Geração de carga de trabalho ciente do desempenho dinâmico do sistema

Lourenço Alves Pereira Júnior[1,2], Regina H. Carlucci Sanatana[2], Marcos José Santana[2], Francisco José Monaco[2]

[1]Instituto Tecnológico de Aeronáutica — ITA
ljr@ifsp.edu.br

[2]Universidade de São Paulo — USP
{ljr,rcs,mjs,monaco}@icmc.usp.br

Abstract. *Workload characterization synthesizes the load behavior of a system by statistical methods whose inputs are commonly from empirical data. Thus, in another moment, one can excite a system with synthetical different workloads that share statistical properties. In this context, to the best of our knowledge, no other study has yet considered the dynamical performance model in the process of workload generation. We take a performance model of a large-scale e-commerce system as a Transfer Function, which allows us to analyze the system behavior in Frequency-domain. We demonstrate cases in which two workloads generated by different distributions lead to similar results and two statistical equivalent workloads lead to inconsistent results.*

Resumo. *A metodologia aplicada à caracterização de carga de trabalho convencional considera dados obtidos empiricamente por meio de métodos estatísticos com a finalidade de criar um modelo capaz de sintetizar o caso real. Dessa forma, pode-se reproduzir a carga original, gerando diferentes instâncias que se equivalem em suas características estatísticas. Neste artigo, apresentamos os benefícios de considerar o desempenho dinâmico de um sistema de* e-commerce *por meio de um modelo de função de transferência. Apresentamos um caso em que cargas aplicadas ao sistema em estudo apresentam resultados diferentes do esperado, e uma explicação para esse fenômeno está na análise no domínio da frequência proporcionada pela abordagem adotada.*

1. Introdução

A caracterização de carga de trabalho em sistemas computacionais é uma área bem estabelecida e que produz resultados significativos para a avaliação de desempenho de sistemas computacionais [Calzarossa et al. 2016]. Essa técnica aliada a Teoria de Filas fomam um arcabouço importante para o planejamento de capacidade de sistemas computacionais. Provêem um modelo analítico capaz de representar, em regime estacionário, características dos sistemas (tais como: taxa de chegada de requisições, tempo de processamento, quantidade de centros de serviço, tamanho da fila, distribuição de probabilidade etc.) e o desempenho do sistema (tempo de resposta, taxa de utilização, *throughput* etc.). Porém, com a efetividade de recursos computacionais ofertados como utilidade (computação em nuvem) [Mell and Grance 2011], um sistema computacional pode aumentar ou diminuir em tempo de execução sua potência. Desse modo, constatamos que os sistemas estão mais suscetíveis a condições de transição (de um estado a antes da

adaptação para um estado b após sua efetivação), e os modelos estácionários não capturam esses momentos.

Dado a importância da geração de carga sintética para o projeto de sistemas computacionais de grande escala, neste trabalho buscamos entender como uma carga de trabalho impacta em um sistema computacional, considerando seu comportamento dinâmico. Nesse sentido, este artigo objetiva demonstrar (1) como cargas estatisticamente diferentes podem impactar de modo semelhante dependendo do comportamento dinâmico do sistema em estudo e (2) como a análise de resposta em frequência pode revelar quais componentes impactam mais no sistema em estudo. Ao apresentar estas questões, este trabalho contribui para o estado da arte ao evidenciar os ganhos obtidos por modelar um sistema computacional com uma Função de Transferência, mostrando qual impacto o ruído presente na entrada do modelo gera na variável de saída.

Em estudos anteriores [Pereira et al. 2017b, Pereira et al. 2017a], demonstramos a importância de se conduzir experimentos para identificação de modelos de Função de Transferência e, a partir disso, como realizar análises de desempenho. Se por um lado o foco foi em como mensurar e prever o comportamento enquanto o sistema encontra-se em estado de transição, neste artigo, verificamos como o entendimento do modelo dinâmico impacta em um sistema em estado estacionário. Demonstramos como uma carga estacionária, gerada a partir de uma função de probabilidade, é processada pelo sistema. Evidenciamos que a absorção da carga depende da dinâmica do processamento de requisições do sistema, e que alterações bruscas de curta duração minimamente impactam na variável de saída — respeitada a resposta em frequência do sistema em estudo.

