**HPE Performance Cluster Manager**

|  |
| --- |
| **HPE Performance Cluster Manager**    **HPE Performance Cluster Manager (HPCM)** delivers an integrated system management solution for Linux®-based high performance computing (HPC) clusters. HPE Performance Cluster Manager provides complete provisioning, management, and monitoring for clusters scaling up to Exascale sized supercomputers. The software enables fast system setup from bare-metal, comprehensive hardware monitoring and management, image management, software updates, power management, and cluster health management. Additionally, it makes scaling HPC clusters easier and efficient while providing integration with a plethora of 3rd party tools for running and managing workloads. HPE Performance Cluster Manager reduces the time and resources spent administering HPC systems - lowering total cost of ownership, increasing productivity and providing a better return on hardware investments. Furthermore, it optimizes cluster performance by proactive monitoring, multiple telemetry visualization GUIs, and customizable alerting techniques.    Initial system setup involves installation of software including the Linux operating system on the administrative node, discovery of hardware components for the cluster nodes and operating system provisioning for all the compute and service nodes in the cluster. HPE Performance Cluster Manager can quickly provision a cluster with thousands of nodes from bare metal - typically within an hour. In addition, new cluster nodes being added to the existing cluster are automatically discovered and configured without requiring system shutdown. For increased resiliency of the cluster, a pool of minimum three admin nodes can be used to achieve a high availability (HA) Quorum. The admin nodes in the HA Quorum share a virtual machine image which can be used to reconfigure the cluster in case of system failure.    Hardware management is comprehensive and secure. HPE Performance Cluster Manager offers fine-grained centralized monitoring and management of clusters to keep it performing at its best. The software collects telemetry from the cluster nodes and stores them in a secure repository. System administrator tasks on the administrative nodes are kept secure from end-user access. When issues are detected, alerts are sent or displayed for the attention of the system administrator. The system administrator can setup automatic reactions to specific alerts such as power capping when a specific temperature is reached in the datacenter. System data can be monitored, and additional analysis of the metrics can be done by visualizing the metrics at a specific point in time or over a historical period - either in system GUI, CLI or via third-party tools such as Ganglia, Nagios, Kibana, or Grafana. HPE Performance Cluster Manager offers several preconfigured dashboards for both hardware and jobs monitoring.    HPE Performance Cluster Manager offers comprehensive cluster health management. With lightweight and interactive health check reports, system admins can diagnose and fix issues in the production environment quickly without seeking help from HPE support teams, thus improving the turn-around time for any cluster health issues. Easy to visualize and powerful console based and graphical dashboards, make cluster health monitoring more efficient. Validation of system diagnostics run in HPE factory on new system after installation at the customer site ensure additional security. Invasive health checks during maintenance windows to uncover potential system issues and node checks before running jobs to prevent their failure ensure performance optimization. Furthermore, the system also offers ongoing monitoring of the whole system via preconfigured Grafana dashboards.    The software also utilizes hooks in job-schedulers like Altair PBS Professional and Slurm to check the health of nodes before running a new job. This helps to improve the overall job performance by avoiding running jobs on unhealthy cluster nodes. The installed software including the bios on the cluster nodes can be compared and flagged for any inconsistencies with versions or missing items. Integrated firmware flashing supports flashing of bios, BMC/iLO, CMC, network adapters and switches.    The HPE Performance Cluster Manager image management system supports a secure software image repository that stores software in multiple formats. Software stored in the image repository can include multiple versions of Linux operating system or other software such as middleware and other applications. Each software image has version control accountability built-in to track changes. Any software image in the repository can be installed on-demand on a cluster node or set of cluster nodes and restored to the original software environment as required.              