than those for the correlator ASIC. The power and size are based on using state-of-the-art techniques from published papers in a 32 nm process. It appears that about 8 filter bank structures capable of producing $2^{18}$ channels can be fit onto each ASIC, so 228 ASICs are needed for all signals.

One possible board-level architecture is to include several of these ASICs on each mezzanine board of a Power-MX motherboard. If we can fit 6 ASIC(s) on each mezzanine board (which requires 40 Gb/s of input and output bandwidth per board), then each POWER-MX motherboard could hold 24-ASIC(s), and we need a total of 10 POWER-MX motherboards. However, the cost and power estimates below are based on construction of a custom PC board that carries 4 ASICs and provides filter banks for 32 channels. Board-level I/O then adds up to 153.6 Gb/s (in and out), which can be handled by 16 10 Gb/s transceivers. To handle the 1822 signals, 57 of these boards are required.

Performance Analysis
The architectures described above fulfil the required specifications of the SKA-1 LOW correlator and channelizer.

Cost and Power
Due to the relatively small number of ASIC devices needed (512 correlator, 228 filter bank), we plan to order additional copies of the multi-project wafers built for prototyping, rather than paying for dedicated mask sets. This results in high per-chip cost (~1200 US$), but lower total cost. The following is a summary of the cost estimates that are given in more detail in Appendix C.

NRE for the correlator ASIC is relatively low because its design is simple and partly complete. It is higher for the filter bank ASIC because that design is not yet well developed and it is expected to be more complicated. In each case, if the ASIC is actually built we assume that it will be used for all three telescopes of SKA1, so only 1/3 of the NRE is allocated to SKA1-low.

The power estimates (see Appendices A and B) include all of the processing chip dissipation (ASICs and FPGAs), multiplied by 2.0 to account for power supply efficiencies and infrastructure and by 1.5 to account for (external) cooling, for a total multiplier of 3.0. The power cost is estimated at the rate of 1.0 €/W/year.
**Correlator**

<table>
<thead>
<tr>
<th>Component</th>
<th>NRE</th>
<th>US$</th>
<th>€</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIC</td>
<td>1.0M</td>
<td>1.0M</td>
<td></td>
</tr>
<tr>
<td>PCB</td>
<td>150k</td>
<td>150k</td>
<td></td>
</tr>
</tbody>
</table>

NRE shared with other parts of SKA1; SKA1-low share 1/3:

- **Total correlator NRE for SKA1-low**: 383k US$ 295k €
- **ASIC production chips (including spares)**: 627k US$ 482k €
- **Board production (excluding ASICs)**: 96k US$ 73.8k €

- **Total power (incl. cooling)**: 1,920 W

- **Cost of power for 10 years**: 19.2k €

**Total Correlator Cost**: 870k €

**Channelizer**

<table>
<thead>
<tr>
<th>Component</th>
<th>NRE</th>
<th>US$</th>
<th>€</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIC</td>
<td>2.0M</td>
<td>2.0M</td>
<td></td>
</tr>
<tr>
<td>PCB</td>
<td>150k</td>
<td>150k</td>
<td></td>
</tr>
</tbody>
</table>

NRE shared with other parts of SKA1; SKA1-low share 1/3:

- **Total correlator NRE for SKA1-low**: 717k US$ 551k €
- **ASIC production chips (including spares)**: 627k US$ 482k €
- **Board production (excluding ASICs)**: 285k US$ 219k €

- **Total power (incl. cooling)**: 1,553 W

- **Cost of power for 10 years**: 15.5k €

**Total Channelizer Cost**: 1,267.5k €

**Other (very rough estimate of other system costs not included above)**

<table>
<thead>
<tr>
<th>Component</th>
<th>NRE</th>
<th>US$</th>
<th>€</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRE</td>
<td>250k</td>
<td>250k</td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>100K</td>
<td>100K</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>375</td>
<td>375</td>
<td>3.8k</td>
</tr>
</tbody>
</table>

**Total Other**: 285.8k €
Appendix A

Correlator ASIC Summary
Larry R. D’Addario, Jet Propulsion Laboratory, California Institute of Technology
29 April 2013

I. INTRODUCTION

This memo summarizes information on low-power ASIC designs that might be applicable to all three telescopes of SKA1. Some of this may be useful in the CSP White Papers now being written, and later in the proposal to the SKAO. Only the cross-correlation engine of a correlator (“X” part) is addressed here.

The minimum-power architecture for large-\(N\) cross correlation was addressed more than two years ago [1]. Since then, logical design of an ASIC that uses the optimum architecture has been nearly completed at JPL. That design is described in a high-level data sheet [2]. It targets an IBM 32 nm SOI process and is synthesized from an ARM standard-cell library along with embedded DRAM and PLL cores from IBM. Here the properties of the chip are elaborated in the context of the SKA1 telescopes and its performance at the board and system level are estimated. Results are scaled to estimate the power in case the design is modified to support wider input data words.

II. PERFORMANCE ESTIMATES

The ASIC is designed to be usable for cross-correlating \(2^N\) signals when \(N\) is between 32 and 1024. It can also be applied outside this range, but it is then much less efficient. A single chip always correlates all baselines, but the bandwidth processed decreases inversely with number of baselines. For cost reasons, the design does not include high-speed SERDES, so we assume here that an FPGA is used at the board level to aggregate I/O from multiple ASIC devices and provide \(\sim 10\) Gbps serial I/O streams into and out of the board. Specifically, we consider a board with 16 ASICs and one Xilinx Virtex 7 FPGA. About 30% of the board-level power is used by the FPGA.

<table>
<thead>
<tr>
<th>(N_{\text{max}})</th>
<th>(B)</th>
<th>(\text{beams})</th>
<th>(\text{Board})</th>
<th>(\text{Boards})</th>
<th>(\text{Estimated Power, W})</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHz</td>
<td>MHz</td>
<td>2b+2b</td>
<td>3b+3b</td>
<td>4b+4b</td>
<td></td>
</tr>
<tr>
<td>SKA1-low</td>
<td>1024</td>
<td>250</td>
<td>1</td>
<td>8</td>
<td>606</td>
</tr>
<tr>
<td>SKA1-mid</td>
<td>256</td>
<td>1000</td>
<td>1</td>
<td>125</td>
<td>165</td>
</tr>
<tr>
<td>SKA1-survey</td>
<td>128</td>
<td>500</td>
<td>36</td>
<td>250</td>
<td>1,018</td>
</tr>
</tbody>
</table>

Table 1 summarizes the results for the SKA1 telescopes. In each case, the maximum number of antennas that can be correlated exceeds the number planned for the telescope and the bandwidth is inferred from the Baseline Design document\(^2\). The bandwidth that each 16-ASIC board can process is given along with the number of boards required to cover the full bandwidth for all beams. The column

---

\(^1\) ©2013 California Institute of Technology; all rights reserved. Government sponsorship acknowledged.

\(^2\) The bandwidth for SKA1-mid is ambiguous in the Baseline Design document. The number of boards and the total power are proportional to bandwidth, so the results can be easily scaled to any other bandwidth.
"2b+2b" under "Estimated Power" is for the ASIC design described in [2], which uses input data samples 4 bits wide, 2b real and 2b imaginary. In the remaining columns, the required power is scaled to hypothetical designs with 6 bit (3b+3b) and 8b (4b+4b) input words. This is done by doubling the input data rate and the internal DRAM size of each ASIC, and assuming that the chip area and power consumption of each complex multiplier-accumulator (CMAC) are proportional to $0.5(1+b^2/16)$, where $b = 4, 6, 8$ is the input word width. That is, it is assumed that half of each CMAC (the multipliers) grows as $b^2$ and the other half (the accumulators) remains unchanged. In the absence of a more detailed analysis, this gives a reasonable estimate.

For the 2b+2b design, the ASIC power is estimated from simulations and the data sheets of instantiated IP, and the FPGA power is estimated by Xilinx's Power Estimator spread sheet. The ASIC power is only its internal dissipation (static and dynamic), and the FPGA power includes only I/O. To account for things not included in these estimates, as well as power supply overhead, the total is doubled to get the numbers in Table 1 (100% overhead and margin).

III. DEVELOPMENT COST

To take the ASIC design from its present state (RTL code and synthesized netlist available) to having prototype chips in hand will cost, very roughly, USD 1.0M. This is based on being included on a multi-project wafer (MPW), which will result in delivery of 40 to 60 chips. The largest costs have the largest uncertainties. Some of them might be lower through universities, or reduced by negotiation with IBM. With adequate funding and availability of skilled personnel, it is possible to have prototype chips in about 12 months, depending on the MPW fabrication schedule (which has only two runs per year). A more comfortable schedule would allow 18 to 24 months.

REFERENCES
Appendix B

ASIC-based Filter Banks
Larry R. D’Addario, Jet Propulsion Laboratory, California Institute of Technology
23 April 2013

I. INTRODUCTION

This memo provides a discussion of possible ASIC implementations of the filter banks (channelizers) needed in all three telescopes of SKA1. Unlike an earlier discussion of correlator ASICs, this is not based on a specific design but rather on published results and rough calculations. It is intended only to give a quantitative estimate of what might be possible. Much more work is needed to produce a practical design.

II. FILTER BANK REQUIREMENTS AND STRUCTURES

The heart of almost any uniform digital filter bank is an FFT, but radio telescopes usually precede the FFT by an FIR filter that is designed to improve the channel bandpass shape. In the most common (but not the only) implementation, the pre-filter has a polyphase structure, where one branch of length \( L \) for each of the \( N \) channels produces a result every \( N \) input samples. The pre-filter outputs are then the inputs to a length-\( N \) FFT. Since the pre-filter weights are generally real, it requires real multiplications and real additions at rate \( 2LB \) for input sample rate \( B \). The radix-\( k \) FFT then consists of \( \log_k N \) pipeline stages, each of which does "butterfly" operations at rate \( B \), where a butterfly requires \( k−1 \) complex multiplications and \( k \) complex additions (although these counts vary slightly depending on the implementation algorithm), so that the rates of real multiplications and additions are \( 4(k−1)B \log_k N \) and \( (4k−2)B \log_k N \), respectively. Almost always, \( k=2 \) or \( k=4 \); in ASIC implementations, radix \( k=4 \) is usually best.

For this discussion, we will fix the pre-filter length \( L \) at 8. Values between 4 and 12 are generally sufficient and the choice is independent of \( N \) for reasons that are beyond the scope of this memo. From the above expressions we find that the arithmetic in the FFT exceeds that in the pre-filter whenever \( N > 16 \) at either \( k=2 \) or \( k=4 \).

Memory requirements are at least as important as arithmetic operation rates. The pre-filter requires a shift register of length \( LN \) (complex) input words. The FFT requires memories totaling \( N−1 \) to \( 2N−2 \) words, depending on the implementation algorithm; in a pipeline structure, this is not a single memory but rather multiple small FIFOs that must operate simultaneously. Similarly, the pre-filter memory consists of \( L \) simultaneously-operating FIFOs of length \( N \). For \( L=8 \) (or for any value from 4 to 12), the memory requirement is dominated by the pre-filter.

According to the Baseline Design document, all three telescopes should have uniform filter banks of length \( N = 2^{18} = 4^9 = 262,144 \). For such a long filter, an implementation on a single chip (whether it is an ASIC, FPGA, CPU, or GPU) would require that the vast majority of active elements on the chip are memory rather than arithmetic logic. Consequently, all implementations rely on off-chip memory. (For GPUs, this usually seems to require use of main memory, common to all cores, rather than local cache. This hurts efficiency.)

The memory requirement for the FFT is eased if it is broken into two stages, where an FFT with, say, \( N_1 = 2^9 \) is cascaded with another having \( N_2 = 2^9 \) producing results identical to an FFT of length

---

3 ©2013 California Institute of Technology; all rights reserved. Government sponsorship acknowledged.
$N_1N_2$. This does not reduce the total memory needed, but it moves most of it into a "corner turner" between the stages, which can be a single memory of $N$ words that does one write and one read per sample. However, this does not ease the memory requirement for the pre-filter.

Another approach involves breaking not only the FFT but also the pre-filter into two stages, so that each stage is a complete filter bank of length $N_1$ or $N_2$. For reasons that are beyond our present scope, this produces major performance problems due to interchannel aliasing. The aliasing can be reduced to an acceptable level (but not eliminated) if the first stage produces outputs that are oversampled by a factor of 1.1 to 1.3. It can do this by overlapping successive blocks of $N_1$ input samples.

III. A STRAWMAN ASIC IMPLEMENTATION

A. Description

Here I propose a flexible filter bank ASIC having the structure shown in Figure 1, which consists of a length-1024 filter bank, a length-256 filter bank, and an on-chip RAM of 1024x256 words, along with two input ports and two output ports. Two controllable switches allow these elements to be configured as either two separate filter banks or cascaded to create one long filter bank. Each of the elements can be separately powered off if it is not being used. The first filter bank can be operated in either normal or overlapped (oversampled) mode. The complete chip consists of as many copies of the Fig. 1 circuit as will fit.

![Diagram](Diagram.png)

**FFT**

At the cost of considerably more control logic, it is probably possible to provide additional flexibility by allowing the length of each filter bank to be reduced by a user-selectable power of 4.

Many things needed in a practical circuit are not shown in Fig. 1, including clocking and additional buffering.

B. Performance

<table>
<thead>
<tr>
<th>Table 1: Strawman ASIC Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Parameters used for analysis</strong></td>
</tr>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>Energy per write+read</td>
</tr>
<tr>
<td>Energy per I/O bit</td>
</tr>
<tr>
<td>Area per MAC</td>
</tr>
<tr>
<td>Area per FFT r4 stage</td>
</tr>
<tr>
<td>Area per Mb of DRAM</td>
</tr>
<tr>
<td>Area per SERDES+xcvr</td>
</tr>
</tbody>
</table>

TOTAL | | | | | | | | 4.73E-10 | 11.24 |
Table 1 is an attempt to estimate the performance that might be achieved by such a chip. Table 1a shows the parameters used. The column labelled "SoA" (state of the art) is a measure of what has been achieved in 65 nm technology, based on the reference cited under "Source". The MAC numbers are for a single real multiplier and adder; these are scaled from the ALMA chip (250 nm, see [1] section A1.B). The memory numbers are for embedded DRAM and are obtained from the CACTI simulator [3]. In all cases, the scaled or simulated numbers are doubled for margin.

For a few of the energy parameters, an FPGA value is given for comparison. The MAC and butterfly energies are based on working correlator and filter bank designs, respectively, in ROACH-1 boards (Virtex 5); here the comparisons are a bit unfair because the FPGA values include all board overhead. The I/O energy is based on 12.5 Gbps transceivers in Virtex 7, using the Xilinx Power Estimator spreadsheet; it is still 6x the 65 nm SoA, even though Virtex 7 is built in 28 nm technology.

Table 1b applies the numbers in 1a to the straw man chip. Here we assume that each input or output word is 4b+4b wide, and that the input and output blocks in Fig. 1 are paired to make two transceivers. Internally, however, we assume wider data words; in particular, the memory is sized at 1 MB = 8 Mb to hold 256K words at 16b+16b each. The total energy per word processed is 0.473 nJ, or 0.473 W at B=1 GHz. This assumes that all elements are operating at full speed, which would not normally be the case (e.g., in cascaded mode only one I/O pair would be used, and in dual mode the memory would not be used.) The total area is a bit more than 11 mm², which should allow at least 8 of these structures to be built on one reasonable-size chip. Notice that more than half of the power and half of the area are used by the memory. Both power and area are conservative estimates for 65 nm technology, and both would be substantially lower in more advanced technologies.

Assuming 8 structures per chip, these results imply the following totals for SKA1:

- **SKA1-low**, 911*2 signals, \( B=250 \text{ MHz} \), \( N=262,144 \): 228 chips, 215.5 W.
- **SKA1-mid**, 254*2 signals, \( B=1 \text{ GHz} \), \( N=262,144 \): 64 chips, 240.3 W.
- **SKA1-survey**, 96*2*36 signals, \( B=500 \text{ MHz} \), \( N=256 \): 432 chips, 388.8 W

(-survey has memory off, both filter banks of each block used separately).

To get system-level power consumption, board and system overhead must be added to these chip-level estimates.

**REFERENCES**


## Appendix C

### ASIC-based Cost and Power Estimates

**SKA1-low**

**ASIC Implementation** All costs in k$US(2013)

### Cost and Power

<table>
<thead>
<tr>
<th>NRE</th>
<th>Quantity</th>
<th>Cost</th>
<th>Power</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SKA1-low</td>
<td>req’d</td>
<td>avai.</td>
<td>k$</td>
<td>W</td>
</tr>
<tr>
<td>Channelizer (256K channels/signal)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIC NRE</td>
<td>2000</td>
<td>667</td>
<td></td>
<td>includes ~50 prototype chips via MPW</td>
</tr>
<tr>
<td>Board NRE</td>
<td>150</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signals to process</td>
<td>1,822</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIC production</td>
<td>228</td>
<td>250</td>
<td>285</td>
<td>5 extra MPW wafers at $57k</td>
</tr>
<tr>
<td>Board production</td>
<td>57</td>
<td>57</td>
<td>171</td>
<td>$1k parts + $2k fab&amp;assy per board</td>
</tr>
<tr>
<td>Total power</td>
<td>1822</td>
<td></td>
<td>1,035</td>
<td>0.284W/signal + 100% overhead</td>
</tr>
<tr>
<td>Correlator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIC NRE</td>
<td>1000</td>
<td>333</td>
<td></td>
<td>includes ~50 proto chips; design partly done</td>
</tr>
<tr>
<td>Board NRE</td>
<td>150</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASICs</td>
<td>512</td>
<td></td>
<td>0.488 MHz/ASIC</td>
<td></td>
</tr>
<tr>
<td>ASIC production</td>
<td>512</td>
<td>550</td>
<td>627</td>
<td>11 extra MPW wafers at $57k</td>
</tr>
<tr>
<td>Board production</td>
<td>32</td>
<td>32</td>
<td>96</td>
<td>$1k parts + $2k fab&amp;assy per board</td>
</tr>
<tr>
<td>Total power</td>
<td>1,280</td>
<td></td>
<td>16 ASICs and FPGA per board +100% overhead</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
<td>things not included above</td>
</tr>
<tr>
<td>NRE</td>
<td>250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost subtotals</td>
<td>1,350</td>
<td></td>
<td>1,279</td>
<td></td>
</tr>
<tr>
<td>System-level power</td>
<td></td>
<td></td>
<td>3,847</td>
<td>Sum of above + 50% for cooling</td>
</tr>
<tr>
<td>Power cost</td>
<td></td>
<td>50</td>
<td>1.3$W/year * 10 years</td>
<td></td>
</tr>
<tr>
<td>Total life cost</td>
<td></td>
<td></td>
<td>2,679</td>
<td></td>
</tr>
</tbody>
</table>
Software Correlator for SKA-Low White Paper
SKA1-LOW-CSP/CBF
Stephen ORD

ABSTRACT
This document establishes foundational feasibility of a software based correlator for the SKA-low element, for the purposes of inclusion in the response to the RFP presented by (NRC/MDA-led) CSP consortium
**Introduction**

This document outlines the technical feasibility of a commercial off-the-shelf (COTS) solution to the hardware requirements of the SKA Central-Signal-Processor. It demonstrates that a system fully capable of meeting all requirements can be built within the permitted budget for both capital costs and operational expenditure.

This document addresses the requirements of SKA1-Low - consisting of 911 stations with an instantaneous bandwidth of 250 MHz and 1 beam per polarization.

It is not the purpose of this document to champion a particular technology / approach, but, since is required that costing estimations are provided we will assume the compute will be provided by the NVIDIA GPU product line and that they will be housed in servers comparable to the IBM iDataPlex form factor (3U, housing 2 GPU per server), connected via 10Gb Ethernet and a 40Gb+ (QDR infiniband).

We will also scale power consumption and compute capability to 2016 but consider that the preconstruction development process will identify COTS hardware and novel applications that will substantially improve upon these predictions.

The system is costed here in current Euros and we believe that the only significant risk is in the available throughput of an enterprise class server at that time.

**Subsystem Description**
Software correlator group

The team assembled to develop the software correlator is already very experienced and in the process of building software correlators for operational instruments. They are:

- **Curtin University/ICRAR.** The team has implemented a GPU based correlator for the MWA, which is an SKA pathfinder. The system consists of 40 blades and 48 GPUs (24 of which are used for correlation). This is an operational system and has been used to scale the performance of the proposed system in this White Paper.

- **ASTRON.** A large software correlator has been in use for 5 years based on BlueGene for LOFAR, a major pathfinder for the SKA. This has refined many of the techniques required for an operational instrument. Following considerable research and benchmarking an upgraded software correlator is being developed based on a similar architecture to this proposal. This new correlator, COBALT, will be available by the end of 2013 for operational use.

- **NCRA.** A software correlator using CPUs has been built and is in use with the existing 30 antenna, 32 MHz BW configuration of the GMRT telescope. Substantial work towards a new software correlator based on NVIDIA devices is being developed for the upgraded GMRT (30 stations, 400 MHz BW). This is a slightly different data architecture, but is based on the same hardware as proposed here. This correlator is scheduled to be on-line in 2014. This will enable comparative studies of performance.

- **Cambridge University.** As the lead for the SDP consortium Cambridge is organizing the resources for the major post processing developments including architecture, hardware and software algorithms. The HPC Services department is part of the Cambridge team, they already manage the largest Intel based cluster in the UK, currently with 128 Tesla GPU sub-cluster and shortly to be upgraded to a 256 GPU (K20) cluster. The data communications uses 56Gb/s Infiniband on the Mellanox range of switches. This is available for testing correlators in a large environment. Further, the facilities will be extended to provide a “computing lab” for a number of different processors and architectures for SDP testing – this will all be available for software correlator performance and implementation testing. There is also great experience in managing a large HPC cluster which guides the implementation.

- **KASI.** While there is little resource presently available, considerable research has already taken place on software correlator architectures.

- **NVIDIA.** As can be seen the current architectures and the implementation described here are based on NVIDIA GPUs. As a company NVIDIA are very active in supporting this work with all the institutions above and the SKA in general. They have committed full time resource for this and SDP work.
at Cambridge and considerable support and advice elsewhere. They will ensure that we have maximum performance from the processors and advise us on the roadmaps for forthcoming devices.

- **CISCO Systems.** As a partner in the Australian SKA-preconstruction effort Cisco Systems will design the networking infrastructure and via a sub-contract with Curtin will examine the possibility of distributing some of the computing load within the network itself.

As can be seen from the foregoing, we will have representative correlators running before the end of Stage 1 of the Preconstruction phase, which of themselves can be scaled to prove SKA 1 correlators – they can be tested as a large system and be integrated closely with the SDP work.

The skills are fully available within this group to deliver the three correlators and beamformers for the SKA Phase 1.

We consider that implementing a software correlator is low-risk, scaleable, cost-effective and flexible solution for the SKA1.

**Technical Description**

We detail a system that meets all of the requirements of the SKA-LOW correlator and can be built today. This is a straw man model that is modular (separate F-stage and X-stage and a simple switch interconnect. It may be more cost effective to buy this as a modular supercomputer solution from a major vendor and to upgrade the compute elements as required e.g. several cabinets of Cray XC30.

There are 911 stations of 289 antennas, each station is producing a single independent dual polarization beam. We will assume that this data can be switched in such a manner that any single beam from a single antenna can be directed to any server.

**Filterbank/Channelizer**

We have been asked to assume a wideband input 250 MHz in width, although the elements will have to be channelized to some degree to form the beams we will assume no channelization has taken place as a worst case.

The F-stage of the correlator can be performed in multiple stages – for example a coarse stage prior to the cross-connect and a fine stage within the X-engine, but for simplicity we will assume that all the F-stage is performed in one step. Although functionally a different element to the X-engine in this model the F-stage will be performed on the same machines as the X-stage. In previous implementations this has been performed on both the host CPU and the compute GPU. In this case however I will assume we are performing all the (whole sample) delay correlation and the full F-stage in a standalone cluster.
Delay Buffering

With a delay tracking buffer required to be sized for +/- 50 km if we perform the delay tracking after the F-stage then there will be a significant drop in correlator sensitivity as the whole sample delay can only be applied in units of FFT length (0.2M samples). It is better to apply the whole sample delay before the FFT calculation. But the buffer size required is trivial (+/- 350,000 8bit samples)

Phase Rotation

It is possible for the phase rotation to be applied after the FFT, in fact this may be desirable as it allows the rotation to be applied at 32bits of precision. This could be applied before the samples are distributed across the interconnect. Alternatively it can be applied in a “phase tracking” operation within the correlator.

Input Rates

A dual polarisation 250 MHz wide beam, 8 bit - digitised is 8Gb/s. We need to capture 911 of these beams: which requires ~130 Infiniband (56 Gb/s) interfaces. We will assume 456 servers, each handling two station, hence capturing 2, dual polarization antenna signals of 10Gb/s each.

If the full F-stage is required to be performed for 4 streams (dual polarization for 2 antennas) the operations required are:

\[ 4 \times 2 \times \left( \frac{250E6}{2} \right) \times \log(262,144) = 36G.MACS \]

This level of compute is within the capabilities of a CPU, it can be performed now: an Intel Xeon E5–2650 is capable of > 120G. MAC. It can of course easily be performed on a GPU.

Output Rate

We are to assume N=262,144 for the input/output of the X-engine. The F-stage native output will be 32 bit floating point. This will be down sampled to 8 bit for output. Assuming no oversampling in the filterbank this gives an input and output rate that is identical to the input.
Cross-connect

There are multiple options here but the most sensible solution is a 500+ port Infiniband 56 Gb/s switch. This will perform the corner turn. Redistributing the 262,144 channels over the 456 machines (576 per server). Each server therefore holds the same 576 channels for a single dual polarization beam for all 911 antennas. These switches are currently available from Mellanox, the SX6536, http://www.mellanox.com/page/products_dyn?product_family=122.

X-engine

The X-engine is required to perform the X correlation for 576 channels (0.55 MHz) for 911 elements. An operation that requires ~7Tflops well within the capability of a future GPU (an Nvidia K10 is available now and is capable of 4.58 Tflops). However in order to provide a conservative estimate we will assume 2 GPU are required per server. For a total of 912.

The full data rate out of the correlator at maximum frequency resolution is 13GB/s per GPU, which is just within the bandwidth of PCIe3.0: which could be carried to SDP on 2 infiniiband links.

However we have been asked in the WP-TOC to consider dropping the number of channels per output polarization to 4096. The reasoning being that full resolution is not likely to be required over the full bandwidth. This implies that the output data rate would be 1.6Gb/s which can trivially be transferred by 10Gb Ethernet.

A spreadsheet that allows the manipulation of these parameters is included in this package.

Beamformer

While the Baseline design does not call for time domain processing on SKA-low the processing system described above can readily perform as a beamformer at no additional cost and using software developed for SKA-mid.

System Summary

The proposed system (for purchase within 2 years) for correlating 911 stations and a single beam from SKA-Low is therefore 456 enterprise class Xeon servers, each housing a dual channel FDR Infiniband card. Each server should hold sufficient GPU cards to supply 7Tflop of computation

The interconnect can be a 500+ port Infiniband switch from Mellanox

The output to the SDP readily fits on a single 10Gb line.
**Risk**

There is little risk in this implementation as all the elements can be purchased now. It will be possible to reduce the required number of servers by a factor of 2 when the GPUs are individually capable of ~7Tflop which is likely by 2016. If this approach was taken a risk in the current outline is the throughput of an enterprise class server.

The Xeon family of processors currently have a memory bandwidth of up to 32 GB/s – with dual processor chipsets achieving double this. We consider that by 2016 it will be routine to process data at the required rate.

**Mitigation**

We have access to multiple high performance compute and networking solutions in the various facilities available to our collaboration. We also have prototype and operational software that can be tasked to perform every aspect of the required compute. We will know by the end of stage one of preconstruction whether this risk can be retired.
Development Schedule

The Plan of Execution is outlined in the WPEP for this PIP.

Hardware:

As a consortium of research groups we have access to software correlators running on a wide variety of platforms, and we will begin testing and development at Cisco and NVIDIA throughout the Stage 1 preconstruction to determine which technology and network infrastructure would likely be most efficient.

We will also be able to retire risks in the design and run full scale (several percent) prototypes both in the lab and even at the MRO site. We have obtained five percent of the Directors Time of the MWA to pursue prototyping.

Software

It is likely that we will devote considerable time to software design, combining the best elements of our software architectures. Optimization will be required depending upon the platform that is chosen. We have therefore enlisted NVIDIA to supply engineering expertise to ensure that our software design is as efficient as possible.

Within our research group there are three operational software correlators in the field. In addition there is significant Research and Development effort that will be directly targeted at prototyping during Stage 2 of preconstruction. I therefore consider that any additional software development cost to actually “construct” the correlator will be minimal. However considerable effort will be directed at the development of the interfaces between the antennas and the system, and between the system and the SDP.
Costing

Construction Expenses

For the correlator we require a 456 node cluster, each node at least 2 Xeon processors, with dual channel infiniband. Comparable machines were recently purchased from Cisco Systems (UC servers) at an original quote of ~10K AUD ea. Which would give an approximate costing of 5.9M AUD. Although the specification does not quite match this is a good indicative cost of an enterprise class server of the required capability (without discounts). In this strawman design the X-servers require an Infiniband card and 2 GPU.

A K series GPU is approximately 3K AUD at the moment. We would need ~912 of them at 2.7 M AUD in total. Once again, in 3 years K series GPU will be a fraction of this cost – but we will likely use Volta – which will likely remain at this price point.

A recent costing from Mellanox for a 500+ port 56Gb/s infiniband switch, recently purchased by Cambridge University was 0.5M GBP it draws 10kW.

Personnel

As discussed in the previous section, by the time of actual construction it is envisaged that the correlator software itself will already be constructed by the end of preconstruction. Of course development of interfaces both at input and output will likely continue, but this effort will be minimal. It may be required by the project office that some software engineering and software project management tasks be performed to ensure maintainability above that produced during research and development.

I will assume 2FTE for 1 year for assembly and maintenance of the software elements.

I will also assume 2FTE for systems management.
**Capital Cost Summary**

(Indicative: See spreadsheet for invertible cost model)

<table>
<thead>
<tr>
<th>Capital Costs</th>
<th>Meuro(2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td>4.7</td>
</tr>
<tr>
<td>Interconnect</td>
<td>1.0</td>
</tr>
<tr>
<td>Total</td>
<td>5.7</td>
</tr>
</tbody>
</table>

2016+: 456 servers for combined F and X stage, 912 GPU, 500+ input 56Gb/s FDR infiniband interconnect.

**Construction Costs**

We have been asked to include power infrastructure estimates. Which is 0.9 MEuro. This number is based upon a total power draw of ~450kW and 2 Euros per Watt.

Also 2FTE for 1 year (0.2M) for software integration.

**Operating Costs**

<table>
<thead>
<tr>
<th>Operating Costs</th>
<th>2016+</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 FTE</td>
<td>0.2</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.5</td>
</tr>
<tr>
<td>Power</td>
<td>1.3</td>
</tr>
<tr>
<td>Total</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Therefore in summary the construction cost if built today would be 5.9 MEuro (plus power infrastructure of 0.9M), and the operating costs are 2 MEuro per year.

**Technology Refresh Costs**

It is likely that a refresh of GPU technology could be performed every 2 to 3 years, for both compute purposes and for power consumption. Every 100W
drop in power load from the GPU would save 50kW in total and thus 50Keuro/year. At 3000Euro per GPU this would cost 2.7MEuro.

However a considerable saving will be made when a single GPU can perform in excess of 7Tflop then the number of servers can be dropped by a factor of 2.

