Bus

- A channel or path between the components in a computer
- Has evolved slowly compared to other computer components
  - Most computers still have Industry Standard Architecture (ISA) bus developed for the original PC in the ’80s
  - Need for long-term compatibility with large number of hardware manufacturers
  - Before multimedia, few peripherals fully utilized the bus speed
- Two key buses
  1. System bus or local bus
     - Connects CPU to memory
  2. ISA or PCI bus
     - Connects to system bus through a bridge
     - Bridge resides on chipsets and integrates data from other buses into system bus
- Dual Independent Bus (DIB)
  - Replaces standard system bus to isolate the path between CPU and memory
  - Replaces single system bus with frontside bus and backside bus
  - Backside bus
    - Provides a direct and fast channel between CPU and Level 2 cache
  - Frontside bus
    - Connects system memory (memory controller) to CPU and other buses to CPU and system memory
- Shared bus
  - Another main bus to connect additional components to the computer
  - Lets multiple devices access the same path to CPU and system memory
- Other buses
  - Extended Industry Standard Architecture (EISA)
    - Original ISA-bus operated at 4.77MHz at 8-bits width
    - Improved to 8MHz at 16-bits width (still called ISA bus)
    - EISA was 32-bits wide at 8MHz
  - Vesa Local Bus (VL-Bus)
    - Video Electronics Standards Association
    - 32-bits at the speed of local bus, which is normally the speed of CPU itself
    - Ties directly into CPU
    - Works with one or two devices but more devices can cause interference to CPU
    - Typically used to connect a graphics card
- PCI bus (Peripheral Component Interconnect)
  - Developed by Intel
  - Provides features of both ISA and VL-Bus
  - Provides direct access to system memory for connected devices, but uses a bridge to connect to the frontside bus and therefore to the CPU
  - Capable of even higher performance than VL-Bus while eliminating the potential for interference with the CPU
– Can connect more devices than VL-Bus, up to five external components
– Can theoretically have more than one PCI bus
– PCI bridge chip
  * Regulates the speed of the PCI bus independently of the CPU’s speed
  * Provides a higher degree of reliability
– Originally operated at 33MHz using a 32-bit-wide path

Serial devices

- Used with a variety of devices, such as printers, terminals, and network

Serial standards

- RS-232
  - Most common standard for serial interface
  - Specifies electrical characteristics and meaning of each signal wire along with the pin assignments on traditional 25-pin (DB-25) serial connector
  - Overkill for real world situations, leading to alternative connectors
  - Uses shielded, twisted pair (STP) cables
  - Signal voltage is $\pm12$ volts, or $\pm5$ volts (common), or $\pm3$ volts
    * Higher voltage is less susceptible to interference
  - Not an electrically balanced system
    * Uses single conductor for data traveling in each direction
    * No reason to use twisted pair
    * STP provides the shield that helps reduce interference
  - If two data lines – TD (transmitted data) and RD (received data) – are placed together on a single line that results in reduced reliability and smaller range
  - Connectors can be male or female
  - Pins are numbered 1 to 13 in top row and 14 to 25 in bottom row, with pins 1 and 14 being the leftmost pins in their rows, in the female
    * Like numbered pins mate in both types of connectors
    * Most of the pins in male connectors are not even installed as they are ignored in real world
  - Two interface configurations for serial equipment
    1. Data terminal equipment (DTE)
    2. Data communications equipment (DCE)
    * DTE and DCE share the same pinouts but specify different interpretation of signals
  - Every device configured as either DTE or DCE
    * Computers/terminals/printers configured as DTE
    * Modems configured as DCE
  - Fine points of DTE and DCE
    * RS-232 pinout for a given connector type is the same, regardless of male or female, and regardless of cable, DCE, or DTE
RS-232 terminology is based on straight-through connection from a DTE device to a DCE device
- TC on DTE end is connected to TC on DCE end
- Signals are named relative to perspective on DTE device
  - Name TD implies “data transmitted from DTE to DCE” even though TD is an input on DCE
- Wiring of DTE to DTE requires tricking each device to think that the other is a DCE
  - Both DTE devices will expect to transmit on TD and receive on RD
  - Cross-connect the wires to achieve this effect
  - Cross three sets of signals: TD and RD, and RTS and CTS; also, each side’s DTR pin must be connected to both the DCD and DSR pins of the peer
- A cable crossed for DTE-to-DTE is called a null modem cable; cable for a modem is called a modem cable or a straight cable

