Infiniband in ALICE HLT

Timo Breitner for the ALICE HLT project

Frankfurt Institute of Advanced Studies

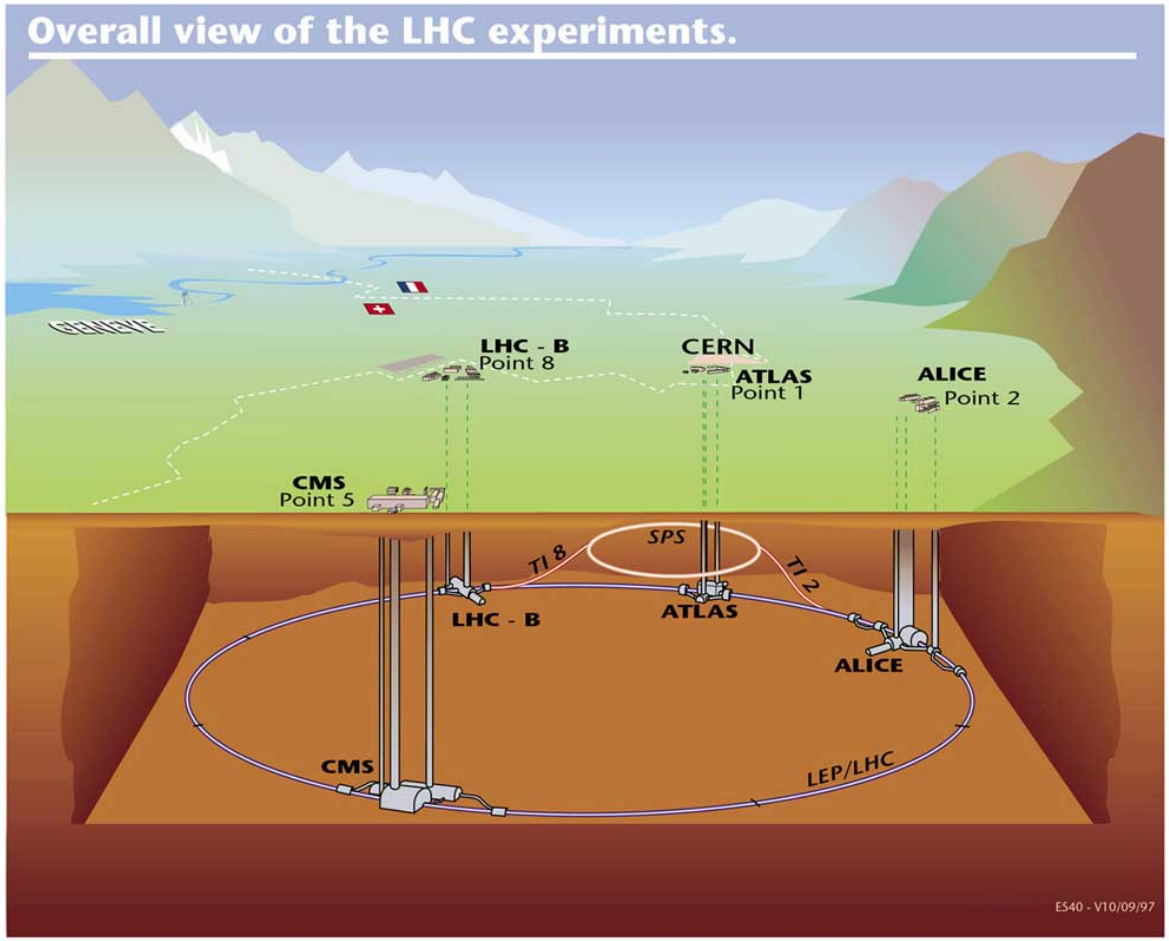
University of Frankfurt, Computer Science Department

tbreitne@cern.ch

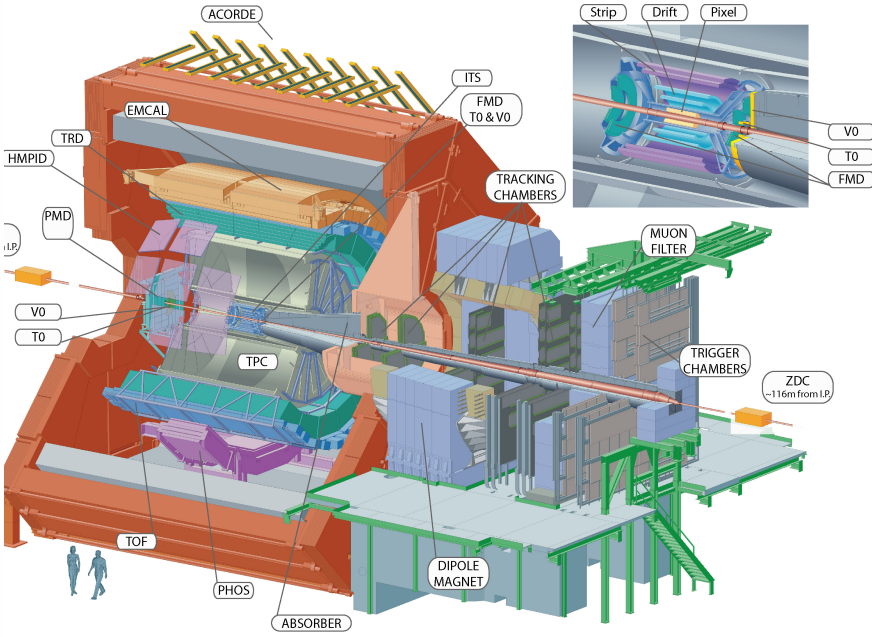
Overview

- Infiniband@ALICE HLT installed summer 2010, in production since late 2010
- HLT software stack based on IP
  - Only IPoIB, (currently) no RDMA
- Topics of this talk
  - HLT layout, bandwidth requirements
  - Motivation for using IB
  - Performance numbers
  - Experience

A Large Ion Collider Experiment

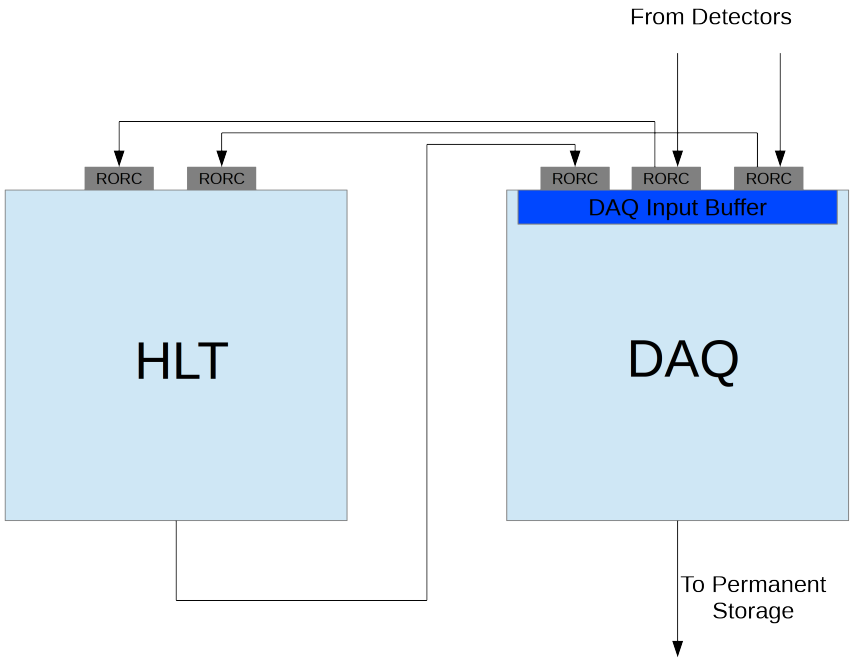


- LHC experiment focused on heavy ion physics



- Avg. event size of central PbPb collision: ~ 70MB
- TPC largest contributor (~ 85%)

