

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

TRABALHO DE FORMATURA

Planejamento de Capacidade Para Web Sites

Orientador: Prof. Mauro Spínola

Roberto Gonçalves Jábali
São Paulo
2000

14 2000
JABALI

Agradecimentos

Inicialmente ao professor Mauro Spínola pela seriedade e comprometimento, sem os quais este trabalho não poderia ser realizado.

A toda a equipe de projeto da Andersen Consulting, representada por Carlos Ferrarini, André Rezende e José Armando Porto, pelo apoio técnico, compreensão e crescimento pessoal que me proporcionaram durante toda a fase de desenvolvimento deste trabalho.

A meus pais, José Munyr e Maria Eugênia por todo o respeito e apoio emocional que tiveram durante toda a minha vida.

À Érica pela superação de tantos momentos difíceis e ausências.

A turma da sala, Alex, Paulo, Luiz, Marcel, Thiago, Leonardo, Renato, Daniel, pelo espírito de ajuda em momentos críticos.

A todos meus companheiros da Bateria da Poli, representados por Fábio, João e Fábio Mollica por acreditarem em um objetivo.

A meus eternos amigos Jefferson, Jean, Marcelo, Fernando, Maurício e Michele que poderiam também ser chamados de irmãos.

Sumário

Na nova economia em que vivemos, é de vital importância que se tenha um conhecimento sobre novas descobertas, principalmente as que podem ter uma influência direta na maneira de viver das pessoas.

A Internet é, sem dúvida, um dos mais importantes temas existentes na atualidade e terá um papel fundamental na integração entre pessoas que antes nem imaginavam um universo tão vasto de informações.

Neste contexto, surge o chamado Comércio Eletrônico que já tem um papel bastante importante em segmentos de mercado que podem tanto envolver exclusivamente empresas, no chamado comércio eletrônico Business to Business, quanto empresas relacionadas ao consumidor Business to Consumer.

Em ambos os casos, já existe uma forte concorrência entre empresas com mesmo ramo de atuação, o que provoca uma busca por alternativas de diferencial competitivo que levem essas empresas, nem sempre muito experientes, a uma posição de destaque no mercado. Dada a facilidade que os usuários da Internet têm em comparar diferentes empresas, a confiabilidade de um site e a rapidez com que este possa atender às solicitações de seus clientes torna-se indispensável para aqueles que realmente queiram estabelecer um negócio de longo prazo.

É justamente sobre a performance de Web Sites que será desenvolvido o tema deste trabalho. O objetivo é que se tenha uma metodologia para se planejar uma infra-estrutura que atenda a requisitos previamente especificados e que possibilite um desempenho adequado do Web Site.

Inicialmente é desenvolvida uma parte teórica sobre o perfil de usuários e sobre o mercado de empresas virtuais. Na sequência faz-se um estudo técnico sobre a Infra-estrutura de Web Sites, principais problemas do sistema e, finalmente, apresenta-se o Modelo de Planejamento de Capacidade (MPC), que servirá como base para o caso prático visto no final do trabalho.

ÍNDICE

Capítulo 1 - Introdução	1
1.1 Proposta de trabalho	2
1.2 Motivação	3
1.3 Foco	4
Capítulo 2 - E-Commerce	6
2.1 Definição	7
2.2 Histórico	8
2.3 Relevância do Foco Escolhido	10
2.3.1 Benefícios do E-Commerce	11
2.3.2 Limitações do E-Commerce	13
2.4 Estudo do Mercado Atual de Empresas Virtuais	15
2.5 Estudo Atual do Perfil de Usuários de Internet no Brasil	15
2.5.1 Perfil de usuários	16
2.5.2 Hábito dos usuários	22
2.5.3 Consumo via Internet	24
Capítulo 3 - Performance em Sistemas	27
3.1 Redes	28
3.1.1 Tipos de rede	28
3.1.2 Protocolos de comunicação	31
3.1.3 O Modelo Computacional Cliente/Servidor	33
3.2 Quando performance é um problema	35
3.2.1 Exemplo de Web Site com problemas de capacidade.	36
3.3 Performance de arquiteturas Cliente/Servidor	38
3.3.1 Gráficos de Processamento de Comunicação	39
3.3.2 Componentes de sistemas Cliente/Servidor que podem ocasionar atrasos:	42
3.3.3 Definição das variáveis de performance que compõem o sistema	47

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

Departamento de Engenharia de Produção - Trabalho de Formatura

3.3.4 Algumas Conclusões Importantes Sobre Performance	48
3.3.5 Métricas para avaliação de performance em sistemas Cliente/Servidor	50
3.4 Particularização: Performance de Web Sites	53
3.4.1 Web Servers	54
3.4.2 Análise de transações Web	59
3.4.3 Métricas para se avaliar o desempenho de Sistemas Web	63
Capítulo 4 - Metodologia para planejamento de capacidade	65
4.1 Capacidade Adequada	66
4.2 O Modelo de Planejamento de Capacidade (MPC)	67
4.2.1 Fase 1 - Delimitar e Entender o Problema	71
4.2.2 Fase 2 - O Modelo de Carga de Trabalho (MCT)	72
4.2.3 Fase 3 - O Modelo de Performance (MP)	74
4.3 Aplicação do MPC a um caso real	77
4.3.1 Fase 1 - Delimitar e Entender o Problema	77
4.3.2 Fase 2 - O Modelo de Carga de Trabalho (MCT)	81
4.3.3 Fase 3 - O Modelo de Performance (MP)	85
4.4 Principais restrições do modelo	92
Capítulo 5 - Conclusões	94
5.1 Revisão Geral	95
5.2 Avaliação de alcance dos objetivos propostos	96
5.3 Estudo da contribuição prática deste projeto	98

ÍNDICE DE FIGURAS

<i>Figura 2.1 – Sexo dos Internautas.....</i>	<i>16</i>
<i>Figura 2.2 – Ocupação dos Usuários.....</i>	<i>17</i>
<i>Figura 2.3 – Posição Hierárquica dos Usuários</i>	<i>17</i>
<i>Figura 2.4 – Ramo de Atuação dos Usuários</i>	<i>18</i>
<i>Figura 2.5 – Grau de Escolaridade dos Usuários.....</i>	<i>19</i>
<i>Figura 2.6 – Renda Familiar.....</i>	<i>20</i>
<i>Figura 2.7 – Usuários por Região do País.....</i>	<i>21</i>
<i>Figura 2.8 – Estado Civil dos Usuários</i>	<i>21</i>
<i>Figura 2.9 – Quantidade de Acessos.....</i>	<i>22</i>
<i>Figura 2.10 – Principal Atividade na Internet.....</i>	<i>23</i>
<i>Figura 2.11 – Usuários por Região do País.....</i>	<i>24</i>
<i>Figura 2.12 – Possibilidade de Consumo via Web</i>	<i>25</i>
<i>Figura 3.1 (a) LAN Bus-based (b) Ring-Based LAN (c) Dual Ring-Based LAN...</i>	<i>30</i>
<i>Figura 3.2 - Camadas de Protocolos de Comunicação Baseados em TCP/IP</i>	<i>33</i>
<i>Figura 3.3 – Servidores: (a) Arquivos (b) Consultas (c) Aplicativos.....</i>	<i>35</i>
<i>Figura 3.4 - Gráfico de Processamento de Informações.....</i>	<i>39</i>
<i>Figura 3.5 - Arquitetura de um Sistema I/O.....</i>	<i>43</i>
<i>Figura 3.6 - Conectividade Cliente/Servidor</i>	<i>45</i>
<i>Figura 3.7 - Filas em Roteadores.....</i>	<i>46</i>
<i>Figura 3.8 - Representação Gráfica para um Recurso e sua Fila de Espera.....</i>	<i>49</i>
<i>Figura 3.9 - Definição de Intervalo de Decisão, de Reação e Tempo de Resposta. .</i>	<i>51</i>
<i>Figura 3.10 - Componentes de Redes Internet ou Intranet.....</i>	<i>53</i>
<i>Figura 3.11 - Interação Cliente/Servidor Utilizando o Protocolo HTTP 1.0 e 1.1 ..</i>	<i>57</i>

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

<i>Figura 3.12 - Anatomia de uma Transação Web</i>	<i>60</i>
<i>Figura 4.1 - O Modelo de Planejamento de Capacidade (MPC)</i>	<i>69</i>
<i>Figura 4.2 – Decomposição Entre Componentes do Sistema</i>	<i>73</i>
<i>Figura 4.3 - A Arquitetura Técnica e as Instalações de Produção.....</i>	<i>79</i>
<i>Figura 4.4 – Recursos que Geram Atrasos no Sistema.....</i>	<i>86</i>
<i>Figura 4.5 - Fluxo de Dados para Cadastramento de Clientes.....</i>	<i>87</i>
<i>Figura 4.6 – Porcentagem Atribuída a Cada Recurso do Sistema</i>	<i>92</i>

ÍNDICE DE TABELAS

<i>Tabela 3.1 - Resultados de Planejamento de Capacidade</i>	<i>37</i>
<i>Tabela 4.1 – Variáveis do Ambiente Estudado</i>	<i>72</i>
<i>Tabela 4.2 – Variáveis do Ambiente Analisado</i>	<i>81</i>
<i>Tabela 4.3 – Levantamento de Dados de Carga de Trabalho</i>	<i>84</i>
<i>Tabela 4.4 - Estimativas de Capacidade.....</i>	<i>85</i>
<i>Tabela 4.5 - Resumo das Estimativas de Capacidade.....</i>	<i>85</i>
<i>Tabela 4.6 – Parâmetros de Processamento na Parte Cliente.....</i>	<i>89</i>
<i>Tabela 4.7 - Tempos de Transferência em Redes</i>	<i>90</i>
<i>Tabela 4.8 - Tempo de Processamento no Servidor.....</i>	<i>91</i>
<i>Tabela 4.9 - Tempo Total Gasto em uma Transação de Efetivação de Compra</i>	<i>91</i>
<i>Tabela 5.1 - Alcance das Metas Propostas.....</i>	<i>97</i>

Capítulo 1 - Introdução

1.1 Proposta de trabalho

Existe, na atualidade, um fenômeno bastante interessante e que jamais alguém haveria pensado que pudesse ocorrer. Este fenômeno tem nome: Internet. Uma rede de comunicação entre computadores que possibilita uma infinidade de novos serviços ao usuário, desde o pagamento de uma simples conta bancária até a compra de automóveis.

Segundo TURBAN [et al.] , a Internet vem apresentado um sensacional crescimento desde a sua criação, com o número de Web Sites dobrando praticamente a cada seis meses. A cada um desses Web Sites pode-se associar uma funcionalidade, ou seja, algo que este tenha a apresentar e que seja, de alguma forma, interessante para o usuário. É de fácil entendimento, por exemplo, que sites funcionando como bancos virtuais tenham milhares de acessos todos os dias. Esses sites adicionam muito valor ao proporcionar uma significativa economia de tempo aos usuários, evitando que estes tenham de ir várias vezes a uma agência. Há também outros tipos de sites que podem, por exemplo, facilitar a pesquisa de preços e compra de CDs, ou mesmo fornecer o caminho completo e detalhado até um determinado destino.

Enfim, existe uma gama enorme de diferentes sites, muitos deles concorrendo entre si, em busca de crescimento no número de clientes e principalmente fidelidade daqueles que uma vez se utilizaram de seus serviços. O aumento no volume de clientes de um site é o que todos procuram, porém atrelado a este crescimento, surge a necessidade de se manter o nível de serviço esperado pelos usuários. Um número elevado de solicitações a um mesmo site pode levar a um prejuízo em seu desempenho global, fazendo com que haja a perda de clientes para a concorrência.

Para que se possa manter o desempenho de um site em um nível aceitável para a maioria de seus usuários, é necessário planejar uma estrutura tecnológica que seja capaz de atender à sua demanda de utilização e que ainda mantenha a qualidade de serviço em situações de aumento de demanda.

Para solucionar este tipo de problema podemos lançar mão do Planejamento de Capacidade de Sistemas, uma área de atuação bastante interessante e que requer a concentração de assuntos relacionados a outras áreas de conhecimento, como, por exemplo, PPCP, Previsão de Demanda, Teoria de Filas, entre outros. No decorrer deste trabalho serão utilizadas por diversas vezes idéias decorrentes destas áreas, sempre com um objetivo comum: Efetuar-se uma correta Previsão de Capacidade para sites Internet.

Como objetivos deste trabalho, podemos citar algumas metas a serem atingidas que irão servir de avaliação para sua contribuição prática.

- Introduzir os conceitos sobre e-commerce, incluindo suas diferentes formas de operação, limitações e funcionalidades que sejam relevantes para possíveis investidores neste ramo.
- Apresentar um perfil dos usuários de Internet, buscando verificar qual o potencial de mercado que possui o comércio eletrônico.
- Mostrar quais os conceitos técnicos envolvidos nas operações realizadas via Internet, incluindo arquiteturas Cliente/Servidor, fluxo de uma transação Web e os diferentes recursos envolvidos no sistema.
- Apresentar uma metodologia de planejamento de capacidade, com uma seqüência de etapas a serem seguidas, de modo a cercar todos os parâmetros envolvidos na sua execução. Aplicar esta metodologia a um caso real.

1.2 Motivação

Dado o grande crescimento do número de Web Sites, a entrada de novos elementos em um mesmo ramo de atividade faz com que haja um aumento na concorrência, o que torna bastante importante a qualidade do serviço oferecido por cada um deles. Segundo MENASCÉ, um dos principais fatores

que podem auxiliar um determinado site a atingir os objetivos de qualidade de serviço esperados é a rapidez e a confiabilidade no atendimento às solicitações feitas pelo usuário. Isso faz com que seja criado um sentimento de lealdade com o site, elevando tanto o número, quanto a "qualidade" dos clientes, esta última podendo ser avaliada pelo valor médio de cada compra, por usuário.

Muitas vezes, no planejamento de implantação de um Web Site, não se tem a exata noção da infra-estrutura necessária para que se tenha um performance de atendimento correta, talvez pela falta de conhecimento da demanda, talvez por não se possuir ferramentas adequadas para a definição de uma base tecnológica coerente.

Estes são os principais fatores que nos levam a realizar, neste trabalho, uma análise quantitativa das principais variáveis que têm influência no desempenho de um Web Site. Através de um estudo baseado em fatos reais, e não em meras previsões intuitivas, sem dúvida torna-se mais fácil e com melhores resultados o planejamento da capacidade de um Site.

1.3 Foco

Existem várias possibilidades de se realizar um planejamento de capacidade para sites Internet, dada a variedade de natureza de operação de cada um deles. Existem sites que trabalham como agenda, outros responsáveis pela divulgação de notícias, sites especializados em esportes, enfim, diversos ramos de atuação.

O foco a ser dado neste trabalho diz respeito aos sites especializados em vendas de mercadorias de caráter não complexo, como CDs, livros, etc. Este foco foi escolhido por envolver uma natureza de negócios similar àquela que aprendemos ao longo do curso de engenharia de produção, relacionada a indústrias atuantes neste setor e que se caracterizam pela produção e/ou distribuição de produtos desta mesma natureza.

O planejamento de capacidade aqui analisado pode ser considerado como uma extensão àquele realizado pelos engenheiros atuantes na área de Planejamento e Controle da Produção em uma empresa.

Existem vários pontos que apresentam similaridades no planejamento da capacidade de uma linha de produção, e de um Web Site:

- Avaliar processos
- Levantar tempos de processamento gastos em recursos
- Identificar os gargalos da produção
- Avaliar a capacidade de máquina (recursos) que compõem a linha (sistema)
- Prever a demanda de utilização de um determinado recurso

Estes pontos são fundamentais na escolha de um tema que, além de apresentar bastante identificação com assuntos estudados durante o curso (PPCP, sistemas de informação, entre outros), possua um caráter atual, projetando as ações do engenheiro conforme as tendências do futuro.

Capítulo 2 - E-Commerce

2.1 Definição

Todo o estudo de planejamento de capacidade realizado neste trabalho terá como foco os Web Sites, mais especificamente aqueles especializados em Comércio Eletrônico. Por este motivo, é importante que sejam apresentadas as diferentes tendências apresentadas por este tipo de negócio, e como elas podem ter alguma influência sobre o planejamento da infra-estrutura necessária para um bom desempenho frente aos concorrentes.

Segundo TURBAN [et al.], a Tecnologia de Informação (TI), juntamente com o Comércio Eletrônico, vêm se configurando como o maior facilitador das atividades de negócio do mundo atual, além de serem um grande catalisador para mudanças organizacionais, gerenciais e de operações dentro das empresas. Esse novo modelo de fazer negócios é, sem dúvida, um dos grandes responsáveis pelo aparecimento de novas empresas e pela sobrevivência de muitas outras, que passaram a se dedicar e a gerar receitas a partir desta nova forma de atuação.

Ainda segundo TURBAN, o E-Commerce é, em resumo, um conceito emergente que descreve o processo de compras, vendas ou trocas de produtos, serviços e informações através de redes de computadores, com destaque para a Internet, a principal delas.

Muitas pessoas enxergam o E-Commerce apenas como transações sendo realizadas entre parceiros de negócio, no entanto, o termo mais correto, dada a variedade de conceitos englobados na definição, seria e-Business, este último sendo mais genérico e relacionado à colaboração entre parceiros, à troca de conhecimentos e informações e à busca por uma melhoria de produtividade através da agilidade proporcionada pelas transações online.

O Comércio Eletrônico pode apresentar uma série de diferenciações relacionadas à natureza de cada transação efetuada, podendo ser encaixado nas seguintes categorias:

- **Business to Business (B2B):** Grande parte do Comércio Eletrônico realizado nos dias atuais é desta natureza. Faz referência às transações realizadas entre empresas.
- **Business to Consumer (B2C):** São as operações de varejo realizadas com um comprador individual. A compra de um CD via Internet é uma operação de Comércio Eletrônico do tipo B2C.
- **Consumer to Consumer (C2C):** É tipo de comércio onde apenas pessoas físicas estão envolvidas. A venda de algum produto através de classificados, ou mesmo a apresentação pessoal de uma pessoa através da Internet são exemplos de Comércio Eletrônico C2C.
- **Comércio Eletrônico não Voltado a Negócios:** É realizado por instituições acadêmicas, sem fins lucrativos e mesmo governamentais, com o intuito de melhorar sua atuação. A efetuação de boletins de ocorrência online é um exemplo.
- **Comércio Eletrônico Intraorganizacional:** Nesta categoria estão englobadas as transações que ocorrem internamente a uma organização, desde a venda de produtos a seus funcionários até o treinamento online.

2.2 Histórico

As aplicações envolvendo o Comércio Eletrônico tiveram o seu início por volta de 1970, com a existência do EFT (Electronic Fund Transfers), uma aplicação que estava basicamente limitada às grandes corporações, Instituições Financeiras, e algumas empresas de menor porte. Veio então o EDI, que introduziu transações de outras naturezas que não financeiras e, com isso, fez com que ocorresse a adesão de empresas de outros ramos de atividade, como a manufatura, varejistas, serviços, entre outros.

Muitas outras aplicações seguiram o exemplo do EDI, desde mercado de ações até a reserva de passagens aéreas online. Estas aplicações eram conhecidas como *aplicações de telecomunicações* e tiveram seu valor estratégico bastante reconhecido.

A partir da popularização da Internet no início da década de noventa e do surgimento de um mercado potencial de milhões de pessoas, foi introduzido o termo *Comércio Eletrônico*. Um grande impulso foi dado ao crescimento deste mercado pelo desenvolvimento de novas arquiteturas de redes, protocolos de comunicação, softwares e hardwares que proporcionassem um desempenho aceitável e coerente com o volume de transações que era esperado.

Um outro fator que contribuiu bastante para este desenvolvimento foi o aumento da concorrência nos setores da chamada “Velha Economia”, que fez com que muitas empresas procurassem um diferencial competitivo que lhes garantisse, pelo menos temporariamente, uma sobrevivência no mercado.

Hoje em dia praticamente todas as empresas de médio e grande porte possuem seu Site na Internet. No futuro, este não será mais um diferencial, passando apenas a ser uma exigência sem a qual a empresa não será capaz de se manter ativa.

O comércio eletrônico, apesar de ainda estar em seus estágios de desenvolvimento, apresenta uma característica bem definida: O caráter multidisciplinar e a junção de diferentes áreas de atuação em busca de um objetivo comum. Como exemplo dessas diferentes áreas, podemos citar:

- **Marketing:** O ganho de mercado dos novos sites está totalmente atrelado a uma estratégia de marketing eficiente. Em um mercado onde surgem vários novos sites diariamente, é importante ter um destaque sobre os outros.
- **Ciências da Computação:** Toda a parte de infra-estrutura, e desenvolvimento de sites está baseada em linguagens de programação e softwares relacionados ao campo técnico de conhecimento.

- **Comportamento dos Consumidores e Psicologia:** Para que se tenha sucesso, principalmente nas transações do tipo B2C, é importante que se tenha em mente as exigências do consumidor e o processo de compra que é por eles realizado.
- **Finanças:** Muitas Instituições financeiras são responsáveis pelo surgimento de novos sites, funcionando como parceiras de investimento.
- **Economia:** O comércio eletrônico é bastante influenciado pela economia global e, por outro lado, também possui influência sobre esta.
- **Leis do trabalho e ética:** Os assuntos legais são de extrema importância principalmente num mercado global. Existem, inclusive, temas polêmicos como a ética na Internet e os direitos de propriedade intelectual.

2.3 Relevância do Foco Escolhido

Para que fique mais evidente a importância do tema a ser abordado neste trabalho, além da pertinência de se adquirir conhecimentos a respeito de um assunto que, sem dúvida, terá uma participação decisiva no futuro das pessoas, faremos uma análise dos benefícios e das limitações associadas ao comércio eletrônico, visando destacar quais são os motivos pelos quais uma empresa, ou o consumidor deve optar pelo comércio eletrônico, e quais são os principais pontos em que se deve ter cautela no momento de decisão.

2.3.1 Benefícios do E-Commerce

Poucas vezes vimos um potencial de crescimento tão sólido quanto o apresentado pela chegada do comércio eletrônico. A inovação proporcionada pelo caráter global da tecnologia, pelos custos baixos, horizonte de mercado de milhões de pessoas resulta em um aparecimento de inúmeras novas oportunidades de desenvolvimento para as empresas, para as pessoas e para a sociedade. Muitos destes benefícios estão aparecendo somente a partir de agora, mas sem dúvida eles irão acompanhar o crescimento do comércio eletrônico.

Os itens a seguir mostram os principais benefícios proporcionados pela presença do comércio eletrônico tanto para empresas quanto para os consumidores:

Benefícios para as Organizações:

- O comércio eletrônico possibilita a expansão dos horizontes da empresa, abrindo as portas inclusive para o mercado externo. As empresas podem facilmente encontrar mais consumidores, melhores fornecedores, melhores distribuidoras, etc, a um custo relativamente baixo.
- O comércio eletrônico diminui de forma considerável os gastos relativos a transações anteriormente efetuadas em papel, como criação, processamento, controle, etc.
- Possibilita uma renovação no processamento das empresas atribuindo ao consumidor a responsabilidade de iniciar o processo de produção, através de seus pedidos. Isso diminui o nível de estoques necessários à produção, dada a rapidez das solicitações realizadas online.
- A produção centrada no consumidor também possibilita a realização de customizações no produto final, de forma que este possa ser feito

"sob medida" para cada cliente, o que é, sem dúvida, um grande diferencial competitivo.

