PUC Minas

LICAP
Laboratório de Inteligência Computacional Aplicada

PLANEJAMENTO DE CAPACIDADE DE SISTEMAS COMPUTACIONAIS

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

UNDERSTANDING CLOUD COMPUTING: . EXPERIMENTATION AND CAPACITY PLANNING.
DANIEL A. MENASCE PAUL NGO.

Equipe MAD

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

O que é Cloud Computing:

Computação na Nuvem possui diferentes significados para diferentes pessoas. Embora a definição básica que envolve todas elas é:

Modalidade de computação caracterizada pela disponibilidade de recursos sobre demanda numa estrutura dinámica e escalável.

O termo *Recurso* pode ser usado para Infra-estrutura, Plataforma, Software, Serviços, ou Armazenamento.

O *Provedor de nuvem* é responsável por tornar os recursos disponíveis sobre demanda de forma eficiente para que as necessidades dos usuáros sejam atendidos dentro de um nível QoS.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Vantagens do Cloud Computing:

1) **Page pelo uso**: As empresas podem evitar gastos de capital usando recursos de nuvem conforme necessário e sobre determinados serviços. Na abordagem Proprietária, o custo total de propriedade (TCO) é uma função de três componentes principais:

- Investimento de capital inicial para hardware, software, rede, infra-estrutura de instalações, incluindo refrigeração e energia;
- Custos operacionais que incluem manutenção de hardware e software, custo de pessoal (incluindo sistema, rede, administradores de banco de dados, analistas de planejamento de capacidade), consumo de energia e depreciação;
- Atualizações de sistema necessárias para lidar com o crescimento da carga de trabalho existente e/ou novas cargas de trabalho.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Vantagens do Cloud Computing:

2) **Não há necessidade de provisionar cargas para horários de pico:** A responsabilidade de suportar cargas de pico em níveis de serviço acordados (SLA, é do provedor em nuvem.

3) **Tempo para Comercializar:** Os usuários de computação em nuvem não precisam adquirir, instalar e testar todas as infraestruturas, incluindo middleware e aplicativos em muitos casos.

4) **Desempenho e disponibilidade consistentes**: Quando os serviços são fornecidos pela nuvem sob SLAs rígidos e específicos em tempo de resposta e disponibilidade, os usuários não precisam se preocupar tanto em manter níveis adequados para essas métricas. Essa responsabilidade é transferida para a nuvem disponibilizando recursos autonomicamente (por exemplo, máquinas virtuais) para acompanhar cargas de trabalho variadas e imprevisíveis.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Desvantagens do Cloud Computing:

1) Privacidade e segurança: Organizações podem ficar preocupados em ter seus dados confidenciais armazenados nas mesmas plataformas de seus concorrentes. Pode haver preocupações em relação à exposição de dados privados de uma empresa ao provedor de nuvem. Em alguns casos, uma empresa pode estar sujeita a vários tipos de regulamentos (por exemplo, LGPD) cuja responsabilidade não pode ser facilmente delegada a um provedor terceirizado.

2) Dependência externa para aplicativos de missão crítica: mesmo quando os provedores de nuvem oferecem garantir SLAs rígidos e pagam multas por não conformidade, os usuários da nuvem podem se preocupar em confiar alguns de seus aplicativos mais críticos a terceiros.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Desvantagens do Cloud Computing:

3) Recuperação de desastres: os usuários de nuvem precisam ter garantias de que o provedor possui planos adequados de backup e recuperação de desastres que evitar a interrupção das atividades de um usuário diante de desastres naturais ou causados pelo homem.