ALICE High Level Trigger

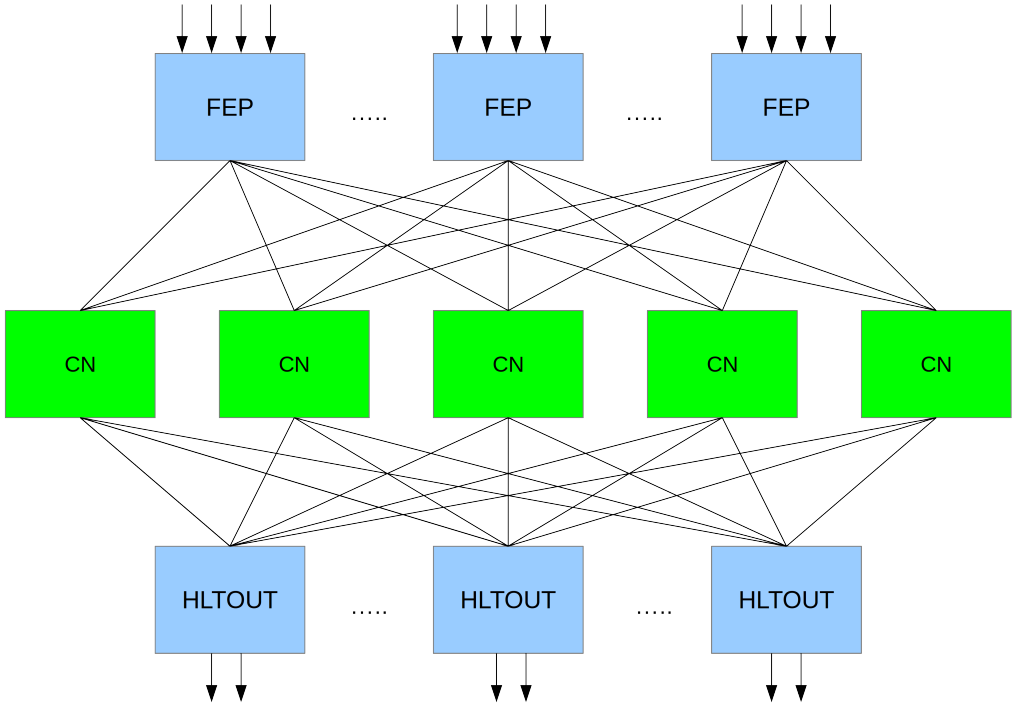


- Purpose: partial/full data reconstruction
  - Trigger/filter
  - Data compression
  - Online QA

- Input: Raw detector data, via optical fibers (DDL), direct copy from DAQ
- Output: Trigger decision, compressed data forwarded to DAQ via optical fibers
- Current std. Operation mode: TPC data reduction
  - Local Pattern Recognition (cluster finding)
  - Huffman encoding
  - Compression factor 4-5, more possible

TPC data flow schematics

Input layer: 54 x Front End Processors
- 2 x AMD Opteron 2378 (Shanghai, 4 cores)
- 4 DDL@160MB/s



Processing layer: 45 x Compute nodes
- 2 x AMD Opteron 6172 (MagnyCours, 12 cores)
- Nvidia GPU GTX480/580
- Track reconstruction, vertexing, data compression

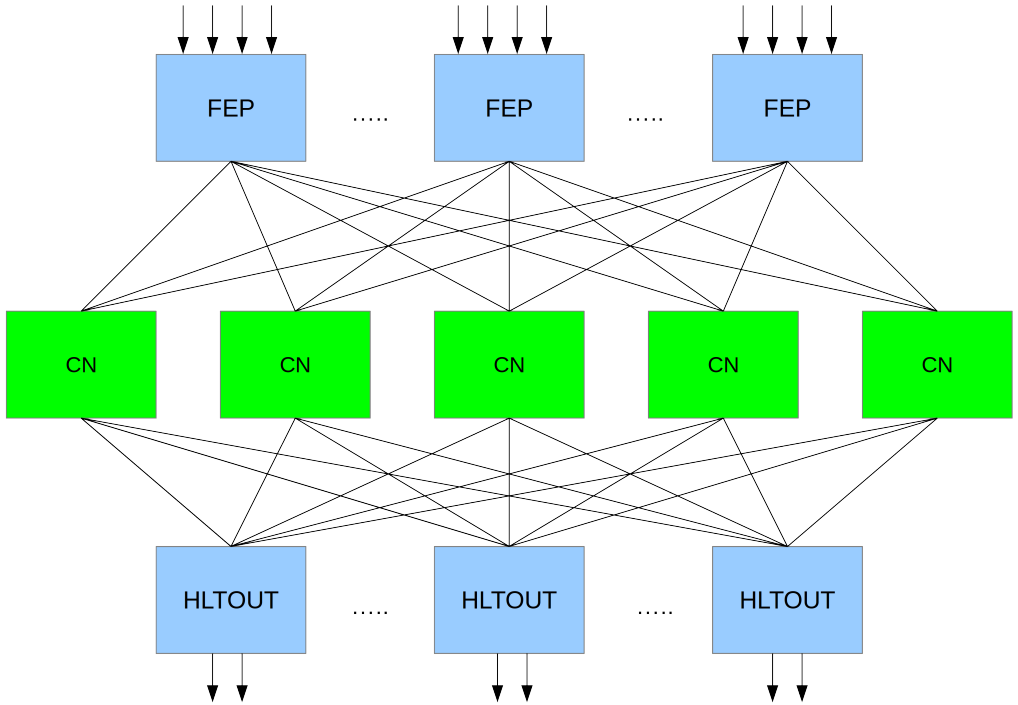
Output layer: 14 x HLTOut nodes
- 2 x AMD Opteron 2378 (Shanghai, 4 cores)
- 2 DDL@160MB/s

Total nodes in system:
100 FEP nodes
80 CN nodes

Bandwidth requirement

- HLT data transport framework based on SHM (intra-node) and TCP/IP (inter-node)
- Initial deployment: 1Gbit Ethernet
  - Enough for first years (pp)
  - Still in use for system management
- Towards PbPb runs 2010:
  - Discussion on suitable network infrastructure
  - Particle multiplicity (event size) not clear
  - Event rate not clear
- Worst case scenario assumed!

