ParaStation MPI

MPICH BoF SC‘23
November 15th, 2023

Simon Pickartz, ParTec AG

PARASTATION MODULO SOFTWARE SUITE

ParTec
MODULAR SUPERCOMPUTING

ParaStation CLUSTERTOOLS

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

ParaStation HEALTHCHECKER

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

ParaStation TICKETSUITE

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

ParaStation MPI

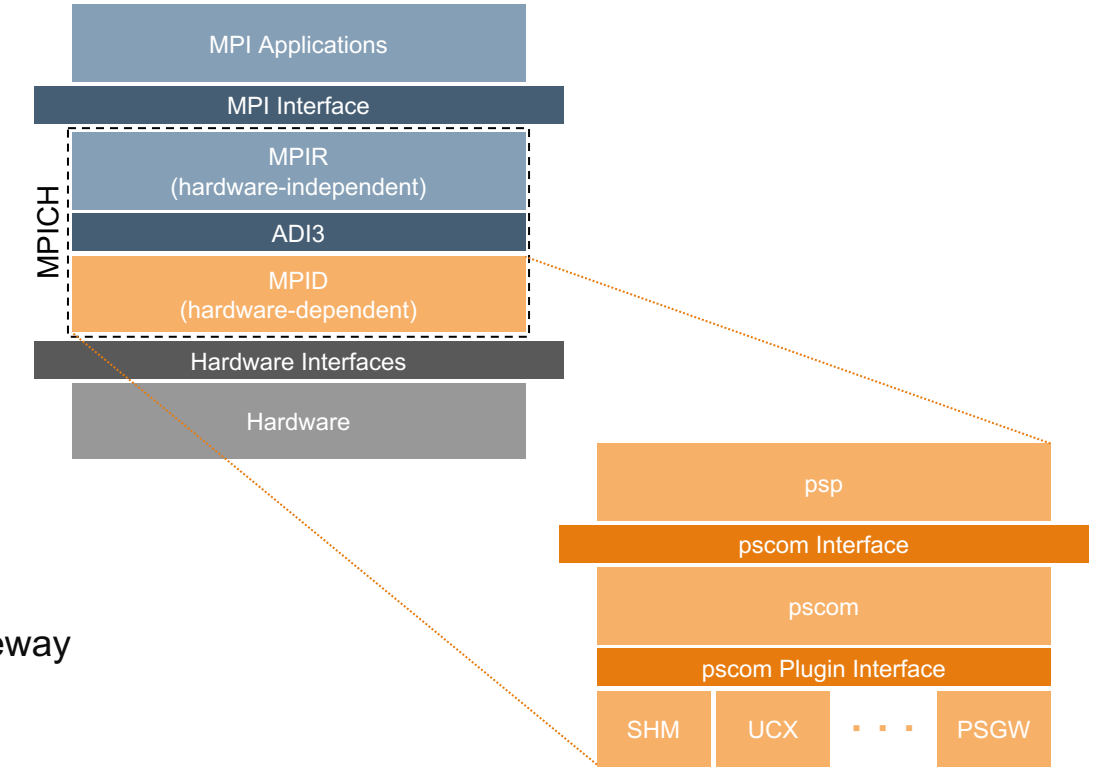
Execution Environment and MPI Library

- MPI-4.0-compliant
- MPICH ABI-compatible
- Supports multiple interconnects in parallel
- Modularity support
- Network bridging
- PMIx support
- Full Slurm integration

