Universidade Federal da Paraíba (UFPB)
Centro de Ciências Aplicadas e Educação (CCAE)
Curso: Bacharelado em Sistemas de Informação
Disciplina: Avaliação de Desempenho de Sistemas
Professor: Marcus Carvalho

Aluno:

Prática 2: Teoria das Filas / Simulação

Nesta atividade de laboratório é apresentado um problema de teoria das filas, envolvendo um simulador de um sistema web. Você deve executar o programa fornecido e avaliar o desempenho do mesmo, a partir das métricas gravadas pelo programa. Você também deve aplicar os conceitos de teoria das filas para resolver as questões de forma analítica. Para responder às questões, você deve escrever um relatório com as suas análises.

Deve ser enviado pelo Google Classroom o relatório, de preferência neste arquivo do Google Docs contendo as respostas das questões com as análises, além de um arquivo *ZIP* com os dados coletados na medição dos programas. É **fortemente recomendado** o uso de gráficos nos relatórios, para a exibição dos dados coletados e para ajudar na sua análise. Tabelas também devem ser usadas para exibir dados quando necessário.

No início do relatório, descreva a configuração da máquina na qual você está executando cada análise de desempenho, com informações sobre a CPU, Memória, Cache, Disco, etc.

Problema

O programa deste laboratório é um simulador de um sistema Web, que possui **n** servidores para processar as $\mathbf{A_0}$ requisições realizadas no tempo de observação **T** escolhido. Para executá-lo, você precisa passar a seguinte sequência de parâmetros:

No Linux

```
java -cp bin:lib/* ServidorWeb <taxa-de-chegada-media>
<tempo-de-servico-medio> <num-servidores> <tempo-observacao>
```

No Windows

```
java -cp bin;lib\* ServidorWeb <taxa-de-chegada-media>
<tempo-de-servico-medio> <num-servidores> <tempo-observacao>
```

Os parâmetros de entrada do programa são:

1. Taxa de chegada de requisições que são realizadas para o sistema web, em requisições por segundo.
2. Tempo médio de serviço para o processamento de uma requisição, em segundos.
3. Número de servidores (**n**) que serão usados para processar as requisições. Na implementação, cada servidor é uma thread do sistema, sendo possível processar até **n** requisições simultaneamente no sistema web.
4. Tempo durante o qual o sistema web será observado e medido na simulação.

Como saída, a execução do programa vai gerar as seguintes saídas na tela:

- **TaxaDeChegadaMedia**: o valor do parâmetro de taxa de chegada média de requisições, que foi passado como entrada.
- **TempoDeServicoMedio**: o valor do parâmetro de tempo serviço médio de requisições, que foi passado como entrada.
- **NumServidores**: o valor do parâmetro de número de servidores, que foi passado como entrada.
- **RequisicoesSubmetidas**: quantidade de requisições que foram submetidas (ou seja, que chegaram) ao servidor web durante o tempo de observação.
- **RequisicoesConcluidas**: quantidade de requisições que foram finalizadas (ou seja, completaram sua execução) durante o tempo de observação.
- **TempoMedioDeResposta**: tempo médio de resposta para as requisições executadas no servidor web.
- **TamanhoMedioDaFila**: tamanho médio da fila de requisições acumuladas no sistema web durante a execução.

Você deve responder às questões abaixo explorando diferentes configurações de parâmetros de entrada, analisando as saídas e usando as técnicas de modelagem analítica (teoria das filas) aprendidas na disciplina.

Especificações da Máquina

Processador: Intel(R) Core(TM) i3-7020U CPU @ 2.30GHz 2.30 GHz

RAM: 12GB

Disco: 1T

Perguntas

1. Qual modelo de filas é mais adequado para analisar o desempenho deste sistema web? Indique quais os parâmetros da notação de Kendall seriam mais adequados e faça um desenho do modelo de filas em questão.

O melhor modelo de filas é o com uma única fila e múltiplos servidores.

