

JuRoPA

Jülich Research on Petaflop Architecture

One Year on

Hugo R. Falter, COO

Lee J Porter, Engineering

Outline

The work of ParTec on JuRoPA (HF)

Overview of JuRoPA (HF)

Basics of scalability (LP)

Expectations & some experiments (LP)

Conclusions (LP)

The work of ParTec on JuRoPA

project consulting	assistance to setup a tender, solving legal questions, preparing MoUs and contracts
manpower	on site system administrators backed by teams of experienced developers and architects
HPC systems consulting	system architects, HPC network specialists and HPC aware software engineers
experience	70 + man years of software development, parallel application debugging installation knowledge (hardware and software) and front-line support experience
flexible / dynamic	A pragmatic approach – always seeking alternatives, and finding workarounds
ParaStationV5 (open source)	ParTec created and maintains the ParaStation cluster middleware – the chosen cluster operating system of JuRoPA

What is ParaStationV5

ParaStation V5 is an integral part of the success of the JuRoPA machine

- ParaStation MPI
- ParaStation Process Management
- ParaStation Provisioning
- ParaStation Accounting
- ParaStation Parallel Debugger
- ParaStation **GridMonitor**
- ParaStation **Healthchecker**

The work of ParTec on JuRoPA (HF)

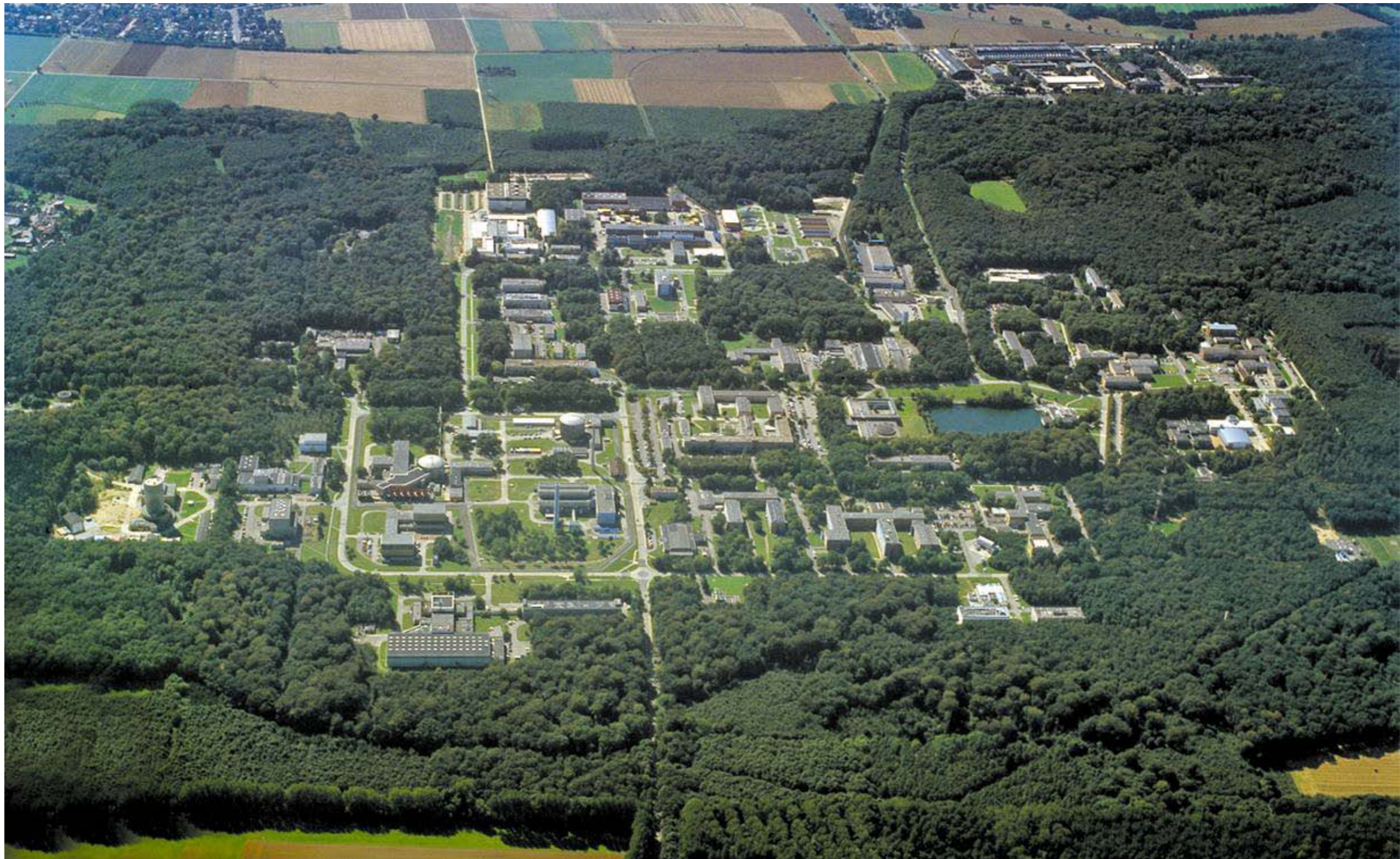
Overview of JuRoPA (HF)

Basics of scalability (LP)

Expectations & some experiments (LP)

Conclusions (LP)

Forschungszentrum Jülich (FZJ)



Jülich Supercomputing Centre (JSC)



