HPC Activities at CERN

Ioannis Agtzidis – CERN IT
Michal Husejko – CERN IT

29 October 2013

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

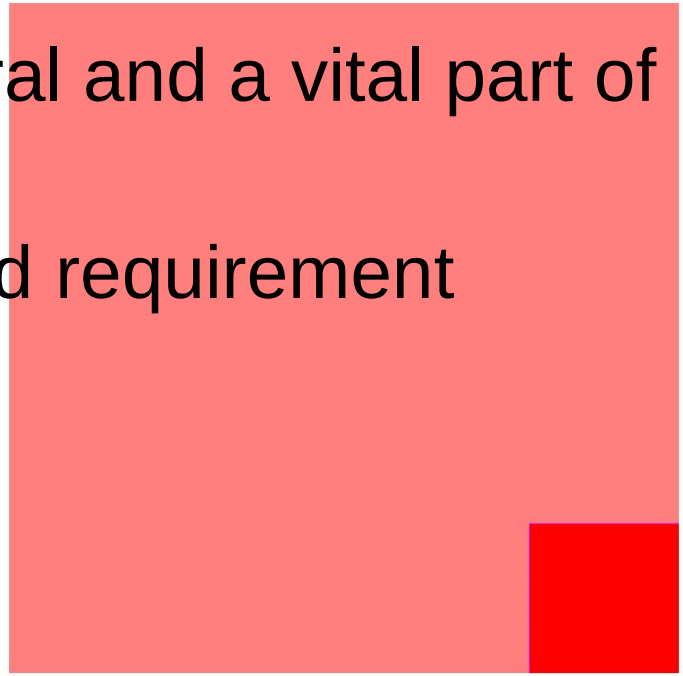
Outline

- Introduction
  - Why HPC
  - Parties served from HPC
- Principles on problem scaling
- Preliminary results
  - Memory-Disk I/O bounded
  - Memory bandwidth bounded
  - Interconnection impact
- Decisions on the subject
- Future steps

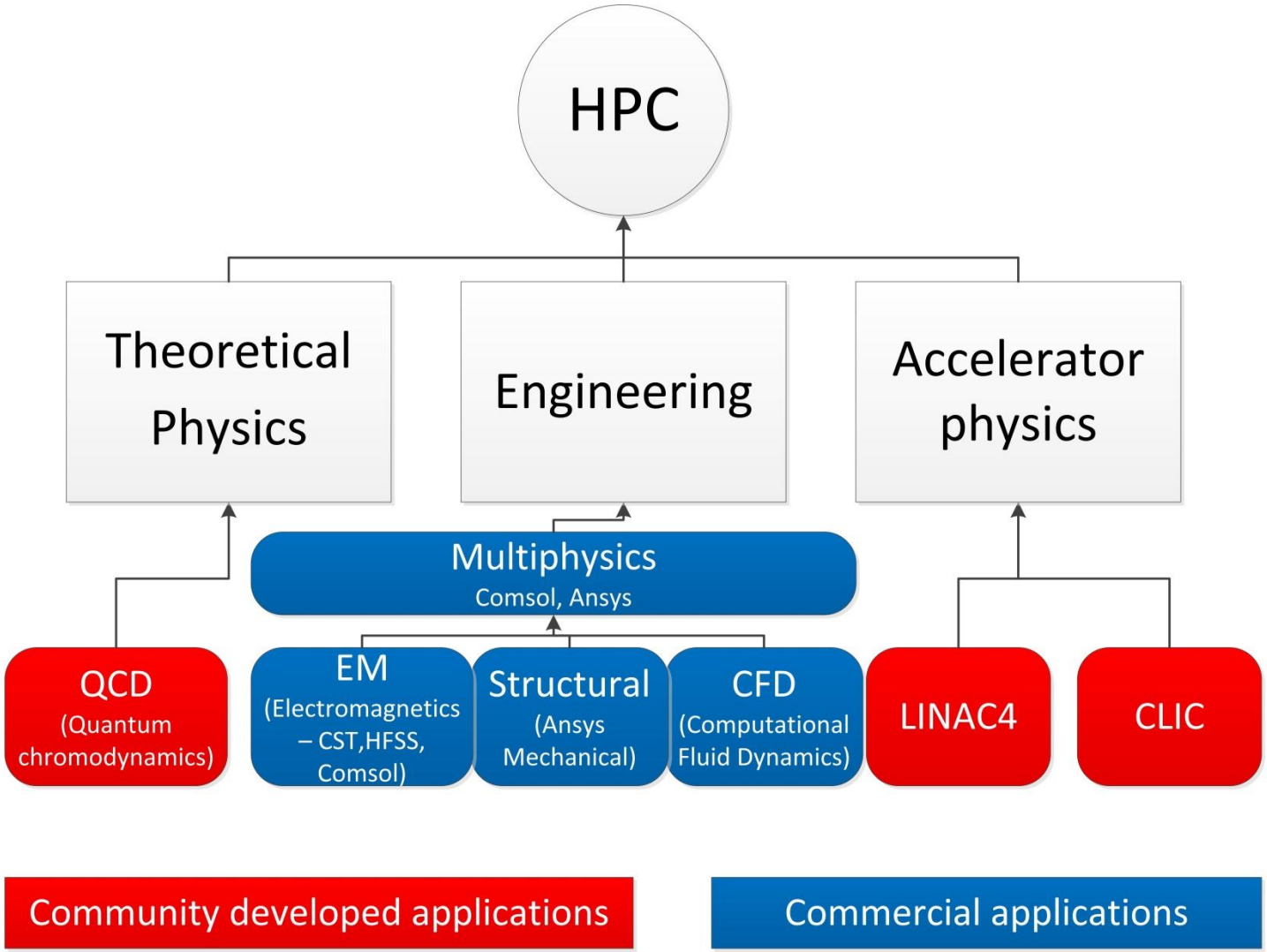
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Why are we interested in HPC?