Notação de Kendall -> M/M/m

Taxa de chegada com distribuição Exponencial

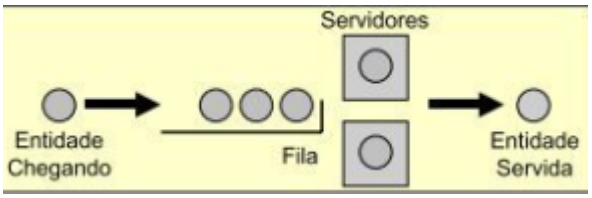
Tempo de serviço em distribuição Exponencial

Múltiplos servidores

Fila sem limitação

População ilimitada

Disciplina de serviço FIFO



2. Como o tempo de resposta das requisições varia ao aumentar a carga do sistema (ou seja, aumentar a taxa de chegada de requisições no sistema)? Você considera esse sistema web escalável? Considere para esta questão o seguinte cenário base de parâmetros de entrada:
    - Taxa de chegada: variável
    - Tempo de serviço: 0.84 segundo
    - Número de servidores: 10
    - Tempo de observação: 30 segundos

Resolva o mesmo problema usando os modelos de teoria das filas e informe como você realizou os cálculos. Compare em seguida os resultados obtidos via simulação com os resultados obtidos pelo modelo.

Primeira temos que verificar até quanto a nossa taxa de chegada pode variar para que o sistema continue em equilíbrio