JuRoPA-JSC

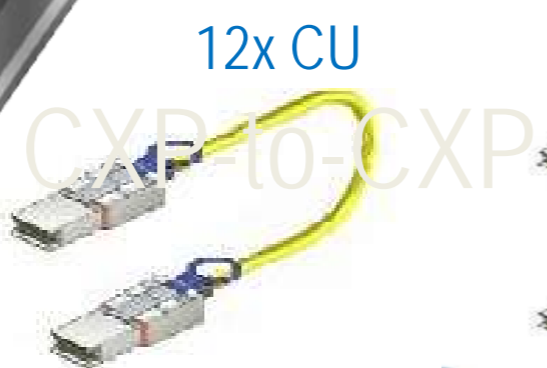


IB HARDWARE – JUROPA CLUSTER

**Sun M9 QDR switch – 24 x 9
12X ports**



**Sun Blade 6048 InfiniBand QDR
Switched Network Express Module**



12x CU

CXP-to-CXP

12x Optical
CXP-to-CXP

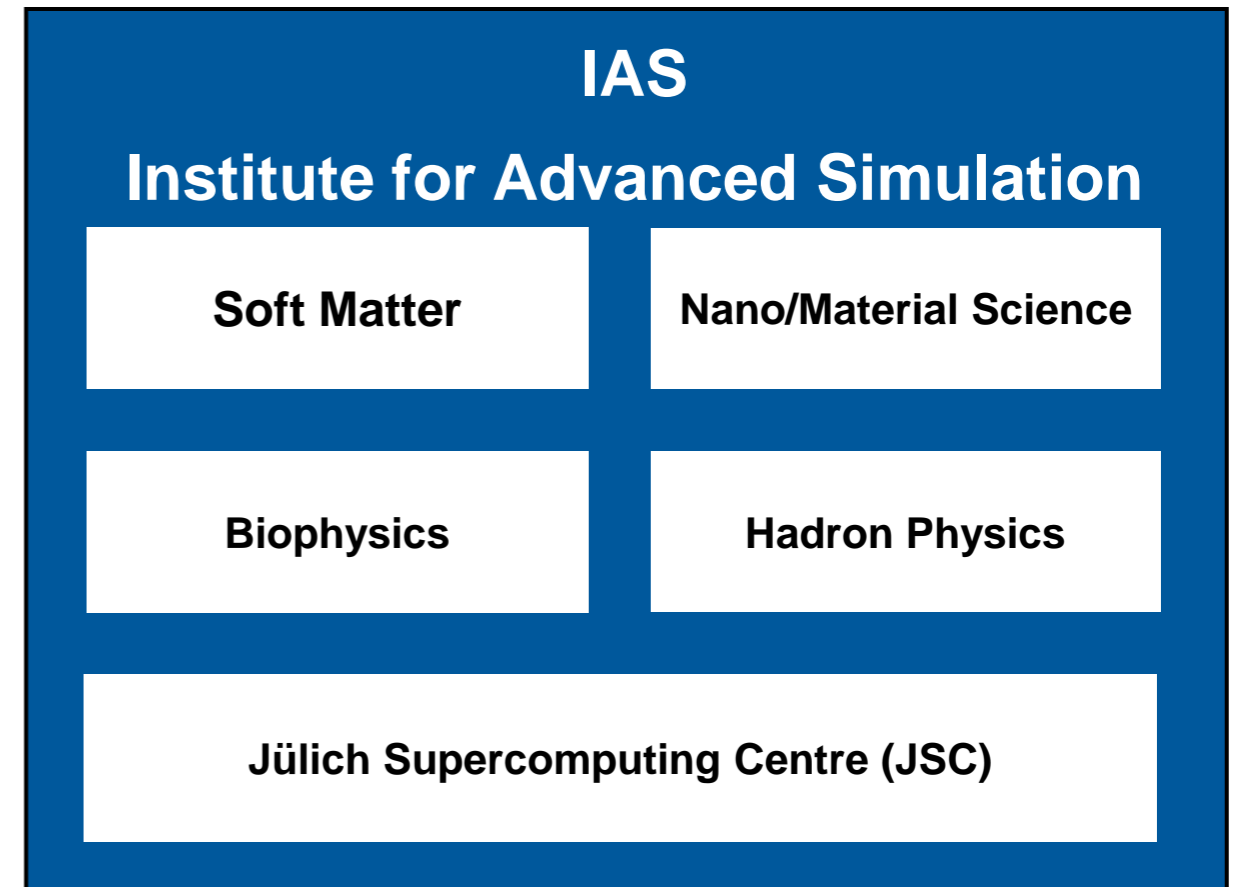


12x Optical/CU Splitter
CXP-to-3 QSFP

JuRoPA - HPC-FF



User Research Fields



- Chemistry
- Life + Environment
- Many Particle Physics
- Material Science
- Elementary Particle Physics
- Soft Matter
- Other

GCS: Gauss Centre for Supercomputing



Germany's Tier-0/1 Supercomputing Complex

- Association of Jülich, Garching and Stuttgart
 - A single joint scientific governance
 - Germany's representative in PRACE
- More information: <http://www.gauss-centre.de>

Overview

JSC:

Sun Blade SB6048

2208 Compute nodes

- 2 Intel Nehalem 2.93 GHz
- 24 GB DDR3, 1066 MHz Memory
- Mellanox IB ConnectX QDR HCA

17664 cores, 207 Tflops peak

HPC-FF:

Bull NovaScale R422-E2

1080 Compute nodes

- 2 Intel Nehalem-EP 2.93 GHz
- 24 GB DDR, 1066 MHz Memory
- Mellanox IB ConnectX QDR HCA

8640 cores, 101 Tflops peak

Lustre FileSystem:

- 2 MDS (Bull R423-E2)
- 8 OSS (Sun Fire 4170)
- DDN S2A9900
- Sun 4400 JBOD

Lustre over IB (530 TB)

