

UNIVERSIDADE ESTADUAL PAULISTA

Campus de Ilha Solteira

Faculdade de Engenharia de Ilha Solteira

João Marcelo Rondina

Sistema de Controle Fuzzy adaptado ao fenômeno de rajadas para
construção de um modelo de carga de trabalho de Serviços Web

Ilha Solteira (SP)

2005

João Marcelo Rondina

Sistema de Controle Fuzzy adaptado ao fenômeno de rajadas para
construção de um modelo de carga de trabalho de Serviços Web

Orientador: Prof. Dr. Aleardo Manacero Jr.

Dissertação de Mestrado elaborada junto ao
Programa de Pós Graduação em Engenharia
Elétrica – Área de Concentração em Sistemas de
Energia Elétrica, como parte dos pré-requisitos
para obtenção do Título de Mestre em
Engenharia Elétrica

Ilha Solteira (SP)

2005

CERTIFICADO DE APROVAÇÃO

TÍTULO: Sistema de Controle Fuzzy Adaptado ao Fenômeno de Rajadas para Construção de um Modelo de Carga de Trabalho em Serviços WEB

AUTOR: JOÃO MARCELO RONDINA

ORIENTADOR: Prof. Dr. ALEARDO MANACERO JUNIOR

Aprovado como parte das exigências para obtenção do Título de MESTRE em ENGENHARIA ELÉTRICA pela Comissão Examinadora:

Prof. Dr. ALEARDO MANACERO JUNIOR

Departamento de Ciênc. Comp. e Estatística / Instituto de Biociências, Letras e Ciências Exatas de São José do Rio Preto

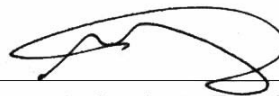
Prof. Dr. CARLOS ROBERTO MINUSSI

Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Prof. Dr. MARCOS LUIZ MUCHERONI

Centro Universitário de Marília - UNIVEM

Data da realização: 21 de dezembro de 2005.



Presidente da Comissão Examinadora
Prof. Dr. ALEARDO MANACERO JUNIOR

Dedico esse trabalho à minha esposa, Ana Maria
e à minha filha Natália.

Agradecimentos

Ao meu orientador, Prof. Aleardo, agradeço pela paciência e dedicação, pelo modo com que soube equilibrar positivamente minhas limitações e potenciais, e assim me conduzir e auxiliar durante todo o caminho trilhado no mestrado. Pode contar sempre com minha gratidão, respeito e amizade.

Aos meus pais, João e Cida, por ter me criado com muito amor e mostrado um exemplo a seguir. Ao meu irmão, Juliano, minha admiração e afeto.

À Dona Cida e Andréa, pela dedicação e carinho dado à minha filha, sempre, em todos esses anos.

Aos meus colegas da FAMERP, Wellington, Landim e Rildo.

À diretoria da FAMERP, em especial ao Dr. João Aris, Dr. José Victor, Dr. Humberto e Dr. Acayaba, pelo apoio.

Ao Kleber, da FAMERP, pela compreensão e apoio durante o mestrado.

Às coordenadoras dos cursos de pós da FAMERP: Elza, Margarete, Neuseli e Maysa, por ter acreditado em meu trabalho.

À Elizabeth Somera, pela pareceria, orientação e apoio nas aulas de didática.

Ao meu amigo Igor, pelos conselhos, brincadeiras, conversas. Inestimável ajuda nas apresentações, e em todos os momentos difíceis durante o mestrado.

Ao meu amigo Ivan, pelo incentivo e compreensão.

Ao meu amigo Luís Ricardo, pelo voto de confiança, há muito tempo atrás.

Aos colegas da FEIS: Coffani, Tiago e Osvaldo.

A todos os colegas do GSPD do IBILCE.

Aos professores e funcionários da FEIS. Em especial, aos professores Marcelo e Aparecido, pelo trabalho realizado na coordenação da pós.

Aos funcionários do IBILCE, em especial a Olga, da secretaria de graduação da computação.

Aos meus alunos do curso de Medicina da FAMERP.

Ao saudoso prof. Odelar e à querida profa. Maria Augusta, do IBILCE.

Epígrafe



Desenho "Astronomia - Olhos Curiosos Voltados ao Céu", de livro alemão sem data.

"...quando não se pode medir, o conhecimento é frágil e insatisfatório".
Lord Kelvin

"A idéia de que planejar significa adivinhar o futuro é simplesmente absurda."
Peter Drucker

SUMÁRIO

1	INTRODUÇÃO.....	1
1.1	CONTEXTUALIZAÇÃO E MOTIVAÇÃO	1
1.2	OBJETIVOS	1
1.3	DESCRIÇÃO DO TRABALHO	2
2	PLANEJAMENTO DE CAPACIDADE.....	4
2.1	CONCEITOS FUNDAMENTAIS	4
2.1.1	<i>Planejamento de Capacidade Para Serviços Web.....</i>	<i>5</i>
2.1.2	<i>A Atividade de Gestão do Desempenho.....</i>	<i>6</i>
2.1.3	<i>Expectativa dos Clientes de Serviços Web</i>	<i>7</i>
2.1.4	<i>Comportamento do Cliente.....</i>	<i>8</i>
2.2	DEFINIÇÃO DE CAPACIDADE ADEQUADA.....	9
2.2.1	<i>Frameworks Para Especificação de Ambientes de Produção.....</i>	<i>11</i>
2.2.2	<i>Ferramentas para Análise do Ambiente de Produção.....</i>	<i>14</i>
2.2.3	<i>Acordos de Nível de Serviço (ANS)</i>	<i>15</i>
2.3	FRAMEWORKS PARA PLANEJAMENTO DE CAPACIDADE PARA SERVIÇOS WEB.....	17
2.3.1	<i>Miopia em Planejamento de Capacidade.....</i>	<i>17</i>
2.3.2	<i>Planejamento de Capacidade de Guerrilha</i>	<i>18</i>
2.3.3	<i>Bancarrota dos Servidores Web</i>	<i>19</i>
2.3.4	<i>Escalabilidade Super Serializada.....</i>	<i>20</i>
2.3.5	<i>Modelos de Custo</i>	<i>21</i>
2.3.6	<i>O Framework para Planejamento de Capacidade de Menascé</i>	<i>26</i>
2.4	CARGAS DE TRABALHO DE SERVIÇOS WEB.....	29
2.4.1	<i>Ocorrência do Fenômeno de Rajadas em Cargas de Trabalho</i>	<i>30</i>
3	MODELOS PARA REPRESENTAÇÃO DA CARGA DE TRABALHO COM O FENÔMENO DE RAJADAS	33
3.1	ABORDAGEM OPERACIONAL PARA O FENÔMENO DE RAJADAS	33
3.1.1	<i>Algoritmo para Cálculo dos Parâmetros a e b.....</i>	<i>35</i>
3.1.2	<i>Leis de Utilização do Recurso e o Fenômeno de Rajadas.....</i>	<i>36</i>
3.1.3	<i>Algoritmo para Cálculo do Coeficiente Proporcional de Rajada a.....</i>	<i>38</i>
3.2	ANÁLISE DE MÚLTIPLAS ESCALAS DE TEMPO.....	39
3.3	PREVISÃO DA CARGA DE TRABALHO FUTURA	40
4	LÓGICA FUZZY NO PLANEJAMENTO DE CAPACIDADE PARA SERVIÇOS WEB....	44
4.1	LÓGICA FUZZY	46
4.2	CONJUNTOS FUZZY	48
4.3	SISTEMAS DE CONTROLE FUZZY	48
4.3.1	<i>Base de Conhecimento.....</i>	<i>49</i>
4.3.2	<i>O Processo de Fuzzificação e as Variáveis Fuzzy.....</i>	<i>50</i>
4.3.3	<i>Inferência em Lógica Fuzzy.....</i>	<i>51</i>
4.3.4	<i>Defuzzificação: Método do Centro de Área.....</i>	<i>53</i>
4.4	CONSIDERAÇÕES FINAIS	54
5	CONTROLE FUZZY PARA MODELAGEM DE CARGA EM SERVIÇOS WEB.....	55
5.1	FUNDAMENTAÇÃO GERAL	55
5.2	SISTEMA DE CONTROLE FUZZY RAJIN.....	57
5.2.1	<i>Definição das Variáveis de Entrada e Saída.....</i>	<i>58</i>
5.2.2	<i>Base de Dados e de Conhecimento.....</i>	<i>60</i>
5.3	CONTROLE FUZZY CSWEB.....	62
5.3.1	<i>Variáveis de Entrada e de Saída</i>	<i>62</i>
5.3.2	<i>Base de Dados e de Conhecimento.....</i>	<i>65</i>
5.4	IMPLEMENTAÇÃO COMPUTACIONAL E CONSIDERAÇÕES FINAIS	66
6	ANÁLISE DE RESULTADOS	69
6.1	AMBIENTE DE TESTE	69
6.1.1	<i>Instrumentação dos Registros Históricos.....</i>	<i>71</i>

6.1.2	<i>Registro do Tempo de Uso da CPU</i>	73
6.2	DEFINIÇÃO DOS PADRÕES DE ANÁLISE DOS REGISTROS HISTÓRICOS	74
6.2.1	<i>Padrão dos Dados Históricos</i>	75
6.2.2	<i>Conclusões Sobre a Análise do Padrão dos Dados Históricos</i>	78
6.3	OCORRÊNCIA DO FENÔMENO DE RAJADAS.....	80
6.3.1	<i>Análise de Rajadas no Serviço GA</i>	82
6.3.2	<i>Análise de Rajadas no Serviço Proxy</i>	84
6.3.3	<i>Conclusões Sobre a Análise da Ocorrência do Fenômeno de Rajadas</i>	87
6.3.4	<i>Utilização do Serviço Web</i>	87
6.4	MODELO DE CARGA DE TRABALHO FUZZY	89
6.4.1	<i>Intensidade de Rajada – Controle RAJIN</i>	90
6.4.2	<i>Utilização do Serviço Web – Controle CSWeb</i>	92
6.5	CONSIDERAÇÕES FINAIS	96
7	CONCLUSÃO	97
7.1	TRABALHOS FUTUROS	99
	REFERÊNCIAS BIBLIOGRÁFICAS	101

LISTA DE FIGURAS

Figura 2.1 – Gerenciamento da Desempenho do Serviço Web.....	6
Figura 2.2 - CBMG - Comportamento do Cliente.....	9
Figura 2.3 - Definição da capacidade adequada em um Serviço Web.....	10
Figura 2.4 - FACPS - Gerenciamento de TI - área de suporte.....	12
Figura 2.5 - Gráfico de utilização de servidor Web – tps.....	14
Figura 2.6 - O “Homúnculo Desempenho”, de Gunther.....	17
Figura 2.7 - Lei da Bancarrota de Servidores (Gunther).....	19
Figura 2.8 – Curva Negra de Gunther.....	20
Figura 2.9- <i>Framework</i> para o Planejamento de Capacidade de Menascé.....	28
Figura 2.10 – Ocorrência do fenômeno de rajadas em cargas de trabalho.....	31
Figura 3.1 – Efeito da intensidade da rajada na capacidade do serviço Web.....	35
Figura 3.2 - Gráficos com os 4 padrões de dados históricos.....	42
Figura 4.1 - Sistema de Controle Fuzzy.....	49
Figura 4.2 - Forma triangular.....	51
Figura 4.3 - Forma Trapezoidal.....	51
Figura 4.4 - Método do Centro de Área.....	54
Figura 5.1 – Controle Fuzzy RAJIN e CSWeb.....	55
Figura 5.2 - Variável de Entrada FRAJA - Parâmetro a.....	59
Figura 5.3 - Variável de Entrada FRAJB - Parâmetro b.....	59
Figura 5.4 - Variável de Saída – RAJ – Intensidade do Fenômeno de Rajadas.....	60
Figura 5.5 – RAJ: Intensidade do Fenômeno de Rajadas.....	64
Figura 5.6 – NHTTP: Requisições http.....	64
Figura 5.7 – Variável de Saída - USW – Utilização do Serviço Web.....	64
Figura 5.8 – Implementação da variável de entrada NHTTP.....	67
Figura 5.9 – Algoritmos do controle fuzzy CSWeb.....	68
Figura 6.1 - Serviços Web analisados - GA e Proxy.....	70
Figura 6.2 - Serviço GA - requisições anuais.....	75
Figura 6.3 - Serviço GA - Requisições http 2004 e 2005.....	76
Figura 6.4 – Requisições http do serviço Proxy.....	77
Figura 6.5 – Requisições http diárias do serviço Proxy.....	78
Figura 6.6 – Serviço GA: requisições http em 26/04 com escala de tempo de minuto.....	82
Figura 6.7- Serviço GA: requisições http em 26/04 com escala de tempo de 1 hora.....	83
Figura 6.8 – Serviço Proxy: requisições http - 14/04 - escala de tempo - 1 hora.....	84
Figura 6.9 - Serviço Proxy: requisições http - 14/04 - escala de tempo: 10 minutos.....	85
Figura 6.10 - Serviço Proxy: requisições http - 11/04 - escala de tempo: 10 minutos.....	86
Figura 6.11 - Serviço Proxy: Requisições http - 11/04 - escala de tempo: 10 minutos.....	86
Figura 6.12 - Ocorrência do Fenômeno de Rajadas.....	91
Figura 6.13 – Processo de defuzzificação - Serviço Web Proxy (11 de Abril).....	94
Figura 6.14 - Processo de defuzzificação - Serviço Web Proxy - 14 de Abril.....	94
Figura 6.15 – Comparação de resultados entre o modelo operacional e fuzzy.....	95

LISTA DE TABELAS

Tabela 2.1 - Exemplo de Planilha para funcionalidades do sistema..	24
Tabela 2.2 - Planilha de Desempenho.	25
Tabela 2.3 - Planilha de Custos.	26
Tabela 3.1 - Técnicas de previsão - enfoque qualitativo e quantitativo.	41
Tabela 4.1- Exemplo de inferência fuzzy	52
Tabela 4.2 - Método de Mamdani	53
Tabela 5.1 - Operadores utilizados nos Controles Fuzzy.	57
Tabela 5.2 - Variáveis de entrada e saída – Controle Fuzzy Rajin.	58
Tabela 5.3 - Base de Conhecimento (Regras e Fatos) - Controle Fuzzy Rajin.	61
Tabela 5.4 - Matriz Regras de Inferência - Controle Fuzzy Rajin.	61
Tabela 5.5 - Variáveis de entrada e saída – Controle Fuzzy CSWeb.	63
Tabela 5.6 - Base de Conhecimento – Controle Fuzzy CSWeb.	65
Tabela 5.7 - Matriz Regras de Inferência - Controle Fuzzy CSWeb.	66
Tabela 6.1 - Configuração de hardware dos serviços Web.	71
Tabela 6.2 - Funcionalidades do serviço GA.	72
Tabela 6.3 - Análise dos Serviços Web.	74
Tabela 6.4 - Fator de rajada da carga de trabalho.	81
Tabela 6.5 – Utilização do Serviço Web.	89
Tabela 6.6 - Resultados entrada e saída controle fuzzy Rajin.	90
Tabela 6.7 – Cálculo de Utilização do Serviço – Abordagem Operacional e Fuzzy	93

Resumo

O objetivo do planejamento de capacidade dos serviços Web é prover sua qualidade. Conhecer a carga de trabalho é um requisito necessário para esse processo de planejamento. A caracterização da carga de trabalho permite produzir um modelo representativo e simples do comportamento da carga. Uma das características da carga de trabalho analisadas, nesse estudo, é a ocorrência do fenômeno de rajadas e seu impacto na capacidade do serviço. Nesse contexto, são relatadas duas abordagens para construção de modelos de carga de trabalho adaptados a esse fenômeno. Um modelo determinístico registra resultados precisos, baseados na técnica da abordagem operacional. Como alternativa, esse estudo propõe um modelo baseado em um sistema de controle fuzzy, que retorna resultados aproximados, porém em menor tempo e sem a necessidade de utilizar registros históricos. Os modelos são utilizados para analisar as cargas de trabalho de dois serviços Web e os resultados obtidos são comparados, descrevendo as vantagens e desvantagens de cada solução.

