Autonomic Condor Clouds

David Wolinsky
ACIS P2P Group
University of Florida

So What's the Big Deal

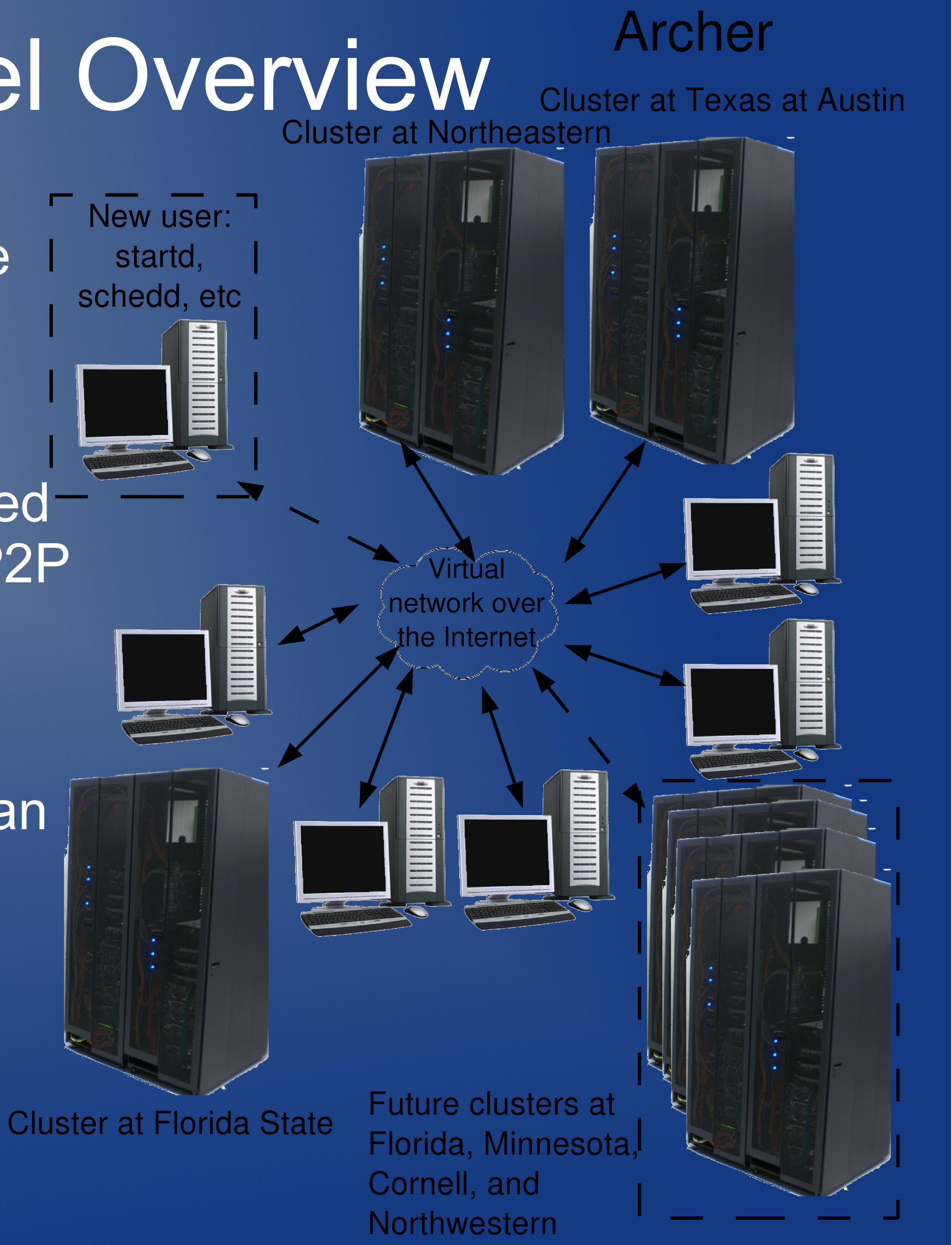
- Support connectivity across the Internet, in constrained locations, and with clouds
- Simplify packaging
- Minimize Condor configuration
- Reduce downtime
- Let's try to make this easy ....

Discussion Goals

- The High-Level Overview
- Self-Configurable Condor Components
- Virtual Networking
- The User Experience (Time allowing)

High level Overview

- Self-contained VM Appliance
- Configuration through virtual floppy
- Fully distributed, decentralized Virtual Private Network via P2P Overlay
- Job scheduling via Condor
- Customization through Debian and Stacked File system


Archer
Cluster at Texas at Austin
Cluster at Northeastern
New user: startd, schedd, etc
Virtual network over the Internet
Cluster at Florida State
Future clusters at Florida, Minnesota, Cornell, and Northwestern

Everything But Virtual Networking

- Virtual Machines (VMs)
  - Support for VMware, VirtualBox, KVM, and Xen
  - Homogenous environment on heterogenous resources
- Configuration
  - Provided by Virtual floppy, 25 KB download
  - P2P overlay info
  - Users identification certificates
  - Condor machine type
- Customization
  - Based upon Debian 4.0, access to apt repositories
  - Stacked file systems allow users to modify image and share only the changes
- Condor .... :-)