- O comércio eletrônico reduz o tempo entre a saída de capital e o recebimento dos produtos.
- Há um aumento de produtividade de máquinas e de pessoal, dada a agilidade proporcionada pela transferência de informações através de redes.
- Existem outros ganhos, como por exemplo, a melhoria na imagem da empresa, encontro de novos parceiros comerciais, acesso mais rápido a informações, que não são facilmente mensurados, porém estão atrelados à aplicação dessa nova tecnologia.

Benefícios para os Consumidores:

- O comércio eletrônico elimina o problema do "Horário Comercial", ou seja, os consumidores podem fazer compras a qualquer hora do dia, inclusive aos sábados, domingos e feriados, além de poderem fazê-las de qualquer computador.
- Abre uma enorme variedade de novas possibilidades de compra, além de facilitar de maneira nunca antes vista a comparação entre diferentes lojas virtuais na busca por um mesmo produto.
- Dado o menor custo que os comerciantes virtuais têm em manter seus produtos, abre-se ao consumidor a possibilidade de encontrar produtos e serviços a um preço mais baixo.
- O consumidor pode obter, em segundos, muito mais informações sobre um produto do que conseguia anteriormente em semanas.

2.3.2 Limitações do E-Commerce

As limitações do comércio eletrônico podem ser divididas em limitações técnicas e limitações não técnicas.

Limitações Técnicas:

- Por estar em seus estágios de desenvolvimento, ainda existem incompatibilidades entre protocolos de comunicação, além de uma certa falta de segurança e confiabilidade nas transações realizadas online.
- Ainda não há largura de banda suficiente para um atendimento ideal das transações realizadas. Esse tema será melhor discutido mais adiante neste trabalho.
- Ainda não há uma padronização com relação aos aplicativos, bancos de dados e outros recursos necessários à realização das transações.

Limitações Não Técnicas:

Além das limitações de ordem técnica, existem ainda outros tipos de limitações que têm impacto no crescimento do comércio eletrônico, e na familiarização de seu uso por parte de uma grande parcela da população. Muitos destes fatores também afetam a adesão de novas empresas a este novo tipo de vendas.

Os principais fatores de ordem não-técnica são:

- **Justificativa de Custos:** O custo de se desenvolver sites que tenham bom desempenho em transações online é relativamente elevado e é também alto o risco de se incorrer em sites não eficientes, principalmente devido à falta de experiências anteriores no

desenvolvimento de sistemas. Além disso, muitos dos benefícios provenientes da implantação de um site de comércio eletrônico são intangíveis, como a melhoria da qualidade de serviço e da imagem da empresa, o que acaba prejudicando a quantificação dos ganhos com este tipo de investimento.

- **Segurança:** Este item é de extrema importância, principalmente na área de comércio eletrônico do tipo B2C, onde os consumidores, por não saber exatamente qual é o método de segurança utilizado, acabam por não efetuar as compras, receosos quanto ao destino de suas informações.
- **Falta de confiança:** Os consumidores, por estarem acostumados a realizar compras tendo um vendedor, um recibo, enfim, toda uma burocracia por trás de cada transação, ainda não confiam nas novas maneiras de se realizar compras online, cujo processo além de ser mais eficiente é menos burocratizado. Em resumo, trocar lojas “físicas” por lojas virtuais ainda gera desconfiança nos consumidores.

Há ainda uma série de outros fatores que limitam o raio de atuação das empresas virtuais e que prejudicam o seu desenvolvimento, porém, como visto acima ao se traçar uma comparação entre os benefícios provenientes da implantação de um site eficiente e as limitações que ainda existem nesse ramo, observamos que é de grande interesse das empresas que se implemente algo que tenha relação com a venda de produtos via Internet, principalmente para aquelas empresas que tem em seu plano estratégico o ganho de mercado e a conquista de consumidores que anteriormente estavam fora de sua área de atuação.

2.4 Estudo do Mercado Atual de Empresas Virtuais

Ao elaborarmos uma pesquisa de mercado relacionada ao Comércio Eletrônico, devem ser levantadas as principais variáveis que tenham uma influência direta no relacionamento entre consumidores, produtos e campanhas de marketing com o objetivo de se entender melhor o processo de compra dos usuários, estabelecer um plano de marketing correto e avaliar o desempenho da estratégia de marketing adotada para ganho de mercado. A pesquisa de mercado deve incluir tópicos como a situação da economia, ramo de atividade, empresas, produtos oferecidos, valor e preço destes, distribuição de mercadorias e comportamento de compra dos consumidores.

2.5 Estudo Atual do Perfil de Usuários de Internet no Brasil

Faremos agora um estudo sobre o perfil de usuários de Internet no Brasil. Essa pesquisa foi realizada pelo *IBOPE*, um renomado instituto de pesquisas, em conjunto com o site *CADE?*, que forneceu todo o apoio tecnológico à sua execução.

As respostas, em número aproximado de 25 mil, foram dadas segundo o preenchimento de um formulário acessado através de um *banner*, posicionado no site *CADE?*.

Os resultados da pesquisa serão analisados segundo os seguintes tópicos:

- Perfil de usuários, incluindo sexo, principal faixa etária, escolaridade, renda, habilidade para falar outra língua e ocupação.
- Hábitos dos usuários, como por exemplo local onde realiza a conexão, quantidade de conexões por dia, principal atividade desempenhada na rede (e-mail, chat, compras, etc.).
- Consumo realizado, ou seja, se já compraram pela rede, ou não, valores gastos, frequência de compra, etc.

2.5.1 Perfil de usuários

Os internautas típicos são homens, mas a participação feminina vem se fortalecendo na Rede. Em novembro de 1996, elas representavam 17% e, em agosto de 1998, são 29%.

Embora a Rede seja acessada por todas as faixas etárias, há uma expansão em torno dos internautas entre 15 e 29 anos.

Elaborado pelo Autor

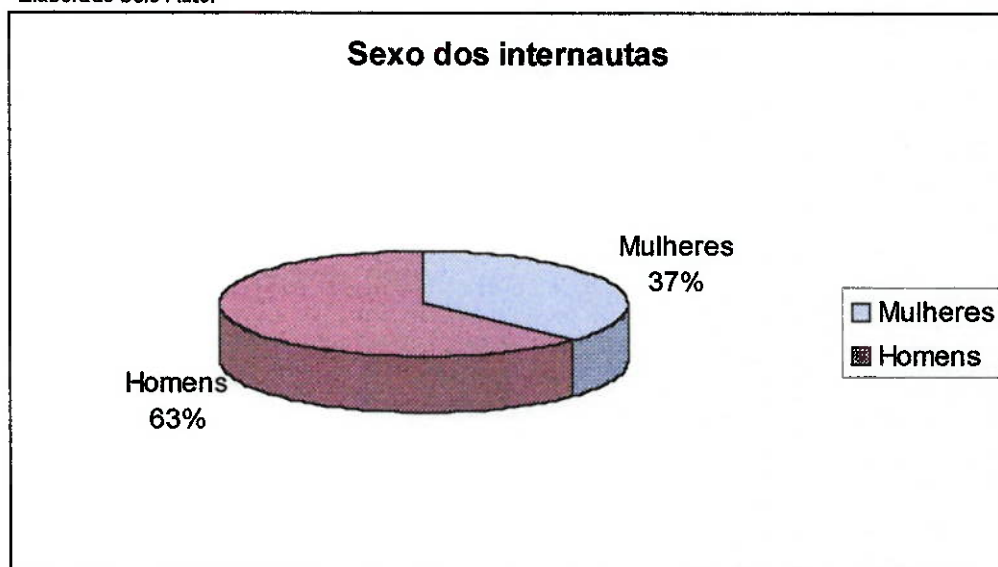


Figura 2.1 – Sexo dos Internautas

A maioria dos internautas é solteira (75%)

A Internet tem perfil qualificado, sendo acessada por 18% de executivos e empresários. Mais da metade dos pesquisados têm renda familiar acima de 20 salários mínimos, sinalizando o topo da pirâmide social, mas há uma tendência à popularização. O principal grupo de internautas possui secundário completo. A língua inglesa é falada por 55%, especialmente pelos mais jovens. A maioria dos usuários estuda, ou trabalha e estuda simultaneamente.

Elaborado pelo Autor

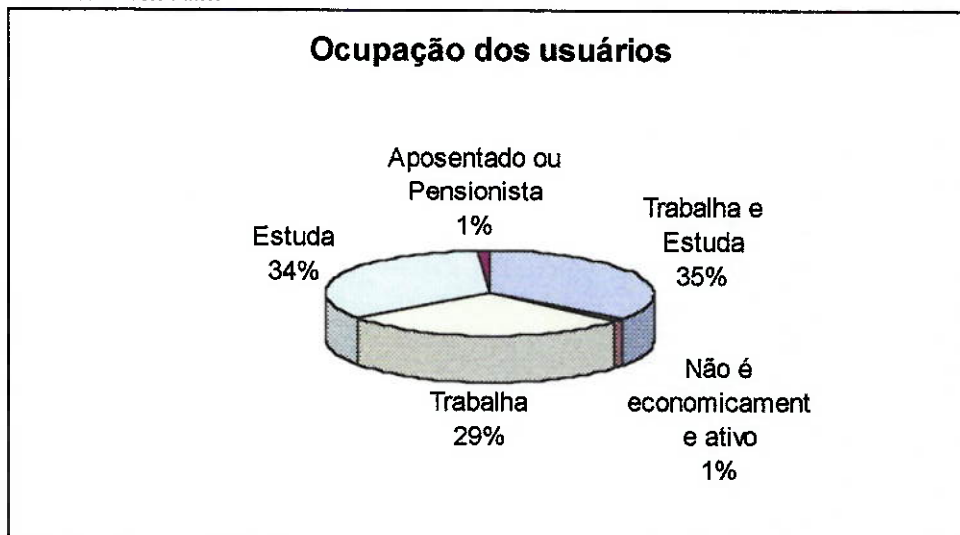


Figura 2.2 – Ocupação dos Usuários

Ressaltando o perfil elitista que está relacionado a usuários de Internet atualmente, a posição ocupada por aqueles que trabalham, possui uma parcela significativa de gerentes e diretores de empresas.

Elaborado pelo Autor

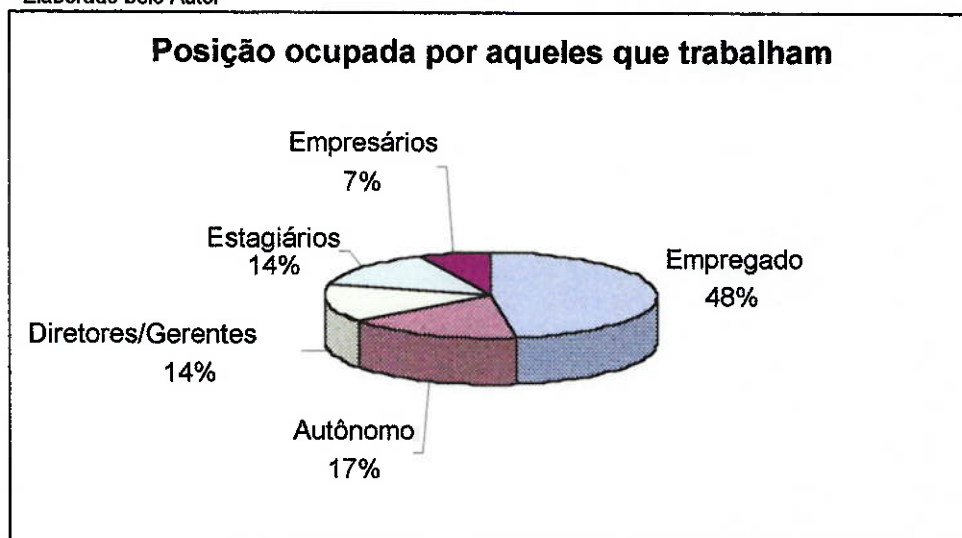


Figura 2.3 – Posição Hierárquica dos Usuários

Com relação ao tipo de profissão desempenhada pelos usuários, podemos dizer que quase um quinto dos internautas trabalha com informática, mas o crescimento de usuários com esse perfil ocupacional vem se reduzindo,

dando espaço para todas as categorias profissionais, especialmente as da área administrativa. A distribuição por ramo de atividade seria segundo o gráfico a seguir:

Extraído de IBOPE INTERACT, 1999



Figura 2.4 – Ramo de Atuação dos Usuários

Em 98, o principal grupo de internautas tem o secundário completo (42%).

Com o ingresso mais intenso e conseqüente expansão do grupo dos jovens entre 20 e 29 anos, diminuiu a proporção de internautas com o superior completo. Ressalta-se, porém, que entre os adultos com 30 anos ou mais, indivíduos com superior completo ou pós-graduados representam mais de 60%.

Uma nova informação é que o grupo de internautas pós-graduado é relevante, tanto para o total (representam 9%), quanto entre os mais velhos. São pós-graduados 18% dos internautas entre 30 e 39 anos, e 22% daqueles com 40 anos ou mais. Colocados em gráfico, teríamos a seguinte distribuição:

Elaborado pelo Autor

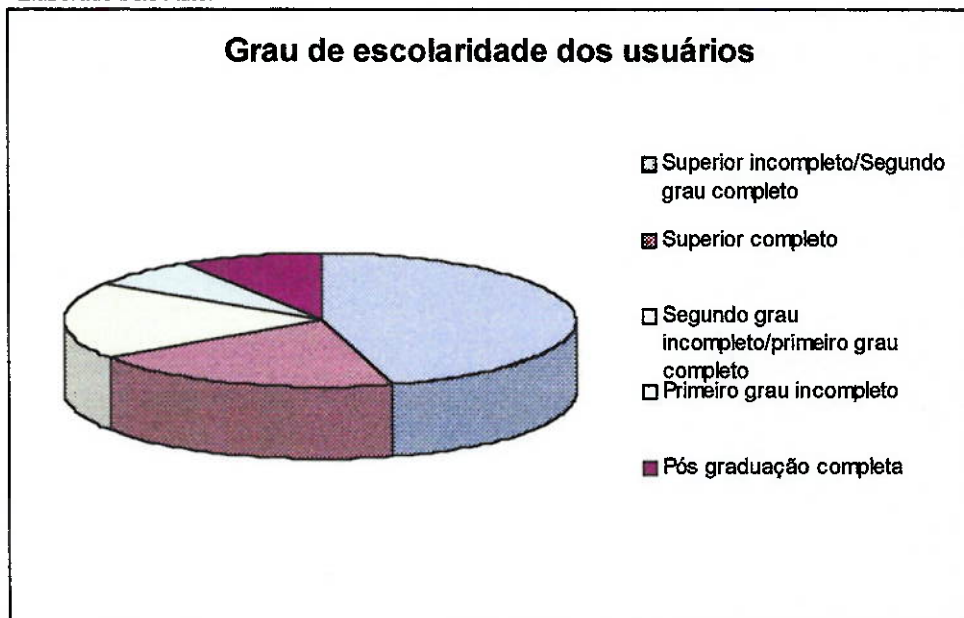


Figura 2.5 – Grau de Escolaridade dos Usuários

A qualificação dos internautas evidencia-se em seu rendimento familiar: mais da metade dos pesquisados têm renda familiar acima de 20 salários mínimos. Em relação às posses e comportamento, 61% deles fizeram viagem aérea nos últimos dois anos e 75% têm cartão de crédito.

Os internautas brasileiros estão no topo da pirâmide social, mas nota-se uma tendência à popularização pelo aumento da participação das camadas de menor renda.

Elaborado pelo Autor

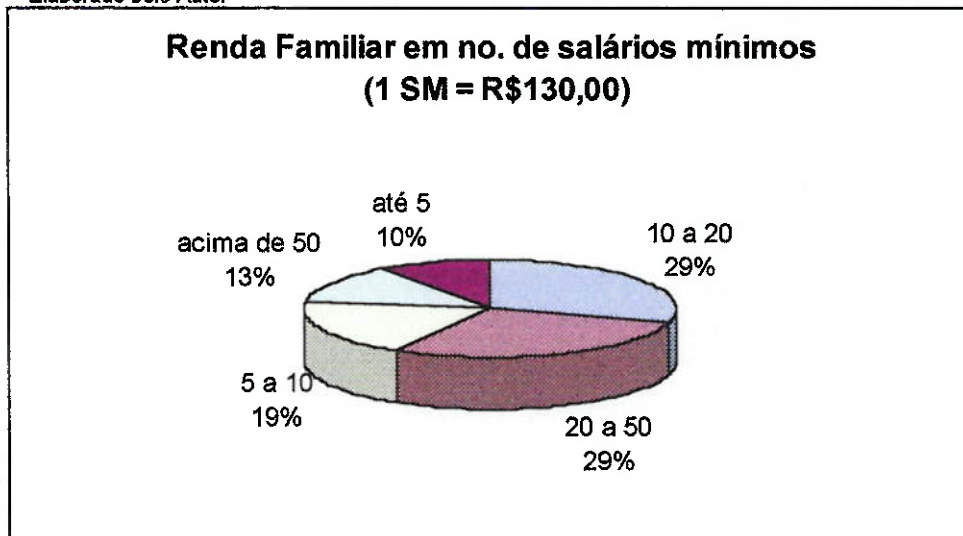


Figura 2.6 – Renda Familiar

Com relação às compras efetuadas pela Internet, é muito importante que se saiba a porcentagem de usuários que possui cartão de crédito, visto que este é o principal meio de pagamento de valores pela Internet. Dentre os usuários de Internet brasileiros 67,6% possuem cartão de crédito, e 32,4% não possuem.

Com relação à localização dos usuários de Internet no Brasil, informação importante quando se pensa em rede de entregas, logística de fornecedores, ou mesmo localização de possíveis armazéns, temos os seguintes dados:

Elaborado pelo Autor



Figura 2.7 – Usuários por Região do País

Por último, temos os dados sobre o estado civil do internauta, dados que revelam que grande parte deles ainda é solteira, podendo, teoricamente, assumir mais riscos com relação a valores.

Elaborado pelo Autor

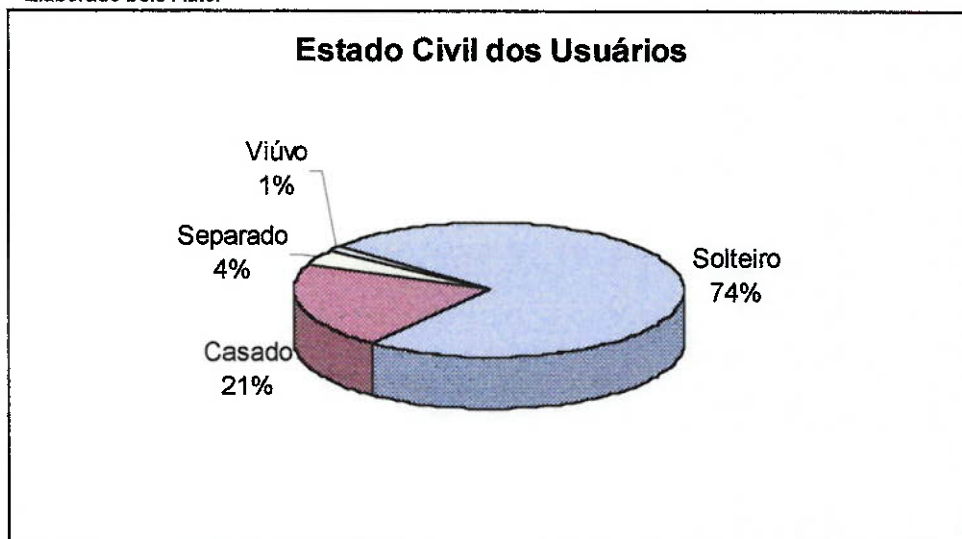


Figura 2.8 – Estado Civil dos Usuários

Análise distanciada revela a fisionomia da comunidade :

- concentrada no Sudeste (57% é de SP, RJ ou MG);
- forte traço masculino (os homens correspondem a 63% dos respondentes);

- no auge de seu vigor e juventude (15 a 29 anos é o principal grupo etário: 68%);
- livre de compromissos sérios (79% solteiros ou separados);
- voltada à sua própria formação (69% estuda, 55% fala inglês);
- e sustento (64% trabalha);
- qualificada para o consumo (renda mensal familiar situada entre 10 e 50 salários mínimos para 59%).

2.5.2 Hábito dos usuários

A relação dos usuários com a Internet, apesar de ainda ser bastante recente (56% está conectada há menos de 2 anos), já é muito intensa e faz parte de seu dia a dia :

Com relação ao local de conexão, vemos que a maioria faz sua conexão através de um computador doméstico, sendo este, em quase metade dos votos pesquisados, compartilhado por uma ou duas pessoas.

Além disso, o número de acessos diários varia conforme os dados do gráfico a seguir:

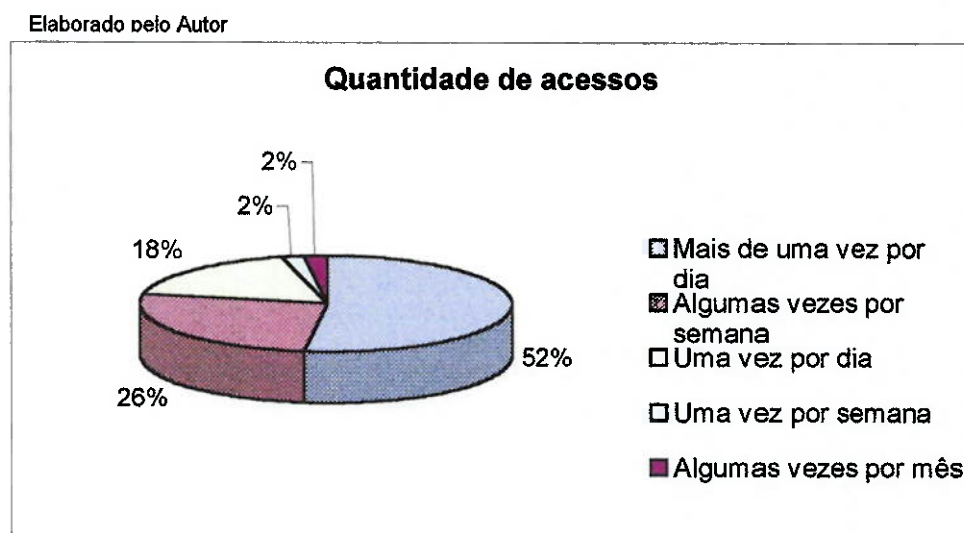


Figura 2.9 – Quantidade de Acessos

Esses dados nos trazem a informação de que grande parte dos usuários (cerca de 70%) acessa a Internet pelo menos uma vez ao dia, o que nos mostra grande potencial de consumo através da rede.

