

# The Next Generation Cluster/Cyberinfrastructure Computing Management Standard -- SMASH

Yung-Chin Fang, Tong Liu  
{Yung-Chin\_Fang; Tong\_Liu}@DELL.COM  
Dell Inc.

Chokchai Leangsuksun  
box@coes.latech.edu  
Louisiana Tech University

Stephen L. Scott  
scottsl@ornl.gov  
Oak Ridge National Laboratory

Both high performance cluster computing (HPCC) and high availability (HA) cluster computing are widely accepted. Over time, a computer center tends to have generations of HPCC and HA clusters from various vendors. This trend has caused resource reliability, availability, serviceability and security (RASS) to become more important than ever. Most environments are managed by the various vendor dependent management frameworks. The complexity of the various cluster hardware, firmware and software stack makes resource management and administration even more challenging and time consuming than before. Currently management interoperability is usually compromised or absent because of the heterogeneous environment. In order to solve this problem and further reduce the direct and indirect total cost of ownership and enhance RASS, industry is defining the Systems Management Architecture for Server Hardware (SMASH) initiative. SMASH is a suite of specifications, which standardize management interfaces across hardware architecture and heterogeneous management frameworks. The specification suite provides an architectural framework that includes unified interfaces, resource discovery, and resource addressing and data model profiles. SMASH addresses complicated RASS management challenges as well as enables heterogeneous Cyberinfrastructure manageability and as a result, will bring cluster computing and Cyberinfrastructure manageability to an even higher level.

## I. THE MANAGEMENT CHALLENGE FROM SCALE OUT TECHNOLOGY

Supercomputing technology provides a unique environment through which researchers can study problems that are otherwise impractical or impossible to address. These range from the arena of geophysics, semiconductor, telecommunication, database, digital content creation, weather and climate research, automotive, software, finance and other R&D to supporting advanced industrial methods with significant economic, health and other benefits, such as designing genome sequence based disease cure computationally rather than through expensive and time-

consuming trial and error experiments. The number of HPCC on Top500 supercomputer list [1] grows from 2.2% to 58.8% in 2 years (2002 to 2004). This rapid percentage growth indicates scale out is the mainstream supercomputing technology.

As this scale out trend for HPCC continues, the number of computers in our computing centers will grow accordingly. On the Top500 supercomputer list, the Intel processor implies the HPCC architecture and this processor class has become the volume leader since Nov/2003 on Top500. Furthermore, as shown in Figure 1, the Intel processor count has grown over 600% in 3 years. With this growth comes the associated growth in management challenges.

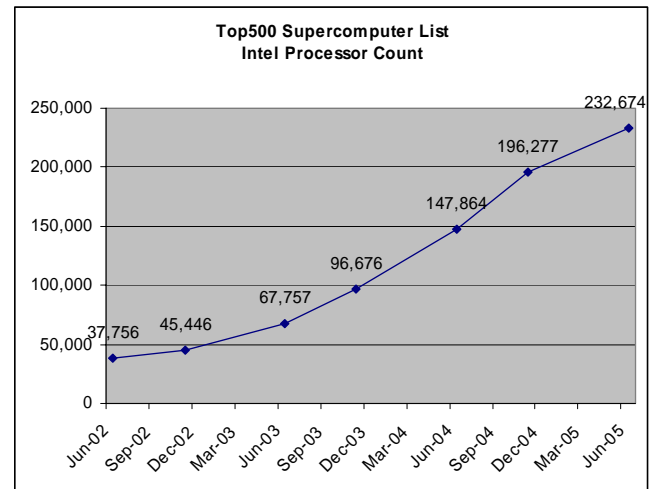


Figure1. Processor Count Growth Momentum

## II. RASS MANAGEABILITY

All the Intel based cluster listed on the June/2005 Top500 list are equipped with at least two hundred processors and many systems are equipped with thousands of processors.

As cluster node count increases, the number of components per cluster also increases and the mean time between failure (MTBF) is reduced. (Note that a smaller MTBF is a bad thing as this is an increase in failure frequency.) How to design and provide reliability, availability, serviceability and security for scaling out heterogeneous computing environments has become a serious challenge. This challenge points to Cyberinfrastructure growth and usability directly. A well-designed comprehensive information structure management framework can enhance computing information structure uptime, utilization and yield, streamline administration operations, and proactively prevent system failure in advance. In many cases, the SMASH implementation may be the enabling technology for cluster computing technology to advance to Cyberinfrastructure computing.

When a significant initial investment in a cluster computing system is equipped with a large number of processors, the direct and indirect administrative total cost of ownership (TCO) has become one of the more important concerns. The direct and indirect TCO is generally in proportion to the number of computing resources and includes many factors such as computing environment depreciation, facility rent, environment maintenance costs, utility bills, system administrator costs and other long-term expenses such as support, retooling, training, etc. SMASH implementation is not only engineered to provide RASS manageability but also designed to reduce direct and indirect cost of ownership.