**Scaling to SKA2**

An expansion of collecting area and increase in baselines can be accommodated as the compute capability of the GPUs improves. There is considerable overhead in input bandwidth and this system could probably ingest and corner turn twice as many stations easily. But each GPU would need to operate at 12 Tflop which will occur around 2018, well on the timescale of SKA2.
Feasibility White Paper
SKA Central Signal Processor

1.3.2.2 SKA.TEL.CSP.CBF-STG1.MID.PIP

A high performance, low-power, scalable solution for the SKA1 MID dish array CSP Correlator and Central Beamformer, based on the PowerMX platform.

Authors:
Brent Carlson, National Research Council of Canada
Francois Kapp, SKA South Africa, South Africa
David Stevens, MDA Corporation, Canada
Rob Halsall, STFC Rutherford Appleton Laboratory, UK
Andrew Ensol, Auckland University of Technology, New Zealand
Giovanni Comoretto, INAF, Italy
Manqing Wu, China Technology Research Group Corp No. 38 Research Institute, China

Date: May 3, 2013
Version: Release
Table of Contents

INTRODUCTION AND CONTEXT .................................................................................................................... 4

SIGNAL PROCESSING DEFINITION AND KEY REQUIREMENTS .................................................................... 6

TECHNICAL DESCRIPTION............................................................................................................................ 7
  PowerMX Simplified Description ................................................................................................................. 7
  SKA1-MID.CBF Technical Description ........................................................................................................ 9
    F-Part .......................................................................................................................................................... 9
    X-part ....................................................................................................................................................... 10
  External/Separate PSS-BF ............................................................................................................................. 13
  External/Separate PST-BF ............................................................................................................................. 14
  Transient Buffer ......................................................................................................................................... 14
  Pulsar Phase Binning .................................................................................................................................. 15
  Additional X-part Pipeline Processing ........................................................................................................ 16
  Cooling Strategy .......................................................................................................................................... 16
  Other Devices and Methods ......................................................................................................................... 16
  Test and Verification Correlators ................................................................................................................ 16
  Risks and Mitigation .................................................................................................................................... 17

ROUGH DEVELOPMENT SCHEDULE ........................................................................................................ 17

PERFORMANCE ANALYSIS ......................................................................................................................... 19

COST and POWER ANALYSIS ..................................................................................................................... 21
  Key Cost and Power Estimates and Parameters .......................................................................................... 21
  BOTTOM LINE—Development Costs (NRE) ............................................................................................... 23
  BOTTOM LINE—OPTION 1 Total Cost Estimates ..................................................................................... 23
  BOTTOM LINE—OPTION 2 Total Cost Estimates: BOTTOM LINE .......................................................... 24

VIABILITY AND SCALABILITY TO SKA2 .................................................................................................... 25

CONCLUSIONS ............................................................................................................................................. 27


Table of Figures

Figure 1  SKA1-MID Central Signal Processor Context Diagram. ................................................................. 4

Figure 2  Simplified block diagram of the SKA1-MID CBF, extracted from the Baseline Design. .............. 5

Figure 3  Simplified PowerMX block and data flow diagram. Each lane of each highway (M1-M4) contains 12 hi-speed serial paths. .................................................................................................................................. 8

Figure 4  F-part pipeline processing. Each highway handles 4 antennas, with the concatenation of 3 PowerMX boards into a triple-board blade. .................................................................................................................................. 9

Figure 5  Triple-PowerMX blade for F-processing in the correlator. ............................................................ 10

Figure 6  Dual-PowerMX blade for correlation and transmission of data to the PSS-BF, or performing that function in a mutually exclusive fashion with the correlator. .............................................................. 11

Figure 7  Shelf side view showing triple-PowerMX F-blades plugging in from one direction, and dual-PowerMX X-blades plugging in from the other direction. The entire correlator fits as a single shelf in one rack (shelf is ~35” W x 12U(21”) H x 47” D). If the correlator, PSS-BF, and PST-BF are mutually exclusive in use, this is the entire SKA1-MID.CBF, although 20 triple-PowerMX blades are required to handle the complete (mutually exclusive) PSS-BF (16 blades) or PST-BF (20 blades) load. Data flow on the dual blade is actually the opposite direction of the triple (it faces the opposite direction shown), however this is easily handled as the PowerMX motherboard is direction agnostic and different mezzanine cards wired for the opposite direction can be installed.................................................................................................. 12
INTRODUCTION AND CONTEXT

The Square Kilometer Array Central Signal Processor (SKA-CSP) is the central processing facility for the telescope. There are 3 telescopes planned to be built to form Phase 1 of the SKA (“SKA1”), and this Feasibility White Paper addresses the SKA1-MID Correlator Beamformer (CBF) for the combined dish array consisting of 64 MeerKAT dishes and 190 SKA dishes located and operated in the Karoo Desert in South Africa.

A context diagram of the CSP, extracted from the SKA Request For Proposal (RFP) Baseline Design (SKATEL-SKO-DD-001, Revision 1, 2013-03-12) is shown in Figure 1 below:

A view into the SKA1-MID CBF, also extracted from the Baseline Design is shown in Figure 2. This White Paper does not explicitly address the Optical Distribution, MeerKAT Interface Box, Local Timing and Clock Distribution, or Local Monitor and Control but it is noted that these are important components of the CBF and must be addressed in due course in the design of the CBF during its development stages.

Local Timing is key of course, and provides the CBF with “perfect virtual time”, a time reference synchronized to the central array timing reference against which every part of the CBF synchronizes its operations and calculations to. This timing reference does not have the coherence requirements of analog-domain receivers and samplers, but simply needs to be good enough to be within data buffer depths used throughout the CBF, and “discrete-time synchronized” to allow for accurate timekeeping and time stamping of data at various points in the Element.

This White Paper proposes how to implement CBF functions for SKA1, showing extensibility to SKA2, using the “PowerMX” common platform “board” that may be packaged and used in a variety of ways.
However, this is notional, and further refinement and moulding will likely occur with the goal of developing and using an “in-SKA” COTS-like common platform solution applicable to the full range of SKA signal processing problems, and therefore benefiting from shared and coherent development effort both in the platform and in the layers of firmware and software residing on it.

**Figure 2** Simplified block diagram of the SKA1-MID CBF, extracted from the Baseline Design.

The “Buffer” in Figure 2 is a Delay Buffer (implemented in, effectively, a FIFO memory structure) with active delay tracking to compensate for observed source wavefront arrival time mismatch at each antenna, as well as absorb differences in data transport time via fiber to the CBF. Each buffer contained within each “antenna handler” in the CBF must be sized for the worst case; with a 200 km aperture, associated wavefront and fiber delays, and the required sample rate, it is quite large. At the output of this buffer a particular photon, received by all of the antennas, will be present at approximately the same time relative to the local timing reference. The word “approximate” is used because the memory can only compensate for wavefront delay—or more correctly, a model of the wavefront delay—to within +/-0.5 samples at the digitizer sample rate. Further “very fine delay” correction, as well as earth-rotation phase correction” is performed after the channelizer with complex mixers performing a “phase-delay” correction to “fringe stop” before correlation and beamforming. This ensures that the peak of the cross-correlation function stays centered within the correlator’s delay window (essentially the size of the FFT), with residuals only due to differences between the delay model and the actual delay.

The “Channelizer” is a filterbank that “splits” the wideband signal into narrower sub-bands before cross-correlation and beamforming. There is a very high probability this will be a “poly-phase FFT” filterbank as it has superior channel purity compared to straight FFT methods, although at some additional cost of processing. The “Corner Turn” re-arranges the data so that in the “Correlator and Central BFM” all of the necessary correlation and beamforming functions can be performed. “Beam Data” is output to the NIP (Non-Imaging Processing) portion of the CSP (pulsar searching and pulsar timing) noting that each beam is on a different delay center on the sky within ~the half-power points of the antenna primary beam (+/-0.5 degrees), and the CBF must perform necessary operations to do so. “Integrated u-v” are cross-correlation coefficients or “visibilities” required by the Science Data Processor (SDP) for imaging.
SIGNAL PROCESSING DEFINITION AND KEY REQUIREMENTS

The SKA1-MID CBF performs all cross-correlation and central beamforming operations required for the SKA1-MID telescope.

The key SKA Element-level requirements of the CBF for SKA1-MID are as follows:

- Receive digital data, likely via fiber\(^1\), from 254 dishes of the combined array. This data is digitized to 8-bits per sample, real sampling, dual-polarization, at nominally 2.5 GHz bandwidth per polarization, therefore requiring \(\sim 80\) Gbps of data transport from each antenna. The exact RF frequency that this signal represents is of relatively small importance to the correlator/beamformer—it must be known and is required for some calculations, but there is no technology-specific requirement the RF frequency places on the correlator/beamformer. Similarly for polarization, from the correlator’s perspective, it is a data stream label.

  The key issues that drive the correlator/beamformer implementations and technologies are bandwidth, array size and number of elements (determining internal buffer depths, technologies, and techniques), required channelization and associated attributes, integration times and modes, and the RFI environment and associated RFI mitigation requirements.

- Correlate every antenna pair in the array, including with themselves. For 254 antennas, there are \(254^2/2 = 32,259\) cross-correlations plus 254 autocorrelations. Each cross-correlation produces the integrated cross-power spectral response of the pair of antennas at 262,144 channels per polarization product, with post-correlation RFI excision, and channel averaging to 4096 channels\(^2\) per polarization product. Two integration times, 1.6 sec (and need not be smaller) for baselines < 10 km, and 0.08 sec (and need not be larger) for baselines > 10 km, are required. Note that baselines < 10 km can be formed from antennas outside the core. Correlation coefficients are “visibilities” that are transported to the SDP for further processing into images of the observed radio source. It is not (believed to be) explicitly stated, but if some pre-processing of visibility data can be performed to ease the load of the SDP, it is highly desirable.

- Two beamformers are required. The “PSS-BF” (Pulsar Search BeamFormer) and the “PST-BF” (Pulsar Timing Beamformer). The PSS-BF is to produce 2100 combined polarization beams across 300 MHz at 20 kHz spectral resolution, to be used by the CSP.NIP.PSS Sub-Element to search for pulsars. In this case, only the \(~125\) antennas in the central \(~900\) m core of the array are used for beamforming. This restriction is necessary so as to have a balance between beam sensitivity (proportional to the number of antennas coherently\(^3\) added together) and beam size (inversely proportional to the combined aperture of the antennas coherently added together) to be able to survey vast stretches of the sky in a reasonable time \(<\ll\) lifetime of a pulsar astronomer).

  The PST-BF is to produce 10 dual-polarization beams across 1.4 GHz of bandwidth, at a minimum spectral resolution of 10 MHz (100 nsec timing resolution) using the entire SKA1-MID array with an extent of \(\sim 200\) km. In this case the formed coherent beams are very small but are

---

\(^1\) Ultimately via long-haul fiber transmission methods, the specifics of which are under the scope of the SADT Element.

\(^2\) This is tentative and requires further science definition and confirmation. EVLA Memo #64 (R. Perley) conclusions seem to indicate that the desired FOV of 0.83 deg\(^2\) at L-band (Baseline Design, Table 9) may not be achievable at this channelization. See http://www.aoc.nrao.edu/evla/geninfo/memoseries/evlamemo64.pdf

\(^3\) Incoherent beamforming is possible, but not addressed in this White Paper.
targeted to specific radio sources, rather than having to survey large stretches of the sky. Absolute precision and accuracy of time stamping of data to the NIP.PST processing engines are important to ensure reliability and correctness of derived results. Some of this is within the control of the CBF in the way that time is accounted for as data transports through the sub-element, and some of this relies on fundamental array timing and issues external to the CBF.

- It is yet to be decided if the correlator, PSS-BF, and PST-BF must be operating concurrently, a matter to be decided by science operational requirements, modes, and cost. The correlator is needed to determine the residuals to form coherent beams, however, if the system (the entire system including antennas, receivers, data transport etc.) is stable enough then it should only be necessary to correlate every so often and otherwise be forming PSS and PST beams. Both such scenarios (OPTION 1 and OPTION 2) are presented to deal with these possibilities in the technical description, performance analysis, and costing sections.

- As the SKA1-MID CBF is located at a remote site in the Karoo desert of South Africa, due attention to product/quality assurance and RAM (Reliability, Availability, and Maintainability) is required since, if the CBF is down, the entire telescope is down. Generally it is accepted that one or a few antennas will be down at any given time, and that therefore, a few baselines’ results will be unavailable for imaging. This effect naturally transfers to the CBF, and therefore cost-effective reasonable measures are needed (to be defined in detail) to ensure operational availability. It is a requirement that the effects of failed/failing parts of the CBF are flagged and don’t contaminate the results, both in delivered visibilities to the SDP, and beams to the NIP engines within the CSP. In some cases, such as internal calculation errors due to a failed piece of silicon, it is only possible to determine faults with off-line testing, or with additional hardware performing “rotating” redundant calculations. 1+1 or N+1 redundancy must be used where necessary to reduce the probability of single point failures taking down the telescope. It is possible, but unlikely, that a fully redundant hot-standby (or cold-standby) CBF be constructed. Further requirements and operational definition is required.

Further requirements are contained within the RFP Baseline Design, and clarified for the purposes of these feasibility White Papers in the “Canada-led SKA CSP Consortium – Technical Proposal White Paper Metrics and TOC – V1.1” dated April 22, 2013.

**TECHNICAL DESCRIPTION**

The SKA1-MID.CBF implementation proposed is native custom hardware-centric, based on use of the “PowerMX” common platform, but could accommodate CPUs and/or GPUs (or other devices) at any point along the processing pipeline if required. PowerMX can be used in isolation in rack-mount “pizza boxes”, as well as concatenated and used in pizza-boxes, mounted and used as blades in shelves, or stacked and concatenated in larger pizza boxes. Combined with huge I/O resources, PowerMX is therefore a highly scalable platform for use across the SKA, both within the CSP and other Elements, with reasonable scalability to SKA2.

**PowerMX Simplified Description**

A simplified diagram showing the main data flows and chip (FPGA, ASIC, or other) sites on the board is shown in Figure 3 (a draft detailed technical description has been developed—it is open source and available on request).
Figure 3  Simplified PowerMX block and data flow diagram. Each lane of each highway (M1-M4) contains 12 hi-speed serial paths.

There are 4 mezzanine sites with high-performance connectors to the PowerMX motherboard and I/Os. Each mezzanine site (90x90 mm) can host a single large device (45x45 mm, and larger) and memory modules (4 shown for each), several smaller chips, or whatever circuitry that fits and is required for the task. If motherboard I/O is not required, the mezzanine card can be extended in one dimension (horizontal in the above figure) to provide additional capability and perhaps custom I/O.

Each site has associated I/O highways known as “M1-M4” (“motorway”—“freeway” sounds better, but is not such common terminology outside North America) in the figure; each highway contains 4 lanes, and each lane contains 12 hi-speed data paths up to 28 Gbps each (to be verified; pending further study, possibly up to 40 Gbps each). Therefore, each highway contains 48 hi-speed data paths, with direction determined by the mezzanine card and I/O modules. In total, along each edge, there are 192 hi-speed paths or 384 in total.

Each lane at each edge of the board routes to a high-speed 12-signal connector, and the connector can host a desired pluggable SNAP-12-style (or whatever is required and fits) transmitter or receiver module, or 2 or more connectors at each edge can be ganged to plug in a user-defined I/O module extending a user-defined distance away from the edge of the board containing user-defined circuitry. There are also 8 lanes between adjacent sites, with direction determined by the mezzanine modules. These 8 lanes allow for full intra-board meshing by connecting/repeating through nearest-neighbour mezzanine cards.

There are also dual 10GBASE-T connections from each site to motherboard RJ-45 connectors providing 8x10G auxiliary I/O paths (not shown).
White Paper: 1.3.2.2 SKA.TEL.CSP.CBF-STG1.MID.PIP (NRC,SKA-SA,MDA,STFC-RAL,AUT-NZ,INAF,CETC)
VRelease May 3, 2013

For M&C, there are 1G or 10 G links from each mezzanine site to a central M&C SoC card (System on a Chip, also not shown), and 1G or 10G Ethernet on fiber or copper from that SoC to the outside world.

Also not shown are clock I/Os with ADC-quality clock delivery to each mezzanine site for radio astronomy applications.

The motherboard is “17” x 7”, and, depending on power and cooling, can fit within a 1U (1.75”) pizza box or other carrier. For this application, mounting PowerMX boards on metal substrates to form multi-board plug in blades seems to be the most attractive.

**SKA1-MID.CBF Technical Description**

Using PowerMX as a building block unit, the SKA1-MID.CBF is proposed as follows.

**F-Part**

The F-Part of the correlator requires the transient capture buffer, Delay Buffer, and the poly-phase filterbank. In analysing requirements and device capability in the 2016 timeframe it was decided that one PowerMX highway would handle the data from 4 antennas, at 80 Gbps per antenna (using 320 Gbps of 3-lane 360 Gbps capacity). Further, it was found that the Delay Buffer memory requirements exceeded on-chip memory by quite a large margin and so for this proposal, there will be 3 PowerMX boards concatenated into a “triple-PowerMX” blade to handle the necessary processing. One highway of such a blade is shown in Figure 4:

![Figure 4 F-part pipeline processing. Each highway handles 4 antennas, with the concatenation of 3 PowerMX boards into a triple-board blade.](image)

The first chip in the pipeline contains the Delay Buffer for 2 antennas (160 Gbps), implemented with 4 (mezzanine mounted) DDR3 modules, one for each polarization of each antenna, requiring 10 Gbytes/s of I/O performance to each single-port module. This same buffer is used as a transient capture buffer. The other 2 antennas’ data are passed (repeated) straight through to the next chip in the pipeline performing the same function. In addition, for the case where the Pulsar Timing beamformer (PST-BF) is external to the correlator, these chips perform the required coarse filterbank and channel-to-channel isolation, and ship the data out the other direction using the spare lane.

The full triple-PowerMX blade configuration, showing data flows for M1 are shown in Figure 5. With 4 antennas processed in each of 4 sites/highways, a single triple-blade can process 16 antennas. 16 such blades—providing 256-antenna capability—are then plugged into a backplane consisting of 4 identical single highway backplanes for meshing to the X-part of the correlator. Backplane drivers attached to the right-most blade are capable of driving 10 Gbps over 40” per hi-speed data path, which this design uses.
throughout (i.e. it is not using 28 Gbps capacity). See the PowerMX technical description for more details.

Figure 5 Triple-PowerMX blade for F-processing in the correlator.

The final (right-most) site/blade performs the fine filterbank function for the 4 antennas, as well as the first stage mesh via nearest-neighbour intra-column connections, and ships the data off to the backplane mesh. With 32 serial lines active, a total of 96 serial transceivers are required for each mezzanine site’s device to handle all of these connections. This may require a more expensive FPGA than budgeted, although the blended cost with the delay function FPGAs might average out to the cost-projected mid-range 20 nm device. Fine filterbank data is used for correlation and for Pulsar Searching beamforming (PSS-BF), whether within the correlator blades (analysis indicates that the correlation and PSS-BF problems are about the same size) or within an external beamformer. 8-bit data is transmitted; 4 bits are used for correlation, whereas the full 8 bits are used for beamforming.

X-part

Plugging into the other side of the backplane are 16 dual-PowerMX blades sized for the correlation problem. Data flow is the opposite direction to that of the F-part from the PowerMX board’s point of view (backplane-to-front panel), requiring different mezzanine boards wired for the different direction. These could have different chips (FPGAs/ASICs) if required to suit the task, and one or two additional boards could be added to the blade if more or advanced SDP pre-processing were required. The dual-PowerMX blade showing data flows for M1 is shown in Figure 6:
Correlation coefficients can be shipped off to the SDP via auxiliary 10GBASE-T data paths (two per PowerMX mezzanine site), or by using the spare lane, providing 120 Gbps (and even higher if necessary) capacity per highway. The fine filterbank data gets pipelined and goes out the front of this blade for connection to the external PSS-BF, or the FPGAs are re-programmed to do beamforming, with beams transmitted out the same path. In this latter case, intra-column connections (not shown in Figure 6, but shown in Figure 3) can be used as the first-stage mesh, with an external fiber or active switch providing the second stage to yield finally “gathered” beams. Note that only 300 MHz of bandwidth on 125 antennas must be beamformed or transmitted to the PSS-BF so there is huge over-capacity in the PowerMX data paths to handle this small amount, even if there are 2100 beams (1.26 Tbps total; spare lane aggregate capacity across 18 blades is 86 Tbps). Processing with 16 dual blades only allows for ~1527 beams, though.

Figure 7 is a diagram showing a side view of the dual-entry shelf. The shelf is ~35“ W x 21” H x 47” D and can fit in one (probably custom) rack, with fans below and above providing cooling airflow top to bottom. The dual blade component side faces the opposite direction shown.
Figure 7 Shelf side view showing triple-PowerMX F-blades plugging in from one direction, and dual-PowerMX X-blades plugging in from the other direction. The entire correlator fits as a single shelf in one rack (shelf is ~35” W x 12U(21”) H x 47” D). If the correlator, PSS-BF, and PST-BF are mutually exclusive in use, this is the entire SKA1-MID.CBF, although 20 triple-PowerMX blades are required to handle the complete (mutually exclusive) PSS-BF (16 blades) or PST-BF (20 blades) load. Data flow on the dual blade is actually the opposite direction of the triple (it faces the opposite direction shown), however this is easily handled as the PowerMX motherboard is direction agnostic and different mezzanine cards wired for the opposite direction can be installed.
If the correlator, PSS-BF, and PST-BF are mutually exclusive in use, one back-to-back shelf in one rack is the entire SKA1-MID.CBF. In this case, a 20-slot backplane with X-part triple-PowerMX blades to handle the PST-BF load are required. This provides 5 blade overcapacity for the PSS-BF and double over-capacity for the correlator (something that may be needed anyway, see below).

4-bit correlation in the X-part is planned to minimize computing and power, but still have reasonable dynamic range and low sensitivity loss. The possibility of a channel saturating with RFI in a time variable fashion is ever present. The plan is to deal with this in the F-part as follows. For a particular packet containing a time burst for a particular frequency channel, power is calculated over the same interval. These power values are available for readout by a controlling CPU (i.e. the PowerMX M&C SoC, or even a mezzanine-mounted microcontroller), which then sets a threshold common across all frequency channels wherein packets are marked as “ok” or “bad” en-route to the correlator based on whether their calculated power is below or above the threshold respectively. This avoids having to set a 4-bit quantizer threshold different for each channel and should handle time and frequency variable RFI.

Back of the envelope calculations indicate that each FPGA, in the 16 x dual-PowerMX blade configuration, for a mid-size 20 nm FPGA should\(^4\) have the on-chip memory (~31 Mbits), and bandwidth to off-chip DDR3 memory (4 SODIMMs at 12.8 Gbytes/sec each) for the full specified frequency resolution at a ~200 msec integration time. If a data invalid counter is used for each spectral channel (applicable to all 4 pol’n products), then the data transfer rate never exceeds that required for correlation coefficients since if a time-burst data packet is flagged as invalid, no correlation is performed and no data transfer (aside from updating the invalid counter) is necessary for that channel. As the minimum required integration time is 80 msec on ~1/2 the baselines it would seem not to be possible with DDR3 memory speeds. If DDR4 memory speeds are not capable, using a 20-slot triple-PowerMX blade (with DDR4 memory) configuration may be necessary, splitting the total correlation problem across 240 FPGAs, rather than 128, increasing the capital cost of the X-part from 16 x $30,124 ~\approx \$0.5M, to 20 x $43,936 ~\approx \$0.9M, and power consumption from 16 x ~620W ~\approx 9.9 kW to 20 x ~1088W ~\approx 22 kW, or ~$12k per year in power. If necessary, a quad-PowerMX blade, 20 slots, with 320 FPGAs could be used increasing the X-part cost further to ~$1.1M (~$0.6M over the 16-slot base) but with no serious system design change. If PST-BF functions are mutually exclusive with the correlator, this could be a moot point as 20 triple-PowerMX blades appear to be required for that function anyway. See tables on pages 19 and 21 for further information on cost and power number estimates.

As the same data going to the correlator finds its way to the PSS beamformer, 8-bit data values are transmitted, with the X-part choosing some upper portion for 4-bit correlation. Here, of course, proper sign extension handling must be in place and is another area to detect saturation of each particular frequency channel. Further study is required to determine how per-channel voltage sample RFI flagging might be employed.

**External/Separate PSS-BF**

In the case where the PSS-BF is to be concurrently operating with the correlator, a separate shelf in a separate rack\(^5\), constructed in a similar manner as for the correlator is proposed. In this case, no re-meshing of the data occurs in the correlator. Plugging into the front of the shelf are 15 triple-PowerMX blades that implement the beamformer. They plug into a backplane that re-meshes them to 6 single-

---

\(^4\) Feasibility will, of course, come down to the details of the implementation.

\(^5\) Could be in the same rack, but as will be seen in the power estimation section, total power is likely an issue.
PowerMX blades plugging in from the other direction. These blades perform final data re-arrangement and formatting ready for transmission to the external NIP.PSS engine. At 300 MHz BW, 2100 beams, 1.26 Tbps of beam data output is generated. Six PowerMX boards have 11.5 Tbps aggregate output capacity so there is no problem there.

Note that due to fine frequency channels and the small 900 m aperture of the array being beamformed, beamforming is entirely phase-delay (sin/cos LUT and complex mix) operations. In this case, the worst-case differential delay between the correlator delay tracking center, for a 0.5 deg offset, amounts to ~26 nsec, allowing for a 2.3 MHz maximum channel width before exceeding +/-22.5 deg of phase error at the edge of the channel. This mixer takes care of delay tracking very fine delay, earth-rotation phase correction, and beamforming.

External/Separate PST-BF

In the case where a separate PST-BF is to be concurrently operating with the correlator, a separate shelf in a separate rack is used. In this case, data sources from the front panels of the F-part of the correlator (see previous sections) and routes into 20 triple-PowerMX blades for processing. This routing must include meshing so that going into each highway (or at minimum each blade) are all of the 254 antennas at a fraction of the bandwidth needed for beamforming. This meshing can be accomplished with the use of COTS 12-fiber MTP-to-LC breakout boxes, cables, and reverse, mounted in the PST-BF rack, or a 19” rack mounted beside it. At 1.4 GHz/poln, 254 antennas, and 8-bit sampling, the total data rate into the PST-BF is 11.4 Tbps. Aggregate capacity out of the F-engine, using the spare lane as shown in Figure 4, is 7.68 Tbps; aggregate capacity into the PST-BF engine is 28.8 Tbps easily providing this ability.

A single-PowerMX blade is plugged into the other side of this smaller shelf, the 20 triple blades meshing through the backplane to this board. This board performs final data gathering and formatting for delivery of the final 10 beams dual-pol’n to the external NIP.PST engine. Total data rate in is 448 Gbps, with a single board having aggregate 1.92 Tbps, easily handling this function.

Note that due to coarse frequency channels (10 MHz) and the large aperture of 200 km of the array being beamformed, beamforming must be true delay beamforming. For each 10 MHz sub-band (oversampled by a factor of 1.2, so really 12 MHz must be handled) of each beam it consists of a 128-deep memory (5.82 usec maximum differential delay to the correlator delay center, 0.5 degree offset), a complex FIR filter (32 taps each for I and Q; if 64 taps are required, 39 triple-PowerMX blades are required for beamforming) made of DSP blocks⁶, followed by a complex mixer (with sin/cos LUTs). This assumes complex, dual-pol’n beams are the final output.

Since data is tapped off from the antennas virtually immediately after the main Delay Tracker, amplitude and/or phase distortion or absolute timing uncertainties in the data path are minimized for this ampl/phase/time sensitive application as the signal goes through fewer processing steps compared to being tapped off of fine filterbank data.

Transient Buffer

The Delay Buffers (shown in Figure 4 and implemented as external DDR3⁷ memory) provide transient capture capability. To avoid interrupting delay tracking, once a trigger is provided data must be read-out and pipelined to some external memory “place”. Each external DDR3 memory is being read at 5

⁶ If a FPGA LUT-based FIR is used it might be possible to increase the number of taps without increasing the number of blades, but in this case ping-pong FIRs are required to avoid disruption of data flow during delay tracking since LUT-base coefficients take longer to update than straight-up multiplier coefficients.

⁷ And, of course, likely DDR4 by the time the design is implemented.
Gbytes/sec, amounting to at least 40 Gbps if not a bit higher each. Existing PowerMX extra lane capacity is not sufficient to handle this, as for the entire highway (i.e. ¼ of the board), the data rate is 320 Gbps (4 antennas, at 80 Gbps per antenna). This burst only lasts until the memory is empty and could, therefore, be buffered and read out more slowly, however, it requires a memory as deep and fast as the Delay Buffer to do this. Also, what if there are multiple rapid triggers? It seems one would want to keep all of that data.

What to do?

Only the largest and most-expensive FPGAs have 8 DDR3 memory controllers and, using the 2 10GBASE-T auxiliary data paths per mezzanine site, a ~1/8 duty cycle would be supported.

A better way is to put two Avago “micropod” devices on each of these Delay Buffer/transient capture mezzanine cards (providing 120 Gbps aggregate capacity each), and route fiber from these to ports on the front panel. Then, all of these from every antenna can be routed to an external transient capture buffer, which could even be a single separate shelf made of 16 single-PowerMX blades stacked with memory and solid-state hard drives for longer-term/higher-capacity data storage.

The possibilities are endless here and can’t be firm’d up until real requirements are defined.

**Pulsar Phase Binning**

A requirement for the correlator is to provide 50 pulsar phase bins with bin widths as narrow as 1 millisecond. In this operating mode, the correlator’s integration time is synchronized with the model of a rotating astronomical source (pulsar) that generates radio pulses on a repetitive basis. Thus, the correlator is essentially taking 50 pictures of the pulsar during its rotation, and continues to integrate into the same bins such that an integrated picture is taken. Not only that, but to make effective use of each bin, it is highly likely that bins should be able to be “bunched” during the pulsar on time, with only a few, and longer integrations, on the pulsar off time.

Each of these pulsar phase bins must be kept in separate memory, and on each 1 millisecond, dumped and integrated with it. Given the large number of baselines (up to 32,768) and frequency channels (262,144 x 4 pol/NS/(16x8) corr chips) ~ = 8192 channels/baseline x 8 bytes per visibility, yields ~ 32,768 x 8192 x 8 ~ 2.1 Gbytes per integration, for a total of ~ 100 Gbytes for 50 bins. For now this is not the case as the White Paper TOC indicates that one can tradeoff frequency channels for phase bins, so it really is only 2.1 Gbytes, easily within external DDR3 module capacity (4 per chip are notionally planned). Only 8192/50 ~ = 164 channels must be saved. Each correlator FPGA/chip must still integrate ~ 20 MHz of bandwidth, on all baselines, for a spectral resolution of ~ 122 kHz.

In thinking about how each correlator chip will work, it will get packets from every antenna for a “time-burst” of a particular frequency channel, and for 1 millisecond binning that time-burst contains 1 millisecond of data for that channel. Since there are 122 channels to be correlated and saved within that 1 msec, then it must be done every 1 msec/122 ~ = 8.2 usec. Therefore, 32,768 baselines x 8 bytes per visibility must be read and written to external DRAM every 8.2 usec. That is a data rate of ~ 32 Gbytes/sec, one way, or 64 Gbytes/sec to read existing data, and integrate and write it back. As 4 DDR3 SDRAM modules are planned for each chip, each one capable of ~ 12.8 Gbytes/sec data transfer (see micron MT16JTF51264HZ-1G6), that is a total capability of 51.2 Gbytes/sec, not quite meeting the requirement. If the number of channels is reduced to ~ 64, then the time to integrate—dependent of how many samples go into that integration—is 1 msec/64 ~ = 15.625 usec, requiring ~32 Gbytes per

---

8 DDR4 will likely be available by the time of implementation, no data sheets could be found on the Micron website.
second read/write rate, which should be possible. Therefore 262,144/100 \approx 2600 \text{ channels per pol'n product should be possible providing a spectral resolution of 2.5 GHz/ 2600 channels } \approx 1 \text{ MHz.}

\textbf{Additional X-part Pipeline Processing}

It should be pointed out that the X-part dual or triple PowerMX blades could be augmented with an additional pipeline stage before output by mounting another PowerMX motherboard with computation-specialized chips on them such as might be required for pre-processing or co-processing before shipping data off to the SDP. Some of this might be more intelligent RFI mitigation, gridding etc. Further investigation is required. Likely a quad-PowerMX blade depth is the limit in this regard.

