JUPITER

Paving the way for Application-centered Exascale Supercomputing in Europe

2024-10-10 I EUPEX FORUM

Member of the Helmholtz Association

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

Gauss Centre for Supercomputing

JÜLICH Forschungszentrum

Shaping Change

APPLICATIONS

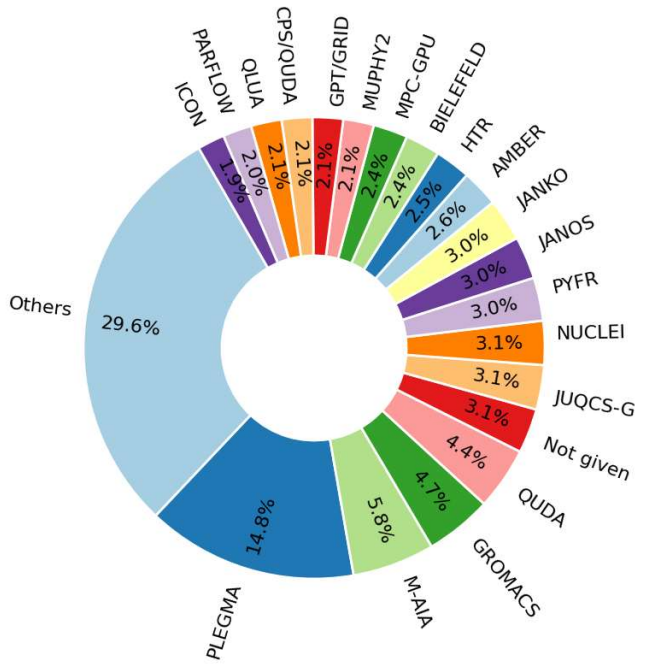
ASSESSING WITH APPLICATIONS

- Theoretical FLOP/s and GB/s are nice; but building machines for users
- → **Applications** core of procurement assessment
- Define representative benchmarks, *ExaBench*
  1. Analyze JSC workload
  2. Select fitting applications
  3. Benchmark*ize* them
  4. Submit as part of specification

  $ Get best machine

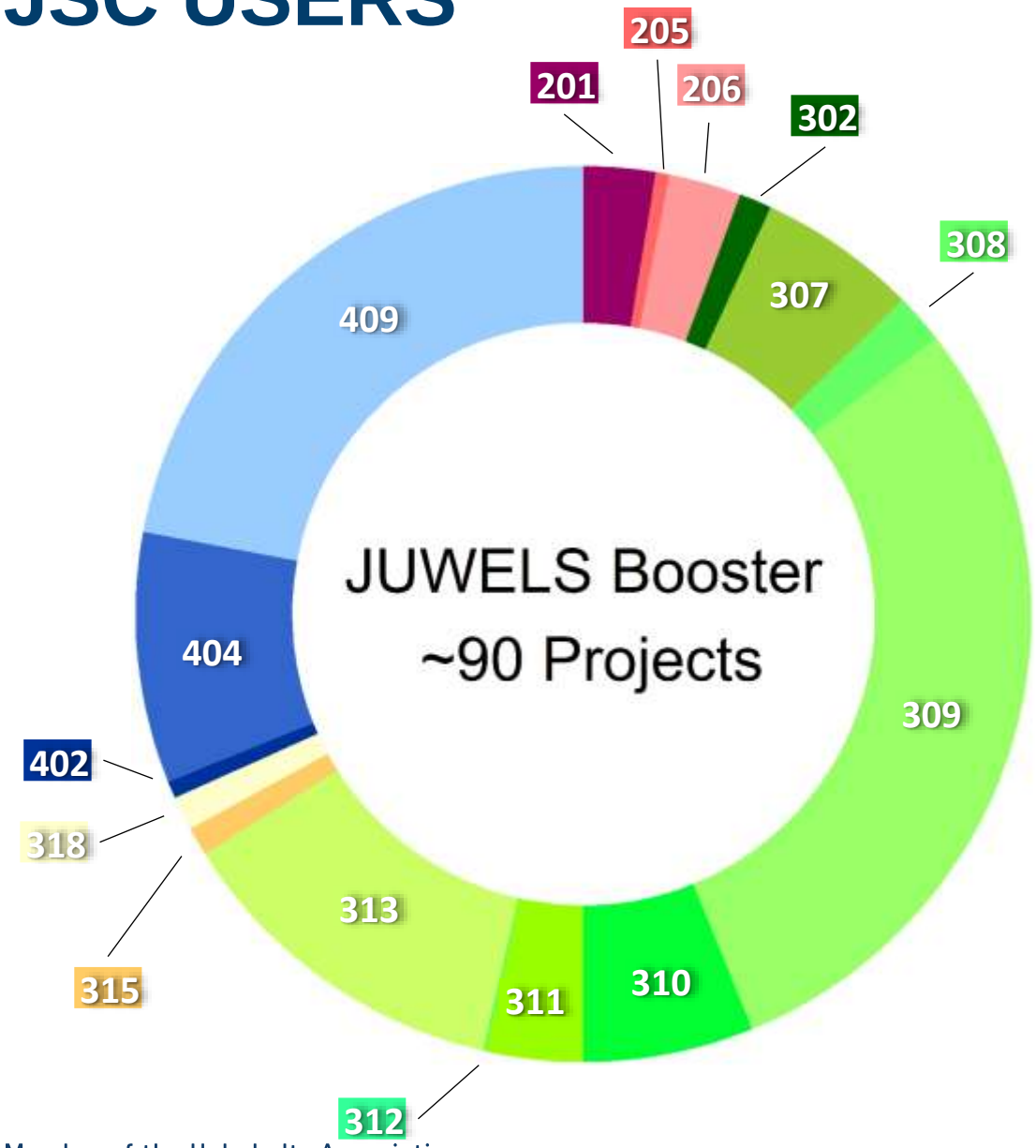
JSC WORKLOAD

- Application Support: Look at proposals, Slurm, …
- Very heterogeneous
- GPU mainly, but also
  - MIC (because good GPU candidates)
  - CPU (because might make change)



GPU Top 20 software 2021 (core-h)

JSC USERS



Reseach Fields

201	Basic Biological and Medical Research
205	Medicine
206	Neurosciences
302	Chemical Solid State and Surface Research
307	Condensed Matter Physics
308	Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas
309	Particles, Nuclei and Fields
310	Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics
311	Astrophysics and Astronomy
312	Mathematics
313	Atmospheric Science, Oceanography and Climate Research
315	Geophysics and Geodesy
318	Water Research
402	Mechanics and Constructive Mechanical Engineering
404	Heat Energy Technology, Thermal Machines, Fluid Mechanics
409	Computer Science

→ Define Benchmarks

APPLICATION SELECTION

- Selection criteria
  - Current workload
  - Future workload
  - Relevance
  - Balance with other applications
    - Domains
    - Programming models
    - Programming languages
    - Profile
  - Available PI/researcher

- Amber
- Arbor
- Chroma
- GROMACS
- ICON
- JUQCS
- nekRS
- ParFlow
- PIConGPU
- QuantumEspresso
- SOMA
- MMoCLIP
- NLP (Megatron)
- ResNet
- *DynQCD*
- *NAStJA*

FURTHER BENCHMARKS

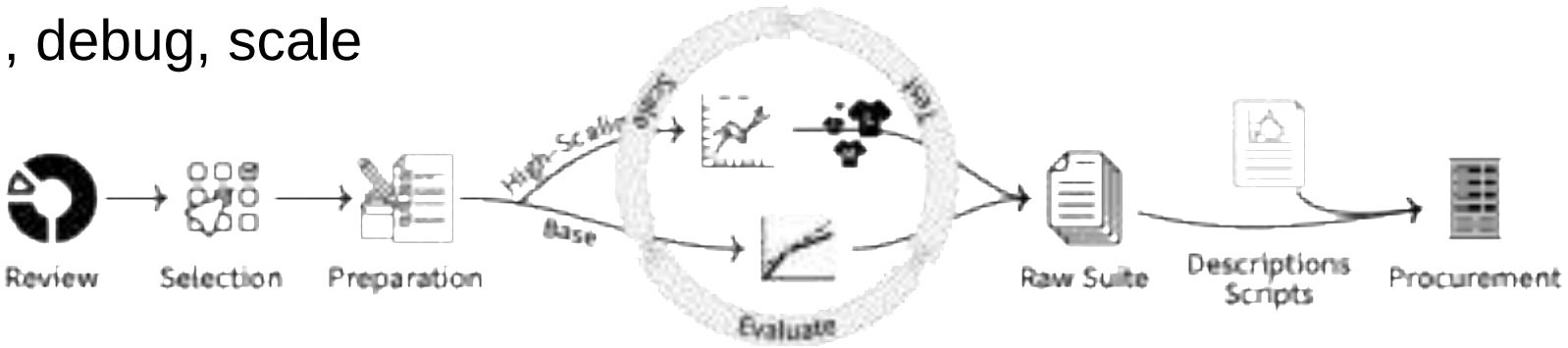
- Augment application (complex) benchmarks with synthetic (simpler) benchmarks
- Application benchmarks: Test complex interplay of usage by real-world applications
- Synthetic benchmark: Test specific feature of system design

- OSU micro-benchmarks *(network/MPI)*
- STREAM CPU, GPU *(Memory)*
- Graph500 *(network)*
- HPCG *(memory, network)*
- HPL *(compute, network)*
- IOR *(storage)*
- Linktest *(network/MPI)*

BENCHMARKIZATION

bˌɛntʃmɑːɹkɪzˈeɪʃən, creating *benchmarks* of mere *applications*

- **Goal**: Version of application for vendors, for which we get (best) result back
- Implications: recipe, rules, verification pre-defined; only small corrections
- Steps
  - Define workload, metric (unit of time)
  - Create JUBE script for reproducibility, uniformity, abstraction
  - Add verification of results
  - *Benchmark* benchmark: Run, debug, scale
  - Add documentation, rules



SUB-BENCHMARKS, VARIANTS

- Type of benchmarks
  - Applications benchmarks
  - Synthetic benchmarks
- Execution targets
  - JUPITER Booster (GPU, CPU)
  - JUPITER Cluster (CPU)
  - MSA
- Application benchmark categories
  - TCO
  - High-Scaling