Alternative connectors

- Have different physical configuration but provide access to same signals as DB-25
- Easy to convert from one connector to another
- Mini DIN-8 variant
  - Found on laptops/workstations, and Macs
  - Almost circular and compact
  - Provide connections for seven signals
- DB-9 variant
  - Nine-pin connector to carry eight most commonly used signals
  - Looks like a smaller version of a DB-25 connector
- RJ-45 variant
  - Eight-wire modular telephone connection
  - Similar to RJ-11 connector used in telephone
  - Most commonly used for Ethernet wiring
  - Compact, self-securing, and cheap
- Yost standard for RJ-45 wiring
  - Maps the pins on an RJ-45 connector to a DB-25
  - Can be used with either DCE or DTE equipment
  - Only need one kind of connector
  - Easy to produce this cable, using only a crimping tool
  - Allows almost universal connectivity, because of uniform interface
  - Make it possible to connect any device to any other device, using null-modems or null-terminals, changing pins on cable connectors, or building special cables

Cable length

- RS-232 standard allows a maximum cable length of 75’ at 9600 bps
Serial device files

- Serial ports are represented by device files in /dev
- Two serial ports – /dev/ttya and /dev/ttyb are built in
  - Solaris keeps these as links to other files/ports
- More than one device file could refer to the same port, but different minor device number
  - Look at the files /dev/term/a and /dev/cua/a
- Behavior of a device is not determined by name but by major and minor device numbers

Software configuration for serial devices

- Serial devices do not require configuration at the kernel level
  - Devices that connect directly to host bus require configuration at kernel level
- Configuration checklist for a new device
  - Hardwired terminal
    - Listen for logins on the terminal’s port
    - Specify speed and parameters of serial connection
  - Dial-in modems
    - Configured similar to hardwired terminals
  - Dial-out modem
    - Add entries to /etc/remote file for use by tip and cu commands

Configuration of hardwired terminals

- Obsolete now
- Need to understand because graphical displays use the same drivers and configuration files as real terminals
- Checklist
  1. Make sure a process is attached to a terminal to accept logins
  2. Information about the terminal should be available once user has logged in
- Login process involves several processes
  - init daemon, started at boot time
  - init forks a getty on each terminal port turned on in /etc/ttyys or /etc/inittab files to set the port’s initial characteristics (speed/parity)
  - On Solaris, getty is replaced by ttymon
- Sequence of events in login process
  - getty provides the prompt and gets user’s login name
  - getty executes login with login name as argument
  - login requests a password and validates it against /etc/passwd
  - login prints motd
login sets up TERM environment variable and execs the default shell, changing shell’s ownership to user

– Shell executes appropriate startup files, provides prompt, and waits for input
– When shell terminates, init execs another getty on the terminal port

• Files for terminal types
  – Not uniformly defined for all systems
  – Called /etc/saf/_sactab (Service Access Controller Table) on Solaris; saf is Service Access Facility
  – saf is a hierarchy that generalizes procedures for service access to uniformly manage
    * login access on local system
    * network access to local services

• Terminal support
  – termcap and terminfo databases
    – terminfo is a database that describes the visual interface of the terminal for use by programs such as vi; the special attributes include items such as reverse video, blinking, and underlining
    – Kept in /etc
    – Contain characteristics of terminals
    – Most important terminal types these days are DEC VT100 and xterm; almost all terminal types emulate one of these two

Special characters and terminal drivers

• Binding of functions to keys can be set with tset and stty (set teletype) commands
  – stty is used to both set and display the terminal settings
    – The stty command affects the serial line parameters, including data rate, start/stop bits, parity, translation of carriage return into linefeed, division of input into lines, definition of special control characters to erase a character in the input buffer, killing input line, and interrupting a running process
  – tset is inherited from the BSD lineage
    – It gets the terminal characteristics from termcap and terminfo