\textbf{Cooling Strategy}

With all of those blades active in a (correlator, PSS-BF, PST-BF, mutually-exclusive use) combined 20-slot rack with dual triple blade shelves, total power dissipation is estimated at \simul\text{ 40 kW within the rack. This seems like an impossible situation, however this rack has a larger cross-sectional area of } \simul\text{ 35" x 54" or } \simul\text{ 13 ft}^2 (\simul\text{ 1.2 m}^2). \text{ By comparison, a 19” server rack might have a cross-sectional area of } \simul\text{ 2.1 ft}^2 (\simul\text{ 16” x 19") or a factor of } \simul\text{ 6 lower. 40 kW in a single “shelf-plane” in this rack is equivalent to 6.6 kW power in a 19” rack. This should not be a problem with air-only, “rack-as-a-duct” cooling with blower fans below and above the shelf and airflow from bottom to top. In this case cool air is provided from the raised floor and evacuated to the top of the rack. No closed-cycle rack cooling systems or liquid cooling schemes should be required.}

\textbf{Other Devices and Methods}

This White Paper provides a baseline implementation description based on FPGA technology. However, other chips may have architecture and performance that is comparable or better than FPGAs for a particular task in the processing pipeline. This can entirely be accommodated by the PowerMX design, where each board, and indeed each mezzanine site, can be populated by different chips.

Some examples of these other chips are the Adaptive Array Systems chip being developed in the UK, the CETC DSP chip from China, and perhaps others not yet known. Additionally, the right FPGA for the job, from the right vendor (Xilinx, Altera, others?) can be optimally chosen for performance and to hopefully increase competition to save capital costs. It is not a free-for-all though, as each chip of course must be integrated into the processing chain, contain firmware or software, and M&C drivers all of which must be designed and supported and so choosing something different should be done for good reason. Still, it opens up the possibility of using the best chip for the job to save money and power in the long run.

A different F-to-X corner-turner/mesh mechanism could be used if deemed appropriate or necessary. One method is to use active 10G switches to provide switching capability, noting that it is likely only necessary to provide switching within the plane of each PowerMX highway (e.g. M1), and not across highways. This limits the interconnect degrees of freedom and might actually be feasible and desirable if requirements demand it. The CSIRO passive fiber mesh described in the next section might also be employed.

\textbf{Test and Verification Correlators}

A test correlator for the purposes of testing antennas could be lashed together using a single rack-mount pizza box PowerMX board as outlined in the PowerMX technical description since the board contains all necessary connections for correlation processing. If that packaging isn’t used, a single or dual-PowerMX blade can be jury-rigged rack-mounted to do the same thing. Probably an existing fully-tested offering using MeerKAT correlator parts is more cost-effective.
A verification correlator for the purposes of testing correlator code and antennas on the SKA site can be lashed together with a dual-PowerMX blade, wherein the F-part is performed on one PowerMX board, and the X-part on the other. To develop and shake-down final code, the F-part can be done piecemeal as each part does not depend on the number of antennas. However, the X-part is best handled initially with a simulated packet generator for the full number of antennas, and then finally with a full system across a backplane emulating the full number of antennas.

**Risks and Mitigation**

Risks associated with the PowerMX board are described in its technical description document. Risks (and mitigation/fallback positions) for the implementation described are believed to be:

- Mesh backplane cannot achieve desired performance of 10 Gbps over ~35”. In this case, the fallback position is to use an external fiber mesh, based on the CSIRO (Hampson et. al. see: [link](http://www.cyberska.org/pg/publications/ham238/read/24827/advancements-in-beamformer-and-correlator-optical-backplane-technology)) passive fiber mesh product. Here, the F-part and X-part could be separated into two racks and back-to-back across a rack aisle from each other. 16 planes of mesh, each one containing 3 such constructs are built to allow 2 fibers from each lane of each F-part to route to each X-part lane for the same highway, thereby duplicating what the backplane would do. This multi-plane mesh could be mounted in an appropriate carrier above the aisle between racks, with fiber routed into the back of each back-to-back facing rack. Alternatively, the mesh could be mounted in a rack in-between the F-part and X-part racks in the same aisle.

- Performance estimates do not properly take into account implementation details that might mean more devices (and therefore boards and racks) must be used. If this happens it means a higher cost, but as PowerMX is highly scalable, it should not cause any “show-stopper” situations. The maximum blade length is probably 4 PowerMX boards (~32”...a pretty large blade!) to fit within the backplane reach of 40” described, however if this is not possible, different external meshing and connectivity possibilities exist, as already described. Additionally, different chips, and even migration to one or more ASICs can occur without a complete architectural re-design; reduced capacity (baselines, bandwidth, channels) can be initially deployed with FPGAs, and then full capacity can follow once the ASIC(s) are available and ready. This brings any ASIC development off the critical path to initial telescope operations.

**ROUGH DEVELOPMENT SCHEDULE**

There are no critical path long-lead time items in this proposal that need to be expedited beyond what the general SKA schedule mandates. For example it doesn’t seem, if the cost and power numbers in this analysis are correct, that ASICs for correlation or beamforming would be cost effective in this case. Even if one or more ASICs are found to be cost effective, design and test at the sub-element level can occur with FPGAs and then ASIC mezzanine cards can be plugged in before final testing and deployment, thus removing ASICs from the critical path.

The main technology item is development, test, and setting in motion volume manufacturing of reliable and robust PowerMX boards and mezzanine cards, as well as setting in place firmware and software IP management and practises that maximize design re-use across projects (sub-elements within the CSP,
and if desired, other consortia that might use the same platform) to minimize development and re-
development effort. And of course, if possible, utilizing existing firmware where possible.

Development of PowerMX can occur concurrently with sub-element system engineering, provided base
specifications (functionality and performance, form factor, mounting holes, connector pinouts, M&C
definition) are in place. Much of this has been done at a draft level in the PowerMX technical
description document. Further feasibility investigation and refinement leading to a “base standard”
during Stage 1 will (if successful) lead to a situation where now multiple parties can start building,
testing, and deploying according to that standard.

The rough development schedule/approach is therefore as follows.

**Stage 1:**

- Develop a draft “base standard” for the PowerMX platform. Much of this would (ideally) have
  contributions from other sub-elements and perhaps consortia efforts. This standard includes
  things like board dimensions; mounting hole locations and attributes; standard I/O connector
  locations, pinouts, and signalling/supply; user-defined I/O areas; power connections and
  attributes; allowable bottom-side component height; and base supervisory M&C requirements
  and mechanisms. This base standard would not include details of application mezzanine cards
  and attributes. This base standard then allows higher level sub-element engineering activities to
  proceed as well as detailed investigation of the mezzanine option and design and construction
  of “PowerHX” (“hardened” i.e. w/o application mezzanine) boards for those who want them in
  the near term.
- Proceed concurrently with PowerMX mezzanine feasibility/performance investigations and
  definitions leading to a draft PowerMX “mezzanine standard” and with sub-element engineering
  investigations (packaging, device utilization, system engineering PAQA, RAM, interfaces etc.).
- Develop the PowerMX “supervisory and M&C standard” to the draft level. This defines what
  any PowerMX or PowerHX board must provide for supervisory and M&C base level capability.
  Things like V+T+P monitoring, control, protection; site/mezzanine card identification and
  programming; M&C methods and protocols from the M&C mezzanine card to each site etc.
- Participate in development of draft plan for firmware and software IP and library sharing.
  Perhaps most of this should come under the .STA work package.

By the end of Stage 1, draft base and mezzanine standards for PowerMX exist, ready for Stage 2. The
base standard may be at V1.0 if prototype and construction by others has been done. The feasibility
and performance of the mezzanine card and standard has been established to set the stage for Stage 2
prototype design and testing. As well, sub-element engineering of the MID-CBF should be to the point
of feasibility leading into more detailed design work of the entire sub-element.

**Stage 2:**

- Carry out further sub-element design work: PowerMX blade electrical, mechanical, thermal, I/O
  modules; shelves; backplanes; racks; cooling; infrastructure specification.
- Carry forward PowerMX into prototyping and testing, including PAQA, mfg, and RAM issues.
- Build and test a triple-PowerMX blade to ensure electrical, mechanical, and thermal
  performance risks/issues are retired.
- Carry forward PowerMX supervisory M&C, application layer, firmware/software IP libraries and
  sharing.
• Further refined device utilization studies, initial designs, interfaces; ready for full testing during construction.

By the end of Stage 2 the sub-element system design is ready for fabrication and testing; PowerMX base, mezzanine, and “supervisory and M&C” standards are at release V1.0 level; PowerMX performance risk is retired. Ready to start pre-production construction and testing of complete sub-element. Poised for technology refresh with new mezzanine cards to be developed and integrated within existing sub-element infrastructure design. Careful and rigorous management and standards of quality are key.

PERFORMANCE ANALYSIS

The following table lists the key metrics used in an Excel spreadsheet in coming up with performance leading to the proposed design in the technical description section. For pulsar phase binning see the description in the section above.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Key Base Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of TMACs/s on a mid-size 2016 FPGA, 100% utilization.</td>
<td>3.45</td>
<td>2 x XC7VX550T @ 600 MHz usable clock speed (2 x 2880 x 600 MHz).</td>
</tr>
<tr>
<td>Power per TMACs/s, FPGA technology</td>
<td>7.5 W</td>
<td>Reference: FPGA power estimator, 50% switching, 25% reduction in power over 28 nm (mfg claims ~50%, so this is conservative)</td>
</tr>
<tr>
<td>Power per DDR3 SODIMM (&gt;10 GB/s) and FPGA I/O power</td>
<td>5 W</td>
<td>Micron 8GB SODIMMs are ~3 W, so seems about right. &lt;8 GB power is less</td>
</tr>
<tr>
<td>Power per tx driver, 10 Gbps</td>
<td>0.2 W</td>
<td>A bit worse than FPGA advertising claims. Need to find reference.</td>
</tr>
<tr>
<td>Power overhead factor for FPGA (multiplier x DSP block power), not including I/O.</td>
<td>1.4</td>
<td>Needs further refinement/investigation.</td>
</tr>
<tr>
<td>FPGA DSP block utilization factor</td>
<td>0.8</td>
<td>As per White Paper TOC.</td>
</tr>
<tr>
<td>LVDC power supply conversion efficiency, each step</td>
<td>0.93</td>
<td>Two steps required, for 0.865 net.</td>
</tr>
<tr>
<td>Total F-part (fine filterbank) processing load, TMACs/s, factoring in 80% FPGA utilization factor.</td>
<td>57.2</td>
<td>Doesn’t include the practical impact of delay tracking with external DDR RAM.</td>
</tr>
<tr>
<td>Memory per antenna for Delay Buffer, 1400 usec depth</td>
<td>112 Mb</td>
<td>Exceeds projected 84 Mb internal FPGA capacity—need to use external DDR SDRAM</td>
</tr>
</tbody>
</table>

---

9 Gaussian noise, properly sampled, switches at 50%; this should be conservative.
**Total X-part processing load, TMACs/s, factoring in 80% FPGA utilization factor.** | 1612.9 | Peter’s number was 2003, but for confusing values. This is calculated for a 2.5 GHz full pol’n correlator.
---|---|---
**Adjusted X-part processing load 4-bit TMACs/s** | 403.2 | Stuffing 4-bit CMAC in a single FPGA DSP block; John Bunton and others.
**PSS BF, total processing load TCMULTs/s, phase-delay correction on 300 MHz, 2100 dual-pol’n beams** | 158 | Complex. Can fit 8-bit cmult into 3 DSP blocks; total of 1.152 CTMACs/s (3.45/3) in 2016 FPGA.
**PST BF, total processing load, TMULTs/s, 32-tap complex delay correction, phase rotation.** | 571.8 | 10 MHz x 1.2 oversampling x (64 taps + 3 mults\(^{10}\)) x 2 pol’n x 254 antennas x 1400MHz/10MHz x 10 beams.
**No. of FPGAs required for F-part, 80% util (calculated = 17, see comments)** | 192 | Due to Delay Buffer and fine filterbank requirements, actually need to use 192; 2/3rds for the Delay Buffers, with 64 for the fine filterbanks. Therefore, there should be large overcapacity in the FFB FPGA.
**No. of FPGAs required for X-part, 80% util** | 117 |
**No. of triple-PowerMX blades, F-part** | 16 | Needs input BW to handle 254 antennas; do 4 dual-pol’n antennas per highway.
**No. of dual-PowerMX blades, X-part (calculated=15)** | 16 | Make an even 16 for mesh numerology. 18 also fits nicely with 3x12 data path numerology.
**No. of FPGAs required for PSS-BF, 80% util** | 171 | 3 DSP blocks can perform an 8-bit complex multiply.
**No. of FPGAs required for PST-BF, 80% util** | 208 |
**No. of triple-PowerMX blades, PSS-BF** | 15 |
**No. of triple-PowerMX blades, PST-BF** | 20 | Based on 6 dual-pol’n, 10 MHz beams with **1.2 oversampling** formed in each FPGA, true delay tracking with 32-tap complex FIRs.
**Power**
**F-part, X-part triple-PowerMX blade, Power ea** | 1088 W | Includes all FPGA, I/O, DDR3 module power, power supply inefficiencies, net ~90 W per mezzanine site. Should be conservative. 50% switching.
**PSS-BF and PST-BF, triple-PowerMX blade,** | 580 W | Reduced data path use, no DDR3

\(^{10}\) An 8-bit complex multiply can be done in 3 DSP blocks (ref: Xilinx Coregen compiler).
Power ea modules, net ~48 W per mezzanine site. Should be conservative. 50% switching

PSS-BF, PST-BF, single-PowerMX blade, Power ea 193 W ~1/3rd triple

COST and POWER ANALYSIS

Key Cost and Power Estimates and Parameters
The following table contains key cost and power estimates and derived parameters captured from the costing and power spreadsheet developed for this purpose. This table does not contain total installed power and infrastructure and operating life overhead—see next sections.

<table>
<thead>
<tr>
<th>Description</th>
<th>Price</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PowerMX Cost estimates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost per mid-size 20 nm FPGA 2016, includes 3.45 TMACs/s and all other chip resources. Usable clock rate of 600 MHz.</td>
<td>$1800</td>
<td>Blended FPGA mfg estimates. For &gt;=18x18 integer multiply and accumulate/adder.</td>
</tr>
<tr>
<td>Cost per Gbps, fiber modules, one direction (10 Gbps technology).</td>
<td>$1.50</td>
<td>Need more investigation, but previous investigation indicated $1/Gbps for transmit/receive pair is achievable (so this is 3X worse). This means a 12-fiber x 10G/fiber tx or rx module is ~$180</td>
</tr>
<tr>
<td>Cost per Gbps, copper, one direction (10 Gbps technology)</td>
<td>$0.30</td>
<td>WAG—needs further investigation. PowerMX I/O connectors are ~$.13</td>
</tr>
<tr>
<td>Turn-key delivered cost of bare PowerMX motherboard, not including fab NRE which is ~$35k.</td>
<td>$4000</td>
<td>8% turnkey parts overhead, 25% conversion (mfg, test, profit), 5% to developer, 15% contingency; based on $500 bare PCB cost.</td>
</tr>
<tr>
<td>Turn-key delivered cost of PowerMX mezzanine card with FPGA</td>
<td>$2328</td>
<td>Includes 4 DDR3 SODIMs</td>
</tr>
<tr>
<td>Turn-key delivered PowerMX M&amp;C mezzanine card</td>
<td>$500</td>
<td></td>
</tr>
<tr>
<td>Turn-key delivered PowerMX backplane driver/transition module</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>Turn-key delivered PowerMX lane repeater</td>
<td>$75</td>
<td></td>
</tr>
<tr>
<td>Single-PowerMX blade packaging, including thermal</td>
<td>$800</td>
<td></td>
</tr>
<tr>
<td>Dual-PowerMX blade packaging, including thermal</td>
<td>$2500</td>
<td>Includes board-to-board repeaters</td>
</tr>
</tbody>
</table>
### Derived Total System Costs

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total turn-key cost, SKA1-MID correlator, with reduced shared beamforming capacity</td>
<td>$1.3 M</td>
<td>Could do 1400 PSS beams 300 MHz, 700 MHz PST all beams, mutually exclusive.</td>
</tr>
<tr>
<td>Total turn-key cost, SKA1-MID correlator and shared PSS/PST beamformers; 36 triple blades, loaded (OPTION 1)</td>
<td>$1.7 M</td>
<td>Does not include external fiber or power cabling, or spares. Correlator, PSS-BF, and PST-BF operate mutually exclusively.</td>
</tr>
<tr>
<td>Total turn-key cost, separate PSS-BF</td>
<td>$0.87 M</td>
<td></td>
</tr>
<tr>
<td>Total turn-key cost, separate PST-BF</td>
<td>$0.98 M</td>
<td></td>
</tr>
<tr>
<td>Total turn-key cost, SKA1-MID correlator, PSS, and PST beamformers, all operating concurrently, with full capacity. (OPTION 2)</td>
<td>$3.2 M</td>
<td>Does not include external fiber or power cabling, or spares.</td>
</tr>
</tbody>
</table>

### Derived Total System Power

<table>
<thead>
<tr>
<th>Description</th>
<th>Power</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPTION 1: Shared correlator and beamformer, full mutually-exclusive capacity</td>
<td>41.5 kW</td>
<td>Doesn’t including cooling power—see next sections</td>
</tr>
<tr>
<td>OPTION 2: Full concurrent capability</td>
<td>49 kW</td>
<td>Doesn’t including cooling power—see next sections</td>
</tr>
</tbody>
</table>
BOTTOM LINE—Development Costs (NRE)

The following table contains estimated development costs. FTE cost of $130,000 Euro/yr was used rather than 100,000, which seems a bit thin. FTE estimates for Stage 1 and Stage 2 were based on the WPEP for this workpackage. The table includes construction and commissioning FTE estimates. Note that this table does not include PowerMX development effort, which will (may) be shared across workpackages and possibly consortia if successful as a common platform.

<table>
<thead>
<tr>
<th>ENGINEERING/DEVELOPMENT EFFORT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1 (FTEs for 1 year)</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Stage 2 (FTEs for 2 years)</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Construction (FTEs for 2 years)</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Commissioning (FTEs for 2 years)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td><strong>TOTAL NRE FTE Effort</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: if PowerMX used for other sub-elements/Elements much of the Stage1, Stage2, and construction effort can be shared*

<table>
<thead>
<tr>
<th>SUB-TOTAL COST OF NRE FTEs (MEuros)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated total prototyping costs, NREs etc for PowerMX with FPGAs and custom fiber module (MEuro)</td>
<td>€ 0.50</td>
</tr>
<tr>
<td><strong>TOTAL ESTIMATED NRE (MEuros)</strong></td>
<td><strong>€ 6.74</strong></td>
</tr>
</tbody>
</table>

Note that this is sub-Element level development wherein the PowerMX motherboard and mezzanine cards are commodity “purchased” items with a shared-cost development and use model. If such development is required strictly within the SKA1-MID CBF alone, then another 1 FTE in Stage 1, and 2 FTEs in Stage 2 are likely required, increasing the cost by ~650k Euros over that shown in the above table. This table includes 1 FTE during construction to get the sub-Element-specific PowerMX application software interface to the point of abstract (e.g. XML) configuration and monitor to tie into the Element-level LMC work. However, of course, predicting software effort is very difficult and will likely come in above any seemingly reasonable estimate.

BOTTOM LINE—OPTION 1 Total Cost Estimates

“OPTION 1” is a combined correlator, PSS-BF, and PST-BF allowing total capability to be provided in a mutually-exclusive fashion. The PST-BF, on the “X-side” of the rack dominates here, requiring 20 triple-PowerMX blades\(^{11}\) to process all 254 antennas, 1.4 GHz/pol’n, 10 dual-pol’n output beams, at an array aperture of 200 km. All White Paper TOC metrics were used, including a 10 year operating lifetime, and 2 FTEs maintenance, but at $130,000 Euro/yr rather than 100,000. A 1.35x dollars to euros conversion factor was used.

---

\(^{11}\) As noted in the calculations on page 11 quad blades could be used if needed by the X-part, increasing the capital cost by ~0.15M Euros.
OPTION 1: Capital cost plus spares (includes power+infrastructure) €1.59
OPTION 1: Total design build estimate, w/o operating costs, includes infrastructure+NRE €8.33

OPTION 1: NRE, CAPITAL, AND OPERATING COSTS (TOTAL LIFE CYCLE)

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost (€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital cost + spares</td>
<td>1.34</td>
</tr>
<tr>
<td>Operating power, raw (kW)</td>
<td>41.41</td>
</tr>
<tr>
<td>Operating power, at stated 97% mains power delivery efficiency (kW)</td>
<td>42.69</td>
</tr>
<tr>
<td>Total operating power, including cooling overhead (kW)</td>
<td>64.03</td>
</tr>
<tr>
<td>Power delivery infrastructure capital cost overhead (MEuro)</td>
<td>0.1281</td>
</tr>
<tr>
<td>Power removal infrastructure capital cost overhead (MEuro)</td>
<td>0.1281</td>
</tr>
<tr>
<td>Operating maintenance cost (personnel) (MEuro)</td>
<td>0.26</td>
</tr>
<tr>
<td>Operating cost (power) for stated lifetime (MEuro)</td>
<td>0.6403</td>
</tr>
</tbody>
</table>

OPTION 1 GRAND TOTAL LIFE CYCLE COST (NRE, capital, operating, maint, spares) €11.57

Note that operating cost includes 5% spares, 2 FTEs, and power, and is ~4% of design-build-install cost. This does not include contingency. **It is prudent to add 30% contingency at this point in time, so the TOTAL LIFE CYCLE COST is more like 15 M-Euros.**

BOTTOM LINE—OPTION 2 Total Cost Estimates: BOTTOM LINE

“OPTION 2” is a total solution that provides concurrent operation of the correlator\(^1\), PSS-BF, and PST-BF. All White Paper TOC metrics were used, including a 10 year operating lifetime, and 2 FTEs maintenance, but at **130,000 Euros/year** rather than 100,000. A 1.35x dollars to euros conversion factor was used.

OPTION 2: Capital cost plus spares (include power+infrastructure) €2.75
OPTION 2: Total design build estimate, w/o operating costs, includes infrastructure+NRE €9.49

OPTION 2: NRE, CAPITAL, AND OPERATING COSTS (TOTAL LIFE CYCLE)

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost (€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital cost + spares</td>
<td>2.45</td>
</tr>
<tr>
<td>Operating power, raw (kW)</td>
<td>48.99</td>
</tr>
<tr>
<td>Operating power, at stated 97% mains power delivery efficiency (kW)</td>
<td>50.51</td>
</tr>
<tr>
<td>Total operating power, including cooling overhead (kW)</td>
<td>75.76</td>
</tr>
<tr>
<td>Power delivery infrastructure capital cost overhead (MEuro)</td>
<td>0.1515</td>
</tr>
<tr>
<td>Power removal infrastructure capital cost overhead (MEuro)</td>
<td>0.1515</td>
</tr>
<tr>
<td>Operating maintenance cost (personnel) (MEuro)</td>
<td>0.26</td>
</tr>
<tr>
<td>Operating cost (power) for stated lifetime (MEuro)</td>
<td>0.76</td>
</tr>
</tbody>
</table>

OPTION 2 GRAND TOTAL LIFE CYCLE COST (NRE, capital, operating, maint, spares) €12.85

Note that operating cost includes 5% spares, 2 FTEs, and power, and is ~3.5% of design-build-install cost. This does not include contingency. **It is prudent to add 30% at this point in time, so the TOTAL LIFE CYCLE COST is more like 16.7 M-Euros.**

---

\(^{12}\) If the X-part of the correlator requires 20, quad-PowerMX blades, cost increases by ~0.45M Euros, as noted in calculations on page 11. This is well within the 30% contingency noted.
VIABILITY AND SCALABILITY TO SKA2

The SKA2 correlator/beamformer, if ever built, will be a formidable machine in comparison to SKA1. Increasing the number of antennas by a factor of 10 increases the correlator by at least a factor of 100, and if high-sensitivity PAFs are viable, by a factor of ~3000. The PSS-BF aperture likely won’t change, but the number of antennas added to the BF will by a factor of ~10. Same for the PST-BF. As argued in previous sections, SKA1-MID.CBF fits in a few racks “in the corner of the room”, SKA2, with the same technology, could fill several rooms. Power costs will be of extreme concern and lowest-power ASICs, even if the NRE is millions will likely be worth the effort in power savings.

It is hard to predict what future technologies might hold, but perhaps there can be some reasonable guesses. Here are some:

- Optical transmission, over any distance, may win the arms race over copper. Optical backplane technology may mature before PCB chip-to-chip optics as there is more to be gained by tackling the biggest problem (longest distance) first.
- SERDES/serial interconnects will rule as they provide the highest connectivity bandwidth. Is there anything better than serial in this regard? Chip-to-chip and optical multi-level encoding might be employed as bit rates get higher. How high is high enough? Do we need 100 Gbps/pair any time soon? Will silicon match that?
- Silicon processing performance will continue to improve, but increasingly 3D die stacking and advanced thermal packaging will likely be required to continue to boost performance (no crystal ball gazing here, this is already happening). Probably higher chip power consumption will be the case, requiring more advanced thermal solutions.
- Optical fiber will continue to be the choice interconnect technology. Is there any hint of anything better? Wormholes in spacetime?

The PowerMX platform was conceived to deal with fast-changing technology as much as possible, and to be scalable to SKA2 by providing a motherboard with so much I/O performance that it will be a long time before silicon performance (for the kinds of things we do) will catch up any time soon (e.g. from an I/O perspective, the entire SKA1-MID.CBF—and NIP—could almost be on one PowerMX board, if silicon were willing). The previous sections used 10G serial interconnects to be conservative; “catalog engineering” indicates PowerMX interconnects should be capable of 28G each, and possibly 40G, so there is likely some room for I/O performance growth, if required. Even if motherboard I/O maxes out, the design is such that I/O and nearest-neighbour connections can be transferred to the mezzanine cards while still using the motherboard for other base functions.

Given these conditions and assumptions, PowerMX might scale to SKA2 in the next 10 years as follows, assuming 4X\(^1\) performance increases for the same power per chip\(^1\) in just FPGAs, 28 Gbps/pair with much better performance-to-power ratio in ASICs:

- 16 antennas per highway, or 64 antennas per PowerMX blade assuming 2.5 GHz/polarization. For 10X antennas it requires 40 triple-PowerMX blades. Power \(\approx 1000 \text{ W/blade} \times 40 \approx 40 \text{ kW.}
In this case each antenna requires 3 x 28 Gbps serial paths for 2.5 GHz/polarization 8-bit sampling, requiring 48 paths per lane, something PowerMX is already equipped with.

---

\(^1\) One prediction that seems to hold is that predictions of the future are often horribly wrong!

\(^1\) This may not be unreasonable; Altera has announced they have an exclusive deal with Intel to use their 14 nm technology, 10 nm probably isn’t that far behind.

\(^1\) This could be a stretch, might want to multiply all of the following power numbers by 2.
100X more baselines, at 4X more processing each, means 25X dual-PowerMX blades (400) for correlation. 400 x ~650W (dual blade) ~ = 260 kW. With ASICs, this might drop by a factor of 10 max.

At 20 blades/rack, that’s 2 racks for the F-part, and at 20 blades/rack, 20 racks for the X-part.

In the F-part, the backplane — copper or fiber, whichever is capable — performs the normal mesh operation, but instead of immediately routing to X-boards on the rear side, copper (or optical)-to-fiber transition modules are plugged in. In this case all fibers from both shelves within a particular slot (384 fibers in total) route, in a point-to-point manner, to one of the 20 X-part shelves’ front-panel (or rear) fiber inputs. Each X blade in that shelf gets 384/20 = ~20 fibers, or ~5 fibers for each of the 4 lanes.

The correlator itself, for 2560 antennas, might be 300 kW and maybe as low as 30 kW with delay tracking, F-part, and X-part ASICs (although that lower power number somehow seems unachievable). PSS-BF and PST-BF requires further study, but data routing should occur in a similar manner as SKA1, with something like Avago Micropod fiber drivers on F-part delay mezzanine cards routing PST-BF data to the front panel as there is no spare lane for it.

This little thought exercise demonstrates that, scaling to SKA2, for 2.5 GHz/polarization single beam may not be a completely wild fantasy.

Keeping the same costs per chip (i.e. FPGAs), and multiplying by 22 (i.e. 22 racks vs 1), assuming the same cost per unit FPGA, puts the cost at 22 x $1.3M ~ = $30M (correlator only, not including NRE, but including power and cooling infrastructure). If PAFs at that bandwidth are used, well, do the math — ASICs will definitely be required! One would think that SDP processing cost, power, and infrastructure will dwarf the correlator/BF for SKA2 in any case.

Other hidden costs and effects migrating from SKA1 to SKA2 might be:

- Real-time processing IP/firmware/software. Can much of this be re-used and migrated to SKA2? Or, does it have to start from scratch. Does it matter?
- Observation/M&C software. If SKA1 is built with some capabilities/flexibilities that likely will never be re-used SKA2, is all of it throw away? Will it have to be built from scratch? There is significant effort in this regard in a robust system. The cost and impact of possibly doing it twice must be considered carefully.

With PowerMX as a platform for SKA1, migration to SKA2 should be able to build on everything already developed and not mean a complete re-design provided firmware IP blocks and libraries, as well as layers of software are built and managed in a rigorous manner so as to be re-usable as much as possible as the system scales up and across generations of technology.
CONCLUSIONS

This white paper has presented a plan for implementation and construction of the “combined” 254-antenna SKA1-mid dish array 2.5 GHz/polarization correlator and beamformer. It is based on the PowerMX platform that is scalable to meet SKA1 and SKA2 needs, as well as keep up with fast paced technology to deliver a cost-effective system for the telescope. It allows for shared development and use with the intention of maximizing involvement of all contributing SKA organizations. PowerMX would normally be populated with FPGAs however, other devices may be used on PowerMX if found to be more cost effective for example ASICs, specialized DSP chips etc.

Two options were costed out even though they both use the same technologies and system design resources. Thus there is only one main approach that is planned to be investigated in Stage 1, and if successful, following up in Stage 2.

The options costed out are as follows:

- **OPTION 1** is a combined correlator/beamformer to be used in the case where the observing model is such that they are not required to be operated concurrently. Total 10-year life-cycle cost, including NRE is estimated at 15M Euros, including a 30% contingency. The total design-build estimate is ~11M Euros, including NRE, spares, power and cooling infrastructure, and 30% contingency. All labour costs were calculated at 130k Euros per FTE-year. Operating power dissipation is estimated at 42 kW.

- **OPTION 2** is a separate correlator/beamformer, sharing functions such that they may be operated concurrently. Total 10-year life-cycle cost, including NRE is estimated at 16.7M Euros, including a 30% contingency. The total design-build estimate is ~12.3M Euros, including NRE, spares, power and cooling infrastructure, and 30% contingency. All labour costs were calculated at 130k Euros per FTE-year. Operating power dissipation is estimated at 50 kW.
TECHNICAL DESCRIPTION

“PowerMX”

A concept for a common FPGA, ASIC, and CPU/GPU co-processor platform for all SKA1 and SKA2 signal processing...

