Introduction to Roll Development

Rocks-A-Palooza III

Rocks Philosophy

- We’ve developed a “cluster compiler”
  - XML framework + XML parser + kickstart file generator
  - Source code + preprocessor + linker
- Think about “programming your cluster”
  - Not “administering your cluster”

Goal of Rolls

- Develop a method to reliably install software on a frontend

- "User-customizable" frontends

- Two established approaches:
  - Add-on method
  - Rocks method

Add-on Method

1. User responsible for installing and configuring base software stack on a frontend
2. After the frontend installation, the user downloads ‘add-on’ packages
3. User installs and configures add-on packages
4. User installs compute nodes

Major issue with add-on method

◆ The state of the frontend before the add-on packages are added/configured is unknown

ROCKS

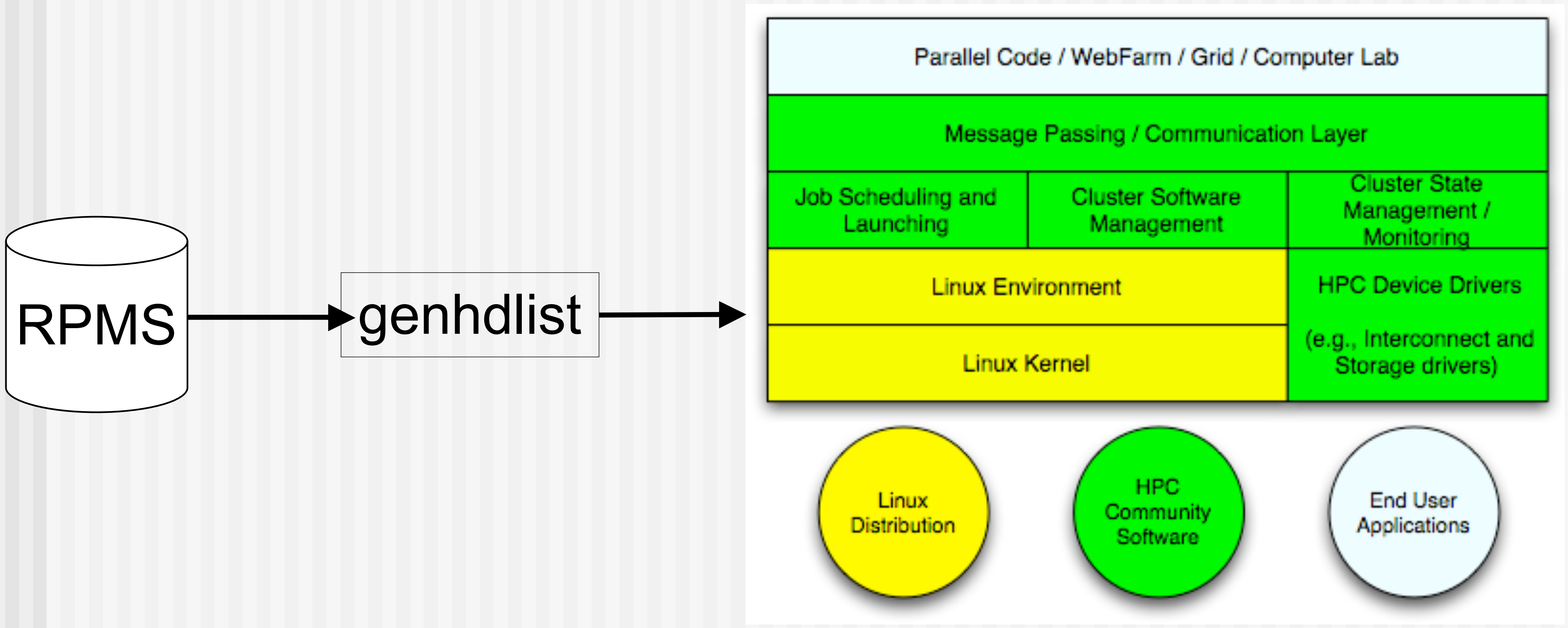
Rocks Method

- To address the major problem with the add-on method, we had the following idea:
  - All non-RedHat packages must be installed and configured in a controlled environment
- A controlled environment has a known state
- We chose the RedHat installation environment for the controlled environment

Goal of Rolls

- This led to modifying the standard RedHat installer in order to accept new packages and configuration
- A tricky proposition
  - A RedHat distribution is a monolithic entity
    - It’s tightly-coupled
    - In RHEL 4, a program called “genhdlist” creates binary files (hdlist and hdlist2) that contain metadata about every RPM in the distribution
- To add/remove/change an RPM, you need to re-run genhdlist
  - Else, the RedHat install will not recognize the package
  - Or worse, it fails during package installation

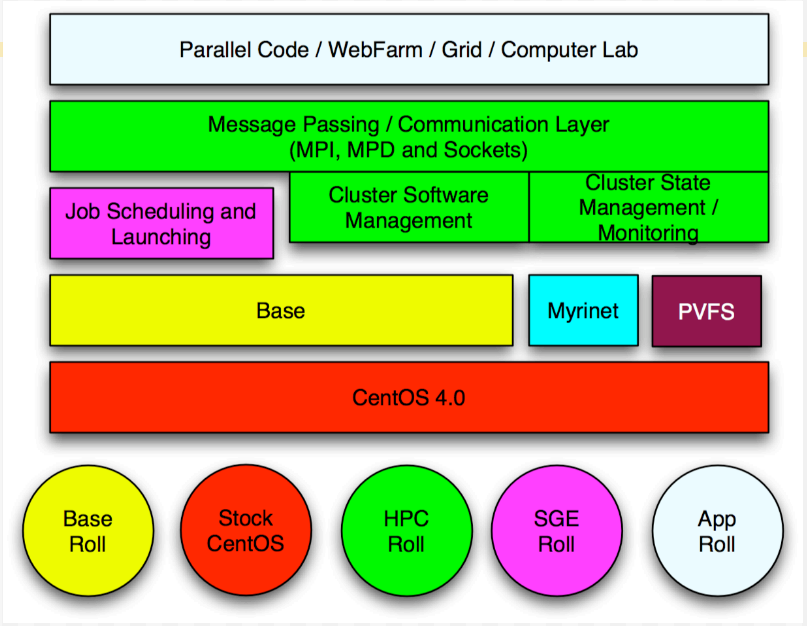
Monolithic Software Stack


RPMS
genhdlist
Parallel Code / WebFarm / Grid / Computer Lab
Message Passing / Communication Layer
Job Scheduling and Launching
Cluster Software Management
Cluster State Management / Monitoring
Linux Environment
Linux Kernel
HPC Device Drivers
(e.g., Interconnect and Storage drivers)
Linux Distribution
HPC Community Software
End User Applications

Goal of Rolls

- Problem: To make the frontend user-customizable at installation time, we needed a mechanism that could accept new packages
- And, we still wanted to leverage the RedHat installer
  - We don’t want to be in the installer business
- Solution: Our implementation makes the RedHat installer “think” it is just installing a monolithic RedHat distribution

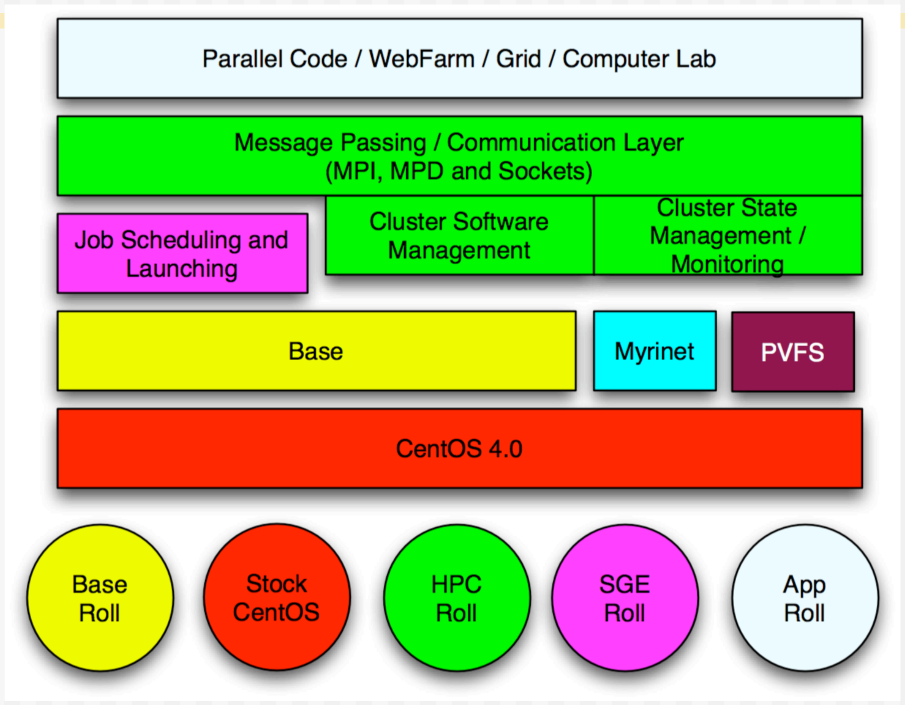
Goal of Rolls



Parallel Code / WebFarm / Grid / Computer Lab
Message Passing / Communication Layer (MPI, MPD and Sockets)
Job Scheduling and Launching
Cluster Software Management
Cluster State Management / Monitoring
Base
Myrinet
PVFS
CentOS 4.0
Base Roll
Stock CentOS
HPC Roll
SGE Roll
App Roll

- How do you make all the packages above look like a monolithic distribution?
  - Easy! Just run “genhdlist” at release time!
- But, how do you do it when some of the above blocks are optional and/or unknown?
  - An “unknown” block is one produced after the release or by a third-party

