Departamento de Informática

Disciplina: Modelagem Analítica do Desempenho de Sistemas de Computação

Introdução e Conceitos Básicos

Prof. Sérgio Colcher
colcher@inf.puc-rio.br

Site

- **Todo o material da disciplina ficará disponível em:**
  - **http://www.inf.puc-rio.br/~inf2511**

Forma de Avaliação

- **Listas (quase semanais)**
  - *Exercícios*
  - *Trabalhos*
    - Implementações
    - Questões a serem resolvidas a partir da leitura de algum artigo
      - *Estado da arte*
      - *Histórico*
- **Uma prova (?)**
  - *Se necessário*

Motivação

IT Systems Ubiquity

- **IT systems are becoming increasingly ubiquitous and help support most aspects of everyday life.**
- **The Internet has helped accelerate the rate at which IT is integrated into most social systems.**
- **People rely on IT systems to address most of their major human and social concerns**
  - ***health, education, entertainment, access to communication services, access to customer support, finances, safety, privacy, access to government services, and travel.***

IT QoS

- **The various concerns of individuals and of the society as a whole may face major breakdowns and incur high costs if IT systems do not meet the Quality of Service (QoS) requirements of performance, availability, security, and maintainability that are expected from them.**
  - ***Examples***
    - A call to the emergency number has to be answered by a dispatcher in a few seconds or human life may be endangered.
    - When the stock market goes through periods of extreme ups and downs, a large number of online traders tend to flock to online trading sites, causing potential problems due to overloaded and non-responsive systems.
      - *The inability to trade in a timely manner may cause substantial financial losses.*
    - During health crises, such as the outbreak of new diseases, people need to get easy and fast access to health insurance companies to obtain authorization to be admitted to a hospital or to undergo a medical procedure.

IT QoS (cont)

- ***Examples (cont)***
  - Major infrastructures, such as the telephone and cellular networks may become overloaded as their capacity to process calls is stretched, impairing the responsiveness of such systems.
    - *This infrastructure has to be properly designed and sized to handle the extraordinary demands of specific ocasions .*

Congestion

- **Most people need to interact with automated or semi-automated customer support systems and expect near immediate response.**
  - ***Unfortunately, it is not uncommon for someone to be placed on hold for dozens of minutes before being connected to a human being who will take care of a problem or provide the needed information.***
    - These situations cause significant frustration and are a major cause for companies to lose customers.
- **The number of people signing up for access to a wide variety of communication services such as wireless and Internet access services is increasing at exponential rates.**
  - ***The growth in traffic has not been met by an adequate growth in system capacity.***
  - ***As a result, callers may hear the unpleasant recording "all circuits are busy, please try your call later," when trying to place a call.***
- **People have come to expect 24 / 7, instantaneous, and extremely reliable services.**

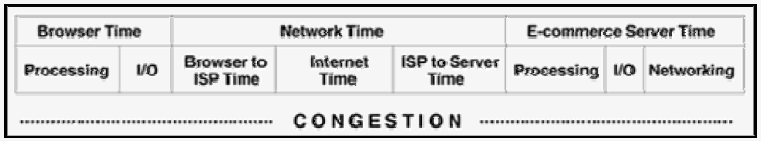
Attributes of an IT System

- IT systems touch people everywhere and every effort must be made to ensure that they operate efficiently, reliably and dependably so that they meet the needs of society and complement the capabilities of users.
- We will briefly discuss the following QoS attributes of an IT system:
  - *response time,*
  - *throughput,*
  - *availability,*
  - *reliability,*
  - *security,*
  - *scalability, and*
  - *extensibility.*

Response Time

- The time it takes a system to react to a request is called the response time.
  - *An example is the time it takes for a page to appear in a browser with the results of a search of the catalog of an online bookstore.*
  - *The response time, usually measured in seconds, and may be broken down into several components.*

