Using ALFA for High Throughput, Distributed Data Transmission in ALICE O² System

ALICE O² Computing System

The ALICE O² (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 O² 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 memory to create a sub-timeframe. Then, the data will be distributed over EPNs for aggregation and additional compression. The O² facility will consist of 256 FLPs and 4 EPNs. Each FLP will be logically connected to each EPN through a high throughput network. The O² farm will receive data from detectors at 27.2 Tb/s, which after processing will be reduced to 720 Gb/s.

Test Platforms

<table>
<thead>
<tr>
<th>Network</th>
<th>Network adapter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 Gb/s</td>
<td>Chelsea TS80</td>
<td>Intel E5-2690</td>
</tr>
<tr>
<td>IB FDR</td>
<td>Mellanox MT27500</td>
<td>Intel E5-2690</td>
</tr>
<tr>
<td>OPA</td>
<td>-</td>
<td>Intel E5-2680v4</td>
</tr>
</tbody>
</table>

ALFA

The ALFA (ALICE-FAIR) is a common layer of the jointly developed framework by ALICE O² and FAIR teams. It is based on message-like approach and standardizes software related tasks such as data transport, state machines, configuration and monitoring more. ALFA supports two data transport libraries: ZeroMQ and nanomsg.

Benchmarks

- ZeroMQ: message-based library supporting a large number of socket patterns that help to create complex, distributed systems.
- nanomsg: fork of ZeroMQ with the ability to plug custom transports, improved thread 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 builds on the top of ALFA (FairMQ library).
- FDT: 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 four libraries ZeroMQ turned out to be more performant reaching significantly higher throughput than nanomsg.

Saturating Receiver - EPN

The plot below presents network throughput in function of block size between 1 FLP device and 4 servers x 60 EPN devices measured on EPN side for two socket configurations: bind (B), connect (C). The tests were performed for O² 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 O² 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 full set-up.