- Current batch service and computing resources can cover most of the use cases for physics computing
- Less than 5% of the applications have special requirements
  - 4000 batch nodes versus 200 nodes
- These applications are an integral and a vital part of physics computing
- These applications need detailed requirement analysis



Parties served from HPC



Problem scaling - Theory

Amdahl's Law

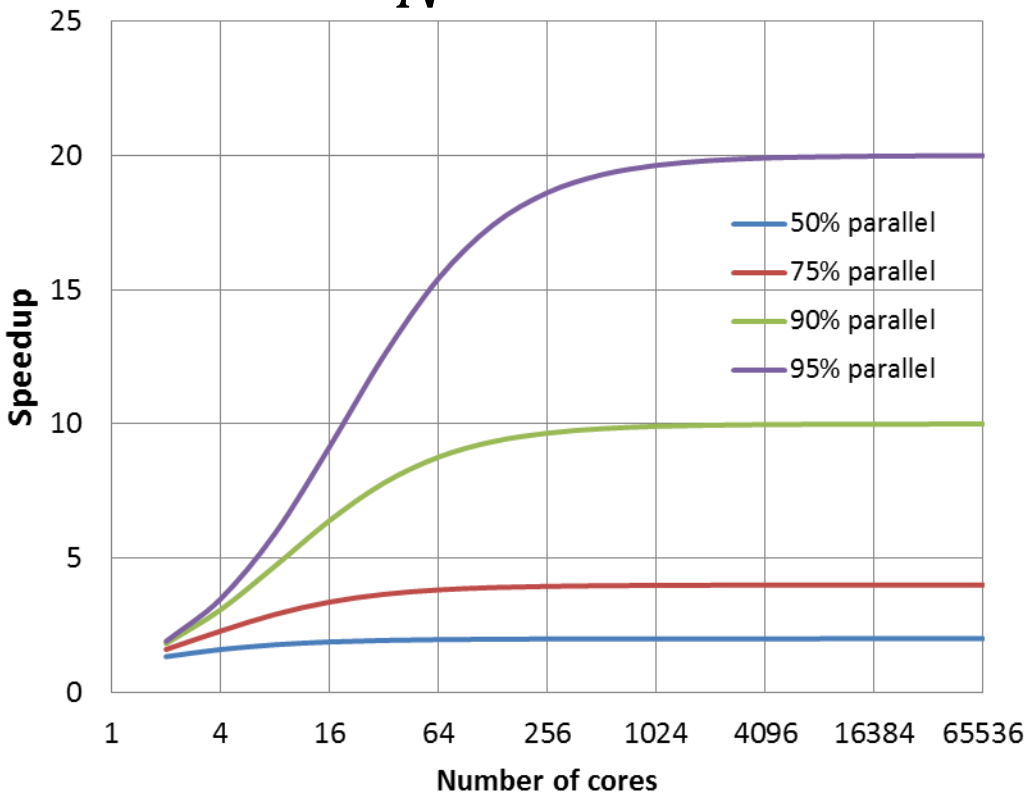
$$S(N) = \frac{1}{(1-P) + \frac{P}{N}}$$