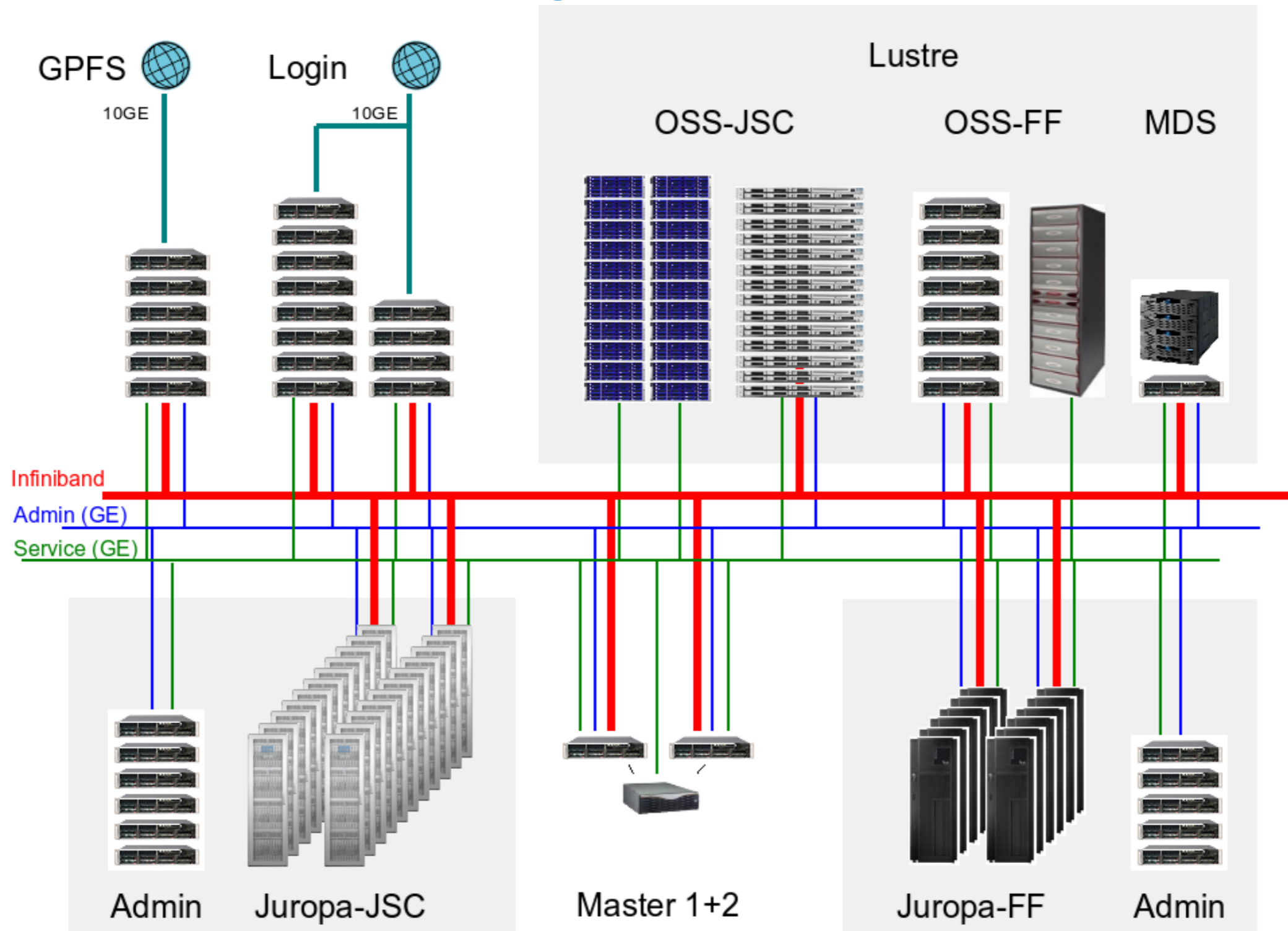


- Novell SUSE Linux Enterprise Linux
- ParaStation Cluster-OS

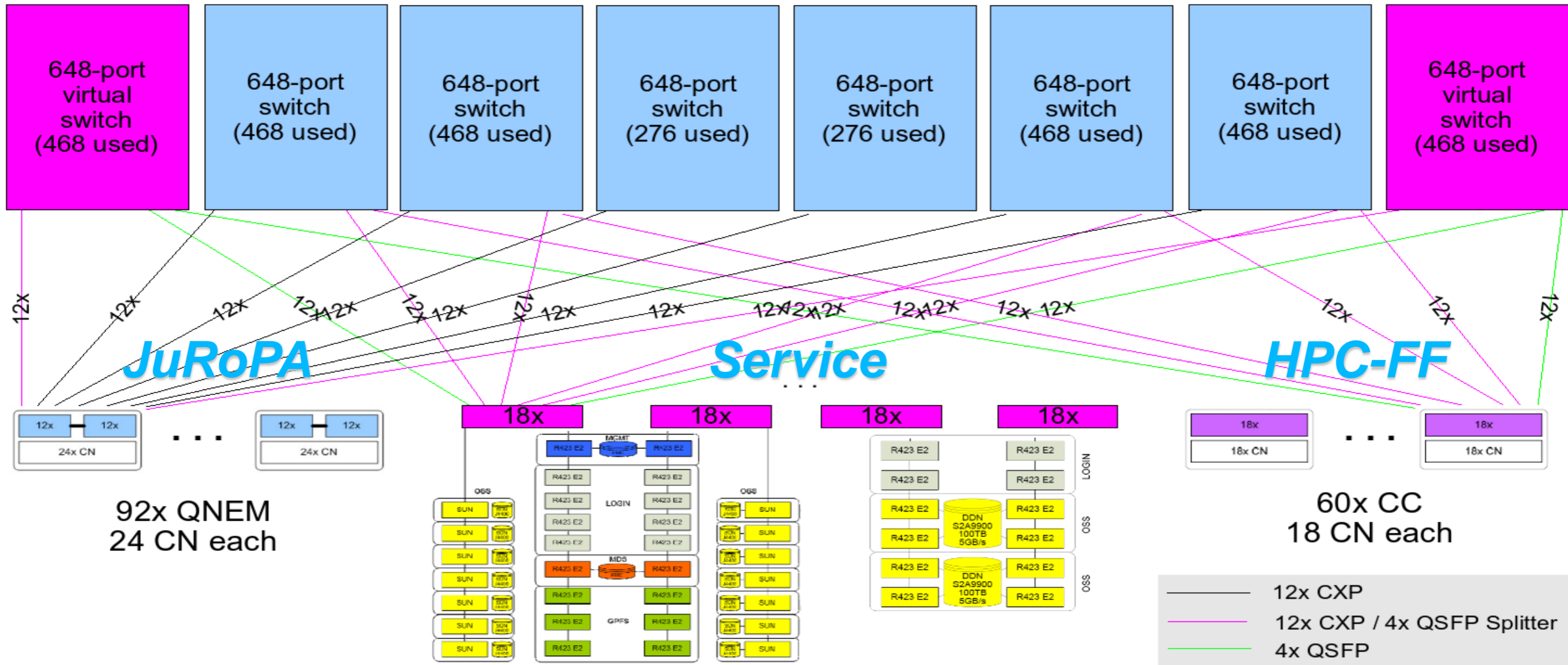
JuRoPA common ingredients

- 2208 + 1080 Compute Nodes
- Intel Nehalem Processor
- Mellanox 4th Gen. QDR InfiniBand interconnect
 - Global fat-tree network
 - Adaptive routing not yet implemented
- ParaStation V5 cluster middle-ware
- Intel development tools (compiler, libraries, etc.)
- Linux OS (starting w/ Novell SLES; moving to Novell SLERT)
- Lustre for I/O (but gateways to GPFS)
- Batch-system: Torque/Moab
- MPI: ParaStation (Intel-MPI, further MPIs from OFED)

Overall design – schematic view



23 x 4 QNEM modules, 24 ports each
6 x M9 switches, 648 ports max. each,



23 x 4 QNEM modules, 24 ports each
6 x M9 switches, 648 ports max. each,
468/276 links used

Mellanox MTS3600 switches (Shark), 36 ports, for service nodes

4 Compute Sets (CS) with 15 Compute Cells (CC) each
CC with 18 Compute Nodes (CN) and 1 Mellanox MTS3600 (Shark) switch each

Virtual 648-port switches constructed from 54x/44x Mellanox MTS3600

Status today

Juropa Machine at **90% aggregate utilization** outside of maintenance slots – utilization comparable to cloud systems

JSC portion is currently **overbooked X 10** – indicates this is a popular machine amongst scientific community

The work of ParTec on JuRoPA (HF)

Overview of JuRoPA (HF)

Basics of scalability (LP)

Expectations & some experiments (LP)

Conclusions (LP)

Scalability - Nehalem?

Newest, fastest, Intel processor

Technically:

- Great memory-bandwidth (up to 32 GB/s per socket) - QPI
- Superior cache structure
 - Large, low latency
 - 1st level: 32 kB – 8-way associative
 - 2nd level: 256 kB – 8-way associative
 - 3rd level: shared 8192 kB – 16-way associative
 - 128 bits into SSE register in one cycle

Results in > 90% efficiency in Linpack

- BlueGene ~82%, other Clusters 75%-80%

Scalability - QDR InfiniBand?

