

# High-Performance Computing and the SMASH Initiative

To enhance management interoperability and help reduce total cost of ownership across heterogeneous nodes in high-performance computing (HPC) clusters, IT organizations can implement systems that comply with the Systems Management Architecture for Server Hardware (SMASH) initiative. The SMASH initiative is a suite of specifications designed to standardize management interfaces for heterogeneous computing environments and to provide an architectural framework that includes unified interfaces, resource discovery, resource addressing, and data model profiles. In this way, SMASH not only addresses complicated administrative challenges, but it also enables HPC clusters to enhance resource utilization and system uptime.

BY YUNG-CHIN FANG AND JON HASS

## Related Categories:

*Cluster management*

*Clustering*

*Dell PowerEdge servers*

*High-performance computing (HPC)*

*Parallel systems*

*Remote management*

*Standards*

*Systems management*

Visit [www.dell.com/powersolutions](http://www.dell.com/powersolutions) for the complete category index.

**H**igh-performance computing (HPC) clusters are widely used in industry, research, and academic segments for geophysics, semiconductor, telecommunication, database, digital content creation, weather and climate research, automotive, software, finance, and other research and development purposes. On the June 2000 TOP500 Supercomputer Sites list, only 2.2 percent of the supercomputer systems were cluster-based. By 2005, 60.8 percent of the supercomputer systems on the TOP500 list were cluster-based, while the percentage of massively parallel processing (MPP) and constellation architectures dropped rapidly from five years earlier.<sup>1</sup> Figure 1 shows the increasing prevalence of cluster architectures among the TOP500 Supercomputer Sites from 2000 to 2005.

## The need for manageability

Besides gaining popularity among architecture types, HPC clusters are also scaling out. All the systems listed on the June 2005 TOP500 Supercomputer Sites list are equipped with at least 200 processors and some systems are equipped with thousands of processors. The direct total cost of ownership (TCO) for an HPC cluster is in proportion to the scale of the data center, which includes computing resource depreciation, facility rent and maintenance costs, utility bills, system administrator costs, and other financial overhead. When an HPC cluster is equipped with hundreds or thousands of processors, such factors can contribute considerably to TCO. A well-designed, comprehensive remote management framework

<sup>1</sup> For more information, see the TOP500 Supercomputer Sites Web site at [www.top500.org](http://www.top500.org). Results are available at [www.top500.org/lists/2000/06](http://www.top500.org/lists/2000/06) for the June 2000 list and at [www.top500.org/lists/2005/06](http://www.top500.org/lists/2005/06) for the June 2005 list.

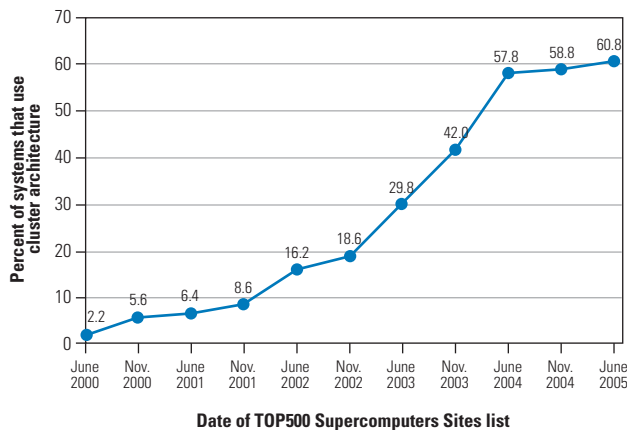


Figure 1. Growing use of cluster architecture among TOP500 Supercomputer Sites

can help reduce TCO by enhancing system uptime, resource utilization, and yield; streamlining administrative operations; and monitoring the health of cluster components to help address warning conditions before they result in system failure.

### The need for interoperability

TCO not only includes direct costs, but it also includes indirect costs such as support, training, and retooling. Data centers commonly operate generations of heterogeneous hardware. Each platform comes with a management framework, which typically contains a rich set of tools and utilities. In many cases, these tools are specialized and adapted to an individual environment, installation, and product in the data center.

