MPI RUNTIMES AT JSC, NOW AND IN THE FUTURE

Which, why and how do they compare in our systems?

08.07.2018 I MUG'18, COLUMBUS (OH) I DAMIAN ALVAREZ

MPI RUNTIMES AT JSC

Outline

- FZJ' mission
- JSC's role
- JSC's vision for Exascale-era computing
- JSC's systems
- MPI runtimes at JSC
- MPI performance at JSC systems
- MVAPICH at JSC

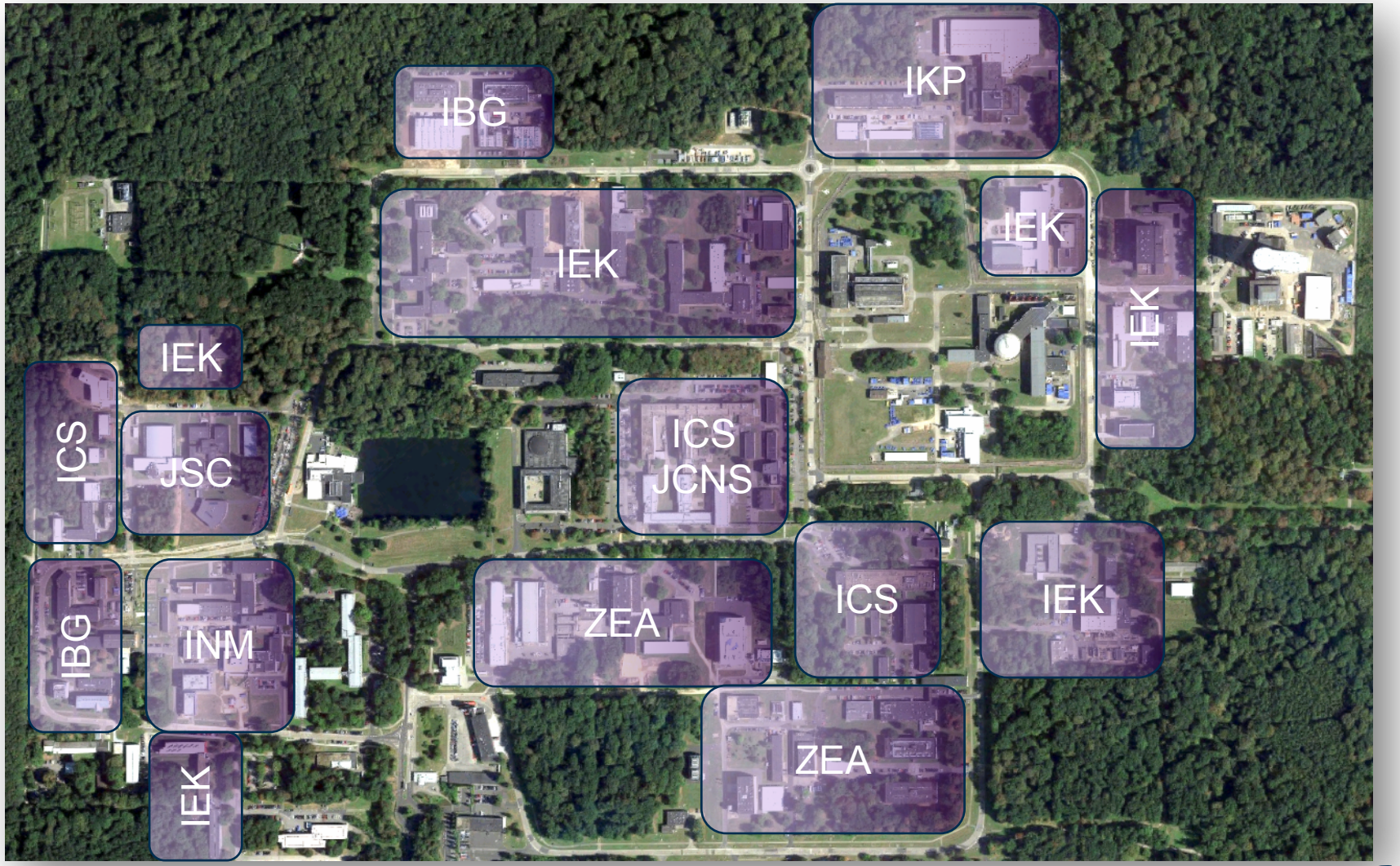
MPI RUNTIMES AT JSC

FZJ's mission

JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

FZJ's mission

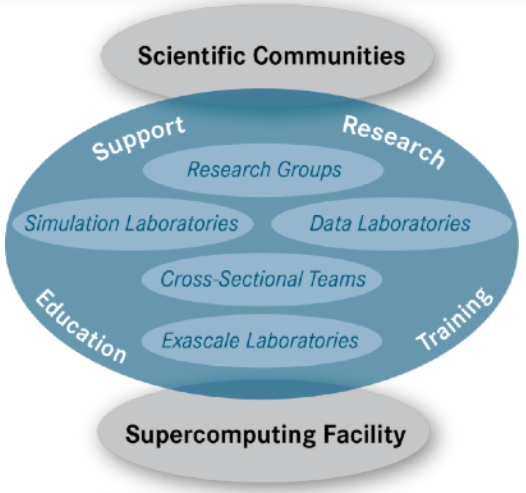


JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

JSC's role

- **Supercomputer operation for:**
  - Centre – FZJ
  - Region – RWTH Aachen University
  - Germany – Gauss Centre for Supercomputing
    John von Neumann Institute for Computing
  - Europe – PRACE, EU projects
- **Application support**
  - Unique support & research environment at JSC
  - Peer review support and coordination
- **R&D work**
  - Methods and algorithms, computational science, performance analysis and tools
  - Scientific Big Data Analytics with HPC
  - Computer architectures, Co-Design
    Exascale Labs together with IBM, Intel, NVIDIA
- **Education and training**

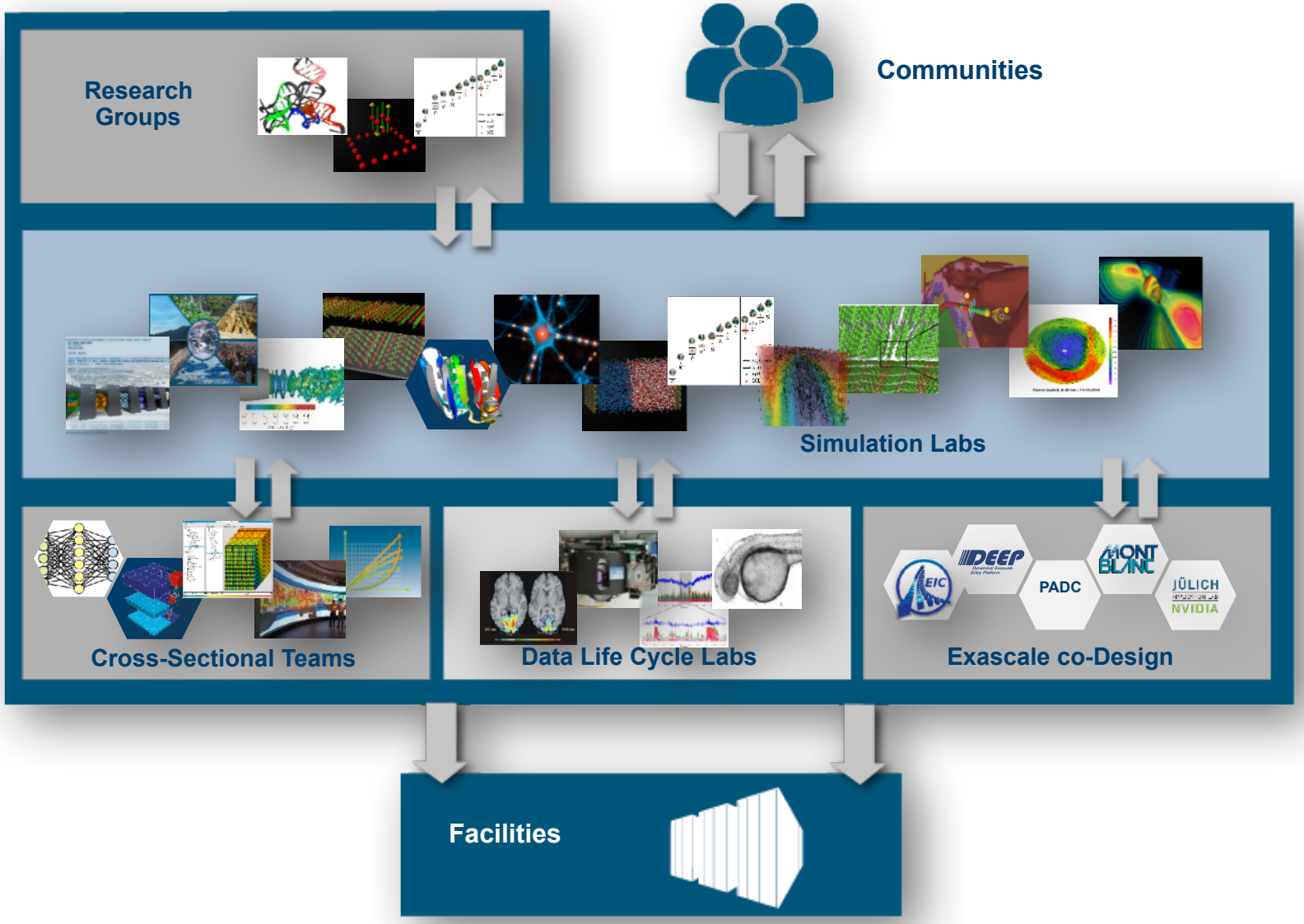


DEEP

JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

JSC's role



JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

JSC's role

Dominant Sites

#	Site	Country	Norm-HPL
1	LLNL	USA	1,466
2	LANL	USA	810
3	ORNL	USA	737
4	SNL	USA	652
5	NSCC Guangzhou	China	449
6	RIKEN AICS	Japan	354
7	**NASA/Ames**	**USA**	**352**
8	**FZ Jülich**	**Germany**	**333**
9	JAMSTEC	Japan	325
10	**NERSC**	**USA**	**303**
11	**ANL**	**USA**	**298**

Steady Sites

Never #1

TOP500

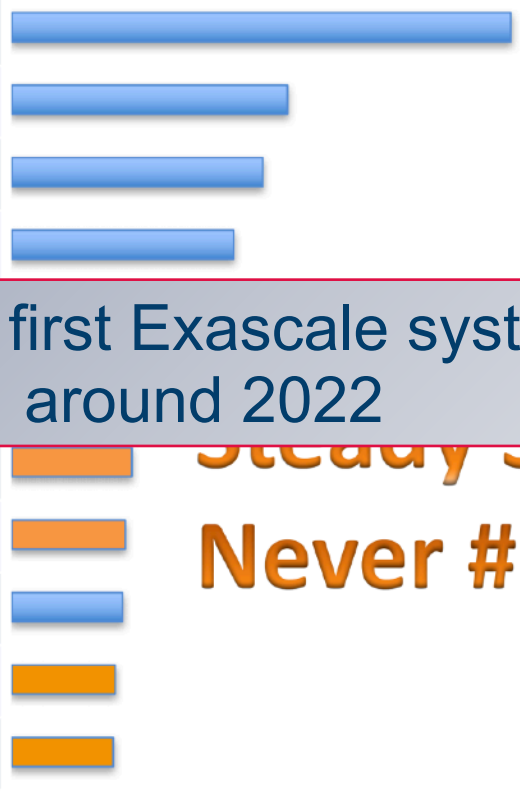
Source: Jack Dongarra, 30 Years of Supercomputing: History of TOP500, February 2018