Breakdown of Response Time



- The Figure shows the three major components of the response time of a search request to an ecommerce site: browser time, network time, and server time.
  - *The browser time includes the processing and I/O time required to send the search request and display the result page.*
  - *The network time component includes the time spent in the transmission from the browser to the user's Internet Service Provider (ISP), the time spent in the Internet, and the time spent in communication between the ISP at the e-commerce site and its server.*
  - *The third component includes all the times involved in processing the request at the e-commerce site, all the I/O time, the networking time internal to the e-commerce site.*
- Any of the three components include the time spent waiting to use various resources (processors, disks, and networks). This is called congestion time.
  - *The congestion time depends on the number of requests being processed by a system.*
    - The higher the number of requests in the system, the higher the congestion time.
      - *We will learn how to compute the congestion time through the use of performance models.*

Throughput

- The rate at which requests are completed from system is called throughput and is measured in operations per unit time.
- The nature of the operation depends on the system in question.

Throughput

- **Examples of systems and corresponding typical throughput metrics are given in the following table**

System	Throughput Metric
OLTP System	Transactions per Second (tps) tpm-C
Web Site	HTTP requests/sec Page Views per Second Bytes/sec
E-commerce Site	Web Interactions Per Second (WIPS) Sessions per Second Searches per Second
Router	Packets per Second (PPS) MB transferred per Second
CPU	Millions of Instructions per Second (MIPS) Floating Point Operations per Second (FLOPS)
Disk	I/Os per Second KB transferred per Second
E-mail Server	Messages Sent Per Second

Throughput Metric

- **When considering a throughput metric, one has to make sure that the operation in question is well-defined.**
  - ***For example, in an Online Transaction Processing (OLTP) system, throughput is generally measured in transactions per second (tps).***
  - ***However, transactions may vary significantly in nature and in the amount of resources they require from the OLTP system.***
  - ***So, in order for the throughput value to be meaningful, one has to characterize the type of transaction considered when reporting the throughput.***
    - In some cases, this characterization is done by referring to a well established industry benchmark.
      - *For example, the Transaction Processing Performance Council (TPC) defines a benchmark for OLTP systems, called TPC-C, that specifies a mix of transactions typical of an order-entry system.*
        - The throughput metric defined by the benchmark measures the number of orders that can be fully processed per minute and is expressed in tpm-C

Throughput Example

- **Assume that an I/O operation at a disk in an OLTP system takes 10 msec on average.**
- **If the disk is constantly busy (i.e., its utilization is 100%), then it will be executing I/O operations continuously at a rate of one I/O operation every 10 msec or 0.01 sec.**
  - ***So, the maximum throughput of the disk is 100 I/Os per second.***
- **But if the rate at which I/O requests are submitted to the disk is less than 100 requests/sec, then its throughput will be equal to the rate at which requests are submitted.**
  - ***This leads to the expression***

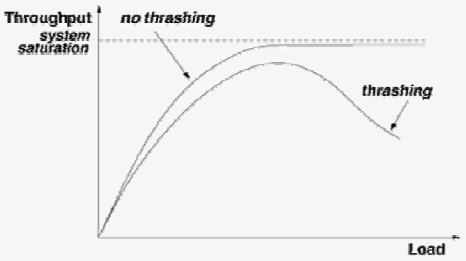
$$Throughput = \min\{ServerCapacity, OfferedWorkload\}$$

Note: This is expression has to be qualified by the assumption that arriving requests do not "change their mind" if the system is busy, as happens routinely in Web sites.

Throughput x Workload

Throughput usually shows an almost linear increase at light loads and then saturates at its maximum value when one of the system resources achieves 100% utilization.

However, in some cases, at high overall loads, throughput can actually decrease as the load increases further. This phenomenon is called thrashing, and its impact on throughput is depicted in the bottom curve.



An example of thrashing occurs when a computer system with insufficient main memory spends a significant amount of CPU cycles and I/O bandwidth to handle page faults as opposed to process the workload.