Configurable Components

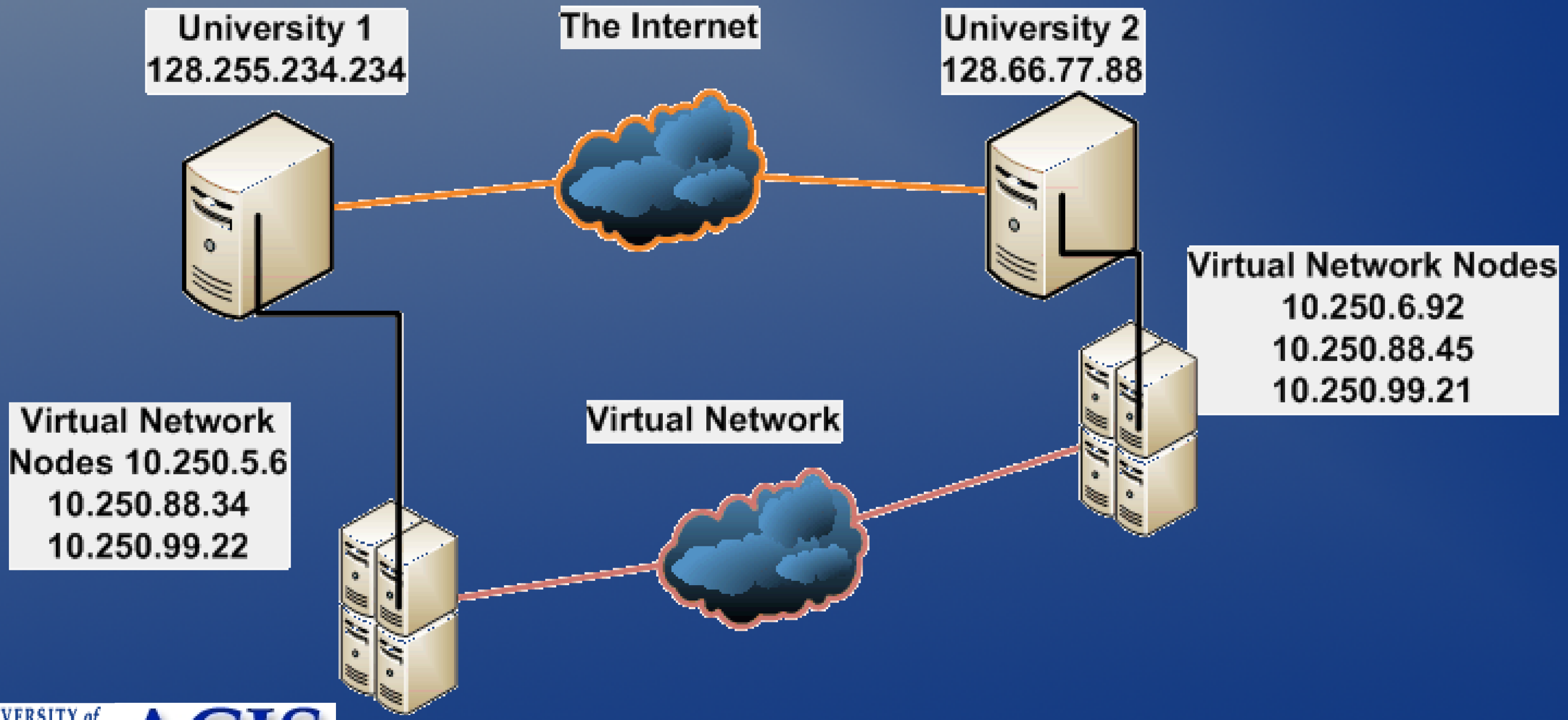
- Daemons to start
  - Worker – execution only - startd
  - Client - submission and execution – startd and schedd
  - Master / Server – negotiator and collector
- Which Master / Server to connect or flock to
- User / Group Resource ownership and preemption
- Client can share files via autofs enabled NFS
- Monitoring and binding to a dynamic IP address

More Opportunities

- Support for VM Universe
- Distributed data storage
- Web portal front-end
- Self-policing security system
- Self-sustaining condor cluster
- Portable (decentralized) Condor System Configurer
- Configurable through Social Networks and User Portals

