A Power-Aware, Application-Based, Performance Study Of Modern Commodity Cluster Interconnection Networks

Torsten Hoefler, Timo Schneider, and Andrew Lumsdaine

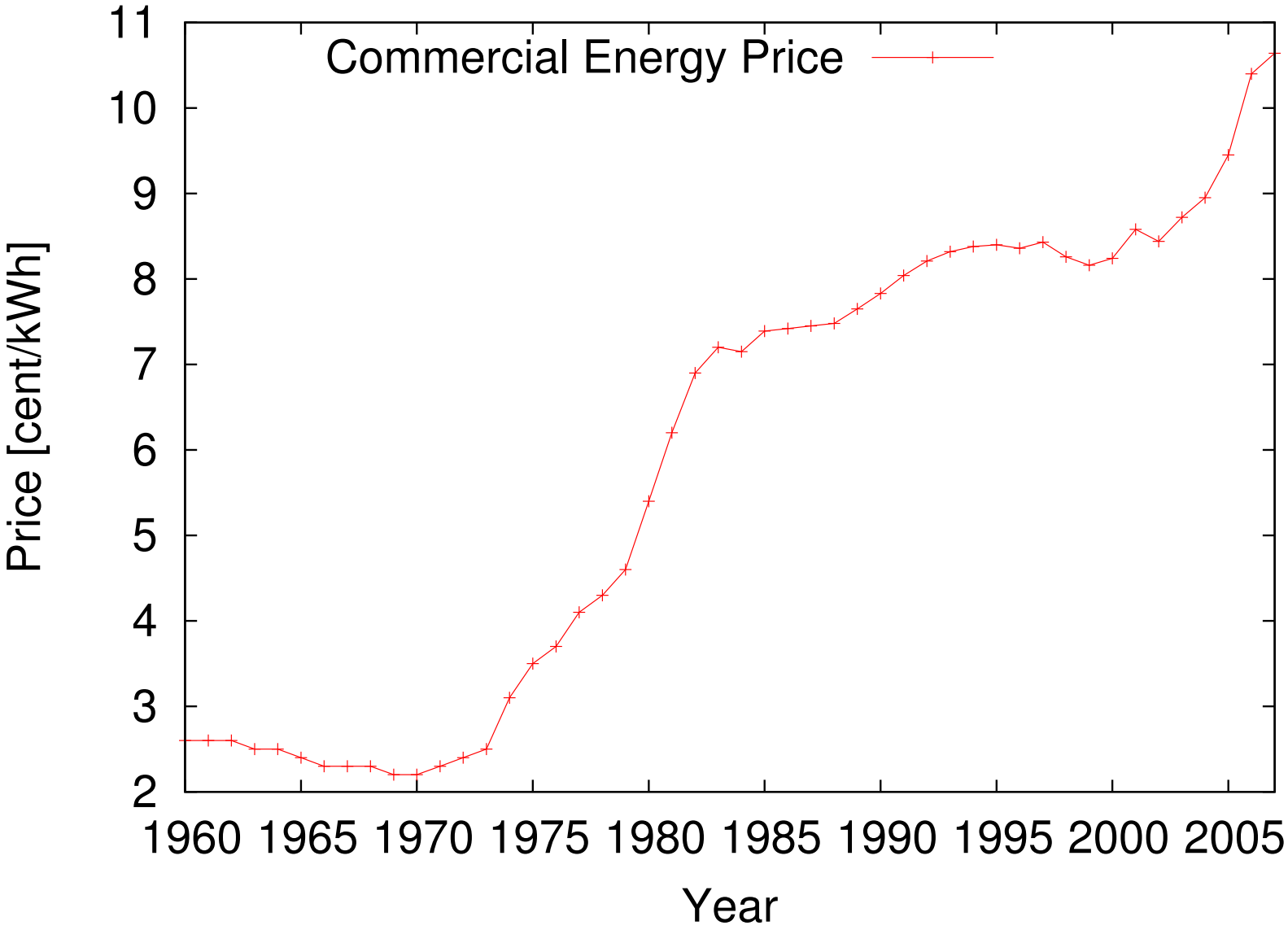
Open Systems Lab
Indiana University
Bloomington, USA

CAC'09 - IPDPS'09

Rome, Italy

May, 25th 2009

Motivation I (economic)



Motivation II (personal)

Motivation III (scientific)

- Interconnection network is the heart of parallel computing
  - How do we compare different network technologies?
  - Microbenchmarks!
  - Often Latency and Bandwidth only
  - **Is this enough to predict application performance?**

- Power consumption is becoming a problem for system designers
  - Green500 list as an addition to Top500
  - Power input (cooling!) major design goal for large systems
  - **What about power efficiency of the network?**