4) Monitoramento e aplicação de SLAs: Negociar, monitorar e aplicar SLAs pode ser um desafio na computação em nuvem porque os recursos e serviços da nuvem são compartilhados por vários usuários e porque os provedores têm pouco controle sobre a intensidade da carga de trabalho dos diferentes aplicativos em nuvem.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de estruturas do Cloud Computing:

I) Nuvem pública (Public Cloud Computing)

- A nuvem pública é disponibilizada via Internet pública, para recursos fornecido por provedores como Amazon Web Services e Microsoft Azure.
- Este modelo possui um menor custo desde que a gestão, atualização e manutenção dos equipamentos são "divididos" entre as empresas que utilizam o serviço.
- A escalabilidade dos recursos é automática, permitindo modificar recursos como memória RAM, espaço de armazenamento, etc.
- Porém, este modelo proporciona menos controle sobre a plataforma. Para garantir QoS, o provedor de serviços regula o nível de personalização e controle avançado que o cliente possui sobre o hardware em que os serviços são executados.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de estruturas do Cloud Computing:

II) Nuvem privada (Private Cloud Computing)

- Os serviços dessa nuvem são oferecidos pela Internet ou rede interna privada.
- A nuvem privada é para empresas que necessitam do completo controle da infraestrutura, sem compartilhamento de recursos ou políticas padronizadas. Um exemplo de segurança, é que todos os dados que trafegam dentro de uma nuvem privada estão isolados em uma rede exclusiva, por meio de VLANs (Virtual Local Area Network serve para criar uma comunicação entre uma ou mais LANs (Local Area Network))
- Porém, os custos tendem a ser mais altos. A empresa será responsável por investimentos em manutenção, gestão, controle e atualização da plataforma.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de estruturas do Cloud Computing:

III) Nuvem híbrida (Hybrid Cloud Computing)

- O ambiente híbrido modelos público e privado. Essa alternativa tem sido amplamente adotada pelas empresas que estão em busca da redução de custos de TI sem abrir mão de proteger os dados cruciais.

- Os dados sigilosos são disponibilizados no ambiente privado, blindado por configurações próprias de firewall e políticas de uso. Os demais serviços de TI são armazenados na nuvem pública, os custos operacionais totais.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de Serviços do Cloud Computing:

Infraestrutura como serviço (IaaS):

Tornam fácil e acessível o provisionamento de recursos do cliente, como servidores, conexões, armazenamento e ferramentas relacionadas. Isso permite que os desenvolvedores criem ambientes de aplicativos do zero de forma rápida e econômica.

As empresas podem aumentar ou diminuir seus recursos dependendo das cargas de trabalho atuais em minutos.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de Serviços do Cloud Computing:

Plataforma como serviço (PaaS):

É a integração entre a infraestrutura e uma plataforma de desenvolvimento comercial para criar e lançar aplicativos ou serviços.

Nuvens PaaS, trabalham em combinação com nuvens IaaS, têm o benefício de simplificar a implantação e a escalabilidade e garantir que os custos sejam linearmente incrementais e razoavelmente previsíveis.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Tipos de Serviços do Cloud Computing:

Software como serviço (SaaS):

Visa o software em nível de aplicativo usado pelos usuários da nuvem para cumprir sua missão. Ele varia de software de escritório a software financeiro, como preparação de impostos e orçamento, etc.

O termo SaaS já existe há algum tempo, mas a computação em nuvem revitalizou o modelo SaaS, reduzindo o custo de produção de um aplicativo SaaS. O SaaS serve como uma ponte na evolução da computação em nuvem em termos de gerenciamento e alocação de recursos.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Planejamento de capacidade para Cloud Computing:

- Para o planejamento de capacidade em nuvem dois pontos de vista devem ser considerados: um do ponto de vista do usuário dos serviços em nuvem, e outro do ponto de vista do provedor.

- Quando os serviços são executados sob demanda, o planejamento de capacidade muda para o provedor dos serviços de nuvem. No entanto, os usuários da nuvem devem ser capazes de negociar SLAs com provedores de serviços de nuvem.

- Como pode haver SLAs para diferentes métricas de QoS, os usuários da nuvem devem considerar o uso da noção de função de utilidade para determinar a utilidade combinada dos serviços de nuvem como uma função dos vários SLAs.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Planejamento de capacidade para Cloud Computing:

- As funções de utilidade são usadas com bastante frequência em economia.
- A seguinte notação para formalizar o problema de seleção ótima de SLAs a serem negociados com o provedor de serviços em nuvem.

