NPACI Rocks Clusters: Tools for Easily Deploying and Maintaining Manageable High-Performance Linux Clusters

Philip M. Papadopoulos, Mason J. Katz, and Greg Bruno

San Diego Supercomputer Center, La Jolla, CA, USA
{phil,mjk,bruno}@sdsc.edu
http://rocks.npaci.edu

Abstract. High-performance computing clusters (commodity hardware with low-latency, high-bandwidth interconnects) based on Linux, are rapidly becoming the dominant computing platform for a wide range of scientific disciplines. Yet, straightforward software installation, maintenance, and health monitoring for large-scale clusters has been a consistent and nagging problem for non-cluster experts. The complexity of managing hardware heterogeneity, tracking security and bug fixes, insuring consistency of software across nodes, and orchestrating wholesale (or forklift) upgrades of Linux OS releases (every 6 months) often discourages would-be cluster users.

The NPACI Rocks toolkit takes a fresh perspective on management and installation of clusters to dramatically simplify this software tracking. The basic notion is that complete (re)installation of OS images on every node is an easy function and the preferred mode of software management. The NPACI Rocks toolkit builds on this simple notion by leveraging existing single-node installation software (Red Hat's Kickstart), scalable services (e.g., NIS, HTTP), automation, and database-driven configuration management (MySQL) to make clusters approachable and maintainable by non-experts. The benefits include straightforward methods to derive user-defined distributions that facilitate testing and system development and methods to easily include the latest upgrades and security enhancements for production environments. Installation performance has good scaling properties with a complete reinstallation (from a single server 100 Mbit http server) of a 96-node cluster taking only 28 minutes. This figure is only 3 times longer than reinstalling just a single node.

The toolkit incorporates the latest Red Hat distribution (including security patches) with additional cluster-specific software. Using the identical software tools that are used to create the base distribution, users can customize and localize Rocks for their site. This flexibility means that the software structure is dynamic enough to meet the needs of cluster-software developers, yet simple enough to allow non-experts to effectively manage clusters. Rocks is a solid infrastructure and is extensible so that the community can adapt the software toolset to incorporate the latest functionality that defines a modern computing cluster. Strong adherence to widely-used (*de facto*) tools allows Rocks to move with the rapid pace of Linux development.

Y. Cotronis and J. Dongarra (Eds.): Euro PVM/MPI 2001, LNCS 2131, pp. 10–11, 2001.

Rocks is designed to build HPC clusters and has direct support for Myrinet, but can support other high-speed networking technologies as well. Our techniques greatly simplify the deployment of Myrinet-connected clusters and these methods will be described in detail during the talk, including how we manage device driver/kernel compatibility, network topology/routing files, and port reservation. We will also give "untuned" Linpack performance numbers using the University of Tennessee HPL suite to illustrate the performance that a generic Rocks cluster would expect to attain with no specialized effort.

Version 2.1 (corresponding to Redhat Version 7.1) of the toolkit is available for download and installation.