System Planning and Capacity Management

Frank Yeong-Sung Lin (林永松)
Department of Information Management
National Taiwan University
Taipei, Taiwan, R.O.C.

Outline

- Introduction
- System planning & capacity management
- Examples
- Summary
- Conclusion

Introduction

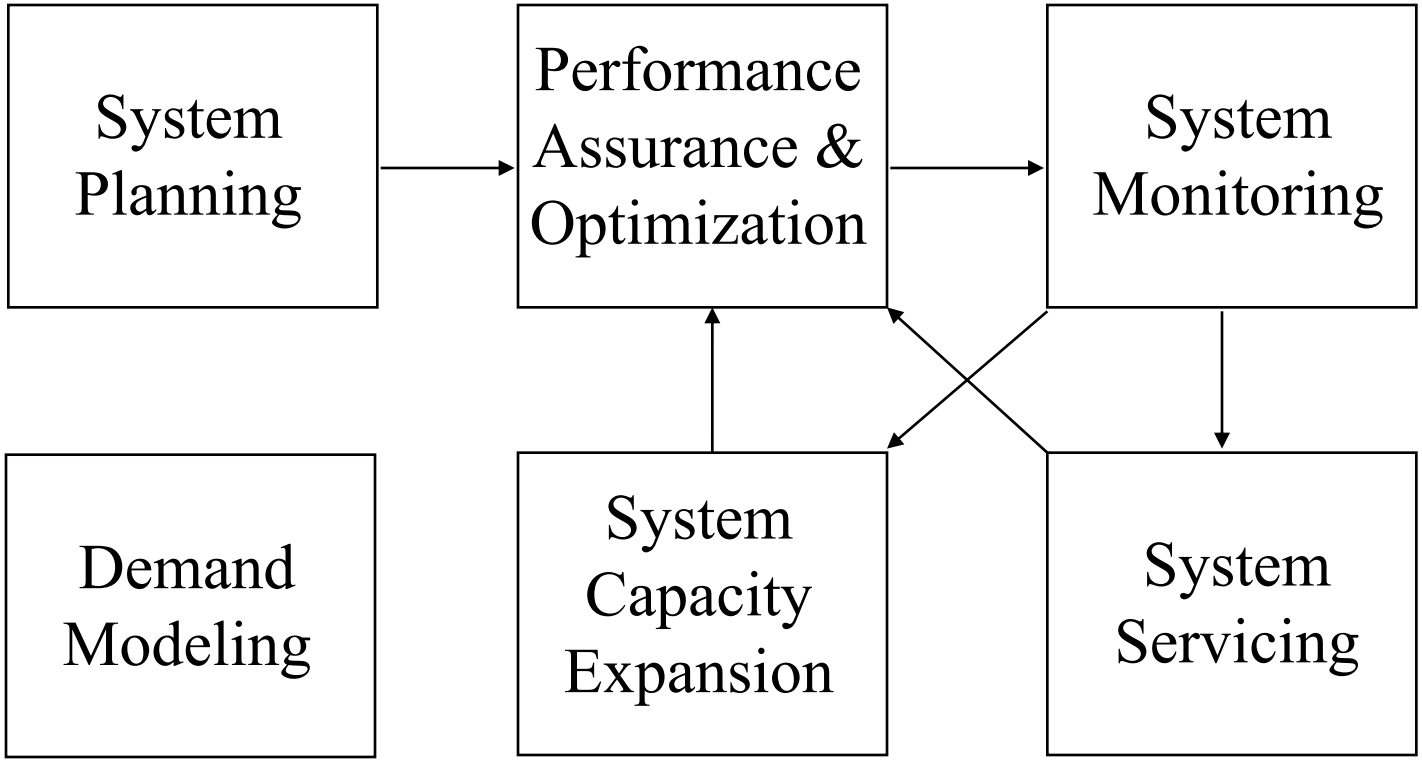
- Motivation
  - complexity of systems
  - needs of decision support systems (DSSs) and operation support systems (OSSs)
- Considerations
  - installation/operation/maintenance cost
  - performance (sanity)
  - integrity

Introduction (cont'd)

- Issues
  - efficiency and effectiveness
  - timeliness (development and response)
  - capacity
  - environment
  - user friendliness
  - integration and reliability
  - cost

System Planning & Capacity Management

- System architecture



System Planning & Capacity Management (cont'd)

- System planning
  - to design a system with the minimum installation and operation cost subject to performance (QoS), survivability/reliability and other constraints
- System performance assurance/optimization
  - for an in-service system, to assure pre-specified QoS requirements and/or to optimize certain performance measures, e.g. to minimize the total system throughput/revenue or to minimize the average cross-system delay

System Planning & Capacity Management (cont'd)

- System monitoring
  - for an in-service system, by using relevant measurements or performance modeling techniques (or a combination of the two) to identify potential performance exceptions and to activate corrective actions
  - to collect relevant measurements for load forecasting purposes (to feed the servicing and the capacity expansion processes)

