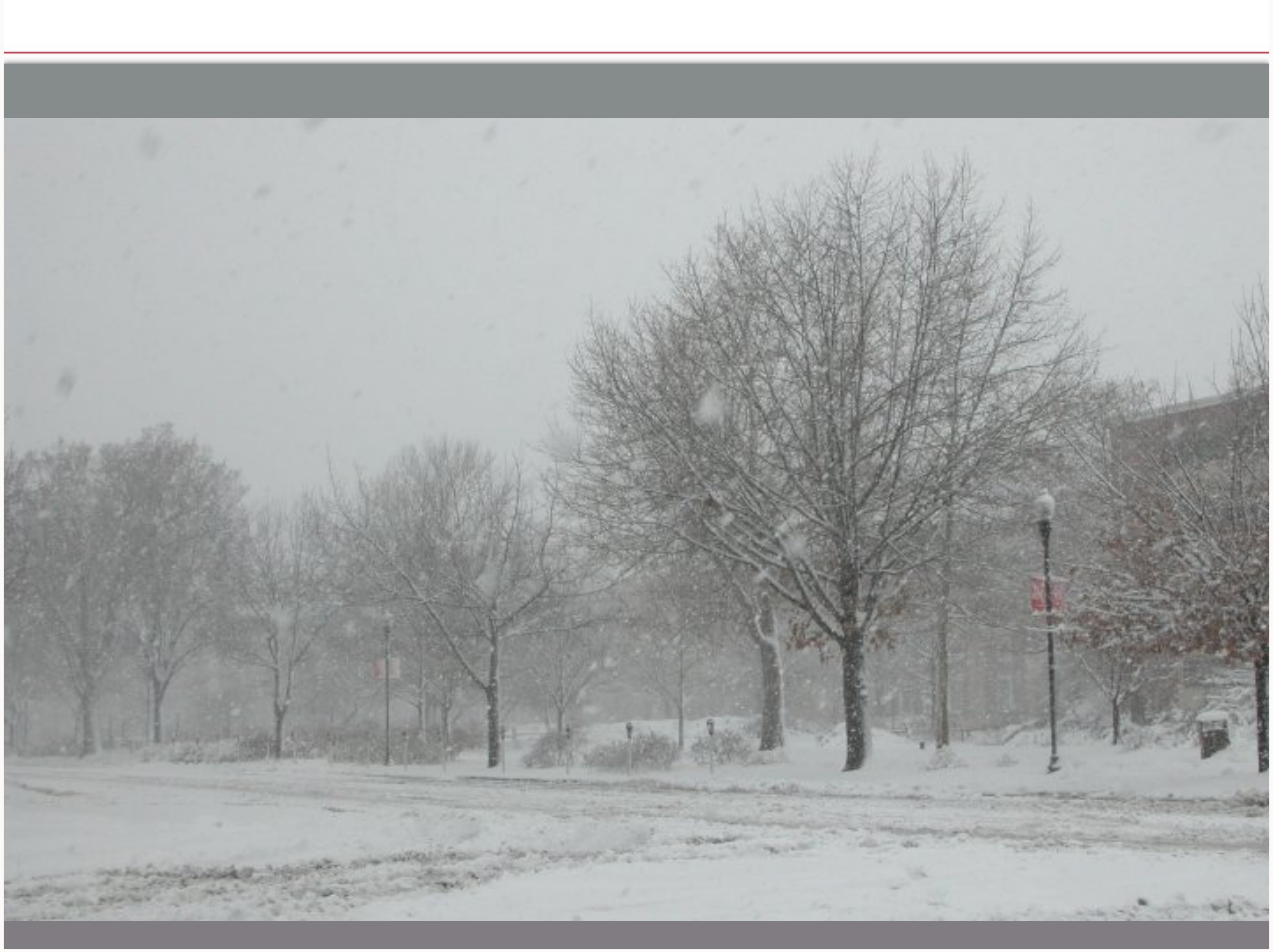
NETWORK LOAD-BALANCING GRIDFTP SERVERS ON THE CHEAP

Garhan Attebury
Holland Computing Center - University of Nebraska-Lincoln



- What’s LVS and why should I care?
- Ohh, that LVS ... so this isn’t anything fancy?
- LVS in a nutshell
- UNL gridftp-hdfs LVS setup (direct routing)
  red-gridftp.unl.edu ... ~12Gbps, high(er) availability than before
- Other uses for LVS at a grid site
  (look Alton, it multitasks!)