Short answer:

- Higher bandwidth (now at 32 Gbit/s)
- Lower latency (less than 1.0 μ sec point-to-point)

Actual reasons:

- Appropriate network topologies: Clos, fat-tree, Ω -network,....
- 4th Gen. QDR silicon has 36-port building blocks (24-ports in earlier versions)
 - Higher port arity \rightarrow fewer hops \rightarrow Reduced complexity – reduced end-to-end latency
- QDR supports adaptive routing (not yet implemented JuRoPA)
 - IB's static routing might lead to congestion

Scalability - Middleware

Cluster Middleware

ParaStationV5

- MPI – can it scale to > 10,000 MPI tasks
- Expose MPI tuning to user environment
- used the *ParaStation* MPI and scheduler to start and manage over 25000 MPI-processes – standard scheduler and MPI failed to scale (unexp. timeouts)
- Communication layer, process management, (pscom) designed to scale. ParaStation MPI uses pscom to start 25000 processes in less than 5 min!
- **ALL OPEN SOURCE**

Scalability - Administration

ParaStation Healthchecker:

special development of a tool to detect broken components in software and hardware at node boot and during job preamble.

IB network diags tools:

near full automatation in IB diagnostics, link monitoring, and active link management.

Problem tracker:

automated hardware / software PR submission with ParTec supported tracking system

ParaStationV5

The work of ParTec on JuRoPA (HF)

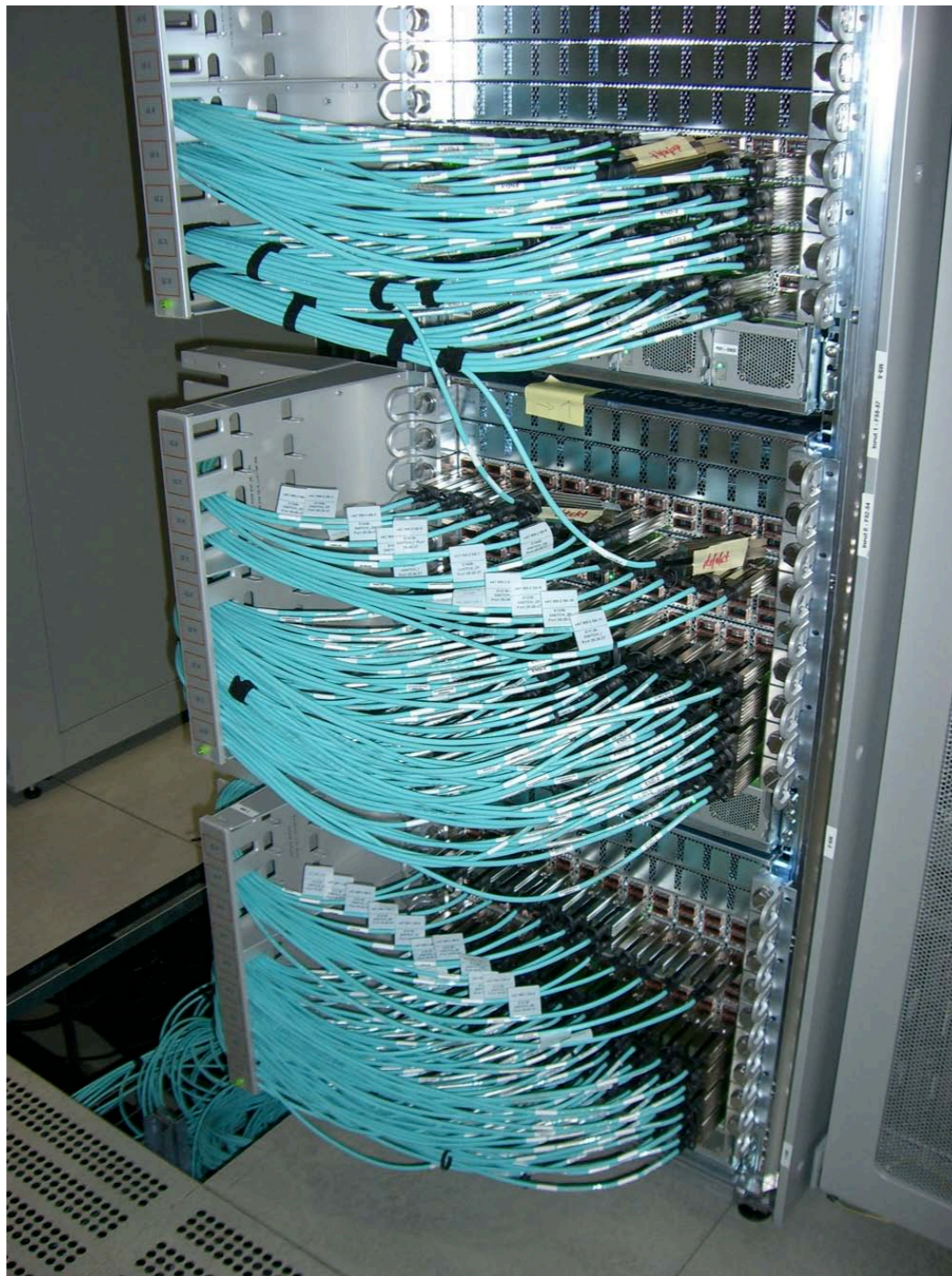
Overview of JuRoPA (HF)

Basics of scalability (LP)

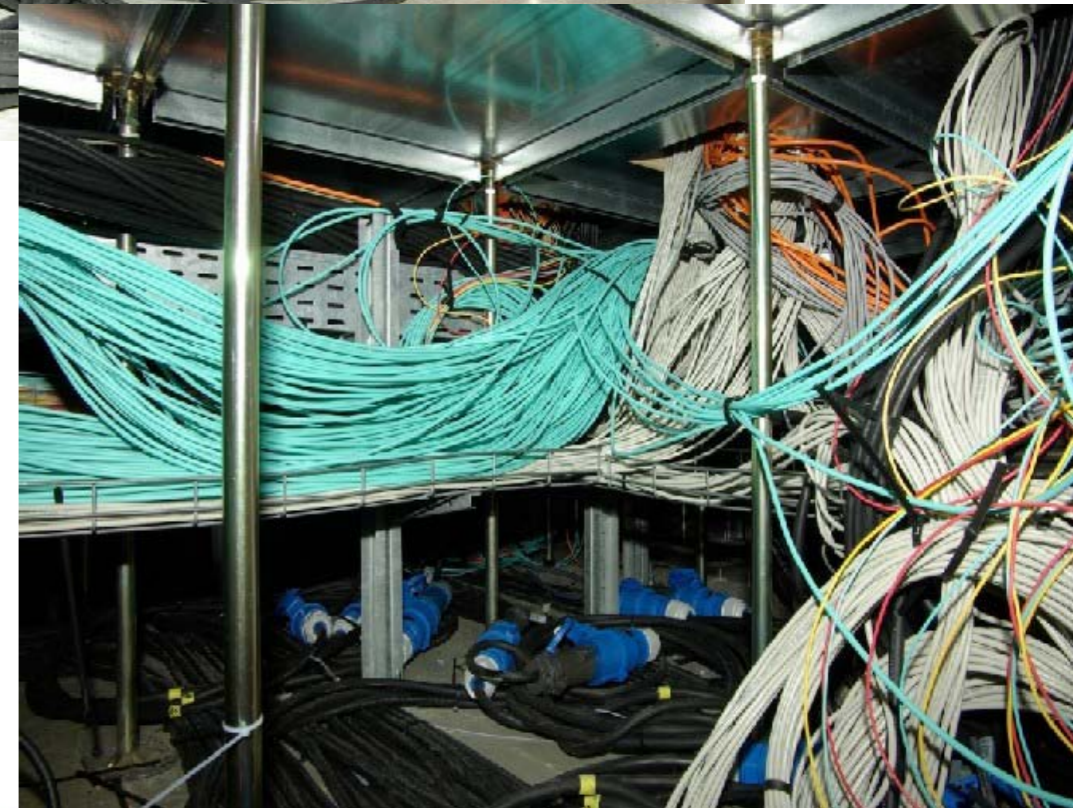
Experiences , Expectations & Experiments (LP)