This may occur because at high loads there are too many processes competing for a fixed amount of main memory. As each process gets less memory for its working set, the page fault rate increases significantly and the throughput decreases. The operating system continuously spends its time handling extra overhead operations (due to increased load), which diminishes the time the CPU can be allocated to processes. This increases the backlog even further, leading to a downward performance spiral that can cripple the system.

Throughput

- An important consideration when evaluating computer systems is to determine the maximum effective throughput of that system and how to achieve it.

Availability

- Imagine that you access an online bookstore and get as a result the page shown below

Availability

- You are likely to become frustrated and may turn to another online bookstore to buy the book you are looking for.
- The consequences of system unavailability can be far more reaching than a loss of customers.
  - *Service interruptions can even threaten lives and property.*

Availability

- Availability is defined as the fraction of time that a system is up and available to its customers. For example, a system with 99.99% availability over a period of thirty days would be unavailable for

$$(1-0,9999)\times(30 days)\times(24h/day)\times(60\,\text{min}/h)=4,32\,\text{min}$$

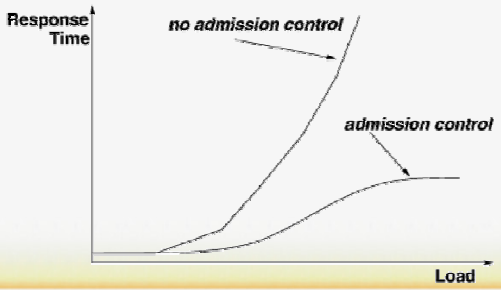
  - *For many systems (e.g., an online bookstore), this level of unavailability would be considered excellent.*
  - *However, for other systems (e.g., defense systems, Emergency services), even 99.99% would be unacceptable.*

Availability

- **The two main reasons for systems to be unavailable are failures and overloads.**
  - ***Failures may prevent users from accessing a system.***
    - For example, the network connection of a Web site may be down and no users may be able to send their requests for information.
  - ***Alternatively, overloads occur when all components are operational but the system does not have enough resources to handle the magnitude of new incoming requests.***
    - This situation usually causes requests to be rejected.
      - *For instance, a Web server may refuse to open a new TCP connection if the maximum number of connections is reached.*
  - ***Failures must be handled rapidly to avoid extended down times.***
    - The first step for failure handling is failure detection.
    - Then, the causes of the failures must be found so that the proper resources (e.g., people and materiel) may be put in place to bring the system back to its normal operational state. T
    - hus, failure handling comprises failure detection, failure diagnosis, and failure recovery.

Controlling Overload: Admission Control

- **One of the reasons for controlling and limiting the number of requests that are handled concurrently by an IT system is to guarantee good quality of service for the requests that are admitted.**
- **This is called admission control and is illustrated in the Figure, which shows two response time curves versus system load.**



Controlling Overload: Admission Control

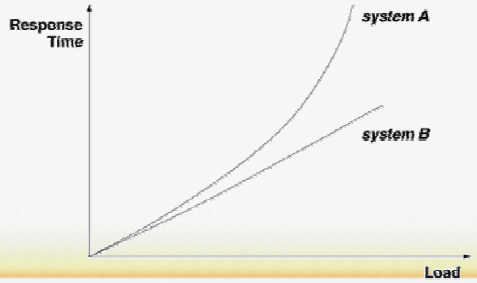
- **If no admission control is used, response time tends to grow exponentially with the load.**
- **In the case of admission control, the number of requests within the system is limited so that response time does not exceed a certain threshold.**
  - ***This is accomplished at the expense of rejecting requests.***
  - ***Thus, while accepted requests experience an acceptable level of service, the reject ones may suffer very large delays to be admitted.***