	Booster			Cluster	MSA
Name	GPU	GPU High-Scale	CPU	CPU	
Amber	✓				
Arbor	✓	✓			
Chroma	✓	✓			
Gromacs	✓				
ICON	✓				
JUQCS	✓	✓			✓
nekRS	✓	✓			
ParFlow	✓				
PIConGPU	✓	✓			
Quantum ESPRESSO	✓				
SOMA	✓				
AI-MMoCLIP	✓				
AI-NLP	✓				
AI-ResNet	✓				
dynQCD				✓	
NAStJA				✓	
Graph500			✓		
HPCG	✓			✓	
HPL	✓			✓	
IOR			✓	✓	
LinkTest			✓	✓	✓
Multi-Flow IP			✓		
OSU	✓		✓	✓	
STREAM	✓			✓	

TCO

Total Cost of Ownership

- Traditional benchmark category
- ***How much of benchmark suite can be run in lifetime of system?*** Also: energy
- Key: same metric for each benchmark
  - Unit: time / s
  - Needed to convert rate → time
- One reference run for formula (e.g. 8 nodes); additional strong-scaled runs (e.g. 4, 16)
- Weights per individual benchmark
- Sophisticated formula for Cluster-Booster combination

HIGH-SCALING

- Give benchmarks a focus on large-scaleness of system
- Compare execution on full* JUWELS Booster to full* JUPITER Booster
  - *: Use 50 PFLOP/s$^{\text{th. peak}}$ part of JUWELS Booster

  → compare to 1000 PFLOP/s$^{\text{th. peak}}$ part of JUPITER Booster
- AKA **20×50 PF category**
- New challenge for us *(yay!)*
  - Design for unknown system, unknown device, unknown memory size
    Introduce 3 memory variants: small ($^2/_4$), medium ($^3/_4$), high ($^4/_4$ JWB A100 memory)
  - Hard to test on scale at JUWELS Booster
  - No way to test on scale required for JUPITER
  - Code issues at scale

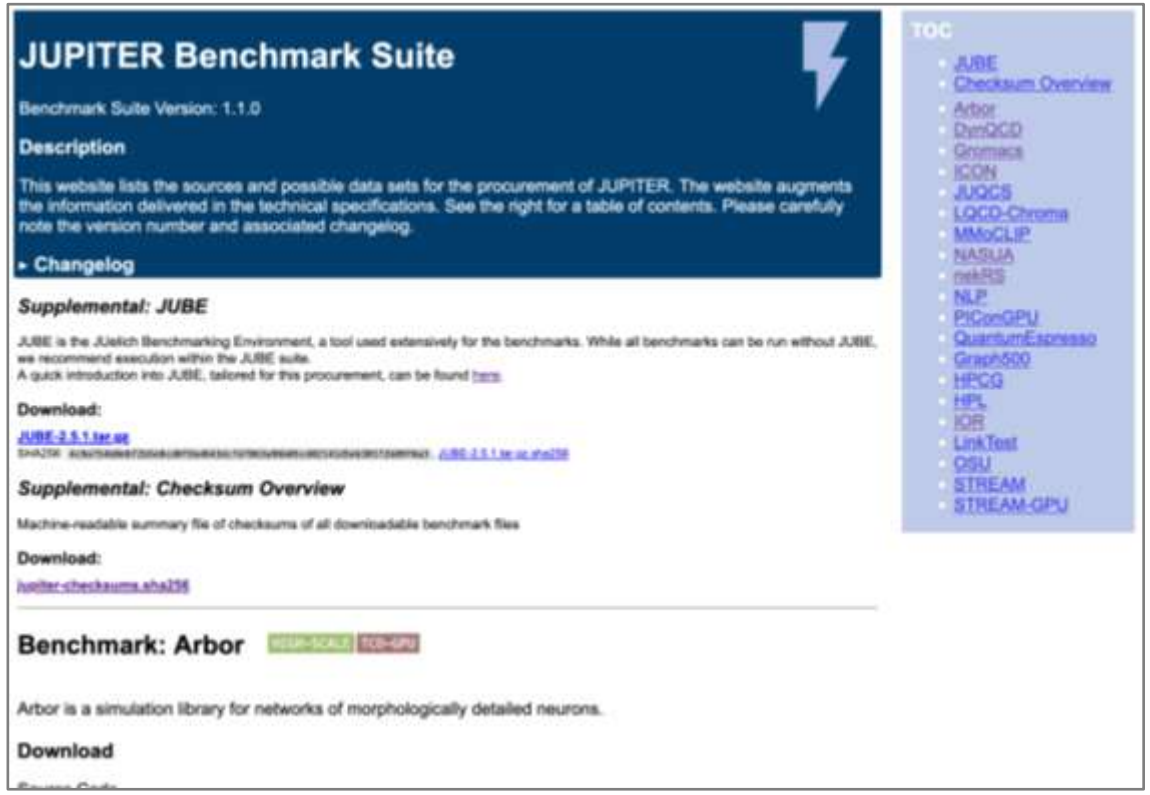
- Arbor
  tiny ($^1/_4$), small, medium, large
- Chroma
  small, medium, large
- JUQCS
  small, large
- nekRS
  small, medium, large
- PIConGPU
  small, medium, large

FINAL BENCHMARK LISTS

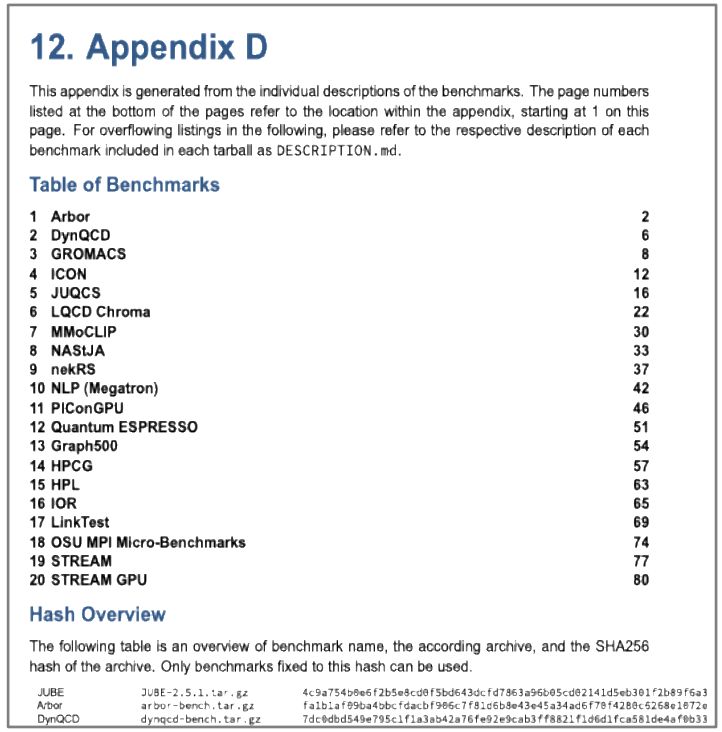
		Booster			Cluster	MSA
Before Dialogue	After Dialogue	GPU	GPU High-Scale	CPU	CPU	
Amber	~~Amber~~	~~✓~~				
Arbor	Arbor	✓	✓			
Chroma	Chroma	✓	✓			
Gromacs	Gromacs (2)	✓				
ICON	ICON (2)	✓				
JUQCS	JUQCS	✓	✓			✓
nekRS	nekRS	✓	✓			
ParFlow	ParFlow	✓				
PIConGPU	PIConGPU	✓	✓			
Quantum ESPRESSO	Quantum ESPRESSO	✓				
SOMA	~~SOMA~~	~~✓~~				
AI-MMoCLIP	AI-MMoCLIP	✓				
AI-NLP	AI-NLP	✓				
AI-ResNet	~~AI-Resnet~~	~~✓~~				
dynQCD	dynQCD				✓	
NAStJA	NAStJA				✓	
Graph500	Graph500			✓		
HPCG	HPCG	✓			✓	
HPL	HPL	✓			✓	
IOR	IOR			✓	✓	
LinkTest	LinkTest			✓	✓	✓
Multi-Flow IP	~~Multi-Flow IP~~			~~✓~~		
OSU	OSU (2)	✓		✓	✓	
STREAM	STREAM	✓			✓	

SUBMITTED FILE, WEBSITE

Rolling release of benchmark (as-early-as-possible) via website; with hashes



Reference description, list of hashes, in attachment of Technical Specification



12. Appendix D

This appendix is generated from the individual descriptions of the benchmarks. The page numbers listed at the bottom of the pages refer to the location within the appendix, starting at 1 on this page. For overflowing listings in the following, please refer to the respective description of each benchmark included in each tarball as `DESCRIPTION.md`.

