LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Fakultät für Physik – Ludwig Maximilians Universität
Klaus Steinberger • Ralph Simmler • Alexander Thomas

HA Cluster using Open Shared Root

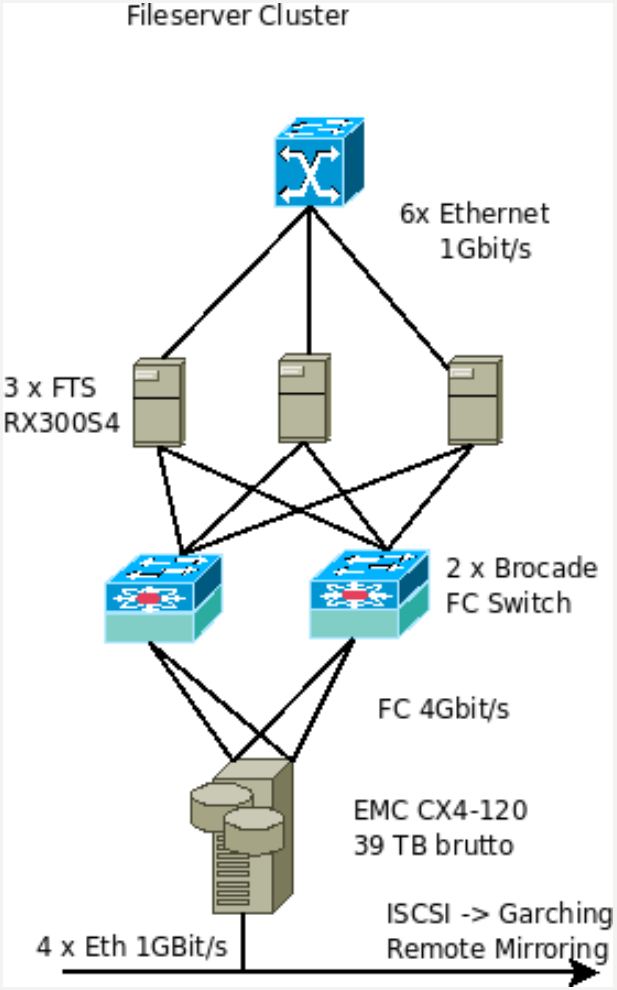
Agenda

- Our Requirements
- The Hardware
- Why Openshared Root
- Technical Details of Openshared Root
- Filesystem and Services
- Conclusion
- Questions

Our Requirements

- Need to deploy a new Storage Server (3k Users)
- Redundant as much as possible and feasible
- Must be Scalable
- Fileservice: NFSV4 and Samba
- Many Services needed besides Fileservice
  - (dhcp, dns, kerberos, Active Directory, Edirectory, Flexlm .....)

- Possible Hardware Solutions:
  - NAS Storage (NetAPP) → very expensive – HA Cluster in front of it needed anyway for non-fileservices
  - ISCSI + Cluster:
    - At start of the decision process only ISCSI solutions with redundancy other 1GBit Ethernet where available → probably too much latency
  - FC San Storage + Cluster
    - We have already some Expertise with SAN + Cluster
    - High performance
      - At least 2 active paths → 2 x 4GBit = 800 Mbyte / sec
    - Further Expansion is affordable.
    - At the end of the decision we got a system with both FC and ISCSI + remote mirroring

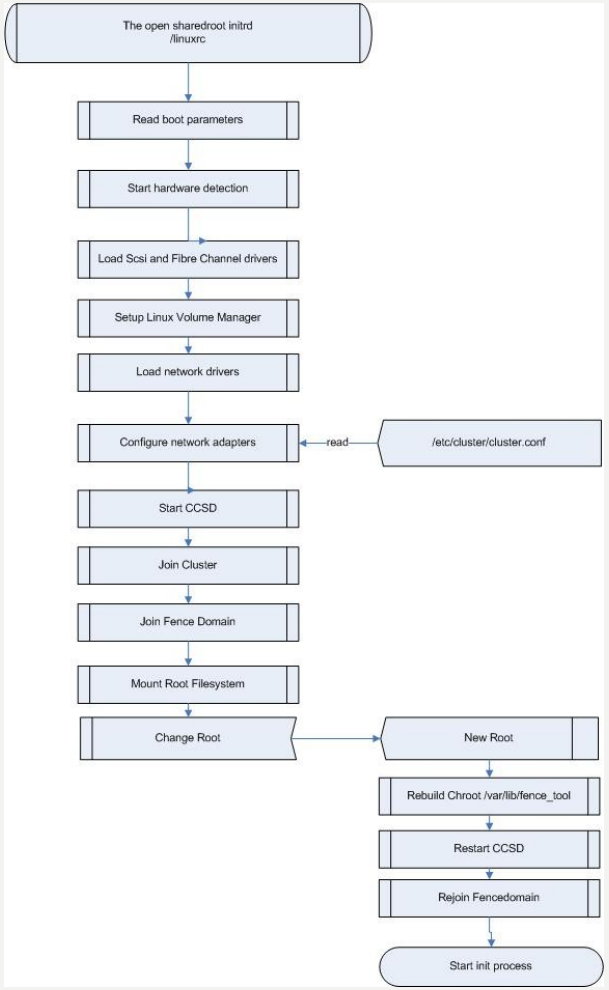


- OS: SL5.3 with Xen
- Server:
  - FTS RX300S4
  - Low Power CPU L5420 @ 2,5 Ghz
  - 32 Gbyte RAM
  - No local disk -> boots from LUN on EMC
- Main EMC CX4-120 with 4 Shelfs
  - 2 Shelfs FC (30 x 0,3 TB = 9TB)
  - 2 Shelfs SATA (30 x 1 TB = 30 TB)
- Secondary EMC at Garching site
  - 1 Shelf 15 Tbyte
  - Mirroring per ISCSI