Experiment Setup

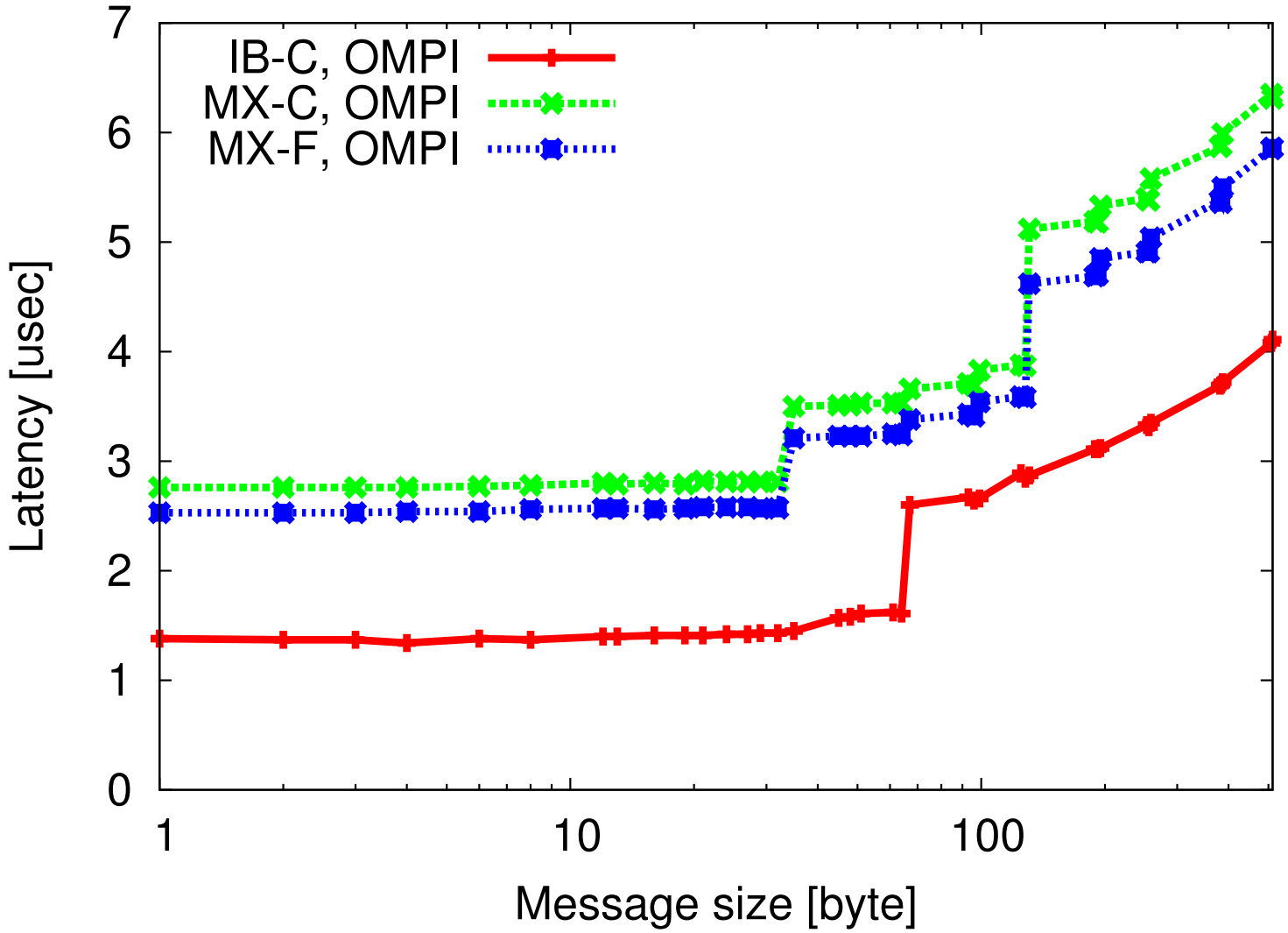
We compare three different network technologies

- Fiber-based Myrinet 10G
- Copper-based Myrinet 10G
- Copper-based ConnectX InfiniBand

We compare latency and bandwidth results (NetPIPE) and application performance on absolutely identical systems.

