



User Session 3

Rocks-A-Palooza III



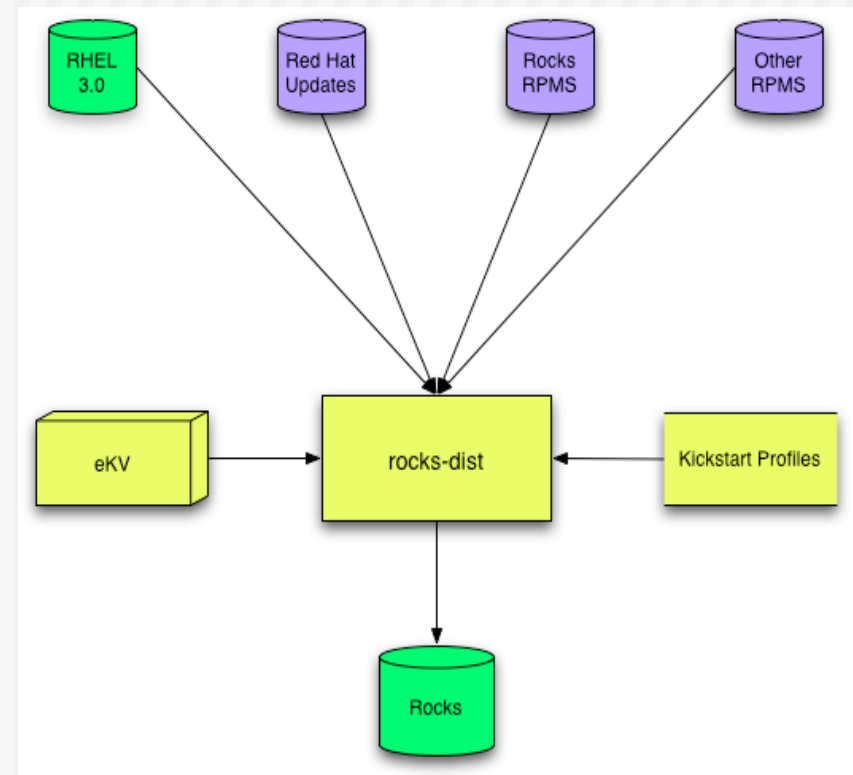
Building on Top of Rocks

Inheritance and Rolls



How Rocks is built

- ◆ Rocks-dist
 - Merges all RPMs
 - Red Hat
 - Rocks
 - Resolves versions
 - Creates Rocks
- ◆ Rocks distribution
 - Looks just like Red Hat
 - Cluster optimized Red Hat





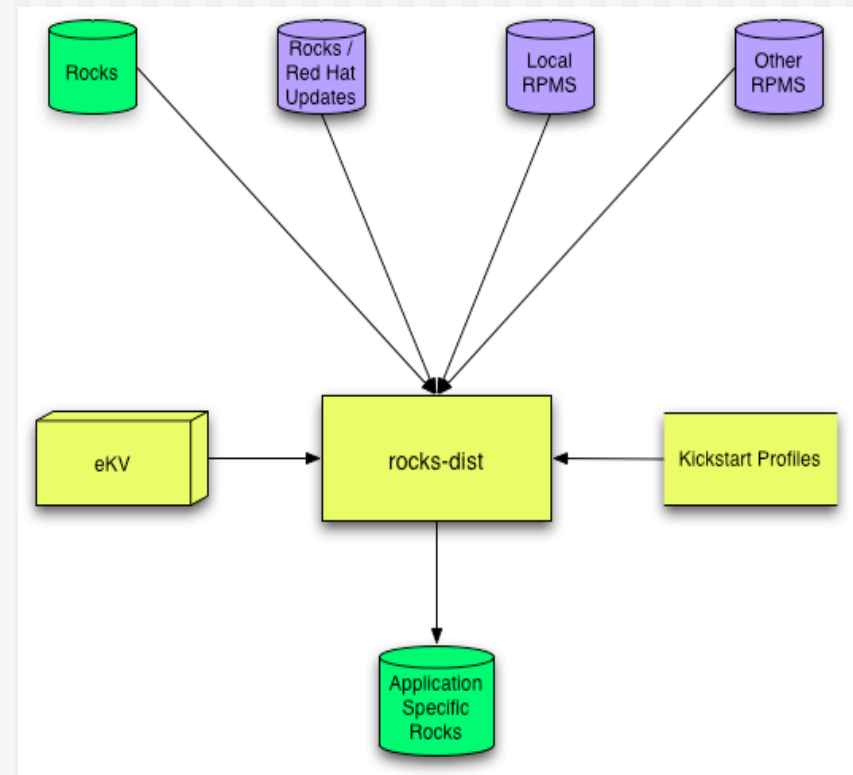
How You Create Your Own Rocks

◆ Rocks-dist

- ⇒ Merges all RPMs
 - Rocks
 - Yours
- ⇒ Resolves versions
- ⇒ Creates Rocks++

◆ Your distribution

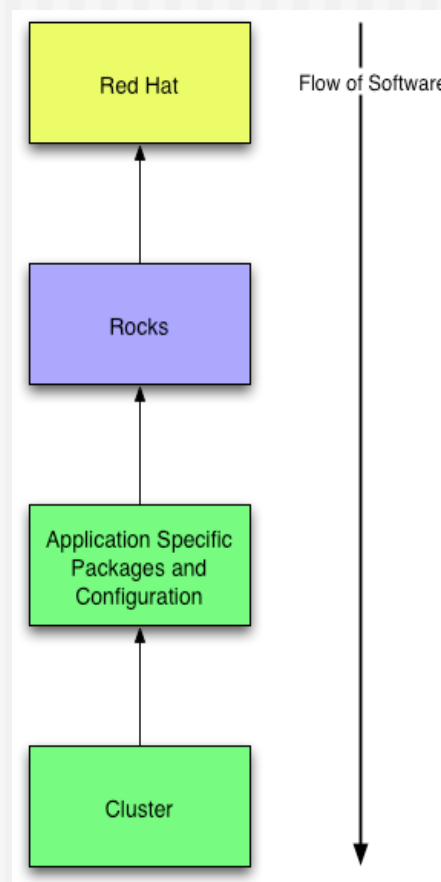
- ⇒ Looks just like Rocks
- ⇒ Application optimized Rocks





Extension Through Inheritance

- ◆ UCSD/SDSC Rocks
 - BIRN
 - GAMESS Portal
 - GEON
 - GriPhyN
 - Camera
 - Optiputer
- ◆ Commercial
 - Other stacks “based” on Rocks
- ◆ Can also override existing functionality
 - Rocks without NFS?
 - Rocks for the desktop?





Need Better Flexibility in Stack

◆ Issues

- Static Stack
 - Cannot redefine
 - Cannot extend
- Monolithic Stack
 - Cannot “opt out”
 - All or nothing solution
 - E.g. PBS not SGE

