ParTec
Modular Supercomputing

Scaling Innovation

Modular Computing & QPUs

Forum Teratec

June 22, 2021

Thomas Moschny
CTO, ParTec AG

ParTec enables HPC

- Strong general purpose cluster specialist for more than two decades
  - *ParaStation research project: 1995 (Univ. of Karlsruhe)*
  - *ParTec founded as a spin-off in 1999*
  - *HPC full service provider since 2004*
- Cooperation with Jülich Supercomputing Centre since 2004
  - *ParaStation Consortium founded in 2005*
- Pioneering and enabling Modular Supercomputing
  - *Since 2010: DEEP Projects*
  - *ParaStation Modulo Software Suite*
- Significant contributions in European research projects
  - *Exascale-related: *-SEA Projects, EUPEX*
  - *Quantum- and AI-related: HPCQS, CoE RAISE*

ParTec enables HPC

- ParaStation Modulo Software Suite
  - *Software for HPC Systems developed for >20 years*
  - *Pioneering the Modular Supercomputing Architecture (MSA) for >10 years*
  - *Extensively used in production environments*
  - *Platform for research projects*
- ParTec Support: on-site (or remote) system operations
  - *System setup and installation*
  - *System maintenance and administration*
  - *General 1st and 2nd level support*
- Co-design and co-development
  - *Transferring results from research projects into production*
  - *Enhancing production systems over their lifetime*

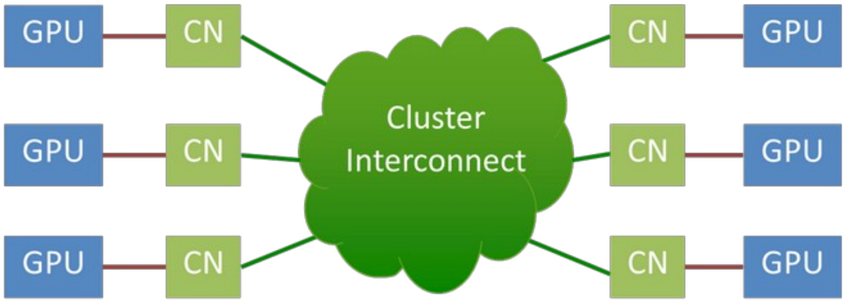
ParaStation Modulo

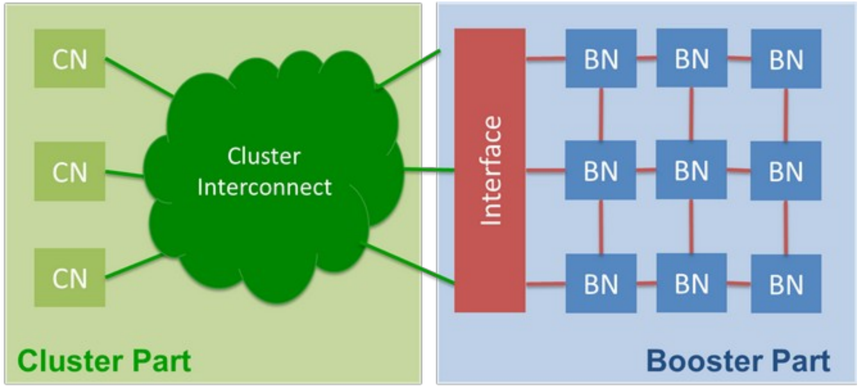
- ParaStation ClusterTools
  - *Tools for provisioning and management*
- ParaStation HealthChecker & TicketSuite
  - *Automated error detection & error handling*
  - *Ensuring integrity of the computing environment*
  - *Keeping track of issues*
  - *Powerful analysis tools*
- ParaStation MPI & Process Management
  - *Runtime environment specifically tuned to the largest distributed memory supercomputers*