Table of Benchmarks

Hash Overview

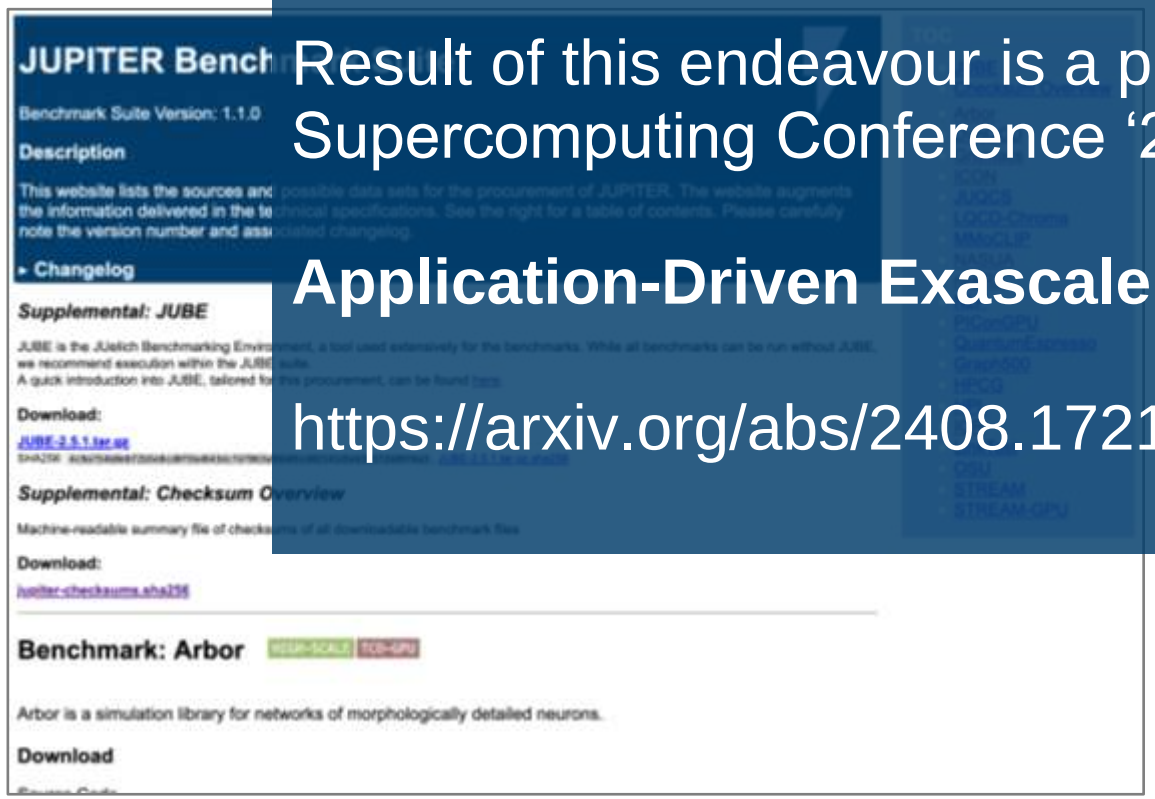
The following table is an overview of benchmark name, the according archive, and the SHA256 hash of the archive. Only benchmarks fixed to this hash can be used.

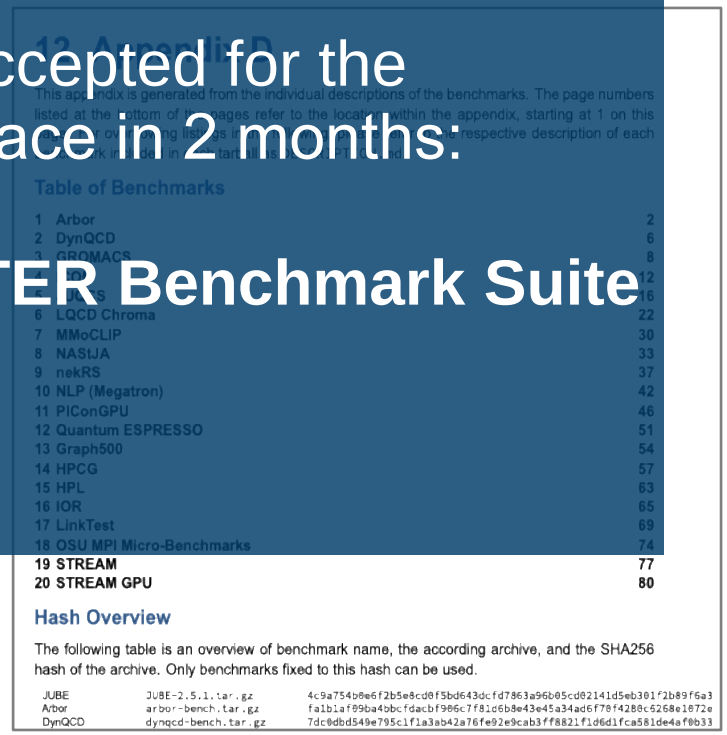
JUBE	JUBE-2.5.1.tar.gz	4c9a754b0e6f2b5e8cd0f5bd643dcfd7863a96b05cd02141d5eb301f2b89f6a3
Arbor	arbor-bench.tar.gz	fa1b1af09ba4bbcfdacbf906c7f81d6b8e43e45a34ad6f70f4280c6268e1072e
DynQCD	dynqcd-bench.tar.gz	7dc0dbd549e795c1f1a3ab42a76fe92e9cab3ff8821f1d6d1fca581de4af0b33

SUBMITTED FILE, WEBSITE

Rolling release of benchmark (as-early-as-possible) via website; with hashes

Reference description, list of hashes, in attachment of Technical Specification



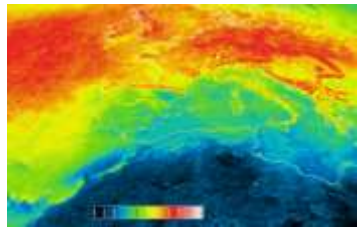


Result of this endeavour is a publication accepted for the Supercomputing Conference ‘24 to take place in 2 months:

Application-Driven Exascale: The JUPITER Benchmark Suite

https://arxiv.org/abs/2408.17211

HPC DEMONSTRATOR APPLICATIONS

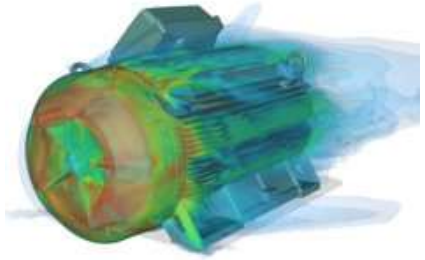


Weather Forecasting

IFS

AI4EO

PyTorch DDP, HPDBSCAN



Engineering

ESPRESO FEM

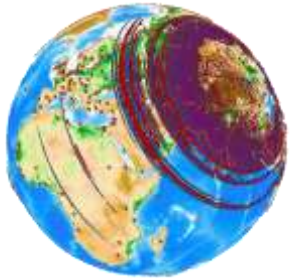
Biology and Health

LIGEN

Precision Agriculture Forecast

Astrophysics

Gadget



Seismology

SPECFEM3D

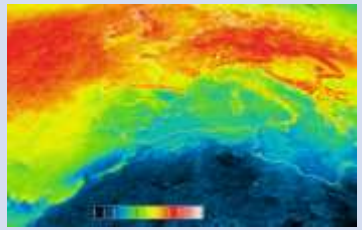
Material Science for Biology

BigDFT

- 8 HPC application demonstrators
  - Have been vectorised and optimized for ARM SVE + HBM
  - Are currently ported and optimized for Grace/Hopper

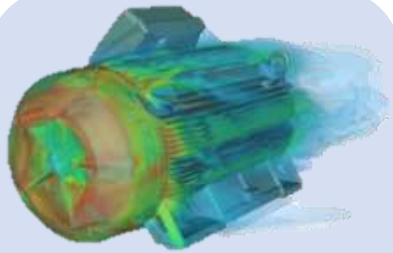
- https://eupex.eu/applications/

HPC DEMONSTRATOR APPLICATIONS



Weather Forecasting
IFS

AI4EO
PyTorch DDP, HPDBSCAN



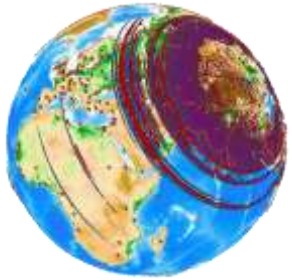
Engineering
ESPRESO FEM

Biology and Health
LIGEN

- None of these codes used during procurement
- BUT common domains represented there

Precision Agriculture Forecast

Astrophysics
Gadget



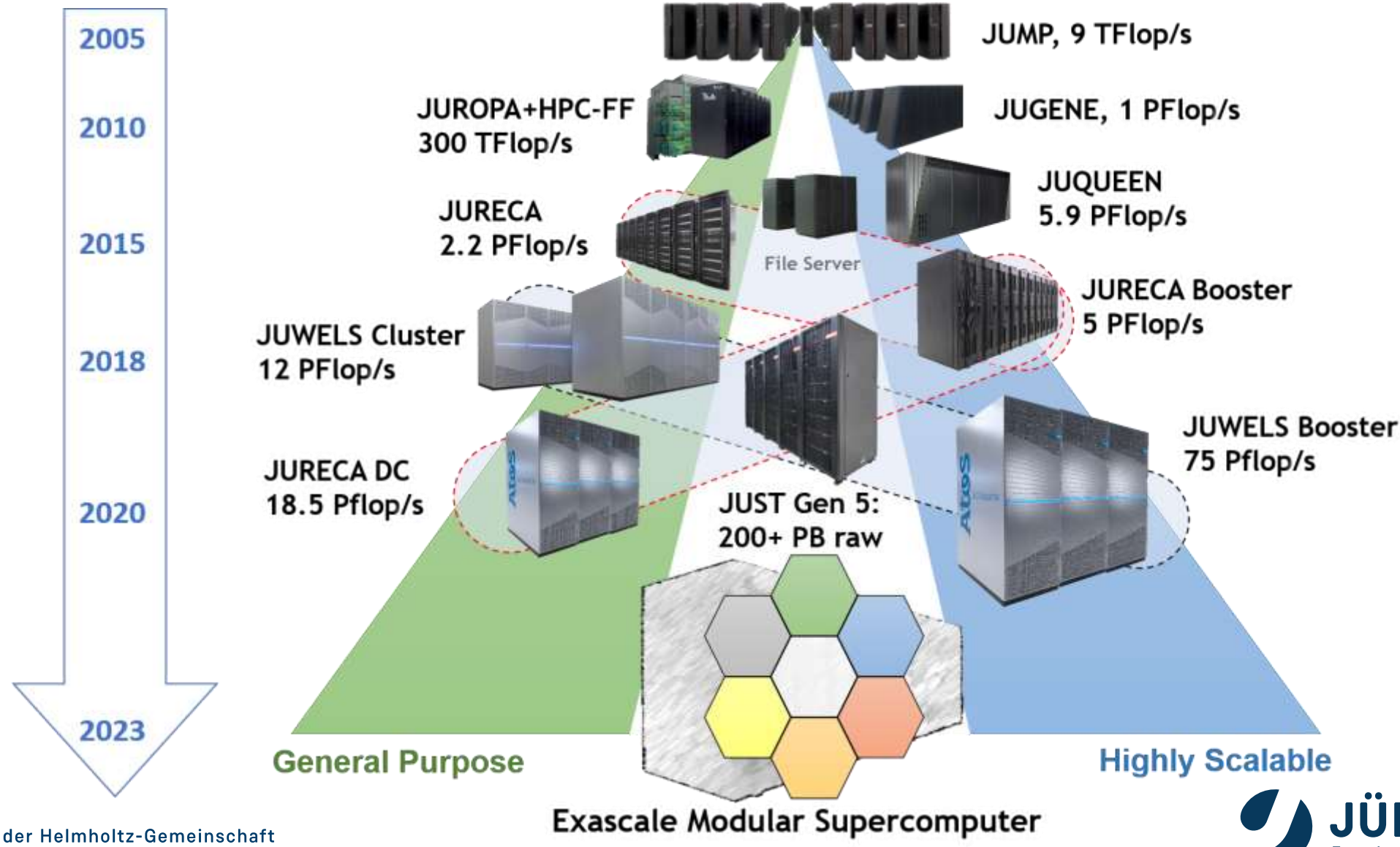
Seismology
SPECFEM3D

Material Science for Biology
BigDFT

- https://eupex.eu/applications/

JUPITER ARCHITECTURE

(DUAL) hardware strategy at JSC



MODULAR SUPERCOMPUTING AT EXASCALE

Modular Supercomputing Architecture (MSA)