HPE Performance Cluster Manager offers comprehensive monitoring and management for nodes, chassis and racks as well as power distribution units (PDU) and cooling hardware. HPE Performance Cluster Manager also offers additional power management features such as:  tools for accurate measurement and prediction of power usage for better capacity planning and setting limits to trigger node level power caps. Step-by-step topology and protocol-aware Power On/Off feature enables controlled start and shut-down of the clustered system. For example, power-on order is rack, chassis, cluster node and power-off order is cluster node, chassis, rack. Power telemetry is collected in watts and the metrics can be saved for analysis and historical reference. In addition, HPE Performance Cluster Manager supports advanced power management features for power capping and power resource management for jobs via integration with the job schedulers. HPE Apollo systems require Apollo Platform Manager (purchased separately) for power capping and rack management. |
| Overall, HPE Performance Cluster Manager delivers a comprehensive cluster management environment with resiliency, security, operational efficiency and scale for HPE Apollo, HPE Cray EX, SGI and ProLiant high performance computing clusters. |
|  |
| **What’s new**  HPE Performance Cluster Manager v1.6 features:   Support for RHEL 8.4, SLES 15 SP3, and Cray OS 2.2   Monitoring and Health-check for AMD GPUs   Real-time health-check of HPE SlingShot fabric   Easier system infrastructure monitoring using Prometheus   Support for HPE Cray EX systems with HPE Slingshot Fabric-11 software   Simplified Quorum HA set-up   Improved system configuration with multiple host name support and generic network configuration |

Standard Features

|  |
| --- |
| **Interface**  HPE Performance Cluster Manager supports a Command Line Interface (CLI) and Graphical User Interface (GUI). The GUI can be customized and allows access to all compute node consoles from a single screen. |
|  |
| **System Setup and Provisioning**  Initial system setup involves installation of software including the Linux operating system on the administrative node, discovery of hardware components for the cluster nodes and operating system provisioning for all the compute and service nodes in the cluster.    HPE Performance Cluster Manager can quickly provision a Linux operating system on a cluster with thousands of nodes from bare metal - typically within an hour. Nodes are provisioned in parallel for maximum system performance in one session - no need to provision in groups of nodes - even for the largest clusters.    HPE Performance Cluster Manager supports these two cluster configurations:   Flat - suitable for systems with less than 512 nodes (HPE Apollo, HPE ProLiant and HPE SGI 8600)   Distributed – for larger systems (as above + HPE Cray EX systems)    The flat configuration is defined as an administrative node with many compute nodes. The distributed configuration is defined as an administrative node with leader nodes that support up to 576 compute nodes each. The distributed configuration enables the cluster to scale-out to the Exascale sized supercomputers. It consists of shared pool of Leader Nodes managing groups of compute nodes in an active-active setup, without RHEL or SLES HA software. There needs to be one leader for 512 nodes. Leaders come in groups of 3 servers (3 for resiliency at the storage level, i.e. 3 leaders for 1.5k nodes, 6 leaders for 3k nodes, etc.). Any leader can manage any compute node. The compute nodes are automatically redistributed on the available leaders. Leader node can be serviced without shutting down compute nodes. All 3 leaders share a common secure respository where all the compute node information is stored and can be retrieved from. During the system setup discovery phase, the hardware components including networking configuration information are recorded in the secure repository. Having the hardware inventory stored in a central location enables easier management of the full cluster using HPE Performance Cluster Manager --- no need to manage separate files with hardware configuration descriptions |
|  |