S: Speedup

P: Parallel part of program

N: Number of cores

Pessimistic approach



Problem scaling - Theory

Gustafson's Law

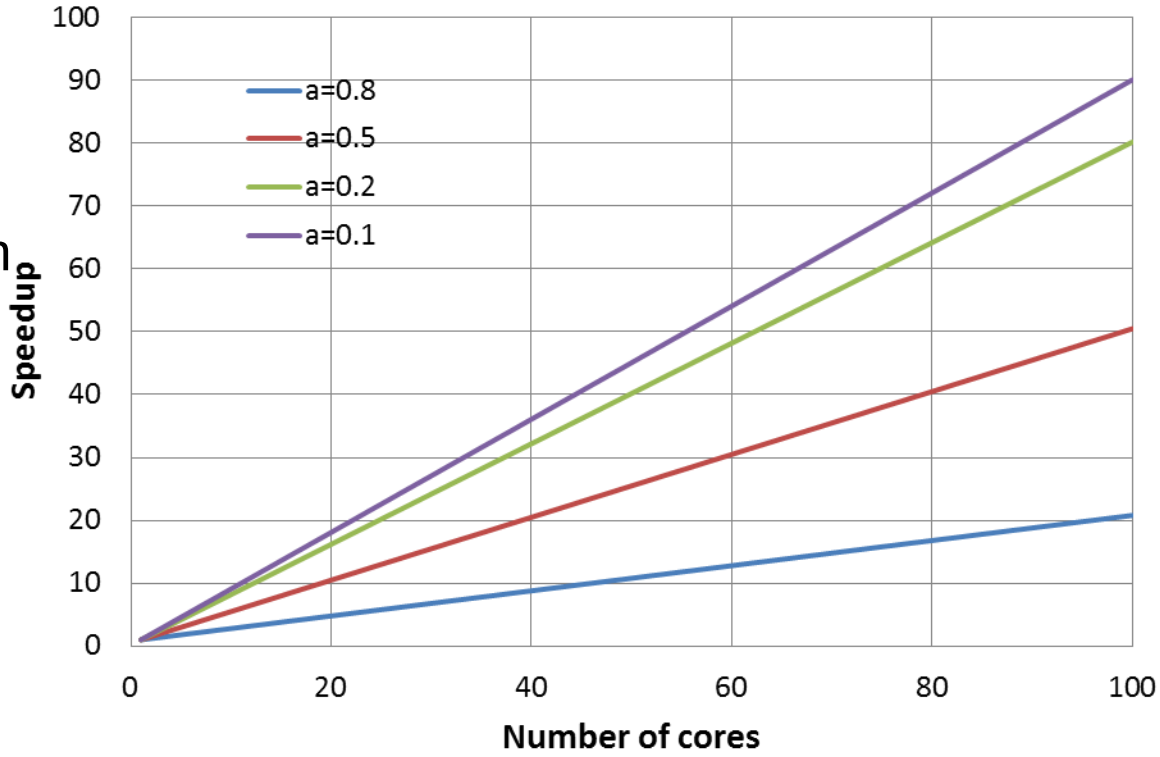
$$S(P) = P - a \cdot (P - 1)$$

S: Speedup

P: Number of Processors

a: Non-parallelised fraction

Optimistic approach

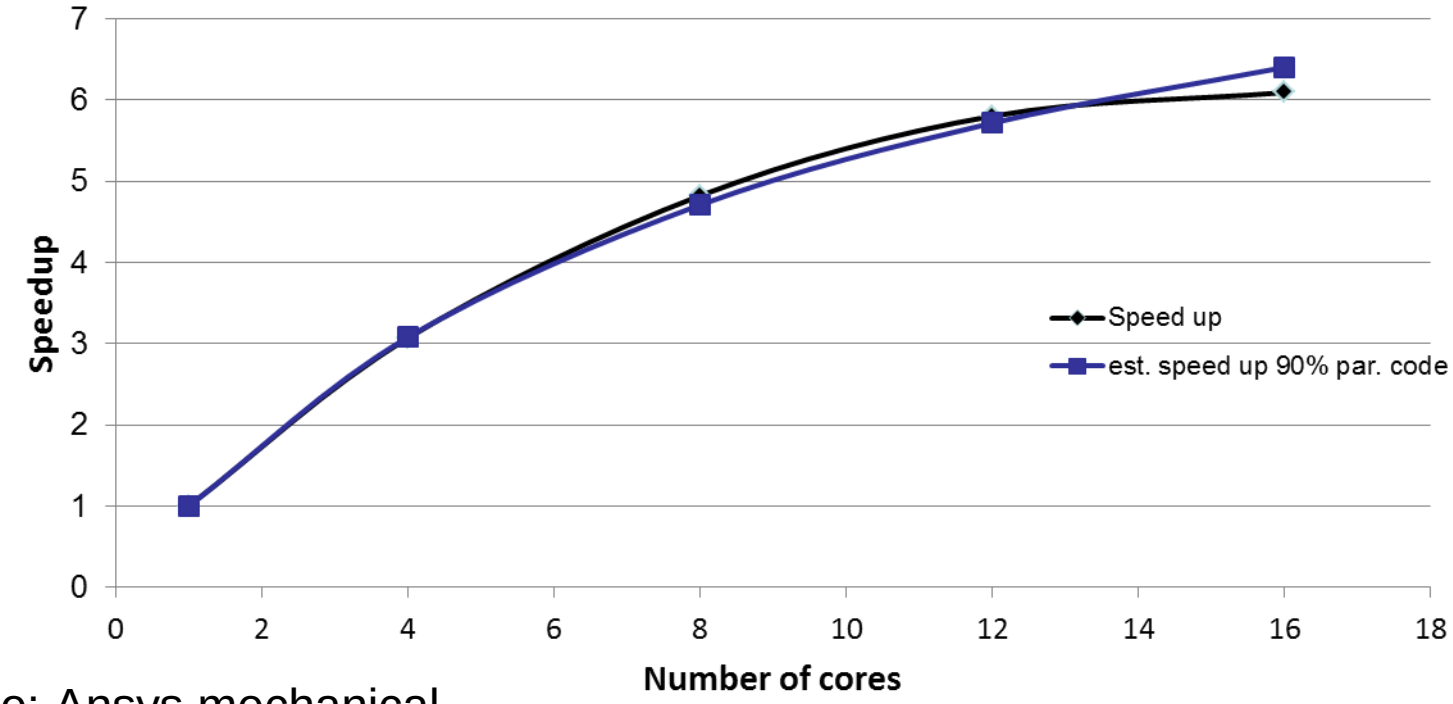


How we gathered statistics

- We use standard tools for measuring system wide statistics
  - iostat, dstat, sar, netstat
  - Intel PCM(Performance Counter Monitor)