Rolls Function and Value


Parallel Code / WebFarm / Grid / Computer Lab
Message Passing / Communication Layer (MPI, MPD and Sockets)
Job Scheduling and Launching
Cluster Software Management
Cluster State Management / Monitoring
Base
Myrinet
PVFS
CentOS 4.0
Base Roll
Stock CentOS
HPC Roll
SGE Roll
App Roll

- Function: Rolls extend/modify stock RedHat
- Value: Third parties can extend/modify Rocks
  - Because Rolls can be optional

The RedHat Installer

Anaconda: RedHat’s Installer

- Open-source python-based installer
- Developed by RedHat
- (Somewhat) object-oriented
  - We extend when we can and insert “shims” when we can’t

Anaconda: RedHat’s Installer

- Key tasks:
  - Probe hardware
  - Ask users for site-specific values
    - E.g., IP addresses and passwords
  - Insert network and storage drivers
    - For network-based installations and to write packages down onto local disk
  - Install packages
    - RPMs
  - Configure services
    - Via shell scripts

Scripted Installation

- Anaconda achieves “lights-out” installation via kickstart mechanism
- It reads a “kickstart file”
  - Description of how to install a node
- One file composed of three key sections:
  - Main: general parameters
  - Packages: list of RPMs to install
  - Post: scripts to configure services

Kickstart File

- Main section

```
rootpw --iscrypted loijgoij5478fj2i9a
zerombr yes
bootloader --location=mbr
lang en_US
langsupport --default en_US
keyboard us
mouse genericps/2
install
reboot
timezone --utc America/Los_Angeles
part
```

Kickstart File

- Packages section

```
%packages --ignoredeps --ignoremissing
@Base
PyXML
atlas
autofs
bc
chkrootkit
contrib-pexpect
contrib-pvfs-config
contrib-python-openssl
```

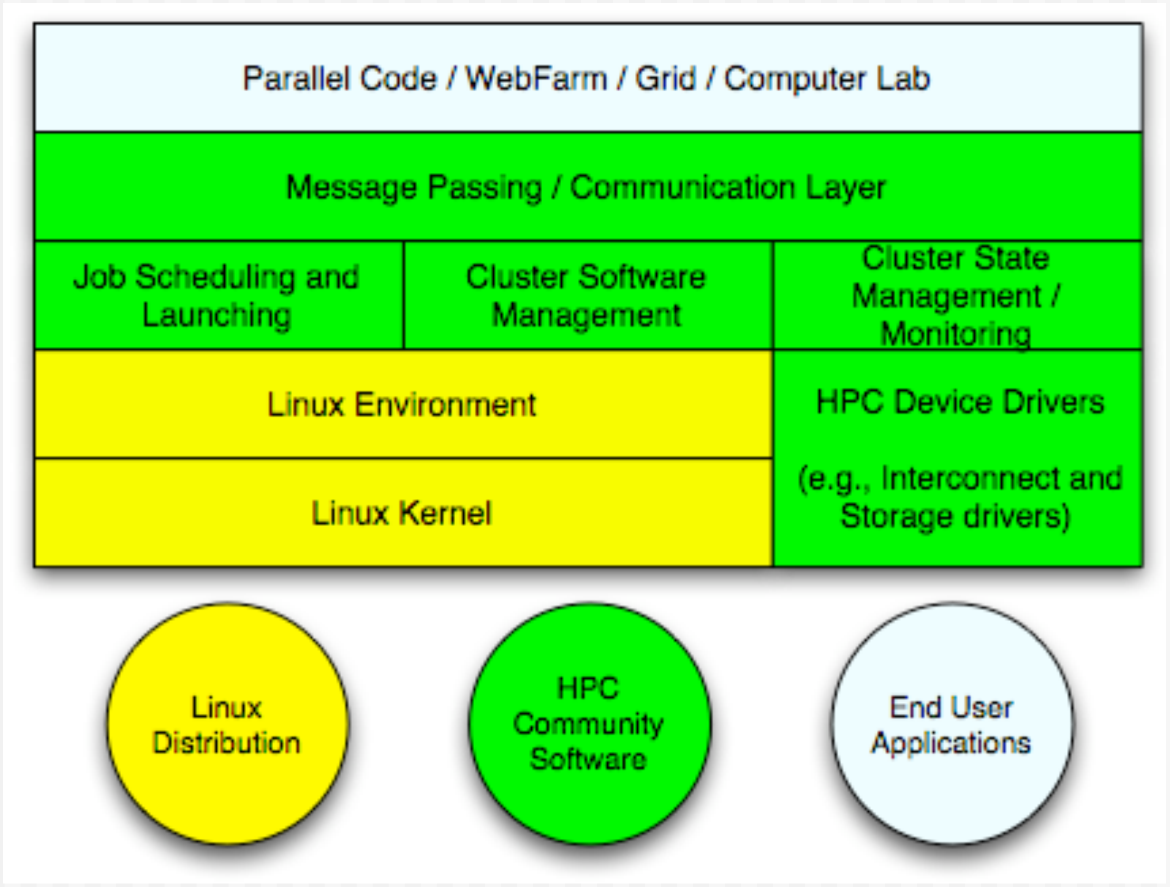
Kickstart File

- Post section

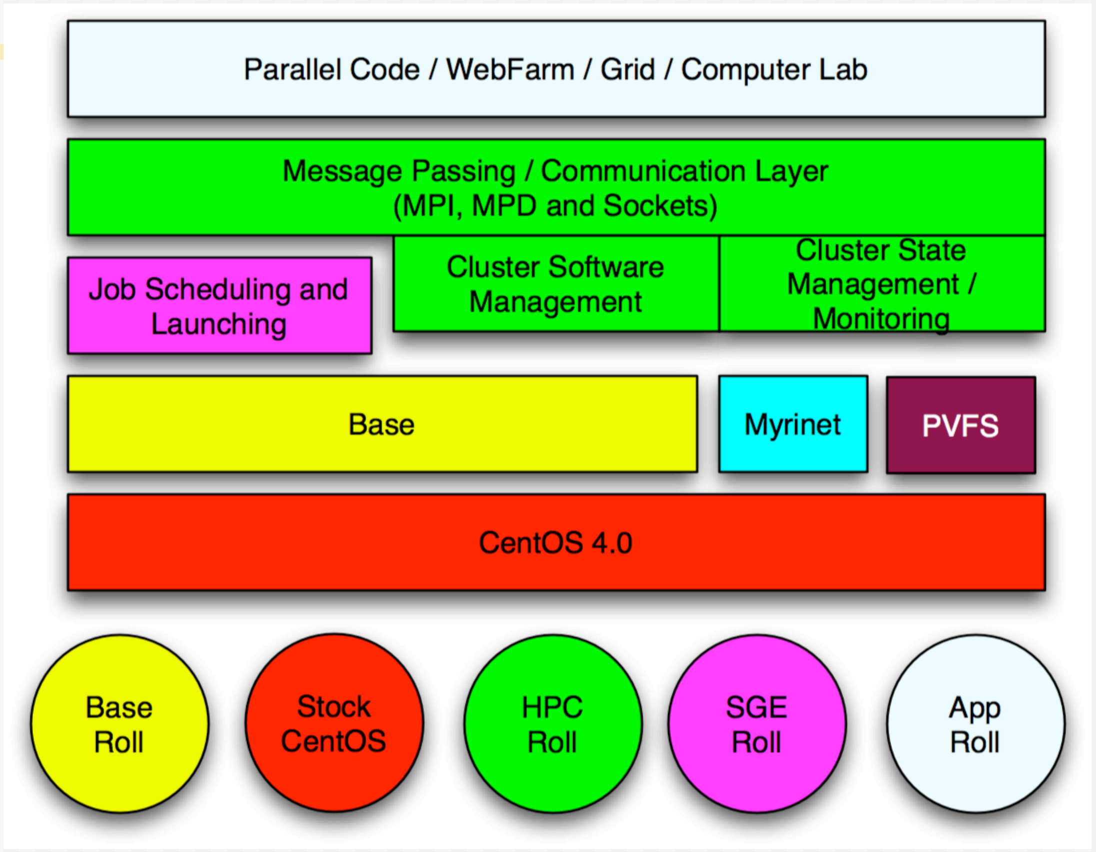
```
%post

cat > /etc/motd << 'EOF'
Rocks Compute Node
EOF
```