Usually, system administrators must learn all the different management frameworks in a heterogeneous environment before using the corresponding remote management frameworks to perform one task—such as remote power cycling of all nodes. Multiple management frameworks require more administrator and user training, higher support costs, and longer training time than a single management framework. Such factors contribute to increased TCO. Many organizations are finding that a unified, interoperable management interface and framework has become a necessity to meet enterprise requirements to scale out HPC clusters quickly, flexibly, and cost-effectively.

### The SMASH specifications suite

In January 2004, the Server Management Working Group (SMWG) was established in the Distributed Management Task Force (DMTF).

The SMWG initiated the Systems Management Architecture for Server Hardware (SMASH)<sup>2</sup> initiative to address the interoperable manageability requirements of small to large-scale heterogeneous computing environments. The DMTF has more than 3,000 active participants across industries and leads the development of management standards and integration technology for enterprise and Internet environments. The DMTF governs several specifications including System Management BIOS (SMBIOS),<sup>3</sup> Common Information Model (CIM),<sup>4</sup> Desktop Management Interface (DMI),<sup>5</sup> Web-Based Enterprise Management (WBEM),<sup>6</sup> and Alert Standard Format (ASF).<sup>7</sup> SMWG members include Dell, HP, IBM, Intel, Newisys, OSA Technologies, Sun, and others. Dell has committed more than 50 professionals to participate in the DMTF and has made significant contributions to the SMASH initiative as well.

SMASH comprises a suite of specifications that deliver architectural semantics, standardized server management protocols, and hardware data model profiles designed to help unify data center management. SMASH includes the Server Management (SM) Command Line Protocol (CLP) specification, the SM Managed Element Addressing specification, the SM CLP-to-CIM Mapping specification, the SM CLP Discovery specification using the Services Locator Protocol (SLP), the scripting specification, and several dozen system and component data model profile specifications.

System administrators can use a consistent SMASH command-line interface (CLI) to monitor and manage heterogeneous cluster hardware resources remotely, update firmware, and perform inventory. The CLI can be used to monitor and remotely manage the health status of components in large heterogeneous clusters to help overcome differences in hardware architecture, OS dependencies, and issues stemming from different command sets and utilities in existing management frameworks from different vendors. Through the CLI, administrators can tailor and automate computer- and data center-specific management tasks such as remotely changing the cluster's BIOS boot order, remotely power cycling hung nodes, and remotely updating firmware in parallel. A SMASH-compliant implementation is designed to improve interoperability and manageability—enabling optimal resource utilization and uptime while helping to reduce the indirect costs of training, support, and retooling. In this way, SMASH can help minimize TCO while enhancing reliability, availability, and manageability.

**SM CLP.** This specification defines the syntax and semantics of a small set of verbs that act consistently on heterogeneous hardware

<sup>2</sup> For more information about SMASH, visit [www.dmtf.org/standards/smash](http://www.dmtf.org/standards/smash).

<sup>3</sup> For more information about SMBIOS, visit [www.dmtf.org/standards/smbios](http://www.dmtf.org/standards/smbios).

<sup>4</sup> For more information about CIM, visit [www.dmtf.org/standards/cim](http://www.dmtf.org/standards/cim).

<sup>5</sup> For more information about DMI, visit [www.dmtf.org/standards/dmi](http://www.dmtf.org/standards/dmi).

<sup>6</sup> For more information about WBEM, visit [www.dmtf.org/standards/wbem](http://www.dmtf.org/standards/wbem).

<sup>7</sup> For more information about ASF, visit [www.dmtf.org/standards/asf](http://www.dmtf.org/standards/asf).

systems and components represented by CIM-based data models. The CLP can be implemented in different ways, including in band, out of band, and via proxy. The CLP and the SMASH architecture are designed to be independent of machine state, OS state, server system architecture, and access method. The variety of ways the CLP can be implemented can facilitate existing local and remote management components without requiring additional memory and CPU resources on compute nodes. The unified command protocol is designed to be user-friendly and simple on both existing and future clusters.