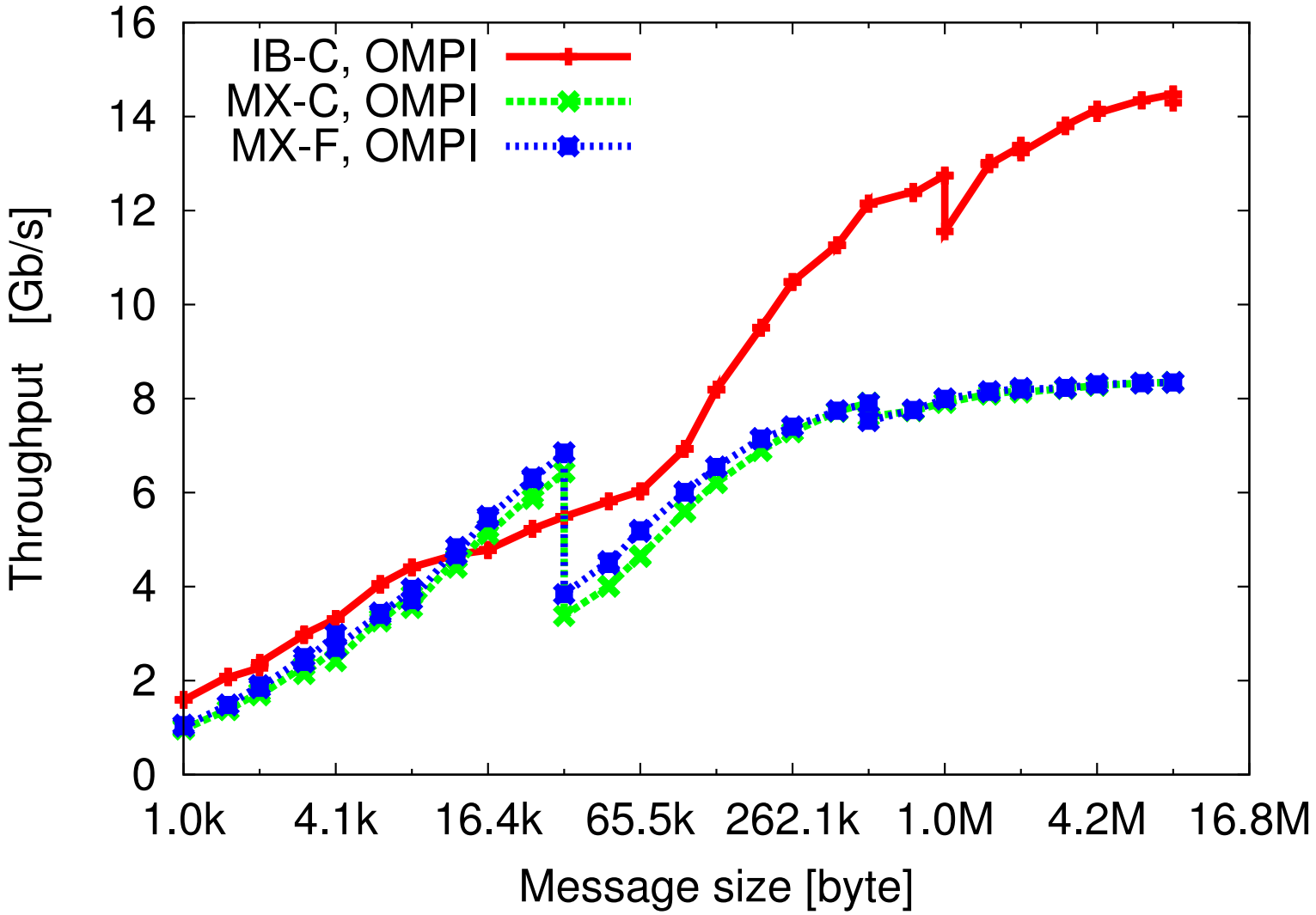
- OpenMPI 1.2.8, OFED 1.3, MX 1.4.3
- SLES 10 SP 2 (Linux 2.6.16)
- 14 nodes, 2 $\times$ 4 Xeons L5420 2.5 GHz
- 4 GiB RAM per core

Microbenchmark Results - Latency



Latency: IB 1.4μs, MX-F 2.5 μs, MX-C 2.8 μs

Microbenchmark Results - Throughput



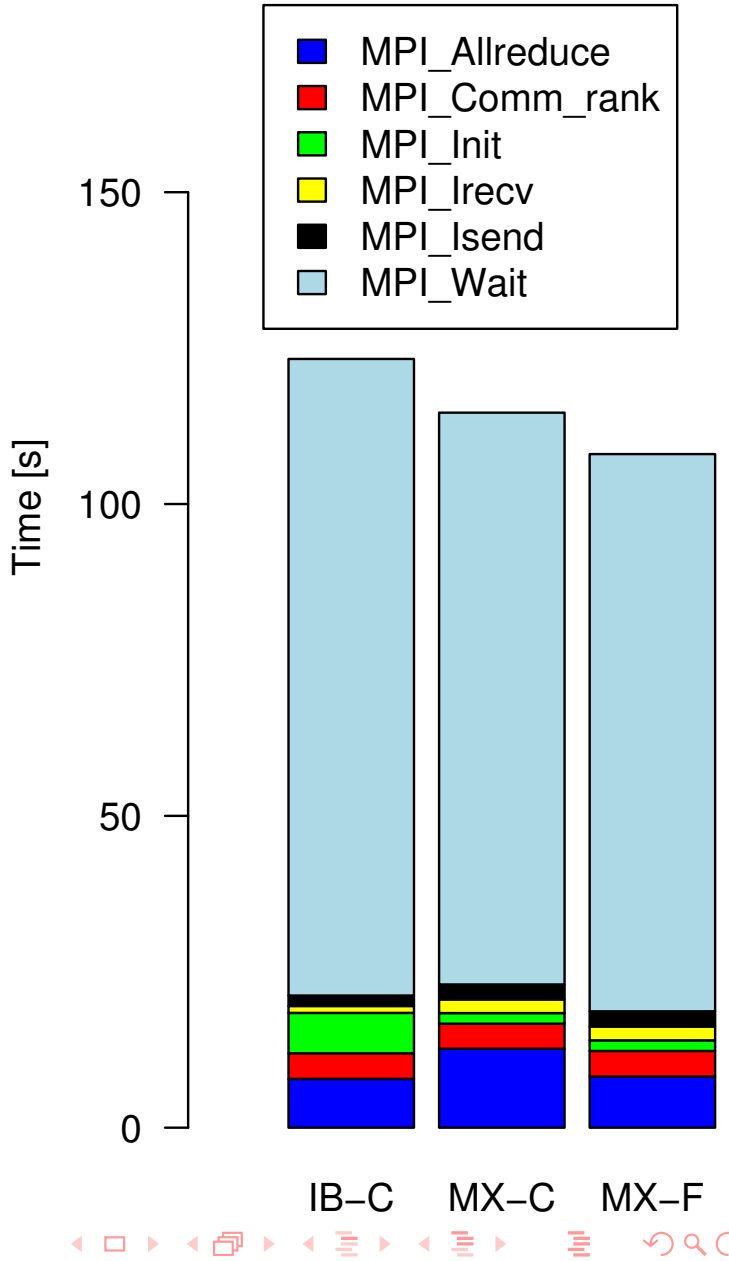
Bandwidth: IB 13.9 Gib/s (86.9%), MX 9.1 Gib/s (91%)

