Teoria das Filas Aplicada à Computação - 2003/1	04/12/03
Aula 1: Motivação	
Instrutor: Berilhes Borges Garcia	Escriba: Berilhes Borges Garcia

1 Qual é o assunto deste curso

Teoria das filas é uma área da matemática, envolvendo análise estocástica, a qual permite prever o desempenho de um sistema computacional (por desempenho nós tipicamente representamos o tempo de serviço médio). A idéia é modelar analiticamente um sistema computacional como se este fosse constituído de recursos (como CPU, disco, etc) e trabalhos que demandam estes recursos. Contenção ocorre quando vários trabalhos requeiram simultaneamente um recurso, o que significa que algum recurso deve esperar. Teoria das filas permite prever estas esperas.
Diferença em relação aos cursos ortodoxos de teoria das filas e processos estocásticos:

- Em primeiro lugar este curso é voltado para aplicação.
- Em segundo lugar a teoria das filas tradicional baseia-se em suposições não realistas com relação a distribuição da carga de trabalho.

Eu espero que ao final do curso nós possamos avaliar de forma consistente o estado da arte da área de modelagem analítica de sistemas computacionais.

1.1 Exemplos da potência da modelagem analítica

Ferramenta de previsão. Exemplo: Uma vez que nós conhecemos a topologia de uma rede, as rotas que os pacotes seguem, a razão média de chegada de pacotes, o tempo de transmissão e a distância entre os roteadores. Então, teoria das filas pode ser utilizada para computar o tempo médio que os pacotes gastam na rede esperando por serviço em um determinado roteador i, a distribuição do tamanho das filas no roteador i, o tempo total que um pacote leva para ir do ponto A para o ponto B na rede, etc.
Ferramenta de projeto. Alguns exemplos.

1.1.1 Exemplo 1 Duplicando a Razão de Chegada

Considere um sistema formado por:
1 CPU, FCFS, Poisson - 3 trabalhos por segundo. Qualquer distribuição do tempo de serviço. Tempo de serviço médio menor que o tempo entre chegadas.
Questão: *Seu chefe diz a você que amanhã começará uma promoção que duplicará a razão de chegada. Logo ele quer que você faça alguma coisa, que ele não sabe muito bem o que é, para assegurar que os trabalhos experimente o mesmo tempo de resposta médio. Isto é, os consumidores deveriam não notar que a razão de chegada média aumentou. Qual deveria ser o aumento da velocidade da CPU de modo a manter o mesmo tempo médio de resposta? Resposta:*

Questão: Porque? Resposta:
Questão: Alguém pode apresentar um argumento, sem nenhuma teoria das filas, que sustente esta afirmação? Resposta:
Questão: Como a resposta muda se a CPU utilizar uma ordem de serviço “time sharing” ao invés de FCFS? Resposta:

1.1.2 Exemplo 2 Sistema Batch

6 trabalhos. Probabilidade de 0,5 do trabalho ir para o servidor 1, e probabilidade 0,5 de ir para o servidor 2. Servidor 1 serve trabalhos a uma razão média de 1 trabalho a cada 3 segundos. Servidor 2 também serve trabalhos a uma razão média de 1 trabalho a cada 3 segundos. Distribuição de serviço irrelevante para este exemplo.
Pergunta: Você substituiu o servidor 1 por um servidor que é duas vezes mais rápido. Como este melhoramento afeta o tempo médio de resposta do sistema? Como este afeta a vazão média (throughput) do sistema? Resposta:
Questão: Suponha que o sistema tenha sido mudado para um sistema aberto, ao invés de um sistema fechado. A modificação melhora o tempo médio de serviço? Resposta:

1.1.3 Exemplo 3 - Um estudo de caso

Considere o sistema iterativo apresentado na figura abaixo. O sistema consiste de alguns terminais. Cada terminal repete o seguinte processo: o usuário gera um trabalho; o usuário espera que o trabalho passe através do subsistema central e retorne um resultado; o usuário então pensa por algum tempo; finalmente o usuário envia um novo trabalho. O subsistema central consiste de uma CPU, um disco lento, e um disco rápido e talvez vários outros dispositivos.
Nós coletamos algumas medidas durante um período de observação de $T = 900\,\mathrm{sec}$: O tempo total que cada dispositivo esteve ocupado durante T. O número total de trabalhos completados em cada dispositivo durante T. O número total de trabalhos completados durante T. O tempo médio de pensamento entre o envio de trabalhos por parte do usuário.
Qual das seguintes mudanças é mais eficaz de modo a aumentar o throughput: Substituir a CPU por uma duas vezes mais rápida? Balancear carga entre o disco lento e o disco rápido? Comprar um segundo disco rápido?

2 Bibliográfia

S. M. Ross, Introduction to Applied Probability Models with Optimization Applications, Dover Publications, 1970.
L. Kleinrock, Queueing Systems, Vol. I: Theory, John Wiley and Sons, Inc., 1975.
L. Kleinrock, Queueing Systems, Vol. II: Computer Applications, John Wiley and Sons, 1975.
A. O. Allen, Probability, Statistics, and Queueing Theory with Computer Science Applications , Academic Press, 1990.
R. Jain, The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling, John Wiley & Sons, 1991.
E. Lazowska, J. Zahorjan, S. Graham, K. Sevcik, Quantitative System Performance , Prentice-Hall, 1984.
D. Bertsekas and R. Gallager, Data Networks , Prentice-Hall, 1992.
A. M. Law and W. D. Kelton, Simulation Modeling and Analysis, 2nd Edition, McGraw-Hill, Inc. 1991.