PALAVRAS-CHAVE: planejamento de capacidade, serviços web, carga de trabalho, fenômeno de rajadas, sistemas de controle fuzzy, abordagem operacional.

Abstract

The objective of the capacity planning of the Web service is to provide its quality. To know the workload is a necessary requirement for this process of planning. The characterization of the workload allows this knowledge and produces a representative and simple model of its behavior. One of the characteristics of the workload analyzed in this study is the occurrence of the burst phenomenon and its impact in the service capacity. In this setting, two approaches for construction of workload models to this phenomenon are reported. A deterministic model records accurate results, based on operational approach technique. Proposing as an alternative, this study considers a model based on a fuzzy control system, which returns approximately results, but in more less time and without the necessity you uses historical data. The models are used to analyze the workload of two Web services, and the results are compared, describing both the advantages and disadvantages of each solution.

KEYWORDS: Capacity planning, Web services, workload, burstiness, fuzzy control systems, operational analysis.

1 Introdução

1.1 Contextualização e Motivação

O planejamento de capacidade dos serviços Web é fundamental para assegurar a sua qualidade, disponibilidade e continuidade. Esse planejamento é realizado com base na análise de diversos aspectos do serviço, como carga de trabalho, desempenho, custo e disponibilidade.

Esse estudo descreve a carga de trabalho dos serviços Web, que é formada por componentes complexos e heterogêneos, os quais dificultam o correto entendimento e caracterização. É fundamental, no planejamento de capacidade para serviços Web, construir um modelo simples e representativo da carga de trabalho.

Nas cargas de trabalho dos serviços Web existem características comuns, chamadas de invariantes, que devem ser representadas no processo de construção do modelo. No conjunto de características invariantes da carga de trabalho, destaca-se a ocorrência do fenômeno de rajadas, por causar gargalos de desempenho do serviço em ocasiões de difícil previsibilidade.

Nesse contexto, os modelos determinísticos são utilizados no processo de caracterização de carga de trabalho para representar o fenômeno de rajadas. A implementação desses modelos depende da existência dos serviços Web e da disponibilidade de seus registros históricos, muitas vezes não padronizados ou incorretamente armazenados. O tempo para processamento dos dados é também outro fator que agrega complexidade à construção desses modelos.

1.2 Objetivos

O objetivo principal desse trabalho é apresentar um sistema de controle fuzzy adaptado ao fenômeno de rajadas e demonstrar sua eficiência em comparação aos modelos determinísticos comumente utilizados, como a abordagem operacional. Espera-se que os resultados obtidos pelo controle fuzzy sejam próximos dos gerados pelos modelos determinísticos

Como diferencial, o controle fuzzy deve produzir as análises da carga de trabalho sem necessitar dos registros históricos, possibilitando sua utilização para construção de estudos de viabilidade na implementação de serviços Web. Também é necessário verificar se os resultados são produzidos em um tempo menor do que o apresentado pelo modelo determinístico.

1.3 Descrição do Trabalho

O capítulo 2 apresenta um levantamento bibliográfico sobre planejamento de capacidade. São apresentados conceitos sobre capacidade adequada, acordos de nível de serviço, ferramentas para análise do ambiente de produção, modelos de custos e comportamento do cliente. Ainda nesse capítulo destaca-se uma seleção de *frameworks* que podem ser utilizados para o planejamento de capacidade.

O capítulo 3 descreve um modelo de carga de trabalho adaptado ao fenômeno de rajadas, implementado por uma técnica chamada abordagem operacional, que permite a análise quantitativa dos registros históricos dos serviços Web, obtendo informações como intensidade e gargalos de desempenho causados por esse fenômeno. Nesse capítulo também são relacionadas outras características importantes na construção do modelo, como a escala de tempo analisada e a evolução da carga de trabalho.

O capítulo 4 apresenta uma definição de lógica fuzzy e sua aplicação em soluções para planejamento de capacidade para serviços Web. São descritos diversos elementos que fazem parte das soluções fuzzy: operadores, variáveis, conjuntos, quantificadores, universo de discurso, formas, defuzzificação, dentre outros.

O capítulo 5 apresenta uma documentação completa sobre as variáveis de entrada e de saída dos controles, operadores de fuzzificação e defuzzificação utilizados nos processos de inferência, regras e fatos presentes na base de conhecimento. Apresenta-se uma solução para construção do modelo fuzzy adaptado ao fenômeno de rajadas, composta por dois controles, denominados RAJIN e CSWeb, desenvolvidos nesse trabalho, servindo respectivamente para o cálculo da intensidade da

ocorrência do fenômeno de rajadas em cargas de trabalho de serviços web e seu impacto na utilização dos mesmos.

No capítulo 6, apresentam-se os resultados obtidos na análise das cargas de trabalho de dois serviços Web em funcionamento em uma Instituição de Ensino. A análise da ocorrência do fenômeno de rajadas é verificada e medida, utilizando a abordagem operacional. O padrão dos dados históricos dos serviços é estudado e descrito na evolução da carga de trabalho. Os resultados obtidos utilizando o controle fuzzy também estão presentes nesse capítulo, que é encerrado com uma comparação do resultado obtido pelas duas soluções.

O capítulo 7 oferece as conclusões sobre os resultados obtidos no estudo, além de algumas recomendações para trabalhos e pesquisas futuras.

2 Planejamento de Capacidade

A principal missão do planejamento de capacidade dos serviços Web é evitar que aconteça o momento de saturação, conhecido como gargalo de desempenho, fornecendo alternativas que indiquem o que deve ser feito para que esse problema nunca ocorra. Desse processo de planejamento, espera-se a predição do comportamento futuro do serviço em resposta a mudanças do comportamento do cliente, aumento de funcionalidades oferecidas e evolução natural da demanda.

Esse capítulo apresenta modelos para custos, definição do comportamento do cliente, de acordos de nível de serviço e da capacidade adequada. *Frameworks* para o planejamento de capacidade também são descritos, mostrando diversas abordagens diferentes para a construção do plano.

2.1 Conceitos Fundamentais

O conceito de Planejamento de Capacidade não é novo. Tem suas origens nos ambientes computacionais do Mainframe, quando os responsáveis pelo planejamento de capacidade se deparavam com situações, na maioria das vezes, muito bem definidas. Nesse contexto, encontram-se relatos de características como picos de utilização de CPU em dois períodos do dia (10 da manhã e 2 da tarde), que por isto receberam a alcunha “10-2” (GUNTHER, 2001).

A maior preocupação do responsável pelo Planejamento de Capacidade é evitar a aquisição de novo hardware pelo maior período de tempo possível, devido a seu alto custo. Um exemplo dessa situação ocorreu em 1.964, quando a IBM introduziu um conceito de arquitetura com escalabilidade que norteou toda a indústria dali por diante, lançando a sua linha de computadores System/360. Estava feita a promessa de uma continuidade nos investimentos em uma arquitetura de computadores Mainframe: você comprava hoje um e poderia comprar, num segundo momento, uma atualização ou um outro computador para acomodar o aumento de carga de trabalho de um serviço.

2.1.1 Planejamento de Capacidade Para Serviços Web

Desde o início da indústria de computadores, busca-se a redução de tamanho dos circuitos eletrônicos digitais. Em 1.965, um dos fundadores da Intel, Gordon Moore, registrou essa notável tendência tecnológica através da chamada *Lei de Moore*, que diz que a quantidade de transistores armazenados que um microchip dobra a cada dezoito meses (THING, 2003).

A revolução do computador pessoal, com o barateamento do hardware, teve repercussão também nos métodos tradicionais de planejamento de capacidade para o Mainframe, centrados em hardware. Esses métodos não eram mais adequados para essa nova realidade (BRETON, 1991).

O cenário atual é caracterizado pela alta complexidade dos sistemas distribuídos. Nesses ambientes encontramos diversos tipos de sistemas operacionais (Linux, NT, Unix), servidores Web, monitores de transação e sistemas gerenciadores de bancos de dados (GUNTHER, 2003); (GUNTHER, 2002); (MENASCÉ; ALMEIDA, 2000a). Além dos ambientes cliente-servidor, clusters e grades computacionais também começam a se tornar, hoje, uma realidade no meio acadêmico e empresarial (FORTES, 2004). Nesse contexto, uma particularidade dos atuais serviços Web é a não existência de um padrão para os aplicativos que compõe seu ambiente cliente-servidor (COCKCROFT; WALKER 2001).

Nesses ambientes, o foco da atuação do planejamento de capacidade está na predição dos gargalos causados pelos aplicativos e suas inter-relações com a capacidade de rede e a arquitetura cliente-servidor dos serviços Web, e não mais tão somente no hardware. O responsável pelo planejamento de capacidade terá de desenvolver modelos para a predição da reação dos serviços Web quando mudanças em três fatores ocorrerem (MENASCÉ; ALMEIDA, 2000a):

1. Evolução natural da carga de trabalho atual do serviço Web
2. Desenvolvimento de novas funcionalidades para o serviço Web
3. Mudanças no comportamento do usuário que utiliza o serviço Web.

2.1.2 A Atividade de Gestão do Desempenho

Alguns autores, como Adrian Crockcroft, situam o planejamento de capacidade como um dos componentes de uma atividade maior, chamada de Gerenciamento de Desempenho. A figura 2.1 mostra esse processo: uma Linha Base recolhe informações, através de medições sobre o desempenho atual do serviço Web. Esses dados são analisados para a produção de um modelo, ou plano, que é utilizado em todo o processo de planejamento de capacidade. Em seguida, o planejamento da capacidade para o serviço Web é implementado e diversos processos de gerenciamento dos recursos de suporte ao serviço Web são implementados ou ajustados. Esse ciclo se repete continuamente para garantir o correto desempenho do serviço Web (COCKCROFT; WALKER 2001).

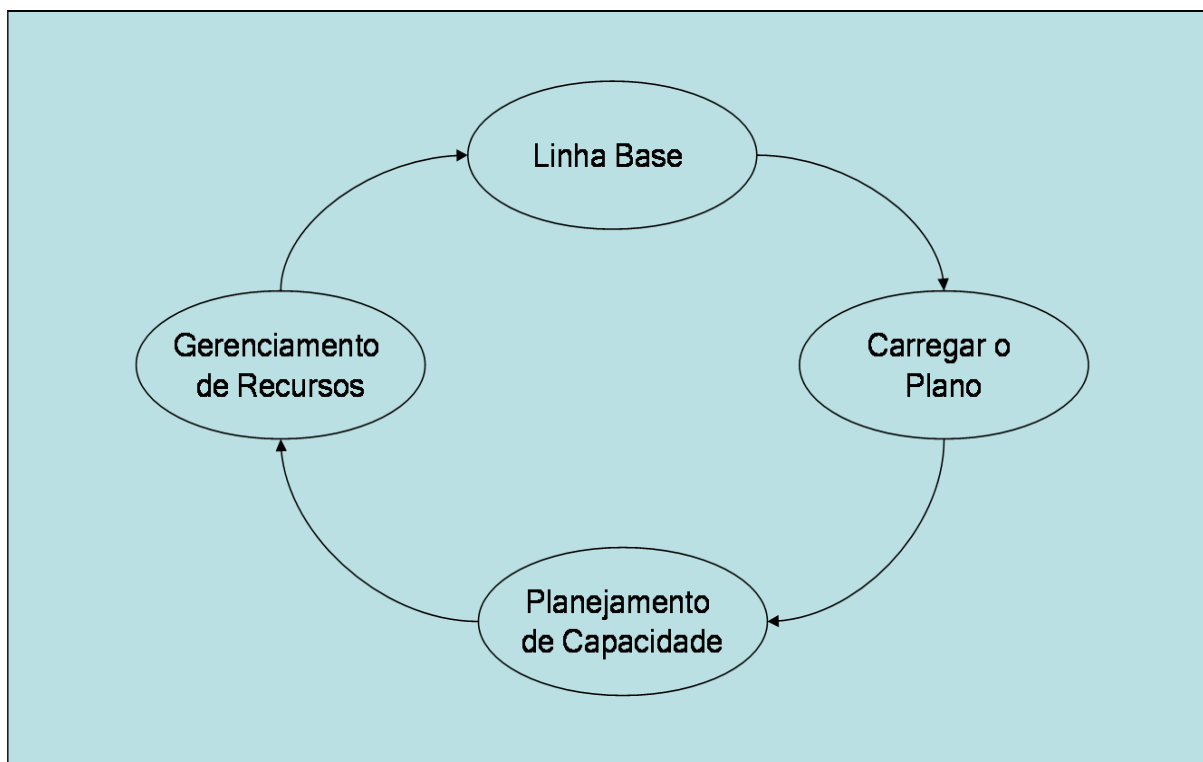


Figura 2.1 – Gerenciamento da Desempenho do Serviço Web.

A indústria de software está focada em ferramentas para gerenciamento e melhoria de desempenho dos diversos sistemas que formam o ambiente cliente servidor (BMC, IBM, HP, dentre outras). Muitos fabricantes de software também desenvolvem pacotes próprios para monitoramento e melhoria de desempenho de seus produtos (ORACLE, Microsoft, dentre outros). Essas

ferramentas de gerenciamento de desempenho atuam no estado atual de um serviço do ambiente cliente-servidor, coletando dados e fazendo análises que permitam “remediar” o gargalo no momento em que o mesmo ocorra (GUNTHER, 2002).

2.1.3 Expectativa dos Clientes de Serviços Web

O cliente de um Sítio da Internet espera velocidade no tempo de resposta, disponibilidade, privacidade das suas informações e segurança nas transações realizadas. Entretanto, a Internet possui características que podem ser definidas como imprevisíveis, como a taxa de requisição de transações por segundos, de um determinado serviço Web. Essas características podem afetar a qualidade do serviço Web e não atender as expectativas dos clientes, que podem não conseguir efetivar uma compra, por exemplo (CALZAROSSA et al., 2000); (MENASCÉ; ALMEIDA, 2002).

A disponibilidade de um serviço Web é apreciada pelos clientes. Os serviços Web, pela sua natureza, divulgam a seus clientes que estão disponíveis a qualquer momento, usando a escala “365x7x24” (365 dias ao ano, 7 dias por semana, 24 horas por dia). Uma das maiores expectativas do cliente é que esse padrão seja mantido. Fato recente na Internet brasileira, datado de Junho de 2004, foi a indisponibilidade, por algumas horas, do serviço Web de consulta à restituição do Imposto de Renda, fornecido pela Receita Federal (RECEITA, 2004).

Serviços Web que apresentem problemas como tempo resposta elevado ou períodos de indisponibilidade, podem causar diversos prejuízos às empresas. Os clientes perdem a motivação em utilizar o serviço Web e não retornam mais ao Sítio da Internet da empresa, causando perda de audiência. A interrupção do processo de compra antes de finalizá-lo é outra atitude comum por parte do cliente, causando perda de receita e conseqüente diminuição dos lucros. A marca da empresa, sua imagem junto a sociedade, pode também ser seriamente prejudicada em virtude desse serviço Web ineficiente: o cliente intui que todo o restante da empresa também é incapaz de atender às suas necessidades (SEYBOLD; MARSHAK, 2000).

2.1.4 Comportamento do Cliente

O comportamento do cliente de um serviço Web pode ser analisado e mapeado. Uma técnica para realizar essa análise do comportamento do cliente é a “CBMG”: Modelo de Grafo com o Comportamento do Cliente (ANDERBERG, 1973); (MENASCÉ; ALMEIDA, 2000a).

