Análise de Desempenho de Sistemas de Computadores Usando a Teoria de Filas

Disciplina: Administração e Segurança de Sistemas de Computadores

Curso: Licenciatura em Engenharia Informática

Docentes: Doutor Eng. Lourino Chemane, Engra. Ivone Cipriano e eng.Délcio Chadreca

DEEL, Faculdade de Engenharia, UEM

Abril 2019

Tópicos

1. Introdução a Teoria de Filas
   - Notação
   - Lei de Little
2. Sistemas Computacionais e Rede de Filas
   - Descrição dos clientes
   - Descrição de Servidores
   - Modelos de Rede de Filas e Teoria de Filas
   - Lei de Little
3. Leis Operacionais
   - Leis de Utilização
4. Exemplos/Exercícios

Motivação: Avaliação do Desempenho de Sistemas de Computadores

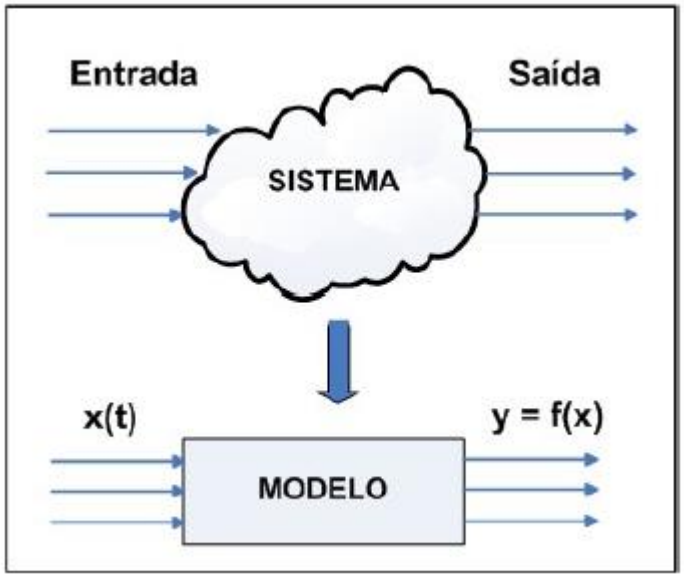
- A avaliação de desempenho está presente em todos os momentos do ciclo de vida de um sistema computacional. Na hora de projectar, produzir ou implantar um sistema, o objectivo final é sempre o mesmo: *escolher de entre diversas alternativas aquela que proporcione o melhor desempenho, com o menor custo possível.*
- Não existe um meio universal com o qual possamos avaliar o desempenho das diversas classes de sistemas computacionais. Cada aplicação possui características próprias, o que faz com que cada estudo de desempenho seja único.
- A caracterização da carga (*workload*), a selecção de métricas, a escolha da técnica de avaliação, tudo isso deve ser feito de maneira específica para cada caso, levando-se em consideração as características do sistema objecto de estudo.

Algumas Aplicações da Análise de Desempenho

- Comparar sistemas entre si e definir qual é o mais adequado aos requisitos de desempenho impostos;
- Prever o desempenho do sistema frente a futuras mudanças na carga ou no próprio sistema;
- Identificar gargalos de desempenho;
- Propor mudanças de configuração de modo a obter um melhor desempenho (*tunning*).

Noções de Sistema e Modelo

- **Sistema**: Segundo a definição do IEEE, um **sistema** é uma combinação de componentes que agem em conjunto para realizar uma função que não seria possível ser feita por qualquer das partes individuais. No nosso caso, o termo sistema será empregue para denotar o conjunto de elementos de hardware e software utilizados para o processamento de uma carga.
- **Modelo:** Um modelo é algo que simula o comportamento de um sistema. É uma abstracção de algo real. A modelagem matemática consiste em definir o conjunto de variáveis de entrada e variáveis de saída associadas a um sistema, bem como estabelecer o relacionamento entre elas.