Rolls High-Level Description

Monolithic Software Stack

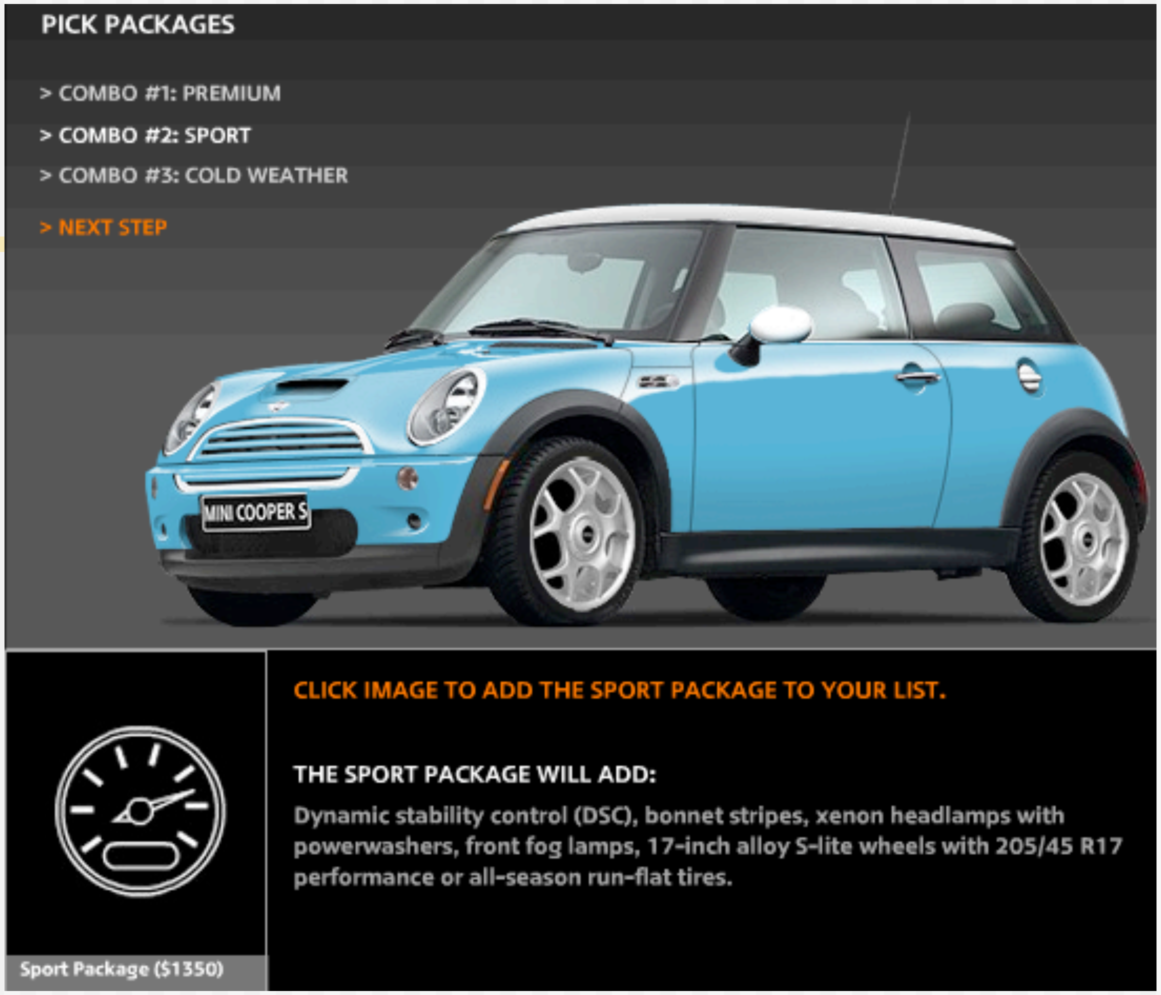

Parallel Code / WebFarm / Grid / Computer Lab
Message Passing / Communication Layer
Job Scheduling and Launching
Cluster Software Management
Cluster State Management / Monitoring
Linux Environment
Linux Kernel
HPC Device Drivers
(e.g., Interconnect and Storage drivers)
Linux Distribution
HPC Community Software
End User Applications

Rolls


Parallel Code / WebFarm / Grid / Computer Lab
Message Passing / Communication Layer (MPI, MPD and Sockets)
Job Scheduling and Launching
Cluster Software Management
Cluster State Management / Monitoring
Base
Myrinet
PVFS
CentOS 4.0
Base Roll
Stock CentOS
HPC Roll
SGE Roll
App Roll

- Dissecting the monolithic software stack

Rolls


PICK PACKAGES

> COMBO #1: PREMIUM

> COMBO #2: SPORT

> COMBO #3: COLD WEATHER

> NEXT STEP

CLICK IMAGE TO ADD THE SPORT PACKAGE TO YOUR LIST.

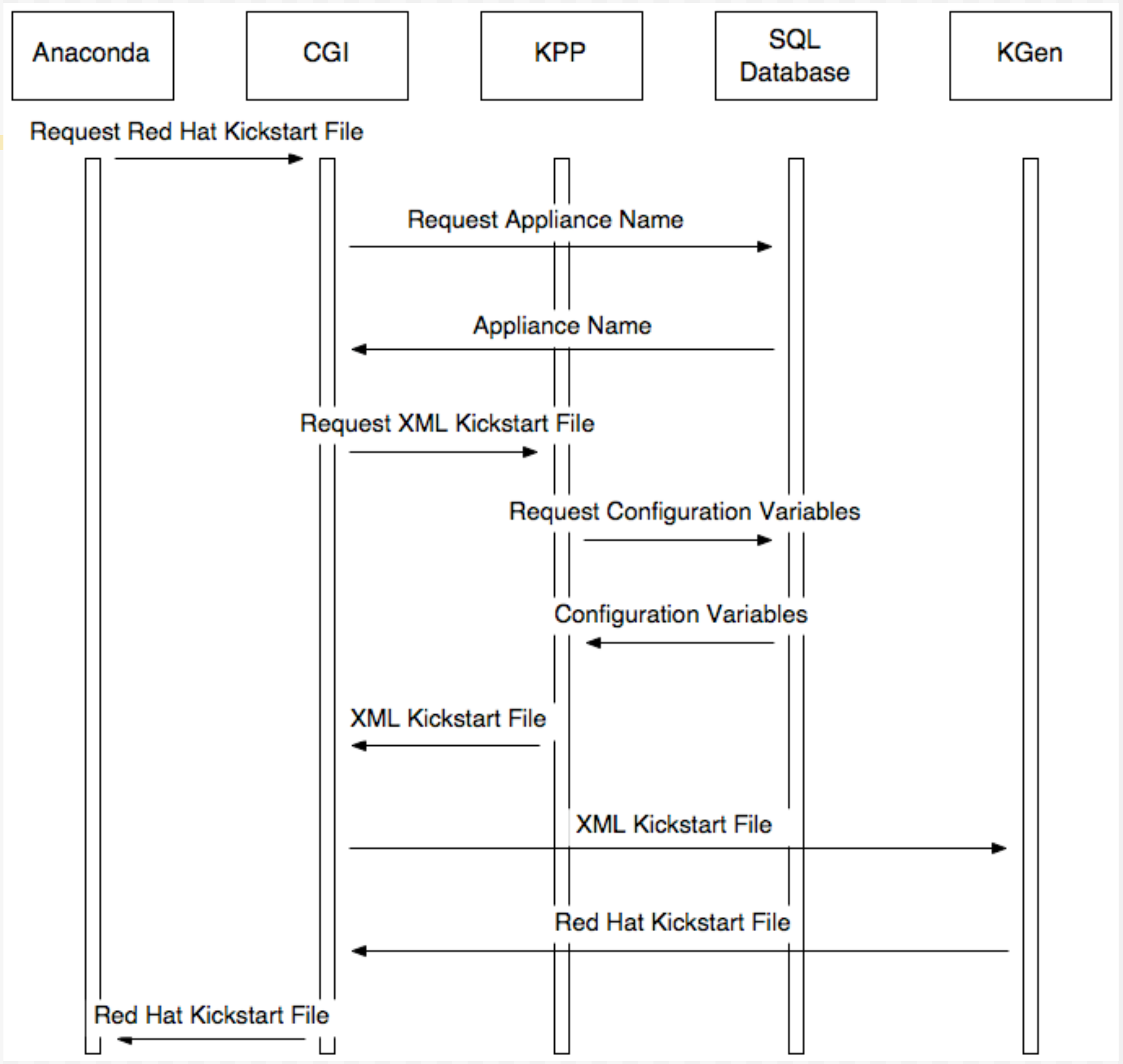
THE SPORT PACKAGE WILL ADD:

Dynamic stability control (DSC), bonnet stripes, xenon headlamps with powerwashers, front fog lamps, 17-inch alloy S-lite wheels with 205/45 R17 performance or all-season run-flat tires.

Sport Package ($1350)

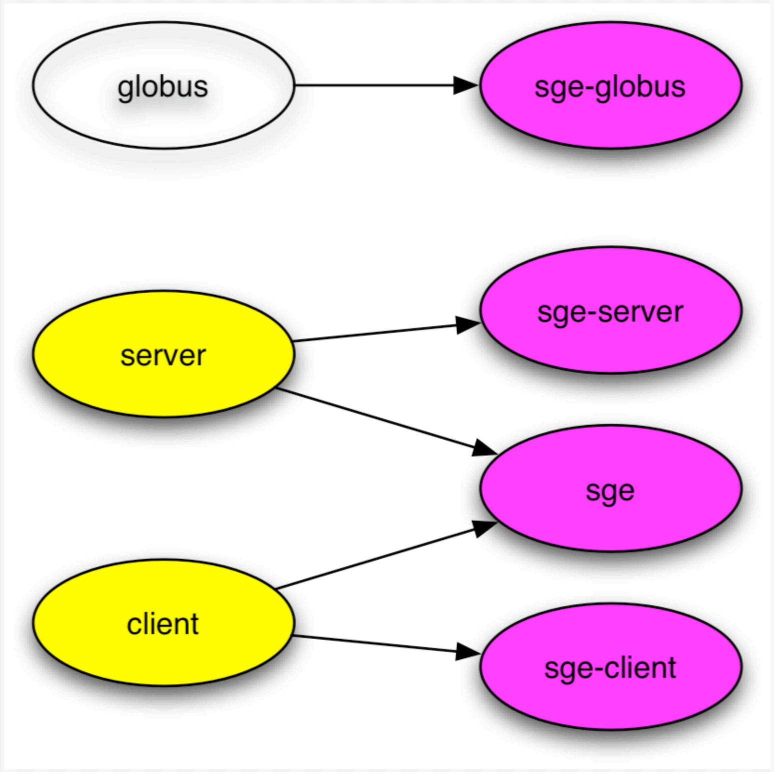
- Think of a roll as a “package” on a car

