HIOS: Heterogeneous I/O for Scale

Viktor Khristenko,[1*] Maria Girone,[1] Mondrian Nuessle,[2] Dirk Frey,[2,3] Ulrich Bruening,[3] Judith Schonbohm[3]

[1]**CERN, 1211, Geneva 23, Switzerland;**
[2]**Extoll GmbH, Rheinvorlandstrasse 5, 68159 Mannheim, Germany;**
[3]**Heidelberg University, Postfach 10 57 60, 69047 Heidelberg, Germany;**
*Corresponding author: viktor.khristenko@cern.ch

ABSTRACT

The project HIOS aims to investigate ways of providing efficient interfaces for I/O intensive applications. As the world is increasingly becoming more and more data driven, with humongous quantities of data constantly being moved around, processed, analysed and stored, the infrastructure employed needs to cope with ever increasing network traffic requirements. The main idea is to explore the use of heterogeneous resources to make them act as I/O intermediary that handles data intensive traffic.

Keywords: Big Data; HPC; FPGA; CPU; GPGPU; TCP/IP.

1. INTRODUCTION

With the arrival of the era of Big Data, the quantities of generated and collected information have grown exponentially. It served as a catalyst for various fields of science and technology that accelerated their further development and allowed for significantly deeper insights in their respective disciplines. For example, in the field of High Energy Physics (HEP) Petabytes of data from the Large Hadron Collider (LHC) are accumulated in order to probe for new type of physics and interactions. Success and abundance of the Artificial Intelligence (AI) algorithms, in essentially everyday of our lives, is built on top of having these enormous quantities of data. For the business side, more data and ability to efficiently extract value out of it created new markets and new types of assets companies are trying to accumulate. For instance, online sales giants can do personalized targeting, which in turn improves customer experience and increases company's revenues – all thanks to more and more data that is being accumulated all the time.

On the other hand, as with any breakthrough, there are associated challenges to be overcome in order to make efficient use of the abundance of data. For the High-Performance Computing (HPC) farms, Data Centres, or any other type of environment, it is crucial to be able to handle this increase in the amounts of data efficiently. One of the recent advances, as a response to higher computing requirements, is the move towards the use of heterogeneous computing resources (e.g. GPGPUs, FPGAs, specialized ASICs), which are more efficient than traditional Central Processing Units (CPUs). However, although this does address the computational challenges of the Big Data era, the problem of having to move the data around constantly remains.

The project HIOS also moves in the direction of heterogeneous computing but tries to use the same principle not to accelerate computations, but to provide a scalable low-latency FPGA-based platform for data-intensive applications that perform heavy data ingestion.

Therefore, during the first phase of ATTRACT, time has been mostly devoted to the following activities:

- Exploring available implementations of the TCP/IP stack for FPGAs. Typically called TCP Offloading Engines (TOE).
- Identifying common Big Data related patterns and what can be further offloaded (e.g. compression/decompression).
- Integration of the existing NAM logic with the TCP/IP network stack.

2. STATE OF THE ART

Although processing data in the network is a hot research topic now, HPC facilities typically do not employ much of this technique currently. One will, of course, find high-bandwidth low latency network connectivity that allows for efficient message passing between the nodes. In other words, for the case of data intensive applications that require ingesting a lot of data (but do not require so much data exchange between the nodes), there is not much help available in terms of hardware. Essentially all the processing happens on the compute node, even though there are clearly interfaces that are common for all data intensive applications that could be factored out (e.g.

caching that lowers application's memory consumption, simple filtering, etc.). To put it another way, the way batch data crunching is currently done is essentially the same as it used to be done 10 years ago, with differences coming for the most part in more powerful CPUs (and/or GPUs), more memory and faster networks, although the amounts of data has exponentially increased.

To add further about the state-of-the-art, Extoll's NAM product is by itself also a new type of platform, currently being extensively tested as part of the DEEP-EST prototype, EU project hosted by the Julich Supercomputing centre to build the next generation of the modular supercomputing architecture. Currently, NAM provides Remote Direct Memory Access (RDMA) functionality, acting essentially as another layer in the memory hierarchy.

