

# Análise e Modelagem de Desempenho de Sistemas de Computação

Profa. Jussara M. Almeida  
1º Semestre de 2014

# Problemas, problemas, problemas

- Um servidor de banco de dados tem uma CPU e 10 discos. Transações para execução no banco de dados chegam a uma taxa de 2 transações por segundo (2 tps). Durante a sua execução, uma transação alterna processamento com acesso a um dos discos, possivelmente mais de uma vez.

Qual o tempo de resposta médio de uma transação? Qual a taxa de processamento (throughput) do servidor?

O que precisa ser feito???

- Medição em sistema real: se ele estiver em operação (pode ser muito intrusivo, necessita ferramental)
- Simulação: precisa de tanto? Com que nível de detalhe?
- Modelagem analítica: precisa de inputs - quais?

# Problemas, problemas, problemas

- Suponha que uma investigação mais detalhada de um log de um servidor de dados revela que transações individuais submetidas ao servidor têm características diferentes, por exemplo:
  - Tempos de processamento muito heterogêneos
    - Consultas triviais X médias X complexas
  - Categorias diferentes: transações online X processamento (noturno) em batch
  - SLAs (acordos de nível de serviço) diferentes
    - Transações online têm que ser atendidas em menos de 1 segundo, em média, enquanto processos batch não têm requisitos estritos

Qual o tempo de resposta médio de uma transação do tipo X?

Qual o throughput the transações do tipo X?

# Problemas, problemas, problemas

- Suponha agora que o mesmo servidor seja usado como parte de uma aplicação cliente/servidor. Estações de trabalho clientes são conectadas ao servidor de banco de dados através de uma rede LAN. Clientes trabalham de forma independente e alternam entre processamento local ("thinking") e "waiting" por uma resposta do servidor.

Qual o tempo de resposta médio de uma transação de um determinado tipo, incluindo tempo de transmissão na LAN? Qual o throughput do servidor?

# Problemas, problemas, problemas

- Suponha agora que o servidor de banco de dados faça parte de um serviço Web utilizado por aplicações de terceiros (backend). Suponha ainda que os gerentes de IT desejam prover uma garantia no tempo de resposta médio das transações para seus clientes (não pode ser superior a 4 segundos, em média). Para tanto, um sistema de controle de admissão é adicionado ao sistema de forma a limitar o número de transações em processamento simultâneo a no máximo 10 .

A atual configuração do servidor irá atender a necessidade dos gerentes?

Se não, o que terá melhor relação custo-benefício: comprar um disco a mais e fazer balanceamento de carga ou comprar uma CPU 2x mais rápida

# Problemas, problemas, problemas

- O servidor de banco de dados faz parte de um serviço Web, composto de um servidor HTTP (front-end) e um servidor de aplicação (3-tier platform). Os gerentes do serviço querem garantir que a probabilidade de uma requisição ao **serviço** tenha um tempo de resposta superior a 5 segundos seja no máximo 0.05:

$$\text{Prob}(R > 5) = 0.05$$

A atual configuração do servidor irá atender a necessidade dos gerentes?

Se não, o que terá melhor relação custo-benefício: comprar um disco a mais e fazer balanceamento de carga ou comprar uma CPU 2x mais rápida

# Problemas, problemas, problemas

Facing sharp competition, consider an automotive parts distribution company that has decided to develop and implement a new call center. The goals of the new call center system are to: 1) foster better relationships with customers, creating customer loyalty and ensuring quality service, 2) improve efficiency and service performance, and 3) identify and explore new sales opportunities.

The IT team has decided to create an integrated customer relationship management system. The proposed new architecture consists of computer telephony integration software that controls inbound calls and routes these calls to service representatives integrated with application software.

# Exemplo de problema que vamos abordar

The call center will be staffed by service representatives who handle customer orders, returns, queries by phone, fax, and email.

Each customer service representative will have immediate access to the following basic functions: 1) historical tracking of all customer contacts, 2) a single view of customer information, 3) records of resolution of past problems, and 4) help functions. Since customers are less tolerant of long calls, the managers require that the system meets sub-second response times on all functions and be operational 24 hours a day; 7 days a week.



# Problemas, problemas, problemas

The company wants to consider various scenarios when designing its call center. The IT team is planning to design, build, test, and deploy the new call center applications in 12 months. Before rollout, management wants assurances that the performance requirements are satisfied. Questions that they might ask:

- Is the system design able to meet the subsecond response time for all functions? Is that feasible?
- What will be the impact of doubling the number of system representatives in the next year?
- Can acceptable performance levels be maintained after integrating the system with the mainframe-based inventory application?
- Is the system capacity adequate to handle up to 1000 calls in the busiest hour and yet preserve the subsecond response time goal?
- How do failures in the database server affect the 24x7 availability goal?
- What is the impact of starting to offer Web-based self-service to customers?

Problemas,  
problemas  
problemas

NYTimes,  
Apr 19, 2007

Justin Sullivan/Getty Images

Douglas Philips scanned his BlackBerry for e-mail messages in San Francisco Wednesday. More than five million users of the wireless devices were affected by an unexplained nationwide loss of service for 10 hours.

By **BRAD STONE**

Published: April 19, 2007

Where were you when the BlackBerrys went out?



Fred R. Conrad/The New York Times

Ilana Raz checked her BlackBerry outside Grand Central Terminal.

On Tuesday night at 8 p.m. Eastern time, technical problems cut off more than five million [BlackBerry](#) users in the United States from their cherished wireless e-mail. Service was restored 10 long, data-starved hours later.

The BlackBerry blackout was grueling to many — and revealed just how professionally and emotionally dependent so many people had become on their pocket-size electronic lifelines.

Stuart Gold was in Phoenix on a business trip when the service went down. Mr. Gold, the marketing director for Omniture, a software firm, noticed ominous red X's next to his outgoing e-mails.

E-MAIL

PRINT

REPRINTS

SAVE

SHARE

ARTICLE TOOLS  
SPONSORED BY

*Waitress*

Part of the problem, though, could be the service's rapid growth: R.I.M. says it has added three million subscribers in the last 12 months, for a total of eight million, in part because of the popularity of its superslim BlackBerry Pearl.

Problemas,  
problemas,  
problemas

NYTimes,  
Feb 9, 2008

The New York Times

## Media & Advertising

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION  
MEDIA & ADVERTISING WORLD BUSINESS SMALL BUSINESS YOUR MONEY DEALBOOK MARKETS



### Nielsen Tells TV Clients It Is Working on Ending Delays in Ratings

By [BILL CARTER](#)  
Published: February 9, 2008

With many of its television clients, including the broadcast networks, expressing increasing frustration with delays in reporting ratings for programs, the Nielsen Company sent a letter on Friday to its customers expressing disappointment with its performance and promising to do better — though not for a while.

[E-MAIL](#)

[PRINT](#)

[REPRINTS](#)

[SAVE](#)

[SHARE](#)

→ Nielsen now plans to upgrade its computer servers, which are at the company technology base in Oldsmar, Fla. But that may take some time. In its letter, the company said, “We expect to see a material reduction in the number of data delays beginning in mid-March with continuous improvement every 20 business days thereafter for the next four months.”