A seguir, na Seção 2 é apresentada a formulação para representação do modelo de desempenho dinâmico como Função de Transferência. A Seção 3 apresenta o modelo considerado para a análise de resposta em frequência. Os trabalhos que se relacionam com o presente são descritos na Seção 4. Por fim, na Seção 5 conclui-se este estudo.

2. Representação do desempenho dinâmico

Para a representação do comportamento dinâmico, é necessário um modelo de desempenho que leve em consideração a variável tempo. Conceitualmente, durante as transições que ocorrem no regime operacional, o desempenho do sistema computacional em estudo é influenciado pelas alterações na carga de trabalho, mas também pelo seu estado interno (i.e: preenchimento de *buffers* e provisionamento de recursos). Dessa forma, precisamos de um modelo matemático capaz de representar como a evolução dinâmica do desempenho ocorre temporalmente. Como apresentado em [Ljung 1999], o modelo autorregressivo atende esses requisitos e pode ser expresso na forma

$$A(q)y(k) = \frac{B(q)}{D(q)}u(k) + \frac{C(q)}{F(q)}v(k), \tag{1}$$

em que $y(k)$ é a variável de desempenho no instante de tempo k (discreto), $u(k)$ é a variável de entrada, $v(k)$ é a representação do ruído. $A(q)$, $B(q)$, $C(q)$ e $D(q)$ e $F(q)$ são polinômios que permitem ponderar como as variáveis irão impactar em $y(k)$; q é um operador de atraso e serve para especificar o estado do sistema, quanto maior o valor de q, maior será a influência de valores passados no valor atual. A razão entre os polinômios é chamada de Função de Transferência.

3. Análise de carga de trabalho sintética

Nesta seção descrevemos o estudo que demonstra a utilização da análise de resposta em frequência como uma ferramenta que auxilia na geração de carga de trabalho em sistemas computacionais. A noção explorada nesta abordagem é que, dependendo da FT do sistema em estudo, a diferença entre duas cargas está na análise de seus espectros. O Diagrama de Bode quantifica como o sistema absorve as flutuações na carga de trabalho, evidenciando como o sistema atenua as frequências da variável de entrada. Dessa forma, as frequências muito atenuadas da entrada provocam pouca influência na variável de saída. Portanto, se duas cargas de trabalho diferem apenas nas componentes de frequência que são muito atenuadas, então, para efeitos práticos, essas duas cargas podem ser consideradas similares.

Iniciamos com a descrição do modelo de desempenho utilizado, seguido pela especificação de um filtro que mantém as frequências que impactam na variável de saída do sistema. A análise segue ao considerar três cargas diferentes: 1) gerada por uma distribuição de probabilidade exponencial com média 300; 2) corresponde à carga 1 (exponencial) filtrada pelo filtro desenvolvido; 3) normal (média 300 e dispersão 67), aproximação da carga 2 (exponencial filtrada).

A Função de Transferência foi o modelo dinâmico adotado para a representação do sistema em estudo, cujos parâmetros foram identificamos como

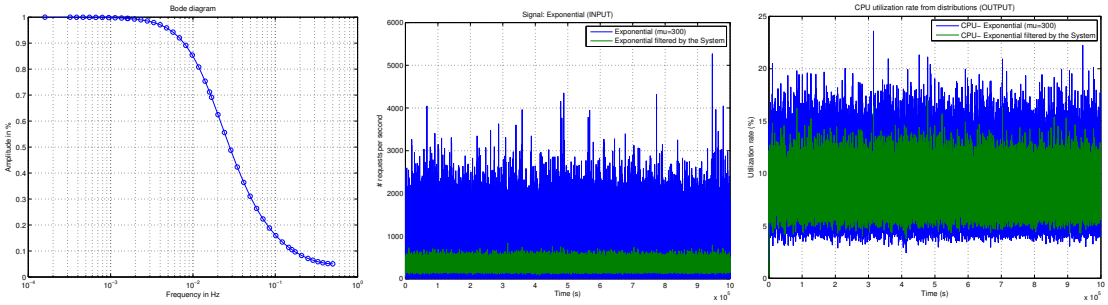
$$G(z) = \frac{b}{1 + az^{-1}} = \frac{0.00273}{1 - 0.9037z^{-1}}. \tag{2}$$

