PDR.02.01.01 Compute platform: Hardware alternatives and developments

Document number……SKA-TEL-SDP-0000019
Context……...COMP
Revision……1.0
Authors……J. Bancroft, A. Ensor, J. Taylor, S. Wu, Y. Zhu, P. C. Broekema, A. Mika
Release Date……2015-02-09
Document Classification……. Unrestricted
Status……... Final

Name	Designation	Affiliation
Chris Broekema	COMP Team Lead	ASTRON
Signature & Date:	**Signature:** P.C. Broekema (Feb 9, 2015) **Email:** broekema@astron.nl	

Name	Designation	Affiliation
Paul Alexander	Project Lead	University of Cambridge
Signature & Date:	**Signature:** Paul Alexander (Feb 9, 2015) **Email:** pa@mrao.cam.ac.uk	

Version	**Date of Issue**	**Prepared by**	**Comments**
1.0	2015-02-09	SDP.COMP	

ORGANISATION DETAILS

Name	Science Data Processor Consortium

Table of Contents

List of abbreviations

ASIC	Application-Specific Integrated Circuit
CDR	Critical Design Review
COTS	Commercial Off-The-Shelf
DDR	Double Data Rate
DMA	Direct Memory Access
DOE	Department of Energy
DSP	Digital Signal Processor
FFT	Fast Fourier Transform
FISC	Functional Instruction Set Computer
FLOP	Floating Point Operation
FLOPS	Floating Point Operations per Second
FPGA	Field Programmable Gate Array
GDDR	Graphics Double Data Rate
GPU	Graphics Processing Unit
HMC	Hybrid Memory Cube
I/O	Input/Output
MIC	Many Integrated Core, now the Intel Xeon Phi product
NRE	Non-recurring Engineering
PCB	Printed Circuit Board
PCI	Peripheral Component Interconnect
PCIe	Peripheral Component Interconnect Express
RISC	Reduced Instruction Set Computing
SDK	Software Development Kit
SDP	Science Data Processor
SKA	Square Kilometre Array
SoC	System on a Chip
SPU	Scientific Processing Unit
USD	USB Secure Digital
VHDL	VHSIC Hardware Description Language
VHSIC	Very High Speed Integrated Circuit

List of Figures

List of Tables

Summary

Purpose of this document

This document is intended as supporting material to the SKA Science Data Processor compute platform sub-element design report [AD-01]. The focus is on relevant hardware developments and alternative design options for the SDP hardware compute platform. Currently available examples of relevant hardware are used as part of the discussion.

Scope of this document

This document is not intended as an exhaustive discussion on all possible design options available for the SKA1. It is instead meant to be an illustrative document showing the wide range of available technology options and development projects that we intend to explore on the road to CDR and beyond.

References

Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.

Reference Number	Reference
AD-01	SKA-TEL-SDP-0000018 P. C. Broekema: Compute platform element sub-system design
AD-02	SKA-TEL-SDP-0000046 J. Taylor: Costing basis of estimate
AD-03	SKA-TEL-SDP-0000054 I. Cooper: Prototyping plan

Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

Reference Number	Reference

Introduction

The Compute Platform sub-element design document [AD-01] provides a Compute Island baseline model using current-day technology based on Xeon-Tesla hardware:

- Dual Intel Xeon E5-2660v2 CPU (10 cores @2.2GHz each)
- 2x 576-768GB DDR3 main memory @1866MHz
- 2x Nvidia Tesla K40 accelerator
    - PCIe v3 x16; 12 GB GDDR5; 4.29 TFLOPS peak SP; 1.43 TFLOPS peak DP
- 10 GbE Ethernet NIC (PCIe v2 x8)
- FDR Infiniband HCA (PCIe v3 x8)
- HGST Ultrastar SSD1600MR 1.6TB Enterprise MLC SSD

For comparative purposes, a Compute Island server matching this configuration would have the following properties:

- Power: 104-305 Watt for the Intel CPU and 235 Watt for each Tesla K40 for a peak power requirement of approximately 800 Watt.
- Processing: Dual Intel Xeon E5-2660v2 CPU (10 cores at 2.2GHz each) are reported by Intel to deliver 352 GFLOP single precision (SP), 176 GFLOPS double precision (DP), while two Nvidia K40 compute 8.6 TFLOPS SP, 2.86 TFLOPS DP for a total of 8.95 TFLOPS SP or 3.04 TFLOPS DP. All of these are peak performance figures.
- Memory: Up to 1.5TB DDR3 and 1.6 TB flash. Additional spinning disk as required.
- Power performance: 11.1 GFLOP/Watt SP or 3.75 GFLOPS/Watt DP

This document describes alternative Compute Island designs using different hardware technology choices, notably ARM-based, FPGA, OpenPOWER, many core and SoC processors, as well as custom ASIC designs. The primary rationale is that these alternative technologies are anticipated to reduce the power requirements for Compute Islands. Given the rapid evolution of computer technologies and the disruptive changes in supercomputer and datacentre designs caused by tem, it is quite possible that the preferred Compute Island design in 2018 may be quite different from the current, 2015 baseline.

Current Exascale Computing Projects

The first Exascale supercomputer is expected to be delivered around 2020. Several countries including China, the United States, Japan and the European Union have announced plans to develop Exascale systems. Exascale feasibility depends on the rise of energy-efficient technology: a growth in processing power but not in energy consumption. In this section, we provide a survey of current Exascale projects which illustrate the different approaches to achieving Exascale computing. We also discuss the relevance of these developments for the SKA.

Exascale computing projects in China

China is preparing to work on a supercomputer with a peak capacity of 100 PFLOPS by 2015 and will try to produce the first Exascale computer in 2020. Tianhe-2, which was built by China's National University of Defense Technology in collaboration with the Chinese IT firm Inspur Group, is now the world's fastest supercomputer, according to the Linpack benchmark reported in the TOP 500 ranking. The theoretical peak performance of Tianhe-2 is 54.9 PFLOPS while the result of its Linpack test is 33.86 PFLOPS. Tianhe-2 is a two-phase project. The first phase was finished in 2013. During the second phase, scheduled to complete by 2015, Tianhe-2 will be upgraded and enhanced to 100 PFLOPS. This is a key milestone toward the goal of the first exascale computer in 2020.

Tianhe-2 adopts a hybrid architecture using a heterogeneous integration of CPUs and MICs (Intel Xeon Phi-s). The main hardware consists of three components:

- Computing system
- Interconnected communication system
- I/O system

Tianhe-2 has 16,000 compute nodes, each comprising two Intel Xeon processors (E5-2692 v2, 12 cores, 2.2 GHz) and three Xeon Phi co-processor cards (31S1P, 57 cores, 1.1 GHz). The double precision peak performance is 211.2 GFLOPS for one CPU and 1.003 TFLOPS for one Xeon Phi card. So the peak performance is 3432 GFLOPS for each compute node. Each compute node has 88 gigabytes of memory. 64 gigabytes come from the DDR3 main memory used by the CPUs and 24 gigabytes from three Xeon Phi cards with 8 gigabytes of memory each.

The interconnected communication system called TH Express-2 uses high-radix network routing chips and high-speed network interface chips to construct the communication network. Both types of chips are proprietary designs.

The I/O system uses a hybrid hierarchical file system called H2FS, which combines node-local storage and shared storage into a single, dynamic namespace to optimise I/O performance in data-intensive applications. H2FS presents a hybrid storage architecture and management scheme which can deal with a mixture of different storage systems including hard disk drive and Flash hybrids.

The storage subsystem contains 256 I/O nodes and 64 storage servers with a total capacity of 12.4 PB, packaged in 24 storage racks. As mentioned above, it is a hybrid hierarchical storage

architecture that consists of node-local storage and global shared storage (based on hard disk drives) to enable a large capacity with high bandwidth and low latency.
The cooling system is a close-coupled chilled water cooling with a customised liquid water cooling unit. It has a high cooling capacity of 80 kW. The power dissipation at full load is 17.8 MW excluding the energy consumption of the cooling system.
Tianhe-2 was designed as a common HPC platform for applications from many different fields, including computational fluid dynamics, life sciences and aerospace. The 88 GB memory per node is sufficient for most applications. However, it is much smaller than the memory requirements of the SKA1 computing platform which exceed 1000 GB per node.
All the 16,000 compute nodes are interconnected by a high-speed, proprietary communication system. If we assume that the SKA will use technology provided by vendors in 2016 or later, and the peak performance of Xeon CPU and Xeon Phi will have at least doubled by then, the peak performance of a supercomputer with an architecture similar to Tianhe-2 will be at least 100 PFLOPS which meets the computing requirements of the SKA1.
The energy consumption of Tianhe-2 is 17.6 MW without and 24 MW including the cooling system. This is much larger than the 5.5 MW power budget of the SKA1. It seems that the majority of the energy consumption comes from the Xeon Phis. Because the power budget of future Xeon Phi products will most likely stay at the same level as today, the energy consumption may still be a big issue when using Xeon Phi as the accelerator of choice in the SKA compute platform.