A análise CBMG do comportamento de um cliente, ao utilizar o serviço Web, pode ser usada também para predição de futuros cenários com situações de saturação do serviço (ARLITT; WILLIAMSON, 1996); (MENASCÉ; ALMEIDA, 2003b); (MENASCÉ; ALMEIDA, 2000a). O processo de análise CBMG pode ser definido no seguinte algoritmo:

1. São definidas e numeradas todas funcionalidades do serviço Web.
2. O arquivo de registros de utilização do servidor Web é analisado. Nesse arquivo, cada sessão de utilização de um cliente registra a frequência com que determinada funcionalidade do serviço Web foi acessada.
3. É feita uma relação entre a utilização das funcionalidades do serviço Web para estados do grafo da análise CBMG. As transições entre estados (funcionalidades) do grafo são analisadas pela frequência com que ocorreram, assim determinando um padrão de comportamento para cada cliente. São representadas todas as possíveis transições que tenham ocorrido.
4. Cada funcionalidade é representada com um estado no grafo e então são representadas todas que possam ocorrer, com base na maneira que o cliente utiliza essas funcionalidades.
5. Um grafo é produzido para uma determinada categoria. Esse grafo representa determinados padrões comuns a todos clientes dessa categoria. No final do processo da análise CBMG, para cada categoria de clientes, é produzido um grafo, com seu modelo de comportamento de utilização do serviço Web.

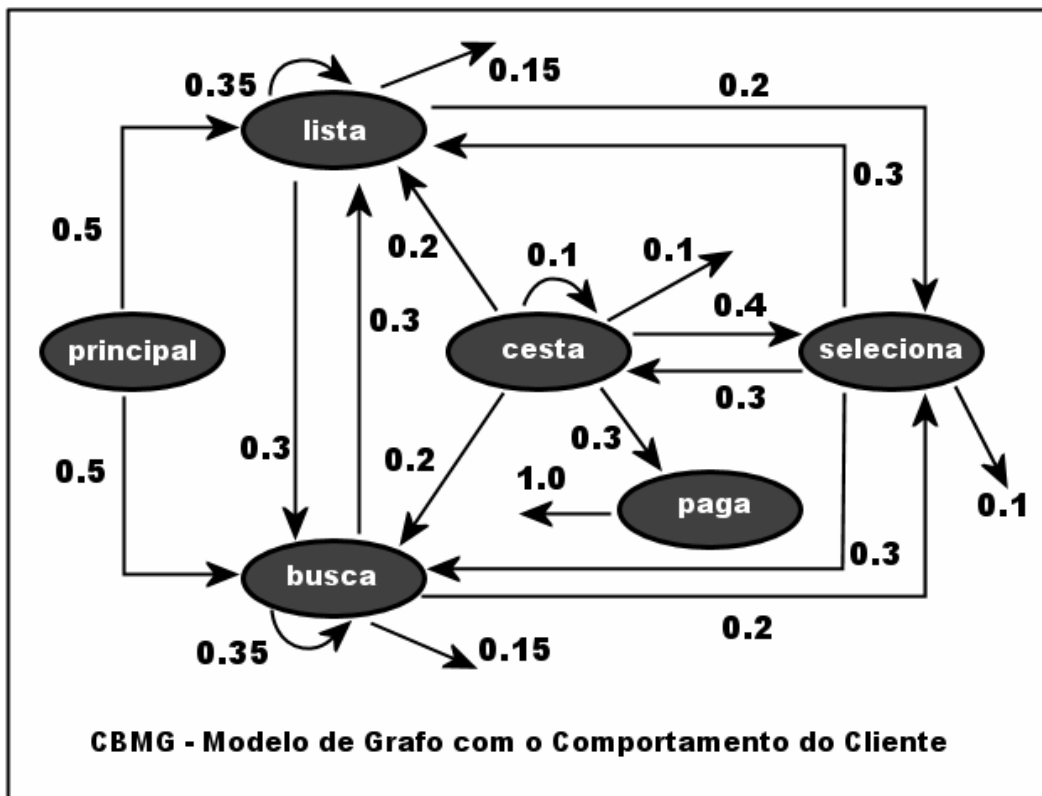


Figura 2.2 - CBMG - Comportamento do Cliente.

A figura 2.2 ilustra o grafo resultante de uma análise CBMG. Nesse exemplo, é registrado o comportamento de uma categoria de cliente de um serviço Web de comércio eletrônico. Cada estado do grafo representa uma determinada função: seleção de produtos, compras e pagamentos, busca de produtos, dentre outras. As transições entre estados e as frequências nas quais elas ocorrem também estão indicadas.

2.2 Definição de Capacidade Adequada

Para falar de planejamento de capacidade deve-se definir, claramente, o que é capacidade adequada. A capacidade adequada especifica qual desempenho esperar de um ambiente computacional cliente-servidor, dada uma carga de trabalho, e qual o custo para manter essa característica. Serve de indicador principal sobre o serviço Web para a gerência de Tecnologia da Informação (MENASCÉ; ALMEIDA, 2003b); (MENASCÉ; ALMEIDA, 2003a).

Pode-se usar três elementos, descritos a seguir, para definir a capacidade adequada em um ambiente de serviços Web. Na figura 2.3 esses elementos mostram como as necessidades de clientes e gerentes são relacionadas para a obtenção da capacidade adequada. Os elementos do modelo são:

1. “Acordos de Nível de Serviço”, que descrevem os acordos feitos entre clientes e gerência de TI.
2. “Restrições de Custo”, que definem valores e viabilidade de investimento para a implementação e manutenção do serviço Web.
3. “Tecnologias e Padrões”, que são os aplicativos e hardware disponíveis no ambiente de produção e serão utilizados para suportar os serviços Web.

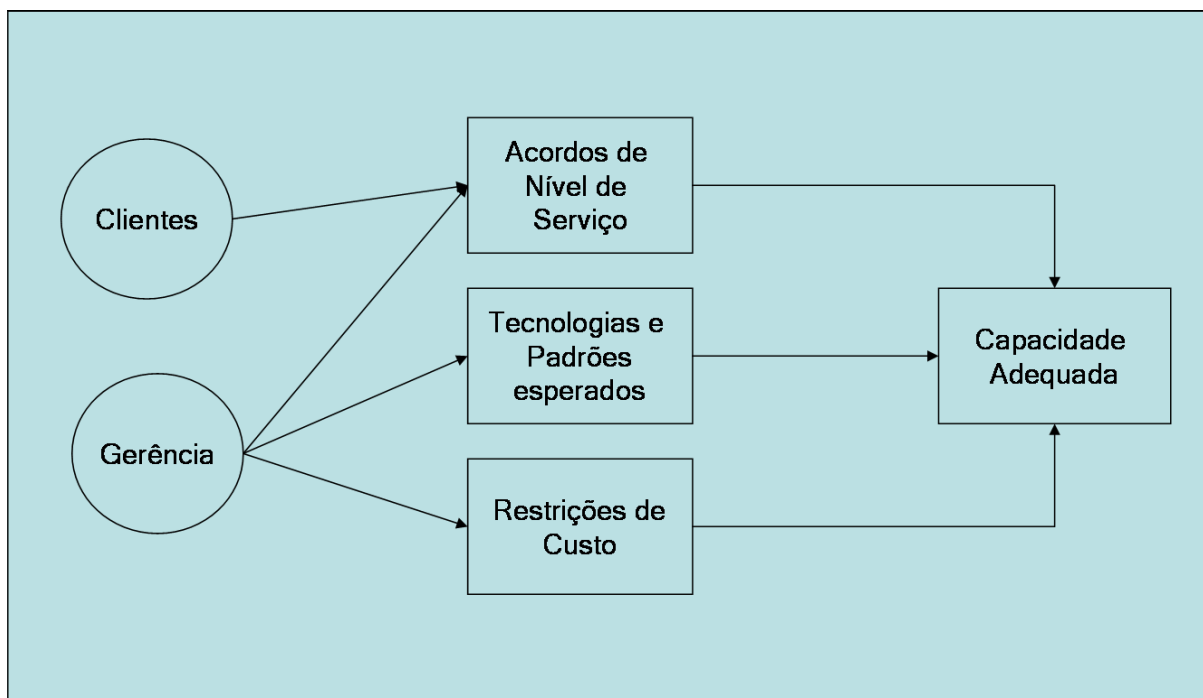


Figura 2.3 - Definição da capacidade adequada em um Serviço Web.

2.2.1 Frameworks Para Especificação de Ambientes de Produção

Para definir corretamente capacidade adequada, é necessário conhecer e especificar o ambiente onde o serviço Web será instalado. Esse ambiente apresenta diversas restrições a hardware e software¹ que podem ser utilizadas nos projetos para desenvolvimento de serviços Web. As restrições do ambiente de produção do serviço Web estão diretamente ligadas a recursos disponíveis para sua implementação, como contrato com determinado fabricante, opção por utilizar determinado sistema operacional, dentre outros, e afetam diretamente a capacidade adequada do serviço Web (MENASCÉ; ALMEIDA, 2003a); (MENASCÉ; ALMEIDA, 2003b).

Atualmente, existem diversos *frameworks* que fornecem modelos para definição de padrões de governança em TI. Todos eles possuem componentes para especificação de ambientes de produção, em particular Centros de Dados com estruturas que suportem serviços Web. O planejamento de capacidade para serviços Web é caracterizado com uma das funções cuja gestão é de responsabilidade de uma Área de Gestão de Desempenho (COCKCROFT; WALKER 2001). Alguns dos *frameworks* mais utilizados atualmente são:

- ISO FCAPS (*International Organization for Standardization – Fault Configuration Application Performance Security*)

A figura 2.4 mostra a proposta do ISO-FCAPS para a estrutura de Tecnologia da Informação. Através de uma hierarquia, definida por função, subfunção e atividades, fica possível determinar qual tarefa deva ser realizada dentro de uma estrutura bem definida. O planejamento de capacidade, em sua maior parte, fica restrito à subfunção “gestão de desempenho”. Algumas tarefas do planejamento de capacidade, entretanto, ficam atribuídas à subfunção “gestão de configuração”.

¹ Na definição de tecnologias e padrões de software, pode haver restrições a fabricantes de sistema operacional, servidores Web, aplicativos de desenvolvimento, sistemas de banco de dados, dentre outros. Muitas vezes, devido a contratos e políticas da própria empresa com esses fabricantes.

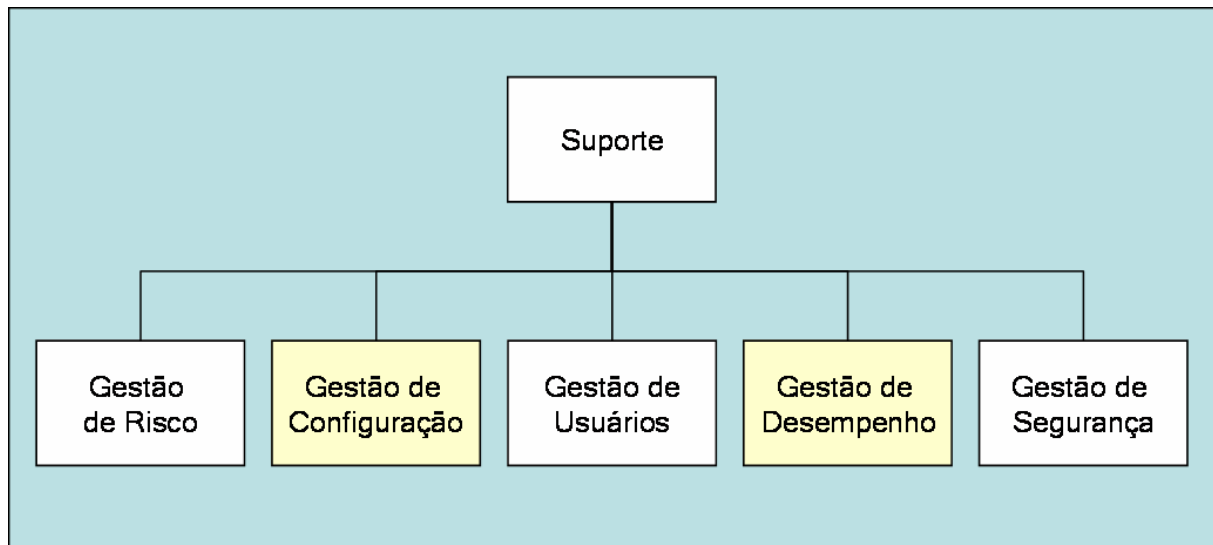


Figura 2.4 - FACPS - Gerenciamento de TI - área de suporte.

No ISO-FCAPS, a função “gestão de desempenho” possui, além do planejamento de capacidade, as seguintes subfunções de gerenciamento: desempenho de serviços, aplicações e computadores, definição de carga de trabalho e de ANS (acordos de níveis de serviço), dentre outros (COCKCROFT; WALKER 2001).

- ITIL (*Information Technology Information Libray*)

No framework ITIL, há uma metodologia para o gerenciamento de desempenho e de capacidade do ambiente de TI, que também passa pela especificação do ambiente de produção que dá suporte ao serviço Web, em cinco diferentes itens (MOLLOY, 2003):

1. Hardware – servidores que hospedam o serviço Web;
2. Software – sistemas operacionais, servidores Web e aplicações;
3. Periféricos – dispositivos de armazenamento e backup;
4. Equipamentos de Rede – Roteadores, Links WAN, links de Rede Local;
5. Recursos Humanos – somente as pessoas cuja responsabilidade ou função esteja diretamente associada a eventos que possam afetar os índices definidos no ANS (Acordos de Nível de Serviço) do Serviço Web;

Seguindo a recomendação da ITIL, o processo de gerenciamento de capacidade encampa algumas das seguintes atividades:

1. Monitoramento do desempenho dos serviços Web;
2. Constante otimização do desempenho para uso mais eficiente dos recursos existentes;
3. Entendimento das demandas atuais sendo atendidas pela área de TI;
4. Produção de um Plano de Capacidade que propicie à área de TI atingir e manter o padrão de qualidade especificado no ANS.

O uso do *framework* ITIL aliado ao gerenciamento de capacidade traz muitos os benefícios. Há o aumento de eficiência e economia de recursos, principalmente financeiros, pela diminuição de mudanças urgentes e desorganizadas. Também há a redução do risco de ocorrência de gargalos em aplicações, através de um efetivo planejamento de capacidade (MOLLOY, 2003).

- *Frameworks* proprietários

As grandes corporações de TI também desenvolveram *frameworks* para a descrição do ambiente de produção dos serviços Web, como o SEM (IBM) e SunReady Roadmap (Sun). Esses *frameworks* são fornecidos geralmente como pacotes de serviços de valor agregado às empresas que adquirem soluções de hardware e software desses fabricantes. O *framework* SRM (*Server Resource Management*), da IBM, provê métricas de itens como CPU, disco e memória, em períodos de tempo de dias, horas, meses. Esses dados podem ser usados para gerenciamento de desempenho e também para planejamento de capacidade (WEB, 2004).

2.2.2 Ferramentas para Análise do Ambiente de Produção

Uma maneira de analisar o ambiente de produção do serviço Web é determinar, graficamente, os períodos de utilização normal e de pico de utilização, através da análise do registro de acesso dos clientes presente no servidor Web. Nessa análise, geralmente diária, a quantidade de transações por segundo (tps) recebidas e atendidas pelo serviço Web é obtida em um cálculo da média da taxa de chegada de requisições, minuto a minuto. A análise do gráfico é rápida e facilita o entendimento da situação atual do serviço Web (GUNTHER, 2001); (MENASCÉ; ALMEIDA, 2003b); (MENASCÉ; ALMEIDA, 2000a).