Security

- **Security is a combination of three basic attributes:**
  - ***Confidentiality: only authorized individuals are allowed access to the relevant information.***
  - ***Data Integrity: information cannot be modified by unauthorized users.***
  - ***Non-repudiation: senders of a message are prevented from denying having sent the message.***
- **To enforce these properties, systems need to implement authentication mechanisms to guarantee that each side in a message exchange is assured that the other is indeed the person they say they are.**
- **Most authentication mechanisms used to provide system security are based on one or more forms of encryption.**
- **Some encryption operations may be very expensive from the computational standpoint.**
- **The tradeoffs between security and performance have been the subject addressed by many research workgroups**

Scalability

- A system is said to be scalable if its performance does not degrade significantly as the number of users, or equivalently, the load on the system increases.
- For example, the response time of system A in increases in a non-linear fashion with the load, while that of system B exhibits a much more controlled growth.
  - *System A is not scalable while system B is.*



Extensibility

- Extensibility is the property of a system to easily evolve to cope with new functional and performance requirements.
  - *It is not uncommon for new functionalities to be required once a new system goes into production.*
    - Even a careful requirements analysis cannot necessarily uncover or anticipate all the needs of system users.
  - *Changes in the environment in which the system has to operate (e.g., new laws and regulations, different business models) may require that the system evolve to adapt to new circumstances.*

System Life Cycle

- Addressing performance problems at the end of system development is a common industrial practice that can lead to using more expensive hardware than originally specified, time consuming performance-tuning procedures, and, in some extreme cases, to a complete system redesign
- It is therefore important to consider performance as an integral part of a computer system life cycle and not as an afterthought.
- The methods used to assure that that QoS requirements are met, once a system is developed, are part of the discipline called Performance Engineering (PE)

Concluding Remarks

- IT systems have become increasingly complex and contain many thousands or even millions of interacting software and hardware components.
  - *Their reach is as encompassing as the air traffic control system for an entire country or an e-commerce system.*
- System designers and analysts often do not take into account QoS requirements when designing and/or analyzing IT systems.
  - *A primary reason for this is a simple lack of awareness about the issues and of the available techniques to consider performance related issues.*

Concluding Remarks

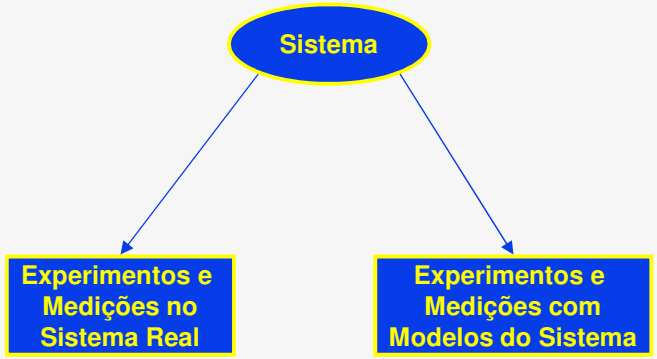
- **We introduced several properties and metrics used to assess the quality of IT systems.**
  - ***Such metrics include response time, throughput, availability, security, scalability, and extensibility.***
- **We also discussed the importance of addressing QoS issues early on in the design stage as opposed to after the system is deployed.**

Modelagem

Sistema

- **Coleção de itens**
  1. ***entre os quais se pode encontrar ou definir alguma relação;***
  2. ***que agem ou interagem para um determinado fim;***
  3. ***que são objeto de estudo ou interesse***
     - definição da fronteira do que é parte do sistema e o que é externo
       - *o que o sistema engloba depende do interesse do estudo ou das questões que se quer responder*

Formas de Estudo de um Sistema



Sistema Real x Modelo

- **Muitas vezes estudar ou medir o sistema real é muito caro, inviável ou até impossível**
  - ***muitas vezes os instrumentos ou ferramentas de medição têm uma ação destrutiva ou interferem de tal forma no sistema que ele deixa de representar o sistema em funcionamento normal***
  - ***o sistema ainda não existe e***
    - o seu dimensionamento pode depender do estudo
    - testar o sistema depois de pronto pode oferecer riscos de segurança
      - *Testar o desempenho do sistema de alarme de uma usina nuclear*
  - ***o sistema não é facilmente observável***
    - medições de fenômenos a nível atômico, experimentos destinados a pesquisas espaciais, ...