Exascale projects in the United States

The U. S. Department of Energy is funding the Exascale Computing Initiative, a government/industry partnership to create commercial exascale computing products by 2023. There are several phases, commencing with the purchase of two 150 PFLOPS systems to be delivered in 2017. One of the systems has the option to be expanded to 300 PFLOPS (http://www.computerworld.com/article/2849250/coming-by-2023-an-exascale-supercomputer-in-the-us.html).

One systems, called Sierra, will be used for military research. The other, called Summit, is planned for 2017 and will be open to the scientific community. Summit will have 3400 compute nodes each containing multiple IBM Power9 CPUs and NVidia Volta GPUs. The memory per node will exceed 512 GB (high-bandwidth memory and DDR4). The interconnects between the CPUs and GPUs will use the NVLINK (5-12x PCIe3), while the Dual Rail EDR-IBs (23 GB/s) will be used as system interconnects. Summit has a power budget of 10 MW(https://www.olcf.ornl.gov/summit/).

Summit meets the memory requirements of a potential SKA1 supercomputer. The node architecture is also similar to the SDP compute node baseline model. However, the energy consumption still remains an issue with Summit's power budget being twice as large as that of the SKA1.

Exascale projects in Japan

Japan is developing plans for a $1.1B investment program aiming to deploy an Exascale capable machine by 2020 (http://www.ci.anl.gov/blog/rick-stevens-race-exascale). The project, called FLAGSHIP 2020, is led by the RIKEN Advanced Institute for Computational Science. However, the initial scope has been modified and the project now aims at building a 200- to 600-PFLOPS system by 2020 instead of an Exascale system(http://www.computerworld.com/article/2849250/coming-by-2023-an-exascale-supercomputer-in-the-us.html). RIKEN selected Fujitsu to develop the basic system design. This new supercomputer will be the successor of the K supercomputer which was the fastest supercomputer in the world in 2011.

The theoretical peak performance of the K is 11.28 PFLOPS. Each computing node contains a single SPARC64 VIIIfx processor (2.0 GHz, 8 cores) and 16 GBs of memory. The interconnect is a proprietary six-dimensional torus called Tofu. An optimised parallel file system is adopted (called the Fujitsu Exabyte File System) which is scalable to several hundred petabytes. The power consumption of the K is 12.66 MW. The annual running costs are estimated at 10 million USD.

Fujitsu have presented their next generation supercomputer system called Post-FX10 at the International Supercomputing conference in 2014. The peak performance of Post-FX10 is expected to be 100 PFLOPS. The Post-FX10 will use SPARC64 Xlfx CPU which includes 32 cores (http://www.hotchips.org/wp-content/uploads/hc_archives/hc26/HC26-11-day1-epub/HC26.11-1-High-Performance-epub/HC26.11.120-SPARC64-Xlfx-Yoshida-Fujitsu-rev1.2.pdf). The double precision peak performance of this CPU will be 1 TFLOPS. The network interfaces are embedded in CPU chips and the multi-dimensional torus network topology inherited from the K computer will be used in Post-FX10 (http://www.pcworld.com/article/2690212/fujitsu-to-design-japanese-exascale-supercomputer.html).

Exascale projects in Europe

In this section, we discuss eight recent European Exascale projects. They are CRESTA (Collaborative Research into Exascale Systemware, Tools & Applications), DEEP (Dynamical Exascale Entry Platform) and its follow-up DEEP-ER (DEEP – Extended Research), Mont-Blanc, EPiGRAM (Exascale ProGramming Models), NUMEXAS (Numerical Methods and Tools for Key Exascale Computing Challenges In Engineering and Applied Sciences), EXA2CT (Exascale Algorithms and Advanced Computational Techniques) and EESI (European Exascale Software Initiative).

The CRESTA project (since 2011) tackles the software challenge in Exascale machines (http://www.cresta-project.eu/). The project has two integrated strands: one focused on enabling a key set of co-design applications for the Exascale, the other focused on building and exploring appropriate systemware for Exascale platforms. The key co-design application areas are:

biomolecular systems, fusion energy, the virtual physiological human, numerical weather prediction and engineering.

The DEEP project (2007-2013) was aimed at developing a novel, Exascale-enabling supercomputing architecture with a matching software stack and a set of optimised applications (http://www.deep-project.eu/deep-project/EN/Home/home_node.html). DEEP focused on both hardware and software. Instead of adding accelerator cards to cluster nodes, DEEP complemented a conventional HPC system with an accelerator cluster, called Booster, to increase its compute performance (see Figure 1). An extrapolation to millions of cores would take the DEEP concept to an Exascale level. The cluster-level heterogeneity of DEEP would allow users to run applications with kernels of high scalability alongside kernels of low scalability concurrently on different sides of the system.

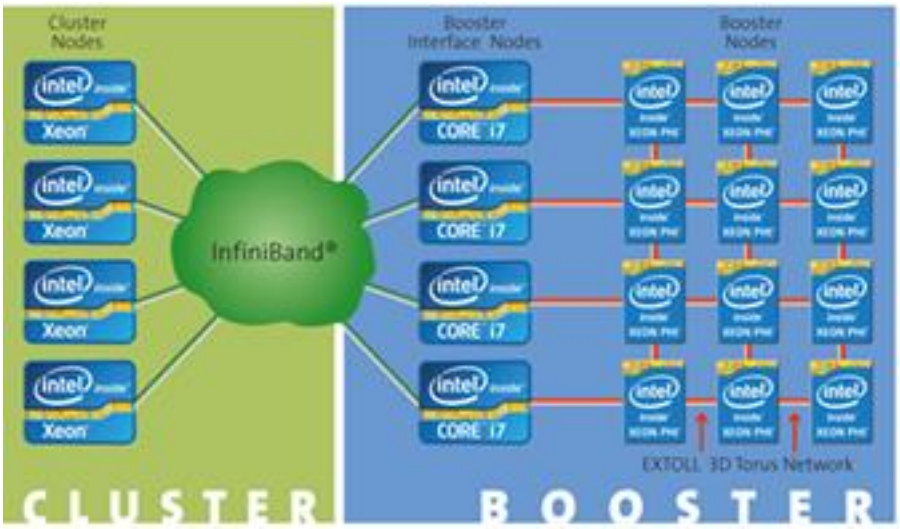

Figure 1: The DEEP hardware architecture.

Having started in October 2013, the DEEP-ER (DEEP – Extended Reach) project will advance the Cluster-Booster architecture developed in DEEP in terms of processor technology, network interconnect and storage. On the software side, the central research topics will be I/O and resiliency. By the end of the project in 2016, a large-scale prototype will be deployed with applications fine-tuned to the underlying software and hardware architectures (www.deep-er.eu).

Since October 2011, the aim of the Mont-Blanc project has been to design a new type of supercomputer architecture based on energy-efficient solutions used in embedded and mobile devices. The Mont-Blanc project is led by the Barcelona Supercomputing Centre and brings together leading European technology providers (ARM, Bull and Gnodal), as well as leading Tier-0 supercomputing organisations in PRACE (Partnership for Advanced Computing in Europe): Juelich, LRZ, GENCI, and CINECA. The Mont-Blanc project has unveiled a prototype which consists of nine blades, each of them containing 15 Samsung Daughter Board nodes. The blades are interconnected using a tree topology with four nodes connected to the outside. Each Samsung Daughter Board node has two ARM Cortex-A15 CPUs, an ARM Mali T-604

GPU, 4 GB LPDDR3 RAM and a 16 GB USD card. The network is 1 Gb Ethernet(https://www.montblanc-project.eu/arm-based-platforms). ARM-based supercomputers can provide a power saving of up to 10% to 20% compared to today's systems. However, it is unclear whether sufficient application software will be available by the time they are delivered.

The objective of the EESI project (http://www.eesi-project.eu/pages/menu/homepage.php), co-funded by the European Commission, is to build a European vision and roadmap to address the challenges of the new generation of massively parallel systems composed of millions of heterogeneous cores which will provide multi-petaflop performances in the next few years and exaflop performances in 2020.

The following three smaller projects focus more on the programming model side of the Exascale challenge:

The aim of the EPiGRAM project is to prepare message passing and partitioned global address space programming models for Exascale systems by fundamentally addressing their main current limitations. The concepts developed will be tested and guided by two applications in the engineering and space weather domains (http://www.epigram-project.eu/).