MPI RUNTIMES AT JSC

JSC's role

Dominant Sites

#	Site	Country	Norm-HPL
1	LLNL	USA	1,466
2	LANL	USA	810
3	ORNL	USA	737
4	SNL	USA	652
5	NSCC Guangzhou		
6	RIKEN AICS		
7	**NASA/Ames**	**USA**	**352**
8	**FZ Jülich**	**Germany**	**333**
9	JAMSTEC	Japan	325
10	**NERSC**	**USA**	**303**
11	**ANL**	**USA**	**298**



JSC will host the first Exascale system in the EU around 2022

Steady Sites

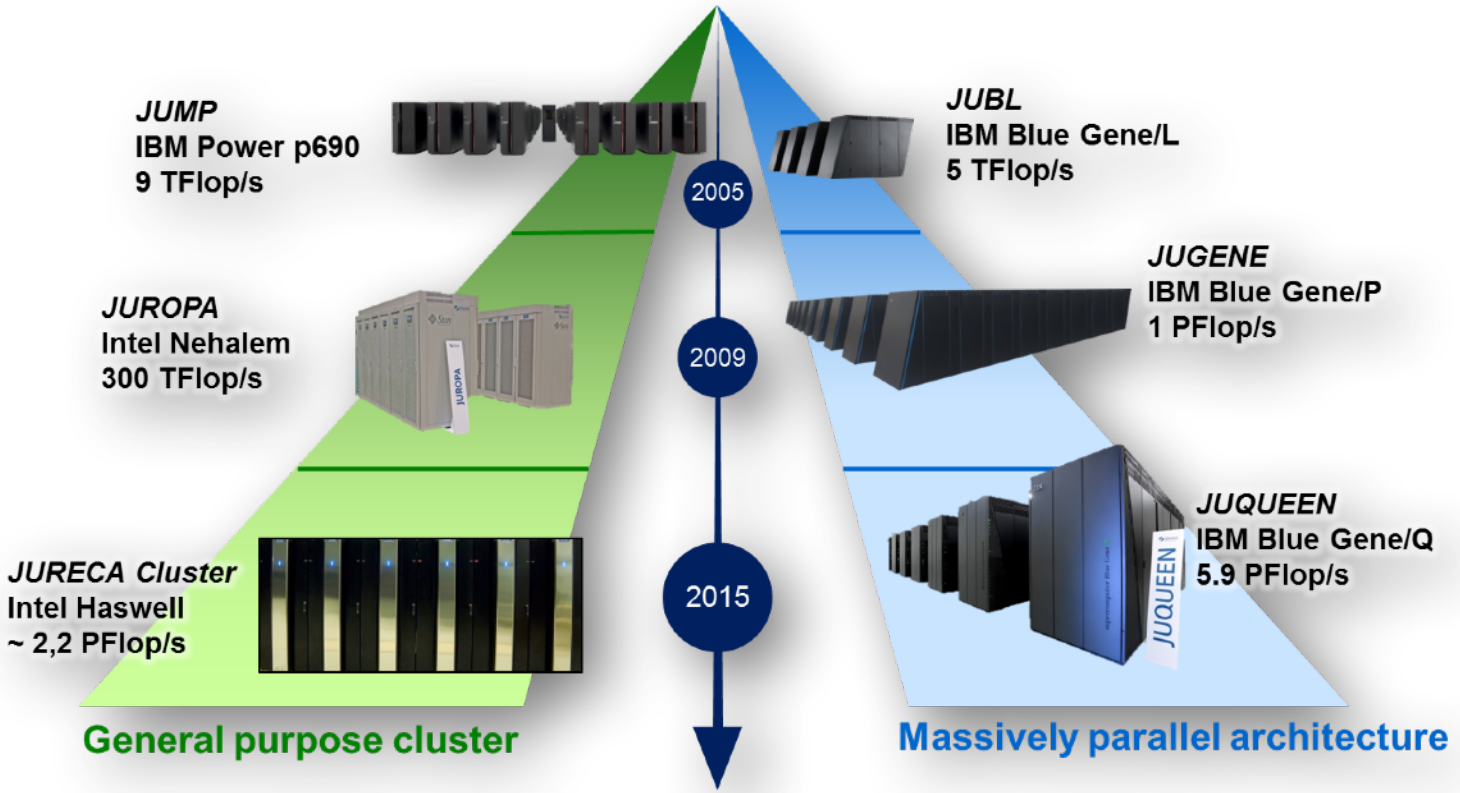
Never #1

TOP500

Source: Jack Dongarra, 30 Years of Supercomputing: History of TOP500, February 2018