| **Hardware Monitoring and Management / System Monitoring**  HPE Performance Cluster Manager collects telemetry from the cluster components like iLO logs, CMC, SBMC, as well as other hardware such as cooling units and storage, and stores them in a secure repository. Some standard pre-configured metrics are CPU load, uptime, memory usage, and I/O speeds. Integration with NVIDIA Datacenter GPU Manager enables monitoring of GPU diagnostics, health, configuration and alerts in a GPU-enabled server. HPE Performance Cluster Manager aggregates the GPU metrics to monitor GPU activity across the cluster.    In addition to the metrics, we use message queuing applications, like Mosquitto and Kafka to channelize the data flow and collate all the telemetry information for persistent storage and analytics in an Elasticsearch cluster. The filtering mechanism in Elasticserach allows Administrators to configure options for producing alerts and additional metrics based on the hardware components in the system.    We also include REST APIs for automation and third party software integrations. |
|  |
| **System Alerts**  HPE Performance Cluster Manager sends alerts in response to certain system conditions or issues that occur on the system. For some alerts, the system administrator can setup automatic responses to alerts triggered when thresholds are reached for selected metrics. The alerts are sent to the CLI or GUI console and by email.    HPE Performance Cluster Manager supports Field Replaceable Unit (FRU) inventory management (serial number, hostname, part number, bmc firmware and date) and alerts (incl. detecting new hardware, notifications when HW changes name/is replaced/pulled out) that enable more detailed communications with HPE Pointnext, improving serviceability of the cluster. |
|  |

|  |  |
| --- | --- |
| **Supported Product**  **Monitoring GUIs**  HPE Performance Cluster Manager makes it possible to view the state of the cluster at-a-glance via cluster monitoring GUI which provides four different views of the state of selected groups of nodes:   In Table View, a table is displayed with rows corresponding to nodes and columns corresponding to metrics.   In Instant View, a "flower" is displayed for each set of user-selectable metrics. Each "petal" of flower corresponds to one of the nodes in the group, and the length of the petal is proportional to the most recently reported value of the metric on that node. Additionally, in Instant View, each flower can include a performance gauge which provides the cumulative value of any metric over the nodes that are displayed.   In Time View, a "tube" is displayed for each set of user-selectable nodes. The "tube" is a 3D display where each cross-section contains the information shown in the corresponding flower at a specific time. The tubes show the metrics over time. The 3D displays can be manipulated to rotate and stretch and includes the ability to change angular perspectives.   In Bar Graph View, the monitored values for the selected groups of nodes appear in the form of a bar graph.    **HPCM offers several preconfigured monitoring dashboards in Grafana, such as**   System Monitoring Dashboard (cluster-level CPU loads and memory, disk read & write data, memory statistics via graphs, as well as info on top 20 performing nodes).   CDU dashboards (offering detailed information about cooling distribution unit such as temperature, humidity, valve controls, dew point & water flow)   Lustre dashboards (showing alert numbers as well as information on minimum, maximum and current statistics for the  hardware incl. IME)   Health Check Dashboard (see Cluster Health Management section)   Scheduler Status Report Dashboard - real-time monitoring of jobs scheduled via either PBS Professional or Slurm showing information such as nodes where the schedulers are running, whether they are active/inactive, jobs statuses & queues, click on Job ID shows CPU utilization data etc.)    For Ansible, HPE Performance Cluster Manager can create Ansible Playbooks with the hardware inventory that can be input and viewed by Ansible. | |
|  | |
| **Hardware Management**  Hardware management is comprehensive and secure. System administrator tasks on the administrative nodes are kept secure from end-user access. Analysis of the hardware metrics is done by visualizing the metrics at a specific point in time or over a historical period. The installed software including the bios on the cluster nodes can be compared and flagged for any inconsistencies with versions or missing items. System management commands can be broadcast to all or selected cluster nodes for operations such as provision, reboot, shutdown and power off. Integrated firmware flashing supports flashing of bios, BMC/iLO, CMC, network adapters and switches.    The Cluster topology supports separate management and compute fabrics which enables the system management and application workloads running on dedicated networks to maintain best bandwidth for the specific tasks or workloads. HPE Performance Cluster Manager securely orchestrates the tasks that run on the Ethernet management fabric versus on the Ethernet, InfiniBand, Omni-Path or HPE Slinghshot compute fabric. | |
|  | |