Abordagem de Análise de Desempenho

- **Definição de objectivos do sistema** consiste em definir exactamente o que será estudado e em que nível.
- **Escolha de métricas** consiste em definir quais critérios serão utilizados para medir o desempenho do sistema. Exemplos de métricas incluem: tempo de resposta, *throughput*, taxa de erros, entre outros;
- **Listagem de parâmetros e selecção de factores** consiste em identificar quais parâmetros afectam o desempenho do sistema e em seguida escolher dentre eles aqueles que sofrerão variação durante o estudo (os factores);
- **Escolha da técnica de avaliação** existem basicamente três técnicas usadas para a avaliação de desempenho de sistemas:
  - Modelagem analítica;
  - Simulações; e
  - Medições.
- A escolha de uma delas depende de uma série de factores que serão discutidos nas próximas aulas;

- **Caracterização da carga (*workload*)** descrever o que (e de que modo) será requisitado do sistema. Por exemplo, em um sistema de base de dados, a carga seria formada por um conjunto de transacções que buscam e/ou actualizam dados;
- **Análise e interpretação de dados** nesse passo, são analisados os dados obtidos em medições e simulações. Mais importantes que a colecta de dados em si, são as conclusões que o analista tira a partir deles;
- **Apresentação de Resultados** por fim cabe ao engenheiro de desempenho comunicar de forma clara os resultados obtidos durante o estudo.

Selecção das Métricas

- Métricas podem ser agrupadas em três categorias:
  - **Reacção: D**iz respeito ao tempo que o sistema leva a dar resposta
  - **Produtividade: R**efere à taxa de execução dos serviços
  - **Utilização: Está** relacionada à utilização de um recurso.
- Algumas das métricas mais utilizadas na avaliação de sistemas computacionais:
  - ***Throughput* (taxa de serviço)** é a taxa em que as requisições são processadas por um sistema. Exemplos:
    - **MIPS** (milhões de instruções por segundo) ou
    - **FLOPS** (instruções de ponto flutuante por segundo), para CPUs;
    - **Transacções por segundo**, para sistemas de Bases de Dados;
    - **bps** (bits por segundo), para redes de comunicação, etc.;

Selecção das Métricas

- Algumas das métricas mais utilizadas na avaliação de sistemas computacionais (Continuação):
  - **Tempo de resposta** corresponde ao intervalo de tempo entre o momento em que uma requisição é submetida ao sistema e o momento em que o sistema responde a requisição;
  - **Utilização** percentagem do tempo em que os recursos de um sistema encontram-se ocupados. O recurso com a maior utilização é dito ser o **gargalo** do sistema;
  - **Taxa de erros** indica a frequência com que o comportamento do sistema se apresenta fora do esperado;
  - **Disponibilidade** a porção de tempo em que o sistema fica à disposição dos usuários para atender às suas requisições;

- A teoria das filas é uma técnica de modelagem analítica bastante utilizada para a avaliação de desempenho de sistemas computacionais.
- O cenário de teoria das filas pode ser descrito desse modo: um recurso (também chamado de **centro de serviço**) é compartilhado por um conjunto de clientes (ou *jobs*) que de tempos em tempos vão a esse recurso para receberem seu serviço.
- Se ao solicitar o serviço o cliente encontrar o recurso disponível, ele será imediatamente atendido. Caso contrário, se o recurso já estiver atendendo outro(s) cliente(s) no momento da solicitação, ele deverá aguardar em uma fila.
- Cenários como esse são muito comuns dentro e fora da computação, o que faz com que a teoria das filas tenha aplicações nas mais diversas áreas, como, redes de comunicação, sistemas telefónicos e, claro, avaliação de desempenho de sistemas computacionais.