Bandwidth requirement



FEP in: 4 DDL * 160 MB/s → 640 MB/s
FEP out: 640 MB/s

Network: 54 * 640 MB/s → 34.56 GB/s

CN in: 34.56 GB/s / 45 nodes → 768 MB/s
CN out: 768 MB/s / 4 → 192 MB/s

Network: 45 * 192 MB/s → 8.64 GB/s

HLTOUT in: 8.64 GB/s / 14 → 617 MB/s

Per-node Bandwidth: up to 800 MB/s
Network bandwidth: up to 45 GB/s

→ 10 Gbit Ethernet OR Infiniband QDR

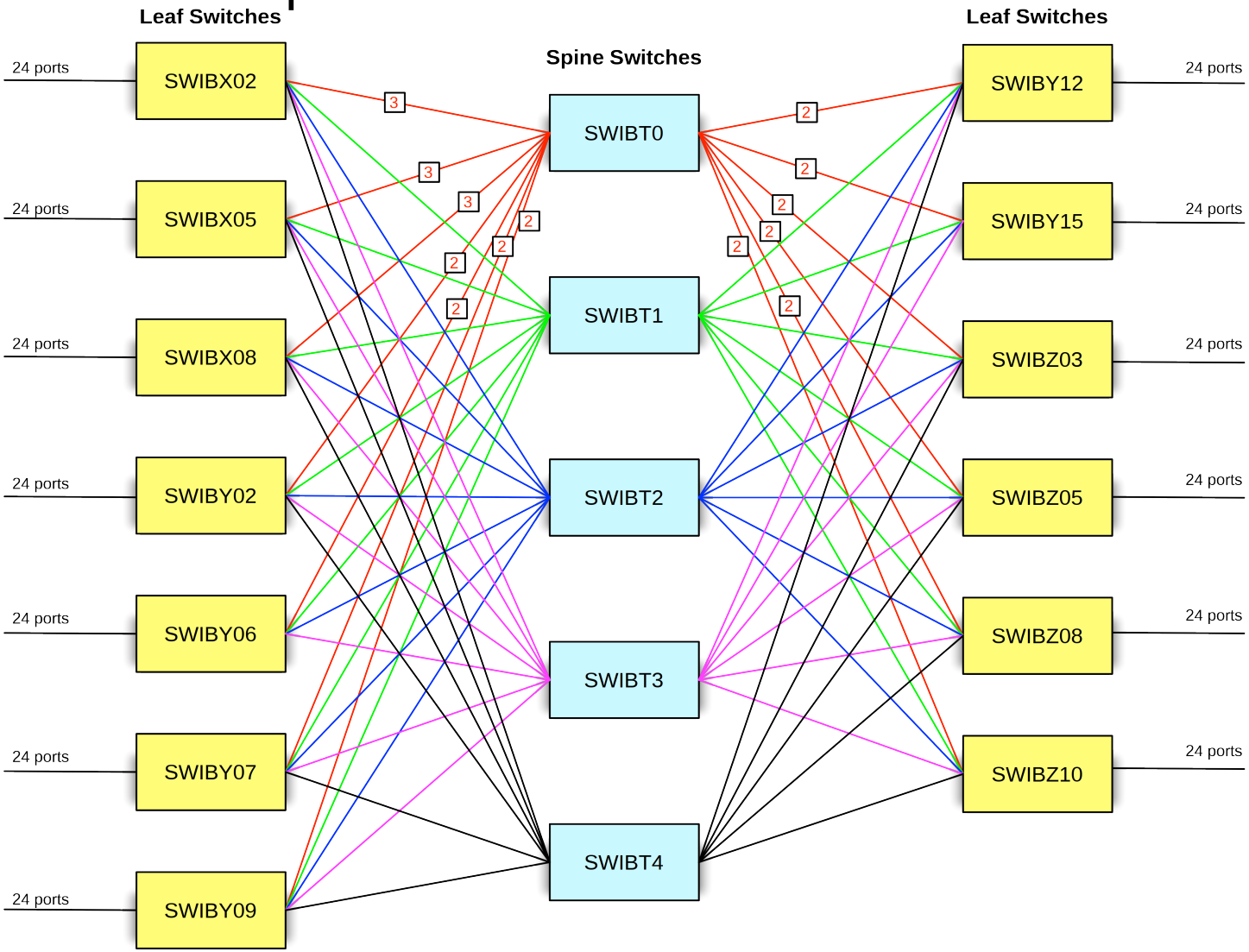
Note: latency is not an issue

Why Infiniband

- Had already some experience
  - Small test setup: backbone net, proof-of-concept
- Price/performance
  - Nominally 4 x faster than 10GBit ethernet (only with “native” IB mode possible)
  - IPoIB: ~1.5 x faster
  - Cheaper (back then, probably still is)
  - Israeli program to fund acquisition of Israeli Hi-tech products by CERN

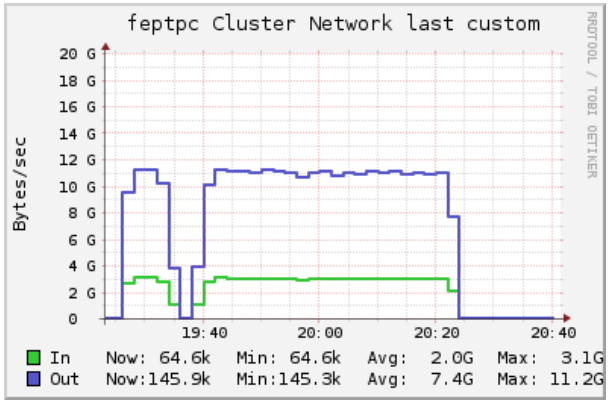
Network Topology

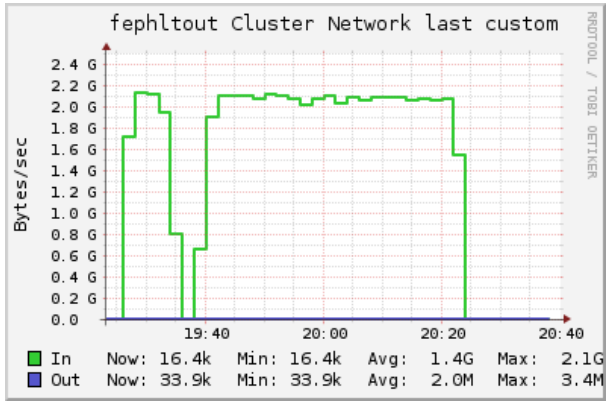
- 5 x 36 port spine switches
- 13 x 36 port leaf switches: 24 node connections, 12 uplinks



- 2/3 uplinks per leaf switch
- Mix of FEP and CN nodes per leaf switch

Performance: PbPb runs 2011





- Typical data rates:
  - Input from TPC: 9 GByte/s
  - Output to DAQ: 2.2 GByte/s
- Maximum data rates:
  - Input from TPC: 14.9 GByte/s
  - Output to DAQ: 3.5 Gbyte/s
- Worst case scenario assumed twice the bandwidth

Test Setup

- 3 types of nodes in use @HLT:
  - Shanghai: AMD Opteron 2378, 2 x 4 cores
  - MagnyCours: AMD Opteron 6172, 2 x 12 cores
  - Nehalem: Intel Xeon E5520, 2 x 4 cores + HT
- All nodes equipped with