→ He and other Nielsen critics want to know why the company did not foresee the vast new demands and plan accordingly.

→ Ms. Herkovic conceded the point. “I don’t think we did as good a job as we should have in anticipating these spikes,” she said.

# Problemas, Problemas, Problemas

## Michael Jackson Invented Pop, Beat the Rap, Crashed Google

BY CLAY DILLOW Fri Jun 26, 2009 at 12:00 PM



The news of Michael Jackson's passing yesterday shocked the world, but perhaps none were as stunned as Google and Twitter. Upon rumors of the pop icon's death in Los Angeles, fans flocked to the Web for confirmation, resulting in an avalanche of searches, Tweets, and page views that crashed Google and brought Twitter to a standstill as the number of Tweets per second doubled in an instant.

The tremendous volume of queries for the term "Michael Jackson" was interpreted by Google as an automated attack, triggering a collapse of the site that lasted for about half an hour. More than one-third of the top 100 Google searches among U.S. users were reportedly M.J.-related during the height of the frenzy.



Meanwhile Twitter managed to remain online, but it might as well have crashed. According to Twitter analysis tool [Twist](#), at one point yesterday afternoon 22.6% of messages posted to Twitter were related to the King of Pop's passing as users barraged the site with questions and rumors, and later fond farewells. At one point, nine of the top ten topics on Twitter involved Jackson's death. The surging volume caused the site to grind to a halt temporarily, and Twitter Search was reportedly running about twenty minutes behind real time. This is the same Twitter the U.S. State Department asked to [remain functional](#) last week so protesters in Iran would be able to

# Problemas, Problemas, Problemas

## The PayPal outage cost its users between 7 and 32 million USD

Posted in [Main](#) on August 4th, 2009 by Pingdom

As you may know, PayPal suffered from downtime yesterday. Overall, the problems lasted approximately 4.5 hours before being fully resolved. Since a significant number of e-commerce sites and online services handle some or all of their transactions through PayPal, how much money did the PayPal outage end up costing its users?



According to eBay (which owns PayPal), about \$2,000 in payments flow through PayPal's systems every second. This means that PayPal processes about **\$7.2 million in payments every hour** for its users.

On its official blog [PayPal has stated](#) that the service was completely down, globally, for about one hour. So the outage cost PayPal users *at least* around \$7 million. But PayPal also admits that it took an additional 3.5 hours after that before the service was fully restored, something we also need to take into account.

**This means that the outage and following service problems cost Paypal users somewhere between \$7 million and \$32 million in lost payments.**

Exactly where the actual cost landed between those two extremes is hard to tell, but judging by the comments we have seen in blogs and on Twitter, the problems were significant even after the initial one-hour service outage.



# Problemas, Problemas, Problemas



EVENTS

GIGAOM.TV

PRO RESEARCH

HOMEAPPLEBROADBANDCLEANTECHCLOUDCOLLABORATIONMOBILEVIDEO

NewTeeVee

Search

## CNN: Inauguration P2P Stream a Success, Despite Backlash

By Janko Roettgers | Feb. 7, 2009, 12:01am PT | 16 Comments

 Tweet 0 Share Like



Sometimes it's tough to be the winner. CNN [attracted by far the largest online audience](#) for its web stream of President Obama's inauguration, delivering a total of 25 million streams on Jan. 20th. It did so by utilizing [Octoshape's P2P plug-in](#) for Adobe's Flash player, a fact that was widely known within the industry but only received very [little coverage](#). Until yesterday, that is, when [Windows Secrets](#) suddenly came out [blasting](#) the network for "deceptive marketing," alleging that CNN hijacked its users' upload bandwidth. Others quickly jumped on the bandwagon, [going so far as to suggest](#) that "the most trusted name in news just stole your computer."

I got in touch with CNN and Octoshape to get both sides of the story, which are, unsurprisingly, slightly different. Both companies reject the claim that they misled viewers. But Mike Wise, technical adviser of the R&D group of CNN parent company Turner Broadcasting System, had an even more important point to make: CNN used P2P, he told me, because it had to.

The inauguration is now seen as one of if not the largest live video event in the history of the Internet, and the folks at CNN believe that traditional CDNs simply weren't up to the task. "Turner anticipated that the resulting stresses on the various CDNs and the Internet in general would make it extremely difficult if not impossible to serve such an unprecedented audience without the use of P2P technologies." Wise told me in an email.



Free Trial



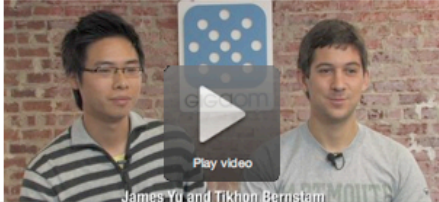
Let the MPAA Speak, There's Nothing To Be Scared Of | TorrentFreak  
<http://t.co/rND73pJ>

JANKO'S POSTS

Netflix tests new web UI for kids. First look

Cord Cutters: Checking out the Boxee iPad app

GIGAOM TV



Play video

James Yu and Tikhon Bernstein

# Problemas, Problemas, Problemas

## Amy Winehouse death gets 20 million Twitter mentions

Lucy Carne | The Sunday Telegraph | July 24, 2011 1:34PM | 25 comments

Share

Recommend

Send

28 people recommend this. Be the first of your friends.

10 retweet

Share



The body of dead English pop star Amy Winehouse is carried from her house in north London.

Picture: AFP Source: The Sunday Telegraph

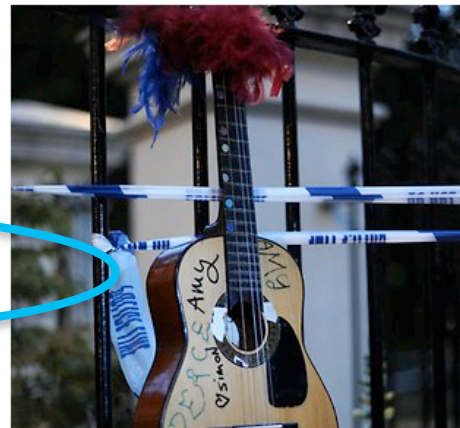
< Prev 1 of 3 Next >

### AMY Winehouse's tragic death triggered an instant outpouring of grief that swept the world.

Within seconds of the confirmation that the British five-time Grammy award-winning singer had died in her London home, fans and celebrities took to Twitter to express their shock and sorrow.

In a wave of mourning, Winehouse became the highest searched name on the Internet and was mentioned by 20 million people on Twitter within minutes of the news breaking of her death.

Celebrities, many of whom were friends with the singer, expressed their immense sadness at her sudden death, believed to be from a drug overdose.



# Problemas, Problemas, Problemas

TECH | 1/16/2014 @ 6:16PM | 2,404 views

## Web Performance Is User Experience

+ Comment Now + Follow Comments

By [Lara Swanson](#)

Think about how you search for things on the web. How quick are you to close a tab and go to the next search engine result if a site takes too long to load? Now consider doing that on your phone while waiting in line for your coffee order—you have even less time, so your expectations for a site to load quickly are even higher.