Detalhes do processo de identificação estão descritos em [Pereira et al. 2017b]. Trata-se de um sistema de *e-commerce* de três camadas: apresentação (representado pelos softwares clientes—*browser*), aplicação (composto de 8 servidores) e dados (banco de dados SQL). A taxa de amostragem é de 1 segundo. A variável de entrada do sistema é a quantidade de requisições por segundo (req/s), a saída é a taxa de utilização média dos servidores de aplicação. O ganho DC é $0,0283$, o que significa que uma requisição impacta na taxa de utilização em $0,02\%$. O tempo de assentamento do sistema é de 39 s.

A partir da FT (Eq. 2) é possível gerar sua representação em Diagrama de Bode. Nele podemos quantificar o quanto do sinal de entrada (req/s) será atenuado no sinal de saída. Se há uma quantidade de variação grande e brusca no valor de entrada, a frequência com que alterações ocorre é alta. Se mudanças no sinal de entrada são duradouras, a frequência é baixa. Pelo Diagrama de Bode (inspecionar Figura 1(a)), percebemos que baixas frequências são absorvidas pelo sistema, e isso significa que quando elas acontecem, novos patamares estacionários são alcançados. Por outro lado, à medida que se aumenta a frequência com que mudanças acontecem, elas tendem a não ter efeito na saída. Frequências altas caracterizam-se por picos e vales seguidos, que, em efeitos práticos, anulam o impacto no sistema.

3.1. Análise das cargas de trabalho sintéticas

Gerou-se um sinal oriundo de um gerador de números pseudo-aleatório (PRNG); o tamanho desse sinal será de 2^{20} segundos (aproximadamente 1 milhão de amostras), ou seja, uma simulação de 291 horas, pois cada amostra corresponde a 1 segundo. Especificamos dois sinais entrada: 1) **Exponencial**: sinal gerado por uma distribuição de probabilidade



(a) Diagrama de Bode do sistema em estudo.

(b) Sinais de entrada: requisições por segundo.

(c) Sinais de saída: taxa de utilização.

Figura 1. O desempenho do sistema Eq. 2 é semelhante ao receber duas cargas de trabalho diferentes que possuem as mesmas componentes de baixa frequência.

exponencial de média 300 e 2) **Filtrado**: produzido pelo filtro que representa a dinâmica do sistema, cuja entrada é o sinal exponencial ([Pereira et al. 2017b]). A Figura 1(b) apresenta os sinais exponencial e filtrado no domínio do tempo. Esses dois sinais foram passados para a Função de Transferência do sistema (Eq. 2), produzindo os sinais de saída apresentados na Figura 1(c).

O sinal exponencial possui média $299,82$ e desvio padrão $299,90$, enquanto o sinal filtrado $299,81$ e $67,33$ para média e desvio padrão, respectivamente. Os sinais de saída para o exponencial possuem média $8,49$ e desvio padrão de $1,90$, ao passo que para a saída do filtrado $8,49$ de média e $1,35$ para desvio padrão. Ao dividir a média dos sinais de saída pela média dos sinais de entrada, obtém-se o ganho de $0,02835$, o que é condizente com o sistema. Ao dividir o desvio padrão dos sinais de entrada, obtém-se $0,22452$ $(67,33/299,9)$, fazendo a mesma operação para os sinais de saída, tem-se como resultado $0,7105$ $(1,35/1,9)$. Isso significa que o sinal filtrado possui uma dispersão correspondente a $22,45\%$ do sinal exponencial original. Porém, nos sinais de saída, o resultado dessa razão aumenta para $70,75\%$, ou seja, boa parte da dispersão é filtrada pelo sistema, o que faz com que os sinais de saída sejam menos dispersos do que os de entrada em cerca de 50% $(70,75 - 22,45)$.

