PGAS in-memory data processing for the Processing Unit of the Upgraded Electronics of the Tile Calorimeter of the ATLAS Detector

Daniel Ohene-Kwofie and Ekow Otoo

School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

E-mail: daniel.ohene-kwofie@cern.ch

Abstract.

The ATLAS detector, operated at the Large Hadron Collider (LHC) records proton-proton collisions at CERN every 50ns resulting in a sustained data flow up to PB/s. The upgraded Tile Calorimeter of the ATLAS experiment will sustain about 5PB/s of digital throughput. These massive data rates require extremely fast data capture and processing. Although there has been a steady increase in the processing speed of CPU/GPGPU assembled for high performance computing, the rate of data input and output, even under parallel I/O, has not kept up with the general increase in computing speeds. The problem then is whether one can implement an I/O subsystem infrastructure capable of meeting the computational speeds of the advanced computing systems at the petascale and exascale level.

We propose a system architecture that leverages the Partitioned Global Address Space (PGAS) model of computing to maintain an in-memory data-store for the Processing Unit (PU) of the upgraded electronics of the Tile Calorimeter which is proposed to be used as a high throughput general purpose co-processor to the sROD of the upgraded Tile Calorimeter. The physical memory of the PUs are aggregated into a large global logical address space using RDMA- capable interconnects such as PCI-Express to enhance data processing throughput.

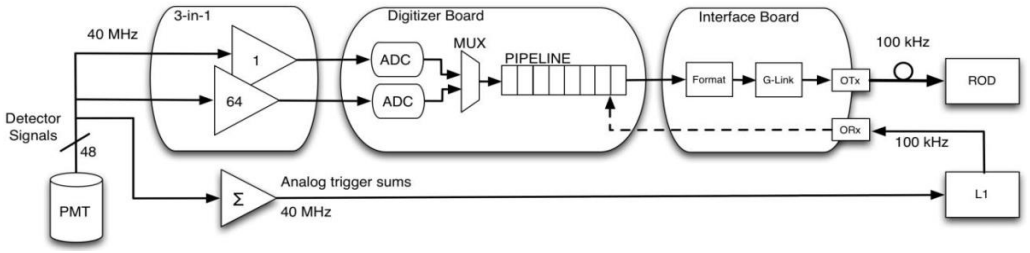
1. Introduction

The Large Hadron Collider, is the most powerful proton-proton collider ever built [1, 2]. The discovery of the Higgs Boson [3, 4] was independently observed by the ATLAS and Compact Muon Solenoid (CMS) detectors. A Toroidal Large Hadron Collider Apparatus (ATLAS) is the largest of all the LHC detectors. It consists of a series of concentric rings: the inner detector, Tile Colorimeters and the Muon Spectrometers. TileCal is the central hadronic calorimeter of the ATLAS experiment at the LHC at CERN. It is primarily used to measure the energy and direction of hadrons and jets as they are produced. The Trigger and Data Acquisition System (TDAQ) is designed for event selection, processing and storage of the read-out data of the detector [5]. This selection mechanism is based on three trigger levels in the data flow that defines the different domains for the read-out electronics in terms of methods and rates for this selection.

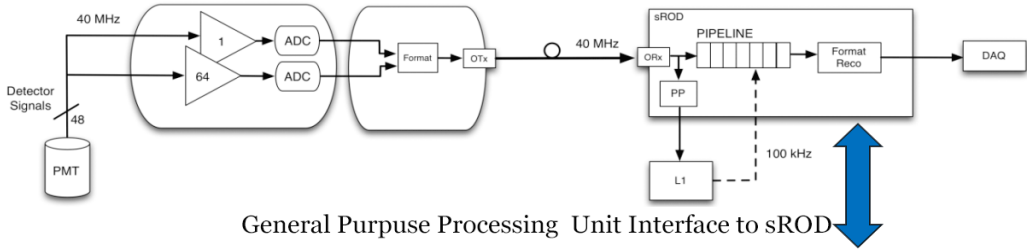
Detector electronics are being upgraded (as part of scheduled upgrade phases on the accelerator and experiments components) to allow an increase in the Level-1 acceptance rate of events from 70kHz to 100kHz [6]. There by subsequently increasing the raw data sent for processing. These upgrades are aimed to extend the physics reach of the LHC measurement programme.