MSA Pre-Exascale @Lux, @It

2× MSA pilots

2011

FP7 - H2020

DEEP Projects

2020

EuroHPC

SEA Projects

2021

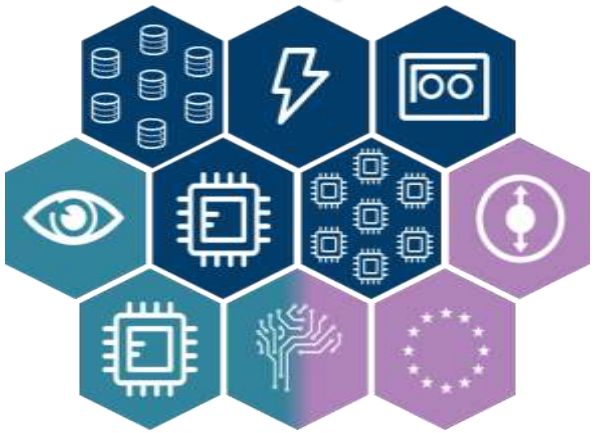
2022

EU PEX

HPC OS

2023

2024



JÜLICH Forschungszentrum

Member of the Helmholtz Association

DISCOVERING JUPITER

- ParTec/Eviden Supercomputer Consortium
- Implementing Modular Supercomputing Architecture
- JUPITER Booster: High scalability; 1 EFLOP/s HPL, >70 EFLOP/s FP8
- JUPITER Cluster: High versatility; 0.5 B/FLOP balance
- Network: 200/400 Gigabit NVIDIA Mellanox InfiniBand NDR
- Storage: 29 PB Flash IBM Storage Scale 6000
- 17 Megawatt Linpack Power Consumption
- Direct Liquid Cooled (36 -> 4x degree) to enable heat-reuse

BullSequana XH3000 DLC

ParTec MODULAR SUPERCOMPUTING EVIDEN NVIDIA SIPEARL

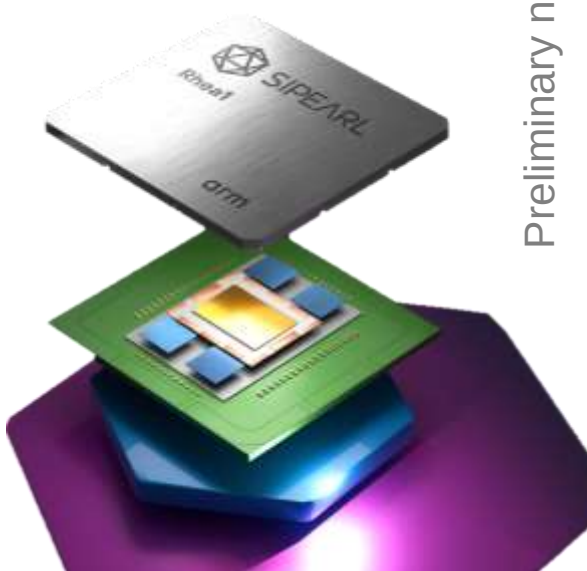
JUPITER MODULES

JUPITER Booster

- ~125 Racks BullSequana XH3000
- Node design
  - ~6000 nodes
  - 4× NVIDIA CG1 per node
- CG1: NVIDIA Grace-Hopper
  - 72 Arm Neoverse V2 cores (4×128b SVE2); 120 GB LPDDR5
  - H100 (132 SMs); 96 GB HBM3
  - NVLink C2C (900 GB/s)

JUPITER Cluster

- ~14 Racks BullSequana XH3000
- Node design
  - ~1300 nodes
  - 2× SiPearl Rhea1 per node
- Rhea1
  - 80 Arm Neoverse V1 cores (2×256b SVE)
  - 256 GB DDR5, 64 GB HBM2e

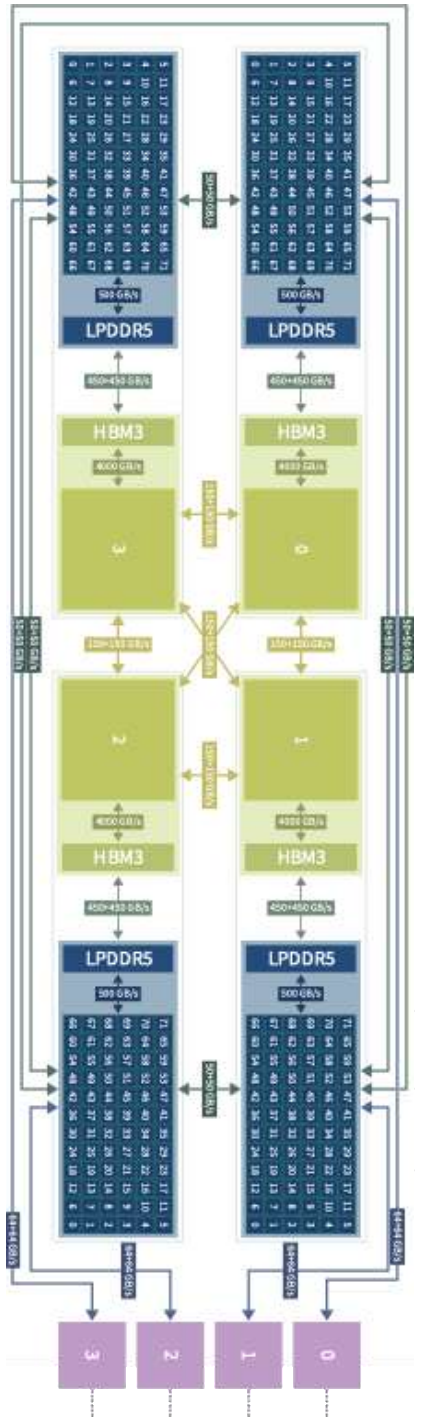


Preliminary numbers, might change during installation

JUPITER – THE BOOSTER

Highly-Scalable Module for HPC and AI workloads

- **1** ExaFLOP/s (FP64, HPL)
- NVIDIA Grace-Hopper CG1
  - ~5900 compute nodes
  - **4×** CG1 chips per compute node
- NVIDIA Mellanox NDR
  - **4×** NDR200 NICs per compute node
- BullSequana XH3000
  - Direct Liquid Cooled blades
  - **2× compute node per blade**

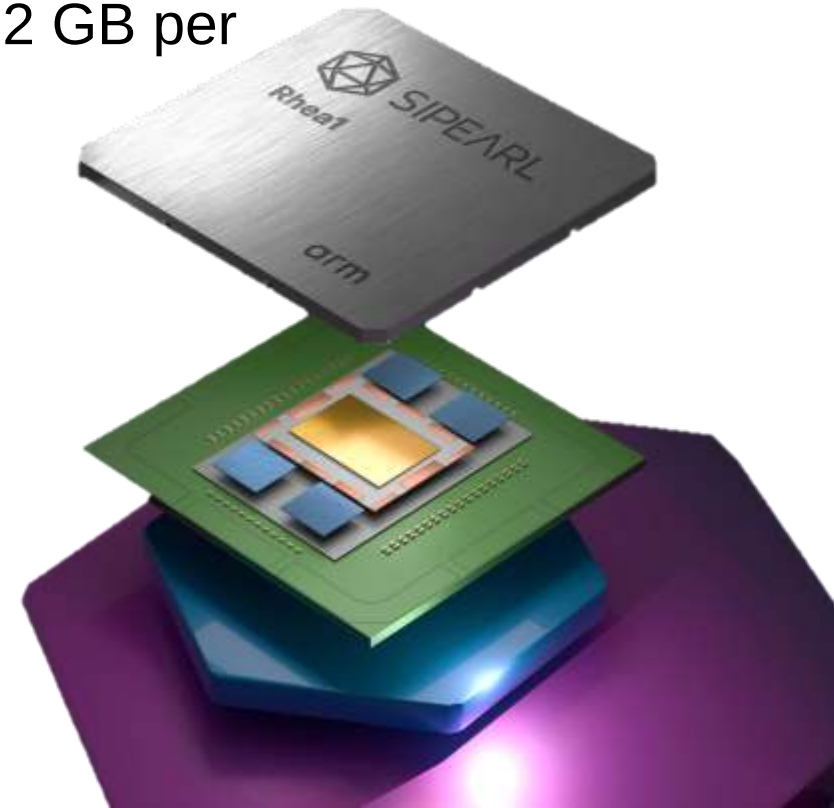


JUPITER – THE CLUSTER

General-Purpose Module for Mixed Workloads

- >5 PetaFLOP/s (FP64, HPL)
- SiPearl Rhea1
  - ~1340 compute nodes
  - 2× CPUs per node
- NVIDIA Mellanox NDR
  - 1× NDR200 NICs per compute node
- BullSequana XH3000
  - Direct Liquid Cooled blades
  - 3× compute nodes per blade

- 80 Neoverse V1 cores
  - 2× 256 SVE each
- 64 GB HBM (128 GB per node)
- 256 GB DDR5 (512 GB per node)