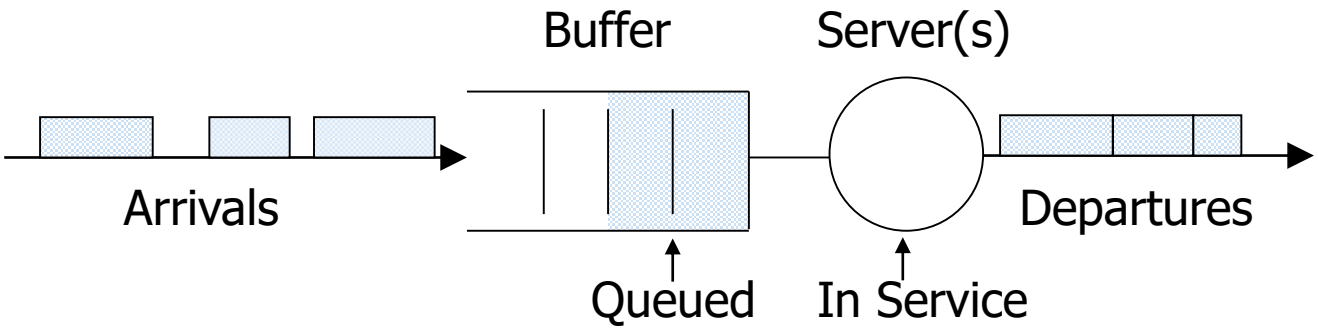
Notação da Teoria de Filas

- A notação hoje amplamente utilizada para descrever filas foi criada por Kendall e é composta de seis parâmetros: A/S/m/B/K/SD
  - **A (*Arrival Process*)** o processo de chegada das requisições ao centro de serviço;
  - **S (*Service Time Distribution*)** a distribuição que determina tempo que o centro precisa para atender as requisições;
  - **M o número de servidores**, que são considerados parte de um mesmo sistema de filas caso sejam todos idênticos;
  - **B (*Buffer*)** o número máximo de clientes que o sistema aceita. Esse número pode ser limitado por questões de espaço ou para limitar o tempo de espera;
  - **K o número de clientes** que podem porventura vir a solicitar serviço do sistema;
  - **SD (*Service Discipline*)** A ordem em que os clientes são atendidos. Alguns exemplos são FIFO (*First-In, First-Out* primeiro a chegar, primeiro a ser servido), LIFO (*Last-In, First-Out* último a chegar, primeiro a ser servido), Round Robin (cada cliente recebe uma mesma quantidade de tempo de serviço no centro), dentre outros;

Notação da Teoria de Filas

- Os processos de chegada e de serviço são representados por uma letra, que indica qual a distribuição que eles seguem:
  - M Exponencial;
  - $M^{[x]}$ Exponencial em rajadas;
  - E_k Erlang, com parâmetro k;
  - H_k Hiper-exponencial, com parâmetro k;
  - D Determinístico;
  - G Geral;
- A letra M, usada para indicar a distribuição exponencial, vem de *memoryless* ou de Markovian, para indicar que valores assumidos no passado não influenciam nos valores futuros.
- Quando vem acompanhado por um sobrescrito, significa que as chegadas (ou serviços) acontecem em rajadas, isto é cada chegada (ou serviço) consiste, na realidade, em um grupo de requisições.
- Nesse caso, x representa o tamanho do grupo, que também é uma variável aleatória.
- A distribuição determinística denota um valor constante.
- Uma distribuição geral significa que não se conhece exactamente qual a distribuição, e o resultado é válido para qualquer distribuição.

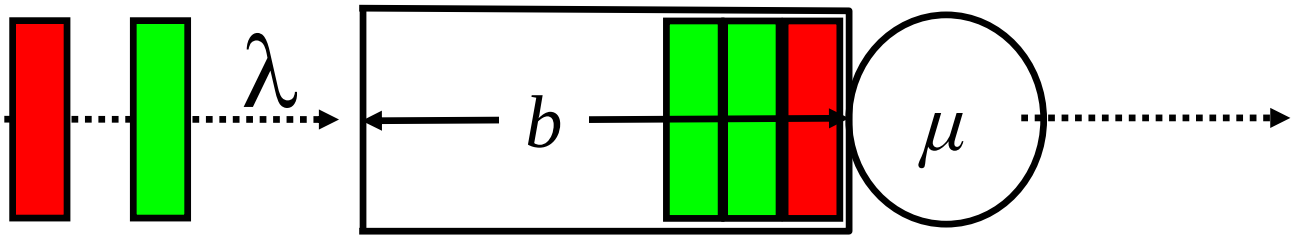
Notação da Teoria de Filas



- Como um exemplo, **M/M/2/50/5000/FIFO**, indica um sistema de fila no qual:
  - Os tempos de chegada e de serviço são exponencialmente distribuídos,
  - tem 2 servidores, e que
  - suporta um máximo de 50 requisições,
  - de uma população total de 5000 clientes,
  - servidos em um regime de primeiro a chegar, primeiro a sair.