1.1. The Off-line Data processing Challenge

Scheduled upgrades to the ATLAS detector anticipated in 2022 will result in a much higher rate of collisions [7] at the LHC, resulting in an increase by 200 times the current rate to over 41Tb/s data output from the TileCal [5]. Storing such massive dataset for off-line processing presents a



(a) Current Architecture



(b) Upgraded Architecture

Figure 1: ATLAS Tile Calorimeter

great challenge and is not desirable. The MAC project at the University Of Witwatersrand is aiming for a cost-effective, and high data throughput Processing Unit (PU), using several consumer ARM processors in a cluster configuration, as general purpose co-processor to augment the read-out system (sROD) of the upgraded TileCal.

A major challenge with processing such large volumes of data is the inpt/output (I/O) sub-system. The continuously growing gap between CPU and I/O speed has resulted in the conspicuous performance gap between the processor speeds and what the storage I/O subsystem can deliver. The solution is to reduce disk accesses and enhance in-memory data processing.

We present a brief overview of a complete in-memory storage system for on-line data processing in the ARM cluster configuration for the upgraded electronics of the TileCal. In-memory data processing provides extremely fast response time and very high throughput, with an average of about $100-1000$ times lower latency for a complete Random Access Memory(RAM) storage than disk-based storage systems and consequently a $100-1000$ times greater throughput [8].

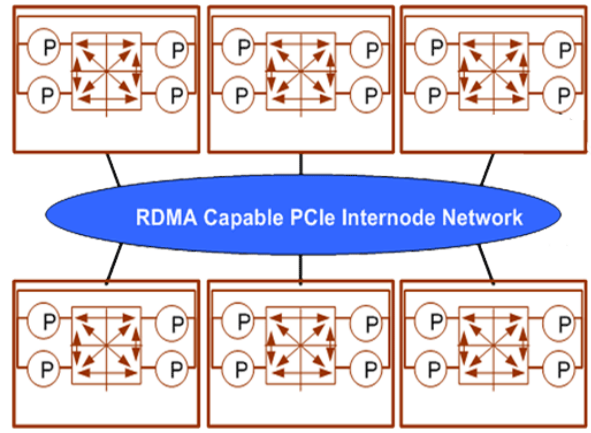
2. PGAS Architecture

The architecture consists of a cluster of ARM processing units, part of whose memories are aggregated to form a combined logical global address space. The general operational architecture includes low cost ARM processing units interconnected via PCI Express interconnects (PCIe). The PCIe interconnect facilitates Remote Data Memory Access (RDMA) and offers very low-latency host-to-host transfers by copying the information directly between the host application memories. This enhances a seamless global logical address space for in-memory data processing as a low-overhead protocol. They are relatively affordable, low power and also provide straightforward, standards-compliant extensions that address multi-host communication and I/O sharing capabilities. Figure 2a illustrates the schematic diagram of the architecture. Part of the on-going research work is to address a number of challenges related to fault-tolerance and high speed access of the memory resident data.

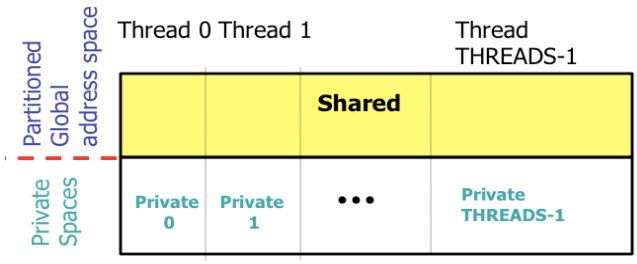
3. The PGAS High Data Throughput Architecture

The Figure 2 and the experiments conducted so far only demonstrates the processing capability of the PU under the PGAS computing model. This is easily transformed into a configuration where the aggregated memories of the ARM processors serve as a large in-memory buffer for data staging. By augmenting the ARM cluster with data input ports and data output ports to external hierarchical storage devices, data can be streamed into the ARM cluster memory using parallel I/O. This in-memory data can then be compressed and streamed out using parallel I/O to a first level of attached solid state disks (SSDs) and subsequently to archival disks. We envisage that the high processing speed of compressing the data and streaming

it out to other high speed devices or memory of other devices, will outweigh the rate that data is streamed into the ARM cluster. Experiments are underway to test this architecture.