Web performance is user experience. Fast page load time builds trust in your site; it yields more returning visitors, more users choosing your site over a competitor's site, and more people trusting your brand. Users expect pages to load in two seconds, and after three seconds, up to 40% of users will abandon your site. Similar results have been noted by major sites like Amazon, who found that 100 milliseconds of additional page load time decreased sales by one percent, and Google, who lost 20% of revenue and traffic due to half a second increase in page load time. Akamai has also reported that 75% of online shoppers who experience an issue such as freezing, crashing, taking too long to load, or having a convoluted checkout process will not buy from that site.

Web performance impacts more than just e-commerce sites; improvements from page speed optimization apply to any kind of site. Users will return to faster sites, evidenced in a study by Google Maps that noted an increase in returning traffic when the Google Maps homepage weight was reduced from 100KB to 80KB. Additionally, page load time is factored into search engine results, bumping faster sites higher in the results list than slower sites.

Page load time also has a significant impact on mobile users' experience. My team at Etsy found an increased bounce rate of 12% on mobile devices when we added 160KB of images to a page. DoubleClick removed one client-side redirect and saw a 12% increase in click-through rate on mobile devices. In another study, researchers found that if Amazon changed all of their images to compressed JPEG files, it would save 20% of the energy needed to load the



# Problemas, Problemas, Problemas

High Performance Web Sites blog

[home](#) | [more St](#)

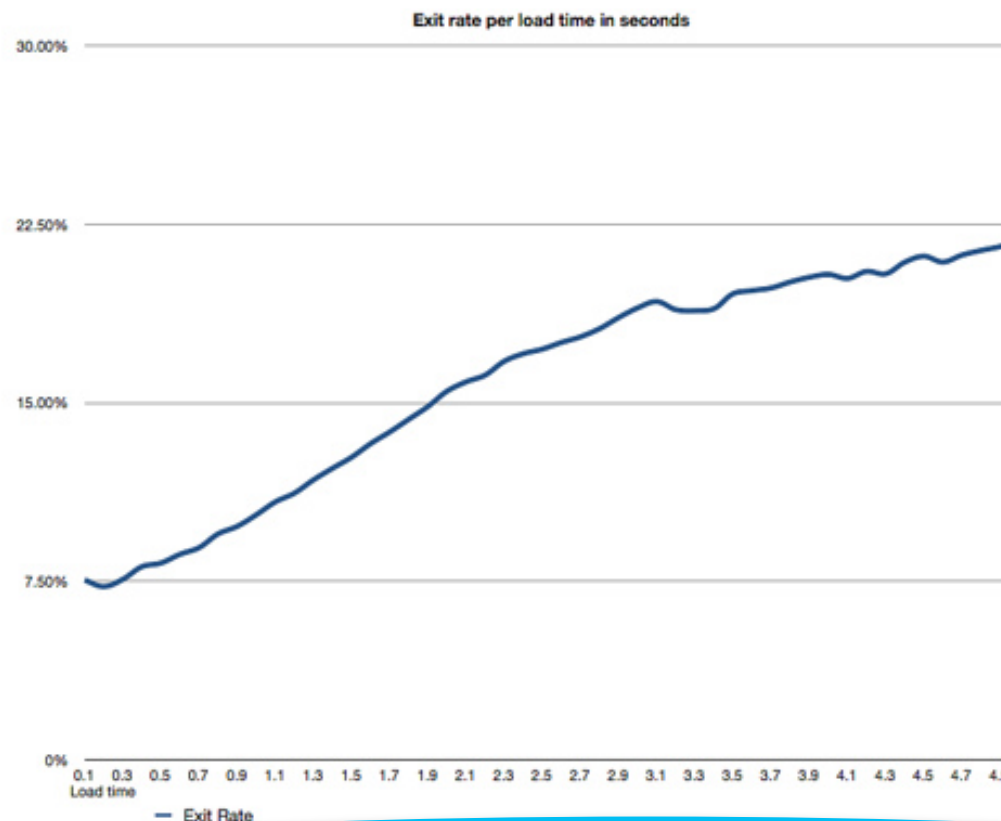
SEARCH

Find

## Wikia: fast pages retain users

July 27, 2009 10:52 PM

At [OSCON](#) last week, I attended Artur Bergman's session about [Varnish - A State of the Art High-Performance Reverse Proxy](#). Artur is the VP of Engineering and Operations at [Wikia](#). He has been singing the praises of Varnish for awhile. It was great to see his experiences and resulting advice in one presentation. But what really caught my eye was his last [slide](#):