Notação da Teoria de Filas

- É comum abreviar a notação, e considerar que não existe limitação de buffer, a população de clientes é infinita e a política de serviço é FIFO. Assim, costuma-se escrever apenas os três primeiros parâmetros para representar uma fila. Os tipos mais comuns de filas são os seguintes:
  - **M/M/1**: esse tipo de fila é usado para representar sistemas de um único processador ou qualquer outro recurso individual;
  - **M/M/m**: esse tipo de fila pode ser usado para modelar centros de serviço compostos por vários dispositivos idênticos, como por exemplo, um subsistema de I/O com vários discos. O sistema é formado por m servidores, cada um com uma taxa de serviço, atendendo a clientes que chegam a uma taxa.



Lei de Little

- A lei de Little é uma das mais importantes leis da teoria das filas.
- Ela estabelece que o número médio de requisições em um sistema é igual ao produto entre a taxa de chegada e o tempo médio que a requisição passa no sistema.
- Apesar de ser uma lei intuitivamente razoável, trata-se de um resultado excepcional, se considerarmos que ele é válido para a grande maioria das situações, não sendo necessário nenhum conhecimento adicional sobre os processos de chegada e saída.
- A única exigência é que o número de requisições que chegam ao sistema seja igual ao número de requisições que têm seus serviços completados. E isso implica que o *throughput* (taxa de serviço) do sistema seja igual à taxa de chegada, donde tiramos:

$$E[N] = \lambda \times E[R]$$

$$E[N] = X \times E[R]$$

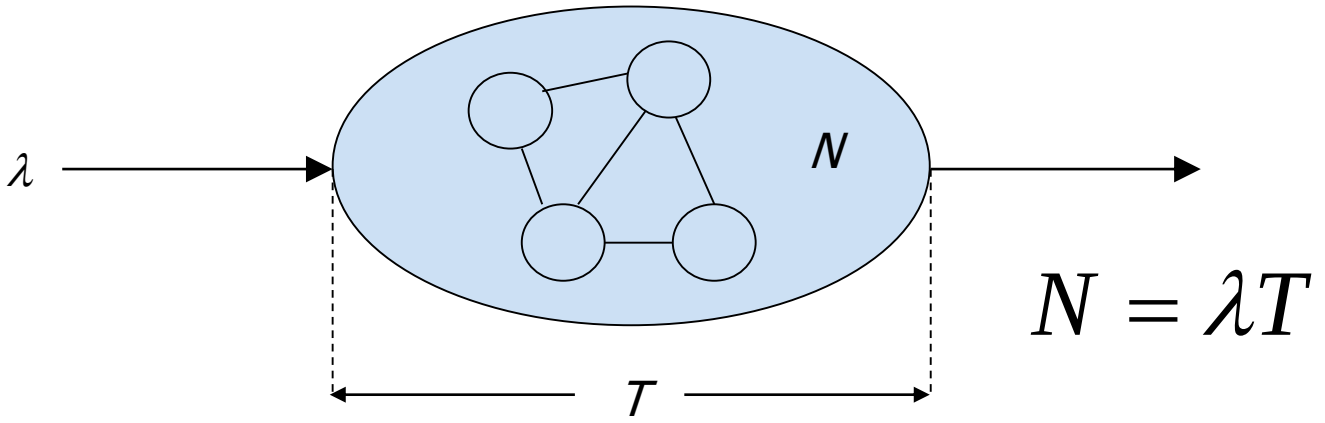
Lei de Little

- Pode-se ainda obter a partir da Lei de Little outros relacionamentos:

$$E[N_q] = \lambda \times E[w_q]$$

$$E[N_S] = \lambda \times E[s]$$

- Onde:
  - $E[N_q]$ Número médio de requisições na fila de espera;
  - $E[w_q]$ Tempo médio de espera na fila;
  - $E[N_s]$ Número médio de requisições recebendo serviço;
  - E[s] Tempo médio de serviço;