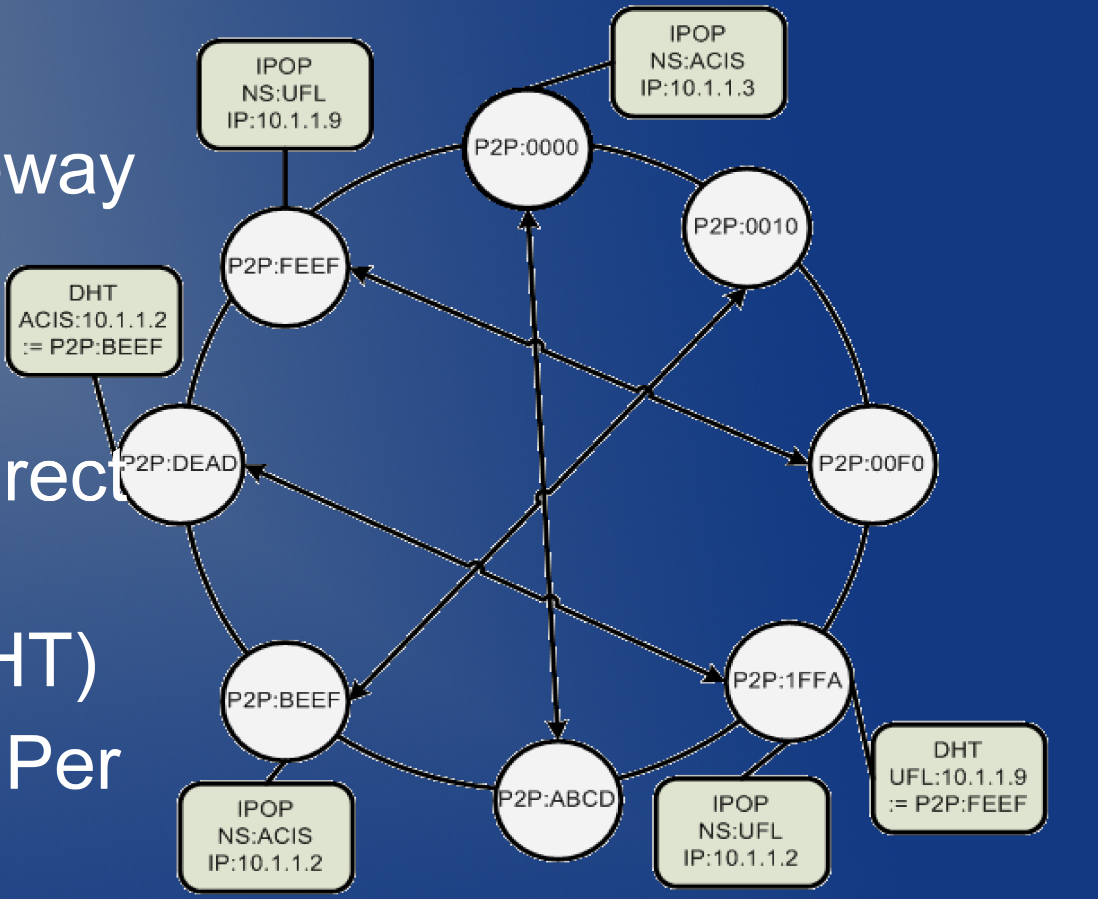
Virtual Networking

- Unified layer 3 (IP) network for all machines
- Cross-site communication without a middleware broker



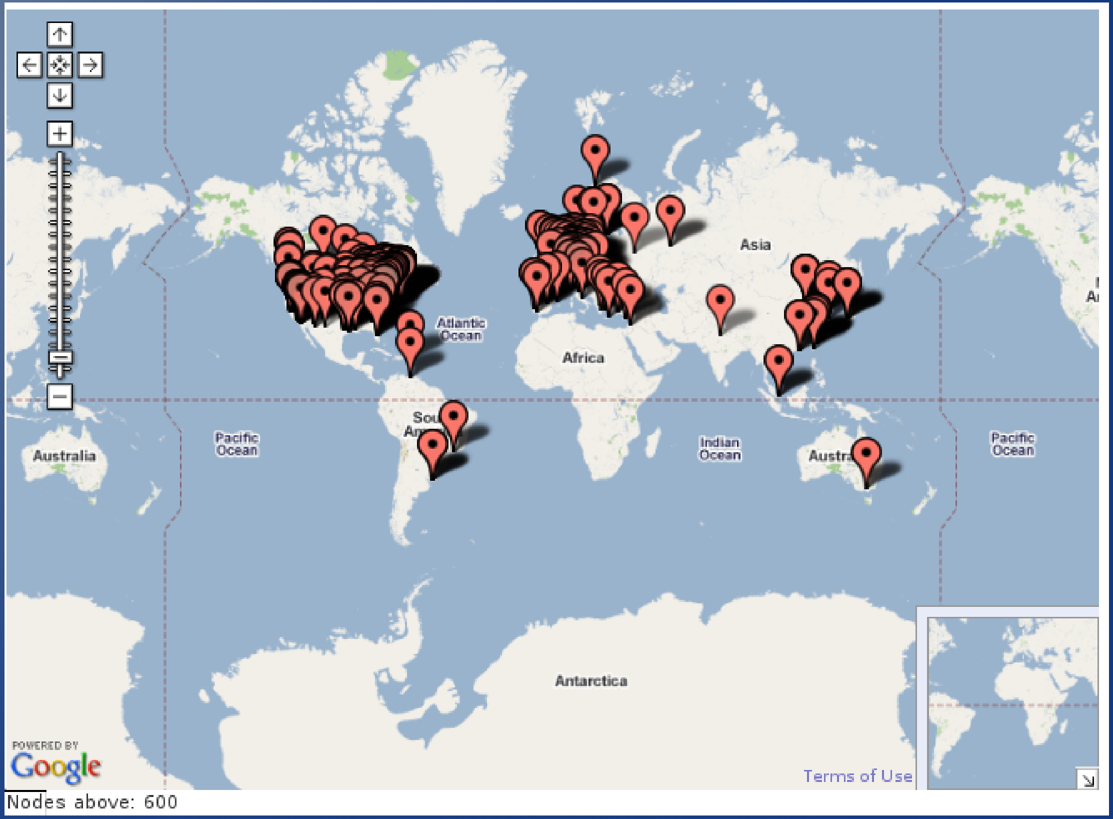
IPOP

- Open Source
- NAT Traversal (STUN)
- Transparent Subnet Gateway
- Structured P2P Network Overlay
- Provides tunneling and direct shortcuts
- Distributed data store (DHT)
- Multiple Virtual Networks Per Overlay



