The Massive Affordable Computing Project: Prototyping of a High Data Throughput Processing Unit

The Massive Affordable Computing Project: Prototyping of a High Data Throughput Processing Unit

Mitchell A. Cox and Bruce Mellado

School of Physics, University of the Witwatersrand. 1 Jan Smuts Avenue, Braamfontein, Johannesburg, South Africa, 2000

E-mail: `mitchell.cox@cern.ch`

Abstract. Scientific experiments are becoming highly data intensive to the point where offline processing of stored data is infeasible. High data throughput computing or High Volume throughput Computing, for future projects is required to deal with terabytes of data per second. Conventional data-centres based on typical server-grade hardware are expensive and are biased towards processing power rather than I/O bandwidth. This system imbalance can be solved with massive parallelism to increase the I/O capabilities, at the expense of excessive processing power and high energy consumption. The Massive Affordable Computing Project aims to use low-cost, ARM System on Chips to address the issue of system balance, affordability and energy efficiency. An ARM-based Processing Unit prototype is currently being developed, with a design goal of 20 Gb/s I/O throughput and significant processing power. Novel use of PCI-Express is used to address the typically limited I/O capabilities of consumer ARM System on Chips.

1. Introduction

Projects such as the Large Hadron Collider (LHC) at CERN and the Square Kilometer Array (SKA) in South Africa generate enormous amounts of raw data which presents a serious computing challenge. After planned upgrades in 2022, the data output from the ATLAS Tile Calorimeter (TileCal) will increase by 200 times to over 40 Tb/s (Terabits/s) [1].

A simple plot, shown in Fig. 1, of the increase in CPU processing power in MIPS (Million Instructions Per Second) and hard drive read-write speed in MB/s (MegaByte/s) over many years clearly demonstrates the fact that hard drive I/O rates are insufficient, and will not become sufficient in the near future, to store the entirety of raw data from modern scientific experiments such as the SKA and LHC [2].

The increase in Ethernet throughput, however, is at a similar rate to the increase in CPU processing power. Based on Amdahl's Laws it has been recommended that approximately one compute instruction per bit of data is required for a balanced system and this relationship is clear when comparing CPU and Ethernet in Fig. 1 [3]. It appears upon inspection that CPU performance and Ethernet throughput are well balanced but in reality high-end Ethernet and other external I/O interconnects in not commonly available except on very high-end and therefore expensive systems. For example, 1 Gb/s Ethernet from 2002 is suitably balanced with a 2002 performance CPU. It is imbalanced when it is coupled with a modern CPU with an order of magnitude higher performance, however this is a very common situation since high-end CPUs are more prevalent than cutting-edge external I/O.

It is important to remember that the balance between I/O and CPU processing power is highly dependant on the workload. For some high-level triggering tasks and event reconstruction in the LHC, for example, the algorithms are very CPU intensive. In other situations, simpler computations such as Optimal Filtering takes place and the system becomes I/O bound.

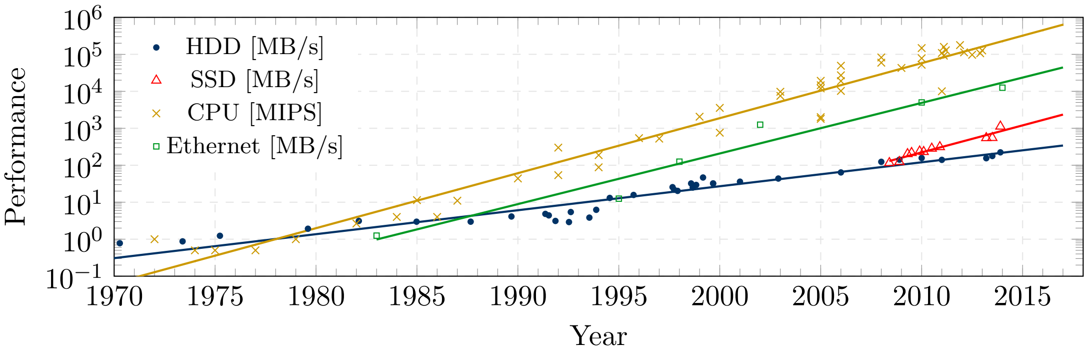

Figure 1: Hard drive (HDD) and Solid State Drive (SSD) and Ethernet 802.3 data throughput and CPU performance on a log scale with time [2, 4].

