DATA STORAGE

R. Kapeller, C. Higgs

Using Linux on the SLS data storage servers has proven to be a stable, well-performing and budget-friendly decision. No major outage has occurred since machine operations started. Security and privacy of data, however, must be improved in the future. Emergency recovery procedures also have to be defined for the server infrastructure.

THE NEED FOR SERVERS

File servers at SLS are used for permanent data storage (control system) and for temporary data storage (data acquisition from beamline experiments). The total amount of accumulated data can reach several hundred Gbytes per day and experiment. Each beamline has a dedicated file server with up to 1 Tbyte of data storage.

WHY LINUX?

Several products from commercial storage manufacturers were evaluated prior to the SLS commissioning phase. Turn-key systems (e.g. Network Appliance) seemed to fulfill our requirements, but by far exceeded our budget. Our high demands in terms of storage and computer availability are similar to those of high-end commercial institutions. However our budget could not support these solutions.

As large-capacity disks became ever cheaper, even personal computers could be (mis) used for high volume data storage. Certainly the data produced at the SLS is too precious to be simply saved locally on each PC, without any safety mechanism.

At this time, we had already decided to use Linux for the SLS operator consoles, for the IOC boot computers and on the development workstations. The idea to use Linux as a server platform was also discussed, but regarded as too know-how hungry.

Independently AIT (the PSI Computing Department), decided to use Linux on Dell servers for high volume data storage. Thus the decision to use Linux at the SLS became much easier. Our selected operating system became Red Hat Linux 6.2.

FLEXIBLE STORAGE MANAGEMENT

At this early stage of the SLS history, it was very difficult to predict storage requirements and therefore a flexible solution was imperative. Red Hat was not yet bundling storage management, but the LVM (Logical Volume Manager) was already available for Linux. With valuable know-how from AIT, a suitable installation was compiled for the Dell server systems.

This standard server installation consisted of:

- Red Hat Linux 6.2 with the linux-kernel-2.4.17 (patch: aacraid, lvm)
- with Red Hat 6.2 Updates (Red Hat Errata)
- and extra software (LVM utils, logcheck, sar)

The Dell platform consists of PowerEdge servers and PowerVault disk arrays. Each array takes up to 14 disks each of 72 Gb and connects via dual-channel SCSI to the server. The SCSI controller on the server allows grouping of disks (container) to any common RAID (Redundant Array of Independent Disks) level. Disk containers are then presented to Linux as block devices.

NFS/SMB Exports	X05DA		X05DB
Linux file system	/dev/vg0/lv0		/dev/vg0/lv1
Linux Logical Volume Manager	Lv0 500GB		Lv1 200GB
	Physical Volume 'vg0'		
Linux SCSI	/dev/sda1		/dev/sdb1
RAID Manager	Container 'A'		Container 'B'
Dell Disk Array	Unit 'A' 350GB		Unit 'B' 350GB

Fig. 1: The Logical Volume Manager allows a flexible storage management, by abstracting the physical devices.

These block devices are used by the LVM to combine, split and present their own abstraction of block devices, which can then be formatted with Linux file systems (Fig. 1). This configuration allows easy resizing of volumes without the need to reformat, and new disk arrays can be quickly added to the existing volumes.

The extra LVM layer does not derogate performance. The real bottleneck in accessing data still remains with NFS (Network File System) (Fig. 2).

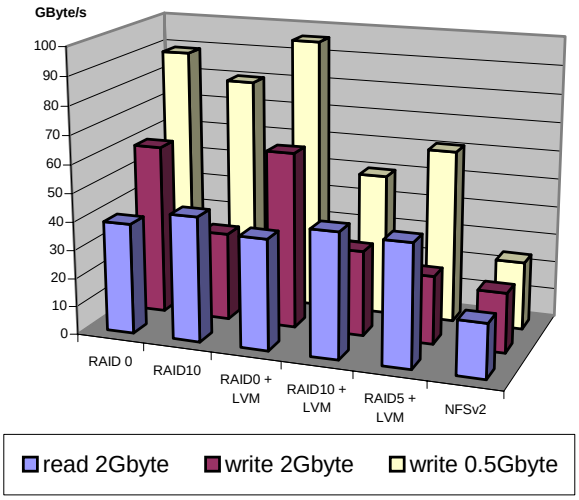

Fig. 2: NFS limits I/O performance, while reading/writing large files of 2 Gbyte resp. 500 Mbyte.

AVAILABILITY

As of December 2002, six Linux file servers (incl. stand-by) are in production. None of these servers has had a system-crash during the past two years. Shutdowns were planned and were normally due to system reconfiguration, upgrades or electrical power maintenance. The average "uptime" of all the SLS file servers is still better than 99.95 %.

A remaining and nagging problem still troubles the client computers after a server's NFS export tables have been modified. This problem manifests itself on the client side with the message *NFS: stale file handle*. However we rarely need to do this and when required, it usually occurs during scheduled SLS machine shutdown. The problem should be solved with a newer Linux release.

SECURITY & FLEXIBILITY

These two characteristics often present incompatibilities. Finding the optimal balance is not always easy. Resource leakages, vague guidelines, scientific spontaneity and technical limits are often hurdles in implementing practical security and effective privacy in an existing environment. Some data needs to be shared amongst many users and accidental data loss may not be fully excluded.

Automatic data backup is performed on the SLS machine system software. Data files generated at the beamlines are the responsibility of each beamline user. A tape library can be used for medium-term storage for the latter. Alternatively beamline users can copy the data to their own disks using our media stations (Windows XT; USB or Firewire disks) provided at each beamline.

FUTURE GOALS

- As SAN/NAS (Storage Area Network/ Network Attached Storage) technology permeates the storage market, we will study how this technology can be applied to the SLS storage requirements.
- As Red Hat no longer supports and delivers security and bug fixes for their 6.2 release, a major software upgrade appears likely during the second half of 2003.
- Defining and implementing a documentation framework for the SLS server infrastructure.
- Planning, implementing and documenting an emergency recovery scenario for every server system. (Perhaps fail-over systems.)
- Review the security concept for the SLS server infrastructure.
- Average uptime for all SLS servers must remain above 99.95 %

REFERENCE

[1] http://www.redhat.com

[2] http://www.sistina.com/products_lvm.htm

[3] S. Hunt et al., *The Beamline Data Acquisition and Control System*, PSI Scientific Report 2001, VII.