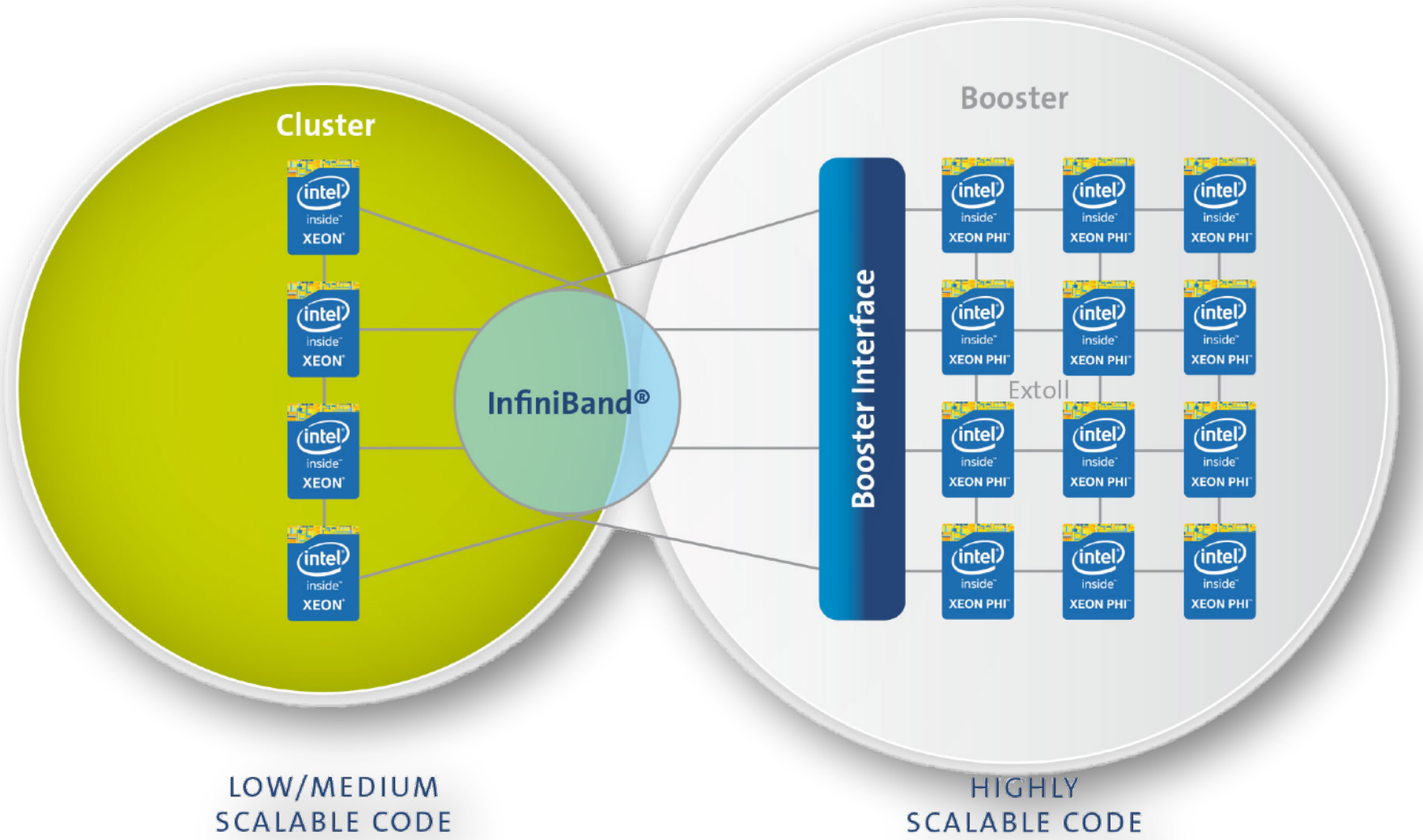
MPI RUNTIMES AT JSC

JSC's vision



MPI RUNTIMES AT JSC

JSC's vision

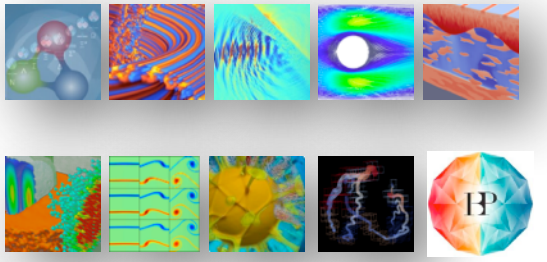


JÜLICH Forschungszentrum

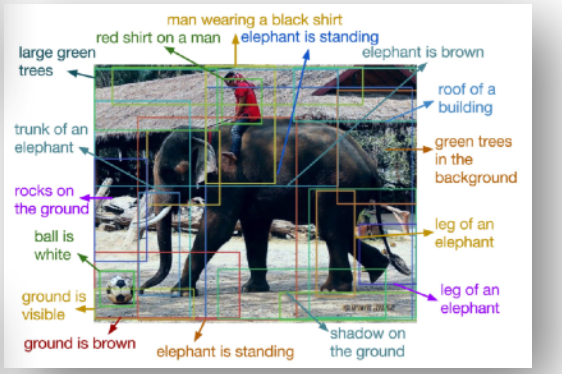
MPI RUNTIMES AT JSC

JSC's vision

Extreme Scale Computing



Big Data Analytics

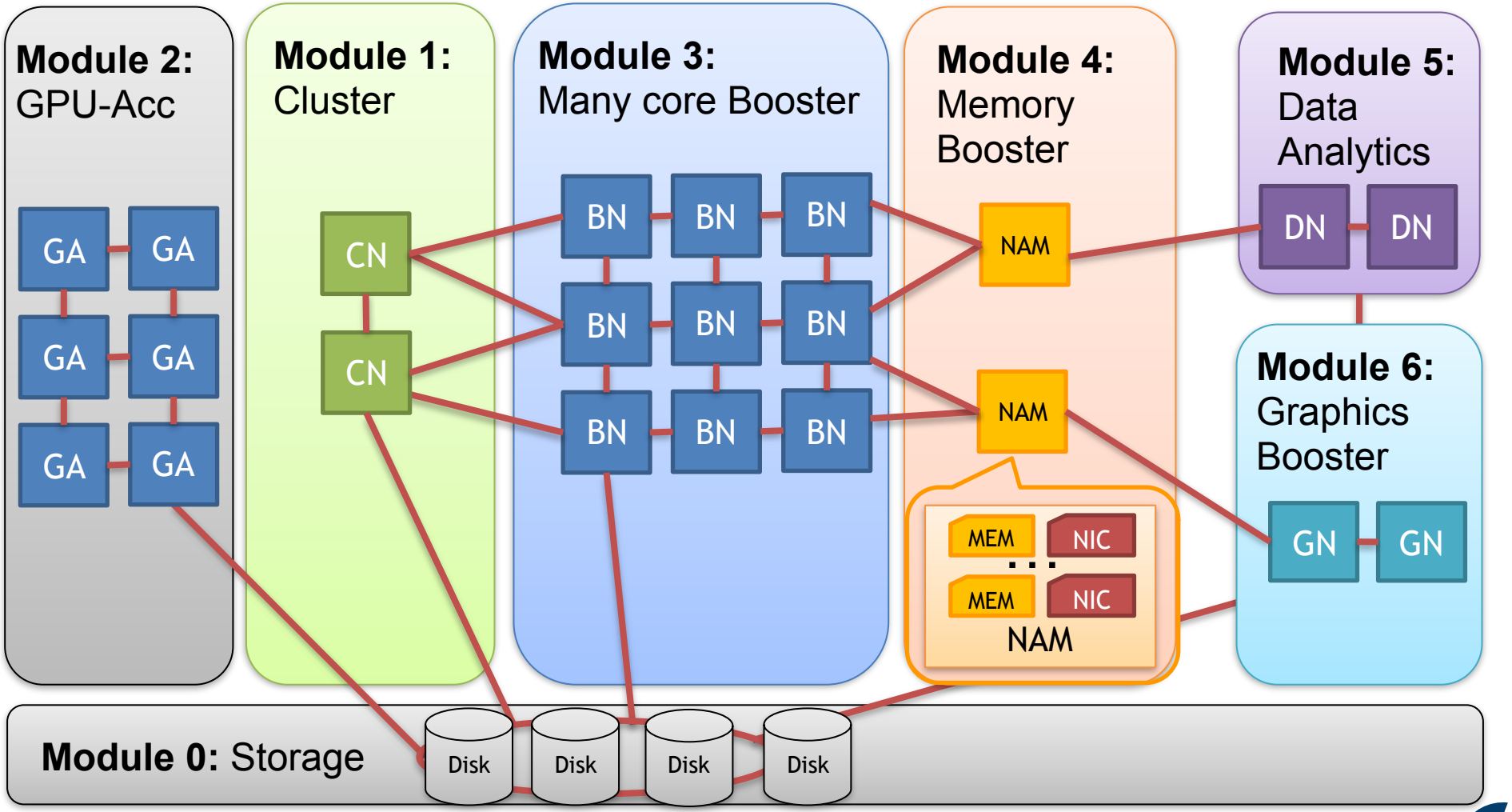


Deep Learning

Interactivity

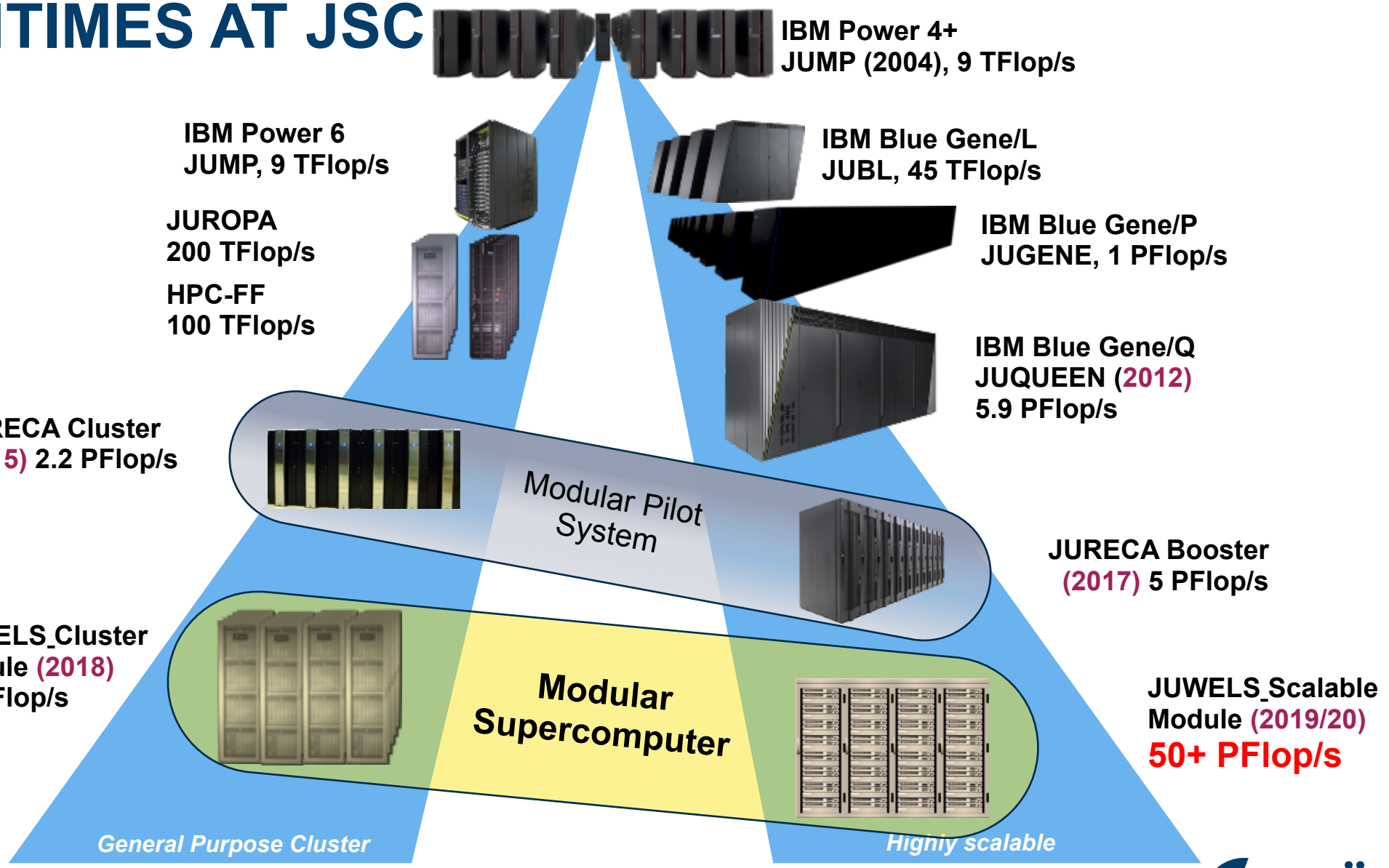
MPI RUNTIMES AT JSC

JSC's vision



MPI RUNTIMES AT JSC

JSC's vision



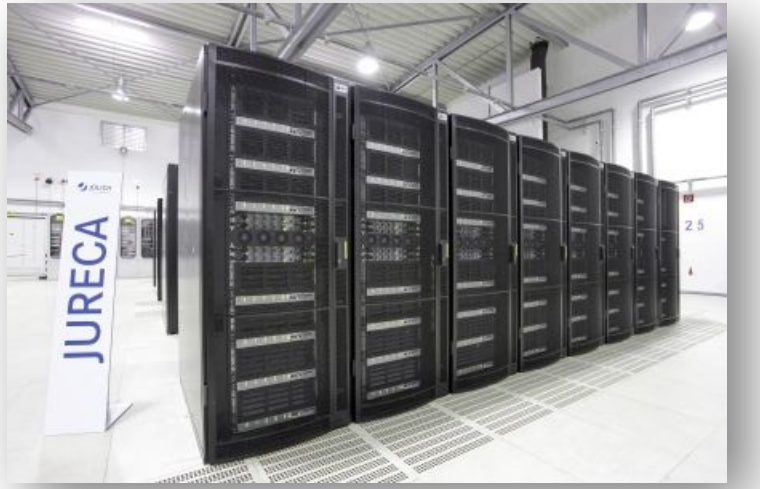
JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

JSC's systems

JURECA

- Dual-socket Intel Haswell (E5-2680 v3)
  - 12× cores/socket
  - 2.5 GHz
  - ≥ 128 GB main memory
- 1,884 compute nodes (45,216 cores)
  - 75 nodes: 2× K80 NVIDIA GPUs
  - 12 nodes: 2 × K40 NVIDIA GPUs
    512 GB main memory
- Peak performance: **2.2 Petaflop/s** (1.7 w/o GPUs)
- Mellanox InfiniBand EDR
- Connected to the GPFS file system on JUST
  - ~15 PByte online disk and
  - 100 PByte offline tape capacity

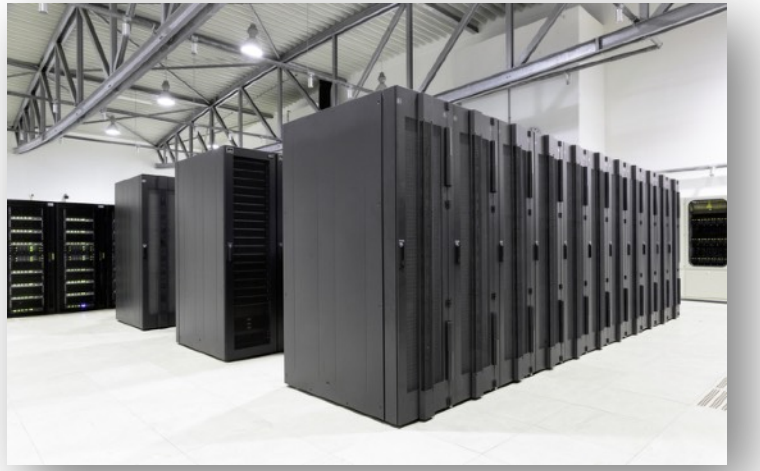


MPI RUNTIMES AT JSC

JSC's systems

JURECA Booster

- Intel Knights Landing (7250-F)
  - 68× cores
  - 1.4 GHz
  - 16 + 64 GB main memory
- 1,640 compute nodes (111,520 cores)
- Peak performance: **~5 Petaflop/s**
- Intel OmniPath
- Connected to GPFS



+JURECA

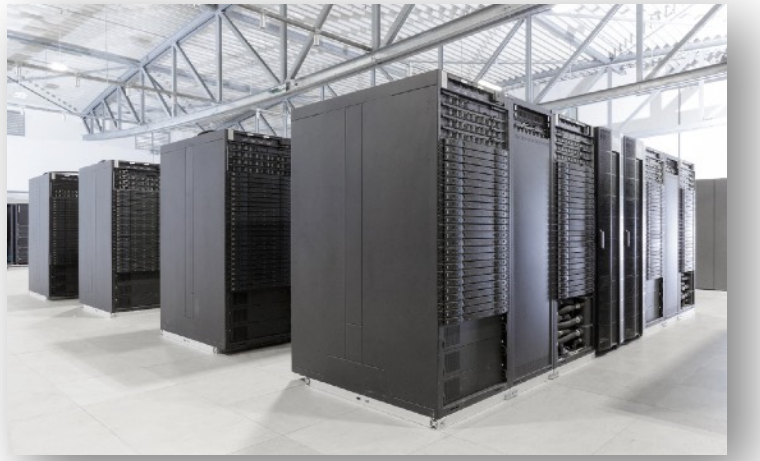
June 2018:
#14 in Europe
#38 worldwide
#66 in Green500