ParaStation MODULO

Maximize job throughput –
Minimize administration effort

System-level Heterogeneity developed in DEEP Projects



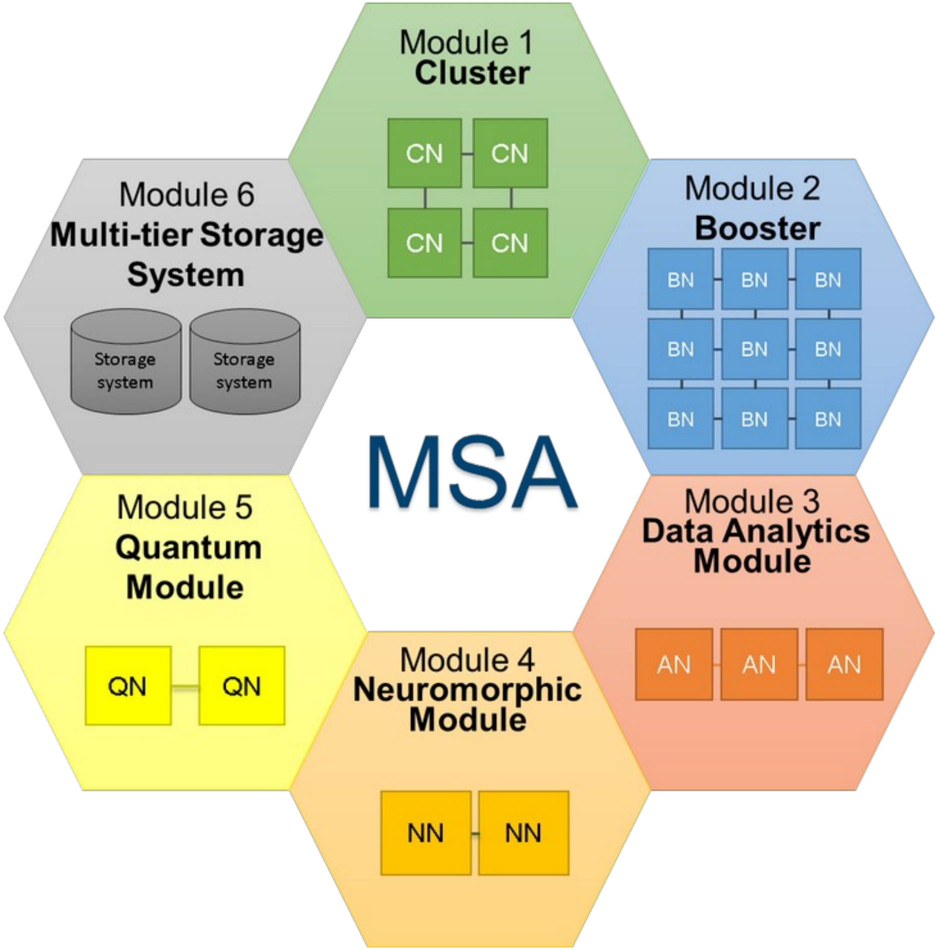


- Accelerated Cluster
  - *Fixed, static ratio and assignment of accelerators to CPUs*
  - *Static management of resources*
  - *Accelerators do not act autonomously*
  - *General-purpose Cluster interconnect*
  - *Programming via local offload interfaces*

- Cluster-Booster Architecture
  - *No fixed ratio or assignment between resources (multicore & manycore nodes)*
  - *Dynamic management and association of resources*
  - *High-throughput network in the Booster*
  - *Programming via MPI and “global” tasking interfaces*