Another interesting benefit that comes as a result of such separation of concerns is with presence of GPUs on the compute nodes. Typically, the way GPUs (or other types of accelerators) are employed for big data applications is that, first, data is read into CPU's main memory from a remote storage, certain I/O related logic is performed by CPU (e.g. compression, decompression) and only then data is further transferred to the GPU for further processing. The benefit here of having an I/O intermediary is that you can bypass CPU and memory involvement completely and transfer data directly to the GPU's memory (already available in the Extoll's NAM platform). To summarize, the idea of the project is to factor out what is common for all Big Data type of workloads and provide a scalable pluggable infrastructure solution that would implement the necessary interfaces efficiently.

3. BREAKTHROUGH CHARACTER OF THE PROJECT

As it has been established in the previous section, a single common aspect of all data intensive applications is that all of them require ingesting a lot of data to the place where computations are to be performed. Furthermore, all types of Big Data workloads perform also various kinds of common operations such as: compression, various types of I/O specific coding and decoding, simple types of filtering.

The main idea of the HIOS project is to offload this I/O specific functionality from the compute nodes onto heterogeneous I/O intermediary units. We try to leverage the existing Extoll's NAM platform that already provides RDMA functionality towards the compute nodes. Introducing such specialized I/O units would essentially allow to remove latency related to ingesting from a remote location and lower the burden on a compute node's memory as aggressive caching will not be needed as much. Furthermore, other logic (e.g. compression, filtering, coding and decoding) can also be moved down to such a specialized unit.

4. PROJECT RESULTS

The core outcome of the first phase of the HIOS project was finding, testing and validating the TCP/IP stack that could potentially be used to integrate with the existing Extoll's NAM solution. The TCP/IP FPGA-based network stack was tested using Xilinx VCU118 evaluation board with 10G Ethernet link.

As was already mentioned, one of the common operations that all data intensive workloads perform, once the data is on the compute node, is the compression and decompression of streamed buffers. Analysis of the most widely used algorithms (mainly looked at LZ4, LZMA, zstd, Deflate, 842B) has been performed and implementation is still ongoing. The selection was made based on what typically gets used for the Big Data applications.

Fig. 1 shows a substantially simplified version of the foreseen architecture once TCP/IP stack is included in the Extoll's NAM implementation. Note, this is not the final version and provided here for illustration purposes only.

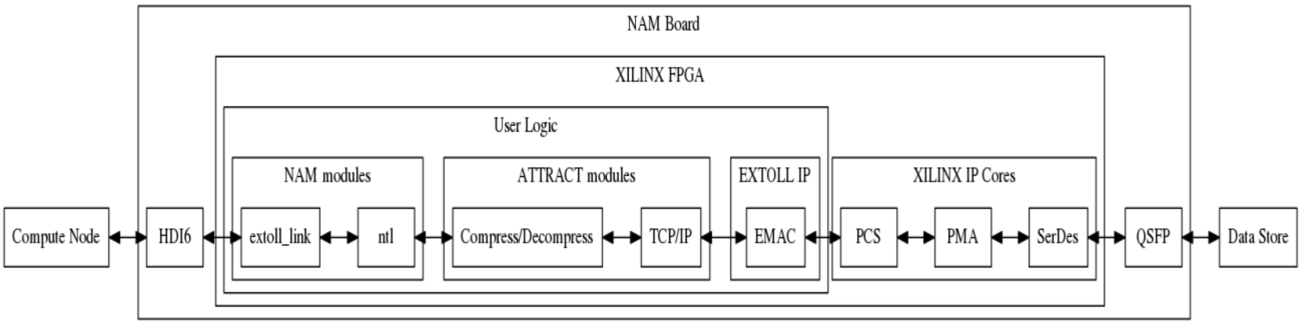

Fig. 1. Simplified version of the envisioned Extoll's NAM solution with TCP/IP stack. Provided for illustrational purposes only

5. FUTURE PROJECT VISION

The HIOS project vision for the ATTRACT Phase 2 is to go from exploration, testing and validation mode to a ready-to-go product that is capable to be integrated at HPC centres and other computing facilities.

