Università della Svizzera italiana | Faculty of Informatics | Institute of Computational Science ICS

Experiences with the Rocks Distribution for Cluster Deployment at USI

Cristian Bianchi, Arne Dirks, **Dorian Krause** and Rolf Krause

Università della Svizzera Italiana
Lugano

hpc-ch forum, Lausanne, May 20, 2010

Rocks Cluster Distribution

- Open-Source Linux Distribution designed for clusters, grid endpoints, clouds, …
- Addresses difficulties of deploying manageable clusters
- Developed mainly at San Diego Supercomputer Center, sponsored by NFS
- www.rockscluster.org

Why Rocks?

- Promises ….
  - Faster installation and updates
  - Stability
  - Large user base, hence community support
  - Easy to set up also for inexperienced users

Rocks Details

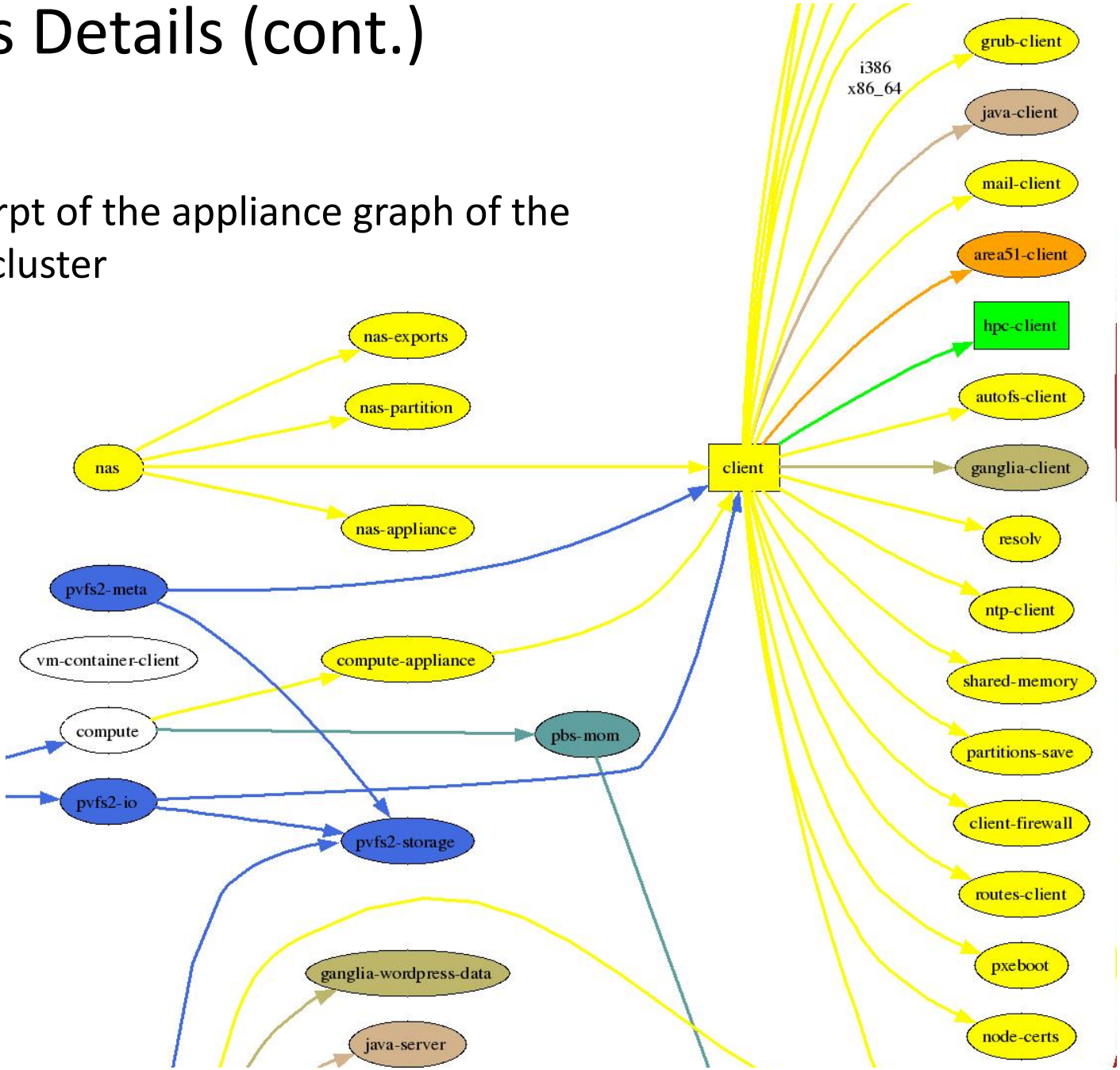
- Views cluster as a collection of **appliances**
  - frontend
  - compute server
  - NAS
  - I/O server, metadata server (e.g., pvfs2)
- User can create new appliance types via XML files
- Each appliance is decomposed into smaller single-purpose configuration modules.