Getting A Kickstart File

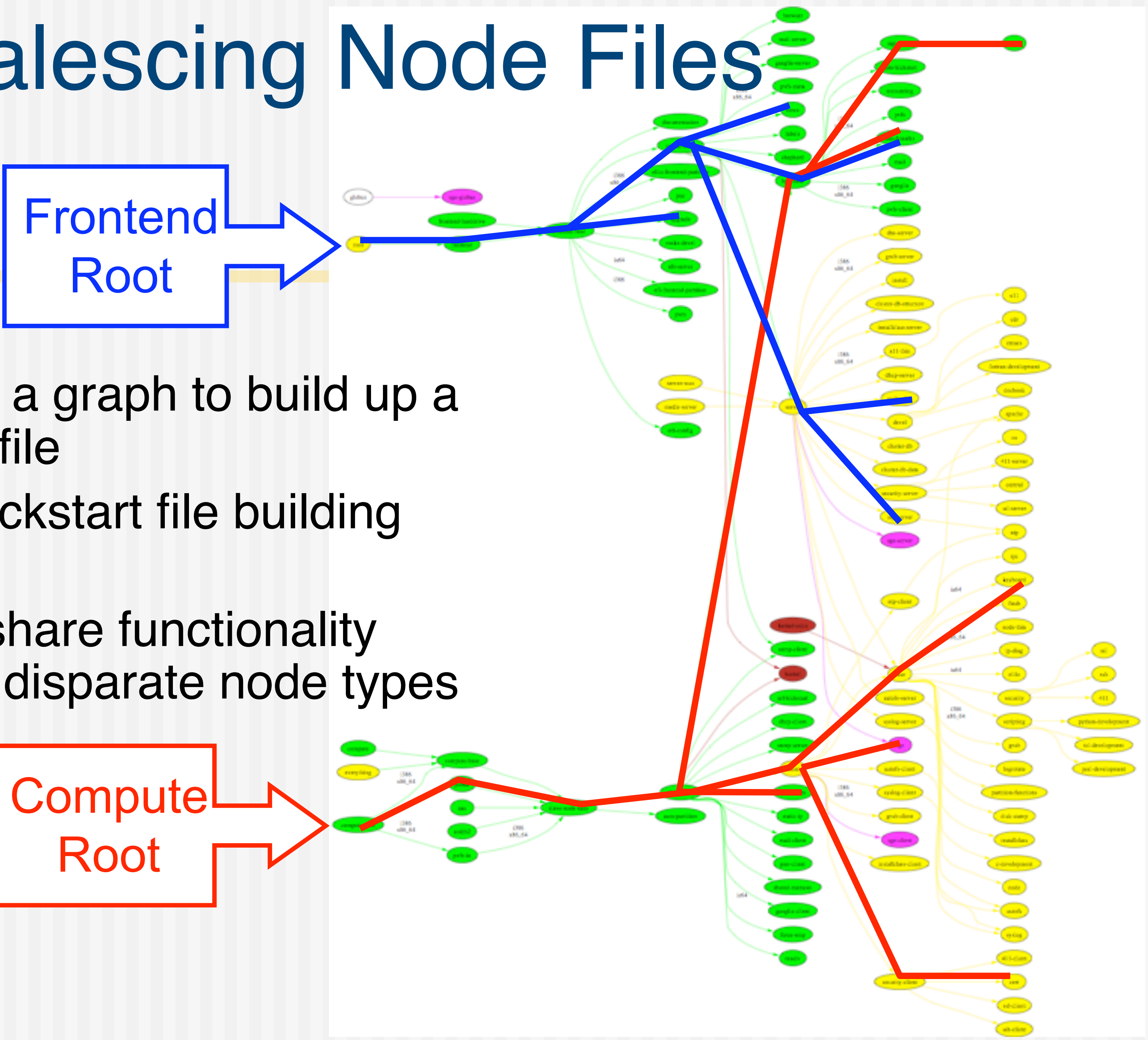


Use Graph Structure to Dissect Distribution

- Use ‘nodes’ and ‘edges’ to build a customized kickstart file
- Nodes contain portion of kickstart file
  - Can have a ‘main’, ‘package’ and ‘post’ section in node file
- Edges used to coalesce node files into one kickstart file



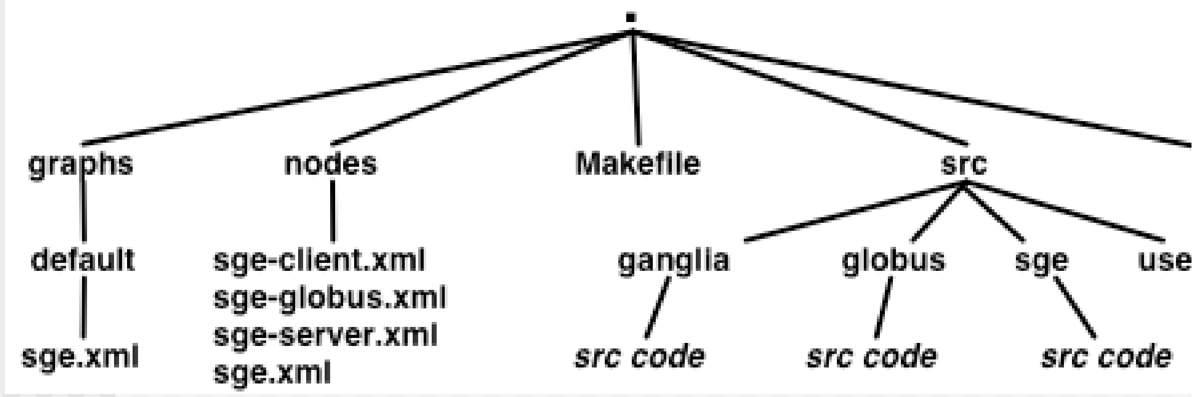
Coalescing Node Files

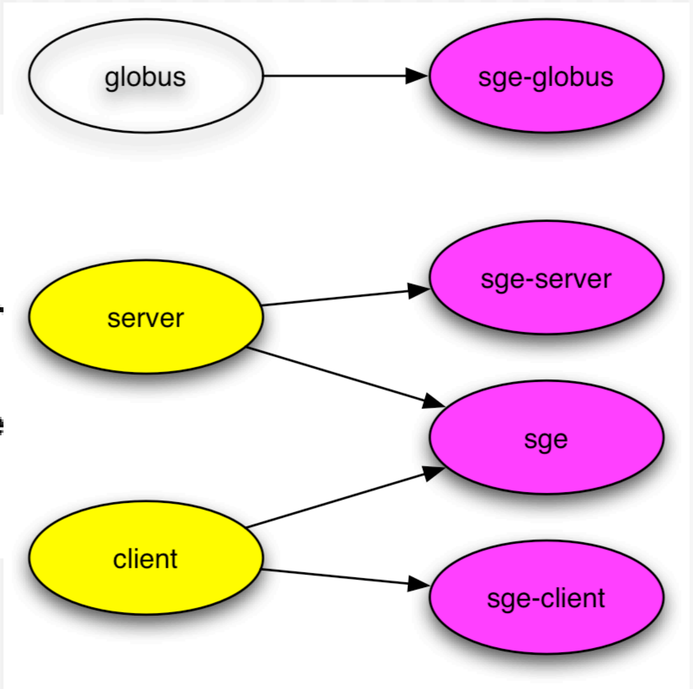


- Traverse a graph to build up a kickstart file
- Makes kickstart file building flexible
- Easy to share functionality between disparate node types