The aim of the NUMEXAS project is to develop, implement and demonstrate the next generation of numerical simulation methods to be run on Exascale computing systems. The NUMEXAS methods and codes will be the main project outcomes (http://www.numexas.eu/).

The EXA2CT project will produce modular open source proto-applications that demonstrate the algorithms and programming techniques developed in the project. The aim is to help boot-strap the creation of genuine Exascale codes (http://www.exa2ct.eu/content/about-project.html).

A start-up company called Optalysis is claiming that they can create an optical solver supercomputer by 2020, which will be a 17-EFLOPS machine. A 340 GFLOPS proof-of-concept model will be launched in January, 2015(http://www.montblanc-project.eu/press-corner/news/united-states-china-europe-and-japan-race-exaflop-supercomputer).

Applicability to the SKA1

One of the biggest challenges in designing an Exascale supercomputer is to limit its energy consumption. The energy budget of the SKA1 is capped at 5.5 MW while the required computing performance is in the order of 100 PFLOPS. Current technologies cannot meet these requirements. The energy consumption of Tianhe-2, the current fastest supercomputer is 17.6 MW excluding and 24 MW including the cooling system which is much higher than the SKA1 power budget.

Intel has announced that revolutionary technologies, such as the 14 nm process, will be used in the next generation Xeon Phi product. The peak performance of the new Xeon Phi will be better than 3 TFLOPS which constitutes a factor three improvement. To further improve the application performance, the new Xeon Phi will be running as a standalone processor to avoid data

movement between the CPU and accelerator. This improved performance might make the Xeon Phil a potential choice for the SKA.

Summit, a heterogeneous system with a CPU+GPU architecture, will achieve a performance of 150 PFLOPS with an energy consumption of 10 MW by 2017. Although the energy consumption of the Summit is nearly twice the SKA1 energy budget, considering the future development of GPUs it is very likely that a supercomputer with a similar architecture will meet the SKA1 requirements by 2020.

New types of supercomputers based on energy-efficient solutions such as used in the Mont-Blanc system will come closer to meeting the SKA1 power requirements.

The Cluster-Booster architecture developed in the DEEP project provides another option for designing the SKA supercomputer. This architecture allows to run less parallel and highly parallel tasks on the cluster and the booster concurrently. It has the advantage of providing more flexibility in sizing the cluster and booster parts. However, it would result in less modular Compute Islands.

The common software that will be developed in the CRESTA, EPiGRAM, NUMEXAS, and EXA2CT projects could prove to be useful for the SKA.

On the use of Integrated CPU and GPU Architectures

Discrete CPUs and loosely interconnected GPUs

The basis for the current cost estimate for the compute node [AD-02] described in the Introduction, is based on a standard accelerated commodity server, a Graphics Processing Unit (GPU), specifically an Nvidia Tesla, in the PCI-Express bus sub-system of the server. This implementation is widely used in HPC systems as a means to deliver considerably more FLOPS per CPU die than is achievable in standard architectures. This is because more silicon area is devoted to the arithmetic pipelines, spread over thousands of cores, which can operate independently on data. As such, given the prevalence of this type of architecture today, it is a reasonable choice for the cost estimate.
While this computational model has been successful, there are some impediments to its adoption. One is the necessary multi-threaded programming model itself. Another one is the disparity in memory bandwidth between the CPU/GPU interconnect and the PCI-E bus.
While these two factors have impeded the wider adoption of such technologies in the general case, they have been surmounted to achieve high raw compute capability for specific application areas, especially when considering power constraints. The programming model, data-flow and synchronisation are shown in Figure 2.

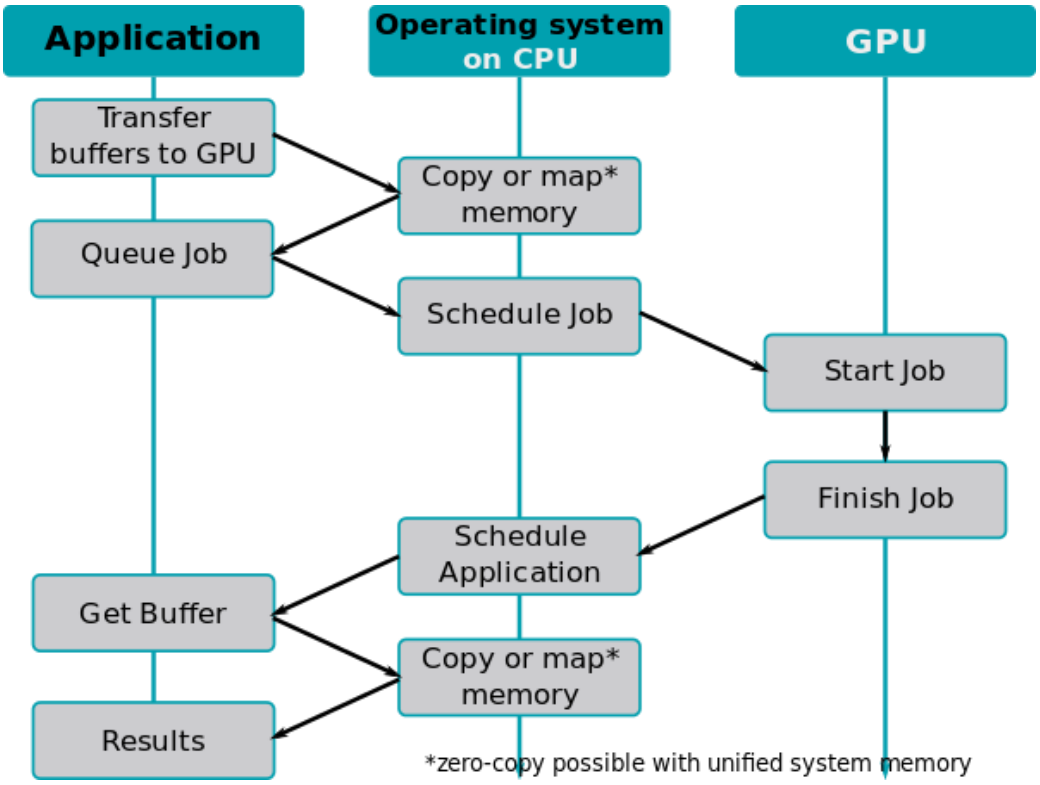

Figure 2: Programming model, data flow and synchronisation of a coupled CPU and GPU architecture.

It is, however, to be expected that new system interconnect bus technologies like NVLink will allow for a tighter integration of CPU and GPU due to significantly higher bandwidth (compared to PCIe GEN3 today) and support of coherence protocols (http://www.nvidia.com/object/nvlink.html).
An architectural approach which has been driven by System-on-Chip (SoC) vendors, most notably in the areas of mobile computing, adopts a different model whereby the CPU and GPU are housed on the same silicon thus reducing the overall power budget. This approach is now being extended to the server domain, incorporating devices from Intel, ARM and AMD and has been described by a number of terms such as co-processors or arithmetic processing unit. While there are a number of implementations being discussed, we mention here the attempts of the Heterogeneous System Architecture foundation (http://www.hsafoundation.com/) .
This proposes a type of computer processor architecture that integrates CPUs and accelerators on the same bus with shared tasking and memory. The platform's stated aim is to reduce communication latency between CPUs and GPUs (and other compute devices), and make these various devices more compatible from a programmer's perspective. This relieves the programmer of the task of planning the moving of data between devices' disjoint memories (as must be done with OpenCL or CUDA) as depicted in Figure 3.

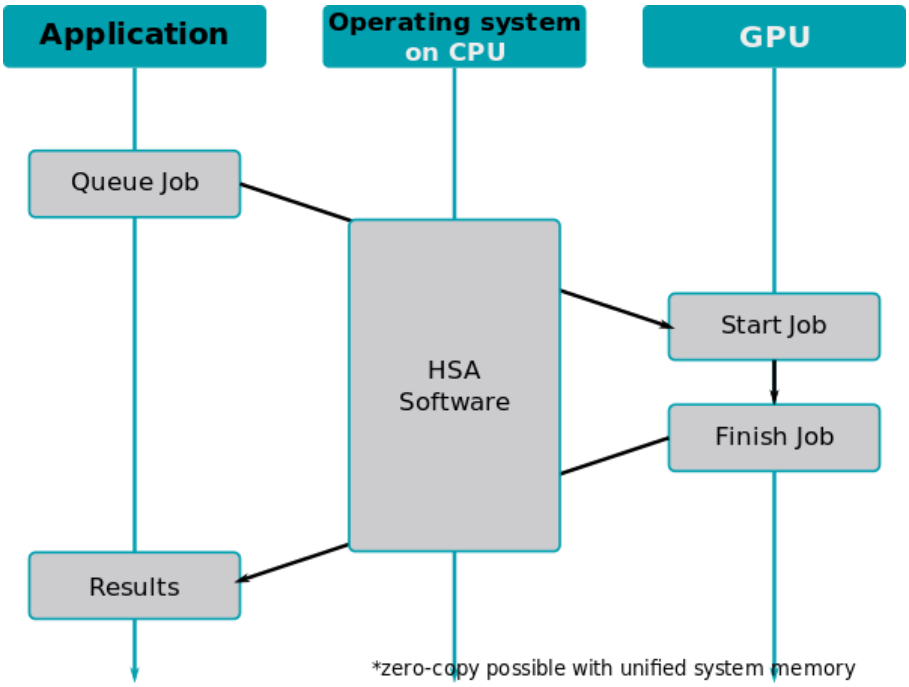

Figure 3: Example of a computer processor architecture integrating CPUs and accelerators on the same bus with shared tasking and memory.

