USING ALFA FOR HIGH THROUGHPUT, DISTRIBUTED DATA TRANSMISSION IN ALICE O^2 SYSTEM

Presented by Adam Wegrzynek
On behalf of the ALICE Collaboration
CERN, Geneva, Switzerland

ALICE O^2 COMPUTING SYSTEM

The ALICE O^2 (Online-Offline) computing system will allow to record Pb-Pb collisions at 50 kHz rate. Some detectors will be read out continuously, without physics triggers. Instead of rejecting events O^2 will compress data by online calibration and partial reconstruction.
The first part of this process will be done in dedicated FPGA cards that receive raw data from detectors. The cards will perform baseline correction, zero suppression, cluster finding and inject the data into FLPs[1] memory to create a sub-timeframe. Then, the data will be distributed over EPNs[2] for aggregation and additional compression.
The O^2 facility will consist of 268 FLPs and 1500 EPNs. Each FLP will be logically connected to each EPN through a high throughput network. The O^2 farm will receive data from detectors at 27.2 Tb/s, which after processing will be reduced to 720 Gb/s.

MOTIVATION

Transferring and processing Tb/s of data inside the O^2 system is a challenge for the network and computing resources. Assuming a throughput of 40 Gb/s, the distance between Ethernet frames is short – 300 ns. During that time the Linux kernel has to go through the whole TCP/IP stack and deliver the data to user space which consumes a large amount of computing resources. This work aims at estimating the CPU needs for data transport inside the O^2 system.
In addition, the high number of nodes and connection in the final set up may cause race conditions that can lead to uneven load balancing and poor scalability. The performed tests allow to validate whether the traffic is distributed evenly over all receivers. It also measures the behaviour of the network in saturation and evaluates scalability from a 1-to-1 to a N-to-M solution.

TEST PLATFORMS

Network	Network adapter	CPU
40 GbE[3]	Chelsio T580	Intel E5-2690
IB[4] FDR (56 Gb/s)	Mellanox MT27500	Intel E5-2690
OPA[5] (100 Gb/s)	-	Intel E5-2680v4

ALFA

The ALFA (ALICE-FAIR) is common layer of the jointly developed framework by ALICE O^2 and FAIR teams. It is based on message-like approach and standardizes software related tasks such as data transport, state machines, frame building, configuration and monitoring and more.
ALFA supports two data transport libraries: ZeroMQ and nanomsg.

BENCHMARKS

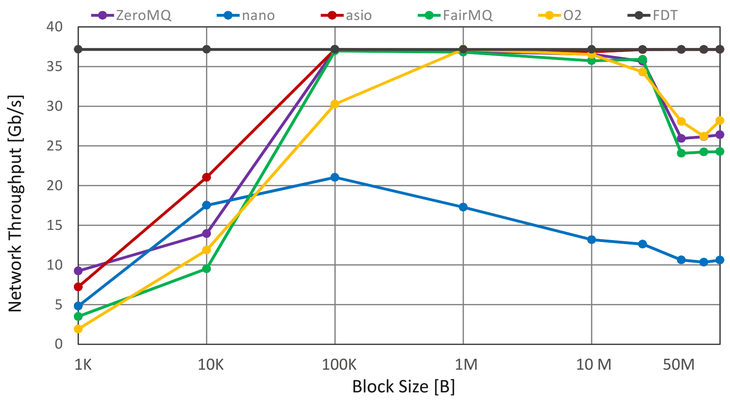
- **ZeroMQ** – message-based library supporting a large number of socket patters that help to create complex, distributed systems.
- **nanomsg** – fork of ZeroMQ with the ability to plug custom transports, improved threading model and state machine.
- **asio** – asynchronous, low level I/O library.
- **FairMQ** – high level transport framework with internal state machine and ability to work on top of a lower level network library such as ZeroMQ and nanomsg.
- **O2** – development version specific to the ALICE framework that is built on the top of ALFA (FairMQ library).
- **FDT**[6] - reference benchmark that measures bandwidth of the network, uses all available CPU cores and multiple TCP streams.

SINGLE FLP, SINGLE EPN

All listed benchmarks where tested on all platforms using a single FLP, single EPN architecture. The measurements allowed to evaluate the performance of each pair of network library and technology, and chose the most efficient combination for further test.