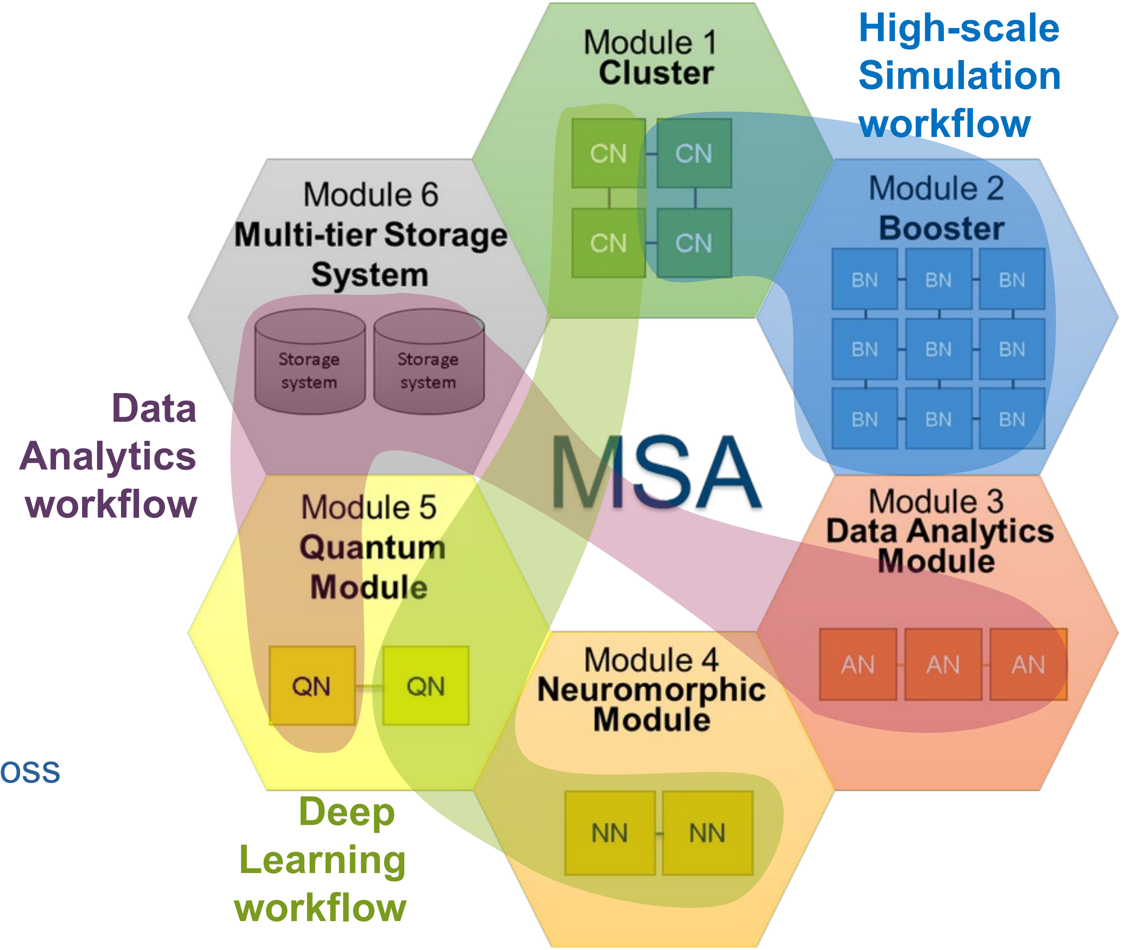
Modular Supercomputing Architecture

- Generalization of the Cluster-Booster Concept
  - *Composability of heterogeneous resources*
  - *Effective resource-sharing*
- Any number of (specialized) modules possible
  - *Cost-effective scaling*
- Fit application diversity
  - *Large-scale simulations*
  - *Data analytics*
  - *Machine/Deep Learning, AI*
  - *Hybrid Quantum Workloads*
- Achieve leading scalability & energy efficiency → Exascale
- Unified SW environment to run applications across all modules
  - *ParaStation Modulo providing a Slurm-based Scheduler*



Modular Supercomputing Architecture

- Generalization of the Cluster-Booster Concept
  - *Composability of heterogeneous resources*
  - *Effective resource-sharing*
- Any number of (specialized) modules possible
  - *Cost-effective scaling*
- Fit application diversity
  - *Large-scale simulations*
  - *Data analytics*
  - *Machine/Deep Learning, AI*
  - *Hybrid Quantum Workloads*
- Achieve leading scalability & energy efficiency → Exascale
- Unified SW environment to run applications across all modules
  - *ParaStation Modulo providing a Slurm-based scheduler*