◆ What we need

- Dynamic Stack
- Component Based Stack
- User / Developer Extensible

PICK PACKAGES

- > COMBO #1: PREMIUM
- > COMBO #2: SPORT
- > COMBO #3: COLD WEATHER
- > NEXT STEP

MINI COOPER S

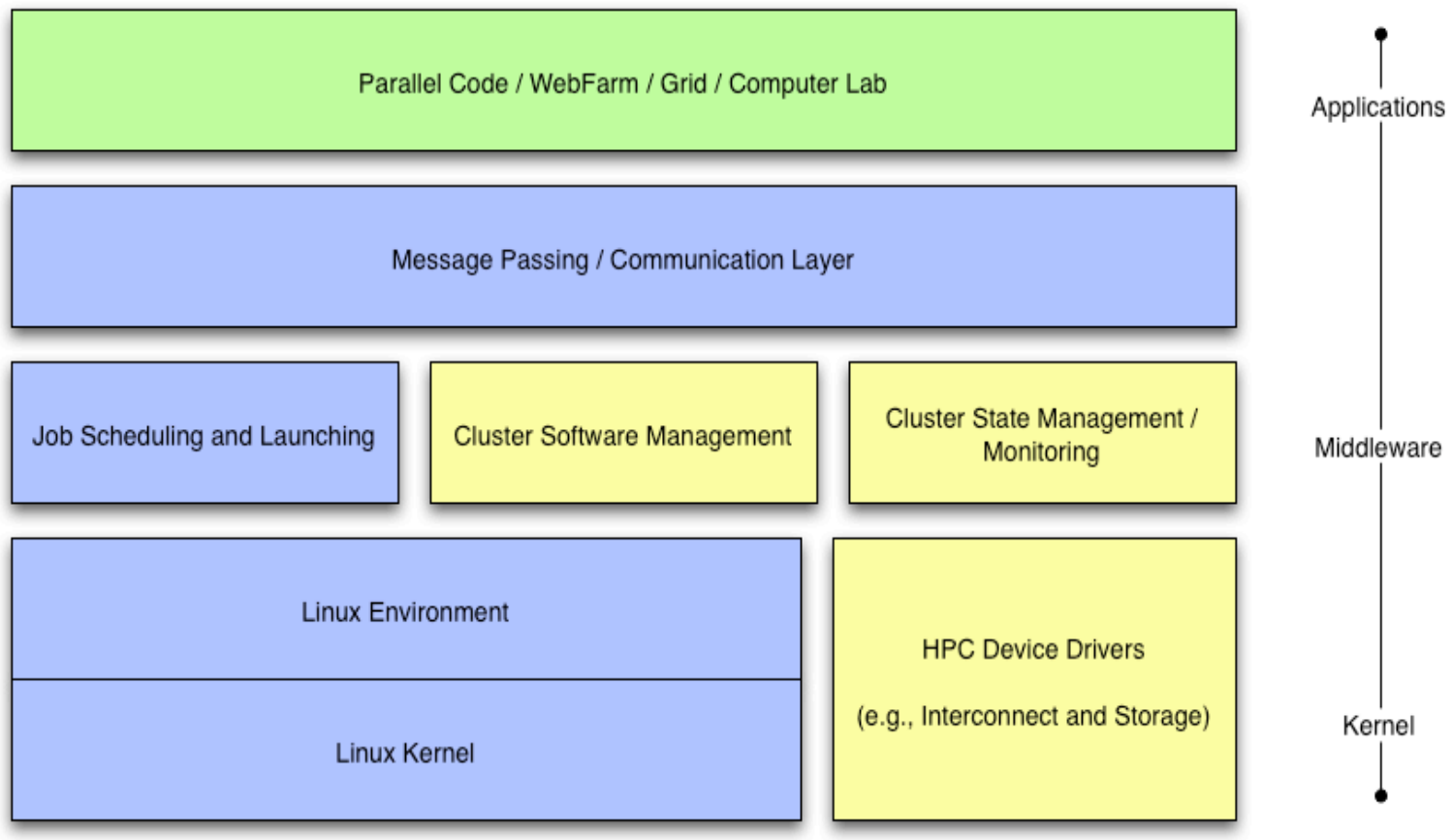
CLICK IMAGE TO ADD THE SPORT PACKAGE TO YOUR LIST.

THE SPORT PACKAGE WILL ADD:
Dynamic stability control (DSC), bonnet stripes, xenon headlamps with powerwashers, front fog lamps, 17-inch alloy 5-lite wheels with 205/45 R17 performance or all-season run-flat tires.

Sport Package (\$1350)



Rolls Break Apart Rocks



Rolls: Modifying a Standard System Installer to Support User-Customizable Cluster Frontend Appliances. Greg Bruno, Mason J. Katz, Federico D. Sacerdoti, and Phil M. Papadopoulos. *IEEE International Conference on Cluster Computing*, San Diego, California, Sep. 2004.



Rocks is What You Make it

◆ Motivation

- ⊖ “I’m concerned Rocks is becoming everything for everyone” - rocks mailing list
- ⊖ “Building a cluster should be like ordering a car. I want the sports package, but not the leather seats, ...” - z4 owning rocks developer
- ⊖ We need to let go of Rocks but hold onto the core
 - Recruit more external open-source developers
 - Only trust ourselves with fundamental architecture and implementation
- ⊖ We wanted to move the SGE but need to still support PBS

◆ Rolls

- ⊖ Optional configuration and software
- ⊖ Just another CD for installed (think application pack)
- ⊖ SGE and PBS are different Rolls
 - User chooses scheduler
 - PBS Roll supported by Norway
 - SGE Roll supported by Singapore (and us)
- ⊖ Rolls give us more flexibility and less work

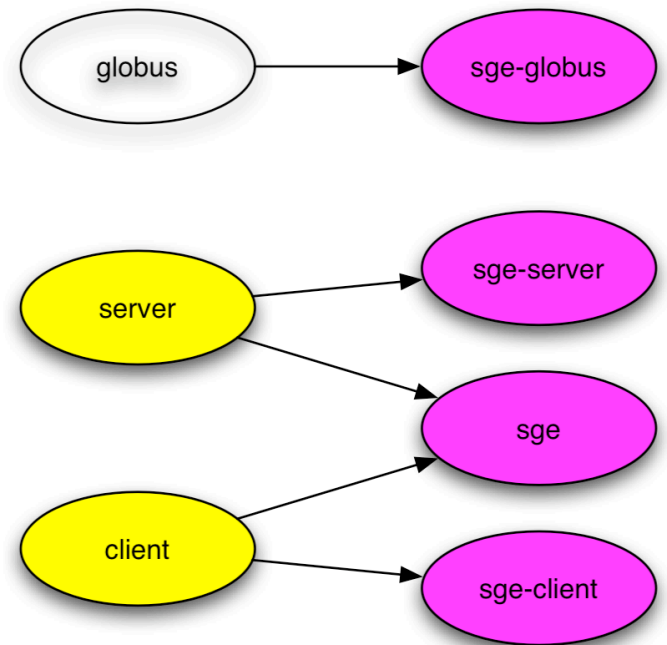
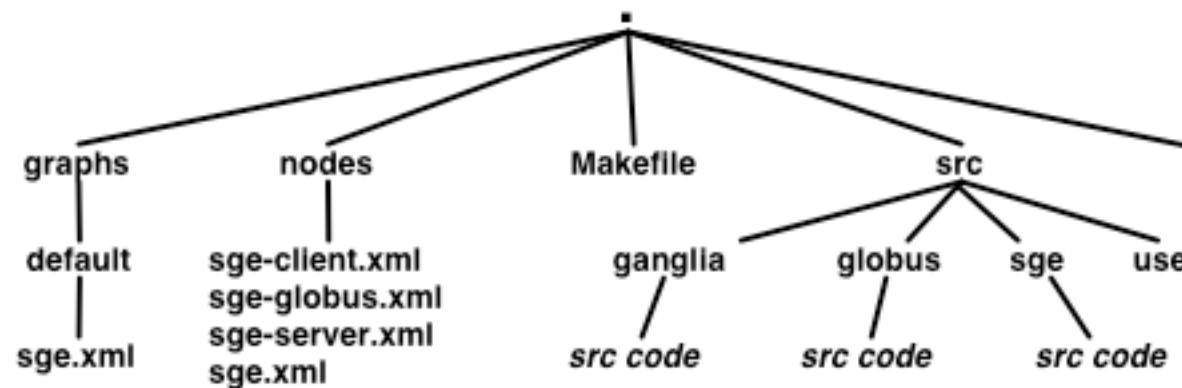
◆ Rocks is done

- ⊖ The core is basically stable and needs continued support
- ⊖ Rolls allow us to develop new ideas
- ⊖ Application Domain specific



