Omni-Path tests

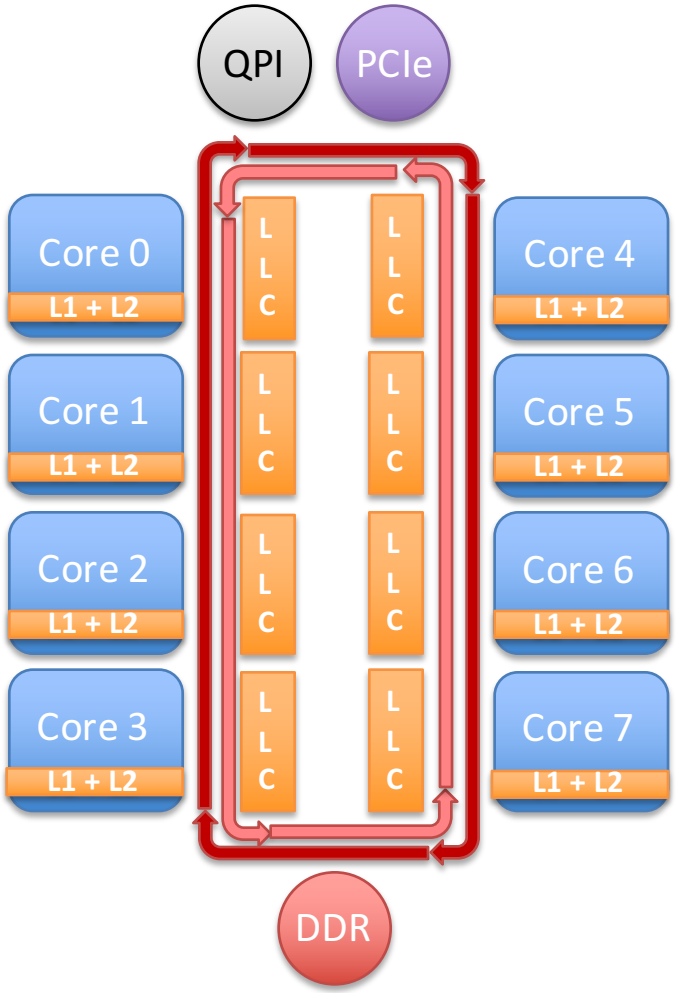
MPI and IPoFabric

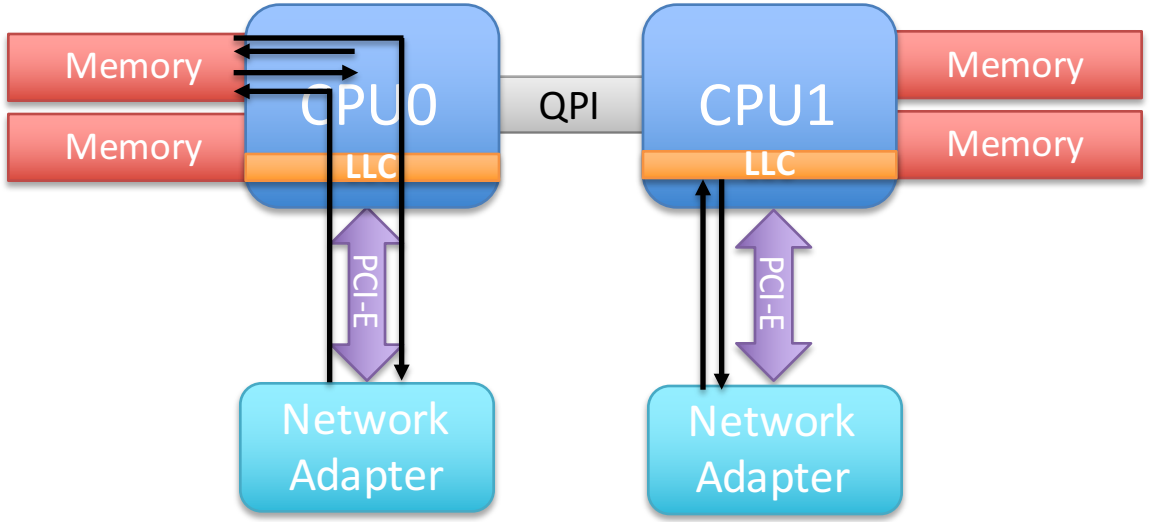
Sylvain Chapeland, Adam Wegrzynek · CERN

Hardware

- 16x E5-2680v4 "Broadwell EP"
  - 12 cores
  - DDR4
- Omni-Path 100Gb interconnect
  - CPU and fabric integration