JUWELS – A Modular Supercomputer

Cluster Module

© Forschungszentrum Jülich

Booster Module

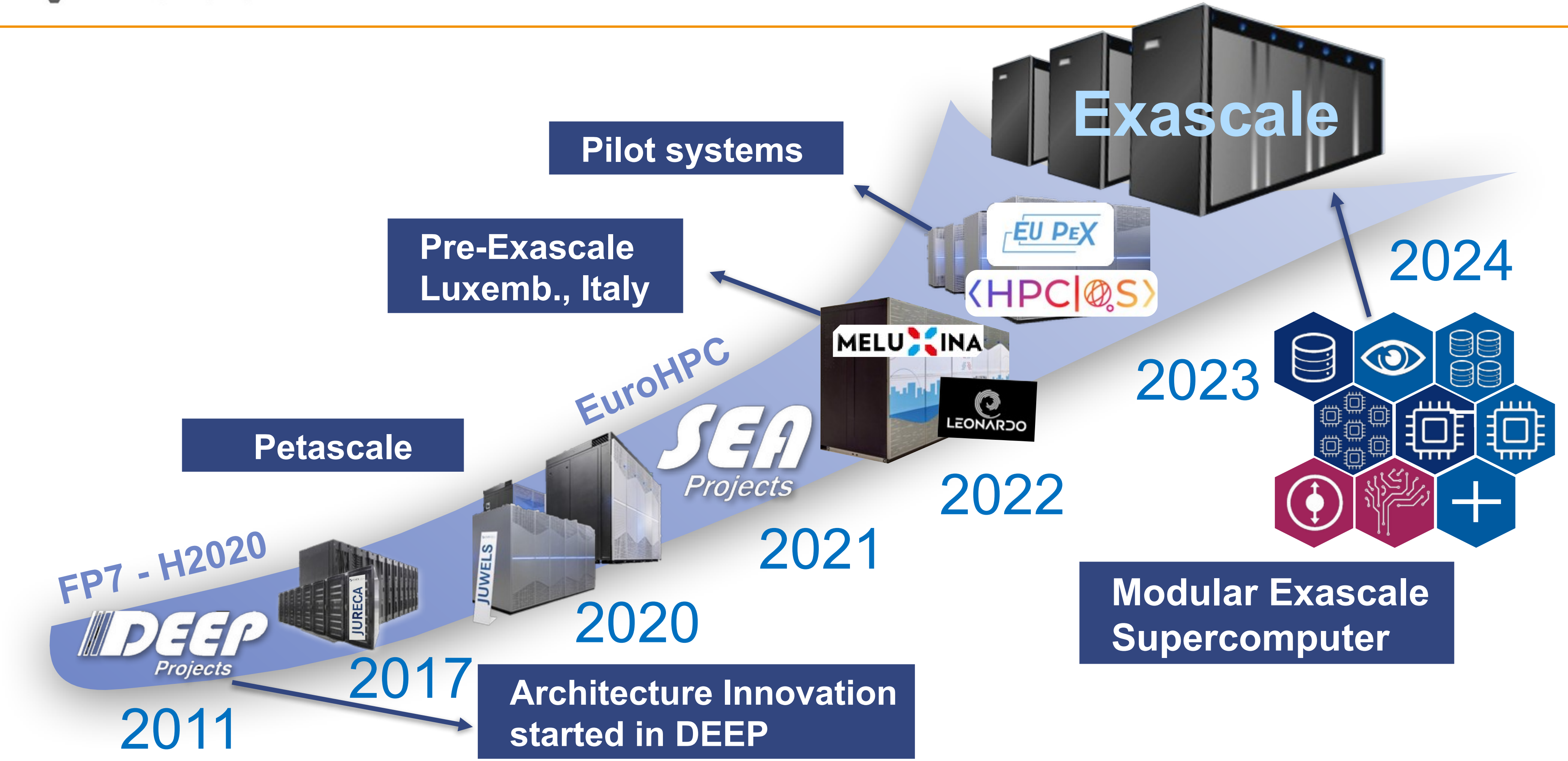
© Forschungszentrum Jülich/Ralf-Uwe Limbach

- 12 PFlop/s peak
- #23 on Top500 list (June 2018)
- 2575 nodes (Bull Sequana X1000)
- Intel Xeon Platinum 8168 / Gold 6148
- Mellanox EDR, ParaStation MPI

- GPU-accelerated module, 70 PFlop/s peak
- #7 on Top500, #3 on Green500 (Nov. 2020)
- 936 nodes (Bull Sequana XH2000)
- 4x NVIDIA A100 GPUs per node
- Quad-rail Mellanox HDR200, ParaStation MPI