Microbenchmark Summary

- Results:
  - IB performs significantly better in nearly all configurations!
  - MX-F is slightly faster than MX-C
  - OMPI's MX eager-rendezvous switching point seems suboptimal
- Projection:
  - IB should deliver higher application performance
  - no data about power consumption yet
- $\Rightarrow$ proceeding to real application runs!
  - three runs with each application/network
  - lowest running time counts
  - all results were very stable ($<$ 3% variance)

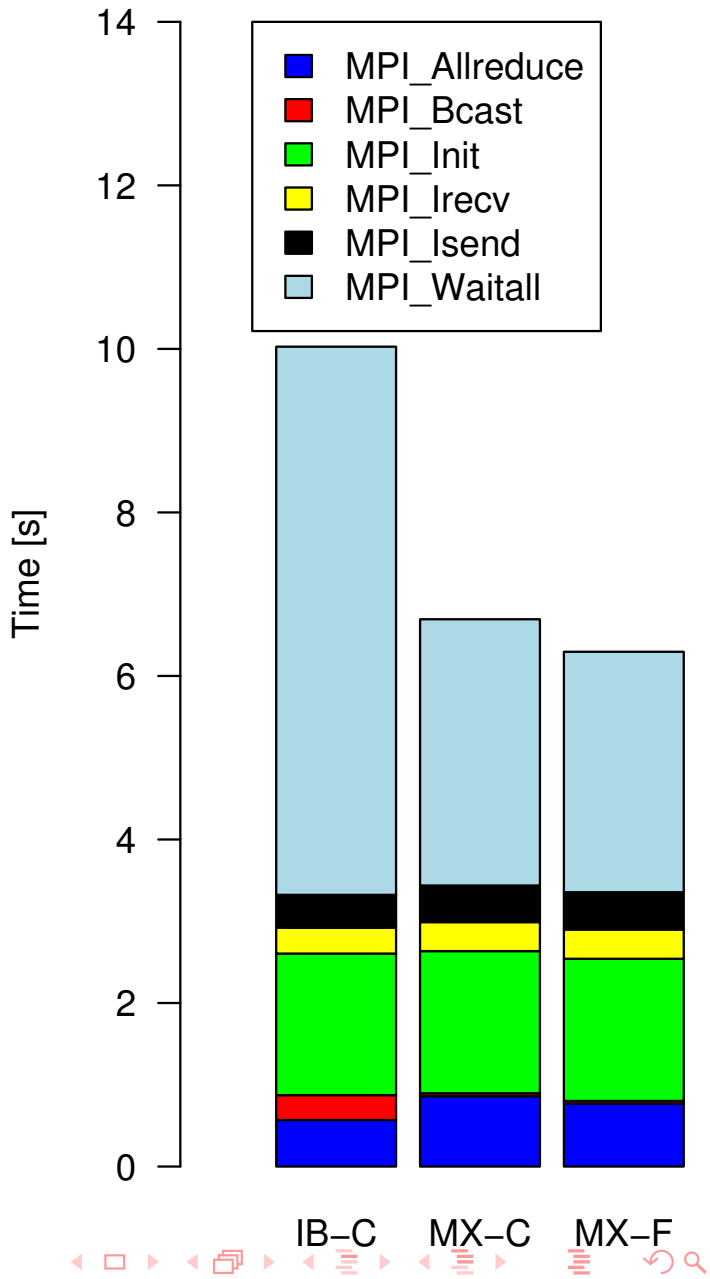
Application Performance - MILC

- Quantum chromodynamics code (nuclear physics)
- Multiple programs
- We used NERSC ”medium” benchmark for su3rmd
- Runtime:
  - IB: 444s (123s MPI)
  - MX-C: 435s (115s MPI)
  - MX-F: 426s (107s MPI)