... if only we'd stop putting them into the servers ...

- ~1PB usable at many CMS Tier2 sites
  - @ 1Gbps, that’s over 4 months (!)
  - @ 10Gbps, around two weeks
  - (when do we get 40/100GbE again?)

- Once upon a time...
  - ~few hundred TB of dCache storage with ~15-20 gridftp “doors”, one on each dCache pool
  - Someone deletes data, retransferred in ~48 hours. 10GbE is good!
  - Often had failures with pool servers getting burned to the ground: Unbalanced, no failure detection, badness all around

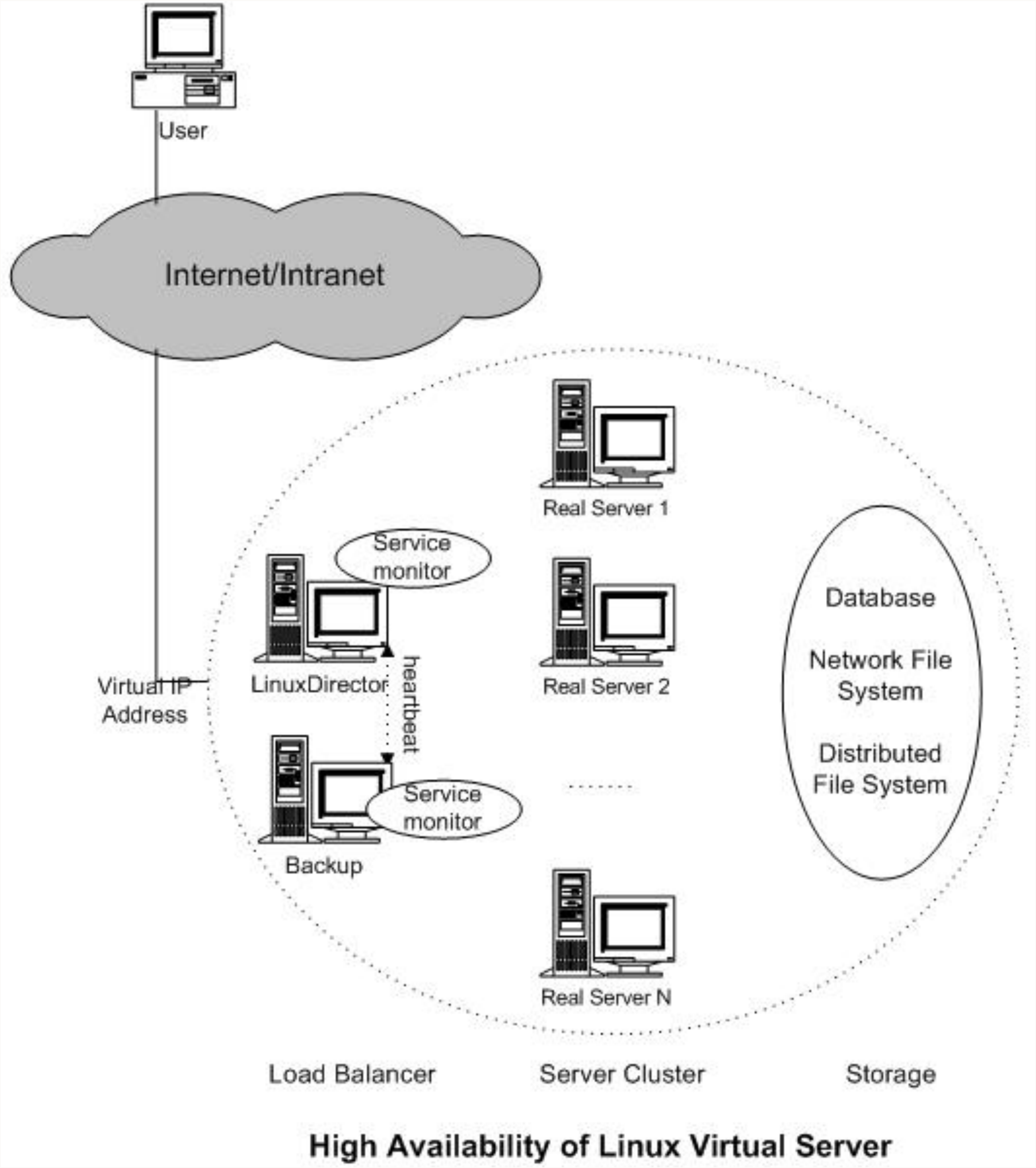
- Things changed a bit ...
    - Initial HDFS deploy with ~400TB
    - 3 dedicated gridftp doors - tried a 10GbE card, but limited to 2.5Gbps due to gridftp-hdfs and hardware limiations
    - Bottlenecked!