A figura 2.5 exibe um gráfico da utilização de um determinado serviço Web (GUNTHER, 2001). A métrica utilizada é a quantidade de transações por segundo (tps). No exemplo, foram analisados três dias de registro dos acessos dos clientes ao servidor Web onde o serviço está disponível. Notam-se picos de utilização que ocorrem sempre após as 0:00 horas. A análise desse tipo de métrica, graficamente, é bastante simples. Conclusões como as do exemplo são imediatas e auxiliam no planejamento de capacidade, pois mostram a utilização atual do servidor e muitas vezes permitem visualizar os momentos de pico de utilização (GUNTHER, 2001); (MENASCÉ et al., 1999).

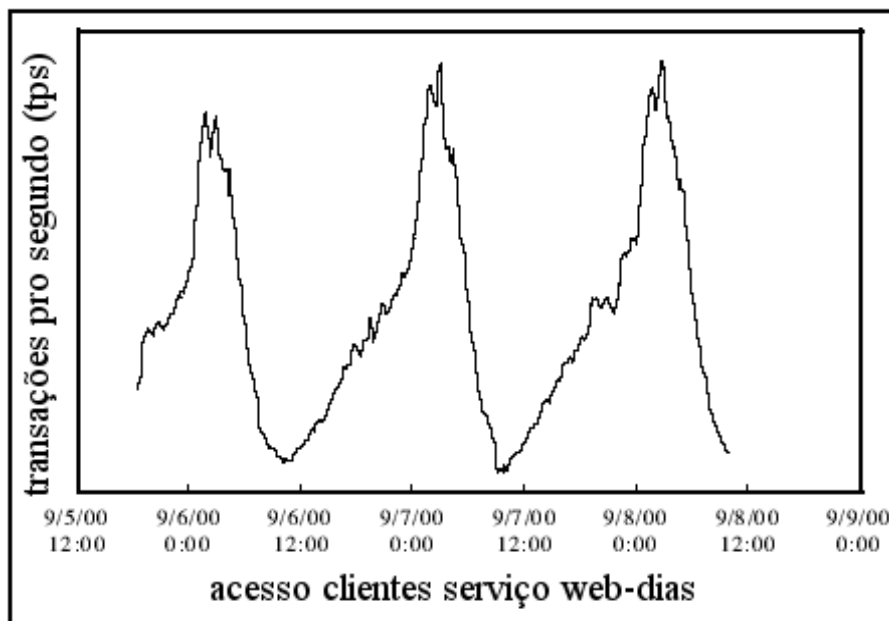


Figura 2.5 - Gráfico de utilização de servidor Web – tps.

Outra representação que também pode ser utilizada são mapas de conectividade, que ilustram os protocolos utilizados e as conexões entre os diversos servidores do ambiente de produção (DILLEY et al., 1998); (MENASCÉ; ALMEIDA, 2003a); (MENASCÉ; ALMEIDA, 2003b). Nessas representações são elencados, de forma gráfica, servidores Web, servidores de Banco de Dados, dispositivos de rede, facilitando o entendimento de como o ambiente de produção que dá suporte ao serviço Web está especificado. Combinado a esses mapas de conectividade podem ser especificados diagramas da configuração dos diversos servidores que formam o ambiente de produção. Memória, disco e CPU são registrado nesses diagramas, que recebem o nome de “Modelos QN” (fila de rede). Esses modelos podem ser usados para especificar ambientes de produção para serviços Web bastante complexos, tais como Sítios de comércio eletrônico (MENASCÉ et al., 2001a); (MENASCÉ; ALMEIDA, 2000a).

2.2.3 Acordos de Nível de Serviço (ANS)

Os ANS² são uma lista de expectativas sobre o padrão de qualidade esperado na definição da capacidade adequada de um serviço Web. Os ANS proporcionam, de certa maneira, uma “ponte de entendimento” entre os clientes e a área de TI. Ao definir os ANS, são explicitadas quais as expectativas do cliente e da gerência de TI para valores obtidos em medições de desempenho, como tempo de resposta, disponibilidade do serviço, taxa de processamento de transações, janelas de manutenção, latência de rede (COCKCROFT; WALKER 2001). Alguns exemplos de definições que podem ser usadas como ANS incluem:

1. Garantir a disponibilidade do serviço Web de requerimento de atestados escolares nos dias úteis das 8:00 horas às 21:00 horas.
2. O serviço deve enviar duzentos e-mails por segundo, utilizando apenas metade de sua capacidade.

² Os acordos de nível de serviço são comumente abreviados, em muitos textos traduzidos, de SLA (*Service Level Agreement*). Nesse texto, essa abreviação foi traduzida para ANS.

3. O serviço dever ter capacidade de atender simultaneamente trinta sessões de usuários e gerenciar uma fila de espera de vinte usuários.

No processo de definição dos ANS, deve-se buscar a resposta a diversas questões. Responder a essas questões pode explicitar antagonismos de expectativa entre usuários e fornecedores do serviço Web, e até apontar para falhas no nível do serviço que afetem de forma crítica a qualidade do serviço Web (BANGA; DRUSCHEL, 1999):

1. Qual a expectativa do cliente quanto ao serviço Web?
2. Qual a expectativa do provedor do serviço Web?
3. O que os clientes esperam do provedor do serviço Web?
4. O que o provedor do serviço Web espera dos clientes?
5. Como o ANS pode alavancar os negócios da empresa, manter o cronograma do projeto e melhorar a disponibilidade do serviço Web?

A ausência de ANS traz para o cliente a percepção de serviço de baixa qualidade. O prestador de serviço muitas vezes faz a oferta de suporte não solicitado ou prestado de maneira inconsistente, fora de padrão, causando sensação de favorecimento e parcialidade.

Um ANS pode falhar quando o canal de comunicação entre a área de TI e os clientes não estiver bem definido, ou quando as expectativas quanto ao serviço são impossíveis de serem atendidas, ou ainda pela dificuldade de quantificação e especificação do que deve ser definido como ANS.

Quando corretamente definidos, os ANS trazem diversos benefícios, atingindo a satisfação do cliente, pois este sabe exatamente o que esperar do serviço Web. Há a definição clara das prioridades do serviço e, por fim, melhor utilização do orçamento e dos recursos humanos e de computação pela área de TI. Todos esses fatores tornam o ANS uma das ferramentas mais importantes na definição de capacidade adequada de um serviço Web (BANGA; DRUSCHEL, 1999).

2.3 Frameworks para Planejamento de Capacidade Para Serviços Web

As próximas subseções apresentam alguns *frameworks* para planejamento de capacidade, a partir dos trabalhos de Menascé (MENASCÉ; ALMEIDA, 2003b), Gunther (GUNTHER, 2002) e Larsen (LARSEN; BLONJARZ, 2000).

2.3.1 Miopia em Planejamento de Capacidade

A gestão do desempenho é vista como processo fundamental para evitar situações de gargalo de aplicações e serviços Web. Muitas empresas que vendem software para monitoramento e gestão do desempenho de sistemas cliente servidor e serviços Web tem como "slogan" frases do tipo: “40 Milhões de dólares perdidos em vendas porque o Sítio Web estava indisponível, com capacidade inadequada para atender as requisições dos clientes”. Toda essa promoção muitas vezes é usada para venda de ferramentas para “apagar incêndios”, ou seja, resolver os problemas de gargalo somente no momento que ocorrerem (GUNTHER, 2002).



Figura 2.6 - O “Homúnculo Desempenho”, de Gunther.

Alguns gerentes de TI não ignoram a importância do planejamento de capacidade para evitar gargalos e prejuízos. A gerência de TI apenas evita os custos de fazer o planejamento de capacidade. Diversos fatores levam a essa postura “miope” dos gerentes de TI. Um deles é a

enorme pressão para o lançamento de novas funcionalidades de serviços Web. Por esses motivos sempre privilegiam manter os projetos dentro de prazos cada vez menores em detrimento do planejamento de capacidade (GUNTHER, 2003). Muitas vezes escolhe-se lançar o serviço Web sem um estudo prévio do impacto das novas funcionalidades no ambiente de produção: mais tarde os possíveis gargalos deverão ser corrigidos utilizando-se de software de monitoramento de desempenho (GUNTHER, 2002).

A miopia em planejamento de capacidade causa prejuízos às empresas. Gunther compara o planejamento de capacidade usando o exemplo do “Homúnculo Desempenho”, descrito na **Figura 2.6**. O planejamento de capacidade está restrito ao torso do “Homúnculo Desempenho”, quando deveria estar associado às mãos: os sistemas distribuídos que suportam os serviços Web devem estar com recursos proporcionais à sua complexidade, e não perdidos na fragilidade e imprevisto do seu torso. A “miopia” em planejamento de capacidade leva aos gerentes de TI direcionar para essa área recursos muitas vezes menores que os para *backup* e segurança (GUNTHER, 2003).

2.3.2 Planejamento de Capacidade de Guerrilha

O Planejamento de Capacidade de Guerrilha é um *framework* defendido por Gunther como melhor opção frente à realidade vivida no desenvolvimento de serviços Web. Sua principal finalidade é oferecer ao Gerente de TI um “senso de direção” sobre as necessidades de recursos para os serviços Web em ambientes que sofrem pressão por prazos curtos e baixo volume de investimentos em projetos de planejamento de capacidade (GUNTHER, 2002; 2003).

Um dos objetivos do planejamento de capacidade de guerrilha é aproveitar todos dados históricos coletados pelas ferramentas de gerenciamento de desempenho. Esses dados, quando produzidos, cobrem só o estado atual do serviço Web. Analisando os dados com outro enfoque, busca-se extrair a essência das características de desempenho, contrapondo-se com a capacidade adequada do ambiente, realizar previsões. Para análise desses dados, usam-se métricas para calcular informações sobre o período de tempo no qual a capacidade do serviço deve aumentar e como avaliar a escalabilidade de um serviço em relação à sua quantidade de cliente. Essas métricas são descritas nas próximas duas seções.

2.3.3 Bancarrota dos Servidores Web

A bancarrota dos Servidores Web é uma métrica de previsão do crescimento do serviço Web. Essa métrica, proposta por Gunther pode ser usada para serviços Web que tem como característica um aumento elevado na audiência ou utilização. Geralmente são serviços que atraem clientes de maneira intensa, na melhor tradição da Internet. O objetivo da métrica é indicar o período de tempo gasto para o volume de processamento consumido pelo serviço Web “dobrar” em relação ao volume atual (GUNTHER, 2001; 2002; 2003).

A equação representa a projeção do uso futuro desses processadores, a partir do número determinado de semanas e do consumo atual:

$$U_{\text{futuro}} = U_{\text{atual}} e^{\lambda w} \quad (2.1)$$

Parâmetros:

W = número de semanas

λ = taxa de crescimento exponencial

U_{atual} = utilização atual do processador

U_{futuro} = utilização futura do processador

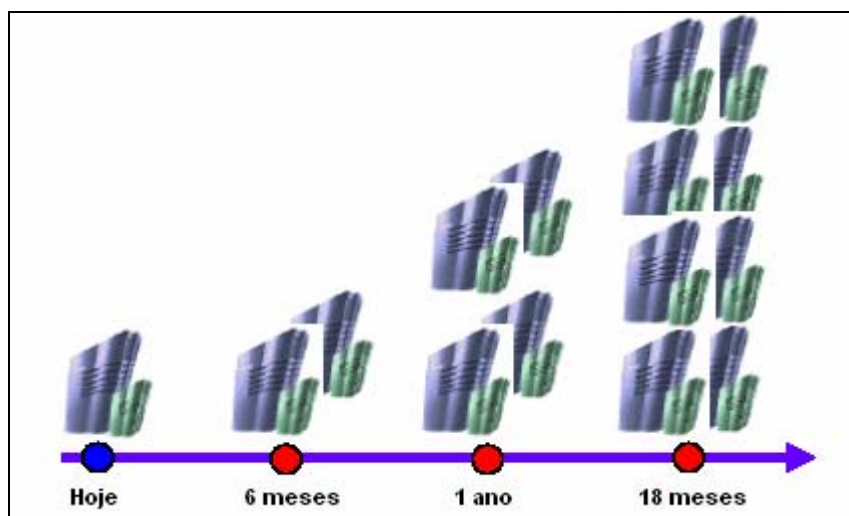


Figura 2.7 - Lei da Bancarrota de Servidores (Gunther).

A Figura 2.7 (GUNTHER, 2002) ilustra a projeção da métrica “Capacidade de processamento futuro”. A cada período determinado de meses os serviços Web com alto grau de crescimento de utilização atingirão o dobro de capacidade de processamento de um servidor Web em comparação à capacidade atual. Mais servidores Web serão necessários para atender o serviço com a mesma qualidade definida no acordo de nível de serviço. Essa lei, apelidada como “Bancarrota dos Servidores Web”, é quatro vezes mais rápida que a Lei de Moore (GUNTHER, 2001).

2.3.4 Escalabilidade Super Serializada

A métrica “Escalabilidade Super Serializada” permite avaliar a escalabilidade de um serviço Web em relação à quantidade de clientes, ao longo do tempo. Essa métrica é proposta por Gunther como parte das ferramentas de planejamento de capacidade de guerrilha. A métrica permite determinar, quantitativamente, a escalabilidade de um serviço Web, usando regressão não linear (GUNTHER, 2001; 2002; 2003).

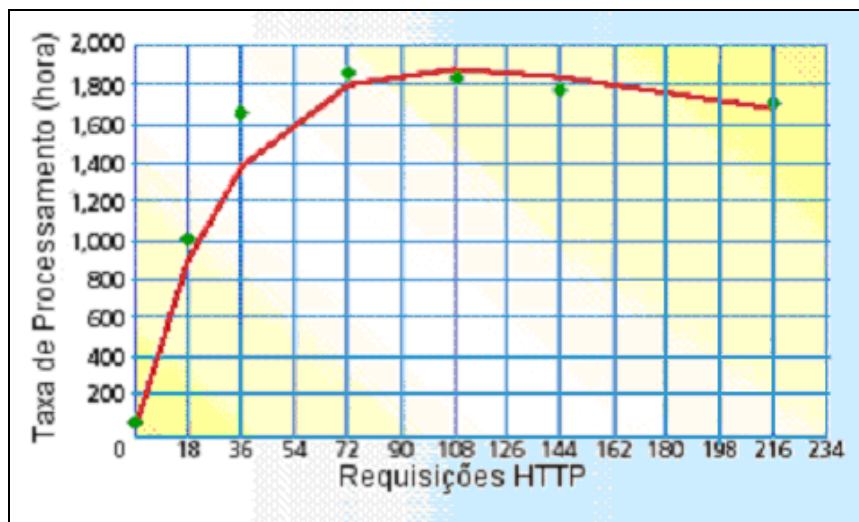


Figura 2.8 – Curva Negra de Gunther.

A Figura 2.8 (GUNTHER, 2002) exibe o gráfico foi chamado por Gunther de “curva negra”, representando a escalabilidade do serviço Web. O eixo x é o número de requisições http recebidas e o eixo y representa a quantidade de transações por hora processadas pelo serviço Web para atender a esses clientes. A Figura 2.8 mostra que, à medida que o número de requisições http aumenta de

zero, a taxa de processamento aumenta de modo linear. Entretanto, há um momento de saturação do serviço Web quando o número de clientes simultâneos chega a 100. A quantidade de transações por hora atinge o gargalo de desempenho de 1800 tph. Para um gerente de TI, essa análise gráfica seria importante para determinar a quantidade máxima de clientes a serem atendidos dentro de uma ANS que estabelecesse um padrão de qualidade mínimo para esses clientes.

A equação (2.2) representa a escalabilidade do serviço Web em função do tempo e da quantidade de clientes (GUNTHER, 2001):

$$S(\alpha, \beta, N) = \frac{N}{\{1 + \alpha \cdot [(N \cdot 1) + \beta \cdot N (N \cdot 1)]\}} \quad (2.2)$$

Parâmetros:

S = função de escalabilidade;

α = tempo de espera associado a execução de componentes de software do serviço Web;

β = tempo de espera associado ao hardware do Servidor Web;

N = número de usuários utilizando o serviço Web.