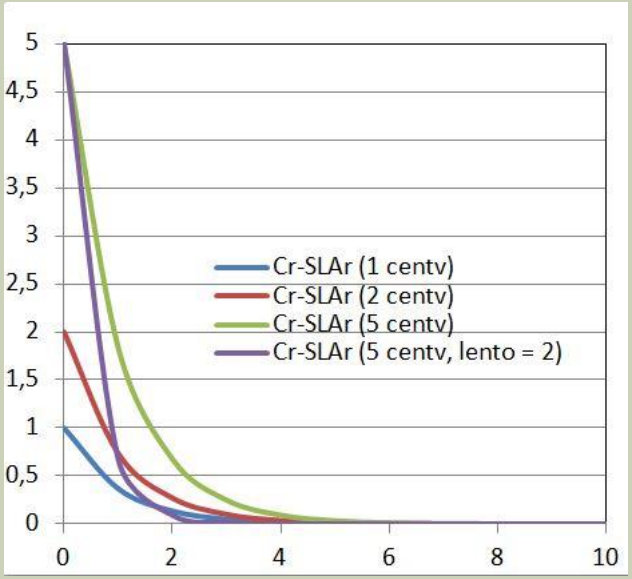
- ***SLAr**: SLA (em segundos) sobre o tempo médio de resposta por transação.*
- ***SLAx**: SLA (em tps) sobre o throughput da transação.*
- ***SLAa**: SLA de disponibilidade na nuvem.*

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

- ***Cr(SLAr)**: custo por transação (em centavos) quando o **SLA** de tempo de resposta negociado for **SLAr**.*

$$C_r(\mathrm{SLA_r}) = \alpha_r e^{-\beta_r\ \mathrm{SLA_r}}$$

SLAr	Cr-SLAr (1 centv)	Cr-SLAr (2 centv)	Cr-SLAr (5 centv)	Cr-SLAr (5 centv, lento = 2)
0	1	2	5	5
1	0,367879441	0,735758882	1,839397206	0,676676416
2	0,135335283	0,270670566	0,676676416	0,091578194
3	0,049787068	0,099574137	0,248935342	0,012393761
4	0,018315639	0,036631278	0,091578194	0,001677313
5	0,006737947	0,013475894	0,033689735	0,000227
6	0,002478752	0,004957504	0,012393761	3,07211E-05
7	0,000911882	0,001823764	0,00455941	4,15764E-06
8	0,000335463	0,000670925	0,001677313	5,62676E-07
9	0,00012341	0,00024682	0,000617049	7,61499E-08
10	4,53999E-05	9,07999E-05	0,000227	1,03058E-08