Preliminary numbers, might change during installation

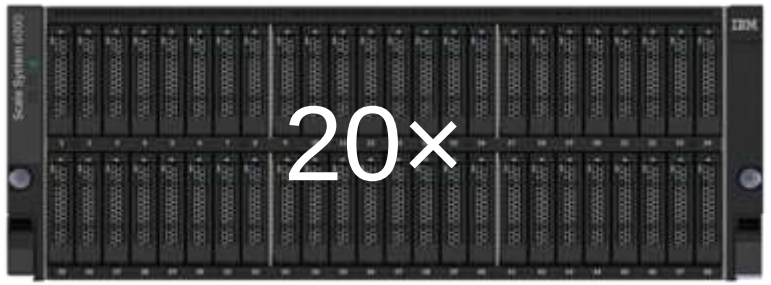
JUPITER – LOGIN/VISUALIZATION

Login Partition and Visualization Capabilities

- Login Nodes
  - Booster: 12 nodes, 1* CG1
  - Cluster: 5 nodes, 2* Rhea1
- Visualization Nodes
  - Booster: 3 nodes, 1* CG1
  - Cluster: 3 nodes, 2* Rhea1 and 2* NVidia A40
- 2* 100Gbit Ethernet for external connectivity

JUPITER – STORAGE (SCRATCH)

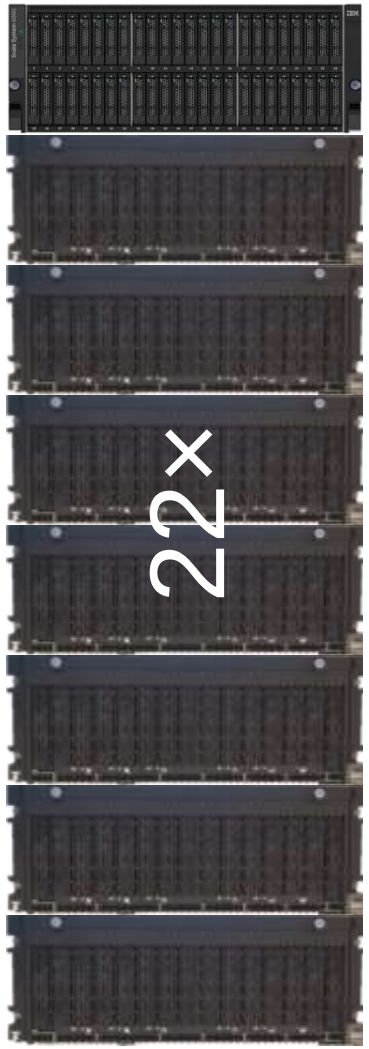
- Gross Capacity: 29 PB; Net Capacity: 21 PB
- Bandwidth: 2.1 TB/s Write, 3.1 TB/s Read
- 20× IBM SSS6000 Building Blocks (40 servers)
  - 2× NDR400 per server
  - 48× 30 TB NVMe drives per block
  - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
  - Integrated into InfiniBand fabric



JUPITER – STORAGE (EXASTORE)

In kind contribution from JSC, not part of the JUPITER procurement

- Gross Capacity: 308 PB; Net Capacity: 210 PB
- Bandwidth: 1.1 TB/s Write, 1.4 TB/s Read
- 22× IBM SSS6000 Building Blocks (44 servers)
  - 2× NDR200 per server
  - 7× JBOD enclosures, each with 91x 22 TB Spinning Disks per block
  - IBM Storage Scale (aka Spectrum Scale/GPFS)
- Manager and Datamover Nodes
- Exclusive for JUPITER
  - Integrated into InfiniBand fabric



NETWORK DESIGN

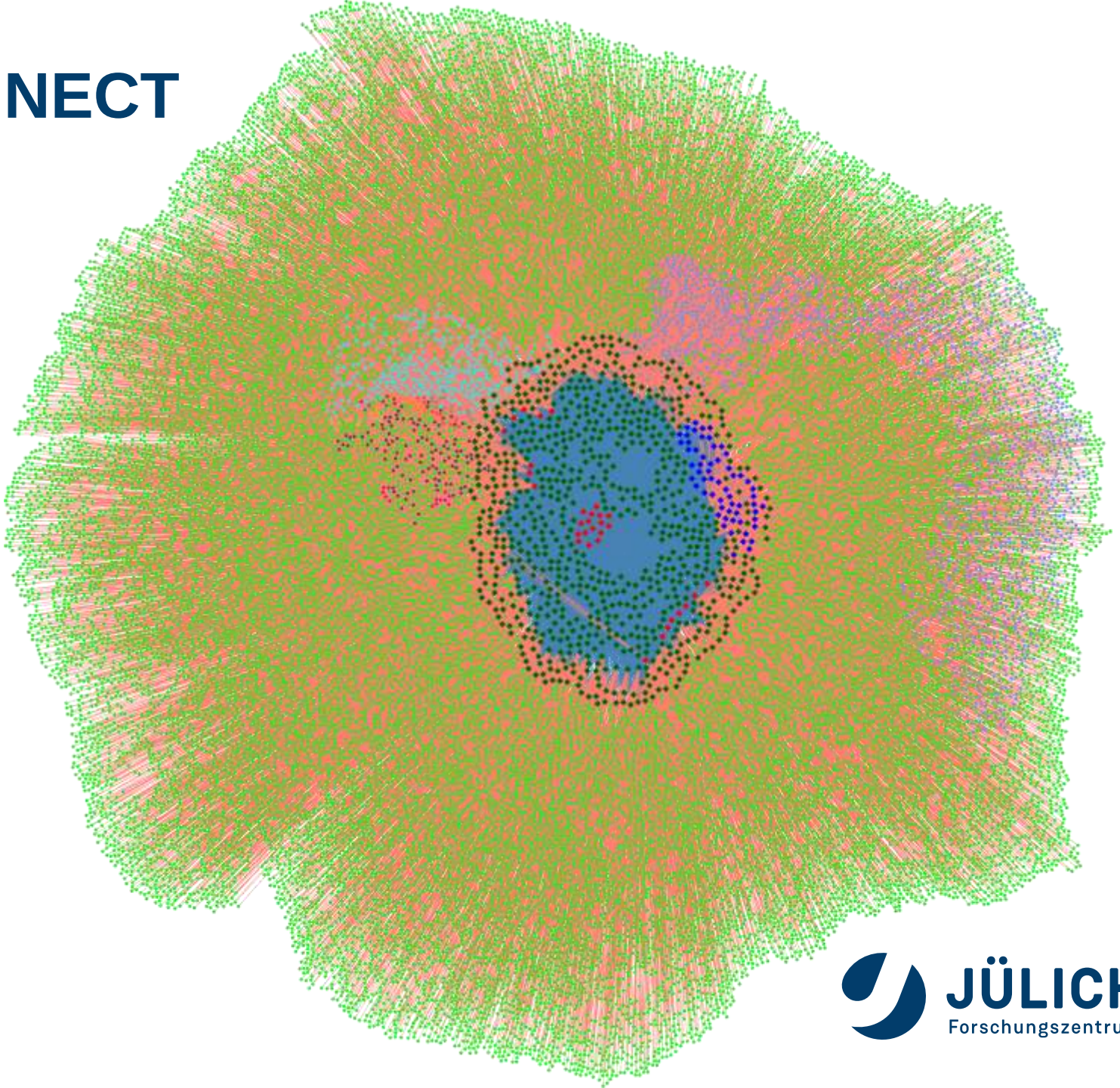
JUPITER – INTERCONNECT

One Network to Rule Them All

- NVIDIA Mellanox InfiniBand NDR/NDR200
  - NVIDIA Quantum-2 switches
  - NVIDIA Connect-X7 HCAs
- Dragonfly+ topology
  - 27 Dragonfly groups
  - Within each group: full fat tree
- 51000 links, 102000 logical ports, 25400 endpoints, 867 switches
- Adaptive Routing
- In-network processing on switch level (SHARPv3), tentatively