- These days ...
    - ~1.2PB usable HDFS storage with 12 dedicated gridftp-hdfs servers and Bestman2 in front
        - Annoying to admin (pre-puppet especially), admin mistakes were common
        - Hard to keep real servers transparent to users, hard to change things around
        - Bestman2 - lot of Java and maintenance for the simple job it does

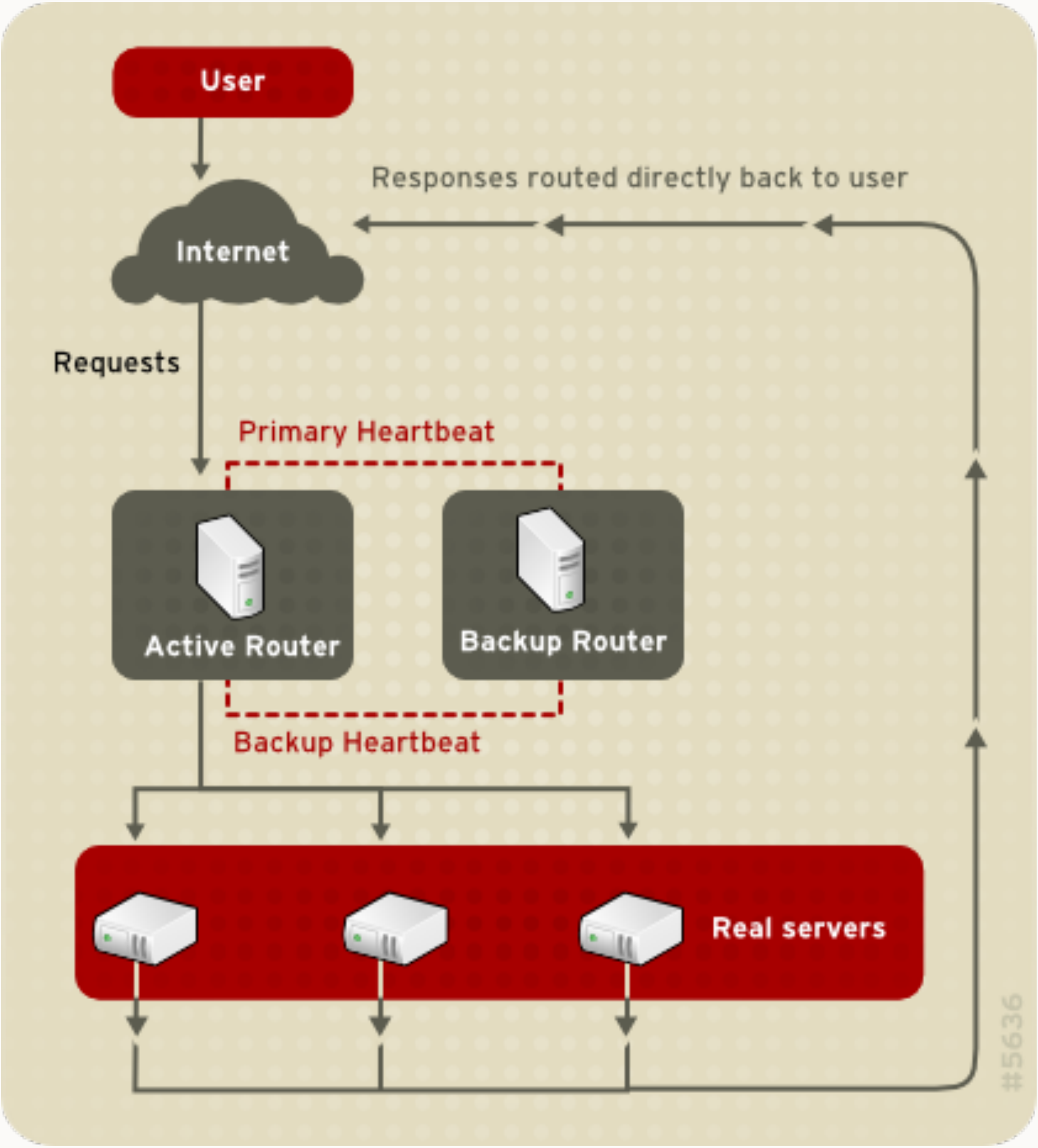
- Simplified with LVS balanced gridftp
  - Single “virtual” address presented to world red-gridftp.unl.edu
  - Automatic detection / removal of failed servers (yes, we could have done this before)
  - Scales linearly-ish (?) + scheduling
  - Removes one reason for SRM

