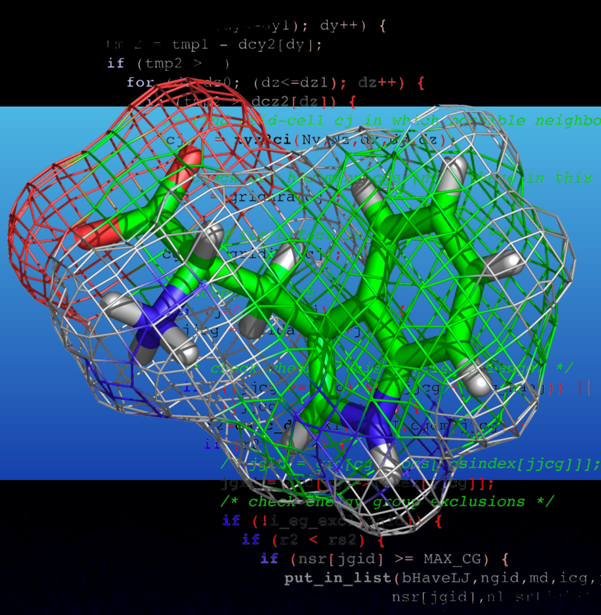
Building Clusters for Gromacs and other HPC applications

Erik Lindahl

lindahl@cbr.su.se

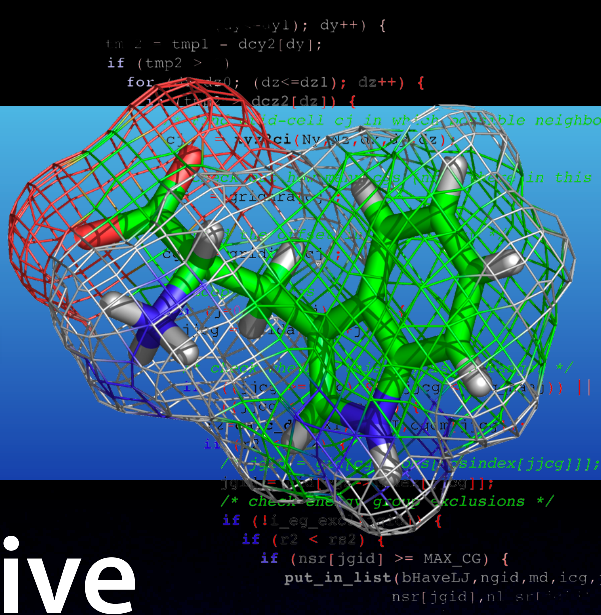
CBR

Outline: Clusters


dy++) {
= tmp1 - dcy2[dy];
if (tmp2 >)
for dz0; (dz<=dz1); dz++) {
dcz2[dz]) {
/* check energy group exclusions */
if (!i_eg_exc
if (r2 < rs2) {
if (nsr[jgid] >= MAX_CG) {
put_in_list(bHaveLJ,ngid,md,icg,
nsr[jgid],nl

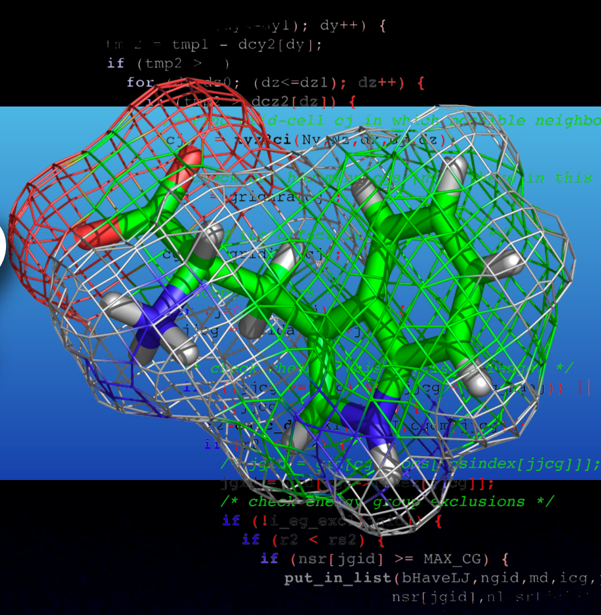
- Clusters vs. small networks of machines
- Why do YOU need a cluster?
- Computer hardware
- Network interconnects
- Storage
- Administration software, queue systems
- Cost vs. performance
- Installation, setup, maintenance

Cluster justifications



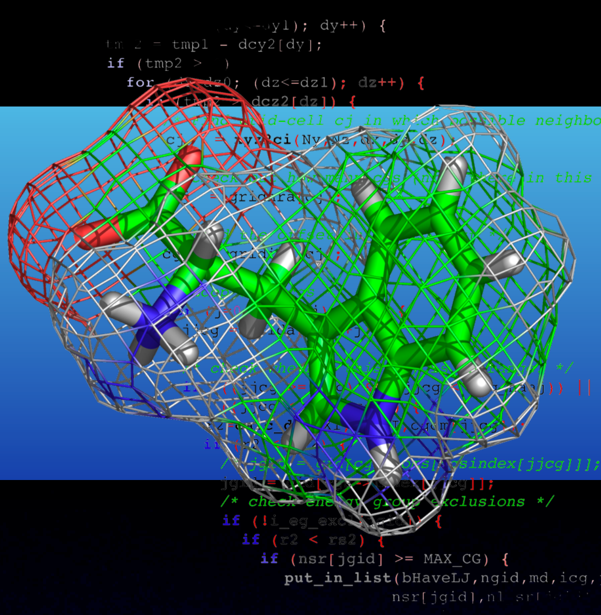
- Performance is unfortunately addictive
- If you don't already, you will soon wish you had a faster computer for simulations
- Dual-dual (4x) core workstations are nice!
- Free energy calculations can use 20-40 independent simulations i parallel
- With several workstations, it can still be a pain to start and check all simulations
- Parallel simulations require dedicated boxes

What is YOUR goal?



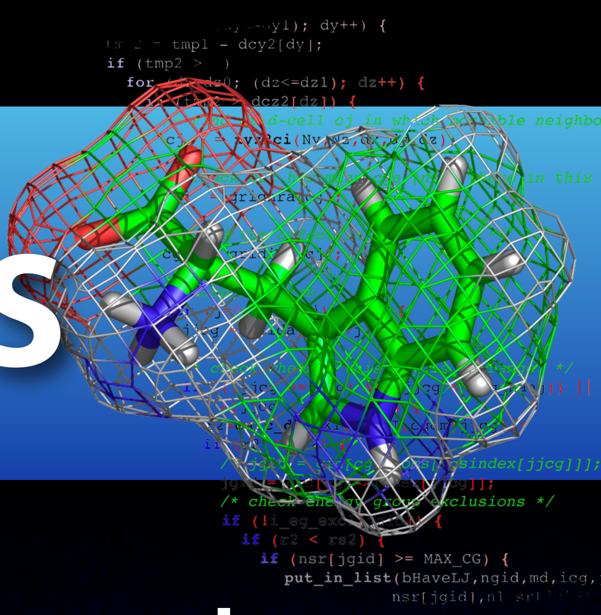
- Running weakly coupled simulations like replica exchange? **Cheap x86 cluster**
- Running 1000’s of independent short simulations to improve sampling, e.g. for free energy calculations? **Cheap x86 cluster**
- Running in parallel over 10-100 processors to create single microsecond trajectories of large systems? **Expensive machine with good interconnect**