P2P Overlay

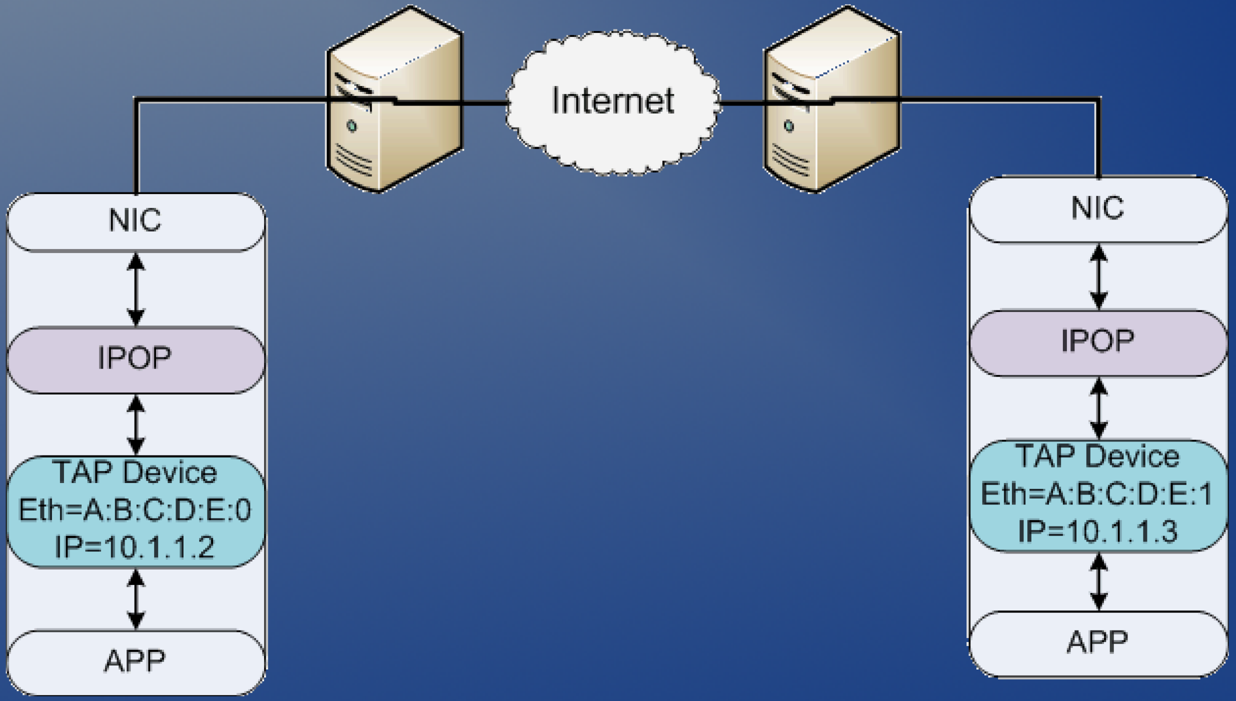
- Several hundred well distributed nodes
- Assist in bootstraping and NAT traversal
- Runs on top of Planetlab



Nodes above: 600

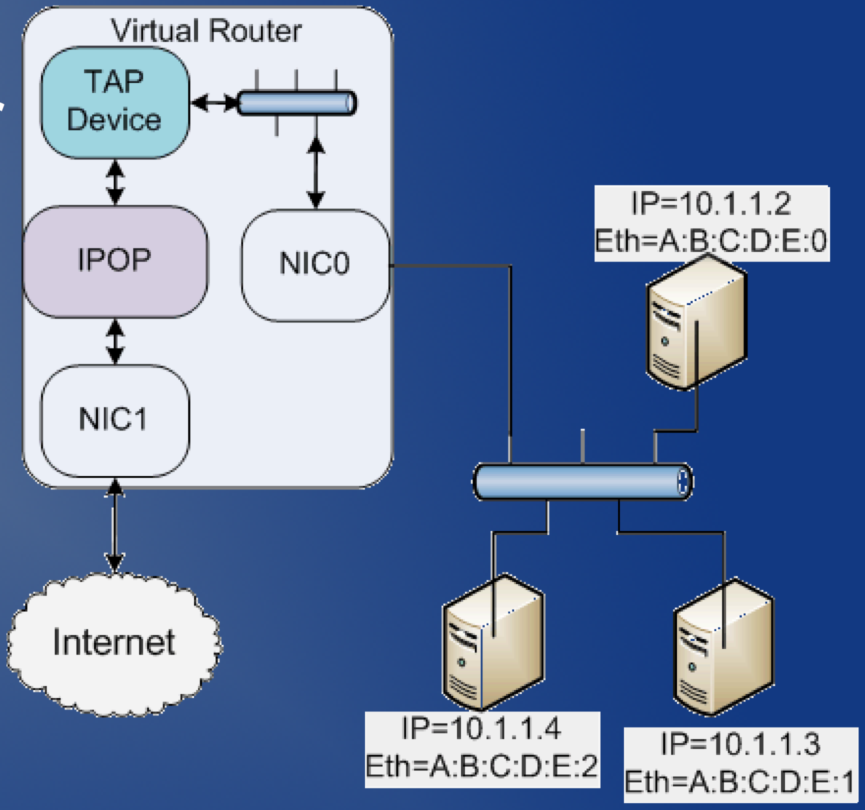
VN Interfaces

- Each machine has VN Interface running locally
- Machine has VN and “Internet” connectivity



