Adding a disk

SA's important tasks

- Connecting and configuring new storage devices
- Making local and remote files available to users
- Monitoring and managing the system's finite disk resources
- Checking for and correcting file system corruption
- Protecting against file corruption, hardware failures, and user errors via a well-planned backup schedule

Disk interfaces

- Interface standards
  - SCSI (The Small Computer Systems Interface)
    - most command and widely supported disk interfaces on servers
  - IDE (Integrated Drive Electronics)
    - simple, low-cost interface for PCs.
  - Fiber Channel
    - high bandwidth and large number of devices
  - USB (The Universal Serial Bus)
    - Has enough bandwidth to support slower disk devices

SCSI interface

- SCSI defines a generic data pipe that can be used by all kinds of peripherals
  - Disks, tape drivers, printers, scanners
- SCSI standard revisions
  - SCSI-1 in 1986
  - SCSI-2 in 1990
    - Add new performance features:
      - Command queuing
      - Scatter-gather I/O
  - SCSI-3
    - Add enhancement
      - Device auto configuration

SCSI interface (Cont)

- Some terms
  - Fast
    - The bus speed is doubled
  - Wide
    - Number of bits is larger
  - Single-ended devices
    - Every other pin is grounded to help reduce crosstalk
  - Differential SCSI
    - Use inverted signal next to each pin in stead of a ground
  - Cable length
    - Due to the signal crosstalk and number of devices, cable length is limited

STA Terms (notes-see below)	Bus Speed, MBytes/Sec. Max.	Bus Width, bits	Max. Bus Lengths, Meters (1)			Max. Device Support
			Single- ended	LVD	HVD	
SCSI-1 (2)	5	8	6	(3)	25	8
Fast SCSI (2)	10	8	3	(3)	25	8
Fast Wide SCSI	20	16	3	(3)	25	16
Ultra SCSI (2)	20	8	1.5	(3)	25	8
Ultra SCSI (2)	20	8	3	-	-	4
Wide Ultra SCSI	40	16	-	(3)	25	16
Wide Ultra SCSI	40	16	1.5	-	-	8
Wide Ultra SCSI	40	16	3	-	-	4
Ultra2 SCSI (2,4)	40	8	(4)	12	25	8
Wide Ultra2 SCSI (4)	80	16	(4)	12	25	16
Ultra3 SCSI or Ultra160 SCSI (4)	160	16	(4)	12	(5)	16
Ultra320 SCSI (6)	320	16	(4)	12	(5)	16

Notes:
The listed maximum bus lengths may be exceeded in Point-to-Point and engineered applications.
Use of the word "Narrow", preceding SCSI, Ultra SCSI, or Ultra2 SCSI is optional.
LVD was not defined in the original SCSI standards for this speed. If all devices on the bus support LVD, then 12-meters operation is possible at this speed. However, if any device on the bus is singled-ended only, then the entire bus switches to single-ended mode and the distances in the single-ended column apply.
Single-ended is not defined for speeds beyond Ultra.
HVD (Differential) is not defined for speeds beyond Ultra2.
After Ultra2 all new speeds are wide only.

Source: http://www.ramelectronics.net/scsi_connectors.ep#scsitypes

SCSI connectors

- Connector Types:
  - Narrow SCSI
    - 50 pins
  - Wide SCSI
    - 68 pins
  - Internal device:
    - 50pin or 68 mini-micro connector attached to a ribbon cable
  - External device
    - high density 50 pin or 68 pin
  - Single Connector Attachment plug: 80 pins that includes bus, power, and all the drives' needs, useful for hot swappable drive arrays.

SCSI devices

- SCSI uses daisy chain configuration,
  - so most external devices have two SCSI ports.
  - Internal SCSI devices are attached to ribbon cable where connectors can be clamped onto the middle.
- Each end of SCSI bus must be terminated
  - Terminator
  - Auto terminating
- Address
  - Each device has a target number that distinguish it from other devices on the bus
    - SCSI controller itself counts as a device
  - Address number is from 0-7 or 0-15
    - Address is arbitrary generally speaking
      - Set the SCSI ID if available on the external device
  - Logical unit number (LUN)
    - Each target can have several logical unit inside it
      - Disk array connected to one SCSI controller

SCSI troubleshooting

- Things can go wrong, keep the following in mind
  - Check no mixed differential and single-ended devices in the chain
  - Check OS discover the new device fine
  - Check terminator on both ends
  - Check the length of cable, including the internal part
  - Never forget that your SCSI controller uses one of the SCSI addresses.