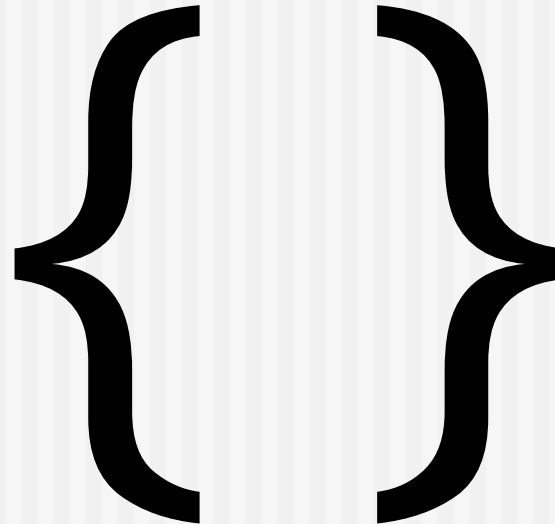
Rolls are sub-graphs

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



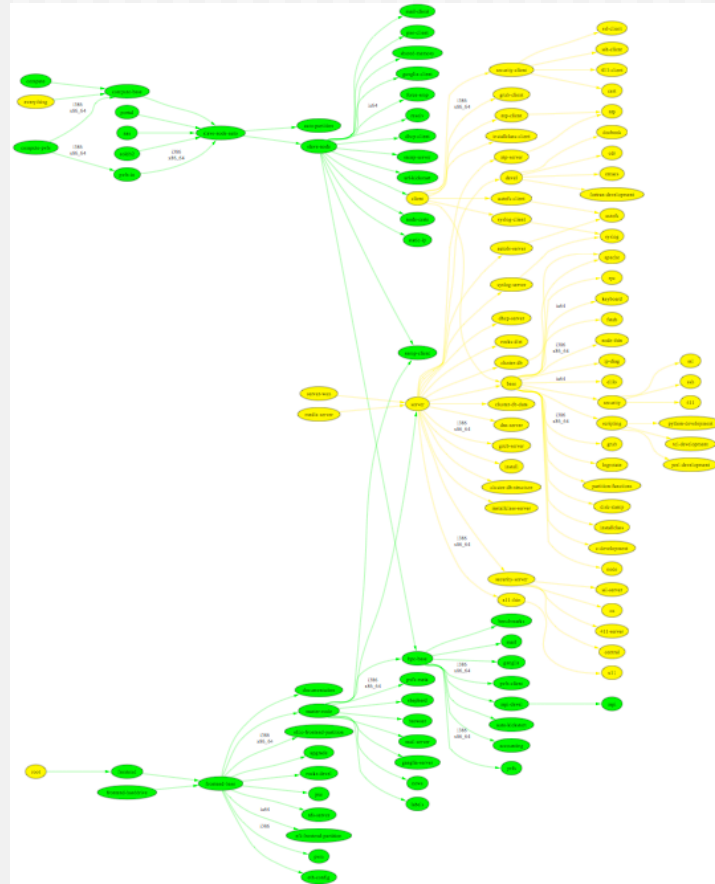


Starting from the empty set



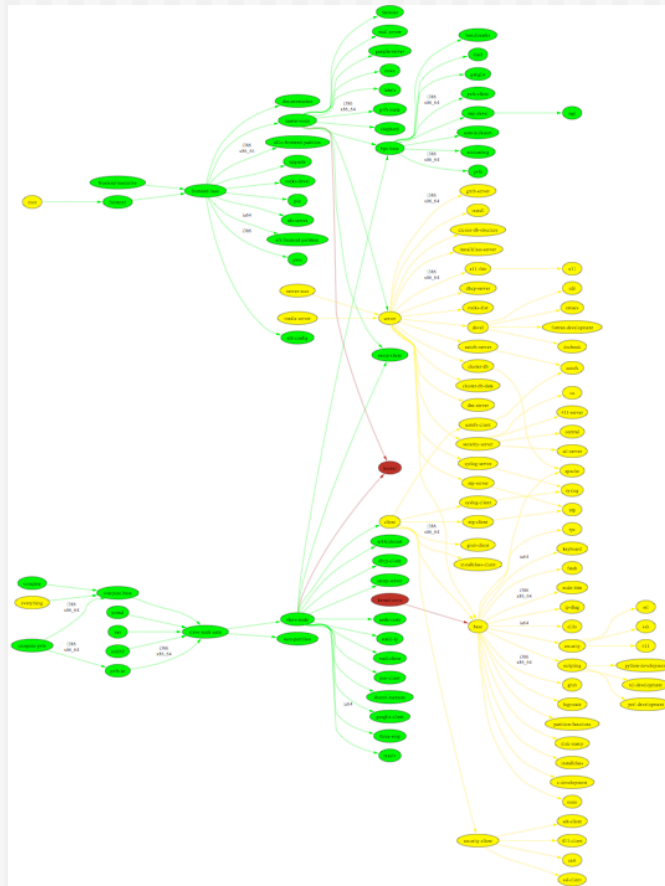


{ base, hpc }



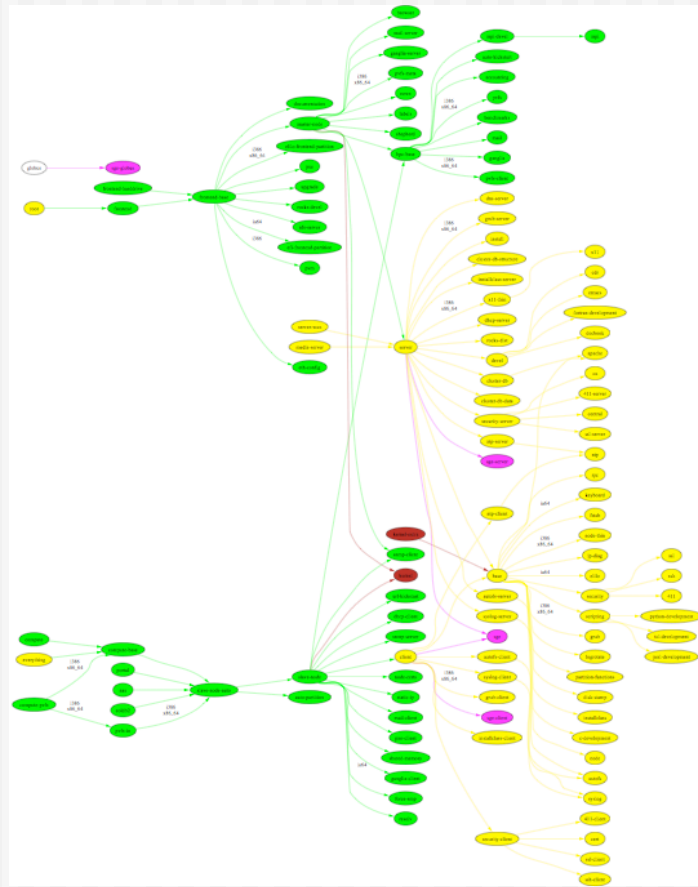


{ base, hpc, kernel }





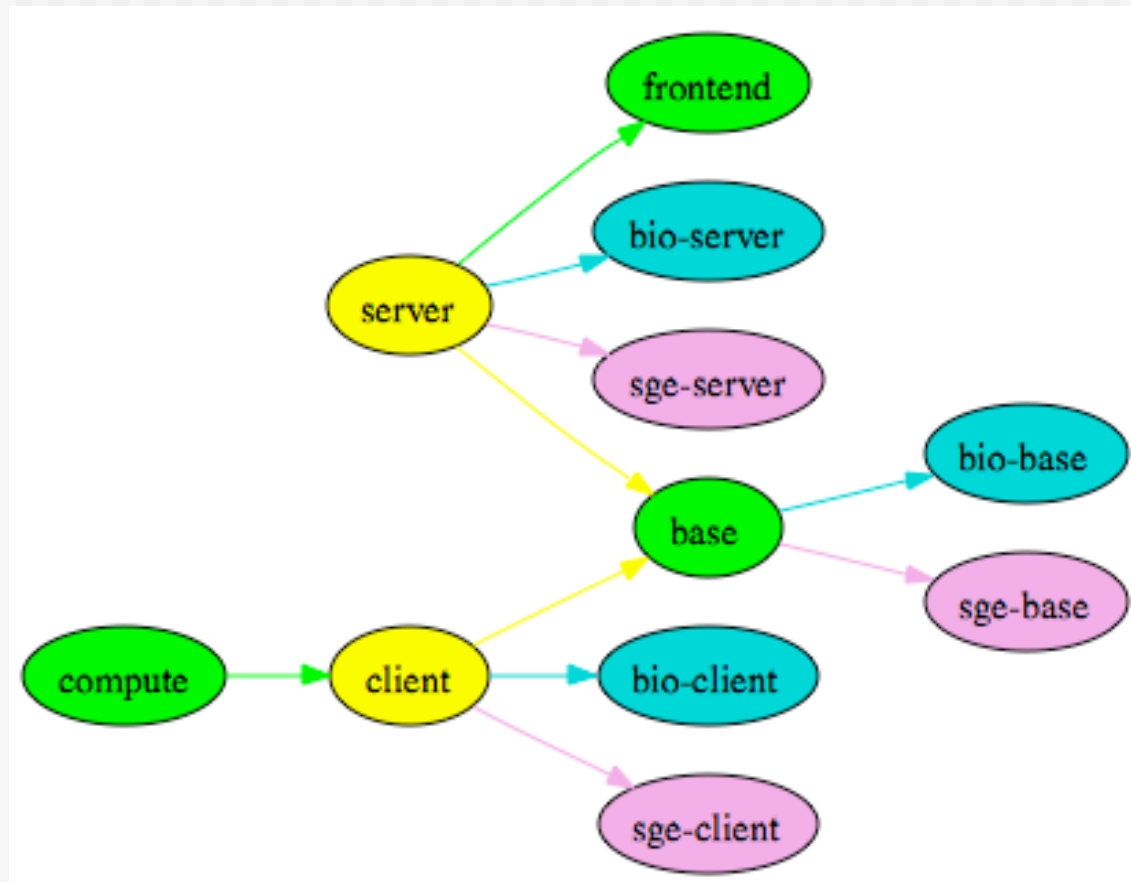
{ base, hpc, kernel, sge }





Simplified Example

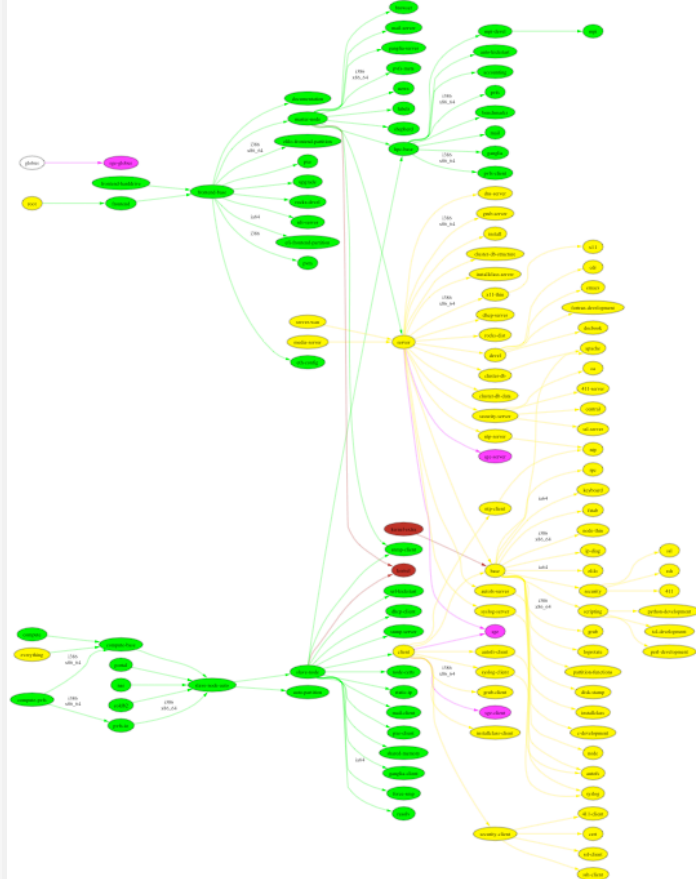
{base, hpc, sge, bio}



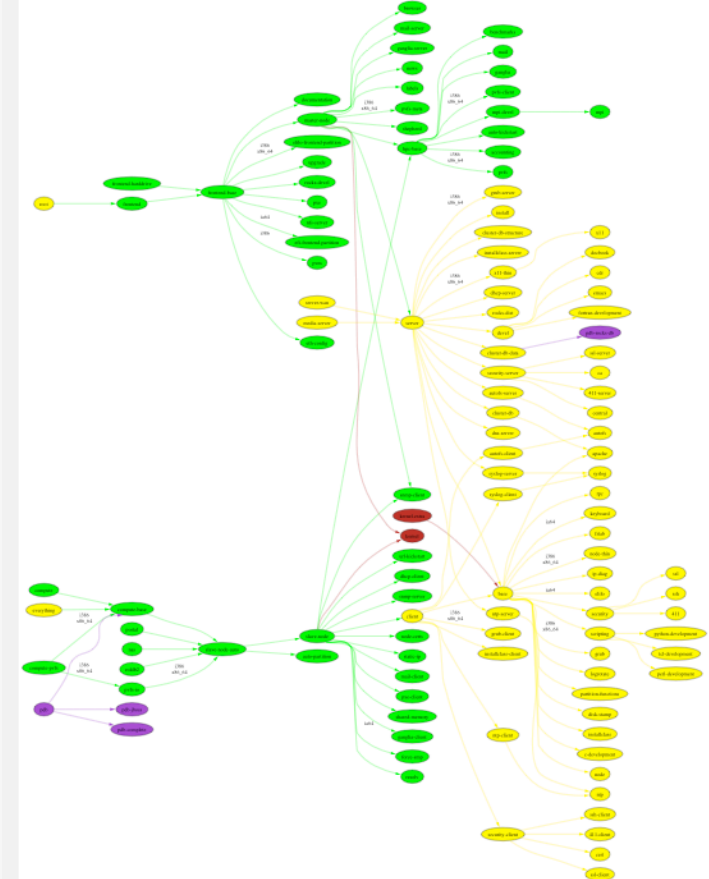


Two different Clusters

MPI Cluster:::{base, hpc, kernel, sge}



