

## Lista de Exercícios 2

### Questão 1 (1+2+2+2+2+1 pontos)

Você foi contratado para avaliar o desempenho de uma plataforma para um serviço recém lançado na Web. A plataforma consiste de três componentes: o servidor Web (*front end*), a aplicação (camada intermediária) e, por fim, o banco de dados (*back end*). Em uma caracterização preliminar da carga esperada para o serviço, descobre-se que os tempos entre chegadas de requisições têm média  $T$  segs e variância  $T^2$  segs<sup>2</sup>. Além disto, sabe-se que os tempos de serviços de cada requisição *no servidor Web e na aplicação* são exponencialmente distribuídos com médias  $1/\mu_W$  e  $1/\mu_A$  segundos, respectivamente, e que o tempo médio de serviço no banco de dados é  $1/\mu_D$  segundos. Por fim, espera-se também que  $p_W\%$  das requisições enviadas ao serviço sejam completamente atendidas pelo servidor Web e que apenas  $p_D\%$  delas exijam processamento no banco de dados. Responda:

- Apresente o modelo de filas representativo do sistema avaliado. Indique claramente os **parâmetros** (e suas distribuições) de cada fila representada. Justifique sua resposta.
- A camada de aplicação pode ser modelada com precisão como uma fila M/M/1? Justifique sua resposta.
- Qual a probabilidade de que uma requisição enviada à camada de aplicação tenha um tempo de residência neste componente superior a  $1/\mu_A$ ? Qual a probabilidade de que ela tenha que esperar mais de  $1/\mu_A$  segundos para ser tratada por este componente? Simplifique ao máximo suas respostas. Se este componente está em equilíbrio, o que você pode dizer a respeito destas duas probabilidades?
- Estimando que o coeficiente de variação dos tempos de serviço no banco de dados seja 2, qual o tempo médio de residência de uma requisição no banco de dados? Simplifique ao máximo sua resposta.
- Descobre-se porém que o coeficiente de variação dos tempos de serviço no banco de dados é 1.01. Neste caso, pode-se afirmar que a probabilidade de haver no mínimo  $c$  clientes neste centro pode ser aproximada por  $\left( \frac{p_D}{T \times \mu_D} \right)^c$ ? Justifique sua resposta
- O que acontece se  $\mu_W > 1/T > \mu_A/(1-p_W)$ ? Discuta claramente o que aconteceria com o sistema, em particular com os centros correspondentes ao servidor Web e à aplicação. O que você faria para melhorar a precisão do modelo de desempenho sabendo que a memória disponível no componente de aplicação é finita?

### Questão 2 (2+2 pontos)

Suponha agora um servidor de banco de dados que atende requisições enviadas por  $N$  aplicações diferentes, cada uma executando *independentemente* em uma plataforma diferente. Sabe-se que os tempos de serviços no servidor de banco de dados seguem uma distribuição Weibull com média  $1/\mu$  segundos e variância  $1/(1.1\mu^2)$  segundos<sup>2</sup>. Além disto, sabe-se também que cada aplicação  $i$  ( $i=1..N$ ), gera um número de requisições por segundo para o servidor de acordo com uma distribuição Poisson, com taxa de requisições por segundo igual a  $\lambda_i$ . Qual o tempo médio de espera por serviço que cada requisição experimenta no servidor de banco de dados? Suponha agora que o SLA estabelecido entre as aplicações e o servidor de banco de dados estabeleça que 99.9% das requisições devem ter um tempo de resposta inferior a  $r$ . Qual a probabilidade de que o SLA seja violado? Responda às duas perguntas com a maior precisão possível em cada caso. Justifique suas escolhas de modelagem.

**Questão 3 (4 pontos)**

Considerando ainda o cenário descrito na questão 1e), suponha que você precise melhorar o desempenho da camada de banco de dados do exemplo acima. Duas opções são cogitadas: (a) substituir a máquina existente por um multiprocessador com  $m=2$  processadores idênticos ao processador utilizado na plataforma original mantendo um escalonador de tarefas único e compartilhado; e (b) substituir a máquina existente por uma com uma CPU duas vezes mais rápida que a original. Se seu objetivo é garantir que a probabilidade de que uma requisição enviada para o banco de dados tenha que esperar para ser processada seja a menor possível, qual opção você escolhe (ignorando os custos associados com cada plataforma)? Assumindo que  $T=0.1$  segundo,  $\mu_D=10$  requisições por segundo e  $p_D = 30\%$ , justifique sua resposta.