MPI RUNTIMES AT JSC

JSC's systems

JUWELS

- Dual-socket Intel Skylake (Xeon Platinum 8168)
  - 24× cores/socket
  - 2.7 GHz
  - ≥ 96 GB main memory
- 2,559 compute nodes (122,448 cores)
  - 48 nodes: 4× V100 NVIDIA GPUs, 192 GB main memory
  - 4 nodes: 1× P100 NVIDIA GPUs, 768 GB main memory
- Peak performance: **12 Petaflop/s** (10.4 w/o GPUs)
- Mellanox InfiniBand EDR
- Connected to the GPFS file system on JUST

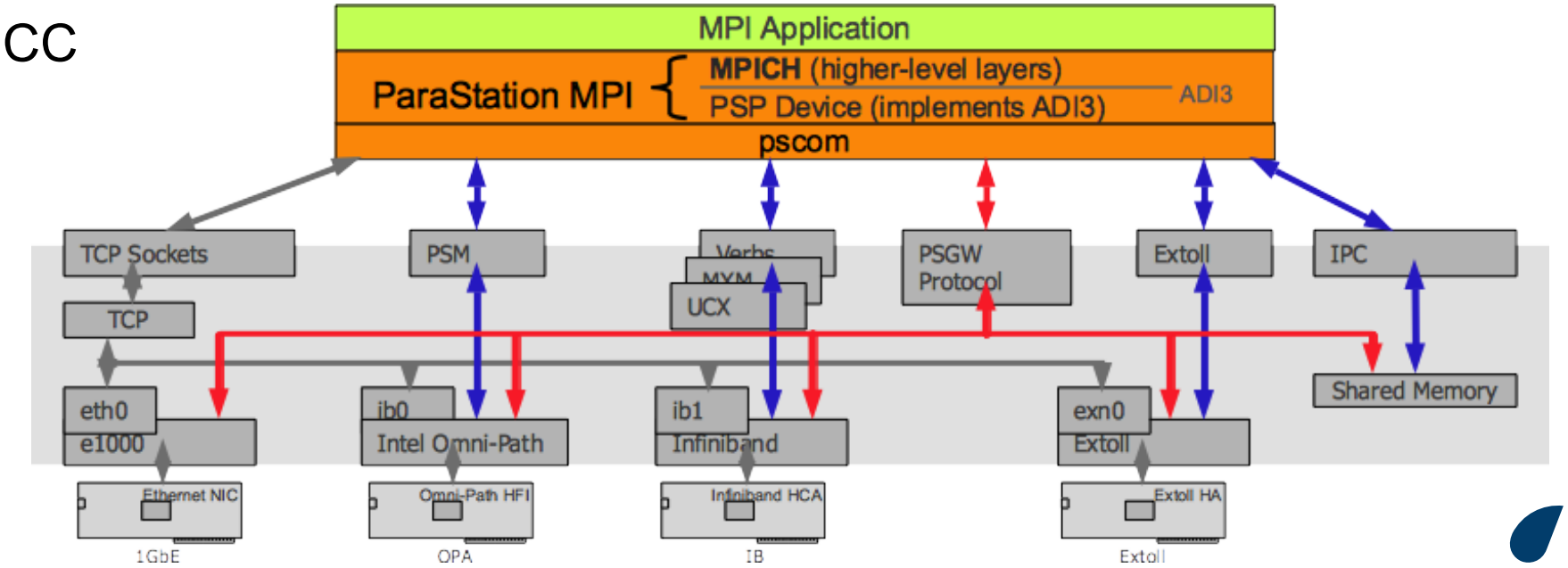


June 2018:
#7 in Europe
#23 worldwide
#29 in Green500

MPI RUNTIMES AT JSC

MPI runtimes

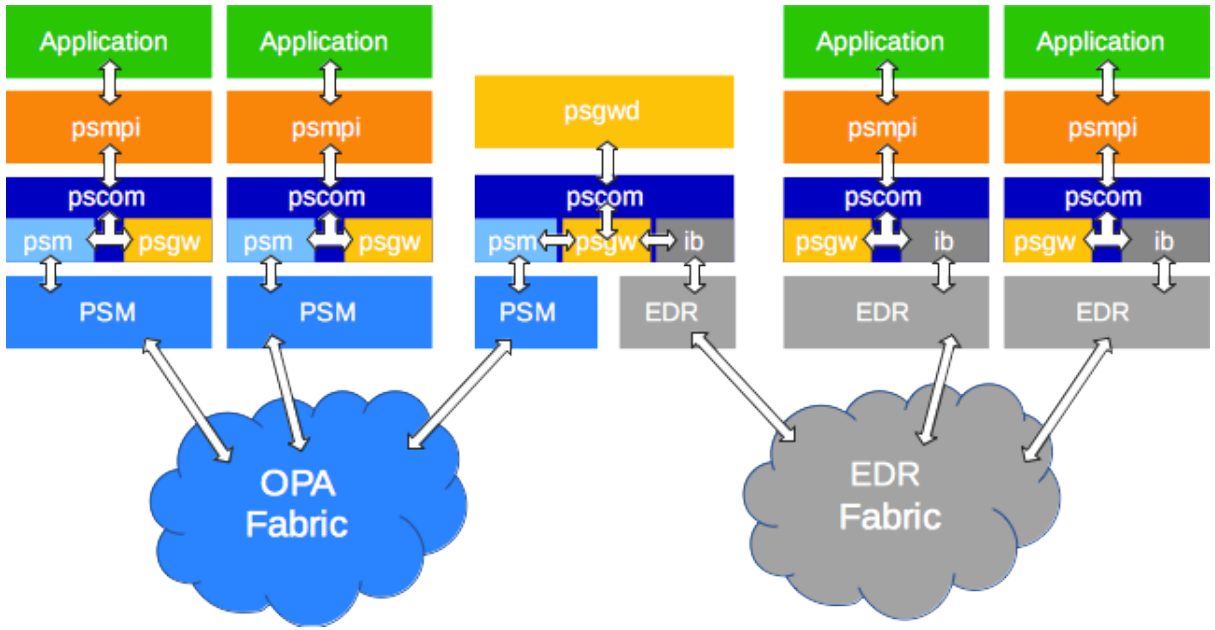
- ParaStationMPI
  - Preferred MPI runtime
  - Developed in a consortium including JSC, ParTec, KIT and University of Wüppertal
  - Based on MPICH
  - Intel + GCC



MPI RUNTIMES AT JSC

MPI runtimes

- ParaStationMPI
  - For the JURECA Cluster-Booster System, the ParaStation MPI Gateway Protocol bridges between Mellanox EDR and Intel Omni-Path
  - In general, the ParaStation MPI Gateway Protocol can connect any two low-level networks supported by *pscom*.
  - Implemented using the *psgw* plugin to pscom, working together with instances of the *psgwd*, the ParaStation MPI Gateway daemon.



MPI RUNTIMES AT JSC

MPI runtimes

- IntelMPI
  - As back up of ParaStationMPI
  - Software stack mirrored between these 2 MPI runtimes
  - Intel compiler

- MVAPICH2-GDR
  - Purely for GPU nodes (but can run in normal nodes)
  - PGI + GCC (and Intel in the past)

MPI RUNTIMES AT JSC

MPI performance

- Microbenchmarks: Intel MPI Benchmarks
  - PingPong
  - Broadcast, gather, scatter and allgather
- ParaStationMPI, Intel MPI 2018.2, Intel MPI 2019 beta, MVAPICH2 and OpenMPI
- JURECA, Booster and JUWELS
- 1 MPI process per node. Average times. No cache disabling.

MPI RUNTIMES AT JSC

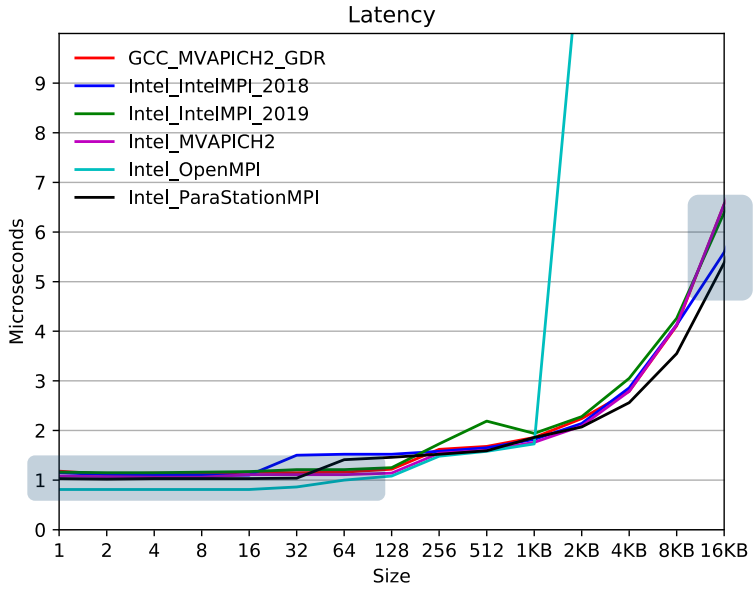
MPI performance

	Jureca	Booster	Juwels
ParaStationMPI	Verbs	PSM2	UCX
Intel 2018	Default (DAPL+Verbs)	Default (PSM2)	Default (DAPL+Verbs)
Intel 2019	Default (libfabric+Verbs)	Default (libfabric+PSM2)	Default (libfabric+verbs)
MVAPICH2-GDR	Verbs	-	-
MVAPICH2	Verbs	PSM2	Verbs
OpenMPI	Default (UCX, Verbs and libfabric+Verbs enabled)	-	-

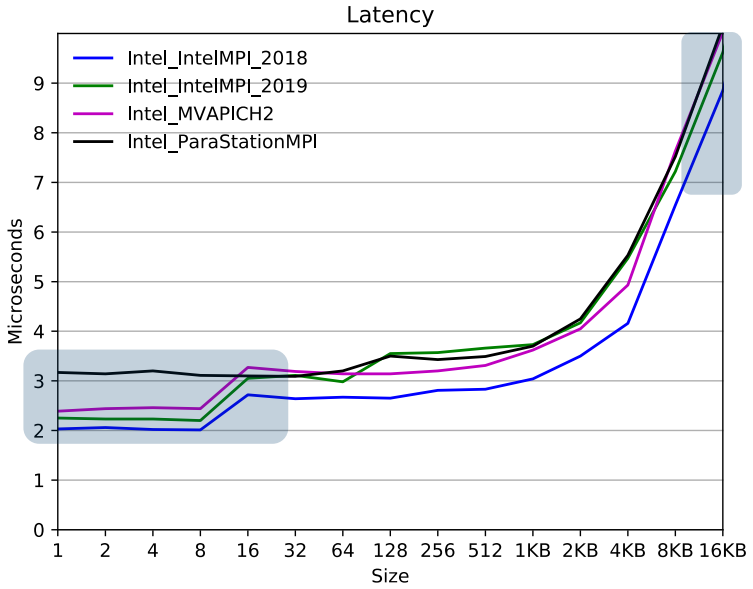
MPI RUNTIMES AT JSC

MPI performance (PingPong)