Quanto às principais atividades desempenhadas por usuário enquanto utiliza a Internet, temos os seguintes dados estatísticos:

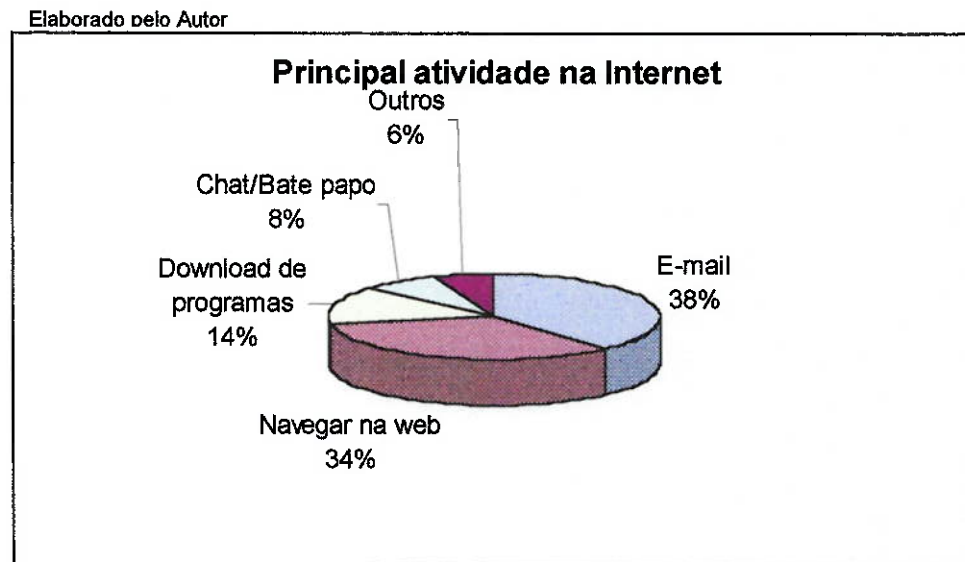


Figura 2.10 – Principal Atividade na Internet

- As atividades mais importantes são o e-mail (38%) e a navegação (34%), dirigida sobretudo à busca por informações sobre produtos ou serviços.
- É extensa a adesão a facilidades de administração do dia a dia : 59% fez entrega online da declaração de IR, e 50% utiliza regularmente o home banking.

Esses dados nos mostram que grande parte dos usuários utiliza a Internet não apenas como instrumento de diversão e entretenimento, mas como ferramenta de comunicação e facilitação do dia a dia.

Além disso podemos destacar o caráter endógeno da população de usuários, de acordo com os seguintes dados:

- Os principais assuntos de interesse de quem navega são: a própria Web, sendo que 83% dos votos consideraram esse assunto muito interessante, seguida da informática, com 71% de consideração.
- a maioria dos novos endereços de site é encontrada na própria rede (70%);

A Internet trouxe consigo muitas alterações nos hábitos diários das pessoas, algumas deixaram de ler revistas, outras não assistem mais televisão, algumas até mesmo perdem tempo de sono, em decorrência de seus acessos à rede. Alguns dados sobre estas mudanças são apresentados a seguir:

Elaborado pelo Autor

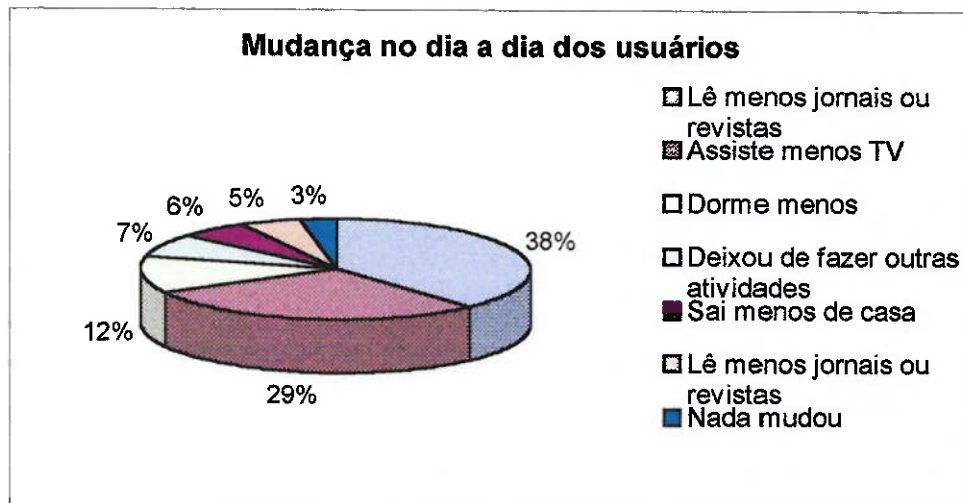


Figura 2.11 – Usuários por Região do País

2.5.3 Consumo via Internet

O Comércio Eletrônico vem ganhando uma fatia importante do mercado consumidor brasileiro, principalmente devido à sua facilidade e rapidez de compra, além de evitar os transtornos decorrentes da locomoção até as lojas. Um dos grandes problemas enfrentados atualmente pelos sites de Comércio Eletrônico, é a rapidez e eficiência na entrega dos produtos solicitados.

Muitas vezes esses prazos dependem de uma gama enorme de fatores, que vão desde uma correta localização dos armazéns e depósitos de produtos, até uma correta logística das rotas de entrega.

Os principais produtos comercializados via Internet nos dias atuais são aqueles chamados de não complexos, incluídos aí os CDs, livros, softwares, entre outros. Os dados apresentados a seguir nos mostram que, apesar de ainda reticentes com relação às compras via Internet, existe um grande potencial de crescimento desse setor, principalmente devido ao interesse mostrado por grande parte dos pesquisados em futuras compras.

No âmbito geral, os números relativos àqueles que comprariam, ou já compraram via Internet, são os seguintes:

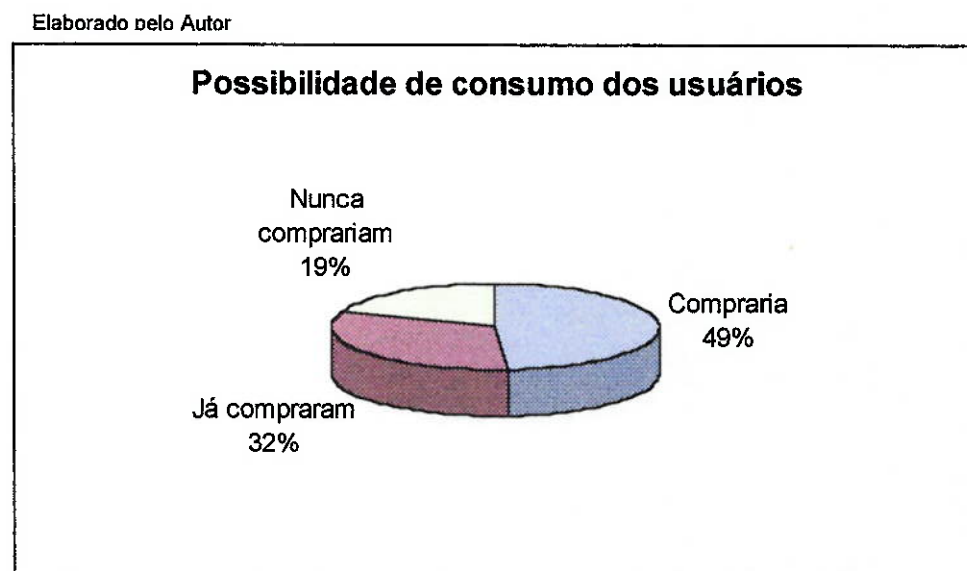


Figura 2.12 – Possibilidade de Consumo via Web

Comprar na Web é também um assunto que ainda desperta muita rejeição entre os internautas (19% declarou que nunca compraria), aparentemente pela complexidade e risco da escolha exclusivamente virtual de produtos desconhecidos, ou mesmo pela insegurança causada pelo fornecimento do número do cartão de crédito na Rede.

Com relação aos principais produtos comprados, como dito anteriormente, temos os seguintes números:

- Dos 32% que já fizeram alguma compra pela rede, 20% comprou CDs, 18% livros e 16% softwares, comprovando a tendência à venda de produtos não complexos.

Vemos também que a frequência de compra online está ligada à introdução ao hábito: 54% comprou de 2 a 5 vezes em 1999, além do volume negociado via rede também apresentar um aumento conforme os dados:

- 44% gastou até R\$ 50,00 a cada compra, 45% gastou entre R\$51,00 e R\$200,00.

Apesar de apresentar crescimento no ramo de consumo, um dos grandes atrativos da Internet ainda é a facilidade de solução de problemas que antes tomavam muito tempo do usuário, como por exemplo utilização de home banking, e entrega da declaração de imposto de renda. Mais da metade dos usuários pesquisados já utiliza home banking, e quase 60% deles utilizou a Internet para entrega da declaração.

Capítulo 3 - Performance em Sistemas

3.1 Redes

As redes de computadores, como são conhecidas atualmente, têm sua origem ligada a um projeto desenvolvido na década de 60 pela Agência de Projetos em Pesquisa Avançada americana, chamada à época de ARPA, pertencente ao Departamento de Defesa Norte Americano. O projeto tinha o nome de ARPANET e consistia em uma troca de informações sobre projetos em andamento. A principal contribuição vinda deste projeto foi a comprovação da utilidade das redes como forma de comunicação entre pessoas, num princípio do que conhecemos hoje por e-mail.

O conceito de Internet foi introduzido em 1983 quando a ARPANET foi dividida em duas redes diferentes: a MILNET, uma rede utilizada para fins militares e uma versão reduzida da ARPANET.

Como veremos mais a frente, a Internet dos dias atuais nada mais é do que um conjunto de redes interligadas através de um protocolo de comunicação e que viveu um processo de crescimento que partiu de 10 pontos no início da ARPANET até aproximadamente 100 milhões de pontos em menos de trinta anos.

3.1.1 Tipos de rede

De acordo com a área geográfica coberta pelas redes, estas podem ser conhecidas como Redes de Longa Distância (chamadas, no inglês, de Wide Area Network ou WANs) ou Redes Locais (Local Area Network ou LANs).

Redes de Longa Distância

As redes de longa distância podem interligar cidades, países ou mesmo continentes. A tecnologia básica utilizada pelas Redes de Longa Distância é chamada de Packet Switching, ou seja, as informações trocadas pelos computadores, chamados também de *hosts*, são quebradas em partes chamadas de *packets* (pacotes), que possuem algumas características como:

- Tamanho máximo
- Cabeçalho, fornecido pelo computador fonte, que possui duas funções:

- Carregar as informações necessárias para que o packet seja corretamente encaminhado da fonte até o destino.
- Manter a ordem lógica necessária à reconstrução do arquivo no computador de destino.

As redes de longa distância consistem basicamente de Packet Switches interligados por links de alta velocidade. Packet Switches são computadores responsáveis por atribuir o caminho correto para os pacotes por eles recebidos. As informações de roteamento são encontradas no cabeçalho dos pacotes e são cruzadas com tabelas de caminho utilizadas para a tomada de decisão. Esse método de roteamento é chamado de *Store and Forward*.

Redes Locais

As redes locais estão limitadas a uma região geográfica muito menor que as redes de longa distância, normalmente um prédio ou um conjunto de prédios interligados. Existem vários tipos de tecnologias utilizadas em redes locais, dentre as quais podemos destacar:

- Ethernet com conexões de 10Mbps ou 100Mbps
- Token Ring de 4 ou 16 Mbps
- Fiber Distributed Data Interconnect (FDDI) de 100Mbps

A tecnologia Ethernet é constituída pela *Topologia de Ônibus (MENASCÉ)*, isso significa que todos os computadores estão interligados a um mesmo cabo coaxial compartilhado. Quando uma informação é transmitida por um dos componentes da rede, ela pode ser recebida por qualquer outro computador interligado, porém como os pacotes contêm o endereço do destino, somente este último terá permissão para arquivar esta informação. Este tipo de comunicação é conhecido como *broadcast communication*. (Fig. 3.1a).

O segundo tipo de rede local, conhecido como *Token Ring*, é constituído, como o próprio nome diz, por um anel de computadores. O computador emissor de informações coloca o pacote na rede e este viaja até o computador de destino, especificado no endereçamento do pacote enviado.

O controle de envio de informações é feito por um dispositivo denominado *Token* cuja posse é necessária para que se possa colocar informações na rede. Depois de enviada a informação, o *Token* é passado ao próximo computador da rede que, se possuir algo a ser enviado, exerce seu direito de uso, senão, simplesmente passa ao próximo componente da rede.

Com o crescimento do número de computadores conectados às redes do tipo *Token Ring*, o tempo de espera pela autorização de envio, bem como a espera pelo envio de informações de outras máquinas passa a comprometer a performance da rede, pois o *Token* deverá percorrer uma distância maior para atingir seu destino.

Outro tipo de rede local é o FDDI, que utiliza cabos de fibra ótica na sua constituição e que, por possibilitar o fluxo de informações em ambas as direções (Fig 3.1c), aumenta a confiabilidade da rede, já que, dada a ocorrência de problema em uma das estações de trabalho, o hardware reconfigura o sistema para uma rede de anel único, o que não compromete o desempenho global da rede.

Extraído de MENASCÉ, 1998

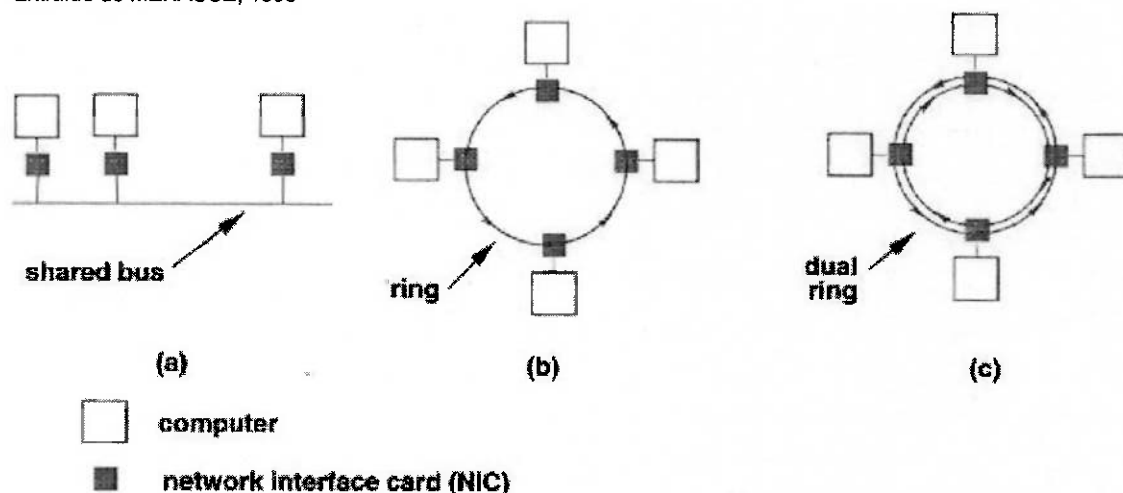


Figura 3.1 (a) LAN Bus-based (b) Ring-Based LAN (c) Dual Ring-Based LAN

A conexão de redes locais a redes de longa distância

As LANs normalmente se conectam às WANs através de conexões de alta velocidade denominadas T1 (1.544Mbps) ou T3 (45 Mbps).

3.1.2 Protocolos de comunicação

Protocolos de comunicação possuem um conjunto de funções que, em conjunto, tornam possível a comunicação e a transferência de informações de um computador até outro. Segundo MENASCÉ, as principais funções de um protocolo de comunicação são:

- Atribuir endereço e rotear pacotes: Garante a entrega dos pacotes de informação ao destino correto.
- Verificar erros, corrigi-los, e sequenciar pacotes: Assegurar que problemas decorrentes de más conexões, pacotes corrompidos ou até mesmo a perda de informação sejam corretamente solucionados.
- Controlar o fluxo de informações: Se o computador emissor envia pacotes a uma velocidade maior do que o computador receptor pode receber, o protocolo de comunicação é responsável pela manutenção de um fluxo compatível de dados.

A conexão entre dois computadores, ditos A e B, pode ser feita basicamente por duas maneiras diferentes: *Connectionless* e *Connection-Oriented*.

A conexão do tipo *connectionless* pode ser entendida como a troca de informações entre dois pontos onde o tamanho máximo da informação se encaixa do tamanho máximo da unidade de dados proporcionada pelo protocolo. Isso significa que a informação não precisa ser quebrada em várias partes distintas, tornando, portanto, o sequenciamento de informações uma tarefa desnecessária.

Já no tipo *connection-oriented*, o tamanho da informação a ser transmitida é bem maior que a unidade de dados permitida pelo protocolo de comunicação, solicitando a quebra em vários pacotes diferentes, o que torna essencial o ordenamento dessas informações.

Veremos, então, os dois principais tipos de protocolo de comunicação utilizados atualmente na Internet:

O **IP**, ou Internet Protocol, é o responsável pelo roteamento de informações através da rede, ou seja, devido ao fato de cada um computadores conectados à rede possuir um *endereço IP*, uma espécie de identificação única na rede, o protocolo de comunicação fica a cargo de atribuir o caminho correto a cada um dos pacotes de informação que passam através dos *roteadores*, computadores responsáveis pelo direcionamento de arquivos.

O protocolo de comunicação IP, não é responsável pelo sequenciamento de informações, ou seja, cada pacote pode seguir por um caminho diferente na rede, até chegar a um mesmo destino, fora de ordem. O ordenamento dessas informações fica a cargo do protocolo que será visto a seguir, o TCP.

O **TCP** é o responsável pelo ordenamento das informações enviadas de um computador a outro e também pela reconstrução de eventuais erros ocorridos durante a transferência de informações.

Juntos, esses dois protocolos de comunicação formam o **TCP/IP** que é, sem dúvida o principal protocolo de comunicação utilizado atualmente na Internet. Existem outros protocolos de comunicação que estão baseados no TCP/IP, conforme visto na figura abaixo.

Um dos mais importantes, e que será visto em detalhes posteriormente é o HTTP (*Hyper Text Transfer Protocol*).

Extraído de MENASCÉ, 1998

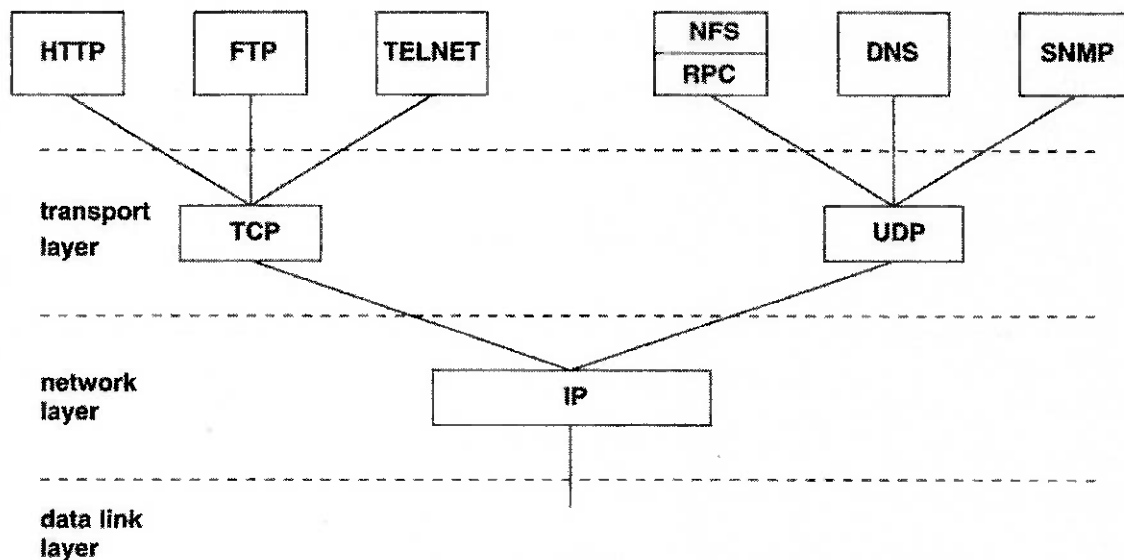


Figura 3.2 - Camadas de Protocolos de Comunicação Baseados em TCP/IP

3.1.3 O Modelo Computacional Cliente/Servidor

A idéia central que motivou a criação do modelo computacional Cliente/Servidor, foi a divisão de processamentos entre essas duas entidades. Cada uma dessas entidades possui as seguintes características:

- **Cliente:** Roda em uma máquina de menor porte, chamada *Desktop*, e geralmente exclusiva para cada usuário, que promove toda a interface gráfica do usuário (GUI) responsável pela captura e apresentação de dados. A parte cliente também fica a cargo de submeter solicitações de processamento para serem executadas em uma ou mais máquinas remotas, além de executar uma parte do código de aplicações.
- **Servidor:** Roda em uma máquina que geralmente possui uma capacidade de processamento bem maior que os Desktops, e com um sistema operacional multiprocessual (UNIX ou Windows NT). Os servidores geralmente não dão início a solicitações, atendo-se apenas a responder àquelas feitas pelo cliente.

Visto que um mesmo servidor pode ser receber solicitações de vários clientes ao mesmo tempo, filas são formadas no servidor de modo a atendê-las em uma seqüência lógica.

Para que o tempo de atendimento a solicitações não fique comprometido, os servidores criam processamentos paralelos, de modo a executar várias solicitações ao mesmo tempo.

Tipos de Servidor

Existem vários tipos de servidores, com diferentes tipo de funções, e cuja aplicação se diferencia na localização de processamento, no cliente, ou no servidor.

Servidores de arquivos são importantes para o compartilhamento de arquivos entre uma rede de computadores, procurando sempre a manutenção de uma versão única de arquivos entre os usuários, como visto na figura 3.3a, a parte gráfica, o aplicativo e a consulta a bases de dados fica na parte cliente, ficando a cargo do servidor apenas o compartilhamento de arquivos.

Servidores de banco de dados são aqueles onde podem ser efetuadas consultas em um ou mais bancos de dados. A principal vantagem deste tipo de servidor é que a consulta é feita no próprio servidor e apenas a resposta à consulta requisitada é enviada ao cliente, o que diminui significativamente o tráfego na rede.

Servidores de aplicativos dão acesso a *procedures* requisitadas pelo cliente, que, por sua vez, geram a lógica de processamento do aplicativo através de uma série de consultas a bancos de dados internos.

Conforme a figura 3.3, da esquerda para a direita o número de solicitações ao servidor diminui, porém a complexidade destas solicitações aumenta.

Extraído de MENASCÉ, 1998

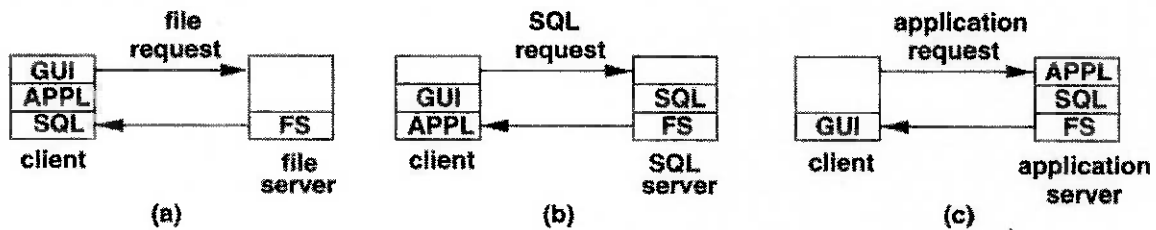


Figura 3.3 – Servidores: (a) Arquivos (b) Consultas (c) Aplicativos

O tipo de servidor que mais nos interessa é o chamado *Web Server* que provê acesso a documentos, imagens, sons e uma série de outros produtos através do chamado protocolo HTTP, e que será visto em maiores detalhes a seguir.

3.2 Quando performance é um problema

Como visto anteriormente, sabemos que a taxa de crescimento apresentada pela Internet é realmente bastante expressiva, com o número de Web Sites praticamente dobrando a cada seis meses.