“At its heart it simply provides high-performance access to abundant chip resources, both processing and I/O, packaged within a medium-granularity form factor.”

Initial Author: Brent Carlson, NRC-Canada
Date: April 10, 2013
Version: DRAFT
EXECUTIVE SUMMARY
The SKA offers significant signal processing challenges in terms of scale of processing and in terms of the development and roll-out timescale, which can span several technology generations. The requirement to deliver reliable and robust systems, and the ever-present pressure to cost-effectively keep up with the technology curve for cost, power, and increased performance benefits only adds to the challenge.

“PowerMX” is a platform conceived to deal with these often conflicting issues. At its heart it simply provides high-performance access to abundant chip resources, both processing and I/O, packaged within a medium-granularity form factor. PowerMX is a motherboard accepting mezzanine plug-in cards that are carriers for the latest chips required for each particular application and with enough user-defined I/O capability that it will likely take a decade or more for silicon performance to catch up. It prescribes a specific intra-board interconnect fabric upon which layers of firmware and software can be built, but allows for inter-board connectivity suitable and scalable for each particular deployed application. It is an “appliance”, analogous to an iPhone (or its ilk) for applications to be built on.

The mezzanine approach provides flexibility but does introduce some risk; a very attractive connector in terms of reliability, lifetime, and performance-lifetime has been found and is being investigated (see page 31). The fallback position if an acceptable connector is not found is to produce “hardened” versions of the board with the same I/O and connectivity, but with specific chip sets designed into the board.

Figure 1 is an annotated diagram of the PowerMX concept showing major functional blocks. There are 4 application mezzanine cards sites, each card ~89 mm (“3.5”) on a side and large enough for a very large (45 mm) chip and memory modules or, as shown, 4 smaller F672 27x27 mm chips and memory modules. Monitor & Control is provided by a smaller mezzanine card; shown on this card is a Xilinx Zynq SoC (System on a Chip) FPGA which mates an FPGA with a hard ARM CPU. “SNAP12-A” “panel-ready modules at 12 serial I/Os each provide the board with its bulk I/O capacity of 384 serial (fiber) links. Dual auxiliary 10G links are provided to each mezzanine card for lower performance I/O such as transfer of integrated coefficients or filter coefficients. The board is ~17.1” (434 mm) x ~6.7” (170 mm) and is therefore suitable for mounting in a 19” rack-mount pizza box; blade mounting/use is also explored on page 25.

The mezzanine approach adds flexibility and the ability for various SKA organizations to contribute at levels they are able in a coherent fashion to maximize use of distributed resources. Motherboards can be designed and built by one or more organizations and then made available in a COTS-like turn-key manner to other user organizations at an attractive price (typically ~5-7% above mfg price—similar to what might be obtained through a licensing/royalty agreement with industry) through one or more EMS (Electronics
Manufacturing Supplier) providers. Other organizations can develop plug-in mezzanine cards to suit particular applications, and use them as well as make them available to other users under a similar business/purchase model. A MSA (Multi-Source Agreement) could be established to allow all of this to happen so that everyone involved benefits from their own and others’ developments. Yet other organizations can develop firmware and software libraries and applications built on the hardware. Finally, sub-element integration and test facilities can gather in all these resources to put together fully integrated and tested sub-elements for delivery to telescope sites.

The ultimate goal with this (or a better) platform architecture is to deliver the lowest total life-cycle cost, lowest-power, reliable, scalable solution, with a long technology lifetime to the SKA while making the best use of distributed development and test resources available to the project.
Figure 1 PowerMX annotated diagram showing major features. Four application mezzanine sites are provided; shown with large 45x45 mm chips, and one with four 27x27 mm chips providing similar aggregate capacity at much lower cost for those compute problems that can be split across devices. The card is ~17” x 7”.
# TABLE OF CONTENTS

EXECUTIVE SUMMARY .................................................................................................................. 2

TABLE OF CONTENTS .................................................................................................................... 5

TABLE OF FIGURES ........................................................................................................................ 6

REVISION HISTORY ....................................................................................................................... 7

KEY FEATURES ............................................................................................................................... 8

THERMAL AND PACKAGING SOLUTIONS ...................................................................................... 12

  Default 19” Rack-mount “Pizza-box” with Air Cooling ................................................................. 12

  19” Rack-mount Pizza-Box with Monolithic Liquid Cooling ......................................................... 14

MULTI-BOARD PACKAGING SOLUTIONS—DAISY CHAINING .................................................... 16

MULTI-BOARD PACKAGING SOLUTIONS—MESHING ................................................................. 20

INTRA-BOARD MESH ..................................................................................................................... 23

BLADE AND BACKPLANE MOUNTING POSSIBILITIES ............................................................... 25

OTHER I/O POSSIBILITIES AND APPLICATIONS ......................................................................... 28

MEZZANINE CARD FLEXIBILITIES ............................................................................................... 30

TECHNICAL DETAILS ...................................................................................................................... 31

  400-pin Meg-Array Connector ..................................................................................................... 31

    Contact-to-Contact and Ball-to-Contact Reliability and Lifetime ........................................... 31

    Differential Pair Performance .................................................................................................. 33

    I/O Capacity and Pinouts .......................................................................................................... 34

    Insertion/Extraction Force and Multi-connector Use ............................................................... 35

    Cost ........................................................................................................................................... 37

  “SNAP12-A” 100-pin Meg-Array Connector ............................................................................. 37

  M&C Mezzanine Meg-Array Connector ....................................................................................... 39

  Clock Network ............................................................................................................................ 40

  48 and 12 V Power ....................................................................................................................... 41

  Monitor & Control ....................................................................................................................... 41

RISKS AND CHALLENGES ........................................................................................................... 42

WAY FORWARD ................................................................................................................................. 43
TABLE OF FIGURES

Figure 1 PowerMX annotated diagram showing major features. Four application mezzanine sites are provided; shown with large 45x45 mm chips, and one with four 27x27 mm chips providing similar aggregate capacity at much lower cost for those compute problems that can be split across devices. The card is “17” x 7”. .................................................. 4

Figure 2 PowerMX motherboard showing all significant I/O and intra-board data paths. .................................................. 11

Figure 3 Notional 1U PowerMX 19” rack-mount pizza-box air-cooled solution. Maximum power for a 50 °C rise is 30 W/mezzanine chip; 17 W max results in a more comfortable T_{max} of about 50 °C. If 1.5U height is used, with correspondingly larger fans and heatsinks, 30 W/chip might be comfortably handled. These fans could be integrated into the front and rear panels; if a fan fails, cables are removed and the panel—with new fans—is replaced. .................................................. 13

Figure 4 Monolithic, integrated, liquid cooling solution. The cooling plate is attached to the mezzanine cards either using the mezzanine card mounting holes or using motherboard mounting holes (not shown). ........................................ 15

Figure 5 PowerMX plates integrated with cooling plates. ................................................................................................................. 16

Figure 6 Dual PowerMX boards daisy-chained and integrated into one pizza-box. ................................................................. 17

Figure 7 Single Meg-Array, 12-channel repeater “printed wiring” connection board. ................................................................. 18

Figure 8 Eight PowerMX boards integrated in a dual stack of 4 with fiber cable interconnect mesh. ........................................ 21

Figure 9 2304-station correlator in single crate using the CSIRO ADE passive fiber mesh. The F, corner-turner, and X parts could be spread across a rack (or multiple racks) in 3 parts with S-style daisy-chaining. Each 12x12 MTP cable optical cross-connect costs ~$1000 (Hampson, verbal quote). ................................................................. 22

Figure 10 Reflex Photonics pluggable “lightABLE” module showing MT cable and retention clip. This module and fiber is likely suitable for high density intra-crate cabling. Other manufacturers have similar offerings. ......................... 23

Figure 11 PowerMX copper backplane transition module consisting of a PCB, 4 mating 100-pin Meg-Array connectors, VSC7227 drivers/receivers, and dual FCI ExaMax right-angle backplane connectors. A similar concept is possible if optical backplane technology becomes mature, in which the copper right-angle PCB connectors are replaced with fiber connectors and drivers/receivers. ................................................................. 25

Figure 12 Single PowerMX board mounted on an aluminum substrate with backplane transition module. The substrate, rather than the PCB, slides into guide rails. The signal path, in profile, is shown in red. Note that the backplane can be split into 4 identical backplanes. ................................................................................................................. 26

Figure 13 PowerMX with ADC transition modules for direct RF/IF input. .................................................................................. 29

Figure 14 Extended mezzanine card. The SNAP12-A I/O module sites are unused. ................................................................. 30

Figure 15 28 Gbps Meg-Array eye diagram, BSC or ESC, no signal conditioning according to OIF short reach specification CEI-28G-SR. ................................................................................................................. 33

Figure 16 PowerMX 3 x 400-pin Meg-Array mezzanine connector pinouts. ............................................................................... 36

Figure 17 Photos of the FCI Meg-Array connector, pin (a), socket (b), pin closeup (c), socket closeup (d), underside BGA (e). ................................................................................................................. 38

Figure 18 “SNAP12-A” 100-pin Meg-Array preliminary pinout. ................................................................................................. 39

Figure 19 M&C mezzanine card preliminary pinout definition. ................................................................................................. 40
## REVISION HISTORY

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Authors</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAFT</td>
<td>April 9, 2013</td>
<td>B. Carlson</td>
<td>First draft release (internal).</td>
</tr>
<tr>
<td>DRAFT</td>
<td>April 10, 2013</td>
<td>B. Carlson</td>
<td>Change name to PowerMX to avoid conflict with commercial product names; first release for “call for co-author” review and comment.</td>
</tr>
</tbody>
</table>
KEY FEATURES

- Medium granularity, “future-proof”\(^1\) design can be used (and re-used) across generations of technologies for all SKA1 and SKA2 signal processing. With this foundation (and associated packaging/cooling options) development can focus on firmware and software applications without always worrying about/spending resources on spinning a new board (and associated changes in connectivity, connectors, form factor etc.) to keep up with technology. A small mezzanine card with a few chips on it has a faster development/deployment time than a large board; different groups can develop different mezzanine cards with different chips and technology generations concurrently or in a “leap-frog” fashion.

- Amenable to initial deployment with FPGAs and eventual migration to ASICs using the same physical and M&C mechanisms.

- Four 3.5” x 3.5” application mezzanine sites. Each site has 96 I/Os to front and rear-panel SNAP12-A (see page 37) modules and 192 I/Os to nearest neighbour modules. This arrangement allows for full-bandwidth intra-board mesh connections with nearest-neighbour performance. Additional fiber or nearest neighbour mezzanine-to-mezzanine connections could be established independent of the motherboard, with appropriate modules or connectors on mezzanine cards.

- Total primary I/O capacity of 384 serial links at (TBC) 28 Gbps/link (possibly up to 40 Gbps/link) for aggregate capacity of 7.95 Tbps (@28G/link).

- Each application mezzanine site uses 3, 400-pin 4 mm FCI Meg-Array connectors. These connectors have recently been characterized for 28 Gbps operation, are rated for a 22 year lifetime, have been in production since ~1998 with ~17 billion installed line units, and cost about $43 per mating pair (DigiKey catalog, MOQ=250, 50 Au). See page 37 for part numbers and costs.

- SNAP12-A modules are “panel ready” for MTP 12-fiber connectors; other individual or aggregate I/O modules can be used/plugged into the 100-pin Meg-Array connectors. These are patterned off industry-standard SNAP-12 modules/connectors but with modified I/O assignment and mechanics for higher performance and direction-agnostic use. See page 37 for further details.

\(^1\) Thanks to Oliver Sinnen from Auckland University of Technology for coining this term.
• The motherboard does not define serial link direction; direction is defined by mezzanine cards and I/O modules. This provides huge flexibility for large fan-in, fan-out, and daisy-chained applications.

• M&C function on a smaller mezzanine card supplied with power in a similar manner as the application mezzanine cards. This could allow flexibility in choice of M&C device and functionality to suit different applications. Provide for standard serial intra-board M&C communications (see Figure 2).

• Motherboard supplies power to each application mezzanine card via high-current (10+ A) “barrel-in-socket” connectors. Using 15 A Linear Tech (LTM4628—see http://cds.linear.com/docs/en/datasheet/4628fd.pdf) programmable supplies, allows for 45 A core supply, and 3 x 15 A I/O supplies with each supply programmable (by the mezzanine card) in the range of 0.6 V to 5 V. Auxiliary +12 V connector allows the mezzanine card to develop its own voltages required by its application. These Linear Tech supplies are very high reliability/lifetime devices.

• Optional 500 W 1/8th brick (Vicor P/N: IB048E120T40N1-00, see: http://cdn.vicorpower.com/documents/datasheets/vibrick/ds_IB048E120T40N1-00.pdf) 48 V to 12 V converter; or, supply +12 V directly. For quiet/mixed-signal applications some or all switching power supplies can be turned off, with each mezzanine card developing its own voltages from the +12 V quiet rail.

• Dual clock I/O SMA connectors with low-jitter fanout buffers/selectors routing clocks to/from mezzanine cards for application-specific clocking. Use low-noise power supplies, low-jitter clock drivers, and high-performance routing to allow for ADC-quality clock distribution to mezzanine and attachment cards (see page 40).

• M&C mezzanine card SoC FPGA provides all supervisory functions including application mezzanine card power sequencing, V+T+P monitoring, and FPGA programming. Supervisory +12V supply (not shown in Figure 1) keeps the M&C function operating even when the main +48 V-to-12 V regulator is turned off. i.e. for simplicity the bulk of the board can be power cycled via network commands rather than via an external dedicated connection. M&C SoC FPGA CPU deadman timer for auto-recovery if the SoC processor crashes. Note that the CPU in an SoC chip is hard-logic, and comes up before the FPGA bitstream is loaded.

• Dual FCI “ExaMax” rear R/A PCB connectors provide total 64 auxiliary I/O connections up to 28 Gbps per pair. This allows for low-cost nearest-neighbour board connections in low I/O daisy-chained applications (e.g. hierarchical beamforming, massive beamforming, correlation).
• Secondary/auxiliary I/O capacity of 40 Gbps/mezzanine card (dual 10G BASE-T) for a total of 160 Gbps. These allow the board to couple to COTS CPU/GPU cards via industry-standard interface for co-processor or FPGA/ASIC accelerator applications.

Figure 2 is a layout diagram of the PowerMX board showing all significant I/O paths. Double arrows do not indicate full-duplex connections but rather that the motherboard is agnostic to signal direction.
Figure 2 PowerMX motherboard showing all significant I/O and intra-board data paths.
THERMAL AND PACKAGING SOLUTIONS

The PowerMX motherboard provides huge I/O and processing power resources for mezzanine cards to utilize. Of course, all of this must be cooled and packaged in a robust manner to allow for development and deployment of real systems. This section provides some examples of packaging and cooling methods that might be employed. Others are likely.

Default 19” Rack-mount “Pizza-box” with Air Cooling

A notional default air cooling solution allowing PowerMX to fit in a 1U height pizza box is shown in Figure 3.

Dual one inch 12 VDC fans (ADDA P/N AD0212MB-K50) are used for entry and exit, each one specified for 2.4 CFM, which over a ~4” x 1” aperture amounts to ~85 LFM each for a (conservative) total of ~170 LFM (CFM to LFM calculator: http://www.csgnetwork.com/cfm2lfmcalc.html). Use an Aavid-Thermalloy 45x45x25 mm, 15x6 fin heatsink (P/N 10-554-50245-C1-HSG) on each 45x45 mm device; since this is a tight space, airflow should be highly collimated², and under these conditions the thermal resistance calculates to ~1.7 °C/W (thermal resistance calculator: http://www.aavid.com/thermal-tools/bonded-fin). At 30 W power dissipation, the temperature rise is 51 °C, or case temperature of 71 °C at an ambient of 20 °C. This is pushing the comfort limit and is likely the maximum power dissipation that could be handled with this configuration. For a Tj max of 50 °C (assuming flip-chip cavity down and low thermal resistance to the case), maximum power dissipation is about 17 W per device. This seems about right based on previous experience with similar heatsinks in the EVLA project, but of course, needs testing and validation.

By comparison, a low profile BGA Fan Heat Sink (Aavid-Thermalloy part number 11-5602-50G), with a mount height of 11.5 mm (leaving a ~13 mm space for air intake—which could be an air-starvation problem), shows about a 3.5 °C/W thermal resistance. This is a bit puzzling as these are supposed to have superior thermal performance, although it is one fan, and fin surface area is significantly smaller that the above case.

A 1.5U height case with correspondingly larger airflows and heatsink fins has cooling area and capacity increase of ~1.87X, which should result in a thermal resistance of ~0.9 °C/W, allowing for a power dissipation of 33 W for a 30 °C temperature rise.

Further investigation is required to determine optimal 1U and 1.5U height air-cooled solutions.

² Care, of course, must be taken to avoid blocking airflow with front or rear fiber cables.
Figure 3 Notional 1U PowerMX 19” rack-mount pizza-box air-cooled solution. Maximum power for a 50 °C rise is ~30 W/mezzanine chip; 17 W max results in a more comfortable $T_{j,\text{max}}$ of about 50 °C. If 1.5U height is used, with correspondingly larger fans and heatsinks, ~30 W/chip might be comfortably handled. These fans could be integrated into the front and rear panels; if a fan fails, cables are removed and the panel—with new fans—is replaced.
19” Rack-mount Pizza-Box with Monolithic Liquid Cooling
A monolithic liquid cooling concept is shown in Figure 4.

In this concept, a liquid plate with input and output cooling lines routing through the back panel, is thermally bonded to mezzanine card devices using a suitable “soft” thermal interface material such as Bergquist Gap Pad 5000S35 (http://www.bergquistcompany.com/thermal_materials/gap_pad/gap-pad-5000S35.htm). For a 45 mm package, and 1 mm (0.040”) thickness, the material has a thermal resistance of ~0.1 °C/W. (This material was used very successfully for this sort of purpose in the EVLA correlator project.) At that thermal resistance and with suitable coolant flow, device power dissipation in excess of 50 W should be easily handled. Also, compared to individual chip liquid cooling plates with lines connecting them in series, this has the advantage of only requiring two external liquid lines per PowerMX, thereby significantly reducing the chance of a leak.

In this solution, though, vertically-mounted memory modules would be problematic to handle, requiring machined cavities in the plate or that they be mounted in the horizontal plane at or below the level of the FPGAs/chips.

A variant of this possible solution is to encase the PowerMX board and mezzanine cards in a ~0.75” thick “case”, thermally bonded to the upper (and for power supplies, lower) surfaces. The liquid cooling plate is then bolted to this externally and may even be integrated in the rack into which PowerMX “plates” are installed. In this case, liquid cooling lines, rather than having to be flexible and removable could be hardwired into fixed cooling plates built into the rack with dual back-to-back (or just single) PowerMX plates in between.

A simplified diagram of this concept is shown in Figure 5. Further investigation would be required to determine how/if mechanically and thermally this would work as well as cable routing, power supply delivery etc. It is an interesting possibility, though.
Monolithic, integrated, liquid cooling solution. The cooling plate is attached to the mezzanine cards either using the mezzanine card mounting holes or using motherboard mounting holes (not shown).
MULTI-BOARD PACKAGING SOLUTIONS—DAISY CHAINING

Previous sections have illustrated how a single PowerMX board can be integrated into a 19” rack mount pizza-box unit. It is perfectly fine to do this, and multiple PowerMX boards can be interconnected in many ways using SNAP12-A pluggable cables (most likely 12-fiber MTP) using just this unit. However, in some cases it may be useful and more cost effective to integrate more than one PowerMX board within a pizza box. A possible method to do this is shown in Figure 6 (not shown is daisy-chained M&C network connections—probably most cost-effectively built into the M&C mezzanine card).

In this example, signal entry is from the bottom of the figure (but could be the opposite direction). An example is a 192-antenna/element correlator (or beamformer); the first stage (in the first PowerMX) is the filterbank and intra-column mesh; the second stage (in the second PowerMX) is the correlator (or beamformer and intra-column mesh if forming wideband beams).
Figure 6 Dual PowerMX boards daisy-chained and integrated into one pizza-box.

In this example, with 28 Gbps/pair of bandwidth, a 192-station correlator/beamformer could be built at 850 MHz/polarization, 8-bit samples (assuming the silicon is capable, of course; with more capable silicon the entire thing from an I/O perspective could fit on one PowerMX board, nicely illustrating the “future-proofness” of the concept).
Rather than using fiber modules and cable for interconnecting two closely spaced boards, “printed wiring” connections are established between the boards using the 100-pin Meg-Array connectors\(^3\). A quad connection board could be used, or a single Meg-Array connection board with a signal repeater/driver could be used (or a quad connection with repeater/drivers) to achieve performance. This little board, provided power by the Meg-Array connector is shown in Figure 7. The Vitesse VSC7227 is a 12-channel, 14.1 Gbps 13x13 mm BGA chip with automatic gain and equalization that could be used for this purpose (see: https://www.vitesse.com/products/product.php?number=VSC7227).

![Figure 7 Single Meg-Array, 12-channel repeater “printed wiring” connection board.](image)

By extension then, further concatenation/daisy-chaining of boards can be performed in the horizontal plane, and as well in the vertical plane (by stacking horizontal planes) using fiber modules and cables to interconnect planes (mounted in an “S-style” daisy-chaining configuration) or by installing the connectors and circuitry of Figure 7 on a flex PCB (FCI’s application data for this connector specifically mentions that it can be mounted on a flex PCB—see section 11.3 of http://portal.fciconnect.com/Comergent//fci/documentation/gs-20-033.pdf).

Horizontal planes of boards can be separated vertically in a number of ways. For a “stack” of (probably) less than 4, mount each PowerMX board to a metal (aluminum) substrate and stack these on top of each other with appropriate-length standoffs. This stack could be rack-mounted or in a standalone box as required for the particular application. The metal substrate provides heat spreading/sinking capability for the Linear Tech power supply modules (assuming they are mounted on the back of the PowerMX board) and allows each board and associated printed-wiring interconnects to be mechanically stabilized and fixed. For larger stacks, an internal horizontal shelf system might be used where the entire shelf becomes one 19” rack-mountable crate. Crate and chip cooling fans (or liquid cooling schemes) are integrated as appropriate.

For the case where multiple boards are daisy-chained for massive beamforming operations (e.g. SKA1-MID beamformer) secondary output (partial or full) beam data paths can be provided off

---

\(^3\) This is a nice feature of this connector: it can be used for a plug-in module or a board-to-board connection.
of the application mezzanine cards directly, for example, using the Avago MicroPOD 12-fiber transmitter module (see: http://www.avagotech.com/pages/en/fiber_optics/parallel_optics/minipod_micropod) coupled into one or more front or rear-panel fiber ports.

There are a myriad number of ways that PowerMX boards can be concatenated and used either integrated into larger boxes and crates or individually packaged and interconnected by external fiber networks. It therefore forms a “lego-block” that can be scaled in many dimensions as required by the application.
MULTI-BOARD PACKAGING SOLUTIONS—MESHING
When more PowerMX boards must be used for bigger problems and meshing/corner-turner operations are involved, there are several methods to provide connectivity. In the simplest case boards are stacked either within a crate or as separate 19” rack-mount entities, and fiber cable connects them together.

A simplified diagram (not containing power delivery or cooling methods) showing a concept for a dual stack of 4 PowerMX boards with fiber interconnect (corner-turner) mesh is shown in Figure 8. Only the cables for a single mezzanine card slot for each board are shown (16 cables); in a fully-connected maximum I/O configuration, 64 cables would be used. MT-connector intra-crate 12-fiber ribbon cables with associated fiber modules (such as the Reflex Photonics modules shown in Figure 10) are probably the best bet for minimum cable mass. Other copper-only connections may be possible, see Samtec http://www.samtec.com/cable-systems/active-optics/on-board-optical/firefly.aspx (ref. provided by Francois Kapp, SKA-South Africa).

While the interconnectivity in this example may seem unwieldy, the capacity (if the silicon is capable) is for a 4x192=768-station correlator (or beamformer) at (28 Gbps/pair), ~850 MHz/polarization 8-bit samples, further illustrating the future-proofness of the concept. If silicon is not capable of this capacity, fewer I/Os and mesh cables are used and the passive fiber interconnect method being used in the CSIRO ADE system (see Hampson, G., Advancements in Beamformer and Correlator Optical Backplane Technology, URSI Boulder meeting, 11-January-2013) can be used within the crate or by separating the F and X (beamforming) parts, spread across a rack.

Other corner-turner methods such as active Ethernet switches (see Kapp et. al. for the MeerKAT DBE) can also be used if active routing flexibility is required. If more X silicon resources are required, multiple PowerMX boards can be daisy chained as previously described, or daisy-chained in a nearest neighbour fashion in columns within a rack using the auxiliary FCI ExaMax connectors. There are many possibilities here.

If meshing needs exceed what can be provided with cable, the CSIRO ADE passive fiber mesh method can be used to provide very large meshes with applicability to SKA2. One example of this is shown in Figure 9. In this figure, connectivity for a 2304-station correlator/beamformer is provided with 16 vertically-mounted passive fiber mesh “cards” each one 12 x 12 MTP cables. The figure shows them integrated into one crate, however, the F, mesh, and X parts can easily be spread across a rack (or multiple racks) using “S-style” daisy-chaining connections. There are many possibilities here as well.
Figure 8 Eight PowerMX boards integrated in a dual stack of 4 with fiber cable interconnect mesh.
Figure 9 2304-station correlator in single crate using the CSIRO ADE passive fiber mesh. The F, corner-turner, and X parts could be spread across a rack (or multiple racks) in 3 parts with S-style daisy-chaining. Each 12x12 MTP cable optical cross-connect costs ~$1000 (Hampson, verbal quote).
INTRA-BOARD MESH

The PowerMX motherboard contains all mezzanine-to-mezzanine connections to allow for full mesh connections required for complete input data path usage for frequency-domain correlator and beamformer applications, in a nearest-neighbour (for performance) fashion. In this case, after the FFT, each mezzanine card device “keeps” 1/4 of the data (i.e. transmits it out one of its own associated transmit modules) and transmits 3/4 of the data to the other mezzanine cards (1/4 to each one), who each in turn transmits that data out one of its own transmit modules. Ideally, there would be direct connections from each mezzanine card to all of its neighbours, however, performance and routing congestion will likely make this direct connect approach impossible.

The number of mezzanine card device SERDES transceivers required to support \( n \) serial inputs (and outputs) (e.g. from the front panel and to the rear panel) is \( n \times 3 \). Current generation FPGAs have maximum 96 SERDES transceivers and so only 32 inputs can be supported. Devices with 72 SERDES channels can support 1/2 the input capacity of 24 serial lines per mezzanine card. It is possible to support all 48 inputs with 96 SERDES devices but each mezzanine card
design needs to be different\textsuperscript{4}, and each one requires 3, 12-channel driver chips (like the VSC7227) to do the necessary signal transmission without using the mezzanine card device (FPGA) and without invoking switching devices, which could be difficult to route and performance limiting. While this is possible, it is unwieldy and likely not feasible. Besides, current generation devices likely cannot implement 48 poly-phase filterbanks at the highest speeds with any substance (i.e. number of spectral channels). This can mean that motherboard data path capacities are over-designed, although this is sometimes a part of being future proof.

Note that the intra-column mesh on the board is there to be used but doesn’t have to be used. External meshing using the CSIRO ADE passive fiber mesh, or some other cross-connect could be used. However, the intra-column mesh does form an important first stage of a larger mesh, requiring only one other second stage (in the orthogonal direction) to be implemented to form a full mesh for even very large meshes as shown in Figure 9. If it is not used, more external meshing capacity is required and it could be more unwieldy than using the intra-board mesh.

For time-domain beamforming, meshing is not required; rather nearest-neighbour connections are used to collect partial sums-of-products to form entire beam sums. In this case the number of beams that can be formed will be limited by the number of nearest-neighbour connections, or full sums can be calculated by downstream boards.

\textsuperscript{4} That is the author’s conclusion; perhaps someone has some clever approach that was missed.
BLADE AND BACKPLANE MOUNTING POSSIBILITIES
The PowerMX I/O design is primarily based on the “bet” that ultimately fiber transmission between chips, with transceivers located near them will be the “way of the future”. This seems reasonable since higher data rates become more difficult to transmit over any distance on copper. However, manufacturers of backplane driver devices and connectors keep stepping up to the plate and extending this “inevitability” even further beyond what could have been imagined just a short time ago—it is not unreasonable to conclude that the jury is still out on this one. For example, the Vitesse VSC7227 device claims the ability to drive 40” of FR-4 material at 10 Gbps data rates, and the FCI ExaMax right-angle PCB connector claims 28 Gbps data rates with a design that virtually eliminates stubs and compensates for differential pair length mismatch over the 90 degree signal path corner of the connector.

Given this “arms race”, it is prudent to see if PowerMX might be mounted and used in a blade/backplane mounting application, or if such an application is beyond reach and better serviced with a completely different design. A copper “backplane transition module” concept, complete with VSC7227 drivers is shown in Figure 11. This, along with mounting PowerMX on an aluminum substrate provides backplane plug-in blade capability as shown in Figure 12. A similar approach could be used if fiber backplane technology matures, with copper right-angle PCB connectors replaced with fiber drivers/receivers and connectors.

![PowerMX copper backplane transition module](image)

Figure 11 PowerMX copper backplane transition module consisting of a PCB, 4 mating 100-pin Meg-Array connectors, VSC7227 drivers/receivers, and dual FCI ExaMax right-angle backplane connectors. A similar concept is possible if optical backplane technology becomes mature, in which the copper right-angle PCB connectors are replaced with fiber connectors and drivers/receivers.
Figure 12 Single PowerMX board mounted on an aluminum substrate with backplane transition module. The substrate, rather than the PCB, slides into guide rails. The signal path, in profile, is shown in red. Note that the backplane can be split into 4 identical backplanes.
Note that in Figure 12 with a full mesh backplane (or 4 separate identical ones—vertical meshing does not have to be performed on the backplane), full total mesh capability is provided since the first stage of the mesh occurs with intra-PowerMX nearest neighbour connections, and the second stage is provided by the backplane.

Positioning the VSC7227 devices close to the backplane connectors provides maximum backplane reach capability. The card shown in Figure 12 in a 25-slot backplane at 1U per slot, would span 43.75” (exceeding the current VSC7227 drive capability at 10 Gbps—future devices might be capable). If possible though, it would provide substantial meshing and computing capability, providing 100 devices of compute horsepower\(^5\) with 4800 28 Gbps inputs (on the front panel); in the limit of silicon capacity an entire 850 MHz/polarization correlator, which far exceeds even SKA2 correlator processing requirements could fit in a single (12U H x 210 mm D x 45” W) crate.

If more compute resources are required, deeper crates can be used with 2 or more PowerMX boards daisy-chained on each blade’s aluminum substrate as illustrated in Figure 6, each board processing part of the problem, and passing processed or unprocessed data through the daisy-chain until it all ends up at the backplane connector ready for the mesh.

Extending this concept further, F-boards (mounted on aluminum substrates) could plug into the front, and mesh to X-boards plugged into the back (with staggered connectors), with each one containing several daisy-chained PowerMX boards (within reason!) to provide massive correlation (or beamforming) capacity (in this case for 24 slots total). X-board front-panel fiber drivers could daisy-chain to other X-boards in other racks as with this capacity the correlator part is enormous and unlikely to fit into one crate, even a large one as described. It would be a very large crate, in one very large rack (~48” wide, several feet (10?) deep, 12U + cooling intake and output high), but it would have enormous capability, mostly with copper interconnects.

All of this with four primary blade ingredients, not including the backplane, crate housing, power delivery, and cooling:

- PowerMX board + mezzanine cards for the application.
- Aluminum substrate.
- Backplane transition module; if split into F and X boards, one transmit module, one receive module.
- Daisy-chain driver as shown in Figure 7.