IDE

- The controller is build into the disk, reduce the interface costs and simplifies the firmware
  - Also called ATA
- Standards
  - ATA-2
    - Faster programmed I/O
    - Direct Memory Access modes
    - Better bus's plug and play features
    - Logical Block Addressing (LBA)
      - Over come the problem that BIOS can only access the first 1024 cylinders of a disk
  - ATA-3
    - Additional reliability
    - more sophisticated power management
    - self monitoring
  - ATA-4
  - Ultra-ATA
    - Extend the bus bandwidth

IDE disks

- IDE disks are used internally
  - Short cable length
    - Ata-2 bus: 18 inches
  - One bus only supports two IDE devices
    - Manufactures provide more than one IDE bus on the motherboards
  - IDE devices are accessed in a connected manner
    - Only one device can be active
    - Spread the device over multiple buses
    - Put fast devices on one bus and slower devices on other buses

IDE devices

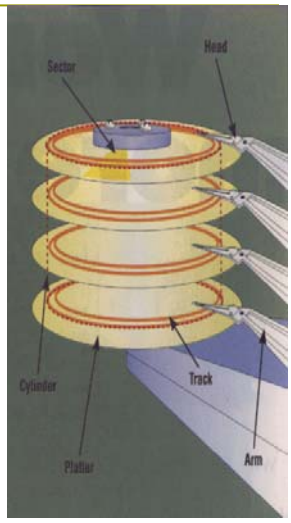
- IDE connector: 40 pin header
- If there are more than one device on an IDE bus
  - Designate one as the master and the other as slave
  - Some IDE devices do not like to be slaves

Which is better, SCSI or IDE

- SCSI beats IDE in every possible technical sense
  - For best possible performance
  - Server and multiuser systems
  - Connecting many devices
  - Particular features of SCSI
    - For example: Hot-pluggable devices
- IDE is cheap and works well for single-user workstation

Disk Geometry

- A stack of platters
  - Coated with a magnetic film on both sides
    - Data can be writing on both side of the platters
  - Rotate at a constant speed
    - 5400RPM – 15000RPM or more
- Heads
  - Move back and forth, floating very close to the surface of the platters
    - Seeking time
    - Disk diameter gets smaller
      - 14 inches 10 years ago, 5 ¼ inches 10 year ago, to 3 ½ inches and smaller …
- Track
- Sector
- Cylinder



The disk installation procedure

- The procedure of adding a disk involves:
  - Connecting the disk to computer
  - Creating device files
  - Formatting the disk
  - Labeling and partitioning the disk
  - Establishing logical volumes
  - Creating Unix filesystems within disk partition
  - Setting up automatic mounting
  - Setting up swapping on swap partitions.

Connecting the disk

- Depends on the interface
  - For IDE
    - Try to configure the system with only one IDE disk per bus
    - Check master/slave settings on each disk
  - For SCSI
    - Terminate both ends of the bus
    - Check cable length
    - Target number does not conflicts

Creating device files

- Devices are presented as *special files* in /dev
  - Devices are either *block* or *character* special files
- Devices are created with mknod just as directories are created with mkdir
  - Devices have *major* and *minor* number
    - The *major* number represents the *device driver*
    - The *minor* number represents the *instance of a device* of the type specified by the major device number

```
brw-r----- 1 root disk 8, 0 2009-08-28 13:29 sda
brw-r----- 1 root disk 8, 1 2009-08-28 13:29 sda1
brw-r----- 1 root disk 8, 2 2009-08-28 13:29 sda2
brw-r----- 1 root disk 8, 3 2009-08-28 13:29 sda3
brw-r----- 1 root disk 8, 4 2009-08-28 13:29 sda4
```

- Many UNIX automatically create all possible device fiels
  - Tru64 5.0 “hwmgr -scan scsi”
  - SunOS 5.x creates devices at boot time; use boot -r

Formatting the disk

- The formatting process
  - Writes address information and timing marks on the platters to delineate each sector
  - Identify “bad blocks”
- All hard disks come preformatted
  - Formatting the disk in the field can take long hours

Partitioning the disk

- Disk must be divided into chunks called partitions or slices
  - Partitions are independent data areas
    - But in reality
      - All system define one partition to be the entire disk.
      - Partition can be defined with overlaps
  - Device driver knows about the disk layout
- Partition Table
  - Kept on disk in a record called the label.
    - The label usually occupies the first few blocks of the disk.