Operated as one Modular System with ParaStation Modulo and Slurm

Modular Supercomputing to Exascale



2011

2017

2020

2021

2022

2023

2024

Petascale

Pre-Exascale
Luxemb., Italy

Pilot systems

Modular Exascale
Supercomputer

Architecture Innovation
started in DEEP

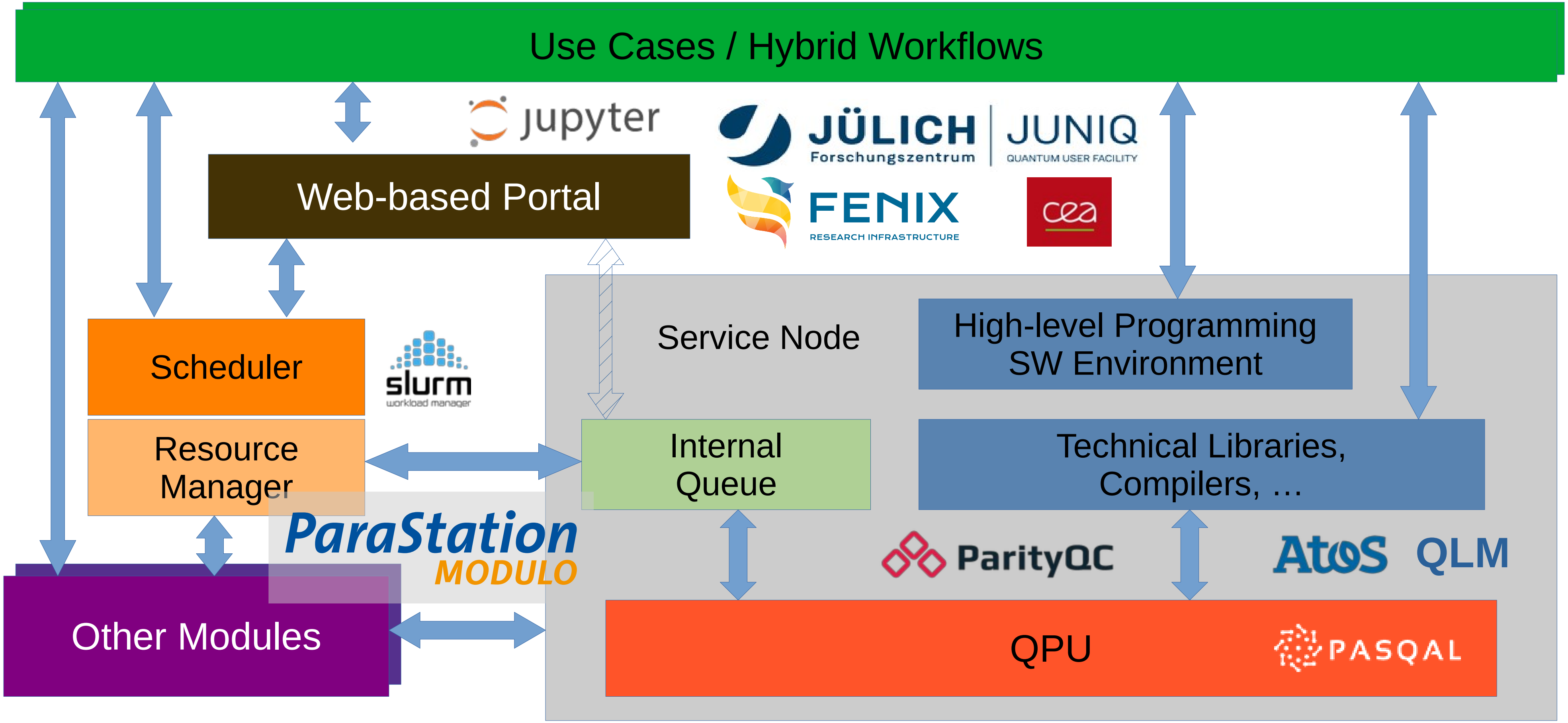
Hybrid Quantum Computing

- Applying the MSA approach
  - *Integrate the QPU as a new module type into the supercomputer*
- Implementation aspects
  - *Integration of the QPU and its front-end into the full management stack of the modular supercomputer, including user and SW management, storage access, provisioning, and more*
  - *Low-latency connection to other modules via federated, high-speed network*
  - *Integration in the scheduling and resource management on the system level*
- Benefit: New usage models
  - *Tightly coupled simulations: benefit from efficient data exchange*
  - *Workflows exhibiting one or more stages on the QPU and doing pre- and post-processing tasks on other modules*
  - *Unified environment (due to the tight integration: user and SW management, storage, …)*

QPU-to-MSA integration challenges

- The QPU is a scarce resource: it cannot be used concurrently by multiple users
  - *Implement a pseudo-shared usage model, e.g., based on time slices*
  - *Enable communication between the internal queue of the QPU and the system-wide scheduler/resource manager via well-defined interfaces*
- Provide “direct” access of the QPU via the web-based portal
  - *Redirect portal requests through the global scheduler/resource manager*
  - *Pseudo-shared usage model as prerequisite*
- Exact requirements/timings depend on the use case and are subject to research