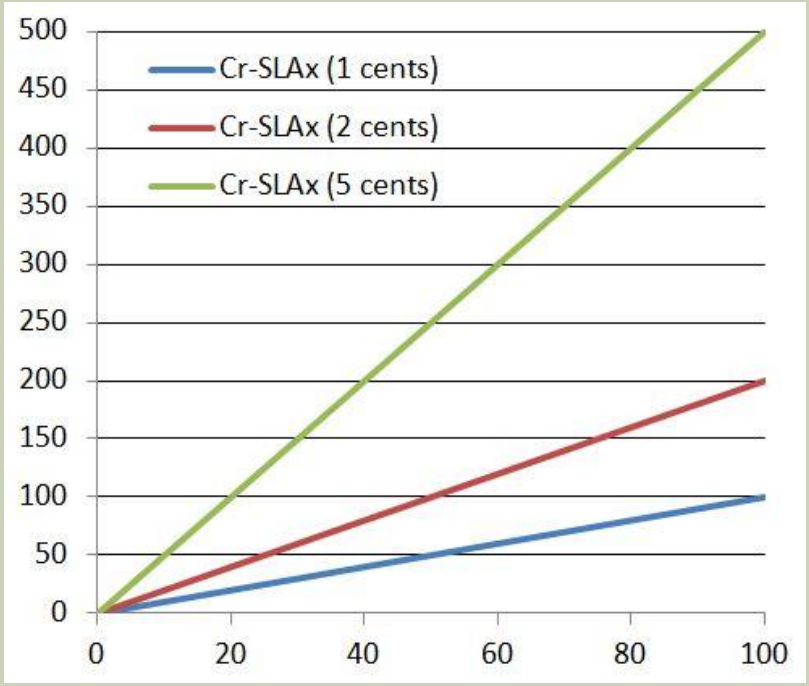


PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

- ***Cx(SLAx)**: custo por transação (em centavos) quando o SLA de throughput negociado for SLAx.*

$$C_x(\mathrm{SLA_x}) \quad = \quad \alpha_x \ \mathrm{SLA_x}$$

SLAx	Cr-SLAx (1 cents)	Cr-SLAx (2 cents)	Cr-SLAx (5 cents)
0	0	0	0
10	10	20	50
20	20	40	100
30	30	60	150
40	40	80	200
50	50	100	250
60	60	120	300
70	70	140	350
80	80	160	400
90	90	180	450
100	100	200	500

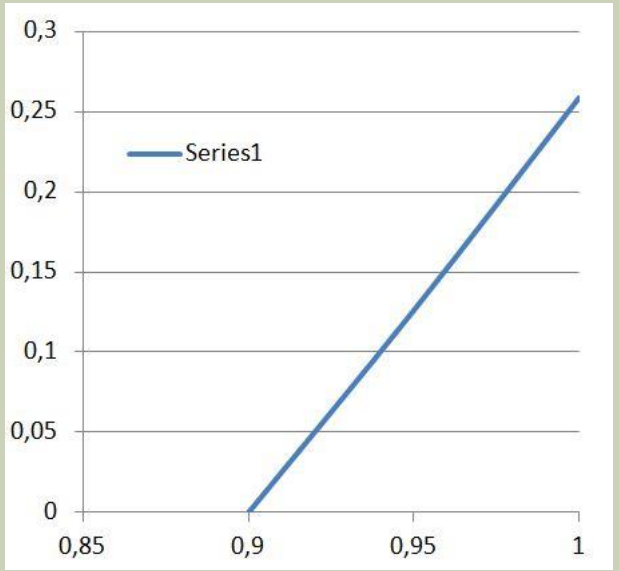


PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

- ***Ca(SLAa)**: custo por transação (em centavos) quando o SLA de disponibilidade negociado for SLAa.*

$$C_a(\mathrm{SLA_a}) = e^{\beta_a\ \mathrm{SLA_a}} - e^{0.9\beta_a},\ \mathrm{SLA_a} \geq 0.9$$

SLAa	Cr-SLAa (beta=1, SLA>0,90)	Cr-SLAa (beta=0,6, SLA>0,80)	Cr-SLAa (beta=0,2, SLA>0,60)
0,1	-1,354432193	-0,554237856	-0,107295512
0,15	-1,297768868	-0,521900118	-0,097042318
0,2	-1,238200353	-0,488577551	-0,086686077
0,25	-1,175577694	-0,454240159	-0,076225755
0,3	-1,109744304	-0,418857039	-0,065660305
0,35	-1,040535563	-0,382396342	-0,05498867
0,4	-0,967778414	-0,344825252	-0,044209784
0,45	-0,891290926	-0,306109951	-0,033322568
0,5	-0,81088184	-0,266215595	-0,022325934
0,55	-0,726350093	-0,225106274	-0,011218781
0,6	-0,637484311	-0,182744988	0
0,65	-0,544062282	-0,139093608	0,011331532
0,7	-0,445850404	-0,094112847	0,022776947
0,75	-0,342603095	-0,047762217	0,034337391
0,8	-0,234062183	0	0,046014019
0,85	-0,119956259	0,049216793	0,057808
0,9	0	0,09993246	0,069720512
0,95	0,126106548	0,152192649	0,081752746
1	0,258678717	0,206044398	0,093905907



PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

- ***U(SLAr):*** ***Função Utilidade para SLA no tempo de resposta***

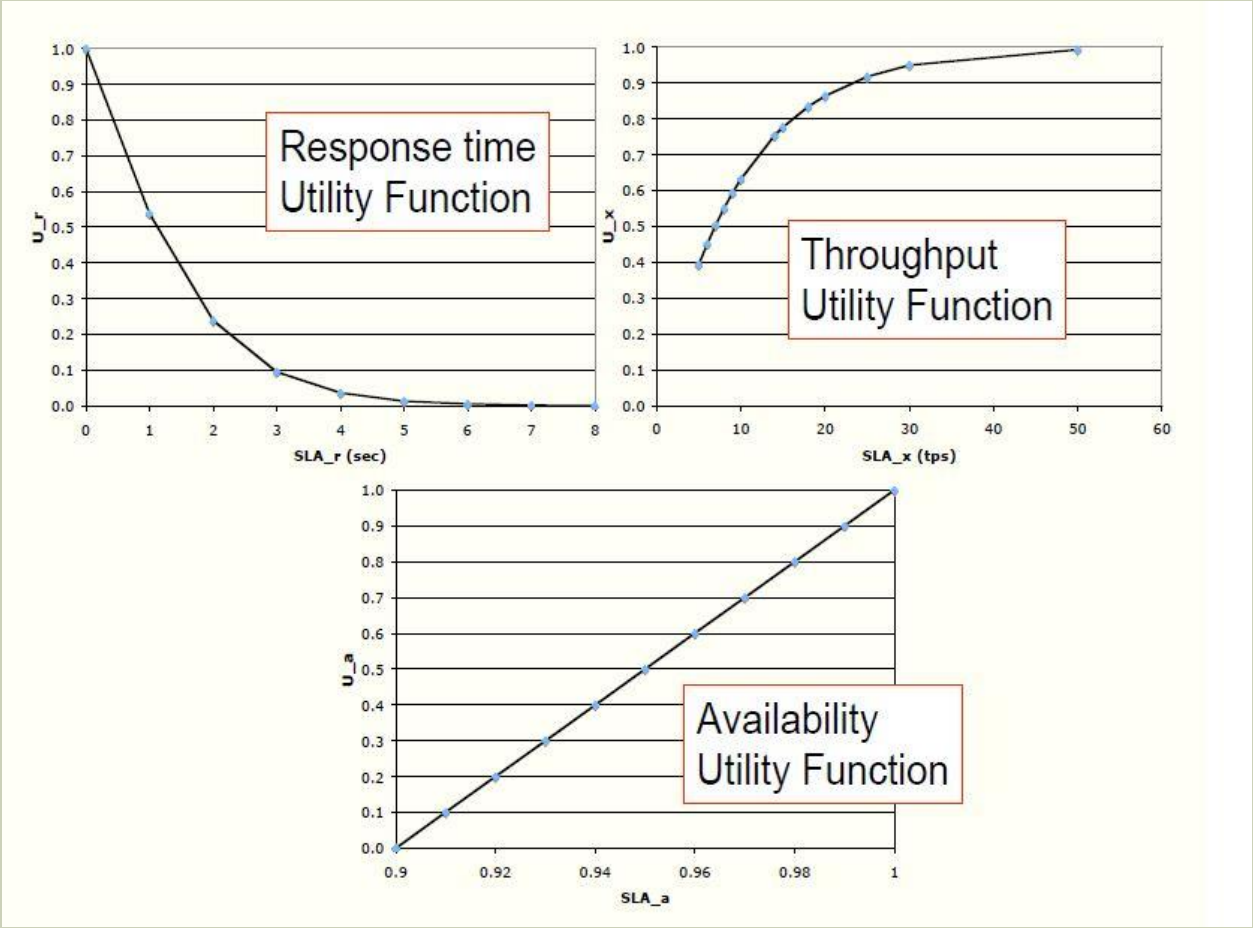
$$\frac{2.0\ e^{-\mathrm{SLA_r}}}{1+e^{-\mathrm{SLA_r}}}$$