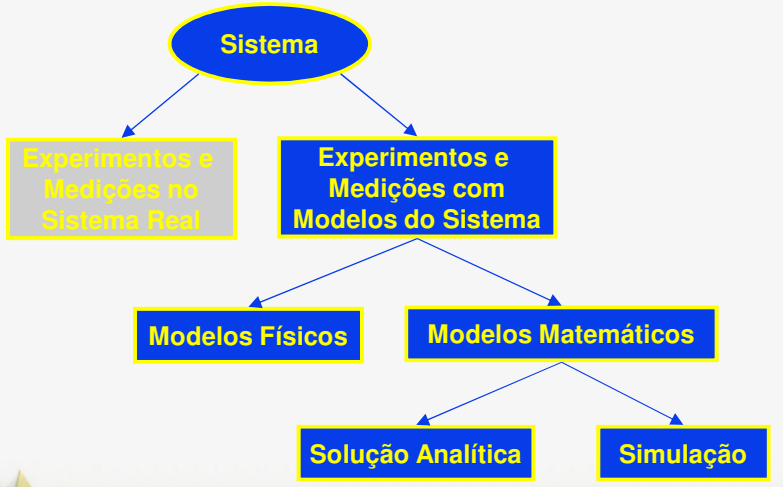
Modelo

- **Descrição de um Sistema**
  - ***Exemplos***
    - Princípios descritos por equações matemáticas
    - Maquetes
    - Programas
  - ***Um modelo é, na verdade, uma abstração do sistema, isto é***
    - A descrição diz respeito tão somente aos aspectos ou o propósito do estudo, medição ou função desejada

Algumas Observações Sobre Modelos e o Processo de Modelagem

- **Modelos são freqüentemente confundidos com o sistema sendo modelado**
  - ***Esse é um abuso de linguagem que podemos até nos dar ao luxo de cometer desde que estejamos completamente conscientes e prevenidos sobre a limitação do modelo em relação ao sistema real***
- **Modelagem é um processo complexo**
  - ***Considerado como uma mistura de ciência e arte***

Formas de Estudo de um Sistema



Tipos de Modelo

- **Modelos Físicos**
  - ***Construção que procura reproduzir ou mimetizar todo o funcionamento observável do sistema***
- **Modelos Matemáticos**
  - ***Para solução analítica***
    - permitem a formulação de equações através das quais o comportamento do sistema pode ser obtido ou medido
      - *Atribui-se valores às variáveis e, em seguida, efetua-se a resolução das equações*
      - *As medidas obtidas são, em geral, de desempenho.*
  - ***Para simulação***
    - permitem a formulação de estruturas que podem ser exercitadas de forma a mimetizar o comportamento do sistema ou obter medidas (de desempenho, p. ex.) sobre o seu funcionamento

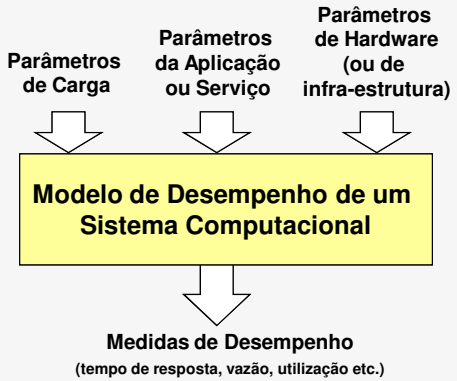
Modelagem Analítica

- **Permitem a formulação de equações através das quais o comportamento do sistema pode ser obtido**
- **Várias técnicas, dentre as quais:**
  - ***Teoria das Filas***
    - Teoria da probabilidade, processos estocásticos etc.
    - "Network Calculus"
      - *Tratamento "determinístico"*
        - Limites ("upper bounds")
      - *Extensões para tratamento estocástico*