- MPI level tracing
  - mpiP
- Data analysis
  - Matlab
  - Excel

Preliminary results 1/8

Almost perfectly complying with Amdahl's law



Use case: Ansys mechanical

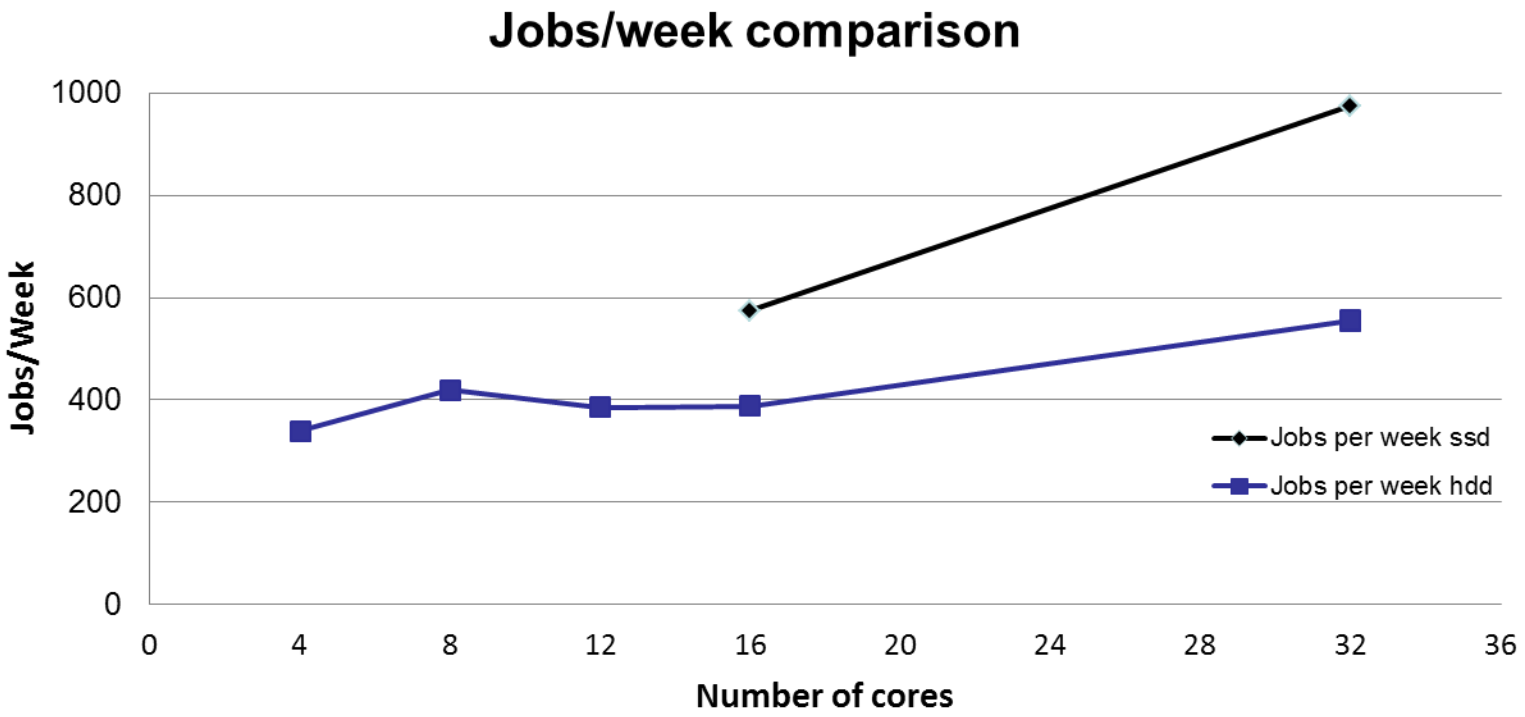
It aligns almost perfectly with Amdahl's law with 90% parallel program

Preliminary results 2/8

- But almost nothing is that ideal
- Many factors play a role
  - How the program is written
  - Disk I/O
  - Memory channel saturation
  - Interconnection bandwidth-latency
  - Communication overhead