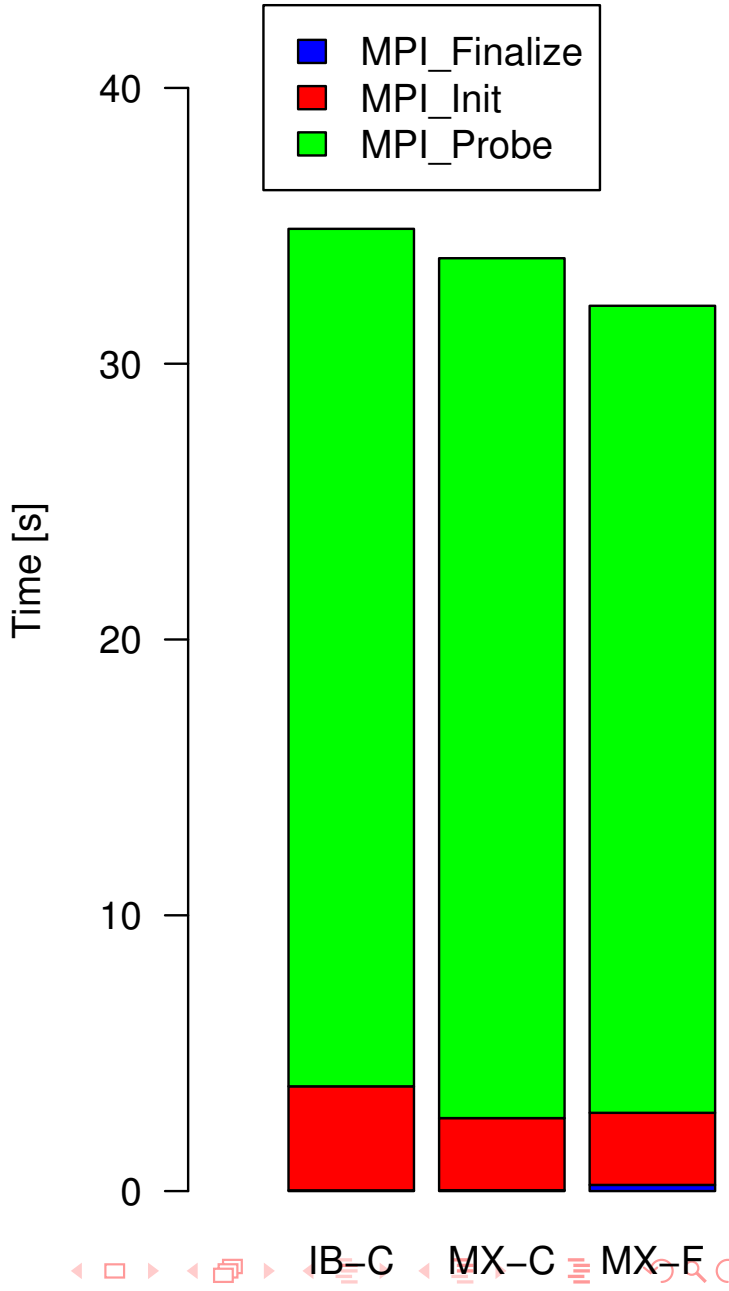
Application Performance - POP

- Ocean circulation simulations
- We used the x1 POP benchmark (32 cores on 14 nodes)
- Runtime:
  - IB: 66s (10s MPI)
  - MX-C: 63s (7s MPI)
  - MX-F: 61s (5s MPI)



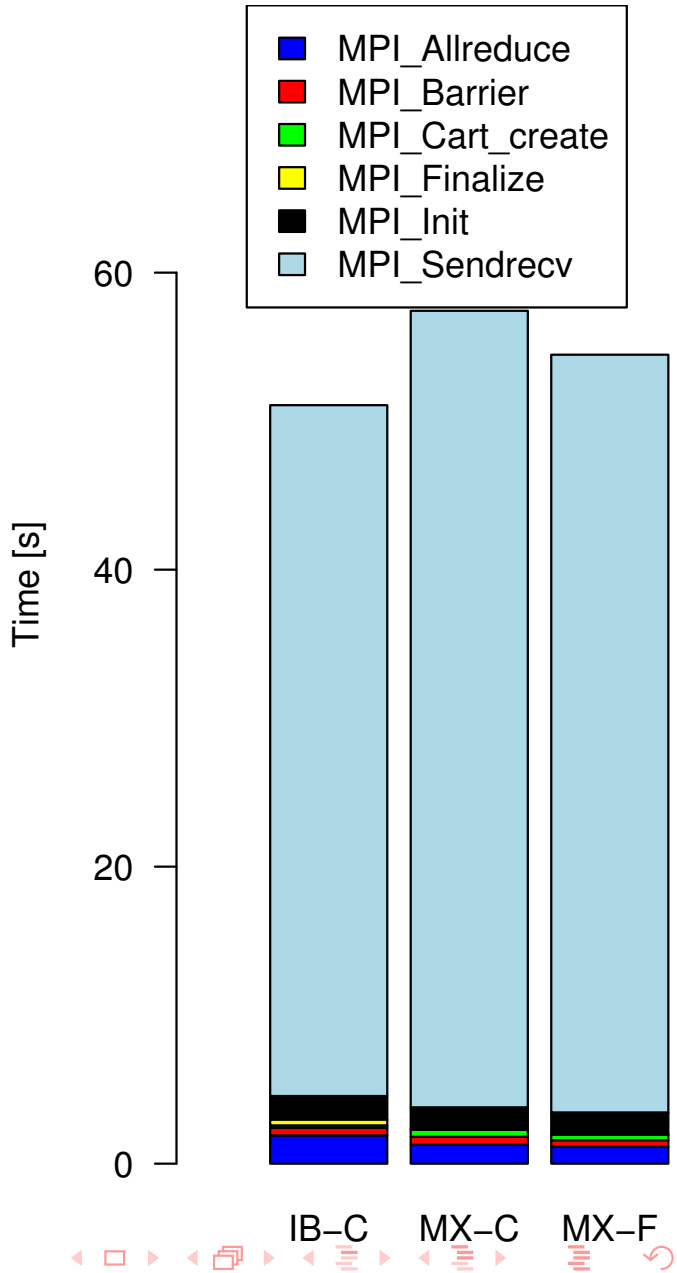
Application Performance - RAxML

- Models evolution by building phylogenetic trees from DNA
- We calculated 112 trees (1 per core) from 50 genome sequences with 5000 base pairs each
- Runtime:
  - IB: 746s (35s MPI)
  - MX-C: 743s (32s MPI)
  - MX-F: 738s (32s MPI)!