MSA Integration of the QPU



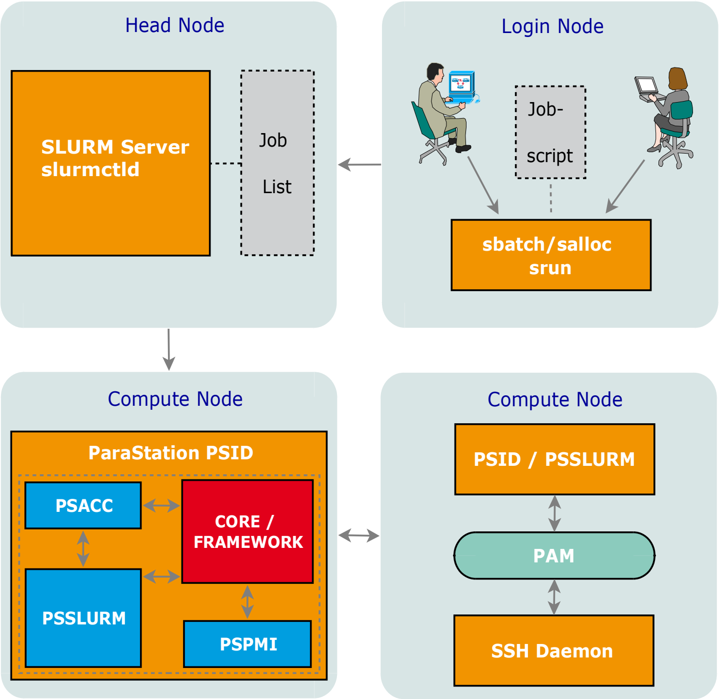
Thank you for your attention!

Questions?

moschny@par-tec.com

ParaStation Process Manager

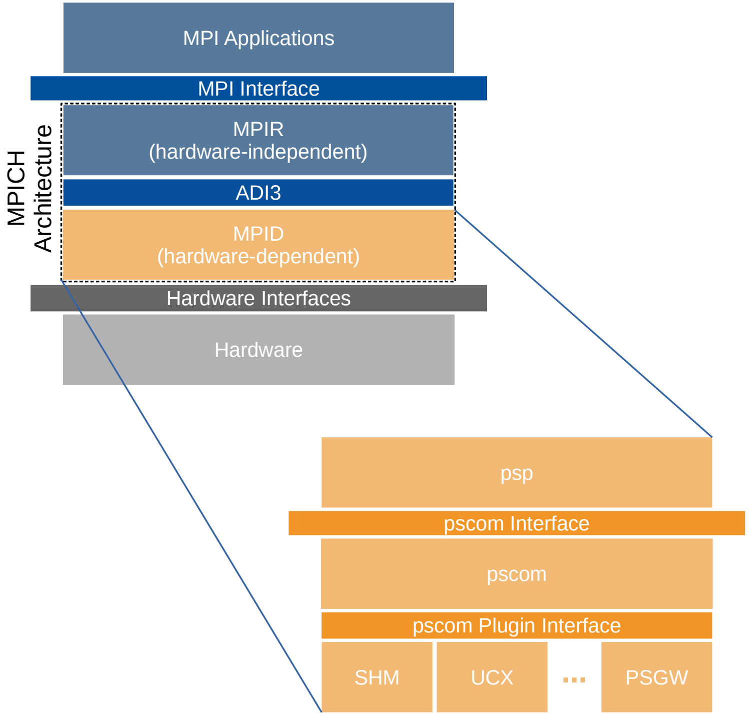
- Scalable network of MPI process management daemons
  - *Running on the computational nodes*
  - *Process startup and control, I/O forwarding, …*
  - *Precise resource monitoring*
  - *Proper cleanup after jobs*
- PSSLURM: Full integration for Slurm
  - *Plugins to the ParaStation Management daemons*
  - *Replace node-local Slurm daemons (also reduces number of daemons)*
  - *Enforces resource limits*



ParaStation MPI

ParaStation MPI Architecture

- Based on MPICH 3.3.2 (MPI-3.1 compliant)
  - *Maintains MPICH ABI compatibility*
  - *Supports MPICH tools (tracing, debugging, …)*
  - *MPICH layers beneath ADI3 are replaced by ParaStation PSP Device*
  - *Powered by pscom low-level communication library: non-blocking p2p semantics*
- Support for various transports and protocols via pscom plugins
  - *Support for InfiniBand, Omni-Path, Extoll, …*
  - *Applications may use multiple transports / plugins at the same time*
  - *Gateway capability via PSGW plugin to bridge transparently between different networks*
  - *CUDA awareness for all transports / CUDA optimization via GPUDirect for UCX, and Extoll*
- Proven to scale up to ~3,500 nodes and ~140,000 processes per job



ParaStation MPI