Protein Databank:::{base, hpc, kernel, pdb}





key point

Minor differences in the graph add up to large functional differences



Where are the Scaling Limits?

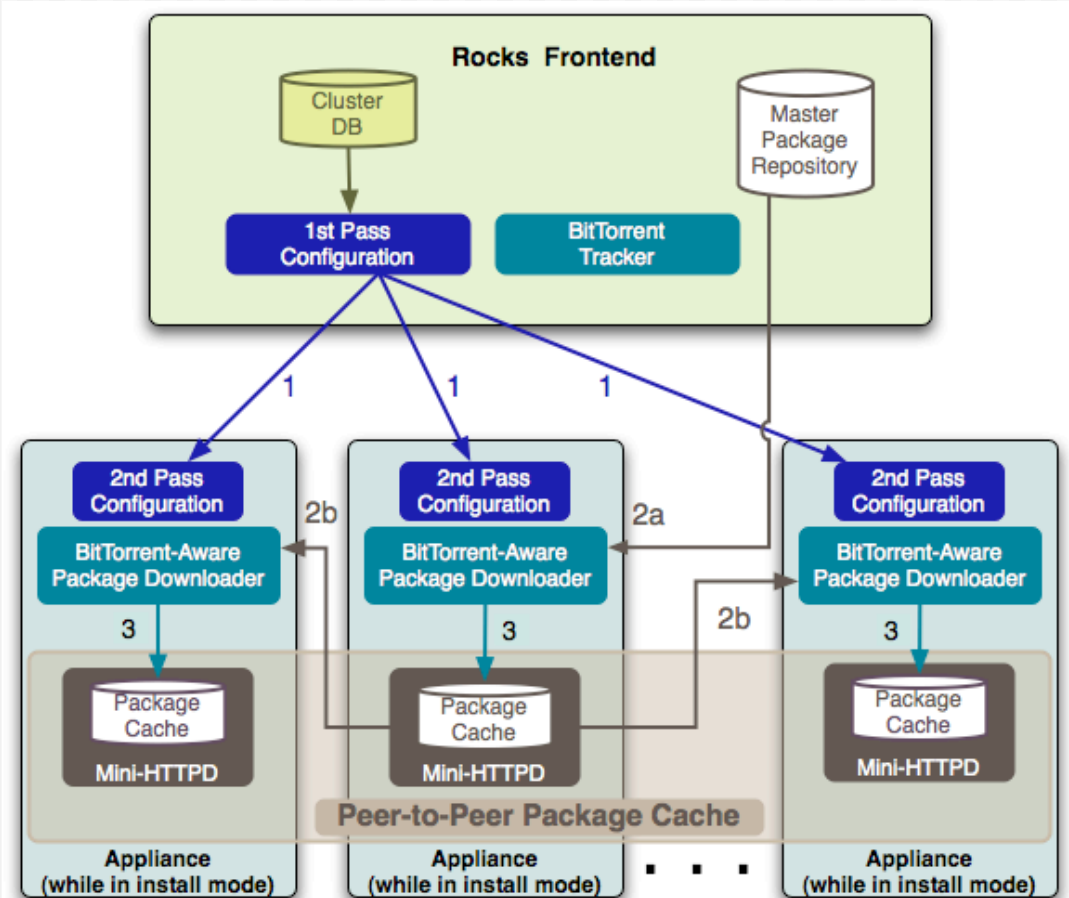
- ◆ Time for Kickstart Generation
 - ⤷ 3 - 4 s / host
 - ⤷ $O(n)$
- ◆ Time to Download Packages
- ◆ Rocks uses HTTP to transport Packages
- ◆ Linux easily serves HTTP files at
 - ⤷ 100MB/sec @ 1Gbit
 - ⤷ 12 MB/Sec@100Mbit
- ◆ Time = $\langle \#nodes \rangle * \langle \text{total MB packages} \rangle / \text{HTTP Speed}$
 - ⤷ Total Packages ~ 350MB