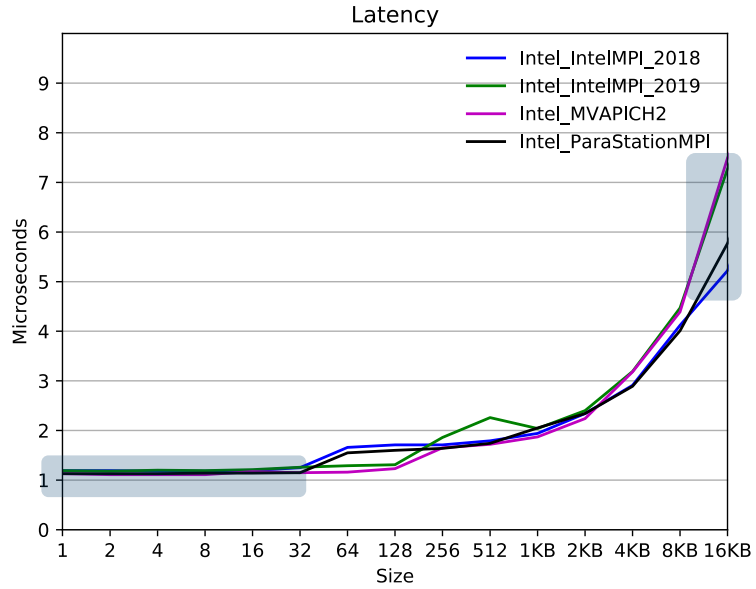
Jureca



Booster



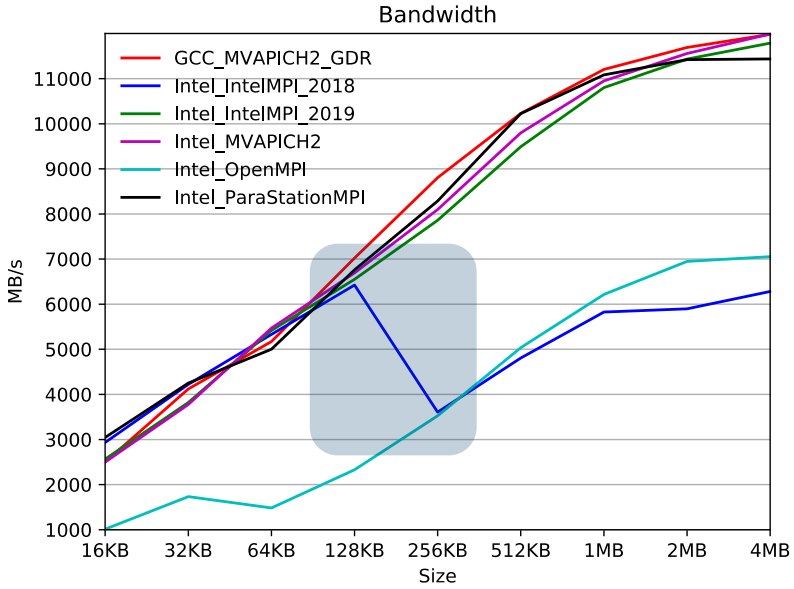
Juwels



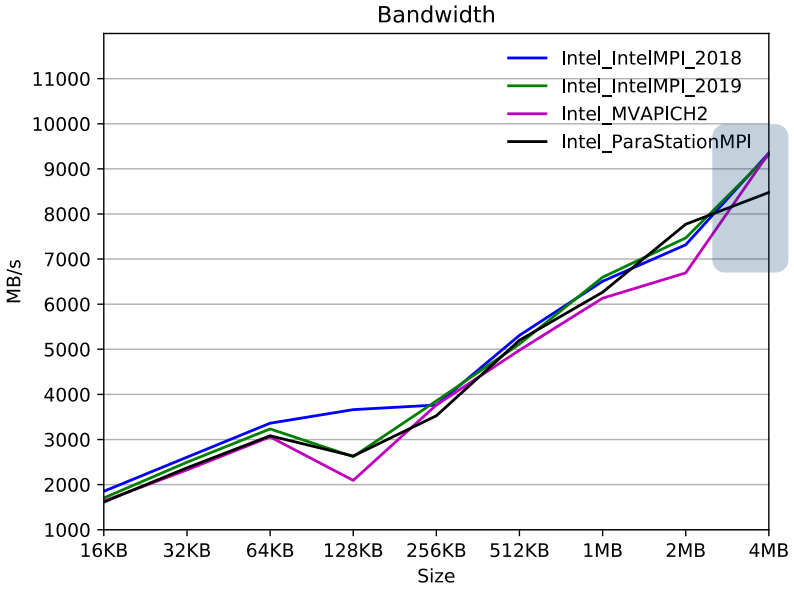
MPI RUNTIMES AT JSC

MPI performance (PingPong)

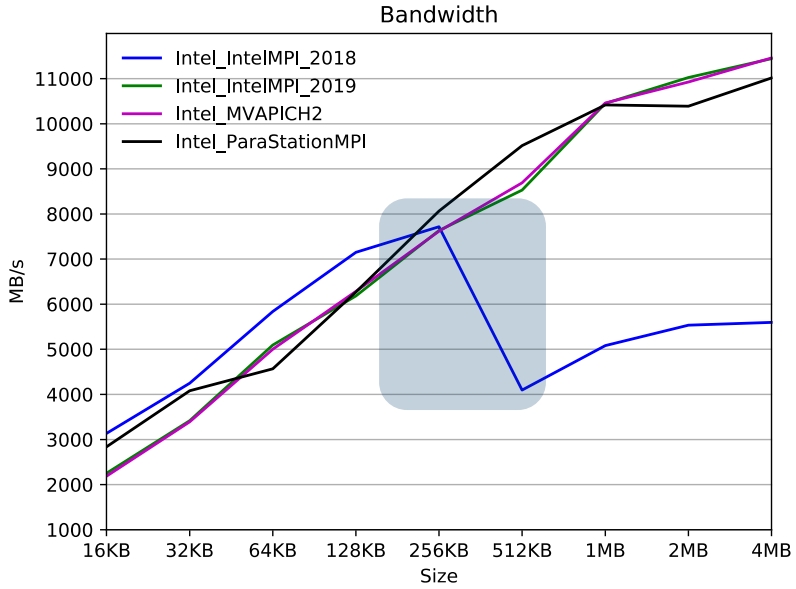
Jureca



Booster



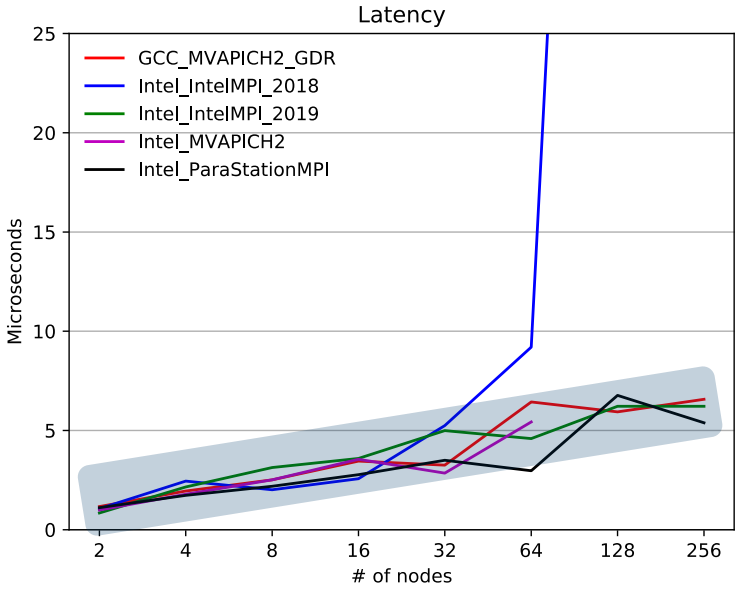
Juwels



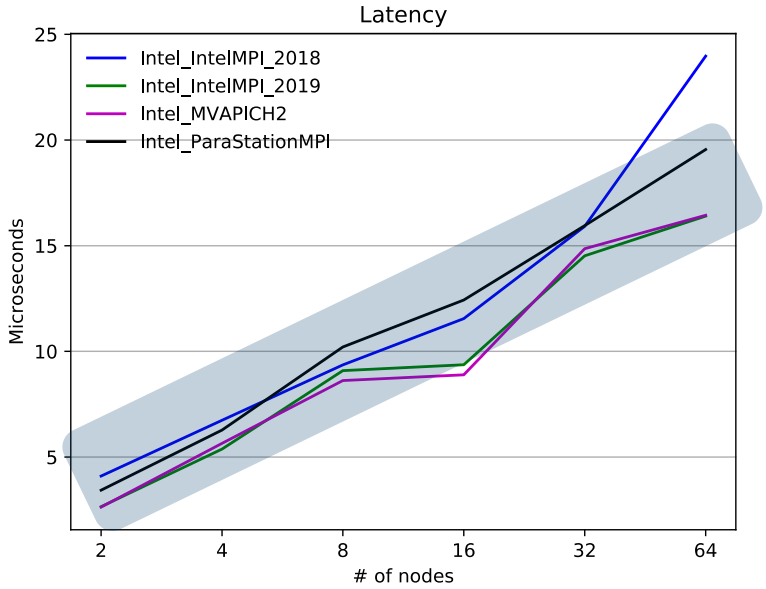
MPI RUNTIMES AT JSC

MPI performance (Broadcast, msg size 1 byte)