### III. INTEROPERABILITY IS REQUIRED

A typical data/computing center has numerous generations of hardware which comprise a rather heterogeneous environment. Generally, each platform comes with a specific management framework including tools and utilities. In many cases, these tools are specialized and adapted to each individual environment, installation and product in the data center. Usually a system administrator must learn all these management frameworks in order to invoke a single functional command across the entire heterogeneous environment. A trivial example of this issue is the remote power cycle of all nodes. Multiple management frameworks lead to higher administrator and user training, increased support costs and longer training time. All these costs increase the system TCO. This increased cost makes a unified interoperable management interface/framework across heterogeneous hardware/management-frameworks such as that provided through SMASH a must have.

### IV. SMASH INTRODUCTION

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) [2] initiative to address the interoperable manageability

requirements of small to large-scale heterogeneous computing environments. The DMTF has over 3,000 active participants across industries and is an industry organization leading the development of management standards and integration technology for enterprise and Internet environments. The DMTF governs a number of specifications including: System Management Basic Input Output System (SMBIOS) [3], Common Information Model (CIM) [4], Desktop Management Interface (DMI) [5], Web Based Enterprise Management (WBEM) [6], Alert Standard Format (ASF) [7] among others. SMWG members include: Dell, HP, IBM, Intel, Newisys, OSA Technologies, Sun and others. Dell has over 50 professionals participating in the DMTF and has made significant contributions to the SMASH initiative.

SMASH is a suite of specifications that deliver architectural semantics, standardized server management protocols and hardware data model profiles to unify the management of data centers. SMASH includes: System Management Command Line Protocol specification (SM CLP), System Management Managed Element Addressing specification, CLP-to-CIM Mapping specification, CLP Discovery using SLP specification, Scripting Specification and several dozen system and component data model profile specifications. Among the advantages include: system administrator can use a consistent SMASH Command Line to remotely monitor and manage heterogeneous cluster hardware resources, update firmware, perform inventory, etc. The command line based interface can be used to remote monitor and manage the health status of large heterogeneous clusters and overcome the hardware architecture differences, OS dependencies and issues with different command sets and utilities of existing management frameworks from the various vendors. The CLI interface can be used to tailor and automate computer and data center specific management tasks such as remote change cluster BIOS boot order, remote power cycle hung nodes, remote parallel firmware update, etc. A SMASH compliant implementation provides valuable interoperability and manageability and these capabilities can enhance computing facilities' utilization and up time, reduce the indirect cost of training, support, and retooling. Thus SMASH can reduce the total cost of ownership, improve reliability, availability and manageability, and provide an increase return on investment.

**The System Management Command Line Protocol specification** defines the syntax and semantics of a small set of verbs that act consistently on system and component heterogeneous hardware 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 command protocol and the SMASH architecture are designed to be independent of machine state, operating system state, server system architecture or access method. The variety of ways the CLP can be implemented can facilitate existing local and remote management components without introducing new or extra memory footprint and CPU utilization to compute managed

nodes. The unified command protocol is designed to be user-friendly, simple, and can be used on existing and future clusters.

**SMASH Managed Element Addressing Specification.** This specification defines the formulation of command target addresses that resemble hierarchical file system naming conventions. 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 services of a system that provides management in accordance to specifications published under the DMTF SMASH initiative. An important aspect of MAP operations management is to support simultaneous sessions through the MAP, thus unleashing the potential of remote parallel management functionality.

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

**SMASH CLP Discovery using SLP Specification.** This specification, leveraging the Services Locator Protocol (SLP), addresses 3 aspects of discovery: 1) How a client discovers which managed elements the MAP manages, 2) Discovering the capabilities of the MAP itself, 3) Discovering the service access points of the MAP. The MAP is a network-accessible service for managing a managed system. A MAP can be instantiated by a management process, a management processor, a service processor or a service process.

**SMASH Profiles.** Server Management profiles provide the object model definitions for manageability content and architecture models for mapping computer hardware to fully connected association graphs in a way 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 ensure that all instances in the system are implemented in a consistent fashion across multiple vendor architectures and offerings. User-friendly views based on profiles are defined to simplify managing system boot, power, storage, firmware change management, system configuration and hardware asset inventory. Server

management Boot Control Profile is an example profile, which can be used to define boot order and some configuration aspects of boot devices.