Seguindo adiante, podemos aproximar o sinal da exponencial filtrada ao de uma normal e novamente passar os dois sinais (normal e filtrado) pela FT. A intuição neste instante, é que o resultado seria equivalente, pois teríamos dois sinais que compartilham de propriedades estatísticas. Foi gerado um sinal com 1 milhão de amostras oriundo de um PRNG com distribuição normal com média 300 e dispersão 67; denominamos esse sinal como normal. Os sinais de entrada filtrado e normal foram passados pelo sistema o que gera o sinal de saída. O sinal normal possui média 299.96 e desvio padrão 66.97. Para o sinal normal, ao dividir a média de sua saída pela média do respectivo sinal de entrada, obtém-se o ganho de $0,02835$, o que é condizente com o sistema. Ao dividir a dispersão dos sinais de entrada (dispersão do normal pelo filtrado), obtém-se $0,99361$, enquanto que, para a mesma operação considerando os sinais de saída, tem-se $0,31719$. O que significa que há uma semelhança estatística entre os sinais de entrada, porém os de saída são diferentes. O que leva à evidência de que duas cargas de trabalho semelhantes apresentam resultados diferentes. O que é contraditório ao esperado.

Para observar a semelhança dos sinais filtrado e normal e a diferença apresentada nos resultados, calculou-se a Função Densidade de Probabilidade (PDF) com a função `ksdensity` disponível no Matlab ©. Essa função recebe uma série de números e estima como seria a PDF da série. Observou-se uma boa aproximação para os sinais de entrada, porém uma distribuição diferente para a saída, muito embora a média seja a aproximadamente a mesma.

Até o momento, tem-se que para duas distribuições diferentes os resultados produzidos foram semelhantes, e que para duas distribuições semelhantes os resultados foram diferentes. Ou seja, o resultado sugere que as estatísticas de média e dispersão da distribuição, nesse caso, influencia pouco nesse sistema. A análise da resposta em frequência que considera tanto o sinal de entrada quanto o comportamento do sistema pode revelar o ponto que explica tais resultados.

Os três sinais considerados até o momento (exponencial, filtrado e normal) foram passados para o algoritmo Fast Fourier Transform (FFT). Do vetor resultante, obteve-se o valor correspondente aos harmônicos que geram o sinal no domínio do tempo. Observou-se que o sinal exponencial é composto pela média (300), correspondente ao harmônico fundamental e um ruído branco, isto é, uma grande quantidade de harmônicos distribuídos aproximadamente de forma isonômica na faixa de frequência. Ao passar o sinal exponencial pelo filtro especificado, parte das das frequências são atenuadas, de modo que as componentes mais altas (próximas a $0,5$ Hz) tendessem a zero, conforme previsto pelo Diagrama de Bode (Figura 1(a)). Há uma concentração de harmônico de baixa ordem.

A FFT do sinal normal difere dos outros dois sinais. É diferente da FFT do sinal exponencial pois as frequências do sinal normal possuem amplitude menor. Muito embora os dois sinais tenham a mesma média, as diferenças fazem com que os dois sinais tenham probabilidade de distribuição diferentes. A exponencial com harmônicos de maior amplitude propiciam picos maiores no domínio do tempo, o que é característico de sua cauda. A FFT do sinal filtrado quando comparada à FFT do sinal normal possui harmônicos de baixa frequência de maior amplitude. Esse fato impacta na saída do sistema, pois a presença de baixas frequências significa uma maior excursão do sinal resultante, e isso foi observado nos resultados, uma vez que a dispersão do sinal de saída do filtrado foi maior que o respectivo do normal.