• You can display all the current settings by typing the command stty -a as follows

```
% stty -a
speed 9600 baud; -parity
rows = 24; columns = 80; ypixels = 316; xpixels = 499;
inpck -istrip imaxbel
crt iexten
erase kill werase rprnt flush lnext susp quit stop eof
? "^U ^W ^R ^O ^V ^Z/`^Y `^C `\ `^S/`^Q `^D
%
```

It indicates that the terminal has 24 rows and 80 columns. In addition, it displays the various special characters associated with the keyboard, such as erase

• The current setting can be changed by issuing the command
  
  `stty parameter value`

  – The number of rows in the terminal can be changed by
% stty rows 40
   – The backspace key can be changed to the key labeled Delete by
   % stty erase Delete
   where the string Delete indicates you to press the key labeled Delete and not to type it out

• If you use a terminal regularly where you prefer to use the Delete key in favor of Backspace, you can add the above command to your .login file

• Meaning of special key sequences in the output of stty is given by

<table>
<thead>
<tr>
<th>String</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>erase</td>
<td>Backup one space and delete character under cursor</td>
</tr>
<tr>
<td>kill</td>
<td>Erase entire command from command line</td>
</tr>
<tr>
<td>intr</td>
<td>Interrupt a currently executing command and return to prompt</td>
</tr>
<tr>
<td>susp</td>
<td>Suspend or halt the currently executing command</td>
</tr>
<tr>
<td>stop</td>
<td>Suspend output scrolling</td>
</tr>
<tr>
<td>start</td>
<td>Resume output scrolling</td>
</tr>
</tbody>
</table>

• The command stty sane will reset the keyboard values to reasonable values; useful if the screen gets garbled

Modems

• Modem converts digital serial signal produced by computer into an analog signal suitable for transmission over a voice line

• External modems
  – RJ-11 jack for standard phone line on analog side
  – RS-232 interface, typically female DB-25, on digital side

• Internal modems
  – Usually seen only on PCs
  – Plug into ISA, PCI, or PCMCIA slot
  – Have an RJ-11 jack sticking out the back of the computer
  – Cheaper than internal modem but more difficult to configure
  – Lack indicator lights for debugging help

• Modulation, error correction, and data compression protocols
  – Baud rate
    * Indicates the speed at which the signal is modulated
    * If there are more than two signal levels, more than one bit can be sent per transition, leading to higher bps
    * Confusing because historically, you could only send only one bit per transition
  – Fastest modems @ 56K V.90 standard
    * 33.6 Kb/s from computer to ISP
    * 53 Kb/s from ISP to computer
  – Error correction required to counter line noise
  – Data compression used to transmit fewer bits

• Dial-out configuration
– /etc/phones or /etc/remote (Solaris)
– tip – Connect to remote system
– cu – Call another Unix system

• Bidirectional modems
  – No standardization but useful in establishing sys admin work connection

Debugging a serial line

• Typical errors
  – Forgetting to tell init to reread its configuration files
  – Forgetting to set soft carrier when using three-wire cables
  – Using a cable with wrong nullness
  – Soldering or crimping DB-25 connectors upside down
  – Connecting a device to the wrong wire due to bad or non-existent wire maps
  – Setting the terminal options incorrectly

• Breakout box
  – Patched into serial line
  – Shows signals on each pin as they pass through the cable
  – Possibly bisexual in positioning

Other common I/O ports

• Parallel ports
  – Faster than serial ports as they transmit 8 bits of data in parallel
  – Require bulkier cabling and connectors
  – Most commonly used for printers
  – Best standard is IEEE-1284
  – Can be set to operate in EPP (enhanced parallel port) or ECP (extended capability port)
    * The two modes are equivalent
    * ECP supports DMA
  – Computers provide female DB-25 connector for parallel port and peripherals tend to have female 36-pin Centronics connector
    * Most parallel cables are male DB-25 to male Centronics
  – Parallel cables can be up to 10 meters long