Muitos dos sites que podem ser acessados hoje em dia apresentam funcionalidades bastante interessantes para os usuários, como por exemplo, a compra de livros, CDs, computadores entre outros, bancos virtuais, que economizam muito tempo do cliente, além de proporcionar um lucro por transação bem maior aos bancos devido a seu menor custo operacional. Segundo GRAY J.[et al.], essas funcionalidades, para que possam ser bem realizadas e trazer o lucro desejado pelos investidores, devem possuir um mínimo de performance aceita pelo usuário, sob pena de perderem a competitividade, e assim grande parte de seu mercado de atuação.

Além disso, a performance de um Web Site torna-se indispensável na manutenção do sucesso destes, principalmente pelo aumento de requisições que ocorre nos estágios de consolidação e ascensão de mercado do site. Sem um correto planejamento, esse aumento no número de solicitações poderia ocasionar uma queda na performance e uma conseqüente fuga de clientes, um sério risco para a sobrevivência num mercado tão competitivo.

3.2.1 Exemplo de Web Site com problemas de capacidade.

Este exemplo foi retirado de MENASCÉ, e reflete algumas consequências provenientes de um Site com problemas operacionais.

Uma empresa vende automóveis pela Internet. Esta empresa possui contratos de parceria com vários revendedores de automóveis, que disponibilizam informações sobre os automóveis a serem comercializados. Essas informações são armazenadas em um banco de dados, que pode sofrer consultas entradas pelos usuários. Os clientes podem, então, efetuar as seguintes requisições:

- Requisitar imagens e documentação dos automóveis.
- Requisitar uma pesquisa no banco de dados da empresa seguindo alguns critérios de busca.
- Entrar com ordens de compra.

Como dito anteriormente, é importante que o desempenho do site seja aceitável, sob pena de ocorrer a perda de vendas. A Empresa pensa em fechar acordos com novos revendedores de automóveis, o que deve trazer um aumento gradual no número de solicitações. Num primeiro momento esse número deve subir em torno de 10%, na segunda fase 20%, e em seguida 30% em relação a atual taxa de solicitações.

A solicitação crítica é a de pesquisa no banco de dados. Foi observado que se o tempo de resposta ficasse entre 4 e 6 segundos, 60% dos usuários cancelariam a solicitação e a venda seria perdida. Se o tempo da resposta excedesse os 6 segundos, então 95% dos usuários cancelariam a solicitação. Considerando que 5% das solicitações geram vendas e que cada venda gera, em média US\$ 18.000, a gerência da empresa quer que sejam respondidas as seguintes perguntas:

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

- O Web Site será capaz de manter o tempo de resposta inferior a 4 segundos, dado o aumento no número de solicitações?
- Se não for capaz, em que ponto terá sua capacidade saturada e por que?
- Quanto, em dinheiro, pode ser perdido, se o Web Site não for capaz de atender às novas solicitações?

Sabe-se que a média diária de procura ao Site, atualmente, é de aproximadamente 92.500 solicitações.

A Tabela 3.1 mostra o quanto pode ser perdido, em decorrência de um site com problemas de capacidade.

Elaborado pelo Autor

	Atual	Atual + 10%	Atual + 20%	Atual + 30%
Solicitações Diárias	92.448,00	101.693,00	110.938,00	120.182,00
Tempo de resposta (Seg)	2,86	3,80	5,67	11,28
Porcentagem de Solicitações Canceladas	-	-	60	95
Vendas Diárias	4.622,00	5.085,00	2.219,00	300,00
Receita Diária (em US\$1.000)	83.203,00	91.254,00	39.938,00	5.408,00
Receita Potencial Diária	83.203,00	91.254,00	99.844,00	108.164,00
Renda Perdida Diariamente	-	-	59.906,00	102.756,00

Tabela 3.1 - Resultados de Planejamento de Capacidade

Ao analisarmos a tabela, vemos que a partir de um aumento de 20% no número de solicitações, passa a haver perda de vendas.

Podemos observar, também, que os valores perdidos devido ao baixo desempenho do site são bastante consideráveis, principalmente quando o aumento previsto for de 30%, o que pode acarretar prejuízos da ordem de US\$100.000,00 por dia.

Este exemplo tem por objetivo mostrar que o Planejamento de Capacidade, além de proporcionar um diferencial competitivo em termos de velocidade de atendimento a solicitações, pode auxiliar no planejamento de expansão de negócios, prevendo futuros aumentos de demanda e verificando quais são as possíveis conseqüências que podem vir de um mau planejamento.

O que veremos mais adiante é como decompor os fatores que influenciam a capacidade de um Web Site, levantar quais são os mais importantes e, analisando-os um a um, verificar o ponto de otimização de capacidade.

3.3 Performance de arquiteturas Cliente/Servidor

Sistemas baseados na arquitetura Cliente/Servidor são compostos de muitos recursos de hardware como Desktops da parte cliente, Servidores com seus discos e processadores, LANs, WANs e Roteadores. Além disso, esses sistemas possuem uma série de softwares de processamento que compartilham a utilização dos recursos de hardware disponíveis. O uso compartilhado desses recursos dá origem àquilo que chamamos de *Filas de Espera*. Uma transação que seja baseada em sistemas Cliente/Servidor pode sofrer uma série de atrasos, que basicamente podem ser decompostos em dois tipos de atraso:

- Atrasos por tempo de serviço: Tempo utilizado para a utilização de recursos tais como discos, redes, etc.
- Atrasos por tempo de espera em fila: Basicamente, o tempo de espera pela disponibilização de um recurso que esteja sendo utilizado por outra transação.

Apresentaremos a seguir o conceito de *Gráficos de Processamento de Comunicação*, utilizados no entendimento do processamento de transações desde o seu envio, pela parte cliente, passando pelo processamento da parte servidor, até a sua recepção final, novamente pela parte cliente.

3.3.1 Gráficos de Processamento de Comunicação

Os gráficos de processamento de comunicação são basicamente compostos por eixos verticais, que representam o tempo, com incremento do topo para a parte inferior.

As linhas sólidas representam o tempo consumido na parte cliente e na parte servidor, enquanto as linhas tracejadas indicam o tempo gasto em redes.

A figura 3.4 representa uma arquitetura Cliente/Servidor de duas partes interligadas por uma rede local.

Extraído de MENASCÉ, 1998

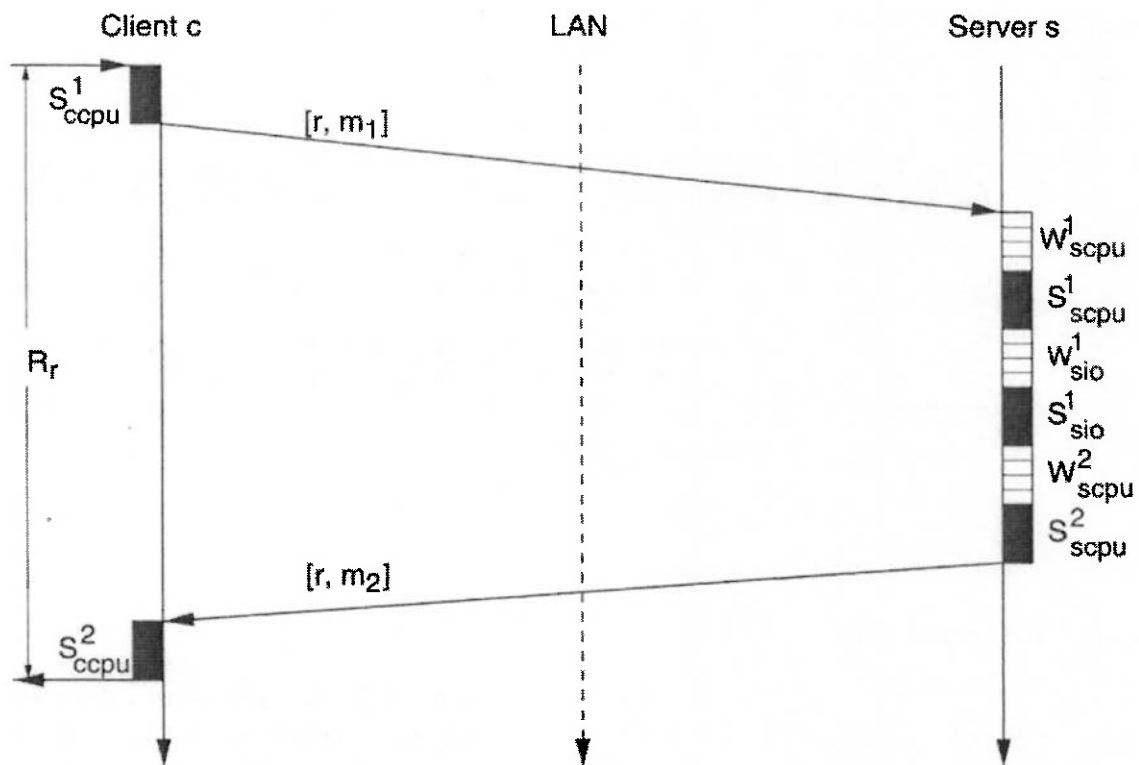


Figura 3.4 - Gráfico de Processamento de Informações

As linhas diagonais indicam as requisições passando, através da LAN, da parte cliente para a parte servidor. Cada uma destas requisições vem identificada por um par do tipo $[id, m]$, sendo id a identificação da solicitação e de sua respectiva resposta, e m o tamanho médio, em bytes, da mensagem

que carrega a solicitação ou sua resposta. Na figura a requisição r possui $m1$ bytes e sua resposta $m2$ bytes.

O tempo de transmissão de requisições do cliente até o servidor é, sendo B a largura da Banda da Rede:

$$T = \frac{m}{B}$$

O tempo total de processamento no servidor pode ser decomposto em dois tipos: *Intervalos de Serviço* e *Intervalos de Espera*.

Intervalos de serviço são representados no Gráfico de Processamento de Informações como retângulos sólidos colocados ao lado dos eixos de tempo e consistem no tempo em que a requisição está sofrendo algum tipo de serviço, seja de processamento ou de disco. A notação utilizada para representar intervalos de serviço obedece aos seguintes padrões:

$$S_i^j$$

Tempo de serviço gasto no recurso i , na $j^{\text{ésima}}$ visita a este recurso. Portanto, na figura 3.4 temos, como exemplo, o tempo de serviço gasto na primeira visita ao CPU do servidor como sendo:

$$S_{SCPU}^1$$

Da mesma forma, o tempo de espera pelo atendimento a requisições que queiram acessar um mesmo recurso (Tempo de Espera na Fila), é representado por:

$$W_i^j$$

O Somatório dos tempos de serviço, por requisição, é chamado *Demanda de Serviço*, e é representado por:

$$D_i^j = \sum_{\text{Visitas } j} S_i^j$$

O Somatório dos tempos de espera, por requisição, é chamado *Tempo de Fila*, e é representado por:

$$Q_i^j = \sum_{\text{Visitas } j} W_i^j$$

O Tempo gasto pelas informações em trânsito, ou seja, em recursos de rede, é representado pela soma dos tamanhos dos arquivos de requisições e respostas ($m1$ e $m2$), divididos pela largura de banda da rede B , ou seja:

$$D_{LAN} = \frac{(m1 + m2)}{B}$$

O Somatório da Demanda de Serviço e dos Tempos de Filas, por requisição, é chamado *Tempo de Residência*, e é representado por:

$$R_i' = D_i + Q_i$$

O tempo de residência é o tempo total gasto por uma requisição, em um determinado recurso.

Finalmente, o tempo gasto por uma requisição R_r , é a soma dos tempos de residência em todos os recursos do sistema. No exemplo da figura 3.4, R_r seria:

$$R_r = R_{CCPU}' + R_{SCPU}' + R_{LAN}'$$

Considerando-se um determinado recurso pertencente a um sistema Cliente/Servidor, que possa ser acessado por muitas vezes por uma mesma requisição antes de enviar uma resposta, o tempo médio de demanda D_i neste recurso é representado por:

$$D_i = V_i * S_i$$

Onde V_i é o número médio de visitas ao recurso i e S_i é o tempo médio de serviço no recurso i .

3.3.2 Componentes de sistemas Cliente/Servidor que podem ocasionar atrasos:

Um sistema baseado na arquitetura Cliente/Servidor, como visto anteriormente, possui uma série de componentes, chamados *Recursos*, que possuem diferentes tipos de operação, executam diferentes tarefas, mas possuem duas similaridades: podem ocasionar atrasos em decorrência do tempo de processamento e em decorrência da formação de filas de espera para processamento.

Segundo MENASCÉ, os principais recursos que podem gerar atrasos dessas duas naturezas são:

- Discos magnéticos, ou seqüência de discos magnéticos.
- Redes
- Roteadores

Veremos, então, quais são os principais pontos que podem gerar atrasos em cada um desses recursos.

Atrasos gerados em discos magnéticos:

Basicamente, os atrasos gerados em serviços nos discos magnéticos são causados pelos deslocamentos mecânicos necessários à localização da informação na superfície do disco.

Os sistemas de troca de informações, chamados de *I/O systems* são formados por vários componentes, além do disco magnético, como pode ser visto na figura 3.5.

Extraído de MENASCÉ, 1998

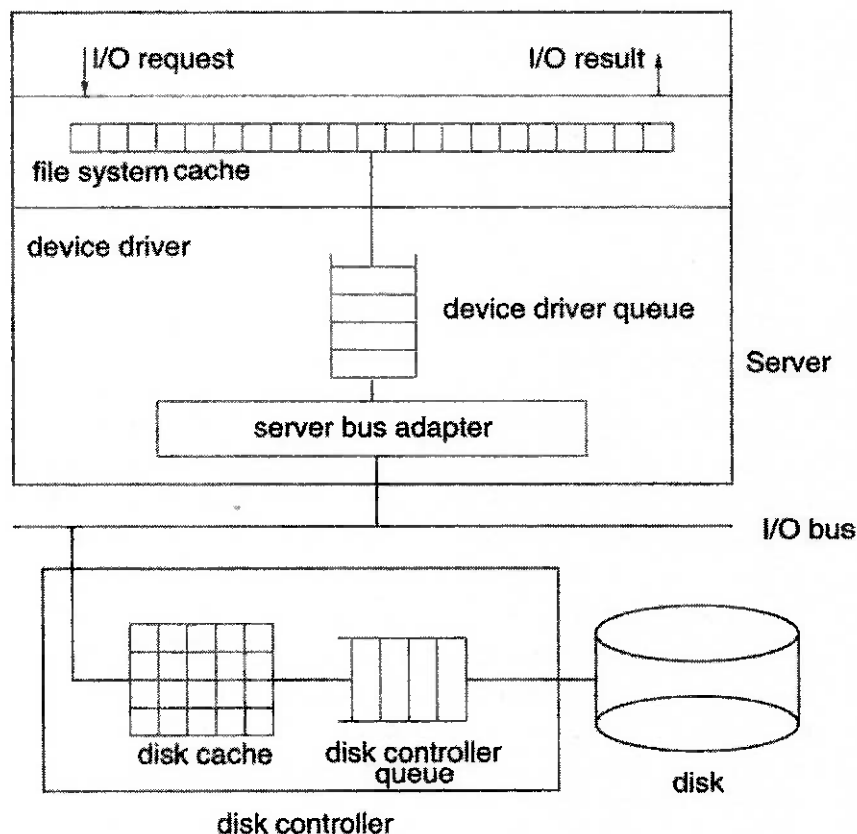


Figura 3.5 - Arquitetura de um Sistema I/O

O fluxo de transferência de informações obedece à seguinte ordem de eventos:

1. Uma solicitação de transferência de informações é passada ao sistema de arquivos

2. Os arquivos mais recentemente utilizados são guardados em uma parte do sistema de acesso rápido chamada *cache*. Se a informação requisitada estiver disponível no cache, não há a necessidade de acesso a disco, o que torna a transação mais rápida.
3. Se não estiver disponível no cache, então a solicitação entra na fila de espera de processamento.
4. É então enviada ao controlador de disco, que gerencia a entrada e saída de solicitações. O controlador pode também prever a futura utilização de algum bloco de informações e, portanto repassá-lo ao cache.
5. Aí, então, a solicitação é atendida através do movimento do braço do disco magnético.
6. As informações são agora enviadas de volta para o sistema de arquivos, que atende a solicitação.

A equação geral que determina o tempo a ser gasto em solicitações a discos magnéticos é a seguinte:

$$S_d = \text{ControllerTime} + P_{\text{perda}} \times (\text{SeekTime} + \text{Latência Rotacional} + \text{Tempo de transferência})$$

Onde os fatores da expressão são:

- **S_d** : Tempo médio de serviço gasto em discos magnéticos
- **ControllerTime**: Tempo gasto pelo controlador para processar uma solicitação de I/O.
- **P_{perda}** : É a probabilidade de uma determinada informação não estar disponível diretamente no cache.
- **SeekTime**: Tempo de posicionamento do braço leitor para a retirada da informação.
- **Latência Rotacional**: Tempo de espera que ocorre entre o momento em que a informação foi encontrada e o início da transferência de informação.

- **Tempo de Transferência:** Tempo despendido na transmissão de um bloco de informações do disco ao controlador de disco.

Atrasos gerados em redes:

O tempo de serviço referente às redes é o tempo gasto para se transmitir as informações através de todas as partes componentes desta rede. Este tempo pode ser melhor entendido como sendo a proporção entre o tamanho do pacote a ser transmitido e a largura de banda da rede.

A título de ilustração, podemos ver na figura 3.6, uma conexão Cliente/Servidor a um servidor de banco de dados através de uma rede composta por uma rede do tipo Ethernet de 10 Mbps, uma conexão de fibra óptica do tipo FDDI de 100Mbps e uma rede local do tipo Token Ring. Para cada uma destas redes, temos uma entidade chamada MTU, ou *Maximum Transmission Unit* que, como visto na figura, é de 1500 bytes para a rede Ethernet, 4472 bytes para o *backbone* FDDI e 4444 bytes para a rede do tipo Token Ring.

Extraído de MENASCÉ, 1998

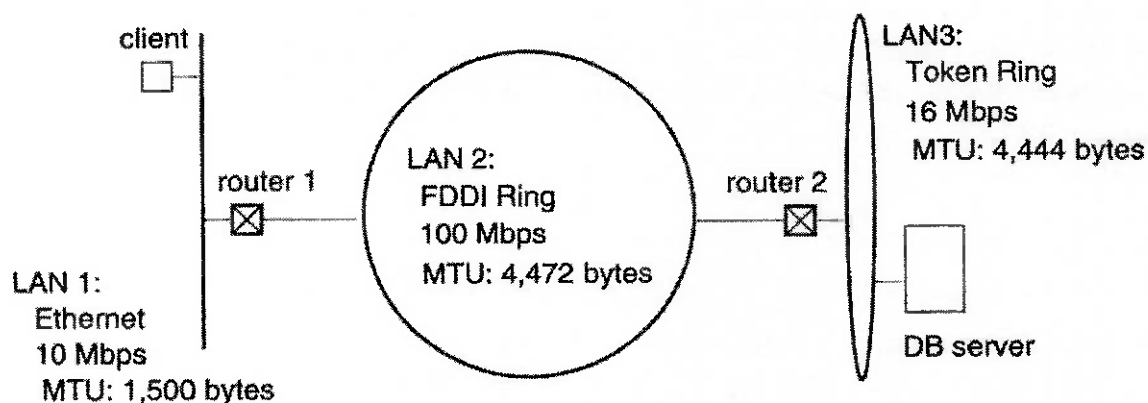


Figura 3.6 - Conectividade Cliente/Servidor

Sendo assim, temos que o tempo de serviço utilizado em redes seria de:

$$S_n = (8 \times \text{Tamanho da mensagem}) / 10^6 \times \text{Largura de banda}$$

Para uma combinação de redes como a mostrada na figura, o tempo total de utilização é:

$$U_n = \sum_{\text{mensagens } j} \lambda_j \times \text{Tempo de Serviço}_n^j$$

onde λ é a taxa de chegada de mensagens do tipo j , e tempo de serviço $_n^j$ é o tempo médio de serviço de uma mensagem do tipo j na rede n .

Atrasos gerados em roteadores:

Um roteador nada mais é do que um processador dedicado à determinação da rota a ser percorrida pelos pacotes de informação (chamados no protocolo IP de datagramas) da origem até seu destino. Estes datagramas são colocados em filas, até que o roteador esteja disponível para, analisando o cabeçalho identificador de cada datagrama e cruzando os dados com tabelas de roteamento, atribuir um direcionamento correto a cada pacote, conforme ilustra a figura 3.7.

Extraído de MENASCÉ, 1998

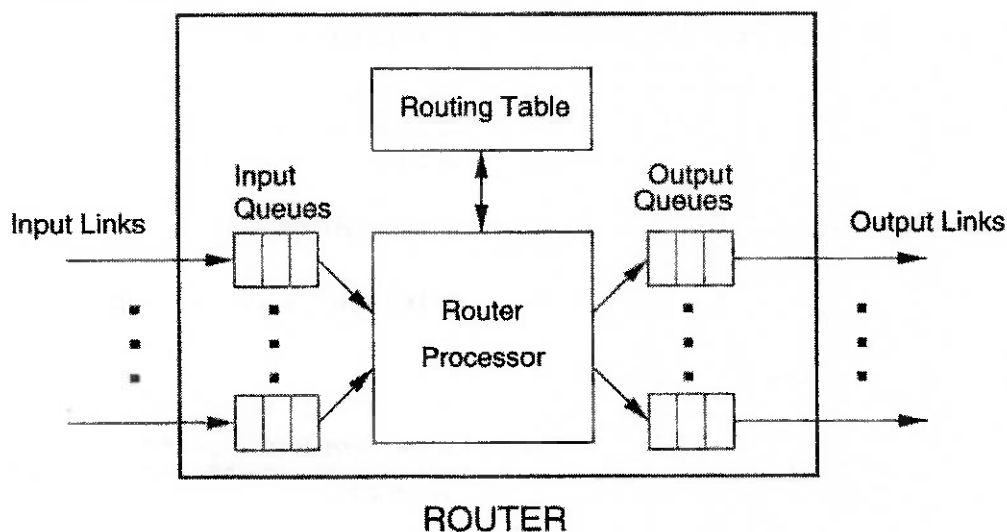


Figura 3.7 - Filas em Roteadores

O tempo de processamento gasto pelo roteador para cada pacote é geralmente fornecido pelos fornecedores e é chamado de *Latência do Roteador*.

Sendo assim, temos que o tempo de serviço total gasto em roteadores pode ser expresso por:

Tempo de serviço do roteador = Ndatagramas x Latência do Roteador

3.3.3 Definição das variáveis de performance que compõem o sistema

Para cada um dos recursos que compõem o sistema Cliente/Servidor, como já dito anteriormente, temos o tempo de espera em filas e o tempo de processamento propriamente dito.

Conforme os diferentes tipos de recursos, podemos ter grandes variações da relação Tempo de espera/Tempo de processamento, como por exemplo nos sistemas em que se têm vários recursos da mesma natureza dedicados a uma mesma tarefa (Processadores).

Essa relação é bastante influenciada pela taxa de chegada de solicitações, visto que à medida que a carga aumenta sobre um mesmo recurso, o número de solicitações que são colocadas em filas também aumenta. Por outro lado, se a demanda não for muito elevada, o tempo de processamento para a ter um peso bem maior na composição do tempo total no sistema.

