Trabalho apresentado no XXXVII CNMAC, S.J. dos Campos - SP, 2017.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Controle de consumo energético de servidores via tráfego pesado e programação dinâmica

Rodolfo Spinelli Teixeira[1]
Faculdade de Engenharia, UFJF, Juiz de Fora, MG
Saul de Castro Leite[2]
Departamento de Ciência da Computação,UFJF, Juiz de Fora, MG

1 Introdução

Cada vez mais serviços computacionais são oferecidos na Internet, como serviços de busca, editores de texto, planilhas eletrônicas, serviços de comércio eletrônico, dentre outros. Todos estes serviços necessitam de grandes centros de computação capazes de atender a demanda de vários usuários. Com o crescente número de usuários, estes centros são cada vez maiores e consequentemente necessitam de mais energia para serem mantidos. Atualmente nos EUA, certa de 1,5% da energia gerada no país é destinada para esses centros os quais são permanentemente ligados para responder picos de demanda [1]. Várias estratégias vem sendo proposta na literatura para a redução do consumo de energia destes centros. Dentre elas está o trabalho [2] que propõe que uma porção dos servidores fiquem mantidos como reservas e só devam ser ligados ou desligados de acordo com a demanda. Um agravante deste problema é que os servidores levam tempo para ligar ou desligar e este tempo deve ser levado em consideração no modelo. Em [2], o autor faz um modelo de filas $M/M/s$ e usa técnicas heurísticas para determinar uma boa estratégia para ligar e desligar os servidores reservas. Neste trabalho, adota-se uma estratégia similar a proposta em [2]. Porém, será usado um modelo de tráfego pesado para um sistema de filas $GI/M/s$ proposto por [3]. Para modelar o tempo que os servidores levam para ligar e desligar, é usado uma Cadeia de Markov a tempo contínuo que representa o estado dos servidores extras.

2 Modelo de Fila

Similar ao que é proposto em [2], considera-se um sistema de filas com n servidores permanentes e m servidores extras, que podem ser ligados e desligados de acordo com a demanda. Uma Cadeia de Markov a tempo contínuo $\{\theta(t), t \geq 0\}$ é usada para modelar o estado dos servidores extras. Para o modelo do número de clientes na fila, usou-se a aproximação em tráfego pesado para uma fila $GI/M/n$ desenvolvida por [3]. Considere um sistema com n servidores com taxa de entrada dada por λ e taxa de serviço μ e defina a seguinte constante: $\rho_n = \lambda/(n\mu)$. Quando n é grande e ρ_n próximo e inferior a 1, o

[1]rodolfo.spinelli@engenharia.ufjf.br
[2]saul.leite@ufjf.edu.br

Teorema 3 de [3] pode ser usado para aproximar o processo do número de clientes na fila $\{Q(t), t \geq 0\}$ pelo processo $\{X_n(t), t \geq 0\}$ o qual é solução da seguinte equação diferencial estocástica: $dX_n(t) = (\lambda - \mu X_n(t) \wedge n)dt + \sqrt{n}\mu(1 + \sigma^2)dW(t)$, em que σ é o coeficiente de variação da distribuição de entrada, $x \wedge y$ denota o mínimo entre x e y. Com o modelo estabelecido, termina-se o problema de controle, que consiste em determinar a política π que minimiza a esperança de $c_1X(t) + c_2N(t)$ em regime estacionário, em que c_1 e c_2 são constantes, $X(t)$ o número de clientes no sistema no tempo t e $N(t)$ representa o número de servidores ligados no tempo t. Para resolver este problema numericamente, o método da Cadeia de Markov Aproximada foi usado, que consiste em discretizar o problema de controle em tempo contínuo e obter um processo de decisão Markoviano (PDM). Este PDM é resolvido utilizando o método de iteração de valores relativos.

3 Resultados

Nesta seção é comparado o método apresentado em [2] (HR) com o desenvolvido aqui (DP). Em todos os senários, tem-se $c_1 = 1$ e $c_2 = 2$, o tempo de serviço possui distribuição exponencial com taxa $\mu = 1$. Nos experimentos 1 e 2, o tempo entre entradas consecutivas possui distribuição exponencial. Os parâmetros para os experimentos 1 e 2 são: $\lambda = 10$, $n = 20$ e $m = 9$ e $\lambda = 4$, $n = 10$ e $m = 5$, respectivamente. Os experimentos 3 e 4 possuem tempo entre entradas consecutivas de clientes com distribuição hiper-exponencial com coeficiente de variação quadrático dado por $\sigma^2 = 10$. Os para os experimentos 3 e 4 são $\lambda = 10$, $n = 20$, $m = 9$ e $\lambda = 4$, $n = 10$, $m = 5$, respectivamente. Os resultados estão apresentados na tabela abaixo. Note que em todos os casos os resultados são melhores quando a metodologia proposta aqui é empregada.

Ex. 1	**U**	**D**	**Custo Médio**	**Ex. 2**	**U**	**D**	**Custo Médio**
DP	19	17	33.066862	DP	10	9	14.123313
HR	19	10	37.356842	HR	9	4	17.309647
Ex. 3	**U**	**D**	**Custo Médio (sim)**	**Ex. 4**	**U**	**D**	**Custo Médio (sim)**
DP	21	17	37.230 (37.227, 37.233)	DP	12	9	16.521 (16.517, 16.525)
HR	19	10	39.020 (39.016, 39.023)	HR	9	4	17.491 (17.487, 17.495)

Agradecimentos

Os autores agradem o apoio financeiro da FAPEMIG, projeto de número: APQ 00945/14.

Referências

[1] US Environmental Protection Agency. *EPA Report on Server and Data Center Energy Efficiency.* 2007.

[2] I. Mitrani. Managing performance and power consumption in a server farm. *Annals of Operations Research,* 202(1):121–134, 2013.

[3] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research,* 29(3):567– 588, 1981.