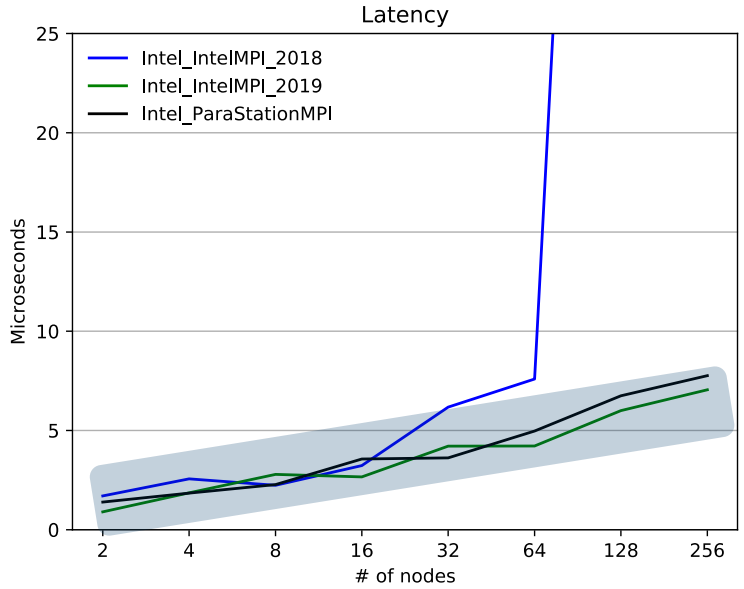
Jureca



Booster



Juwels

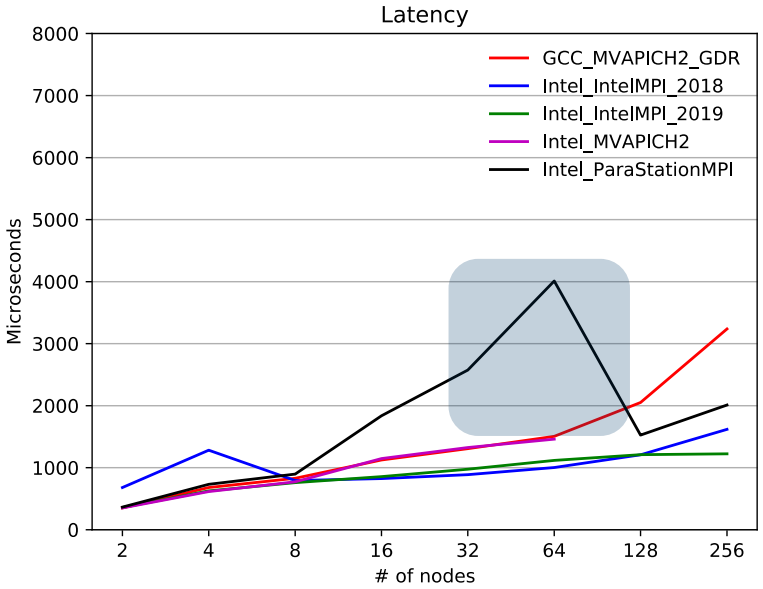


JÜLICH Forschungszentrum

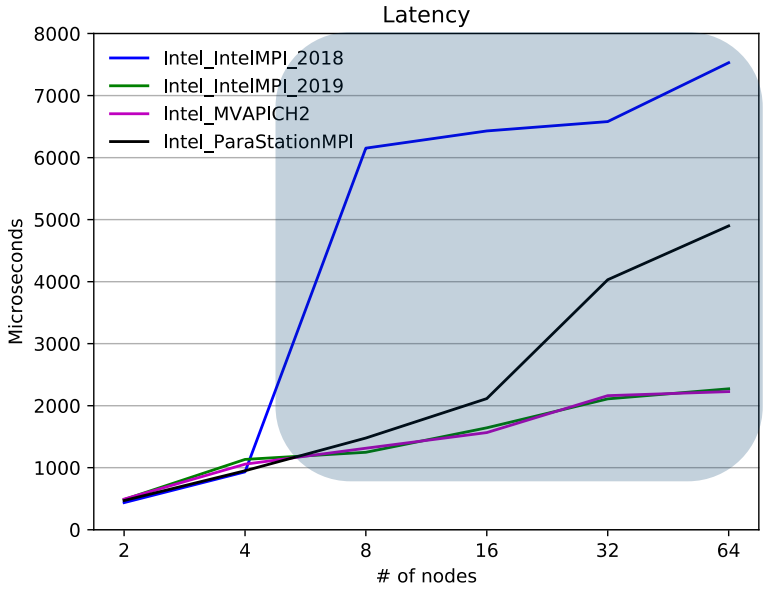
MPI RUNTIMES AT JSC

MPI performance (Broadcast, msg size 4MB)

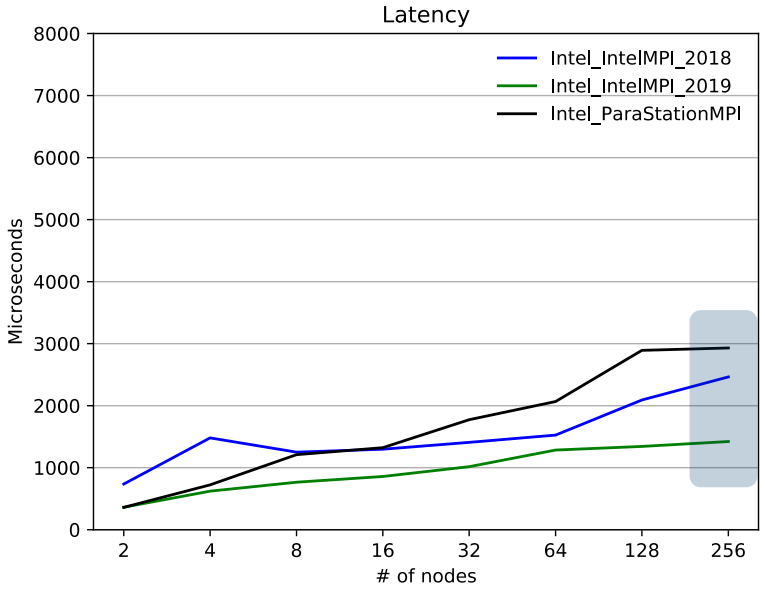
Jureca



Booster



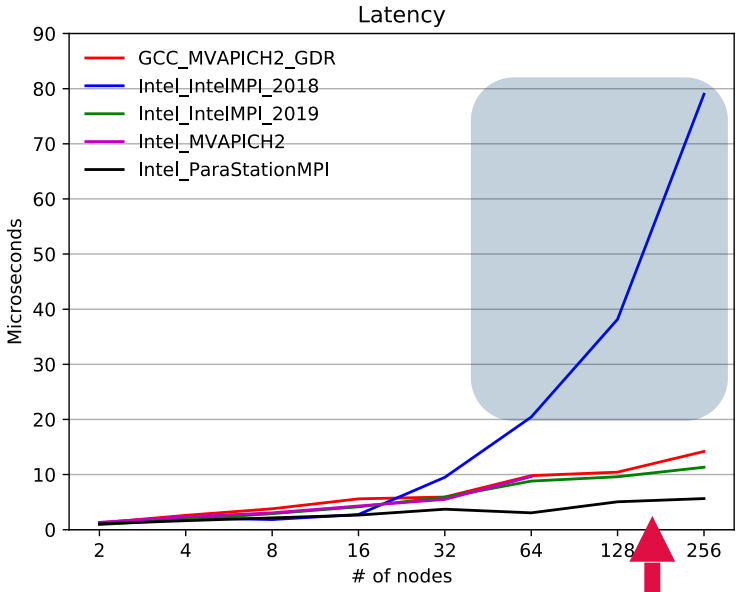
Juwels



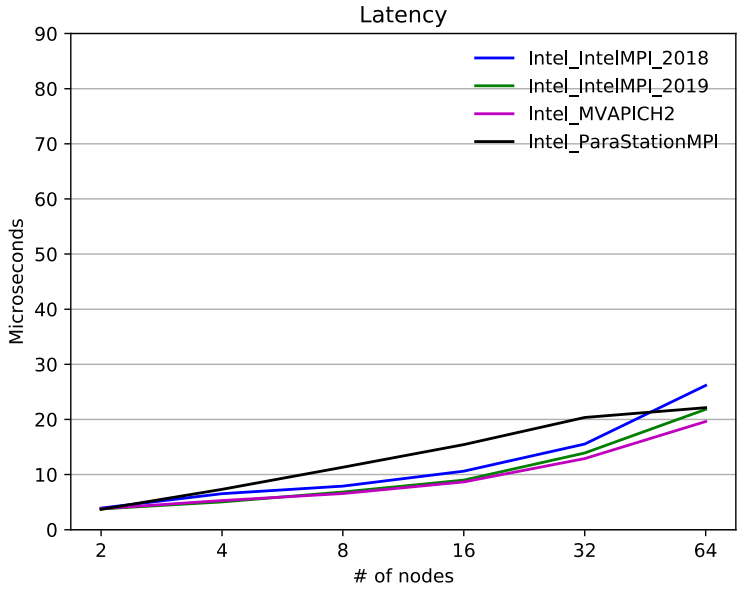
MPI RUNTIMES AT JSC

MPI performance (Scatter, msg size 1 byte)

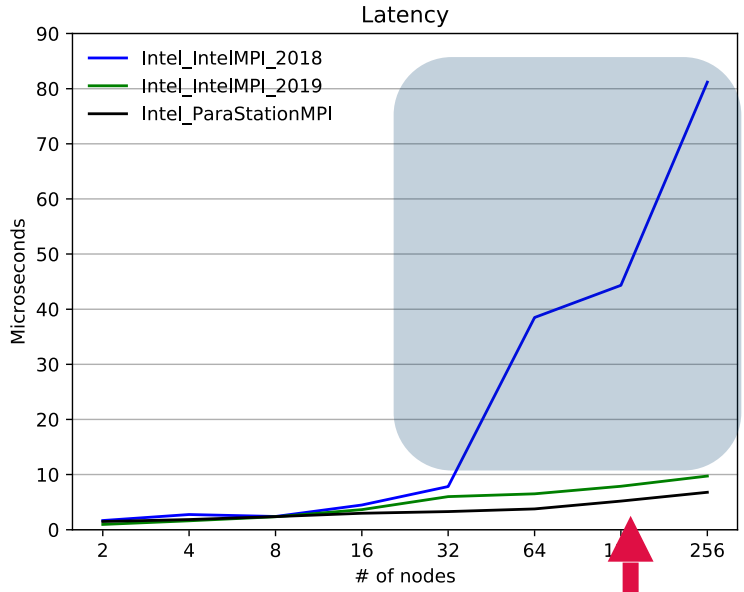
Jureca



Booster



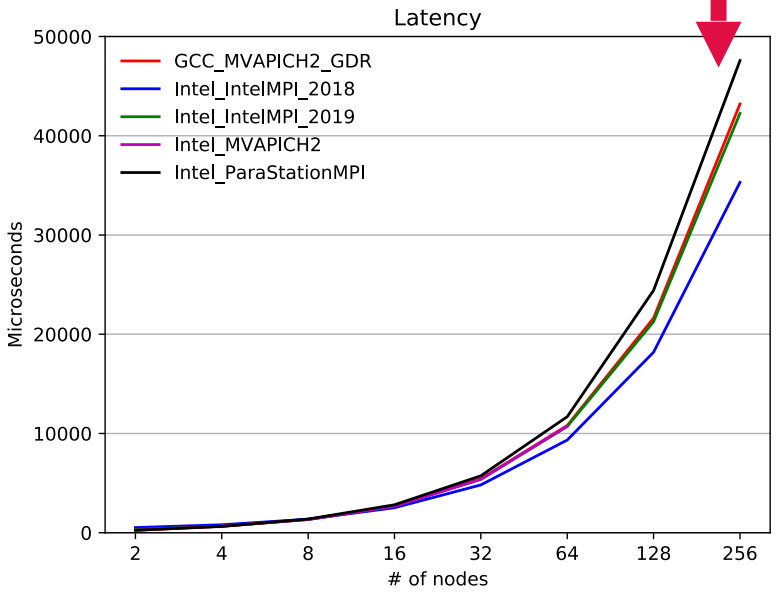
Juwels



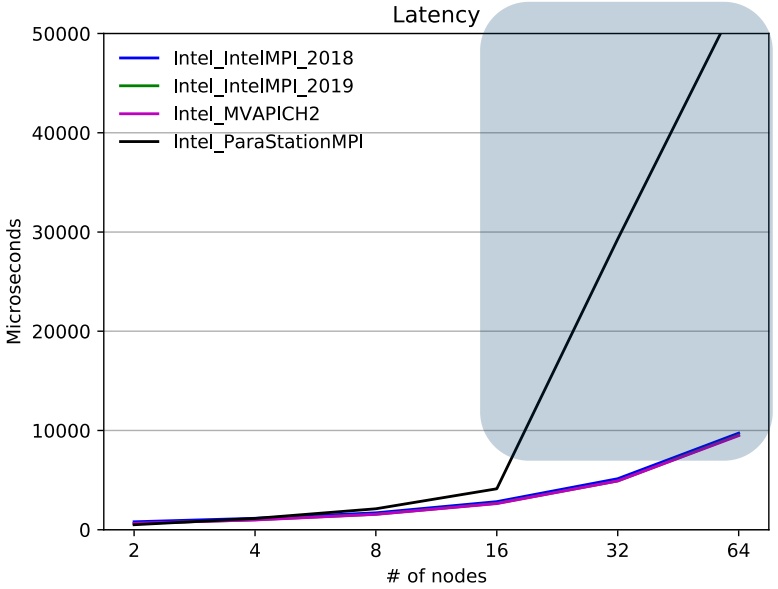
MPI RUNTIMES AT JSC

MPI performance (Scatter, msg size 2MB)