4. Trabalhos relacionados

Não encontramos outros trabalhos que consideram Função de Transferência (FT) como modelo de desempenho e empregue Análise de Resposta em Frequência para entender como o sistema processa uma carga de trabalho caracterizada. A aplicação comum de FT tem sido empregada como base para implementação de controladores de recursos baseados em teoria de controle [Shevtsov et al. 2017]. No entanto, a FT pode ser empregada como modelo de desempenho complementado as melhores práticas correntes [Yang and Liu 2012]. Nesse sentido, há trabalhos que focam nessa temática [Pereira et al. 2017b, Pereira et al. 2015, Mamani et al. 2015, Pereira et al. 2017a], porém nenhum deles apresentou o impacto da resposta em frequência de um sistema em regime estacionário, como apresentado neste artigo. Nosso trabalho avança o estado da arte por prover uma visão da FT como modelo de desempenho e analisar como é impacto de cargas de trabalho estacionárias em um sistema computaci-

onal previamente modelado.

5. Conclusões

A conclusão desse estudo é que o fato de haver casos em que o sistema é excitado por (1) distribuições diferentes que geram resultados semelhantes e (2) distribuições semelhantes que geram resultados diferentes, depende não somente da distribuição estatística utilizada, mas como o sistema absorve a carga de trabalho imposta a ele. Os resultados evidenciam a importância da abordagem da modelagem de sistemas computacionais através de modelos dinâmicos e destaca o uso da análise em frequência, pelos Diagrama de Bode e FFT, para obter comportamentos importantes a respeito do desempenho do sistema. Como trabalhos futuros, pretende-se avaliar avaliar sistemas com modelos não lineares (NARX) e de múltiplas entradas e saídas (MIMO).

Agradecimentos

Agradecemos CAPES, FAPESP, CNPq, LaSDPC/USP e IFSP pelo apoio financeiro.

Referências

Calzarossa, M. C., Massari, L., and Tessera, D. (2016). Workload characterization: A survey revisited. *ACM Comput. Surv.*, 48(3):48:1–48:43.

Ljung, L. (1999). *System Identification: Theory for the User*. Pearson Education.

Mamani, E. L. C., Pereira, L. A., Santana, M. J., Santana, R. H. C., Nobile, P. N., and Monaco, F. J. (2015). Transient performance evaluation of cloud computing applications and dynamic resource control in large-scale distributed systems. In *High Performance Computing Simulation (HPCS), 2015 International Conference on*, pages 246–253.

Mell, P. M. and Grance, T. (2011). Sp 800-145. the nist definition of cloud computing. Technical report, National Institute of Standards & Technology, Gaithersburg, USA.

Pereira, L. A., dos Santos de Souza, F. L., Mamani, E., and Monaco, F. J. (2017a). Geração de carga de trabalho transiente para aplicações de e-commerce multicamadas. In *XIV Workshop de Computação em Clouds e Aplicações*, Belém, PA. SBC.

Pereira, L. A., Mamani, E., Santana, R., Santana, M., and Monaco, F. J. (2017b). Análise de resposta em frequência para modelagem e geração de carga de trabalho em aplicações de e-commerce. In *XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 888–901.

Pereira, L. A., Mamani, E. L. C., Santana, M. J., Santana, R. H. C., Nobile, P. N., and Monaco, F. J. (2015). Non-stationary simulation of computer systems and dynamic performance evaluation: A concern-based approach and case study on cloud computing. In *Computer Architecture and High Performance Computing (SBAC-PAD), 2015 27th International Symposium on*, pages 130–137.

Shevtsov, S., Berekmeri, M., Weyns, D., and Maggio, M. (2017). Control-theoretical software adaptation: A systematic literature review. *IEEE Transactions on Software Engineering*, PP(99):1–1.

Yang, F. and Liu, J. (2012). Simulation-based transfer function modeling for transient analysis of general queueing systems. *European Journal of Operational Research*.