

# Performance Modeling of Web Servers

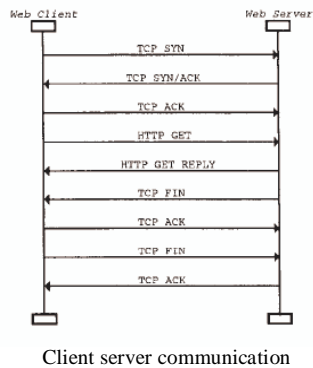
Sheila Srinivas

April 28, 2004

CS 756

1

## Web Server



- ❑ Requests arrive at listener process and are dispatched to one of a pool of server processes/threads.
- ❑ Request for HTML or image is retrieved and sent back.
- ❑ Request for dynamic content (CGI): the server process creates a child process which runs a CGI script and returns output for the server to send back.

CS 756

2

## When Performance is a Problem

- ❑ Rapid growth of the Internet.
  - ❑ Popular web sites receive millions of hits/day, resulting in high response times.
  - ❑ Trade-off between average latencies and percentage of requests rejected.
  - ❑ End user performance constrained by the bottleneck component.
- Solution: Adequate Capacity Planning.

CS 756

3

## Capacity Planning

- ❑ It is the process of predicting when future load levels will saturate the system and of determining a cost effective way of delaying or overcoming it.
  - Must consider the evolution of workload
  - Desired service levels
- ❑ Methodology:
  - Workload characterization, model validation.
  - Performance model development, validation.
  - Cost model development, cost/performance analysis.

CS 756

4

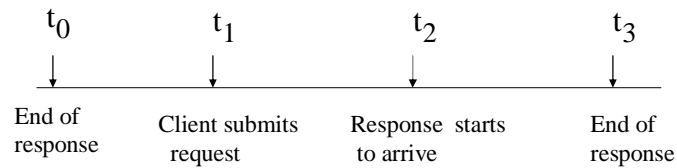
## Workload Characterization

- ❑ From [2], [3]. Studies suggest that web traffic is bursty at many timescales (self-similar). The following distributions are heavy tailed:
  - File sizes and hence transmission time
  - Think time
- ❑ Web workload exhibits:
  - Concentration of references
  - Temporal locality
- ❑ Workload can be partitioned into classes depending on the type of request.

CS 756

5

## Performance metrics



Think time:  $t_1 - t_0$ , reaction time:  $t_2 - t_1$ , response time:  $t_3 - t_1$

- ❑ Average response time
- ❑ Average throughput
- ❑ Average queue length
- ❑ blocking probability

CS 756

6

## Performance modeling

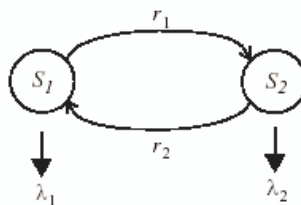
- ❑ Models are used to predict performance.
- ❑ System-level models
  - Simplest model:  $M/M/1/K*FCFS$
  - Other variants:  $M/G/1/K*FCFS$ ,  
 $M/D/1/K*FCFS$ ,  $M/G/1/K*PS$
  - $MMPP/G/1/K*PS$
- ❑ Component-level models
  - Queuing Networks
  - Layered Queuing Networks

CS 756

7

## Markov Modulated Poisson Process

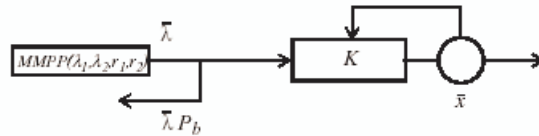
- ❑ Doubly stochastic Poisson process whose rate varies according to the state of a Markov chain.
- ❑ Modulation introduces correlations between successive inter-arrival times.
- ❑ A two state Markov chain is denoted MMPP-2 or switched Poisson process.
- ❑ Superposition of MMPP's is also MMPP.



CS 756

8

## MMPP/G/1/K\*PS model [4]



- ❑ Arrival process is two-state MMPP with parameter  $\lambda_1, \lambda_2, r_1, r_2$ .
- ❑ Service time is general.
- ❑ Scheduling is processor sharing.
- ❑ Service can handle at most  $k$  requests at a time.

CS 756

9

## MMPP/G/1/K\*PS model [4]

- ❑  $\lambda_{(\text{mean})} = (\lambda_1 r_2 + \lambda_2 r_1) / (r_2 + r_1)$
- ❑ The rate of blocked requests is  $\lambda_{(\text{mean})} \cdot P_b$
- ❑ MMPP parameters:
  - $r_2 = 0.05, r_1 = 0.95$
  - $\lambda_1 = 0.75 \lambda_{(\text{mean})}$
  - $\lambda_2 = ((r_2 + r_1) \lambda_{(\text{mean})} - \lambda_1 r_2) / r_1$
- ❑ The average response time, throughput and blocking probability are obtained through simulations and validated by measurements.

CS 756

10

## Session based MMPP/G/1 model

- ❑ From [5]. Session based workload model.
- ❑ When server is overloaded, a dropped request leads to incomplete sessions.
- ❑ Severe loss of throughput in terms of completed sessions, while maintaining throughput in requests/sec.
- ❑ Admission control: accept new session only if server can guarantee its successful completion.