On-die integration of CPUs and GPUs

The current OpenPOWER roadmap shows that new high-speed bus technologies will allow for a tight integration of discrete CPUs and GPUs. This does not only feature a higher bandwidth between the host and the device but also a single coherent memory address space.

Impact on SDP

While the Host + Accelerator model is mature and therefore a reasonable choice for the cost estimate, the evolving technology around integrated CPUs and GPUs is of considerable interest. Notwithstanding the potential performance of such devices both in terms of FLOPS and memory bandwidth, the ease of the programming model and the simplicity of the interface between memory and other I/O devices may significantly improve the buffer performance within a Compute Island and ease the programming of pipelines.

Low-power alternatives

There are numerous processors that can be considered “low-power” alternatives to the conventional HPC market that has been dominated by the x86, Tesla GPU, PowerPC and more recently Xeon Phi architectures. However, one has to bear in mind that it is not the instruction set architecture (ISA) but rather the design priorities and platform characteristics that define the energy consumed. Prototyping our specific application set is obviously essential to getting a handle on realistic numbers for the SKA.

ARM-based processors

ARM-based processors (http://www.arm.com/) follow a Reduced Instruction Set Computing (RISC) design that consumes less power than x86 processors. They have been the main architecture for battery-powered smart phones, tablet and embedded system processors. This architecture has recently started to also make inroads in the server and data centre markets, with particular interest in the newer 64-bit ARM architecture.

Notable ARM processor releases in 2014 intended for the server and data centre markets include the 64-bit AMD Opteron A100 CPU (http://www.amd.com/en-us/products/server) with four or eight ARMv8-A cores, the AppliedMicro 64-bit HeliX (https://www.apm.com/products/embedded/helix-family/) with two, four or eight ARMv8-A cores and the AppliedMicro 64-bit X-Gene processors (https://www.apm.com/products/data-center/x-gene-family/) with up to eight cores. HeliX is developed for the embedded, while X-Gene for the server market.
In November 2014 Cray announced that they will explore ARM-based clusters using 48-core Cavium processors (http://investors.cray.com/phoenix.zhtml?c=98390&p=irol-newsArticle&ID=1990117). In October, 2014 Hewlett-Packard announced the first two ARM server models to reach the market from Project Moonshot (http://www8.hp.com/us/en/products/servers/moonshot/):

- the ProLiant m400 general-purpose memory-optimised machine intended for cloud providers with a 64-bit 8-core AppliedMicro X-Gene ARM,
- the ProLiant m800 intended for real-time signal processing applications with four 32-bit TI 66AK2H SoC, each with four ARM A15 cores and eight C66x DSPs that have native floating point support.

A 4.3U chassis can accommodate 45 m400 or m800 server cartridges. A single-precision, 1024-point complex-to-complex FFT has been benchmarked in a C66x DSP and reported in the Texas Instruments DSPLIB Test Report to execute in 0.86 µs and consume only 8.6 µJoules (by comparison an Nvidia C2070 executes the FFT faster in 0.16 µs but consumes a much larger 36 µJoules) (see e.g., .http://www.sagivtech.com/contentManagment/uploadedFiles/fileGallery/Multi_core_DSPs_vs_GPUs_TI_for_distribution.pdf)

It has to be pointed out though that an ARM microserver is typically less compute capable than a conventional x86 server. E.g. Figure 10 at http://www.extremetech.com/extreme/188396-the-final-isa-showdown-is-arm-x86-or-mips-intrinsically-more-power-efficient/2 provides a good indication of the complexity in the arguments of the ARM vs. x86 comparison.

System on a Chip processors

Unlike Nvidia's Tesla series (with the Tesla, Fermi and Kepler architectures) of high-end GPUs aimed at the HPC market, their Tegra series of SoCs targets the mobile and embedded market with a mix of both ARM and GPU cores (http://www.nvidia.com/object/tegra.html). A Tegra was originally selected in 2011 to be used in the Mount Blanc project (http://www.montblanc-project.eu/) to build an energy-efficient HPC prototype, although a Samsung Exynos was chosen in 2012 for later designs. The Nvidia Tegra K1 SoC, as available on the Jetson TK1

development kit, has a 192-core Kepler GPU and 4-Plus-1 quad core ARM Cortex-A15 CPU. The 20 nm successor to the K1 is the Nvidia X1, released in January 2015. It has a 256-core Maxwell GPU and 4-Plus-4 eight core ARM A57-A53 CPU. The X1 has a 512 GFLOP peak performance for 32-bit floating point operations and 1024 GFLOP for 16-bit floats.
Samsung's Exynos SoC are also ARM-based and primarily intended for the mobile market (http://www.samsung.com/global/business/semiconductor/minisite/Exynos/w/). The Exynos 5 Octa range is particularly interesting as it uses the ARM big.LITTLE 4-Plus-4 architecture with four slower low-power ARMv8-A cores and four more powerful cores, allowing lower priority or less computationally intensive tasks to be scheduled on the LITTLE cores to reduce power consumption.

Kalray MPPA

The Kalray MPPA processor (http://www.kalrayinc.com/) is promoted for its power efficiency in compute-intensive applications. The MPPA consists of multiple clusters on chip interconnected with a 2D torus mesh, where each cluster has cores with L2 shared memory. The Kalray MPPA-256 "Andey" many-core processor (available since Q1 2013 in 28 nm CMOS from the Taiwan Semiconductor Manufacturing Company Limited) with 16 clusters each containing 16 cores and 2 MB memory, resulting in (comparatively very large) 32 MB on-chip memory. This large memory and intra-network make the processor particularly attractive for applications such as gridding and FFTs.

SKA prototyping reported in SKA-TEL-CSP-0000142 has demonstrated the ability of the MPPA-256 Andey to perform a 2^{18}-point 1D FFT including window function (floating point and fixed point) within a (entirely on-chip) single cluster in approximately 8 ms. This can be extended to a power-efficient 2048×2048 2D FFT for the Imaging pipeline entirely on-chip, without the need (and associated power cost) of using off-chip memory. The rationale for performing an FFT entirely on-chip rather than utilising RAM is that by 2018 floating point operations are estimated to consume about 10.6 pJ/op on 11 nm lithography technology, whereas the cost of reading from (or writing to) DRAM will only improve modestly to 1000 pJ unless more energy-efficient memory technology is developed (see, e.g. *Exascale Computing Technology Challenges* (2011) by Shalf, Dosanjh, Morrison).

The MPPA-256 "Boston" due in Q1 2015 offers twice the performance of Andey for both integer and single precision operations. Also on the Kalray public roadmap (http://www.kalray.eu/products/mppa-manycore/) for 2015 is the MPPA-1024 "Coolidge", which will increase the number of clusters by a factor of four compared to "Boston", thereby accommodating at least a 4096×4096 2D FFT on-chip.The Kalray ACC02 board includes four MPPA processors.

Tilera TILE

Another low-power alternative processor is the Tilera TILE as available on the TILEncore-Gx72 with the TILE-Gx8072 processor and 72 cores (http://www.tilera.com/). Unfortunately there is not enough information available on this processor yet for a serious consideration for the SDP compute platform.

Epiphany processor

The Adapteva Epiphany is a much smaller and less-mature alternative to an ARM architecture that also uses a RISC core design (http://www.adapteva.com/). The Epiphany is a many core co-processor that arranges its cores in an interconnected mesh and currently achieves up to an impressive 70 GFLOP/Watt performance. It is available as the 16-core Epiphany III and from Q3 2014 as the 64-core Epiphany IV with Zynq ARM/FPGA host Parallela development boards. Adapteva has publicly announced plans to advance the RISC Epiphany processor to 4096 cores by 2016.