Intel Data Direct I/O on E5 architecture





See: Direct Cache Access for High Bandwidth Network I/O

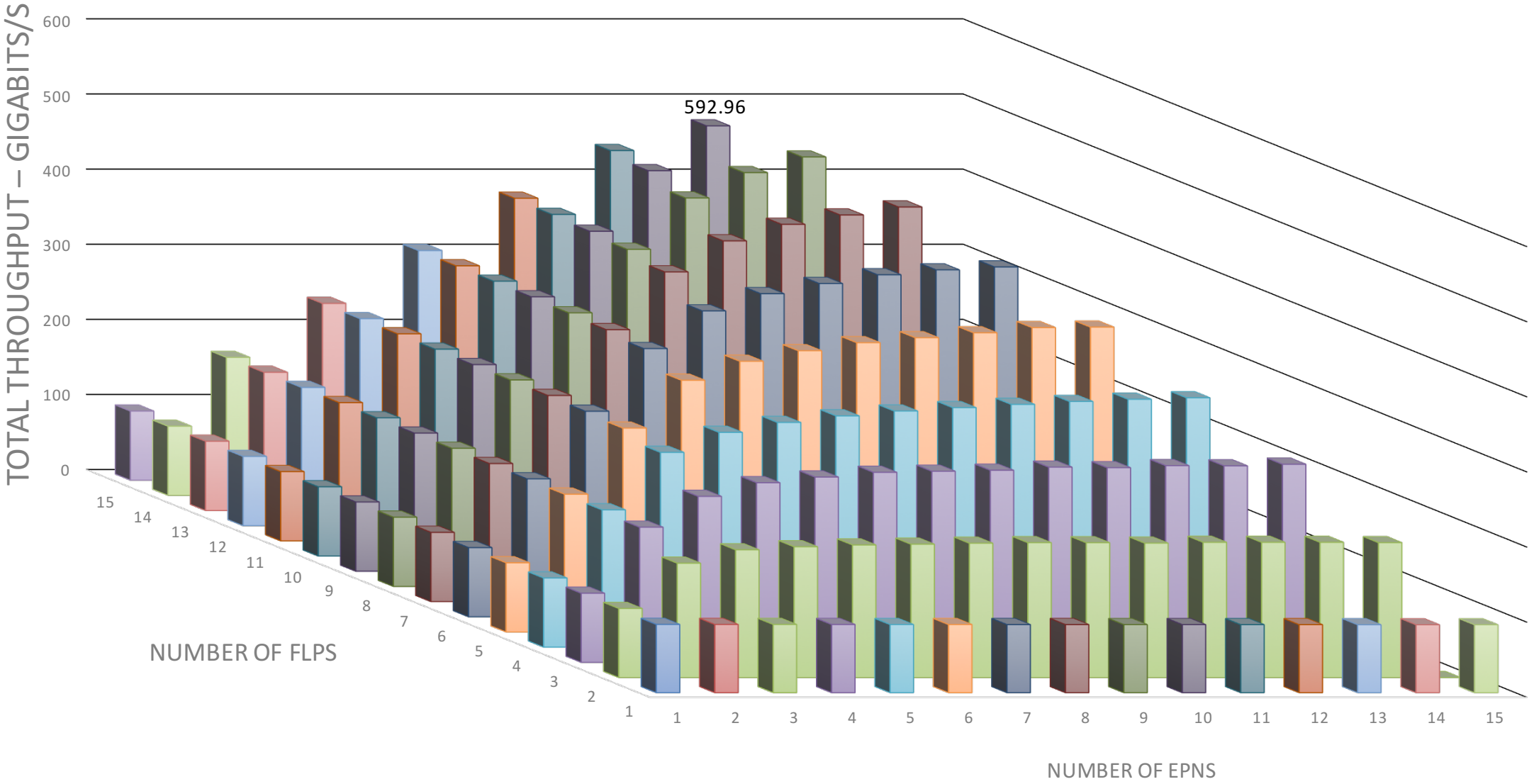
MPI: Simple “event building”

- Software:
  - x FLPs send event fragments to y EPNs
    - EPNs used round-robin
    - First fragment of event N+1 sent to EPN i+1 only after all fragments of event N received by EPN i (global synchronization barrier)
    - Event fragment size configurable, normal distribution +/-10% around nominal size
  - Implementation using MPI for transport
    - MPI_Send, MPI_Irecv, MPI_Waitall, MPI_Barrier
    - ~200 lines of c++ code
    - Compiled with Intel mpiicpc
- Test protocol: measure per-link and aggregate bandwidth
  - Different number of FLPs/EPNs (X+Y<16)
  - Different event fragment size (100 bytes … 1 GBytes)
  - Starting mpirun with 1 process per host, e.g. : *mpirun -ppn 1 …*