PARASTATION MPI

ARCHITECTURE

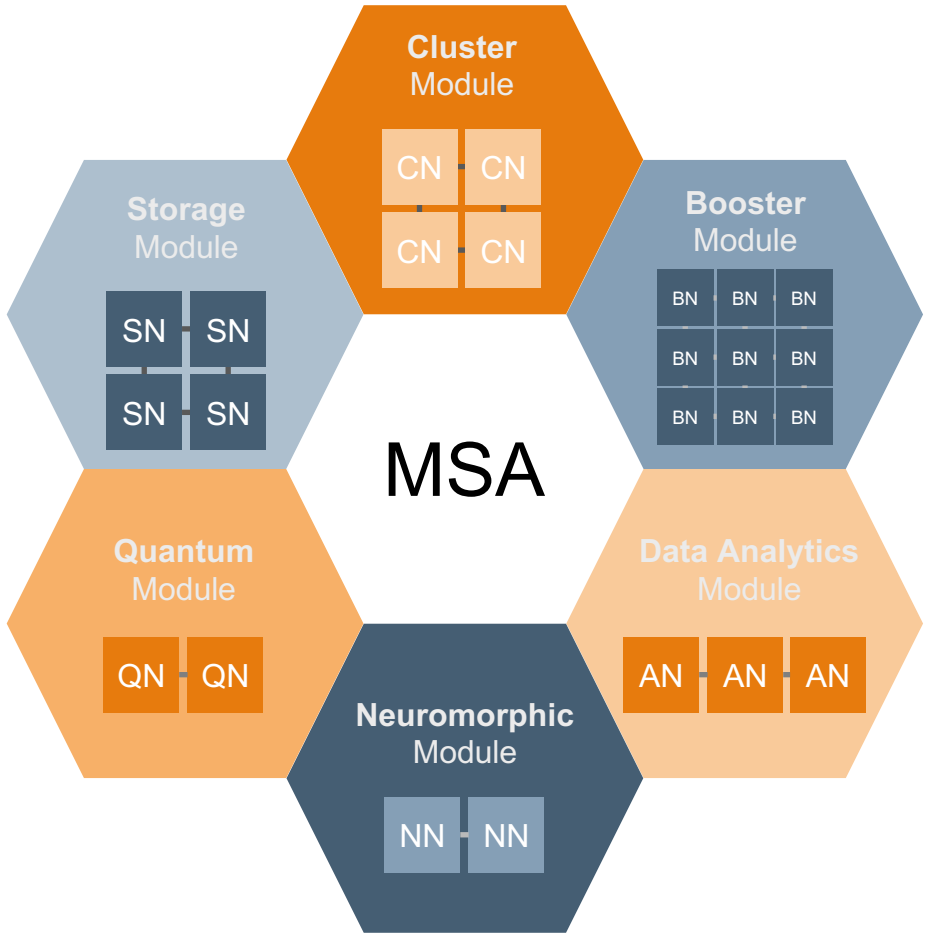
- Based on MPICH 4.1.2
  - Support MPICH tools for tracing, debugging, etc.
  - Integrates into MPICH on the MPID layer by implementing an ADI3 device
  - The PSP Device is powered by pscom – a low-level point-to-point communication library
  - Support the MPICH ABI Compatibility Initiative
- Support 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