Relevance for the SKA

Although the low power processors achieve very good power efficiency, they are typically designed to accommodate comparatively low I/O rates, so data transfers can easily become a performance bottleneck, particularly with problems that cannot easily exploit data parallelism. This could be alleviated in the Imaging Pipeline using schemes such as the Priapress visibility data lossless compression scheme described in SKA-TEL-CSP-0000112 with prototyping reported in SKA-TEL-CSP-0000113 to reduce the required I/O by around 34%. The processors are often not designed to maintain peak performance either, so heat extraction needs to be carefully considered when clustering many processors together in a low-power island design.

The iotadrive is an ARM-GPU hybrid system under development by Nyriad Limited (http://www.nyriad.com/) that uses dual Nvidia Tegra X1 processors and 8 TB solid state storage. A Compute Island could be designed using nine iotadrives with 18x10 Gb Ethernet NIC, 18x12 Gb FCOE ports and with 25.6 GB direct memory bus interconnects between iotadrives. The Compute Island node would have the following properties:

- Power: each iotadrive consumes up to 55 Watt, requiring a peak 500 Watt for the Compute Island (the baseline Xeon-Tesla Compute Island design requires approximately 800 Watt).
- Processing: excluding the ARM cores, the dual X1-s in each iotadrive contribute together 1 TFLOP single precision peak performance for a total of 9 TFLOP (the baseline also has a total of 9 TFLOP single precision).
- Memory: 288 GB RAM and 72 TB high speed SDD 4 GB/s RAM (the baseline has a larger 1.5TB DDR3 but a lower 5.2 TB flash).
- Power performance: 18 GFLOP/Watt (the baseline has 11.3 GFLOP/Watt).

If single precision calculations were acceptable in the SDP, a 100,000 iotadrive Compute Island design for the SDP using 2015 technologies would achieve a theoretical single precision performance of 100 PFLOP, have 800 PB SDD storage, require 40 m^3 rack space when stacked while consuming a maximum of 5.5 MWatt.

Custom accelerators (FPGAs, ASICs)

Compared to commodity CPUs and GPUs, FPGAs (field-programmable gate arrays) and ASICs (application-specific integrated circuits) provide very impressive I/O capabilities and comparatively good compute power performance. This typically comes at the expense of much greater NRE and development effort, and comparatively less flexibility with iterative software development.

Examples of FPGA acceleration for HPC projects

FPGAs are used in various areas, including various industries, design activities, financial services, scientific research, digital content creation and movie animation. Some specific examples where FPGAs have been used in relevant applications are:

LOFAR

The LOw Frequency ARray (LOFAR) in the Netherlands, centred on Exloo, is an aperture array radio telescope. It consists of 48 stations, each containing a phased array of antennas. Antenna data is first processed locally in a station for data reduction, before it is sent to the central data processor located at the University of Groningen.

Currently, the LOFAR station processing is implemented on 48 Xilinx Virtex-4 SX35 FPGAs, a platform introduced in 2004. An equivalent hardware platform, based on technologies introduced in 2012 and capable of real-time station processing, would require either a multi-socket system with at least seven Intel i7-3820 processors containing a total of 56 cores or three Xilinx Virtex-7 x690T FPGAs.

Maxwell – a 64 FPGA Supercomputer

The High Performance Computing Alliance was founded in early 2005 to take forward the ideas of an FPGA-based supercomputer. The alliance partners are Algotronix, Alpha Data, EPCC at the University of Edinburgh, the Institute for System Level Integration, Nallatech and Xilinx. The project was facilitated and partly funded by the Scottish Enterprise Industries.

The machine itself – Maxwell – was completed in the first part of 2007 and subsequently went on to win the silver medal in the BT Flagship Award for Innovation at the 2007 British Computer Society Project Excellence Awards.

Maxwell is a 32-way IBM BladeCenter containing 64 Xilinx Virtex-4s configured in two flavours. It is a world-leading High-performance Reconfigurable Computing research platform.

The need for power-efficient FPGA computation in the SKA

Table 1 illustrates the power limits for the three SKA1 telescopes. Three figures are given for each: an estimated, a “likely” and a “best case” power limit. The estimated power limit was computed using the SDP parametric model based on values from the cost model. The “likely” power limit is based on an overall power budget that can be considered realistic given the current state of the design. The “best case” power limit is the absolute highest power that the SDP will have available if all current unknowns are replaced by the most favourable assumptions. These power limits are those measured at the building entrance, i.e. including cooling, losses, auxiliaries, etc.

Telescope	Required Power Estimate (MW)	Likely Power Limit (MW)	“Best Case” Power Limit (MW)
SKA1 Mid	4.2	2.5	5
SKA1 Survey	5.9	2	4
SKA1 Low	2.4	0.75	1.5

Table 1: The power caps of the three SKA1 telescopes

According to the information in Table 1, the SKA SDP power envelope is likely to exceed even the best case power limit if the baseline design is being followed.
As argued earlier in this document, the use of FPGAs often results in better power efficiency. In the following table, we made a projection of performance and power consumption of current technologies as well as technologies expected in 2022. It appears that an FPGA-based solution using 2022 technology might be capable of meeting the power budget of the SKA SDP.

	Current status of TianHe-2 (#1 on top500)	Solution with current CPU+MIC/GPU	Solution with CPU+MIC/GPU in 2022	Solution with FPGAs in 2022
Computing	**54.9 PFLOPS**	**149 PFLOPS**	**149 PFLOPS**	**149**

Performance	**(Theoretical) 33.9 PFLOPS (Measured)**			**PFLOPS**
Power Consumption	**17.8 MW**	**48.3 MW**	**12.7 MW**	**3.06 MW**

Table 2: Power consumption of a CPU+MIC/GPU and an FPGA-based SDP architecture based on current technology and projected to 2022.

The power budgets in Table 2 were derived as follows: assuming the same architecture as used in Tianhe-2, the number one computer on the Top 500 list in 2014, 149 PFLOPS will require a power budget of 48.3 MW. According to the cost model [AD-02], the power consumption will be 12.7 MW in 2022.

Provided that we still use the same algorithms as today, the computation complexity will be dominated by FFT and (de)gridding which constitute ~80% of the total computations. Using FPGAs to accelerate FFT and (de)gridding in five years' time, the power consumption could possibly be reduced to ~3 MW. However, power efficiency is highly dependent on the application at hand, so this number has to be taken with a grain of salt. The next section provides a comparison of the expected compute capability and power efficiency of many-core CPUs, GPUs and FPGAs.

Processor Power Comparison

Figure 4 shows the single-precision compute capability of some current and planned many-core CPU, GPU, and FPGA processors (obtained from publicly available specifications and technology roadmaps up to 2016), along with their compute power efficiency. For each technology the trend is of rapidly increasing compute capabilities with each generation and more modest improvements in compute power efficiencies. Although performance is highly application specific, in general the GPU processors have the greatest compute capabilities per processor, followed by FPGAs, but FPGAs have the potential for greater power efficiency. The CPU processors demonstrate a great variety of power efficiencies and lower overall compute capabilities, being designed for general and flexible computing requirements. However, for a problem that can be effectively distributed across many processors, such as gridding, some CPU processors may be quite competitive in terms of overall power consumption.

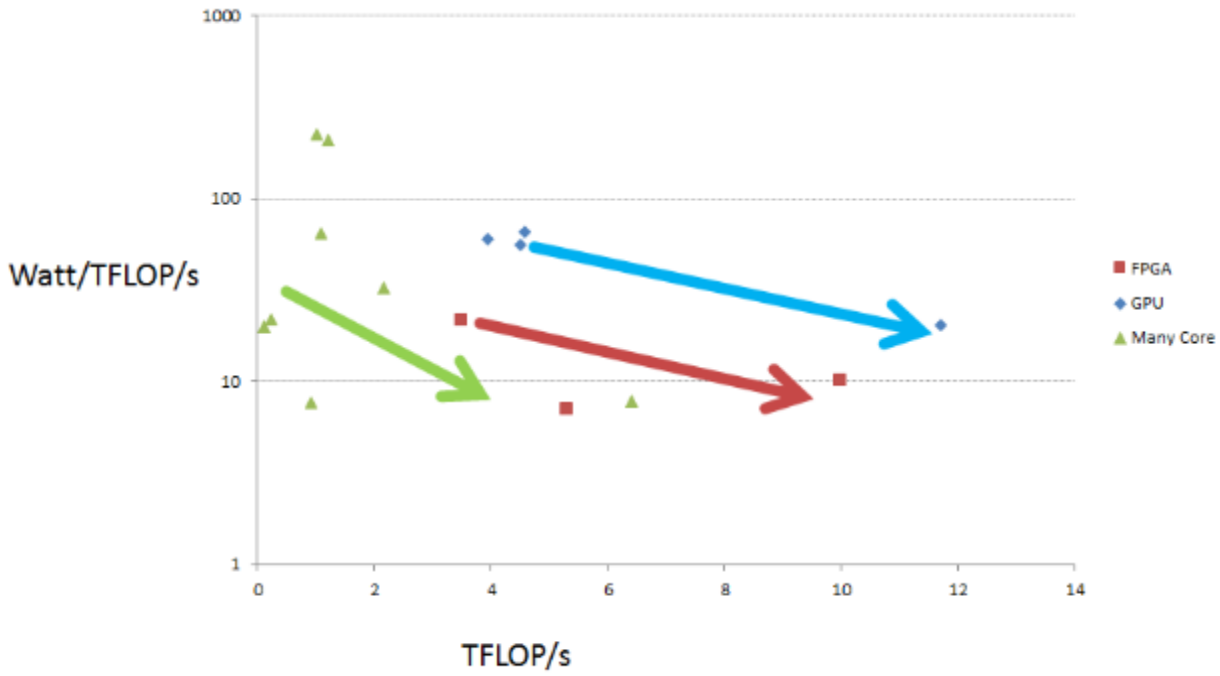

Figure 4: Compute efficiency comparison for current and roadmap CPU (many core)/GPU/FPGA processors. The arrows indicate trends across technology generations.