JUPITER – INTERCONNECT

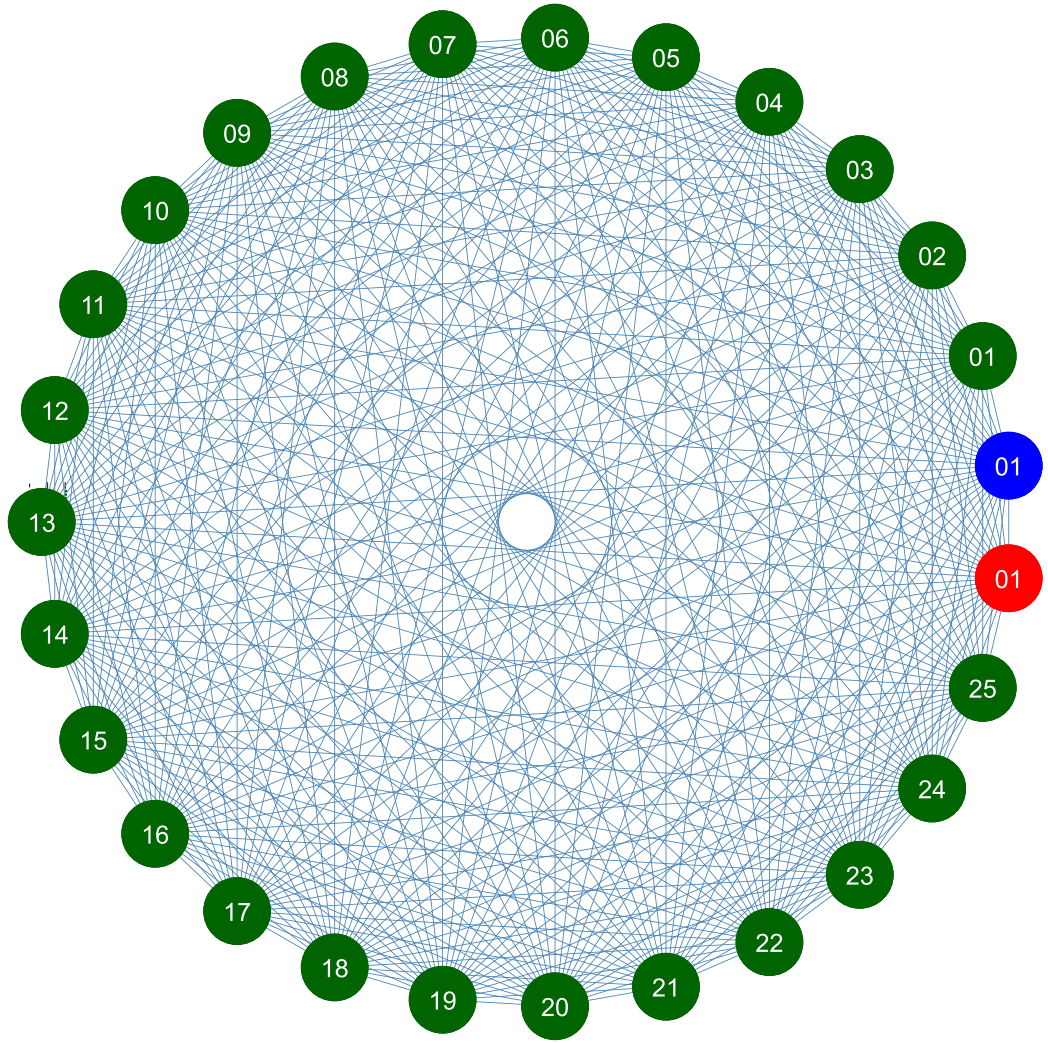
One Network to Rule Them All



JÜLICH
Forschungszentrum

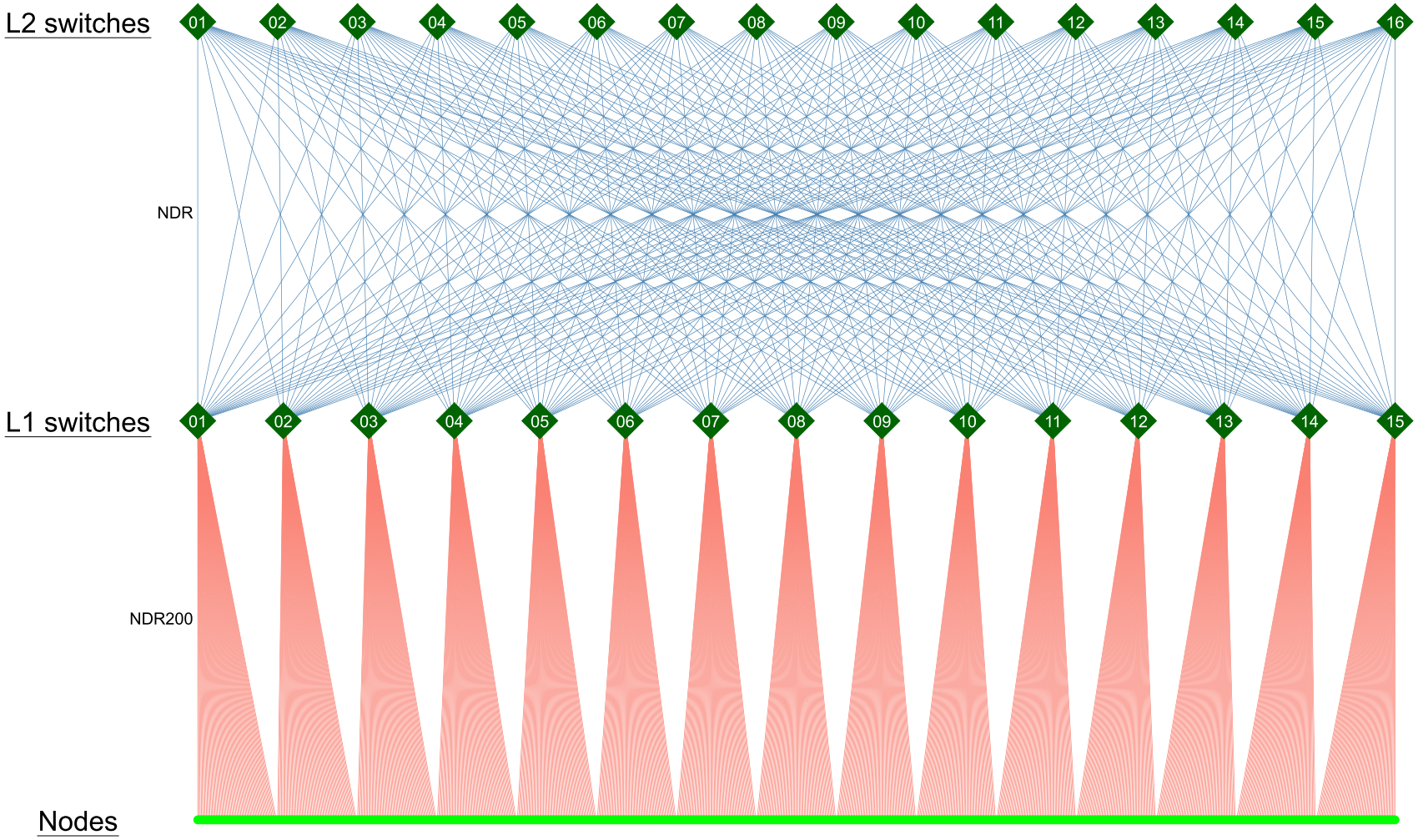
JUPITER – INTERCONNECT

One Network to Rule Them All



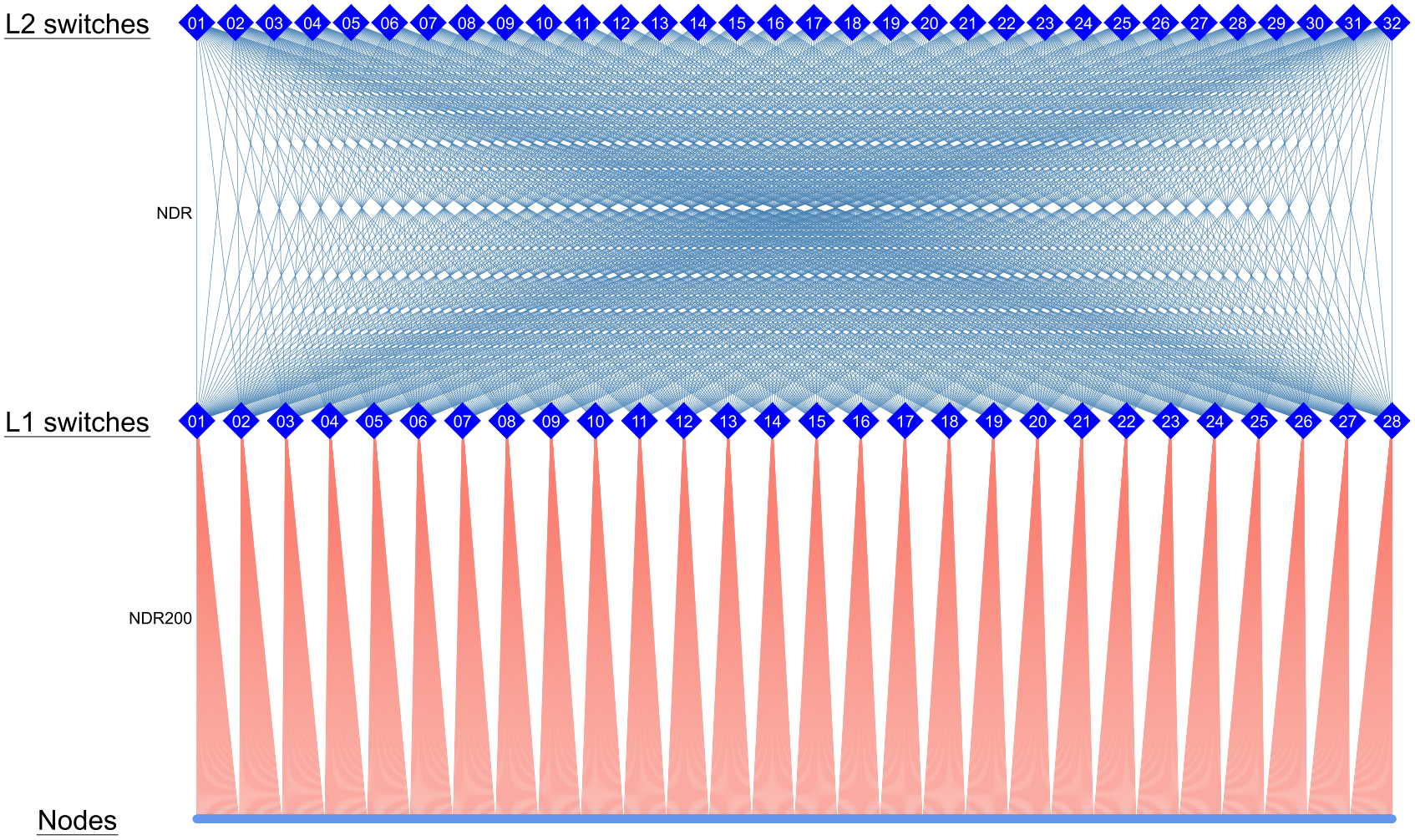
JUPITER – INTERCONNECT

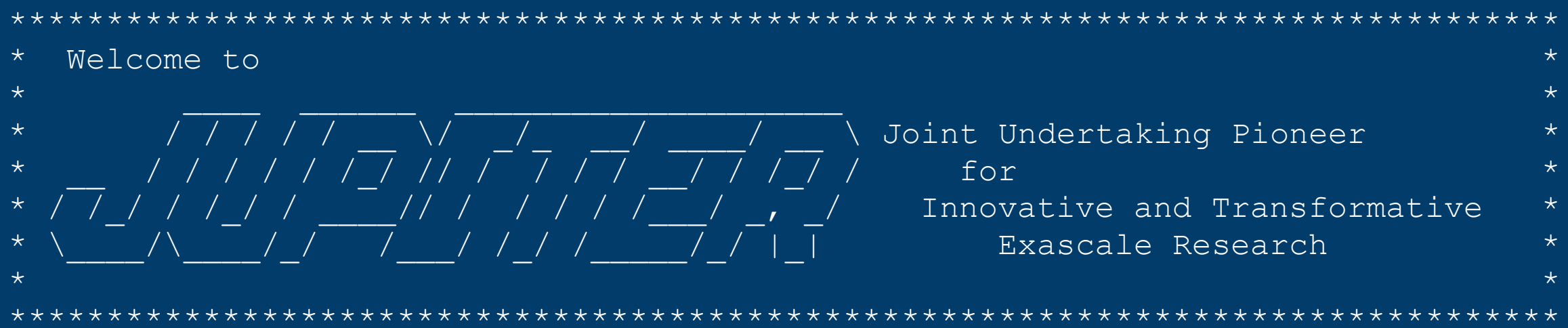
One Network to Rule Them All



JUPITER – INTERCONNECT

One Network to Rule Them All





SYSTEM MANAGEMENT

JUPITER Management Stack

3 main pillars/actors

SMC xScale
Core part of the stack. Vast majority of components come from here.
Developed by Eviden
Heavily based on open source and cloud technologies