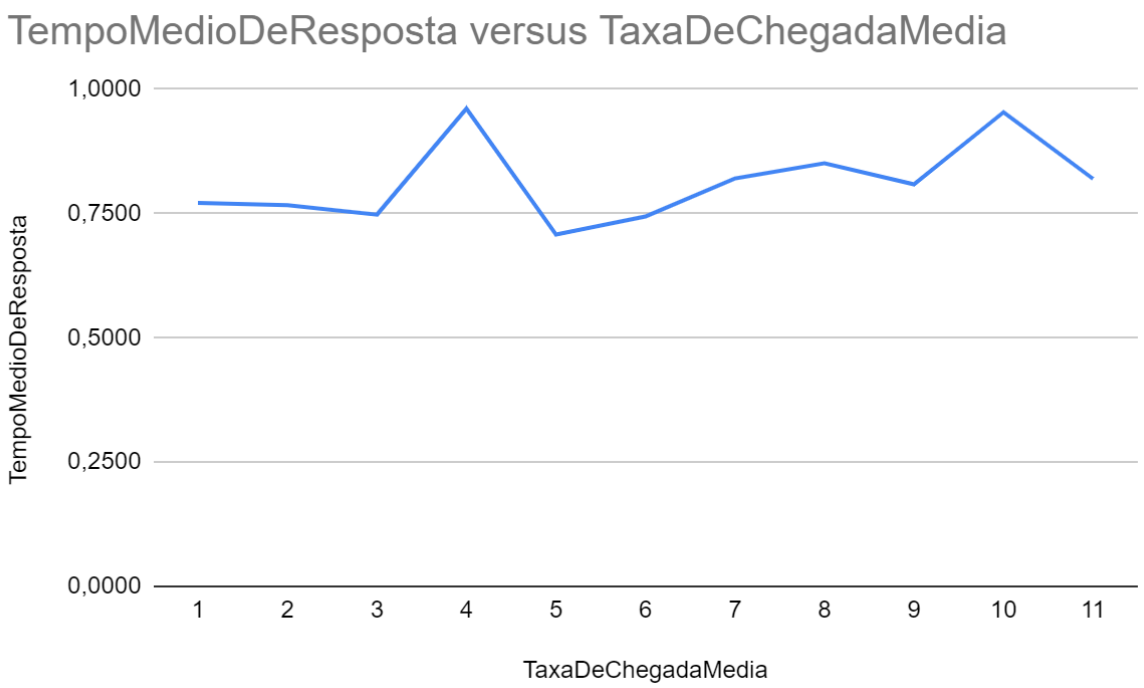
$y < m\mu$

y < 10 * 1/0.84

y < 10 * 1.19

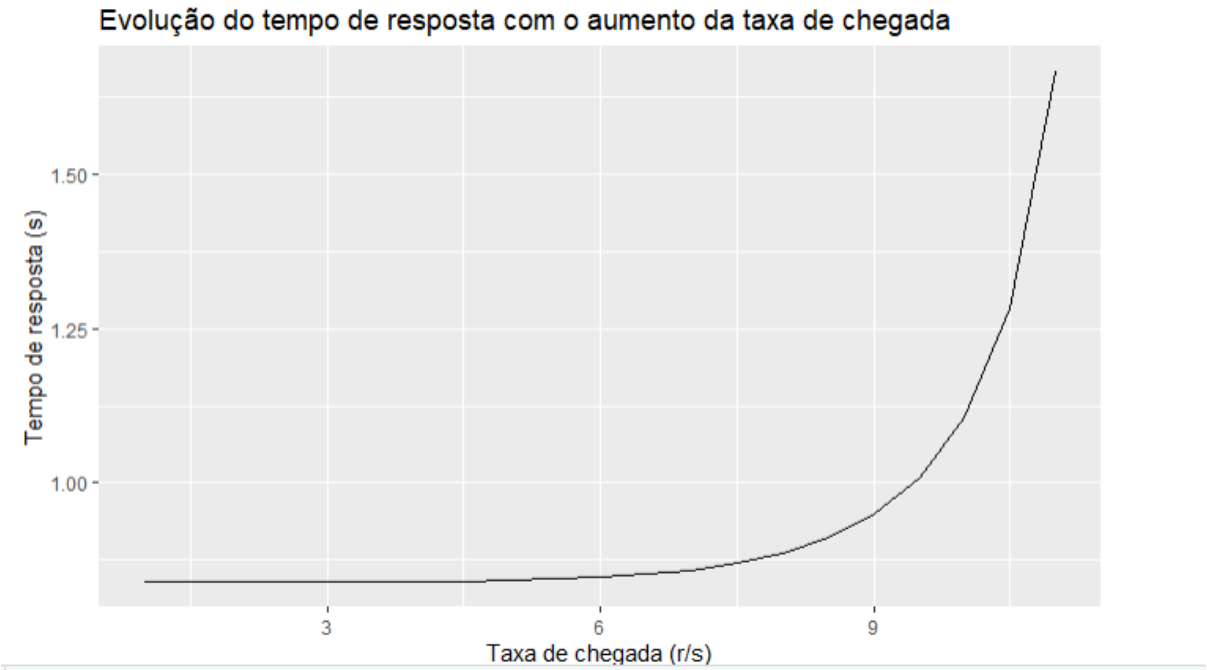
y < 11.9

Agora sabendo que o a taxa de entrada pode variar enquanto for menor que 11.9, termos os seguintes dados, acompanhe o gráfico.



Para saber se um sistema é escalável espera-se que o tempo de resposta não aumente significativamente quanto a carga for aumentada, esperamos que ele se mantenha constante ou que o aumento seja pouco. Observando o gráfico podemos observar que a taxa de chegada varia de 1 a 11 e que o tempo de resposta tem uma variação de 0,75 a 0.85. Devido a isso esse sistema não é escalável.

Isso fica mais claro quando observamos o gráfico gerado utilizando o modelo de teoria das filas em R, acompanhe:



Podemos ver uma grande variação do tempo de resposta de acordo com o aumento da carga.

3. Qual a quantidade mínima de servidores necessários para obter um tempo de resposta médio menor que 1 segundo? Analise diferentes cenários de simulação para dar sua resposta. Considere para esta questão o seguinte cenário base de parâmetros de entrada:
   - Taxa de chegada: 9,5 requisições por segundo
   - Tempo de serviço: 0.84 segundo
   - Número de servidores: variável
   - Tempo de observação: 30 segundos

   Resolve o mesmo problema usando os modelos de teoria das filas. Compare em seguida os resultados obtidos via simulação com os resultados obtidos pelo modelo.

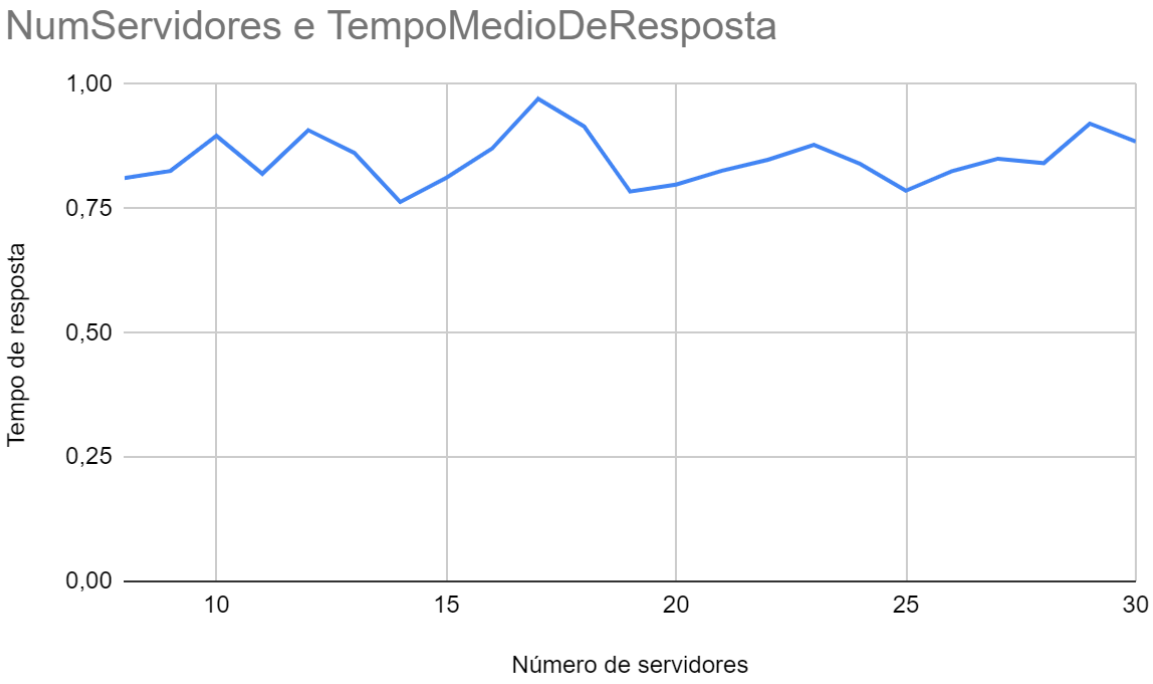
Primeiro temos que saber se até que ponto o nosso sistema está em equilíbrio, para isso vamos utilizar a forma:

$y < m\mu$

$9.5 < m * 1.19$

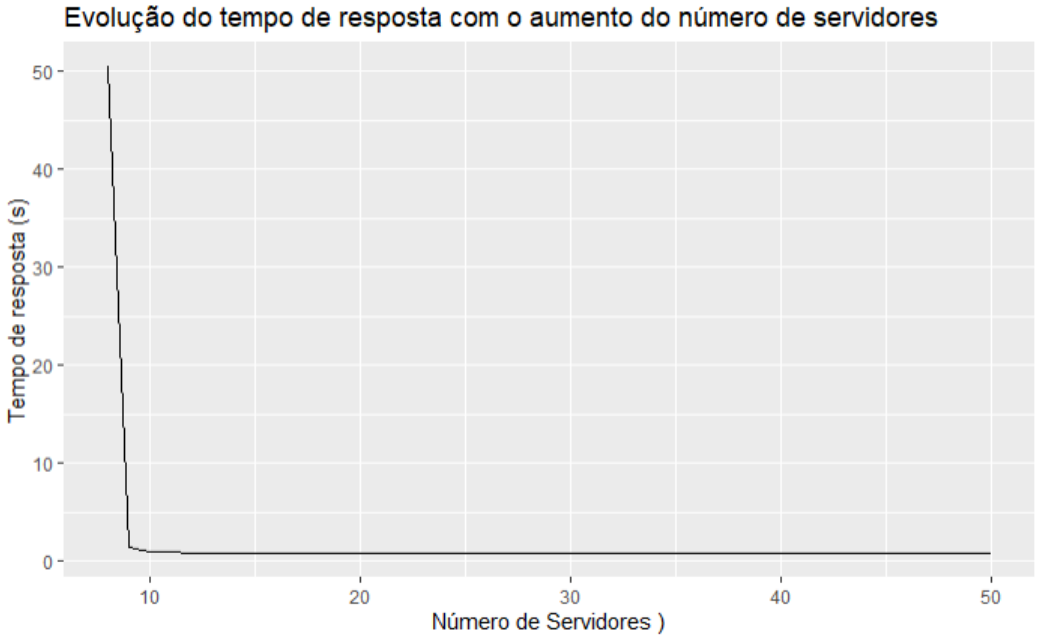
$7.9 < m$

Então para que o sistema esteja em equilíbrio ele tem que ter mais que 7.9 servidores, sabendo disso vamos observar os dados do simulador:



De acordo com o gráfico gerado, podemos observar que com qualquer número de servidores o tempo de resposta está abaixo de 1s, ele está variando com mínimo de 0.76 com 14 servidores e máxima de 0.97 com 17 servidores.

Já observando os valores gerados no modelo de filas em R podemos observar uma pequena diferença, observe:



Observe que com 8 servidores tem um tempo de resposta muito grande, que é 50s, e ele vai diminuindo com o aumento do número de servidores, até chegar em uma

constância, acima de 11 servidores o tempo de resposta já está abaixo de 1s, mas ele fica constante mesmo a partir de 15 servidores , que fica com um tempo de resposta por volta dos 0.84 segundos.