FPGA vendors and system designs

Xilinx (http://www.xilinx.com/) and Altera (http://www.altera.com/) are currently dominating the FPGA market. Xilinx offers a series of FPGA chips that implement the ARMv7 instruction set. Examples include the the Virtex high-end series, Kintex mid-range series, the less-expensive Artix series, and the Zynq series. Similarly, Altera offers the Stratix high-end series, Arria mid-range series and the less-expensive Cyclone series. Altera also offers an ARMv7 IP core to create a SoC.

Both manufacturers provide IP cores for common signal processing applications. In addition to traditional hardware description languages such as VHDL or Verilog, both manufacturers now also support OpenCL to facilitate application development. Altera provides OpenCL code examples for large FFTs athttp://www.altera.com/support/examples/opencl/opencl.html. FPGAs are often incorporated on custom PCB designs that include DDR3 or DDR4 RAM or more recently Hybrid Memory Cube that can accommodate greater data bandwidths and memory densities. FPGAs are also available in packaged systems - notably the Maxeler MPC-X2000 1U system with eight FPGAs (https://www.maxeler.com/) and the Convey Wolverine R720 2U system with at least four Xilinx Virtex7 FPGAs (http://www.conveycomputer.com/). Both of these are utilised in HPC and data centre environments.

Heterogeneous FPGA-CPU approaches for data centres include Microsoft Catapult (http://research.microsoft.com/pubs/212001/Catapult_ISCA_2014.pdf) using Stratix FPGA and

Xeon Sandy Bridge CPU, and CAPI (http://newsroom.altera.com/press-releases/nr-ibm-capi.htm) which is compatible with Altera FPGAs, the Power8 CPU and will be supported by future generations of IBM POWER processors.

Candidate hardware/software tools for FPGAs

Both Altera OpenCL SDK and Xilinx SDAccel are candidate tools for developing applications for FPGAs. They intend to provide a CPU/GPU-like development experience on FPGAs.

In November, 2014, Xilinx launched their SDAccel Development Environment. According to the release notes, SDAAccel offers the first CPU/GPU-like development experience on FPGAs and the first CPU/GPU-like run-time and update experience. The SDAccel Development Environment targets host systems based on x86 server processors and provides commercial off-the-shelf (COTS) plug-in PCIe cards that add FPGA functionality. In a compression benchmark, SDAccel is claimed to have produced hardware accelerators that are over three times faster and over three times smaller than accelerators generated by competing tools. Area and throughput is comparable or better than hand-optimised register transfer level implementations.

In August 2012, Altera announced OpenCL SDK. By "freezing" many elements of the architecture, Altera has created a "sandbox" in which users can utilise the FPGA without doing any register transfer level design. The results are seemingly impressive for the FPGA-based approach in contrast to the CPU and GPU alternatives.

Altera's OpenCL implementation is currently x86-specific from a system CPU standpoint. The interface between the CPU and FPGA is also currently restricted to PCI Express, and the FPGA portion of the design must also include particular memory controllers; the Altera SDK automatically creates these fixed function blocks in the process of generating the FPGA design's bitstream.

ASICs

FPGAs balance flexibility, performance and power efficiency. However, their power efficiency is usually lower by an order of magnitude than that of ASICs. The advantages of FPGAs can be further dwarfed if SKA applications require high-precision floating-point calculations since double floating-point calculation circuits on FPGAs require a significantly larger area than on ASICs.

To cut NRE costs, co-design methodology is currently widely used in ASIC design. Most of the hardware logic can be reused in different applications, which lowers the NRE costs. Regarding flexibility, ASICs can cover a variety of applications by replacing the co-design module. Though they introduce a longer design period and larger cost, they might still be a competitive candidate for such a large-scale scientific facility as the SKA.

The Scientific Processing Unit

The SPU (Scientific Processing Unit) is a highly-efficient many-core processor ASIC. It was designed for parallel scientific computing by the same China Academy of Science team who invented the Godson-T many-core processor which was built on a many-core microarchitecture (Godson-T: An Efficient Many-Core Processor Exploring Thread-Level Parallelism. IEEE Micro 32(2): 38-47, 2012). The SPU is the next generation of the Godson-T many-core processor and currently is still in an early stage of development. The tapout is planned for the beginning of 2016.

There were two primary objectives in the SPU design. One was to increase the efficiency of science and engineering applications, which is normally no more than 10% of the compute capacity in current mainstream HPC platforms (x86+GPU/MIC). The other objective was to greatly improve the energy efficiency, especially for memory-bound applications.

The first generation SPU is shown in Figure 5. The SPU chip consists of 4 FISC (Functional Instruction Set Computer) general cores, and each FISC core has an accelerator (ACC) which consists of a two-dimensional 16*16 processing elements (PE) array. Each PE operates at 1.5 GHz, offering 3 TFLOPS of double-precision peak performance, while the maximum power consumption is less than 60 W. The SPU will have up to 16 GB of GDDR5 memory, with 384 GB/sec of memory bandwidth. SPU is a co-design architecture for some important kernels such as stencil/FFT making it suitable for these kind of computations.

Each FISC core is a medium-sized general purpose processor, whose performance is similar to that of an ARM Cortex-A15. It has special instructions to control the ACC. It also has a normal cache hierarchy containing both L1 and L2 caches. All the FISC cores are connected to GDDR controllers with a crossbar.

Each ACC has a multi-bank local buffer (BUF), which has a separate address space and can be accessed by both the ACC and the adjacent FISC core. The FISC core can move data from the memory to the local buffer in an asynchronous mode similar to DMA.

Within the ACC, each PE node has one Floating-Point Multiply-Accumulator unit and some local buffers and registers. Instructions on the ACCs are executed in a data flow mode, producing a higher efficiency than the traditional control-flow execution mode.