• USB – Universal Serial Bus – ports
  – First standard published in 1995
  – Great properties for a low speed bus
    * Up to 127 devices can be connected
    * Cables have only four wires – power, ground, and two signal wires
    * Connectors and connector genders are standardized
Connectors are small, and cables are thin and flexible
Devices can be connected and disconnected without powering down
Signaling speeds up to 12 Md/s
Legacy serial and parallel devices can be connected with adapters

Adding a disk

Disk interfaces

- **SCSI interface**
  - Small Computer Systems Interface
  - Supports multiple disks on a bus with various speeds and communication styles

- **IDE interface**
  - Integrated Drive Electronics
  - Standard architecture for all modern disks
  - Simple, low cost interface for PCs
  - Puts the hardware controller in the same box as the platters
  - Relatively high level protocol for communication between computer and disks
  - Disks are medium speed, high capacity, and low cost
  - Useful with only four or fewer devices on workstations

- **Fiber channel**
  - Serial interface with high bandwidth, with speeds of more than 100 MB/s
  - Supports a large number of devices simultaneously
  - Uses fiber optic or twinaxial copper cables
  - Devices identified by a hard-wired ID number called a World Wide Name

- **USB**
  - Useful for slower disk devices such as CD-ROMs

- **SCSI interface**
  - Developed in 1986
  - SCSI disks are more expensive but much higher quality compared to IDE
    - SCSI disks can rotate at 10,000 to 15,000 RPM
    - Typically come with a 5 year warranty and 1,200,000 hours mean time to failure
    - The performance has been matched recently by serial ATA standard but still requires a separate dedicated cable from controller card to each disk
  - Implemented on several chipsets
  - SCSI support may be available on the CPU or peripheral board
  - Defines a generic data pipe for many types of peripherals
    - Mostly used for disks, tapes, scanners, and printers
    - Computer and peripherals are treated as peers; no master/slave relationship
    - Bus is busy only when actually transferring data; not when disks are moving heads to go to correct sector
– Does not specify the construction or lay out of disk; only the communication aspects
– scsi-2
  * Allows for command queuing
    · Disk can start to move head even while the data from previous command is being transferred
  * Devices can reorder I/O requests to optimize throughput
  * Scatter-gather I/O to permit DMA from discontiguous memory regions
  * “Fast” SCSI or “wide” SCSI
    · Indicate double speed on the bus, or number of bits transferred simultaneously is larger (16 or 32, instead of 8)
    · Wide SCSI chains can support up to 16 devices, instead of 8
– scsi-3
  * Family of standards
  * Includes specifications for physical media, including traditional parallel buses and high speed serial media such as Fibre Channel and IEEE 1394 (Firewire)
  * Defines SCSI command sets and introduces enhancements to support device autoconfiguration, multimedia applications, and new device types
– Putting an older device on a newer bus can slow down the entire bus, and affect maximum cable length
– Noise issues
  * Normal single-ended SCSI has every other pin grounded to reduce cross talk among signals
  * Limits the cable length to 6m on SCSI-1, 3m on SCSI-2, and 1.5m on UltraSCSI
– Differential SCSI
  * Differential signaling puts an inverted signal next to each pin instead of a ground, making net voltage zero and reducing noise significantly
  * Increases the cable length to 25m for SCSI-2 and 12m for UltraSCSI
  * Incompatible with nondifferential devices; does not allow mixing of the two device types
    · Requires differential controller, disk, cable, and terminator
– Connectors
  * Narrow/standard SCSI devices have 50 pins
  * Wide SCSI devices have 68 pins
  * Apple reduced the 50 pins to 25
    · Tied all the ground lines together
    · Shoehorned the bus onto a DB-25 connector
– SCSI bus
  * Use daisy chain configuration
  * Most external devices have two identical and interchangeable SCSI ports
  * Internal SCSI devices are attached to a ribbon cable
    · Only one SCSI port required because connectors can be clamped onto the middle of ribbon cable
  * When using a ribbon cable, pin 1 of SCSI bus must be connected to pin 1 on hard drive
  * Each end of SCSI bus must have a terminating resistor or terminator
    · Terminator absorbs signals as they reach end of bus
    · Noise is not reflected back onto the bus
    · Terminators can be small plugs snapped into devices
    · Some devices may be autoterminating
    · One end of the bus terminates into the host computer; on SCSI controller or an internal SCSi drive
    · Improper termination on either end can cause problems
- SCSI address or target number
  - Required for each device to distinguish the devices
  - Start at 0 and go to 7 (standard SCSI) or 15 (wide SCSI)
  - SCSI controller itself is a device, typically 7
    - It stays as 7 on wide bus to ensure backward compatibility
  - SCSI address is completely arbitrary
    - Determines the device’s priority on the bus
    - Some systems pick disk with lowest target number to be boot device; others require boot disk to be target 0
  - Device address can be set with an external thumbwheel on the device, or by using DIP switches and jumpers