Os parâmetros α e β podem ser determinados através de ferramentas de cálculo de regressão não linear, disponíveis em planilhas de cálculo.

2.3.5 Modelos de Custo

Serviços Web podem ser facilmente desenvolvidos. Muitas ferramentas de software possibilitam o desenvolvimento de um serviço Web de maneira muito simples e rápida. Pela facilidade encontrada no desenvolvimento de um serviço Web, muitas organizações implementam novas funcionalidades e até mesmo novas aplicações sem uma análise prévia de seu impacto. Dessa maneira, os serviços Web quase sempre não atingem os objetivos de sua organização e acabam custando mais do que o esperado. Sua manutenção é cara e complexa, gargalos de desempenho são frequentes e os acordos de nível de serviço inexistem.

Um dos grandes problemas de um serviço Web é o mito que sua implementação trará, em curto prazo, grande economia à organização. Quase sempre ocorre o contrário: o oferecimento de um serviço Web aumenta gastos da empresa durante seu período inicial (LARSEN; BLONIAZ, 2000), pois:

1. A informação não deve ficar desatualizada, demandando renovação constante dos dados oferecidos aos clientes;
2. O custo dos recursos humanos usados para construir e operar o serviço Web costuma ser muito maior do que o custo dos equipamentos do Sítio;
3. Especificar acordos de nível de serviço usando indicadores de desempenho é muito mais fácil e usual do que especificar indicadores de custo;
4. Havendo gargalos no ambiente de produção do serviço Web, o custo para ajuste da capacidade adequada é alto.

Larsen e Bloniarz (LARSEN; BLONIAZ, 2000) desenvolveram uma metodologia que pode ser usada para organizações que estudam a expansão de seus serviços Web atuais ou o oferecimento de um novo serviço. O objetivo desse método é identificar e especificar diversos cenários possíveis na implementação do serviço Web, com enfoque principalmente em viabilidade financeira, e então escolher o cenário mais promissor.

A solução é constituída por três ferramentas, que permitem especificar o serviço Web, através de processos e níveis de serviço:

1. Planilha de Funcionalidades;
2. Planilha de Desempenho;
3. Planilha de Custos.

A) Planilha de Funcionalidades

Definir as funcionalidades que o serviço Web deverá oferecer é o primeiro passo do método de Larsen e Bloniarz (2000). Nessa planilha, as funcionalidades são especificadas de maneira simples, classificadas em 3 níveis de complexidade esperada pelo serviço: modesto, moderado e elaborado. Essa planilha é produzida na fase inicial do desenvolvimento do serviço Web e permite o melhor entendimento dos processos que devem ser analisados e mensurados nas fases seguintes.

Também são especificadas outras informações correlacionadas ao serviço Web e suas funcionalidades. A **Tabela 2.1** ilustra uma pequena amostra da planilha de funcionalidades. Neste exemplo, foram relacionadas apenas as linhas que registram:

1. Quais das funcionalidades do sistema de informação da empresa deverão ser incluídas no serviço Web. Essa escolha é fundamental para que o serviço seja relevante ao cliente.
2. As operações que os clientes poderão efetuar, e a relevância de cada operação dentro do serviço Web. Essas operações estão alinhadas diretamente com as funcionalidades que serão oferecidas pelo serviço.
3. Quais fontes da empresa serão estratégicas para a gestão das informações utilizadas e geradas pelo serviço Web. Essas fontes geralmente são áreas operacionais dentro do organograma da Empresa.

Tabela 2.1 - Exemplo de Planilha para funcionalidades do sistema..

	Modesto	Moderado	Elaborado
Que operações os clientes poderão efetuar?	cadastro	preferências	compras obtenção de crédito
Quais funcionalidades do sistema serão incluídas no serviço Web?	inclusão novos clientes	cadastro de preferências	seleção de produtos; pagamentos
Quais fontes de informação (interna e externa) devem ser coordenadas?	marketing	-	Vendas logística produção

B) Planilha de Desempenho

No segundo passo do método de Larsen e Bloniarz (2000), são detalhados os níveis de desempenho esperado do serviço Web. Algumas variáveis são definidas pelas áreas de TI e de Negócios (Vendas, Marketing) da empresa para serem acompanhadas em prazos determinados de tempo. Para essas variáveis são propostas metas em três níveis: modesto, moderado e elaborado. As metas definem de maneira clara os objetivos que o serviço Web deve alcançar para atingir sua viabilidade econômica.

A Tabela 2.2 ilustra um pequeno exemplo de uma planilha de desempenho, através do registro de uma variável. No exemplo, a base de novos clientes por mês é a variável registrada e atribuída, uma meta para incremento de seu valor, na ordem de 5%, 10% e 15%, nos níveis modesto, moderado e elaborado. A variável “base de clientes” é definida pela área de negócios da empresa. A planilha de desempenho permite definir, com clareza, as expectativas da empresa em relação ao serviço Web que está desenvolvendo.

Tabela 2.2 - Planilha de Desempenho.

Variável: Base de Clientes		
Medida: Número de novos clientes / mês (melhorar)		
Meta Modesta	Meta Moderada	Meta Elaborada
Aumento da Média em 5%	Aumento da Média em 10%	Aumento da Média em 15%

C) Planilha de Custos

No terceiro e último passo do método de Larsen e Bloniarz (2000) é definida a Planilha de Custos. Nesse modelo são definidos diversos cenários que permitem avaliar o custo-benefício do serviço Web. Essa terceira planilha é especialmente útil em organizações centradas primariamente em custos, pois permite especificar detalhadamente todos os custos inerentes à implantação do serviço Web e sua manutenção futura. Os custos de implantação e manutenção estão divididos em cinco categorias:

1. Suporte ao cliente;
2. Acesso para empregados e outros usuários;
3. Cultura Organizacional;
4. Infra-estrutura do ambiente de produção do serviço Web;
5. Desenvolvimento e manutenção do conteúdo.

Para cada categoria de despesa e seus itens, são estabelecidas três projeções, em função do custo, para o primeiro ano e para os anos seguintes. Essas projeções classificam-se em “modesto”, “moderado” e “elaborado”.

Há um destaque para a especificação dos custos relacionados a recursos humanos. São especificados custos de treinamento, administração de redes e aplicações, gerenciamento de servidores, dentre outros. Esse enfoque no detalhamento dos gastos com recursos humanos auxilia na correta estimativa de custos de uma das maiores fontes dos gastos nos projetos de desenvolvimento de serviços Web.

A Tabela 2.3 registra uma análise de custo esperado para alguns parâmetros que compõem a infra-estrutura necessária para oferecer o serviço Web (recursos humanos, software, hardware). No exemplo os parâmetros são referenciados como categorias de despesas e algumas delas podem ter diversos itens, como o caso dos recursos humanos.

Tabela 2.3 - Planilha de Custos.

	MODESTO		MODERADO		ELABORADO	
	1º ano	Anos seguintes	1º ano	Anos seguintes	1º ano	Anos seguintes
Infra-estrutura do Ambiente de Produção						
Hardware						
Software						
<i>Recursos Humanos</i>						
Treinamento para a Equipe						
Administração de Rede e Sistemas						
Gerenciamento de Servidor Web						

2.3.6 O Framework para Planejamento de Capacidade de Menascé

Daniel Menascé (2003) e colaboradores desenvolveram um *framework* completo para o Planejamento de Capacidade para Serviços Web. O processo utiliza principalmente técnicas quantitativas, baseadas em três tipos de modelos: carga de trabalho, desempenho e disponibilidade. Esse *framework* é bastante robusto e cobre todas as etapas do planejamento de capacidade para serviços Web.

A Figura 2.9 (MENASCÉ; ALMEIDA, 2003b) mostra o ciclo de vida do planejamento de capacidade proposto por Menascé:

1. O primeiro passo é o conhecimento do ambiente atual do serviço Web e da estrutura que dá suporte a esse serviço. No passo seguinte, uma análise desse serviço recolhe dados que caracterizam a carga de trabalho atual, produzindo no final desse processo um modelo de carga de trabalho do serviço Web. Esse modelo de carga de trabalho é usado como insumo para, no passo seguinte, desenvolver-se um modelo de carga de desempenho.
2. O modelo de desempenho é usado na predição do comportamento do serviço Web em diversos cenários, em que a carga de trabalho poderá variar num momento futuro. Esse processo de predição é fundamental dentro do planejamento de capacidade, pois oferece ao gerente de TI alternativas e subsídios para decisões a serem tomadas antes do serviço Web atingir gargalos e momentos de saturação.
3. Paralelamente ao modelo de carga de trabalho e desempenho é desenvolvido um modelo de custo. Esse modelo especifica os diversos custos envolvidos na implementação de ajustes no serviço Web e também auxilia à tomada de decisão do gerente de TI na opção ao modelo de melhor custo - benefício.

O processo de planejamento de capacidade desse *framework* resulta de alguns produtos. Os mais importantes são o modelo de carga de trabalho, modelo de desempenho e o modelo de custo. Além desses modelos, são obtidos um plano de ação para configuração do ambiente de suporte ao serviço Web, um plano de investimentos em novo hardware e software (para manutenção da capacidade adequada do serviço) e um plano de pessoal, que permite definir custos com os recursos humanos necessários para o suporte adequado ao serviço Web.

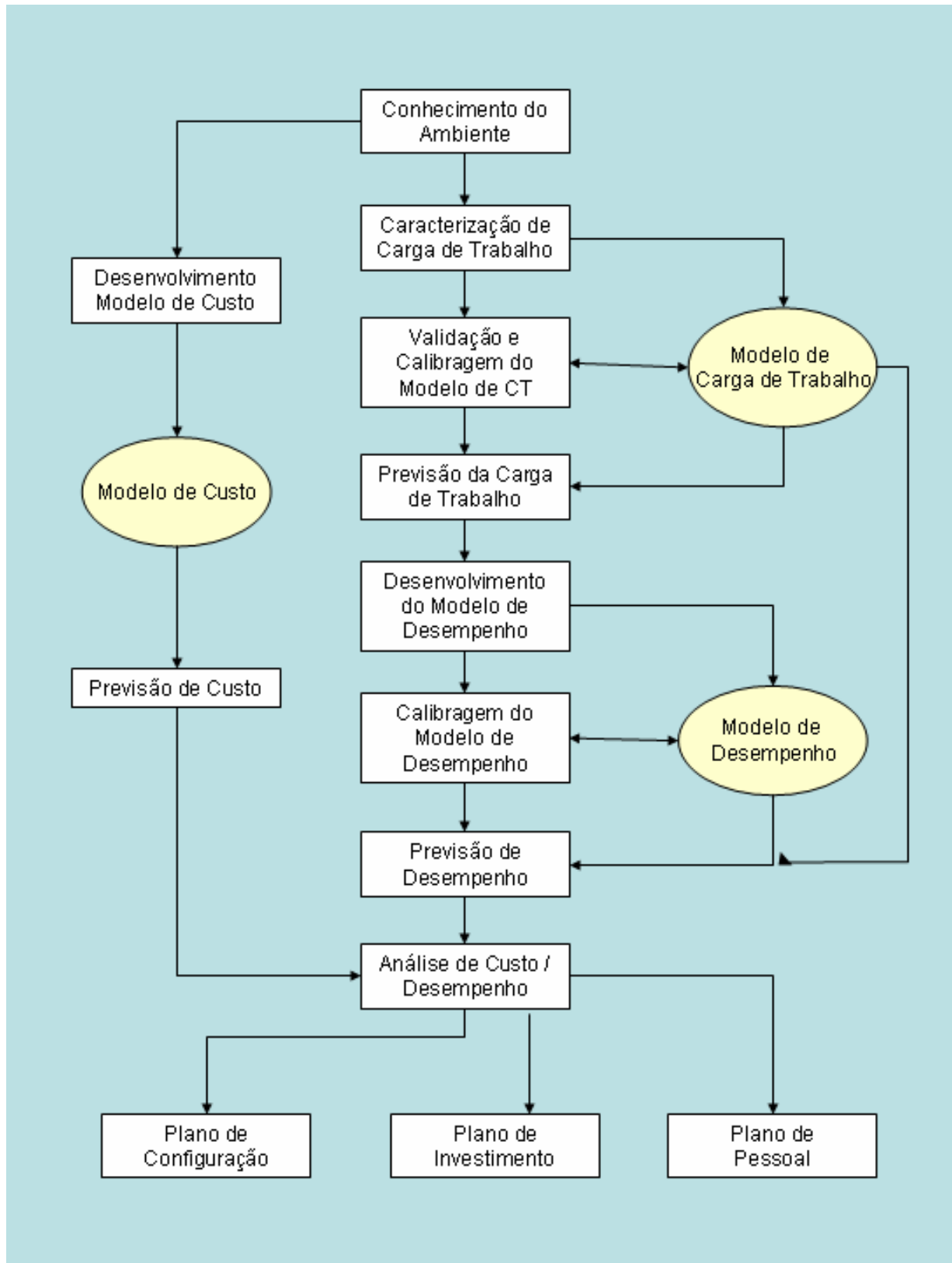


Figura 2.9- *Framework* para o Planejamento de Capacidade de Menascé.

2.4 Cargas de Trabalho de Serviços Web

Para os negócios de uma empresa é desejável que seus serviços Web sejam flexíveis e tenham escalabilidade para suportar altos volumes de requisições, com qualidade e capacidade adequada. Estudos do comportamento de clientes dos serviços Web indicam pouca paciência ao utilizar serviços com tempo de resposta inadequado (GUNTHER, 2002) . A demanda do cliente é diretamente relacionada à carga de trabalho que deve ser atendida pelo serviço Web. Para evitar situações de gargalo de desempenho, que possibilitem a perda de clientes e de receita, é importante antecipar eventuais crescimentos e mudanças da carga de trabalho (MENASCÉ; ALMEIDA, 2003b).

Eventos como escândalos políticos, acidentes, guerras, catástrofes, datas especiais como Natal, ofertas temporárias, lançamentos de novos produtos e campanhas na TV podem aumentar a demanda de clientes por um determinado serviço Web e com isso aumentar a carga de trabalho do respectivo serviço. Alguns desses eventos podem ser classificados como *imprevisíveis*, como desastres naturais. Outros eventos são *previsíveis*, como lançamento de um produto, datas festivas, anúncios na TV (MENASCÉ; ALMEIDA, 2000a).

A carga de trabalho de um serviço Web pode ser definida como todas as requisições de seus clientes recebidas e processadas em um determinado período de tempo. O desempenho de um serviço Web depende diretamente da carga de trabalho que é capaz de processar e é afetado por sua mudança ao longo do tempo (MENASCÉ; ALMEIDA, 2000b).

Mostrou-se no capítulo 2 o momento de saturação do serviço Web face ao aumento das sessões e requisições http dos clientes. Essa saturação, denominada “curva negra” por Gunther, acontece em decorrência do aumento da carga de trabalho até o ponto em que o serviço atinge sua capacidade máxima. Nesse cenário, chamado de gargalo de desempenho, podem acontecer situações como a negação de serviço (DoS) e o colapso do serviço Web (GUNTHER, 2004).

Conhecer a carga de trabalho é requisito fundamental na atividade de Planejamento de Capacidade de um serviço Web. A carga de trabalho pode ser entendida e analisada por diversas maneiras, o que é chamado caracterização. A caracterização da carga de trabalho de um serviço

Web pode ser descrita em diferentes níveis, para ajudar a sua compreensão. Esses níveis vão do plano de visão corporativa até o plano de visão técnico (MENASCÉ; ALMEIDA, 2003b).