CS 756

11

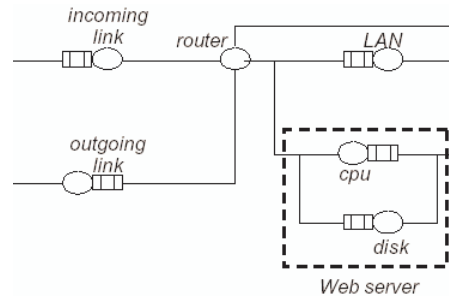
## Session based MMPP/G/1 model

- ❑ Model:
  - Session lengths are exponentially distributed.
  - Think time between requests of the same session is exponentially distributed.
  - Service time is linearly proportional to the requested file size (general distribution).
  - Queue size: 1024 entries.
- ❑ If observed utilization of server is above 95%, reject all new sessions. Admit new sessions if utilization is below 95%.
- ❑ Rejection overhead is quite small.

CS 756

12

## Queuing Network Models



- From [1].
- Single class or multi-class open QN model.
- Different classes are associated with requests for documents of different size.

CS 756

13

## Queuing Network Models

- Input traffic burstiness is taken into account by inflating the service demand of the components using a burstiness factor.
- Time interval  $t$  is divided into  $n$  subintervals of duration  $t/n$  called epochs. Then  $b = (\text{number of epochs where } \lambda_k > \lambda)/n$ 
  - $\lambda_k$  arrival rate of requests in epoch  $k$
  - $\lambda$  average arrival rate during  $t$
- Input parameters: service time of the components to process one request of class  $r$ , arrival rate.
- Response time  $R(i)$ , utilization  $U(i)$  and number of customers  $N(i)$  can be obtained by analysis or by using an iterative algorithm like Mean Value Analysis.

CS 756

14

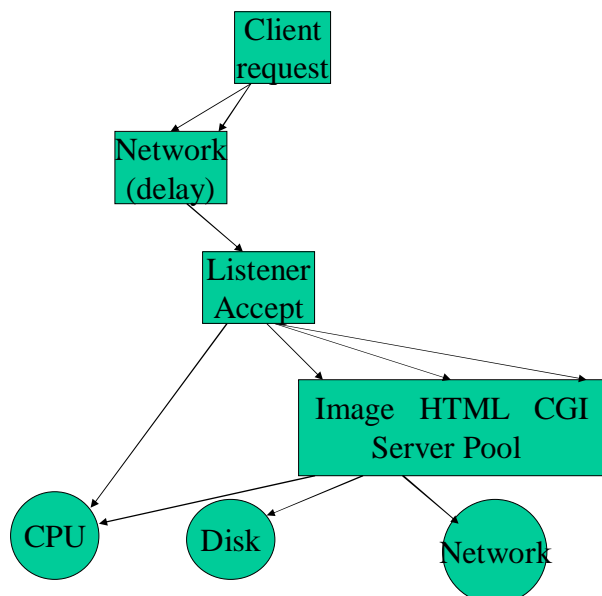
## Layered Queuing Models

- ❑ From [6], [7]. Extended QN models that consider contention for software processes as well as physical resources.
- ❑ Used to model distributed application systems where a process can suffer queuing delays both at its node's devices and at its software servers.
- ❑ Statistically identical processes form a group or class.
- ❑ Groups at level L are permitted to visit only groups at level L-1. Hardware devices are at the lowest level.
- ❑ Assumption: No cycles in the graph.
- ❑ Model parameters like arrival rate, average number of visits, service times are estimated from measurements.
- ❑ Method of layers, an iterative technique based on approximate MVA is used to find the output parameters.

CS 756

15

## Layered Queuing Model for web server [6]



CS 756

16

## References

- ❑ [1] Menasce & Almeida, Capacity Planning for Web Performance: metrics, models, and methods, Prentice Hall, 98.
- ❑ [2] Mark E. Crovella and Azer Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", IEEE/ACM Trans. Networking, Vol. 5, No. 6, Dec 1997
- ❑ [3] Arlitt Martin F. Arlitt, Carey L. Williamson, "Internet Web servers: workload characterization and performance implications", IEEE/ACM Trans. Networking, Vol. 5, No. 5, Oct 1997
- ❑ [4] Cao, Andersson, Nyberg, Kibl, "Performance modeling of an Apache Web Server with Bursty Arrival Traffic", Lund Institute of Technology, Sweden, 2003.
- ❑ [5] Cherkasova, Phaal, "Session Based Admission Control: a Mechanism for Improving Performance of Commercial Web Sites", IEEE, 1999.
- ❑ [6] Dilley : J. Dilley, R. Friedrich, T. Jin, and J. Rolia, "Web server performance measurement and modeling techniques," Performance Evaluation, vol. 33, pp. 5--26, 1998.
- ❑ [7] Rolia, Sevcik, "The Method of Layers", IEEE Trans. on Software Engg, Vol 21, No. 8, August 1995.