Conclusions (LP)

IB bring-up – cabling impressions



IB bring-up – cabling impressions



Some experiences

System was not expected to scale during first phase

- Besides Linpack
- We expected OS jitter problems – to be solved with co-development projects.

We were conservative

- Don't allow jobs with more than 512 nodes (4096 core)

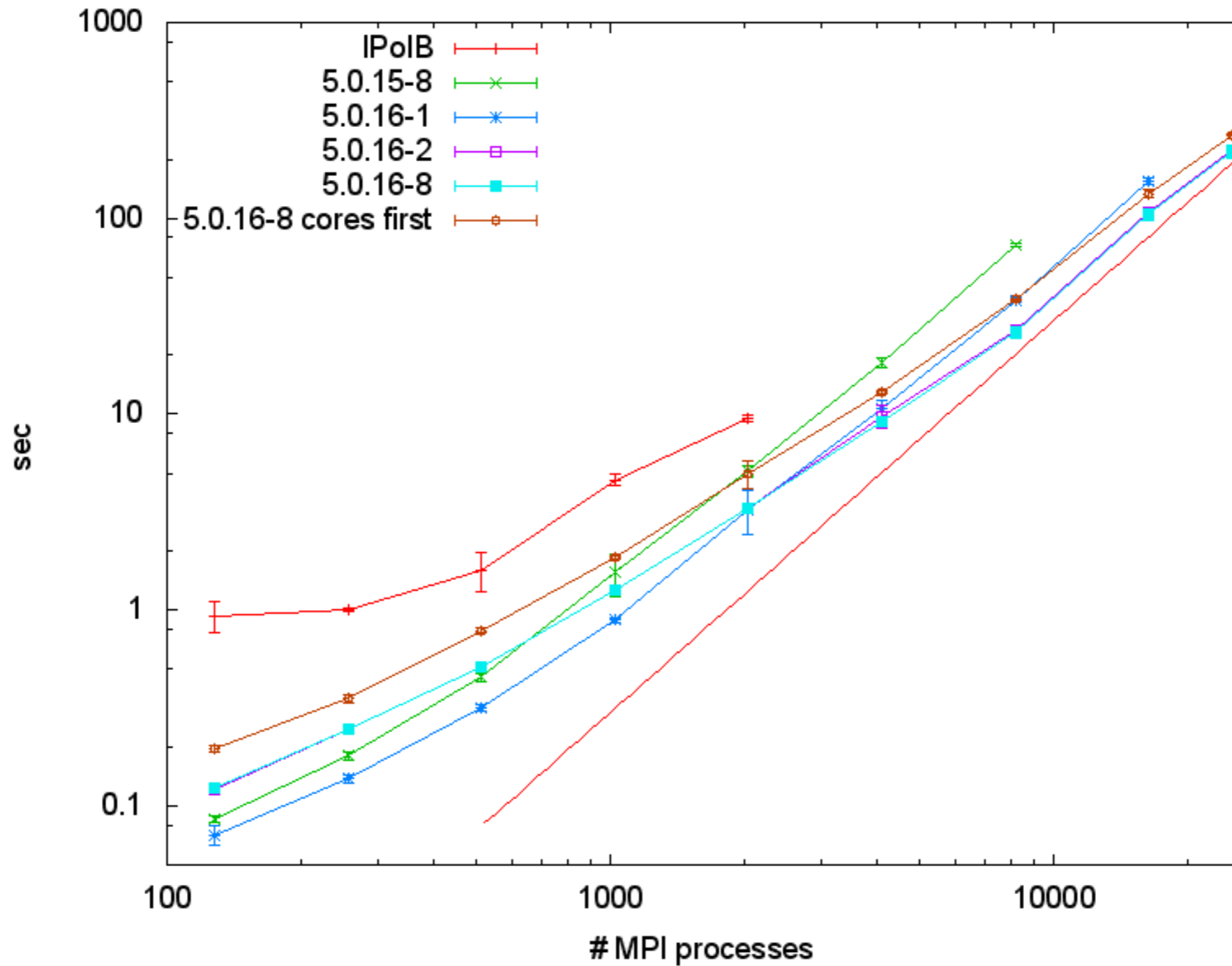
Nevertheless, users are impatient and adventurous

- Push applications to the allowed limits
- Try to scale beyond this limits
- Investigate scalability

