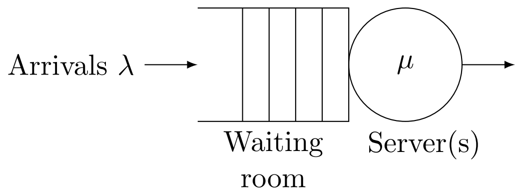
Capacity planning

Franck Jeannot, *Montreal, April 2015, A04, v1.1*

Capacity planning in the IT world is fundamental and despite an important amount of studies in this field and in queueing theory, few IT companies are using best practices and spend the proper time on modelisation of their needs and monitoring of their resources.

1 Queueing theory

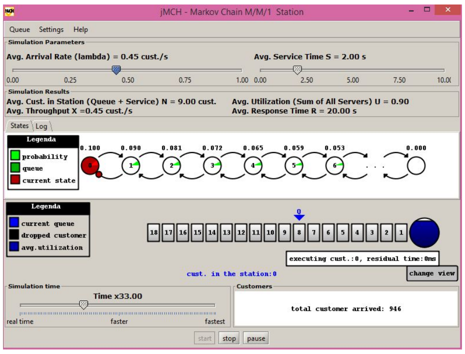
Queueing theory is the mathematical study of waiting lines, or queues. In queueing theory a model is constructed so that queue lengths and waiting time can be predicted. Queueing theory is generally considered a branch of operations research because the results are often used when making business decisions about the resources needed to provide a service [2].
Consider a single station queueing system as shown in below figure. This is also called a single stage queue. There is a single waiting line and one or more servers. A typical example can be found at a bank or post office. Arriving customers enter the queueing system and wait in the waiting area if a server is not free (otherwise they go straight to a server). When a server becomes free, one customer is selected and service begins. Upon service completion, the customer departs the system. Few key assumptions are needed to analyze the basic queueing system[3].



In order to standardize description for queues, Kendall developed a notation with five fields: **AP/ST/NS/Cap/SD**. In the **Kendall** notation, **AP** denotes arrival process characterized by the inter-arrival distribution, **ST** denotes the service time distribution, **NS** is the number of servers in the system, **Cap** is the maximum number of customers in the whole system (with a default value of infinite), and SD denotes service discipline which describes the service order such as First Come First Served (**FCFS**) which is the default, Last Come First Served (**LCFS**), Random Order of Service (**ROS**), Shortest Processing Time First (**SPTF**), etc....

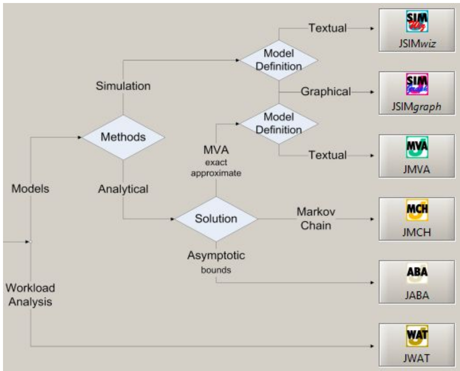
2 Capacity planning

Capacity planning is the process of determining the production capacity needed by an organization to meet changing demands for its products [2]. In the IT ecosystems the goal is to anticipate the need of resources essentially.
Capacity Planning for Server Farms is complex and is based on Multiserver priority queues.
What do per-class mean response times look like for a multi-server system? How do these compare with those for a single-server system? In this case the difficulty is the need for a Markov chain which grows unboundedly in m dimensions, where m is the number of classes.



3 Application and JMT

Ongoing work in the software performance modeling community has significantly stressed the importance of developing automated or semi-automated frameworks for performance optimization and management of complex applications.
Java Modelling Tools(JMT) is a suite of open source applications for performance evaluation and workload characterization of computer and communication systems based on queueing networks.



An example of very interesting tool to approach Markov chains is JMCH.
JMCH application is a graphical simulator of **M/M/c** and **M/M/c/K** queues. The simulation state is visualized both on the queue buffer and on a Markov model representing the system state [5]. The Java Modeling Tools suite (JMT) is proposed as tool to visually help software and system performance engineers to predict the performance of a system and quickly answer what-if questions. JMT is released as an open source tool suite. JMT consists of six applications that communicate using XML with a core algorithmic module composed by the simulation engine (JSIMengine) and by a library of analytical functions for performance model evaluation.

References

[1] Wikipedia capacity planning. http://en.wikipedia.org/wiki/Capacity_planning.

[2] Wikipedia queueing theory. http://en.wikipedia.org/wiki/Queueing_theory.

[3] N. Gautam. *OPERATIONS RESEARCH AND MANAGEMENT SCIENCE HANDBOOK*. A. Ravi Ravindran, Dept. of Industrial and Systems Engineering, Texas A&M University, College Station, 2006. Chapter 9.

[4] Giuseppe Serazzi Giuliano Casale. *Quantitative System Evaluation with Java Modeling Tools (Tutorial Paper)*.

[5] G. Serazzi. M. Bertoli, G. Casale. User-friendly approach to capacity planning studies with java modelling tools. pages 1–9. SIMUTools, ACM, 2009.

[6] Larry W Dowdy Daniel A Menascé Virgílio, AF Almeida. *Capacity planning and performance modeling: from mainframes to client-server systems*. Prentice-Hall, Inc., 1994.