| **Image Management**  Software images are stored in a secure repository that supports multiple formats including RPM, ISO, remote repository and gold image. The repository provides flexibility to easily access the software to install on the system. Each image has specific identification information such as the timestamp of when it was last updated, who updated, version of software, etc. Updates to images are stored as changes to the originals to significantly save on disk space, while maintaining a full revision history. Software images can be loaded into the repository from software download, physical media (DVDs) or captured from running nodes.              The software images can be installed on all or select cluster nodes on-demand. The software image repository makes management of software on a large or complex cluster much easier. For example, if an operating system upgrade has issues, the system can be quickly reverted (re-deploy) to the last working operating system version because the software is available in the software image repository. Another advantage is that the repository can store multiple and different versions of the Linux OS distributions. The cluster’s compute nodes can support a mix of SLES, RHEL , Cray OS, TOSS or CentOS depending on the workload’s requirements.    HPE Performance Cluster Manager provisions software via multicast or BitTorrent protocols. Multicast provisioning is secure and protected with an authentication key. The provisioning process is reliable and does not require manual intervention after network communication interruptions because HPE Performance Cluster Manager will automatically restart the provisioning process and continue until the provisioning task is completed. The provisioning process also performs quickly and can provision and re-provision software on thousands of compute nodes within an hour. | |
|  | |
| **Software Updates**  The provisioning supports diskless (nfs/tmpfs) and disk-full cluster compute nodes. For diskless compute nodes, the administrative node/leader nodes manage the operating system images.    Selected cluster compute nodes can be provisioned with a different Linux operating system distribution as required by the end-user’s workloads. For example, the system is running RHEL 7, but a job requires SLES 12. The system administrator can provision selected cluster compute nodes with SLES 12 and re-provision the cluster nodes back to RHEL 7 after the job has completed. | |
|  | |
| **Power and Cooling Management**  HPE Performance Cluster Manager offers comprehensive power monitoring and management for the nodes, chassis and racks as well as power distribution units (PDU) and cooling hardware.  For HPE PDUs and cooling hardware (CDUs, HPE ARCS and in 1.5 also direct liquid cooling for HPE Apollo systems), monitoring  data are displayed in special GUI which also displays alerts for preconfigured events (i.e. water leakage, temperature changes,  lock/unlock doors, power supply failure). Alerts can be selectively configured to send email to multiple groups. The GUI also provides plotting for trending and historical analysis.  HPCM also offers preconfigured at-a-glance dashboard for monitoring of power and cooling devices in Grafana. HPCM (1.5 onwards) offers real-time anomaly detection on interface hardware such as CDUs and cooling racks using AIOps techniques.    In addition, the cluster manager offers these capabilities:    **Power On/Off control**  Power control is topology and protocol aware and Power On/Off is staged sequentially as following: Power on: rack, chassis, nodes. Power off: nodes, chassis, rack. The feature is also aware of whether node is booted with an operating system or not.    **Power consumption and analysis**  HPE Performance Cluster Manager aggregates power and energy at available measurement points (rack AC, bulk DC, nodes) and reports power data reported in watts. Up to three months of power metrics are retained for reporting and analysis.    **Power capping**  System administrators can set limits to trigger a power cap based on environmental thresholds (power or thermal), datacenter power capacity or for other reasons (i.e. workloads, planned brownouts, time of day).    **Power Management for GPUs**  HPE Performance Cluster Manager is integrated with NVIDIA Datacenter GPU Manager (DCGM). For power management on servers with NVIDIA GPUs, DCGM supports GPU power monitoring and power capping. HPE Performance Cluster Manager enables aggregation of the DCGM GPU power management across the cluster.              **Power resource management for jobs**  The HPE Performance Cluster Manager API for power resource management for jobs enables access to power metrics for CPU and GPU processors. Altair PBS Professional Power Awareness feature and Slurm power management use the power metrics from the cluster management tool to define per job power envelopes, contain the system global power consumption, apply per user, per group, per job class power containment and account for consumed energy. | |