A carga de trabalho gerada na Internet é diversificada, tornando difícil sua análise e entendimento devido a grande quantidade de características apresentadas: uso de CPU, memória, I/O, quantidade de clientes e conexões abertas. Entretanto, existem algumas características comuns em toda carga de trabalho de serviços Web, chamadas de invariantes. As invariantes podem ser o comportamento da requisição do usuário, a temporalidade e popularidade dos arquivos, a ocorrência do fenômeno de rajadas, dentre outras (BARFORD; CROVELLA, 1998); (CALZAROSSA et al., 2000).

2.4.1 Ocorrência do Fenômeno de Rajadas em Cargas de Trabalho

O comportamento da demanda na forma de rajadas é um fenômeno específico da carga de trabalho de serviços Web. Diversos estudos observaram que o tráfego gerado pela requisição HTTP do usuário é em forma de rajadas (BARFORD; CROVELLA, 1998); (CALZAROSSA et al., 2000); (MENASCÉ, 2000); (MENASCÉ; ALMEIDA, 2000b). O fenômeno de rajadas é também chamado de demanda imprevisível, devido a dificuldade da previsão de sua ocorrência.

O desempenho do serviço Web é afetado por essa carga de trabalho que chega com alta variabilidade em um determinado espaço de tempo. Um estudo de Wang et. al. (WANG et al., 2003) mostra como a capacidade do serviço pode ser afetada pela ocorrência do fenômeno de rajadas. Fazendo a análise de um site de comércio eletrônico típico, Wang observou que:

Os registros históricos das requisições agrupados em escalas de 1 minuto, eram os que melhor representavam as características da carga de trabalho. Esses registros foram retirados de um período de tempo de 24 horas.

O tempo de resposta do serviço não apresenta uma forte correlação com a quantidade de requisições dos clientes. Há períodos do dia em que a capacidade é adequada com tempo de resposta mínimo

para as requisições dos clientes. Mesmo com a quantidade de requisições por minuto variando entre uma faixa de 0 a 170, a média do tempo de resposta permanece inalterada.

Em alguns intervalos, entretanto, a taxa de tempo de resposta aumenta drasticamente, causando um gargalo de desempenho no serviço Web. Wang conclui que nesse serviço o gargalo na taxa de tempo de resposta está mais diretamente relacionado com a ocorrência do fenômeno de rajadas que com a taxa de chegadas de requisições (WANG et al., 2003).

Uma forma de observar a ocorrência das rajadas é a inspeção visual de gráficos de sessão do usuário. A Figura 2.10 reproduz um modelo de gráfico padrão utilizado para inspeção visual do fenômeno de rajadas encontrado em diversos estudos (ARLITT et al., 2001); (CALZAROSSA et al., 2000); (MENASCÉ; ALMEIDA, 2000b); (MENASCÉ et al., 1999). Esse modelo utiliza dados históricos com os acessos http dos clientes de serviços Web. O eixo Y representa o número de requisições que chegam em um serviço Web, que podem ser medidos em diferentes escalas de tempo. O eixo X representa a chegada das requisições ao longo do período analisado (ALMEIDA et al., 2002).

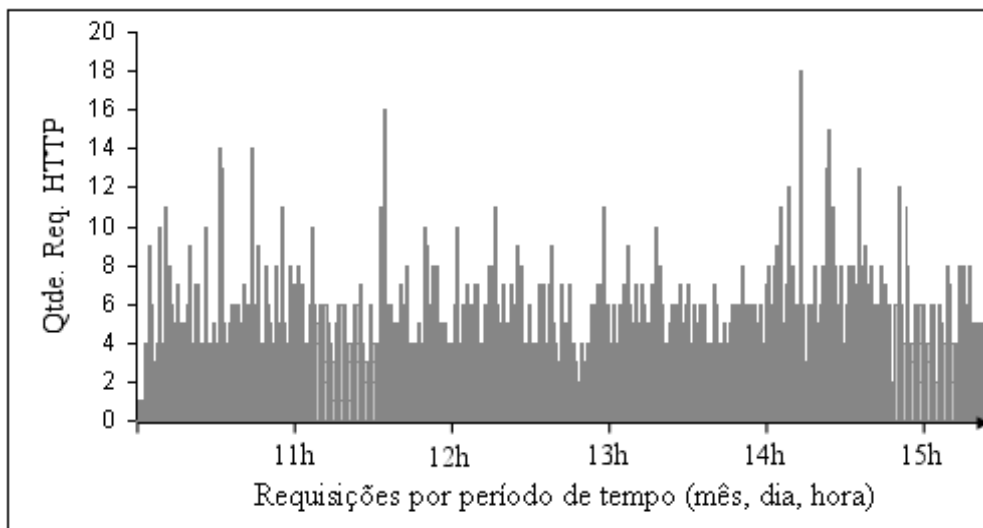


Figura 2.10 – Ocorrência do fenômeno de rajadas em cargas de trabalho.

A inspeção visual dos gráficos de sessão do usuário permite observar, além do fenômeno de rajadas, características de requisição do usuário tais como: taxas de pico de requisição, número de conexões simultâneas, evolução da carga de trabalho do negócio, tráfego semanal e diário, dentre

outros (MENASCÉ; ALMEIDA, 2000a). Embora simples em sua implementação, a inspeção visual é imprecisa.

O próximo capítulo descreve um modelo determinístico que pode ser usado com maior rigor para análise da rajada, determinando parâmetros que medem sua intensidade e proporção de tempo de ocorrência.

3 Modelos para Representação da Carga de Trabalho com o Fenômeno de Rajadas

No capítulo anterior mostrou-se que um fenômeno característico da carga de trabalho de serviços Web é o comportamento da demanda em forma de rajadas. A capacidade de processamento das requisições de um serviço Web diminui proporcionalmente ao aumento da intensidade da rajada. Essa intensidade pode ser registrada por uma métrica chamada fator de rajada (BARFORD; CROVELLA, 1998); (CALZAROSSA et al., 2000); (MENASCÉ; ALMEIDA, 2000b).

3.1 Abordagem Operacional para o Fenômeno de Rajadas

Uma abordagem operacional pode ser utilizada na análise da carga de trabalho de um serviço Web para determinar a ocorrência do fenômeno de rajadas, sua intensidade e seu conseqüente impacto na utilização do serviço Web. Nessa análise são necessários os registros históricos do serviço Web, obtidos através de um arquivo de registros de requisições http dos usuários. A análise dos registros históricos das requisições é feita em apenas um determinado intervalo de tempo T que pode ser definido em uma escala de segundos, minutos ou horas (BUZEN, 1978); (DENNING; BUZEN, 1994).

O primeiro passo para a construção do modelo de carga de trabalho através da abordagem operacional é a definição da taxa média de chegada de requisições http λ em um período de tempo T. Pode-se calcular o valor λ a partir da equação (3.1):

$$\lambda = \frac{L}{T} \quad (3.1)$$

Em que:

L representa o registro histórico da quantidade de requisições http dos clientes do serviço Web.

T é intervalo de tempo no qual o serviço Web é analisado, geralmente definido em segundos ou em minutos.

Obtendo λ , tem-se a informação de quantas requisições, por unidade de tempo, o serviço Web recebeu em média no intervalo que está sendo analisado. Essa informação é fundamental para calcular o fator de rajada que o serviço Web tenha apresentado nesse intervalo de tempo.

O segundo passo para a construção do modelo é a definição do fator de rajada, representada pelos parâmetros (a,b) . Esse fator registra a rajada ocorrida em um determinado intervalo de tempo T. Pode-se definir os parâmetros (a,b) assim:

- O parâmetro a é a razão entre a taxa de requisição acima da média e a taxa média de requisição (λ).
- O parâmetro b é a fração de tempo em que a taxa de chegada instantânea da requisição do usuário excede a taxa média de requisição λ . O parâmetro b pode assumir valores entre 0 e 1, e geralmente é representado em porcentagem, de 0% a 100%.

A intensidade do fator de rajada, mesmo sendo mínima, pode afetar o desempenho do serviço Web. Por exemplo, um estudo de Banga e Drusche (BANGA; DRUSCHEL, 1997) mostra a ocorrência de uma rajada ($a = 6$) em um pequeno intervalo de tempo ($b=5\%$) que reduz a capacidade de um serviço Web entre 12% a 20%.

A Figura 3.1 ilustra como a intensidade da rajada registrada nos parâmetros (a,b) afeta a capacidade de um serviço Web. Para quantidades de requisições http iguais (eixo X), mas com intensidade de rajada variada, o serviço apresenta capacidade de processamento diferenciada. A capacidade de processamento do serviço (eixo Y) diminui conforme a intensidade do fator de rajada aumenta.

O exemplo da Figura 3.1, presente em estudos de Banga et al. (BANGA; DRUSCHEL, 1997; 1999), Y ilustra algumas variações do fator de rajada com o parâmetro $a = 6$ e o parâmetro b entre 0% e 15%. Como resultado, a capacidade de processamento de transações de um serviço Web (dada em transações por segundo) pode ser afetada com a perda entre 20% a 40% do desempenho.

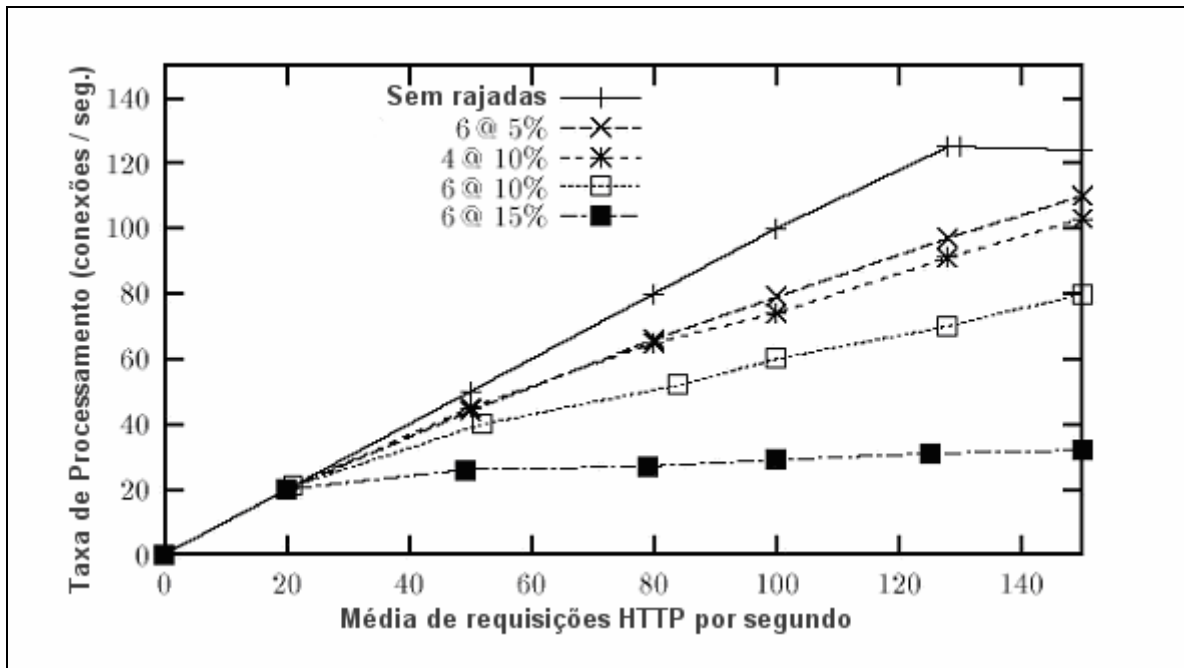


Figura 3.1 – Efeito da intensidade da rajada na capacidade do serviço Web.

3.1.1 Algoritmo para Cálculo dos Parâmetros a e b.

Essa seção apresenta um algoritmo simples que pode ser utilizado para o cálculo do fator de rajada. O cálculo dos parâmetros (a,b) utiliza os registros históricos L dos clientes de um serviço Web.

1. Para ser analisado, o intervalo de tempo T é dividido em n subintervalos de tamanhos iguais. Cada subintervalo é chamado de *época*, sendo a sua duração determinada pela equação (3.2).

$$k = \frac{T}{n} \quad (3.2)$$

2. $A(k)$ é um vetor que representa o número de requisições dos usuários durante k épocas.

3. A^+ é o número total de requisições http recebidas que ultrapassam a taxa média de requisições λ .
4. A^- é o número total de requisições http recebidas no intervalo de tempo T que não ultrapassam a taxa média de requisições λ .
5. O número total de requisições http recebidas no intervalo de tempo T é dado pela equação (3.3):

$$L = A^+ + A^- \quad (3.3)$$

6. λ_k é a taxa de requisição em uma determinada época k .
7. k^+ a quantidade de épocas k em que $\lambda_k > \lambda$.
8. O parâmetro b é calculado pela equação (3.4):

$$b = \frac{k^+}{n} \quad (3.4)$$

9. O parâmetro a é calculado pela equação (3.5):

$$a = \frac{A^+}{b * L} \quad (3.5)$$

3.1.2 Leis de Utilização do Recurso e o Fenômeno de Rajadas

A taxa máxima de processamento do serviço Web diminui quando o fator de rajada aumenta. Essa alteração influencia diretamente a taxa de utilização do serviço, definida como uma fração do intervalo de tempo analisado no qual os recursos do sistema foram ocupados. Em contrapartida, um aumento da capacidade da taxa de processamento do serviço por unidade de tempo tende a diminuir a utilização do serviço. Esse aumento de capacidade ocorre em função de uma otimização

realizada no sistema ou diminuição da demanda gerada pela carga de trabalho do recurso (ARLITT et al., 2001); (MENASCÉ et al., 2004).

Esse modelo de comportamento da carga de trabalho pode ser representado pela Lei de Utilização do Recurso, que é definida pela equação (3.6).

$$U = X * S \quad (3.6)$$

Por definição:

- U é a taxa de utilização do recurso (cpu, disco, memória), definida como a fração do tempo em que o recurso está ocupado. O resultado da utilização pode ser convertido e representado em forma de porcentagem.
- X é a taxa de processamento média do sistema, atendida em uma determinada unidade de tempo (throughput).
- S é o tempo médio de duração para a execução de uma transação ou serviço.

A lei da utilização do recurso pode ser adaptada para um serviço Web, como aparece na seguinte equação:

$$U = \lambda * \mu \quad (3.7)$$

Em que:

- μ é a taxa média da espera pelo processamento das requisições http (B) em uma determinada unidade de tempo (T). Essa taxa pode ser definida também como a demanda de serviço, conforme definida na equação (3.8).

$$\mu = \frac{B}{T} \quad (3.8)$$

- λ é a taxa média de chegada de requisições http (L) em uma determinada unidade de tempo (T). Nesse caso, λ é igual à X, que é a taxa média de processamento do serviço Web para um determinado intervalo de tempo (*throughput*), conforme definida na equação (3.9):

$$\lambda = \frac{L}{T} \quad (3.9)$$

O fenômeno da carga de trabalho em forma de rajadas também pode ser representado no modelo de utilização de recurso do serviço Web (MENASCÉ; ALMEIDA, 2003b). A demanda pelo serviço Web, representada pela taxa média de requisições http, é dada pela equação (3.10):

$$D = D_f + \alpha * b \quad (3.10)$$

Em que:

D_f é um componente da demanda que não é afetado pelo fenômeno da carga de trabalho em forma de rajadas.

$(\alpha * b)$ representa o aumento da demanda do serviço Web gerado pela carga de trabalho quando ocorre a rajada. O fator b é usado com um termo proporcional $\alpha > 0$ para ajustar a representação desse gargalo de desempenho causado pela rajada. O coeficiente proporcional de rajada α pode ser determinado de forma semelhante ao fator de rajada b , através da análise dos registros históricos da carga de trabalho. A seção seguinte apresenta um algoritmo para o cálculo do coeficiente proporcional de rajada.