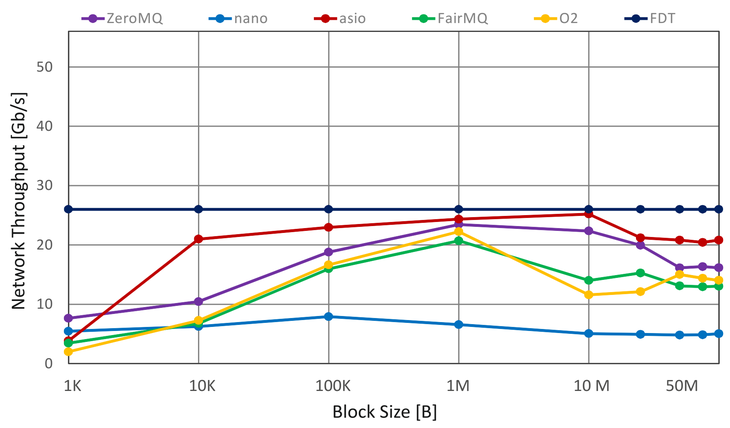
ETHERNET

- Nanomsg - version 1.0 fixes issue that blocked transferring larger data blocks (>1MB).
- Throughput decrease for block sizes larger than 25 MB for all benchmarks except asio.



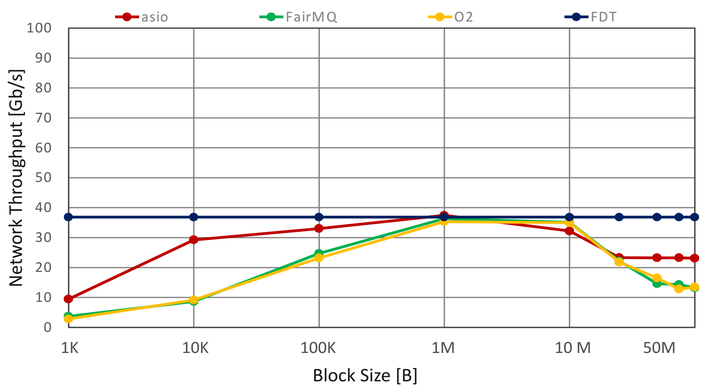
IPoIB (INFINIBAND)

- Decreasing throughput for data blocks larger than 25MB is observed.
- IPoIB has large overhead. The throughput is limited to 25 Gb/s out of 56 Gb/s of available bandwidth.



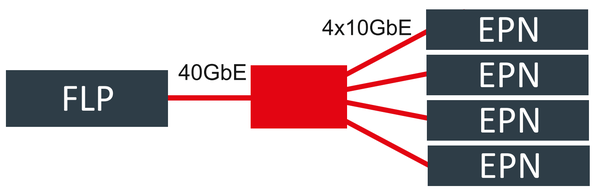
IPoFABRIC (OMNI-PATH)

- Decreasing throughput for messages larger than 25MB is observed.
- The overhead of IPoFabric is even larger than for IB (only 37.5 Gb/s out of available 100 Gb/s).

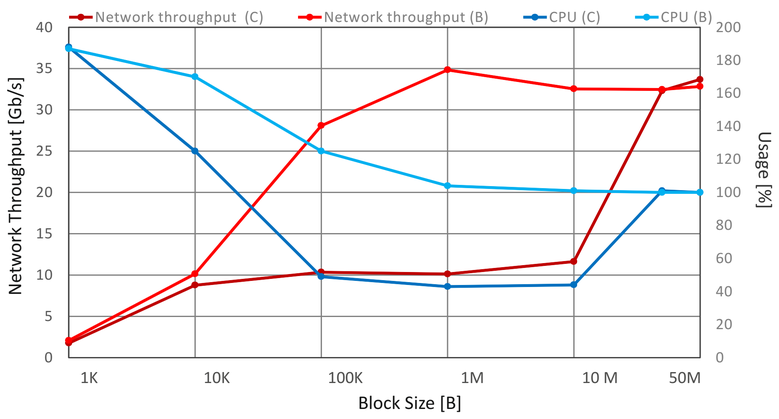


Since the benchmarks require TCP/IP stack, Ethernet, which natively encapsulates TCP/IP, gave the best results. Among the MQ libraries ZeroMQ turned out to be more performant reaching significantly higher throughput than nanomsg.