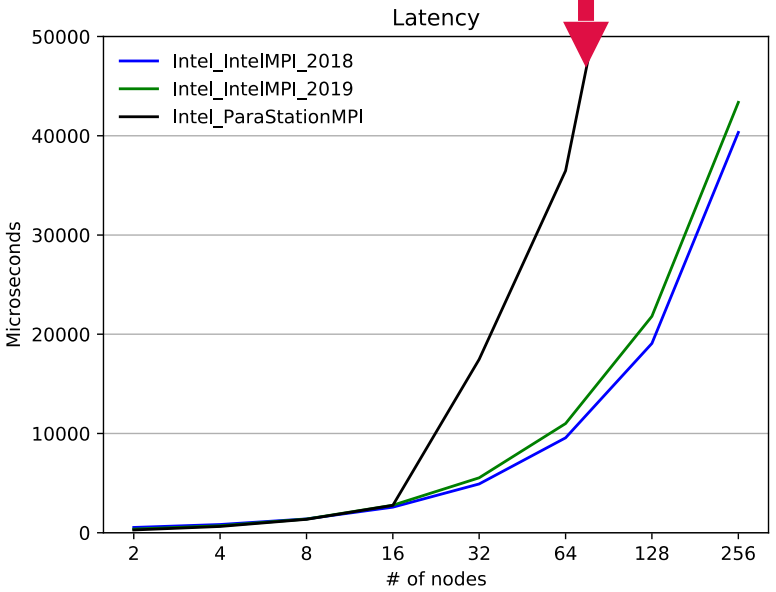
Jureca



Booster



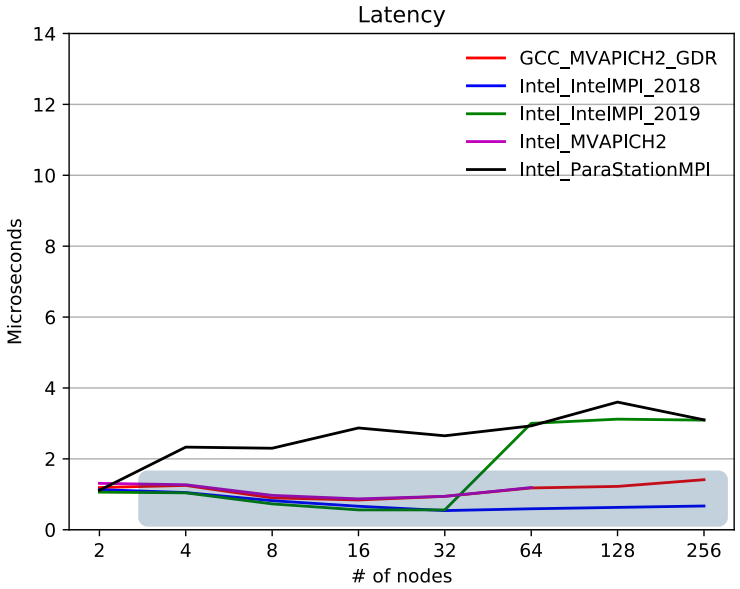
Juwels



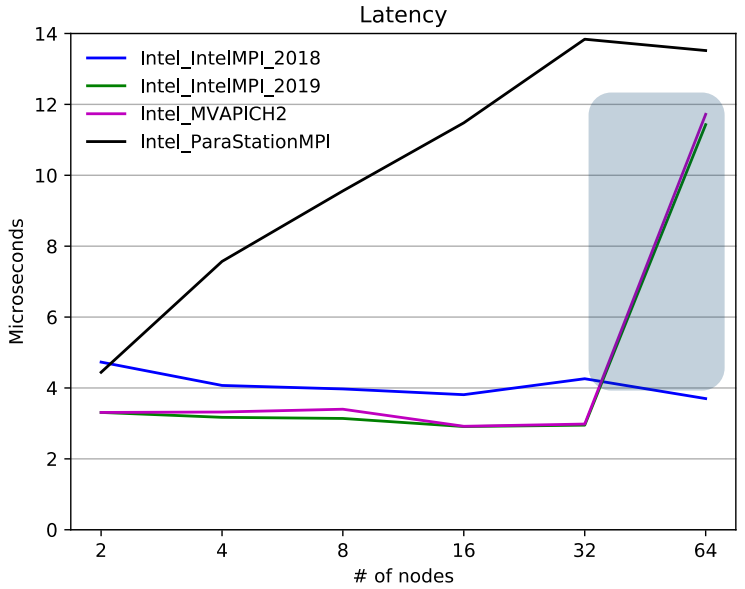
MPI RUNTIMES AT JSC

MPI performance (Gather, msg size 1 byte)

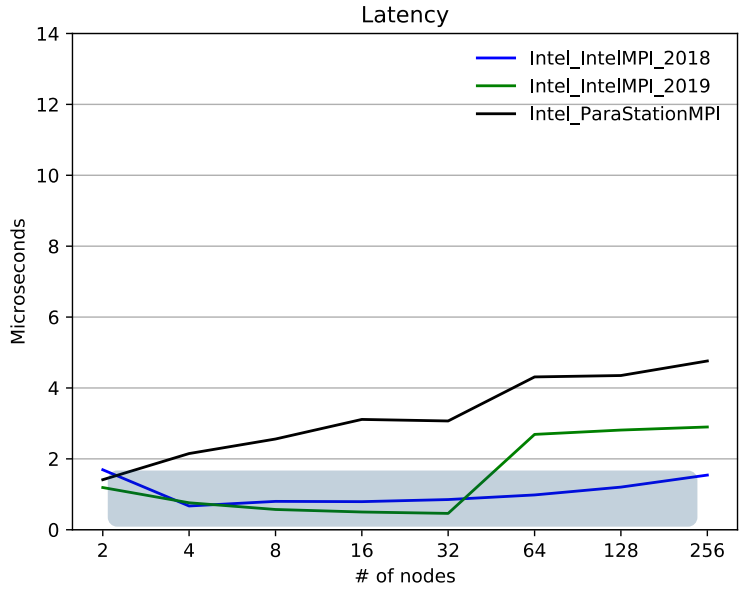
Jureca



Booster



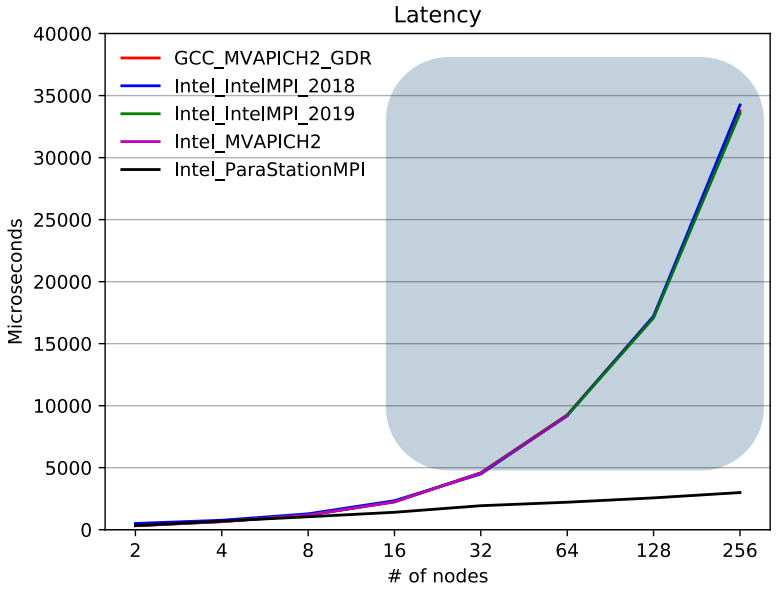
Juwels



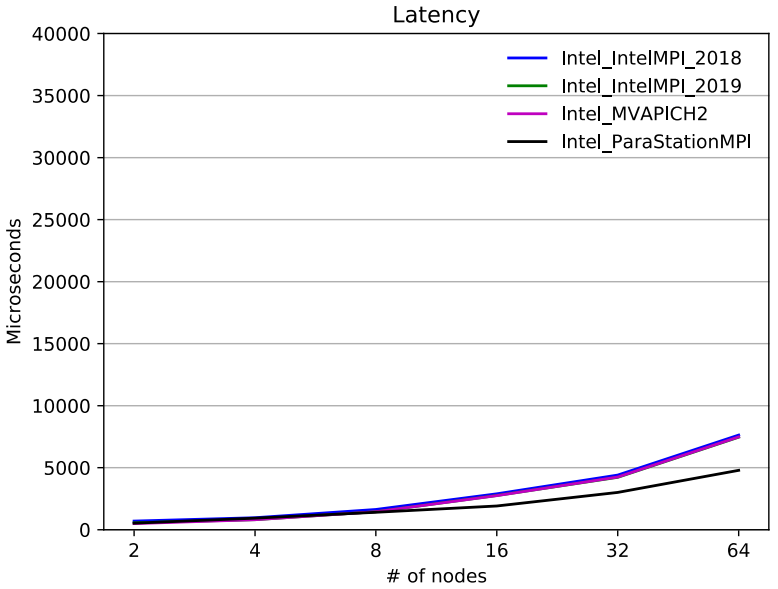
MPI RUNTIMES AT JSC

MPI performance (Gather, msg size 2MB)