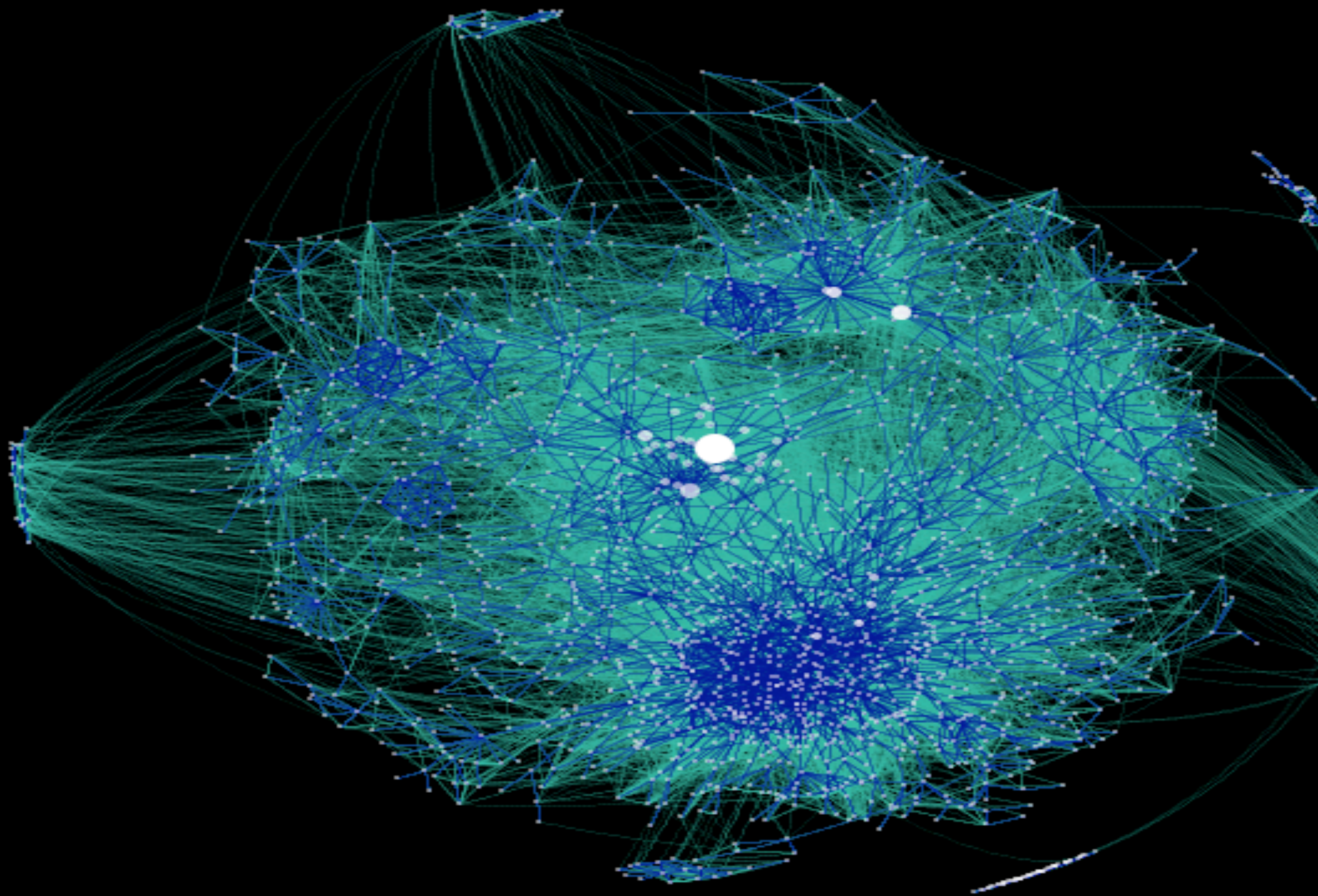


Wikia measures *exit rate* - the percentage of users that leave the site from a given page. Here they show that exit rate drops as pages get faster. The exit rate goes from ~15% for a 2 second page to ~10% for a 1 second page. This is another data point to add to the list of stats from [Velocity](#) that show that faster pages is not only better for users, it's better for business.

# Principais Desafios

- Complexidade e diversidade de aplicações, serviços, plataformas e infraestruturas





# Principais Desafios

- **Grandes volumes de dados:**
  - 2,7 bilhões de usuários da Internet (39% população)
  - 1,2 bilhões de usuários ativos/mês no Facebook
  - 240 milhões de usuários ativos/mês no Twitter
  - 500 milhões de tweets por dia
  - + 1 billion visitas/mês no YouTube
- Dados raramente disponíveis publicamente
- Coleta, processamento e armazenamento:
  - Amostragem
  - Representatividade

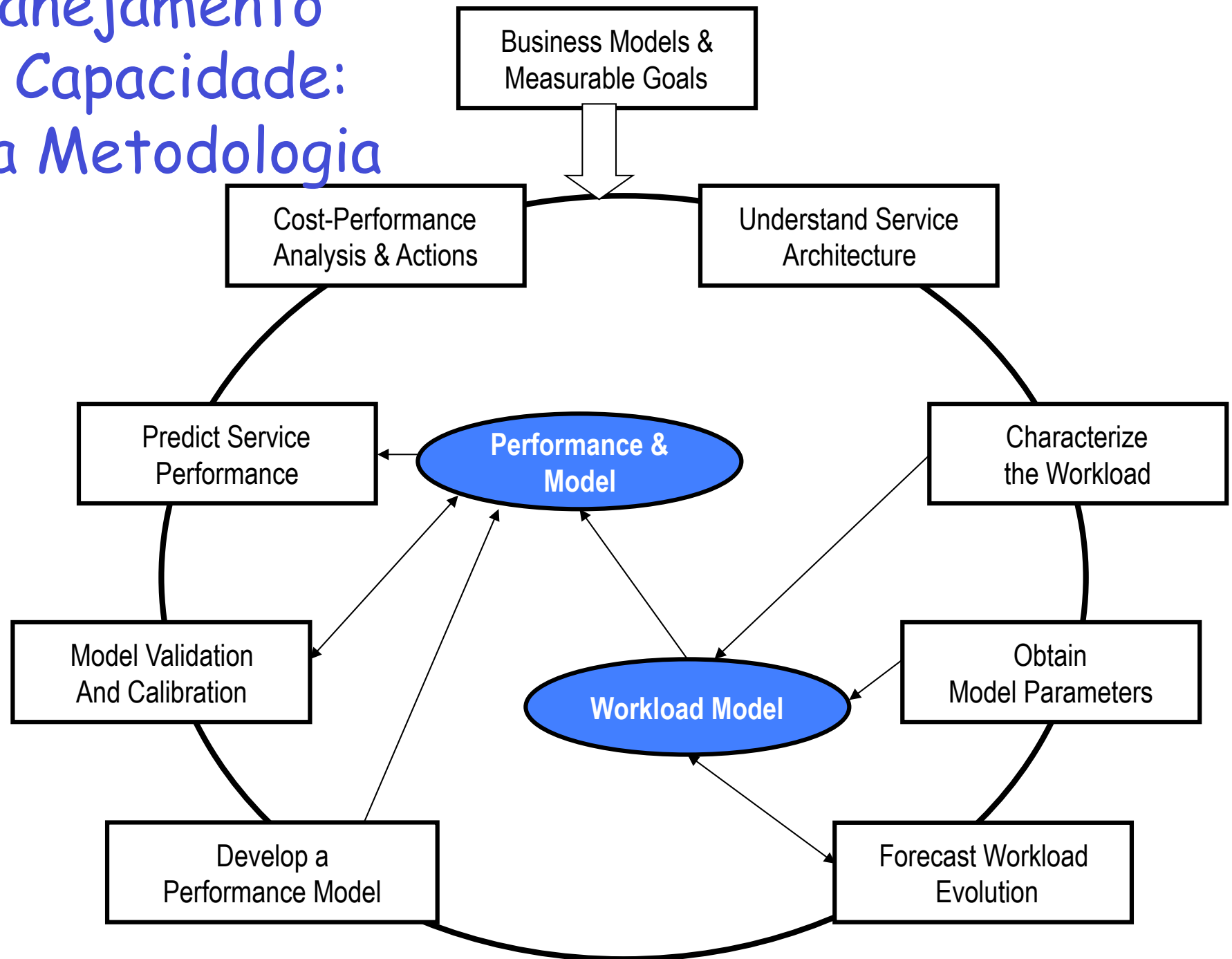


# Principais Desafios

- Padrões heterogêneos e dinâmicos
  - Evolução temporal
  - Alta variabilidade e mudanças drásticas (flash crowds)
  - Interpretação sujeita a alto grau de subjetividade



# Planejamento de Capacidade: Uma Metodologia





# A disciplina

- Objetivo: aprender técnicas que permitam responder perguntas como as mostradas com baixo custo e precisão aceitável.