- Subaddressing or logical unit number
  - May be used to indicate several logical units on the same target number
  - Several disks on a single SCSI controller
  - Seldom used

- Things to check
  - Listing of current devices, including internal devices
  - No intermixing of differential and single-ended devices
  - No duplicate target numbers or SCSI address conflicts
  - Location of target in the bus, autoterminating or otherwise
  - Length of cable should account for internal devices and older SCSI devices
  - One address for SCSI controller

- IDE interface
  - Also called ATA (Advanced Technology Attachment)
  - Designed to be simple and inexpensive
  - Mostly found on PCs and low cost workstations
  - Controller built into disk to reduce interface cost and simplify firmware
  - ATA-2
    - Fast programmed I/O and DMA modes
    - Provides plug-and-play features
    - Logical block addressing
      - Overcomes a problem that prevented BIOSes from accessing more than first 1024 cylinders on disk
      - This constraint limited disk sizes to 504MB
      - BIOS manages part of bootstrapping
      - Create a small bootable partition within the first 1024 cylinders to account for older BIOS
      - Once kernel is up and running, BIOS is not needed and so, the entire disk can be accessed
    - Logical block addressing gets rid of cylinder-head-sector addressing in favor of a linear addressing scheme
  - ATA-3
    - Additional reliability, more sophisticated power management, and self-monitoring capabilities
  - Ultra-ATA
    - Adds high performance modes called Ultra DMA/33 and Ultra DMA/66 to extend the bus bandwidth from 16MB/s to 33MB/s and 66MB/s
  - ATA-4
    - Merges ATA-3 with ATA Packet Interface ATAPI
    - ATAPI allows CD-ROM and tape drives to work on an IDE bus
- IDE disks are almost always internal
- Maximum cable length limited to 18 inches
- IDE bus can only accommodate two devices; most manufacturers give more than one IDE bus on motherboard
- Devices accessed in a connected manner
  * Only one device can be active at a time
  * Performance demands devices be spread over multiple buses
  * Put fast devices (disks) on one bus and slow devices (CD-ROM/tape) on another to prevent slower devices from hindering the faster ones
- IDE connector
  * 40-pin header to connect drive to interface card with ribbon cable
  * Ultra DMA/66 use different cable providing more ground pins to reduce electrical noise
  * Pin 1 on the drive should go to pin 1 on the interface card
    - Pin 1 is usually marked with a small “1” on one side of connector
    - Pin 1 is usually the one closest to power connector
    - Pin 1 on ribbon cable is usually marked in red
- Connecting more than one device on IDE bus
  * One device is designated as master, other as slave
- Check list
  * You can put new IDE drives on old cards and old drives on new cards, with a degradation in performance
  * Cable length is very short
  * Old BIOS
  * Well-designed drivers for DMA and programmed I/O considerations

- Comparison between SCSI and IDE
  - SCSI is usually better but at a premium price