3.1.3 Algoritmo para Cálculo do Coeficiente Proporcional de Rajada α

Essa seção apresenta um algoritmo simples que pode ser utilizado para o cálculo do coeficiente proporcional de rajada α . É necessária a utilização dos registros históricos da carga de trabalho do serviço Web para a realização deste cálculo:

1. Obter L (dados históricos correspondentes ao arquivo de registro de requisições http dos usuários do serviço Web no intervalo de tempo definido por T);

2. Obter dois subintervalos de L (L_1 e L_2) com duração $T/2$;
3. Calcular o fator de rajada b_1 para L_1 ;
4. Calcular o fator de rajada b_2 para L_2 ;
5. Calcular a taxa de processamento para o intervalo T_1 e T_2 , obtido pela equação (3.9);
6. Calcular o coeficiente proporcional de rajada α , segundo a equação (3.11).

$$\alpha = \frac{\left(\frac{U_1}{\lambda_1}\right) - \left(\frac{U_2}{\lambda_2}\right)}{b_1 - b_2} \quad (3.11)$$

3.2 Análise de Múltiplas Escalas de Tempo

Estudos recentes (ALMEIDA et al., 2002); (ARLITT et al., 2001); (MENASCÉ; ALMEIDA, 2003a); (WANG et al., 2003) constataram que a escala de tempo utilizada na análise dos registros históricos influi na caracterização da carga de trabalho. A análise feita, a partir de uma única escala de tempo, pode ficar incompleta ou inconsistente. Escalas de tempo muito grandes são inapropriadas para a avaliação do desempenho de um serviço Web. Escalas de tempo muito reduzidas podem distorcer a ocorrência de eventos como o fenômeno de rajadas.

A escala de tempo está diretamente associada às características da sessão do usuário para um determinado serviço Web. Assim, a sessão do usuário pode ser definida como o intervalo de tempo no qual as requisições http são enviadas para o serviço Web. O envio dessas requisições é intercalado por intervalos de inatividade, chamado de momento de pensar. O padrão de indústria atual para tempo de inatividade é o intervalo de 30 minutos (1800 segundos). Esse intervalo é usado como critério para delimitar o intervalo de inatividade do cliente: após o término desse período o cliente é desconectado do serviço.

Um exemplo desse problema é visto em Arlitt et. al. (ARLITT et al., 2001), na análise de um serviço de comércio eletrônico com alto volume de requisições, em que se conclui que uma escala

de tempo de 15 minutos (900 segundos) pode representar adequadamente o tempo de sessão de um cliente e ser adequadamente utilizada no estudo da carga de trabalho. Essa conclusão é feita a partir da determinação da quantidade média de 30 requisições recebidas por um cliente durante uma sessão.

Almeida et al. (ALMEIDA et al., 2002); (MENASCÉ; ALMEIDA, 2003a) apontam a necessidade de utilizar múltiplas escalas de tempo na análise dos diferentes aspectos da carga de trabalho. Em uma análise feita em dois serviços de comércio eletrônico, diversas novas características da carga de trabalho foram identificadas, algumas delas relacionadas diretamente a escala de tempo:

- A maior parte das sessões de usuários apresenta um período inferior a 1000 segundos (16 minutos);
- 88% das sessões tem menos de 10 requisições;
- Há uma intensa correlação entre a escala de tempo e o processo de chegadas de requisições;
- Essa correlação é mais intensa quando analisada a chegada de requisições em um intervalo de tempo delimitado em poucos minutos.

3.3 Previsão da Carga de Trabalho Futura

A previsão da carga de trabalho futura de um serviço Web é uma parte fundamental dentro do planejamento de capacidade. Conhecer previamente a evolução da carga de trabalho permite manter o nível de serviço na qualidade desejada pelos clientes do serviço Web. A previsão de carga de trabalho é constituída por um conjunto de suposições e cenários que antecipam o crescimento ou diminuição da demanda em períodos específicos, além de obter subsídios e respostas para situações imprevistas (MENASCÉ et al., 2001b).

Para a previsão da carga de trabalho futura podem ser adotadas duas abordagens diferentes: qualitativa ou quantitativa. A abordagem qualitativa é usada quando não há nenhum dado histórico sobre a carga de trabalho à disposição. Em contraponto, a abordagem quantitativa faz uso de uma

quantidade disponível de dados históricos da carga de trabalho para estimar a evolução futura da mesma (MENASCÉ; ALMEIDA, 2003b); (MENASCÉ; ALMEIDA, 2000a).

A Tabela 3.1 mostra algumas técnicas e metodologias disponíveis para as análises de enfoque qualitativo e quantitativo. As análises baseadas no enfoque quantitativo utilizam modelos determinísticos. Em contrapartida, o enfoque qualitativo se baseia mais na opinião de especialistas e usuários.

Tabela 3.1 - Técnicas de previsão - enfoque qualitativo e quantitativo.

Enfoque	Técnicas
Qualitativo	Opinião de especialistas Método Delphi Intuição
Quantitativo	Regressão Linear Médias Móveis Suavização exponencial

Os dados históricos são uma condição necessária para desenvolver estratégias com o enfoque quantitativo. Essa coleção de dados históricos da carga de trabalho na ordem cronológica é chamada de Sequência Temporal. A natureza dos dados históricos pode ser determinada através de inspeção visual em amostras de períodos de registro. Três padrões de dados históricos podem ser detectados (BARFORD; CROVELLA, 1998); (MENASCÉ; ALMEIDA, 2002):

1. Sazonal ou Cíclico: padrões que apresentam flutuações similares, com picos de utilização. A diferença entre os dois é a periodicidade de flutuação, exibida com mais evidência no padrão sazonal.
2. com Tendência: indica a variação monotônica da carga de trabalho, tanto de crescimento quanto de diminuição.
3. Estacionário: esse padrão não apresenta nenhum sinal de grande variação ao longo da seqüência temporal.

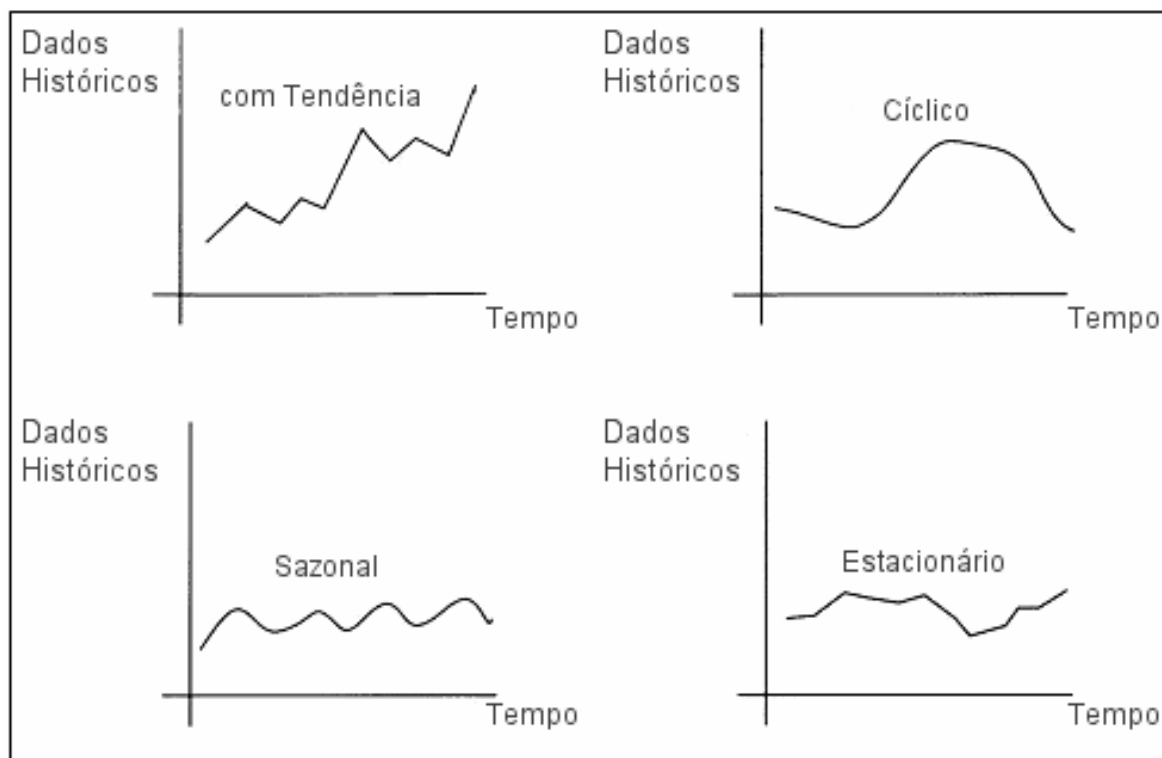


Figura 3.2 - Gráficos com os 4 padrões de dados históricos.

A Figura 3.2 ilustra os diferentes tipos de padrões visuais que podem ser tipificados quando analisamos os dados históricos. O conhecimento prévio do padrão dos dados históricos da carga de trabalho permite uma escolha correta das técnicas de previsão disponíveis para estimar a evolução futura, como regressão linear, médias móveis e suavização exponencial (MENASCÉ; ALMEIDA, 2003b).

O enfoque qualitativo pode ser utilizado na estimativa da carga de trabalho na ausência de registros históricos. Técnicas como média das opiniões do grupo, consenso do grupo e o método Delphi são utilizados para capturar e estruturar o conhecimento do comportamento da carga de trabalho. Opiniões de especialistas passam por um processo de refinamento e são utilizadas na construção de cenários de evolução da carga de trabalho de curto a médio prazo. As opiniões apresentadas por usuários do serviço Web também podem ser consideradas. Essas informações são obtidas por preenchimento de questionários e entrevistas (MENASCÉ et al., 1999).

O processo de previsão da carga de trabalho futura apresenta algumas dificuldades, em ambos os enfoques:

1. No enfoque qualitativo nem sempre é clara e correta a compreensão dos componentes da carga de trabalho por parte dos usuários e até mesmo por profissionais envolvidos na atividade de projeto e suporte aos serviços Web (GUNTHER, 2003).
2. No enfoque quantitativo, o processamento de dados históricos é necessário, entretanto, essas informações muitas vezes não estão corretamente registradas ou se apresentam apenas parcialmente disponíveis. Essa análise exige alta capacidade computacional para o processamento dos registros históricos de serviços Web com grande quantidade de clientes e acessos (MENASCÉ et al., 2004).
3. A ocorrência do fenômeno de rajadas, que surgem em momentos aleatórios consumindo toda a capacidade do sistema, são de difícil previsão, tanto nas técnicas e modelos quantitativos quanto qualitativos, pelas mesmas razões afirmadas nos itens 1 e 2.

4 Lógica Fuzzy no Planejamento de Capacidade para Serviços Web

A lógica fuzzy, também chamada lógica nebulosa, é adequada para ser utilizada em ambientes em que a definição do problema apresenta imprecisões, é não-linear e cheia de incertezas (REED; AYDT, 1999); (SILVEIRA, 2005); (ZADEH, 1987a). Essa característica dos sistemas fuzzy facilita aos clientes e responsáveis pelo suporte do serviço Web definir variáveis, valores, estados e características qualitativas da carga de trabalho.

Atualmente as aplicações para suporte ao Planejamento de Capacidade que mais se beneficiam de implementações com lógica fuzzy são as que utilizam controles para solução de problemas não-lineares e otimização. Uma aplicação que utiliza um controle fuzzy para a otimização dos lucros obtidos na prestação de um serviço Web foi proposta por Diao (DIAO et al., 2002). Nessa aplicação, que utiliza as variáveis obtidas a partir do acordo de nível de serviço estabelecido (SLA) entre o fornecedor do serviço e os clientes, o controle fuzzy produz respostas para manter a qualidade do serviço (QoS) dentro do aceitável.

A carga de trabalho de um serviço Web apresenta características específicas como o fenômeno das rajadas (ARLITT; WILLIAMSON, 1996); (BARFORD; CROVELLA, 1998) e distribuição de cauda longa (CALZAROSSA et al., 2000), que podem afetar diretamente o desempenho e a qualidade do serviço. Aplicações de otimização utilizando sistemas fuzzy foram propostas por Buckley et al. (BUCKLEY et al., 2004a); (BUCKLEY et al., 2004b) para diversos modelos de desempenho de serviços Web, nas seguintes funções:

- Minimizar o tempo de resposta de um serviço Web considerando o fenômeno das rajadas;
- Modelos de desempenho utilizando simulação de taxa de chegada de requisição de usuários usando probabilidades fuzzy;
- Modelos de desempenho adaptados ao fenômeno de rajadas e distribuição de cauda longa.

Um estudo de Dan Reed (REED; AYDT, 1999) mostra um controle fuzzy atuando em conjunto com um sistema supervisor de desempenho para um determinado serviço ou aplicação. O processo de inferência fuzzy recebe diversas variáveis de desempenho, como utilização de disco, memória, processador, e obtém uma resposta que pode ser usada pelo mesmo sistema supervisor para otimização de desempenho. Algumas particularidades fazem do controle fuzzy uma boa opção como solução para problemas em que:

- As medidas de desempenho sejam imprecisas, facilitando a utilização mais de termos qualitativos como “alta utilização do processador” do que números precisos.
- A complexidade do domínio do problema, caracterizado por ambientes de múltiplos recursos, como memória, I/O, disco, processador, rede, seja elevada.
- Múltiplos comportamentos relativos ao desempenho do serviço irão ocorrer em um espaço de tempo, caracterizando muitas vezes uma não-linearidade.

No capítulo anterior mostrou-se que as cargas de trabalho dos serviços Web apresentam características muito particulares, chamadas de invariantes. A ocorrência do fenômeno de rajadas de requisições de clientes é uma invariante que afeta diretamente o desempenho do serviço Web. A abordagem operacional apresentada na seção 3.1 permite representar e analisar a carga de trabalho com a ocorrência de tais fenômenos, utilizando os registros históricos das requisições dos clientes.

Uma das dificuldades encontradas na construção de modelos de carga de trabalho como o estudado na abordagem operacional é a necessidade de registros históricos. Esses modelos são classificados como quantitativos. Entretanto, muitas vezes os responsáveis pelo suporte ao serviço Web não registram essas requisições ou não sabem como analisá-las (GUNTHER, 2002). Uma alternativa é a utilização dos modelos qualitativos, que possibilitem a representação do comportamento de uma carga de trabalho sem necessitar de dados históricos (MENASCÉ; ALMEIDA, 2003b); (GUNTHER, 2002); (MENASCÉ; ALMEIDA, 2000a).

Esse estudo apresenta uma solução baseada em lógica fuzzy como uma alternativa à abordagem operacional para a construção de um modelo de carga de trabalho adaptada ao fenômeno de rajadas, viável e adequado para:

- Análises da carga de trabalho de serviços Web sem nenhum registro histórico disponível. Nesse contexto, a opinião de especialistas em desenvolvimento e suporte ao serviço Web e também dos clientes seria a única fonte de informação sobre o seu desempenho.
- Empresas que não tenham software de monitoramento de desempenho do serviço Web, por seu alto custo e complexidade, nem recursos humanos treinados em planejamento de capacidade e tampouco contam com a possibilidade de contratação de serviços de consultoria externa dessa natureza.
- Produção de resultados rápidos e confiáveis, semelhantes aos obtidos na abordagem quantitativa oferecida pelo modelo operacional, com razoável margem de erro. Esses resultados, mais do que precisão absoluta, devem oferecer uma direção aproximada para análise das implicações que a variação da carga de trabalho com efeito de rajadas trará para o desempenho do serviço, e possam ser utilizados pela gerência de tecnologia da informação para oferecer um serviço de qualidade e desempenho de acordo com as expectativas dos clientes.
- Situações de urgência, para subsidiar a tomada de decisão gerencial, em que uma definição obtida de forma simples e rápida é preferível a uma análise precisa, mas demorada.
- Situações de predição de carga de trabalho e desempenho futuro do serviço Web sob as perspectivas de aumento de requisições de clientes e crescimento da intensidade do fenômeno de rajadas.

