Performance Modeling of Servers - Queueing Systems (cont.)

Kai Shen

Dept. of Computer Science, University of Rochester

Basic Queueing Theory

- Stochastic processes
- Markov processes/chains
- Birth-death processes
- Poisson processes

Consider n(t) in an M/M/1 queue

- h = r/u (traffic intensity or utilization)
- then $p_n = (1-h) \times h^n$
- E(n) = h/(1-h)

How realistic is the M/M/1 model?

- IID Exponentially distributed service time?
- Service time plots will be re-posted after assignment #2

How realistic is the M/M/1 model?

- IID exponentially-distributed inter-arrival time (Poisson arrival process)?
- True if all requests are made from a large number of independent users.
- No true with:
  - multiple related requests in user session
  - multiple related low-level requests as result of a single high-level request (embedded objects in an HTML page request)
- Realistic arrival model:
  - user sessions arrival follows Poisson process
  - with each session, multiple requests arrive follow some usage model

M/G/1 Model

- No assumption of service time distribution
- Request arrival follows the Poisson process
- E(n) can be derived from the traffic intensity (utilization) and coefficient of variation of the service time

Operational Laws

- Operational laws
  - those that don't require assumptions that cannot be validated through black-box operational observations
  - operational metrics: arrival and departure time of requests (no information on corresponding pairs), probably system busy periods, …
- Little's Law
  - mean number in system = arrival rate x mean time in system
  - assumption: number of arrivals equals number of completions

Mean Value Analysis (MVA)

- Sometimes things are easier if we just care about the mean values
  - just need to know E(n), not p_n for all possible n's
- R = S (1+Q)?
  - R – mean response time;
  - S – mean service time;
  - Q – mean queue length, or E(n).
- If so, then Q = T R = T S (1+Q) = h (1+Q);
- Thus Q = h / (1-h).
  - T – throughput;
  - h – traffic intensive, or utilization.

MVA on Closed Queueing Networks

- A closed system with a single queue is not very interesting
  - Q = n
- A closed system with bunch of queues (closed queueing network)
  - CPU – Disk - …
  - Multi-stage services
- $R_n = S (1+Q_{n-1})$ [Reiser and Lavenberg 1980]
  - R_n – mean response time when there are n requests;
  - S – mean service time;
  - Q_n – mean queue length when there are n requests.
- If we also know the relation between Q_n and R_n, we can then calculate them recursively.

Disclaimer

- Some materials in these slides were developed from the book "The Art of Computer Systems Performance Analysis", R. Jain, 1991, Wiley.