(a) Schematic Diagram of the Architecture



(b) PGAS Memory layout

Figure 2: Schematic diagram showing operational layout of the architecture (a). (b) shows the memory layout indicating the global shared memory as well as the private address space. Data is distributed to enhance locality and thus ensure greater throughput.

4. Preliminary Evaluations

Preliminary investigations conducted show promising results. We benchmark the system with the NASA Advanced Supercomputing (NAS) Parallel Benchmarks [9]. This benchmark was designed not just for parallel-aware algorithmic and software methods but also to provide an easy verifiability of correctness of results and performance figures. The evaluations are run using the Fast Fourier Transform(FFT) algorithm which solves a 3D partial differential equation using an FFT-based spectral method [9], also requiring long range communication. FFT performs three one-dimensional(1-D) FFT's, one for each dimension.

The evaluation was done on 4 nodes of the Wits Tegra k1 (2.3GHz Quad-Core ARM Cortex-A15) cluster with 2GB of memory each and 1Gbp Ethernet interconnect between nodes.

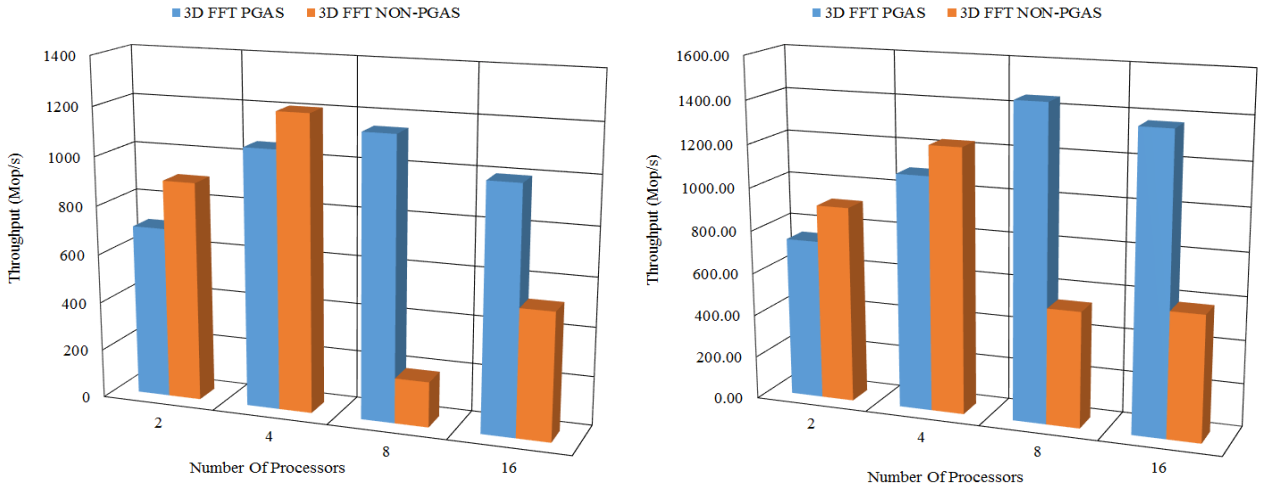
We ran the benchmark with varying workloads as well as varying number of threads. Each workload is ran 6 times and the resulting throughput in MFLOPS (floating point operations per second) reported. A maximum of 4 threads per node are spawned and the dimensions of the FFT is varied from small ($64\times64\times64$ 3D grid) to large ($256\times256\times128$ 3D grid). The User Datagram Protocol(UDP) is used as the inter-process communication protocol between cluster nodes. UPC supports UDP, Message Passing Interface (MPI) as well as RDMA capable interconnects(e.g infiniband, PCIe, 10GbE, etc.) for inter-node communication. The current experimental setup did not have RDMA and therefore UDP was used since it is usually faster than MPI as far as inter-node communication over Ethernet is concerned.

Figures 3a and 3b show the results for the data processing throughput in MFLOPS with varying number of threads and workloads. Generally, as the number of threads increase, there is a corresponding increase in the data processing throughput since less processing is done per thread. We observe a significant and better performance increase with PGAS as depicted in the figures (about $3\times$ more throughput than the NON-PGAS FFT).

Figure 4 shows the average latencies for each run of the experiment. There is a general increase in latency as workload size increases as expected. We also observe that as the number of threads increase latency drops a little and begins to increase after 8 threads. This is due to increase in inter-node communication. PGAS performs much better with lower latencies as compared to its NON-PGAS variant.

