ParaStation MPI

MPICH BoF SC‘22
November 16th, 2022

Simon Pickartz, ParTec AG

PARASTATION MODULO SOFTWARE SUITE

ParTec
MODULAR SUPERCOMPUTING

ParaStation CLUSTERTOOLS

ParaStation HEALTHCHECKER

ParaStation TICKETSUITE

ParaStation MPI

Tools for Provisioning and Management

- System management CLI
- Image management
- Rolling updates
- Stateless & stateful booting
- Post-install configuration
- Slurm integration
- Distributed database for system configuration
- HealthChecker integration

Integrity of the Computing Environment

- Automated error detection & error handling
- Various hook-in points
- No interference with jobs
- TicketSuite integration
- Highly configurable

- 100+ tests (HW/SW):
- Node/System/Fabric level

Issue Tracking on System Level

- Manual and automatic ticket creation
- Prioritization
- Routing/Triage
- Documentation and central information hub
- Maintenance planning
- Interfaces with external ticketing systems

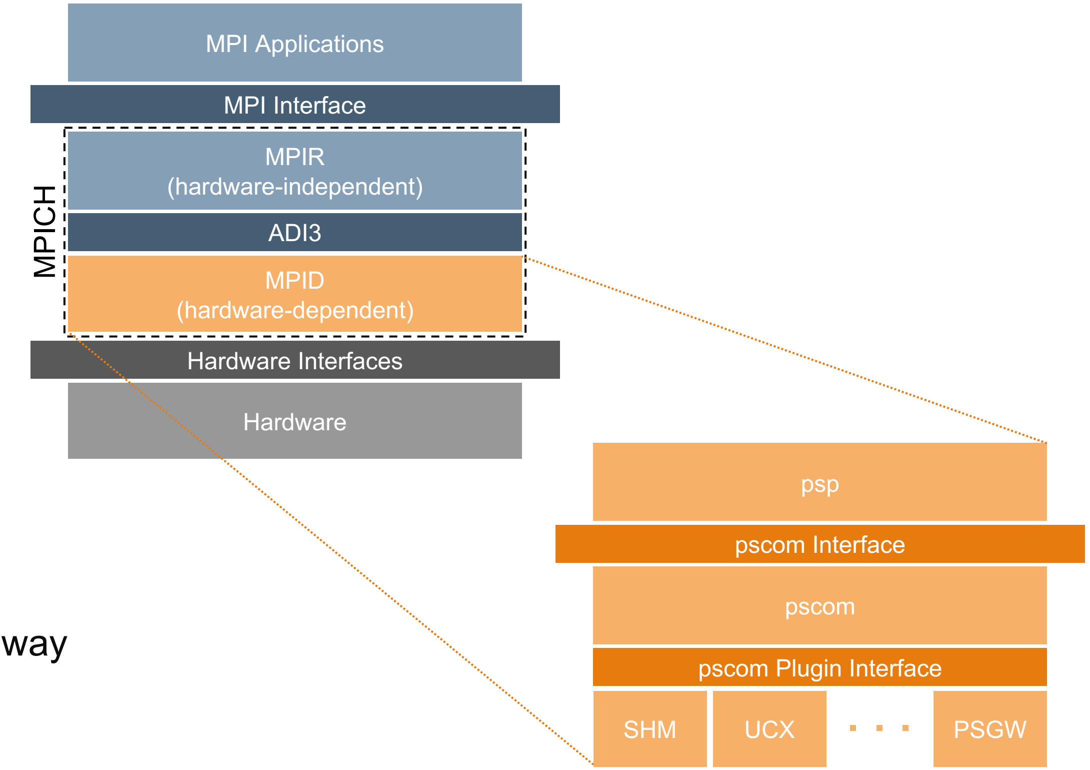
Execution Environment and MPI Library

- MPI 3.1 compliant (MPI 4 support soon)
- MPICH ABI compatible
- Supports multiple interconnects in parallel
- Modularity support
- Network bridging
- PMIx support
- Full Slurm integration