Exemplos Típicos de Modelagem

- **Garantias de desempenho em redes de comunicação**
  - ***Planejamento em redes Comutação de Pacotes e suas "vertentes"***
    - Internet
      - *Serviços diferenciados e Integrados*
    - ATM
      - *Controle de Admissão*
  - ***O problema do "playout buffer"***
    - Compensação da variação estatística do retardo
- **Problemas de escalonamento**
  - ***Algoritmos para escalonamento de processos com requisitos de desempenho, deadlines etc.***
    - Controle de admissão
- **Planejamento de capacidade em servidores**
  - ***Quantidade de disco, memória, CPU, cache etc.***

Modelo de Desempenho



- IT systems are becoming increasingly ubiquitous and help support most aspects of everyday life.
- The Internet has helped accelerate the rate at which IT is integrated into most social systems.
- People rely on IT systems to address most of their major human and social concerns
  - *health, education, entertainment, access to communication services, access to customer support, finances, safety, privacy, access to government services, and travel.*
- The various concerns of individuals and of the society as a whole may face major breakdowns and incur high costs if IT systems do not meet the Quality of Service (QoS) requirements of performance, availability, security, and maintainability that are expected from them.

Exemplo: Sistema de Atendimento Bancário

- Ao chegar no banco, cada cliente deve esperar em uma fila (única) até que chegue a sua vez de ser atendido.
- Quando chega a sua vez, o cliente se aproxima do guichê de um dos *n* atendentes (caixas), realiza suas operações e, quando se sente satisfeito, vai embora dando oportunidade para que o próximo possa ser atendido.
- Medidas de desempenho desejadas:
  - *Tempo médio de resposta*
    - Tempo médio desde que chega até ir embora
  - *Tempo médio de espera*
  - *Utilização dos caixas*
    - Porcentagem do tempo em que um caixa permanece ocupado
  - *Vazão dos Caixas*
    - Número médio de clientes servidos por unidade de tempo.

Elementos de Modelagem

- Recursos
  - *Representam algum insumo disponível no sistema*
  - *Podem ser*
    - Ativos
      - *Prestam serviços*
        - CPU, Disco, caixa de um banco, enlace de transmissão de uma rede etc.
    - Passivos
      - *São consumidos ou utilizados pelo sistemas*
        - Memória, etc.
- Usuários
  - *Representam as entidades que trafegam pelo sistema consumindo seus recursos*
    - Processos, transações, clientes que chegam a um banco

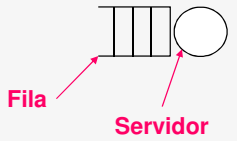
Elementos de Modelagem

- Freqüentemente, recursos de um sistema computacional são, de alguma forma, compartilhados
  - *Caixas de banco, Discos, CPU, Enlaces físicos de uma rede etc.*
- Os elementos utilizados na modelagem devem refletir a existência de *contenção* ou *espera* pelos recursos

Modelos Baseados em Redes de Filas

Redes de Filas

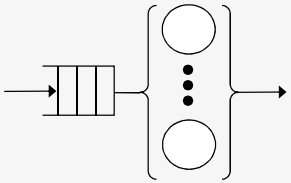
- **Uma rede de filas consiste de**
  - ***Uma ou mais entidades denominadas Centros de Serviço***
    - Cada centro de serviço consiste de
      - *Um ou mais servidores*
      - *Uma ou mais áreas de espera (filas)*
        - Obs.: é comum também chamar-se o centro de serviço inteiro de fila
  - ***Exemplo de rede de filas***
    - Caixa de banco atendendo clientes
      - *Um único centro de serviços*



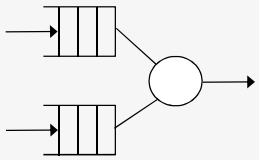
- **Um usuário corresponde a uma entidade que circula pelo sistema modelado**
  - ***Exemplo: cliente do banco***