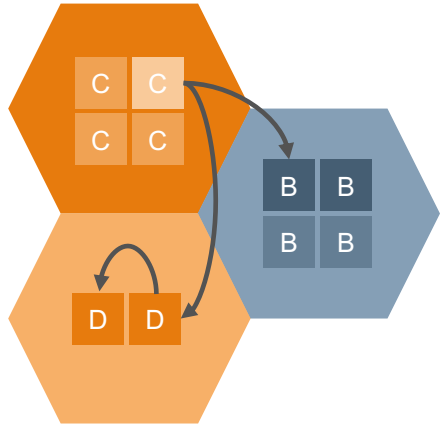
MODULAR SUPERCOMPUTING ARCHITECTURE

- 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

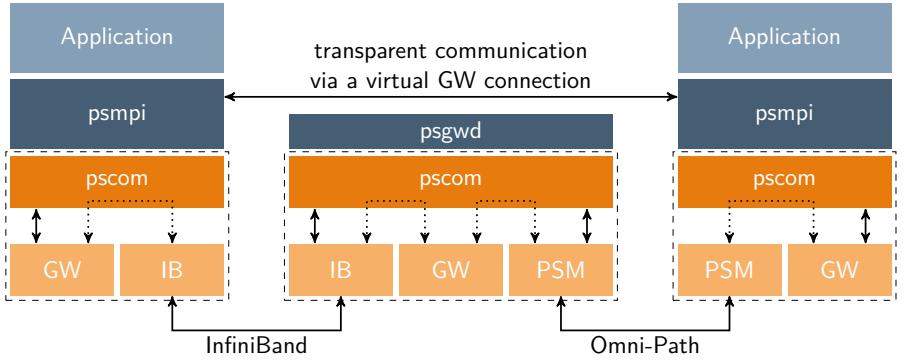


MSA AWARENESS

- 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



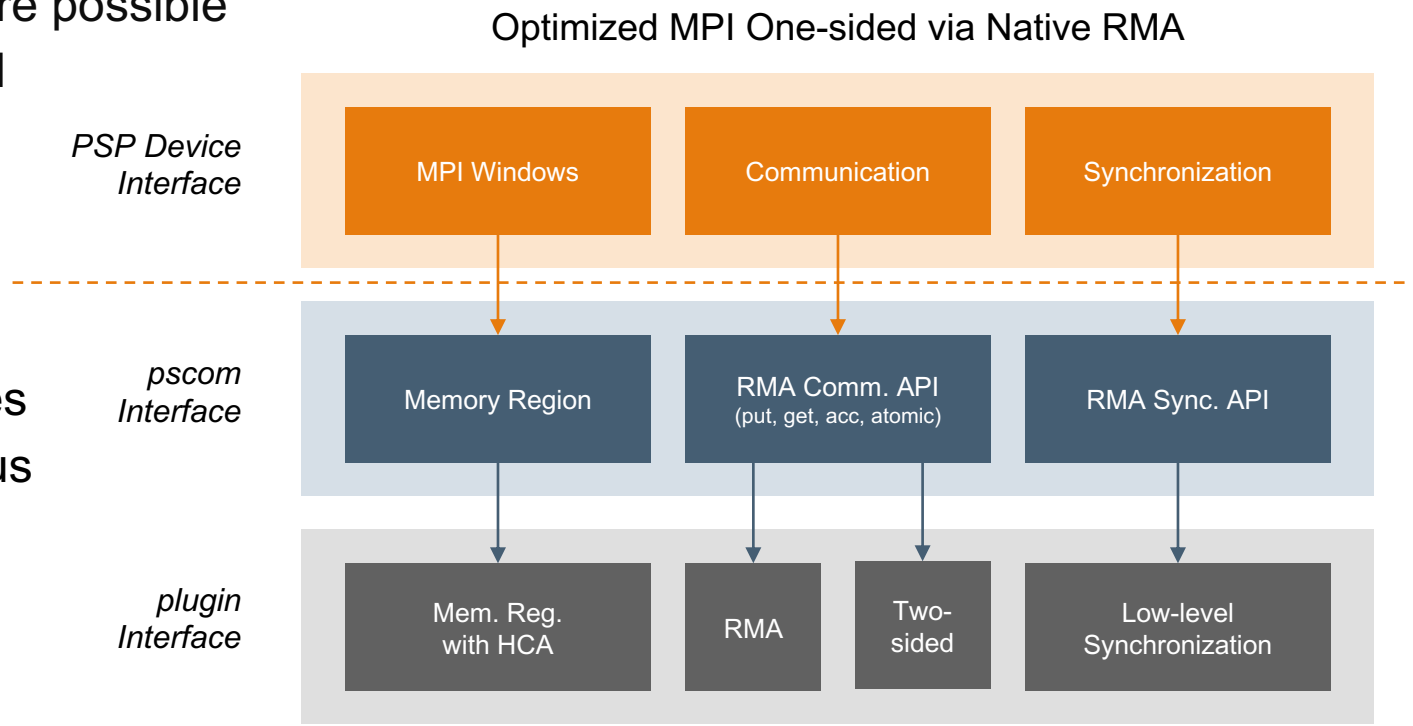
Hierarchical
(MSA-aware)