PARASTATION MPI

ARCHITECTURE

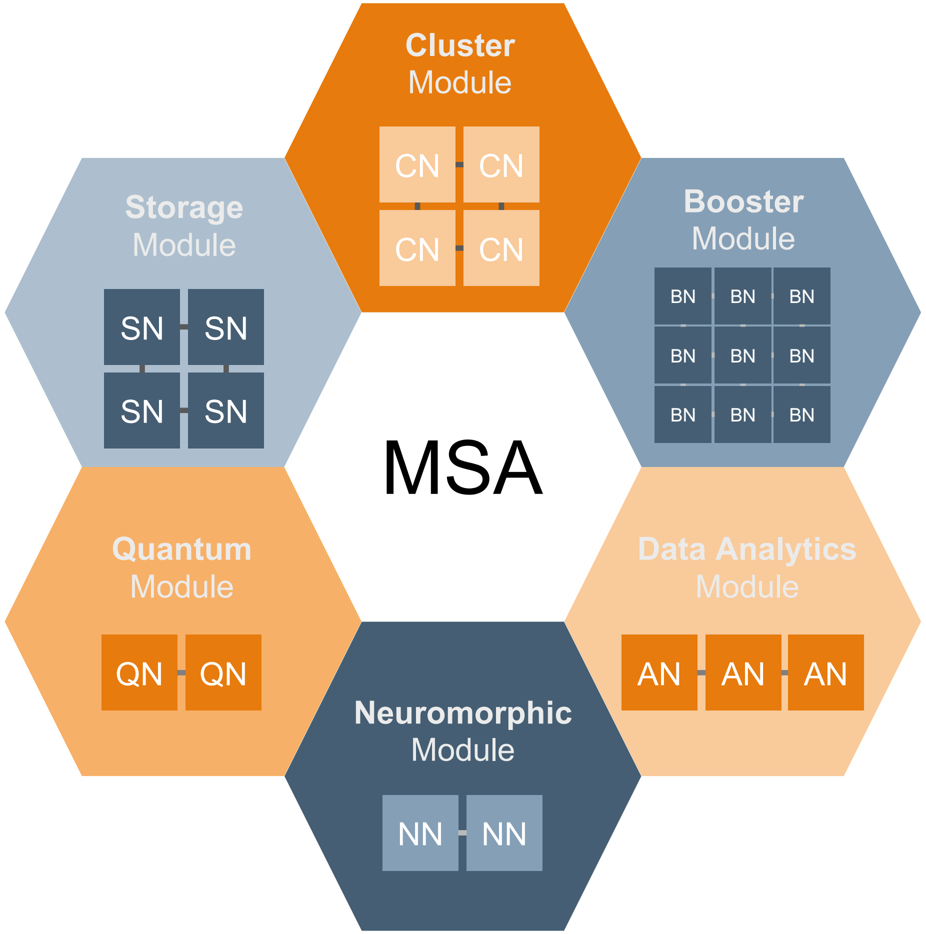
- Based on MPICH 3.4.3 (MPICH 4 coming very soon!)
  - Support MPICH tools for tracing, debugging, etc.
  - Integrates into MPICH on the MPID layer by implementing an ADI3 device
  - The PSP Device is powered by pscom – a low-level point-to-point communication library
  - Support the MPICH ABI Compatibility Initiative
- Support for various transports / protocols via pscom plugins
  - Support for InfiniBand, Omni-Path, BXI, etc.
  - Concurrent usage of different transports
  - Transparent bridging between any pair of networks enabled by gateway capabilities
- Proven to scale up to ~3,500 nodes and ~140,000 processes per job



ParaStation MPI

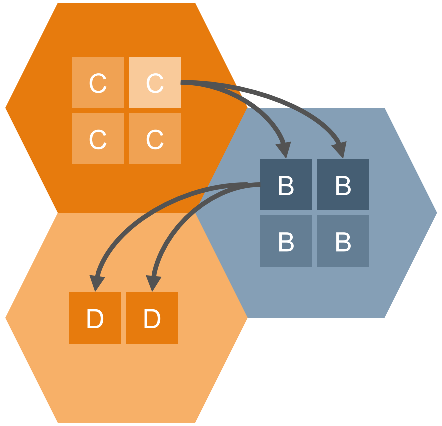
MODULAR SUPERCOMPUTING ARCHITECTURE

- Generalization of the Cluster-Booster Concept
  - Heterogeneity on the system level
  - Effective resource sharing
- Any number of (specialized) modules possible
  - Cost-effective scaling
  - Extensibility of existing modular systems by adding modules
- Fit application diversity
  - Large-scale simulations
  - Data analytics
  - Machine/Deep Learning, AI
  - Hybrid-quantum Workloads
- Achieve leading scalability and energy efficiency
  - Exascale-ready!
- Unified software environment for running across all modules
  - Enabled by the ParaStation Modulo software suite