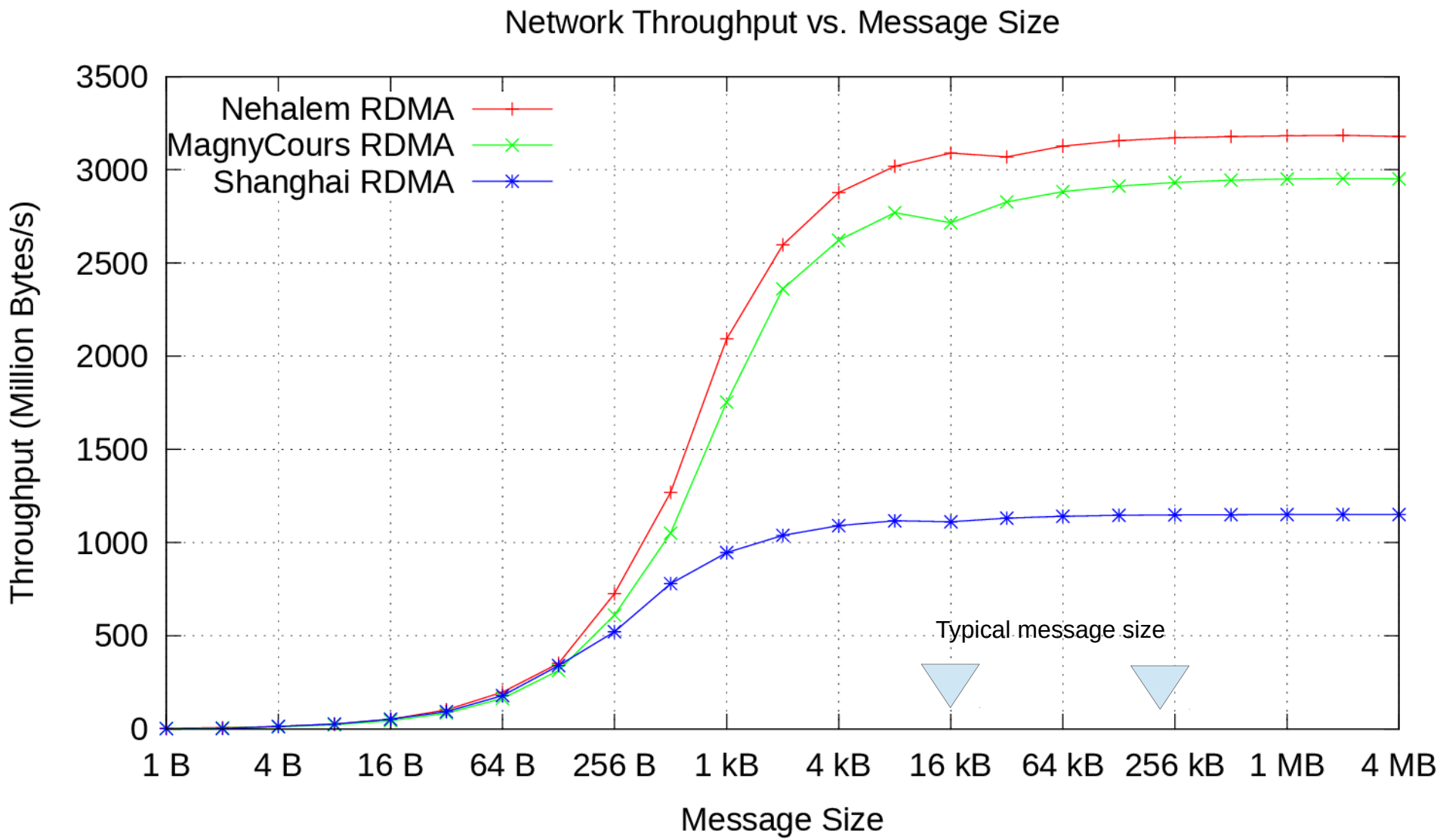
  Mellanox ConnectX PCIe 2.0 QDR

  (Onboard or add-on card)
- osu micro-benchmark suite, MVAPICH2

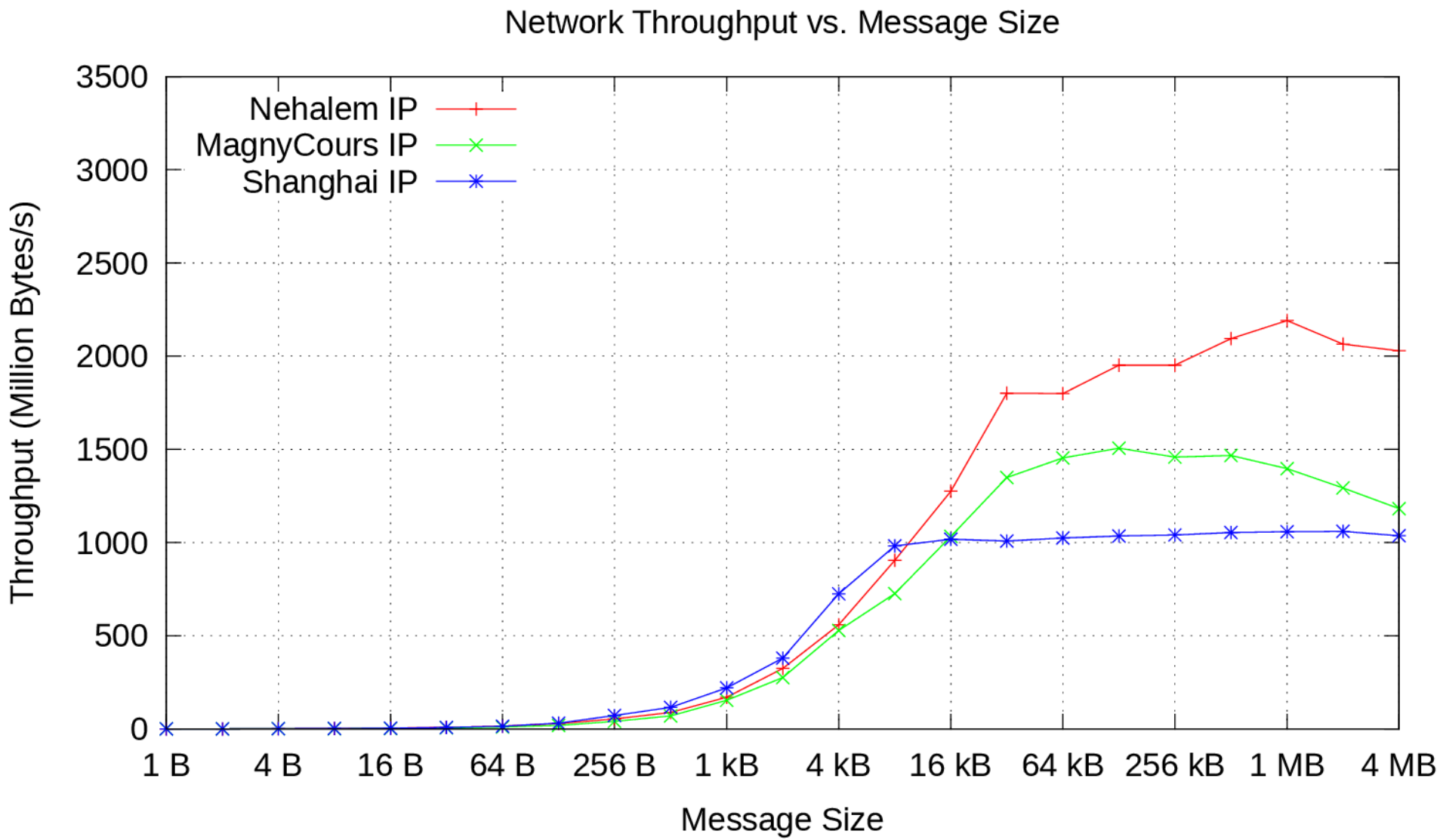
(osu_bw, osu_mbw_mr)
http://mvapich.cse.ohio-state.edu/benchmarks/