- SAS – Serial Attached SCSI
  - Upcoming replacement for SCSI
  - Part of SCSI family with Firewire and Fibre Channel
  - Supported by SCSI Trade Association, comprised of big vendors including IBM, Intel, HP, Fujitsu, and Maxtor among others
  - Offers full duplex connection over small cables
  - Can address thousands of devices per port at transfer rates of 3Gbps and more
  - Has the scalability, manageability, and data availability services of traditional parallel SCSI
  - Good for data centers because of scalability and reliability
  - Shares the common electrical and physical connection interface with serial ATA
    * SAS and serial ATA drives can be mixed in servers to lower the total cost of ownership

- Check out http://www.t10.org for more information on interfaces

**Disk geometry**

- Modern software need not know the physical construction of the drive
- Disk drive
Spinning platters coated with magnetic film in a hermetically sealed box for reliability
Data read/written by head that changes orientation of magnetic particles on the platters
Platters stacked on top of one another
  * Disk may use both sides of platters to store data, or one side to store data and the other to store positioning information
Track/Sector/Cylinder
Seek time/Latency
One head for each surface
Current diameter of disks has shrunk to \(3\frac{1}{2}\) inch or less
Platters rotate at constant speed
  * Heads move back and forth laterally
  * Heads float close to surface but never touch it
  * Head crash
Rotational speed
  * 7200 RPM common, with current fastest listed at 15,000 RPM with a transfer rate of 860Mbps (Maxtor Atlas 15K)
  * Higher rotational speeds imply smaller latency or rotational delay and higher bandwidth for data transfer but may introduce thermal problems
Zone sectoring
  * Not used in original BSD filesystem
  * Tracks on the outside contain more sectors than on the inside
  * Cylinder-head-sector (CHS) addressing scheme
    · Artificial scheme made up to fit the size of the disk
    · Hides the internal layout of disk from software

Overview of disk installation procedure

- Connecting the disk
  - IDE disk
    * Try to configure system with only one disk per bus
    * Check cable orientation and master/slave settings on each disk
  - SCSI disk
    * Properly terminate both ends of bus
    * Check cable length for maximum limits
    * Ensure that there is no conflict in target number

- Creating device entries
  - Need device files in `/dev` pointing to disk for access
  - Need both block-special files and character-special files
    * Block-special files are used for mounting filesystems
    * Character-special files are used to back up and check filesystem integrity
  - Set permissions on device files to be restrictive
    * Writing randomly can destroy a filesystem
    * Allow read/write access for owner (root) and readonly access for group (operator)
- Allows `dump` to be run by operators without root privilege but prevents non-root users from reading from the raw device

- **Formatting the disk**
  - Difference between formatted and unformatted capacity of disk
  - Megabyte (1 million $2^{20}$)
  - Formatting procedure
    - Writes address information and timing marks on platters to delineate sectors
    - Identifies bad blocks or imperfections in the media that cannot be reliably read/written
    - SCSI drives have bad block management built-in, so no need to worry about it
  - Hard disks come preformatted
    - Factory formatting is more precise
    - Avoid low-level formatting if not required
    - If there are read/write errors, check for cabling, termination, and address problems; if problems persist, get a new disk rather than reformat
    - IDE disks are not designed to be formatted outside factory
    - SCSI disks can be formatted by using `format` command on Solaris
      - May verify the integrity of the disk by writing random patterns and reading them back
  - Disks are designed to withstand constant activity; you won’t wear them out easily

- **Labeling and partitioning the disk**
  - Partitions or slices
    - Allow disk to be treated as a set of independent data areas
    - Allows boot blocks and partition table to be hidden from high-level software
    - Only device driver knows about the layout of entire disk; other software works with cleaned-up abstraction of partitions
  - Advantage of partitions
    - Easier backups
    - Improve performance
    - Limit users to only a part of the entire disk
    - Confine potential damage from runaway processes
  - Disk label
    - Record to keep the partition table
    - Occupies the first few blocks of the disk
    - Contains enough information to bootstrap the system
  - Image of entire disk
    - A special partition
    - Allows user commands to access the disk directly through a normal device file
    - User-level process can write the disk’s label or duplicate its contents to a backup disk by using the `dd` command
  - At least the following partitions exist on all systems
    - `root`
      - Everything needed to bring up the system in single-user mode
      - Second copy often stored on another disk for emergency
    - `swap`
      - Stores pages of virtual memory
    - `home` (may be called by some other name)
· User directories, data files, or anything else a user may want to keep online