Algumas das variáveis apresentadas aqui já foram previamente definidas, porém não formalmente. Segundo MENASCÉ, será referenciado como *Fila* a fila de espera propriamente dita mais o recurso a ela atrelado.

Sendo assim, as principais variáveis de performance a serem analisadas neste trabalho serão:

- V_i = Número médio de visitas à fila i no sistema Cliente/Servidor
- S_i = Tempo médio de serviço gasto no recurso i , por visita.

- W_i = Tempo médio de espera, por visita, de uma transação Cliente/Servidor na fila i .
- R_i = Definido como o tempo total de espera na fila i , ou seja, $R_i = S_i + W_i$.
- λ_i = Taxa média de chegada de solicitações à fila i .
- X_i = Taxa média de saída da fila i , ou seja, o número de solicitações completas que saem do recurso i por unidade de tempo. Por hipótese será assumido que o tempo de observação será longo o suficiente para que o número de chegada e de saída de solicitações seja praticamente o mesmo, o que chamaremos de *Hipótese de Equilíbrio de Fluxo*. Isso significa que $\lambda_i = X_i$.
- X_0 = Número de transações completas por unidade de tempo.
- N_i^w = Número médio de transações na fila de espera da fila i .
- N_i^s = Número médio de transações recebendo serviço em um dos recursos da fila i . Se esse recurso for único, então N_i^s é um número entre zero e um que pode ser interpretado como a fração do tempo em que determinado recurso está ocupado, ou seja, a utilização do recurso.
- N_i = Número médio de transações presentes no sistema i (em fila ou recebendo serviço), portanto $N_i = N_i^s + N_i^w$.

3.3.4 Algumas Conclusões Importantes Sobre Performance

A partir dos dados expostos acima, já podem ser tiradas algumas conclusões gerais relativas a performance. Elas são:

Lei da Utilização:

Supondo uma fila de recurso único como a mostrada na figura abaixo:

Elaborado pelo Autor

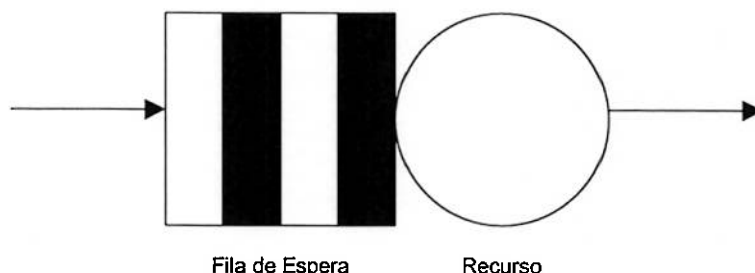


Figura 3.8 - Representação Gráfica para um Recurso e sua Fila de Espera

Sendo a utilização do recurso i definida como a fração de tempo em que este recurso está ocupado, se monitorarmos esse recurso i por T segundos e verificarmos que este recurso esteve ocupado por B segundos, então a sua utilização seria: $U_i = B_i/T$. Podemos assumir que neste mesmo intervalo de T segundos, tivemos C_0 transações completadas a partir desta fila. Isso significa que a saída média da fila é: $X_i = C_0/T$. Se combinarmos estas duas expressões, chegaremos ao seguinte resultado:

$$U_i = \frac{B_i}{T} = \left(\frac{B_i}{C_0}\right) \times X_i = X_i \times S_i$$

Na situação de equilíbrio, ou seja, com o sistema em regime, temos que $X_i = \lambda_i$, portanto a lei da utilização nos diz que a utilização de um determinado recurso é dada pela taxa de chegada de solicitações multiplicada pelo tempo médio gasto para o processamento neste mesmo recurso, ou:

$$U_i = \lambda_i \times S_i$$

3.3.5 Métricas para avaliação de performance em sistemas Cliente/Servidor

Para que se possa ter uma correta avaliação sobre o desempenho de um sistema Cliente/Servidor é necessário se colocar algum método de quantificação em pauta. Para este tipo de sistema, as principais métricas utilizadas são:

- Tempo de Resposta
- Saída Média do Sistema
- Custo

Tempo de resposta

A variável Tempo de Resposta é de extrema importância tanto para os usuários finais, que escolherão os sites que melhor e mais rapidamente atenderem aos seus desejos de compra, quanto para os controladores de sites que tenham como base para a geração de receita o desempenho e o correto atendimento às solicitações de compra.

Para uma análise e uma correta definição do Tempo de Resposta de um sistema, vamos tomar por base a figura 3.9.

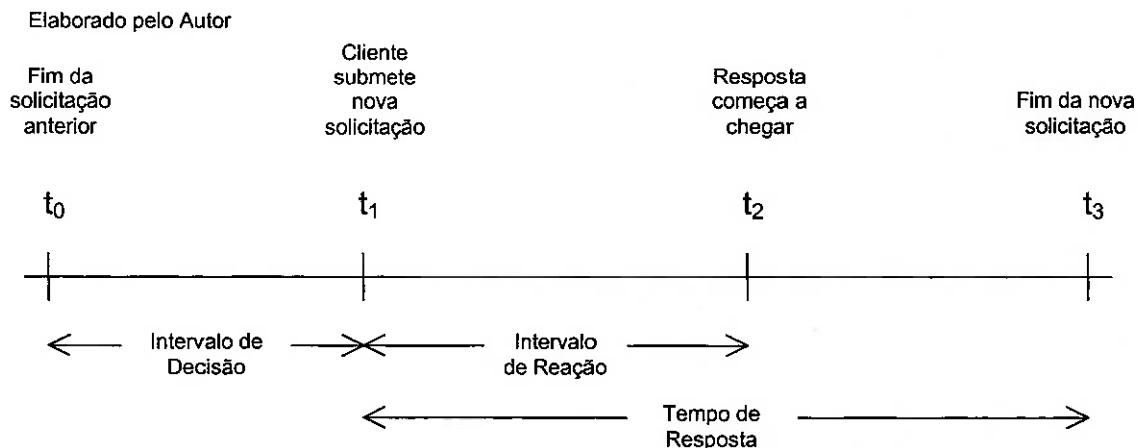


Figura 3.9 - Definição de Intervalo de Decisão, de Reação e Tempo de Resposta.

Supondo um usuário conectado à Internet submetendo solicitações a um servidor Web, temos as seguintes situações:

No momento t_0 a resposta a uma solicitação anteriormente feita pelo usuário acaba de ser completada. No intervalo de tempo t_1-t_0 o usuário está tomando uma decisão sobre qual a próxima solicitação a ser enviada. Este intervalo de tempo é chamado de *intervalo de decisão*.

No momento t_1 , o usuário submete uma solicitação ao servidor que será processada e retornará no instante t_2 . O intervalo de tempo t_2-t_1 é chamado de *Intervalo de reação*, onde o servidor está processando a transação solicitada pelo usuário.

No instante t_3 , a solicitação feita em t_1 estará sendo completada, com todas as informações requisitadas disponíveis para o usuário.

O *Tempo de Resposta* de um Web Server será considerado como o intervalo t_3-t_1 , ou seja, o intervalo compreendido entre o momento em que o usuário entra com a solicitação até o momento em que esta solicitação esteja completamente atendida, com as informações disponíveis para utilização.

Saída Média do Sistema

Se o Tempo de Resposta de um sistema é importante tanto para clientes quanto para a gerência estratégica, a Saída Média de um sistema Cliente/Servidor é uma variável de grande importância para os

Administradores de Sistema, dada a preocupação na manutenção de um sistema que atenda as necessidades dos usuários com eficiência.

A unidade utilizada para a quantificação da saída média de um sistema varia segundo a natureza do Servidor em questão. Para servidores de bancos de dados, por exemplo, o número de transações atendidas por segundo (TPS) é a unidade de medida utilizada. Para servidores de arquivos, a saída é medida através do número de operações de I/O realizadas por segundo.

De acordo com o foco dado a este trabalho, estaremos verificando as transações realizadas em Web Servers. Para este tipo de servidores a saída média do sistema é quantificada em HTTPops/s (operações do tipo http por segundo).

Custo

Ao pensarmos em custo como fator de decisão, devemos sempre associá-lo a algum tipo de métrica de performance, por exemplo, Tempo de Resposta ou Saída Média. Como exemplo podemos ter um custo de hardware e software que gere um determinado número de transações. Neste sistema teremos então um determinado custo por transação efetuada. Sistema que apresentem um custo menor sem perda de performance devem ser priorizados.

3.4 Particularização: Performance de Web Sites

Após a definição das principais variáveis que afetam um sistema Cliente/Servidor, passaremos agora a uma análise mais centrada em redes Internet e Intranet, que nada mais são do que casos particulares de sistemas Cliente/Servidor.

Num rápido questionamento, pode-se verificar que velocidade é o principal problema que afeta os usuários com relação à Internet. De fato, nas atuais conexões feitas à Internet, leva-se um tempo exacerbado para se efetuar consultas, baixar arquivos, enfim, para uma correta interação com os Web Sites.

As redes que compõem a Internet podem ser definidas como a junção padrão de três elementos: Os servidores, que são repositórios de informações, os clientes, que, como dito anteriormente, são os desktops para entrada de solicitações e para recebimento de respostas, e as redes, que fazem o papel de interligação entre os dois primeiros componentes. Um esquema geral dos elementos básicos que compõem uma rede Internet ou Intranet é mostrado a seguir:

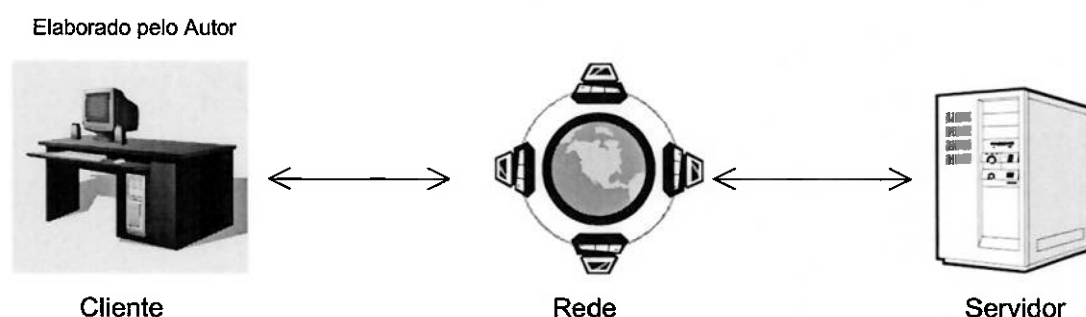


Figura 3.10 - Componentes de Redes Internet ou Intranet

Os servidores são os principais componentes de redes Internet/Intranet. São eles os responsáveis pelo atendimento às solicitações de imagens, texto, gráficos, etc, além de promoverem a correta interação multimídia entre estes elementos. De acordo com a função a que se destina, diferentes tipos de

servidores estarão sendo utilizados. Existem vários tipos de servidores em uso na Internet hoje em dia, dentre os quais podemos citar os Web Sites propriamente ditos, os servidores do tipo proxy, cache e os chamados servidores espelho. Veremos em detalhes os diferentes tipos de servidor um pouco mais adiante.

A outra “ponta” do sistema, a parte Cliente da Internet e das Intranets é baseada em um *Browser* cuja função é interagir com o servidor através de solicitações, e que funciona em conjunto a outras aplicações tais como o e-mail e outros protocolos de transferência de arquivos como o ftp e o telnet. O browser permite que um desktop comunique-se com outros clientes em uma rede TCP/IP e em qualquer servidor na Web que possua um endereço HTTP. Segundo TURBAN, dada a importância atualmente atribuída ao comércio realizado através da Internet, o desempenho dos serviços oferecidos via Internet cada vez mais se transforma em um diferencial competitivo e deve ser tratado como um ponto crítico para o sucesso da empresa. Quanto maior o volume de informação disponibilizado por uma empresa através da rede, maior o número de solicitações a que estará sujeito este site, e quanto maior o número de solicitações de um site, maior a probabilidade de o tempo de espera do usuário se tornar inaceitável. Um grande diferencial que o comércio realizado pela rede tem sobre o comércio de rua é a facilidade que o cliente tem para pesquisar preços e comparar as diferentes opções que tem disponíveis para realizar uma compra.

3.4.1 Web Servers

De maneira simplificada, um servidor web é uma combinação de uma plataforma de hardware, de um sistema operacional, de softwares e aplicações rodando na parte servidor, e de um determinado conteúdo a ser apresentado na rede. Todos estes componentes têm influência no desempenho dos servidores. Serão discutidos, a seguir, os aspectos específicos de cada componente que afetam o comportamento operacional de um Web Server.

A linguagem HTML

Os Documentos que podem ser acessados via Internet são escritos em uma língua chamada Hypertext Markup Language (HTML). O HTML permite a estruturação dos documentos indicando títulos, ênfase, ligações a outros documentos, e assim por diante. Imagens e outros objetos multimedia podem também ser incluídos em documentos HTML. Assim, o HTML permite que se integre vídeo, aplicações de software, e objetos multimedia na Web. A maioria dos documentos em HTML consiste em textos e imagens. Existem também links que indicam quais imagens serão introduzidas em um documento. Essas imagens têm grande impacto no desempenho do servidor. Quando um browser analisa os dados recebidos, reconhece os links associados às imagens e solicita automaticamente as combinações de imagens indicadas por esses links. Na verdade, um documento HTML com imagens é uma combinação de vários objetos, que geram múltiplos pedidos separados ao servidor. Apesar de o usuário ver apenas um documento, o que o servidor interpreta é uma série de diferentes solicitações. A fim de melhorar a percepção do desempenho pelo usuário final, alguns browsers solicitam as imagens antes da chegada do texto completo. Um único click por parte do usuário pode gerar um grande número de solicitações diferentes.

Combinação de Protocolos de Comunicação (TCP/IP e HTTP)

O protocolo de comunicação HTTP opera em conjunto com o protocolo TCP/IP, constituindo o que podemos chamar de *Transações Web*, ou seja, um conjunto de passos a serem seguidos, de modo a que se possa obter algum tipo de informação de um servidor Web.

Cada transação Web consiste em uma solicitação feita pelo cliente ao servidor, seguida de uma resposta dada pelo servidor ao cliente. Cada solicitação HTTP inclui alguns componentes que formam as características desta solicitação: Um *método* que informa qual a ação a ser realizada pelo servidor, uma *URL* (Uniform Resource Locator) que identifica o nome da

informação requisitada, além de outros componentes como o tipo de documento que o cliente quer que seja aceito, etc.

Segundo MENASCÉ, os passos envolvidos em uma transação Web são:

1. Mapear o nome do servidor a um endereço IP (Identidade de cada máquina na rede).
2. Estabelecer uma conexão através do protocolo TCP/IP com o servidor.
3. Transmitir a solicitação (método + URL + outras informações)
4. Receber a resposta
5. Fechar a conexão TCP/IP.

O protocolo HTTP não estende o conceito de sessões abertas de comunicação, além de uma única solicitação completa, ou seja, para cada solicitação que se faz ao servidor uma sessão é aberta e encerrada, no momento em que se recebe a resposta do servidor.

Essa característica do protocolo HTTP traz eficiência às conexões, pois devido ao enorme número de clientes que podem estar conectados ao mesmo tempo em um site, o Web Server não seria capaz de manter um histórico de cada um dos clientes conectados, ou de cada solicitação feita anteriormente.

Para que possa ocorrer uma transferência de informações através de solicitações Cliente/Servidor existem algumas "formalidades" que devem ser realizadas entre as partes com o objetivo de estabelecer as conexões.

A figura 3.11 mostra quais são essas formalidades:

Adaptado de MENASCÉ, 1998

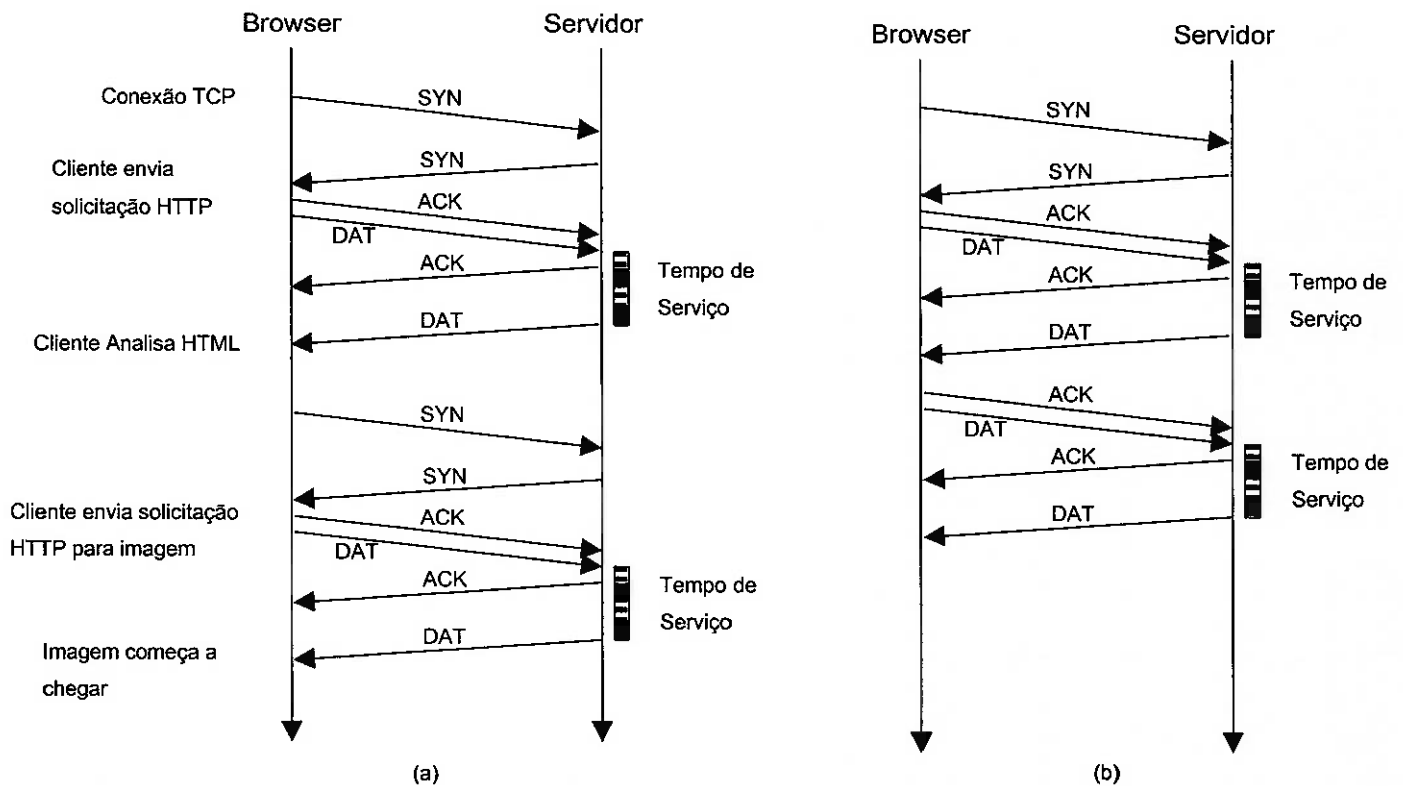


Figura 3.11 - Interação Cliente/Servidor Utilizando o Protocolo HTTP 1.0 e 1.1

Existem atrasos obrigatórios causados pela interação entre os dois protocolos de comunicação e que diferem da primeira versão do protocolo HTTP (chamada HTTP 1.0) para a segunda versão (HTTP 1.1).

Como visto na figura acima, para que seja efetuada uma troca de informações entre cliente e servidor, os seguintes passos devem ser seguidos, esses passos acabam gerando uma série de atrasos obrigatórios, enumerados a seguir:

1. O cliente abre uma conexão TCP com o servidor, que resulta do envio de um pacote de sincronismo (SYN) e um de Acknowledgment (ACK)
2. O cliente envia uma solicitação HTTP para o servidor, que a analisa, executa a ação requisitada e envia de volta o resultado da solicitação. Em seguida o servidor finaliza a conexão através do envio de um

pacote do tipo FIN para o cliente. Estes pacotes não estão representados na figura acima, pois não há a necessidade de seu aguardo para que o cliente possa prosseguir.

3. O cliente analisa o documento HTML enviado pelo servidor e então abre uma nova conexão TCP que novamente necessita de arquivos de SYN e ACK, e solicita, por exemplo uma imagem que deseja obter.
4. A imagem solicitada começa a chegar para o cliente.

Para cada nova solicitação o cliente deve abrir uma nova conexão TCP, o que acaba gerando uma sobrecarga de setup nas conexões. O cliente tem que esperar no mínimo 4 RTTs (Round Trip Time) para começar a visualizar a imagem que deseja. Para cada imagem adicional, pelo menos dois RTTs são necessários, um para estabelecer a conexão TCP e outro para obter a imagem.

Na versão mais recente do HTTP, ilustrado na figura 3.11(b), apenas uma conexão TCP é utilizada para solicitações consecutivas, ou seja, várias solicitações HTTP podem ser realizadas sob uma mesma conexão TCP, o que diminui o tempo de espera por uma imagem.

Hardware e Sistemas Operacionais

Em sua essência, a performance de um servidor Web depende do comportamento de seus elementos básicos, como a plataforma de hardware e o sistema operacional. Com relação ao hardware os seguintes componentes afetam a performance de um servidor:

- Processadores: Velocidade de cada um deles e a quantidade disponível
- Memória: Capacidade

- Sistemas de discos magnéticos: Rapidez de acesso e capacidade
- Largura de banda da placa de rede

Com relação ao sistema operacional utilizado, a disponibilidade, a confiabilidade e a robustez de cada um dos sistemas disponíveis no mercado devem ser alguns dos critérios utilizados no momento de escolha entre cada um deles.

Conteúdo

O verdadeiro valor de um Web Site está na relevância e na qualidade do conteúdo apresentado aos usuários, e na capacidade de superar as expectativas daquele que procura, neste Web Site, algum tipo de informação. Este conteúdo, que pode ser apresentado de várias formas diferentes, como páginas HTML, imagens, sons, etc. o que tem grande impacto na performance dos Web Servers que os contém, e conseqüentemente na rapidez da resposta fornecida ao usuário.

Nos casos em que se possui um conteúdo que seja muito visitado, pode-se optar por “espelhar” este conteúdo nos chamados *Mirror* servers, que auxiliam no redirecionamento de tráfego, evitando gargalos em redes.

3.4.2 Análise de transações Web

Para que possa ser realizado um correto planejamento em relação a mudanças na estrutura tecnológica de uma arquitetura Web, deve-se primeiro analisar quais os principais pontos no sistema responsáveis pelos atrasos nas conexões, e pela demora no recebimento das informações solicitadas pelo cliente.

Utilizando como base os Gráficos de Processamento de Comunicação, já apresentados anteriormente neste trabalho, vamos analisar em detalhes os passos de uma transação Web, os tempos demandados por cada etapa do

processo e verificar quais são os principais gargalos que impactam o tempo total gasto em uma transação.

Fluxo de uma Transação Web

Dado o enorme número de componentes de um sistema como a Internet, faremos a análise de uma transação tendo como foco exclusivamente as três principais entidades que compõem uma rede e que possibilitam a obtenção de informações através de solicitações: Cliente, Servidor e Rede. Como na parte cliente as transações Web são sempre realizadas através de um Browser, nossa análise será então feita sobre essas três partes: Browser, Rede e Servidor.

A figura 3.12 servirá de guia para o estudo do fluxo de uma transação Web.