\(^{5}\) Using the largest Virtex-7 devices with ~2.5 PMACs/s each, provides an unimaginable 0.25 PMACs/s compute capacity in a single crate. With 2 PowerMX boards concatenated 0.5 PMACs/s are possible.
One could argue that there are too many pieces to fit together, and that is a fair argument; however, the compute capacity, flexibility, and parameter space are enormous and can tackle a huge range of signal processing problems on a massive scale, in a relatively compact configuration. Assembly of the blade, from pre-tested sub-assembled modules should be fairly quick and inexpensive for a production worker to handle (~1 hr). The whole blade could then be functionally tested in a test bed before tag-and-bag shipping to a customer.

**OTHER I/O POSSIBILITIES AND APPLICATIONS**

Following the design concept of the backplane transition module of Figure 11, other transition or I/O modules for various applications could be developed and integrated with PowerMX in one or more of the packaging options discussed, and others not yet conceived.

An example application is shown in Figure 13. In this example, ADC transition modules contain A/D converters to allow direct RF/IF input to the complete board assembly. Shown are SMA connectors allowing 32 inputs—with SMAs on both sides and using both edges—128 inputs should be possible. Processed data can exit mezzanine cards via microPODs as previously discussed.

The ADC transition module concept would benefit from the ability to distribute an ADC-quality clock to transition modules via the front (or rear) panel clock inputs and application mezzanine card. Careful design of the clock network on the board (as described briefly on page 40), multiplier PLLs on the transition modules, and quiet clock net power supply design should allow for this. For ultra-low-noise mixed-signal applications, the board can be supplied with quiet +12V and each application mezzanine, M&C mezzanine, and I/O module can develop its own voltages as required, rather than using motherboard switching power supplies.

In the same manner as Figure 13 other transition modules can be developed (mixed-and-matched as required) and used for other purposes. Some possibilities are as follows:

- RF over fiber receivers and ADCs.
- Long-haul fiber transceivers.
- Clock fanout and distribution on coax or fiber; using 12-fiber modules, 384 clock outputs could be generated. This can be useful for AA-low station clock distribution or PAF clock distribution where ADCs are located near receiver elements rather than using RFOF.
Figure 13  PowerMX with ADC transition modules for direct RF/IF input.
MEZZANINE CARD FLEXIBILITIES

The application mezzanine card shown in all previous diagrams is ~3.5” x 3.5” and can carry a very large chip, associated memory modules, or smaller chips as shown in Figure 1. Most I/O is via the 3, 400-pin Meg-Array connectors to the motherboard, but secondary I/O can be obtained by using, for example, Avago microPOD fiber modules directly on the mezzanine card. Direct adjacent mezzanine-to-mezzanine copper connections may be established if a suitable connector or connection mechanism can be found (perhaps even direct solder-in “splices”).

In the case where the mezzanine card design and perhaps performance does not require (or can’t use) the SNAP12-A I/O modules, the mezzanine card can be extended to the edge of the motherboard or beyond to provide specialized I/O capability. This is a nice side effect of using the same Meg-Array connector stack height for the mezzanine card and the SNAP12-A I/O module. A simple example of such a mezzanine card is shown in Figure 14. Other possibilities exist, even including concatenating PowerMX modules together by bypassing one of the I/O module sites. Not shown in this diagram are additional mounting screw hole locations that could be used to attach to the SNAP12-A mounting holes and even aluminum substrate holes.

Figure 14 Extended mezzanine card. The SNAP12-A I/O module sites are unused.
TECHNICAL DETAILS
As in any concept, the “devil is in the details”, and there are several risks and challenges to overcome if PowerMX is to be possible as described in previous sections. This section discusses these issues.

400-pin Meg-Array Connector
The crux of the PowerMX concept, and the key to the flexibility it provides lies in the 4 mm stack 400-pin FCI Meg-Array connector (see Figure 17). While the PowerMX form factor and I/O could be used in a “hardened” (i.e. non-mezzanine) manner, much of its advantage comes from the mezzanine concept. There are several issues to be addressed:

- Connector-to-connector contact reliability and lifetime.
- Connector ball-to-contact reliability and lifetime.
- Performance (differential pair serial data rate).
- I/O capacity to meet functional and performance requirements (i.e. pinouts).
- Insertion/extraction force.
- Multi-connector use.
- Cost.

Each of these concerns will be briefly addressed here but of course full investigation, prototype testing and qualification is required. There is compelling evidence, however, that each of these issues can be adequately addressed providing some reasonable probability that it is an appropriate connector for this application.

See:
http://portal.fciconnect.com/portal/page/portal/fciconnect/mezzselect?s=highspeedmezzanine/Meg-Array

or:

http://portal.fciconnect.com/portal/page/portal/fciconnect/mezzselectmegarraytechnicaldocuments

for webpages with links to the full set of documentation for this connector.

Contact-to-Contact and Ball-to-Contact Reliability and Lifetime
On the FCI website extensive information is provided in this regard for this connector. Some highlights from that documentation are as follows:

- Time-tested reliability record includes Telcordia GR-1217-CORE and NPS-25298-2 use options. Telcordia is (tested) for Central Office like environments, and NPS-25298-2 is
(tested) for more harsh environments. The connector has been in production since ~1998.

- 17 Billion + lines shipped to customer satisfaction.
- Precious metal plating options satisfy varying customer needs for environmental resistance.

The increase in the number of contacts in a PowerMX board due to these connectors will be substantial and fundamentally this decreases reliability. This is mitigated by a factor of ~2 noting that ~1/2 the contacts are grounds (see Figure 16). Additionally, the connector has had a long product lifetime and has undergone extensive reliability testing and characterization. I’m attempting to obtain field reliability reports from FCI to see what the defect rate is in actual installed units. FCI’s FAE also stated that production of this connector is retained in the U.S. factory to ensure product quality and reliability.

While the increase in the number of contacts does fundamentally degrade reliability by some factor, there is a secondary mitigating effect that may have some importance. These are:

- It is easy to determine whether the motherboard or a particular mezzanine card is at fault simply by swapping out mezzanine cards and testing. This results in faster mean time to repair for a failed assembly.
- Lower cost-probability product of a destructive failure, in particular where large pin-count BGA devices are involved. If the motherboard or a mezzanine card must be reworked due to a solder-point or device failure the other unattached boards are not at risk.
The above statements cannot be made for a “hardened” version of the board\(^6\) and often times it literally comes down to guesswork when trying to decide which device contact is at fault if an X-ray or other image does not clearly spell it out. There will still be guesswork here, but less hardware is subject to heat cycle stress to test each guess.

**Differential Pair Performance**

Recently (Q1 2012), the connector was (simulation) tested and qualified, and subsequently advertised for 28 Gbps per pair operation, per OIF short reach specification CEI-28G-SR.

- According to the standard, 300 mm of 5 mil traces on FR-4 material; one connector at 2 spacings along the transmission line, two sets of vias—one before and one after the connector.
- Each active pair surrounded by grounds, both broad-side coupled (BSC) and edge-side coupled (ESC) were tested.
- No signal conditioning (i.e. equalization or gain) at the transmitter or receiver.

An example eye-diagram from this report is shown below in Figure 15:

![Eye-diagram](image)

**Figure 15** 28 Gbps Meg-Array eye diagram, BSC or ESC, no signal conditioning according to OIF short reach specification CEI-28G-SR.


---

\(^6\) Where chips are designed into the motherboard.
There are some differences between the PowerMX application and the test setup for the standard:

- There are two Meg-Array connections in PowerMX.
- For routing density, stripline transmission lines are likely required, degrading performance.
- Higher active-pair density (see Figure 16) than the standard. The FCI performance report does state that given the results achieved, higher density is likely feasible.
- PowerMX should have much shorter total transmission line lengths—probably 1/3rd of that in the standard. 5 mil traces may not be possible, however.
- SERDES transceivers have very high performance signal conditioning capabilities. Something that looks like a completely closed eye in a simulation or by probing with an oscilloscope can be recovered with these techniques. (I believe) this is mostly the case if inter-symbol interference due to reflections along transmission lines don’t dominate.

Given these differences, both better and worse than the standard, simulation testing and, if promising, eventual prototype testing is required to establish feasibility.

It should also be noted that to achieve this kind of performance the PowerMX motherboard and mezzanine cards will need to be carefully designed and modeled before fabrication and test. The 3D geometry of the entire signal path must be taken into account and carefully designed using the latest blind, buried, and microvia technology. “Best practise” design may no longer be enough to achieve performance. Probing with an oscilloscope to see the signal will be obsolete and device “chip-scope” (e.g. Xilinx) capabilities will have to be employed to determine eye characteristics.

I/O Capacity and Pinouts
A preliminary pinout assignment of the 400-pin Meg-Array connector has been developed to ascertain concept feasibility. The preliminary connector pinout diagram is shown in Figure 16. I/O connectivity and M&C functionality (just barely) fits within the number of pins and with grounding (or AC-grounding) of orthogonal adjacent pairs. Pinouts include the following functionality:

- 288 differential pairs for primary high-speed SERDES connections. 96 of these are for SNAP12-A module I/O, and the rest are for nearest-neighbour connections to allow full mesh capability across the board at maximum I/O use.
- 4 differential pairs for auxiliary 10G I/O. Notionally these route to 10G PHYs on the motherboard and connect to RJ-45 10G BASE-T connectors.
- 2 differential pairs for clock I/O.
- 2 differential pairs for serial M&C I/O to the M&C mezzanine connector.
• 4 voltage sense and 4 voltage set lines (voltage set lines, if open, set voltage outputs to minimum to avoid chip damage). These allow the motherboard Linear Tech power supply module voltages to be set and maintain required voltages on the mezzanine card.

• 4 voltage monitor lines and 2 temperature monitor lines. These double as AC ground pins (requiring ~100 nF capacitors to ground) and allow the M&C mezzanine card to monitor all voltages and 2 temperatures on the mezzanine card.

• A single “SD” pin for deadman thermal overload protection. Normally ground, and if open, shuts down all motherboard power supply outputs for this location.

• SDA and SCL for I2C bus communications to the board. Each mezzanine card self-identifies to the M&C mezzanine card before M&C performs any actions such as power sequencing (the self-identify circuit on the mezzanine card is powered off the +12V rail...) This function is similar to that provided by the FMC (FPGA Mezzanine Card) standard (ANSI/VITA 57.1). Each different mezzanine card will have a different ID, allowing drivers to be loaded and executed particular to that card. Other I2C devices may be on this bus on the mezzanine card.

• A single reset (RST) pin.

• JTAG (TMS, TDI, TDO, TCK) lines to allow the mezzanine cards’ FPGA(s)’ boot devices (and other devices) to be programmed. JTAG programming of devices is standard throughout the industry. JTAG lines can also be used to allow board JTAG testing after manufacture. Again, this functionality is similar to that provided by FMC modules.

**Insertion/Extraction Force and Multi-connector Use**

The FCI application specification (see: [http://portal.fciconnect.com/Comergernt//fci/documentation/gs-20-033.pdf](http://portal.fciconnect.com/Comergernt//fci/documentation/gs-20-033.pdf)) covers these issues. Multiple connectors may be operated in the manner proposed for PowerMX and FCI provides specific ball-pad layout and tolerance requirements to do so. Insertion/extraction force is kept minimum since they are “jack-knife” mated rather than direct co-planar mated.

The sample 400-pin connectors I obtained from FCI mated in this manner with the force akin to “squeezing a piece of chewing gum between your fingers”. Still, an important test is to build a site and mate/un-mate with 3 connectors to see if any issues show up. Refer to photos in Figure 17 for close-ups of the pin and dual-blade sockets. It is interesting that the body of the pin connector has slots for the blade sockets to fit into, presumably to help retain contact force without purely relying on metallic spring.

---

7 The FCI FAE also verified this was the case and thought there should be no problem with 3 connectors.
Figure 16 PowerMX 3 x 400-pin MegArray mezzanine connector pinouts.
Likely the mezzanine card pin & (motherboard) barrel high-current power supply connectors will have the largest extraction force as previous experience has shown that during extraction there is a large static coefficient of friction to overcome before they release. This may require special tooling or even attachment screws at the same end with a jack-screw insertion/extraction mechanism. At the very least an extraction tool, or appropriate method, will likely be required to avoid damaging any pins.

**Cost**
As mentioned previously the cost of a connector pair is ~$43 per mating pair, MOQ=250, for the DigiKey on-line catalog for the 50 Au parts.

Part numbers are:

- **Plug**, Telcordia GXT, RoHS: 84740-292LF.
- **Receptacle**, Telcordia GXT, RoHS: 74221-291LF.

**“SNAP12-A” 100-pin Meg-Array Connector**
This connector is identical in performance and reliability to the 400-pin connector of the previous section. It is used in the “SNAP-12” industry standard for (fiber) I/O modules (see: [http://www.physik.unizh.ch/~avollhar/snap12msa_051502.pdf](http://www.physik.unizh.ch/~avollhar/snap12msa_051502.pdf)).

In the SNAP 12 MSA mounting holes and pinouts are such that the transmit and receive modules are different and not amenable to higher performance (28 Gbps) operation. A proposed change in pinouts is therefore in order, as well to make the PowerMX board agnostic to signal direction, to allow for module-based power supply programming, and with +12V power pins to allow for power delivery to auxiliary modules as illustrated in Figure 11 and Figure 13. This, of course, means custom modules; a budgetary quote from Reflex Photonics indicated a ~$150k NRE to take their existing “LightABLE” module and package into a SNAP 12 form factor, with custom pinouts.

A preliminary pinout definition of the 100-pin Meg-Array connector is shown in Figure 18.

---

Figure 17 Photos of the FCI Meg-Array connector, pin (a), socket (b), pin closeup (c), socket closeup (d), underside BGA (e).
Figure 18 “SNAP12-A” 100-pin Meg-Array preliminary pinout.

It contains 12 differential pairs in a more optimized-for-signal-integrity arrangement; two M&C pins (which could be I2C instead of simple Fault and EN); VCC pins with Rset and Vsense pins to allow the module to set VCC supplied by the motherboard; and finally 3 +12 V pins to allow for the module to develop its own voltages if required. Each pin is rated for 0.45 A (30 °C rise), all contacts energized, and 2 A with just one contact energized. This allows a comfortable 8 W to be delivered to each module at +12 V at 50% current derating (more +12 V pins could be defined if desired).

The disadvantage of this connector and module is that it is not front-panel pluggable. Two points here: 1) front-panel pluggable modules might not be able to achieve 28 (and 40) Gbps per pair due to the inherent nature of the right-angle PCB connector they use (although the FCI ExaMax connector is a counterargument to this assertion), and possibly 12-pair density; 2) it may be possible to build into the front/rear panels a module retention mechanism such that replacement of a module requires power off, removal of cable, and removal/replacement of the panel and failed module, rather than completely removing the entire board assembly. The advantage, as previously stated, is that large mezzanine cards can be used if I/O modules aren’t (Figure 14).

**M&C Mezzanine Meg-Array Connector**

Putting M&C circuitry on a mezzanine card helps to “future-proof” the PowerMX motherboard and to allow for different options/choices for M&C. A preliminary pinout diagram of a 240-pin Meg-Array connector for this card is shown below:
Figure 19  M&C mezzanine card preliminary pinout definition.

Further detailed definition and work is required to nail down all M&C functions and granularity. A 300-in connector may be required.

**Clock Network**

Clocking is of ultimate importance in any design. With asynchronous networks, it is less of an issue since clocks are provided locally and rate-matching schemes ensures there is no one “clock master” anywhere.

In a radio astronomy system this is not the case. It is a coherent system and there must be a master timing reference to which all receivers and samplers in the array are synchronized. After sampling, coherent clocking is relaxed significantly and timing jitter need only be good enough to be within the processing system’s buffers and within each receivers locking capability. It is true that everything after sampling could be completely asynchronous and all processing elements derive clocks and time tags from incoming data.

However, it is useful to deliver a master timing signal to all processing elements to be used (or not used) as each one sees fit; this reference can be considered “perfect virtual time” at the array phase center, and all modeling (delay models, phase models, timetags/timestamps) are referenced to this time. Each processing entity (board, chip) need only consider its incoming data and this time reference to do everything it needs to do at the right time.

Additionally, if PowerMX is used for coherent mixed-signal processing (Figure 13) or coherent clock distribution (due to its huge fanout capacity), an ADC-quality clock distribution mechanism is desirable. It is also desirable to make the PowerMX motherboard clock distribution network agnostic to the clock signal, save for maximum frequency, such that it does
not depend on any particular clocking signal content and that each mezzanine card does what it wants with the clock.

Figure 1 show the PowerMX board with 4 right-angle PCB-mount SMA\(^9\) connectors, two on each side of the board. The baseline plan is that the front and back functions the same (i.e. one input, one output), with amplifiers/buffers and selectors allowing for distributing a select clock to each mezzanine card:

- 1 of 2 selector on input (front or back) and then fanout to the 5 mezzanine cards (including M&C). i.e. one SMA input is selected to be the clock source, but it can be switched to the alternate clock source if necessary.
- 1 of 5 selector (from the mezzanine cards), and transmit independently to each (front/back) output.

The lowest jitter, lowest skew, highest-frequency amplifiers/buffers will be used, powered by low-noise regulators off of the +12 V rail. This allows for multi-purpose use of the clock. If the clock signal contains embedded coding, it is up to each mezzanine card to use PLLs etc. to recover the clock and embedded signal as required.

**48 and 12 V Power**

The idea here is that the board can take -48 VDC or 12 VDC thus allowing the user to decide how to supply power. If the 48 V to 12 V converter is not installed, 12 VDC can be applied to the 12 V connectors directly; if the 48 V converter is installed, the 12 VDC lines could be run to another board not containing a 48 V converter. Access for 48 V is notionally via the front and back panels via blind-mate right-angle PCB connectors. Access to 12 V is notionally via straight header connectors on the board supporting mating connectors and cabling. Further investigation and use case study is required to defined the best general-use configuration.

**Monitor & Control**

The vision for M&C encompasses both supervisory and application-specific activities.

A layer of supervisory M&C must be built and supported to provide application-level users a common base upon which to build applications. This includes all V+T+P control and monitor activities, M&C mezzanine card SoC processor boot, identification of mezzanine cards, downloading appropriate drivers/parameters (such as power sequencing) at the supervisory level, programming mezzanine card bitstreams if applicable etc.

---

\(^9\) SMA is a likely standard; a small fiber module could be used instead if appropriate and robust. Further study required.
TECHNICAL DESCRIPTION: “PowerMX” VDRAFT April 10, 2013

(Note: not shown in any previous drawings is the likely advantageous ability to have a front-panel-settable X, Y coordinate (8? 16? bits each) for the board such that higher-level software can then use it to determine where in the SKA it resides, and what its function is.)

At the application level, the M&C SoC CPU identifies what each mezzanine site is and downloads appropriate drivers and software for connection/communication with host computers. For “future-proofness” it is likely the case the M&C to mezzanine communications are more along the lines of “poke and peek” to registers (with “packet” interrupts if required), whereas host to M&C SoC CPU are more along the XML or more abstract level of communication (although this is not strictly required). For example for FPGA to ASIC migration where the FPGA provides only partial capability (i.e. number of beams) host software does not have to change if it is built to be flexible in numbers of beams processed by a site, with details abstracted at the PowerMX SoC CPU level, but downloading ASIC (vs FPGA) device drivers.

There can even be SoCs or microcontrollers on application mezzanine cards.

RISKS AND CHALLENGES

There are several risks and challenges in implementing PowerMX as described. In order of decreasing severity, they are believed to be:

- Critical path high-speed signal integrity analysis and simulation. If at least 10 Gbps/pair is not achievable it is dead. If 28 Gbps/pair is not achievable, it is probably dead or might have to bank on not-yet-defined higher-speed PCB capability. If 40 Gbps/pair is not achievable it is still a viable concept. If 40 Gbps/pair is achievable it is a very “future-proof” concept. Both motherboard and mezzanine card modeling using real PCB build stackups/capability is required. This is the first order of business to establish feasibility before moving on.
- Critical path routing density, and PCB build cost estimates. To demonstrate the board can be routed, well, a board (or portion) must be layed-out and routed. Both motherboard and (representative) mezzanine cards must be designed.
- 400-pin Meg-Array issues: 3 connectors in parallel with barrel-and-socket power supply pins—mechanical model needs to be built and tested; contact reliability—acquire field contact reliability data from FCI if possible, study/scrutinize FCI reports in more detail, contact current and past users/references if possible. Build and test real prototype.
- Thermal and packaging solutions. Some notions of these have been presented. Further detailed mechanical design and development of thermal solutions is required to determine feasibility. Build and test.
- Refinement of functionality, performance, and use cases—are there show-stoppers making it inapplicable to large segments of SKA signal processing?
• Turn-key manufacturing and test. There don’t seem to be issues of concern here, but they must be explored. The board can by boundary-scan or functionally tested by plugging in loopback boards on the SNAP12-A sites and nominal functional mezzanine cards. For longevity, should the motherboard have the Meg-Array pin or socket connectors? How does this impact test mezzanine card lifetime and cost?

• At the assembly and sub-element test level, it seems that PowerMX boards and appropriate firmware can be used to test other (UUT) boards. Is this really the case, or is other surrounding test equipment required?

**WAY FORWARD**

Each of the risks and challenges from the previous section needs to be tackled. The nice thing is that they can be tackled in parallel with multiple interested groups working on them concurrently. Stage 1 of the SKA pre-construction phase would be tackling design and study of each of these in addition to seeing how the PowerMX platform might be used for each applicable SKA sub-element. Stage 2, if Stage 1 investigations are successful, would build and test prototypes to verify Stage 1 investigations. Also in Stage 2, Element-level system engineering can be engaged to plan for deployment of one or more packaging options into sub-elements.

There is much work to be done to establish the feasibility of PowerMX. This feasibility investigation can be distributed across organizations working concurrently on different aspects of the problem. The vision is of one platform deployed and used in different ways with multiple organizations providing expertise and development effort at levels that engage them and they are interested in. It is a vision of a coherent approach to SKA signal processing development, deployment, and operations.
SKA1-Survey Physical Implementation Proposal
Feasibility White Paper

A. Ensor\textsuperscript{1}, J. Bunton\textsuperscript{2}, B. Carlson\textsuperscript{8}, TN. Chan\textsuperscript{4}, G. Comoretto\textsuperscript{6},
L. D’Addario\textsuperscript{7}, S. Gulyaev\textsuperscript{5}, R. Hasan\textsuperscript{2}, G. Kloss\textsuperscript{1}, M. Mahmoud\textsuperscript{5},
O. Sinnen\textsuperscript{9}, and K. Zarb-Adami\textsuperscript{10}

\textsuperscript{1}AUT University
\textsuperscript{2}Center for Research in Analog and VLSI Microsystem Design, Massey University
\textsuperscript{3}Commonwealth Scientific and Industrial Research Organisation
\textsuperscript{4}Compucon NZ
\textsuperscript{5}Institute for Radio Astronomy and Space Research
\textsuperscript{6}Instituto Nazionale di Astrofisica
\textsuperscript{7}Jet Propulsion Laboratory
\textsuperscript{8}National Research Council Canada
\textsuperscript{9}University of Auckland
\textsuperscript{10}University of Oxford

3 May 2013

Abstract

This white paper establishes the foundational feasibility for a number of potential correlator solutions for the SKA1-Survey telescope. The specifications for SKA1-Survey given in the Baseline Design document SKA-TEL-SKO-DD-001-1 are used to investigate ASIC, FPGA, and COTS technologies that can be used to meet the processing demands of the SKA. A preliminary cost analysis is undertaken for some of the technologies expected to be commercially available by 2016 and their relative advantages and drawbacks are discussed.
1 Introduction

According to the SKA Baseline Design Document, the SKA1-Survey will be located within the boundary of the Boolardy station in Western Australia. It will primarily conduct surveys of large fractions of the sky, mapping the sky both in spectral lines and continuum. SKA1-Survey will be a mixed array of 36 12m diameter dishes of the ASKAP array and 60 15m SKA1 dishes. The SKA1 antennas will be arranged in a core with a diameter of about 2km to add density to the existing ASKAP array in the core. Three spiral arms will extend to a radius of about 25km from the centre.

The SKA1 dishes will be equipped with Phased Array Feeds (PAFs), similar or identical to those planned for the ASKAP antennas. SKA1-Survey will cover the continuous frequency range from 650MHz to 1670MHz in a single dual-polarised 500MHz wide instantaneous bandwidth. Signals from the PAFs will be beamformed at the antenna to form 36 dual polarisation beams. In forming the beams the first operation is a (coarse) channelizer that breaks the PAF signal into 500 1MHz channels. A complex "weight and add" is sufficient to form the beams for each of these channels. This beam data is transported to a Central Signal Processor (CSP) facility where it is further channelized to 2kHz channels and all cross-correlations between corresponding beams for all pairs of antennas are formed. Output data from the correlator in the form of visibilities will be transported to the Science Data Processor (SDP) centre in Perth where it is transformed into images and spectral cubes.

The Context Diagram for the Central Signal Processor associated with the SKA1-Survey is show in Figure 25 (Page 74) of the Baseline Design, and included here as Figure 1.

The main inputs are the data for 36 beams from the 96 survey antennas and the main output are the visibilities generated by the correlator which are sent to the Science Data Processing located in Perth. In normal operation the system also receives a time reference signal in addition to the timing data that are embedded in the Phased Array Feed (PAF) beam data. Optional features that will be included in the design are a single external transient trigger and the output of an array beam. No consideration is given to spigots for third party equipment as the function and form of the requirements for this equipment is currently undefined. It is assumed that the system is housed in racks that provide power and cooling.

The main input to the CSP is the 36 beams from each of the 96 antennas. Each beam is a dual polarisation beam with a bandwidth of 500MHz, so the total bandwidth over all the signals from a single antenna is $36 \times 10^9$Hz. It is assumed the antenna is using a frequency domain beamformer, the channel width at which beamforming occurs is 1MHz, and each sample is a complex 8+8 bits. This data must be processed by a second stage (fine) filterbank in the CSP to bring the data to the final frequency resolution of approximately 2kHz. The data from the antennas are oversampled to avoid aliasing. An oversampling ratio of $\frac{8}{7}$ is assumed in this paper, an engineering trade study is needed to determine the oversampling ratio that will be used. The input into the SKA1-Survey CSP is:

$$36 \times 10^9\text{Hz} \cdot (8 + 8\text{ bits}) \cdot \frac{8}{7}\text{oversampling} = 658 \times 10^9\text{b/s} = 76.6\text{GB/s}$$
to the CSP from each of the 96 antennas, or a total data rate of $63.2 \times 10^{12} \text{b/s} = 7357 \text{GB/s}$ from all antennas. If 40Gb/s links are used to transport this data where each link achieves a throughput of 96% of capacity, and the $\frac{3B}{10B}$ encoding is taken into account, then 20 links are sufficient to transport the data from each of the 96 antennas. The fall back position is to use 56Gb/s links, and it is expected that next generation FPGAs will support the 14Gb/s I/O needed for these links. If this technology is sufficiently mature then the number links from each antenna is reduced to 16. As will be seen later a simple correlator structure is possible if 16 links are used.

As the beamformers for the PAFs are assumed to be frequency domain beamformers, each individual module will be processing part of the total bandwidth and producing all beams for that bandwidth. Each data link carries data for all 36 beams for up to 32MHz of bandwidth. The actual number of links and precise correlator structure will ultimately depend on the actual performance of the next generation processors selected and the current number should only be considered as an indication of what is possible.

---

$^1$In this paper 1GB is taken to be $8 \times 2^{30}$ bits, likewise for MB and TB
2 Subsystem Description

2.1 Function Architecture

The basic functional architecture for CSP Survey is shown in Figure 26 (Page 75) of the Baseline Design, and included here as Figure 2. As we will see below it requires revision.

![Functional Architecture Diagram](image)

Figure 2: CSP SKA1-Survey Functional Architecture according to the Baseline Design. For more details see Figure 3 where the revised architecture is provided.

Additional functions that must be included are integrator, cross-connect, delay correction and correction for fringe rotation. The channelizer is a major task and for each input it must process 32 1MHz frequency channels for 36 dual polarisation beams, a total of 2304 signals. A 512 channel filter bank is needed to take the 1MHz data to approximately 2kHz.

2.2 Buffer

The memory requirement for this could consume excessive resources in a processor such as an FPGA. Instead the corner turn and buffer operation can be combined. Small corner turn operations before and after the buffer as well as differences in the read and write sequences can be used to reorder the input data in 1152 sequential data streams. Each data stream carries data for a single dual polarisation beam for one 1MHz channel. All data are processed in sequential order and each filterbank in the channelizer processes a single dual polarisation 1MHz channel at any one time.

The buffer is in DRAM and operates as a circular buffer. The memory address control for this buffer has the buffer freeze signal connected to it and a buffer freeze operation
is a modification to the buffer addressing. The fine filterbank will process sufficient data to output about 512 samples of each 2kHz channel before starting processing on another 1MHz channel. This corresponds to 256ms of data. In the buffer memory this data must be at least double buffered so storage equivalent to 512ms of data is required. This is equivalent to:

$$512\text{ms} \cdot 63.2 \times 10^{12}\text{b/s} = 32.4 \times 10^{12}\text{b} = 3.68\text{TB}$$

The buffer is also used to implement delay correction. The delay correction for each beam is different to the others. This arises because each beam is pointing to a different part of the sky. Hence, it is not possible to apply a single delay correction before the antenna data is processed in the beamformer. It must be applied to the 1MHz channelized data. The method of applying the correction is to modify the read address from the buffer. The precision of the correction is approximately 1µs. As the maximum baseline length in SKA1-Survey is 50km, if the antennas have field of view to the horizon then a delay of up to 167µs is required to accommodate the worst case wavefront travel time. Presuming the differences in fibre delay are up to three times this requires a total delay of up to 667µs of samples to be buffered. This is equivalent to:

$$667\mu\text{s} \cdot 63.2 \times 10^{12}\text{b/s} = 42.1 \times 10^{15}\text{b} = 4.91\text{GB}$$

storage required in the buffer to implement delay correction. This storage required in the buffer to implement delay correction is negligible compared to the corner turn buffering.

### 2.3 Channelizer

Fine delay correction is applied to the data at the output of the channelizer. It takes the form of a phase slope that is applied across the 1MHz bands. This slope is the same for all 1MHz channels for a given beam. This property is used to reduce the memory requirement for the fine delay correction. This system will also implement the correction for fringe rotation.

The use of two cascaded filterbanks reduces the on FPGA memory requirements. For a polyphase filterbank the amount of on FPGA storage is proportional to the length of the filter and for the 250,000 channels the storage would be 5MB for data and half that for the filter coefficient. The same frequency resolution is achieved with two cascaded 512 channel filter bank and storage requirements are reduced to about 30kB. However, second stage filterbank can only process a single 1MHz channel at a time. This requires a corner turn operation ahead of the second filterbank so that it processes successive blocks of data. Each block has 1000 time samples for a single 1 MHz channel for a single dual polarisation beam.