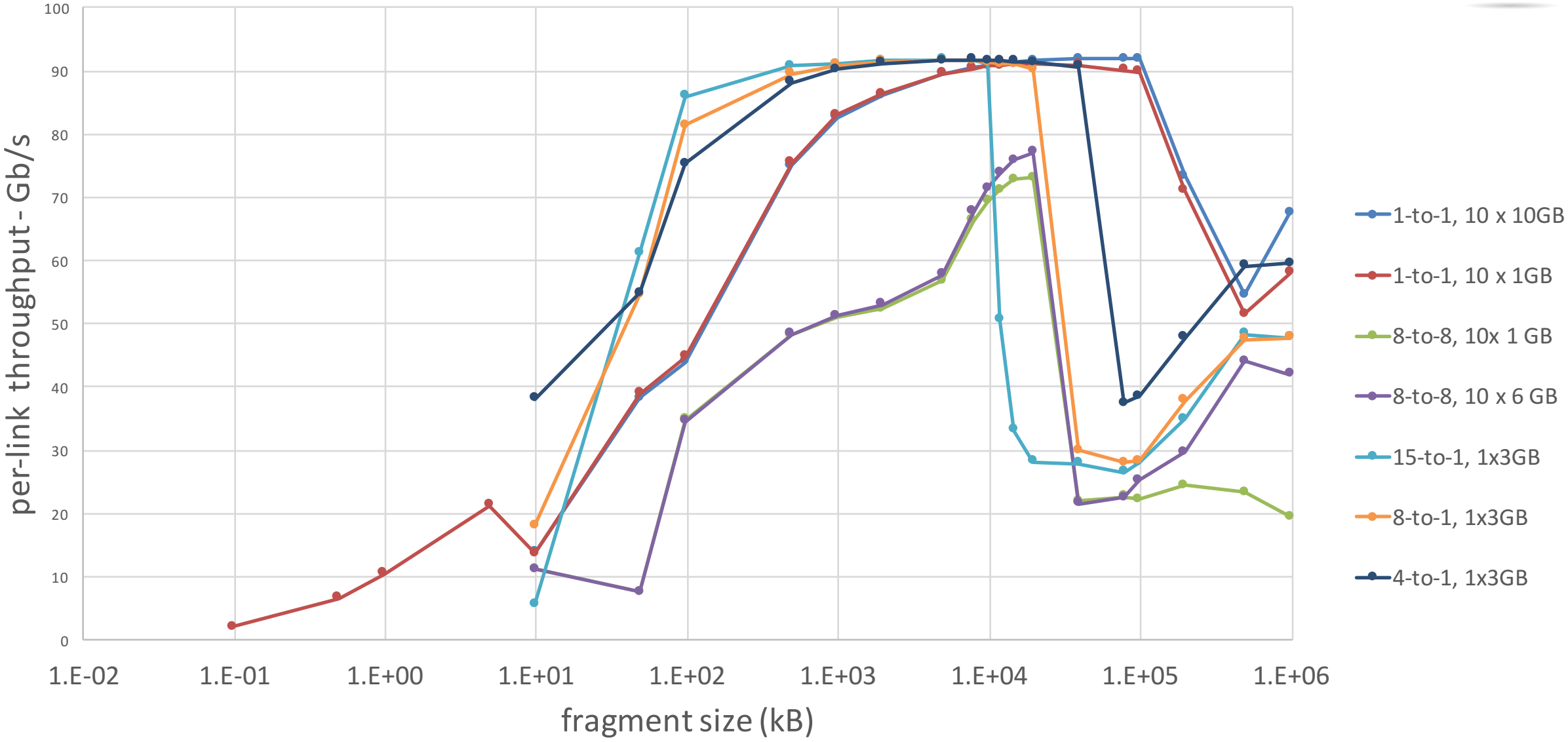
MPI: Observations

- Can easily get close to wire speed: ~92 Gb/s routinely measured
- Very smooth and almost linear scaling operation in all tested configs
  - Link sharing handled well, no traffic collision conflicts even in 15-to-1 operation
- Good aggregate transfer when using all hosts
  - Up to 592Gb/s for a 9 FLP to 7 EPN config, i.e. 85Gb/s per EPN
  - Despite simple transport code (global sync dead time, no overlapping transfers)
- Optimal fragment size 1-10 MB
  - Usual latency effects limitations for smaller blocks
  - Probably some cache-related effects for larger blocks (intel ‘shortcut’ from cache to network does not work any more)
- Easy transport coding with MPI