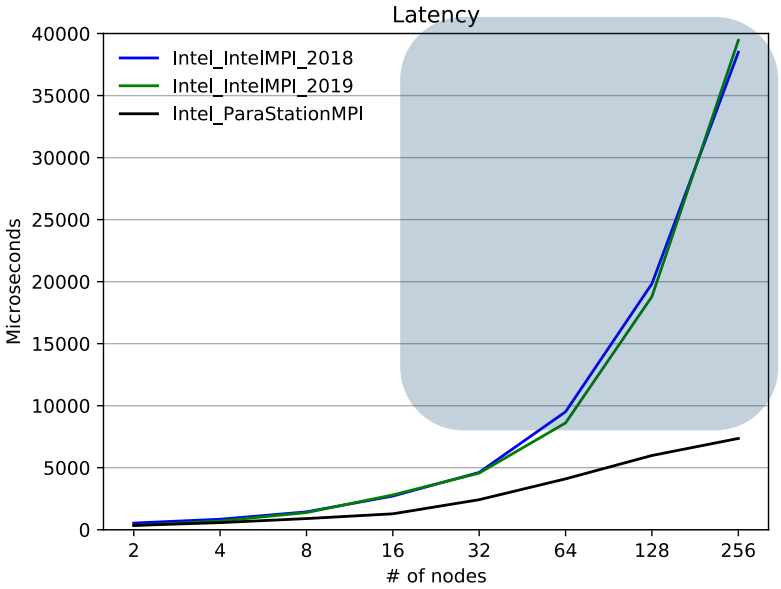
Jureca



Booster



Juwels



MPI RUNTIMES AT JSC

MPI performance

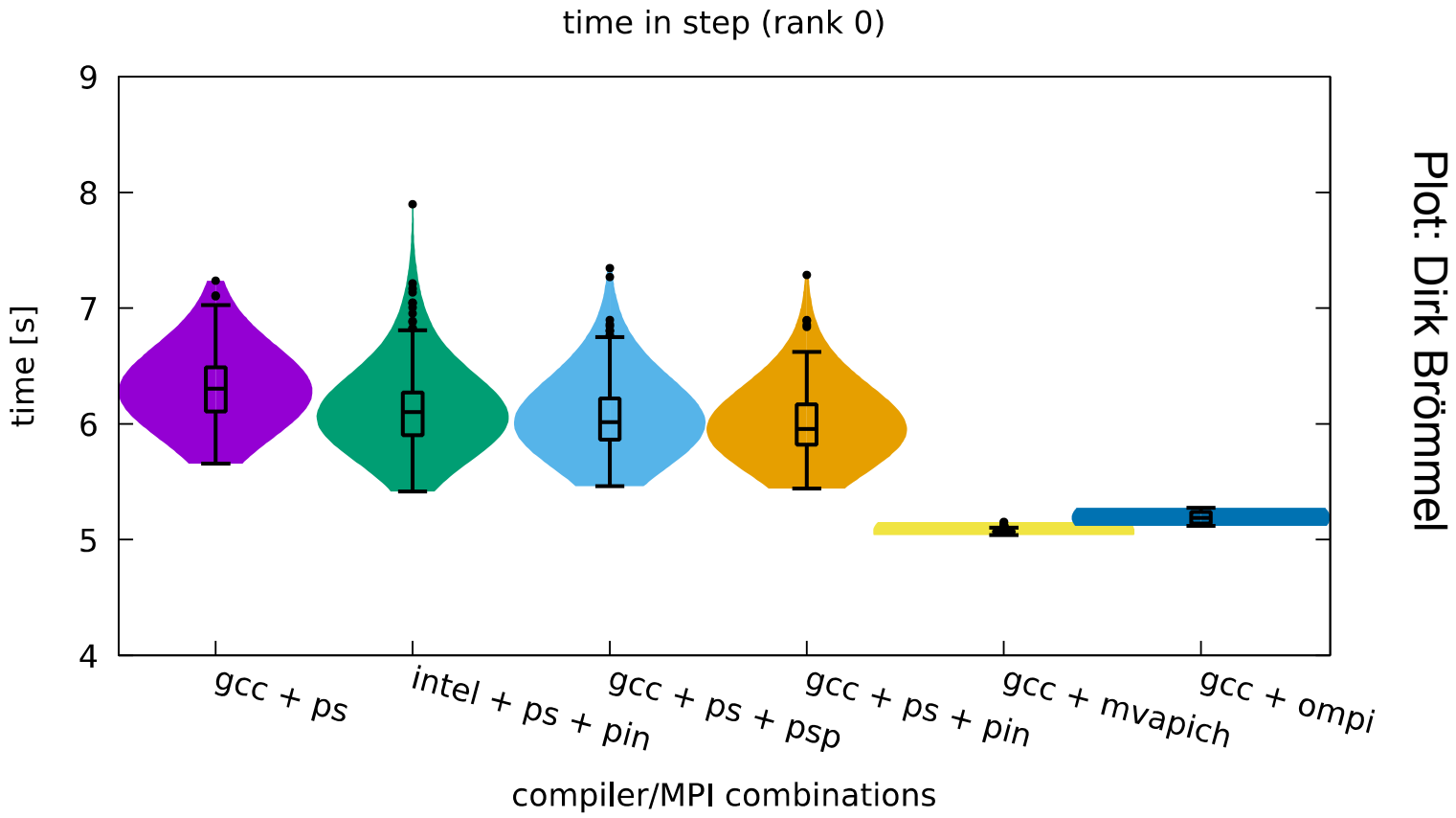
PEPC performance

- N-Body code
- Asynchronous communication based in a communicating thread spawned using pthreads
- 2 algorithms tested
  - Single particle processing with pthreads
  - Tile processing using OpenMP tasks
- Experiments in JURECA using 16 nodes, 4 MPI ranks per node and 12 computing threads per MPI rank (SMT enabled)
- Violin plots show average time per timestep and its distribution

MPI RUNTIMES AT JSC

MPI performance

- PEPC performance with single particle processing

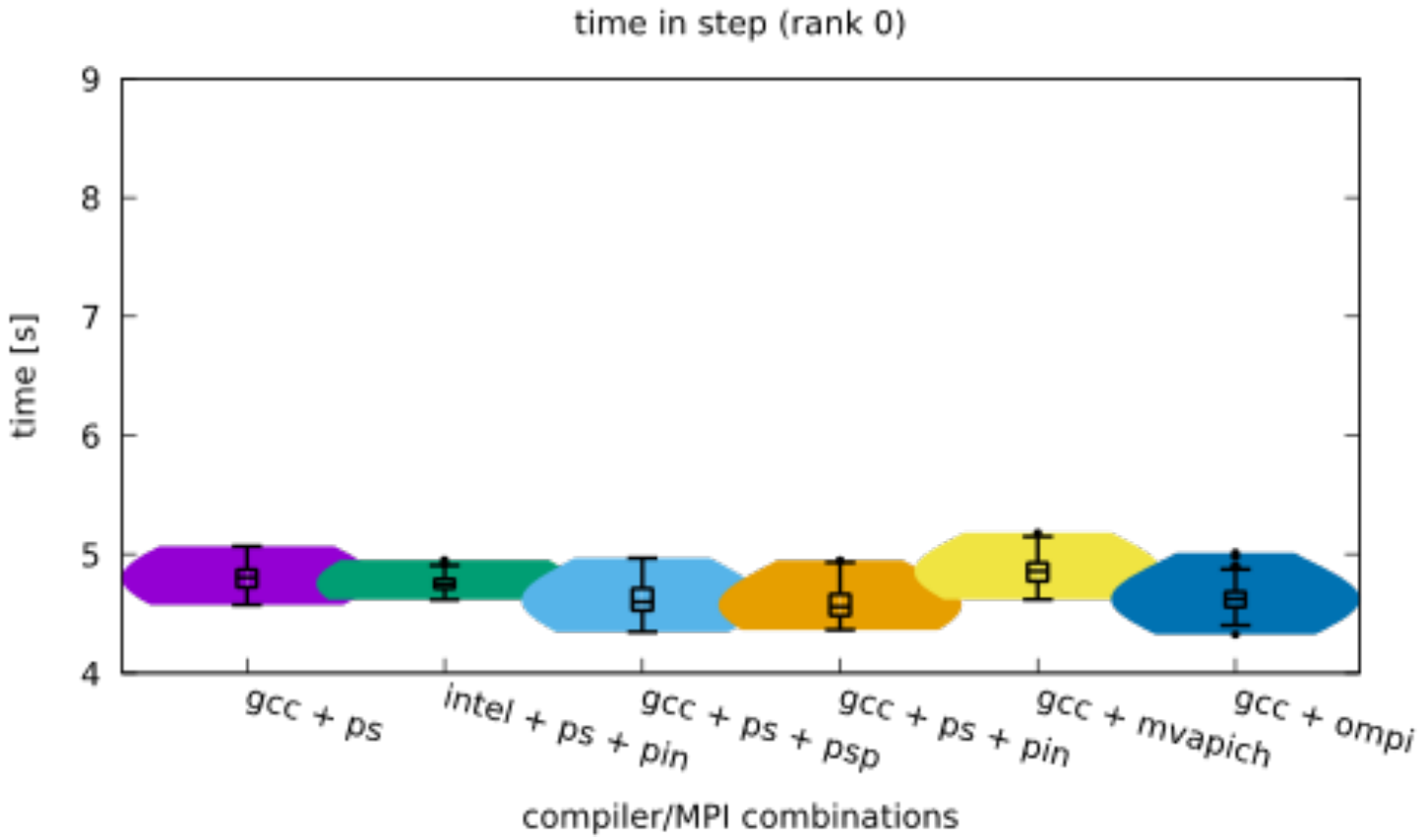


JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

MPI performance

- PEPC performance with tile processing



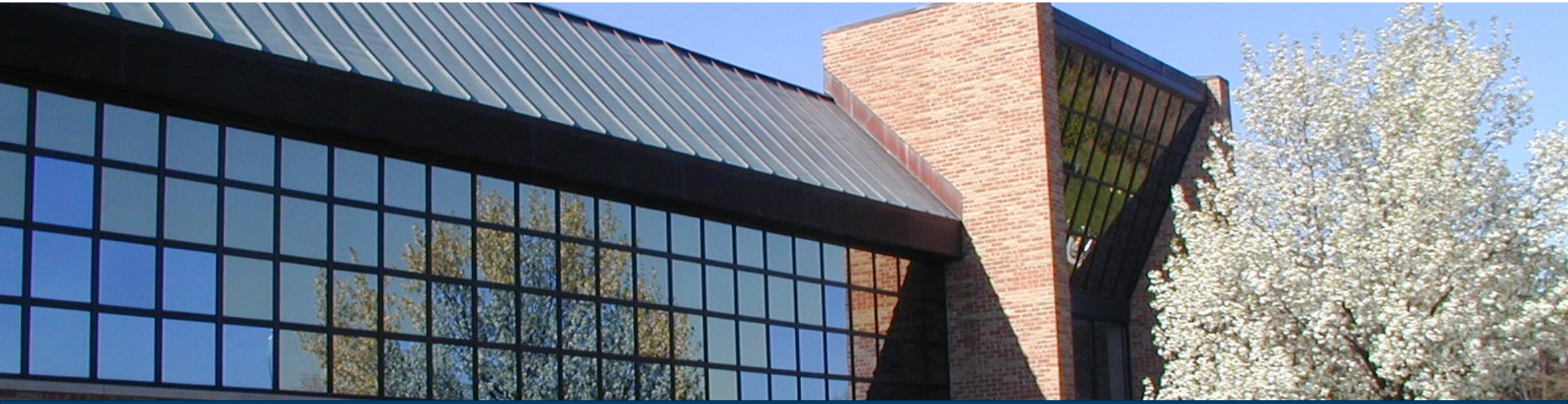
Plot: Dirk Brömmel

JÜLICH Forschungszentrum

MPI RUNTIMES AT JSC

MVAPICH at JSC

- Generally it's been a positive experience. But (as the guy in charge of installing all software in our systems)….
- The fact that MVAPICH2-GDR is a binary only package poses some annoyances
  - Need to ask the MVAPICH2 team for binaries for particular compiler versions
  - No icc/icpc/ifort version
  - Patching of compiler wrappers (xFLAGS used during compilation leak into them)
- The Deep Learning community is pushing for OpenMPI as CUDA-Aware MPI of choice in our systems



THANK YOU!