- So this “LVS” ... how do we get and use it?
  - “Official” site
    http://www.linuxvirtualserver.org/
  - RedHat docs aren’t a bad starting point
    (... but avoid Piranha in the end, not needed)
    http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6/html/Load_Balancer_Administration/index.html
  - Using LVS to balance over gridftp servers is nothing new [2005]:
    http://www.austintek.com/LVS/LVS-HOWTO/HOWTO/LVS-HOWTO.performance.html#9.6G
  - Lots of old docs, most still relevant, but overly complex

- What you get with LVS
    - Load balancing (via director)
    - High(er) availability
    - Flexibility

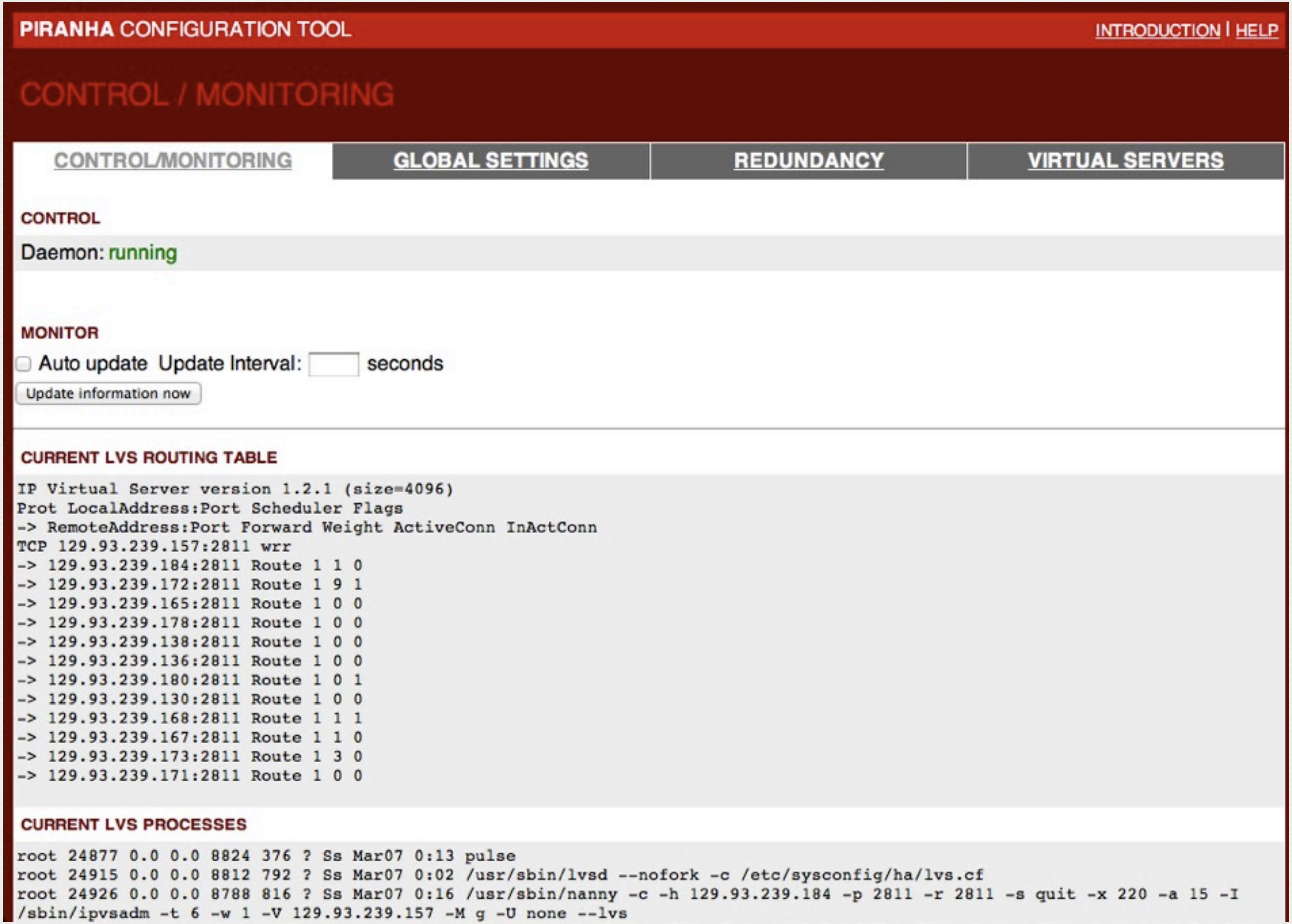

User
Internet/Intranet
Virtual IP Address
LinuxDirector
Service monitor
heartbeat
Backup
Service monitor
Real Server 1
Real Server 2
....
Real Server N
Database
Network File System
Distributed File System
Load Balancer
Server Cluster
Storage
High Availability of Linux Virtual Server

- Direct routing method
- Single layer 2 network, Single GigE interface / box
- ARP problem
- Alternative methods
  - NAT
  - Tunneling



- Setup Overview
  - **Director(s)**
    - /etc/sysconfig/ha/lvs.cf
      Piranha, horrible, but not a bad start
      (hint: you don’t need it!)
    - Services: **pulse** -> **lvsd** -> **nanny**(s)
    - Pulse manages the shared ‘virtual’ IPs setup/teardown
    - High availability, many options - **pulse** “heartbeating daemon” (really?)
  - **Real Servers**
    - IP alias for VIP
    - ARP problem

- On the director:
  - On RHEL, `yum install piranha` will get you all you need
  - Configure lvs.cf (stock file with comments pre-piranha!)
  - Distribute lvs.cf
  - Fire up **pulse** service, stare at /var/log/messages
  - Test failover



