High-Speed Network Communication with NETIO

Jörn Schumacher
CERN and Paderborn University
jorn.schumacher@cern.ch
ATLAS Data Acquisition:
- 60 TB of raw data produced every second
- 1000s of PCs with more than 40'000 cores
- 10000s of applications sending and receiving messages
- Core network: 25 GB/s switched in core routers

Today’s DAQ systems rely heavily on high-speed network technology
DAQ Networks in High-Energy Physics

ATLAS Data Acquisition:
- 60 TB of raw data produced every second
- 1000s of PCs with more than 40'000 cores
- 10000s of applications sending and receiving messages
- Core network: 25 GB/s switched in core

We need fast network technologies to build DAQ systems
HPC Network Market in 2016

Interconnect Families in Top500 List in July 2016

- Infiniband: 40.8%
- 10G Ethernet: 8.4%
- Custom: 14%
- 1G Ethernet: 35.4%
- Proprietary

Ethernet and Infiniband are the two dominating technologies in the HPC market.

ATLAS, ALICE, CMS and LHCb all use Ethernet and/or Infiniband in their DAQ systems.

Source: top500.org
Programming Models (subjective)

- Message Queues
- POSIX Sockets
- MPI
- Verbs (native Infiniband API)
Programming Models (subjective)

- Verbs (native Infiniband API)
  - "Hello world": >800 LOC
- MPI
- Message Queues
- POSIX Sockets
Programming Models (subjective)

- **Verbs (native Infiniband API)**
  - "Hello world": >800 LOC
- **MPI**
  - Targeted at HPC domain
  - (single-program-multiple-data paradigm)
- **Message Queues**
- **POSIX Sockets**
Programming Models (subjective)

Throughput

Suitability

Verbs (native Infiniband API)
“Hello world”: >800 LOC

MPI
Targeted at HPC domain
(single-program-multiple-data paradigm)

Message Queues

POSIX Sockets
Many things have to be implemented by hand
Programming Models (subjective)

Throughput

Verbs (native Infiniband API)
“Hello world”: >800 LOC

MPI
Targeted at HPC domain
(single-program-multiple-data paradigm)

POSIX Sockets
Many things have to be implemented by hand

Message Queues
Only Ethernet

Suitability
Wishlist for a Software Network Library

Support different network technologies (Ethernet, Infiniband and more) seamlessly

High performance (high throughput, low latency)

Message-based, high-level user interface with use-cases for typical DAQ applications

Adjust to future technology developments

Tailored for our applications (performance)

Tailored for our applications (programming interface)
See talk by Soo Ryu on FELIX
NetI\O

Support different network technologies (Ethernet, Infiniband and more) seamlessly

High performance (high throughput, low latency)

Message-based, high-level user interface with use-cases for typical DAQ applications
NetI0: Libfabric Backend

Image source: OFI Working Group
NetIO: Libfabric Backend

Image source: OFI Working Group
NetIO

Support different network technologies (Ethernet, Infiniband and more) seamlessly

High performance (high throughput, low latency)

Message-based, high-level user interface with use-cases for typical DAQ applications
1) Low-Latency Send/Receive

2) High-Throughput Send/Receive

3) Publish/Subscribe
int main(int argc, char** argv) {
    netio::context ctx("posix");
    std::thread bg_thread([&ctx](){
        ctx.event_loop()->run_forever();
    });

    netio::subscribe_socket socket(&ctx);
    socket.subscribe(SUBSCRIPTION_TAG, netio::endpoint("10.113.142.1", 1234));

    while(1) {
        netio::message m;
        socket.recv(m);
        std::cout << m.data_copy().data() << std::endl;
    }
}
NetIO

Support different network technologies (Ethernet, Infiniband and more) seamlessly

High performance (high throughput, low latency)

Message-based, high-level user interface with use-cases for typical DAQ applications

Benchmark Platform:
Intel Haswell CPUs @ 2.4 GHz
40G Ethernet
56G Infiniband FDR
NetI/O Development Status

Actively in use in the ATLAS FELIX project

High-Throughput sockets (point-to-point and publish/subscribe) work reliably

Low-latency sockets need more tuning

Add support for more backends (Intel OmniPath)
If you are interested in NetI/O, please contact Joern.Schumacher@cern.ch