

ParaStation V5

What is ParaStationV5?

ParaStationV5 is a comprehensive middleware-solution for compute clusters and farms. It provides a stable, reliable platform for commodity clusters and produces measurable productivity gains. User-friendly interfaces and cluster provisioning tools greatly simplify cluster administration.

ParaStationV5 consists of several components which divide the complexity of cluster operations into several pieces to make life easy for users and administrators alike:

ParaStation MPI is a high performance MPI implementation with proven scalability in excess of 24,000 processes on more than 3,000 compute nodes, **ParaStation HealthChecker** takes care of the cluster's integrity in both hardware and software aspects, **ParaStation TicketSuite** is a ticketing system with a wiki page, allowing to track problems from when they appear until their solution and allows creating a knowledgebase for your cluster, **ParaStation GridMonitor** provides a graphical interface providing real-time information about the system's state, workload, etc.

ParaStation ClusterTools is a set of utilities making it easy to administer all node related tasks: Installation, replacement, update of nodes, and more

Communication

ParaStationV5 provides a standard interface for parallel applications - MPI (MPI-2). Intra-node communication is handled via a shared memory model (shmem), while inter-node communication can be achieved using any of the following interconnects:

- Myrinet (legacy only),
- Fast Ethernet, Gigabit Ethernet,
- 10G-Ethernet
- Infiniband (verbs)
- InfiniPath (psm)
- Extoll (rma)
- Quadrics QsNet

Optimized protocol-stack for Ethernet communication (P4Sock)

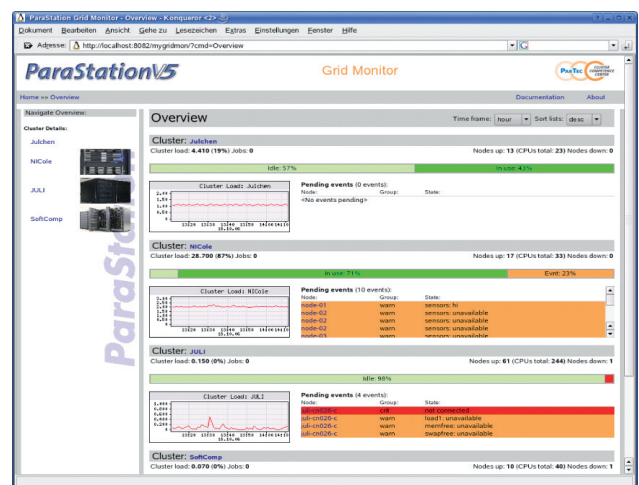
The P4Sock protocol for Ethernet is developed and optimized for cluster communication. It provides a secure, high performance, low latency inter-process data transfer protocol.

TCP-Bypass provides higher performance

TCP-Bypass routes standard TCP socket based communication directly to the high-performance P4Sock protocol layer. It is not necessary to modify or re-link the application.

The job-clean-up functionality cancels a job automatically if a hardware problem occurs. All processes belonging to a dedicated job are terminated and no orphan processes remain.

The pre-emption functionality allows operators to suspend running jobs, and allow higher priority job to proceed. The suspended jobs can be restarted by operator request.



Features:

- MPI-2 (Ethernet, Myrinet, Infiniband, InfiniPath, Extoll, P4sock, shmem, QsNet)
- Parallel Jobs: No need for users to provide node lists
- A single system view for managing multiple clusters
- Capable of managing heterogeneous clusters
- Single-point of cluster management
- Process management for parallel and serial jobs
- Administration command console (psadmin)
- Parallel shell (psh)
- High-performance parallel file copy
- Installation-Management (**ParaStation ClusterTools**)
- Pre/Post job node health diagnosis interaction with the batch system (**ParaStation HealthChecker**)
- Extensible system monitoring tools (**ParaStation GridMonitor**) collecting cluster parameters and

device values from compute nodes, switches, UPS, RAID etc.

- Comprehensive graphical user interface
- Event notification for „out-of-range“ system values
- Automated system installation tool
- Job accounting information
- Integration with standard batch systems (PBS/Torque/Moab/Maui/SGE)
- Multi platform support
- Multi vendor support
- Independent of Linux distribution

Daemon concept

- **ParaStation** daemon runs on each node (no rsh / ssh login necessary)
- Sophisticated process-management to start, monitor and close processes
- Distributed management logic
- Job suspend and resume
- fault tolerant – built in fail over

Monitoring

ParaStation GridMonitor allows cluster administrators to work with the command line or use a browser based user interface to monitor the following cluster state:

- Displays all running jobs and applications
- Displays the current load on the cluster
- Displays all relevant hardware parameters, devices, users and processes
- IPMI compatible
- Event notification

ParaStation HealthChecker

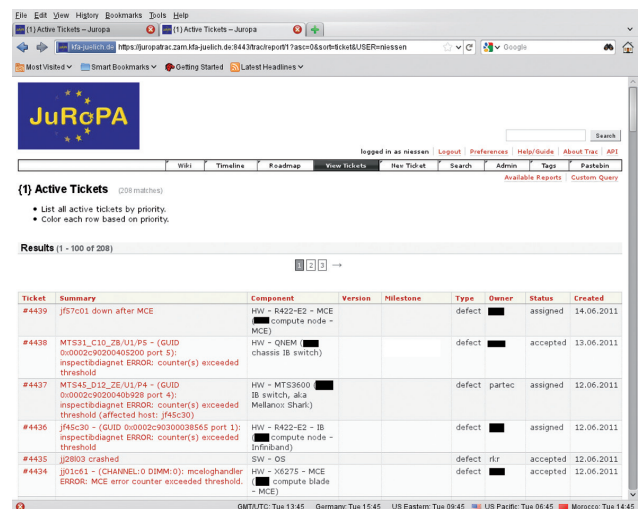
Surveys cluster health and removes faulty nodes from production, ensuring job stability

- Freely customizable and configurable
- Tells batch system about bad nodes
- Lightweight

ParaStation TicketSuite (see picture)

Trac/Wiki based web interface for issue tracking.

- Cross referencing of tickets
- Interfaces with batch system
- Reporting tools (statistics about node failures, software issues, ...)
- Mail interface



HW-Platforms

- Intel / AMD x86, x86-64, IA64 (legacy only),

SW-Platforms

- SuSE Linux Enterprise Server (SLES 10 and 11),
- openSUSE (11.x)
- RedHat based distributions (RedHat, CentOS, Scientific Linux)

Proven Reliability and Performance

ParaStation V5 helped to place JuRoPA cluster at Jülich Supercomputer Centre (Germany) into the global Top 10 (top500.org).

3288 compute nodes (26304 cores)

274,8 Teraflops sustained Linpack performance (June 2009)

Mellanox QDR Infiniband

ParaStation Operation and Management

ParaStation MPI SLES11 Linux

Intel Compiler and Development Tools

Lustre Filesystem

ParaStation MPI delivered proven scalability running more than 25,000 MPI tasks without any threading library with parallel efficiencies in excess of 91,6%.

Certifications



Intel Cluster Ready