Good models abstract out some of the complexities of real systems while retaining what is essential to meet the model goals

- Foco:
  - Modelagem analítica de desempenho: fórmulas **simples**, que fornecem resultados **rápidos** e com **precisão adequada** para vários **problemas práticos**.
  - Caracterização e modelagem de cargas de trabalho

# A disciplina

- Técnicas:
  - Modelagem de Desempenho
    - Leis fundamentais
    - Análise de valores limites
    - Análise de valores médios
    - Filas M/M/1, M/G/1
  - Modelagem de Cargas
    - Caracterização, benchmarks
- Discussão de artigos atuais (?)
  - Apresentações/seminários por alunos
- Avaliação (tentativa):
  - Projeto : 40 pontos
  - 2 provas : 50 pontos
  - Participação (seminários, discussão, listas de exercícios): 10 pontos

# Quem anda estudando "performance"?

- Teórica: construção e desenvolvimento de modelos analíticos
- Experimental: monitoramento, medição e teste
- Aplicações
  - Engenharia de software: requisitos não funcionais
  - Arquitetura de computadores
  - Redes, sistemas paralelos e distribuídos
  - Sistemas Operacionais
  - Recuperação de Informação
  - Internet-Web (profissão de Web analyst)
    - Data centers, cloud computing
    - Logística digital
  - Aplicações críticas: corporativas, segurança pública

# O que é desempenho?

[Aurélio]

desempenho. [Dev. de desempenhar.] *s.m.*,

1. Ato ou efeito de desempenhar(-se).
2. *Mil.* Conjunto de características ou de possibilidades de atuação de uma aeronave, tais como velocidade de cruzeiro, velocidade de pouso, capacidade de carga, autonomia de vôo, etc.
3. *Teat.* *V. interpretação.*

# O que é desempenho?

[Webster]

per.for.mance n p\*(r)-'for-m\*n(t)s\

1. a: the execution of an action  
b: something accomplished : **DEED, FEAT**
2. : the fulfillment of a claim, promise, or request:  
**IMPLEMENTATION**
3. a: the action of representing a character in a play  
b: a public presentation or exhibition
4. a: the ability to perform: **EFFICIENCY**  
b: the manner in which a mechanism performs
5. : the manner of reacting to stimuli: **BEHAVIOR**

# O quê, como e por quê?

- O que é desempenho?
  - Medida da *capacidade de resposta* de um sistema
- Como analisar o desempenho de um sistema?
  - Medições, análise quantitativa
  - Métricas de desempenho
- Por quê?
  - Aspecto chave no projeto, compra, gerenciamento e otimização de um sistema de computação
    - Ex: desempenho de sistemas da Web é crítico para vida contemporânea



# Para quê?

- Identificar pontos de contenção e/ou de desperdício de um sistema
- Avaliar múltiplas alternativas de projeto e/ou implementação de uma aplicação
- Determinar configuração ótima dos parâmetros de um sistema
- Fazer previsão de desempenho em função do crescimento da carga
- Planejamento de capacidade
- Validação  
(identificação de erros de implementação)

# Complexidade

- Ambientes complexos e dinâmicos
  - Sistema = hardware + SO + aplicações
  - Surgimento de novas aplicações
  - Determinação de premissas/escopo é essencial
- Dependência da carga:
  - Comportamento de usuário é imprevisível (??)
  - Mudanças drásticas: flash crowds (11/9, morte do Michael Jackson)
- Requisitos (QoS)
  - *Service Level Agreement (SLA)*
    - Tempo de resposta médio inferior a 1 segundo
    - Taxa de processamento superior a X transações por seg (quanto no Google?)

# Google: Complexidade

- 2008:
  - Searches on Google no mundo inteiro: 48.7 bi
  - Searches/hour on Google em Julho: 65 mi.
  - Searches on Microsoft no mundo inteiro: 2.3 bi
  - Searches/hour on Microsoft em Julho: 3.1 mi



- Mais recente:
  - <http://www.alexa.com/siteinfo/google.com>
  - Ranking in 2014: Google: 1<sup>st</sup> / MS Bing: 23<sup>rd</sup> (world-wide)

# Qual a meta?



"Users really respond to speed."

Marissa Mayer, Google

"Worry about the data first  
before you worry about the  
algorithm"

Peter Norvig, Google

+500 ms      →      -20% traffic  
@Google

+100 ms      →      -1% sales  
@Amazon



Observar  
Medir  
Aprender  
Otimizar



# Desempenho

Figure 2: Effect of distance on throughput and download times.

| Distance from Server to User | Network Latency | Typical Packet Loss | Throughput (quality)  | 4GB DVD Download Time |
|------------------------------|-----------------|---------------------|-----------------------|-----------------------|
| Local: <100 mi.              | 1.6ms           | 0.6%                | 44Mbps (HDTV)         | 12 min.               |
| Regional: 500–1,000 mi.      | 16ms            | 0.7%                | 4Mbps (not quite DVD) | 2.2 hrs.              |
| Cross-continent: ~3,000 mi.  | 48ms            | 1.0%                | 1Mbps (not quite TV)  | 8.2 hrs.              |
| Multi-continent: ~6,000 mi.  | 96ms            | 1.4%                | 0.4Mbps (poor)        | 20 hrs.               |

## Improving Performance on the Internet

Tom Leighton, AKAMAI, Comm. of the ACM Feb 2009

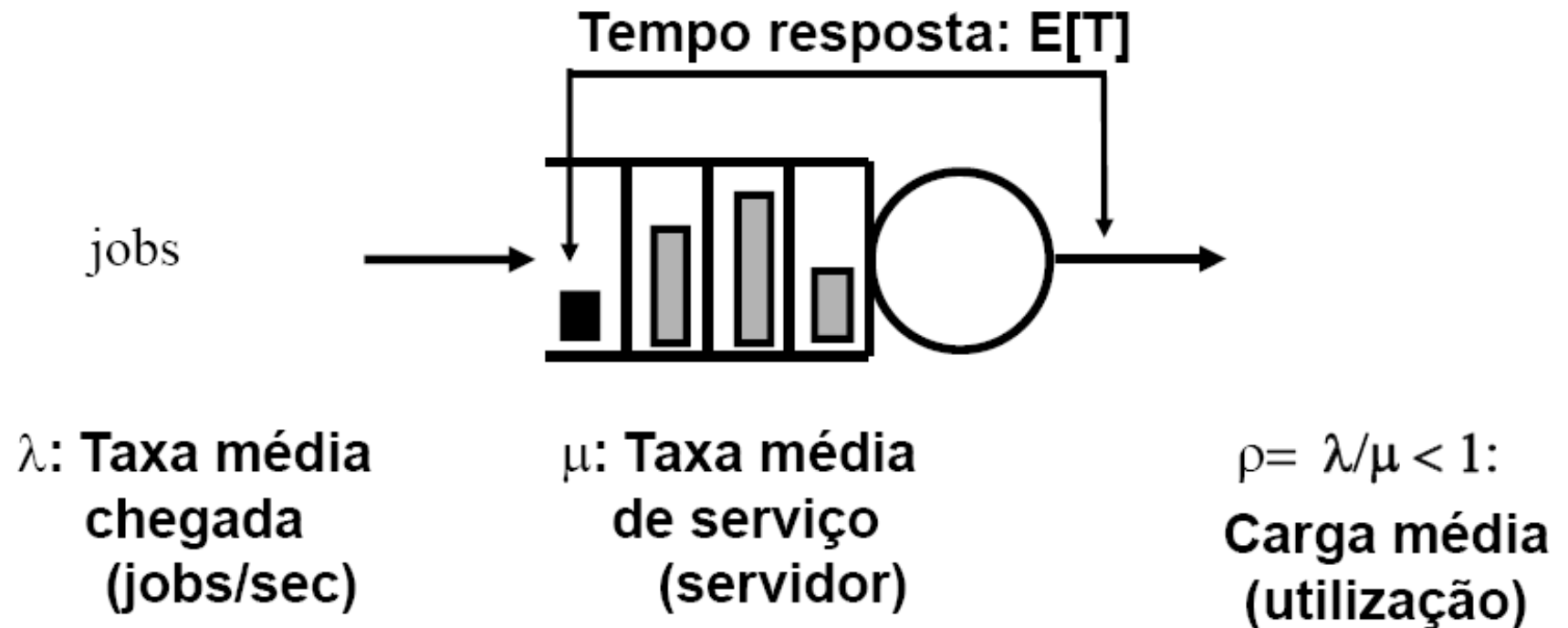
When it comes to achieving performance, reliability, and scalability for commercial-grade Web applications, where is the biggest bottleneck? In many cases today, we see that the limiting bottleneck is the middle mile, or the time data spends traveling back and forth across the Internet, between origin server and end user.