	128 Nodes	1024 Nodes
100 Mbit	3700s (1hr)	9 hours
1 Gbit	460s (8 min)	1 hour



Avalanche Installer

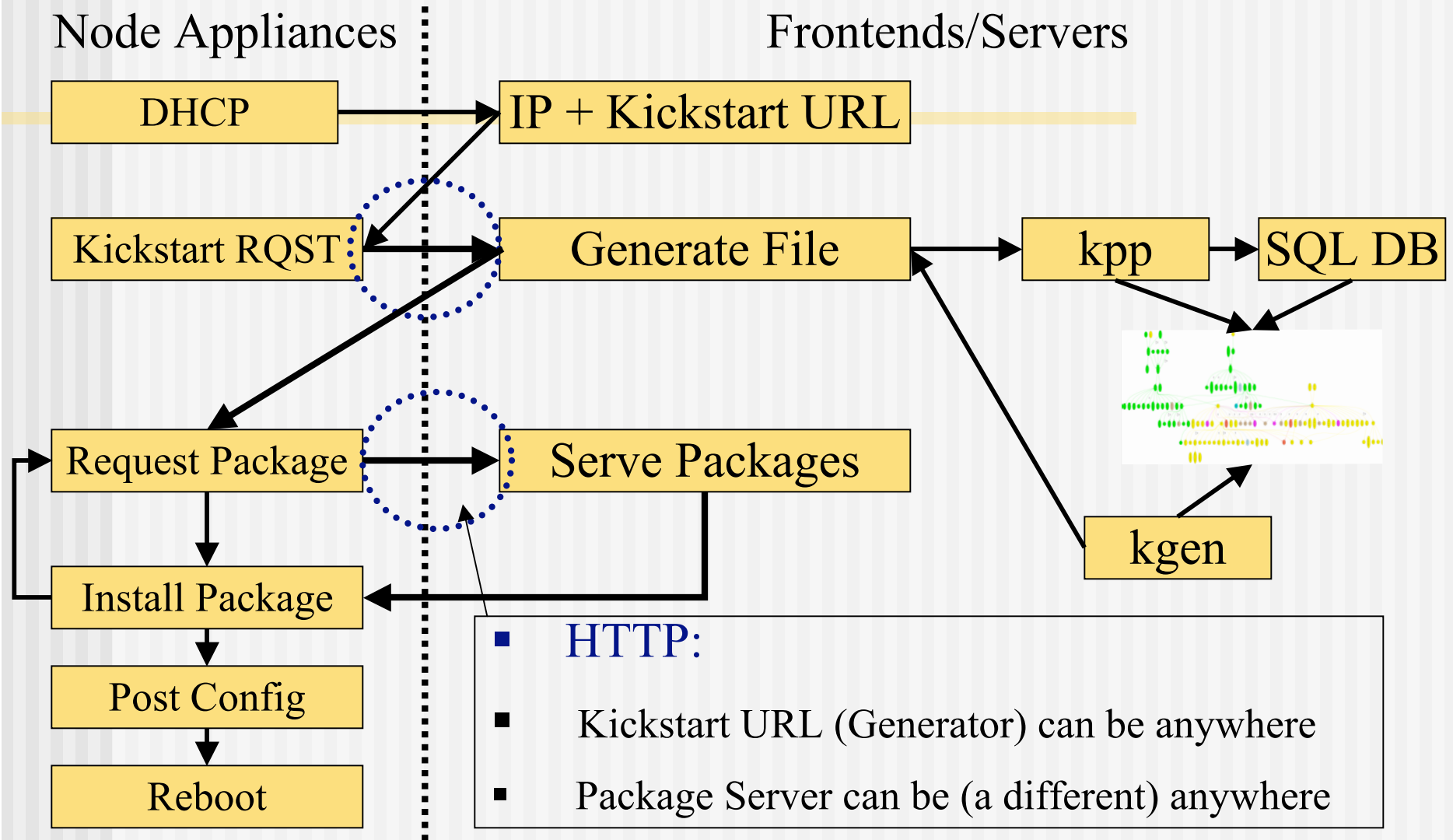
- ◆ Install nodes from a peer-to-peer package cache
- ◆ Takes advantage of switched networks to unload the frontend
- ◆ Kickstart generation is split between frontend and nodes
- ◆ Backoff mechanisms keep the frontend load under control
- ◆ Zero administration





Pre-Avalanche

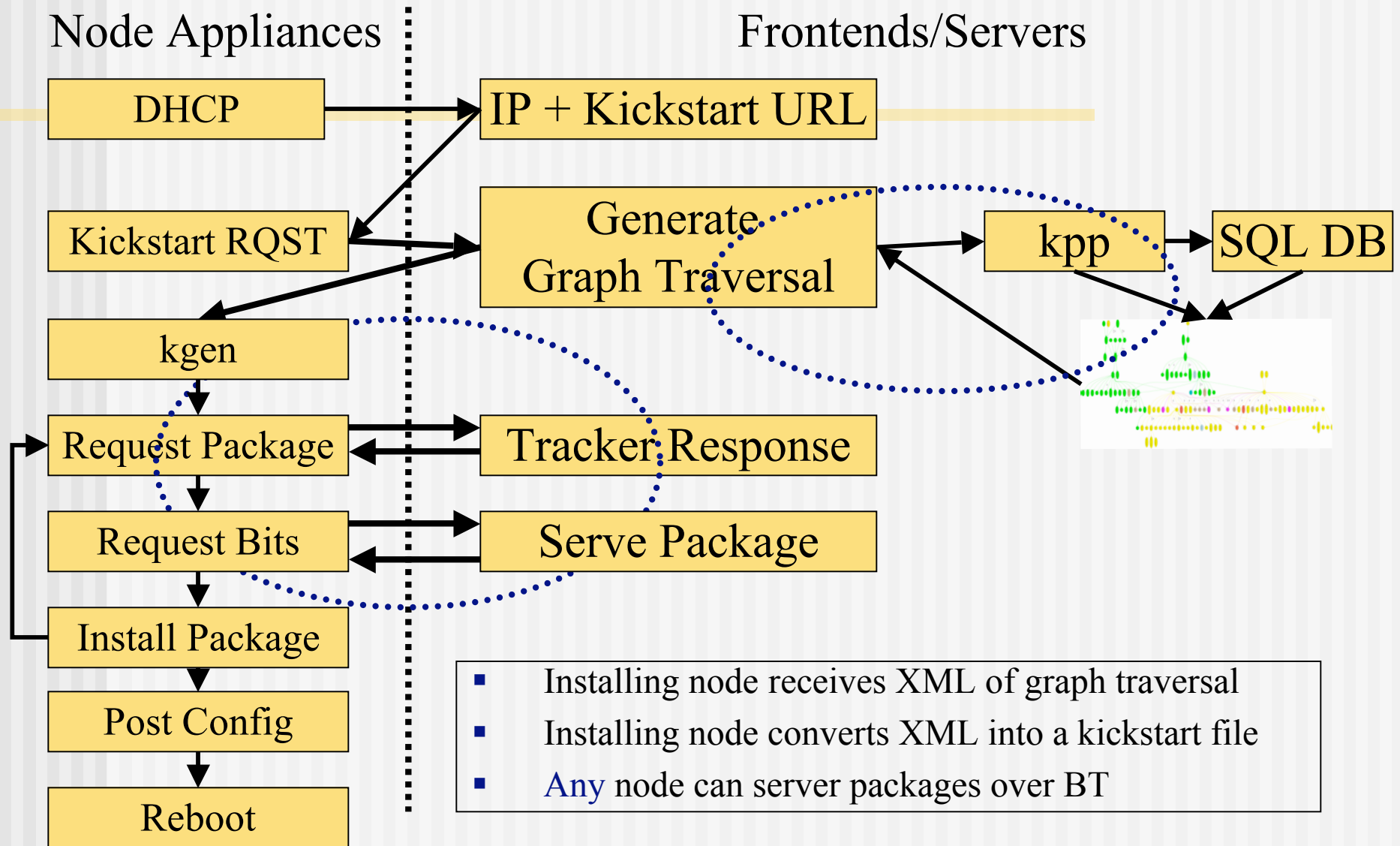
Space-Time and HTTP





Avalanche

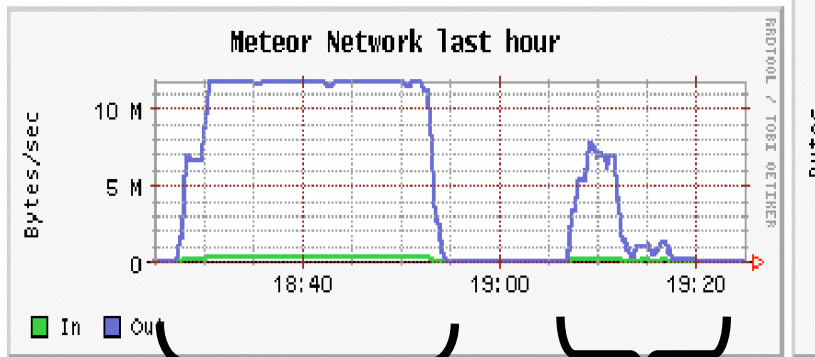
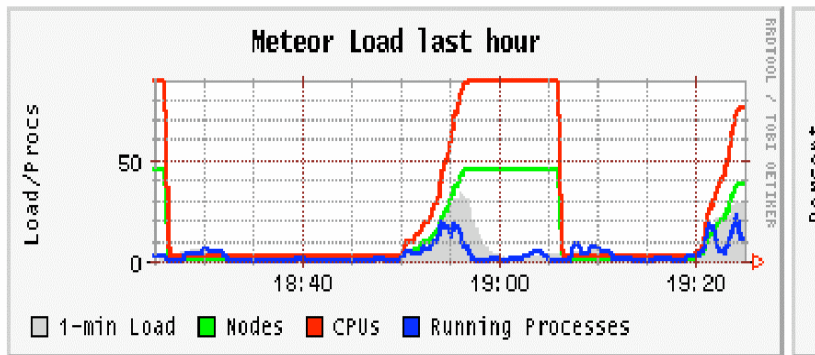
Space-Time and HTTP





A Glimpse at Performance

Overview of Meteor



HTTP-
Only

Avalanche

- ◆ 45 Nodes – 100 Mbit
 - ⊕ Old and Slow!
 - ⊕ 350MB (Slim Compute Node)
- ◆ Pre-avalanche:
 - ⊕ Estimate: 1600s
 - ⊕ Actual: 1700s
- ◆ Avalanche:
 - ⊕ Estimate: 900s
 - ⊕ Actual: 1000s
- ◆ Avalanche is significantly quicker – and reduces load on the frontend