A specialised triggering and data acquisition system is currently employed by the LHC to reduce the amount of data produced to a manageable quantity for offline storage. This solution is not always suitable and so a paradigm shift is necessary to deal with future workloads and new projects.

The cost, energy efficiency, processing performance and I/O throughput of the computing system to achieve this task is vitally important to the success of future big science projects. Current x86-based microprocessors such as those commonly found in personal computers and servers are often biased towards processing performance and not I/O throughput and are therefore less-suitable for cost-effective high data throughput applications due to the necessity for massive parallelism.

High Volume throughput Computing (HVC) provides a suitable paradigm for high data throughput streaming applications [5]. HVC is a datacenter based computing paradigm where the focus is on loosely-coupled throughput-oriented workloads in terms of either requests (service type applications), processed data (big data applications) or the maximum number of simultaneous subscribers (interactive real-time applications). The definition does not include data-intensive MPI workloads since these are suitably covered by High Performance Computing (HPC).

One of the first steps to the development of an effective HVC system is a high data throughput Processing Unit (PU). This PU should be well balanced in terms of CPU performance and I/O throughput and latency to maximise energy efficiency and cost. The primary aim of the Massive Affordable Computing Project at the University of the Witwatersrand, Johannesburg in South Africa is to develop such a PU.

ARM System on Chips (SoCs) are found in almost all mobile devices due to their low energy consumption, high performance and low cost and are the basis for the PU under development [6]. Section 2 provides a brief overview of the performance of four different ARM-based platforms with various ARM SoCs as well as their specifications. In Section 3 an overview of an architecture for the PU is presented. Finally, a PCI-Express test setup is described and preliminary results are given in Section 4 and Section 5 concludes.

Table 1: CPU benchmark results of the ARM platforms at a fixed clock frequency of as close to 1 GHz as possible.

	Cortex-A7	**Cortex-A9**	**Cortex-A15**	**Cortex-A15**
Core Revision	r0p4	r2p2	r3p2	r3p3
System on Chip	AllWinner A20	Freescale i.MX6Q	Samsung 5410	NVIDIA TK1
Clock (MHz)	1008	996	1000	1090
Cores	2	4	4	4
Feature Size (nm)	40	40	28	28
SP GFLOPS	1.76	5.12	10.56	11.70
DP GFLOPS	0.70	2.40	6.04	6.26
CoreMark	4858	11327	14994	16689
Load Power (W)	2.85	5.03	7.48	6.11
Idle Power (W)	1.16	2.02	1.97	2.26
Calc. Power (W)	1.69	3.01	5.51	3.85
DP GFLOPS/W	0.42	0.80	1.10	1.63
Ethernet (Mb/s)	100	1000	100	1000
PCIe (Gb/s)	-	5	-	40

2. ARM System on Chips

ARM System on Chips (SoCs) are low cost, energy efficient and high performance which has led to their extensive use in mobile devices. Several ARM platforms have been tested by the group and a summary of the specifications and of the CPU benchmark results is presented in Tab. 1.

It is clear that the CPU performance per Watt is increasing with newer generation ARM SoCs. This is naturally driven by consumer demand for faster mobile devices. The exact cost of the SoCs is not known because it is either under Non-Disclosure Agreement (NDA) or is based on order quantity. It can be deduced from the cost of development platforms and the commodity nature of the SoCs presented that the cost is low.

The Freescale i.MX6Q and NVIDIA Tegra-K1 are particularly suitable for a high data throughput PU because of their PCI-Express I/O interface. The available Ethernet interfaces on commodity ARM SoCs are not feasible for multi-gigabit I/O. The Tegra-K1 also has excellent CPU performance and power efficiency due to the small silicon feature size.

3. Processing Unit

A diagram of the PU architecture that is currently being prototyped is shown in Fig. 2. Several ARM SoCs will be integrated onto a single circuit board to minimise duplicated components and form a compact solution. The SoCs will be connected via a PCI-Express switch for a high data throughput interconnect. Finally, a network processor SoC, such as one typically found in high-end network routers, will be used to bridge the internal PCIe network to an external, industry standard Ethernet-based network. The external network can also be another PCIe interface if the PU is integrated closely with an FPGA, for example.

Multiple 10 Gb/s XAUI or SFP+ Ethernet connections via a single server-grade (and therefore more expensive) SoC will enable high data throughput I/O. The use of affordable commodity SoCs will enable significant computational power. The combination of high data throughput I/O and affordable, energy efficient computing forms an excellent candidate for a HVC PU.