- Booting from single LUN → Easy Mirroring
- Single System Image (just disk of course)
  - Always same software versions on all nodes
  - One point of administration (hopefully)
- Simple Expansion with more nodes
- Hopefully pays us back ease of administration
- A Test Installation of CentOS based Installation DVD went well
- Last but not least: It's GPL

- Initrd does the hard job
  - Hardware detection
  - Setup LVM
  - Configure network
  - Start Cluster
  - Mount root filesystem (GFS)
  - Change Root
  - Restart some of the cluster services
- Go on with normal startup

(Image taken from www.opensharedroot.org)



- Boot Configuration inside cluster.conf
- <com_info> Block - Essentials for node
- Define GFS Root
  - <rootvolume name="/dev/vg_democluster/lv_sharedroot" fstype="gfs" />
- Define Network Devices (at least cluster net)
  - <eth name="bond0" ip="10.1.2.3" mask="255.255.255.0" />
  - <eth name="eth0" mac="11:22:33:44:55:66" master="bond0" slave="yes" />
  - ....
- Some more
  - Syslogserver, scsi failover, fenceackserver

- Some solution for Node Dependent Data needed

- Context Dependent Symbolic Links
  - com-mkcdslinfrastructure - creates the CDSL infrastructure
  - com-mkcdsl - creates CDSL Links
  - com-rmcdsl - removes CDSL Links
  - com-cdsinvchk - checks CDSL Inventory

- How they are done
  - Would be nice to have CDSL Links in the kernel, but patches are not in mainstream.
  - Just for node dependency CDSL could be emulated by bind mount

- /cluster/cdsl/{nodeid} – node specific files
- Bind mount: /cluster/cdsl/{nodeid} → /cdsl.local
- /var → /cdsl.local/var
- Node independent Files are symlinks back
- /var/lib → /cluster/shared/var/lib
- Inventory of CDSL Links in a XML file

- Watch out cron jobs
  - Especially don't run yum-update and rpm Watcher in parallel on nodes
  - Solution: Use a Script to get locks for cron Jobs
- Watch out for changes on node dependent files after updates and installations
  - Check with com-cdslinvchk
- Linux is quite happy without Swap, but ....
  - Had to add swap to avoid Fencing wars caused by OOM Killer
- Some services suffer from GFS locks → tmpfs helps
  - Xenstored (in 5.3 already uses tmpfs)
  - Samba (tdb's are painfully slow with GFS locks) → ctdb ?
  - Raising plock_rate_limit may help too

- Candidates (both cluster and single node)
  - GPFS → proprietary
  - ZFS → License incompatible with GPL → under Linux not production grade
  - ReiserFS → unclear status (gone into jail ;-)))
  - XFS → too many people complain, support under SL ??
  - Ext3 → reliable and fast
  - Ext4 → very promising but too early?
  - GFS1 → meanwhile reliable, good support under SL
    - Very fast on large files, slow on little ones
  - GFS2 → now supported by TUV but (currently) slower than GFS1?
- Strategic decision: A single VM with ext3/4 or a GFS as a cluster filesystem?
- Scalability was a main point → GFS1

- Isolation of physical Cluster → XEN
- Use rgmanager for failover of VM's
- Fileservices inside a virtual OSR Cluster
  - Three node cluster spread over physical hardware
  - Runs both NFSV4 and Samba
- Run any other Service in dedicated VM's
  - Linux → Just easy
  - Windows (Active Directory, Terminal Server)
    - HVM Network/Disk Performance very slow
    - GPLPV drivers help much (around 600 Mbit/s)
    - Trouble with Windows 2008 → Driver Signing Policy
    - But Hackers will help: ReadyDriverPlus

- Physical Cluster running well since two months
- Virtual Fileserver Cluster running in test
- Kerberos running and provisioned
  - But currently we suffer from a licensing issue with Novell IDM driver
- Active Directory provisioned with user data
- License Server VM → Production (origin + comsol)
- We are late on schedule → originally: 4/09 now 8/09
  - Some learning effects with OSR
  - Some troubles with EMC CX4 (Mainly setup issues: Multipath, Mirroring)
  - More troubles with Active Directory (ever tried to join over firewalls?)
  - Complete changeover to Kerberos and AD needs more time ....
  - Last but not least: Too many work for the number of admins :-((

- Our HPC Cluster (around 80 nodes) currently runs an older Debian version → has to be upgraded
- New chairs got a promise for HPC Clusters → 100-200 new nodes ?
- Discussion about OSR over NFS for cluster nodes.
  - Local disk will be used for swap and short time data.
  - Would ease the maintenance of the cluster
  - Performance issues?

- For an 3 node Cluster we think the assets and drawbacks for OSR are very close
- But expansion is easy → OSR pays back
- Gained experience with OSR → plans for HPC Cluster

More about Open Shared Root:

http://www.opensharedroot.org/

- Questions ?