CONTROL / MONITORING

CONTROL

Daemon: running

MONITOR

□ Auto update Update Interval: ____ seconds

CURRENT LVS ROUTING TABLE

```
IP Virtual Server version 1.2.1 (size=4096)
Prot LocalAddress:Port Scheduler Flags
-> RemoteAddress:Port Forward Weight ActiveConn InActConn
TCP 129.93.239.157:2811 wrr
-> 129.93.239.184:2811 Route 1 1 0
-> 129.93.239.172:2811 Route 1 9 1
-> 129.93.239.165:2811 Route 1 0 0
-> 129.93.239.178:2811 Route 1 0 0
-> 129.93.239.138:2811 Route 1 0 0
-> 129.93.239.136:2811 Route 1 0 0
-> 129.93.239.180:2811 Route 1 0 1
-> 129.93.239.130:2811 Route 1 0 0
-> 129.93.239.168:2811 Route 1 1 1
-> 129.93.239.167:2811 Route 1 1 0
-> 129.93.239.173:2811 Route 1 3 0
-> 129.93.239.171:2811 Route 1 0 0
```

CURRENT LVS PROCESSES

```
root 24877 0.0 0.0 8824 376 ? Ss Mar07 0:13 pulse
root 24915 0.0 0.0 8812 792 ? Ss Mar07 0:02 /usr/sbin/lvsd --nofork -c /etc/sysconfig/ha/lvs.cf
root 24926 0.0 0.0 8788 816 ? Ss Mar07 0:16 /usr/sbin/nanny -c -h 129.93.239.184 -p 2811 -r 2811 -s quit -x 220 -a 15 -I
/sbin/ipvsadm -t 6 -w 1 -V 129.93.239.157 -M g -U none --lvs
```

- Real servers
  - IP alias for “virtual” address

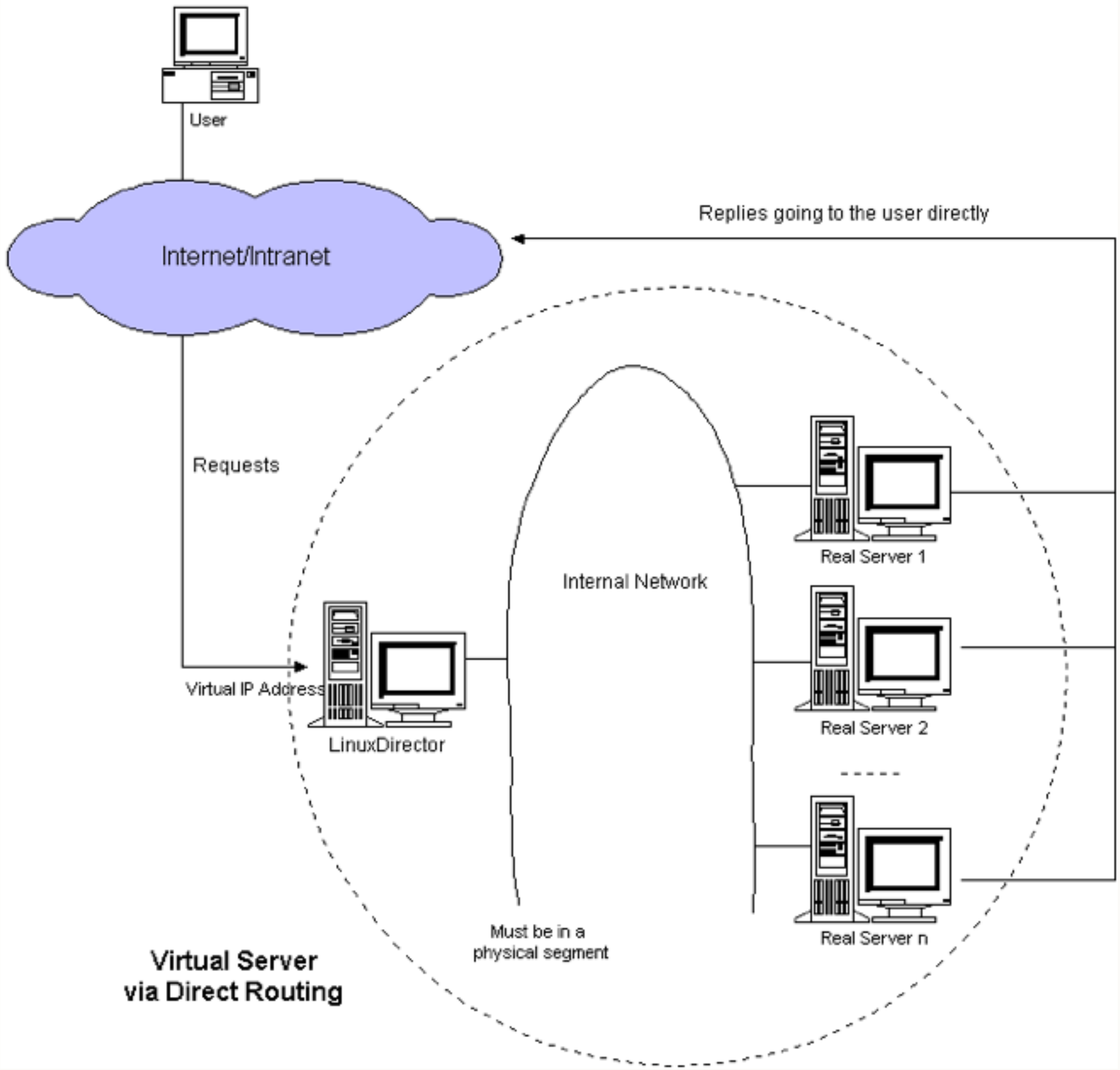
    ```
    ifconfig eth0:1 129.93.239.157 netmask \
    255.255.255.192 broadcast 129.93.239.191 up
    ```