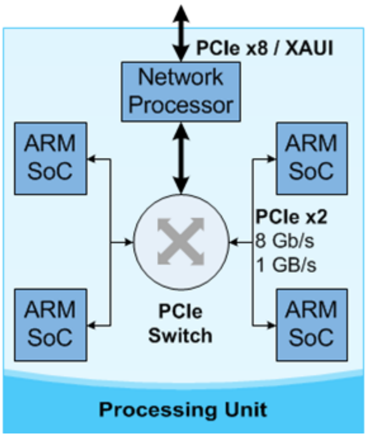

Figure 2: Block diagram of PU architecture showing four hypothetical ARM SoCs, each with PCIe x2 connectivity and an external XAUI (10 Gb/s Ethernet) or PCIe interface.

4. PCI-Express Testing

PCI-Express throughput tests have been performed on a pair of Freescale i.MX6 quad-core ARM Cortex-A9 SoCs clocked at 1 GHz, located on Wandboard development boards [7]. The results are presented in Tab. 2 and a photo of the custom test setup designed by the author is in Fig. 3. Three tests were run to ascertain the maximum data throughput that can be obtained from the i.MX6 SoC: a simple CPU based memcpy command and two Direct Memory Access (DMA) transfers, initiated by the Endpoint (EP) or slave and the Root Complex (RC) which is the host. Unfortunately the i.MX6Q SoC does not have a DMA unit on the PCIe controller and so the Image Processing Unit DMA unit was used instead. This is a workaround provided by the manufacturer.

Table 2: PCI-Express throughput results of a i.MX6 (Wandboard) pair.

	CPU memcpy	**DMA (EP)**	**DMA (RC)**
Read (MB/s)	$94.8 \pm 1.1\%$	$174.1 \pm 0.3\%$	$236.4 \pm 0.2\%$
Write (MB/s)	$283.3 \pm 0.3\%$	$352.2 \pm 0.3\%$	$357.9 \pm 0.4\%$

The theoretical maximum throughput for the PCI-Express Gen 2 x1 link that was used is 500 MB/s. The best result is using DMA initiated by the RC but it is only 72% of the theoretical maximum. The RC-mode drivers are more optimized than the EP-mode drivers due to limited manufacturer support for EP-mode. The read results are lower than write because of overheads to initiate the read. The PU architecture will take these differences into account and use a data push rather than a pull based approach.

A Wandboard to standard PCI-Express x1 slot adapter has also been designed and manufactured by the author. A photo of this is shown in Fig. 4. This will enable future testing of multiple SoCs connected via a PCI-Express switch to closely prototype the proposed PU architecture from Section 3.

5. Discussion, Conclusions and Future Work

High data throughput computing, or more formally High Volume throughput Computing (HVC), is required for projects such as the LHC and SKA which produce enormous amounts of raw

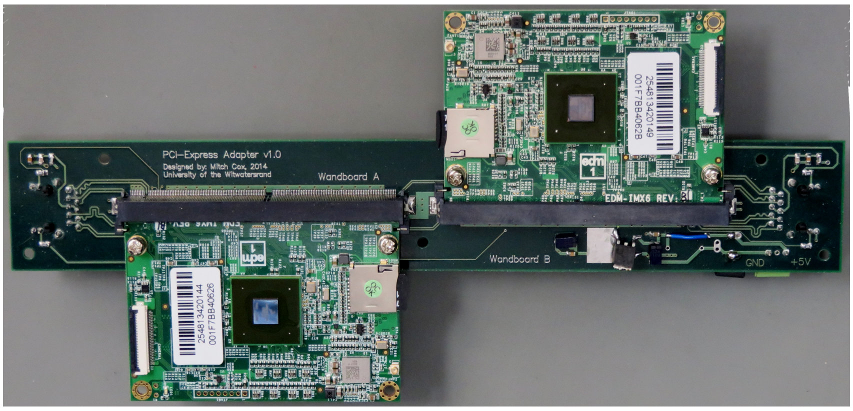

Figure 3: PCI-Express test setup for a pair of i.MX6 SoCs (Wandboards).