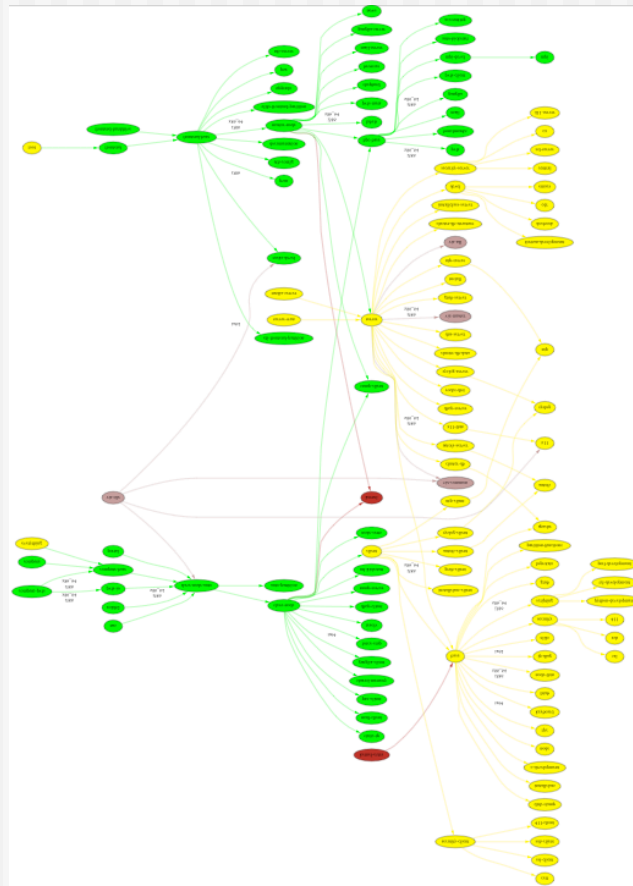


OptIPortal

viz roll



{ base, hpc, kernel, viz }





Early Work: NCSA

◆ LCD Cluster

- ➔ Custom framing
- ➔ One PC / tile
- ➔ Portable (luggable)
- ➔ SC 2001 Demo

◆ NCSA Software

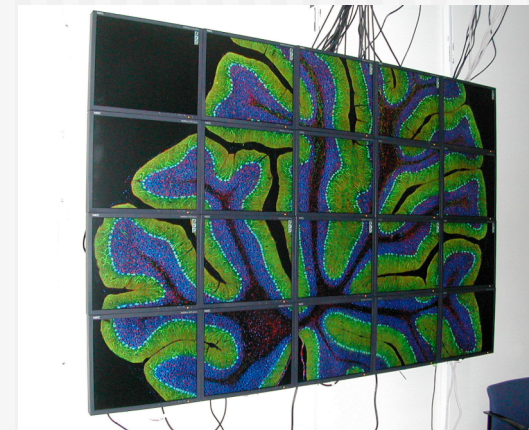
- ➔ Pixel Blaster
- ➔ Display Wall In-A-Box
 - OSCAR based
 - Never fully released





NCMIR

- ◆ Using Rocks
- ◆ Hand configured a visualization cluster
- ◆ “Administered the machine to the point of instability”
 - David Lee
- ◆ Automation is needed





COTS Vis: GeoWall

- ◆ LCD Clusters
 - One PC / tile
 - Gigabit Ethernet
 - Optional Stereo Glasses
 - Portable
 - Commercial Frame (Reason)
- ◆ Applications
 - Large remote sensing
 - Volume Rendering
 - Seismic Interpretation
 - Brain mapping (NCMIR)
- ◆ Electronic Visualization Lab
 - Jason Leigh (UIC)





OptIPortal (SAGE)





One Node per Display





OptIPortal





Nodes Behind the Wall





Genomic Map (cgview)





Building a Rocks Clusters



Young Frankenstein - Gene Wilder, Peter Boyle



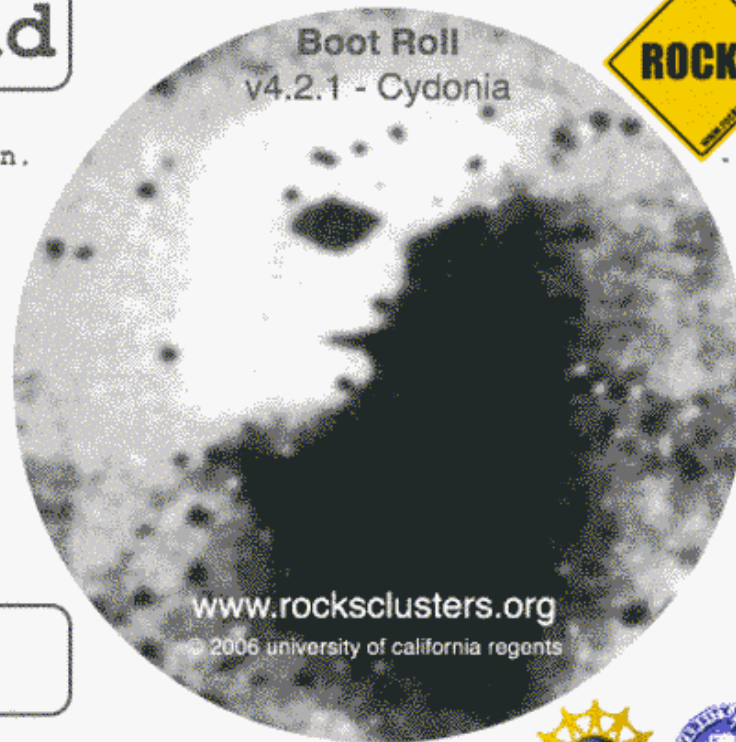
Frontend

```
# frontend  
For a new installation.
```

```
# frontend rescue  
To boot into rescue  
mode.
```

Client

```
do nothing (default)
```





Welcome to Rocks



Selected Rolls

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 roll server into the *Hostname of Roll Server* field and then click the *Download* button. This will query the roll server and all the rolls that the roll server has available will be displayed. Click the *selected* checkbox for each roll you will to install from the roll server.

When you have completed your roll selections, click the *Next* button to proceed to cluster input screens (e.g., IP address selection, root password setup, etc.).

Select Your Rolls

Local Rolls

CD/DVD-based Roll

Network-based Rolls

Hostname of Roll Server

Download

Next



Welcome to Rocks



Selected Rolls

Insert the Roll CD/DVD

No rolls have been selected.

Continue

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 roll server into the *Hostname of Roll Server* field and then click the *Download* button. This will query the roll server and all the rolls that the roll server has available will be displayed. Click the *selected* checkbox for each roll you will to install from the roll server.

When you have completed your roll selections, click the *Next* button to proceed to cluster input screens (e.g., IP address selection, root password setup, etc.).



Welcome to Rocks



Selected Rolls

Selected	Roll Name	Version	Arch
<input type="checkbox"/>	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 roll server into the *Hostname of Roll Server* field and then click the *Download* button. This will query the roll server and all the rolls that the roll server has available will be displayed. Click the *selected* checkbox for each roll you will to install from the roll server.

When you have completed your roll selections, click the *Next* button to proceed to cluster input screens (e.g., IP address selection, root password setup, etc.).



Welcome to Rocks



Selected Rolls

Roll Name	Version	Arch
kernel	4.2	x86_64

Select Your Rolls

Local Rolls

CD/DVD-based Roll

Network-based Rolls

Hostname of Roll Server

Download

Next



Welcome to Rocks



Selected Rolls

Roll Name	Version	Arch
base	4.2	x86_64
hpc	4.2	x86_64
kernel	4.2	x86_64
os	4.2	x86_64
web-server	4.2	x86_64

Select Your Rolls

Local Rolls

CD/DVD-based Roll

Network-based Rolls

Hostname of Roll Server

Download

Next



Welcome to Rocks



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	<input type="text" value="cluster.hpc.org"/>
Cluster Name	<input type="text" value="Our Cluster"/>
Certificate Organization	<input type="text" value="SDSC"/>
Certificate Locality	<input type="text" value="San Diego"/>
Certificate State	<input type="text" value="California"/>
Certificate Country	<input type="text" value="US"/>
Contact	<input type="text" value="admin@place.org"/>
URL	<input type="text" value="http://www.place.org/"/>
Latitude/Longitude	<input type="text" value="N32.87 W117.22"/>

Back

Next



Welcome to Rocks



Help

IP address:

Enter the IP address for eth0. This is the interface that connects the frontend to the compute nodes.

Netmask:

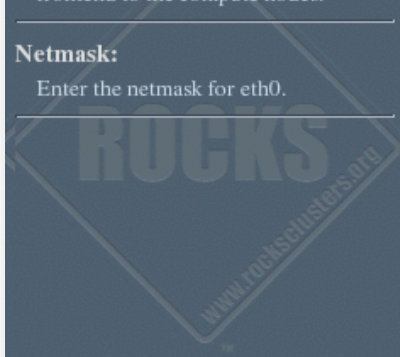
Enter the netmask for eth0.

Ethernet Configuration for eth0

IP address	<input type="text" value="10.1.1.1"/>
Netmask	<input type="text" value="255.0.0.0"/>

[Back](#)

[Next](#)





Welcome to Rocks



Help

IP address:

Enter the IP address for eth1. This is the interface that connects the frontend to the outside network.