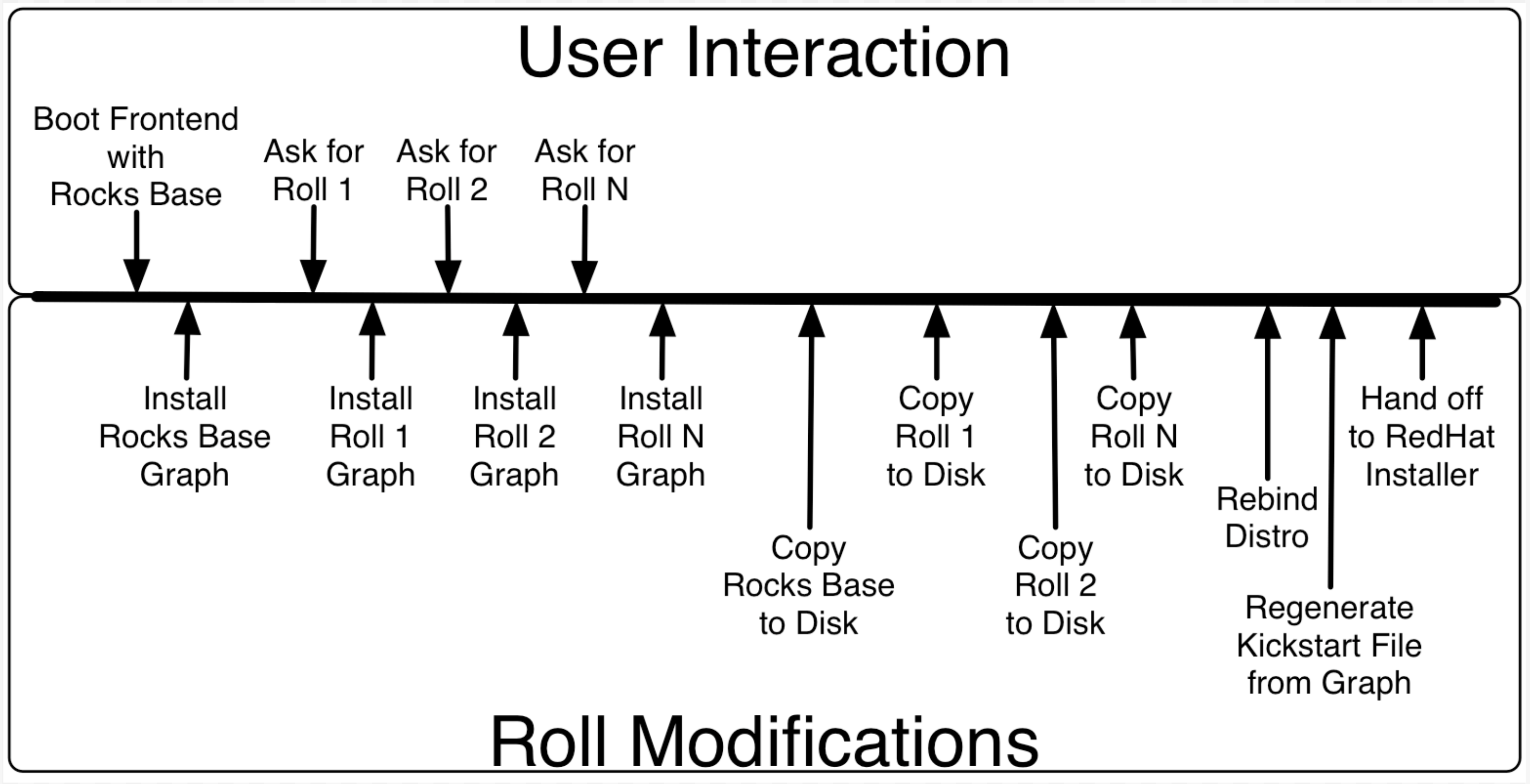
Why We Use A Graph

- A graph makes it easy to ‘splice’ in new nodes
- Each Roll contains its own nodes and splices them into the system graph file

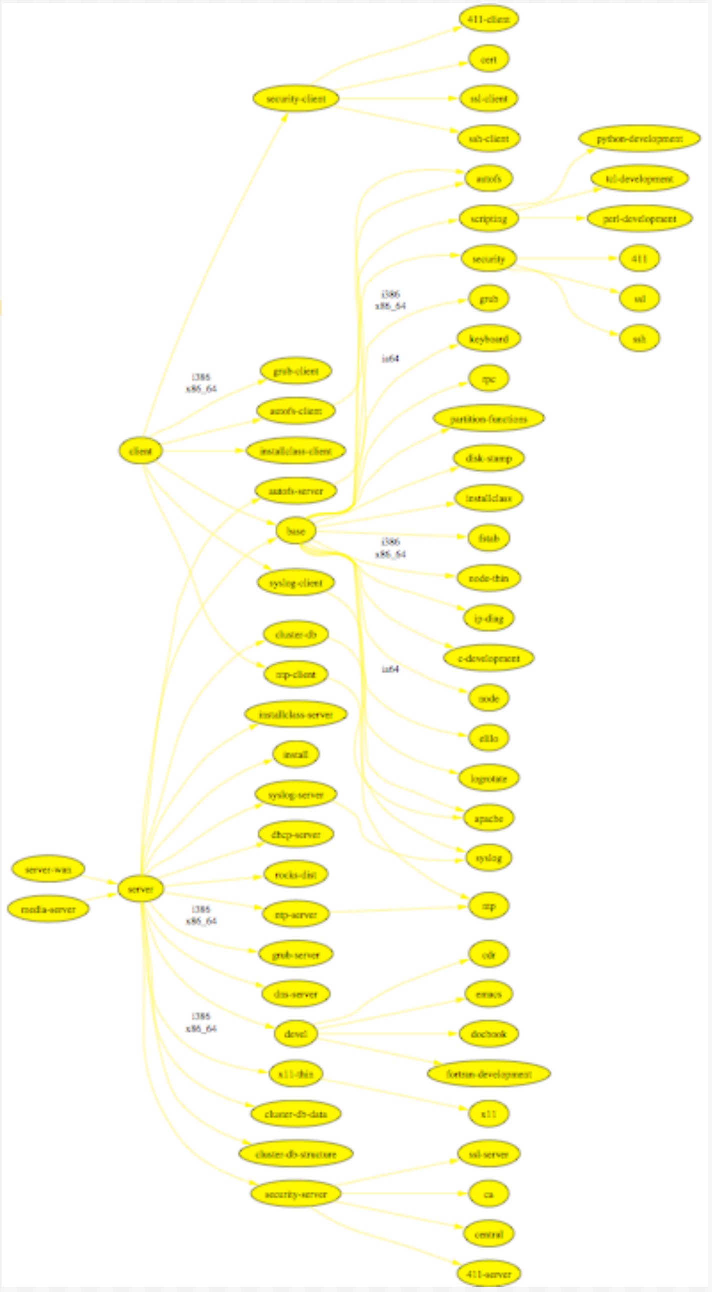




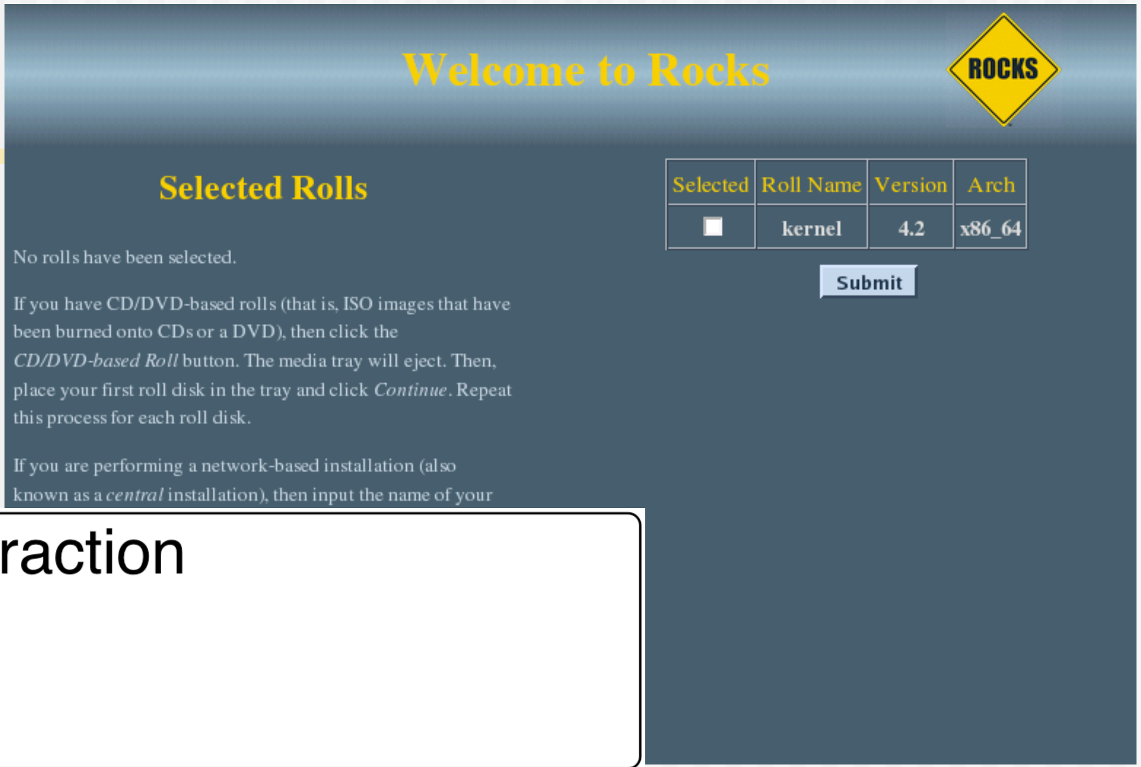
Rocks Extensions Installation Timeline



Install Rocks Base Graph



Anaconda Modified to Accept Rolls



Welcome to Rocks

Selected Rolls

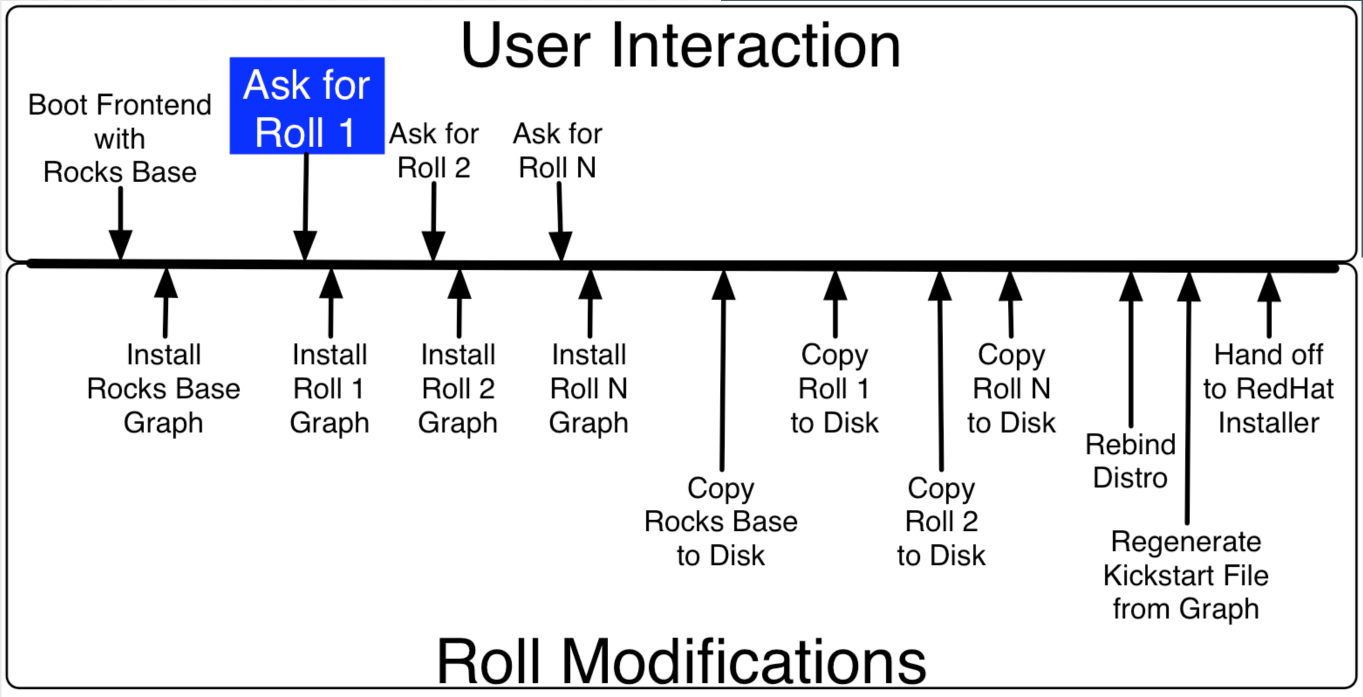
Selected	Roll Name	Version	Arch
☐	kernel	4.2	x86_64

