WICSA 07 Tutorial

Performance Analysis of Distributed Software Systems: Approaches based on Queueing Theory
Varsha Apte
January 7, 2007, (Half Day – AM)

A distributed software system uses a complex system of resources which work together to process a request. A typical web-based request flows through various servers including Web servers, database servers, Java application servers, etc, deployed on various hardware platforms. Such a request encounters various forms of delays at and between these servers: communication delay, processing delay and queueing delay. Queueing delay is incurred at every point where there is any contention for resources, e.g. for acquiring a thread, or the CPU, or the lock to a log file. Queueing delay depends on the rate at which requests arrive for that particular resource, which in turn depends on the user behavior, the flow of the request, and the deployment of the servers. Given the number and type of soft and hard resources that make up a distributed system, it is a non-trivial task to quantify these delays. To address this problem, a number of methodologies and tools have been proposed, which allow a distributed system to be specified at a high level, and which generate and solve an underlying model using queueing theory techniques, to answer questions such as what the response time of a request is, what the bottleneck server is, and so on.

In this tutorial we will review the state-of-the art in methods and tools for modeling and analyzing distributed software systems. This includes:

- Motivating examples of web-based multi-tier server systems
- Queueing systems primer (M/M/c/K, M/G/1 etc).
- Simple examples of application of queueing theory to software servers
- Introductory example of the "layered queueing network" method
- Overview of generalized software performance modeling methods and tools
- Real-life applicability of modeling methodologies: comparisons with measured performance.