5. Conclusions

Management, efficient access and analysis of the Petabytes of data, that is likely to be generated and/or used in the upgraded ATLAS TileCal present extremely challenging tasks. I/O bottlenecks in processing such huge amounts of data require techniques that utilise higher levels of the memory hierarchy to enhance data throughput.



(a) small workloads (b) large workloads

Figure 3: Data Throughput

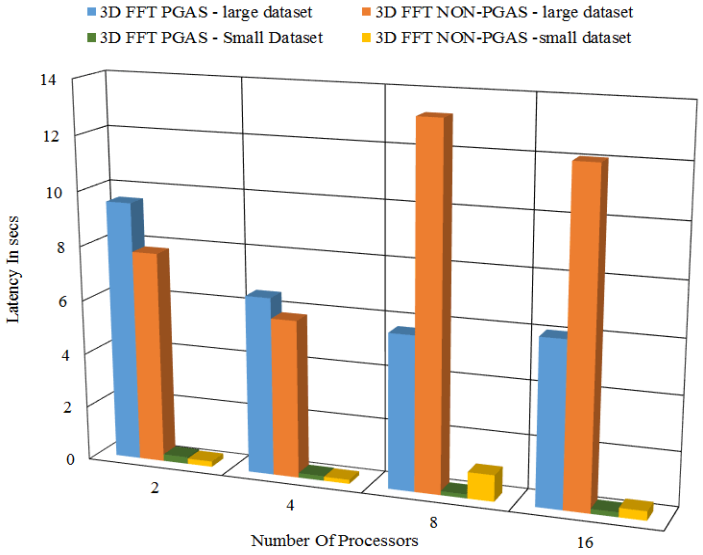


Figure 4: Average latencies for varying workloads

The PGAS model presents an efficient way to process data in a distributed environment by providing a global partitioned shared address space for in-memory data. This enhances the data processing throughput . Additionally, the use of low cost CPUs such as ARM or Atom PUs with PCIe interconnects, ensure low power consumption (high performance per watt) and thus cost effective alternative for data processing in ATLAS TileCal.

Future work anticipated includes further rigorous experimentation using the ARM PUs with RDMA capable PCIe intra/interconnects for kernel bypass applications.

References

[1] CERN Communication Group, "The CERN LHC faq Guide." `http://cds.cern.ch/record/1165534/files/CERN-Brochure-2009-003-Eng.pdf`, February 2009.

[2] CERN, "The Large Hadron Collider." `http://home.web.cern.ch/topics/large-hadron-collider`, 2014.

[3] J. T. Baines and et al., "An overview of the atlas high-level trigger dataflow and supervision," *Nuclear Science, IEEE Transactions on*, vol. 51, pp. 361–366, June 2004.

[4] R. Reed, F. Carrio, P. Moreno, C. Solans, J. Souza., and A. Valero, "A revised high voltage board for the consolidation of front end electronics on the tile calorimeter of the atlas detector at the lhc," in *South African Institute of Physics Conference,SAIP 2013 Poster Presentation*, (University Of the Witwatersrand, Johannesburg, South Africa), Jul 2013.

[5] F. Carrioa and et. al, "The sROD module for the ATLAS Tile Calorimeter Phase-II Upgrade Demonstrator," *Journal of Instrumentation, JINST*, vol. C02019, no. Issue 09, 2014.

[6] The ATLAS Collaboration, "Letter of Intent for the Phase I Upgrade of the ATLAS." `https://cdsweb.cern.ch/record/1402470`, 2012.

[7] The ATLAS Collaboration, "Letter of Intent for the Phase II Upgrade of the ATLAS." `https://cdsweb.cern.ch/record/1502664`, 2012.

[8] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazières, S. Mitra, A. Narayanan, D. Ongaro, G. Parulkar, M. Rosenblum, S. M. Rumble, E. Stratmann, and R. Stutsman, "The Case for RAMCloud," *Commun. ACM*, vol. 54, pp. 121–130, July 2011.

[9] O. Serres, N. Andreev, C. Francois, A. Abhishek, A. Smita, B. Veysel, Y. Yiyi, S. Chauvin, F. Vroman, and T. El-Ghazawi, "UPC NAS Parallel Benchmarks." `http://threads.hpcl.gwu.edu/sites/npb-upc`, April 2011.