Submit

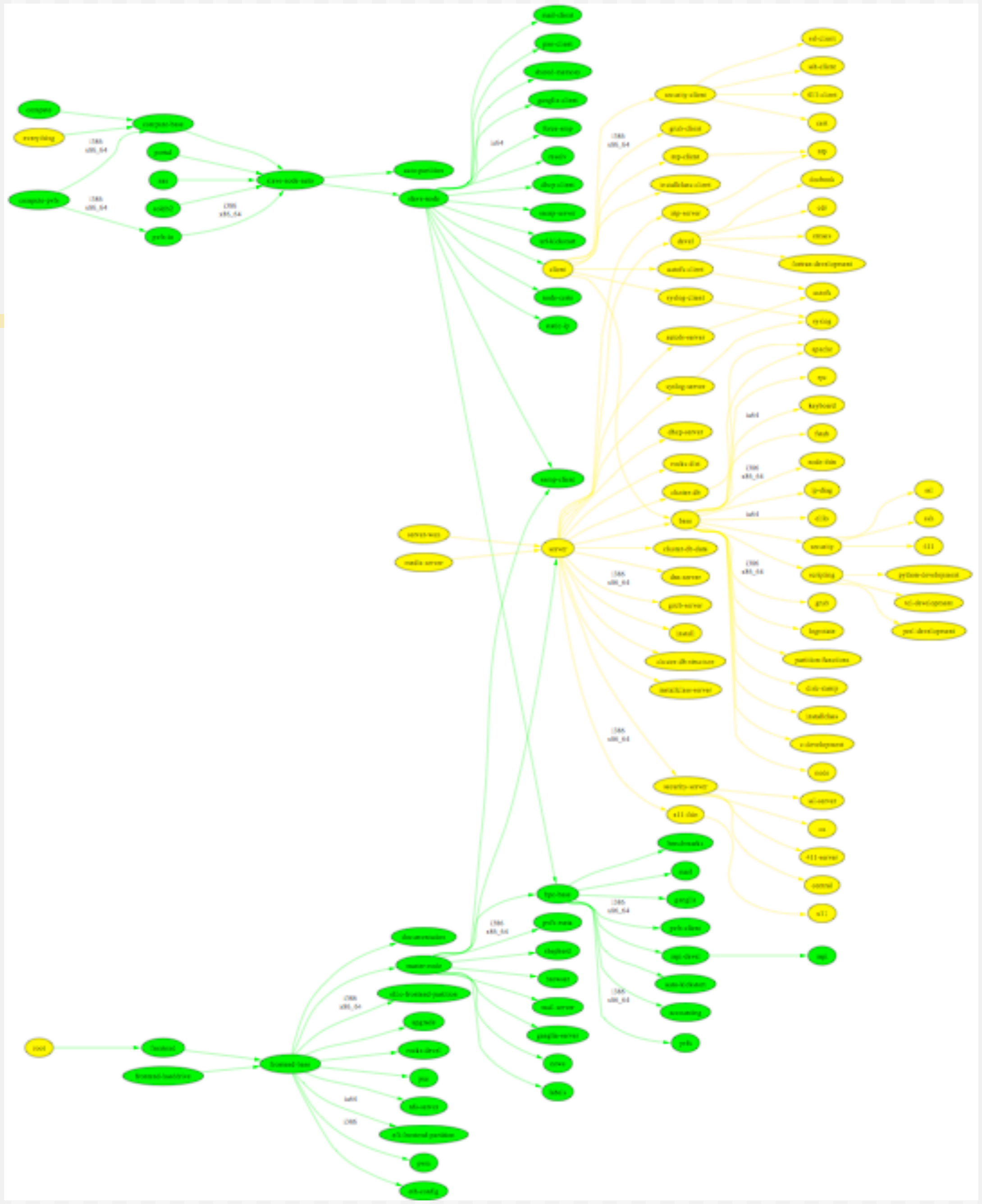
No rolls have been selected.

If you have CD/DVD-based rolls (that is, ISO images that have been burned onto CDs or a DVD), then click the *CD/DVD-based Roll* button. The media tray will eject. Then, place your first roll disk in the tray and click *Continue*. Repeat this process for each roll disk.

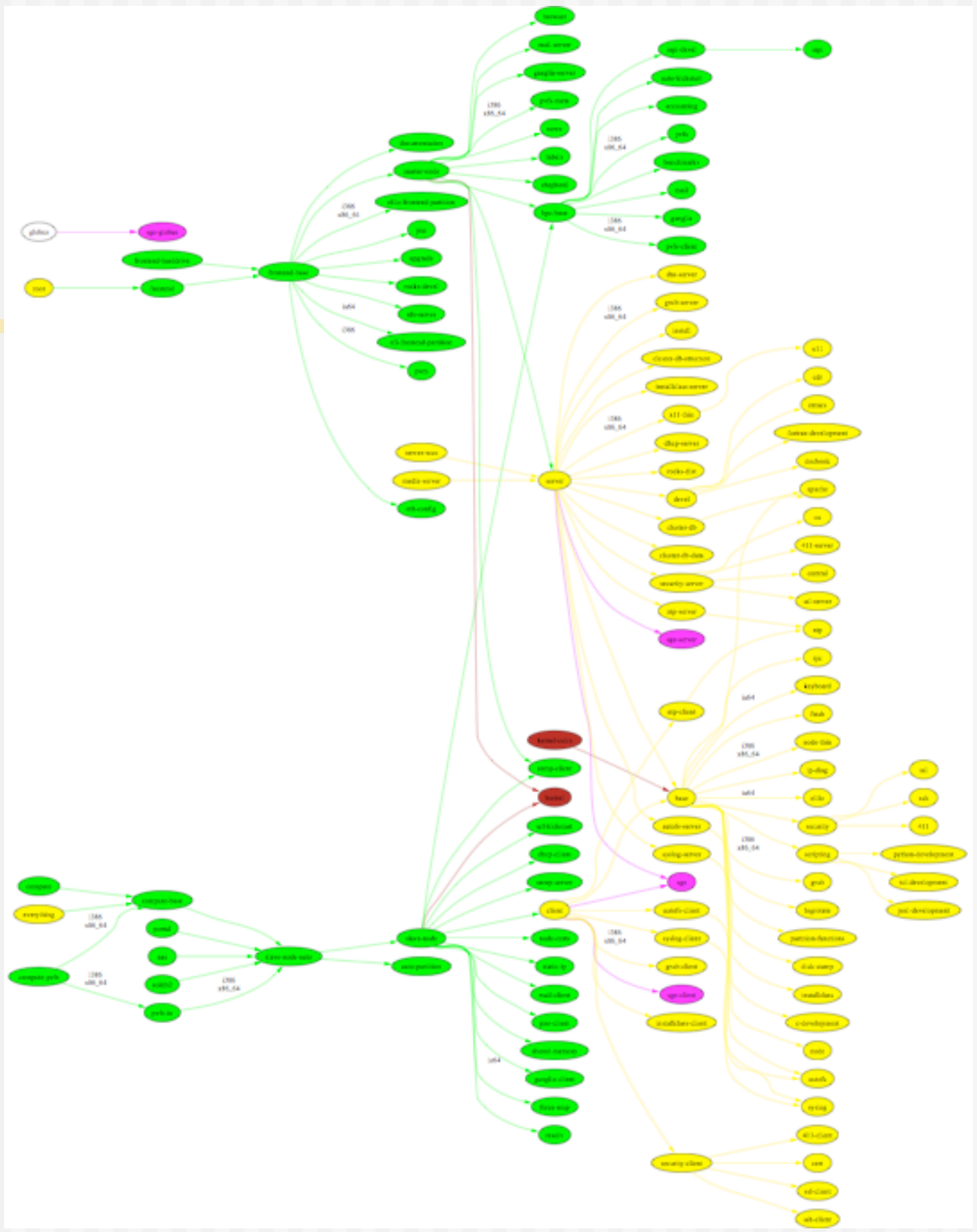
If you are performing a network-based installation (also known as a *central* installation), then input the name of your