Let's look at some examples

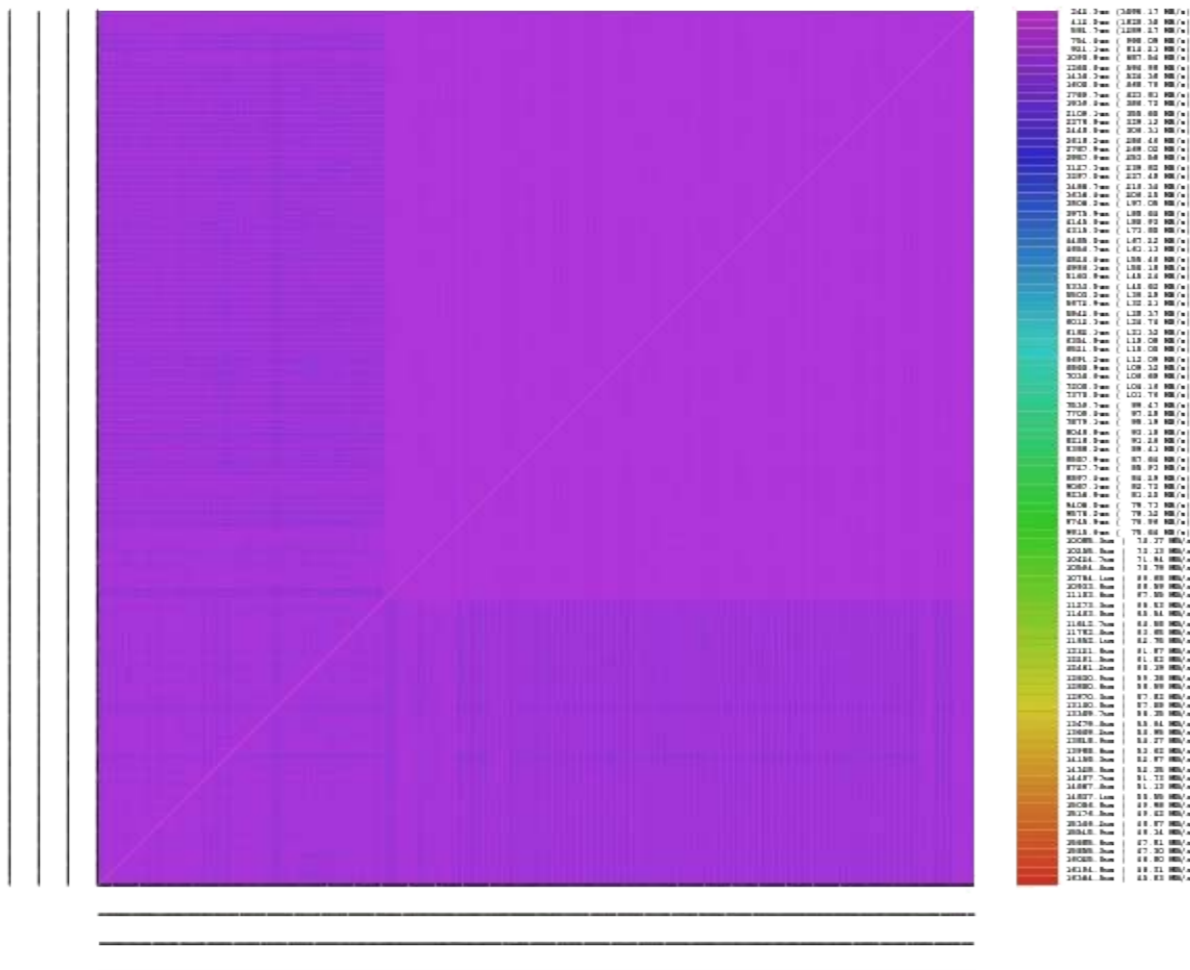
- Synthetic benchmarks (link-test)
- Real-world applications

Application startup results

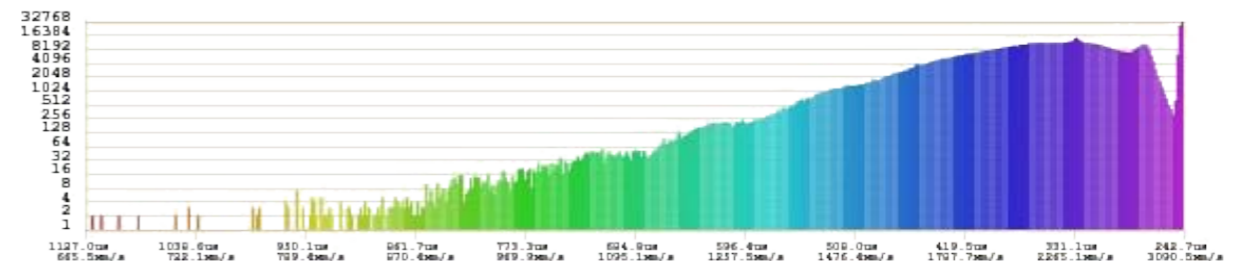
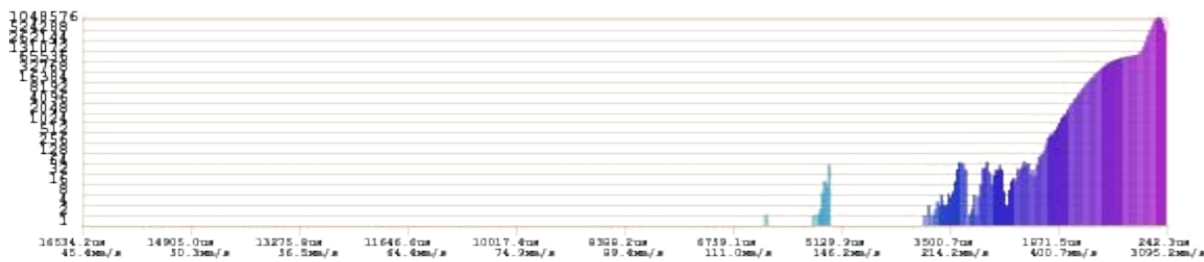
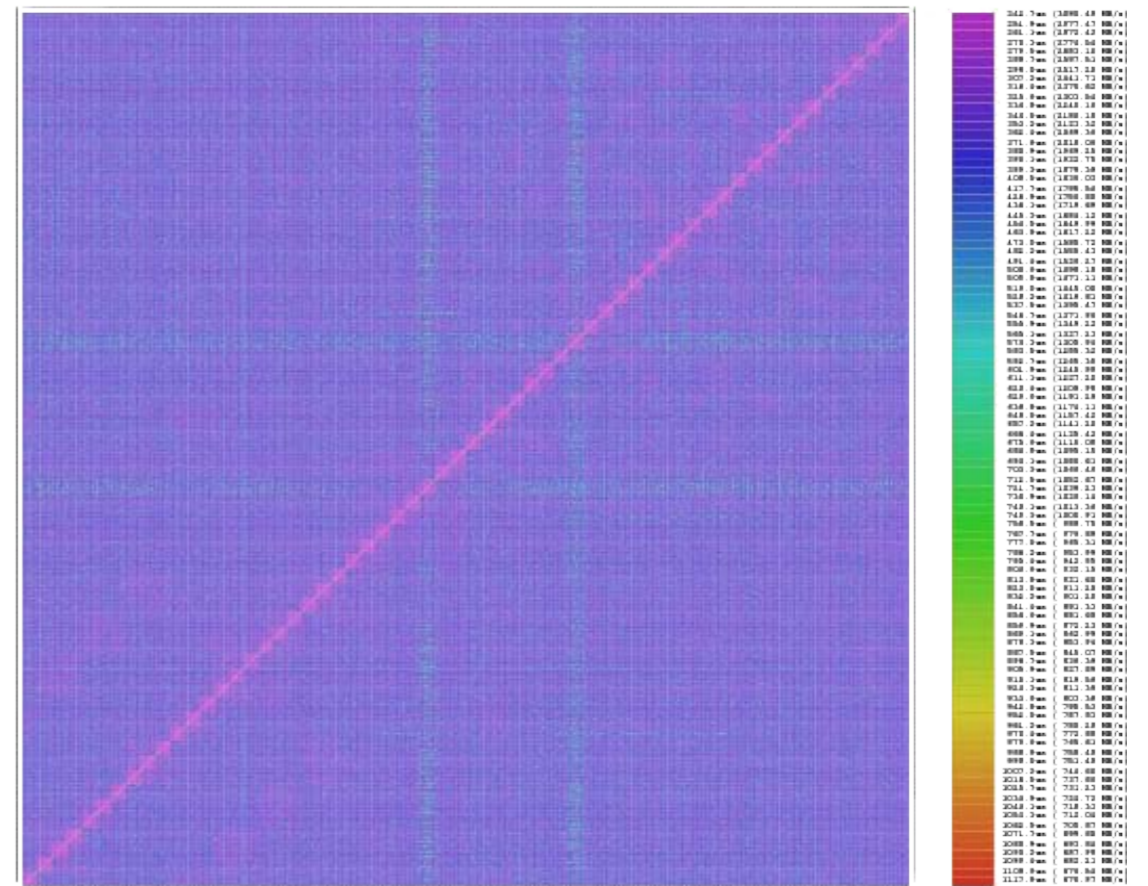