Partitioning the disk (cont)

- Partitions can vary, but at least you will have:
  - Root partition
    - Everything needs to bring the system to single user mode
  - Swap partition
    - Pages of virtual meory
  - User partition
    - Home dirs, data files, etc.

Partitioning the disk (cont)

- Some other considerations
  - Mirror the root filesystem on another disk partition
  - Add swap space if add memory to allow kernel crash dump
  - Split swap among several disks
  - Split file systems among several disks
  - Don't make partition bigger than the capacity of your backup device
  - Create a separate file system /tmp
  - Create a separate /var

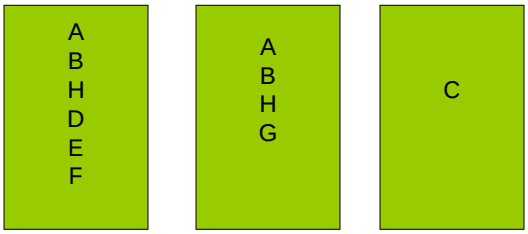
Partitioning the disk (cont)

- Commands for partitioning/formatting drives
  - SunOS 4.x/5.x – format
  - Linux - fdisk
- Commands for examining partition information
  - SunOS 5.x – prtvtoc
  - Linux – fdisk

Example: BSD partitioning

- BSD systems usually use 8 partitions

Partition	Cylinders	Typical use
a	0-15	/
b	16-86	Swap
c	0-841	Entire disc
d	391-407	Alternate Root
e	408-727	
f	728-841	
g	391-841	/usr
h	87-390	



Creating UNIX filesystems

- After partitioning, you can create a filesystem
  - Use command **newfs**
- Most UNIX file system consists of five structural components:
  - A set of inodes storage cells
  - A set of scattered “superblocks”
  - A map of the disk blocks in the file system
  - A block usage summary
  - A set of data blocks

Creating UNIX filesystems (cont)

- Superblock
  - Is a record that describes the characteristics of the file system.
  - Contains:
    - Information about the length of a disk block
    - The size and location of the inode tables
    - The disk block map and usage information
    - The size of the cylinder groups
  - Copied to scattered locations on disks
    - use "**newfs –N** " to see where superblocks are located
  - In-memory copy
    - Flushed to disk copy every 30 seconds

File system types

Use	AIX	FreeBSD	HP-UX	Linux	Solaris	Tru64
Default local	Jfs or jfs2	Ufs	vxfs	Vxfs	Ufs	Ufs or advfs
NFS	nfs	nfs	nfs	nfs	nfs	Nfs
CD-ROM	cdrfs	Cd9660	cdfs	Iso9660	hsfs	Cdfs
swap	Not needed	swap	Swap, swapfs	swap	swap	Not needed
/proc	procfs	procfs	Not supported	procfs	procfs	Procs
RAM-based		mfs		Ramfs,tmpfs	tmpfs	mfs

Mounting the filesystem

- Mounting is the process that makes a disk's content available to the system, merging it into the system directory tree
- Manually mount a file system

  # mount block-special-file mount-point

  Example:
  #mkdir /apg
  #mount /dev/dsk/c1t2d0 /apg

  Manually mount a file system readonly
  #mount –r /dev/dsk/c1d1s7 /mnt

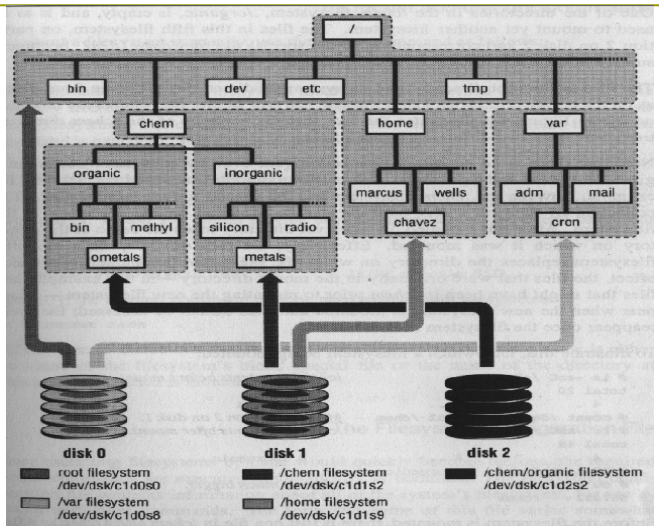
  List all currently mounted filesystem
  #mount