**SM Managed Element Addressing.** This specification defines the formulation of command target addresses that resemble the naming conventions of hierarchical file systems. It specifies the user-friendly class tags and implied association classes that may be used to construct paths to address any managed element appearing within the scope of the manageability access point (MAP). The MAP is a collection of system services that provide management in accordance with specifications published under the DMTF SMASH initiative. An important aspect of MAP operations management is the capability to support simultaneous sessions through the MAP, thus unleashing the potential of remote parallel management functionality.

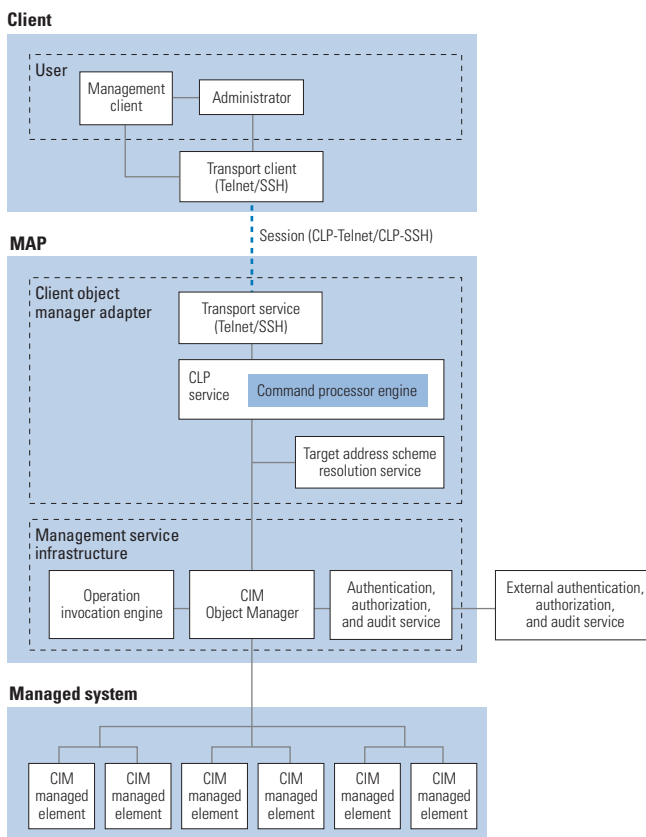


Figure 2. Example SMASH architecture

**SM CLP-to-CIM Mapping.** CIM provides a common definition of management information for systems, components, networks, applications, and services, and it allows for vendor extensions. CIM common definitions enable vendors to exchange semantically rich management information between systems throughout management fabrics. The CLP-to-CIM Mapping specification details how the CLP command verbs manipulate or act on the CIM, thus enabling a WBEM/CIM-compliant interface to the hardware-level management provided by the Intelligent Platform Management Interface (IPMI), ASF, and other device-level hardware management interfaces. The mapping also enables the CLP to potentially apply to existing CIM-based management frameworks.

**SM CLP Discovery using SLP.** This specification, leveraging SLP, addresses three aspects of discovery: how a client discovers which managed elements the MAP manages, the discovery of the capabilities of the MAP itself, and the discovery of the service access points of the MAP. The MAP is a network-accessible service for managing a system. A MAP can be instantiated by a management process, a management processor, a service process, or a service processor.

**SMASH profiles.** Server management profiles provide the object model definitions for manageability content and the architecture models for mapping computer hardware to fully connected association graphs in a manner that is consistent between different implementations. Profiles describe the legal classes and associations that can be used to model system and component hardware. The profiles can be reused and combined in different combinations to help ensure that all instances in the system are implemented in a consistent manner across multiple vendor architectures and offerings. User-friendly views based on profiles are defined to help simplify managing system boot, power, storage, firmware change management, system configuration, and hardware asset inventory. The Boot Control Profile is an example SMASH profile; it can be used to define boot order and other configuration aspects of boot devices.