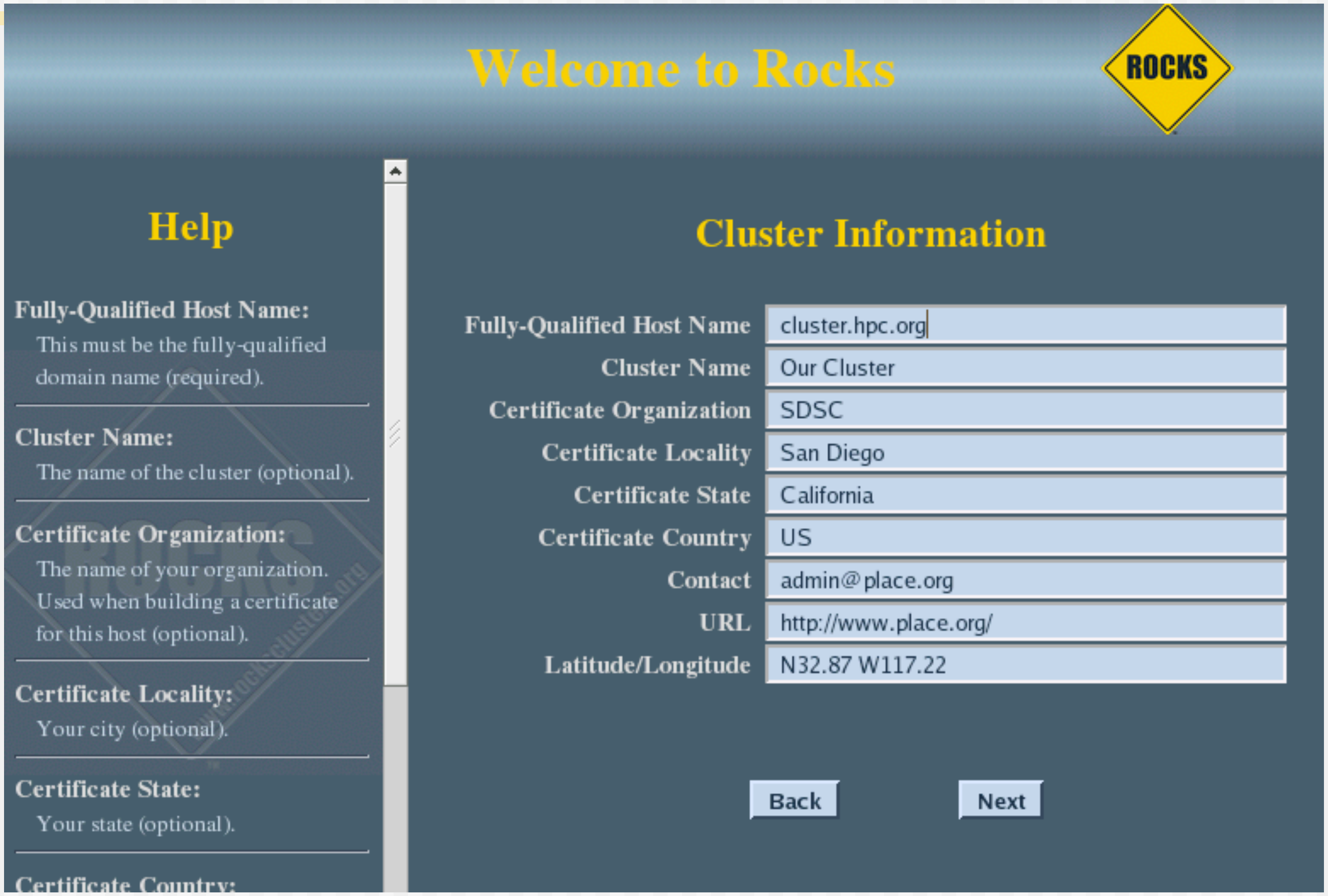
Install Roll Graph



Base + All Rolls



Anaconda Modified to Display New User Input Screens



Help

Fully-Qualified Host Name:
This must be the fully-qualified domain name (required).

Cluster Name:
The name of the cluster (optional).

Certificate Organization:
The name of your organization. Used when building a certificate for this host (optional).

Certificate Locality:
Your city (optional).

Certificate State:
Your state (optional).

Certificate Country:

Cluster Information

Fully-Qualified Host Name	cluster.hpc.org
Cluster Name	Our Cluster
Certificate Organization	SDSC
Certificate Locality	San Diego
Certificate State	California
Certificate Country	US
Contact	admin@place.org
URL	http://www.place.org/
Latitude/Longitude	N32.87 W117.22

Anaconda Modified to Display New User Input Screens

- How we do it:
  - Place a shim in Anaconda to call our screens instead of the ‘betanag’ RedHat screen

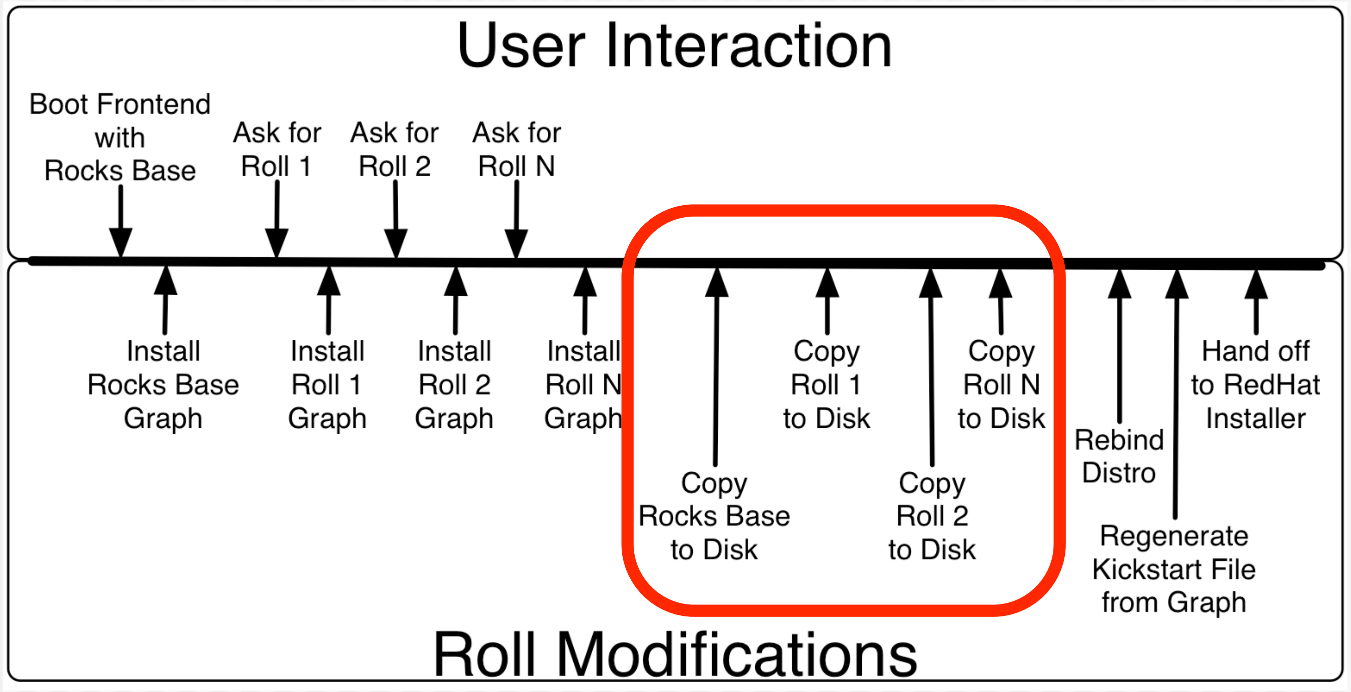
```
index = 0
for key in installSteps:
        if key[0] == "betanag":
                break
        index = index + 1

installSteps[index] = ("rockswindows", ("id.configFileData",))

stepToClass["rockswindows"] = ("ksclass",
                "RocksWelcomeWindowGUI")
```

Anaconda Modified to Display New User Input Screens

- Inside an XML node file, you'll see:

