pro|business group

samba eXperience 2008

The Simple High Available Linux File Server

Schlomo Schapiro
Principal Consultant
Leitung Virtualisierung & Open Source

18.04.2008

Agenda

- Background Information
- Possible Solutions
- The Simple High Available Linux File Server
- Benefits

The Customer Case

- Landesvermessung und Geobasisinformation Brandenburg
  - State office for geo-information and survey
  - Collecting, storing and developing data relating to location on the surface of the earth.
  - Responsible for the production and publication of the official maps and for keeping the official land register of the Federal State of Brandenburg.
  - Located in Potsdam
  - Heavily IT dependant
  - HP main hardware vendor
    - 3 EVA (~40 TB)
    - > 100 servers

The Problem Situation

- Many systems store data on many file servers and locally
- Many complex work flows with lots of FTP and file copy between Unix/Linux servers and Windows systems
- Sometimes large data sets, performance issues
- Local user accounts on many systems (Unix/Linux)

- Solution: Storage Consolidation
- Ongoing project

Storage Consolidation Goals

- Unified storage for Unix/Linux and Windows
- Centralized storage
- High Availability & Disaster Recovery
- Good protocol support
- CIFS:
  Windows ACL, Access Based Enumeration, DFS Replication, AD Integration
- NFS:
  Version 2 and 3, Posix ACL, AD Integration with RFC2307
- FTP, RSYNC, SCP

Decision Process & Criteria

- Economical & technical arguments
- Important criteria
  - Investment & operating costs for 3 years (till 2010)
  - Availability & reliability (solutions without storage use EVA 8k)
  - „Quality“ of CIFS and NFS implementation, interoperability
  - Ease of use, simple management
  - Efficient backup with CommVault Galaxy, preferably LAN-free
- Optional criteria
  - Support asynchronous protocols (RSYNC, FTP …)
  - Integrated replication backup to secondary independent storage
  - Strengthen future storage strategy

Agenda

- Background Information
- Possible Solutions
- The Simple High Available Linux File Server
- Benefits

Solutions

- Windows or Linux file server
- NetApp FAS3040
- IBM SOFS (Scale Out File Services)

IBM SOFS

- Good
  - Highly redundant (many redundant servers)
  - Easy management (Web GUI)
  - ILM solution integrated
  - Acceptable cost
- Bad
  - IBM solution, needs new storage, no combination with HP
  - Backup with TSM or via NFS/CIFS
  - Storage strategy needs to be changed to IBM
  - No internal replication to secondary storage
  - Incomplete CIFS support (Samba)

NetApp FAS3040

- Good
  - Perfect CIFS implementation
  - Easy management (Web GUI)
  - Very redundant (RAID-DP, 2 servers, NVRAM ...)
  - Backup via NDMP (LAN-free)
  - Good replication (only to other FAS)
  - Innovative solutions with multiple FAS (replication & availability)
- Bad
  - NFS ACLs only with NFSv4
  - costly
  - Long-term storage strategy should be changed to NetApp

Windows File Server

- Good
  - Perfect CIFS implementation
  - DFS and replication to other Windows file servers
  - Backup with LAN-free agent
- Bad
  - Very bad NFS support (slow, ACLs)
  - Asynchronous protocols (FTP, SCP, RSYNC ...)
  - Would you trust Windows with 30TB ?