– Splitting disks into partitions
  * If you have multiple disks, make a copy of root on another disk and make sure that you can boot off of it
  * As you add memory, add more swap space
    · For normal use, you must have at least as much swap space as memory
    · Allows a kernel crash dump to fit in swap area in case of system panic
    · Typical rule of thumb is to have three times the amount of main memory as swap space
  * Splitting swap space across disks improves performance
    · Put busy filesystems on different disks for better performance
  * Keep partition smaller than the capacity of backup device
  * Cluster information that changes quickly on few partitions that are backed up frequently
  * Create separate partition for /tmp
    · Limits temporary files to finite size
    · No need to back up temporary files
  * /var partition
    · Keeps log files
    · If part of root, can fill up root and bring machine to a halt

- Establishing logical volumes
  – Logical volume manager
    * Provides a supercharged version of disk partitioning
    * Lets you group multiple disks or partitions into a logical volume (metadisk) that appears as a single virtual disk
  – Concatenation
    * Keeps each device’s physical blocks together
    * Lines up devices one after another
  – Striping
    * Adjacent virtual blocks are actually spread over multiple physical disks
    * Can provide higher bandwidth and lower latency by reducing single-disk bottlenecks
  – RAID5 – Redundant Array of Inexpensive Disks
    * Striping with an extra checksum
    * Checksum provides redundancy so that logical volume’s data can be reconstructed if one of the disks goes bad
    * Also found in smart disk arrays known as hardware RAID
  – Mirroring
    * Mirrored volume is associated with another volume of same size
    * Whenever data is written to one side of mirror, it is duplicated on the other
    * Reads are split among the volumes to increase performance
    * Provides fail-safe volume in case of hardware failure on one volume, with no interruption in service
    * Two sides of mirror must be resynchronized when problem is fixed
  – Veritas
    * Logical volume manager supported on both Solaris and HP-UX
    * Solstice Disksuite is another volume manager from Sun
    * Vinum is open source volume manager for FreeBSD, inspired by Veritas
    * Linux LVM