Cluster hardware



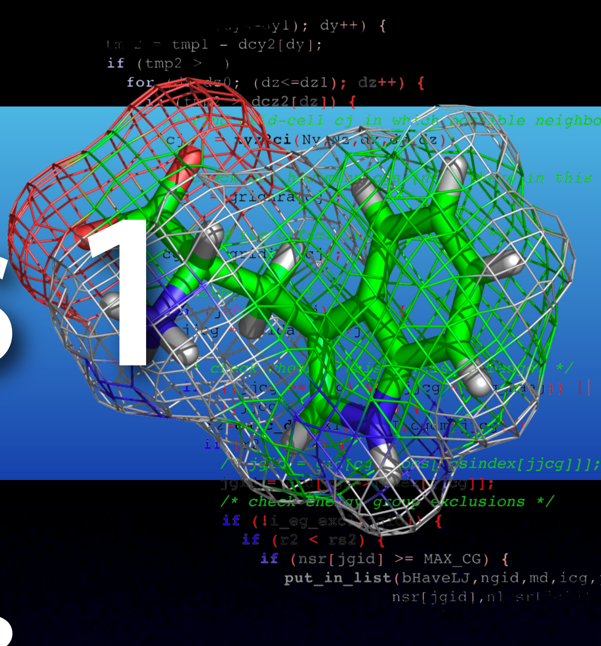
- Gromacs has handtuned assembly kernels for x86 (Intel, AMD) processors
- PowerPC,Sun,BlueGene not competitive on performance/$ (for Gromacs, at least)
- Gromacs is mostly floating-point (CPU) bound and only uses limited memory
- 64-bit is obvious today (~10% faster)
- Maximize the number of CPU cores per node, save on the memory

Current alternatives



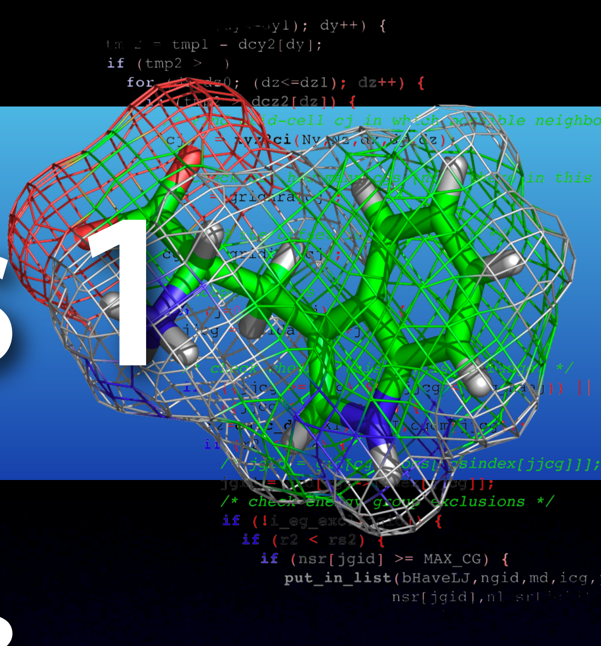
- AMD: Dual-core Opterons perform fine, and have very good memory bandwidth. However, SSE instructions take 2 cycles
- Intel: New (Core2) CPUs are amazing - all SSE instructions finish in 1 cycle!
  - Woodcrest (dual core) is currently the highest-performing Gromacs CPU
  - Clovertown ('quad' core, really 2x dual) are slightly worse *per core*, but better total throughput performance per cost
- True quad cores in 2H 2007 will be amazing!

Other requirements 1



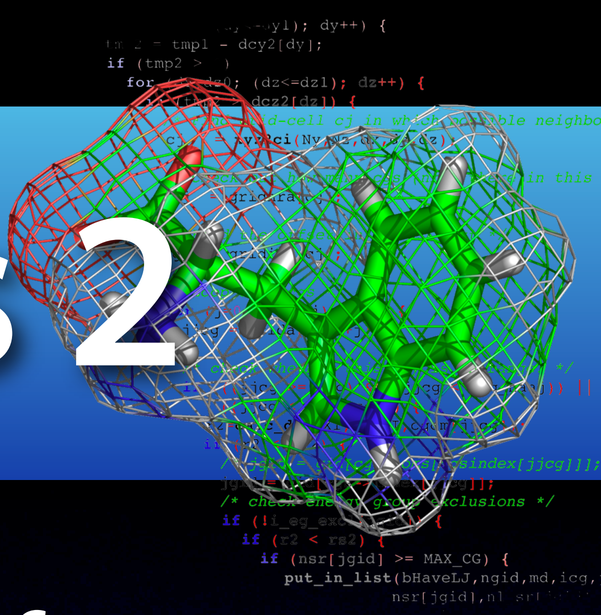
- Gromacs normally uses 256MB to 1GB per process, depending on the system
  - 8GB is fine on a dual quad-core system
- Graphics performance doesn’t matter (for now - we’re working on GPU code...)
- Disk performance doesn’t matter, use cheap 7200 rpm SATA disks
- Store data on one machine/fileserver - use small disks on the others

Other requirements 1



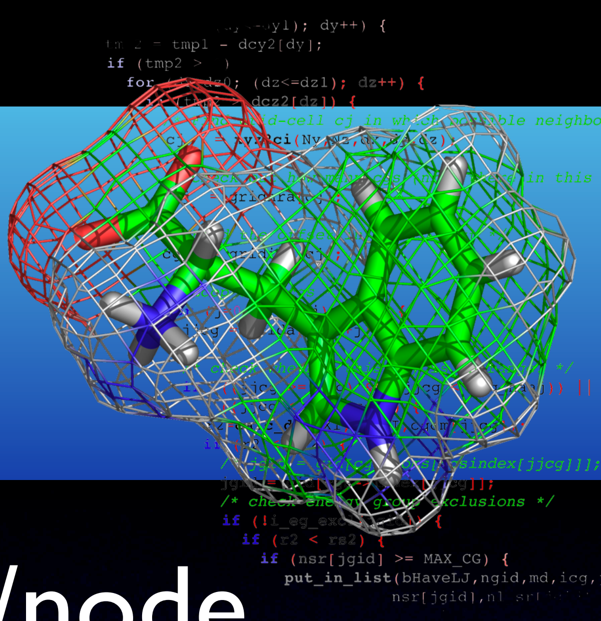
- Gromacs normally uses 256MB to 1GB per process, depending on the system
  - 8GB is fine on a dual quad-core system
- Graphics performance doesn't matter (for now - we're working on GPU code...)
- Disk performance doesn't matter, use cheap 7200 rpm SATA disks
- Store data on one machine/fileserver - use small disks on the others

Other requirements 2



- Mac OS X (e.g. 4 core MacPro) is great for workstations, but for clusters you want really cheap standard x86-64 machines
- Linux operating system - use a free distribution instead of commercial ones!
  - Frequently cheaper to *pay* for MS windows...
- Buy from a vendor that will still be in business when/if you have problems
- Remove all options you don’t need!

Example: 5-10 nodes



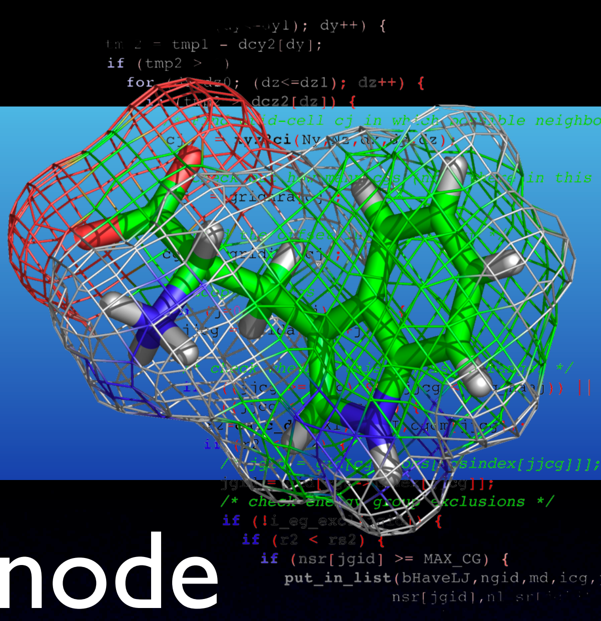
DELL precision 490

HP xw6400

500W/node

- Dual Xeon5355 quad core CPUs @ 2.66GHz
- 8GB Memory @ 667MHz
- 80GB, 7200 rpm SATA disk
- Gigabit ethernet built-in
- 3 year warranty (next business day, on-site)
- List price: Roughly $5000 ($600/core)