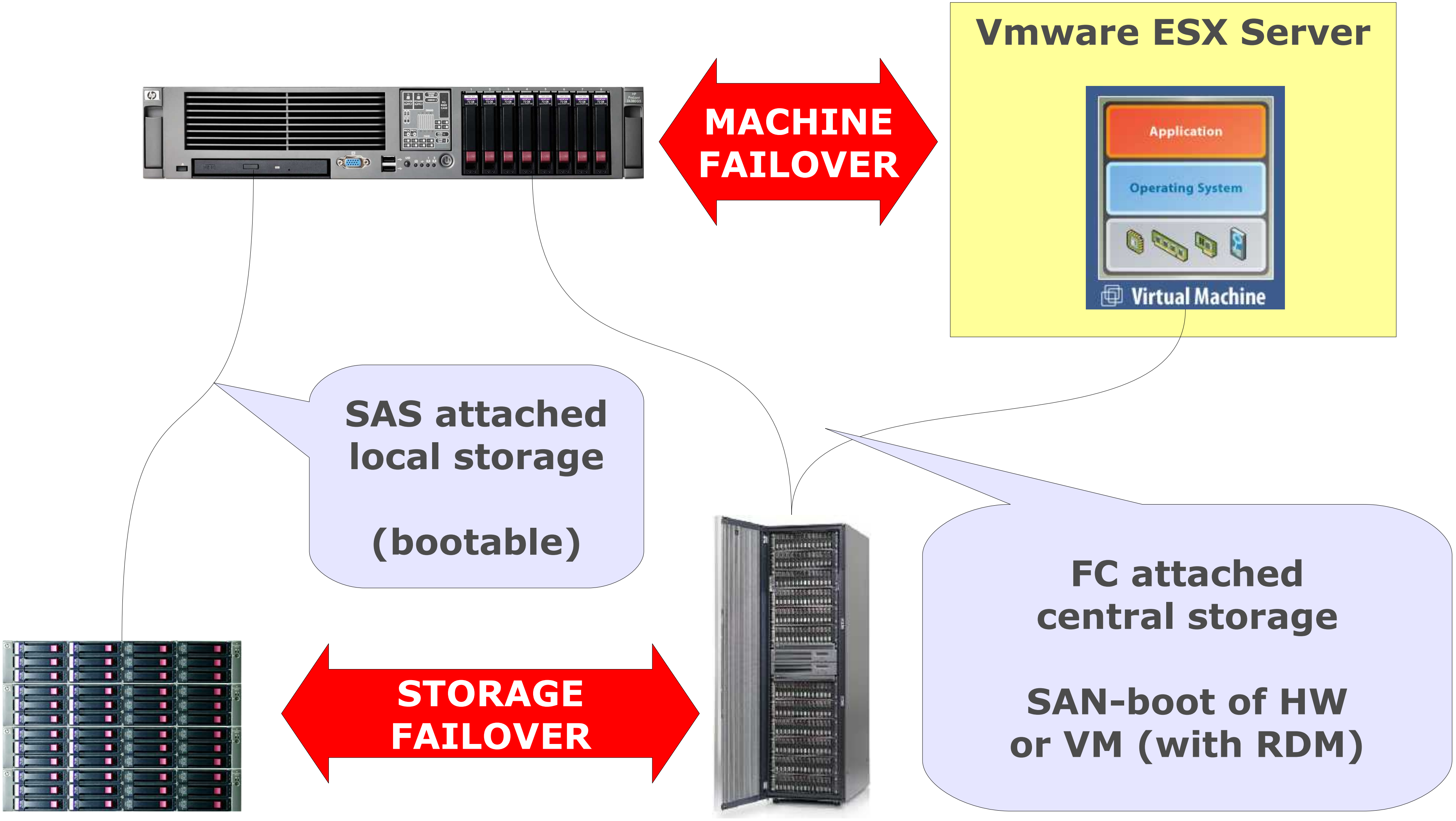
Linux File Server

- Good
  - Very redundant
  - NFS and CIFS ACL integration
  - Richest protocol support
  - Backup with LAN-free agent
  - Integrated replication to secondary storage
  - No implication for storage strategy
  - Linux know-how can be used
- Bad
  - Incomplete CIFS support (Samba)
  - Shell-level management

Agenda

- **Background Information**
- **Possible Solutions**
- **The Simple High Available Linux File Server**
- **Benefits**

The Simple High Available Linux File Server



Vmware ESX Server

MACHINE FAILOVER

SAS attached local storage

(bootable)