### Example SMASH architecture

Administrators issue standardized CLP commands to either a management client or a MAP directly. The MAP includes a CLP command/response protocol and processor engine: a text command message is transmitted from the administrator over the transport service, such as Telnet or Secure Shell (SSH), to the MAP; the MAP receives the command, authenticates the use/command privilege, and processes the command by converting it to a CIM model manipulation (see Figure 2). The command target address is resolved to the appropriate instance of the CIM object representing the component being manipulated. When the model is manipulated, a CIM provider is engaged and it translates the manipulation to a native hardware command. The response from a managed system or component is then transmitted from the MAP back to the client/administrator in

CLP verb	Definition
help	Retrieves context-sensitive help (same as the <code>-help</code> option with the addition of help for targets)
cd	Sets the current default target (navigates the target address space of the MAP)
show	Shows the values of a property or the contents of a collection/target
set	Sets a property or a group of properties to a specific value
exit	Terminates a CLP session
reset	Resets the target
start	Starts the target
stop	Stops the target

Figure 3. Example SMASH CLP command verbs

human-readable text or XML format. Figure 3 shows a few examples of SMASH CLP command verbs.

The CLP command/response message output can be in human-readable text, or for further processing, XML or comma-separated value (CSV) format. For example, the following command shows the current sensor reading of power supply 3 in system 1:

```
Prompt> show -display associations -o
format=clpxml /system1/powersup3
```

The assigned output format is XML. System administrators can then direct the XML information to other applications for further processing to meet specific needs.


### The implication of SMASH for HPC clusters

In the HPC cluster deployment phase, the CLP can be scripted to boot up an additional HPC cluster remotely—and properly, to avoid a power surge that could damage the system. Administrators can use the Boot Control Profile to define the boot order. For example, the boot order can specify that the first boot should be from the network via the Preboot Execution Environment (PXE) to remotely deploy the OS and predefined software stack to all nodes, while the second boot can be from a local hard drive to complete the necessary cluster configuration. Because HPC clusters usually consist of a large number of nodes, remote diagnostic services can also be invoked in this phase to examine cluster-wide hardware health status. Configuration activities can be invoked to stage cluster hardware and to stabilize the cluster operating state. The Power Control Profile, Boot Control Profile, and Diagnostics Profile can help reduce deployment time. The number of days saved during deployment time can translate into additional days of production time and fewer days of overall hardware depreciation time—thereby enhancing cost-effectiveness.

In the HPC cluster operational phase, SMASH can be used to monitor and remotely manage the hardware health status of

heterogeneous HPC clusters and grids to help prevent hardware failure and the need to rerun parallel jobs as well as to help reduce recovery time. For example, when a SMASH-compliant, one-to-many management console reports an unusual memory-bit error count or a nonfatal SMART (Self-Monitoring, Analysis, and Reporting Tools) error in the hard drive, system administrators can respond by launching runtime job migration or check-pointing to preserve the current computing process, suspend the job, swap out the potential problem hardware, and restart the job. SMASH also can be used to reduce cluster maintenance time by facilitating activities such as remote, parallel firmware updates of heterogeneous clusters and remote power management for post-OS updates and upgrades. The CLP also can be tightly integrated with existing job schedulers to enhance overall hardware utilization, because new job scheduling schemes such as temperature- and power-aware scheduling or run-time environment-sensitive scheduling become achievable.

### Enhanced HPC cluster deployment, management, and operations

SMASH-compliant implementations are designed to help solve many of the management difficulties prevalent in heterogeneous HPC clusters. By enhancing the efficiency of an HPC cluster's deployment and operational phases, SMASH-compliant systems can help minimize hardware depreciation and operational costs while enabling administrators to enhance resource utilization and system uptime—responding quickly, flexibly, and cost-effectively to business-critical scale-out requirements. In addition, administrators who do not have to learn and use multiple management frameworks can spend less time training and more time producing results. 

**Yung-Chin Fang** is a senior consultant in the Scalable Systems Group at Dell. He specializes in HPC systems, advanced HPC architecture, and cyberinfrastructure management. Yung-Chin has published more than 30 conference papers and articles on these topics. He also participates in HPC cluster-related open source groups as a Dell representative.

**Jon Hass** is a software architect with the Dell Systems Management Architecture and Standards team. He is vice-chair of the DMTF CIM Core Model Working Group and is the Dell representative on the DMTF Technical Committee. He is also the chair of the IPMI CIM Mapping Committee of the IPMI Forum.

#### FOR MORE INFORMATION

**SMASH specifications:**  
[www.dmtf.org/standards/smash](http://www.dmtf.org/standards/smash)