VN Routers

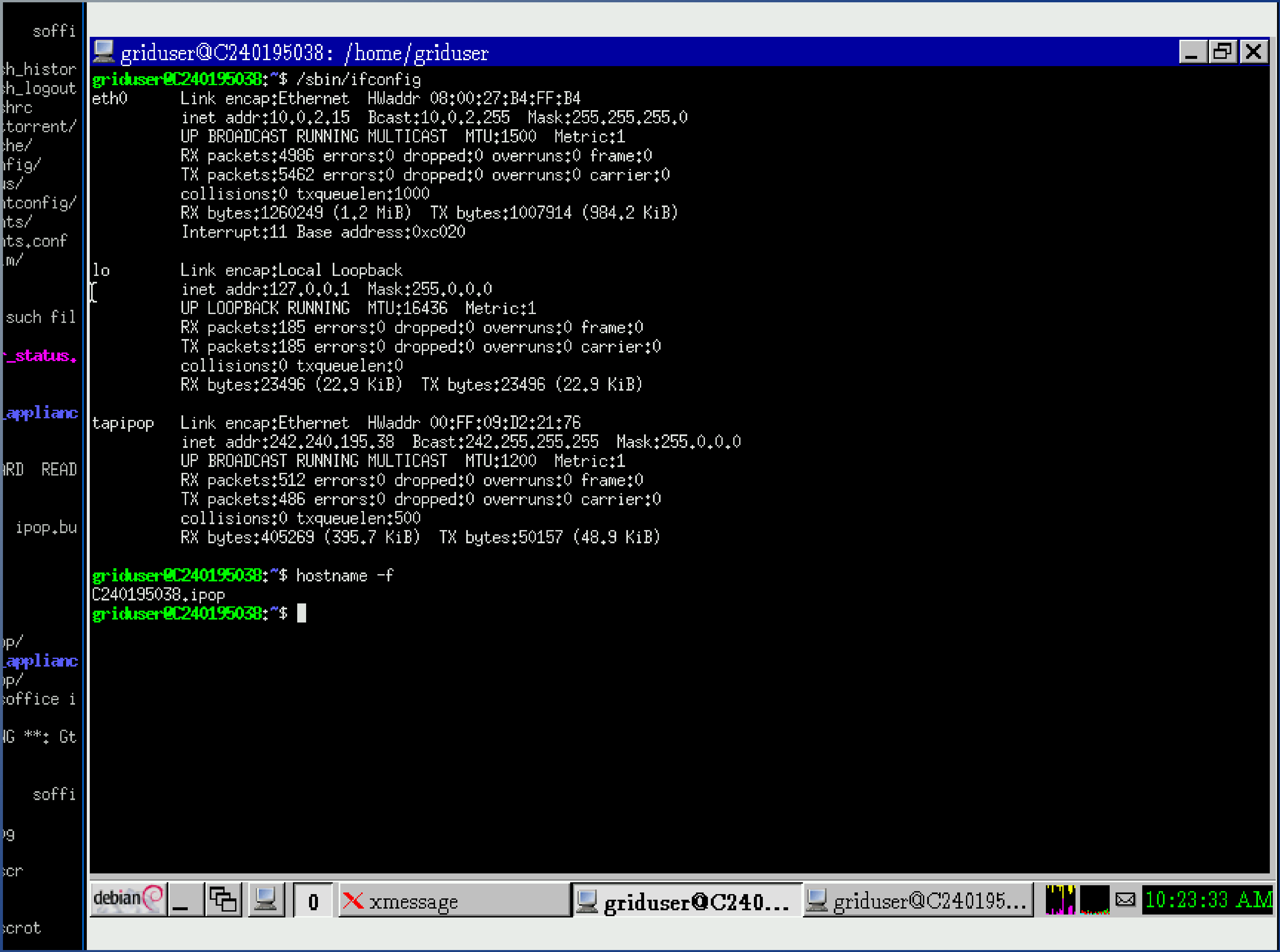
- Single VN instance (Router) for entire cluster
- Limited to no resource configuration
- Isolated VN overhead
- May have "Internet" connectivity if there is an "Internet" router