$$N = \lambda T$$

Leis Operacionais

- Muitos problemas de teoria das filas podem ser resolvidos por meio de algumas relações simples, sem que para isso seja necessário um conhecimento exacto sobre as distribuições dos tempos de chegada e de serviço.
- Algumas dessas relações, conhecidas como **leis operacionais**.
- Seja um sistema modelado como um único centro de serviço (isto é, o sistema consiste em um único recurso). Se observarmos esse sistema por um determinado período de tempo, podemos mensurar as seguintes quantidades:
  - T o tempo total de observação do sistema;
  - A o número total de requisições que chegaram ao sistema;
  - C o número total de requisições processadas pelo sistema;
  - B o total de tempo em que o sistema encontra-se ocupado atendendo às requisições;

Leis Operacionais

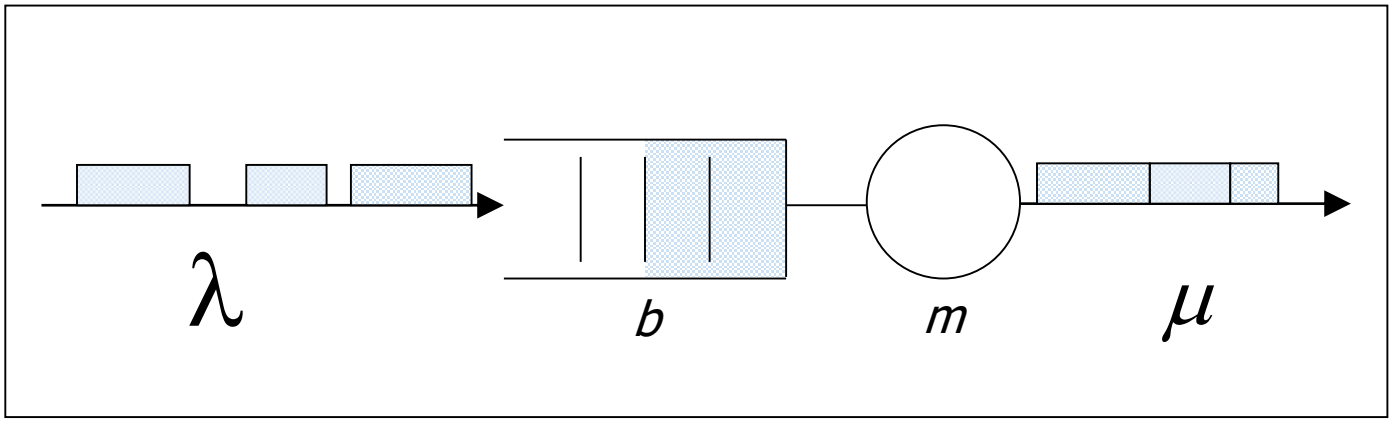
- Esses valores podem ser obtidos directamente, por meio de simples monitorização do sistema daí o termo ***operacional***.
- Eles servem de variáveis independentes em equações que permitem estabelecer um conjunto adicional de quantidades:
- λ a taxa de chegada de requisições. λ = A/T;
- X o *throughput* ou taxa de processamento. X = C/T (ou μ = C/T);
- U utilização do sistema. U = B/T;
- S o tempo de serviço médio de uma requisição. S = B/C;

- **Lei de Utilização**: Outra forma de exprimir a utilização de um recurso. Esta relação é chamada de lei de Utilização