Set up automatic mounting

- File system can be set up to be mounted automatically at boot time
- SunOS, Tru64 Unix, HP-UX and Linux: /etc/fstab
  - format:

    block-special-file mount-loc type opts dump-freq pass-number
    - type: vxfs, advfs, ext2,ext3, nfs, ufs, swap, ...
    - opts: rw, ro, suid, nosuid, noauto, ...
- Solaris: /etc/vfstab
  - Format:

    block_spfile char_spfile mount-loc type fsck-pass automount? Opts
- Mount only require either the mount point or special file name as argument
  - #mount /chem.
  - #mount /dev/disk1d



Umounting a file system

- umount only require either the mount point or special file name as argument
  - #umount /chem.
  - #umount /dev/disk1d
- file system must be inactive
  - Determine who is using the filesystem: **fuser**
    - Identify user, process id
    - Options: -u, -k

Using fsck to validate a filesystem

- The fsck utility checks the filesystem's consistency, reports any problems it finds and optionally repair them
  - Comparing the block free list against the disk addresses stored in the inodes
  - Comparing the inode free list against inodes in directory entryies
- The five most common types of damage are:
  - Unreferenced inodes
  - Inexplicably large link counts
  - Unused data blocks not recorded in the block maps
  - Data blocks listed as free that are also used in a file
  - Incorrect summary information in the superblock
- Fsck's scope is limited to repairing the structure and its component data structure.

Using fsck (cont)

- Disks are normally checked at boot time with fsck –p, which examines all local file system listed in /etc/fstab
  - Under BSD, fsck is run automatically
  - Under System V, fsck is run at boottime on filesystems only if they were not dismounted cleanly
  - If journaling is enabled, fsck simply rolls up the log to the last consistent state.
- Manually run fsck if there is serious error that fsck needs human intervention
  - You may be bring into single user mode
  - Check root partition first.
  - Rerun fsck until the filesystem comes up clean.
  - Do a good backup using dd to copy the whole disk

Extending UNIX filesystems

- Hard way:
  - Backup the data
  - Remove the partition
  - Create new partition
  - Create larger file system
  - Restore the data

Checking file system space

- Command
  - Report file system usage: df
  - Estimate file space usage: du
- CASE: Copy a database
  - Check the file system size on customer site
  - Create a file system big enough
  - Copy the Oracle database data files
    - In the middle of copying, the file system is full, why?
      - Spare file

Establishing logical volumes

- Logical Volume manager
  - Group multiple disks or partitions into virtual disks
- Why LVM?
  - Managing large hard disk farms without disrupting services
  - Creating single logical volume from multiple physical volumes
  - Resizing volume size online

LVM commands for HP-UX

- Identify physical volumes
  - pvcreate
- Manage Volume groups
  - vgcreate/vgextend/vgdisplay
- Manage logical volumes
  - lvcreate/lvdisplay
- Create new file system
  - newfs

Topics

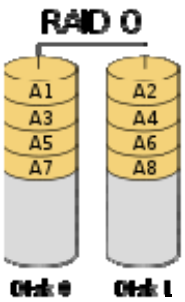
- Redundant Array of Inexpensive Disks (RAID)
- NAS (Network Area Storage)
- SAN (Switch Area Network)

RAID

- Redundant Array of Inexpensive Disks
  - Fault tolerance
  - Performance
- Commonly used on servers
- Raid Level
  - Raid 0, 1, 2, 3, 4, 5, 6
  - Refer
    - http://en.wikipedia.org/wiki/RAID