Example: 100 nodes



PowerEdge 1950

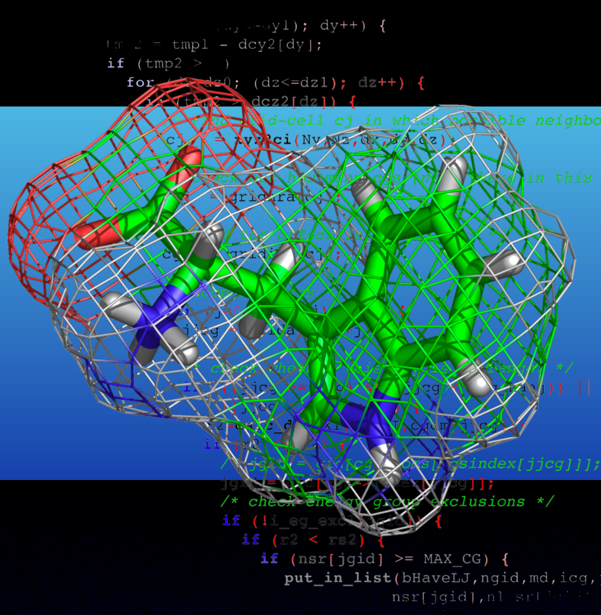
ProLiant DL140

400W/node

40kW power is 350,000 kWh/year You will also need cooling!

- Same basic config: Dual quad-core, 2.66GHz, 8GB
- Requires racks and mounting rails (cheap)
- 1U height - you can fit 42 servers (336 cores) per rack
- Comes without operating system (no Windows tax!)
- Remote management over IPMI (2x gigabit)
- List price: Roughly $5500 ($700/core)

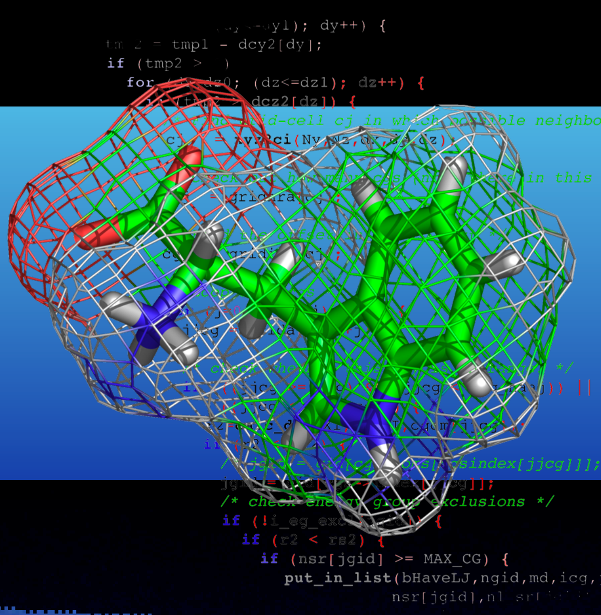
Cheap network

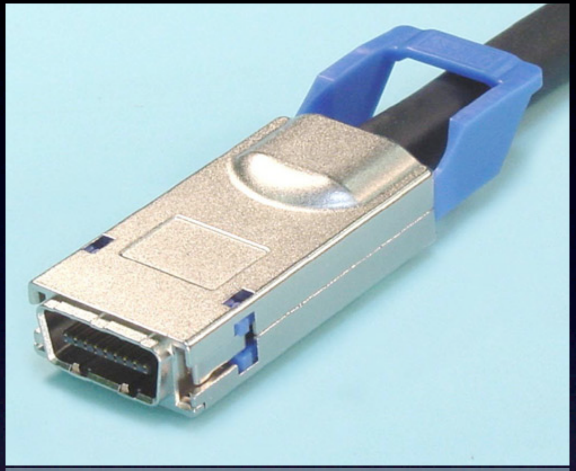


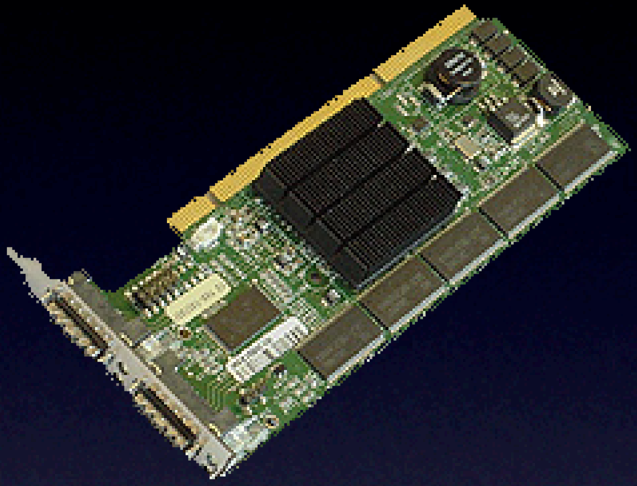
~$700-$1500

- 48-port Gigabit ethernet switch (can often be stacked to make 96-192 ports)
    - 1 gbit/s bandwidth, 100 µs latency
- Gigabit built-in on the nodes, cables cheap
- Good for throughput clusters, limited parallelization scaling between nodes
- Parallelization still works great over the 8 cores in a single node!

Fast: Infiniband





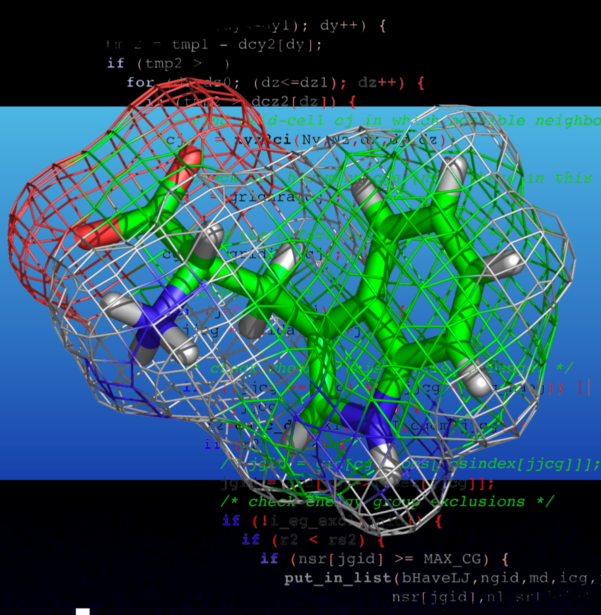


- 10(SDR)-20(DDR)Gbit/s,1-5 µs latency
- Host adapters: $500 per card (SDR)
- Infiniband switch: $7500 for 24 ports SDR
- Cables: $150-500 depending on length (heavy)
- Amazing performance
- DDR IB currently limited by PCI-E bandwidth!
- Alternative: Myrinet

Mother-of-all-switches

- Cisco 7024D
- 288 Port DDR IB switch
- Weight: 200lbs!
- Internal cross-section bandwidth: 11.5 Tb/s
- Port-to-port latency: 200ns
- List price: $359,995.00

