Server Computing – How it is different

Design and Optimization Goals for Server Systems

- Capacity
- Scalability
- Integrity and security
- Availability
- Access to large amounts of data
- Systems management
- Autonomic capabilities

Capacity

- The potential or suitability for holding, storing, or accommodating
  - sufficient disk storage to allow fast access
  - a less expensive and secure media for long-term storage (tape)
- The facility or power to produce, perform, deploy, or simply process
  - sufficient computing capacity to run the programs that will process the data
  - Data could be stored and processed on a single server or dispersed and stored on many servers

Service level agreement (SLA)

- Agreement between a service provider and a recipient (server owner vs. business unit)
  - several SLAs for various aspects of a business
  - capacity management based on precise SLA
- SLA is baseline against for capcity demands
  - Examples from a SLA:
    - 95% of ATM transactions are completed in less than one second.
    - 90% of daily reports are completed by 6 a.m.

SLA implications

Example: ATM transaction
1. Insert the card.
2. Type in the PIN number.
3. Select from the menu.
4. Enter the amount.
5. Collect the cash.
6. Collect the card.
7. Print a receipt.

- Operations on ATM:
  - validating the card, prompting for input, and sending the request to the server.
- Server
  - check if the card is lost or stolen,
  - Implement cryptographic process
  - check the account, verify available funds, note withdrawal
  - no update to account until cash has been dispensed
- ATM signals that money has been dispensed
  - account can be debited.
- SLA involves ATM, communications network, and server.

Planning for downtime

- Must have capacity for (un-) planned outages
- Software
  - require maintenance activity
  - implement new levels or to fix problems
  - Continuity of service can be provided by sharing the critical workload across physical machines
- Careful choice of the time to take the system down (a scheduled downtime)
  - make it possible to meet the SLAs from the remaining systems.

Planning for downtime (contd.)

- Hardware
  - Two or more physical machines are required to provide continuity of service in the event of a machine loss
  - due to failure, maintenance, or repair, that requires the whole machine.
  - Server processes sharing the critical workload have to be placed on different physical machines.

Scalability

- Ability of a system to continue to function well as it is changed in size or volume
  - hardware, software, or a distributed system
  - ability to retain performance levels when adding processors, memory, and storage
- A scalable system can efficiently adapt to work
  - with larger or smaller networks performing tasks of varying complexity.

Integrity and Security

- Data security
  - protection against unauthorized access, transfer, denial of service, modification, or destruction,
  - whether accidental or intentional
- Well defined security objectives
  - Security administrator, security policy
  - Security manager
- System-provided interfaces enforce authority rules.

Availability

- Reliability, Availability, and Serviceability
  - often grouped together as RAS
  - Key features for data processing.
- System „exhibits RAS characteristics“ means
  - Architecture places a high priority on the system remaining in service at all times.
  - RAS is a central design feature of all aspects of a computer system, including the applications.

Reliability - RAS (contd.)

- Reliability
  - The system's hardware components have extensive self-checking and self-recovery.
  - The system's software reliability results from extensive testing and the ability to make quick updates for detected problems.

Availability – RAS (contd.)

- Availability
  - The system can recover from a failed component without impacting the rest of the running system.
  - hardware recovery (by automatically replacing failed elements with spares)
  - software recovery (layers of error recovery provided by the operating system).
- Mean time between failures (MTBF)
  - refers to the availability of a computer system.

Serviceability – RAS (contd.)

- Serviceability
  - The system can determine why a failure occurred.
  - This allows for the replacement of elements (hardware and software)
  - Impacting as little of the operational system as possible
  - well-defined units of replacement, either hardware or software.

Access to large amounts of data

- Storing vs. Processing data.
  - Tape media
    - good for sequential processing of data, but is of no value for random processing.
  - Disk media
    - sequential or random processing,
    - Very large disks are not generally a good solution because this usually results in overloading I/O system.
    - A larger number of smaller disks gives better response times to I/O requests.
- Bandwidth is crucial.

Systems management

- A collection of disciplines to monitor and control a system's behavior.
  - performance, workload, configuration, operations, problem management, network, storage, security, and change management techniques.
  - performed by the operating system or appropriate subsystems
  - Specialized tools marketed by various software companies.
- Good systems management plays a vital role

Autonomic capabilities

- Analogy to the autonomic central nervous system in the human body,
  - adjusts to many situations automatically without any external help.
  - A good way to handle IT complexity is to create computer systems that can respond to changes in their environment,
- Reduced need for human maintenance
  - fixing, and debugging of computer systems.