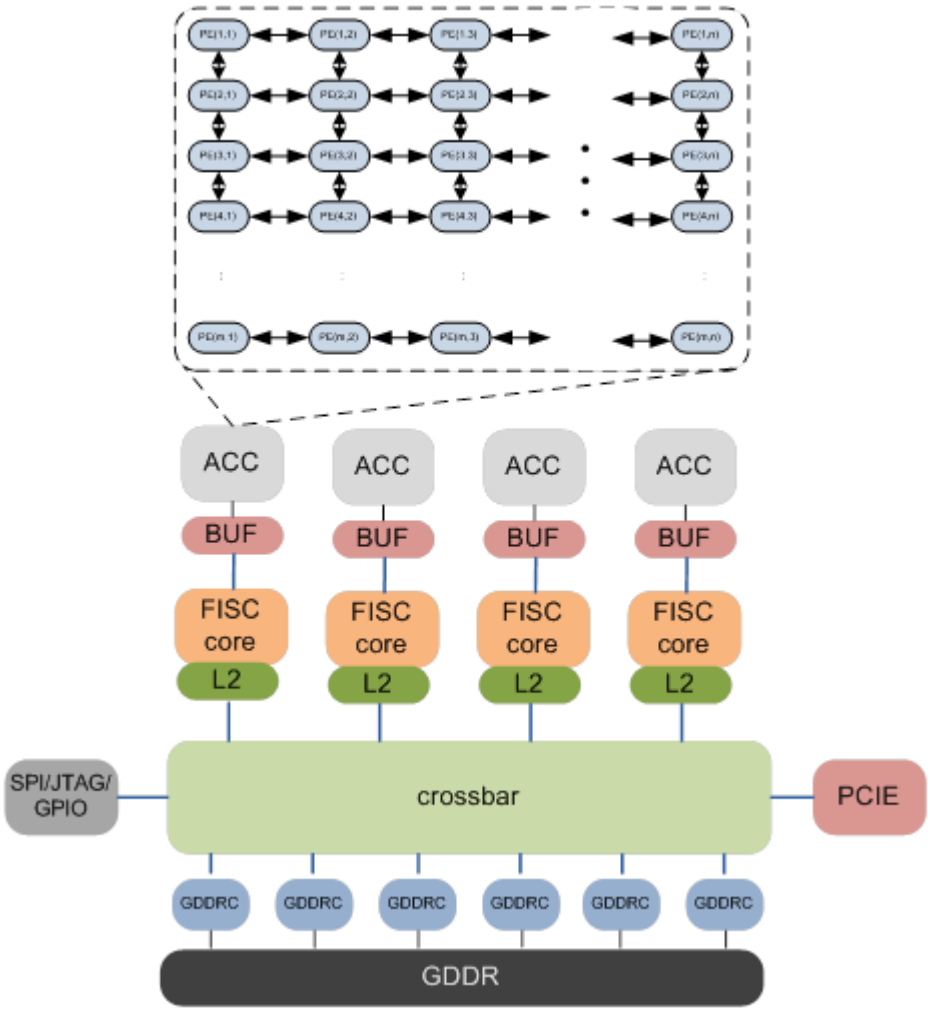

Figure 5: The first generation SPU architecture.

The performance and compute capability of the SPU are expected to be about the same as for next generation GPU/MICs. However, the SPU provides a much better power efficiency and application efficiency. Table 3 shows a comparison between the SPU and next generation GPU/MICs. Please note that efficiency figures will depend on the application at hand. The results shown in Table 3 were derived based on the computation of a typical scientific application, i.e. the seven-point stencil. The efficiency of the stencil kernel is usually no more than 20% on GPU/MIC-based architectures.

	SPU	Next Generation GPU/MIC
Peak Performance	3 TFLOPS	3 TFLOPS
Power	60 W	200-300 W
Performance / Watt	50 GFLOPS/W	10-15 GFLOPS/W

Efficiency	>40%	<20%

Table 3: Comparison between the SPU and next generation GPU/MICs based on the computation of the seven-point stencil.

In summary, the SPU has the following goals:

- High efficiency: The FISC architecture makes the SPU easy to program and the data flow mode ensures high execution efficiency.
- Low Power: The ACC unit can run at variable speeds and it can also be dynamically powered off.
- Elastic: The ACC unit is co-designed with application kernels and can be replaced when different applications are needed.

Printed circuit boards

Processors must be mounted on printed circuit boards equipped with a power supply, I/O bus (such as PCIe or Infiniband), and often also a monitor and control processor. Commodity CPUs and GPUs are typically purchased already mounted on a manufacturer-provided board, but most FPGA and ASIC solutions as well as SoC processors intended for many embedded applications, require a PCB (printed circuit board) design. Although a custom PCB design provides the greatest flexibility for combining processors, memory and bus topology on a board, it involves a significant NRE cost and introduces risk compared to a full commodity solution.

PowerMX

An alternative to a fully customised board option for incorporating FPGA, ASIC and/or embedded processors into an island design is the PowerMX platform (http://www.powermx.org/), a PCB specification being led by the National Research Council of Canada, specifically designed to handle the extreme data rates of the SKA. PowerMX is intended to exploit the work NRE invested in the SKA1-Mid correlator and beamformer system and provide greater commonality as a PCB design across the SKA. This means it could be used as the board for FPGA, ASIC or other non-COTS processors in a Compute Island design, with most of the usual board NRE costs already paid outside the SDP. The SKA could essentially buy processors at wholesale prices which could be competitive with COTS in volume. PowerMX is essentially a flexible COTS vendor-neutral board being developed as an in-SKA specification. It features up to four mezzanines for mounting daughterboards, power supply and general infrastructure. Figure 6 shows a bare motherboard which provides a nearest-neighbour systolic array interconnection between the mezzanines, although a full mesh interconnect specification has also been developed. Designs are underway for two FPGA-based daughterboard variants and a board realisation with 32 I/O 20 Gbps links (2.5 times that of the proposed PCIe version 4.0 rate, and five times that of the current PCIe version 3.0).

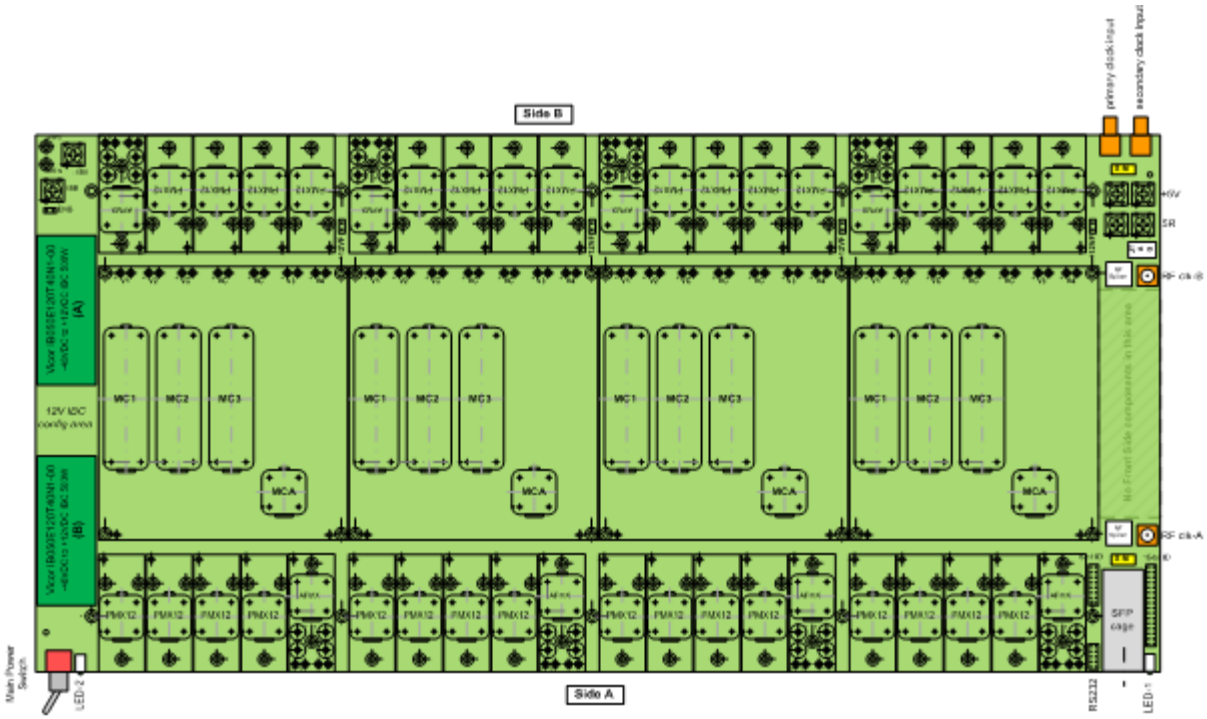

Figure 6: PowerMX SX4-1 Motherboard.

Besides accommodating FPGA and DDR or HMC memory in the mezzanine, the design can be tailored to accommodate other processors or hybrid arrangements. This provides SDP implementers with a greater ability to customise the connectivity and heterogeneity of an SDP Compute Island than possible with a purely COTS solution.

Alternatives to FPGAs and ASICs

Custom accelerator designs

CPUs, FPGAs and ASICs can be combined in a reconfigurable interconnect network. An example topology is shown below.

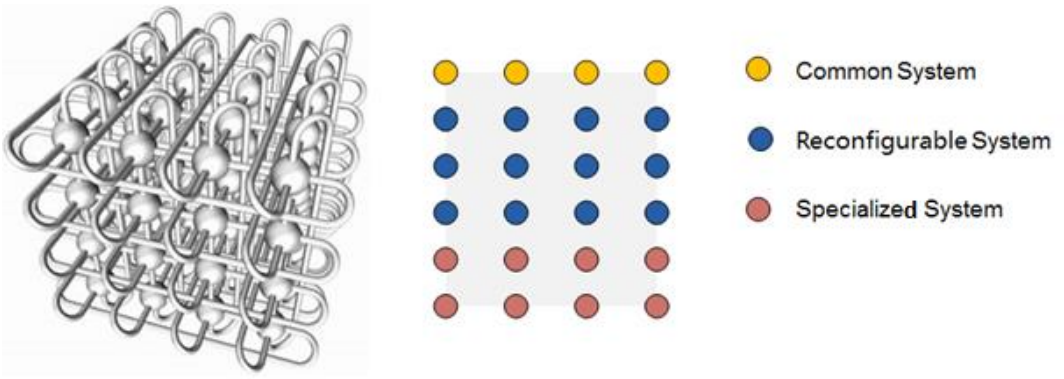

Figure 7: Example topology of a reconfigurable interconnect network combining CPUs, FPGAs and ASICs.