SATURATING SENDER - FLP

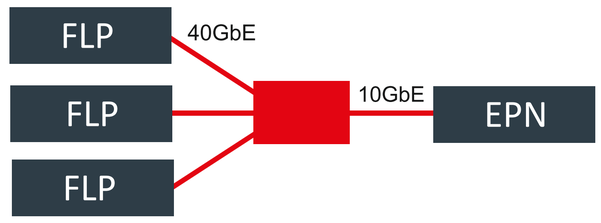


The plot below presents network throughput in function of block size between 1 FLP device and 4 servers x 60 EPN devices measured on FLP side for two socket configurations: *bind* (B), *connect* (C). Test were performed for O^2 benchmark (ZeroMQ) and

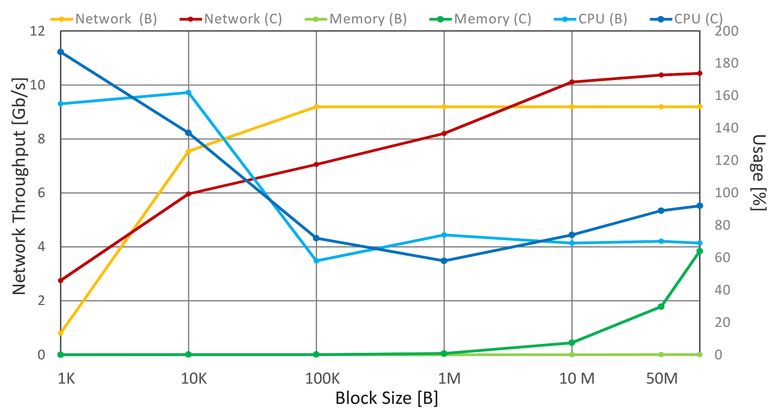


Differences in network throughput between two socket configurations are caused by blocking send in *connect* (C) configuration.

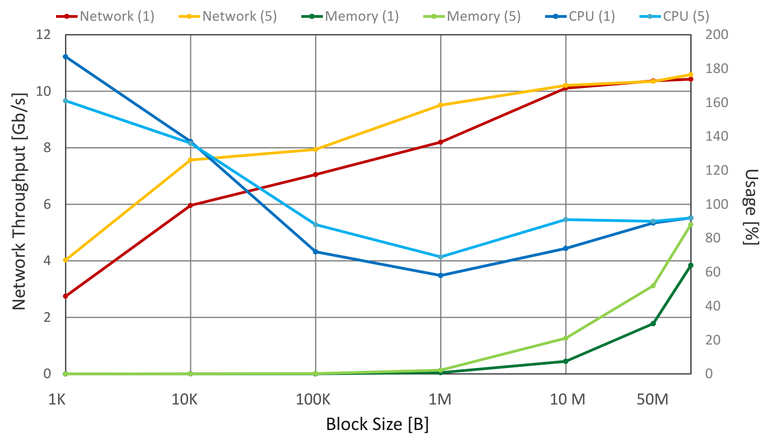
SATURATING RECEIVER- EPN



The plot below presents network throughput in function of block size between 3 servers x 80 FLP devices and 1 EPN device measured on the EPN side for two socket configurations: *bind* (B), *connect* (C). The tests were performed for O^2 benchmark (ZeroMQ) and 10/40GbE. CPU and network throughput results are quite similar for both socket configurations. The difference in memory usage comes from the fact that ZeroMQ creates buffers on the EPN side in *connect* configuration only.



Moreover, the plot below presents measurements for different buffer sizes: (1) maximum one block, (5) up to 5 blocks can be buffered. Presented plot concerns *bind* socket configuration only.



CONCLUSION

- ALFA framework servers its purpose and can be successfully adopted for O^2 needs.
- Ethernet with its long-serving TCP/IP stack reached its maximum speed.
- IPoIB and IPoFabric use less than half of the available bandwidth.
- Two cores (one physical) are enough to receive or transmit data blocks 10KB-100MB in size from single FLP to multiple EPNs or from multiple FLPs to single EPN.
- ZeroMQ and nanomsg:
  - Software buffering in ZeroMQ happens only on *connect* side.
  - *Connect* sockets of ZeroMQ-based benchmarks work in blocking mode unless asynchronous version called explicitly.
  - nanomsg performance is not as good as other libraries but its modular architecture allows to plug custom transports.

FUTURE WORK

Further measurements are foreseen to be performed during consolidation phase with nearly final set-up.

[1] *First Level Processors*
[2] *Event Processing Node*
[3] *Gigabit Ethernet*
[4] *InfiniBand*
[5] *OmniPath*
[6] *Fast Data Transfer*