System Planning & Capacity Management (cont'd)

- System servicing
  - using corrective actions to alleviate the performance exceptions identified by the monitoring process
  - three typical approaches
    - » load balancing
    - » resource reallocation
    - » sizing (minimal-cost capacity augmentation to satisfy the current demand)

System Planning & Capacity Management (cont'd)

- System capacity expansion
  - for an in-service system, to determine the capacity augmentation strategy at each decision stage over a pre-specified time horizon such that the total cost, considering the effect of economies of scale and composite cost of money, is minimized

Performance Considerations

- Performance/service objectives/constraints
  - throughput
  - peak delay
  - mean delay
  - delay jitter
  - tail distribution of delay (percentile type)
  - call set-up delay
  - call blocking probability
  - packet/cell loss probability
  - interference
  - availability/reliability/survivability

Performance Considerations (cont'd)

- Performance evaluation
  - traffic measurements
    - » call/packet/cell counts
    - » packet/cell loss counts
    - » call blocked counts
    - » delay counts are usually not directly available
  - performance modeling
    - » to derive performance measures from available traffic measurements & appropriate queueing models
    - » optimization is used to derive performance bounds from imperfect information for engineering purposes

Performance Considerations (cont'd)

- Performance evaluation (cont'd)
  - introduction to queueing theories
    - » components of queueing systems
      - probability density function (pdf) of interarrival times
      - pdf of service times
      - the number of servers
      - the queueing disciplines
      - the amount of buffer

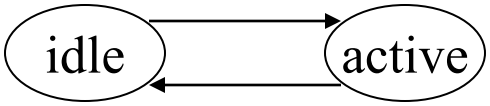
Performance Considerations (cont'd)

- Performance evaluation (cont'd)
  - introduction to queueing theories (cont'd)
    - » notation
      - M: exponential probability density
      - D: deterministic
      - G: general

        e.g. $M/M/1$, $M/M/m/m$, $M/D/1/K$, $G/G/m$
    - » Little's result $N = T\lambda$.
    - » $M/G/1$ queues are fully solvable (P-K formula).
    - » $GI/GI/1$ queues can be approximately analyzed by using the first two moments of the interarrival times and the service times.
    - » $M/M/m/m$ queueing models can be used to analyze the call blocking probability (Erlang B formula).

Performance Considerations (cont'd)

- Notion of equivalent bandwidth
  - Reference: R. Guerin et al, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", IEEE Journal on Selected Areas in Communications, 9(7), Sep. 1991
  - The approximation for the equivalent capacity is based on a fluid-flow model, which focuses on the representation of traffic source.
  - A traffic source is modeled by a two-state Markov source, characterized by the connection metric vector (R_{peak}, ρ , b)



    » R_{peak}: the peak rate of the connection
    » b: the mean of burst period (the mean of times during which the source is active)
    » ρ: utilization (fraction of time the source is active)

Performance Considerations (cont'd)

- Notion of equivalent bandwidth (cont'd)
  - We wish to determine the bandwidth to allocate to the associated connection in isolation.
  - The distribution of the buffer contents, when such a source is feeding a buffer served by a constant rate server, can be derived using standard techniques.
  - From this distribution, it is then possible to determine the equivalent capacity $\hat{c}$, needed to achieve a given buffer overflow probability.
  - Assuming a finite buffer of size x and overflow probability ε (the PDU loss requirement), the equivalent capacity is obtained by

$$\hat{c} = \frac{\alpha b(1-\rho)R_{peak} - x + \sqrt{[\alpha b(1-\rho)R_{peak} - x]^2 + 4x\alpha b\rho(1-\rho)R_{peak}}}{2\alpha b(1-\rho)}$$

$$where\ \alpha = \ln(1/\varepsilon)$$

Cost Considerations

- Deployment cost
  - fixed cost
    - » real estate
    - » other infrastructure and basic components
  - variable cost
    - » transmission/switching capacity
    - » processing/storage capability
- Operational cost
  - maintenance cost
  - personnel cost

Data in Support of System Planning & Capacity Management

- Location data
  - candidate locations and corresponding real estate costs
- Load/Resource requirements
  - end-to-end (preferred) or system element demand
- QoS requirements

Data in Support of System Planning & Capacity Management (cont'd)

- System element cost structure
  - cost of system elements considering pricing of available system element types and economies of scale
- System element characteristics
  - load-service curves of each system element
- Performance objectives
  - user or system performance objectives specified by requirements or service contracts

Examples for System Planning & Capacity Management

- Maximization of cloud computing system survivability under malicious and intelligent attacks
- Design of a cloud computing system

Summary

- Architecture and functionality of system planning & capacity management (SPCM) are presented.
- Examples are given.

Conclusion

- Information (technology) is power!
- Cloud computing is a trend of the information era.
- Cloud computing system planning & capacity management is crucial for reliable and efficient information processing, acquisition, exchange and distribution.
- Information over planned and well managed cloud computing systems is even more powerful!

Q&A