Preliminary results 3/8

Impact of disk I/O

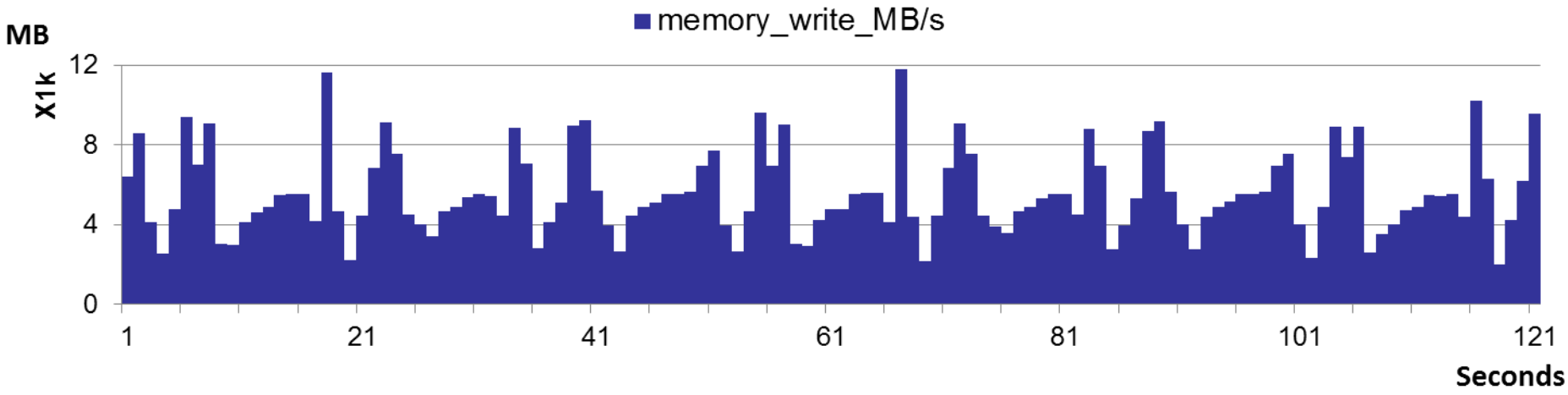


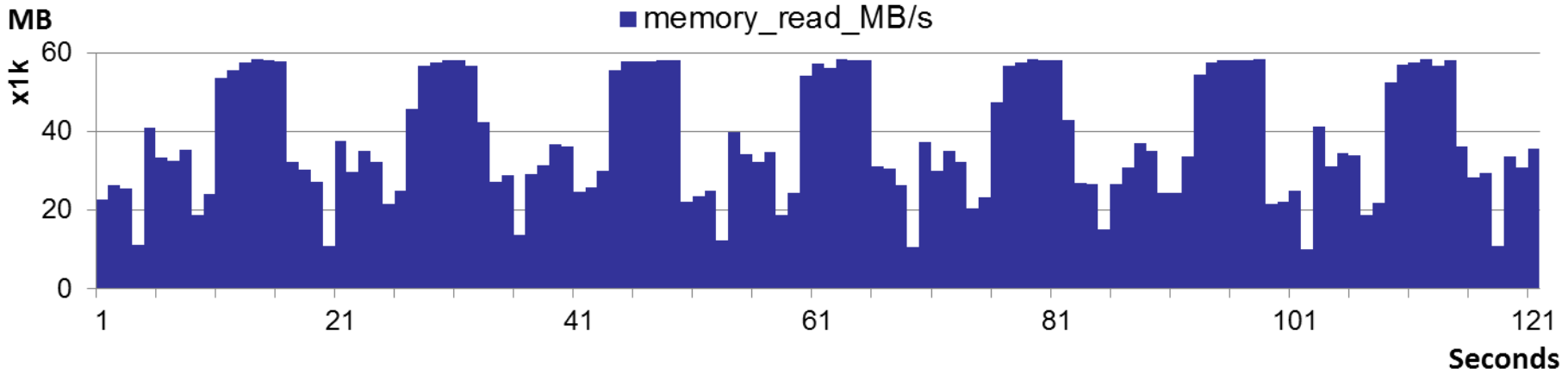
Use case: Ansys Mechanical

SSDs have higher bandwidth and higher IOPS

Preliminary results 4/8

Hit the maximum of memory bandwidth

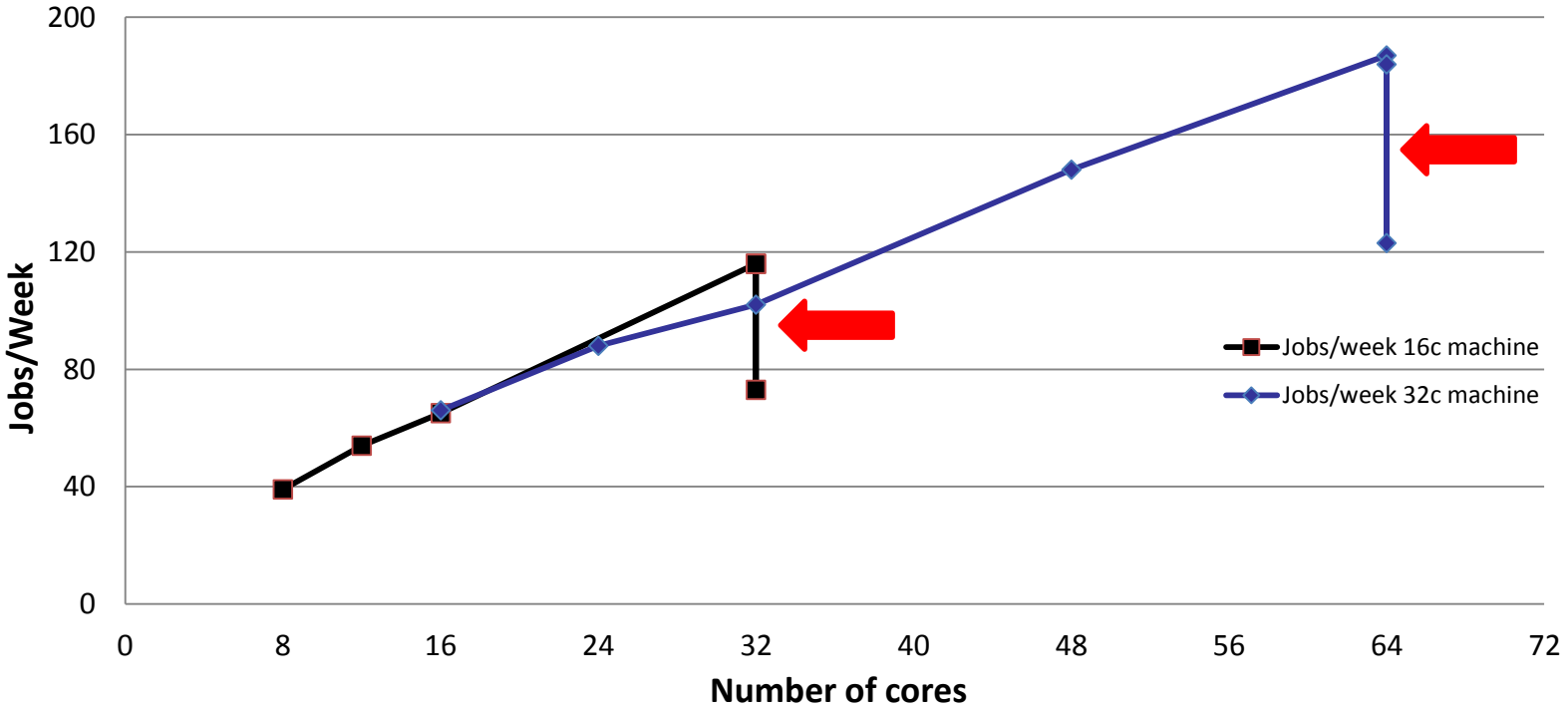




Use case: Fluent

Preliminary results 5/8

Compare 1Gb/s vs 10 Gb/s interconnect