Storage

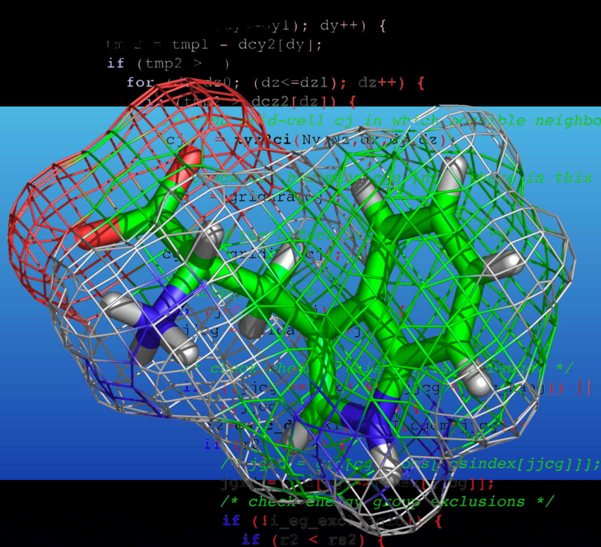


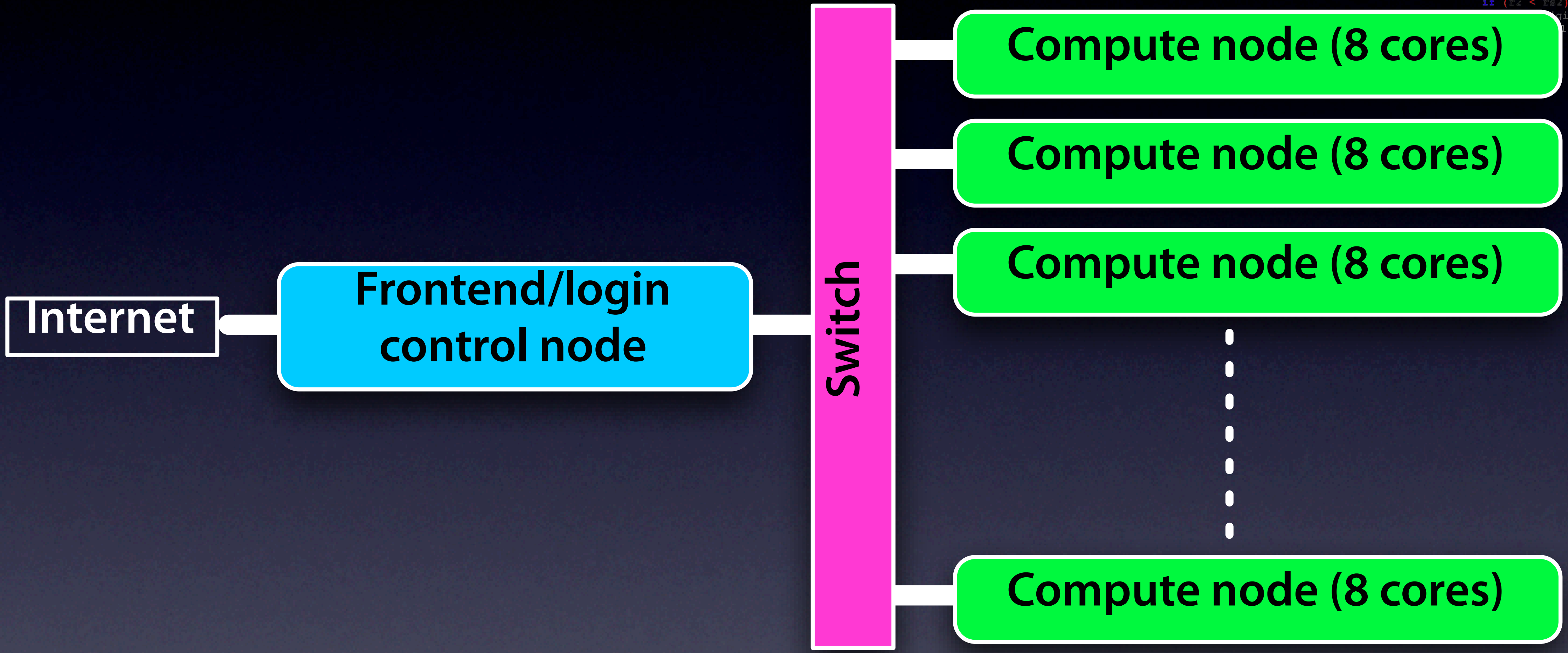
- Low-end: Buy several 1TB SATA disks for the master node, run as RAID5, use as NFS server. But: RAID is *not* backup!
- Medium-level: Dedicated NFS fileserver(s) with 5-10 disks each in RAID5
- High-end: Lustre parallel file system with separate ‘metadata server‘ and ‘object storage servers’ (up to 600MB/s)
- Lustre software is GPL, support costs

lustre

www.clusterfs.com

Network topology





Internet

Frontend/login control node

Switch

Compute node (8 cores)

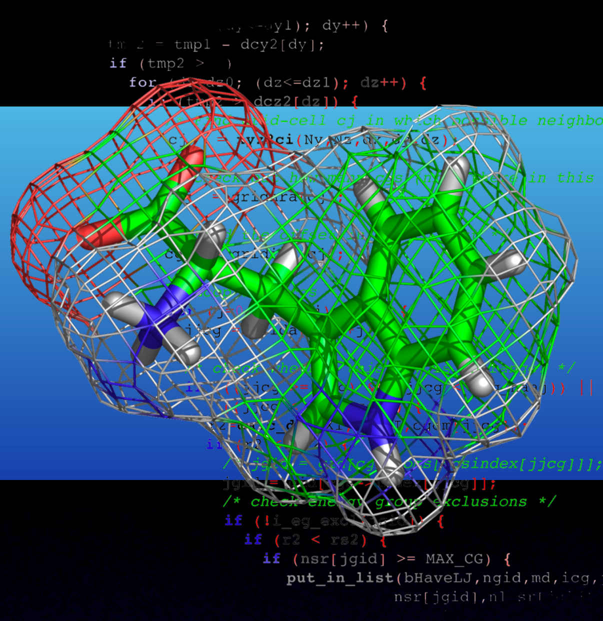
Compute node (8 cores)

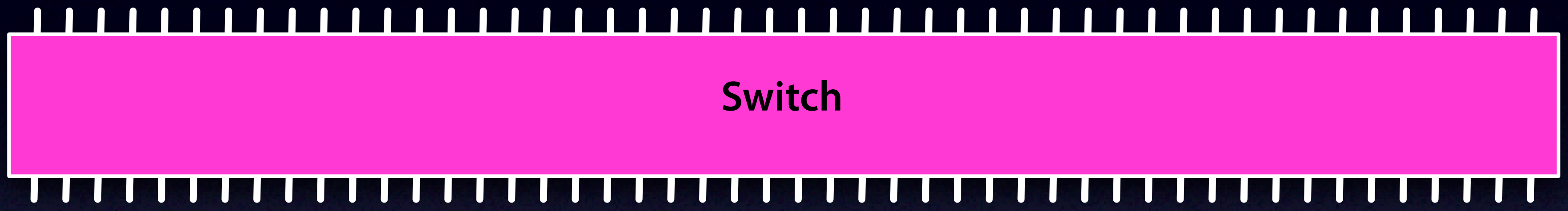
Compute node (8 cores)

Compute node (8 cores)

- Good idea with separate frontend machine
- Does not need to be a 8-core machine!
- Cheap box for $1500 (2 cores, 1GB) will be fine