The channelizer and fine delay have a computational requirement for each complex sample consisting of 6 CMAC for the FIR part of the filterbank, 3 CMAC for the 512 point FFT, and 1 CMAC for the fringe rotation, so 10CMAC in total per complex input sample. The data rate into the channelizer is given by:

$$96\text{ antennas} \cdot 36\text{ beams} \cdot 2\text{ polarisations} \cdot 500 \times 10^6\text{ samples/s} \cdot \frac{8}{7}\text{ oversampling}$$
or $3.95 \times 10^{12}$ complex samples per second. Hence, the compute load for the filterbank, FFT and delay/fringe corrections is $39.5 \text{TCMAC/s}$. With $8+8$ bit complex samples the data rate into the channelizer is:

$$3.95 \times 10^{12} \text{samples/s} \cdot (8 + 8 \text{ bits}) = 63.2 \times 10^{12} \text{b/s} = 7357 \text{GB/s}$$

Although not shown on the functional architecture diagram the context diagram indicates that a VLBI array beam is also required. Assuming the VLBI array beam is implemented on the array beam then the data for the correlator can be critically sampled. This reduces the compute load of the correlator at the expense of a small decrease in correlator efficiency. A trade study is needed to determine the balance between oversampling and correlator efficiency. In this paper it is assumed no oversampling is needed for the correlator data. This reduces the data rate after the channelizer to $55.3 \times 10^{12} \text{b/s}$ ($6437 \text{GB/s}$).

### 2.4 Correlator

At this point the data are ready for the correlator operation. But the antenna data are spread across multiple processing modules. A cross connect\(^2\) operation is now needed to bring data to each module from the same frequency channels in all antennas. After the cross connect each subsystem and subcomponent in the correlator has all data needed to implement the correlation operation.

In SKA1-Survey there are $96 \times 95 \div 2 = 4560$ baselines, and $2 \times 2 = 4$ polarization pairs to correlate per baseline. Presuming critically sampled data, each of the 36 beams has $500 \times 10^6$ samples/s, so the correlator must perform:

$$36 \text{ beams} \cdot 500 \times 10^6 \text{ samples/s} \cdot 4560 \text{ baselines} \cdot 4 \text{ polarization pairs}$$

CMAC per second, which is $328 \text{TCMAC/s}$. This corresponds to the Baseline Design (Page 72) estimate of $1313 \text{TMAC/s}$ (computed for 2.5 GHz bandwidth). Each correlation product is presumed to give an $8+8$ bit output, so the correlator outputs $328 \times 10^{12}$ correlations/s \cdot (8 + 8 bits) = $5.25 \times 10^{15} \text{b/s}$ ($597 \text{TB/s}$), which is immediately integrated (so reducing the data rate).

### 2.5 Integrator

The integrator simply performs $328 \times 10^{12}$ complex accumulates per second to reduce the noise to signal ratio. This is accumulated over a minimum dump time, but can not be too long so that the rotation of the Earth has a significant effect on the baseline. After integration of the correlator data it is dumped to SDP.

The Baseline Design Page 71 states the array configuration for SKA1-Survey as given in Table 1. This information can be used to obtain Table 2, the number of baselines that are within different lengths $B_{\text{max}}$ and the minimum dump time $\Delta t$ for a station diameter

\(^2\)Note a cross connect need no storage as it simply redirects where data are transported to. In contrast a corner turn needs storage, as it typically absorbs all data presented at a given time and redirects that data to a single output.
Table 1: Array configuration for SKA1-Survey

<table>
<thead>
<tr>
<th>Antenna Position</th>
<th>Number Antennas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antennas within 400m of core</td>
<td>11</td>
</tr>
<tr>
<td>Antennas between 400m and 1000m of core</td>
<td>18</td>
</tr>
<tr>
<td>Antennas between 1000m and 2500m of core</td>
<td>31</td>
</tr>
<tr>
<td>Antennas between 2500m and 4000m of core</td>
<td>15</td>
</tr>
<tr>
<td>Antennas between 4000m and 25000m of core</td>
<td>21</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>96</strong></td>
</tr>
</tbody>
</table>

\[ \Delta t = 1200 \frac{D_{\text{station}}}{B_{\text{max}}} \]

Table 2 is consistent with the Baseline Design requirement to use a minimum dump time

Table 2: Minimum dump times for different baselines

<table>
<thead>
<tr>
<th>Baseline (m)</th>
<th>Number Baselines</th>
<th>Minimum Dump Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-800</td>
<td>55</td>
<td>22.50</td>
</tr>
<tr>
<td>800-1400</td>
<td>198</td>
<td>12.86</td>
</tr>
<tr>
<td>1400-2000</td>
<td>153</td>
<td>9.00</td>
</tr>
<tr>
<td>2000-2900</td>
<td>341</td>
<td>6.21</td>
</tr>
<tr>
<td>2900-3500</td>
<td>558</td>
<td>5.14</td>
</tr>
<tr>
<td>3500-4400</td>
<td>165</td>
<td>4.09</td>
</tr>
<tr>
<td>4400-5000</td>
<td>735</td>
<td>3.60</td>
</tr>
<tr>
<td>5000-6500</td>
<td>465</td>
<td>2.77</td>
</tr>
<tr>
<td>6500-8000</td>
<td>105</td>
<td>2.25</td>
</tr>
<tr>
<td>8000-25400</td>
<td>231</td>
<td>0.71</td>
</tr>
<tr>
<td>25400-26000</td>
<td>378</td>
<td>0.69</td>
</tr>
<tr>
<td>26000-27500</td>
<td>651</td>
<td>0.65</td>
</tr>
<tr>
<td>27500-29000</td>
<td>315</td>
<td>0.62</td>
</tr>
<tr>
<td>29000-50000</td>
<td>210</td>
<td>0.36</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4560</strong></td>
<td></td>
</tr>
</tbody>
</table>

\(D_{\text{station}}\) (taken to be 15m for SKA1 dishes, although the ASKAP dishes are actually 12m diameter), given by:

Table 2 is consistent with the Baseline Design requirement to use a minimum dump time of 1.4s for 10km baselines, and 0.3s for 50km baselines. If baseline-specific dump times are used in the integrator then the mean number of dumps per second can be calculated to be 0.7927 dumps per second (based on average 1.26s minimum dump time calculated from Table 2). A more accurate value will be possible once the SKA1-Survey antenna geometry is known. However, it should be remembered that using dump times that vary with the baseline length, while convenient for reducing the amount of data sent to the
SDP, can be impractical for some integrator technologies, increasing the cost more than the savings they provide to the SDP.

Presuming an average for each baseline and polarization pair of 0.7927 dumps per second, and 32 + 32 bit summations results in:

$$36 \text{ beams} \cdot 4560 \text{ baselines} \cdot 4 \text{ polarization pairs} \cdot 0.7927 \text{ dumps/s} \cdot (32 + 32 \text{ bits}),$$

or $$33.31 \times 10^6 \text{b/s (3.97MB/s)}$$ dumped to the SDP per channel. For a 500MHz signal with 2kHz frequency resolution requires 250000 frequency channels, so using $$2^{18}$$ frequency channels gives a total integrator output of $$8.73 \times 10^{12} \text{b/s (1017GB/s)}$$. Using the Baseline Design requirement of a minimum dump time of of 1.4s (presuming 10km baselines) gives instead 915GB/s, or of 0.3s (presuming 50km baselines) gives 4275GB/s (close to the Baseline Design values of 934GB/s and 4670GB/s respectively). With integrator technologies that cannot easily accommodate variable dump times, using a minimum dump time of 0.36s (2.78 dumps per second) corresponding to the longest Survey baseline gives $$30.6 \times 10^{12} \text{b/s (3560GB/s)}$$ dumped to the SDP.

The integrator must have sufficient memory to hold:

$$36 \text{ beams} \cdot 4560 \text{ baselines} \cdot 4 \text{ polarization pairs} \cdot 2^{18} \text{ channels}$$

32 + 32 bits accumulators, requiring at least $$11 \times 10^{12} \text{b (1283GB)}$$ storage. However, performing accumulations of 8 + 8 bit samples at a rate of $$500 \times 10^6 \text{ samples/s}$$ would potentially overflow a 32 + 32 bit accumulator after 34ms. If the full 8-bit accuracy is to be retained then a hierarchy of two accumulators can be used (with a modest increase in the required CMAC), which doubles the memory requirements of the integrator.

### 2.6 Subsystem Summary

Figure 3 summarises the important numbers for SKA1-Survey that are used in this section (see 2.1-2.5). It also shows the main elements of the CSP SKA1-Survey, data throughput between these elements, computational requirements and memory requirements for each element (if not negligible). Connections with other SKA systems (Timing, Telescope Manager, Dishes, SDP) are indicated. Note that for consistency with ASKAP the coarse channelization is presumed to take place at the PAF rather than at the central signal processor. Its processing requirements are similar to those of the fine channelizer described earlier.
Figure 3: CSP SKA1-Survey Revised Functional Architecture
3 Technologies

As SKA1-Survey extends the ASKAP array by a factor of three it is natural to consider the correlator technology that is already employed in ASKAP, namely the Redback correlator, and its evolution. In addition, several other significant technologies are considered. The PowerMX platform is intended as an “in-SKA” common motherboard which accepts mezzanine plugin cards for FPGA, ASIC, and CPU/GPU processors, and intended to span several generations of processor technologies. COTS homogeneous and heterogeneous CPUS are also considered because software based correlators such as DiFX have shown their viability. Also COTS GPU video cards are considered, which have been widely used in HPC and increasingly so for radio astronomy signal processing.
3.1 Redback

The proposed processing module is based on the Redback-3 module used in the ASKAP beamformer. It will be built on 1U chassis with ATX power supply. At the rear is a set of fans to cool the hardware, whereas all data, clock and timing connections are at the front of the module. Also at the front is the command and control board which implements shelf control and data downloads. Behind the front panel is a fibre management module. The front panel is perforated at greater than 30% to allow air flow from the rear to the front. The Redback board is mounted in the remaining space and uses FPGA processors by default. Each FPGA has four DDR4 and two pairs of 12 multimode fibre optical RX and TX. For connection to the data from stations or antennas each FPGA also has connections to four QSFP+ cages. A possible layout of the module with all components in proportion is shown in Figure 4. The QSFP+, DRAM and multimode fibre optical transmitter are all socketed and can be populated as needed either to match the performance of lower computation capacity FPGAs or systems with a lower I/O intensity. The I/O intensity is inversely proportional to the number of antennas. The I/O intensity of this board is designed to meet the needs of SKA1-Survey. SKA1-Mid needs about one quarter of the I/O on this board and SKA1-Low about one tenth.

![Figure 4: Possible layout of proposed Redback-5 processing module.](image)

The command and control has been separated out as another board. This board has an FPGA and multimode fibre transmitters and receivers for timing, clocks, command and control and data downloads. There are three SFP+ card cages for 1 or 10G Ethernet and
two optical receivers for clock and timing. The SFP+ slots are populated as needed. In the simplest configuration it has a single 1GE link to provide command and control. More links can be added for higher capacity data product upload or downloads. The eight fibres for these interfaces are amalgamated onto a single 12 fibre ribbon. The only electrical connection is a JTAG input and output at the front of the board. This connection is used for testing purposes. All JTAG capable devices are accessible from these connectors on the front panel. An additional function of the command and control board is shelf control. It monitors temperatures in the 1U chassis and uses this to control the fan speeds. If the temperature increases to dangerous levels for the Redback board it deprograms the FPGA and if this is insufficient it commands the ATX power supply to power down the processor board. While the Redback processor is powered down the command and control board can still monitor its temperature and report this to local monitor and control.

It is estimated that a design can be implemented in 14nm FPGA devices. The estimated performance of a midrange FPGA is at this process step is 3.5TMAC/s at a cost of €800. Power dissipation is 10.5W/TMAC/s and this includes serial I/O and DDR4 I/O. This gives the six FPGAs on a board a compute capacity of 21TMAC/s or 5.25TCMAC/s. At 80% usage this becomes 4.2TCMAC/s. I/O per FPGA will also increase from just over 30 to about 100. For this paper each FPGA is presumed to have the following connections.

Table 3: FPGA SERDES connections

<table>
<thead>
<tr>
<th></th>
<th>SERDES per link</th>
<th>FPGAs connected to</th>
<th>Links per FPGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>For paired FPGA</td>
<td>20</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>For triples</td>
<td>14</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>For other FPGAs</td>
<td>7</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>For 40Gb/s link</td>
<td>16</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>For multi-mode fibre</td>
<td>24</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td><strong>Total Links</strong></td>
<td></td>
<td></td>
<td><strong>102</strong></td>
</tr>
</tbody>
</table>

If this exceeds the available SERDES then the fall back position is to provide only 7 links between every pair of FPGAs. This reduced the count to 75. Each link can run at up to 14Gb/s and possibly 28Gb/s but here 10Gb/s is assumed. This has already been proven on the Redback-3 system. With these connection the FPGAs can implement a full cross connect simultaneously for the 40Gb/s input data and the data from the 12 multimode fibre receiver. The 40Gb/s link is WDM with four optical frequencies. Each frequency and each multimode fibre have been demonstrated to support 10Gb/s each. Operation at 14Gb/s and higher is possible but this has not been demonstrated by the authors at this time. Each FPGA has 4 DRAM. This will be DDR4 at the time of the hardware build. This is expected to operate at 2.066GHz giving a transfer rate of up to 132Gb/s per DRAM. If the DRAM achieves 80% usage then the total I/O is 2.54 Tb/s for the module.

The estimated performance of the Redback processing module is given in Table 4.
### Table 4: Redback Performance Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute capacity (18bit)</td>
<td>5.25 TCMAC/s (4.2 TCMAC/s at 80% usage)</td>
</tr>
<tr>
<td>DRAM (DDR4) I/O</td>
<td>3.173 Tb/s (2.54 Tb/s at 80%) unidirectional, halve for R/W</td>
</tr>
<tr>
<td>DRAM memory</td>
<td>24 DDR4 (196 byte with 7GB DIMMS)</td>
</tr>
<tr>
<td>Multi-mode fibre bidirectional I/O on Redback board</td>
<td>1400 Gb/s at 10 Gb/s per fibre</td>
</tr>
<tr>
<td></td>
<td>2000 Gb/s at 14 Gb/s per fibre</td>
</tr>
<tr>
<td>Single mode fibre bidirectional I/O on Redback board</td>
<td>960 Gb/s with QSFP+ at 40 Gb/s</td>
</tr>
<tr>
<td></td>
<td>1344 Gb/s with QSFP+ at 56 Gb/s</td>
</tr>
<tr>
<td>Data download/upload</td>
<td>Up to 3 links 1 or 10 Gb/s each</td>
</tr>
</tbody>
</table>

### 3.2 PowerMX

PowerMX is a concept for a common FPGA, ASIC, and CPU/GPU coprocessor platform for all SKA1 and SKA2 signal processing. At its heart PowerMX simply provides high-performance access to abundant chip resources, both processing and I/O, packaged within a medium-granularity form factor. PowerMX is a motherboard accepting mezzanine plug-in cards that are carriers for the latest chips required for each particular application. It has enough user-defined I/O capability that it will likely take a decade or more for silicon performance to catch up. It prescribes a specific intraboard interconnect fabric upon which layers of firmware and software can be built, but allows for interboard connectivity suitable and scalable for each particular deployed application.

Figure 5 is an annotated diagram of the PowerMX concept showing major functional blocks. There are 4 application mezzanine cards sites, each card about 89mm on a side and large enough for a very large (45mm) chip and memory modules or 4 smaller F672 27 × 27mm chips and memory modules. Monitor and Control is provided by a smaller mezzanine card; shown on this card is a Xilinx Zynq SoC (System on a Chip) FPGA which mates an FPGA with a hard ARM CPU. “SNAP12-A” panel-ready modules at 12 serial I/Os each provide the board with its bulk I/O capacity of 384 serial (fiber) links. Dual auxiliary 10G links are provided to each mezzanine card for lower performance I/O such as transfer of integrated coefficients or filter coefficients. The board is about 434 × 170mm and is therefore suitable for mounting in a 19inch rack-mount pizza box; blade mounting/use is also considered.

### Table 5: PowerMX Performance Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>—</td>
<td>€1630</td>
<td>—</td>
<td>10W</td>
<td>—</td>
<td>&gt; 1.92 Tbps</td>
</tr>
</tbody>
</table>

13
3.3 Application-Specific Integrated Circuit

An application-specific integrated circuit (ASIC) is an integrated circuit that is custom designed for a specific purpose. The hardware complexity of ASICs are usually expressed in GE (gate equivalents, or logic gates, a term in digital VLSI Design) unlike the number of chips in case of CPU, GPU or FPGA. FPGA implementations are also sometimes specified in logic gates as it can be moved into next generation higher gate density products. This is because any specified ASIC implementation in GEs is usually considered to be foundry-independent, technology scalable, and portable to next generation process technology. For example, an ASIC with 250K GEs that can be implemented on a 6mm × 6mm chip in a 130nm CMOS technology (2006), can be easily implemented on a much smaller 3mm × 3mm die, in for example, a 65nm CMOS process (2008). One GE is usually considered to be a 4 transistor logic gate (for example, a 4 MOSFET XOR gate). This subsection describes one preliminary ASIC chip design for the CSP.

Each chip will use at least 45nm CMOS in at most 10 mm × 10 mm. Prototype chips will cost at most €15,000. Production cost will be much less (depending on volume needed). Package and I/O pins PGA257 pin grid array have 257 pins, including 250 I/O data pins. The clock speed is 2GHz. Each MAC (16 bit pipelined multiplier accumulator) at 2GHz can perform 2 billion MACs per second. Each chip will have 500000 MACs so that the chip can compute $10^{15}$ MACs per second. So, maximum compute speed is 1000 TMAC/sec. It will be mostly fully pipelined dynamic DFF possibly with some vendor-
independent DRAM and SRAM (avoiding large DRAM) for intermediate storage.

Considering the pre-construction phase for SKA Survey well into 2016, and a continuation of Moore’s Law of Integrated Circuit density, it will be possible to easily integrate 1 billion logic gates on a 12mm × 12mm die for Planar (or 3D FinFET) CMOS technologies nearing the physical scaling limit at around 10nm node. TSMC foundry has already achieved up to 18nm node (18nm MOSFET channel length). At such densities at least 500000 MAC (multiplier accumulators) can be co-fabricated on a single chip (and maybe more depending on the multiplier algorithm used). This will easily allow a survey correlation of 36 dual polarized beams from 96 antennas in a fully pipelined implementation using at least a SKA-Survey ASIC chipset. Considering a power dissipation per logic gate of 2femtoWatt (using pass-transistor logic, just NMOS or just PMOS instead
of classical CMOS logic circuits) the total power dissipation by one billion on-chip core logic gates will be only around 2μW (assuming 0.6V power supply at 2016 technology near 10nm). Making allowances for substrate leakage, I/O pad driving (1 to 3pF off-chip load) the total power dissipation will be under 0.1W. A strategy similar to that of fabless semiconductor houses should be undertaken whereby dependency on any particular fabrication foundry can be avoided. In addition the circuit design should be easily scalable for supply voltage and process technology. The ASIC implementation can be emulated in FPGA boards (available from Altera or Xilinx) which contain over 100 million logic gates (in 2013 technology). One important trade-off in the ASIC implementation is memory versus processing units. By using a fully pipelined architecture the need for memory (particularly DRAM) can be reduced at the cost of more processing units (MACs). For technologies approaching 10nm node, the overall cumulative cost of more on-chip DRAM will outweigh the cost of more processing units due to several reasons:

1. ASIC implementations requiring significant DRAM overhead will be difficult to conveniently emulate on FPGA boards.

2. DRAM at technologies near 10nm will incur more foundry dependency than present day technologies (due to deteriorating gate and bulk leakage issues with finer deep nano-metric nodes).

3. With more MACs faster processing speed is possible without significant intermediate storage which incur multiple cycles and refresh cycle dissipations.

4. These are poorer wafer yield with significant on-chip DRAM, raising unit cost.

5. Process migration (portability) will incur additional cost overhead due to dependency on a foundry for DRAM verification. Hence, any DRAM used should be conservative in design so as to retain reliability with cross-foundry process migration.

An estimation of hardware cost can be made in GE (number of gates) for an ASIC implementation of SKA1-Survey which can be placed on a PowerMX mother-board. SKA’s Survey Telescope CSP is estimated to consist of 96 dishes with 36 double polarized beams. The bandwidth is 500 MHz along with 262144 frequency channels. The data is 8-bit complex with each complex correlation product requiring 4 multiplications. Based on the above, the ball-park hardware cost for a fully pipelined ASIC Survey channelizer implementation in GE (logic gates equivalent) is as follows:

- Cross-Correlator: 50 million GE.
- Total of all complex point FFTs (or, bit-slice DFTs using distributed arithmetic): 10 million GE.
- Total of all Polyphase Filters (also, possibly using distributed arithmetic): 20 million GE.
- Total: 80 million GE.
This doesn’t include front end ATODs, intermediate geometric delay elements and corner turning over head. Now even if we increase this number by a factor of 2 to 160 million GE, assuming dissipation by one logic gate to be 2nanoWatt (conservative estimate at say 65nm node CMOS for pure ASIC), the power dissipation by the pipelined CSP is 0.32Watt.

The Altera DN7020k10 can be used to emulate the ASIC solution. The DN7020k10 is a complete logic prototyping system that enables ASIC or IP designers to prototype logic and memory designs for a fraction of the cost of existing solutions. A single DN7020k10 configured with 20 Altera Stratix IV 4SE820 FPGAs can emulate up to 130 million gates of logic as measured by a conservative ASIC gate-counting standard. All FPGA resources are available to the target application.

Table 6: ASIC Performance Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom</td>
<td>2016</td>
<td>€1500</td>
<td>1000 TMAC/s</td>
<td>0.32 W</td>
<td>—</td>
<td>500 Gb/s</td>
</tr>
</tbody>
</table>

3.4 Field-Programmable Gate Array

An FPGA (Field Programmable Gate Array) is a semiconductor device (chip) that consists of inherently undefined, reconfigurable logic. In contrast to general purpose processors (CPU) or graphic processing units (GPU), their programming (or configuration) does not define how certain existing blocks behave, but rather define the functionally and the interconnection network from scratch.

To achieve that, FPGAs consists of a large number of configurable logic gates (hundreds of thousands) paired with registers, digital single processing (DSP) blocks (e.g. 18 x 18 bits multiplier), embedded fast static memory (SRAM) and high-speed transceivers to connect to the outside. All these components are connected with a large and dense routing fabric. The programming (or configuration) of the FPGA establishes the functionally of the logic blocks and the connection and routing of all the components. Like CPUs, FPGAs can also interface to external DRAM memory or other peripherals.

FPGAs are heavily used in current radio astronomy installations due to their high performance, low power in comparison to CPUs and due to their shorter development time and flexibility in comparison to application specific integrated circuits (ASICs). The Redback cards are an example of such employment in the ASKAP project.

Several manufacturers offer FPGAs with a wide range or characteristics, sizes and costs. In the context of the SKA project, only two manufacturers (Altera and Xilinx) seem appropriate due to the performance of their FPGA devises, the development tools and infrastructure for their products and the relative security in long term availability of their products.

The performance of an FPGA cannot be specified in the same simple way as for CPUs and GPUs, which is usually based on software benchmarks. Due to the extreme
configuration flexibility of FPGAs, provided functionality can be extremely tuned to the targeted problem. The major points that influence the performance of FPGAs are: number of (equivalent) logic gates, number of DSP blocks, amount of embedded memory, number/bandwidth of high-speed transceivers and the maximum operation frequency. Performance statements made by manufacturers in terms of GMAC/s are usually measured with very simple FIR filter applications.

FPGAs from the two major manufacturers Altera and Xilinx are produced with similar semiconductor processes and node sizes as used for modern CPUs and GPUs. Current generation is 28nm and will be 20/14nm within the pre-construction scope. The maximum designed power consumption is with around 50 Watt significantly below that of current generation CPUs and GPUs (120-150 Watt). The following table specifies characteristics of the high end series of both Altera and Xilinx. In large scale designs such as the SKA project often mid-range devices are used for their better price/performance ratio. It can be expected that the next mid-range generation will be comparable or exceed the current high end.

Table 7: Examples of mid-range FPGA devices

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-range</td>
<td>2016</td>
<td>€800</td>
<td>3.5 TMAC/s</td>
<td>37 W</td>
<td>84 Mb</td>
<td>2800 Gb/s</td>
</tr>
</tbody>
</table>

Advantages available as building blocks for custom designs/mezzanine boards: successful track record in radio astronomy applications very low power consumption compared to high end CPUs and GPUs more flexibility than ASICs high performance I/O capabilities good development tools and environment for board design and programming established technology, at least two manufacturers Disadvantages programming FPGAs is more complex, error prone and time consuming than CPUs and GPUs appropriate boards hosting FPGAs might not be available as COTS not as power efficient as ASICs

### 3.5 Graphics Processing Unit

A COTS GPU is a technology option for CSP due to many readily realizable benefits such as rapid software development cycle, easy upgrade path, easy phased deployment, low NRE cost, easy reconfigurability, and excellent fault tolerance as elaborated in SKA Memo 139. In the present context, GPU includes system on chip (SOC) for the reason that COTS GPU has to work in ISA/Von Neumann environment. For justifying GPGPU as a COTS option, this white paper focuses on Nvidia GPGPU for further discussions.

The channelizer may suit a direct hardware implementation more than GPU because of its serial signal conditioning nature.

There are 2 implementation options for the correlator and integrator:

- GPU as a mezzanine card of PowerMX
- GPU as a truly COTS standalone hardware sub-system
The first option would defeat some benefits provided by and meant for COTS as the GPU/SOC technology belongs to a third party and designing a mezzanine card for PowerMX does not qualify as COTS solution. Whilst we can retain this option for further discussion, this section focuses on the second option from this point onwards.

Nvidia and AMD are the two global suppliers of General Purpose Graphics Processing Unit (GPGPU) chips in the market and both qualify as COTS products. In the professional, enterprise, and research market sectors, Nvidia is estimated to have over 80% of the market share and has developed an infrastructure called CUDA (Compute Unified Device Architecture) with a reasonably large developer community running in parallel to OpenCL. Owing to the existing developer infrastructure, more technical information about Nvidia GPGPU is available than AMD. Also, #1 (Titan) and #6 (Tian-he1A) Supercomputers of the Top500 in the world as of November 2012 were based on Nvidia GPU whereas AMD GPU appeared at #52. For justifying GPGPU as a COTS option, this white paper focuses on Nvidia GPGPU for further discussions.

Nvidia has disclosed in March 2013 the roadmap for the next 2 generations of Tesla technology with minimal technical data. These 2 generations are called Maxwell for 2014 and Volta for 2016. The last generation is Fermi and its release appealed to the radio astronomy computing community back in 2009. The current generation is called Kepler and was released in 2012 with a raw peak processing capability of 2.7 times that of Fermi.

Kepler is based on 32nm technology node. Current Intel CPU models namely Ivy Bridge (released in 2012) and Haswell (to be released publicly in June 2013) are based on 22nm. TSMC- the largest contract foundry has disclosed success on 16nm for ARMv8 micro-architecture. Intel has also disclosed 14nm for Broadwell in 2014 or 2015. We can safely expect 14nm for Nvidia GPGPU by the end of 2016.

In terms of micro-architecture, there are 3 new features most published on Kepler. They are dynamic parallelism, Hyper Q, and RDMA (Remote Direct Memory Access). Dynamic Parallelism allows the GPU to take control of allocating work to other GPU cores away from the CPU (Fermi did not allow this). Hyper Q provides up to 32 CPU to GPU simultaneous connections (Fermi has 1). RDMA allows memory transfer over PCIe to be executed outside the jurisdiction of the CPU. A General Electric whitepaper published in 2012 suggested the use of this feature in their products. For SKA CSP purposes, RDMA has a role to play and this will be explained more below.

Nvidia has offered unified memory to Maxwell which is scheduled to be released in 2014/15. On the surface, unified memory refers to the access of Main Memory (RAM) and GPU Global Memory (GDDR) by both CPU and GPU. This is already happening on Intel Ivy Bridge and AMD APU as the GPU is integrated on the same die of the CPU and shares the same internal cache and external main memory. Nvidia has mentioned that a memory size as large as 16GB is accessible by the GPU, which is relevant to SKA CSP Corner Turning.

Nvidia has offered stacked memory to Volta which is scheduled to be available in 2016/17. At present, GPU memory is external to the GPU die but on the same card-GPU cores access GPU global memory with up to 6 controllers with a data bandwidth of 384bits (64 × 6) with a theoretical peak data transfer rate of 250GB/s. Stacked Memory refers to having the global memory mounted on top of the GPU cores presumably with more controllers. Nvidia has mentioned a figure of 1TB/s.
Tegra4 from Nvidia is a popular SOC for smart phones. It consists of 5 CPU cores, 72 GeForce cores, memory, and various hardware registers and controllers to complete a system. Nvidia has published the next 2 generations called Logan and Parker. In essence, these future generations will incorporate GPGPU cores. Logan will have Kepler cores, and Parker will have Maxwell cores. Nvidia is aiming to have Tegra level CPU cores in Maxwell and Volta. That is, we can expect these next generations to be SOC as well.

PCI Express bandwidth plays the limiting role in the SKA1 CSP design. PCIe is managed by PCI-SIG. Current version is 3.0 with close to 16GB/s per direction for 16 lanes (aggregate) of data bandwidth. PCI-SIG has announced v4.0 for doubling the bandwidth to 32GB/s featuring 16GT/s. Final specifications are expected to be released in 2014/2015.

Table 8: GPGPU Performance Parameters (processing performance in double precision)

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Power</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fermi GF100</td>
<td>2010</td>
<td>€1500</td>
<td>1.33 TFLOP/s</td>
<td>8 GB/s</td>
</tr>
<tr>
<td>Kepler GK110</td>
<td>2012</td>
<td>€1800</td>
<td>3.95 TFLOP/s</td>
<td>15.8 GB/s</td>
</tr>
<tr>
<td>Maxwell</td>
<td>2014</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Volta</td>
<td>2016</td>
<td>€1800</td>
<td>11.7 TFLOP/s</td>
<td>31.5 GB/s</td>
</tr>
</tbody>
</table>

3.6 Homogeneous Central Processing Units

The type of processing required is more of the kind as classically available in normal CPUs. Due to the recent advent of massively parallel multi-core processors, a design decision to use CPU-based processing needs to be re-evaluated. Under the bottom line both the ratios of Watt/GFLOP and cost/GFLOP are essential for these types of deployment.

This section evaluates the suitability of the recent technologies presented by Intel (Xeon Phi co-processor board) and Adapteva (Epiphany processor range).

3.6.1 Intel “Many Integrated Cores” (Xeon Phi)

The Intel Xeon Phi is a co-processor board (connected via PCIe 2.0) with 60 cores. It is built to be used as a “cluster on a board” companion to general purpose x86-based systems, and it also runs its own operating system internally (for managing, scheduling, etc.), which the host system does not need to be exposed to.

Xeon Phi seems like an interesting technology, and one can put a few (e.g. 4) of these boards into a single high end x86-based system. It should be quite comparable (both in price as well as capability) to GPU-based systems, with the difference of being much easier to code for. A potential problem lies similarly to GPU boards in the bottle neck of the PCIe bus for data transfer, as Intel currently does not offer a current PCIe 3.0 bus connection. A bigger problem that is foreseeable is future scalability. The cores are arranged on a ring bus topography, so it will be hard to rapidly scale up compute power.
in a way like GPUs do. However, if the problem is throughput bound, this might not be an issue. Generally, Intel is positioning the Xeon Phi to compete against GPU boards for HPC processing, which is reflected as well in the pricing as well as power consumption.

Table 9: Xeon Phi Specification Summary (processing performance in double precision).

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>5110P</td>
<td>2012</td>
<td>€3200</td>
<td>1.01 TFLOP/s</td>
<td>225 W</td>
<td>8 GB</td>
<td>8 GB/s</td>
</tr>
<tr>
<td>7120P</td>
<td>2013</td>
<td>€3200</td>
<td>1.2 TFLOP/s</td>
<td>&lt; 300 W</td>
<td>8 GB</td>
<td>8 GB/s</td>
</tr>
</tbody>
</table>

Specifications (for Xeon Phi 5110P)

- 60 cores of modified P54C design (original Pentium)
- Clock speed 1.05 GHz
- 8 GB of GDDR5 memory at a higher 320 GB/s memory bandwidth
- 30 MB of cache
- 8 GB/s bus bandwidth (7 GHz PCIe 2.0)

3.6.2 Adapteva Epiphany

The Epiphany-IV CPU is a massively multi-core CPU with 64 32-bit RISC cores. It is a “cluster on a chip” co-processor to other systems. The current Epiphany-IV evolution step is produced in a 28nm process succeeding the 16-core Epiphany-III in 65nm process, while both are using the same architectural style.