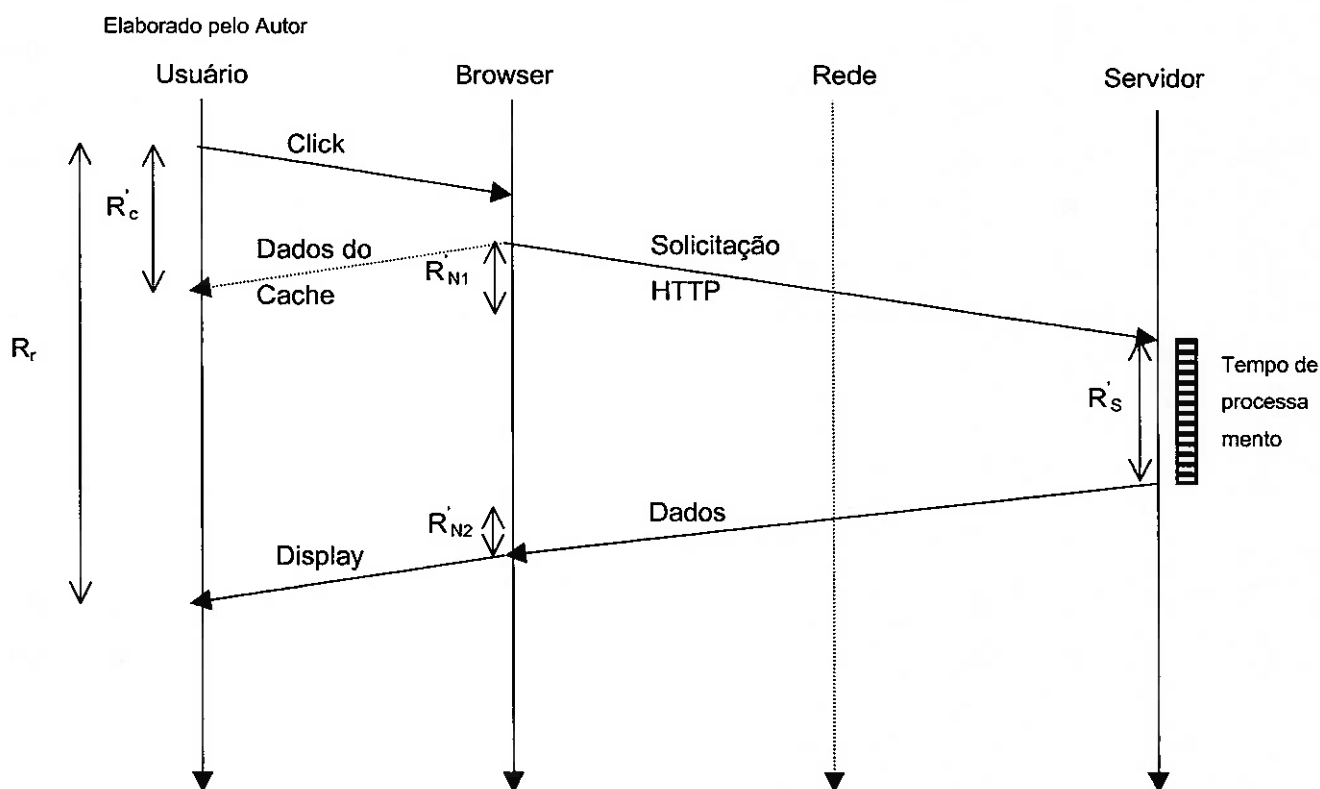


Figura 3.12 - Anatomia de uma Transação Web

1. Browser:

- O usuário clica em um hyperlink e faz a solicitação a um documento.
- O Browser procura pelo documento especificado no Cache local, e o retorna no caso de encontrá-lo. Neste caso, o tempo de resposta ao usuário é dado por R'_{cache} .
- No caso de não encontrar o documento no Cache, os seguintes passos ocorrem:
 - O browser mapeia o endereço IP correspondente ao DNS (Domain Name Service)
 - Abre uma conexão TCP com o servidor, definida pela URL do link
 - O cliente envia uma solicitação HTTP ao servidor.
 - Ao receber a resposta do servidor, o Browser mostra as informações solicitadas na tela, já formatadas.

2. Rede

A rede que interliga a parte cliente à parte servidor é responsável por alguns dos atrasos vistos na figura acima, como por exemplo a transferência de informações entre o cliente e o servidor, representado na figura por R'_{N1} e R'_{N2} . Estes atrasos são gerados pelos inúmeros componentes que existem no caminho, como modems, roteadores, etc...Passaremos a chamar a partir de agora o tempo gasto por uma solicitação HTTP na rede de R'_{rede} . Sendo assim, temos que $R'_{\text{rede}} = R'_{N1} + R'_{N2}$.

3. Servidor

- Uma solicitação chega do cliente
- O servidor analisa a solicitação, de acordo com o protocolo HTTP
- O servidor executa o método descrito

- Se a solicitação for para a busca de um documento (GET) o servidor procura o arquivo em seu sistema. O arquivo pode estar no cache ou nos discos.
- Após encontrar os arquivos, o servidor disponibiliza as informações para envio ao cliente colocando-os na porta de rede.
- No momento em que a transferência se encerrar, o servidor fecha a conexão

Considerando-se todo o ciclo de uma transação Web, o tempo total de processamento pode, basicamente, obedecer duas regras:

Se o documento solicitado for encontrado no cache do cliente o tempo total de processamento da transação será dado por:

$$R_r = R'_{cache}$$

Se, por outro lado, o documento não for encontrado no cache local, o tempo total de processamento da solicitação será dado pela soma de todos os tempo de processamento do sistema, ou seja:

$$R_r = R'_{Browser} + R'_{Rede} + R'_{Servidor}$$

Se considerarmos um número suficientemente grande de solicitações, dito N_t e considerando-se que desse total, N_c transações podem ser atendidas por arquivos encontrados no cache local temos que $P_c = N_c / N_t$ é a probabilidade de um documento ser encontrado no cache local. Sendo assim podemos definir o tempo médio total de processamento de uma solicitação como sendo:

$$\bar{R}_r = P_c \times R'_{Cache} + (1 - P_c) \times R_r$$

Ao longo de todo o caminho percorrido por uma transação Web, existe uma série de componentes que podem ter influência sobre o tempo total de processamento. Conforme o número de clientes e de servidores vai aumentando, a performance do sistema passa a sofrer com as restrições impostas por esses componentes. Esses componentes que limitam a performance geral do sistema são chamados de *Gargalos* e são elementos chave na análise do tempo de processamento, já que são eles que devem sofrer algum tipo de melhoria, ao se objetivar um ganho na performance geral.

Para que se possa realizar uma análise do tempo total de processamento de uma transação, basta, portanto, que se calcule a saída média dos componentes do sistema e que se atribua uma atenção especial àquele que apresentar o menor valor para este indicador.

3.4.3 Métricas para se avaliar o desempenho de Sistemas Web

O desempenho de um sistema Web até este momento foi analisado sob um aspecto qualitativo, relativo a quais seriam os tipos de atraso mais comumente encontrados nas redes Internet e Intranet. Serão apresentados, nesta etapa, os principais indicadores para a avaliação quantitativa deste desempenho e para a avaliação de eventuais pontos a serem melhorados no sistema.

Segundo MENASCÉ, a performance pode ser analisada sob diferentes pontos de vista. O usuário deseja que suas solicitações sejam atendidas com a maior rapidez possível, por outro lado, os administradores de Web Sites, os chamados Webmasters, têm por objetivo o maior número de transações atendidas por unidade de tempo, ou seja, ele deseja que a saída média de seu Web Site seja a maior possível.

Para atender estes diferentes tipos de demanda, serão analisados: o tempo de espera pelo processamento de uma transação, chamado de *Tempo de Latência* e a Saída Média do Sistema.

A saída média do sistema pode ser avaliada pelo número de solicitações HTTP atendidas pelo servidor em termos de *Conexões HTTP por Segundo*, porém, dada a variabilidade entre as solicitações feitas aos servidores, a saída média de um sistema é algumas vezes medida em *bits por segundo* (bps).

O Tempo de Latência será medido pelo *Tempo de Resposta de uma Solicitação*, dado pelo tempo de processamento no servidor somado ao tempo de transferência de informações através da rede mais o tempo de processamento na parte cliente (formatando informações, por exemplo).

Em resumo, as variáveis a serem mensuradas, para avaliação do desempenho de um sistema Web serão:

- Número de conexões por segundo
- Bits, ou Mbits por segundo, de acordo com a ordem de grandeza.
- Tempo de resposta de uma solicitação

Capítulo 4 - Metodologia para planejamento de capacidade

Realizar um planejamento de capacidade que seja de real utilidade e que possa ser implementado de uma maneira sistemática, é uma tarefa que exige o delineamento de etapas consecutivas a serem seguidas de modo a se atingir um objetivo comum: a Avaliação da Performance.

Apresentaremos, a seguir, um modelo de Planejamento de Capacidade baseado nos conceitos vistos anteriormente neste trabalho e que visa enquadrar todo o processo de se planejar a capacidade de um sistema em uma moldura a ser seguida passo a passo.

Antes de entrarmos no modelo, colocaremos uma definição clara sobre o que realmente significa uma capacidade adequada, ou seja, quais são os objetivos a serem atingidos no final do modelo, para que este tenha sucesso.

4.1 Capacidade Adequada

Existem vários casos em que quantias muito elevadas são gastas na implantação de um sistema de informações, e valores iguais ou ainda maiores são empregados na manutenção e na atualização do sistema.

Isso pode ocorrer quando, na fase de planejamento do sistema, não foi feita uma correta especificação sobre quais requisitos o sistema deveria atender e de que maneira isso seria realizado. Na Engenharia de Produção, aprendemos que quanto mais se avança nas fases de um projeto, planejamento, estruturação, implementação, etc., mais altos são os custos de se corrigir erros.

Por esse motivo, é muito importante que se tenha em mente quais são os requisitos a serem atendidos pelo sistema de modo a se ter uma capacidade adequada.

Segundo MENASCÉ, *Capacidade Adequada de um Sistema* pode ser definida através de dois elementos básicos:

- **Nível de Serviço Aceitável:** Ao se decidir pela implementação de um novo sistema de informações, os responsáveis devem verificar junto

aos usuários finais quais são os níveis de serviço esperados, em termos das métricas apresentadas anteriormente. O tempo de resposta de um servidor, até que a página esteja totalmente disponível para o usuário não ficar além dos quatro segundos; um servidor ter a saída média mínima aceitável de 5000 solicitações HTTP por segundo são exemplos claros de níveis de serviço. A capacidade adequada seria, então, manter os indicadores especificados em valores inferiores ao nível de serviço esperados pelos usuários, buscando inclusive uma superação das expectativas.

- **Especificar as tecnologias a serem utilizadas:** Para se atender às restrições impostas pelos níveis de serviço, existem muitas maneiras de se especificar um sistema, com diferentes tipos de hardware, software, diferentes tipos de rede, etc. As empresas podem optar por um determinado tipo de infra-estrutura tecnológica pelo relacionamento que possuem com possíveis fornecedores e/ou pela padronização já existente no mercado, por exemplo, o uso do protocolo TCP/IP para comunicação. De qualquer maneira, é muito importante que existam equipamentos capazes de suprir a demanda tecnológica e que estes estejam corretamente especificados pelos responsáveis.

Além destes dois elementos, existe ainda o fator *Custo* de implantação da solução, porém este fator, por acharmos que não tem maior relevância para o escopo deste trabalho, não será considerado em nossa análise.

4.2 O Modelo de Planejamento de Capacidade (MPC)

Como dito anteriormente, o modelo de planejamento de capacidade apresentado será baseado em uma metodologia composta por regras bem definidas a serem seguidas. Este modelo é uma adaptação do modelo de planejamento de capacidade proposto por MENASCÉ, adaptado para as necessidades deste projeto. Esta metodologia está baseada em dois

qual?
degr.
seg?

modelos, chamados de modelos auxiliares, que servirão como base para a sequência de passos da metodologia principal.

Nesta primeira etapa será dada prioridade ao estudo do modelo como um todo, buscando uma visão geral do que será a base para um planejamento de capacidade baseado em um método.

Em seguida, iremos decompor o MPC, realizando uma análise de cada uma das partes que o compõe, levando em conta a importância de cada uma delas para o resultado final proposto.

Um esquema geral do Modelo de Planejamento de Capacidade é mostrado na figura 4.1.

Elaborado pelo Autor

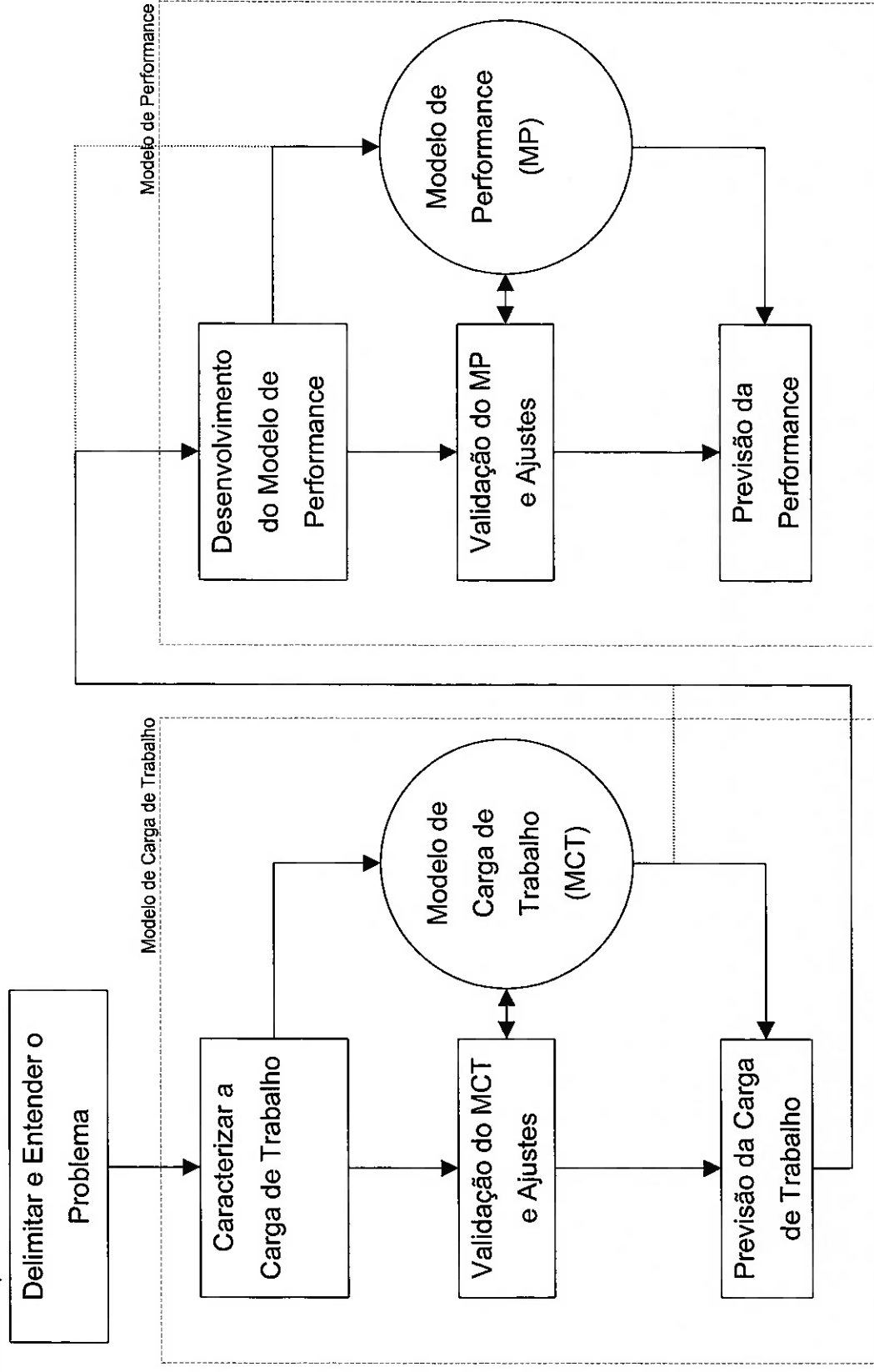


Figura 4.1 - O Modelo de Planejamento de Capacidade (MPC)

Como podemos ver no esquema do MPC, existem basicamente duas partes que compõem o modelo, sendo uma delas chamada genericamente de *Modelo de Carga de Trabalho (MCT)* e a outra chamada de *Modelo de Performance (MP)*. Fazendo-se uma análise geral do modelo vemos que ele é composto de três fases distintas, as quais podem ser chamadas:

- Delimitar e Entender o Problema
- Modelo de Carga de Trabalho (MCT)
- Modelo de Performance (MP)

Na primeira etapa, o modelo tem como entrada uma análise do ambiente onde será desenvolvido o projeto, incluindo especificações de hardware, software e redes, bem como os níveis de serviço atualmente existentes no sistema.

Com este ponto de partida, pode-se passar ao estudo da carga de trabalho a que estará sujeito o novo sistema, através do desenvolvimento do modelo de carga de trabalho (MCT). Este modelo estará sujeito a uma validação para verificar o impacto das abstrações feitas durante a modelagem. Nesta segunda fase faz-se também a previsão da carga de trabalho futura e as variações que acontecerão, de acordo com dados históricos e com o planejamento estratégico da empresa.

Na terceira e última fase, teremos a elaboração e o desenvolvimento do modelo de performance do sistema. Isto significa verificar se o sistema atenderá aos requisitos de nível de serviço especificados, de acordo com as métricas inerentes a cada um deles, por exemplo, tempo de resposta, ou número de solicitações atendidas. É nesta fase que se pode avaliar o correto Planejamento de Capacidade e os benefícios trazidos pelo novo sistema.

A partir desta visão geral do modelo, passaremos, agora, a analisar cada uma dessas fases em detalhes.

4.2.1 Fase 1 - Delimitar e Entender o Problema

O modelo todo tem início em uma parte chamada de Delimitar e Entender o Problema, que consiste no estudo aprofundado sobre quais serão os componentes do sistema em questão, que se pretende utilizar, ou que já estão sendo utilizados e o desempenho atual apresentado pelo sistema. Nesta parte são definidos que tipos de Hardware, Software, redes, protocolos de rede, etc. estão envolvidos no ambiente, além das especificações de nível de serviço existentes, utilização dos componentes, estrutura de gerenciamento, etc. Todos estes dados deverão ser coletados utilizando-se de instrumentos de coleta de dados como, por exemplo, questionários, entrevistas, levantamentos realizados junto aos usuários nas áreas, etc. Alguns elementos a serem considerados nesta etapa inicial estão descritos na tabela 4.1:

Elaborado pelo Autor

Elementos	Descrição
Plataforma utilizada pela parte Cliente	Quantidade e tipo
Plataforma em que está baseado o Servidor	Quantidade, tipo, configuração e funções
Aplicativos	Quais são os utilizados com maior frequência e os mais importantes
Conectividade das Redes	Levantar os diagramas de conectividade mostrando todas as redes locais, remotas, roteadores, servidores e o número de clientes em cada uma das redes.
Protocolos de comunicação	Lista dos protocolos utilizados pelas redes
Níveis de serviço aceitáveis	Os atuais Níveis de Serviço demandados. Se não existirem, podem ser considerados os níveis padrão aceitos pelo mercado.
Gerenciamento das redes	Estrutura de suporte às redes

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

Procedimentos de compra	Levantamento da burocracia necessária para aquisição de equipamentos e para instalação dos mesmos.
-------------------------	--

Tabela 4.1 – Variáveis do Ambiente Estudado

Tendo realizado estes levantamentos, pode-se passar a uma análise da demanda a que estará submetida a nova infra-estrutura tecnológica, através do modelo de carga de trabalho.

4.2.2 Fase 2 - O Modelo de Carga de Trabalho (MCT)

Como visto no MPC o modelo de carga de trabalho é composto por três passos distintos analisados em detalhes a seguir.

Caracterizar a carga de trabalho

Esta fase terá como objetivo descrever a carga de trabalho global do sistema a partir da decomposição primária da carga total para os componentes principais do sistema. Em seguida a carga de cada um dos componentes é analisada com relação aos componentes que formam esse elemento do sistema, e assim por diante até atingirmos um nível suficientemente analítico para que possamos verificar a intensidade demasiada de carga em um componente e caracterizá-los através de um indicador de volume de trabalho, como, por exemplo, número de transações, chegada média de solicitações, etc. A partir desses indicadores, podemos concluir quais são os principais parâmetros que denotam a super ou a sub-utilização de um determinado componente. Um exemplo de decomposição entre componentes de um sistema podem ser vistos na figura 4.2:

Adaptado de MENASCÉ, 1998

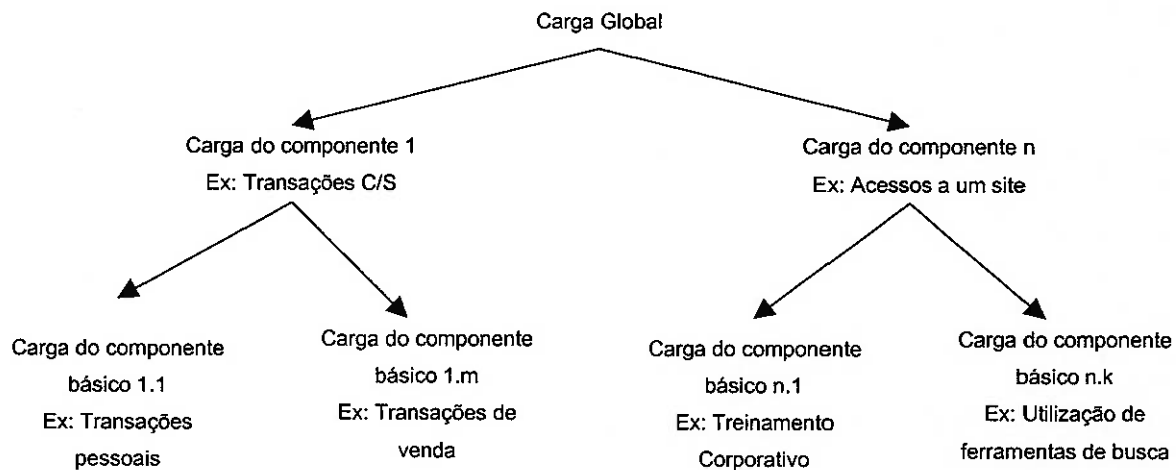


Figura 4.2 – Decomposição Entre Componentes do Sistema

O levantamento destes parâmetros geralmente é feito de forma indireta, ou seja, através da medição de uma consequência atrelada ao parâmetro principal. Se quisermos, por exemplo, medir o número de sessões de um treinamento corporativo por dia, ou o número de documentos HTTP baixados por dia. Com o número de solicitações e o tempo gasto em cada solicitação pode-se calcular o tempo total gasto em um determinado componente básico de um sistema. E assim prosseguir para os outros componentes.

Validação do MCT e Ajustes

O passo seguinte consiste na validação do modelo de carga de trabalho proposto, isto é, para que se possa modelar uma situação que na verdade é real, algumas abstrações devem ser feitas, por exemplo, facilitação na coleta de dados através de um número finito de medições, a ineficiência computacional no processo de modelagem, a inviabilidade de se medir com a precisão necessária determinado componente, etc. Estes são fatores que acabam prejudicando a precisão do modelo. Para que o modelo seja considerado aceito, faz-se uma comparação entre a carga real utilizada e uma carga imaginária, que será imputada no sistema para testes. Se os

resultados da carga de teste forem precisos o bastante para se ter resultados entre 10 e 30% próximos aos daqueles obtidos com cargas reais, então o modelo é considerado válido.