Switch topologies

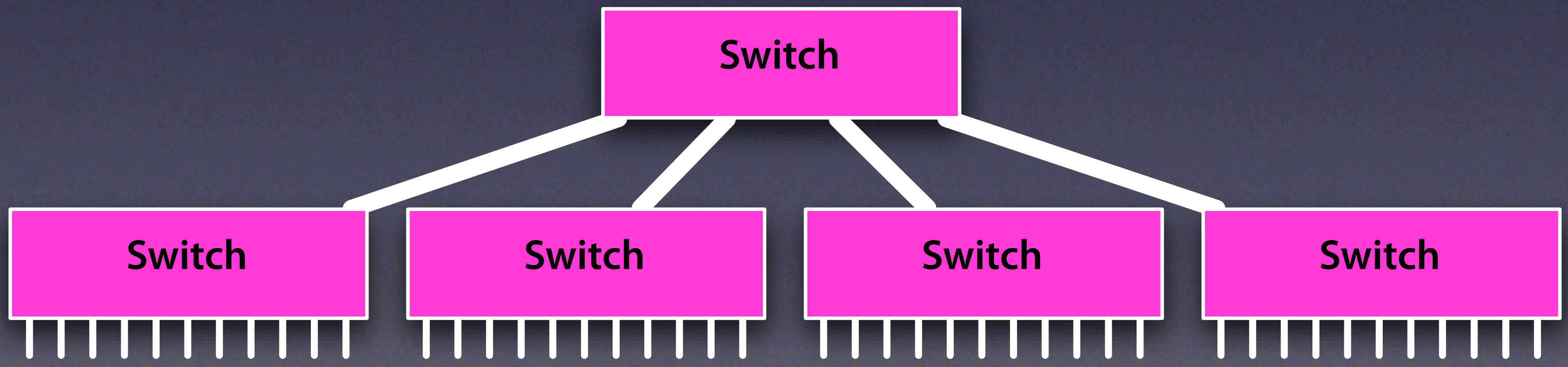




Fast

Low-latency

Expensive

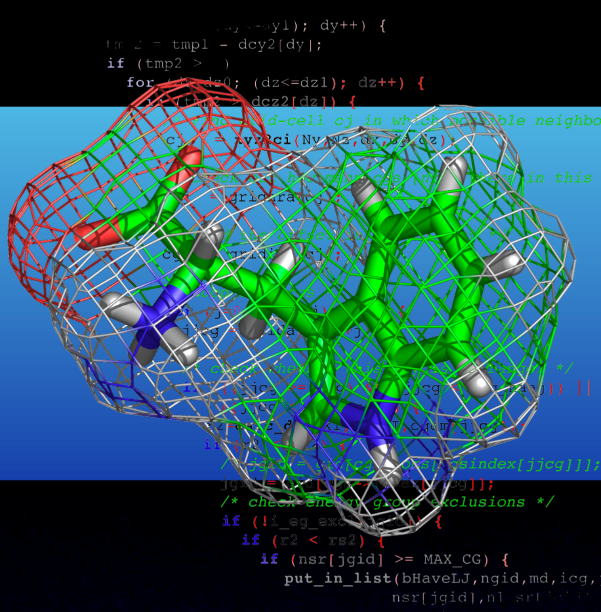


Bottlenecks

Higher-latency

(Much) Cheaper

OS & Software

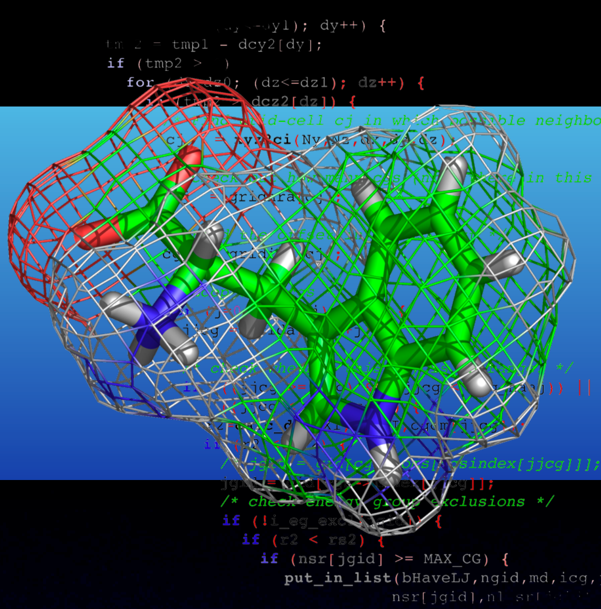


- It's a *very* good idea to use a dedicated scientific Linux cluster distribution
- Recommendation: ROCKS Linux http://www.rocksclusters.org
- Rocks uses CentOS, which is built from Redhat-distributed source code packages
- Comes with everything & documentation!
- Cost: $0 (Developed by SDSC)

ROCKS

www.rocksclusters.org

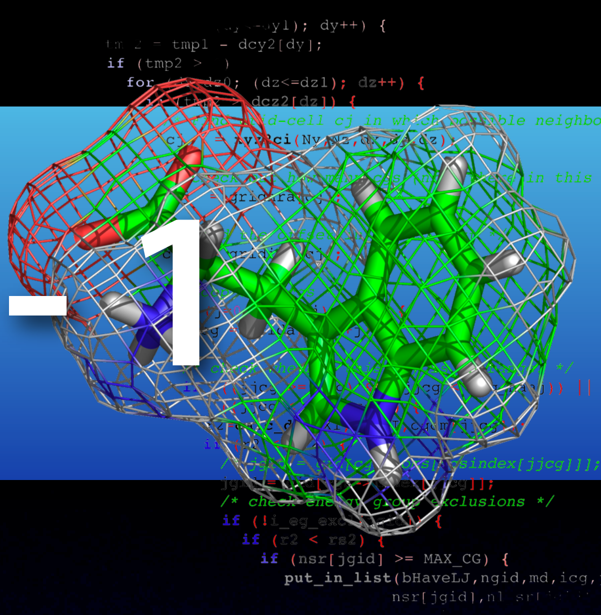
Rocks "Rolls"



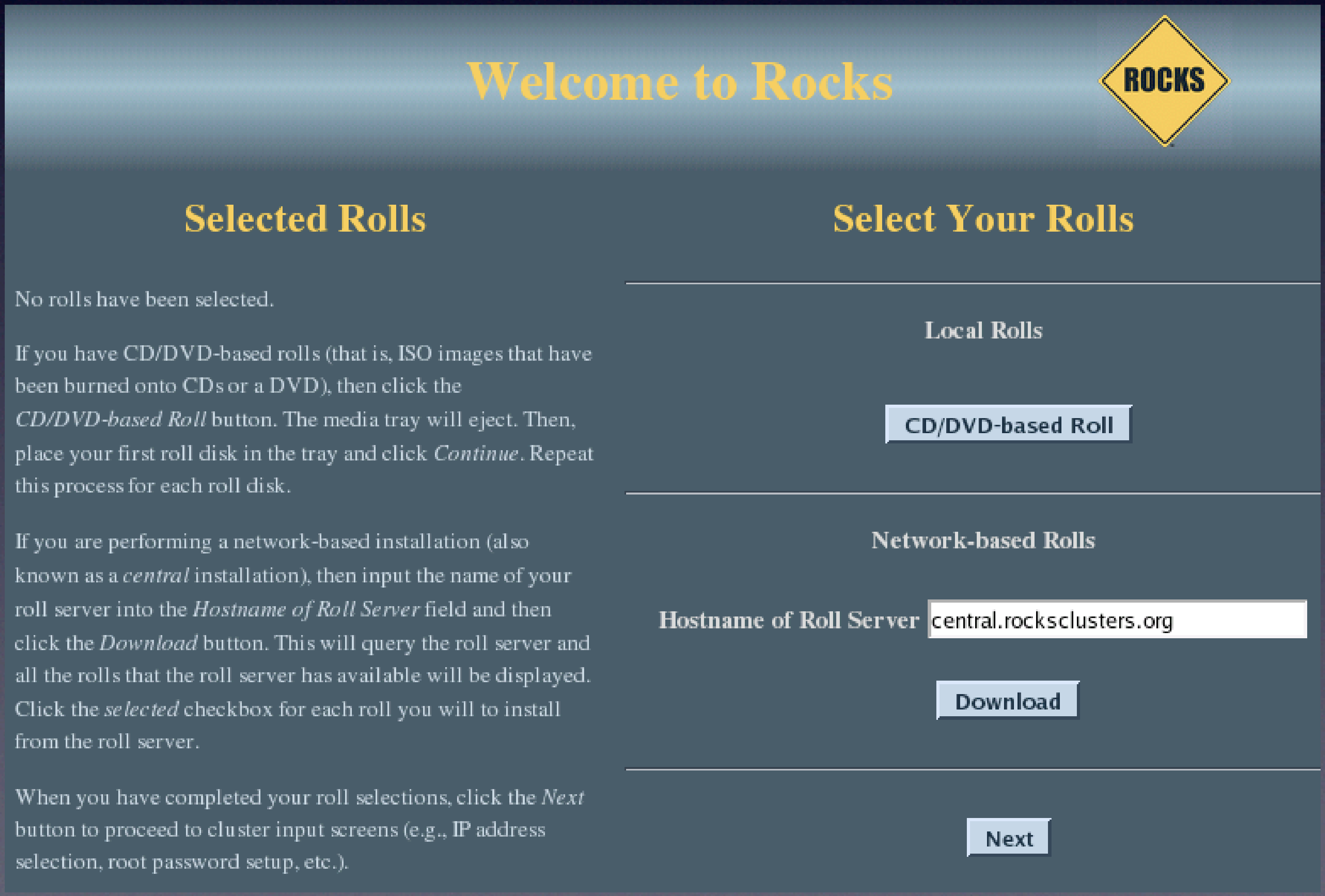
- Rocks comes as a basic distribution, and then you can add functionality by installing additonal "rolls" on the frontend

- Uses bittorrent to install nodes
- New! Rocks Bio Roll: HMMer, BLAST, ClustalW, MPI_Blast, and... Gromacs!
- Precompiled parallel MPI binaries
  - Automatically available on all nodes

Rocks crash course - 1



- Download DVD/CD images
- Insert into frontend, boot, select rolls



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.).