Netmask:

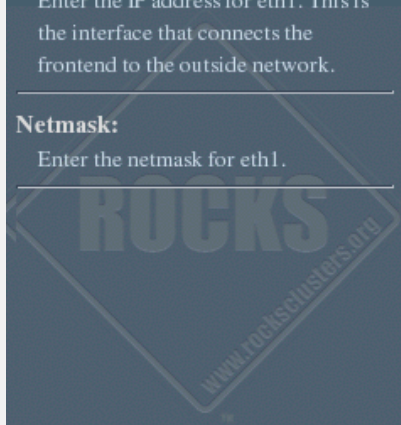
Enter the netmask for eth1.

Ethernet Configuration for eth1

IP address
Netmask

[Back](#)

[Next](#)





Welcome to Rocks



Help

Gateway:

The IP address of your public gateway.

DNS Servers:

Supply a comma separated list of your DNS servers.

Miscellaneous Network Settings

Gateway
DNS Servers

[Back](#)

[Next](#)





Welcome to Rocks



Help

Password:

The root password for your cluster.



Root Password

Password
Confirm

Back

Next



Welcome to Rocks



Help

Time Zone:

Select a timezone for your cluster.

NTP Server:

Input a Network Time Protocol (NTP) server that will keep the clock on your frontend in sync.

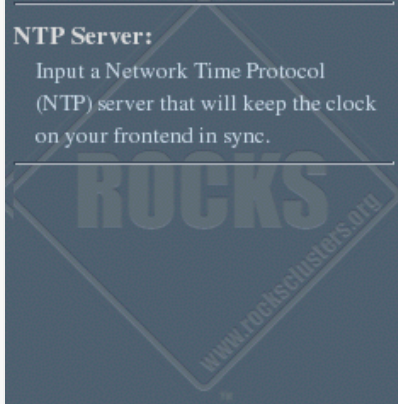
Time Configuration

Time Zone

NTP Server

[Back](#)

[Next](#)





Welcome to Rocks



Help

Auto Partitioning:

The first disk on this machine will be partitioned in the default manner. See the documentation at www.rocksclusters.org for details on the default partitioning scheme.

Manual Partitioning:

The user will be required to set all partitioning information for this machine. A subsequent installation screen will allow you to enter your partitioning information.

Disk Partitioning

Auto Partitioning

Manual Partitioning


Back

Next



Manual Partition

not for new users

www.rocksclusters.org 

Disk Setup

Choose where you would like Rocks to be installed.

If you do not know how to partition your system or if you need help with using the manual partitioning tools, refer to the product documentation.

If you used automatic partitioning, you can either accept the current partition settings (click **Next**), or modify the setup using the manual partitioning tool.

If you are manually partitioning your system, you can see your current hard drive(s) and partitions displayed below. Use the partitioning tool to add, edit,

Drive `/dev/hda` (76317 MB) (Model: WDC WD800BB-22JHC0)

hda1	hda2	hda5
8001 MB	4000	63318 MB

Device	Mount Point/RAID/Volume	Type	Format	Size (MB)	Start	End
▼ Hard Drives						
▼ /dev/hda						
/dev/hda1	/	ext3	✓	8001	1	1020
/dev/hda2	/var	ext3	✓	4001	1021	1530
/dev/hda3		swap		996	1531	1657
▼ /dev/hda4						
/dev/hda5	/export	ext3		63319	1658	9729

Hide RAID device/LVM Volume Group members



www.rockclusters.org



Installing Packages

We have gathered all the information needed to install Rocks on the system. It may take a while to install everything, depending on how many packages need to be installed.

Install Roll

Put Roll disk 'kernel - Disk 1' in the drive

OK

sters.org

Hide Help

Release Notes

Back

Next



www.rockclusters.org



Installing Packages

We have gathered all the information needed to install Rocks on the system. It may take a while to install everything, depending on how many packages need to be installed.

Welcome to CentOS 4 !

Thank you for installing CentOS 4.

CentOS is an Enterprise-class Linux Distribution derived from sources freely provided to the public by a prominent North American Enterprise Linux vendor. CentOS conforms fully with the upstream vendors redistribution policy and aims to be 100% binary compatible. (CentOS mainly changes packages to remove upstream vendor branding and artwork.)

More Info: <http://www.centos.org/>



Installing redhat-logos-1.1.26-1.centos4.1.noarch (8 MB)
Red Hat-related icons and pictures.



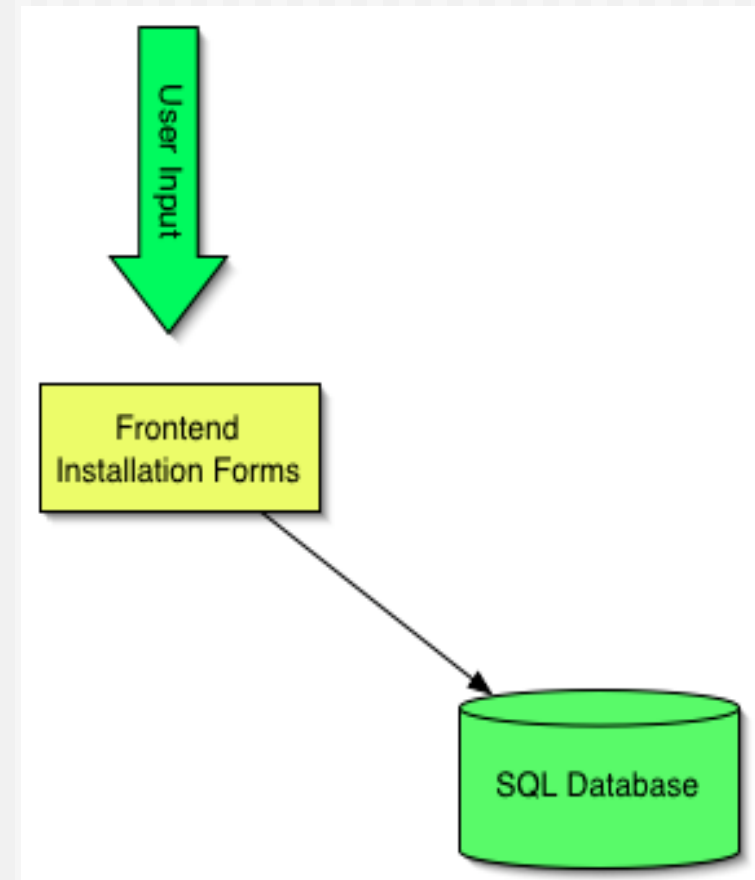
key point

First time cluster builders should stay as close as possible to the defaults



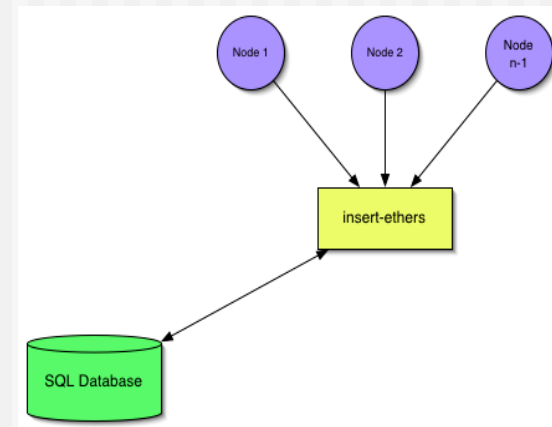
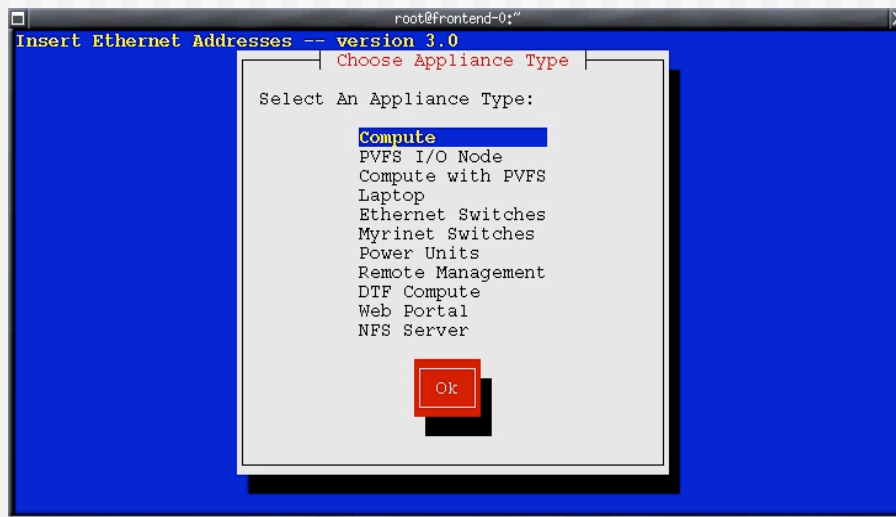
Interactive Screen

- ◆ Fill out the screens we just talked about
- ◆ Use the provided network information
- ◆ Choose your own password
- ◆ All information goes into the cluster database