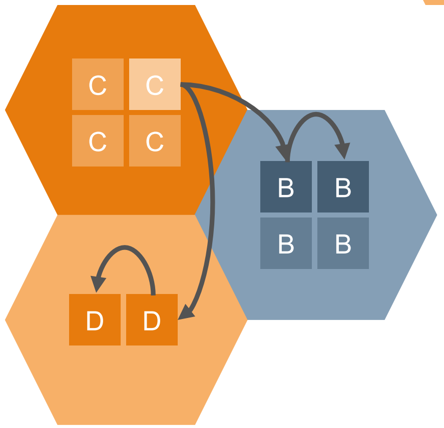


MSA AWARENESS

- Support for multi-level hierarchy-aware collectives
  - Optimize communication patterns to the topology of the MSA
  - Assumption: Inter-module communication is the bottleneck
  - Dynamically update the communication patterns (experimental)
- API extensions for accessing modularity information
  - New MPI split type for communicators (`MPIX_COMM_TYPE_MODULE`)
  - Provide the module id via the `MPI_INFO_ENV` object
- MPI Network Bridging
  - Connect any pair of interconnect and protocol
  - Transparent to the application layer



Binary Tree
(default)



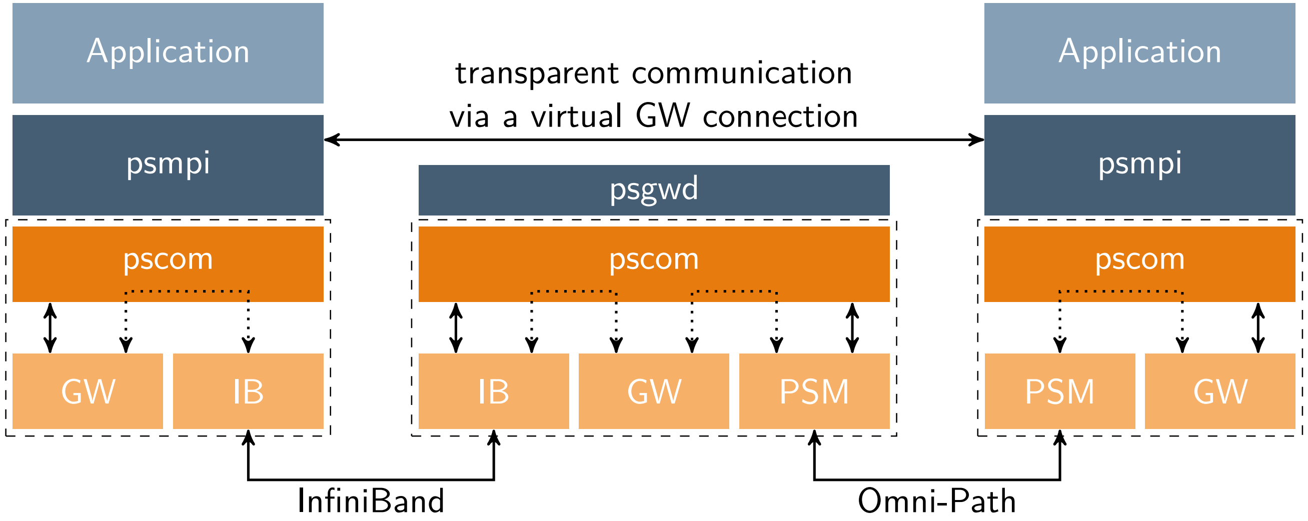
Hierarchical
(MSA-aware)

MPI NETWORK BRIDGING

- Transparent communication across networks
  - Use a gateway when two processes are not directly connected through the same network
  - Bridging between any pair of interconnects supported by pscom (e.g., InfiniBand, Omni-Path, BXI, etc.)