|  | |
| **Security**  HPE Performance Cluster Manager is designed to keep cluster management tasks secure. Some of the features include:   Central management of the system through the secure administrative node. In a distributed configuration, the admin and leader node ssh access is secure and not accessible by users and jobs. In addition, compute node root ssh keys are kept separate from admin nodes.   All system configuration information (i.e. switches settings) are stored in a secure central repository with certificates and keys protecting access to the repository. Similarly, monitoring data are also stored in a secure central database.   Provisioning via HPE Performance Cluster Manager is secure: Transfer of secret information (ssh, passwords, etc) occurs before provisioning. There is no image corruption during multicast provisioning because image is encrypted with an authentication key.   HPE Performance Cluster Manager supports separate management and data networks where management of the system runs on a dedicated secure Ethernet management network. | |
|  | |
| **Cluster Health Management**  Cluster health management provides an overall health check for the whole system. Some of the features include:   Validate system diagnostics run in HPE factory on new system after installation at the customer site   Invasive tests included in HPE Performance Cluster Manager to be used during maintenance window or to diagnose root cause of failures   Self-Service diagnostics with light weight and interactive reports for quick inspection and fixing of issues at e customer site without waiting for HPE support team.   System health check before running job with a workload scheduler tool to avoid initiating jobs on unhealthy cluster nodes   Real-time alerts management - HPCM offers at-a-glance Grafana cluster health dashboard aggregating alerts and severity information on cluster in overall, fabric, cooling hardware with links to more detailed reports/monitoring interfaces. | |
|  | |
| **Operating System Support**  HPE Performance Cluster Manager software supports Red Hat® Enterprise Linux® 7 and 8, SUSE Linux Enterprise Server 15 on the management nodes requiring Quorum HA. CentOS 8 is also supported on the management nodes that do not require Quorum HA. The cluster compute nodes can run Red Hat Enterprise Linux 7 and 8, SUSE Linux Enterprise Server 12 and 15, CentOS 8, Cray OS 2.2 and TOSS 7. Additionally, the HPC versions of RHEL and SLES are also supported on the compute nodes. The administrative and leader nodes have to be on the same Operating system with the same version. | |
|  | |

|  |
| --- |
| **Management Integration**  HPE Performance Cluster Manager offers CLI, GUI and API for integration with other software management tools (i.e. workload managers, fabric software and orchestration tools).    HPE Performance Cluster offers integration with the following solutions:   Altair PBS Professional®   Slurm® by SchedMD   Mellanox Unified Fabric Manager®   Altair Grid Engine   Ganglia monitoring software   Ansible Playbooks and inventory files   NVIDIA Datacenter GPU Manager   Arm Performance Reports   Grafana   Kafka   Alerta |
|  |
| **Hardware Requirements**  HPE Performance Cluster Manager software is supported on the following Gen9, Gen10 and Gen10 Plus platforms:   HPE Cray EX systems   HPE SGI 8600 and HPE Apollo9000   HPE Apollo 2000, 6000, 6500 systems   HPE ProLiant DL360 / 380 / 385 / 580 servers   HPE Moonshot Systems |
|  |
| **Migration to HPE Performance Cluster Manager**  **Starting with Cluster Management Utility**  Customers using Cluster Management Utility on the administrative node with an active HPE Pointnext support contract are entitled to migrate to HPE Performance Cluster Manager. In order to migrate, the current system must be running Cluster Management Utility v8.2. More information about the migration can be found in the QuickStar guide [**here.**](https://www.hpe.com/support/hpcm-cmu-mig-001)    **Starting with SGI Management Suite**  Customers using SGI Management Suite on the administrative node with an active HPE Pointnext support contract are entitled to migrate to HPE Performance Cluster Manager In order to migrate, the current system must be running SGI Management Suite v.3.5.    For more information, please see the HPE Performance Cluster Manager Installation Guide. HPE Pointnext can assist with the migration. Please contact your HPE sales rep for additional information on migrating to HPE Performance Cluster Manager. |