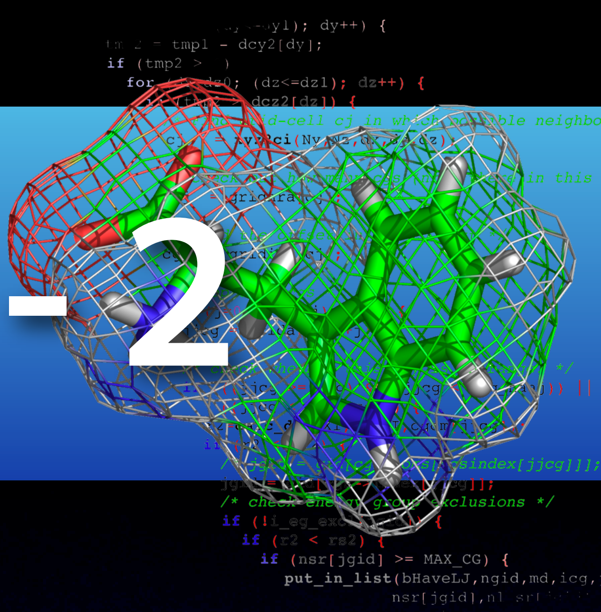
Select Your Rolls

Local Rolls

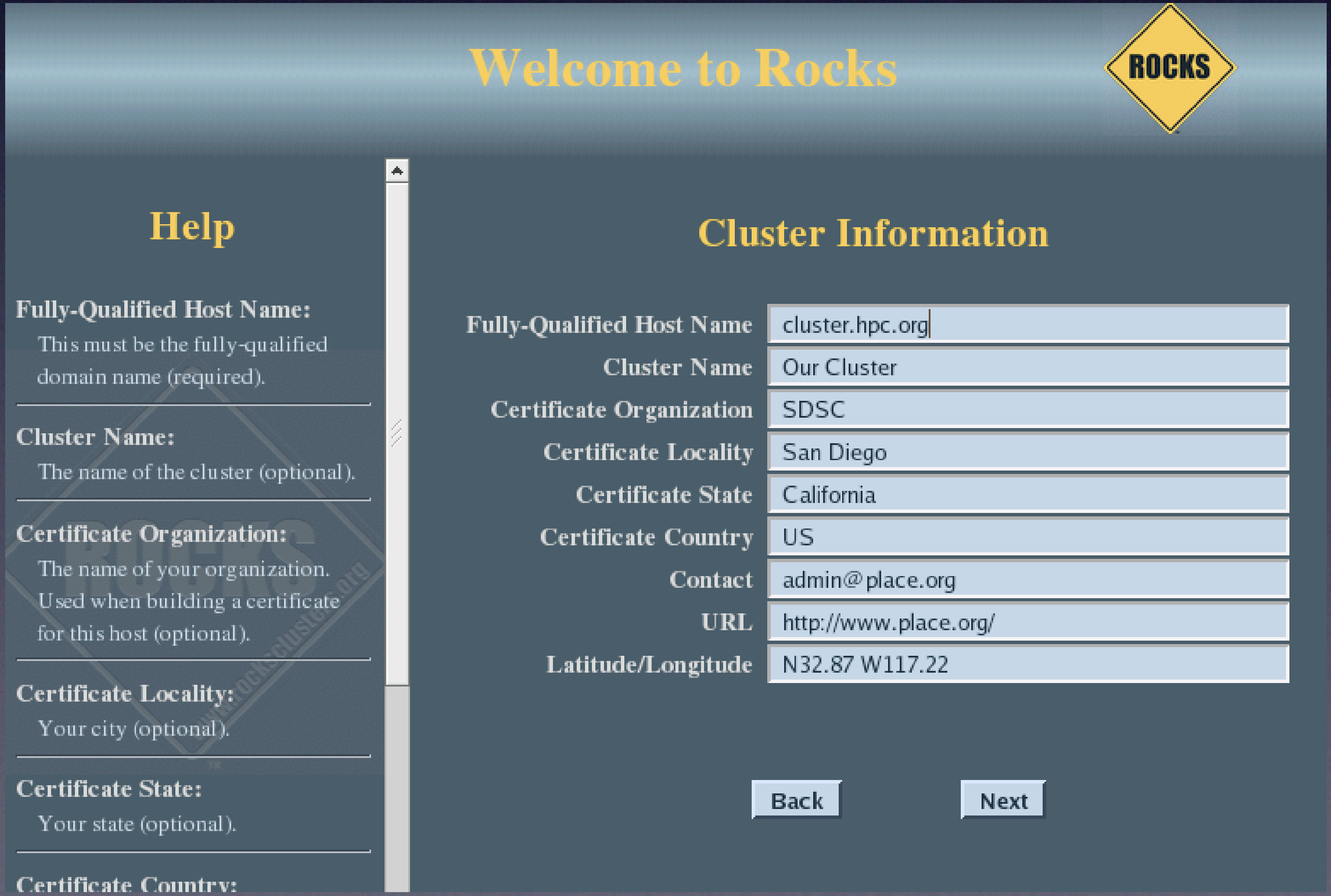
Network-based Rolls

Hostname of Roll Server: central.rocksclusters.org

Rocks crash course - 2



- Give your cluster a name, IP, etc.
- Basically a vanilla Linux installation



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).

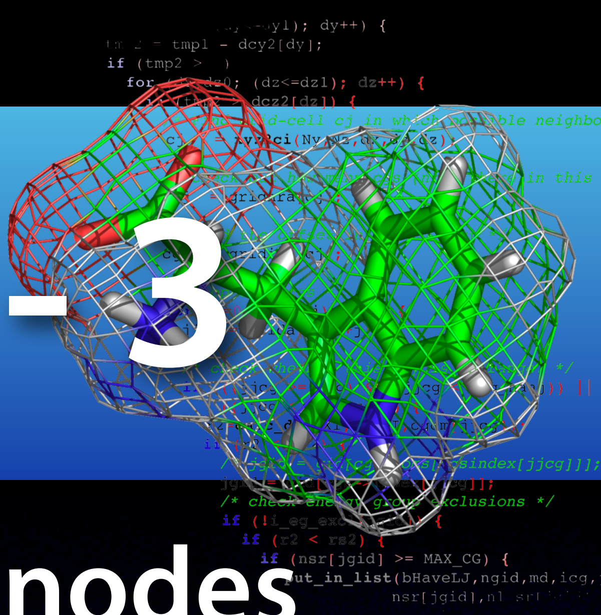
Certificate State:
Your state (optional).