Rocks Details (cont.)

- Rocks leverages Red Hat's **Kickstart** and **Anaconda**
  - Allows to re-use Kickstart files and hide details from the user
- Module description and inheritance described by XML files
  - /export/rocks/install/rocks-dist/x86_64/build/nodes contains appliances, packages, service configurations …
  - /export/rocks/install/rocks-dist/x86_64/build/graphs specifies inheritance diagram

Rocks Details (cont.)

Excerpt of the appliance graph of the cub cluster



Rocks Details (cont.)

```
 ...
<package>openssh</package>
 <package>openssh-clients</package>
 <package>openssh-server</package>
 <package>openssh-askpass</package>

 <package>xorg-x11-xauth</package>

<post>

<file name="/etc/ssh/ssh_config">
Host *
      CheckHostIP          no
      ForwardX11           yes
      ForwardAgent          yes
      StrictHostKeyChecking  no
      UsePrivilegedPort      no
      Protocol             2,1
</file>
...
```

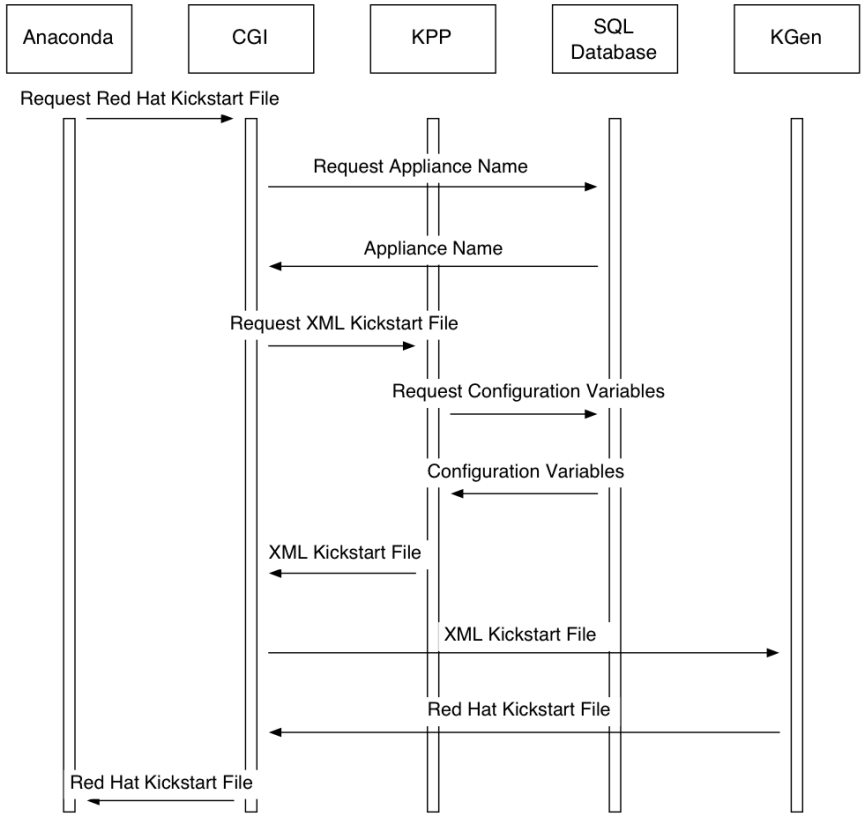
Rocks Details (cont.)