JUPITER Management Stack

3 main pillars/actors

SMC xScale	ParaStation
Core part of the stack. Vast majority of components come from here.	Enhancement of the core
Developed by Eviden	Developed by ParTec
Heavily based on open source and cloud technologies	Integrates ParTec tools in SMCx to streamline their support workflows

JUPITER Management Stack

3 main pillars/actors

SMC xScale	ParaStation	xOPS
Core part of the stack. Vast majority of components come from here.	Enhancement of the core	Enhancement of the core
Developed by Eviden	Developed by ParTec	Developed by JSC
Heavily based on open source and cloud technologies	Integrates ParTec tools in SMCx to streamline their support workflows	Extensive set of Ansible roles for HPC, targeting JSC's requirements and needs

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security Performance Stability HW support	

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security Performance Stability HW support	
Management Storage	Ceph	Multi-use Scalable Performance	

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security Performance Stability HW support	
Management Storage	Ceph	Multi-use Performance Scalable	IBM

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security, Performance, Stability, HW support	
Management Storage	Ceph	Multi-use, Performance, Scalable	IBM
Management Plane	Kubernetes	Scalable, 0 downtime, Flexible, Open	kubernetes

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security, Performance, Stability, HW support	
Management Storage	Ceph	Multi-use, Performance, Scalable	IBM
Management Plane	Kubernetes	Scalable, 0 downtime, Flexible, Open	kubernetes
Configuration Management	Ansible	Standard, Easy to extend, Open	

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security, Performance, Stability, HW support	
Management Storage	Ceph	Multi-use, Performance, Scalable	IBM
Management Plane	Kubernetes	Scalable, 0 downtime, Flexible, Open	kubernetes
Configuration Management	Ansible	Standard, Easy to extend, Open	
Boot Image(s) Management	ImageBuilder	ARM / x86 support, Tracking, Integration	

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Operating System	Linux	Security, Performance, Stability, HW support	
Management Storage	Ceph	Multi-use, Performance, Scalable	IBM
Management Plane	Kubernetes	Scalable, 0 downtime, Flexible, Open	kubernetes
Configuration Management	Ansible	Standard, Easy to extend, Open	
Boot Image(s) Management	ImageBuilder	ARM / x86 support, Tracking, Integration	
Container(s)	UBI Universal Binary Images	Standard, Security, Consistency	

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable API Known	SCHEDMD The Slurm Company; ParaStation PSSLURM

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable Known API	SchedMD The Slurm Company ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance Scalable Data security	IBM

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable Known API	SchedMD The Slurm Company; ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance Scalable Data security	IBM
MPI Runtime	Message Passing Interface	Stable GPU-support Performance Bug-free	ParaStation MPI; EVIDEN Open MPI

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable, Known API	SchedMD The Slurm Company; ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance, Scalable, Data security	IBM
MPI Runtime	Message Passing Interface	Stable, GPU-support, Performance, Bug-free	ParaStation MPI; EVIDEN Open MPI
GPU Support	CUDA, HPC SDK	Memory management, Performance, Integration	NVIDIA

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable, Known API	SchedMD The Slurm Company; ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance, Scalable, Data security	IBM
MPI Runtime	Message Passing Interface	Stable, GPU-support, Performance, Bug-free	ParaStation MPI; EVIDEN Open MPI
GPU Support	CUDA, HPC SDK	Memory management, Performance, Integration	NVIDIA; Red Hat

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable, Known API	SchedMD The Slurm Company; ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance, Scalable, Data security	IBM
MPI Runtime	Message Passing Interface	Stable, GPU-support, Performance, Bug-free	ParaStation MPI; EVIDEN Open MPI
GPU Support	CUDA, HPC SDK	Memory management, Performance, Integration	NVIDIA;
Monitoring & Logging	Prometheus + Thanos, Syslog + Fluentd	Usable, Scalable, Handle data storm	Prometheus; Thanos; fluentd

Slide courtesy of EVIDEN

JUPITER Management Stack – key areas

	Technology	Challenges	Provider
Resource Manager	Slurm	Scalable, Known API	SchedMD The Slurm Company; ParaStation PSSLURM
Parallel Storage	Storage Scale System (GPFS)	Performance, Scalable, Data security	IBM
MPI Runtime	Message Passing Interface	Stable, GPU-support, Performance, Bug-free	ParaStation MPI; EVIDEN Open MPI
GPU Support	CUDA, HPC SDK	Memory management, Performance, Integration	NVIDIA;
Monitoring & Logging	Prometheus + Thanos, Syslog + Fluentd	Usable, Scalable, Handle data storm	Prometheus; Thanos; fluentd
Reference Database	Data Center Information Management	Automation, API, Coherent	netbox