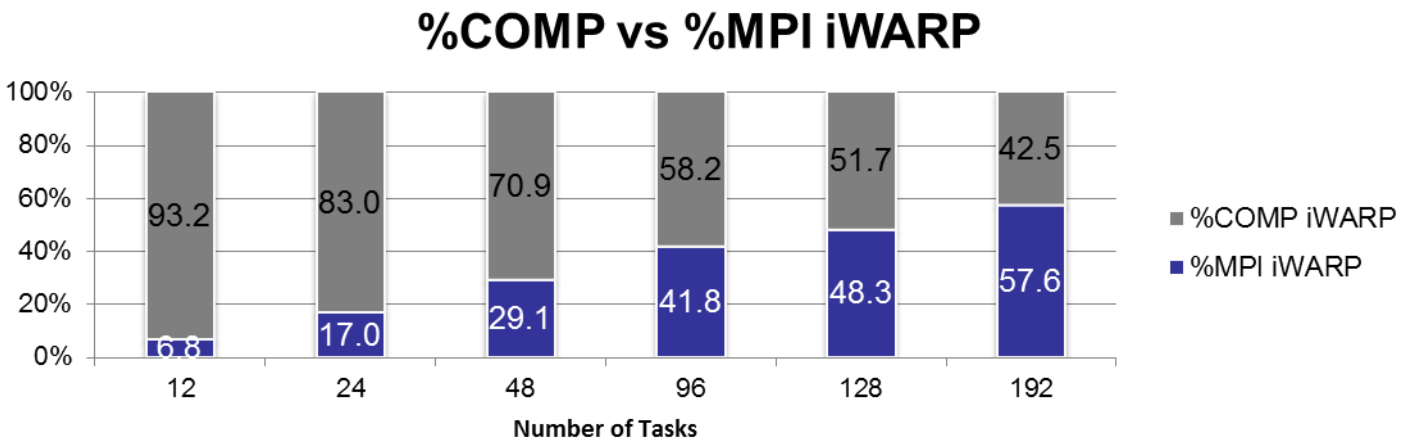
Use case: Fluent

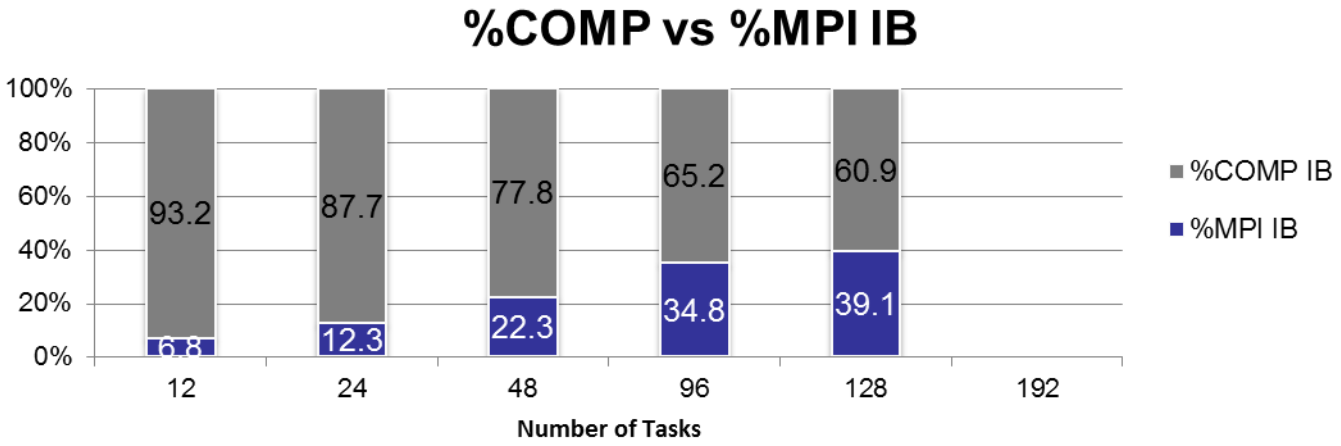
Significant drop in the number of Jobs/Week as we change to 1Gb/s interconnect

Preliminary results 6/8

Impact of communication as we increase the number of nodes & Impact of the interconnect

The percentage of communication increases as we increase the number of nodes.