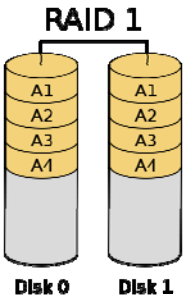
Raid 0

- Striped Disk Array
  - Spreading out blocks of each file across multiple disk drives
  - Improve the performance
    - Disks can seek independently
  - No fault tolerance
    - One drive fails then all data in array is lost



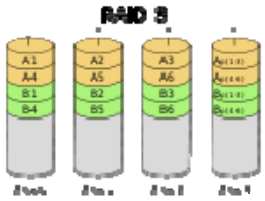
RAID 1

- Disk mirroring
  - Twice read transaction rate of single disks
  - The same write transaction rate
  - Fault tolerance for 1 disk failure
  - Space efficiency: Half disk space



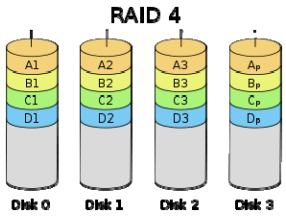
RAID 3

- Byte-level striping with a dedicated parity disk
  - The strip is a lot smaller than a filesystem block
  - Parity handling
    - Read the old data block
    - Read the old parity block
    - Compare the old data and new data, and change the parity data accordingly
    - Write the new data block
    - Write the new parity block
  - Require minimum 3 disks



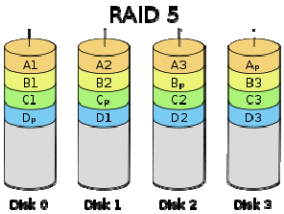
RAID 4

- Block-level stripping with a dedicated parity disk
  - Serve multiple read requests simultaneously
    - Read A1, B2, C3
  - Write the parity disks becoming the bottleneck
    - Write A1, B2
      - Write to disk 0 A1, disk 1 B2
      - Parity write to disk 3 block Ap and Bp



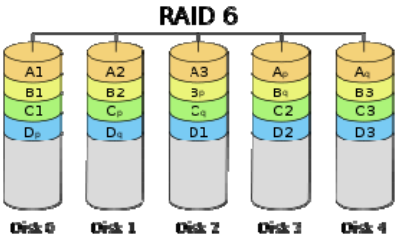
RAID 5

- Block-level string with parity data distributed across all disks
  - Read can not done at the same time
    - Read A1, B2
  - Write the parity disks bottleneck is relieved
    - Write A1, B2
      - Write to disk 0 A1, disk 1 B2
      - Parity write to disk 3 block Ap and disk 2 Bp
  - The data on the failed disk can be reconstructed using the parity data and block data from the survived disks



RAID 6

- Striped set with dual parity
  - Provide fault tolerance from two disk failures
  - Minimum 4 disks
  - Space efficiency: n - 2



Hardware vs. Software RAID

- Hardware Raid
  - Controller card
    - Hardware generated parity
  - High availability
    - Redundant hardware support
- Software Raid
  - Unix utilities, included in OS
    - No addition cost of controller
    - No need to install, configure or manage a hardware raid controller.
  - Disadvantages:
    - Slow performance, especially with parity
    - Boot volume limitation
    - Level support limited
    - No advance feature like hot spares and drive swapping
    - Operating system, software compatibility issue and reliability concerns