STORAGE FAILOVER

FC attached central storage

SAN-boot of HW or VM (with RDM)

Components

- Hardware (all from HP)
  - DL 380 G5
  - EVA 8000 (30TB)
  - 4x MSA60 (36TB)
  - several DL servers as ESX server (used for data center)
- Software
  - SuSE Enterprise Linux 10 SP1
  - Samba, NFS
  - LVM, LVM Snapshots (smbsnap)
  - rsync
  - Virtual machine on VMware ESX 3 (normally switched off)

Challenges

- SAN boot with multipathing (DM-MPIO)
  - Possible with SLES10SP1
  - dm-multipath already in initrd
  - installation via VM
  - See paper in i'X 04/2008 p. 142
  - Dual boot hardware and virtual machine (drivers ...)
  - Prevent accidental boot of virtual machine (ISO image)
- Manage local storage
  - Automated cloning of production system to local storage
  - Modifications to boot from local storage (RAID-1) and mount local storage in place of SAN storage
  - Nightly rsync of all data from SAN to local storage

Benefits

- 2 dimensions of redundancy
  - Hardware and virtual machine run the same system & data – fail over without data loss
  - SAN storage replicated with rsync to local storage
- Recovery times ~ 5 min for hardware or storage failure
- Instant disaster recovery – even with many TB of data
- Very simple system – no complex cluster configuration
- Fail-over: Reboot HW (storage) or boot VM (hardware)
- Administrator carries full responsibility
- Very affordable solution – no extra costs for HA

System & Samba Setup

- Local storage
  - GPT (>2TB)
  - System on RAID-1 (MD)
  - LILO (GPT, MD)
- Everything via LVM
- rsync with sanity checks
- AD Integration RFC2307
- Volume Shadow Copy
- Map BUILTIN Accounts

```
passdb backend = tdbsam
smb ports = 445
disable netbios = Yes
name resolve order = wins
inherit acls = Yes
hide unreadable = Yes
idmap backend = ad
idmap uid = 100-20000000
idmap gid = 100-20000000
winbind enum users = Yes
winbind enum groups = Yes
winbind use default domain = Yes
winbind nss info = rfc2307
use sendfile = yes
```

Performance Tuning

```
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
net.ipv4.tcp_no_metrics_save = 1
net.ipv4.tcp_moderate_rcvbuf = 1
net.core.netdev_max_backlog = 2500
```

- Benchmarks (1GBit):
  - 125 MB/s (NFS)
  - 100 MB/s (CIFS) (>2 streams)
- Use XFS (works well on SLES)
- sysctl.conf (also on client)
- USE_KERNEL_NFSD_NUMBER="16"
- Bonding for redundant network connection
- Jumbo Frames had no measurable effect on throughput, CPU usage reduced by ~50%

Outlook

- Automated fail-over:
  - Heartbeat in initrd before mounting /
  - Monitor storage and network – difficult decisions
- MD or LVM mirroring between SAN and local storage (No true disaster recovery !) as an alternative to rsync replication
- Multipathing SAN boot with RHEL/CentOS, Ubuntu ...
- „Virtual Cold-Standby Server“ can be used for other systems

- This is mostly an idea and way of thinking
- Send me an email for implementation details

Questions & Answers

More Open Source Software (schapiro.org/schlomo/projects)

- Relax & Recover (Linux Disaster Recovery)
- RSYNC BACKUP MADE EASY (Backup Software with Hardlinks)
- OpenVPN Gateway Builder (Linux Routers w/ central management)
- easyVCB (VMware VI3 Backup, w.i.p.)

Schlomo Schapiro
Principal Consultant
Leitung Virtualisierung und Open Source

sschapiro@probusiness.de
+49 160 97846168

probusiness Berlin AG
Potsdamer Platz 11
D-10785 Berlin

berlin@probusiness.de
+49 30 259378 0