Link-test – by Wolfgang Frings - JSC

pingpong_results_bin.sion



pingpong_results_bin.sion



length_of_message:	786432 bytes (768.00 KBytes)	number_of_tasks:	3230
number_of_messages:	3	Execution order:	Parallel
Alltoall:	1	Mixing PE rank:	Yes
Min Value:	242.3us (3095.17 MB/s)	Alltoall Min Value:	762.9us (1 Byte)
Max Value:	16534.2us (45.36 MB/s)	Alltoall Max Value:	24158.0us (1 Byte)
Avg Value:	484.1us (1549.22 MB/s)	Alltoall Avg Value:	1037.4us (1 Byte)

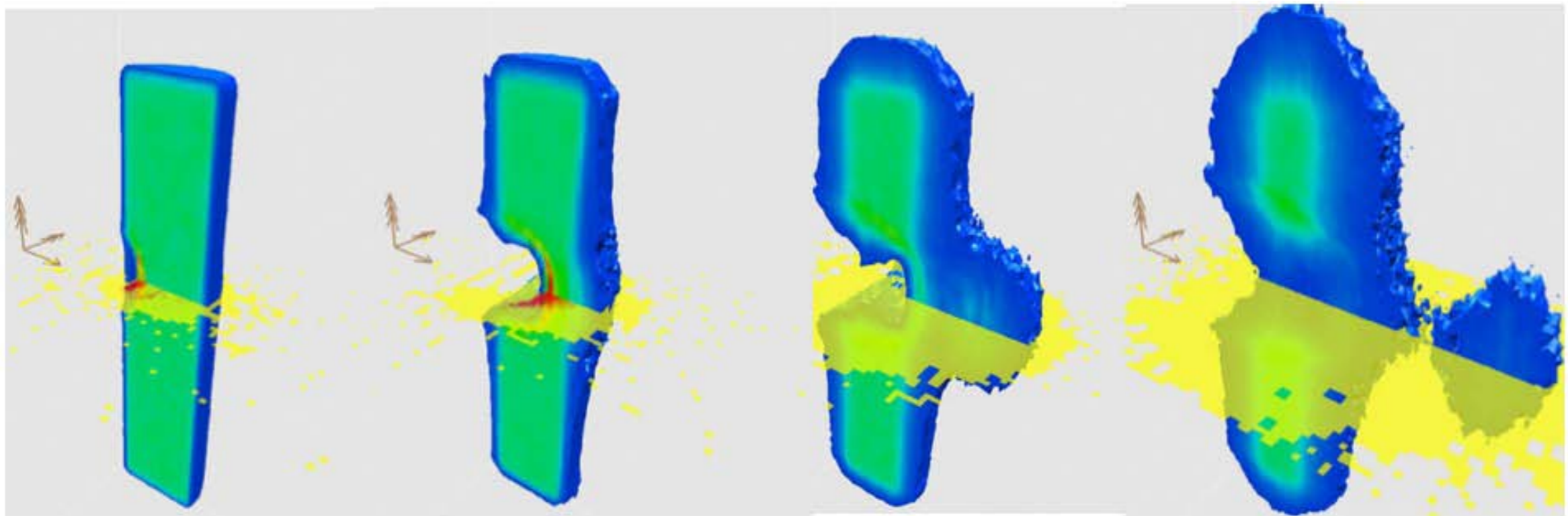
length_of_message:	786432 bytes (768.00 KBytes)	number_of_tasks:	1062
number_of_messages:	20	Execution order:	Parallel
Alltoall:	1	Mixing PE rank:	Yes
Min Value:	242.7us (3090.49 MB/s)	Alltoall Min Value:	234.8us (1 Byte)
Max Value:	1127.0us (665.48 MB/s)	Alltoall Max Value:	14258.9us (1 Byte)
Avg Value:	359.9us (2083.83 MB/s)	Alltoall Avg Value:	343.9us (1 Byte)

Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Jülich GmbH

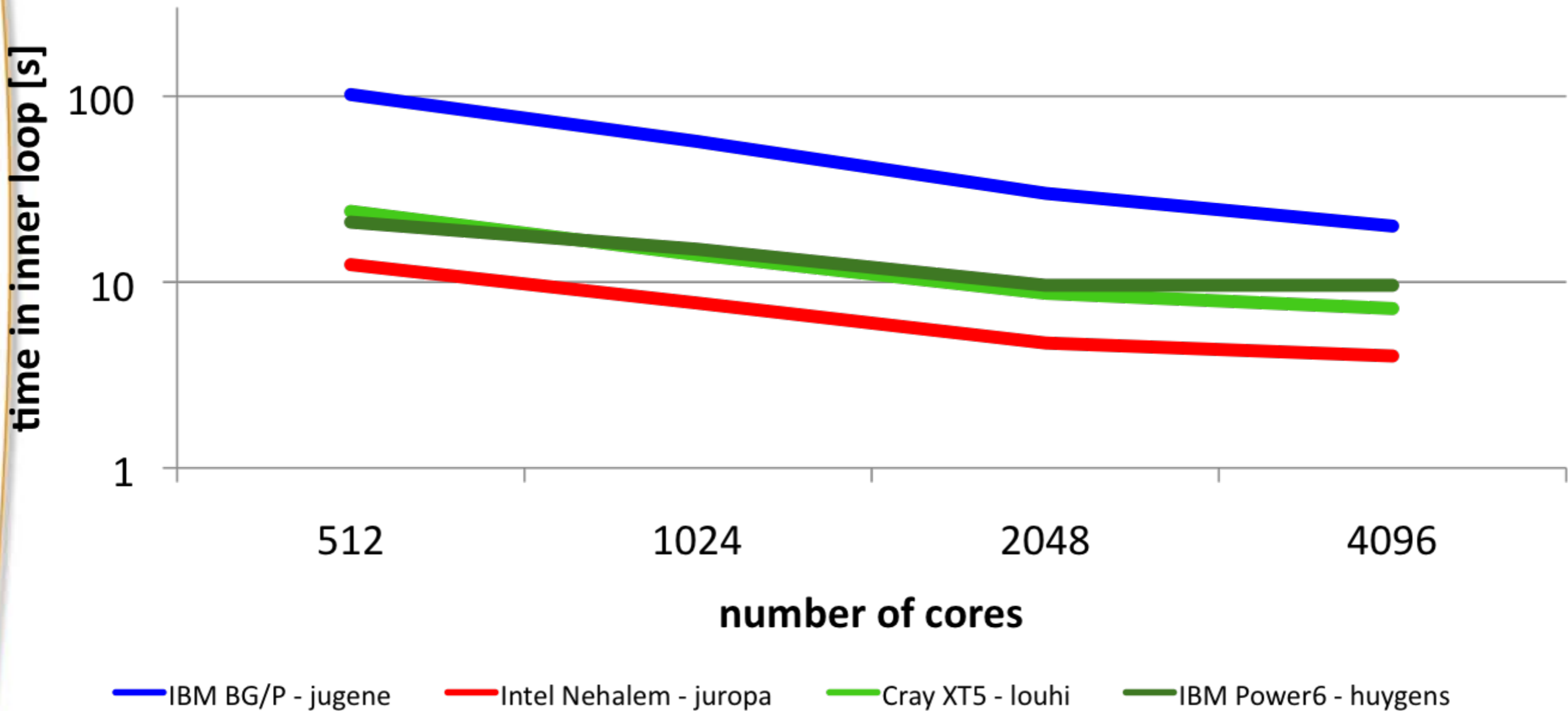
Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Jülich GmbH