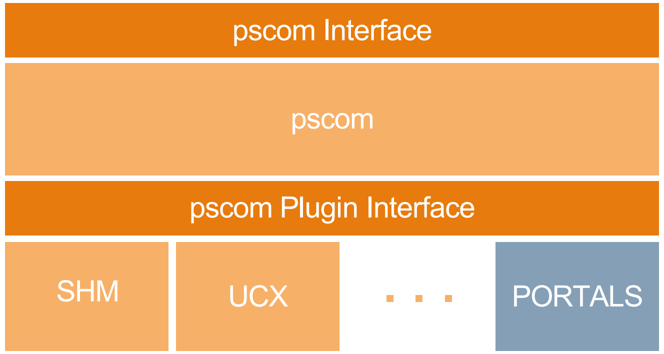
- Static routing
  - Use the same gateway for different destinations
  - Virtual GW connections provide full transparency to the application layer

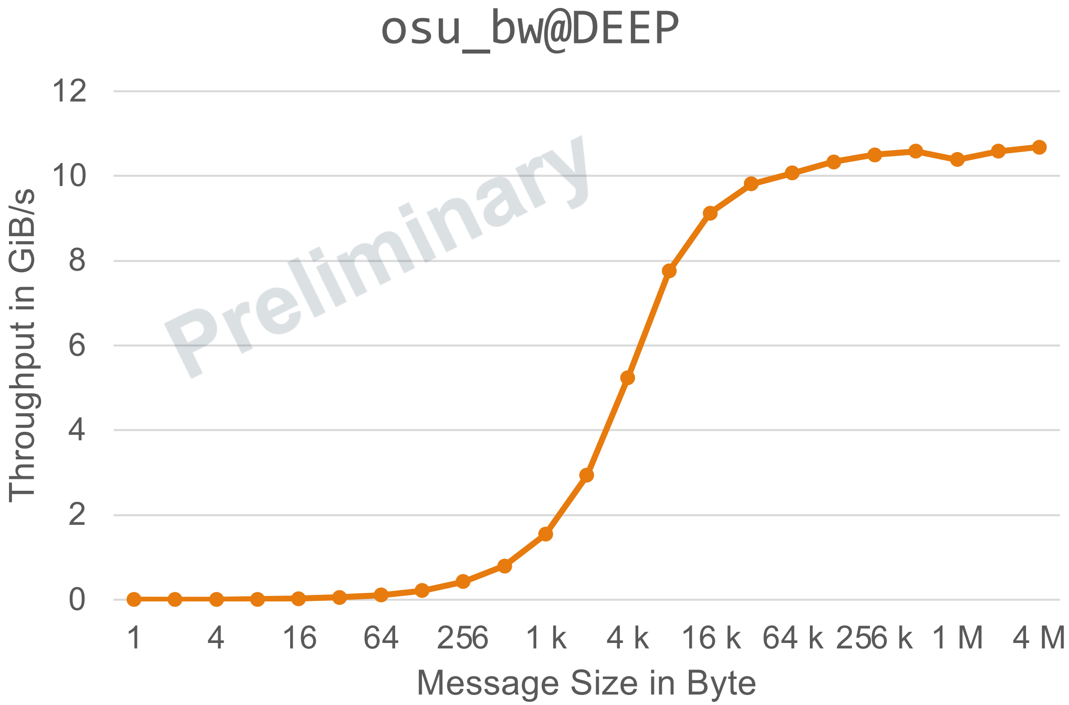
- Successfully deployed in production environments
  - Implemented first for the JURECA Cluster-Booster System
  - Bridging between Mellanox EDR and Intel Omni-Path



SUPPORT FOR BXI NETWORKS

- Integrated as a new plugin into pscom
  - Benefits from existing infrastructure
  - Support for transparent network bridging in federated networks
- Communication modes
  - Low-latency *eager* communication for short messages
  - High-throughput *rendezvous* communication for mid-size to large messages
- Fine-tuning via environment variables





- Intel® Xeon® Gold 5122
- 4 Nodes
- 1 Socket per Node
- 4 Cores per Socket
- 48 GiB Ram per Node
- BXI 1.3 Interconnect

WHAT‘S NEXT?

CURRENT AND FUTURE DEVELOPMENTS

OPTIMIZATION

- Performance optimizations (e.g., further improve BXI support)
- Expose low-level RMA for improved one-sided communication
- Extend support for hierarchical collectives (e.g., UCC support)

MPI-4

- Integration of MPICH 4.x upstream sources
- Improve/extend MPI-4 support
- Tighter integration with the process manager (e.g., for the provision of psets)
- Bring developments upstream

MALLEABILITY

- Dynamic resizing of jobs
- Support for application-driven *(active)* and scheduler-driven *(passive)* malleability
- Leverage PMIx (e.g., `PMIx_Allocation_request`)
- Build upon the MPI Sessions interface

THANK YOU FOR YOUR ATTENTION

QUESTIONS

ParTec AG, Possartstr. 20, D-81679 München – www.par-tec.com

{pickartz, moschny, clauss}@par-tec.com