Slide courtesy of EVIDEN

JUPITER MANAGEMENT STACK

A view from the vendor consortium

Ticketing Servers

SEMS
Smart Energy Management Suite

SDMS
Smart Data Management Suite

SMMS
Smart Maintenance Management Suite

IBMS
Interconnect faBric Management Suite

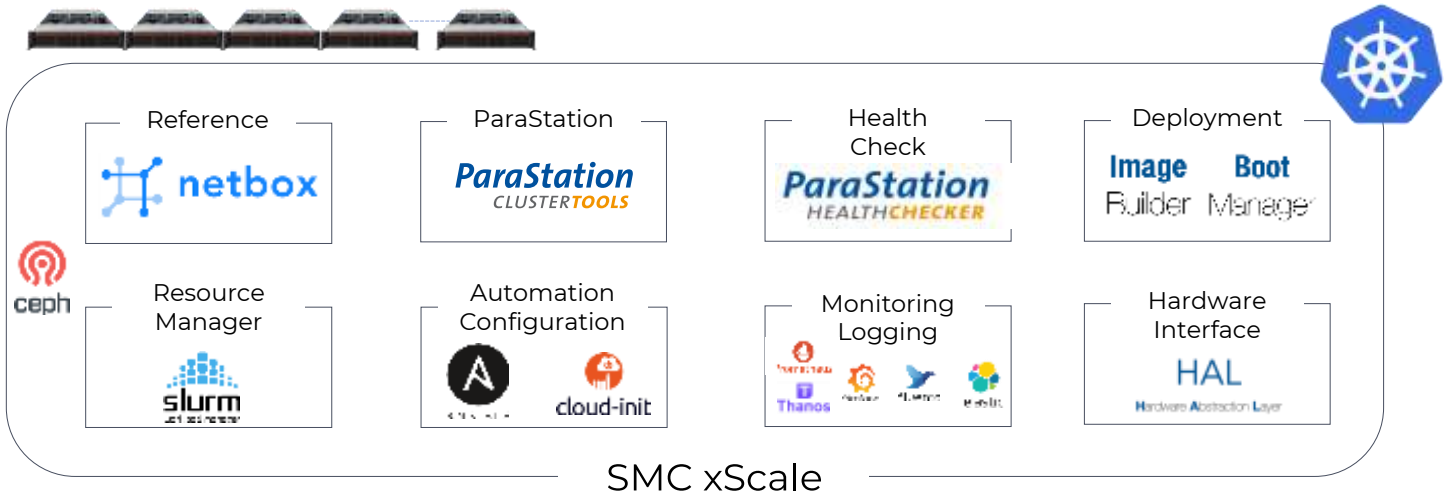
Red Hat Enterprise Linux 9

NVIDIA

SIPEARL

Management Servers

Red Hat Enterprise Linux 8



ParaStation

ParaStation MPI

OpenMPI

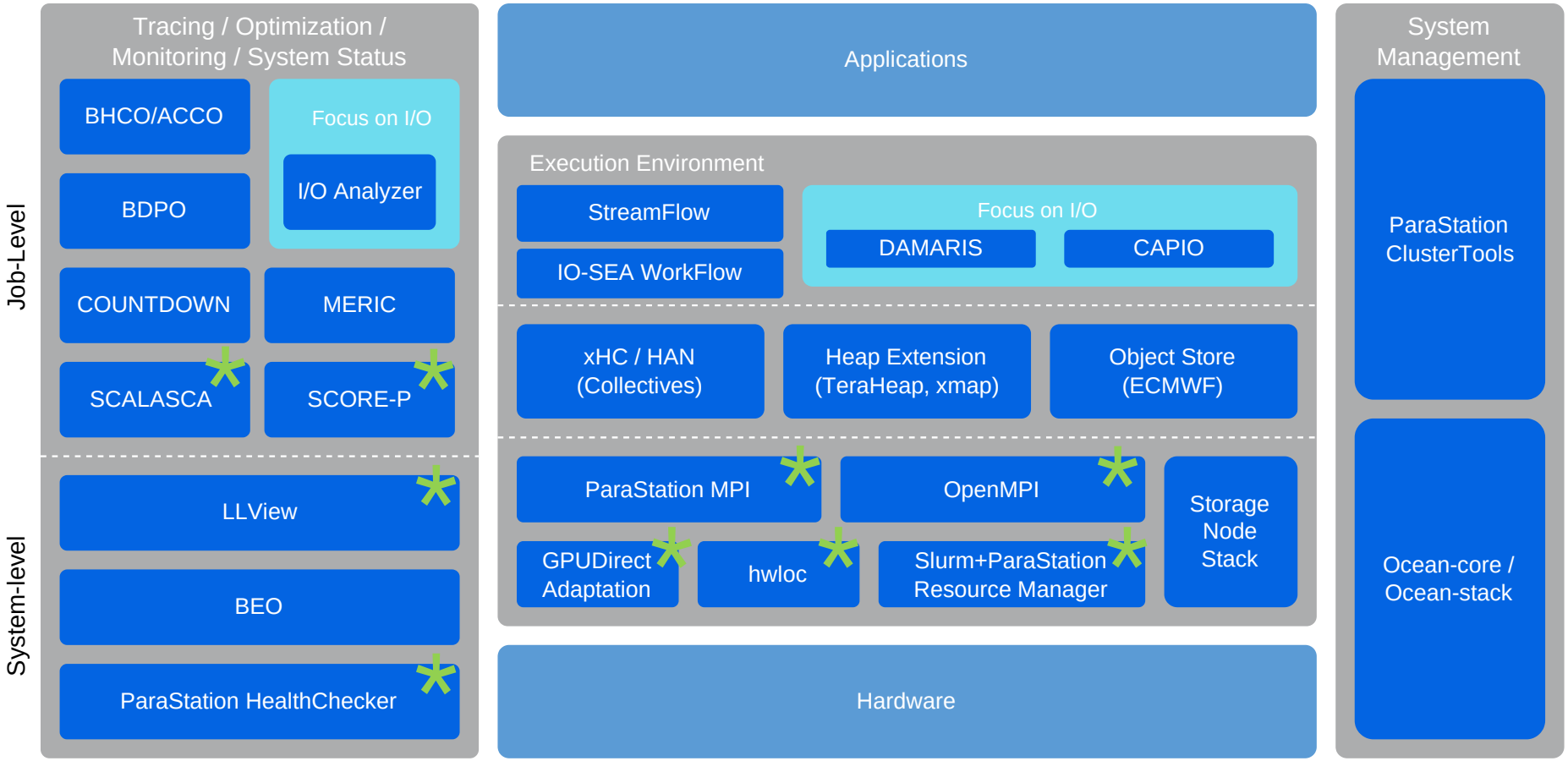
NVIDIA HPC

arm COMPILER

Slide courtesy of EVIDEN

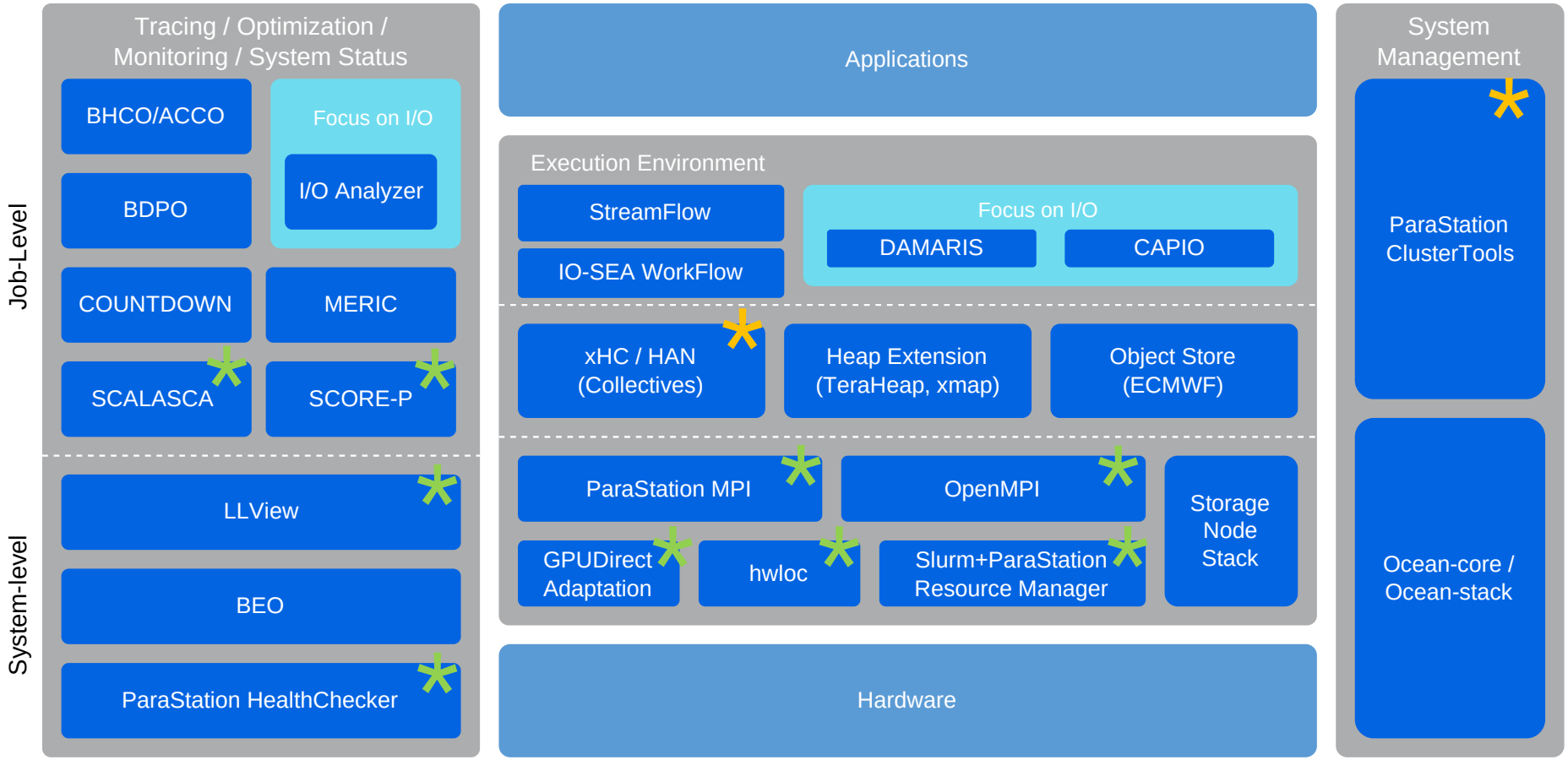
JÜLICH Forschungszentrum

WP5 – EUPEX Software Stack Overview



* Fully used and supported in JUPITER

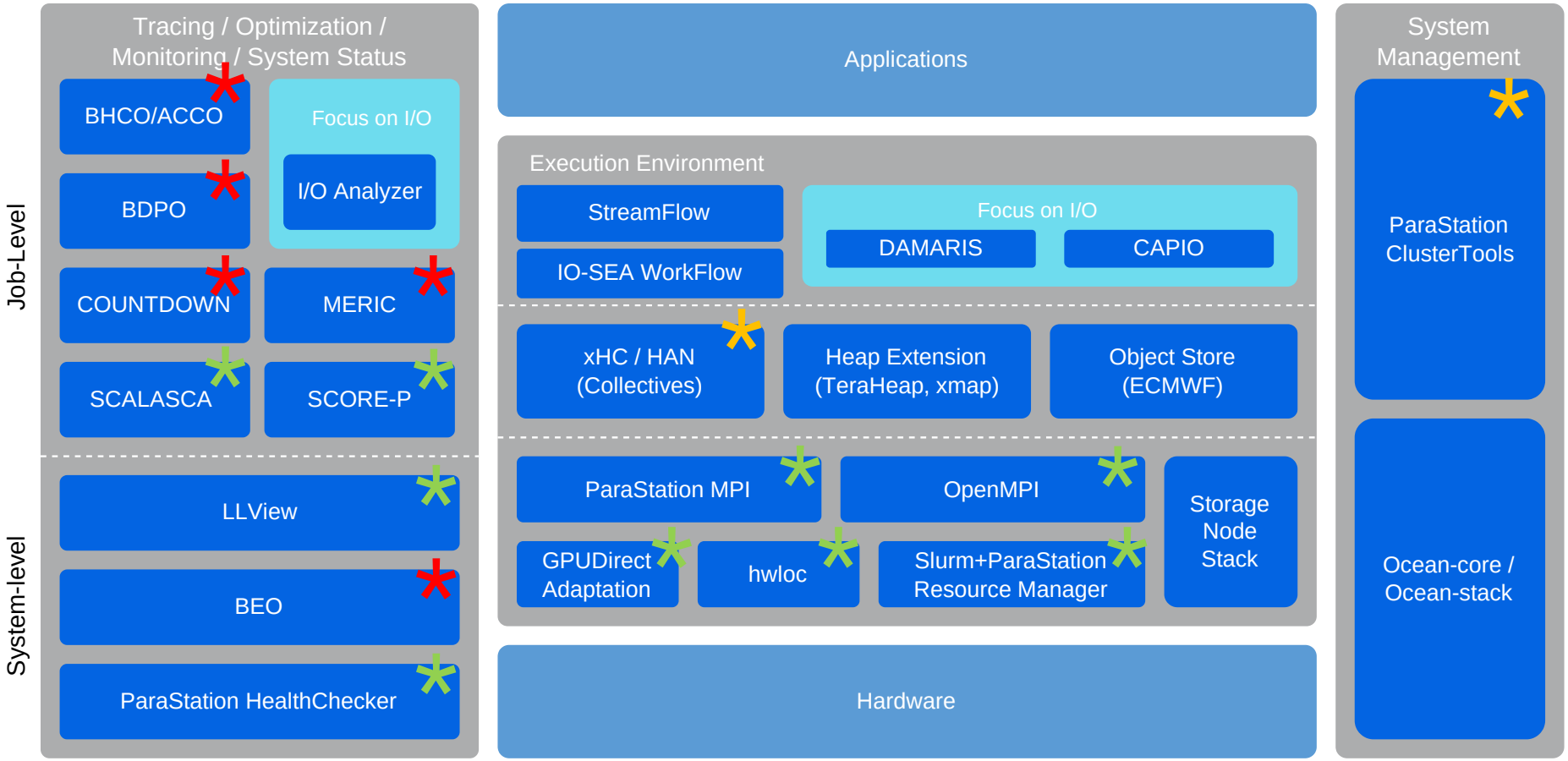
WP5 – EUPEX Software Stack Overview



* Fully used and supported in JUPITER

* Part of it used and supported in JUPITER

WP5 – EUPEX Software Stack Overview



* Fully used and supported in JUPITER

* Part of it used and supported in JUPITER

* Eventually available in JUPITER

JUPITER

The Arrival of Exascale in Europe

fz-juelich.de/jupiter | #exa_jupiter

EuroHPC Joint Undertaking

Federal Ministry of Education and Research

Ministry of Culture and Science of the State of North Rhine-Westphalia

GCS Gauss Centre for Supercomputing

JOINING FORCES

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

fz-juelich.de/jupiter