PEPC

Physical Plasma N-body code
Simulation of laser-plasma interaction
Complexity reduction from $O(N^2)$ to $O(N \log(N))$
Scale up to 8k cores
Developed by P. Gibbon (JSC)



PEPC performance

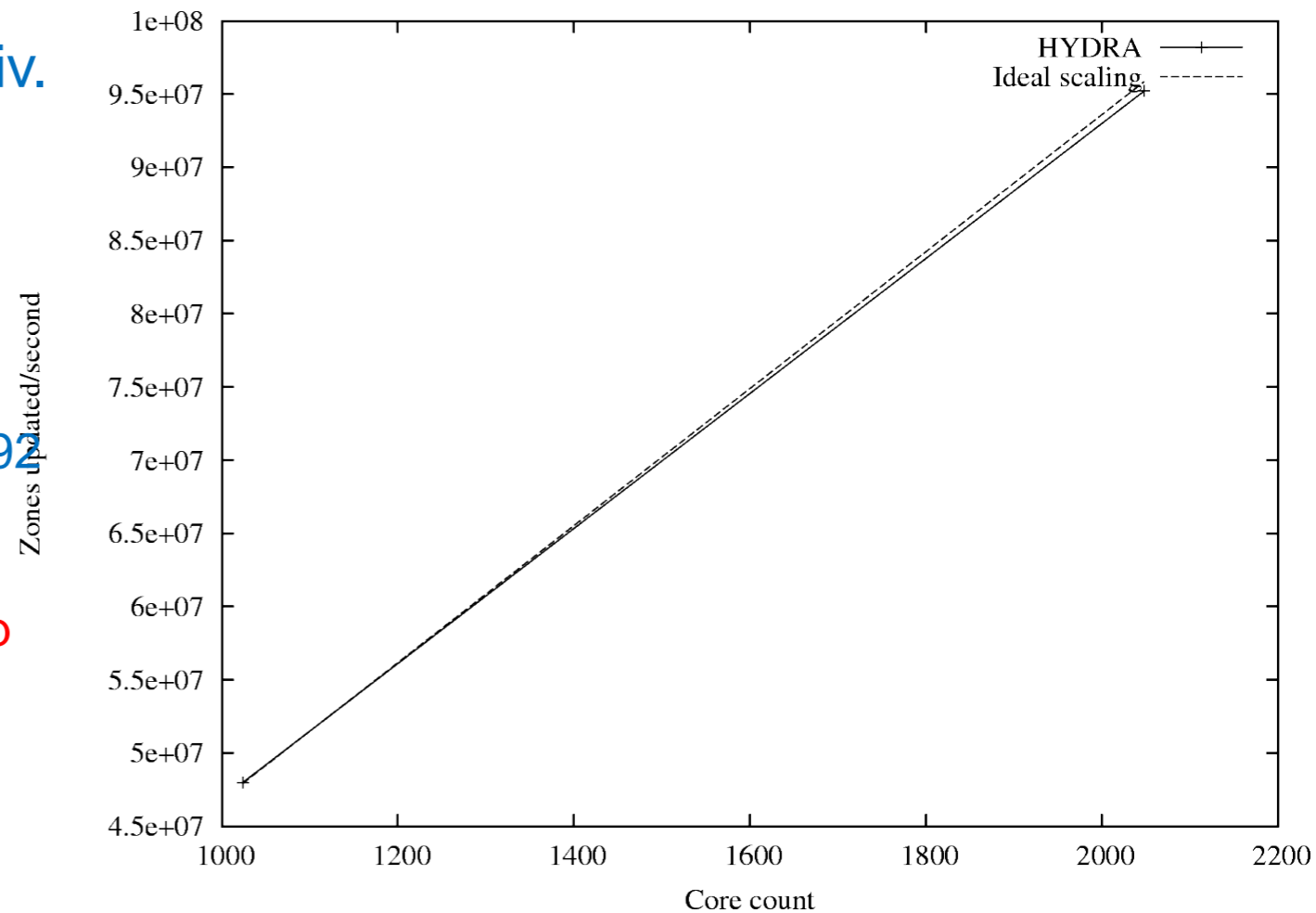


HYDRA – a astrophysics fluid code (PRACE)

T. Downes, M. Browne – Dublin City Univ.

Three prototypes were examined as part of this project: the Cray XT5 at CSC, and the JUROPA and JUGENE systems at FZJ.

HYDRA was shown to scale remarkably well on JUGENE – exhibiting strong scaling [...] from 8192 cores up to the full system (294912 cores) with approximately 73% efficiency. HYDRA **also exhibited excellent strong scaling on JUROPA up to 2048 cores but less impressive scaling on the XT5 system.**



Machine (cores)	JUGENE (294912)	JuRoPA (2048)	LOUHI (1440)
cost-to-solution kW h / time-step	0.52	0.38	0.57

The work of ParTec on JuRoPA (HF)

Overview of JuRoPA (HF)

Basics of scalability (LP)

Experiences , Expectations & Experiments (LP)

Conclusions (LP)

Conclusions

JuRoPA already has some ingredients to assure scalability

- Good all round InfiniBand topology for throughput cluster
- Scalable services
- ParaStation MPI

Some room for improvement

- Progress co-development – OS Jitter resolution
- MPI connections based on UD – reduced memory footprint (ParTec development- in Beta testing)
- Implement Adaptive routing
- Implement collective offloading

Some very good results already

- Many applications scale well – at least better than Cray XT5
- JuRoPA greener than JUGENE for the HYDRA code

Thank you !

www.par-tec.com