Epiphany-IV is not an appliance that can be directly used with generic computers, but a co-processor that can be used to build systems. However, with the “Parallela computer” Adapteva does offer an off-the-shelf system which is based on a dual-core ARM with an Epiphany CPU daughter board. But the real power for the correlator design could be derived through “packaging” these chips together with each other and/or FPGAs or other chips on custom-built mezzanine boards. Particularly the intended pricing and low power consumption of the Epiphany-range of CPUs is interesting. Adapteva is already planning the availability of chips with 1024 and 4096 cores (in the next few years).

Specifications (for 64-core Epiphany-IV E64G401)

- 64 32-bit RISC cores
- 800 MHz clock speed
- 32 kB memory per core (2 MB total distributed, shared, on-chip memory)
Figure 7: A possible way to package Epiphany CPUs with other compute hardware using its four eLink buses. The example on the bottom quite nicely outlines the potential for crafting a hybrid implementation suitable for the purposes of a software correlator.

Table 10: Epiphany Specification Summary (processing performance in single precision).

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>E16G301</td>
<td>2012</td>
<td>€ 80</td>
<td>32 GFLOP/s</td>
<td>&lt; 2 W</td>
<td>512 KB</td>
<td>4 × 2 GB/s</td>
</tr>
<tr>
<td>E64G401</td>
<td>2013</td>
<td>€ 80</td>
<td>100 GFLOP/s</td>
<td>&lt; 2 W</td>
<td>2 MB</td>
<td>4 × 2 GB/s</td>
</tr>
<tr>
<td>1024 Core</td>
<td>2014</td>
<td>€ 80</td>
<td>1.6 TFLOP/s</td>
<td>≈ 15 W</td>
<td>32 MB</td>
<td>4 × 2 GB/s</td>
</tr>
<tr>
<td>4096 Core</td>
<td>2016</td>
<td>€ 80</td>
<td>6.4 TFLOP/s</td>
<td>≈ 50 W</td>
<td>128 MB</td>
<td>4 × 2 GB/s</td>
</tr>
</tbody>
</table>

- 1.6 TB/s local memory bandwidth (8 GB/s per each of core’s four internal links)
- 102 GB/s network-on-chip bisection bandwidth, cores connected via internal “network” (in four directions on square lattice)
- Connectivity: Source synchronous LVDS off chip links for host or direct chip-to-chip interfacing; Chip-to-chip links for integrating up to 64 chips on a single board; 324-ball 15 × 15 mm flip-chip BGA

3.7 Heterogeneous Central Processing Units

Heterogeneous CPUs aim to achieve performance and efficiency by effective use of core specialization versus replication. The STI (Sony, Toshiba and IBM) Cell BE (Broadband Engine) was possibly the first to demonstrate the viability of using heterogeneous multicore CPUs for correlating radio astronomy signals. A crucial lesson learned from applying to signal processing applications was the importance of accelerator cores I/O and its implications for achieving good Flops/Watt.

The correlation of a pair of baseline samples is a product of two complex values followed by a summation. Consider using 8 bit + 8 bit representation for each complex number then the correlation computation will require 4 CMACs operations and 4 bytes of storage. The arithmetic intensity in this case is $1 \left(\frac{\text{CMACs}}{\text{bytes}}\right)$, which is the number of operations performed per byte that needs to be loaded from memory. Essentially a low arithmetic intensity
imposes a great strain on I/O with respect to the accelerator performing the operations. In practice an \( m \times n \) tile is loaded from memory into the accelerator and the relationship between arithmetic intensity and integration time can be readily observed from:

\[
\text{Arithmetic Intensity} = \frac{4mnI}{4(m + n)I + 8mn}
\]

where \( m \times n \) is a tile of baselines and \( I \) is the number of integrations. Thus increasing \( m \times n \) will increase I/O transfer loads however increasing \( I \) will contribute to lessen the output rate from the accelerator. By lessening the burden on the I/O bus better accelerator utilization can be achieved leading to a higher Flops/Watt ratio and therefore greater energy efficiency gains are made towards greater scales.

### 3.7.1 STI Cell BE

The only physical implementation of the Cell BE CPU was the 2008 release of the PowerXCell 8i, which encapsulates a low power 64-bit duo-core PowerPC processor and eight 128 bit RISC cores. The PowerXCell 8i technology reached the end of its development life cycle in 2009, however IBM has announced that it may release future Cell BE families. As the PowerXCell 8i prioritizes greater bandwidth over latency by favouring greater computation throughput this makes it ideal for low arithmetic intensity real-time signal processing.

#### Table 11: PowerXCell 8i Specification Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Cost</th>
<th>Processing Performance</th>
<th>Power Usage</th>
<th>Memory</th>
<th>I/O Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>PowerXCell 8i</td>
<td>2008</td>
<td>€35</td>
<td>230.4 GFLOP/s</td>
<td>55 W</td>
<td>2048 KB</td>
<td>75 GB/s</td>
</tr>
</tbody>
</table>
4 Correlator Option: Redback and FPGA 7 to 8 bit Correlator

4.1 7bit Correlations

The performance stated in Section 3.1 assumes 18-bit multiplication but the correlator requires fewer bits, nominally 8. This allows a number of approaches to improving the performance. For Altera FPGAs each pair of 18 bit multipliers can implement three 8-bit multiplies. This improves the performance of the Redback processing module to 7.88 TCMAC/s (6.3 TCMAC/s at 80% usage). The performance is doubled if 7-bit correlations are sufficient. A 7-bit correlator is 99.97% efficient when run with a signal with an rms of 17.5 quantisation levels. It has good linearity, within 0.1%, for rms signal levels of 2-28. For efficiency greater than 99.9% the level is limited to the range 9 to 22, a 7.8 dB range. Clipping only occurs on 3 sigma or greater excursions making them rare events.

The method used is to change to a sign magnitude representation and implement two 6-bit unsigned multiplies in the single 18-bit multiplier. The bit placement is illustrated in Figure 8.

![Figure 8: Bit ordering for two 6bit unsigned multiplies, the ten bit product $a \cdot d$ in bits 11-22 and $b \cdot d$ in bits 0-11](image)

It is seen that the two 12 bit unsigned results are in the bottom 23 bits of the multiplier output and that the top bit of the bottom result $b \cdot d$ is added to the top result. This mixing can be unscrambled by noting that the bottom bit of $a \cdot d$ is easily calculated. If bit 11 is equal to this then the top bit of $b \cdot d$ was a zero completing the calculation of $b \cdot d$ and $a \cdot d$ is correct. If bit 11 is not equal to the expected result for bit 0 of $a \cdot d$ then the top bit of $b \cdot d$ is 1 and 1 needs to be subtracted from the result for $a \cdot d$ in bits 11 to 22 to get the correct value of $a \cdot d$. This correction can be applied in the following adder.
This method allows two 18-bit multipliers to form all four unsigned multiplies. After this the products are summed in an adder/subtracter with an adder used if the signs are the same and a subtracter if they are different. This gives a twos complement result whose sign may be incorrect. This is corrected with a second adder/subtracter when the result is accumulated. Compared to a two complement implementation the only change is that adders have become adder/subtracters and there is a small amount of logic based on the sign bits of the input to determine the operation of the adder/subtracters. For a Xilinx implementation a $25 \times 18$ bit multiplier is available and the same technique allows full 8 bit complex multiplies to be implemented in two $25 \times 18$ bit multipliers. This technique doubles the processing performance for correlation to 10.5 TCMAC/s (7 to 8 bit) for the Redback processing module or 8.4 TCMAC/s at 80% usage. Using a similar technique to the one above a four bit correlation can be implemented using single 18-bit multiplier. This has been implemented on the SKAMP correlator. The output of the multiply has all the parts of the final result but further processing is required to get the final result. This results in a multiplier that is logic limited not multiplier limited. SKAMP managed to implement 128 4-bit complex multipliers in an FPGA that had 192 18-bit multipliers. This suggest that the technique can increase the performance of a correlator by a factor of 2.7. This is a 35% improvement over the previous scheme. But it does not improve the antenna based processing. Adding this processing leaves an overall improvement of about 25%.

Problems with a 4-bit correlator are correlator efficiency and correlator gain variation with signal level. The correlator efficiency is 98.8%, this is the equivalent of losing a dish from the array. Factoring this into the cost adds about €0.5M for the dish and feed systems and adds about 15kW to the power consumption. This is sufficient to cancel out the cost improvement possible with a 4bit FPGA correlator for SKA1-Survey. Hence in the following a 7bit correlator is assumed with the following performance.

Table 12: Redback Performance Parameters Using 7bit Correlator

<table>
<thead>
<tr>
<th>Compute capacity (18bit)</th>
<th>5.25 TCMAC/s (4.2 TCMAC/s at 80% usage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute capacity (7-8bit)</td>
<td>10.5 TCMAC/s (8.4 TCMAC/s at 80% usage)</td>
</tr>
<tr>
<td>DRAM (DDR4) I/O</td>
<td>3.173 Tbs/s (2.54 Tbs/s at 80%) unidirectional, halve for R/W</td>
</tr>
<tr>
<td>DRAM memory</td>
<td>24 DDR4 (196 byte with 8GB DIMMS)</td>
</tr>
<tr>
<td>Multi-mode fibre bidirectional I/O on Redback board</td>
<td>1400 Gb/s at 10 Gb/s per fibre</td>
</tr>
<tr>
<td></td>
<td>2000 Gb/s at 14 Gb/s per fibre</td>
</tr>
<tr>
<td>Single mode fibre bidirectional I/O on Redback board</td>
<td>960 Gb/s with QSFP+ at 40 Gb/s</td>
</tr>
<tr>
<td></td>
<td>1344 Gb/s with QSFP+ at 56 Gb/s</td>
</tr>
<tr>
<td>Data download/upload</td>
<td>Up to 3 links 1 or 10 Gb/s each</td>
</tr>
</tbody>
</table>
4.2 Redback Implementation

The total processing load is 39.5 TCMAC/s for the channelizer and 328 TCMAC/s at 7 to 8 bits for the correlator. The correlator can implement two 7 to 8 bit correlations in the resources for an 18 bit CMAC this gives and equivalent compute load of 203.5 TCMAC/s, 18 bit. Each Redback-4 module can process 4.2 TCMAC/s at 80% usage. Dividing this into the compute load gives a system which requires 48 modules to meet the compute load requirements.

The DRAM memory has two forms of I/O one is for the buffer which requires a rate twice that of the input data or $126.2 \times 10^{12}$b/s and the correlator. Assuming a correlator integration of 1024 samples then the accumulation for a 2kHz channels takes 0.5 seconds. The number of correlations is

$$250000 \text{ channels} \cdot \text{Stokes parameters} \cdot \frac{96 \times 95}{2} \text{ baselines} \times 36 \text{ beams} = 164 \times 10^9.$$

Each accumulation requires a 64 bit read and write giving a data rate to the DRAM of $42 \times 10^{12}$b/s. The total DRAM I/O required is $168 \times 10^{12}$b/s. Each module has DRAM I/O of $2.54 \times 10^{12}$b/s. To meet the DRAM I/O requirements the correlator needs 67 modules. The DRAM must buffer the data for the 1024 samples at 2kHz. This is at least a double buffer so the total storage required is for 1 second of data. This is $63.2 \times 10^{12}$ bits or 7.2Tbytes. Each module, with 8GB DRAMs has 0.196Tbytes of storage. To meet this requirement the system needs 37 modules. To meet the requirements for compute capacity, DRAM I/O and DRAM capacity at least 67 modules are required. The data rate from a single antenna is $63.2 \times 10^{12}/96 = 658 \times 10^9$b/s (613Gb/s). If 40 Gb/s links are used at 90% capacity then the number of links per antenna is 17. Thus the system could be implemented as 17 units each with 4 Redback-4 modules, which gives a total of 68 modules. Each set of 4 modules has 96 QSFP+ slots which is sufficient for a single fibre input from each of the 96 antennas. This input carries data for 30MHz of data for all 36 beams. The total bandwidth that can be processed is 510MHz. The VLBI tied array beam can be implemented directly on the input data. Each FPGA has four 40Gb/s link attached. The VLBI array beam requires data for one beam or about 4 Gb/s of data per FPGA. The sample rate is 30 MS/s (complex) for each of the four inputs and a single complex multiply is required per sample giving a compute load of 120 MCMAC/s. This is negligible compared to the other processing. It is suggest that the beamforming be implemented as a ring beamformer on the 1MHz input data. This could be linked to the three other modules so the final output is the sum of data from all 96 antennas for 30 MHz of bandwidth. Outputs from all 17 sets of modules are required to cover the full 510 MHz. After the channeliser the data must be cross connected between the 4 processor modules. The first cross connect is on the board. After this cross connect each FPGA has 5 MHz of data for 24 antennas. This data is from the channelizer and is critically sampled the total data rate is $55.3 \times 10^{12}$b/s. Divided into the 408 processing FPGAs gives is $126\text{Gb/s per FPGA}$. The FPGA keep one quarter of this, 1.25 MHz, and distributes the remaining 94.5Gb, 3.75MHz, to the other 3 modules in a set. This achieved by using 12 fibre transmitters at 10Gb/s each, which are cross connected by a passive optical cross connect so the each of the other modules in the set receives 4 fibres carrying a 1.25MHz of bandwidth from the three other modules on a 12 fibre ribbon.
This connects directly to the FPGA processing that 1.25 MHz of data. This completes the cross connect and the FPGA can now process the correlations. The maximum rate from the correlator is $164 \times 10^9$ correlations in 0.3 second. Each correlation is a 64 bit word which gives a data rate of $35 \times 10^{12}$b/s. This would give a dump rate of $86 \times 10^9$ b/s per FPGAs. This requires nine 10 Gb/s links per FPGA.

Table 13: Redback processing module resource usage

<table>
<thead>
<tr>
<th>Resource</th>
<th>Available</th>
<th>Actual used</th>
<th>Fraction usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>40Gb/s link</td>
<td>24</td>
<td>24</td>
<td>100%</td>
</tr>
<tr>
<td>Multimode fibre links</td>
<td>144</td>
<td>126</td>
<td>87.5%</td>
</tr>
<tr>
<td>DRAM I/O per FPGA</td>
<td>3173Gb/s</td>
<td>2470Gb/s</td>
<td>78%</td>
</tr>
<tr>
<td>Processing</td>
<td>5.25TCMAC/s</td>
<td>3.0TCMAC/s (18bit)</td>
<td>57%</td>
</tr>
</tbody>
</table>

4.3 Risks

The design is based on an existing design with all data links already proven to operate at 10Gb/s. In terms of technology only the DDR4 DRAM is unproven. Without physical testing it is not known if this DRAM can operate at 2GHz. A fall back position is to reduce the clock rate or increase the number of modules.

There is a risk that the next generation 14nm FPGAs will not be ready on time. The fall back position is to use 20 nm FPGAs which will have about half the performance. This will increase the number modules required by about 50%. This design would use about 75% of compute resources and less of all the other resources.

There is a risk that the FPGA dissipation will exceed the capacity of the module fans to remove the heat. Based on the current design this is a low probability. The fallback position is to use water cooling.

The performance of all other parts of the system is already proven.

4.4 Development Schedule and Manpower Requirements

First installation use is required at the end of 2016. Time from layout to a production batch was 20 months for systems this design is based on. This was achieved with the effort of 1FTE who also was building a parallel system and coding the command and control firmware. So layout needs to start before May 2015. Layout can start as soon as details of the FPGA are available which is expected by the beginning of 2014. This leaves a window of more than a year for start of board layout. The projected schedule could be

1. First two quarters of 2015 develop document design
2. Third quarter of 2015 layout of boards, processing, and command and control.
3. Fourth quarter board testing and revision. Engineering samples of FPGAs are needed at this time.
4. First three quarters of 2016 design testing, revision and verification. Development of all procurement for production

5. Last quarter 2016 production.

Development of the firmware will be in parallel with the hardware. Prototype firmware exist for ASKAP and APERTIF. This will shorten the design cycle. With this background each firmware component needs about 3FTE months to implement. Testbeds and documentation will add about 6FTE months. There are six or seven modules so the total effort over 20 months is 60FTE months or 3 programmers. This will take place in parallel to the hardware development. But it is noted that many of the modules are common across the three correlators. This is expected to halve the effort attributable to the SKA1-Survey correlator to 30 FTE months. Ten prototypes will add about €0.15M.

4.5 Performance Analysis

The above design for meets the requirements of the baseline design. Table 13 shows the margin by which the resources meet the requirements.

4.6 Cost

Table 14: Estimated cost Redback with FPGA for SKA1-Survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€100k</td>
<td>€323k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>9</td>
<td>39</td>
</tr>
<tr>
<td>Number processors</td>
<td>54</td>
<td>234</td>
</tr>
<tr>
<td>Production cost</td>
<td>€108k</td>
<td>€370k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>2.5kW</td>
<td>10.8kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>4.5kW</td>
<td>19.3kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€12.3k</td>
<td>€53.1k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€16k</td>
<td>€68k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€2135k</td>
<td>€6225k</td>
</tr>
</tbody>
</table>

4.7 Scalability to SKA2

SKA2 is required approximately two process generations after SKA1. This will increase FPGA performance by a factor of 4. There should also be a 25% increase in clock speed giving a total improvement of 5

The antenna based processing increases by 10 and the correlation by 100. Thus the antenna based processing goes from 20% to 2% of the load. The total increase in processing
load is 82. The I/O intensity also decreases by a factor 10 and the design becomes compute limited. The current compute load of 57% will go to 80%. Thus the number of modules using the current design increases by a factor of 58. If FPGA were to increase in computational capacity by 5 then the system size increase is reduced to a factor of 12. The size increases from 68 1U modules to 816 modules at a cost of about €11M for the capital cost and total operating cost over 10 years of about €12M.

The number of modules has scaled in proportional to the number of antennas. So a change to 100Gb/s is not needed but it can be used to improve the design. The FPGA system is estimated to cost about €23M over a 10 year life span and occupy about 25 cabinets making it a possibility for SKA2.

The cost of the correlator is such that an ASIC solution should be explored. The ASIC is one generation behind but is expected to have 4 times the computation capacity. This could possibly reduce the cost of the system by a factor of 4. But DRAM I/O has increased by 10 for the buffer and 100 for the correlator. These buffer were in the ratio of 3:1 is now changed to 1:3. So the memory requirements increase by about 33. A factor of 4 improvements in DRAM I/O might be possible. This still means the system has 8 times as many DRAM. The design already devotes a significant area to DRAM so it must be expect the design is at least 8 times larger in terms of DRAM and FPGA. In implementing a system to meet the I/O requirements we have built everything needed for an FPGA based correlator. It would seem that unless there is a solution to the DRAM I/O bottleneck an ASIC solution will not reduce cost.
5 Correlator Option: PowerMX with FPGAs

This section investigates the feasibility of using the PowerMX platform discussed in Section 3.2 with mid-size 2016 FPGA processors discussed in Section 3.4.

In this and the following sections it is presumed that:

- Each of the twelve data paths in a lane are 10Gbps, so there is $1.92 \times 10^{12}$ bps along each of the two edges of a PowerMX board.

- The FPGA are assumed to have processing capability $1.15 \text{ TCMAC/s}$, I/O $2.8 \text{Tb/s}$, memory 84MB, power 7.5W per TMAC, cost €2430 or €3150 mounted on mezzanine card, FPGA block utilization is 80% so effectively 0.92TCMAC/s.

The Fine Channelizer and Fine Delay are presumed to require input bandwidth $63.2 \times 10^{12}$ b/s, output bandwidth $55.3 \times 10^{12}$ b/s, compute $39.5 \text{ TCMAC/s}$, and negligible memory requirements (however, the Corner Turn memory requirements far exceed the FPGA internal capacity, so requires external memory).

The I/O requirements could be met with 33 boards (unless faster data paths were available). The processing requirements could be met with 43 FPGA, each handling $1.47 \times 10^{12}$ b/s. The economic cost could be reduced by using a larger number of lower-end FPGA for the channelization. To make the different options comparable only fine channelization is considered here, so 33 boards are used with 66 lower-end (half cost) FPGA. However, by instead using 33 triple PowerMX blade configuration the coarse channelization, corner turn, and fine channelization requirements can be met (increasing the channelizer costing by approximately a factor of 2.5).

The Correlator and Integrator are presumed to require input bandwidth $55.3 \times 10^{12}$ b/s, output bandwidth $30.6 \times 10^{12}$ b/s, compute $328 \text{ TCMAC/s}$, and memory $22 \times 10^{12}$ bits.

The I/O requirements could be met with 45 boards (unless faster data paths were available). The processing requirements could be met with 357 FPGA, each handling $0.155 \times 10^{12}$ b/s. However, using the same 7bit correlation approach as described in Section 4.1 halves the requirements to 179 FPGA with still comfortable I/O with each FPGA. This could be achieved by utilizing 23 dual PowerMX blades and a total of 184 FPGA.

As with the Redback option, ASIC processors should be investigated for SKA2. It should be noted that if the PowerMX data paths are increased from 10Gbps to 40Gbps this will help alleviate the board I/O bottleneck.
Table 15: Estimated cost PowerMX with FPGA for SKA1-Survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€100k</td>
<td>€323k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>33</td>
<td>46</td>
</tr>
<tr>
<td>Number processors</td>
<td>66</td>
<td>184</td>
</tr>
<tr>
<td>Production cost</td>
<td>€246k</td>
<td>€569k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>1.8kW</td>
<td>8.6kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>3.2kW</td>
<td>15.5kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€8.7k</td>
<td>€42.5k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€11k</td>
<td>€54k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€2230k</td>
<td>€6298k</td>
</tr>
</tbody>
</table>
6 Correlator Option: GPU Boards

This section investigates the feasibility of using GPU Boards discussed in Section 3.2.

Any solution for SKA1-Survey that depends on PCIe will have a severe I/O bottleneck. PCIe v4.0 allows up to 31.5GB/s input and 31.5GB/s output. For the Fine Channelizer input of $63.2 \times 10^{12}$ b/s this places a lower limit of 234 devices required over PCIe. Similarly, the correlator input of $55.3 \times 10^{12}$ b/s requires at least 204 devices. Interestingly, over 112 Volta GPU boards (with 340W apiece) would be needed to meeting the processing requirements of the Correlator and Integrator. They do however have the advantage that the Integrator could almost work entirely within the high-bandwidth GPU memory, can perform baseline-specific integrations, and have spare processing capability to potentially preprocess data for the SDP.

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€70k</td>
<td>€180k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>234</td>
<td>204</td>
</tr>
<tr>
<td>Number processors</td>
<td>234</td>
<td>204</td>
</tr>
<tr>
<td>Production cost</td>
<td>€472k</td>
<td>€416k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>55kW</td>
<td>48kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>98.4kW</td>
<td>85.8kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€271k</td>
<td>€236k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€344k</td>
<td>€300k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€5546k</td>
<td>€8306k</td>
</tr>
</tbody>
</table>
This section investigates the feasibility of using the PowerMX platform discussed in Section 3.2 with an example of homogeneous CPU processors discussed in Section 3.6.

The Epiphany 1024 core processor due for release in 2014 is a massively multicore processor that is estimated for this paper to require about 15Watt. Its most significant drawback is its very low 8GB/s I/O. Epiphany processors are usually made available on mezzanine, so it is assumed here that the eventual Epiphany 1024 core processor can be mounted on a single 400pin Meg-Array, so that a single PowerMX board can accommodate 16 processors. Although very I/O bound, the I/O requirements of the Fine Channelizer could be met with 1728 Epiphany, and of the Correlator and Integrator by 960 Epiphany. This number of processors each processing at 85% efficiency can provide the 328TCMAC/s required, without attempting to condense any floating point operations.

Table 17: Estimated cost PowerMX with Epiphany 1024 core for SKA1-Survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€35k</td>
<td>€135k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>108</td>
<td>60</td>
</tr>
<tr>
<td>Number processors</td>
<td>1728</td>
<td>960</td>
</tr>
<tr>
<td>Production cost</td>
<td>€688k</td>
<td>€379k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>27.0kW</td>
<td>15.0kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>48.3kW</td>
<td>26.8kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€132.8k</td>
<td>€73.8k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€169k</td>
<td>€94k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€4084k</td>
<td>€6293k</td>
</tr>
</tbody>
</table>

The Epiphany 1024 core appears to be feasible for the Correlator and Integrator, but not for channelization, although in 2014 they still fall well short of the required memory so require external DRAM. However, the flexibility of the CPU can be used to perform baseline-specific integrations to potentially halve the data rate output to the SDP.
8 Correlator Option: PowerMX with Heterogeneous CPUs

This section investigates the feasibility of using the PowerMX platform discussed in Section 3.2 with one of the heterogeneous CPU processors discussed in Section 3.7.

The PowerXCell processor is interesting in that although already a deprecated technology (and not openly available on a mezzanine) it has very good off-chip I/O of 75GB/s, superior to current PCIe specifications. Unfortunately its processing capability of only 230.4GFLOP/s in 2008 means that it cannot possibly handle the computational requirements of an SKA correlator.

Table 18: Estimated cost PowerMX with PowerXCell for SKA1-Survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€35k</td>
<td>€135k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>43</td>
<td>356</td>
</tr>
<tr>
<td>Number processors</td>
<td>688</td>
<td>5696</td>
</tr>
<tr>
<td>Production cost</td>
<td>€259k</td>
<td>€1832k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>38.6kW</td>
<td>319.7kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>69.1kW</td>
<td>571.6kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€189.9k</td>
<td>€1571.8k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€242k</td>
<td>€2000k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€4336k</td>
<td>€25585k</td>
</tr>
</tbody>
</table>

Interestingly, if the PowerXCell evolution had continued its current performance would have made it a very competitive solution. With significant developments in heterogeneous processors soon to emerge they might provide a viable solution by 2016.
9 Correlator Option: ASIC

It would take some significant effort to accurately determine the viability and cost of a potential ASIC correlator solution, so this section provides a very superficial analysis using PowerMX with the ASIC design from Section 3.3 (although that ASIC was not designed specifically to be mounted on a PowerMX board).

Although prototyping costs for ASIC are known, the full production cost for over 1500 ASIC chips is not known at this early time (estimated at €1000 each). The potential viability of a fully pipelined correlator warrants a far more detailed analysis, particularly for SKA2, but also potentially also for SKA1.

Table 19: Estimated cost PowerMX with ASIC for SKA1-Survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Fine Channelizer</th>
<th>Correlator and Integrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design cost</td>
<td>€385k</td>
<td>€415k</td>
</tr>
<tr>
<td>Personnel cost</td>
<td>€780k</td>
<td>€3900k</td>
</tr>
<tr>
<td>Number boards</td>
<td>33</td>
<td>84</td>
</tr>
<tr>
<td>Number processors</td>
<td>264</td>
<td>1344</td>
</tr>
<tr>
<td>Production cost</td>
<td>€464k</td>
<td>€1817k</td>
</tr>
<tr>
<td>Boards and processors power</td>
<td>0.5kW</td>
<td>1.5kW</td>
</tr>
<tr>
<td>Power to site</td>
<td>0.8kW</td>
<td>2.7kW</td>
</tr>
<tr>
<td>Power costs per year</td>
<td>€2.2k</td>
<td>€7.3k</td>
</tr>
<tr>
<td>Power/heat infrastructure</td>
<td>€2.9k</td>
<td>€9.3k</td>
</tr>
<tr>
<td>Total ten year cost</td>
<td>€2656k</td>
<td>€7219k</td>
</tr>
</tbody>
</table>
This paper concludes that processor I/O is a critical feature for any potential SKA1-Survey CSP PIP, indeed probably more so than for SKA1-Low and SKA1-Mid. It has attempted to demonstrate that there are several viable solutions for SKA1-Survey CSP. In particular, two boards, Redback 5 and PowerMX, both appear to have fairly similar costings, within the very approximate estimates undertaken here.

FPGA-based solutions, and ASIC solutions depending on their NRE costs and timeline, mounted on either board should provide viable solutions to both channelization and to correlation. However, there is also the possibility that some CPU solutions may exist, particularly low-power multicore CPU for correlation, but potentially also for channelization by processors expected to be released in the near future. Potential GPU solutions that fail to achieve better I/O do not appear at this stage to be as viable as the other solutions investigated, possibly unless their importance to SDP preprocessing becomes evident or power requirements become lower than on the vague GPU road maps.

The PIP group proposes to progress with a determination of a preferred board for use in SKA1-Survey, with the other as a fallback position, to progress using FPGA processors but also further this preliminary investigation into ASIC and multicore CPU processors. It is intended that two or three potential processors will be carried forward at this stage, with a preferred solution identified by SDR/PDR.
Preamble:

The scientific goal of the Pulsar Search Sub-element (PSS) is to enable the search of the entire sky visible from the site of the SKA telescopes for pulsars\(^1\). The elements that make up the three telescopes of the SKA are widely distributed, and thus have a small filling factor. This means that obtaining the maximum sensitivity requires the sum of as many as possible of these elements coherently, this results in a narrow beam on the sky, and thus results in a very slow survey speed. To compensate for that a large number of beams are formed to dramatically speed up the survey. In the case of the SKA-mid under consideration here, the baseline design document indicates that a maximum number of about 2222 beams are required. This leads to the major challenge of searching for pulsars with the SKA as we will discuss in more detail below.

Subsystem Description and Signal Processing Definition/Scope:

The PSS is mooted to fit into the CSP Element immediately after the tied-array beamformer\(^2\) as indicated in Figure 1, and before the data are output into the SDP. It should be noted here that the exact boundary between CSP and SDP elements in the PSS is not well defined and some aspects of the processing may occur in SDP computing resources. However we will consider here a structure as outline in Figure 1 from the baseline design provided by the SPO.

In the remainder of this document we consider that the data that arrives at the PSS is time-ordered beam-formed data with all frequencies present, this requires that as part of the beamformer the data are reordered. We also consider from here that each beam can be considered as a separate data stream and that any processing element can accept one or more streams, and in our initial model there are two beams per processing node. The basic processing steps that the data must undergo are shown in Figure 1 which is taken from the baseline design. While this is a

\(^1\) We note that although not currently a key science goal, the non-imaging processing should also enable the possibility to search the data for fast transients.

\(^2\) We use the tied-array beamformer here to refer to the coherent or in-coherent sum of widely distributed elements (e.g. dishes or stations) to distinguish from other beamformers in the system.
reasonable representation of the signal processing chain, we have made our slightly more detailed version that we think clarifies the situation and that can be seen in Figure 2.

![Signal chain for pulsar searching as taken from the baseline design.](image)

Let us now consider each of the steps described in Figure 2 in turn:

**Sum polarisations and magnitude**: The signals that come from the beamformer will be channelized but they will still be complex values. For the pulsar searching we only need to consider Stokes I, and so we can simply sum the polarisations and square-law detect them.

**Radio Frequency Interference Mitigation**: It is assumed in the signal processing chain that some RFI mitigation will occur before, or during, the beamforming process. However there will clearly still be RFI present in the data that may only be seen with the sensitivity that comes with the combination of all the elements. There are two places in the signal chain where further RFI mitigation could take place, this is before and after the summation of the polarisations, the advantage of the former is simply that one would still be working with the complex data with higher time resolution and separate polarisations. The exact algorithm to be implemented would be a matter for research, but the processing load would be small compared to the subsequent steps.

