Contribution ID: **488** Type: **Oral**

High-Throughput Network Communication with NetIO

Tuesday, 11 October 2016 14:15 (15 minutes)

HPC network technologies like Infiniband, TrueScale or OmniPath provide low-latency and high-throughput communication between hosts, which makes them attractive options for data-acquisition systems in large-scale high-energy physics experiments. Like HPC networks, data acquisition networks are local and include a well specified number of systems. Unfortunately traditional network communication APIs for HPC clusters like MPI or PGAS exclusively target the HPC community and are not well suited for data acquisition applications. It is possible to build distributed data acquisition applications using low-level system APIs like Infiniband Verbs, but it requires non negligible effort and expert knowledge.

On the other hand, message services like 0MQ have gained popularity in the HEP community. Such APIs facilitate the building of distributed applications with a high-level approach and provide good performance. Unfortunately their usage usually limits developers to TCP/IP-based networks. While it is possible to operate a TCP/IP stack on top of Infiniband and OmniPath, this approach may not be very efficient compared to direct use of native APIs.

NetIO is a simple, novel asynchronous message service that can operate on Ethernet, Infiniband and similar network fabrics. In our presentation we describe the design and implementation of NetIO, evaluate its use in comparison to other approaches and show performance studies.

NetIO supports different high-level programming models and typical workloads of HEP applications. The ATLAS front end link exchange project successfully uses NetIO as its central communication platform.

The NetIO architecture consists of two layers:

- The outer layer provides users with a choice of several socket types for different message-based communication patterns. At the moment NetIO features a low-latency point-to-point send/receive socket pair, a high-throughput point-to-point send/receive socket pair, and a high-throughput publish/subscribe socket pair.
- The inner layer is pluggable and provides a basic send/receive socket pair to the upper layer to provide a consistent, uniform API across different network technologies.

There are currently two working backends for NetIO:

- The Ethernet backend is based on TCP/IP and POSIX sockets.
- The Infiniband backend relies on libfabric with the Verbs provider from the OpenFabrics Interfaces Working Group.

The libfabric package also supports other fabric technologies like iWarp, Cisco usNic, Cray GNI, Mellanox MXM and others. Via PSM and PSM2 it also natively supports Intel TrueScale and Intel OmniPath. Since libfabric is already used for the Infiniband backend, we do not foresee major challenges for porting NetIO to OmniPath, and a native OmniPath backend is currently under development.

Tertiary Keyword (Optional)

Computing middleware

Secondary Keyword (Optional)

Distributed data handling

Primary Keyword (Mandatory)

Network systems and solutions

Primary author: PANDURO VAZQUEZ, Jose Guillermo (Royal Holloway, University of London)

Co-author: SCHUMACHER, Jorn (University of Paderborn (DE))

Presenter: SCHUMACHER, Jorn (University of Paderborn (DE))

Session Classification: Track 6: Infrastructures

Track Classification: Track 6: Infrastructures