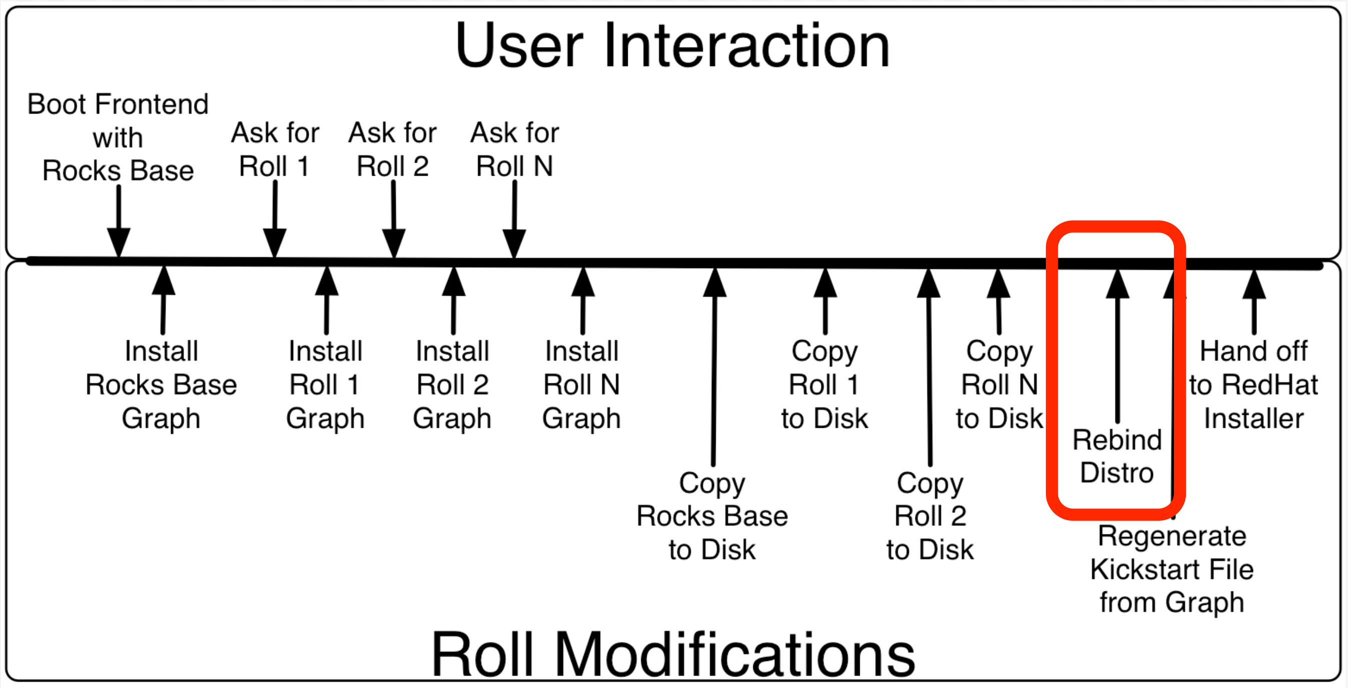
```
<screen>    <title>Root Password</title>

        <code>
                <!-- the 'validate' functions are in this file -->
                <include file="javascript/password.js"/>
        </code>

        <variable>
                <label>Password</label>
                <name>Private_PureRootPassword</name>
                <type>password</type>
                <size>20</size>
                <value><var name="Private_PureRootPassword"/></value>
                <help>The root password for your cluster.</help>
        </variable>

        <variable>
                <label>Confirm</label>
                <name>Confirm_Private_PureRootPassword</name>
                <type>password</type>
                <size>20</size>
                <value><var name="Confirm_Private_PureRootPassword"/></value>
                <validate>confirm_password</validate>
        </variable>

</screen>
```

Copy Media To Local Disk



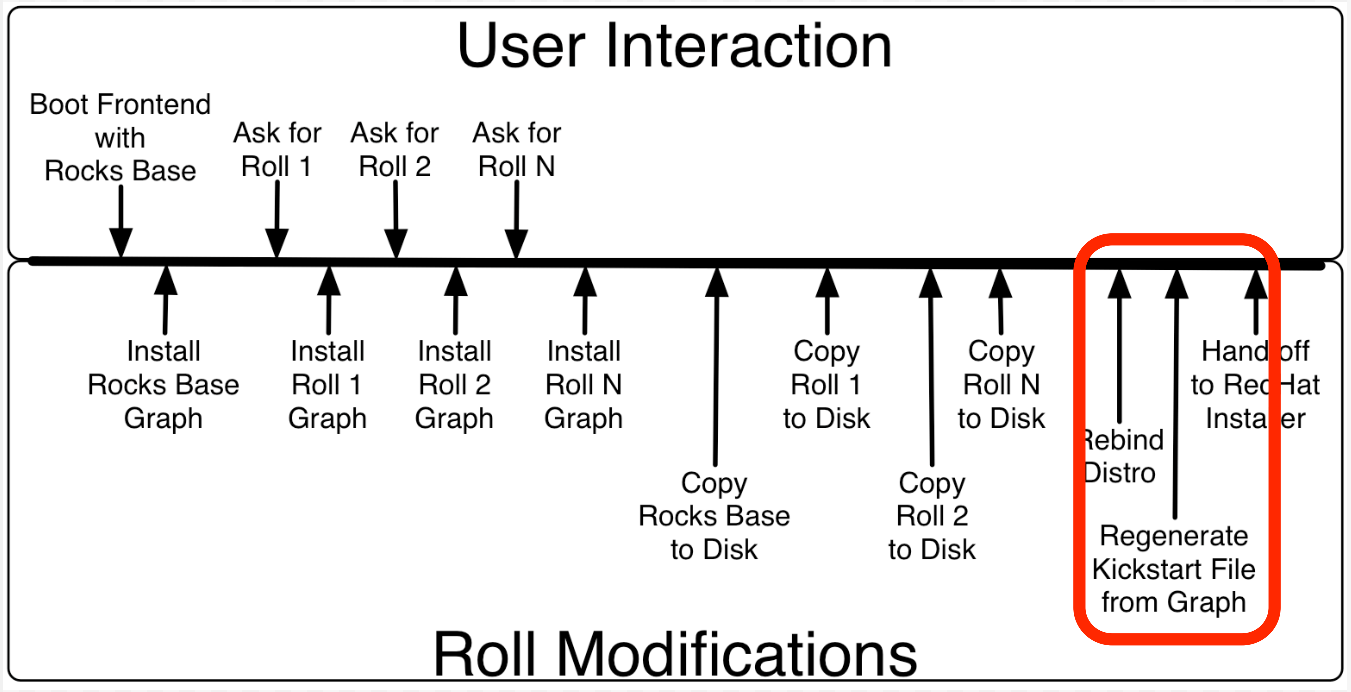
- Base and all user-supplied Rolls are copied to local disk
  - These packages are used to install compute nodes

Rebind Distro



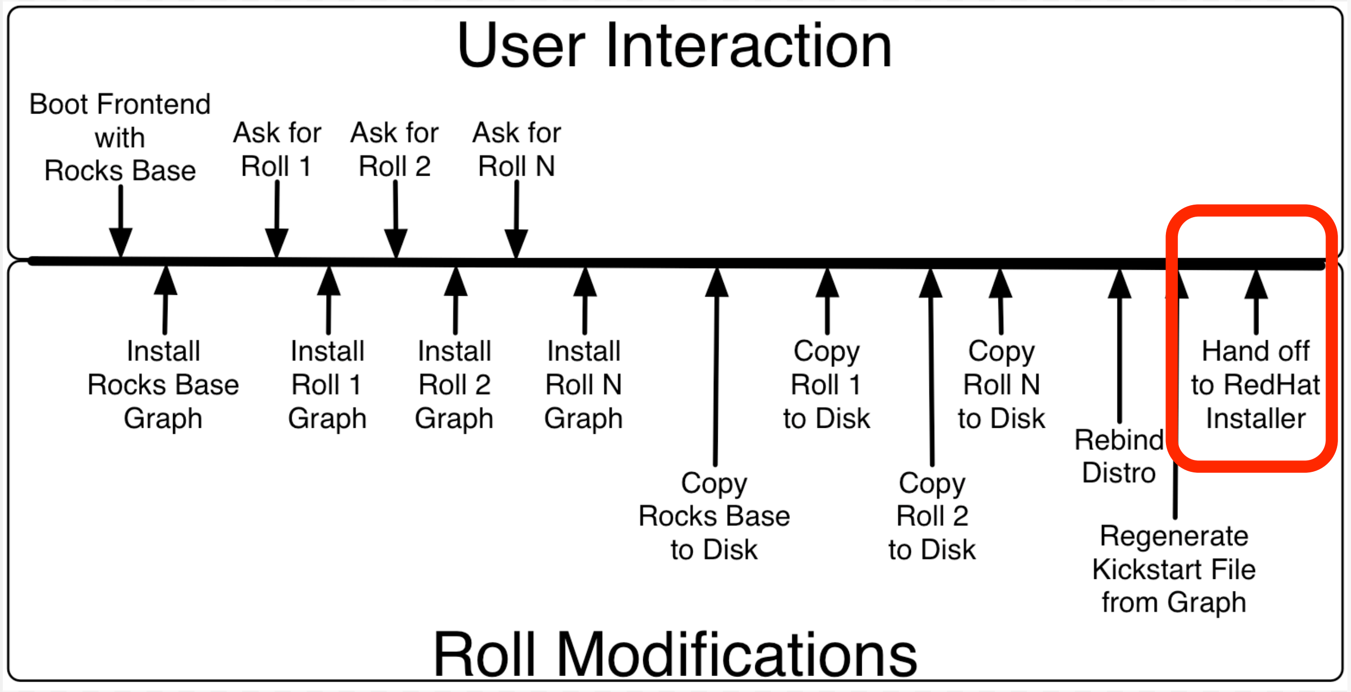
- Merge base with rolls into one RedHat-compliant distribution
  - This takes the dissected distro and tightly binds it
    - Note: We actually install the frontend off the local hard disk (not the CD media)

Rebuild the Kickstart File



- Traverse the final graph using the node 'frontend' as the root
  - Allows us to customize a frontend configuration at install time

Hand Off To RedHat



- Anaconda has no idea what hit it!
- The remainder of the installation looks like a standard RedHat installation (just with more packages and cluster-specific configuration)

Near Future

Rocks Futures

- Rocks 4.3
  - Rocks command line
    - General form:
      - rocks <verb> <modifier> <component> <host1> <host2>
    - For example:
      - rocks-partition --list --nodename compute-0-0
    - Becomes:
      - rocks list host partition compute-0-0
  - Viz Roll x86_64 version
    - Now using all the bits!

Rocks Futures

- Rocks 4.3
  - PXE First
    - Change compute node boot order from:
      - CD, Hard Disk, PXE
    - To:
      - CD, PXE, Hard Disk
    - Enables easy ways in which to:
      - Execute 'memtest86' on compute nodes
      - Flash BIOS
      - 'Headless' installs on groups of nodes
- Release: End of June 2007

Rocks Futures

- Rocks 5.0
  - Base OS will be RHEL 5
    - Key technology in RHEL 5 is Xen
- Release: December 2007 (at the earliest)