  - ARP

    ```
    yum -y install arptables_jf

    arptables -A IN -d 129.93.239.157 -j DROP
    arptables -A OUT -s 129.93.239.157 -j mangle --mangle-ip-s \
    `ifconfig eth0 | sed -n 's/.*inet \addr:\([0-9.]\+\)\s.*/\1/p'`


    service arptables_jf save
    chkconfig --level 2345 arptables_jf on
    ```


User
Internet/Intranet
Replies going to the user directly
Requests
Internal Network
Real Server 1
Real Server 2
Virtual IP Address
LinuxDirector
Real Server n
Must be in a
physical segment
Virtual Server
via Direct Routing

- Scheduling Algorithms
    - Round Robin
    - Weighted Round Robin
    - Least-Connection
    - Weighted Least-Connection (**default**)
    - Locality-Based Least-Connection
    - Destination Hash
    - Source Hash
    - Shortest Expected Delay
    - Never Queue

- Least connection? Perhaps bad as some transfers/sites are faster than others
- Round Robin? Does what it says
- Weighted RR -- UNL currently uses this, though all servers are equal weights at the moment
- DST/SRC Hashing? Possible to control which servers get traffic from certain sites? Might be useful ...

- Monitoring
  - Trivial “send” and “expect” options
    It’s FTP, expect a 220 on login, and simply send a ‘quit’ to gracefully close connection
  - No way for simple 220 quit check to know if arptables is correct - could use external monitoring script

- Certificates?
  - Some gridftp clients aren’t picky, some do “expected hostname doesn’t match cert...”
  - /etc/sysconfig/globus-gridftp-server

    ```
    export X509_USER_CERT=/etc/grid-security/red-gridftp-hostcert.pem
    export X509_USER_KEY=/etc/grid-security/red-gridftp-hostkey.pem
    ```

- Other things to do with your LVS setup
  - Web caches (CMSSW supports internal RR, OSG does not)
  - SRM? Well, sure, I suppose so.
  - CEs? If only so simple...