In Figure 7, CPUs, FPGAs and ASICs are represented as common, reconfigurable and specialised systems, respectively. Computing nodes within a Compute Island could be connected by a 3D torus interconnect. Such an interconnect has been chosen before by major HPC supercomputers on the TOP500 list, e.g. 3D torus networks were used by IBM's Blue Gene/L and Blue Gene/P and the Cray XT3. The IBM Blue Gene/Q uses a five-dimensional torus network. Fujitsu's K computer and the PRIMEHPC FX10 use a proprietary six-dimensional torus interconnect called Tofu.
Each computing node contains a small number of new-generation FPGAs, e.g. Xilinx Virtex-7 558T (whose power efficiency is twice that of the Virtex-6). These can be connected in a 2D cross bar as depicted in Figure 8.

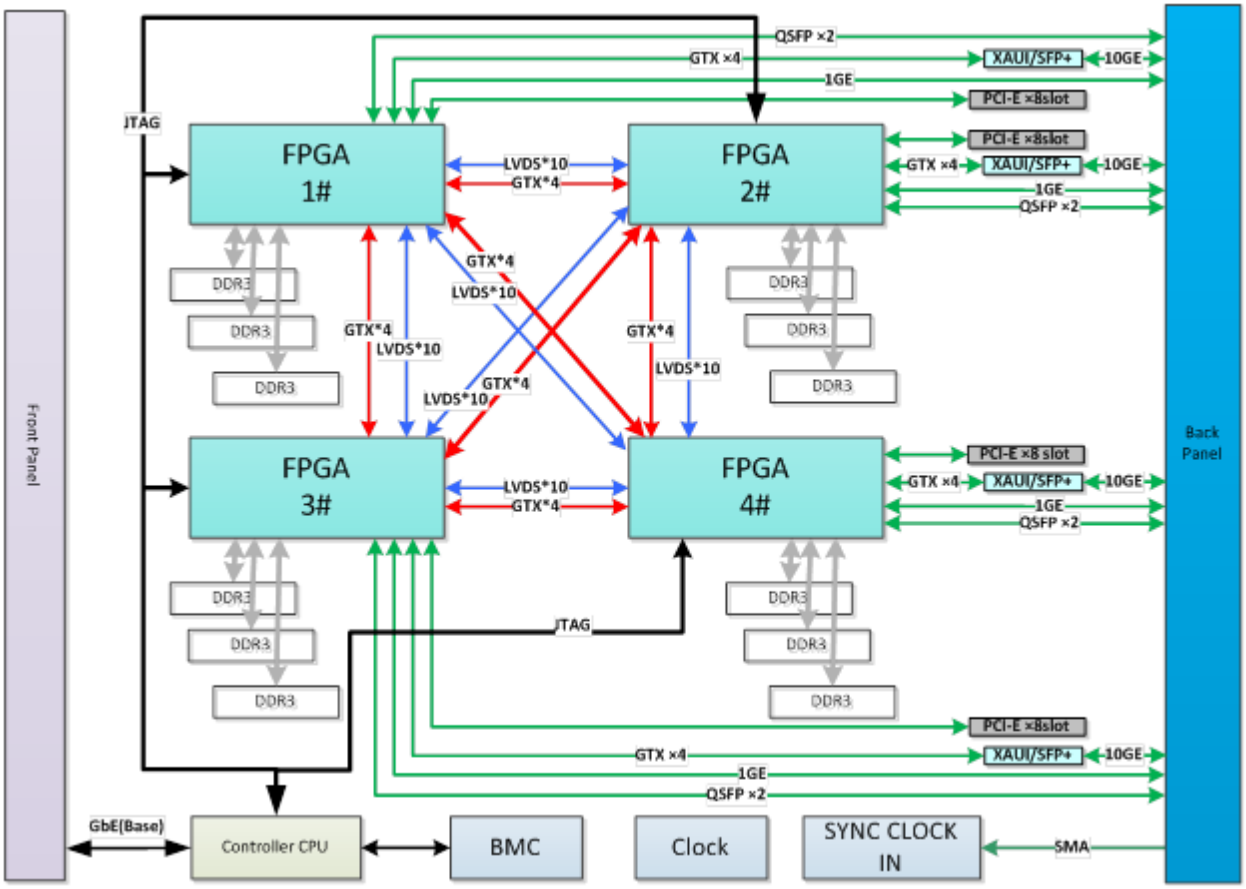

Figure 8: 2D cross bar connection for a custom accelerator design based on FPGAs.

Alternatives for network technologies

Interconnects in high-end HPC systems are currently mainly based on Infiniband (45% of Top500 systems as of November 2014) and Ethernet (38%). There are a number of network technologies, which are expected to become available in 2015 and the years thereafter:

Intel is working on the Storm Lake Fabric technology which will be integrated, e.g. in the next generation of Xeon Phi (https://www.sics.se/sites/default/files/pub/sics.se/avinash_final_sweden_many_core_day_keynote_-_avinash_final_-_clean.pdf).

In 2015 a first ASIC-based implementation of the EXTOLL network architecture is expected to emerge (http://www.extoll.de). It features seven network links with a nominal bandwidth of 12*10 Gbit/s each and supports different network topologies including a 3-dimensional torus.

Bull announced that it is working on a new generation of HPC interconnect technology, without disclosing details (http://www.bull.com/p/register.php?id=262).

Assessing the technical benefits of these new technologies for SDP would require further analysis. Such analysis should also address the sustainability of these technology roadmaps as this could have significant impact on SDP upgrades during the lifetime of the SKA. With these new technologies becoming available and thus competition of HPC interconnect technologies growing, one would expect positive effects on pricing.

Significant impact on network costs may also result from early efforts of different suppliers to enhance high-bandwidth Ethernet technologies such that these would become more suitable for HPC architectures (which, in particular, requires the reduction of latencies). The resulting solutions would partially be based on technologies (e.g. for the physical layer) which have a much broader market than HPC interconnects.

Conclusions

Prototyping

The technologies mentioned in this document need to be prototyped in order to verify their suitability for our applications. In particular, the expected computational efficiency is of interest. This will be done on the road to CDR as described in AD-03.

Technology developments

We need to track technological developments in the next couple of years to make sure we are aware of new directions the industry is taking. We'll do this by strengthening our links with industry and organising a series of workshops with a wide range of vendors. Attending relevant conferences and raising the industry's awareness of the SKA project are also part of our strategy.

Open architecture lab

We're in the process of establishing the Open Architecture Lab (OAL) as a vehicle to support both of the aforementioned activities. The foundations for the OAL have already been laid and it will kicked off in the beginning of 2015.

PDR02-01-01Hardwarealternativesan (1)

EchoSign Document History February 09, 2015

Created:	February 09, 2015
By:	Verity Allan (vla22@mrao.cam.ac.uk)
Status:	SIGNED
Transaction ID:	XJEEY5J7A7C7A3V

“PDR02-01-01Hardwarealternativesanddevelopments (1)” History

Document created by Verity Allan (vla22@mrao.cam.ac.uk)
February 09, 2015 - 2:03 PM GMT - IP address: 131.111.185.15

Document emailed to P.C. Broekema (broekema@astron.nl) for signature
February 09, 2015 - 2:03 PM GMT

Document viewed by P.C. Broekema (broekema@astron.nl)
February 09, 2015 - 2:31 PM GMT - IP address: 192.87.1.200

P.C. Broekema (broekema@astron.nl) verified identity with Google web identity Chris Broekema (https://www.google.com/profiles/118359904325355782244)
February 09, 2015 - 2:33 PM GMT

Document e-signed by P.C. Broekema (broekema@astron.nl)
Signature Date: February 09, 2015 - 2:33 PM GMT - Time Source: server - IP address: 192.87.1.200

Document emailed to Paul Alexander (pa@mrao.cam.ac.uk) for signature
February 09, 2015 - 2:33 PM GMT

Document viewed by Paul Alexander (pa@mrao.cam.ac.uk)
February 09, 2015 - 6:38 PM GMT - IP address: 131.111.185.15

Document e-signed by Paul Alexander (pa@mrao.cam.ac.uk)
Signature Date: February 09, 2015 - 6:38 PM GMT - Time Source: server - IP address: 131.111.185.15

Signed document emailed to Paul Alexander (pa@mrao.cam.ac.uk), P.C. Broekema (broekema@astron.nl) and Verity Allan (vla22@mrao.cam.ac.uk)
February 09, 2015 - 6:38 PM GMT