$$U = \frac{B}{T} = \frac{B}{C} \times \frac{C}{T} \Rightarrow$$

$$\mathbf{U = X \times S}$$

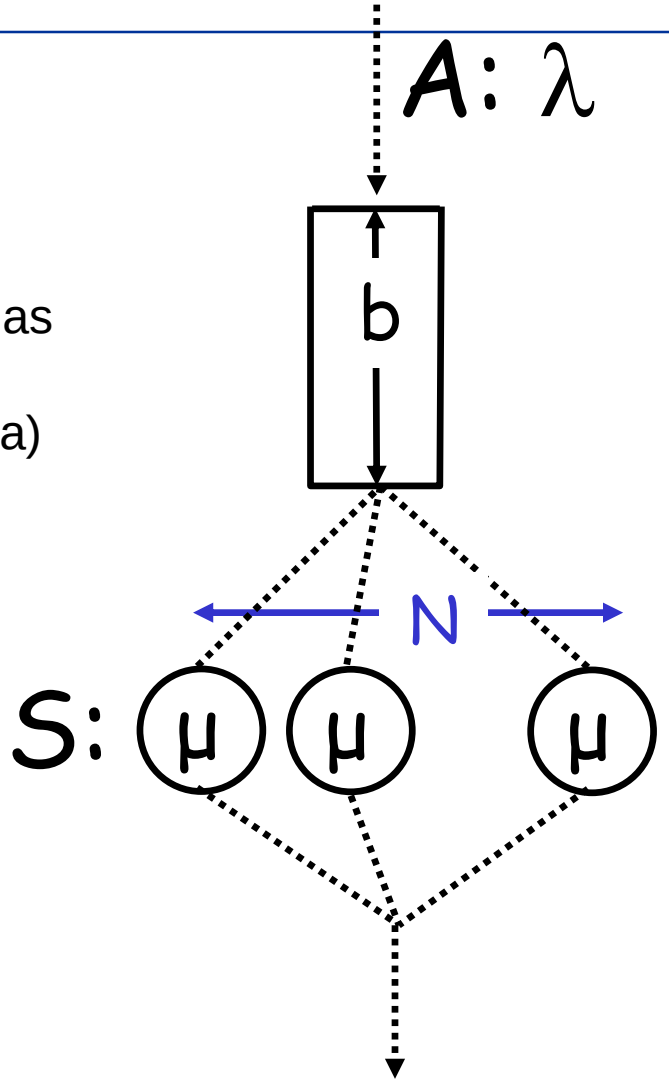
Caracteristicas da Fila



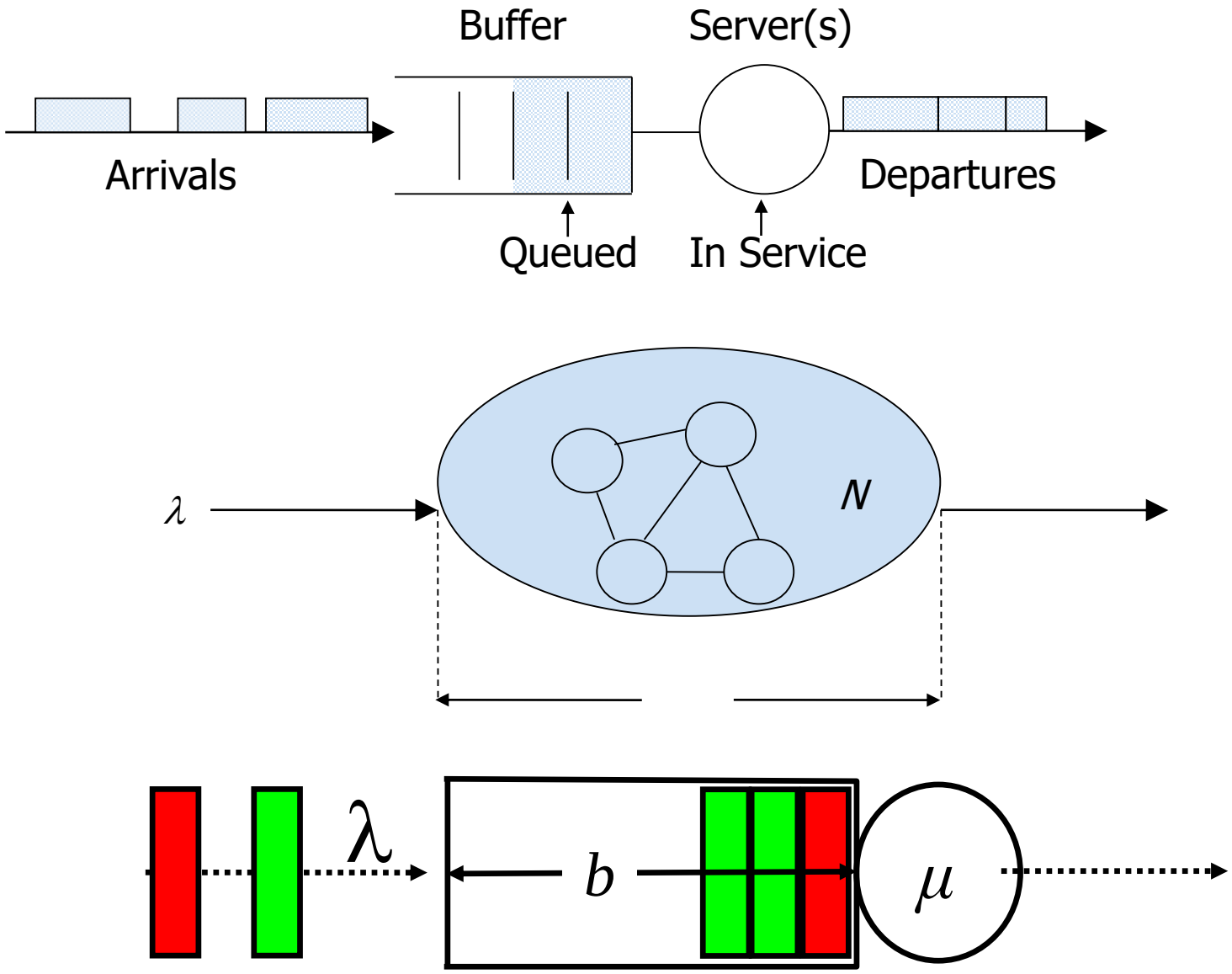
- Número de Servidores m: um, múltiplo, infinito
- Tamanho do *Buffer* b
- A Ordem em que os clientes são atendidos (*scheduling*): FCFS, LCFS, etc.
- Taxa de chegada de Pacotes (clientes) λ
- Taxa de disponibilização de serviço μ