Previsão da Carga de Trabalho

A próxima etapa é muito importante na decisão da Capacidade do novo sistema e é particularmente difícil de ser medida em sistemas Internet ou Intranet. A previsão da carga de trabalho é função do planejamento estratégico de crescimento da empresa, e conseqüentemente da variação no número de solicitações enviadas ao Web Server. Esta previsão envolve também dados históricos da empresa e a tradução de como esta tendência nos negócios estará relacionada à demanda tecnológica da empresa.

Com os dados referentes à demanda de utilização do sistema, pode-se comparar os requisitos mínimos de hardware que devem ser apresentados e aqueles disponíveis para utilização. Uma primeira adequação da infra estrutura disponível aos requisitos é um bom indicador que a performance global será aceitável, o que será comprovado no modelo de performance.

4.2.3 Fase 3 - O Modelo de Performance (MP)

Semelhante ao MCT o modelo de performance possui três passos que devem ser seguidos, sendo o último deles a verificação de todo o trabalho de planejamento de capacidade, por se tratar de uma prévia daquilo que será o novo modo de trabalho do sistema.

Desenvolvimento do Modelo de Performance

O desenvolvimento de um modelo de performance está intimamente ligado ao processo de previsão das medidas de performance de um sistema, segundo certos parâmetros de medição. Algumas medidas de performance utilizadas são: tempo de resposta, saída média do sistema, utilização de um

certo recurso, etc. Os parâmetros relativos a um modelo de performance podem ser divididos em:

- **Parâmetros de Sistema:** São os parâmetros relativos ao sistema como um todo como, por exemplo, o protocolo utilizado nas conexões, o balanceamento de conexões através da divisão entre servidores espelho, etc.
- **Parâmetros de Recursos:** São características intrínsecas de um recurso que podem afetar a performance. Exemplos de parâmetros de recurso são tempos de procura em disco, largura de banda das redes, velocidade de processamento, etc.
- **Parâmetros de Carga de Trabalho:** São derivados da caracterização da carga de trabalho e podem ser divididos em:
 - **Parâmetros de Intensidade de Carga:** Nos fornecem uma idéia da carga imposta ao sistema como um todo, por exemplo, o número de solicitações entradas por dia ou o número de clientes utilizando uma aplicação específica.
 - **Parâmetros de Demanda de Serviço:** Especificam o tempo total de serviço demandado por cada componente básico, em cada recurso. Exemplos são: tempo de processamento no servidor, o tempo total de transmissão das respostas para a rede número 2, etc.

Existem basicamente dois modelos de performance que podem ser utilizados como base para a previsão do desempenho de um sistema: Modelos de Simulação e Modelos Analíticos. Ambos os modelos devem trabalhar respeitando os tempos de utilização e as filas formadas em cada recurso componente do sistema. O quão analítico será o modelo depende do propósito a que se destina e à disponibilidade de informações para sua elaboração.

Validação do MP

A exemplo do MCT, o modelo de performance será válido se os resultados obtidos em cargas de teste estiverem contidos em uma margem de erro de 10 a 30% em relação ao sistema real.

Previsão da Performance

Este é o ponto de conclusão do modelo, onde os principais parâmetros serão analisados e os pontos críticos ficarão evidentes.

Para se prever a performance de um sistema, é necessário que solucione o modelo de performance que representa o sistema. Os modelos analíticos são baseados em fórmulas matemáticas que geram métricas de performance a partir dos parâmetros. Modelos de simulação são programas de computadores que nos informam dinamicamente o comportamento de cada recurso pertencente ao sistema e, com a união destes recursos, nos mostram um fluxo de informações completo. Estatísticas a respeito da formação de filas, tempo de utilização, entre outras informações de cada recurso são disponibilizadas pelo programa e nos ajudam a identificar qual recurso apresenta maior impacto sobre a performance global do sistema.

Existem várias alternativas de se identificar a performance de um sistema. Quanto maior o grau de especificação dos componentes de um sistema, maior a precisão da resposta final. No entanto, à medida que cresce o número de componentes estudados, aumenta bastante o esforço necessário para uma melhora de informação não tão significativa. Deve-se, então procurar um balanceamento entre a precisão do modelo e o esforço necessário para se conseguir a informação desejada.

4.3 Aplicação do MPC a um caso real

Como forma de atribuir um caráter mais prático ao modelo proposto, faremos agora sua aplicação a um caso real de planejamento de capacidade, verificando a carga de trabalho a que está submetido o sistema, construindo um modelo de performance que verifique o atendimento aos padrões estabelecidos e comprovando que a estrutura do Site está adequada para o volume de solicitações submetido.

Com o objetivo de manter o posicionamento estratégico adotado pela empresa e evitar a consulta dos dados aqui apresentados por seus concorrentes, seus diretores solicitaram que o nome da empresa não fosse divulgado, o que de maneira alguma irá prejudicar o objetivo a que se propõe este trabalho, que é verificar a aplicação do MPC a um caso real.

A empresa aqui estudada atua no setor de vendas de produtos pela Internet possuindo uma participação considerável no mercado em que atua. Tem por política a implantação de uma infra-estrutura tecnológica de ponta, buscando sempre se manter equalizada, ou mesmo à frente de seus concorrentes.

Faremos, então, a aplicação do MPC à empresa em questão.

4.3.1 Fase 1 - Delimitar e Entender o Problema

Como especificado no modelo, é nesta etapa que deve ser entendido todo o ambiente de produção que será analisado, buscando o máximo de informações sobre o mesmo.

Nosso estudo terá início na apresentação da infra-estrutura tecnológica utilizada pelo Web Site para sua operação. Dado que na infra-estrutura original existiam componentes utilizados apenas pela parte administrativa da empresa, o que apresentaremos aqui será uma simplificação da arquitetura técnica utilizada considerando apenas os componentes utilizados na parte operacional do site.

A arquitetura técnica operacional está esquematizada na figura 4.3:

Elaborado pelo Autor

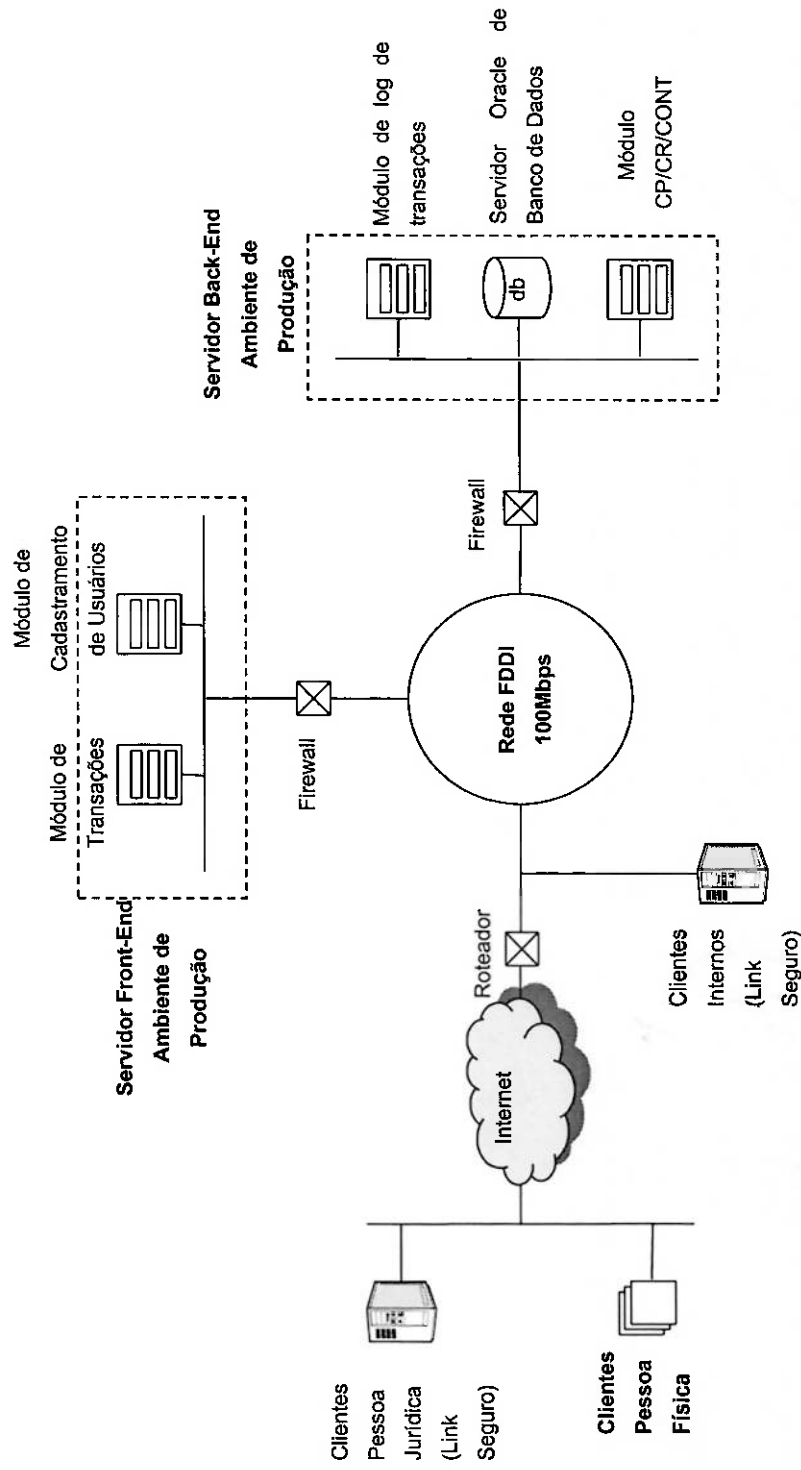


Figura 4.3 - A Arquitetura Técnica e as Instalações de Produção

A arquitetura técnica, seguindo o modelo de arquiteturas Cliente/Servidor visto anteriormente, é basicamente formada por três componentes de natureza semelhante:

- Os clientes, representados pelos clientes do tipo pessoa física, os clientes do tipo pessoa jurídica, que possuem um link de compra dedicado e com conexão segura, e os clientes internos.
- A rede, que, de maneira simplificada, é formada apenas por uma rede do tipo Fiber Distributed Data Interconnect (FDDI)
- Os servidores, que podem ser agrupados em: Front-End Server e Back End Server.

Seguindo a metodologia recomendada para o levantamento dos elementos funcionais de cada uma dessas partes, temos:

Elaborado pelo Autor

Elementos	Descrição
Plataforma Utilizada pelo Cliente	• A plataforma utilizada pela parte cliente é basicamente o Windows em suas várias versões (95, 98, 2000, NT).
Plataforma Utilizada pelos Servidores	• Dada a necessidade de se manter o site em operação. Tanto no Front-End Server quanto no Back-End Server a plataforma utilizada é UNIX, por ser um sistema operacional que garante a utilização contínua do servidor. • As máquinas utilizadas como servidores são do tipo SUN E450 com dois processadores e 1Gb de memória em cada e espaço para armazenamento de dados de 144Gb.
Principais Aplicativos da parte Cliente	• Na parte cliente, os principais aplicativos são os Web Browsers tradicionais: Internet Explorer e Netscape Navigator utilizados para a navegação no site e para a entrada de solicitações

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

Principais Aplicativos Utilizados no Servidor Front End	- No servidor Front-End é utilizado um pacote para sites de comércio eletrônico chamado Open Market Transact versão 4. Esse aplicativo é padrão para sites que efetuam transações via Web e possibilita, entre outras, as seguintes funcionalidades: ➤ Captura segura de ordens de compra ➤ Cadastramento de Usuários ➤ Autorização e processamento de ordens de pagamento online ➤ Cálculo de taxas e fretes ➤ Entrega de produtos digitais via Web ➤ Integração com banco de dados para geração de relatórios e análise
Principais Aplicativos Utilizados no Servidor Back End	No servidor Back_End além da parte do Transact relativa a Contas a Pagar/Receber e Contabilidade e log das transações realizadas, é utilizado um banco de dados que é acessado pelo Front-End Server para consultas relativas às vendas.
Conectividade das Redes	Internamente ao site, temos basicamente uma rede do tipo FDDI com largura de banda de 100Mbps que interliga as solicitações entradas pelos clientes (internos e externos), o servidor Front-End e o Servidor Back-End.
Protocolos de Comunicação	Conforme o padrão de utilização em redes o protocolo de transmissão e comunicação utilizado é o TCP/IP.
Níveis de Serviço Aceitáveis	Tempo total de processamento de uma transação, de qualquer natureza, inferior a 5 segundos.

Tabela 4.2 – Variáveis do Ambiente Analisado

Tendo este primeiro contato com o ambiente a ser estudado, pode-se passar a uma fase de previsão da carga de trabalho a que o sistema está submetido. Essa previsão será realizada através da construção do MCT, o modelo de carga de trabalho.

4.3.2 Fase 2 - O Modelo de Carga de Trabalho (MCT)

O modelo de carga de trabalho tem por objetivo caracterizar a demanda a que está submetido o sistema e verificar se este possui os requisitos de infra-

estrutura que atendam a estes requisitos. Se as instalações não forem coerentes com a utilização do sistema, nem é recomendada a passagem para a terceira fase, já que claramente não serão atendidos os requisitos de performance desejado pelos usuários ou pelos administradores do sistema.

Caracterizar a carga de trabalho

No sistema em questão, a carga de trabalho está relacionada a basicamente três tipos de operações que podem ser realizadas pelo usuário:

- Registro de seus dados pessoais
- Criação de uma cesta de compras
- Efetivação de uma compra

Cada um destas solicitações é tratada em locais diferentes do sistema, de acordo com o tratamento dado pelos servidores à transação. Segundo a natureza de cada uma delas, o trabalho está distribuído da seguinte maneira entre os servidores:

- **Front End Server:** Neste servidor estão presentes os módulos do Transact referentes a Efetivação de Transações e Cadastramento de Usuários. Além dos acessos ao sistema local de arquivos, neste servidor também há o processamento de CGIs, que possibilitam algumas funcionalidades não encontradas no aplicativo Open Market Transact.
- **Back End Server:** Neste servidor estão englobados os módulos de Log de Transação e de Contas a Pagar/Receber/Contabilidade. Porém a parte que mais nos interessa neste servidor é o Banco de Dados Oracle nele contido. Este banco de dados é acessado diretamente pelo Front-End Server, mais especificamente pelo módulo de Efetivação de Transações.

Validação do MCT e Ajustes

Pelo fato de termos os dados aqui apresentados fornecidos pela própria empresa estudada, consideraremos que se encontram dentro da margem de erro citada na apresentação do modelo. As simplificações realizadas não terão impacto nos resultados obtidos.

Previsão da Carga de Trabalho

A previsão da carga de trabalho pode ser realizada através do levantamento de alguns dados importantes que podem ser observados na tabela 4.3.

Elaborado pelo Autor

	Item	Valor	Fonte
Demanda			
D1	Número anual de transações efetivadas	300000	Estimativas
Compras			
C1	Número de itens por Ordem de Compra	2	Estimativas
C2	Número de Páginas Disponíveis para os Usuários	150	Média de páginas disponíveis aos clientes
C3	Número de Objetos por Página	20	Desenvolvimento das páginas
C4	Número de Cliques por visita	100	Histórico de compras
C5	Tamanho médio de um objeto na página (image/html)kbytes	5	Desenvolvimento das páginas
Carga de Trabalho por usuário			
W1	Média de horas de consumo por dia	14	Histórico de compras
W2	Porcentagem de clientes que criam uma cesta de compras	20%	Histórico de compras
W3	Porcentagem de clientes que possuem uma cesta e realmente querem comprar	10%	Histórico de compras
W4	Porcentagem de usuários registrados	5%	Histórico de compras
Transações – Banco de Dados Oracle			
T1	Número de transações no Banco de Dados para adicionar um produto à cesta de compras	35	Arquivo de Log de Transações
T2	Número de transações no Banco de Dados para o preenchimento de registro de usuário	55	Arquivo de Log de Transações
T3	Número de transações no Banco de Dados para efetuar uma compra	82	Arquivo de Log de Transações
T4	Número de processo CGI por ação de compra	1	Arquivo de Log de Transações
T5	Memória utilizada por processo CGI (Mbytes)	25	Estimativa
Dados Técnicos - Banco de Dados Oracle			
B1	Memória demandada pelo processamento do banco de dados(Mbytes)	200	Média Estatística

Grupos de trabalho

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

B2	Memória utilizada pelo acesso ao banco(Mbytes)	20	Média Estatística
B3	Número de acessos ao banco de dados	5	Estimativa
B4	Espaço necessário para armazenamento das tabelas do banco de dados(Mbytes)	50000	Documentação Oracle
B5	Espaço extra para administração do banco de dados	25000	Documentação Oracle

Tabela 4.3 – Levantamento de Dados de Carga de Trabalho

Esses dados nos servem como parâmetros para o levantamento dos requisitos mínimos que o sistema deve possuir, em termos de infra-estrutura, para apresentar uma performance satisfatória. Esses requisitos são vistos na tabela 4.4:

Elaborado pelo Autor

	Item	Cálculo	Valor
	Compras		
CO1	Número anual de transações que utilizam preenchimento completo de ordem de compra	D1	300.000
CO2	Número de clientes por ano que realizam transações	D1	300.000
CO3	Número de clientes que criam cestas de compras	A4/W3	3.000.000
CO4	Número de horas de compra por ano	365*W1	5.110
CO5	Número e usuários utilizando o Front End Server por hora	CO3/CO4	587
CO6	Número máximo de usuários concorrentes utilizando o Front End Server	CO5/60	10
CO7	Número total de usuários por ano que visitam o site	D1/(W2*W3)	15.000.000
	Largura de Banda		
LB1	Volume de dados baixado por um potencial cliente(Kbytes)	C4*C5	500
LB2	Volume de dados baixado anualmente(Kbytes)	LB1*CO7	7.500.000.000
LB3	Largura de banda(Kbits/s)	8*LB2/(CO4*3600)	3.262
	Espaço em disco		
ED1	Número de arquivos disponíveis	C2*C3	3.000
ED2	Espaço em disco necessário (Kbytes)	ED1*C5	15.000
	Memória para Front End Server		
MF1	Número de processos CGI simultâneos	T4*CO6	10
MF2	Memória utilizada pelos processos CGI (Mbytes)	MF1*T5	250
	Banco de Dados Oracle		
BD1	Espaço em disco para as instâncias Oracle(Mbytes)	B4+B5	75.000
	Memória para o processamento do Banco Oracle		
BD2	Total de memória necessária para o processamento	B1+(B2*B3)	300
	Transações Oracle		

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

BD3	Transações no banco para se adicionar um Item à cesta de compras, por hora	$(T1 \cdot C1 \cdot CO3) / CO4$	41.096
BD4	Transações no banco relacionadas a compra, por hora	$CO2 \cdot T3 / CO4$	4.814
BD5	Consulta às informações dos clientes, por hora	$CO2 \cdot W4 \cdot T2 / CO4$	161
BD6	Total de transações no Banco de Dados, por segundo	$SOMA(BD3;BD5)/3600$	13

Tabela 4.4 - Estimativas de Capacidade

Em resumo, as informações mais importantes que podemos retirar do levantamento realizado são:

Elaborado pelo Autor

	Front End Server	Back End Server (Oracle)
Espaço em disco(Mbytes)	15	75000
Memória (Mbytes)	250	300
Largura de banda(Mbits/s)	3,3	

Tabela 4.5 - Resumo das Estimativas de Capacidade

Verifica-se que existe uma demanda muito mais elevada por espaço em disco no Back End Server, devido ao espaço ocupado pelas tabelas do banco de dados Oracle. A utilização de memória é basicamente a mesma tanto para o Front-End Server, quanto para o Back-End Server.

Tendo realizado o levantamento da carga de trabalho, podemos partir para a verificação a respeito do atendimento, ou não, dos requisitos de performance descritos na fase de entendimento do ambiente. Esta verificação será realizada no Modelo de Performance.

4.3.3 Fase 3 - O Modelo de Performance (MP)

Desenvolvimento do Modelo de Performance

Por ser um processo análogo entre os tipos de solicitações entradas pela parte cliente, o modelo de performance a ser estudado aqui será referente a apenas um destes tipos de operações. A operação a ser analisada em detalhes será a efetivação de compra, por envolver praticamente todos os recursos disponíveis no sistema: Clientes, que realizam a entrada de

solicitações, Rede, para transferência de dados e Servidores, aí incluídos o Front-End Server que possui o módulo do Transact referente a efetivação de compras e o Back-End Server, onde as informações são guardadas através de registros no banco de Dados Oracle.

A operação de efetivação de compra envolve as seguintes etapas:

- Entrada de solicitação de compra pelo usuário
- Tráfego pela rede até o Front-End Server
- Processamento das informações no Front-End Server
- Envio das informações validadas para o Banco de Dados através da rede
- Efetivação das transações de compra no Banco de Dados
- Envio de resposta ao cliente.