Application Performance - WPP

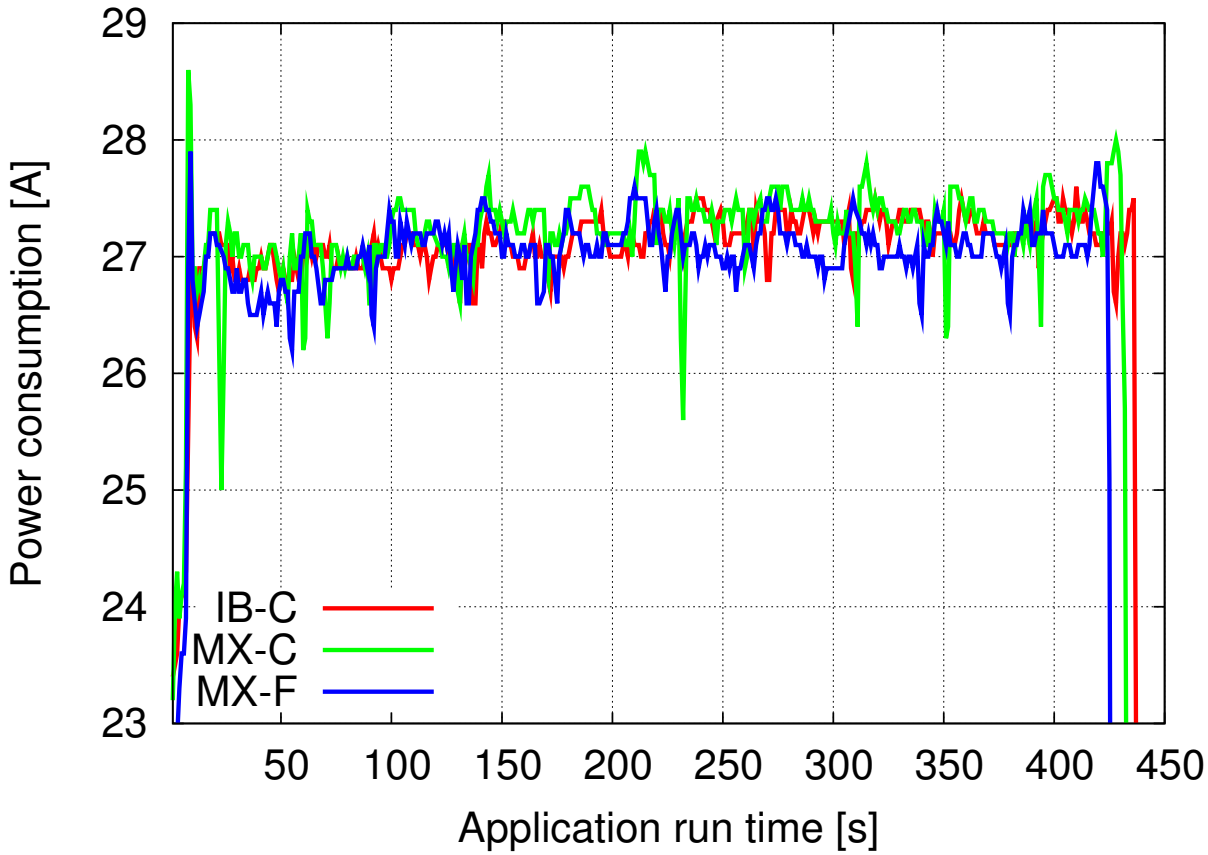
- Simulates time-dependent elastic and viscoelastic propagation of waves which occur during earth quakes and explosions
- 3D seismic modelling with finite difference methods
- 30k $\times$ 30k $\times$ 17k grid, single wave source (LOH1 example) on 112 cores
- Runtime:
  - IB: 702s (51s MPI)
  - MX-C: 706s (57s MPI)
  - MX-F: 701s (53s MPI)!