UF UNIVERSITY of FLORIDA ACIS

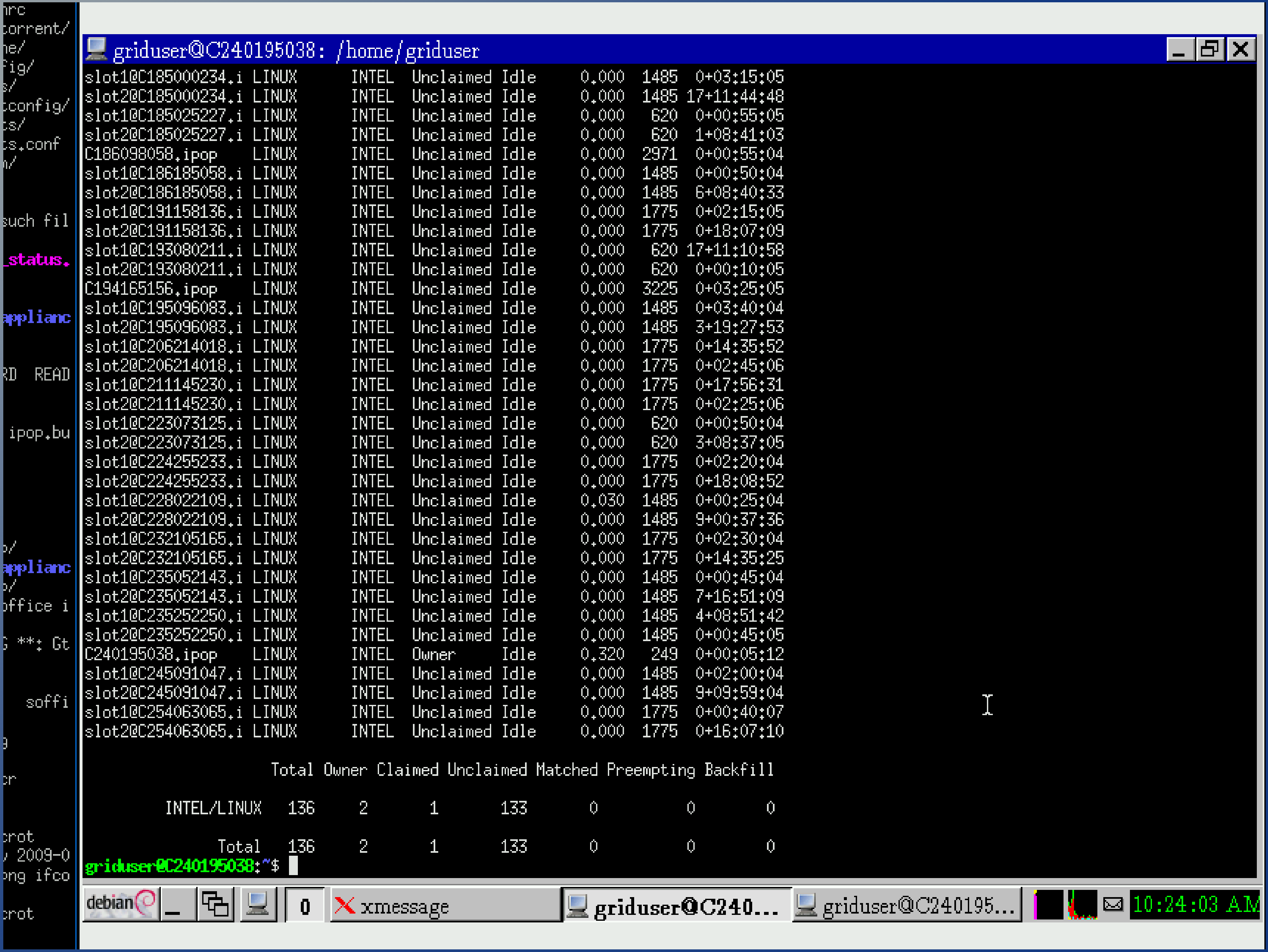
The User Experience

- Time permitting video - http://www.youtube.com/watch?v=1XDvITdayhs
- Otherwise slides –
  - Boot VM and obtain IP Addresses
  - Condor access
  - Direct connectivity (i.e. low ping overheads)