Utilizando-se a notação anteriormente apresentada neste trabalho, os principais geradores de atraso no sistema seriam:

Elaborado pelo Autor

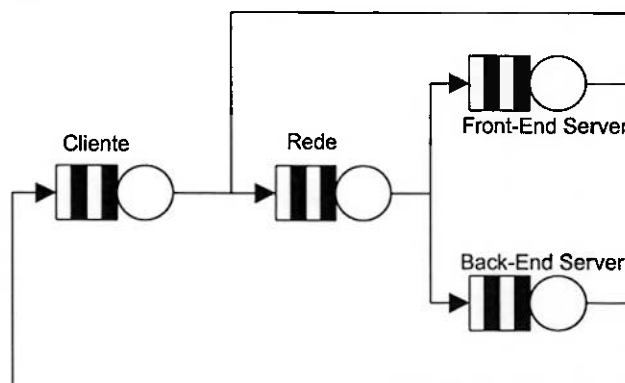


Figura 4.4 – Recursos que Geram Atrasos no Sistema

O fluxo da transação de efetivação de compra pode ser representado, através de um Gráfico de Processamento de Comunicação, da seguinte maneira:

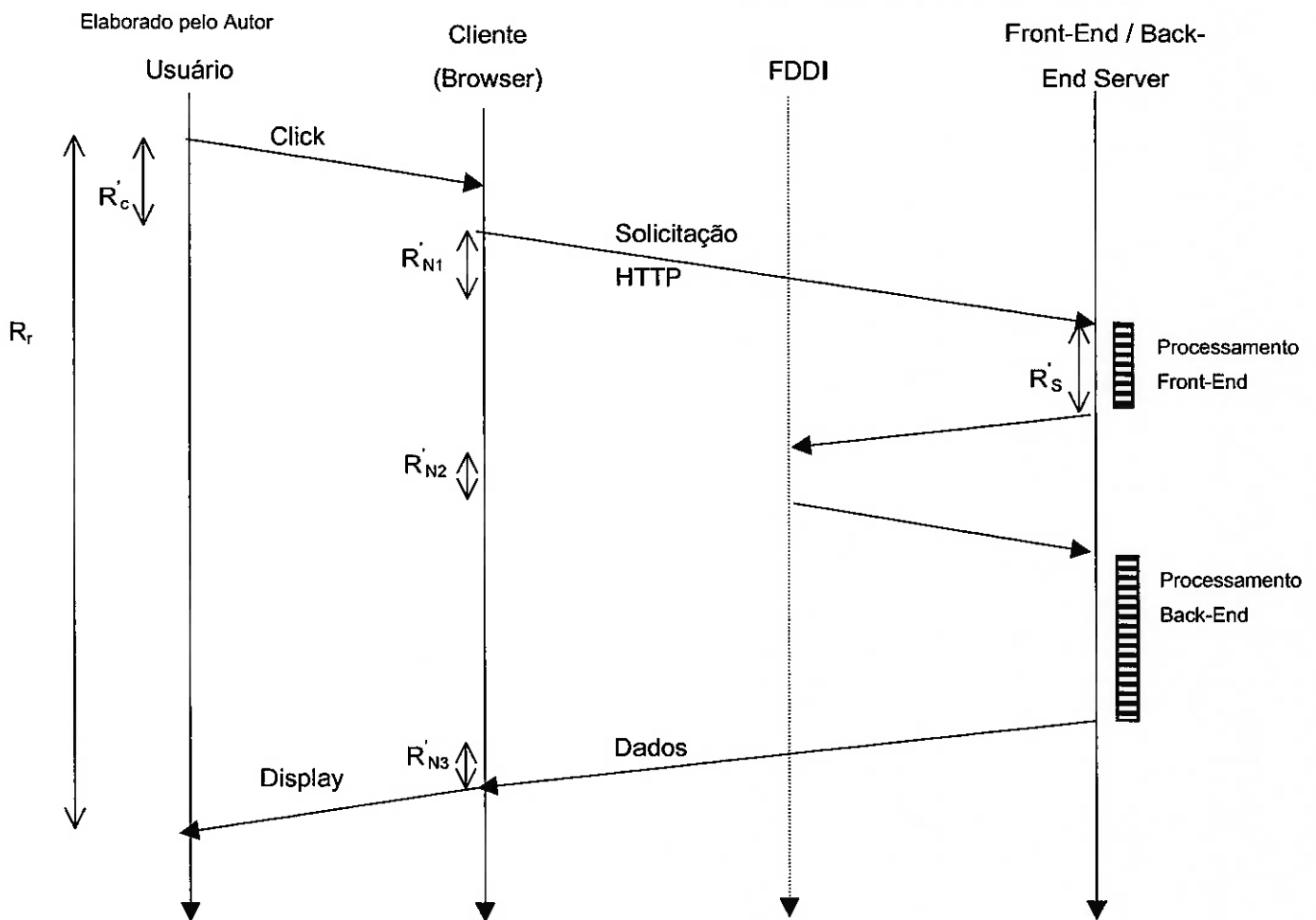


Figura 4.5 - Fluxo de Dados para Cadastramento de Clientes

Em cada uma das partes envolvidas no processamento da transação há uma certa quantidade de tempo demandada para se completar esta etapa. Associando a cada uma delas uma variável de tempo de processamento, teremos:

- R'_c : Tempo gasto com processamentos na parte cliente, como por exemplo, formatação de informações enviadas pelo servidor
- R'_n : Tempo gasto com o tráfego de dados na rede, inclui o tempo de envio da solicitação do cliente até o Front-End Server, o acesso ao Back-End Server feito pelo Front-End e o reenvio de informações do Back-End de volta ao cliente. Respectivamente R'_{n1} , R'_{n2} , R'_{n3} . Portanto: $R'_n = R'_{n1} + R'_{n2} + R'_{n3}$.

- R'_s : Tempo de processamento no servidor, incluindo o tempo gasto no Front-End Server e no Back-End Server. Portanto: $R'_s = R'_{s(F)} + R'_{s(B)}$

Sendo assim, podemos definir nosso Modelo de Performance da seguinte maneira:

$$R_r = R'_c + R'_n + R'_s$$

Onde R_r é o tempo total de processamento de uma transação de efetivação de compra.

Validação do MP

As simplificações efetuadas durante a aplicação do modelo, tais como a de desconsiderar alguns recursos de menor impacto sobre o resultado (o "firewall", por exemplo) não terão influência sobre os resultados obtidos.

Uma validação mais ampla poderá ser obtida com a aplicação do modelo a um conjunto maior de casos.

Assim como no MCT, não será possível traçar uma comparação entre o modelo aqui apresentado e o modelo real, utilizado para produção. Novamente, sendo os dados fornecidos diretamente pela empresa estudada no caso, e, efetuando-se um grau de simplificações razoavelmente baixo, , não teremos comprometimentos sérios dos resultados obtidos. Há também o fato de se realizar o estudo apenas como ilustração do modelo apresentado, não devendo dar uma atenção exagerada aos resultados obtidos, mas sim à maneira como se faz a aplicação do modelo.

Previsão da Performance

Como forma de verificação do atendimento às regras de performance estabelecidas na fase 1 da aplicação do modelo, passaremos agora à solução do modelo de performance.

Dadas as simplificações realizadas no modelo, o cálculo da performance do sistema, para a transação de efetivação de compra, será feito através da soma direta dos tempos de processamento referentes ao Modelo de Performance, como mostrado anteriormente. Efetuando-se o cálculo dos parâmetros da equação, temos:

R'_c : Tempo de processamento na parte cliente:

O tempo de processamento gasto em processamento na parte cliente refere-se ao tempo necessário para se analisar, formatar e mostrar as informações recebidas, normalmente páginas HTML. Basicamente, este tempo é dado pelo tempo gasto em acesso a disco na parte cliente. Vimos que cada cliente ao visitar o site, faz o download, em média, de 500 Kbytes de dados. A equação abaixo, já apresentada neste trabalho, nos dá o tempo de processamento em um disco magnético:

$$S_d = \text{ControllerTime} + P_{\text{perda}} \times (\text{SeekTime} + \text{Latência Rotacional} + \text{Tempo de transferência})$$

Supondo uma taxa de transmissão de um bloco de informações do disco ao controlador de disco de 10Mb/s, temos

Elaborado pelo Autor

	Valor	Fonte
Controller Time	0,10ms	Documentação
P_{perda}	80%	Estimativa
SeekTime	3,75ms	Documentação
Latência Rotacional	3,00ms	Documentação
Tempo de transferência	50,00ms	$T = \text{Tamanho do bloco} / \text{Taxa de Transferência em bytes/s}$

Tabela 4.6 – Parâmetros de Processamento na Parte Cliente

$$S_d = R'_c = 0,1 + 0,80 \times (3,75 + 3 + 50) = 45,4\text{ms}$$

R'_n : Tempo de transferência através da rede:

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

Nesta parte serão considerados os seguintes tamanhos de solicitações e taxas de transferência de arquivos:

Elaborado pelo Autor

	Tamanho da solicitação (Kbits)	Rede utilizada	Taxa de Transferência (média em Kbits/s)	Tempo de transferência (ms)
R_{n1}	80	Internet (modem)	96	830
R_{n2}	800	FDDI	100000	8
R_{n3}	120	Internet(modem)	96	1250

Tabela 4.7 - Tempos de Transferência em Redes

A partir dos dados da tabela acima, podemos concluir que o tempo total de utilização em redes é de:

$$R'_n = 830 + 8 + 1250 = 2088\text{ms}$$

R'_s : Tempo de processamento nos servidores:

O tempo de processamento gasto nos servidores deve basicamente os processamentos CGI gerados em cada efetivação de compra e pelo acesso ao banco de dados Oracle, realizados para a validação da transação de compra.

Da tabela 4.3, vemos que cada transação de efetivação de compra envolve:

- Geração de 1 processamento CGI.
- Geração de 82 transações de diferentes funções no banco de dados Oracle.

Os tempos demandados para o processamento CGI e para as transações realizadas no banco podem ser vistos na tabela abaixo:

	Tempo (média em ms)	Qtde	Tempo total (ms)
Processamento CGI (Front End Server)	30	1	30
Transações no Banco (tempo por transação)	1,2	82	98

Tabela 4.8 - Tempo de Processamento no Servidor

Além disso, vimos que, em média, 10 usuários estão concorrendo para a utilização do Front-End Server, número que será utilizado também para a utilização do Back-End Server. Sendo assim, podemos calcular o tempo médio gasto em processamento no servidor, como sendo:

$$R'_s = 10 \cdot (R'_{s(F)} + R'_{s(B)}) = 10 \cdot (30 + 98) = 1280 \text{ ms}$$

Podemos, então, calcular o tempo total de processamento de uma transação de efetivação de compra, através da tabela abaixo:

Elaborado pelo Autor

	Componente	Tempo (ms)
R_c	Tempo gasto no cliente	45
R_n	Tempo gasto em tráfego na rede	2088
R'_s	Tempo gasto em processamento no servidor	1280
R_T	Tempo Total Da Transação	3413

Tabela 4.9 - Tempo Total Gasto em uma Transação de Efetivação de Compra

Como pode ser visto, o tempo total para ser realizada uma transação de efetivação de compra é de aproximadamente 3,5 segundos. Este valor se encontra dentro do nível de serviço esperado, que era de 5 segundos, comprovando que a empresa analisada possui a infra-estrutura necessária para apresentar um bom desempenho a seus clientes.

Um estudo sobre a porcentagem do tempo de processamento devida a cada um dos componentes do sistema é realizado através do gráfico abaixo:

Elaborado pelo Autor

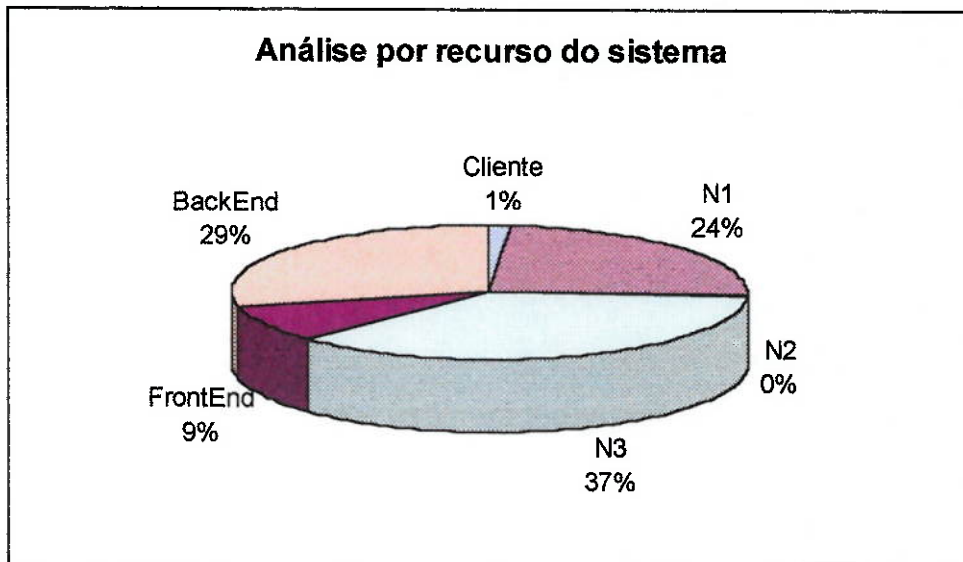


Figura 4.6 – Porcentagem Atribuída a Cada Recurso do Sistema

Vemos que aproximadamente 60% do tempo gasto na transação é perdido no tráfego de dados através da rede, no envio de solicitações para processamento (N1) e na devolução das informações processadas ao cliente (N3). Isto significa que a maior parcela de tempo de processamento está na parte do sistema que não está sob influência da empresa, as redes e roteadores que formam a Internet.

Além disso, uma parcela importante do tempo de processamento é devida ao processamento no Back End Server dado o elevado número de transações realizado no Banco de Dados para validação da efetivação de compra.

4.4 Principais restrições do modelo

O Modelo de Planejamento de Capacidade mostra-se bastante eficiente para a avaliação da carga de trabalho e da performance esperada por um sistema. No entanto, existem algumas restrições que devem ser consideradas no momento de sua utilização:

- O modelo apresentado foi aplicado para apenas um dos tipos de transação realizadas pelo Web Site, no caso, efetivação de compra. Seria

- O modelo apresentado foi aplicado para apenas um dos tipos de transação realizadas pelo Web Site, no caso, efetivação de compra. Seria interessante, em trabalhos futuros aplicar o modelo para várias transações, analisando como seria a interação entre os diferentes recursos do sistema e o impacto verificado no tempo total de processamento.
- A caracterização da carga de trabalho deve ser realizada com a fundamentação de dados que tenham a precisão necessária para a obtenção de resultados coerentes. Um desvio na fase de levantamento de dados pode comprometer todo o trabalho.
- A quantidade de recursos a ser estudada em cada sistema é bastante elevada. Deve-se buscar um ponto de equilíbrio entre a quantidade demasiada de recursos e a falta de informações importantes devido a recursos que possam ter influência no resultado, e que não tenham sido analisados. Essa decisão deve ser tomada na primeira fase do modelo, onde se procura conhecer o ambiente estudado.
- Na Internet, a demanda de um determinado Web Site é extremamente variável, o que dificulta muito a previsão de utilização dos recursos do sistema. Para uma correta análise do modelo, deve-se procurar sempre o caso onde os recursos sejam mais exigidos, buscando atender bem os clientes, mesmo em momentos de demanda máxima.
- A medição dos tempos de atraso gerados em cada recurso é uma tarefa complicada pelo fato de serem intervalos extremamente curtos e nem sempre diretamente ligados ao recurso em questão.

Capítulo 5 - Conclusões

5.1 Revisão Geral

Ao longo deste trabalho, foi desenvolvida uma linha de raciocínio que tinha por objetivo tanto o esclarecimento de um assunto novo como o e-commerce quanto mostrar a importância que terá na vida das empresas num futuro próximo.

Na primeira etapa, como forma de esclarecer aspectos como a relevância do tema estudado e, ainda, apresentar conceitos que pudessem servir de referência à metodologia e à aplicação prática apresentadas na parte final do trabalho, foram introduzidas considerações teóricas pudessem mostrar o impacto que o comércio eletrônico tem no dia dia das pessoas. Nesta fase, além de um breve histórico sobre o tema, foram mostradas vantagens e desvantagens que vêm atreladas à prática do comércio eletrônico.

Como forma de mostrar os principais hábitos daqueles que se utilizam do comércio eletrônico, foram mostrados, também, alguns dados sobre o perfil dos usuários de Internet com ênfase nos usuários que realizam compras pela Rede, como e quando fazem uso desta ferramenta de consumo e qual é o principal segmento de mercado para as empresas que desejam entrar na prática deste tipo de comércio.

Ainda na parte teórica e atribuindo um caráter mais técnico ao trabalho foram apresentados os conceitos sobre performance em sistemas, de que maneira ela pode ter influência sobre o resultado das empresas e como deve ser feito um planejamento de médio prazo prevendo-se a criação de gargalos de performance nos sistemas utilizados. Nesta etapa foi também apresentado um exemplo das principais perdas que podem ocorrer quando um Web Site não possui os requisitos de performance necessários a uma operação adequada.

A seguir foram introduzidas algumas notações que teriam utilização no caso prático apresentado, como os Gráficos de Processamento de Comunicação e, utilizando-os como base, pudemos detalhar os principais componentes de um sistema que utilize arquitetura de rede do tipo Cliente/Servidor e os tipos de atraso que cada um destes componentes pode gerar no processamento de uma solicitação. Particularizando os conceitos de arquitetura

Cliente/Servidor apresentados, foi estudada em maiores detalhes a performance em Web Sites, mostrando-se, inclusive, como é o fluxo de uma transação via Internet.

Como parte central do trabalho, o Modelo de Planejamento de Capacidade foi proposto mostrando-se suas etapas e quais os passos necessários para a realização de cada uma delas.

5.2 Avaliação de alcance dos objetivos propostos

Este trabalho tinha por objetivo global realizar um estudo sobre a capacidade de atendimento de Web Sites segundo a infra-estrutura tecnológica disponível. Como dito anteriormente, um correto planejamento de capacidade levará a empresa a uma otimização da infra-estrutura e à minimização do tempo de retorno sobre este investimento realizado.

Como descrito na proposta de trabalho, existiam algumas metas a serem cumpridas para que se chegasse ao resultado prático esperado para a realização deste trabalho.

A tabela 5.1 mostra quais eram essas metas, e analisa como cada uma delas pôde ser atingida.

Elaborado pelo Autor

<i>Meta</i>	<i>Status</i>
Introduzir os conceitos sobre e-commerce, incluindo suas diferentes formas de operação, limitações e funcionalidades que sejam relevantes para possíveis investidores neste ramo.	O capítulo 2 nos traz conceitos sobre o desenvolvimento do comércio eletrônico, suas diferentes formas e principais competências envolvidas em sua operação. Além disso nos mostra quais são as mais importantes funcionalidades e limitações que podem ter influência em uma decisão de investimento.

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
Departamento de Engenharia de Produção - Trabalho de Formatura

Apresentar um perfil dos usuários de Internet, buscando verificar qual o potencial desse mercado que possui o comércio eletrônico.	Ainda no capítulo 2, é mostrada uma pesquisa sobre os perfil dos usuários que nos traz conclusões importantes a respeito do comportamento de compra desses usuários, poder aquisitivo, segmento de mercado mais influenciado, etc. Um estudo mais aprofundado sobre técnicas de gestão de mercado seria de grande valor para uma possível investida no setor de comércio eletrônico
Mostrar quais os conceitos técnicos envolvidos nas operações realizadas via Internet, incluindo arquiteturas Cliente/Servidor, fluxo de uma transação Web e os diferentes recursos envolvidos no sistema.	No capítulo 3, são introduzidos os conceitos sobre performance em sistemas e de que maneira cada recurso pertencente ao sistema analisado pode ter influência no resultado global apurado. Existem outros conceitos envolvidos neste tema que, por apresentar um caráter extremamente técnico, estariam fora do escopo deste trabalho. Na bibliografia existem obras dedicadas exclusivamente a conceitos de performance e que podem trazer maiores informações aos interessados.
Apresentar uma metodologia de planejamento de capacidade, com uma seqüência de etapas a serem seguidas, de modo a cercar todos os parâmetros envolvidos na sua execução. Aplicar esta metodologia a um caso real.	Foi apresentado o MPC - Modelo de Planejamento de Capacidade, que possui tarefas bastante definidas que levam a uma estimativa sobre a capacidade necessária a ser instalada, para atender de forma satisfatória os clientes. Alguns passos não puderam ser realizados da melhor forma, devido a algumas limitações de aplicação. Esses passos, no entanto, não tiveram influência significativa no resultado geral do modelo.

Tabela 5.1 - Alcance das Metas Propostas

Vemos que, de forma geral, as metas puderam ser atingidas com sucesso, algumas podendo apresentar algum tipo de aprofundamento, outras apresentando limitações que não tiveram influência significativa no resultado global obtido.

A revisão bibliográfica foi mostrada de forma a evidenciar a importância que o comércio eletrônico têm, e terá, no mercado de vendas e ainda mostrar que o potencial de consumo é bastante interessante nesta área.

A parte prática do trabalho, focada na aplicação no MPC a um caso real, teve como objetivo passar uma idéia de como o processamento de transações via Internet se adapta às especificações de performance passadas pelos usuários. Nesta etapa foram feitas considerações sobre os requisitos de infraestrutura necessários para que o site tenha uma capacidade adequada de atendimento e sobre o tempo de processamento de uma transação. Alguns recursos do sistema, como, por exemplo, os roteadores e firewalls não foram incluídos no estudo, dada dificuldade de análise de seus tempos de processamento e impacto pouco significativo no resultado final. Em uma análise mais elaborada, estes elementos devem ser considerados, já que problemas podem ocorrer na análise caso algum destes elementos venha a apresentar condições anormais de operação.

5.3 Estudo da contribuição prática deste projeto

Cada vez mais torna-se indispensável um conhecimento aprofundado sobre assuntos relacionados à Internet. Atualmente, grande parte das transações bancárias são realizadas via Internet e, num futuro próximo, grande parte das transações comerciais, incluindo-se as realizadas entre pessoas físicas e entre pessoas jurídicas, terá seu modo de operar ligado à Internet.

Para o engenheiro de produção, é indispensável estar atualizado sobre novas maneiras de se realizar negócios, novas tecnologias, novas formas de gestão, enfim, sobre as tendências de seu ramo de atuação e as implicações a elas atreladas.

Uma das grandes contribuições práticas deste trabalho diz respeito a essa introdução a novas tecnologias. A revisão bibliográfica e os conceitos de performance em sistemas tiveram como objetivo, além de proporcionar um conhecimento teórico sobre a parte técnica relativa ao assunto, uma visão genérica sobre quais os principais componentes de mercado ligados à área.

Um aprofundamento sobre este relacionamento entre novas tecnologias e o mercado a que se destinam poderia ser obtido através de estudos de marketing ligados à parte de sistemas, considerando questões como, por exemplo, de que forma parametrizar um sistema que possa tratar cada cliente de maneira isolada para que possam retornar ao site e realizar novas compras, quais os produtos que mais se encaixam no perfil do segmento de mercado A ou B, etc. Já existe uma área de atuação específica para o relacionamento com o cliente, chamada de CRM (Customer Relationship Management), onde vários trabalhos, incluindo softwares dedicados a este assunto, vêm sendo desenvolvidos com bastante sucesso.

Uma outra contribuição deste trabalho, talvez a sua principal, e que possui caráter bastante prático, está ligada à elaboração de uma metodologia de trabalho que possa cercar grande parte dos passos relacionados ao Planejamento de Capacidade de um Web Site.

Para o engenheiro, é também muito importante que se tenha uma metodologia de trabalho que possa ter aplicação fácil, eficiente e que traga resultados importantes.

O modelo aqui apresentado foi aplicado para apenas um caso possível de transação de negócio realizada via Internet. Existem muitos outros tipos de operações que podem ser realizadas e, principalmente pelo fato de não podermos fazer um simples somatório dos tempos de processamento das transações, seria bastante interessante realizar um estudo sobre qual a melhor forma de se otimizar um sistema que possua variedade de transações, incluindo prioridades, filas de execução de tarefas, etc.

BIBLIOGRAFIA

GRAY, J [et al.]. Transaction Processing: Concepts and Techniques. São Francisco. Morgan Kaufmann, 1993.

HAMILTON, S., E-Commerce for the 21st Century, IEEEComputer Soc.,1997.

IBOPE INTERACT. 4^a Pesquisa Cadê/Ibope Perfil de Usuários de Internet, www.ibope.com.br, 1999.

KX - Knowledge Exchange. Banco de Dados de Informações da Andersen Consulting do Brasil, São Paulo, 2000.

MENASCÉ, D. [et al.]. Capacity Planning for Web Performance: metrics, models and methods. New Jersey, Prentice-Hall, 1998.

NETO, P.R.B., Aplicações de Modelos Probabilísticos da Pesquisa Operacional. São Paulo, Departamento de Engenharia de Produção, Escola Politécnica, USP.

PERLMAN, R., Interconnections. Reading, Addison-Wesley 1994.

SANTORO, M., Planejamento, Programação e Controle da Produção. São Paulo, 1999. Departamento de Engenharia de Produção, Escola Politécnica, USP.

STEVENS, W., TCP/IP Illustrated, Vol 3. Reading, Addison-Wesley 1996.

TURBAN, E. [et al.]. Electronic Commerce: a managerial perspective. New Jersey, Prentice-Hall, 2000.