5.1. Technology Scaling

As it has already been mentioned earlier, the HIOS project is based on the Extoll's NAM solution that is currently being actively tested at the DEEP-EST HPC facility. Therefore, one of the core values of the HIOS project is that scaling was envisioned since the very beginning. Therefore, the next steps would essentially boil down to: first, complete and validate the full NAM, TCP/IP network stack and necessary add-on interfaces, which requires both verification/simulation and also using test setups; second, find an HPC (e.g. DEEP-EST) facility and test out the functionality. Given the DEEP-EST facility already employs Extoll's NAMs, it should be relatively easy to apply changes to the system in order to make use of the new functionality.

5.2. Project Synergies and Outreach

The idea is to mainly reach out to organizations that are actively working on applying similar techniques. For example, there are plenty of research groups that do "processing in the network" type of research. The goal is to bridge people from different backgrounds (e.g. FPGA development vs Big Data Analytics) and work together to achieve ATTRACT Phase 2 objectives.

So far, there have been no contacts with other ATTRACT Phase 1 funded projects. However, this option will be further explored during the final conference.

5.3. Technology application and demonstration cases

The expected benefits for the society remain unchanged since what has been proposed at the beginning of the project. The fundamental goal of all data intensive applications is to gain insights from the vast quantities of available data. The way we help achieve this goal is by reducing the time to insight. In more concrete terms, the HIOS project aims to provide a faster turn-around for anyone performing heavy data ingestion. For example:

- *For fundamental research, quicker turn-around reduces time to discoveries of the fundamental building blocks of nature;*
- *For medical domain, reduced time to computational solution leads to faster discovery of virus spreads, analysis of patients' symptoms, and diagnosis of diseases;*

Another potential benefit is the integration with existing infrastructure, like HPC centres, which in turn leads to lower maintenance costs and better reuse of the existing computing environments.

5.4. Technology commercialization

Since the work done in ATTRACT Phase 1 was performed based on the already available Extoll's platform, any future commercialization will be done in line with current procedures used by Extoll for these matters.

5.5. Envisioned risks

One of the main risks associated with the product being developed is that it tries to use 2 types of fabrics, Extoll and Ethernet, which is both costly and potentially inefficient for the future. One of the possible solutions to this problem is that Extoll links could be removed completely, but then RDMA must be done over Ethernet. This is precisely what RDMA over Converged Ethernet (RoCE v2) protocol can be used for. The main benefit would be the applicability of the product as Ethernet is the type of fabric that essentially is the default standard almost everywhere nowadays. However, that requires further development and testing.

5.6. Liaison with Student Teams and Socio-Economic Study

The team of the HIOS project is happy to work and interact with students. Since Heidelberg University is part of the project and some of the investigations are done by students, this interaction will be carried out further.

Any type of information sharing, and dissemination is welcomed by the team of the HIOS project. Using various blogs is a proven channel to communicate the technological developments. Furthermore, we would be happy to participate in any type of social study and provide the necessary information for that.

6. ACKNOWLEDGEMENT

Authors thank CERN EP-CMD group for providing access to the necessary evaluation hardware, during the difficult conditions with teleworking only policy.

This project has received funding from the ATTRACT project funded by the EC under Grant Agreement 777222"

7. REFERENCES

[1] Schmidt J. & Bruning U., 2015, openHMC – A configurable Open-source Hybrid Memory Cube Controller, 10th IEEE International Conference on Reconfigurable Computing and FPGAs, 2015, Mayan Riviera, Mexico.
[2] Schmidt J., 2016, NAM: Network Attached Memory, The International Conference for High Performance Computing, Networking, Storage and Analysis, 2016, Salt Lake City
[3] Sidler D. et al., 2015, Scalable 10 Gbps TCP/IP Stack Architecture for Reconfigurable Hardware, 2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines, Vancouver, 2015, pp. 36-43