Transparent Network Bridging

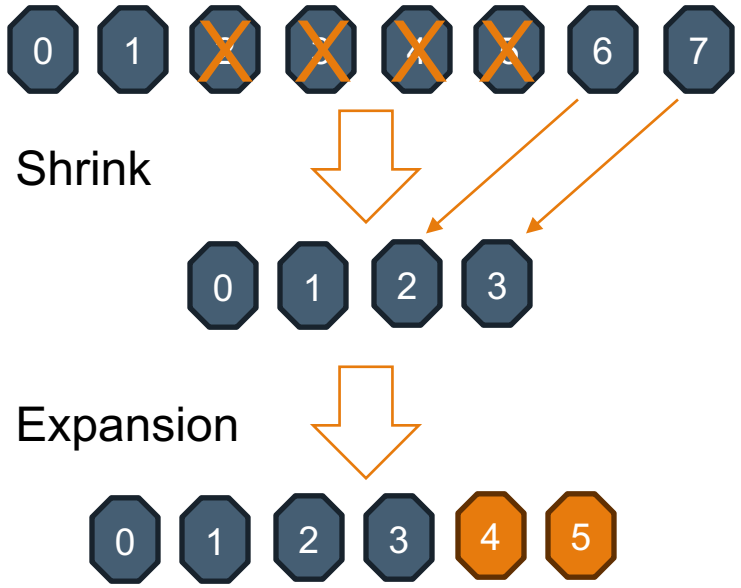
MPI ONE-SIDED VIA NATIVE RMA

- Optimize MPI one-sided communication
  - Leverage hardware capabilities where possible
  - Avoid overheads of two-sided-based implementations
- Implementation on the pscom level
  - Provide upper layers (i.e., PSP) with direct access to hardware capabilities
  - Generic RMA interface for the various transports supported by pscom
  - Provide two-sided-based fallback



MALLEABILITY FOR MPI

- Dynamic resource adaptations within an MPI application
  - Adding or removing of HPC resources during job run time
  - Ensure maximum MPI standard compliance
  - Exploit MPI-4 features (e.g., MPI Sessions)
  - Dense, monotonic MPI rank numbering (i.e., no gaps or overlaps)
- Usage Models
  - Job-initiated (according to current job needs)
  - Scheduler-initiated (maximize system utilization)
  - Externally initiated (based on application models)
- Initially, focus on Job-initiated malleability



DEEP-SEA
https://deep-projects.eu/

MALLEABILITY IN PARASTATION MPI

MALLEABILITY-RELATED DEVELOPMENTS

MPI SESSIONS

- Reference counting & non-standard strict session finalization
- Decoupling from MPI world model
- Re-initialization of the MPI library
- Error handling

PMIx SUPPORT

- PMIx Process Sets for MPI Sessions
- PMIx Spawn support

COMMUNITY ENGAGEMENT

- Included in ParaStation MPI as of release 5.9.2-1 (Sep '23)
- Many enhancements already merged upstream

WHAT'S NEXT?

CURRENT AND FUTURE DEVELOPMENTS

Malleability

- Improve and test MPI extensions for malleability
- Tight integration with the ParaStation Process Manager via PMIx

Optimizations

- Extend support for hierarchical collectives (e.g., UCC support)
- Performance optimizations (e.g., further improve BXI support)
- Improve RMA synchronization

MPI-4.1

- Integration of MPICH 4.2 upstream sources
- Provide MPI-4.1 support

Standardization

- MPI Extensions
- PMIx Extensions

THANK YOU FOR YOUR ATTENTION

QUESTIONS

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

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