Power Measurements

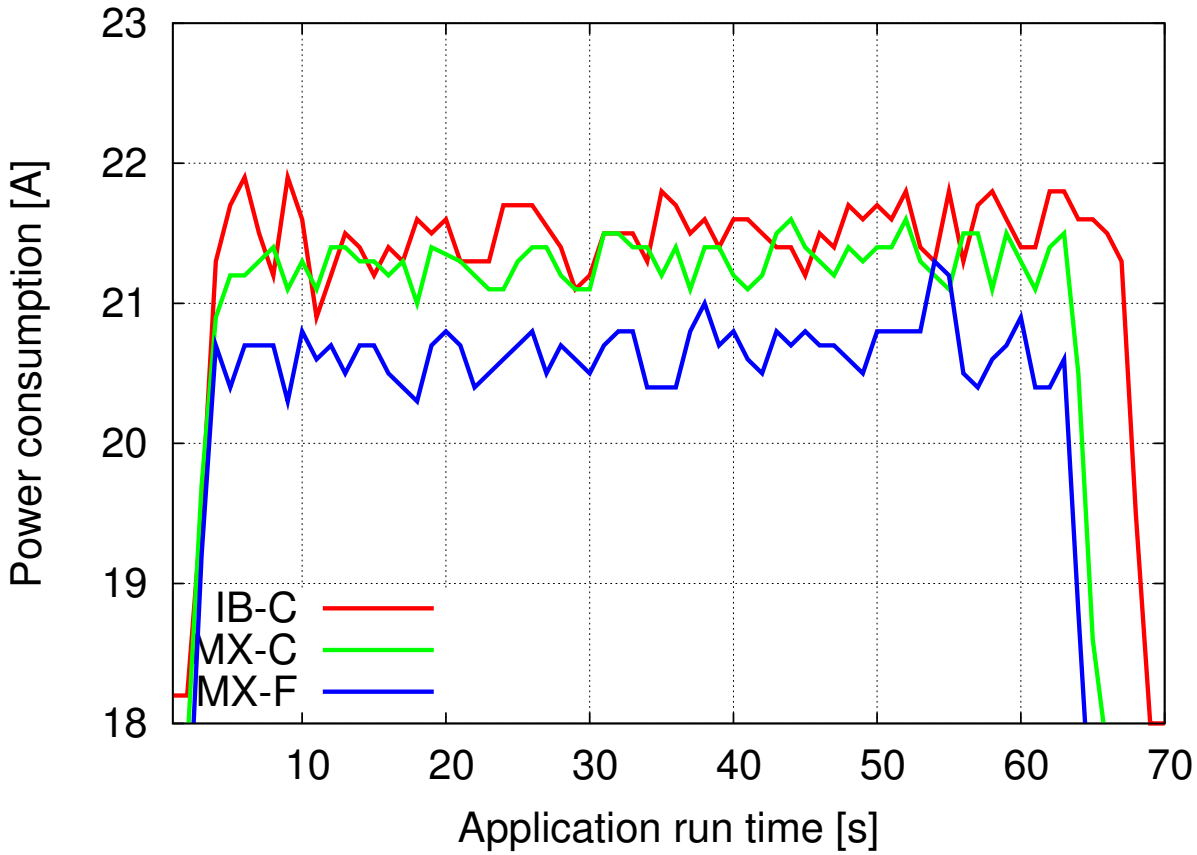
- Methodology:
  - two APC 7800 PDUs, resolution 0.1 A (120 V)
  - data sampled every second via SNMP
  - compute total power consumption as discrete integral
- Base Data:
  - idle system: IB 17.7 A, MX-C 17.3 A, MX-F 16.9 A
  - IB switch: Cisco TopSpin SFS 7000D 0.48 A
  - MX switch: 0.75 A (0.45 A w/o fan)
- 4 nodes idle vs. 8 MiB message-stream:
  - IB: 3.9 A / 5.0 A
  - MX-C: 3.77 A / 4.95 A (PML OB1)
  - MX-C: 3.77 A / 4.75 A (MTL MX)

Power Consumption - MILC



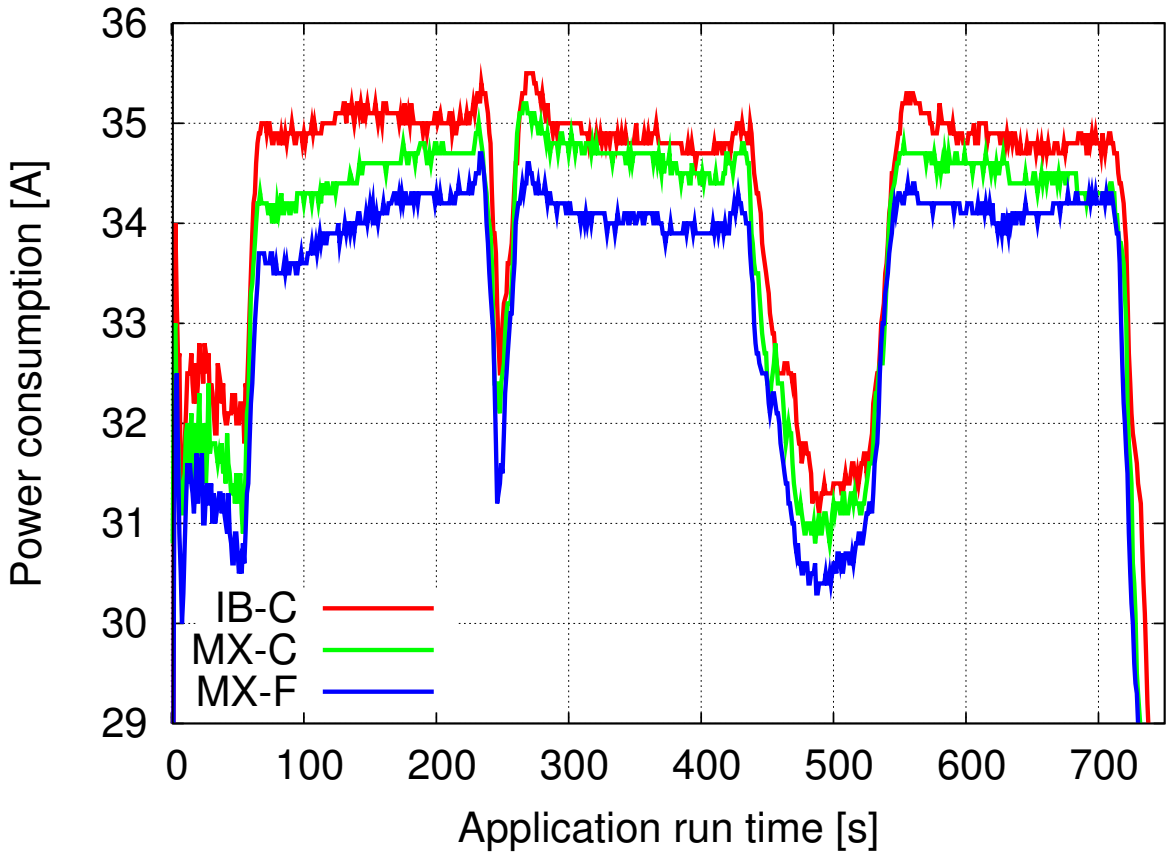
Energy: IB 3.879 kWh, MX-C 0.1% less, MX-F 1.5% less

