Reliable network booting of cluster computers

Matthew Steggink

July 2nd, 2008

Theory

Research question

Test methods

Observations

Alternative booting

Conclusion and future work

Questions

Network booting

- Booting off the network instead of local disk

Network booting

- Booting off the network instead of local disk
- Easily deploy new computers;

Network booting

- Booting off the network instead of local disk
- Easily deploy new computers;
- Centralized image management;

Network booting

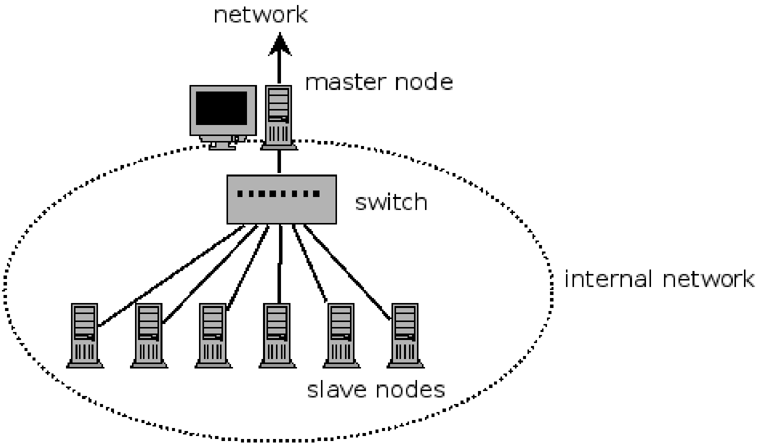
- Booting off the network instead of local disk
- Easily deploy new computers;
- Centralized image management;
- Possibility of diskless computers;

Network booting

- Booting off the network instead of local disk
- Easily deploy new computers;
- Centralized image management;
- Possibility of diskless computers;

- Involves DHCP, ARP and TFTP
- Currently used for network booting: PXELinux

The setup

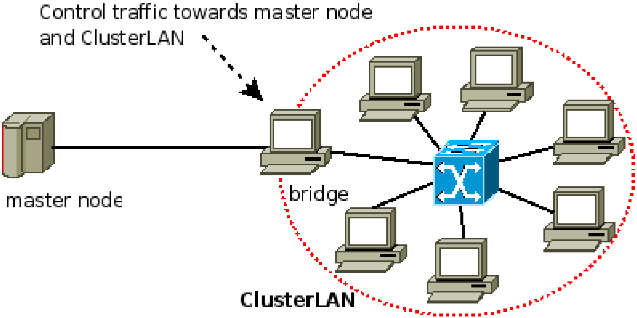


Research question

When booting a large number of clients, some will not complete the boot process

- An analysis of the failing points;
- Determine the cause of the failing clients;
- Search for a solution;

Testing


Control traffic towards master node and ClusterLAN
master node
bridge
ClusterLAN

Shape the traffic

- Limit the traffic to simulate network characteristics
- Two options to shape the traffic

Shape the traffic

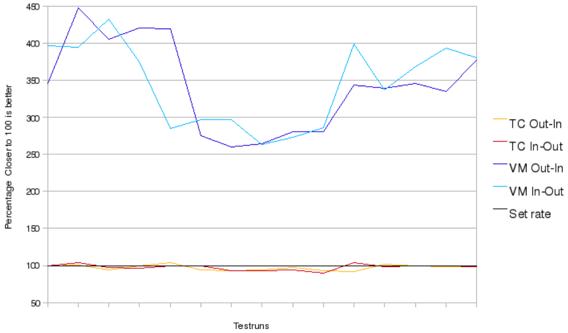
- Limit the traffic to simulate network characteristics
- Two options to shape the traffic
  1. VMWare Teams
  2. Traffic Control in Linux: Token Bucket Filter

Shape the traffic

- Limit the traffic to simulate network characteristics
- Two options to shape the traffic
  1. VMWare Teams
  2. Traffic Control in Linux: Token Bucket Filter
- Limit traffic and set the rates lower to find a failing point

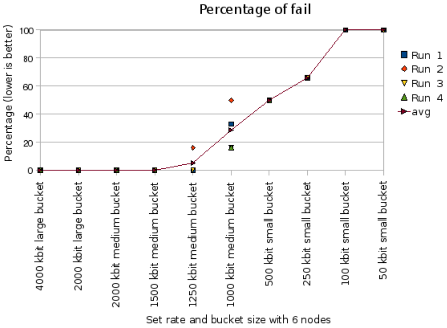
Observations - Traffic control

- VMware teaming does not shape accurately
- TC shapes more reliable



Observations - Fail point

- Too much packet loss and not enough bandwidth



Identified problems

- DHCP
  - No DHCP Offers, No boot file

Identified problems

- DHCP
  - No DHCP Offers, No boot file
- ARP
  - ARP Timeout

Identified problems

- DHCP
  - No DHCP Offers, No boot file
- ARP
  - ARP Timeout
- TFTP
  - TFTP Timeout, Read timeout, illegal operation, server does not support tsize

Identified problems

- DHCP
  - No DHCP Offers, No boot file
- ARP
  - ARP Timeout
- TFTP
  - TFTP Timeout, Read timeout, illegal operation, server does not support tsize
- During downloading (TFTP)
  - `Loading vmlinuz... boot failed`

Booting by TCP / HTTP using gPXE

- gPXE is an open source project
- TCP has delivery reliablity because of re-transmissions with acknowledgments
- Two deployment methods

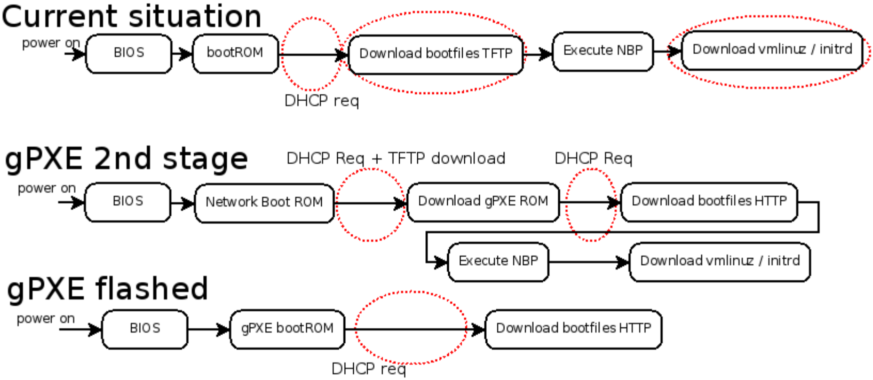
Booting by TCP / HTTP using gPXE

- gPXE is an open source project
- TCP has delivery reliablity because of re-transmissions with acknowledgments
- Two deployment methods
  1. gPXE flashed into the boot ROM
  2. gPXE used as a second stage loader

gPXE results

- gPXE is easy to use, only a few extra lines of code
- No alterations to the clients are needed
- It was compatible with mainstream boot ROM's (Tested: Intel, Broadcom, Nvidia)
- Connections are more reliable; no connections have been aborted during testing
- Disadvantage at this point:
    - Introduces a second DHCP transaction

Situations compared



Conclusion

- gPXE is ready to deploy with only minor alterations;
- The current setup should not use TFTP;
- Connections are more reliable with gPXE and TCP/HTTP;
- Results:
  - DHCP is still the bottleneck
  - TFTP bottlenecks have been solved

Future work

- Take out the second DHCP session
- There might be a better performing DHCP server

Questions

- Matthew Steggink
  `matthew.steggink@os3.nl`