Servers

Saturday, November 03, 2007
7:56 PM

- Hardware for servers is different from hardware for workstations.
  - Extensibility: Servers have more slots and more space.
  - More CPU performance: Multiple CPUs and advanced hardware features. CPUs also faster because they are supposed to last a long time before upgrade.
  - High-performance I/O
  - Rack mountable: Designed for efficient space utilization and cooling in a rack.
  - No side-access needs
  - High availability options: Dual power supply, RAIDs, multiple network connectors, and hot-swap components.
  - Maintenance contracts
  - Remote management options: serial port accesses, internal thermometer or hardware monitoring systems.

- Three product lines
  - Home
    - Components are specified in general terms --> Vendor can choose the cheapest one to sell.
  - Business
    - Lowest total cost of ownership.
    - Takes longer to become obsolete.
    - Leased rather than purchased.
  - Server
    - Lowest cost per performance metric.
    - Increase price for increased uptime, and decreased mean time to repair (MTTR)
    - Chassis might be more expensive.

- Maintenance contracts
  - Non-critical server: next-day or 2-day response time.
  - Large groups of similar servers: purchase spare kits.
  - Critical host: Stack spare kits, or maintenance contract with same-day response
  - Large variety of models from the same member: Have a local technician, or regional spare kits store.
  - Highly critical host: on-site technician and duplicate machine.

- Server maintenance
  - Put them in data center with appropriate HVAC.
  - Mount the rack immediately before installing OS and other hardware.
  - Provide remote console access. (KVM switch, serial terminal)
  - Mirror boot disk of the server. Use RAID.

- More tips
  - Use appliances for specific tasks: DNS, firewalls.
  - Have redundant power supply.
    - Each power supply should have a separate power cord.
    - Each power supply should draw power from separate UPS or different power line.
  - Make an administrative network.

- Full redundancy vs. n+1 redundancy
  - n+1 redundancy = one component fail, the redundant component swaps in.
  - Full redundancy = two identical sets of hardware. If one fails, the other takes over.
  - Load sharing = a type of full redundancy. Two systems work concurrently and share load.
  - If one fails, the other takes all the load.

- Use many inexpensive servers.
  - Make massively scaled farm of servers.
  - If one fails, leave it be. You can clean up once the systems starts going downhill.
  - For example, blade servers.