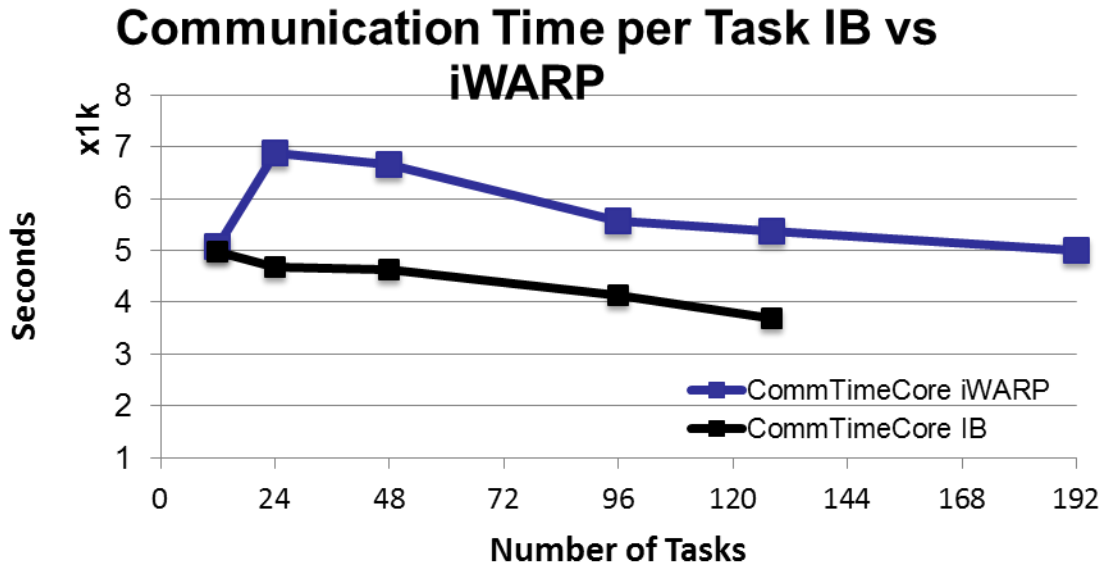


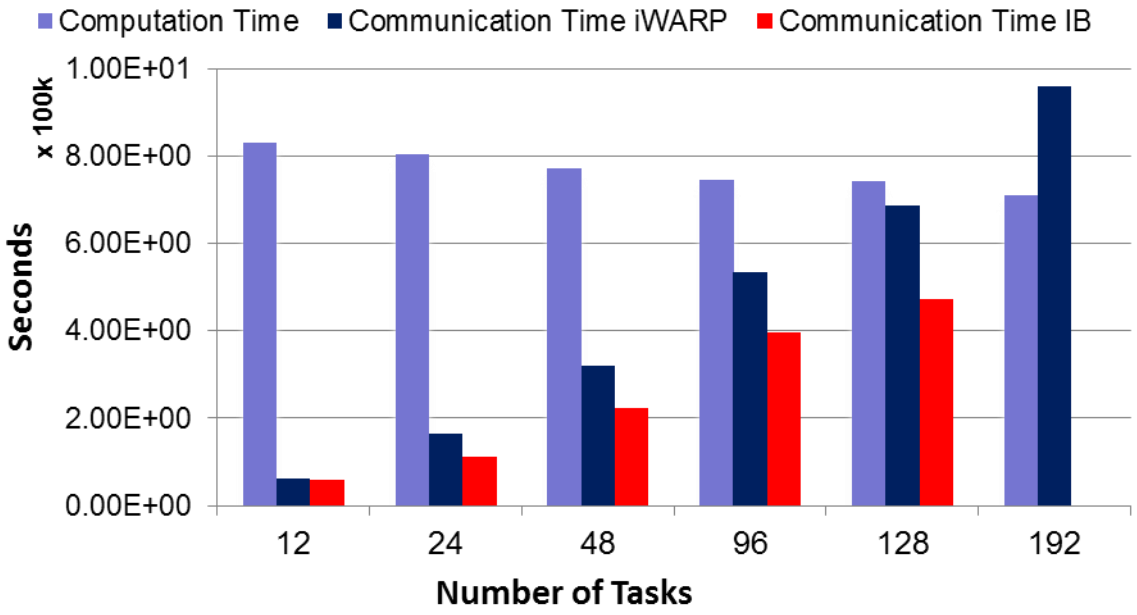


- Lattice QCD application
- Measurements with MpiP

Preliminary results 7/8

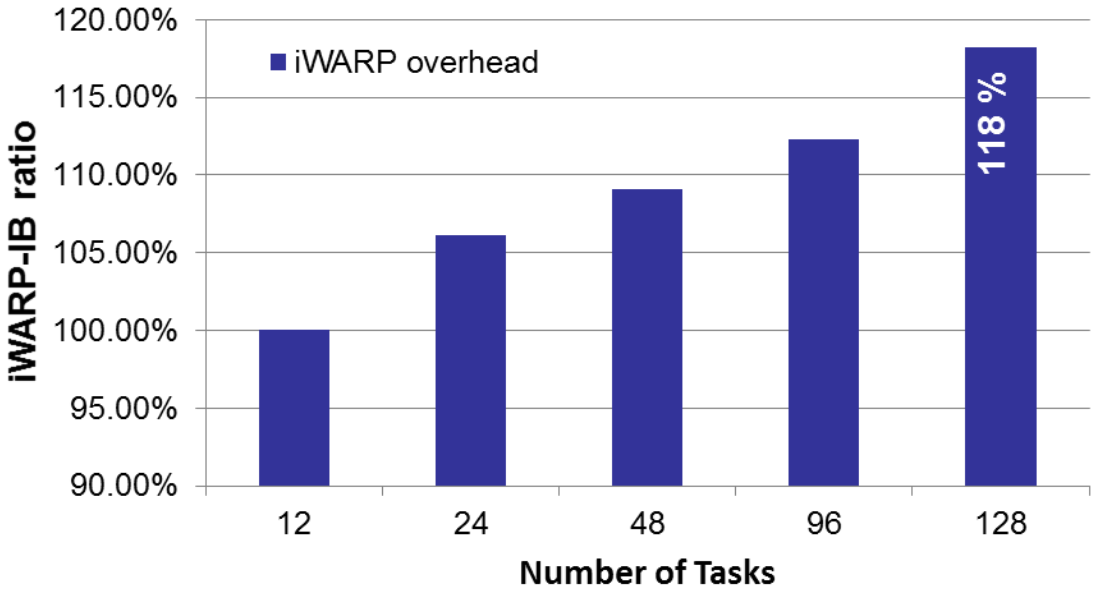
The computation time is the same for both technologies, but the communication time increases faster on the iWARP cluster.





Preliminary results 8/8

The overhead increases as the number of tasks increases.
It stays below 20% for up to 11 nodes (128 tasks), which represents real world use cases.



Synthetic benchmarks favour Infiniband but real world examples show low latency Ethernet is "*good enough*".

	Infiniband	Ethernet
Latency (µs)	1.2	7.1
Bandwidth (Gb/s)	40	10

- MPI: mvapich2-1.9 compiled from source for iWARP and mvapich2-1.7 for Infiniband provided by Intel
- Benchmarks: mvapich2 OSU benchmarks version 4.0.1 osu_latency (1Byte message) and osu_bw (16kB message)

What are we going to use after all?

- We are going to use equipment that we already have
- Interconnected with 10Gb low latency Ethernet
- Everything under single non blocking switch

There are some constraints:

- Certain technologies are used in the computing facilities
- Support staff and services
- Licensing (For commercial applications)

Future Steps

1. iWARP MPI scaling from 20 to 60 nodes
2. iWARP performance tuning though NIC/MPI parameters
3. More detailed analysis of MPI (MPI + PAPI)
4. Provide more tools to users for performance analysis

Thank you

Questions?