**Incoherent dedispersion**: As radio signals propagate through the interstellar medium they interact with the free electrons which cause a frequency dependent delay to the group velocity which disperses the signal. For a pulsed signal, like that from a pulsar, this means that the lower frequencies arrive later than for high frequencies and if we simply summed the data over the wide bandwidths used then the pulse would be significantly smeared, either reducing sensitivity, or extreme cases making the pulse invisible. To correct for this affect we can use a process of
dedispersion, where the wide bands are divided into narrow frequency channels. The signal in each channel is corrected for the delay caused by dispersion, and they can then be summed to reveal the true pulse. However, in the case of a survey the amount of dispersion, called the DM, is unknown for any given source, and so one needs to make a search over a range of possible DM values. The exact number of DMs will be discussed in more detail below. The algorithm for how to do this is well understood, however there is still plenty of scope for optimisation for specific processing solutions.

**Acceleration processing:** One of the key science goals of the SKA is to find pulsars in relativistic orbits with other massive bodies, explicitly neutron stars or black holes. The relativistic aspect of this is required so that they are excellent test beds for theories of gravity. This also leads to a problem though in trying to detect them. The basic pulsar search model performs a Fourier transform of the dedispersed time series generated above and then forms a power spectrum. This is then searched for significant frequency peaks corresponding to the spin period of the pulsar. However this assumes that the frequency of the pulsar doesn't vary during the observation. If the pulsar is moving with a high velocity during the observation, then this will cause the frequency to drift and in some cases this will cause the peak to no longer stand out above the noise and not be detected. There are a number of algorithms that can partially correct for this affect and this is what is meant by acceleration processing. We will not address specifics here as this is part of the investigations in stage 1, but the corrections can be made in either the time or frequency domain, so this could be placed here or within the harmonic folding part.
**FFT & Harmonic Sum & Threshold detection:** As already alluded to, the search for pulsars is usually carried out in the frequency domain, so a Fourier transform is performed on each of the dedispersed time series to produce a power spectrum. These power spectra are then whitened to remove any red noise contribution and then searched for significant peaks. As pulsar pulses have a non-sinusoidal appearance and are typically significantly narrower than the pulse period, the Fourier transform generates a number of harmonics. To improve the search sensitivity, the power spectrum is stretched by a factor of two, up to 32 times, and then added to the original spectrum, resulting in the power in harmonically related frequencies to add up, improving the detection significance. At each stage of the harmonic sum significant peaks are searched for.

**Candidate Detection:** After the process of identifying peaks in the power spectra across all DMs and harmonic folds are completed there will be many detections of the same pulsar at many different dispersion measures, and/or slightly different periods. It is therefore necessary to run some sort of “sifting algorithm” on the data to determine the best period and DM combination for a given candidate. In the case of the SKA, because there will be a large number of beams available it will be possible to also compare candidates across beams, and this will help distinguish RFI,
likely in many beams, from real sources, seen in one, or a few beams. In the SKA where we have acceleration searching, it will also be necessary to sift through to find the best acceleration.

Once this list of candidates is identified it is necessary to run a further set of diagnostics on the data to help determine whether or not the source is really a pulsar. This requires that at this stage there is a version of the data available which still has some frequency resolution available, this is because the data will now be folded and dedispersed at a range of values close to the discovery period. This is to determine whether or not the signal-to-noise ratio of the candidate varies in expected ways as a function of dispersion measure and period, and in the case of binaries, period derivative. As is likely, if we are using some form of machine learning, it is at this stage that a set of scores, relating the observed properties of the candidate to those expected for a pulsar, are calculated.

**Candidate Data:** As there will be many millions of candidate pulsars found in each observing session, it isn’t possible to keep the “raw” data for all of the candidates and so it is likely that candidate selection using some form of machine learning will be implemented that will identify pulsars without direct human intervention. Candidates which pass the machine learning phase will have some subset of the raw data saved, this will be used for human inspection, but also as a first data point in establishing a timing solution for the pulsar. It should be noted that the machine learning part, and perhaps also the Candidate detection stages (or even earlier stages) may run on the Science Data Processor hardware.

**Processing Requirements:**

The processing requirements for almost all of the steps outlined in the previous section are outlined in the baseline design. For ease of reference we recreate Table 24 from the baseline design here and will address each element in turn. The table also introduces some consideration of the processing requirements as well which we will also address later.

<table>
<thead>
<tr>
<th>GPU</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection to 2016</td>
<td>~10 T Flops</td>
<td>1 T Flops</td>
</tr>
<tr>
<td>Peak processing (Single Precision)</td>
<td>~10 Gby/s</td>
<td>Conservative estimate</td>
</tr>
<tr>
<td>Assumed processing efficiency</td>
<td>10%</td>
<td>250W</td>
</tr>
<tr>
<td>GPU processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCIe v 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU Dissipation (estimated)</td>
<td>250W</td>
<td></td>
</tr>
<tr>
<td>GPUs hosted per server</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Hosting server</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No processors per hosting Server</td>
<td>2</td>
<td>multi-core x86 (~ 8 cores)</td>
</tr>
<tr>
<td>Rack size</td>
<td>2U</td>
<td></td>
</tr>
<tr>
<td>No supported GPU cards</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Estimated dissipation per server + 2GPUs</td>
<td>750W</td>
<td></td>
</tr>
<tr>
<td>Feature</td>
<td>Specification</td>
<td>Notes</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Rack power limitation</td>
<td>~20 kW</td>
<td>Assumed</td>
</tr>
<tr>
<td>Number of servers per rack</td>
<td>~20</td>
<td></td>
</tr>
<tr>
<td>Communication fabric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switch aggregate bandwidth</td>
<td>~75 T b/s</td>
<td>Non blocking</td>
</tr>
<tr>
<td>No. ports</td>
<td>648</td>
<td>56 Gbps ports</td>
</tr>
<tr>
<td>Switch size</td>
<td>29U</td>
<td></td>
</tr>
<tr>
<td>Switch Dissipation</td>
<td>10kW</td>
<td></td>
</tr>
<tr>
<td>Cabinet footprint</td>
<td>1000 x 650mm</td>
<td>Estimate TBC</td>
</tr>
<tr>
<td>Processing per rack</td>
<td>~40 T Multiplies</td>
<td>~20 kW per rack**</td>
</tr>
<tr>
<td>Beamformer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beamformer load 835 beams*</td>
<td>254 T Ops/s</td>
<td>Greater than 5° of GP</td>
</tr>
<tr>
<td>Beamformer O/P rate 835 beams**</td>
<td>4 T bits/s</td>
<td></td>
</tr>
<tr>
<td>Beamformer load 2222 beams*</td>
<td>680 T Ops/s</td>
<td>Less than 5° of GP</td>
</tr>
<tr>
<td>Beamformer O/P rate 2222 beams**</td>
<td>11T bits/s</td>
<td></td>
</tr>
<tr>
<td>Beamformer load</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beamformer O/P rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dedispersion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Number of Dispersion Measures</td>
<td>~16113</td>
<td>See Pulsar search region table</td>
</tr>
<tr>
<td>Dedispersion load**</td>
<td>4 T Ops/s</td>
<td>Optimized dedispersion</td>
</tr>
<tr>
<td>Dedispersion O/P rate</td>
<td>2 T bits/s</td>
<td>4 bit assumed</td>
</tr>
<tr>
<td>Binary Search</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number trial accelerations</td>
<td>120</td>
<td>For +/- 100 m s^-2</td>
</tr>
<tr>
<td>Binary search load***</td>
<td>~9.6 P Ops/s</td>
<td></td>
</tr>
<tr>
<td>Binary search O/P rate</td>
<td>130 T bits/s</td>
<td>4 bit assumed</td>
</tr>
<tr>
<td>Pre-Whitening &amp; Normalisation****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak load</td>
<td>TBD Tops/s</td>
<td></td>
</tr>
<tr>
<td>Average O/P rate</td>
<td>130 T bits/s</td>
<td></td>
</tr>
<tr>
<td>Harmonic sum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of harmonics</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Processing load****</td>
<td>TBD T Ops/s</td>
<td></td>
</tr>
<tr>
<td>Average O/P rate</td>
<td>16 T bits/s</td>
<td></td>
</tr>
<tr>
<td>Threshold Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S/N ratio</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Detections</td>
<td>TBD</td>
<td></td>
</tr>
<tr>
<td>Processing load</td>
<td>25 T Ops/s</td>
<td>TBC</td>
</tr>
<tr>
<td>Average O/P rate</td>
<td>TBD</td>
<td>Depends on S/N and RFI</td>
</tr>
<tr>
<td>Candidate Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Pulsars</td>
<td>9800</td>
<td>TBC</td>
</tr>
<tr>
<td>MSPs</td>
<td>1240</td>
<td>TBC</td>
</tr>
<tr>
<td>Processing load TBD</td>
<td>TBD</td>
<td></td>
</tr>
<tr>
<td>SKA survey processing load peak</td>
<td>~10 P Mult/s</td>
<td>Based on acceleration processing load</td>
</tr>
<tr>
<td>Number of processing cabinets (peak option)</td>
<td>~250</td>
<td>~5 MW**</td>
</tr>
</tbody>
</table>

Table 24: Recreation of Table 24 from the baseline design.

*Nbms * Ndish * B * 2ny * 2pol
**[Nbm* Ndm * log2(Ndm)] / tsamp optimised (Taylor tree piecewise linear in ~ 64 sub-bands. [max Nbm* Ndm * Nsamp] / tsamp
***Ndm* Nbms* Nacc * Nsamp * log2(Nsamp) / tobs
**** running mean and rms across blocks of frequency cells subtracted from each cell
*****Ndm* Nbms* Nacc * 1 / tobs * log2(nh) * Nsamp
******Ndm* Nbms* Nacc * 1 / tobs * Nsamp
*******Based on road map from Nvidia & Real-time, fast radio transient searches with GPU de-dispersion paper
******Excluding PSU efficiency and Cooling

Data Rates:
The DRM indicates that the desired time resolution for the pulsar search is 50 µs and that is used in the baseline design. The assumed observation duration in the baseline design is given as 600 seconds and this is based on an all-sky survey taking 2 years of observing time, and is arrived at in combination with the number of beams. It is also assumed in the baseline design that the required channel width Δf is governed by the time resolution and thus is 1/t_res that is, 20 kHz. As the product of Δf and t_res is 1, this means that there is no averaging of the data in the beamformer, nor in the magnitude and polarisation summation step. The data rate coming from the beamformer is thus 2222 beams * 2 polarisations X 2 X 300 Msamples X 8 bits = 21.3 Tbits/s, if we assume we can work with 4 bits then this reduces to the 11 Tbits/s mentioned in the table 24.

We note that this assumption of 15000 frequency channels does enable maintaining the full time resolution of 50 µs out to a DM of over 827 as discussed in the baseline design. However, this large number of frequency channels turns out to be a real game changer in terms of making the processing problem tractable or otherwise. The major problem is the following, the dispersive delay across the frequency band out to the maximum desired DM of 3000 corresponds to about 2.5 seconds given the central observing frequency and bandwidth given in the baseline design. This introduces the need to buffer about 3 Gbytes (based on floats) to dedisperse the data for the first time sample, which may be an issue for the fastest accessible RAM available. Along with this the computational complexity scales as N_tramps x N_dms x N_chans. Hence reducing the number of channels reduces processing requirements.

This is why we present an alternative model where we consider that the data are divided into 4000 frequency channels, corresponding to 75 kHz wide channels. This allows us to achieve a DM of 220 for the maximum time resolution\(^3\), we consider that the maximum number of DMs required is still large, at about 16000, although we note that not all DMs have the same time resolution and this is reflected in the down stream processing requirements.

These are important considerations as they have a significant flow on effect on the data size, throughput calculations and processing calculations. We present below some consideration of why 15000 channels is not feasible and we proceed with an assumption of 4000 channels and a corresponding sample time of 53.2 µs (which corresponds to 4 * 1/Δf). This is sufficiently close to the original sampling interval that we consider that the rest of the numbers given in Table 24 hold.

**Technical Considerations/Prototypes:**

\(^3\) Assuming the same formula as used in the baseline design \(DM_{\text{max}} = t_{\text{res}}(\mu s) \times f_c^3(\text{GHz})/ (8300 \times \Delta f(\text{GHz}))\).

\(^4\) This is still a factor of 5 better than other all-sky surveys performed to date.
We are considering here a solution which is based around the concept of a single processing element which processes two beams in parallel, corresponding to a total of 1111 nodes for the maximum number of beams. It may be possible, depending on input bandwidth limitations, that allow for more beams to be sent to a single processing entity. We consider a large switch, or interconnect between the beamformer and the PSS layer to distribute the data. Again, options for having the processing even closer to the beamformer layer will be considered during Stage 1.

We propose that each processing entity will consist of a host and a number of accelerators. The exact match of accelerators to a specific solution will be made during Stage 1 and will be based on the results of the prototyping that we will undertake. However it is likely that both an fpga and GPU accelerator will be included in each host. The main considerations that will determine the relative weighting between the use of the different accelerators will be related to the power consumption, processing power and matching the different processing tasks described above to the different architectures. Consideration will also be given to extensibility.

One working model for the system has the dedispersion taking place on an fpga based accelerator, and the remainder of the searching, including acceleration trials, candidate detection and candidate filtering using machine learning, take place on the GPUs. However we note that extremely rapid dedispersion codes have already been developed by members of our team for GPUs and so it may be that the fpgas are better suited to the FFT’s and acceleration trials. We have funding to undertake significant prototyping during Stage 1 and this will be used to compare at least 3 different configurations of accelerators and hosts and different combinations of the pipeline running on each.

**Achieving the required performance:**

To assess the requirements needed to achieve the required performance we have considered both a combination of existing knowledge, simulations and straight mapping of required peak performance, as given in Table 24, to the available processing resources. A key observation is that the required resources are not just about the number of FLOPs that are available but the complexity of the algorithm and the required memory bandwidth and total data size. We are preparing appendix documents to support these estimates, but we summarise the main points below.

**GPUs**

The NVIDIA roadmap outlines two iterations of NVIDIA GPUs by 2016 (Maxwell and then Volta). Maxwell will offer unified virtual memory; Volta will make use of stacked DRAM technology. The latter is significant because the technology should deliver a bandwidth to main graphics memory in excess of 1TB/s [8].
Assuming the ratio of (compute power)/(memory bandwidth) remains about constant over the period until 2016 and given Volta’s expected main graphics memory bandwidth in excess of 1TB/s, we can make a conservative estimate of Volta’s peak processing power.

The current NVIDIA Kepler GPU2 has a ratio of \((3950 \text{ GFLOPs}) / (250 \text{ GB/s}) = 15.8\). Using this and the above 1TB/s memory bandwidth figure, a conservative estimate of Volta’s peak processing power is \((15-18 \text{ TFLOPs}) / (1 \text{ TB/s})\). This is \(\sim 1.5x\) larger than the estimate in Table 24. A second way to estimate Volta’s peak processing power is to use the figure of 24 DP GFLOPs per watt taken from NVIDIA’s roadmap. Noting that the power consumption of GPUs for scientific computing will be limited by the thermal envelope of server designers (225-250 Watts) we can calculate a second approximation. We might expect the single precision performance to be about double the double precision performance hence, \(2 \times 24 \text{ DP per Watt} \times 225 \text{ Watts} \sim 11 \text{TFLOPs}\). If we consider the upper limit this gives a figure of 12TFLOPs. This is in line with the estimate in Table 24.

Looking at the trend of increasing main memory capacity we see that: Tesla 1.5 GB (C870) and 4 GB (C1060) Fermi 3 GB (M2050) and 6 GB (M2090) Kepler 6 GB (K20X). Hence it would seem reasonable to expect a Maxwell GPU to have in the region of 12GB of main memory and a Volta GPU to have in the region of 24GB of main memory.

We then ran simulations on existing GPU hardware, where we recreated a data set matching the specifications outlined above and performed the requisite number of dispersion trials to reach the maximum DM of 3000. This used the optimised code developed by one of us, Wes Armour. This showed that on the Kepler K20C we achieved 50% of peak flop performance, 71% of peak shared memory bandwidth and 10% of peak main memory bandwidth. This indicates that, assuming these numbers hold, and the Volta achieves performance of 12 TFlops (with a 40-50% efficiency as shown by the tests) then this matches well to achieving the required dedispersion in real time.

The largest processing load comes from the requirement to do acceleration searching. In Table 24 it is estimated that this requires of the order of 4 TFlops per beam. Assuming similar efficiency to that for the dedispersion algorithm we find that this is dimensioned such that it should fit on a Volta generation GPU. However, of course as the data is being processed in a streaming fashion a total of two GPUs would be needed if the GPU were the only accelerator that was to be used.

**FPGAs**

To consider the characteristics of 2016 technology we have considered here reasonable and conservative extrapolations based on the Altera FPGA family. Although we have chosen a specific vendor, we are not suggesting that this is the actual chip that we would purchase.
If we consider a large Stratix 5, the 5SGSD8 chip for example with the parameters as described in Table 1. With FPGAs, the operating frequency and the power consumption are not predefined characteristics. They are design specific and cannot be predicted in advance. However, it is possible to have a rough estimate for a realistic scenario or an upper limit estimate for the worst case scenario. To do this, Altera provides an ‘early power estimation’ excel tool. Specifying a hardware usage of about 80% and a system frequency of 400MHz, with the biggest heat sink and highest air flow at 25°C ambient temperature. The toggle rate (how often transistors switch) has been increased from the default 12.5% to more than 20% to get the highest possible power consumption. This quantity is limited by the maximum temperature that the device can run at safely (100°C for the chosen device). The results show that the chosen FPGA cannot dissipate more than 57 Watts. Therefore, it is safe to assume a 60W maximum power consumption.

Table 1 Note that the Peak GOPs is based on a total of 5889 multiplier-adders and assuming a 400 MHz clock.

<table>
<thead>
<tr>
<th>Internal Memory (MB)</th>
<th>5SGSD8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max External Memory Bandwidth (GB/s)</td>
<td>51.2</td>
</tr>
<tr>
<td>PCIe interface</td>
<td>8xgen3</td>
</tr>
<tr>
<td>Additional I/O bandwidth</td>
<td>676 Gps</td>
</tr>
<tr>
<td>Peak GFLOPs</td>
<td>1568</td>
</tr>
<tr>
<td>Peak GOPS</td>
<td>4711</td>
</tr>
<tr>
<td>Max Power Consumption (Watts)</td>
<td>60</td>
</tr>
<tr>
<td>Idle Power Consumption (Watts)</td>
<td>6</td>
</tr>
<tr>
<td>GFLOPs/watt</td>
<td>26.13</td>
</tr>
<tr>
<td>GOPs/watt</td>
<td>78.52</td>
</tr>
</tbody>
</table>

If we now consider scaling these numbers for the next generation of FPGAs, we worked from an Altera press release which indicated that it will deliver more than 5 TFLOPs (single precision). If we then consider that the same ratio of FLOPS to OPS can be made as seen for the current generation of FPGAs, i.e. a factor 3, then we see that it is capable of about 15 TOPs. That means that there is sufficient processing power to perform the acceleration processing, the most demanding processing requirement by far, that is required for two beams on a single chip, thereby nicely matching with the proposed 1111 nodes for 2222 beams. If we assume a 20% increase in the power consumption from Stratix 5 to Stratix 10 (the next generation Altera FPGAs), we find that the maximum power consumption of the Stratix 10 FPGA is 72 Watts. If we also assume a hardware usage of no more than 80% then the actual power consumption would be at most 59 Watts.
**Development Schedule:**

The major components, excluding disruptive technologies as outlined below, in our design are based on predominantly COTS solutions. Our development schedule is thus driven by the information that we gain from the prototyping stages that we perform in Stage 1. It is planned that at PDR the design, in terms of balance of CPUs, GPUs and FPGAs is established as is the exact number of nodes per beam. In Stage 2 (i.e. up to CDR) exact choices of components will be made for these elements with all the necessary performance data tested and verified where possible.

**Firmware and Software Development:**

The non-imaging processing requires significant programming effort. For all the processing stages algorithms exist. At all stages there is scope for improving algorithms such that they better match the chosen architectures and/or that they can be done more efficiently, both in terms of operations but also power use. The one area that is presently most uncertain is the area of acceleration processing where a number of different algorithms are possible. As indicated in Figure 2 this can happen either in the time or frequency domain, or some combination of both. This will require research and some choices made in Stages 1 & 2. In all likelihood both will be implemented and the most efficient used.

In the case of GPUs there are existing codes to perform the dedispersion, searching (including the fft) and acceleration steps however they will clearly need work to be made sufficiently efficient and also to match them to the appropriate generation of hardware that is ultimately chosen. In the case of the fpgas the basic building blocks exist for dedispersion and searching (including the fft), but they have not been combined into an end-to-end pipeline. Acceleration processing has not yet been implemented on fpgas.

Estimating the total software and firmware development effort required for the end-to-end system is difficult to do as the detailed scope of where the boundaries to the processing still need to be defined. However we estimate that delivering a robust and processing system for SKA Phase 1 will require:

- Pipelining – 7.5 FTE years
- GPU programming – 7 FTE years
- FPGA programming – 8 FTE years
- Candidate detection and Filtering – 5 FTE years
- Interfaces (Databases, Incoming Data) – 9 FTE years

These estimates of the software implementation costs are based on the assumption that the major algorithm development will already have taken place in the pre-construction phase. As discussed above a number of these algorithms are already in place and so our estimates are based on prior experience of porting code developed
for a certain generation of hardware to the following generation and have it work optimally. This is true therefore for the GPU and FPGA programming and the candidate processing, however the pipelining and interfaces will have less opportunity for testing before the full systems is in place. As elsewhere we have built in a 30% overhead on the staff requirements.

Risks & Mitigation:

The main risks that are foreseen are:

- **Technological roadmaps are not met.** We believe that our projections and requirements are sufficiently conservative, that a very large disparity would be needed before our performance goals could not be met.
- **Algorithm Development Targets are not met.** The majority of the algorithms are mature and so there are no show-stoppers. Achieving the required performance figures is challenging, but active development is already ongoing and we have built in sufficient porting time for next generation accelerators.
- **Power Budget not able to be met.** At the moment the design consists of GPUs, CPUs and FPGAs, if the power budget is too high then we will look to transfer data processing to the less power consuming devices. Or we would rescope as outlined in general below.

In all cases we will do as much risk mitigation as possible through the development of the prototypes. We note also that as a modular system that there are fallback positions be they, less accelerations, less dispersion measures or less beams that are able to be processed for SKA1. We note for example, that in the current design that two beams are sent to each node, if performance was an issue then this could be reduced to 1. This would mean that survey speed would be affected though. Other options would be to process less dispersion measures, for example, this would lower the overall data rate, and as it scales directly into the acceleration processing then even a modest decrease would significantly reduce the required processing load. These fall back positions are all science driven, and the most ambitious of these is the acceleration processing which is a significant multiplier, again, here a decrease of up to 50% may be possible without too much loss of science parameter space.

Disruptive technologies:

For this analysis we have used known technologies which have been conservatively projected through to 2016. It is quite possible that during the Stage 1 and Stage 2 phases that there may be new technologies to come available which could alter this landscape. The ones which we think are most likely are ASICS, which may be able to provide reduced power consumption and/or the development of common platform
board, such as PowerMX, which might dramatically lower the price of the fpga board, and potentially other, costs through economies of scale or simply through tailor made designs. As this issue spans the entire CSP consortium, we will make sure to keep abreast of the developments in these areas and, where appropriate incorporate them into the design. Where this happens on a sufficiently short timescale we will be able to include them in the prototyping phase as well.

**Projection to SKA2:**

While there will be an increase in the number of elements present in the array(s) in phase 2, the data from these will still be combined to form tied-array beams in the beamformer. The obvious scaling to an improved system in phase 2 is therefore to increase the number of beams that it is possible to form and from the point of view of NIP, to process. One projection would be an approximate 5 fold increase in the number beams to achieve almost the maximum available field of view. It may also be possible to use an increased bandwidth which would result in a greater number of frequency channels and dispersion measures which needed to be searched. Taking this into consideration results in about an order of magnitude increase in processing power requirements, this suggests that a replacement beamforming system for phase 2 would be of a similar cost to that for phase 1, assuming a four year difference in technology freeze dates between phase 1 and phase 2.

Also, if there was no large beamforming capability available in the Australian site in phase 1, then it would be highly desirable, and possible, to have one located there too in phase 2.
Introduction:

The Square Kilometre Array (SKA) Central Signal Processor (CSP) includes a sub-system for non-imaging processing that has been divided into Pulsar Survey (PSS) and Pulsar Timing (PST) components. This document describes a candidate physical implementation of the Pulsar Timing sub-system of the CSP.

High-precision pulsar timing is required to achieve one of the two scientific objectives of SKA Phase 1. At the anticipated level of sensitivity of SKA1-Mid (System Equivalent Flux Density of 1.7 Jy; Section 2.2 SKA-TEL-SKO-DD-001) timing precision for the majority of current pulsar timing array sources will no longer be dominated by thermal noise. Rather, timing will be limited by systematic error due to

1) instability of the pulsar emission [1];
2) turbulence in the interstellar medium [2,3];
3) radio frequency interference; and
4) instrumental distortion [4].

Plausible solutions to these problems may require the computation of higher-order statistics, such as cyclic spectra [5] and spectral kurtosis [6] estimates; they may also motivate parallel development of pulsar timing instrumentation for SKA1-low [7]. Therefore, these issues have the potential to increase the computational requirements of the pulsar timing signal processor.

These issues will be studied in greater detail during the pre-construction phase, beginning with a critical review of the baseline design that will inform the subsequent requirements analysis and preliminary design activities. This feasibility white paper describes a standard pulsar data reduction pipeline that includes

- conversion from 8-bit to floating-point representation, with corrupt data flagged invalid;
- division of the observed radio frequency band into sub-channels (e.g. using a filter bank);
- phase-coherent dispersion removal [8];
- formation of the Stokes parameters; and
- integration of phase-resolved averages (known as pulse profiles).

The above data-reduction pipeline is currently implemented by open-source digital signal processing software for pulsar astronomy (dspsr.sourceforge.net) [9]. This high-performance code fully exploits both multi-core CPU and general-purpose GPU technologies and outputs data products using the PSRFITS file format [10], a standard that has been adopted by the international community.
Sub-system description:

The Baseline Design (Table 14 SKA-TEL-SKO-DD-001) specifies that the Pulsar Timing sub-system

- is required only for SKA-mid
  - Band 2: 800 to 1600 MHz
  - Band 3: 1.6 to 3.0 GHz
- will receive 10 simultaneous beams;
- must perform coherent dedispersion up to a maximum DM of 3000;
- must have a timing resolution of 100 ns (i.e. bandwidth of 10 MHz).

The last requirement is a reflection of the misconception that 100 ns timing resolution is required to achieve 100 ns timing precision. Therefore, the PST backend will sub-divide each input 10 MHz sub-band into 16 frequency channels using a software filterbank. The resulting time resolution of 1.6 µs is sufficient to retain all of the timing information in an average pulse profile [11].

Signal processing requirements:

The following spreadsheet provides an order of magnitude estimate of the floating point operations per second (flops) required to process a single phased array beam.

<table>
<thead>
<tr>
<th>SKA-mid</th>
<th>band 2</th>
<th>band 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centre Frequency (MHz)</td>
<td>1200</td>
<td>2300</td>
</tr>
<tr>
<td>Bandwidth (MHz)</td>
<td>800</td>
<td>1400</td>
</tr>
<tr>
<td>Sub-bandwidth (MHz)</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Channel Bandwidth (MHz)</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Filterbank Channels</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td><strong>Dedispersion</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum DM</td>
<td>3000</td>
<td>3000</td>
</tr>
<tr>
<td>DM smearing (ms)</td>
<td>30.40</td>
<td>3.80</td>
</tr>
<tr>
<td>minimum FFT length</td>
<td>65536</td>
<td>8192</td>
</tr>
<tr>
<td><strong>Gflops</strong></td>
<td>288</td>
<td>420</td>
</tr>
</tbody>
</table>

Technical description:

Our preliminary benchmarks indicate that the Band 3 problem can be solved in 3.3 times real-time on a single NVIDIA Kepler GPU. To leave ample room for further development (e.g. RFI mitigation), we propose that 4 GPUs [today's technology] are required per phased array beam. Using a conservative extrapolation to 2016, the problem will be readily solved using only 2 GPUs. The I/O capacity required to receive a 1.4 GHz band (2 polarizations, 8 bits/sample, 44.8 Gbit/s) is readily achieved [today] using FDR Infiniband. A simple, scalable model includes a single 1U server with 2 GPUs and an Infiniband Network Interface Card (NIC) for each phased-array beam.
Cost:

Assuming that all results are immediately sent to the Science Data Processor, there is no need to store large volumes of data in the PST sub-system of the CSP. It is assumed that rack space will be provided on site and that the vendor/supplier of the compute nodes will install the machines, connect the network, install the operating system, and configure the cluster. It is also assumed that the beamformers will send Infiniband packets. Therefore, the PST solution consists of an Infiniband switch and ten compute nodes + one spare. A rough cost estimate (including 10 years of operation, power, maintenance, etc.) is around 450,000 Euro. This estimate does not include the cost of replacing failed equipment.

<table>
<thead>
<tr>
<th>Capital Expenditure</th>
<th>Cost (kEuro)</th>
<th>Power (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compute Node</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 RU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 x GPU (Volta)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 x NIC (Infiniband FDR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Infiniband Switch</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 port FDR</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-total</strong></td>
<td><strong>135</strong></td>
<td>6100</td>
</tr>
<tr>
<td><strong>Testing and Commissioning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Engineer (2 months)</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Domain Expert (2 months)</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td><strong>Sub-total</strong></td>
<td><strong>33</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Operating Cost</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Operation</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Mains Efficiency</td>
<td>97%</td>
<td></td>
</tr>
<tr>
<td>Operating Budget</td>
<td>5%</td>
<td>84</td>
</tr>
<tr>
<td>Cost per Watt per year</td>
<td>1.00</td>
<td>63</td>
</tr>
<tr>
<td>Heat removal</td>
<td>50%</td>
<td>31</td>
</tr>
<tr>
<td>Power Delivery</td>
<td>2.00</td>
<td>12</td>
</tr>
<tr>
<td>Maintenance</td>
<td>5%</td>
<td>5</td>
</tr>
<tr>
<td><strong>Sub-total</strong></td>
<td><strong>196</strong></td>
<td></td>
</tr>
<tr>
<td>SUB-TOTAL</td>
<td><strong>364</strong></td>
<td></td>
</tr>
<tr>
<td>Contingency</td>
<td>20%</td>
<td>73</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>437</strong></td>
<td></td>
</tr>
</tbody>
</table>
**Risk:**

IF there is demonstrated scientific demand for more simultaneous beams THEN there may be need for a larger cluster.

IF the critical review of the baseline design exposes the need for more computationally intensive algorithms (such as cyclic spectroscopy) THEN a GPU-based solution may not be sufficient.

IF the beam-former cannot send Infiniband packets, THEN some hardware development may be required to get the data into the Infiniband switch. (A switch may also be unnecessary if each beam-former is connected directly to each compute node; however, such a design would be less convenient in the case of node failure.)

IF NVIDIA does not produce the most competitive GPU at the time of purchase THEN the dpsr GPU code will have to be ported from CUDA to OpenCL.

IF full-time usage of the COTS equipment decreases the mean time between failure THEN having a single spare compute node may not be sufficient (mitigation: purchase a good warranty from vendor)

IF the COTS equipment does not live up to its advertised heat specification THEN it may be necessary to devise custom cooling solutions (mitigation: operate a test unit from vendor)

IF any element in the signal path (e.g. amplifier, A/D converter, polyphase filterbank) introduces a non-linear response THEN significant development effort will go into modeling and correcting the instrumental response

**References:**

[9] van Straten & Bailes. 2011, PASA, 28, 1