Centros de Serviço

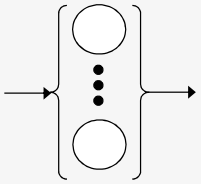
1 classe
1 servidor



1 classe
n servidores



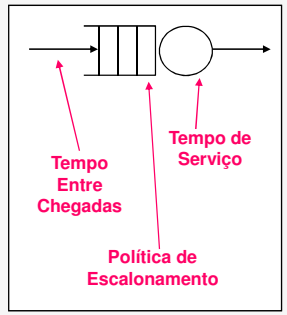
2 classes
1 servidor



Servidor infinito

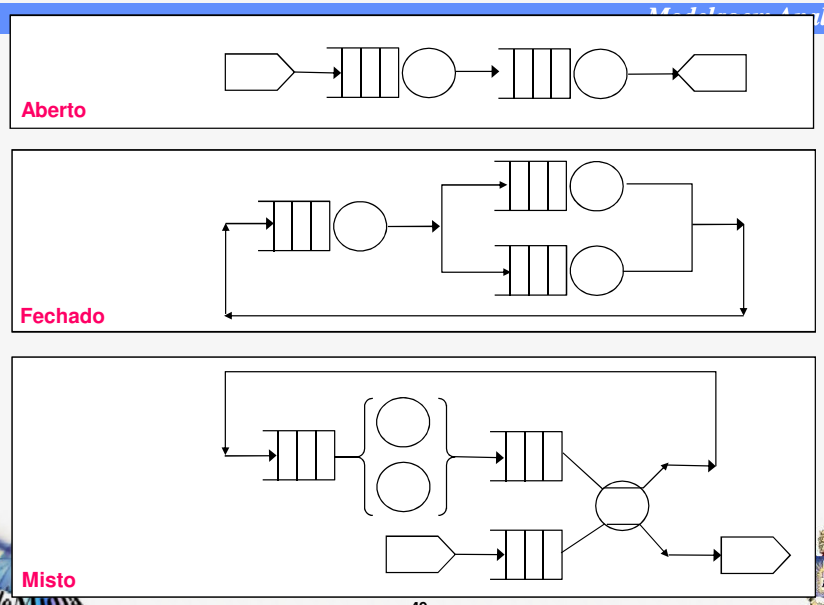
Centros de Serviço

- **Usuários chegam à área de espera**
  - ***Apresentam um padrão de chegadas***
    - Tempo entre chegadas
      - *Pode ser caracterizado por uma variável aleatória*
  - ***Demandando um serviço***
    - Servidor tem um padrão de tempo de atendimento
      - *Pode ser caracterizado por uma variável aleatória*
      - *Pode depender da classe de serviço e do usuário específico*
      - *Pode depender do tamanho da fila*
      - *...*

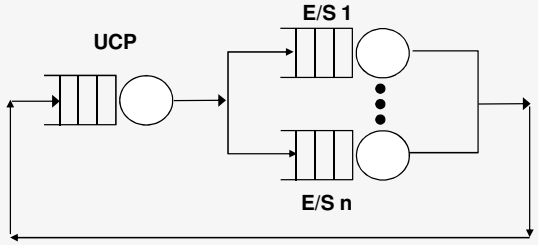


- **A área de espera pode ser servida seguindo diferentes políticas ou algoritmos de escalonamento**
  - ***Com prioridades, preemptivamente, ...***