**Example SMASH Architecture**

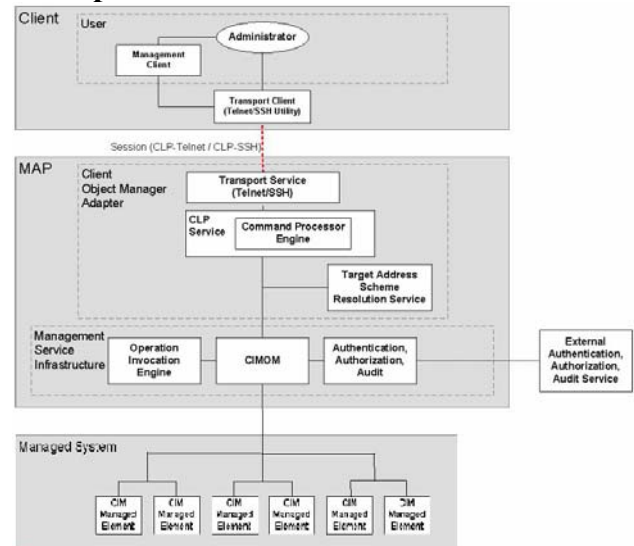


Figure 2. Example SMASH architecture

An administrator issues standardized CLP commands to either a management client or a MAP directly. The MAP includes a CLP command/response protocol and processor engine, which means that a text command message is transmitted from administrator over the transport service (e.g. Telnet or SSH) to the MAP, the MAP receives the command, authenticates the use and command privilege and processes the command by converting it to a CIM model manipulation. 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 that 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 or administrator in human readable or XML formats. The following are few examples of SMASH CLP command verbs:

CLP Verb	Definition
help	Used to get context sensitive help. The functionality is the same as the - help option with the addition of help for targets.
cd	Used to set the Current Default Target (navigate the target address space of the MAP).
show	Used to show values of a property or contents of a collection or target.
set	Used to set a property or set of properties to a specific value.
exit	Used to terminate a CLP session.
reset	Used to reset the target

start	Used to start the target
stop	Used to stop the target

The CLP command response message output is selectable for human readable text, XML or CSV structured formats for further processing. Following example means: show the current sensor reading of power supply #3 of system1; the assigned output format is XML. The system administrator can then pipe the XML information to other applications for further processing as their specific needs require.

**Prompt> show –display associations –o  
format=clpxml /system1/powersup3**

## V. MORE SMASH USE CASE

In the cluster deployment phase, the CLP can be scripted to remote boot up a new cluster properly to avoid surge spike power system issues during boot up time. The Boot Control Profile can be utilized to define the first power up boot order. For example, the first boot may be from the network via PXE to remotely deploy the operating system and pre-defined computing software stack while the second boot can be defined via Boot Control Profile to boot from the local hard drive and proceed with post configuration processing. Since a high performance computing cluster usually consists of large number of nodes, remote diagnostics services can also be invoked in this phase to examine the cluster wide hardware health status. Configuration activities can be invoked to stage cluster hardware and stabilize the cluster-operating state. The Power Control Profile, Boot Control Profile, and Diagnostics Profiles can reduce deployment time and increase cost effectiveness. N days of deployment time saved equals N days more production time and N days less overall hardware depreciation time and also means less system staff time spent on the deployment process. Many, if not all, of these features will apply to high availability clustering with respect to the cluster component.

In the cluster operational phase, SMASH will be used to remotely monitor and manage the hardware health status of heterogeneous clusters and Grids in order to avoid hardware failure in advance and save parallel job re-run or recovery time. For example: when a SMASH compliant one-to-many management console reports an unusual memory bit error count or hard drive un-fatal SMART error, the system administrator (or an automated system monitor) can respond by launching runtime job migration or check-pointing to preserve current computing progress, suspend

the job, swap out the potential problem hardware and restart the job elsewhere on the cluster. SMASH may also be used to reduce cluster maintenance times by facilitating activities such as remote update of heterogeneous clusters' firmware in parallel and remote power management for post OS upgrades. The CLP can also be tightly integrated with existing job schedulers to provide even better overall hardware utilization. New job scheduling schemes such as those that are temperature and power aware or runtime environment sensitive schemes become feasible. This also provides the common ground foundation for very large scale Cyberinfrastructure management.

## VI. SMASH IMPACTS

SMASH compliant implementations will help to solve many of today's heterogeneous cluster management difficulties, enhance deployment and operational phase efficiency and reducing the total cost of ownership. SMASH makes hardware independent RASS manageability, scalable manageability, computing info structure sensitive job scheduler all become possible. Furthermore, SMASH is also one of the key enablers for cluster-computing advance to the Cyberinfrastructure computing era. This derived Cyberinfrastructure manageability may enable scientist to investigate even larger scale problems and bring higher precision results faster than ever.

## VII. ACKNOWLEDGEMENTS

Leangsuksun's research is supported by Department of Energy contract DE-FG02-05ER25659 and Center for Entrepreneurship and Information Technology, Louisiana Tech University. Scott's research is supported by the Mathematics, Information and Computational Sciences Office, Office of Advanced Scientific Computing Research, Office of Science, U. S. Department of Energy, under contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.

## Reference

- [1] <http://www.top500.org>
- [2] <http://www.dmtf.org/standards/smash>
- [3] <http://www.dmtf.org/standards/smbios>
- [4] <http://www.dmtf.org/standards/cim>
- [5] <http://www.dmtf.org/standards/dmi>
- [6] <http://www.dmtf.org/standards/wbem>
- [7] <http://www.dmtf.org/standards/asf>