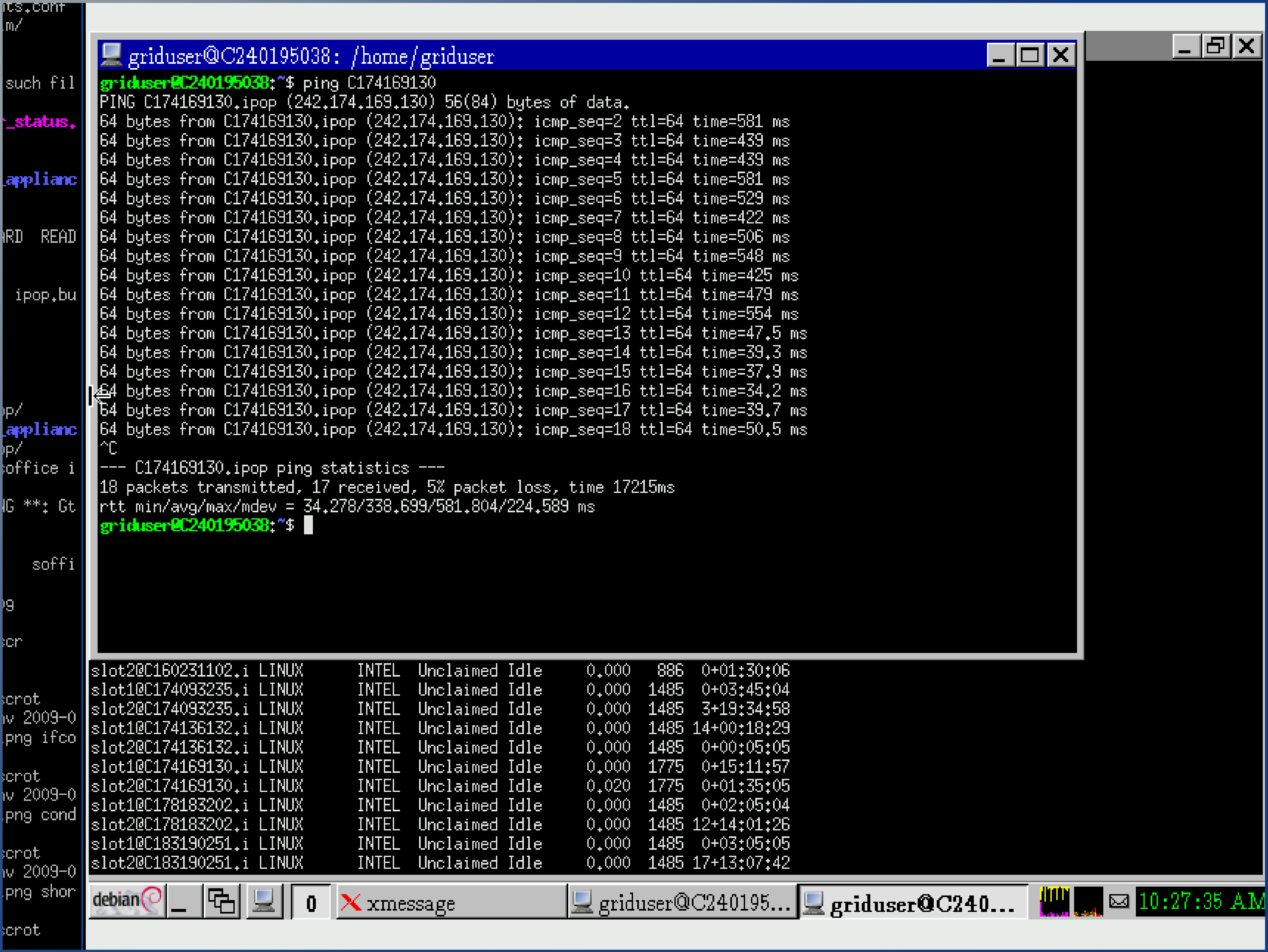
```
griduser@C240195038:~$ /sbin/ifconfig
eth0      Link encap:Ethernet  HWaddr 08:00:27:B4:FF:B4
          inet addr:10.0.2.15  Bcast:10.0.2.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:4986 errors:0 dropped:0 overruns:0 frame:0
          TX packets:5462 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:1260249 (1.2 MiB)  TX bytes:1007914 (984.2 KiB)
          Interrupt:11 Base address:0xc020

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:185 errors:0 dropped:0 overruns:0 frame:0
          TX packets:185 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:23496 (22.9 KiB)  TX bytes:23496 (22.9 KiB)

tapipop   Link encap:Ethernet  HWaddr 00:FF:09:D2:21:76
          inet addr:242.240.195.38  Bcast:242.255.255.255  Mask:255.0.0.0
          UP BROADCAST RUNNING MULTICAST  MTU:1200  Metric:1
          RX packets:512 errors:0 dropped:0 overruns:0 frame:0
          TX packets:486 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:500
          RX bytes:405269 (395.7 KiB)  TX bytes:50157 (48.9 KiB)

griduser@C240195038:~$ hostname -f
C240195038.ipop
griduser@C240195038:~$ 
```

```
    soffi
h_histor
h_logout
hrc
torrent/
he/
fig/
s/
tconfig/
ts/
ts.conf
m/

 such fil
_status.

_applianc

RD  READ

 ipop.bu

p/
_applianc
p/
office i
G **: Gt

    soffi
g
cr

crot
```



```
slot1@C185000234.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+03:15:05
slot2@C185000234.i LINUX      INTEL  Unclaimed Idle     0.000  1485 17+11:44:48
slot1@C185025227.i LINUX      INTEL  Unclaimed Idle     0.000   620  0+00:55:05
slot2@C185025227.i LINUX      INTEL  Unclaimed Idle     0.000   620  1+08:41:03
C186098058.ipop    LINUX      INTEL  Unclaimed Idle     0.000  2971  0+00:55:04
slot1@C186185058.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+00:50:04
slot2@C186185058.i LINUX      INTEL  Unclaimed Idle     0.000  1485  6+08:40:33
slot1@C191158136.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+02:15:05
slot2@C191158136.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+18:07:09
slot1@C193080211.i LINUX      INTEL  Unclaimed Idle     0.000   620 17+11:10:58
slot2@C193080211.i LINUX      INTEL  Unclaimed Idle     0.000   620  0+00:10:05
C194165156.ipop    LINUX      INTEL  Unclaimed Idle     0.000  3225  0+03:25:05
slot1@C195096083.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+03:40:04
slot2@C195096083.i LINUX      INTEL  Unclaimed Idle     0.000  1485  3+19:27:53
slot1@C206214018.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+14:35:52
slot2@C206214018.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+02:45:06
slot1@C211145230.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+17:56:31
slot2@C211145230.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+02:25:06
slot1@C223073125.i LINUX      INTEL  Unclaimed Idle     0.000   620  0+00:50:04
slot2@C223073125.i LINUX      INTEL  Unclaimed Idle     0.000   620  3+08:37:05
slot1@C224255233.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+02:20:04
slot2@C224255233.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+18:08:52
slot1@C228022109.i LINUX      INTEL  Unclaimed Idle     0.030  1485  0+00:25:04
slot2@C228022109.i LINUX      INTEL  Unclaimed Idle     0.000  1485  9+00:37:36
slot1@C232105165.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+02:30:04
slot2@C232105165.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+14:35:25
slot1@C235052143.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+00:45:04
slot2@C235052143.i LINUX      INTEL  Unclaimed Idle     0.000  1485  7+16:51:09
slot1@C235252250.i LINUX      INTEL  Unclaimed Idle     0.000  1485  4+08:51:42
slot2@C235252250.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+00:45:05
C240195038.ipop    LINUX      INTEL  Owner     Idle     0.320   249  0+00:05:12
slot1@C245091047.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+02:00:04
slot2@C245091047.i LINUX      INTEL  Unclaimed Idle     0.000  1485  9+09:59:04
slot1@C254063065.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+00:40:07
slot2@C254063065.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+16:07:10

                     Total Owner Claimed Unclaimed Matched Preempting Backfill

         INTEL/LINUX   136     2       1       133       0          0        0

               Total   136     2       1       133       0          0        0
griduser@C240195038:~$
```

```
rc
orrent/
e/
ig/
/
config/
s/
s.conf
/

uch fil

_status.

applianc

D  READ

 ipop.bu

/
applianc
/
ffice i
 **: Gt

  soffi



rot
 2009-0
ng ifco
rot
```