Certificate Country:

Cluster Information

Fully-Qualified Host Name	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

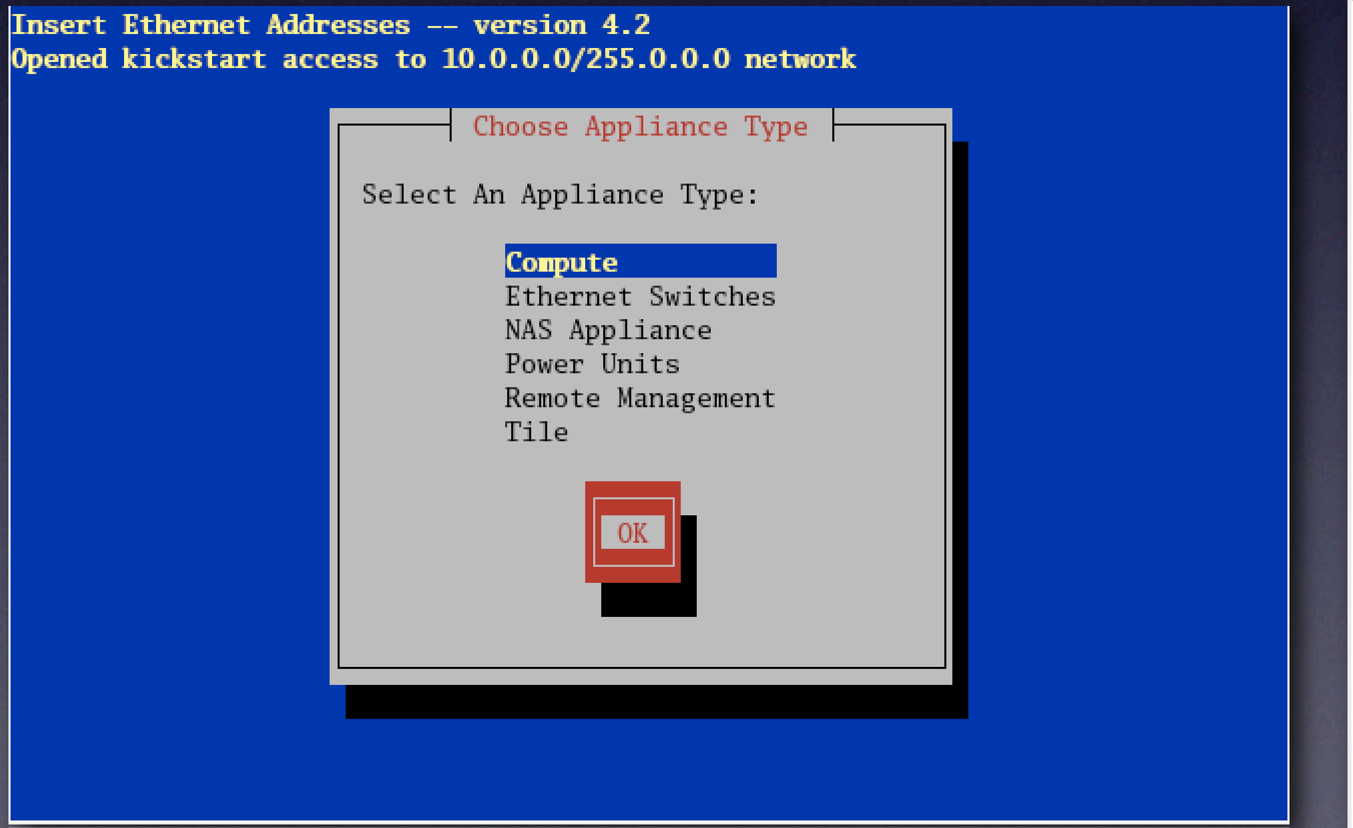
Rocks crash course - 3



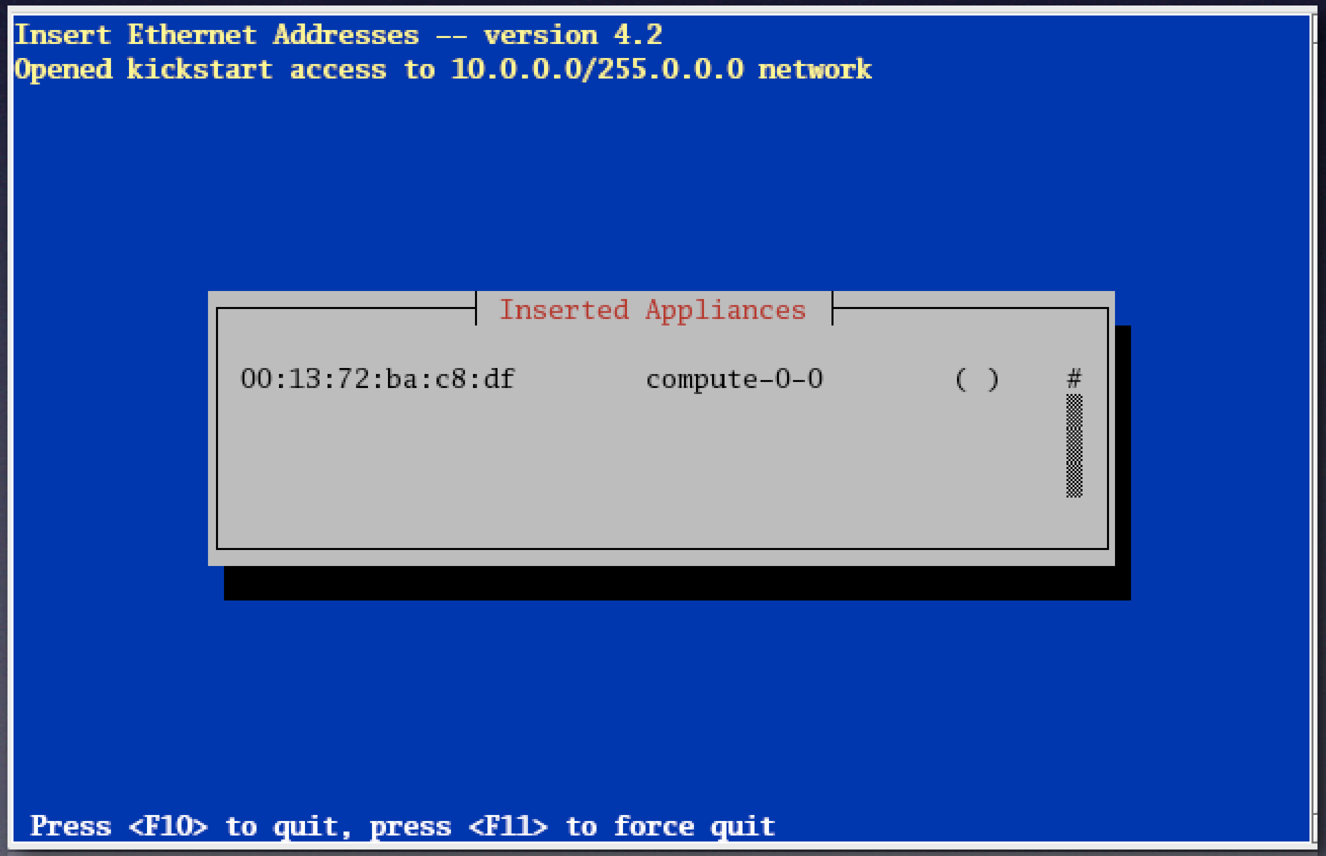
- Post-install: Tell it you want to add nodes

```
#> insert-ethers
```

- Start nodes one by one, tell them to boot from the network before CD or hard disk



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

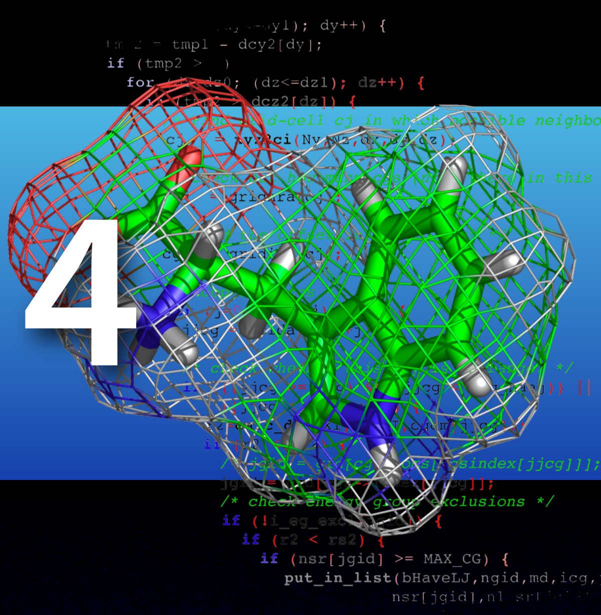


```
Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network
```

```
00:13:72:ba:c8:df        compute-0-0        ( )
```