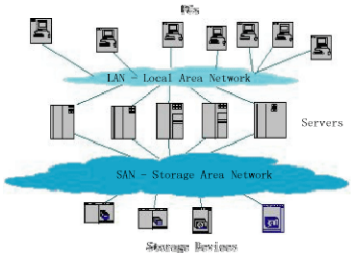
Example: Smart Array

Smart Array	6404	6402	» 5304	» 5302	» 5312
SCSI Protocols Supported	Ultra320, Ultra3, Ultra2	Ultra320, Ultra3, Ultra2	Ultra3, Ultra2, Wide Ultra	Ultra3, Ultra2, Wide Ultra	Ultra3, Ultra2
Maximum Channel Transfer Rate (MB/s)	1280 MB/s total, 320 MB/s per channel	640 MB/s total, 320 MB/s per channel	640 MB/s total, 160 MB/s per channel	320 MB/s total, 160 MB/s per channel	320 MB/s total, 160 MB/s per channel
Channels	4	2 (upgradeable to 4)	4	2 (upgradeable to 4)	2
SCSI Ports (Internal/External)	2/4	2/2 (Upgradeable to 2/4)	2/4	2/2 (Upgradeable to 2/4)	2/2
Maximum Drives	56	28 (upgradeable to 56)	56	28 (upgradeable to 56)	28
Cache	256MB	128MB (upgradeable to 256MB)	128MB(upgradeable to 256MB)	256MB	128MB(upgradeable to 256MB)
Recovery ROM	Yes	Yes	Yes	Yes	Yes
RAID Support	0, 1, 1+0, 5, ADG	0, 1, 1+0, 5, ADG	0, 1, 1+0, 5 ADG	0, 1, 1+0, 5 (upgradeable to ADG)	0, 1, 1+0, 5
Maximum Logical Volume Size	2TB	2TB	2TB	2TB	2TB
PCI Bus	64-bit, 133MHz PCI-X	64-bit, 133MHz PCI-X	64-bit, 66MHz PCI	64-bit, 66MHz PCI	64-bit, 133MHz PCI-X

SAN

- A SAN, Storage Area Network (SAN),
  - high-speed network of shared storage devices
- What is a storage device?
  - Contains nothing but a disk or disks for storing data.
- A SAN's architecture works in a way that makes all storage devices available to all servers on a LAN or WAN.
  - the server merely acts as a pathway between the end user and the stored data.

SAN architecture example



Why a SAN?

- Unprecedented levels of flexibility in system management and configuration
  - The storage can be easily increased, changed or re-assigned.
  - As more storage devices are added to a SAN, they too will be accessible from any server in the larger network.
  - Servers can be added and removed from a SAM while their data remains in the SAN
  - Multiple servers can access the same storage for more consistent and rapid processing.
- SAN brings enterprise-level availability to open system servers.
  - SAN improve staff efficiency by supporting a variety of operating systems, servers and operational needs.
- Because stored data does not reside directly on any of a network's servers, server power is utilized for business applications, and network capacity is released to the end user.
- Support high-performance backup and rapid restores.
- Multiple compute servers and backup servers can access a common storage pool. SAN offers configurations that emphasize connectivity, performance, resilience to outage, or all three.

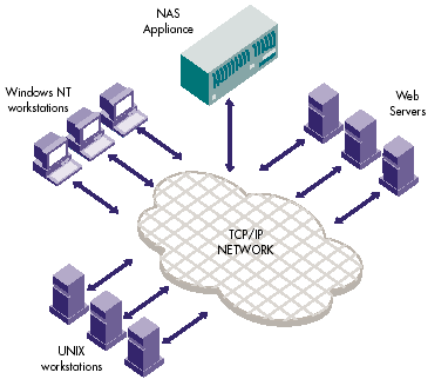
Example: Hadley group's SAN

- EMC Symmetrix Storage Array
  - close to 300 78G drive with total 24TB
- EMC Connectrix SAN switch
  - 32 ports SCSI switch
  - 64 ports fiber switch
- Usage:
  - Oracle data storage at the file system level for all servers
  - Sybase data storage at the raw device level for all server
  - All other file systems (except root) for all servers
  - Serve the data storage for EMC Celerra File server
  - Replicate data
- Problems
  - Not working with Tru64 4.0g with more than 7 devices
  - Creating fake device names and multiple reboots on Tru64 4.0x

NAS

- Is designed to provide data access at the file level.
- Is based on TCP/IP
- Provides Special OS

NAS architecture example



Example: Hadley group's NAS

- Net Appliance
  - Two disk shelves
  - Total 150 G
- Data:
  - Home dir
  - Source code repository
  - Snapshots
- Problems:
  - Network bottleneck
  - Could not be extended more
  - Local tape backup took too long

SAN and NAS comparison

	SAN	NAS
Protocol	• Fibre Channel • Fibre Channel-to-SCSI	• TCP/IP
Applications	• Mission-critical transaction-based database application processing • Centralized data backup • Disaster recovery operations • Storage consolidation	• File sharing in NFS and CIFS • Small-block data transfer over long distances • Limited read-only database access
Advantages	• High availability • Data transfer reliability • Reduced traffic on the primary network • Configuration flexibility • High performance • High scalability • Centralized management • Multiple vendor offerings	• Few distance limitations • Simplified addition of file sharing capacity • Easy deployment and maintenance