Point-to-point performance

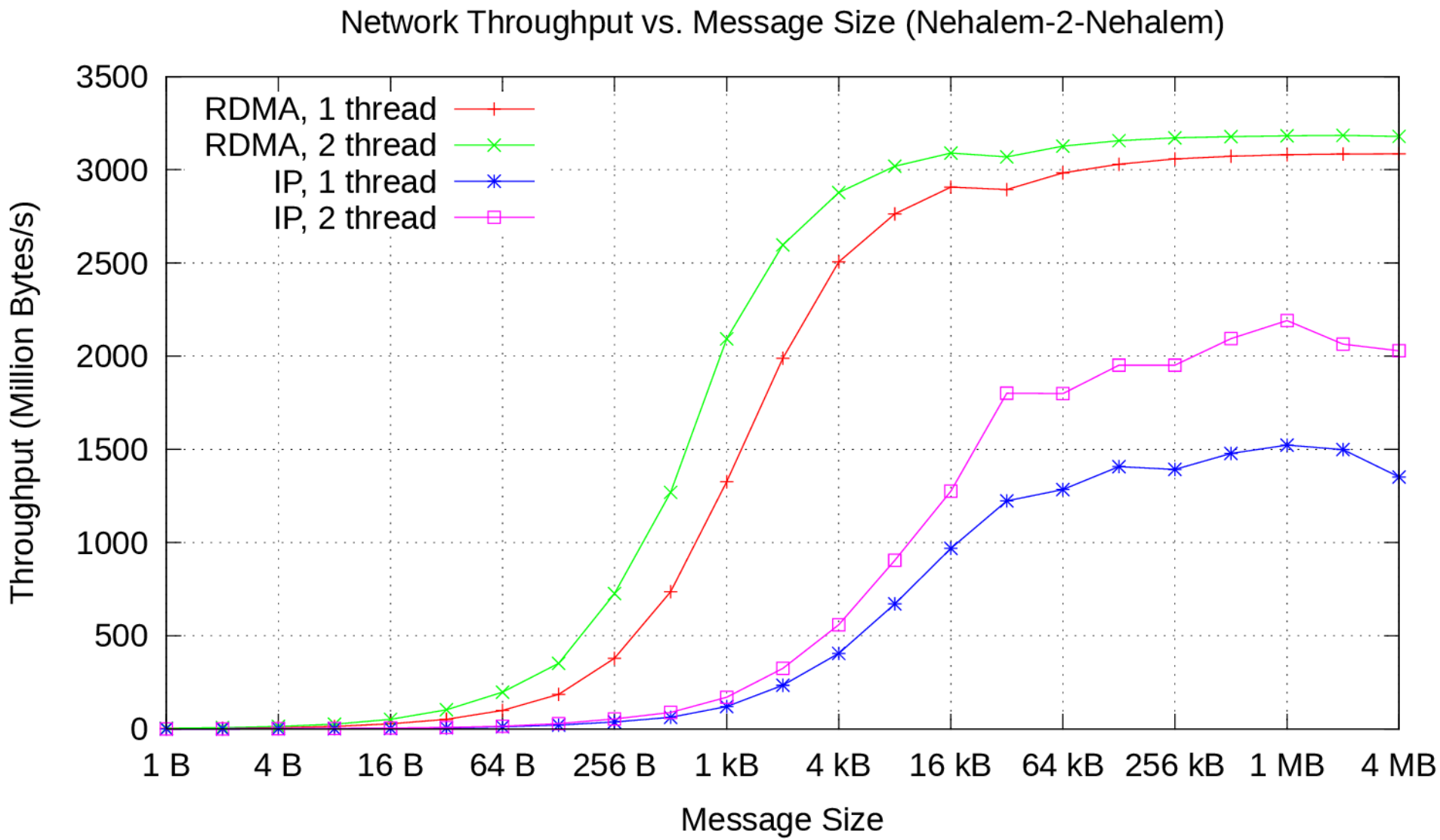
Performance: RDMA



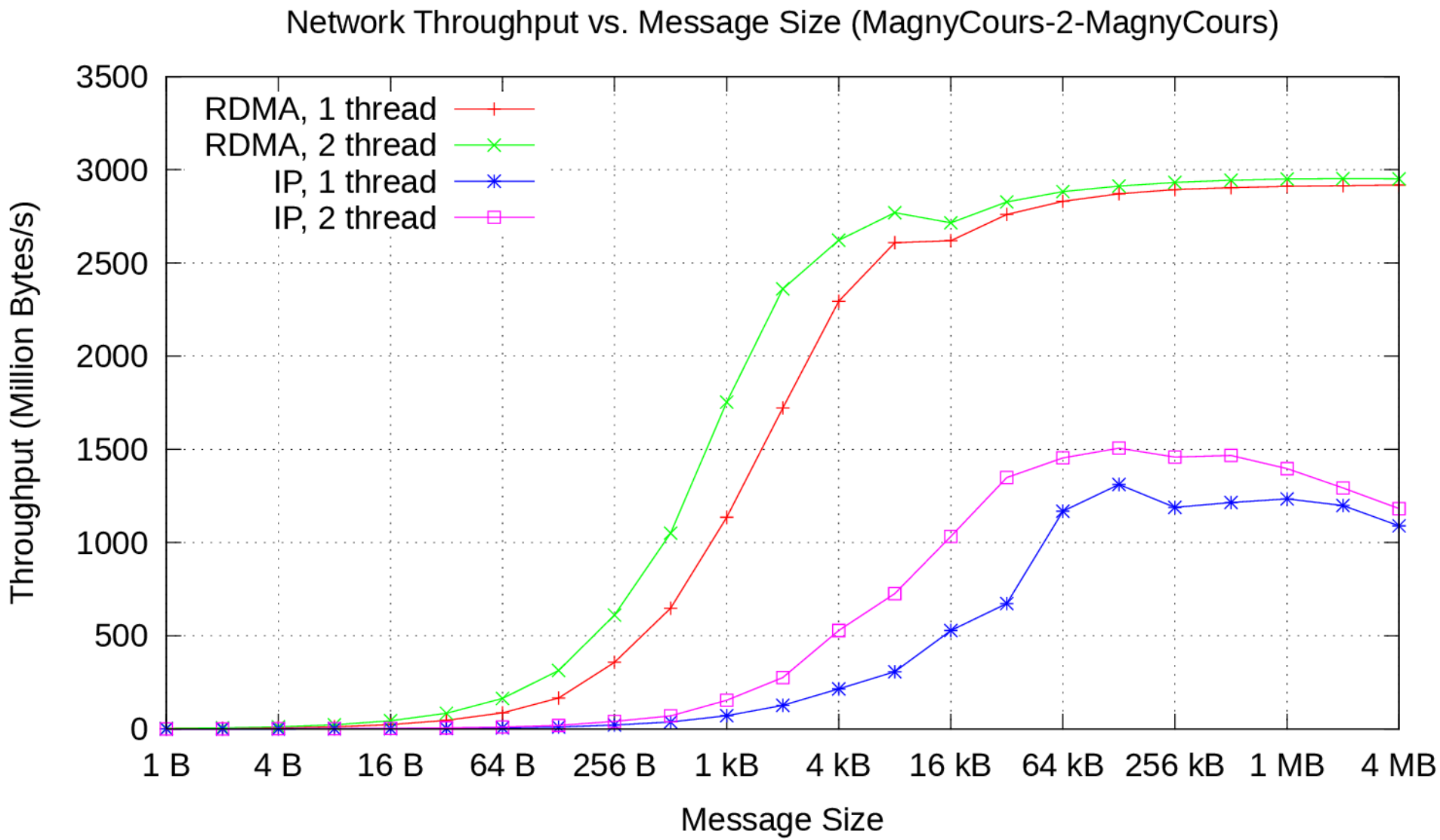
Performance: IPoIB



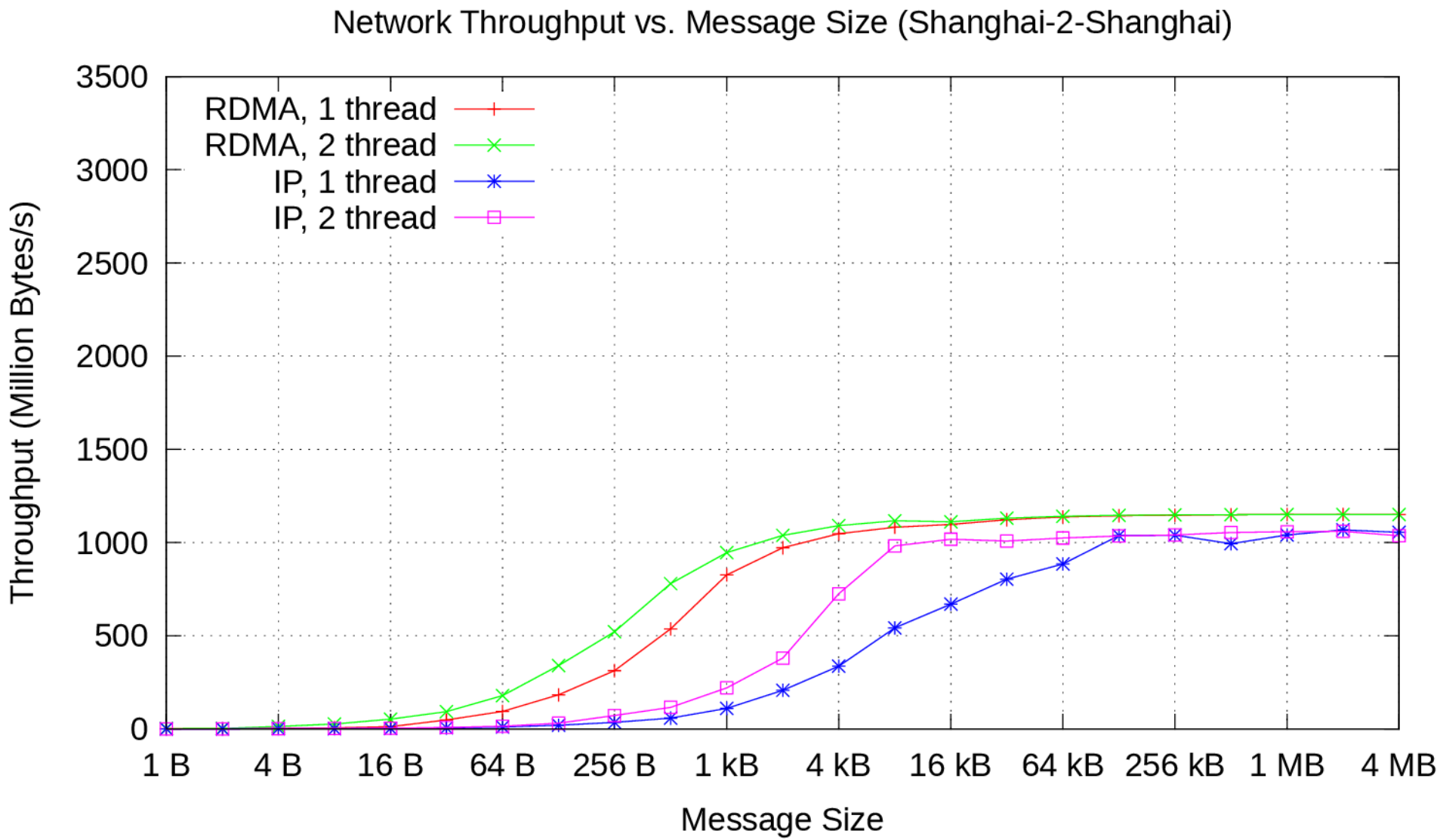
Performance: Nehalem



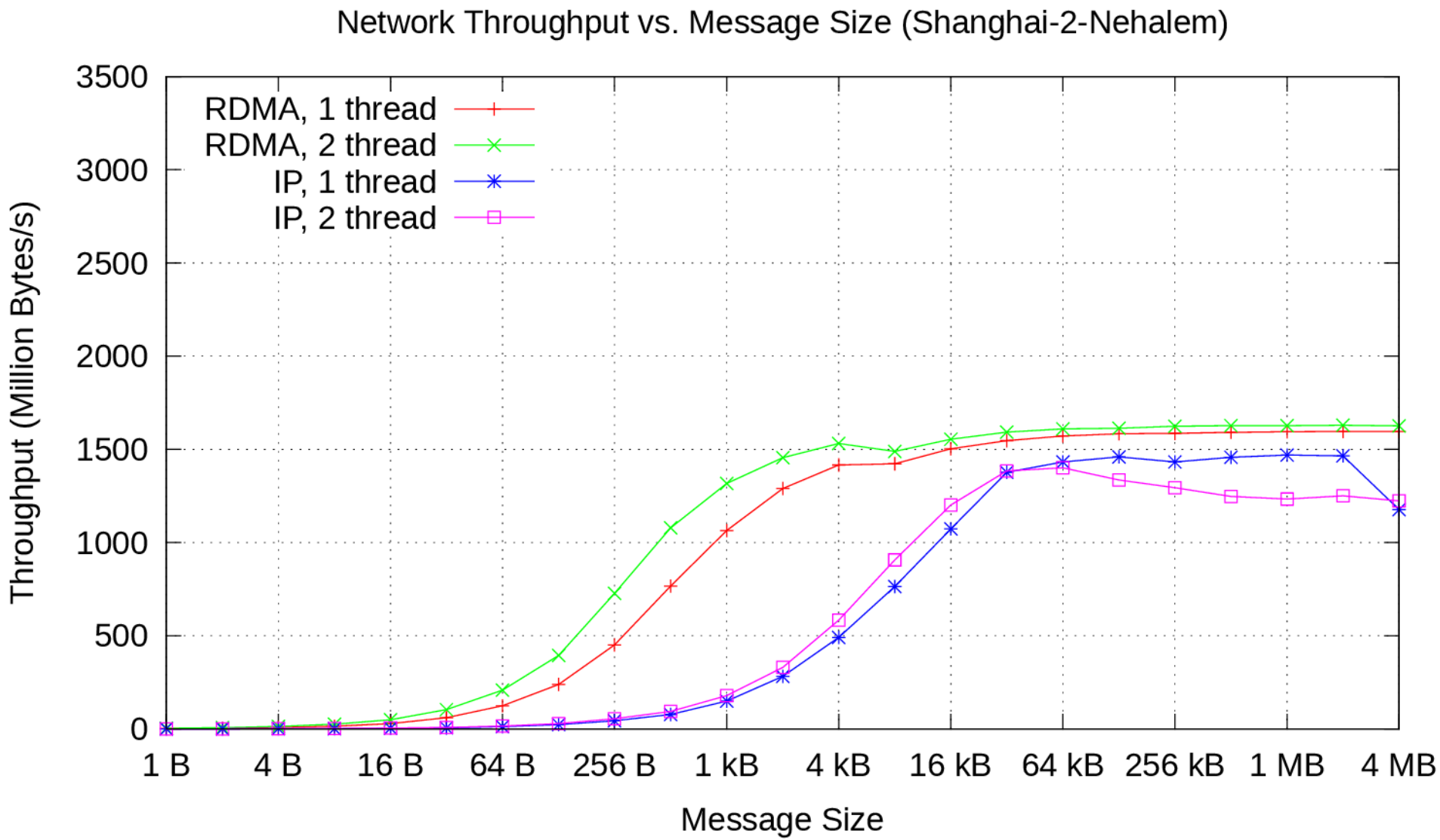
Performance: MagnyCours



Performance: Shanghai

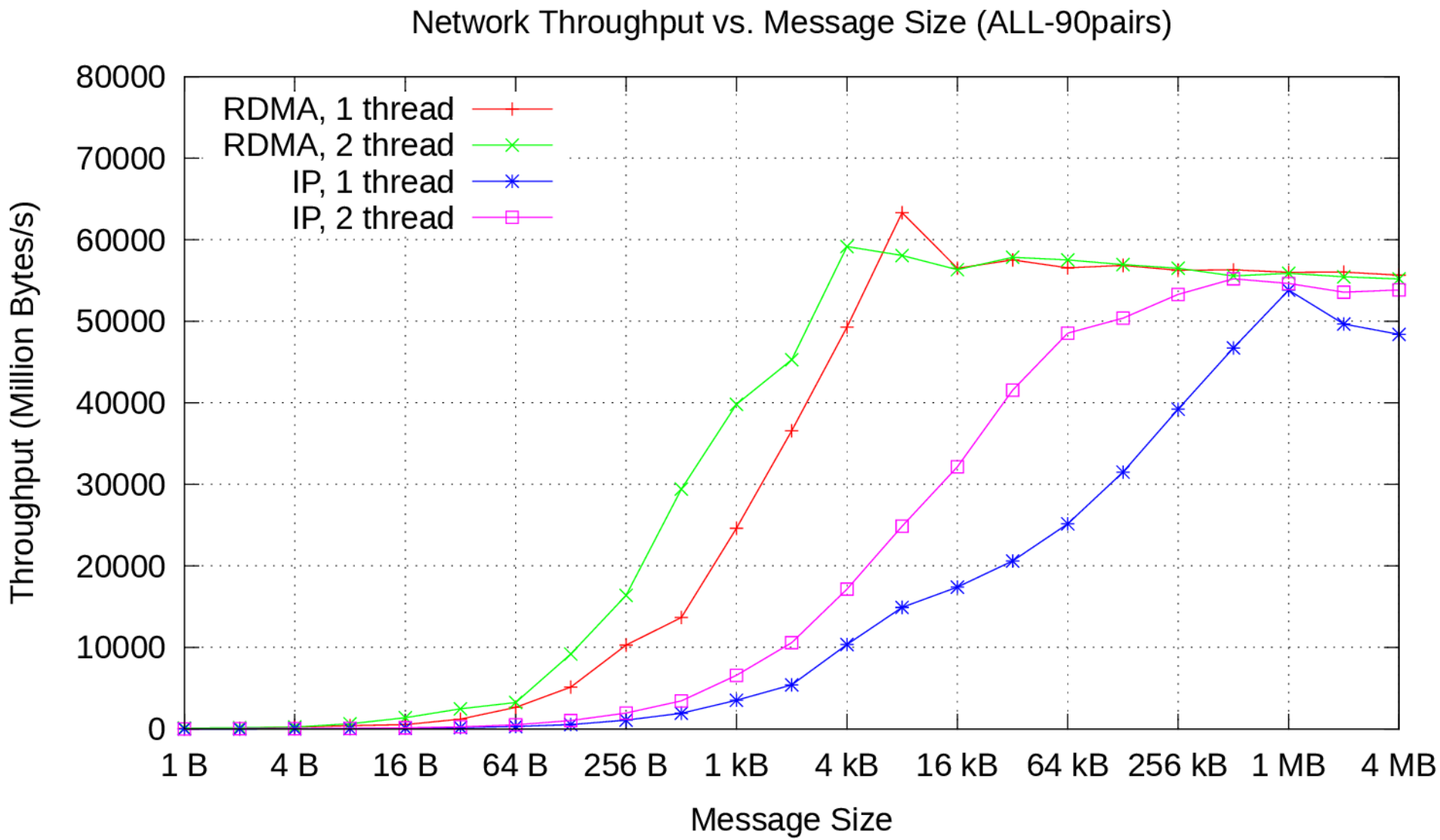


Performance: Shanghai

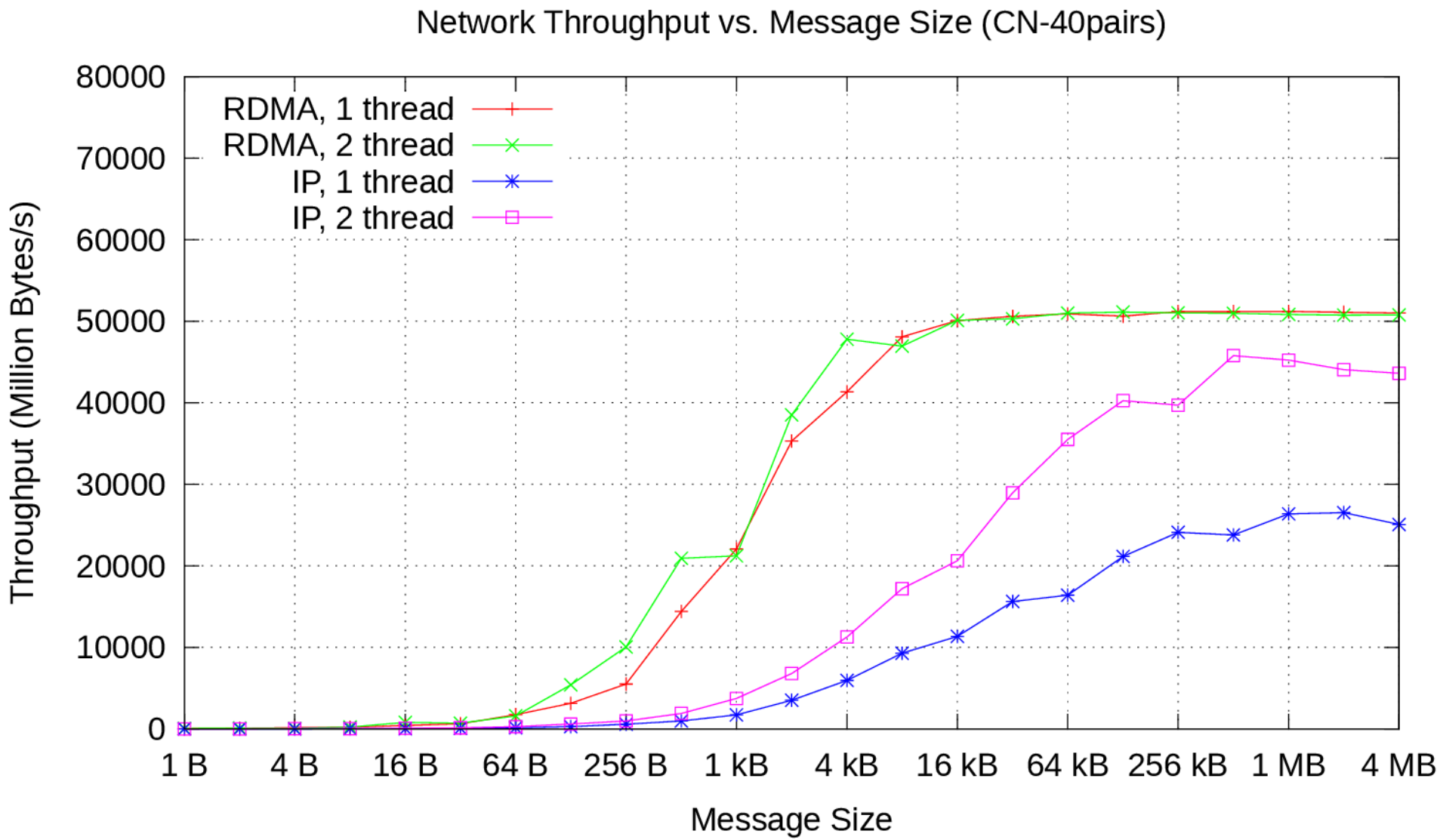


Aggregate bandwidth performance

Performance: Aggregate BW



Performance: Aggregate BW



Experience

- General experience very good
  - Reliable hardware, no failures/losses
  - Massive network bandwidth, nowhere close to saturation in current situation