- XML Files are combined with MySQL database to add site specific variables

```
...
<package>openssh</package>
 <package>openssh-clients</package>
 <package>openssh-server</package>
 <package>openssh-askpass</package>

 <package>xorg-x11-xauth</package>

<post>

<file name="/etc/ssh/ssh_config">
Host *
     CheckHostIP         no
     ForwardX11          yes
...
```

```
mysql> select * from app_globals;
+----+------+------------+-----------+-----------------------------+-----------------------------------------+
| ID | Site | Membership | Service   | Component                   | Value                                   |
+----+------+------------+-----------+-----------------------------+-----------------------------------------+
|  1 |    0 |          0 | Kickstart | PrivateKickstartBasedir     | install                                 |
|  2 |    0 |          0 | Kickstart | DistroDir                   | /export/rocks                           |
|  3 |    0 |          0 | Info      | ClusterURL                  | http://www.usi.ch                       |
+----+------+------------+-----------+-----------------------------+-----------------------------------------+
```

Rocks Details (cont.)



Picture taken from:
Leveraging Standard Core Technologies to Programmatically Build Linux Cluster Appliances.
Mason J. Katz, Philip M. Papadopoulos and Greg Bruno, Cluster 2002: IEEE Int. Conf. on Cluster Computing

- insert-ethers on frontend captures DHCP request
  - User can choose appliance type
- Must be done sequentially, one-by-one

Rocks Details (cont.)

- **Rolls:**
  - ISO image with packages and configuration
  - Integrates into the rocks distribution
- **Standard rolls:**
  - area 51, hpc, os, torque, ganglia, pvfs2, sge, viz
- Other rolls (e.g., TotalView, PBS Professional, Intel Developer Roll, PGI Roll, …) available for purchase from Clustercorp
- **Admin tools**: shoot-node, tentakel, cluster-fork, …

Bladecenter H cluster

- 3× IBM Bladecenter H
  - 3×14 = 42 LS22 servers
  - 2× Quad-core Opteron 2384, 2.7 GHz per server
  - 16 GB RAM per server (2 GB per core)
  - 4x DDR Infiniband
- IBM x3665 login node
  - 2× Quad-core Opteron 2384 HE

Software

- Rocks Cluster V. 5.1 x86_64 based on Cent OS 5 linux
- Installed in June 2009 for one BladeCenter H
  - First installation attempts failed with an Anaconda error which we were able to work around
- Upgrade to three BladeCenters painless
- Installed rolls:

```
mysql> select * from rolls;
+------+------------+---------+--------+-------+---------+
| Site | Name       | Version | Arch   | OS    | Enabled |
+------+------------+---------+--------+-------+---------+
|    0 | area51     | 5.1     | x86_64 | linux | yes     |
|    0 | base       | 5.1     | x86_64 | linux | yes     |
|    0 | ganglia    | 5.1     | x86_64 | linux | yes     |
|    0 | hpc        | 5.1     | x86_64 | linux | yes     |
|    0 | java       | 5.1     | x86_64 | linux | yes     |
|    0 | kernel     | 5.1     | x86_64 | linux | yes     |
|    0 | os         | 5.1     | x86_64 | linux | yes     |
|    0 | web-server | 5.1     | x86_64 | linux | yes     |
|    0 | torque     | 5.1.0   | x86_64 | linux | yes     |
|    0 | pvfs2      | 5.1     | x86_64 | linux | yes     |
|    0 | mlnx-ofed  | 5.1     | x86_64 | linux | yes     |
+------+------------+---------+--------+-------+---------+
11 rows in set (0.01 sec)
```

Experiences

- No big problems experienced except for initial installation issues
  - Reliable system over the last ≈ 9 months
- Initially on some nodes a restart of `gmond` (ganglia) and `pbs_mom` was necessary from time to time
  - Fixed after restart of the complete system
- By default, nodes are reinstalled after a hard reboot
  - Significantly increased down-time
  - We disabled this feature

Impressions

- **Pros:**
  - Open source, freely available
  - Functional, stable
  - Easy to use
- **Cons:**
  - Lack of documentation
  - Some inconsistencies (e.g. MPI)

- **Future (more challenging) plans:**
  - Patch kernel for PAPI support
  - New parallel filesystem

Some questions

- Comparison with e.g. xCAT, Bright Cluster Manager, etc?
- How to limit the amount of physical memory available for all/a group of processes?
- Can we expect a gain in efficiency by using a stripped-down Linux kernel even at this small scale?

Thank you very much for your kind attention