MPI: Event building throughput



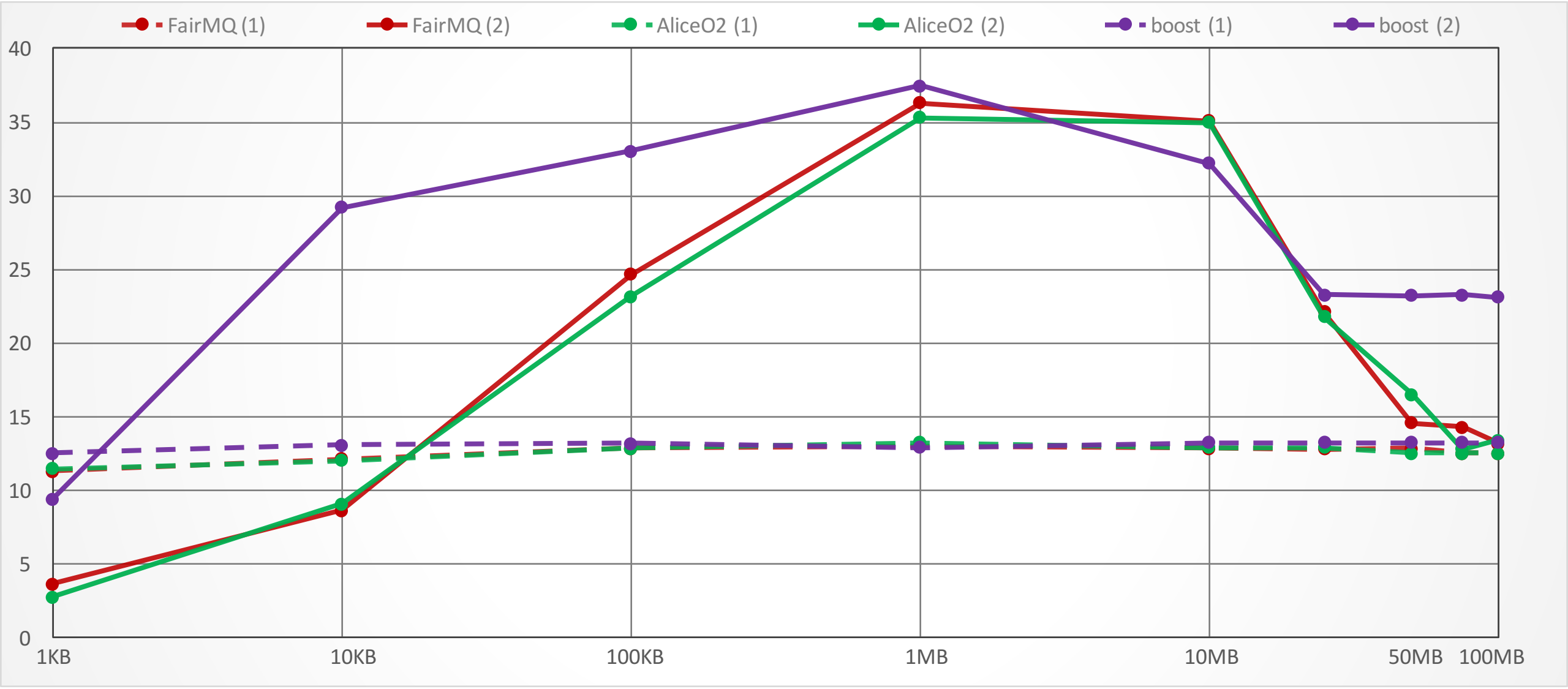
MPI: transfer rate



IPoverFabric: Benchmarks

- FairMQ using ZeroMQ (FMQ)
  - bsampler and sink from FairRoot
- Memory pre-allocation (Zero-copy)
  - AliceO^2 FLP and EPN modified devices using FairMQ (AliceO2)
  - Boost::Asio sender and receiver (boost)

IPoverFabric



IPoverFabric: Conclusion

- For block size in the 1-10 MB: Asio and ZMQ able to transfer data at 32-37 Gb/s with 1 core
- For block size in the 50-100 MB: Asio able to transfer data at 25 Gb/s and ZMQ at 13-17 Gb/s with 1 core
- Large overhead due to IPoFabric – way better performance with MPI