Add Compute Node with Insert-ethers

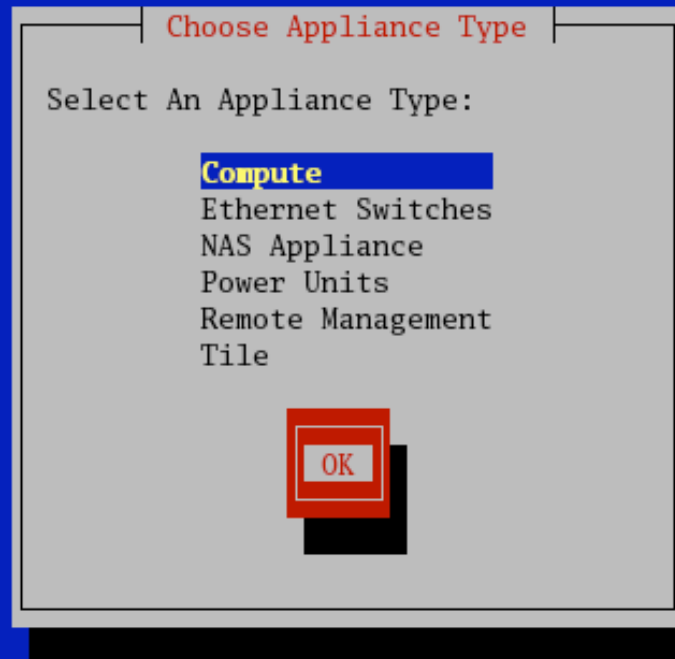


- ◆ Collect the Ethernet MAC address of cluster nodes
- ◆ Only done once, during integration
- ◆ Populates cluster database



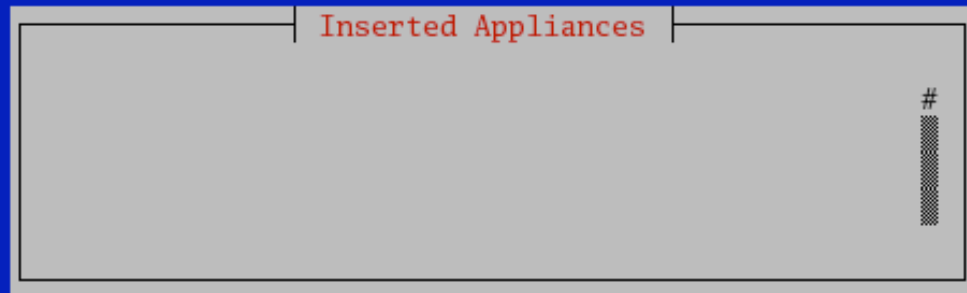
Adding Compute Nodes

Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network





Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network



Press <F10> to quit, press <F11> to force quit



Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network

Inserted Appliances
Discovered New Appliance

Discovered a new appliance with MAC (00:13:72:ba:c8:df)

Press <F10> to quit, press <F11> to force quit



Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network

Inserted Appliances			
00:13:72:ba:c8:df	compute-0-0	()	#

Press <F10> to quit, press <F11> to force quit



Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network

Inserted Appliances			
00:13:72:ba:c8:df	compute-0-0	(*)	#

Press <F10> to quit, press <F11> to force quit



Open Lab



[adult swim]



rockstar.rockclusters.org

- ◆ ssh access (no telnet)
- ◆ Account
 - ⇒ Username: rap-01, rap-02, ...
 - ⇒ Password: amdocks
- ◆ User level access only





Simple MPI Program

```
1: #include <stdio.h>
2: #include "mpi.h"
3:
4: int
5: main(int argc, char *argv[])
6: {
7:     int    numprocs;
8:     int    myid;
9:     int    namelen;
10:    char    processor_name[MPI_MAX_PROCESSOR_NAME];
11:
12:    MPI_Init(&argc, &argv);
13:
14:    MPI_Comm_size(MPI_COMM_WORLD, &numprocs);
15:    MPI_Comm_rank(MPI_COMM_WORLD, &myid);
16:    MPI_Get_processor_name(processor_name, &namelen);
17:
18:    fprintf(stderr, "Process %d on %s\n", myid, processor_name);
19:
20:    MPI_Barrier(MPI_COMM_WORLD);
21:
22:    sleep(120);
23:
24:    MPI_Finalize();
25: }
```



Simple MPI/SGE Submit Script

```
#!/bin/bash
#
#$ -cwd
#$ -j y
#$ -S /bin/bash

MPI_DIR=/opt/mpich/gnu

$MPI_DIR/bin/mpirun -np $NSLOTS -machinefile $TMPDIR/machines hello
```



Compile / Run

◆ Compile

⇒ `/opt/mpich/gnu/bin/mpicc -o hello hello.c`

◆ Run

⇒ `qsub -pe mpich 2 hello.sh`

◆ Monitor

⇒ `qstat`



Example Run

```
mjk@rocks-52:~ — bash (tty1)

[mjk@rocks-52 mjk]$ /opt/mpich/gnu/bin/mpicc -o hello hello.c
[mjk@rocks-52 mjk]$ qsub -pe mpich 2 hello.sh
your job 4773 ("hello.sh") has been submitted
[mjk@rocks-52 mjk]$ qstat
job-ID prior name      user      state submit/start at    queue      master  ja-task-ID
-----
  4773    0 hello.sh  mjk      qw    05/17/2005 15:23:30
[mjk@rocks-52 mjk]$ qstat
job-ID prior name      user      state submit/start at    queue      master  ja-task-ID
-----
  4773    0 hello.sh  mjk      r     05/17/2005 15:23:41 compute-0- SLAVE
  4773    0 hello.sh  mjk      r     05/17/2005 15:23:41 compute-0- MASTER
           0 hello.sh  mjk      r     05/17/2005 15:23:41 compute-0- SLAVE
[mjk@rocks-52 mjk]$ ls -l hello.sh.*
-rw-r--r--  1 mjk      mjk           62 May 17 15:23 hello.sh.o4773
-rw-r--r--  1 mjk      mjk          106 May 17 15:23 hello.sh.po4773
[mjk@rocks-52 mjk]$ cat hello.sh.o4773
Process 0 on rocks-62.sdsc.edu
Process 1 on rocks-62.sdsc.edu
[mjk@rocks-52 mjk]$ qstat
[mjk@rocks-52 mjk]$ hostname
rocks-52.sdsc.edu
[mjk@rocks-52 mjk]$
```




HPL.dat

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
1000 Ns
1            # of NBs
64 NBs
1            # of process grids (P x Q)
1 Ps
2 Qs
16.0        threshold
3            # of panel fact
0 1 2       PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
8            NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
2            RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
1            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
1            DEPTHS (>=0)
2            SWAP (0=bin-exch,1=long,2=mix)
80          swapping threshold
0           L1 in (0=transposed,1=no-transposed) form
0           U in (0=transposed,1=no-transposed) form
1           Equilibration (0=no,1=yes)
8           memory alignment in double (> 0)
```



Example HPL Run

```
mjk@rocks-52:~ -- bash (tty1)
[mjk@rocks-52 mjk]$ cp /var/www/html/rocks-documentation/3.3.0/examples/HPL.dat .
[mjk@rocks-52 mjk]$ qsub -pe mpich 2 hpl.sh
your job 4776 ("hpl.sh") has been submitted
[mjk@rocks-52 mjk]$ qstat
job-ID prior name      user      state submit/start at   queue      master  ja-task-ID
-----
  4776   0 hpl.sh    mjk       qw    05/17/2005 18:11:43
[mjk@rocks-52 mjk]$ qstat
[mjk@rocks-52 mjk]$ cat hpl.sh.o4776
=====
HPLinpack 1.0 -- High-Performance Linpack benchmark -- September 27, 2000
Written by A. Petitet and R. Clint Whaley, Innovative Computing Labs., UTK
=====

An explanation of the input/output parameters follows:
T/V   : Wall time / encoded variant.
N     : The order of the coefficient matrix A.
NB    : The partitioning blocking factor.
P     : The number of process rows.
Q     : The number of process columns.
Time  : Time in seconds to solve the linear system.
Gflops : Rate of execution for solving the linear system.

The following parameter values will be used:

N      : 1000
NB     : 64
P      : 1
Q      : 2
PFACT  : Left   Crout   Right
NBMIN  : 8
NDIV   : 2
```



Linpack Scaling

- ◆ Then edit 'HPL.dat' and change:
 - 1 Ps
 - ⇒ To:
 - 2 Ps
 - ⇒ The number of processors Linpack uses is $P * Q$
- ◆ To make Linpack use more memory (and increase performance), edit 'HPL.dat' and change:
 - 1000 Ns
 - ⇒ To:
 - 4000 Ns
 - ⇒ Linpack operates on an $N * N$ matrix
- ◆ Submit the (larger) job:
 - ⇒ `qsub qsub-test.sh`



Others Tasks

◆ Globus

- See grid roll usersguide
- Setup user keys
- `globus-job-run localhost /bin/hostname`
- `globus-job-run localhost/jobmanager-sge`

◆ Adding RPMs to nodes

- See usersguide for graph instructions

◆ Rebuild with Central/CDROM