- Creating Unix filesystems
– Need to add some overhead before making filesystem ready to use
– Use the command `newfs` or `mkfs` to install a filesystem within a partition
  * `newfs` is a front-end for `mkfs`
– Berkeley Fast File System
  * Designed for 4.2BSD and used by most modern versions of Unix
  * Five structural components
    1. A set of inode storage cells
    2. A set of scattered superblocks
    3. A map of disk blocks in the filesystem
    4. A block usage summary
    5. A set of data blocks
– Filesystem partition
  * Divided into cylinder groups of from 1 to 32 cylinders each
  * Structures such as inode tables are allocated among cylinder groups such that blocks that are accessed together can be stored close to each other on the disk
– Inodes
  * Fixed length table entries
  * Hold information for each file
  * Number of inodes fixed for the partition when the partition is created
– Superblock
  * Record to describe the characteristics of the filesystem
  * Length of disk block, size and location of inode table, disk block map and usage information, size of cylinder groups, and other important parameters
  * Several copies are maintained in scattered locations for redundancy
  * `fsck` can be told to use an alternative superblock when rebuilding a damaged filesystem
    · Location of backup superblocks can be determined with `newfs -N`
    · Block 32 always holds a backup superblock
  * Unix keeps an in-memory copy and several on-disk copies of superblock
    · `sync` command flushes the cached superblock on all disk copies
    · Periodic save minimizes the amount of damage if the machine crashes without `sync`ing the filesystem
    · `sync` also flushes modified inodes and cached data blocks
    · Most system `sync` every 30 seconds to minimize the amount of data lost in a crash
– Disk block map
  * Table of free blocks on the filesystem
  * Map is examined to devise an efficient layout scheme for new files
– Block usage summary
  * Records basic information about blocks already in use
• Setting up automatic mounting
  – Must have a valid filesystem on the partition being mounted
    * Never `mount` or `fsck` virtual filesystems such as `proc` or `swap`
  – Verify new filesystem by mounting manually as
    ```shell
    mount /dev/sd1a /mnt
    ```
    and checking for `lost+found` and the size of the filesystem through `df`
- lost+found directory
  - Created automatically on every filesystem
  - Used by fsck to repair the filesystem
  - Stores “unlinked” files without having to allocate additional directory entries on an unstable filesystem
- /etc/fstab
  - Lists device names and mount points of all system’s disks
  - Six fields per line, separated by whitespace, with one filesystem per line
    1. Device name – local or remote
       - Notation server:/export indicates /export directory on server
       - Can also specify a virtual filesystem instead of a device name (proc or swap)
    2. Mount point
    3. Type of filesystem – ufs on Solaris, ext2 on Linux, vxfs or hfs on HP-UX
    4. Mount point options, with default rw
    5. Dump frequency value to be used by backup products
    6. Pass in which fsck checks the filesystem
       - Filesystems with same value are checked concurrently
       - Two filesystems on same disk should not have the same value or the head will seek back and forth, resulting in performance degradation
  - Renamed to /etc/vfstab on Solaris, with different field structure, with a – to indicate no entry in the field
    1. Device name
    2. Device to fsck (specified as raw device)
    3. Mount point
    4. Type of filesystem
    5. fsck pass
    6. Mount at boot
    7. Mount point options
  - mount -a mounts all filesystems listed in fstab
  - mount command reads fstab sequentially mounting in order
    - /usr/local must follow /usr in fstab
- umount follows a similar rule in reverse
  - Any filesystem with files open cannot be unmounted
  - Check for open files by using the command fuser

- Enabling swapping
  - Raw partitions (without filesystem) used as swap space
  - Kernel does not use filesystem to keep track of swap’s contents
    - Kernel maintains its own simplified mapping from memory blocks to disk blocks
  - Possible to swap a file in a filesystem partition but slower performance
    - User can designate his own swap area by using mmap and munmap
  - For best performance, split swap area over several drives, preferably over several SCSI buses
  - On Solaris systems, swap area can be listed in vfstab as
    - /dev/dsk/c0t0d0s1 - - swap - no - swap - /tmp tmpfs - yes -
  - Swapping is enabled by running the command swapon
Device Management

**fsck: Check and repair filesystems**

- Filesystems in Unix are very robust but may get damaged in a number of ways
  - Kernel panic or power failure
  - Kernel buffers data blocks and summary information, splitting the most recent image of filesystem between disk and memory
  - Memory image is lost during crash
  - Buffered blocks are lost to most recently saved version on disk
- Fixing filesystems using *fsck*
  - *fsck* – filesystem consistency check
  - Used to fix minor damage to the filesystem; works well for most of the common problems
- Common types of damage
  - Unreferenced inodes
  - inexplicably large link counts
  - Unused data blocks not recorded in disk maps
  - Data blocks listed as free but used in a file
  - Incorrect summary information in superblock
- Journaling filesystem
  - Also called logging filesystem, and may replace *fsck* for efficiency
  - Available on Solaris UFS and HP-UX VXFS
  - Write metadata to a sequential log file to be flushed to disk before each command returns
  - Metadata is transferred later to its home from log file
  - Allows log file to be rolled up to most recent consistency point in the event of system crash
  - No need to cross-check the entire file system
- Rerun *fsck* if it makes corrections to filesystem
- At boot time, disks are checked with *fsck -p*
  - All filesystems listed in *fstab* are checked and repaired
  - If filesystem is unmounted cleanly, it may not be checked
  - Journaling filesystem is rolled to last consistent state
- *fsck* may request human intervention for some errors
  - Blocks claimed by more than one file
  - Blocks claimed outside the range of filesystem
  - Link counts that are too small
  - Unaccounted blocks
  - Directories referring to unallocated inodes
  - Any format errors