- Several issues to be resolved during commissioning:
  - Hardware installation (cables)
  - Software/OS setup
  - Firmware

Issues: H/W installation

- Detailed plan of counting room layout and switch location required
  - Cables are rather expensive (esp. longer ones)
  - Copper cables very inflexible (esp. longer ones)
  - Copper cables are quite heavy and can break HCA
- With FDR not so much of a problem
  - Copper cables only up to 3m
  - Fibers for anything beyond (easier to handle, more expensive)

Issues: Software setup

- OFED stack easy to install
- Lots of tunables (OS settings), some are crucial
- OFED packages provide install scripts setting up most of the important things
  - Only RHEL, SLES officially supported
  - Track config files with Chef/Puppet
- Network setup issue during boot:
  - Takes rather long for link state to become “active”
  - Necessary to delay any network access

Issues: Firmware

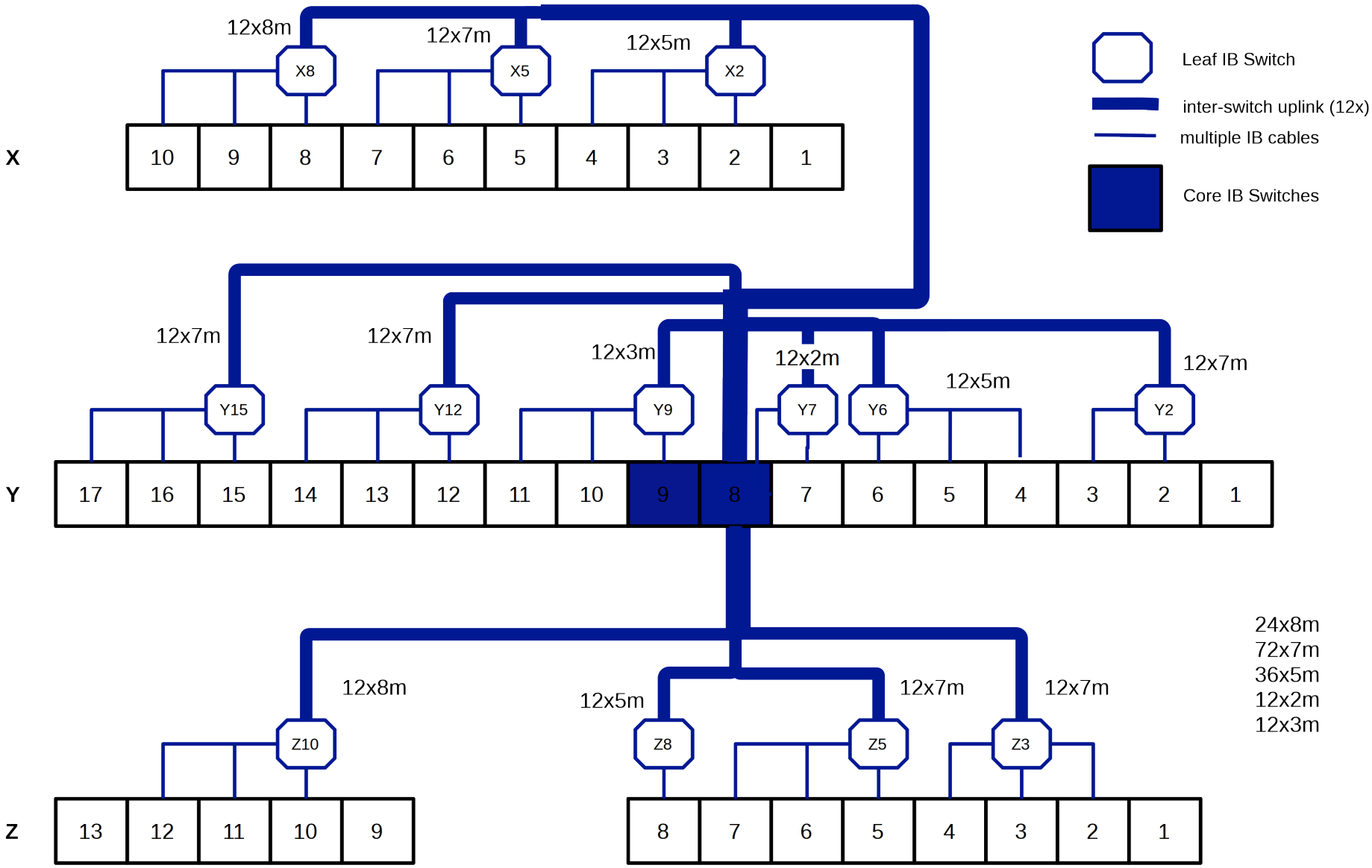
- Issue experienced during HPC cluster setup at GSI
- Lots of issues with FDR network, large error counts in switches
- Firmware problem: link training was not robust enough/failed
- Had to manually set link speed to QDR
  Issue now solved by firmware fix
- Firmware update (switch and HCA) first thing to do
- Mellanox firmware does not fit 3rd party vendor devices