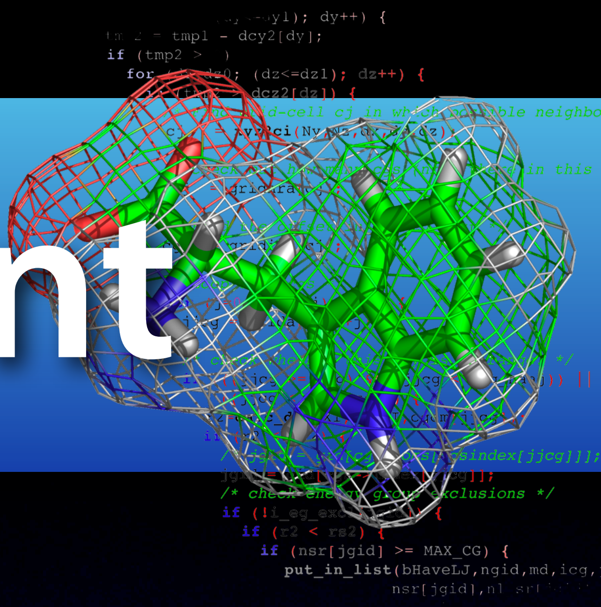
```
Press <F10> to quit, press <F11> to force quit
```

Rocks crash course - 4


if (tmp2 >
for (dz=dz0; (dz<=dz1); dz++) {
/* check the energy group exclusions */
if (nsr[jgid] >= MAX_CG) {
put_in_list(bHaveLJ,ngid,md,icg,
nsr[jgid],nl_sr

- Never debug problems on nodes - just restart them and they will reinstall!
- If that doesn’t solve it, the hardware is faulty - use your 3-year warranty!
- Nodes will report automatically to the frontend when they are up and running
- Rocks comes with packages for the Torque & Maui batch queue system / scheduler
  Show queue: `showq`   Start jobs: `qsub`

Remote management



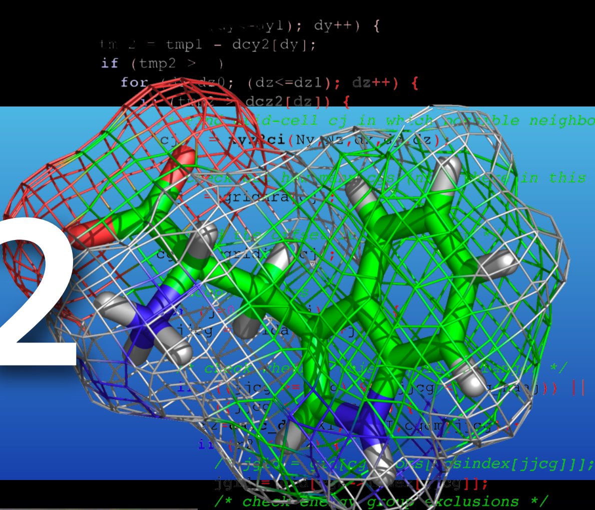
- Most rackmounted nodes come with built-in support for IPMI remote control
- Uses ethernet connection to frontend
- Nodes can be powered on/off
- Check: temperature, fans, warnings, etc.
- Console redirection (edit BIOS remotely)
- Absolutely necessary for >50 nodes...

Installing a new cluster

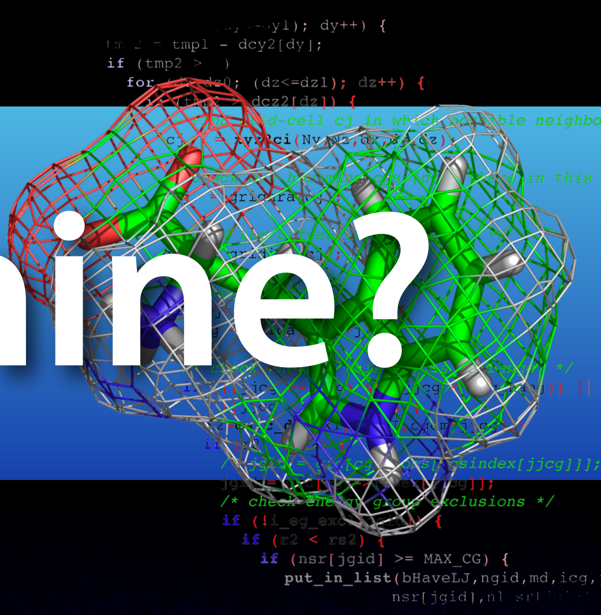
Old Bio-X cluster at Stanford Univ:
300 nodes, 2U each
SMP - 2 * Xeon @ 2.8GHz
1Gb / node
Ethernet (mixed Gb/100Mb)
Small NFS file server
Used Rocks (zero-admin cluster)
Decommissioned Jan 2007

Lessons:
Automate everything possible... things *will* break
Network becomes bottleneck to fileserver
Don't mix 100Mbit and gigabit (packet loss)
Invest in expensive switches

Virtual tour of Bio-X2

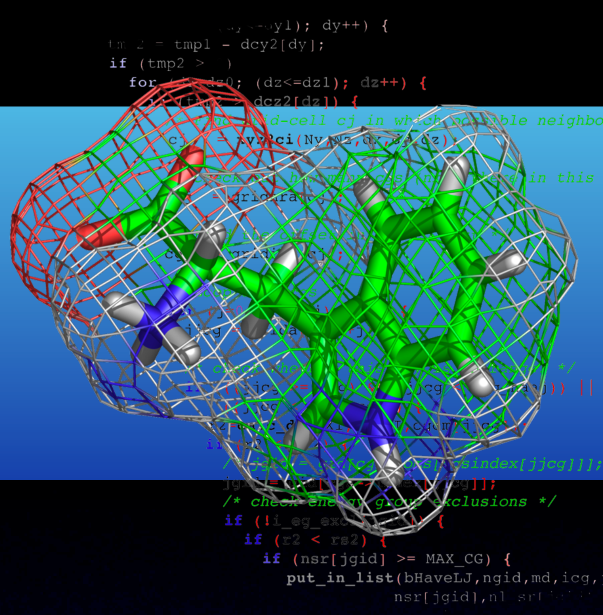


Ultimate dream machine?



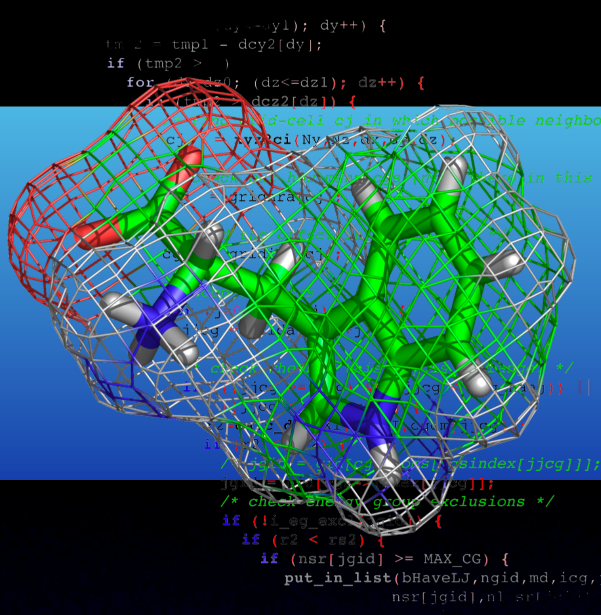
- Cray XT-4 with Opteron processors
- If cost matters, you can’t afford it :-)
- 3D torus interconnect
  (fast to 6 nearest neighbors in 3D)
- Proprietary Cray system bus design
- Bandwidth: 60Gbit/second over *each* link
- Machine just being installed at CSC
- Gromacs: 1.1TFLOPS on 384 nodes y-day!

Summary



- Linux/x86 is quite mature both for cheap mini-clusters and supercomputers
- Many cores, save on memory
- Scaling: infiniband Throughput: gigabit
- Free linux cluster distro: ROCKS
- Gromacs comes pre-compiled on the Bio Roll
- Automated administration really helps
- No big deal to install a Linux cluster today!

Acknowledgments



- Tanya Raschke
- Kilian Cavalotti
- Michael Levitt, Vijay Pande, Russ Altman, and others on Bio-X2 grant
- NSF - paying for the machine