Stateless Clustering Using OSCAR and PERCEUS

Abhishek Kulkarni and Andrew Lumsdaine
Open Systems Laboratory, Indiana University

The 6th Annual Symposium on OSCAR and HPC Cluster Systems
University of Laval
Quebec City, Quebec, Canada
Organization of the talk

- Current state of OSCAR
- Node provisioning in OSCAR
  - Supporting a new provisioning scheme
- Integrating OSCAR and PERCEUS
  - Introduction to PERCEUS
  - Architecture and design
  - Overview of implementation
  - Issues faced during integration
- Lessons learned
  - Need for a generic provisioning framework
Current state of OSCAR

- OSCAR 5.0
  - released Nov 06

- OSCAR 5.1
  - Introduction of the new OPKG infrastructure
  - Unstable crispy branch

- Ongoing merge of branch 5.1 and trunk

- Over 200,000 downloads

- Towards OSCAR 6.0
  - OSCARV, Diskless Clusters, Decouple core infrastructure from external software
Upcoming developments

- Configurator extension
- XOSCAR
- Universal monitoring framework
- Repositories management
- OSCAR V2M extension
- API validator tool
- NFS mountpoints in OSCAR
OSCAR Components

- Core packages
  - OPD, OPKGC, Core libs, CLI, GUI, yume ...
- Provisioning packages
  - SystemInstallation Suite (SIS)
- Administration packages
  - Switcher, C3, netbootmgr, sync_files + opium
- Monitoring packages
  - Ganglia, Nagios
- Libraries, resource managers and utilities
  - TORQUE, Maui, OpenMPI, MPICH
Provisioning

- Deploy a complete computing environment on the nodes in a cluster
  - Operating system
  - Middleware
  - Libraries
  - HPC applications
  - Data

- Provisioning in OSCAR
  - System Installation Suite (SIS)
Node Provisioning in OSCAR

- **SystemInstallation Suite (SIS)**
  - **SystemInstaller**
    - Client node image building utility
    - Build images from package list
  - **SystemImager**
    - Utility for image propagation
    - Automates Linux installation
  - **SystemConfigurator**
    - Automatically configure networking and bootstrapping
    - Covers up differences in Linux distribution and architecture
System Installation Suite

Image source: Sean Dague, IBM, System Installation Suite
Node Provisioning in OSCAR

- Define image
  - Client node disk partitioning
  - Package lists
  - Network configuration
- Build image
- Install image on clients
New Provisioning Scheme

- No observed performance differences between diskfull and diskless clusters\(^1\)
- Issues with diskfull clustering
  - Power consumption
  - Heat dissipation
  - Hard disk failure
  - Less MTBF
- Diskless clusters are faster to deploy and easier to manage

---

Stateless Clustering

- Centralized management paradigm for the client nodes
- Serves a fresh non-persistent file system to the nodes on every reboot
- Utilizes the advances in
  - high-speed interconnects
  - Per-node physical memory
  - Centralized storage infrastructure
- Light-weight client node images usually optimized for computation
Introduction to PERCEUS

- Successor to Warewulf, one of the de-facto industry standards for diskless clustering
- Large scale provisioning of stateless nodes
- Hybrid NFS-Ramdisk filesystem approach
- Single point of administration
- Certified as Intel Cluster Ready™
Architectural Overview

- **Database**
  - Maintains cluster configuration

- **Perceus master**
  - Administers and manages the Perceus client nodes

- **VNFS capsules**
  - Necessary information required for provisioning nodes

- **Slave nodes**
  - Primarily used for computation
Provisioning in Perceus

- Two-stage process
  - Compute node boots the Perceus OS
  - Perceus OS spawns the runtime OS kernel
- Nodes request VNFS capsule from master
- Virtual Node File System (VNFS)
  - Template image used to provision stateless nodes
  - A live root filesystem in the form of an image or archive
  - Packaged with configuration scripts and utilities to form a VNFS capsule
Integrating OSCAR and PERCEUS

- Thin-OSCAR is deprecated
- Fills much-needed niche in cluster computing
- Utilizes the meta-packaging format to leverage OSCAR core infrastructure
- Maintains maximum integrity of both the clustering toolkits
- Lots of issues to be dealt with
Architecture
Implementation Overview

- OSCAR acts as a front-end for the installation and management of the cluster
- Ability to tweak Perceus configuration using OSCAR Configurator API
- Perceus completely handles provisioning and system-level services used for interacting with compute nodes
- Replication of the cluster configuration database
Implementation

- Perceus OPKG
  - Perceus binary installation package
  - Scripts to initialize and configure Perceus to a working cluster environment
  - Perceus documentation

- Building Perceus VNFS Image
  - Utilizes Perceus scripts to build a VNFS image
  - Customizing these images with OPKGS

- OSCAR-Perceus Wrapper class
Status of the integration

- Vanilla cluster installation supporting basic cluster tools and MPI libraries using CLI
- Pending support for additional packages
- Disables features in OSCAR which are now provided by Perceus
  - Reduced flexibility in network configuration
- DB-bridge being reworked upon due to changes in Perceus DB backend in v1.4
- Tried and tested on RHEL only
Issues faced

- OSCAR and Perceus under continuous development
  - Pending merges of trunk and branches
  - Introduction of new features with upcoming releases
- Replication of system-level services and cluster configuration data
- No clean API for interaction between OSCAR and Perceus
- Towards a generic provisioning framework for OSCAR?
Generic Provisioning Framework

- Support for various provisioning components
  - Diskfull
  - Diskless
  - Virtualization
- Plugs into OSCAR using OCA
- Identifies commonality between various provisioning schemes
- Component-based architecture
A Closer Look

- Adds a layer of abstraction between OSCAR core components and SIS
- Provisioning schemes have in common
  - A way of
    - Defining images
    - Defining nodes or clients
    - Building and customizing images
    - Deploying images to the nodes
  - Storing cluster configuration data useful for provisioning
  - Minimal monitoring framework
OSCAR Provisioning component

- Interacts with the core OSCAR framework using a provisioning API
- Workflow defined as XML file describing the interaction and dependency between various provisioning events
- Implementation of these interfaces is found in available provisioning scheme components, e.g., Perceus OCA
Perceus OCA

- Perceus OPKG
  - Binary installation package
  - Additional scripts

- Interaction API
  - Images
    - List
    - Build
    - Deploy
  - Nodes
    - Define parameters
    - Network configuration
Conclusions

- Integration of OSCAR and Perceus results in added complexity and redundancy
- A better, more integrated approach is needed to support alternate provisioning schemes using OSCAR. This can be achieved by introducing an added layer of abstraction in the core framework
- Supporting various provisioning schemes would result in adoption of OSCAR over a wider range of cluster architectures
Thanks

- OSCAR community
- Infiscale, and the Perceus developers
- Open Systems Lab (OSL) guys
Questions?

adkulkar@cs.indiana.edu