Summary and Outlook

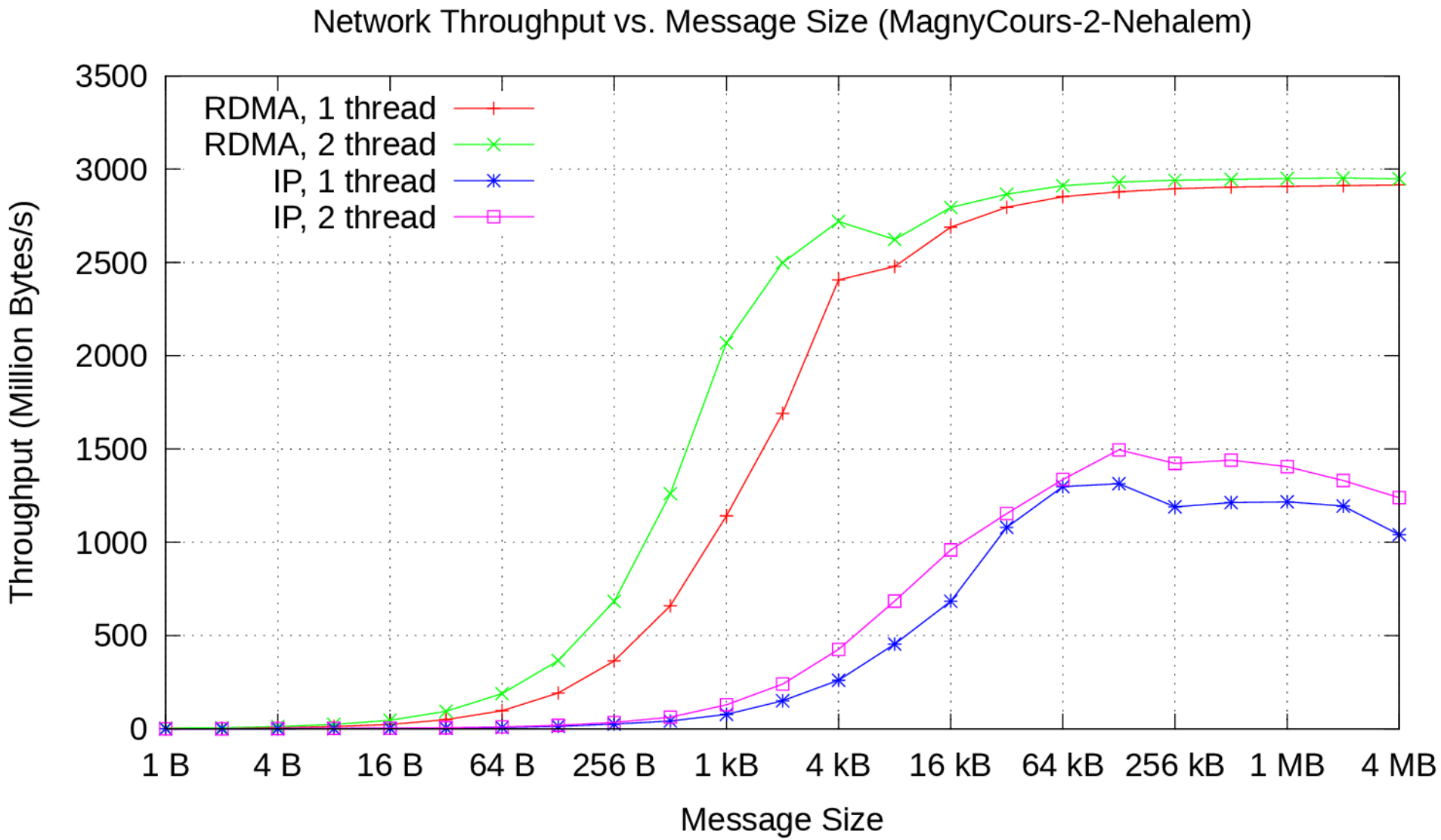
- Reliable and more than enough bandwidth for our current needs
- Some difficulties during setup
- Even more bandwidth possible by using more efficient protocol (RDMA)
- Future ALICE upgrade with combined DAQ/HLT system has much higher requirements (up to 1 TB/s)
- IB also an option there, but has to compete with other technologies in both price and performance

Backup

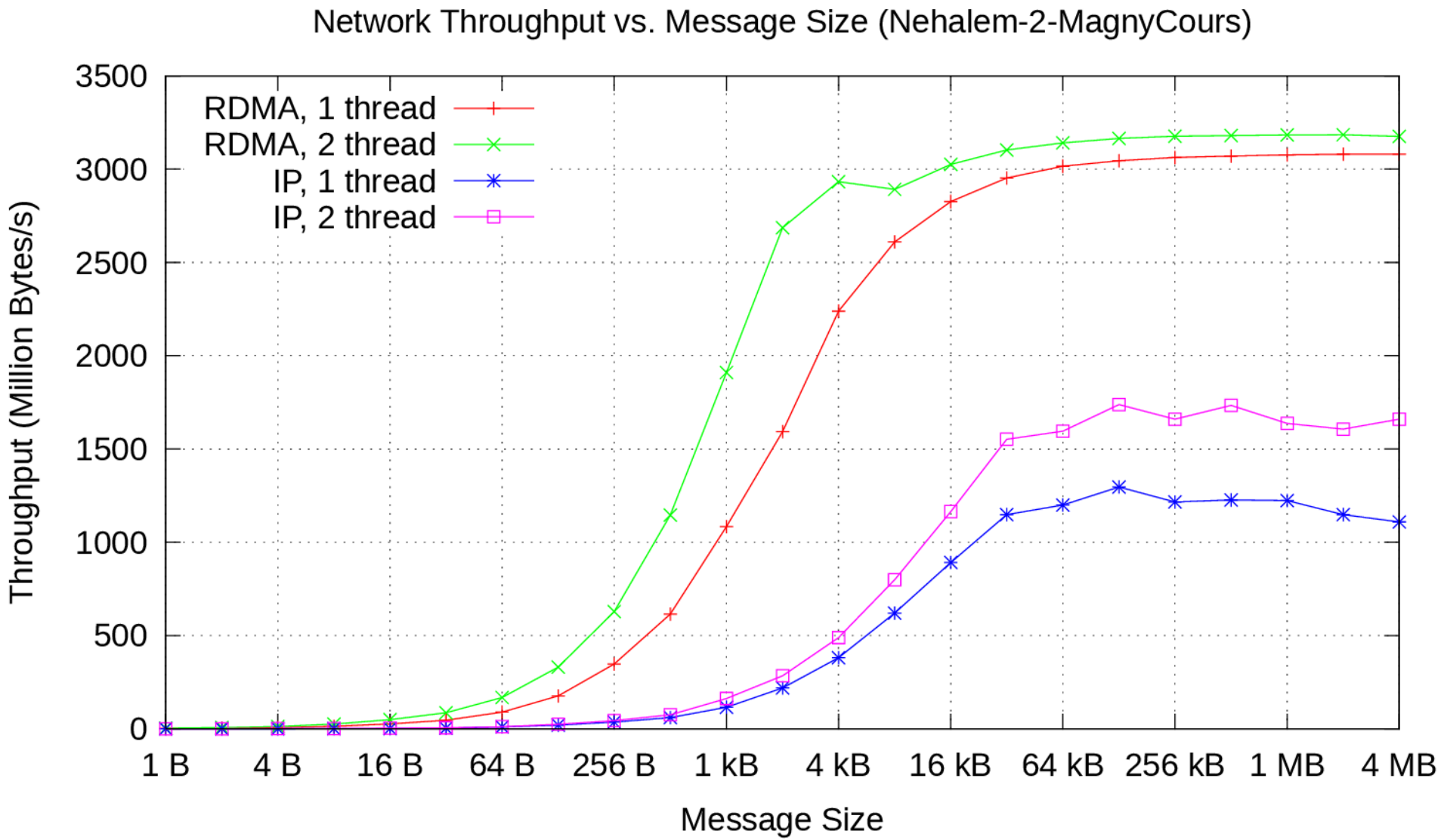
Counting Room Layout



Performance: Aggregate BWs



Performance: Aggregate BWs



Performance: Aggregate BWs