Notação de Kendall's: A/S/N[/b]

- Notação para tipo de filas
- A Descreve o processo de chegada
  - M = Markov/*Memoryless*; Poisson
  - D = determinístico (Tempo constante entre as chegadas)
  - G = Geral (qualquer outra forma de chegada)
- S Descreve o processo de serviço
  - M [exponencial],D,G: como acima
- N É o número de processadores paralelos
- b (opcional) É o tamanho da fila
  - Descarta pacotes quando está cheio



Diagramas de Filas de Sistemas de Computadores



Descrição da Fila: Examples

- M/M/1: Chegadas de *Poisson* (*arrivals*), tempos de serviço com distribuição exponencial, um servidor, *buffer* infinito
- M/M/m: o mesmo como acima mas com *m* servidores
- M/G/1: Chegadas de Poisson, tempos de serviço com distribuição Geral, um servidor, *buffer* infinito
- */D/∞ : Sistema com atraso constante

Distribuição Exponencial

- Função continua com variável x tem distribuição exponencial com parâmetro m se:
-

$$F_X(x) = P\{X \le x\} = \begin{cases} 1 - e^{-\mu x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- Então:

$$f_X(x) = \begin{cases} \mu e^{-\mu x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- Pode Modelar:
  - Intervalo entre as chegadas (*Inter-arrival times*) para processos de chegada de Poisson
  - Tempo de serviço (*Service Process Time*)

Exemplos/Exercícios

Exemplo 1: Mensagens chegam de forma independente num sistema de computadores a uma taxa de 10 por minuto. Os seus cumprimentos são exponencialmente distribuídos com uma média de 3600 caracteres. Elas são transmitidas num canal de 9600 bps. Um caracter tem um cumprimento de 8 bits.

a)Qual é a taxa de serviço?

b)Determine o tempo médio de serviço.

c)Qual é o taxa de chegada A?

d)Qual é a utilização do servidor?

e)Qual é a probabilidade de que há duas (2) mensagens no sistema?

f)Qual é o número médio de mensagens na fila?

g)Qual é o número médio de mensagens no sistema?

h)Qual é o tempo médio de espera na fila?

i)Determine o tempo médio que a mensagem leva no sistema?

End

Lourino Chemane
Contacto: chemane@infopol.gov.mz