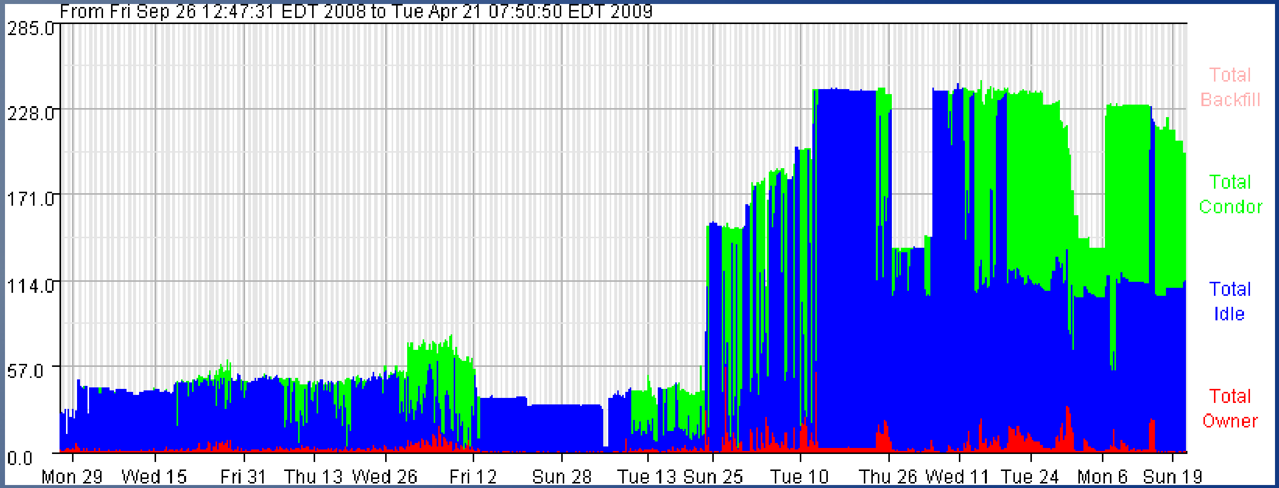
```
griduser@C240195038:~$ ping C174169130
PING C174169130.ipop (242.174.169.130) 56(84) bytes of data.
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=2 ttl=64 time=581 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=3 ttl=64 time=439 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=4 ttl=64 time=439 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=5 ttl=64 time=581 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=6 ttl=64 time=529 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=7 ttl=64 time=422 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=8 ttl=64 time=506 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=9 ttl=64 time=548 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=10 ttl=64 time=425 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=11 ttl=64 time=479 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=12 ttl=64 time=554 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=13 ttl=64 time=47.5 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=14 ttl=64 time=39.3 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=15 ttl=64 time=37.9 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=16 ttl=64 time=34.2 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=17 ttl=64 time=39.7 ms
64 bytes from C174169130.ipop (242.174.169.130): icmp_seq=18 ttl=64 time=50.5 ms
^C
--- C174169130.ipop ping statistics ---
18 packets transmitted, 17 received, 5% packet loss, time 17215ms
rtt min/avg/max/mdev = 34.278/338.699/581.804/224.589 ms
griduser@C240195038:~$
```

```
slot2@C160231102.i LINUX      INTEL  Unclaimed Idle     0.000   886  0+01:30:06
slot1@C174093235.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+03:45:04
slot2@C174093235.i LINUX      INTEL  Unclaimed Idle     0.000  1485  3+19:34:58
slot1@C174136132.i LINUX      INTEL  Unclaimed Idle     0.000  1485 14+00:18:29
slot2@C174136132.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+00:05:05
slot1@C174169130.i LINUX      INTEL  Unclaimed Idle     0.000  1775  0+15:11:57
slot2@C174169130.i LINUX      INTEL  Unclaimed Idle     0.020  1775  0+01:35:05
slot1@C178183202.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+02:05:04
slot2@C178183202.i LINUX      INTEL  Unclaimed Idle     0.000  1485 12+14:01:26
slot1@C183190251.i LINUX      INTEL  Unclaimed Idle     0.000  1485  0+03:05:05
slot2@C183190251.i LINUX      INTEL  Unclaimed Idle     0.000  1485 17+13:07:42
```

```
nts.conf
m/
such fil
_status.
_applianc
RD  READ
 ipop.bu
p/
_applianc
p/
soffice i
G **: Gt
   soffi
g
cr
crot
w 2009-0
.png ifco
crot
w 2009-0
.png cond
crot
w 2009-0
.png shor
crot
```

Where We Are Now

- CondorView
- Tutorials
- Hadoop Appliance (Uses Condor to
- Organize)
- Archer
  - Computer Architecture research/education
  - Hundreds of cores distributed over 6 universities in the US over 3 years
- Over 100,000 CPU Hours used since October



Projects Using IPOP

- Purdue – BoilerGrid
- Clemson – Campus Grids
- FCCN in Portugal – Hadoop-based Web Indexing - GaPPa

Questions

- Our Projects
  - Grid Appliance, Archer – grid-appliance.org
  - IPOP – ipop-project.org
- Acknowledgements
  - ACIS P2P Group
  - Condor Group – Ben Burnett and Alain Roy
  - Effort sponsored by the NSF under grants OCI-0721867 and CNS-0751112.
  - Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

UF UNIVERSITY of FLORIDA ACIS

Simics w/ NFS

- Simics requires license server – runs on condor master – Time sync issue :(
- Can use LARGE (many GB) input files
- Expensive to transfer over the wide area!
- NFS allows block level access, only transfer what you need, and data cache, so if a job comes again, we reduce our transfer overhead
- From http://www.grid-appliance.org/wiki/index.php/Archer:Simics

User Feedback

The best thing we like about Archer is its computing power and easy accessibility and usability. We can easily access 50 or more Intel Xeon cores remotely for our computing needs to accomplish something we cannot do before. For example, in our recent data prefetching and cache management studies, we used more than 10000 simulation hours on Archer.

Besides the huge computational resources, Archer is unbelievably easy to use. All the needed software is packaged in a virtual machine, including the IPOP, which provides communication among the grid. It is all transparent to users. Users just need to download the Grid Appliance package, and is ready to go. Archer even provides a couple of popular CPU simulators by default, like Simics, SimpleScalar and PTLsim, etc. Our group uses Simics frequently and glad to see it is available on Archer. Besides, Archer employs Condor to manage all the tasks and resources, which makes it easy to deploy/monitor the tasks and need not worry about the resources. With the newly added feature of NFS interface, we can do more in a customized way. It allows mounting a local virtual disk to the grid, and sharing user-specific files, i.e., large Simics checkpoints. All the grid nodes can access the files shared in the NFS file system. This feature helps build our own simulation environments efficiently.

- Jih-Kwon Peir