ParaStation CLUSTERTOOLS

ParaStation ClusterTools

Software Product Detailed Description

Product: ParaStation ClusterTools

Date: May 2012

Document number: PSCT-1.0-2en

This software product description documents the functionality provided by the ParaStation ClusterTools as well as the system prerequisites required for installation and operation, licensing scheme and other useful information.

Overview

The ParaStation ClusterTools as part of the *ParaStationV5* cluster suite offer a set of tools to ease the different tasks of setting up and maintaining a high performance compute cluster. It supports a variety of Linux distributions and system configurations.

Beside a minimal setup for the compute nodes and a complete setup for the master node, a couple of packages are installed by default, transforming the bunch of servers into a high performance compute cluster:

- ParaStation MPI as a robust environment to run parallel and serial jobs within the cluster.
- Torque/Maui for batch processing of jobs within the cluster.
- ParaStation GridMonitor to monitor all parts and activities of the system.
- NFS to provide system-wide file systems for user and scratch data.
- SystemImager as a basis for system installation and updates.

Basic installation and configuration

Setting up a HPC cluster from scratch requires careful preparation - lots of packages have to be installed and configurations must be adjusted to form a homogeneous cluster. The ParaStation ClusterTools simplify these tasks to do essentially three different things:

- Set up a master node to provide all the information and services necessary to run a cluster.
- Set up one compute node ("golden client") to provide the basic computational infrastructure.
- Install and update many compute nodes with a previously defined setup.
- Alternatively the compute nodes can be booted as diskless nodes thus rendering a golden client unnessary.

Furthermore, the ParaStation ClusterTools may be used to update compute nodes to new or modified configurations.

The entire cluster is managed from a so called master node. This node runs all the necessary services to organize all nodes into a homogeneous high performance compute cluster. The master node and the compute nodes are interconnected by an administration network, typically Gigabit Ethernet using one internal NIC. Providing NAT, the master node also acts as a gateway for the compute nodes. Usually only the master node is connected to a higher-ranking network.

On a typical HPC cluster an additional data network (MPI network) running application data will be configured. The MPI network is based mostly on advanced technologies like Infiniband.

The system installation and configuration of the compute nodes is based on images, pre-defined or set up on one or more golden clients. The images are maintained and distributed using the SystemImager suite, unless a diskless setup is used. On top of this, the ParaStation ClusterTools defines an additional layer describing the entire cluster.

Flexible node type concept

The image based installation and update mechanism allows the definition and configuration of an unlimited number of specialized service nodes.

For small installations the master node may also serve as a login node and storage server. However, large HPC clusters call for dedicated service nodes. The following node types are used at various customer installations:

- login nodes, allowing job submission and post processing for users, typically configured with additional external network connection.
- I/O servers providing storage access to all nodes, e.g. NFS server, Lustre MDS, OSS servers.
- Infiniband subnet manager servers.
- additional admin nodes allowing the implementation of a cascaded structure of the network used for installation and update of thousands of compute nodes (sub-cluster concept).
- Customer defined servers for special tasks, e.g. GPFS gateways.

A failover configuration of the master node for high availability is supported.

Prerequisites

The ParaStation ClusterTools currently supports the following distributions for the master node's setup:

- SuSE Linux Enterprise Server 11 (incl. SP2)
- openSUSE 12.1
- Redhat Enterprise Linux 5 and 6
- CentOS 5 and 6
- Scientific Linux 5 and 6

The master node may be set up using the typical installation tools provided by the distribution (e.g. yast in the case of SuSE or kickstart for RedHat) and should include a typical server set up.

At least 70GB of disk space is required for the installation of a ParaStation ClusterTools master node. For safety reasons, it's highly recommended to use RAID-based disk systems.

The following services will be installed and enabled on the master node:

- NFS server: to provide cluster-wide software repositories and to distribute user and scratch file systems (/home, /scratch, ...).
- bind: domain name service for name resolution within the cluster.
- NIS server: to enable cluster-wide user authentication. Alternatively LDAP may be used.
- Mail server: to forward mails directed to users on the cluster or master node.
- dhcp server, tftp server, syslinux (pxe boot environment), rsync, slpd: services required for network booting and installation of the compute nodes.

Some additional packages must be available for automatic installation. These packages are included in the ParaStation ClusterTools install media.

All compute nodes must be set up to do a PXE boot first. Tools for collecting MAC addresses from the compute nodes are provided, or hardware vendor supplied lists can be used.

The hardware vendor has to provide a BMC supporting IPMI 2.0 and appropriate tools to get/set BIOS parameters.

Setting up and maintaining compute nodes

The following commands are available to install, configure and maintain the compute nodes.

For large clusters (thousands of nodes) the commands support a cascaded setup, i.e. more than one admin node controlling the compute node images can be defined (sub-cluster concept).

Command	
psnodes-addnew	Adds one or more compute nodes to the list of known nodes. Typically used for a bunch of nodes at a time, e.g. an entire rack. Three methods of obtaining the required MAC addresses are supported: • for blade systems the MAC addresses are gathered automatically from the blade chassis • a file containing the addresses (supplied by the manufacturer) • automatic collection of MAC addresses after power cycling the nodes sequentially
psnodes-replace	Replaces the MAC address of a known node and updates the configuration files. Typically used after node repair.
psnodes-reinstall	Re-installs a node, e.g. with a new image.
psnodes-getimage	Synchronizes an image stored on the master node with a Golden Client.
psnodes-update	Synchronizes selected or all compute nodes (except Golden Clients) with the related image stored on the master node. Can be configured to run automatically on node reboot.
pssyslog	Continuously displays one or more log files of compute nodes.
psconsole	Provides the system console of a compute node using SOL based on IPMI.
pspowercycle, pspowerdown	Node power control.
psipmi	Get/set IPMI related information.
pschglog	Add and list change log entries.
pslshw	Print short information about a node's hardware components.
pssetbios, psgetbios	Set or get BIOS settings.
psnodes-admin	Lists, checks or manipulates the ParaStation ClusterTools's node database.
pstiblink	Test the route between the running node and the specified Infiniband port.
psmaintenance	Reports, sets and clears the time planned for the next maintenance and a next action flag on specified nodes. Nodes sharing blades or enclosures (“Buddy nodes”) are supported.
pschecknodes	Runs extensive checks on the specified nodes and sets the nodes online on successful completion. Additionally the next action flag is cleared, the corresponding tickets are updated and a change log entry is written.

License

ParaStation ClusterTools is not freeware. Details can be found in the ParaStation license agreement on www.par-tec.com.

Support

After signing a support contract, support for all packages is granted for the agreed period. The maximum response time is next business day. Support is available by telephone, email, and/or remote login. On-site support at the installation site is not included.

The support comprises all *ParaStationV5* components as well as the open source software utilized (if applicable). Other open source software tools that have been provided free of charge are only supported if resources are available, a general claim cannot be advanced on the basis of this support agreement.

Scope of delivery

ParaStation ClusterTools comprises the following components:

- Software RPM packages (pscluster, pscluster-admin, pscluster-base),
- Documentation (ParaStation ClusterTools Administrator's Guide) in electronic format,
- Support as agreed in the support contract.

Copyright

Further information

For further information about ParaStation and related products visit http://www.par-tec.com or send an email to sales@par-tec.com.