- ***U(SLAx):*** ***Função Utilidade para SLA na taxa de processamento***

$$(1-e^{-0.1\ \mathrm{SLA_x}})$$

- ***U(SLAx):*** ***Função Utilidade para SLA na disponibilidade***

$$(10\ \mathrm{SLA_a}-9)$$



PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Planejamento de capacidade para Cloud Computing:

- **A função de Utilidade é dada por:**

$$U = w_r \frac{2.0\ e^{-\mathrm{SLA_r}}}{1 + e^{-\mathrm{SLA_r}}} + w_x(1 - e^{-0.1\ \mathrm{SLA_x}}) + w_a(10\ \mathrm{SLA_a} - 9)$$

Wr + wx + wa = 1 *U ϵ [0,1]*

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Planejamento de capacidade para Cloud Computing:

- Um usuário de nuvem agora enfrenta o problema de selecionar os valores ideais dos SLAs que maximizam a função de utilidade sujeita a SLA e restrições de custo. Isso pode ser expresso como o problema de otimização de restrição não linear mostrado abaixo:

$$
\begin{aligned}
maximize\ U \quad &= \quad f(\mathrm{SLA_r}, \mathrm{SLA_x}, \mathrm{SLA_a}) \\
subject\ to & \\
\gamma_r^{\min} \quad &\leq \mathrm{SLA_r} \leq \gamma_r^{\max} \\
\gamma_x^{\min} \quad &\leq \mathrm{SLA_x} \leq \gamma_x^{\max} \\
\gamma_a^{\min} \quad &\leq \mathrm{SLA_a} \leq \gamma_a^{\max} \\
C_r(\mathrm{SLA_r}) \quad &+ \quad C_x(\mathrm{SLA_x}) + C_a(\mathrm{SLA_a}) \leq C_{max}
\end{aligned}
$$

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Planejamento de capacidade para Cloud Computing:

- O problema de otimização acima indica que os valores dos SLAs para tempo de resposta, throughput e disponibilidade devem ser obtidos para que a função de utilidade seja maximizada. Os SLAs possuem restrições (valores mínimos e máximos), que podem ser inerentes ao que o provedor de nuvem pode oferecer. Há também uma restrição de custo máximo Cmax.

PLANEJAMENTO DE CAPACIDADE PARA CLOUD COMPUTING

Exemplo, considere:

- ***αr = 0,4***
- ***βr = 0,1***
- ***αx = 0,03***
- ***βa = 0,8***

- ***1 s ≤ SLAr ≤ 4 s.***
- ***5 tps ≤ SLAx ≤ infinito***
- ***0,92 ≤ SLAa ≤ 0,99***

- ***wr + wx + wa = 1***

C_{max}	Utility	SLA_r	SLA_x	SLA_a
0.70	0.641	1.000	5.625	0.999
0.60	0.438	3.543	5.000	0.999
0.55	0.366	4.000	5.000	0.978
0.50	0.279	4.000	5.000	0.949

Quando se está disposto a gastar 0,70 cents por transação, o SLAr é o melhor possível (1 segundo), o SLAx está um pouco acima do pior valor possível de 5 tps e a disponibilidade é a melhor possível (0,999).
Como o usuário sempres está menos disposto a gastar mais dinheiro por transação, os SLAs a serem negociados com a nuvem mudam. Por exemplo, para Cmax = 0,60 cents, o usuário precisará se contentar com um SLAr de 3,543 segundos. O SLAx diminui para seu pior valor (ou seja, 5,0 tps). Como o Cmax é reduzido, SLAs piores precisam ser negociados conforme mostra a tabela.
Conforme visto na tabela, o valor da utilidade da nuvem diminui à medida que Cmax diminui.