# It is all about queues...



Fila ➡ gargalos (contenção) ➡ espera

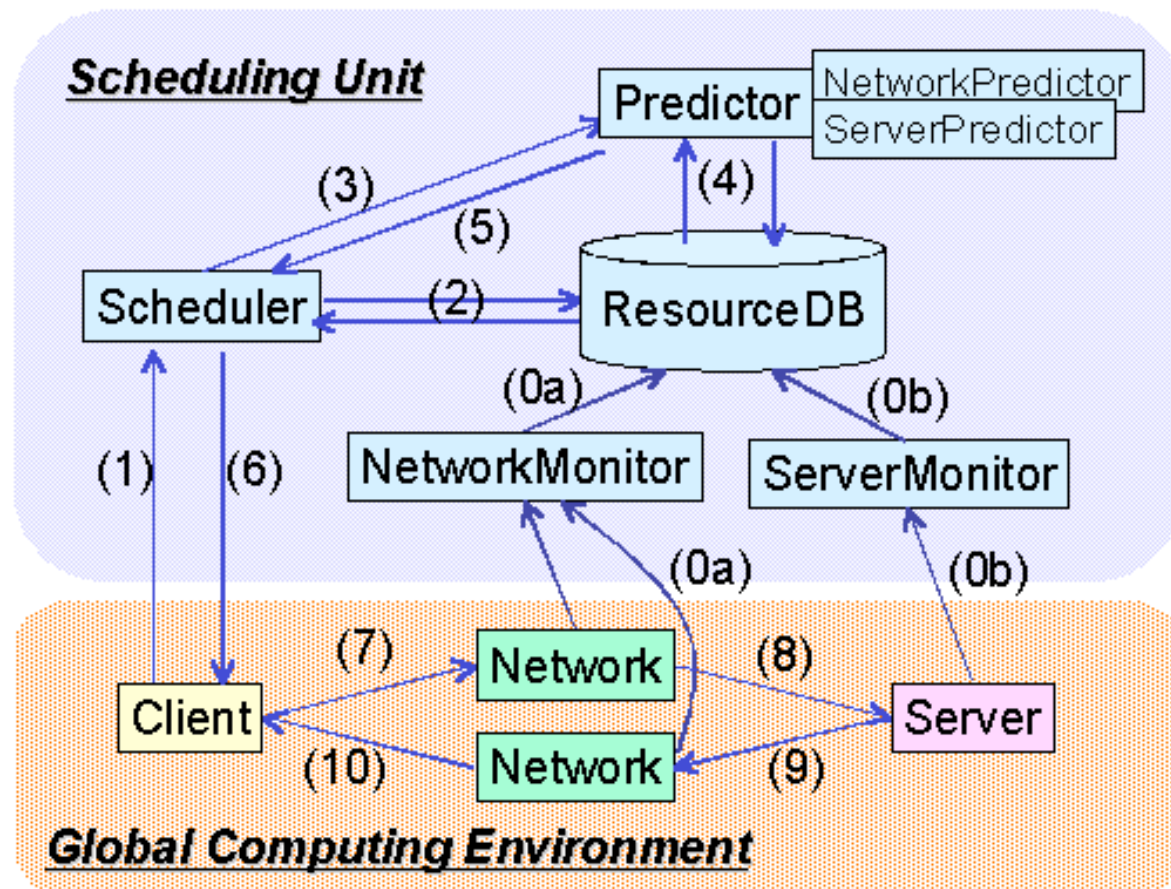
# O Modelo Básico



Servidor: roteadores  
Processadores  
Locks de Banco de Dados  
Web Servers  
Pontos de acesso a WiFi  
Threads

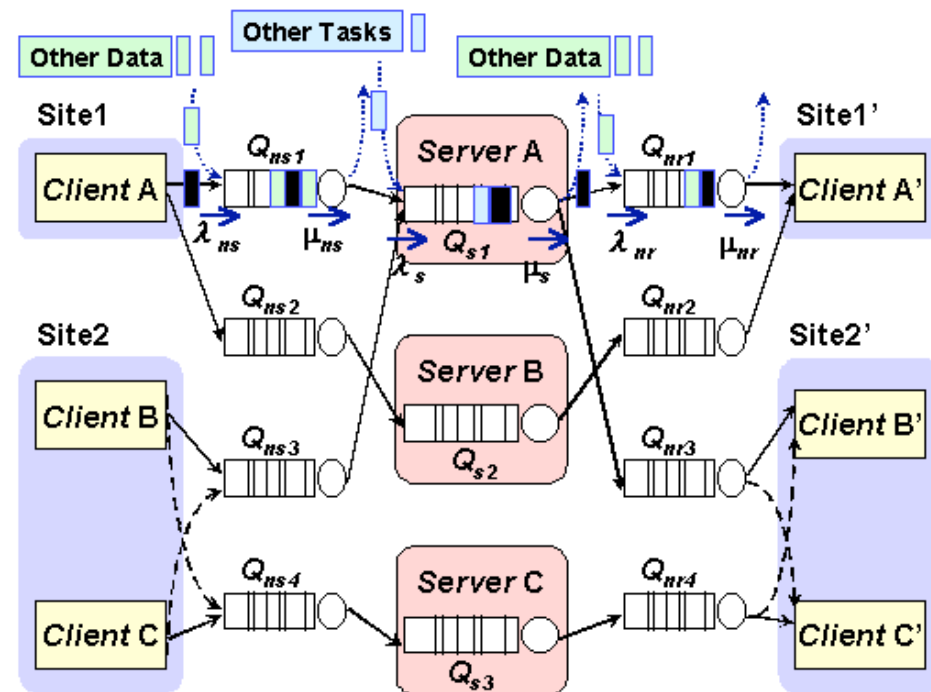
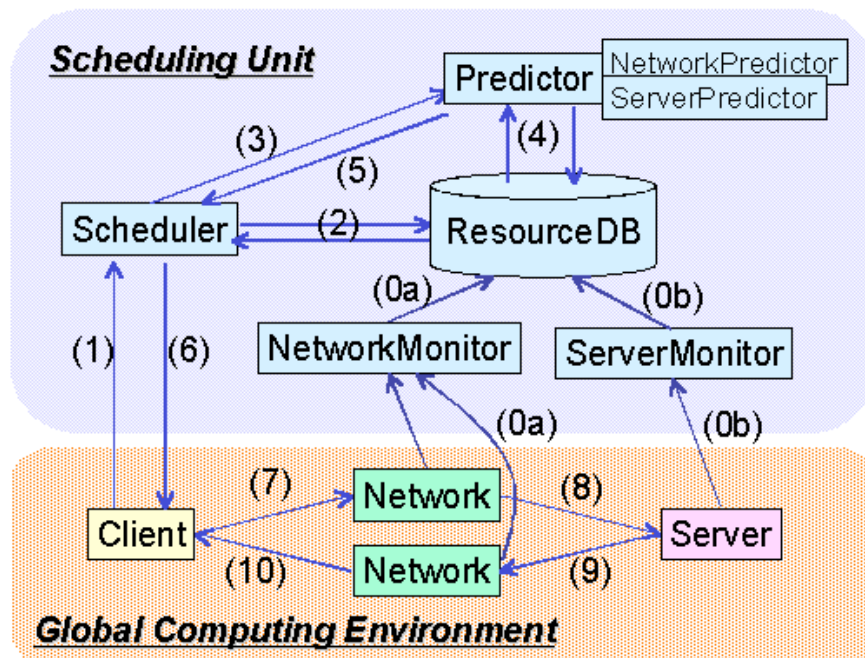
....