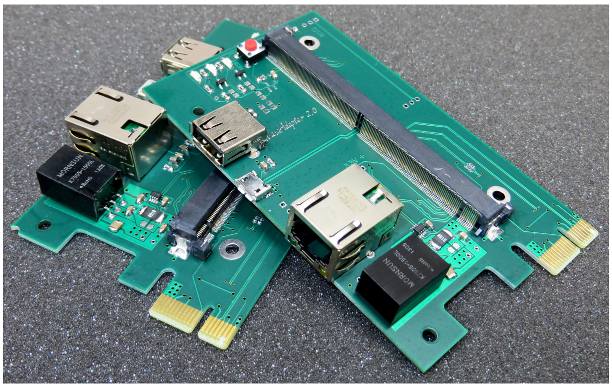

Figure 4: Photo of two Wandboard (via the EDM Connector) to PCI-Express adapters without Wandboards attached.

data. A general purpose ARM System on Chip based processing unit is being developed at the University of the Witwatersrand, Johannesburg which hopes to enable affordable and energy efficient HVC.

In the current prototype, a PCI-Express x1 interface will be used for the raw data transfer between several ARM Cortex-A9 SoCs and a single high-end network processor. This network processor bridges the internal PCIe network to multiple standard external 10 Gb/s Ethernet connections. PCI-Express is superior to Ethernet in energy efficiency, I/O throughput and latency, especially in light of the fact that commodity ARM SoCs do not support Ethernet faster than 1 Gb/s, however PCI-Express is not suitable for longer distance communications.

Initial throughput measurements presented for a pair of Freescale i.MX6 quad-core Cortex-A9 SoCs are 72% of the theoretical maximum 500 MB/s for the available x1 link. Twelve of these SoCs would therefore be connected in parallel to provide 40 Gb/s throughput at a power consumption of less than 100 W. As a proof of concept the Cortex-A9 prototype aims to provide 20 Gb/s aggregated throughput.

The more recent NVIDIA Tegra-K1 SoC supports 40 Gb/s PCI-Express and significantly higher processing power, as well as integrated CUDA GPU processing support. Several of these SoCs will likely be used in the final prototype in future.

The next stage of research by the author will be to test a small PCIe cluster of Cortex-A9 SoCs. An adapter board to connect the Cortex-A9 SoC development board (Wandboard) to a standard PCI-Express x1 connector has been designed and manufactured. A PCI-Express switch development motherboard will be used to connect up to eight devices together. Linux driver development to emulate and Ethernet device over PCIe is currently in progress.

The use of multiple energy efficient commodity ARM SoCs interconnected via PCI-Express and a single higher-end SoC for external communications via multiple 10 Gb/s Ethernet connections is theoretically well suited as a HVC Processing Unit. Future big science experiments may be jeopardised by prohibitive data processing costs but the processing unit presented in this paper may be a possible solution to this problem with its high data throughput, energy efficient and affordable computing capabilities.

Acknowledgements

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. We would also like to acknowledge the School of Physics, the Faculty of Science and the Research Office at the University of the Witwatersrand, Johannesburg. The Massive Affordable Computing Project team and specifically Robert Reed is acknowledged for some of the results presented in this paper.

References

[1] Carrió F *et al.* 2014 *Journal of Instrumentation* **9** C02019–C02019 ISSN 1748-0221 URL `http://stacks.iop.org/1748-0221/9/i=02/a=C02019`

[2] Dursi J 2012 Parallel I/O doesn't have to be so hard: The ADIOS Library Tech. rep. SciNet URL `http://wiki.scinethpc.ca/wiki/images/8/8c/Adios-techtalk-may2012.pdf`

[3] Szalay A S *et al.* 2010 *ACM SIGOPS Operating Systems Review* **44** 71 ISSN 01635980 URL `http://dl.acm.org/citation.cfm?id=1740390.1740407`

[4] Wikipedia 2014 IEEE 802.3 — Wikipedia, The Free Encyclopedia accessed: 16 September 2014 URL `http://en.wikipedia.org/wiki/IEEE_802.3`

[5] Zhan J *et al.* 2012 High Volume Throughput Computing: Identifying and Characterizing Throughput Oriented Workloads in Data Centers *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum* (IEEE) pp 1712–1721 ISBN 978-1-4673-0974-5 URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6270846`

[6] Rajovic N *et al.* 2013 *Journal of Computational Science* **4** 439–443 ISSN 18777503 URL `http://www.sciencedirect.com/science/article/pii/S1877750313000148`

[7] Wandboardorg 2012 Wandboard - Freescale i.MX6 ARM Cortex-A9 Opensource Community Development Board accessed: 18 February 2014 URL `http://www.wandboard.org/`