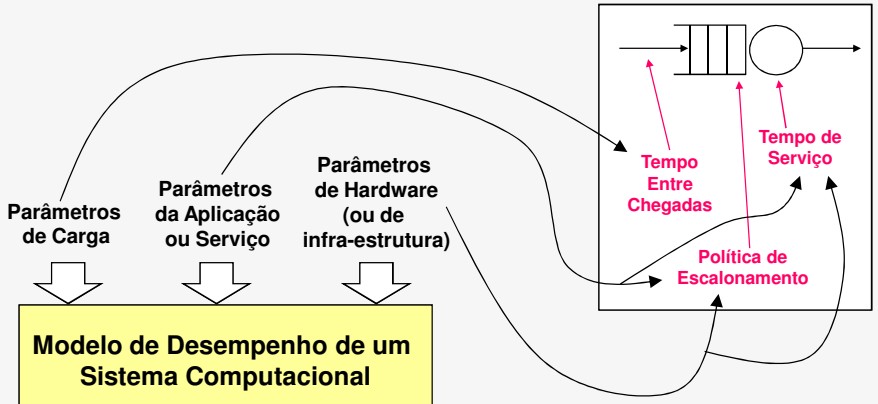
Tipos de Modelos de Redes de Filas



Exemplo: Sistema de Computação em *batch*



Modelo de Desempenho



Bibliografia

- **Network Calculus**
  - Network Calculus : A Theory of Deterministic Queuing Systems for the Internet, ***Jean-Yves Le Boudec e Patrick Thiran, Lecture Notes in Computer Science 2050, Springer-Verlag, 2002***
  - ***Vários papers distribuídos ao longo do curso***
- **Teoria das Filas**
  - Queueing Systems Vol 1 e 2, ***Leonard Kleinrock, John Wiley & Sons Inc, 1975***
  - Modeling and Analysis of Stochastic Systems, ***Vidyadhar G. Kulkarni, Chapman&Hall, 1995***
  - Elements of Queueing Theory: Palm martingale Calculus and Stochastic Recurrences, ***François Baccelli e Pierre Brémaud, Springer, Second Ed., 2003***
- **Análise de Desempenho em Geral e Planejamento de Capacidade**
  - The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, ***Simulation and Modeling, Raj Jain, John Wiley & Sons Inc, 1991***
  - Capacity Planning for Web Services: Metrics, Models, and Methods, ***Daniel A. Menascé e Virgílio A. F. Almeida, Prentice Hall, 2001***
  - Capacity Planning and Performance Modeling: From Mainframes to Client-Server Systems, ***Daniel A. Menascé, Virgílio A. F. Almeida e Larry W. Dowdy***

Bibliografia (Cont.)

- **Simulação**
  - Simulation Modeling and Analysis, ***Averil M. Law e W. David Kelton, 3a. Ed., McGrawHill 2000.***
  - Discrete-Event System Simulation, ***Jerry Banks, John S. Carson II and Barry L. Nelson, 2nd. Ed., Prentice-Hall 1999.***
- **Tecnologia Básica de Redes**
  - Redes de Computadores: Das LANs, MANs e WANs às Redes ATM, ***Luiz F. G. Soares, Guido Lemos e Sérgio Colcher, 2a. Edição, Ed. Campus, 1995***
- **Análise de Desempenho de Redes**
  - Queueing Systems ***Vol. 2, Leonard Kleinrock, John Wiley & Sons Inc, 1975***
  - Data Networks, ***Dimitri Bertsekas e Robert Galager, 2a. Ed, Prentice Hall, 1992***
  - Computer Networks and Systems: Queueing Theory and Performance Evaluation, ***Thomas G. Robertazzi, 3a. Ed, Springer Verlag, 2000***
  - Introduction to IP and ATM Design and Performance with Applications Analysis Software, ***J. M. Pitts e J. A. Schormans, 2a. Ed., Wiley 2000.***

Bibliografia (Cont.)

- **Básico de Modelos Probabilísticos e Processos Estocásticos**
  - Queueing Systems ***Vol 1, Leonard Kleinrock, John Wiley & Sons Inc, 1975 (Apêndices).***
  - Probability and Random Process for Electrical Engineers, ***Yannis Viniotis, McGrawHill 1998.***
  - Stochastic Models, ***In Handbooks in Operations Research and Management Science, Vol 2, DP Heyman and M.J.Sobel (Eds), North-Holland, 1990.***
  - An Introduction to Probability Theory and its Applications, Vol 1 e 2, ***William Feller, John Wiley & Sons, 1968.***