Power Consumption - POP



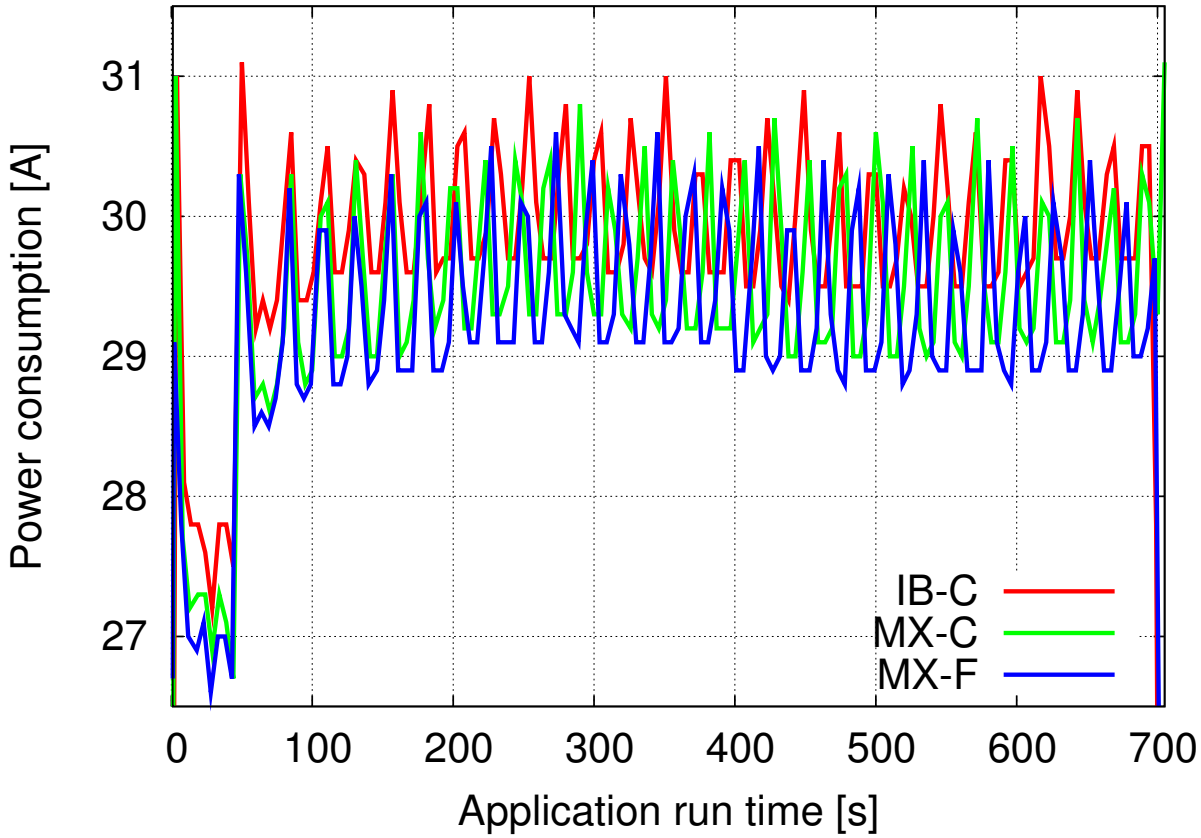
Energy: IB 0.458 KWh, MX-C 4.6% less, MX-F 11.3% less

Power Consumption - RAxML



Energy: IB 8.315 kWh, MX-C 1.8% less, MX-F 3.6% less

Power Consumption - WPP



Energy: IB 6.807 KWh, MX-C 0.4% less, MX-F 1.4% less

Conclusions

- Microbenchmarks and simple metrics such as latency and bandwidth are not accurate performance predictors.
- Other factors influence performance of parallel applications, for example tag matching in hardware, memory registration and cache pollution.
- The network fabric can have an important impact on power consumption, up to 11% in our experiments.

Future Work

- more power aware network fabric comparisons should performed (not by us)
- study influence of the driver stack on application performance

Thanks

Thanks for your attention!

Questions?