# Mapeando um sistema de computação em um sistema de filas

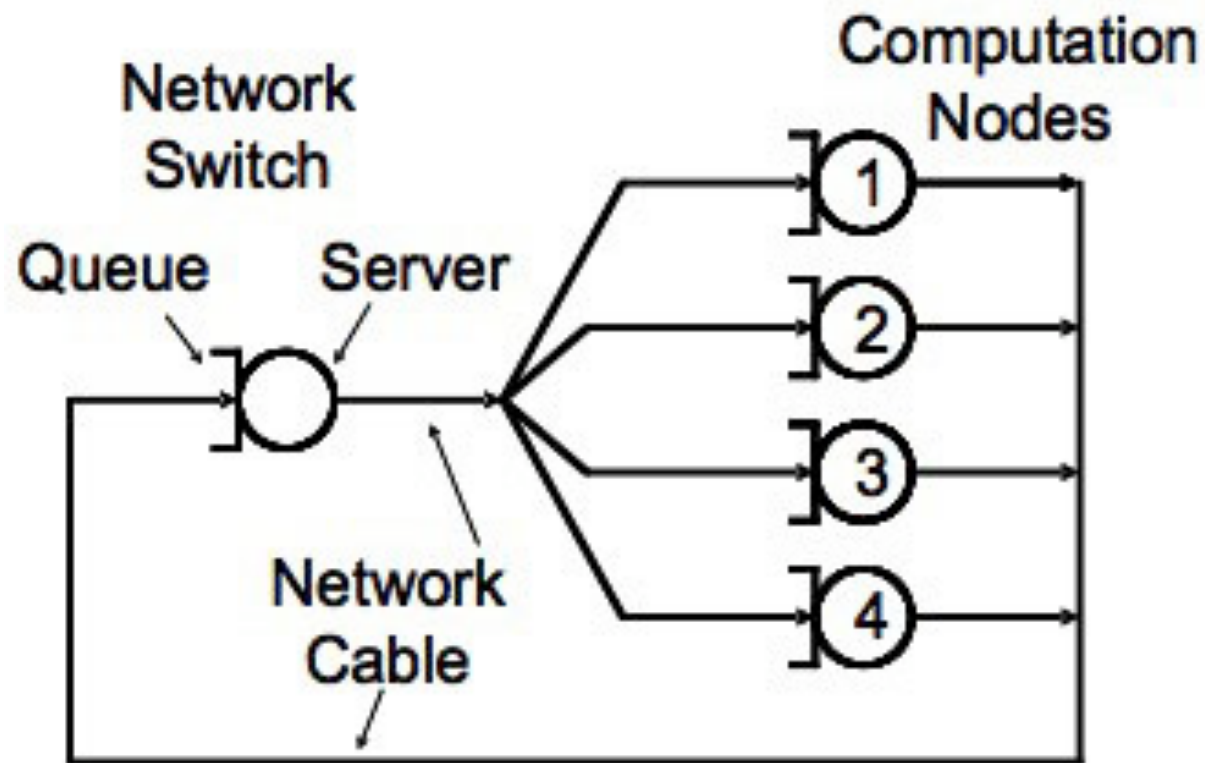




# Mapeando um sistema de computação em um sistema de filas

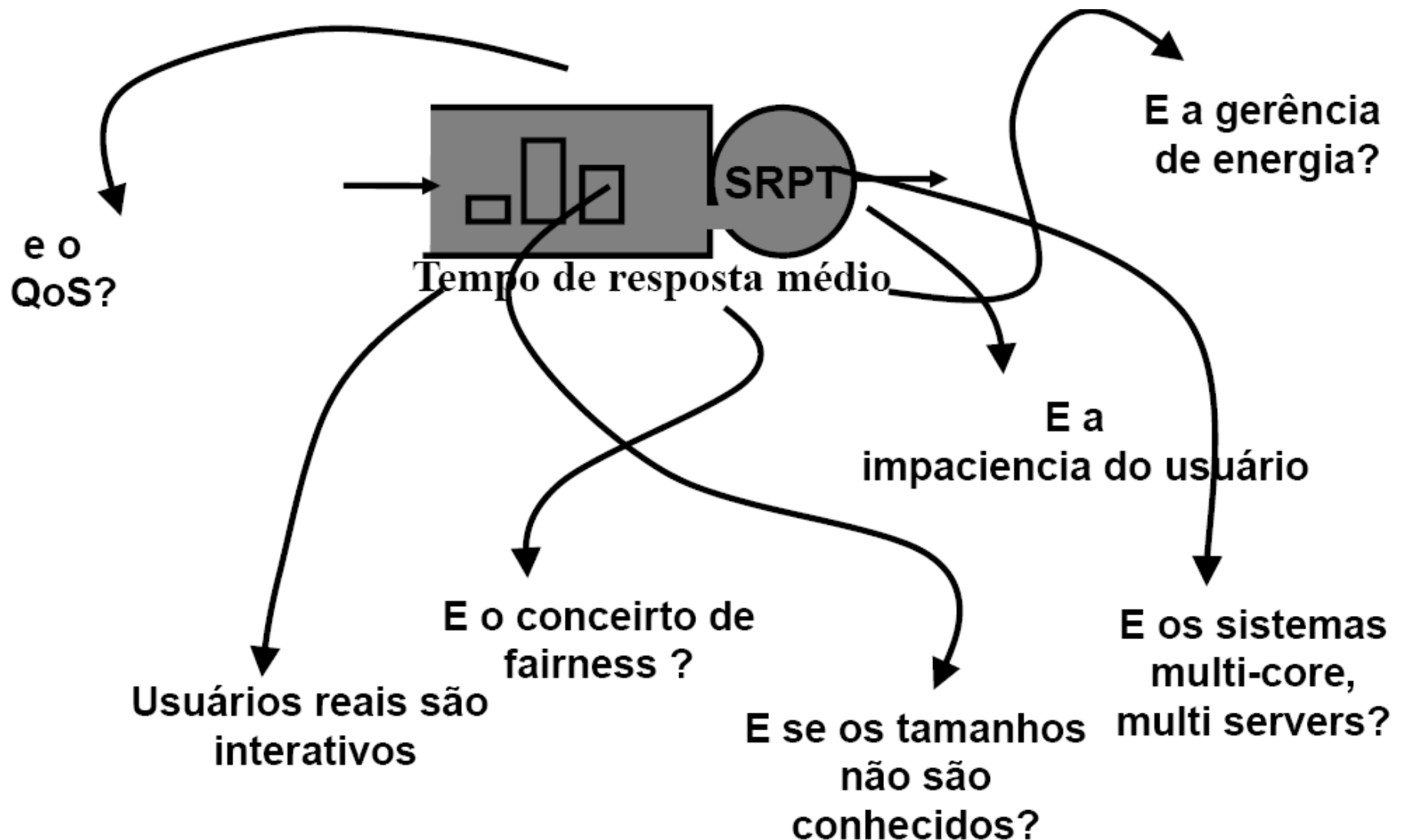


# Um outro exemplo...

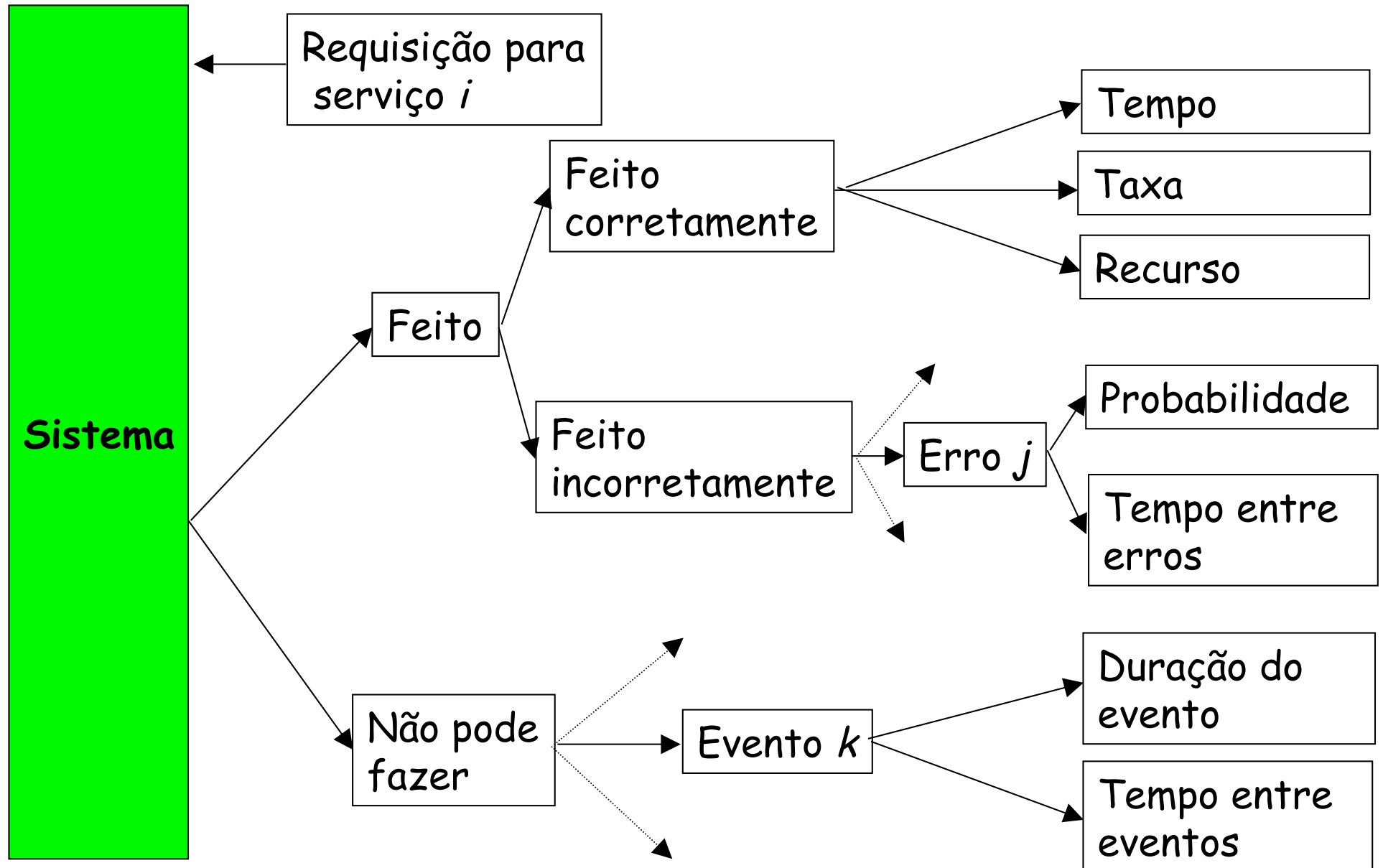




# Desafios do mundo real ...



# Métricas de Desempenho



# Métricas de Desempenho

- Tempo de resposta (total, por componente)
- Taxa de processamento (*throughput*)
  - Transações/s, HTTP requests/s, sessões/s, I/O/s, etc.
- Utilização
- Escalabilidade
- Disponibilidade (falhas ou sobre-carga)
  - Controle de admissão
  - Disponibilidade vs. tempo de resposta
- Confiabilidade
- Custo

# Métricas de Desempenho

- Segurança: confidencialidade, integridade
  - autenticação
- Extensibilidade: fácil evolução para incluir novos requisitos funcionais e de desempenho
- A escolha da(s) métrica(s) mais adequada(s) é chave para o sucesso de uma avaliação de desempenho
  - Eliminar redundâncias (tempo de espera e tamanho da fila)
  - Evitar métricas que têm alta variabilidade (ex: razão de duas variáveis)
  - O conjunto de métricas deve ser completo (ex: protocolo + rápido causa maior # de desconexões)
  - As métricas devem ser realísticas e mensuráveis!
- Métricas devem ser possíveis de se obter!

# Métricas de Desempenho: Exemplos

<http://www.internethealthreport.com>

# Avaliação de Desempenho

- Modelagem analítica:
  - Conjunto de fórmulas e/ou algoritmos que determinam as medidas de desempenho (métricas) em função de um conjunto de parâmetros de carga (entrada)
  - Modelos determinísticos: análise de limites e de valores médios
  - Modelos probabilísticos: processos estocásticos, modelos de filas
- Modelos de simulação
- Experimentação em sistemas reais
  - Instrumentação
  - Monitoração

# Para casa e Projeto

- Para casa: 28/08 (em sala)
  - Ler 1º capítulo de "Performance by Design", by Menasce, Almeida and Dowdy
  - Escrever resenha (até 2 páginas) sobre o capítulo
  - Entregar no dia 18/02, em sala
- Proposta de projeto: 4/03
  - Contextualização/escopo e definição do problema
  - Perguntas que pretende responder
  - Cenários, parâmetros e métricas de avaliação