As próximas seções apresentam conceitos sobre lógica fuzzy e sistemas de controle fuzzy, utilizados para a implementação dessa solução.

4.1 Lógica Fuzzy

Lógica Fuzzy é uma metodologia para solução de problemas com variáveis e enunciados ambíguos, com expressões contendo imprecisões e incertezas (“talvez”), com presença de ruídos e

com informação incompleta. São baseados em regras lingüísticas e envolvem soluções de problemas em que o raciocínio é executado de forma aproximada (ZADEH, 1987a).

Em muitos aspectos os sistemas fuzzy procuram aproximar a decisão computacional da decisão humana, que é capaz de realizar inferências a partir de informações imprecisas. As premissas imprecisas recebidas pelo sistema fuzzy obtêm conclusões baseadas em inferências não triviais que não poderiam ser obtidas facilmente quando utilizadas técnicas convencionais, como a lógica de 1ª ordem ou a teoria das probabilidades (WANG, 1996).

O domínio dos resultados obtidos por sistemas que utilizam lógica booleana sempre apresenta apenas dois valores, verdadeiro ou falso. Em contraponto, o domínio em sistemas que utilizam lógica fuzzy são multivalorados, reconhecendo diversos valores dentro de um intervalo numérico definido por $(0,1)$. A possibilidade de traduzir os termos difusos da comunicação humana e os diversos ruídos e incertezas de problemas reais para valores que possam ser representados em computadores e obter resultados concisos faz dos sistemas fuzzy úteis em diversas aplicações, que vão desde sistemas de apoio ao diagnóstico médico até chips fuzzy desenvolvidos para eletrodomésticos. Pela natureza de sua aplicação os sistemas fuzzy podem ser divididos em quatro categorias (MUNAKATA, 1994):

1. Reconhecimento de padrões em pequenas quantidades de dados: imagens, voz, textos, processamento de sinais.
2. Controle: é a mais desenvolvida atualmente, com diversas aplicações implementadas. As aplicações utilizam sistemas de controle fuzzy, também conhecidos como controladores nebulosos, para efetuar um mapeamento de variáveis de entrada e inferência de variáveis de saída em um determinado processo.
3. Otimização e análise quantitativa: utilizam sistemas fuzzy, muitas vezes combinados com redes neurais (neuro-fuzzy) e algoritmos genéticos, para resolver problemas de otimização em programação linear e não-linear, programação dinâmica, dentre outras.

4. Análise de Informações: banco de dados utilizando técnicas como mineração de dados (*data mining*), manipulando grandes volumes de dados.

4.2 Conjuntos Fuzzy

O conceito de conjunto fuzzy é fundamental para o entendimento de um sistema fuzzy. Em contraponto à teoria de conjuntos clássica, que denota claramente se um elemento pertence ou não pertence a um conjunto, os conjuntos fuzzy flexibilizam essa noção de pertinência. A função que define a pertinência de um elemento em um conjunto fuzzy atribui um grau de pertinência a esse elemento (BUCKLEY et al., 2004a).

Seja Ω um conjunto clássico e A um conjunto nebuloso subconjunto de Ω . A função $A(x)$, apresentada na equação (4.1), retorna o grau de pertinência de x em relação ao conjunto nebuloso A .

$$A(x) : x \rightarrow [0,1]; x \in \Omega \quad (4.1)$$

4.3 Sistemas de Controle Fuzzy

Basicamente, um sistema de controle fuzzy dispõe de uma estrutura de funcionamento como a mostrada na Figura 4.1. Os valores das variáveis de entrada são obtidos por um sensor ou dispositivo de entrada. Uma interface de “fuzzificação” transforma os valores de entrada em números fuzzy. Os dados necessários para essa conversão estão registrados na base de conhecimento. Esses números são analisados por um processo de inferência, acessada uma base de conhecimento com as regras e fatos ativados pelas variáveis de entrada e uma resposta é gerada. Essa resposta passa por um processo de “defuzzificação” e, então, é remetido a um atuador, que pode ser um outro sistema ou um dispositivo de saída acoplado a outro processo ou sistema computacional. O processo de “defuzzificação” também utiliza a base de conhecimento, onde estão armazenados dados sobre o universo de discurso e os valores dos quantificadores fuzzy (REED; AYDT, 1999) (WANG, 1996).

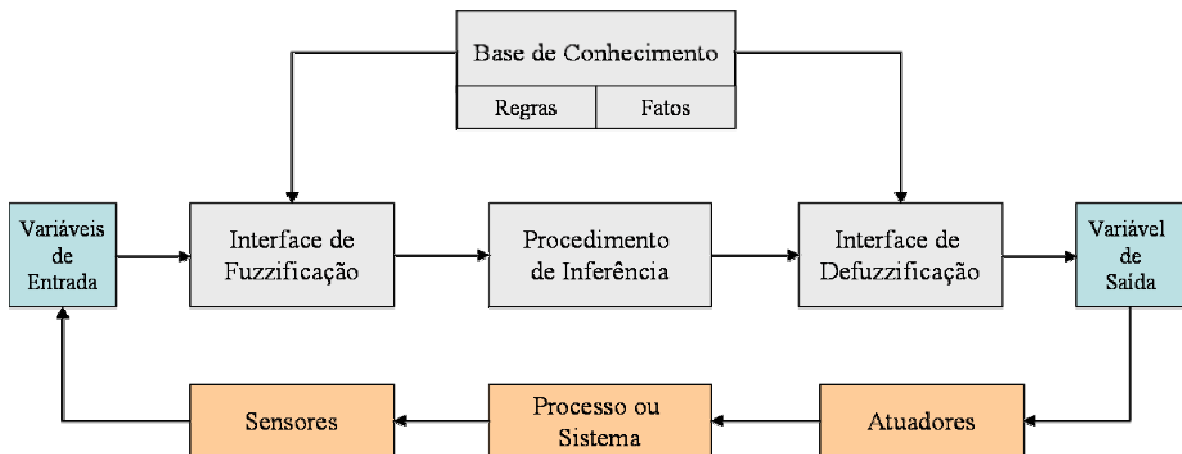


Figura 4.1 - Sistema de Controle Fuzzy.

4.3.1 Base de Conhecimento

Um sistema de controle fuzzy oferece, aos seus usuários, respostas sobre questões que estão representadas em sua *base de conhecimento*, que é formada por uma coleção de *regras* e *fatos*. As informações da base de conhecimento podem ser obtidas através de estimativas de especialistas ou do uso de constantes e probabilidades geradas por análise de dados históricos (BUCKLEY et al., 2004a).

A resposta ou saída obtida por um sistema fuzzy, através de sua base de conhecimento, é associada a um valor, denominado de *fator de certeza*, pertencente a um intervalo $(-1,1)$. Sendo a base de conhecimento formada por fatos e regras, cada uma delas pode ter um *fator de certeza* associado. Uma resposta do sistema associada ao *fator de certeza* 1 indica total concordância. Da mesma forma, uma resposta associada ao valor -1 indica total discordância (ZADEH, 1987b).

Regra associada a um *fator de certeza*, conforme exemplo:

- Se X é A, então, Y é B com $FC = \beta$
- Se X é pequeno, então, Y é grande com $FC = 0.8$

Fato associado a um fator de certeza, conforme exemplo:

- $X \in A'$ com $FC = 0.5$
- $Y \in A'$ com $FC = \gamma$

Além do fator de certeza, a representação das regras e fatos utiliza *variáveis lingüísticas* que possibilitam ilustrar a incerteza de maneira clara. As variáveis lingüísticas são representadas mais por palavras que números, sendo formadas por:

- predicados fuzzy: termos primários que especificam as variáveis sendo medidas.
- quantificadores fuzzy: muito, pouco, alto, baixo, rápido, lento, negativo, positivo, quente, frio, muito alto, extremamente frio, dentre outros.
- conectivos “E” e “OU” e a negação “NÃO”.
- marcadores, como parênteses, adição e subtração.

4.3.2 O Processo de Fuzzificação e as Variáveis Fuzzy

Um controlador fuzzy recebe algumas variáveis de entrada e infere um valor para uma variável de saída. Dada a incerteza do domínio do problema, as variáveis de entrada e saída são representadas como números fuzzy. Esse processo de conversão de um sinal de entrada em um número fuzzy é chamado de fuzzificação.

As variáveis fuzzy podem ser definidos em intervalos de pertinência e em diversas formas (*shapes*). Dentre as formas mais comuns, destacam-se as formas triangulares e trapezoidais. Tal qual o fator de certeza, os valores de pertinência dos números fuzzy podem ser obtidos pela análise estatística dos registros históricos disponíveis no domínio do problema ou pela opinião de especialistas (BUCKLEY et al., 2004a); (MUNAKATA, 1994)

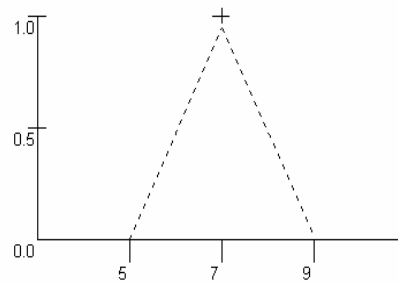


Figura 4.2 - Forma triangular.

A Figura 4.2 representa uma variável fuzzy na chamada forma triangular. Essa forma pode ser definida a partir de três números $a < b < c$ que formam a base do triângulo no intervalo $[a,c]$ e o vértice em b .

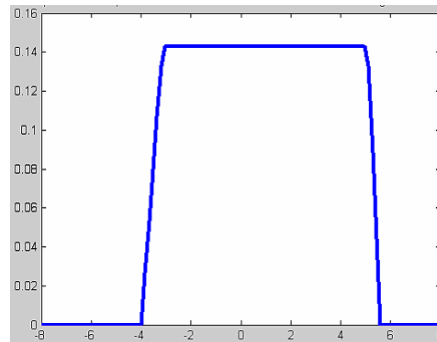


Figura 4.3 - Forma Trapezoidal.

A Figura 4.3 mostra uma variável fuzzy no formato trapezoidal definido pelos números $(a < b < c < d)$. O topo é delimitado no intervalo $[a,d]$.

4.3.3 Inferência em Lógica Fuzzy

A capacidade de inferência é uma das principais características de um controle fuzzy. Basicamente, após analisar as diversas variáveis de entrada, o processo de inferência tem como resultado final uma região de controle nebulosa relacionada com a saída do processo. Todo o suporte ao gerenciamento da incerteza, através do processo de inferência, é oferecido pela base de conhecimento do sistema fuzzy (ZADEH, 1987b).

O controle fuzzy utiliza um algoritmo para realizar o processo de inferência, que pode ser assim definido em 2 passos:

- 1) Encontrar todas as regras ativadas e determinar a saída nebulosa para cada uma.
- 2) Combinar todas as saídas nebulosas.

O exemplo de inferência da Tabela 4.1 mostra que após a análise da entrada de uma variável X, duas regras podem ser ativadas pela confirmação de validade da sentença que define a regra. A validação pode ser feita através de uma consulta à base de conhecimento do controle fuzzy que define o resultado de saída de cada regra.

Tabela 4.1- Exemplo de inferência fuzzy

Regra 1 - Se X é F, então, Y é G

Entrada: X é F

Resultado do processo de inferência: Y é G

Regra 2 - Se (X é A2) e (Y é B2), então, Z é C1

Entrada: X é A2

Entrada: Y é B2

Resultado do processo de inferência: Z é C1

Após serem detectadas, todas as regras ativas são combinadas com uma relação fuzzy simples, união ou interseção. Essas combinações podem ser a de Mamdani (norma T) ou de Gödel (norma S) (ZADEH, 1973; 1987). O método de Mamdani testa a relevância de cada regra ativa e seleciona as mais importantes (MAMDANI, ASSILIAN; 1999). Uma das características desse método é a sua facilidade de implementação, descrita a seguir.

Sejam dois conjuntos fuzzy X e Y , que podem ter universos de discurso distintos, combinados pela seguinte equação:

$$X \Rightarrow Y \equiv X \circ .\min Y \quad (4.2)$$

Na equação 4.2, $\circ .\min$ é produto da função “min” aplicada em cada elemento do produto cartesiano de X e Y . O exemplo da tabela 4.2 mostra o resultado de $\circ .\min$ no vetor x e y , representados nas linhas e colunas.

Tabela 4.2 - Método de Mamdani

$\circ .\min$	y1	y2	...	yn
x1	$x1 \wedge y1$	$x1 \wedge y2$...	$x1 \wedge yn$
x2	$x2 \wedge y1$	$x2 \wedge y2$...	$x2 \wedge yn$
...
xn	$xn \wedge y1$	$xn \wedge y2$...	$xn \wedge yn$

4.3.4 Defuzzificação: Método do Centro de Área

A resposta dada por um controle fuzzy é representada por um número legível pelo domínio do problema. O processo de defuzzificação analisa o resultado obtido na composição das regras que levam a esse número. Segundo Wang (WANG, 1996), os principais defuzzificadores são: centro de gravidade, centro ponderado e máximo. O método do centro de massa é o mais usado, e favorece a regra com maior participação na área resultante da composição das regras ativas.

A Figura 4.4 mostra a área obtida pela composição das regras de um controlador fuzzy. O valor de saída foi defuzzificado pela técnica do centro de massa e é representado por uma marca no centróide do gráfico. A projeção do centróide, na coordenada horizontal, define o valor da variável de saída. No exemplo da Figura 4.4 esse ponto está bem próximo do valor zero.

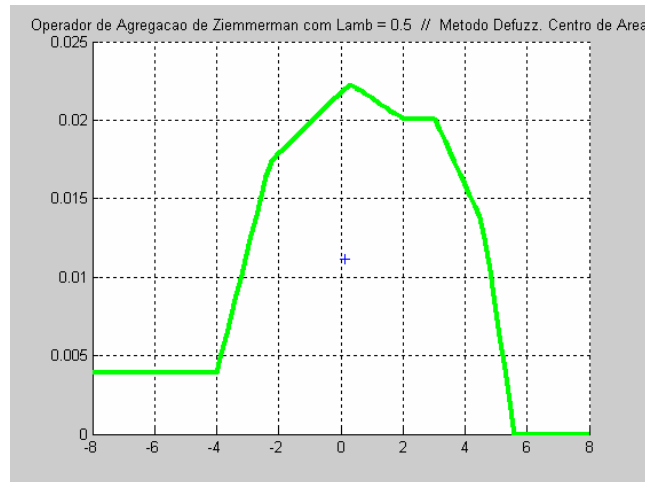


Figura 4.4 - Método do Centro de Área

A equação 4.3 calcula o centro de área de um determinado conjunto fuzzy definido pela variável de saída. O universo de discurso determina os pontos representados por y_i , e a função $\mu(y_i)$ define o valor de pertinência desse ponto para o conjunto fuzzy.

$$y(t) = \frac{\sum_i \mu(y_i) y_i}{\sum_i \mu(y_i)} \quad (4.3)$$

4.4 Considerações Finais

Neste capítulo foram apresentados os conceitos de conjuntos fuzzy e sistemas de controle fuzzy. Destacou-se a base de conhecimento, que é utilizada nos processos de inferência e conversão de números fuzzy, e duas formas usadas para a conversão de números fuzzy, a trapezoidal e triangular. Mostrou-se um método para agregação, Mandani, e para defuzzificação, Centro de Área.

No próximo capítulo será descritas a documentação dos controles fuzzy RAJIN e CSWeb, utilizados na implementação do modelo de carga de trabalho com ocorrência do fenômeno de rajadas.

5 Controle Fuzzy para Modelagem de Carga em Serviços Web

O capítulo 4 abordou conceitos sobre sistemas de controle fuzzy e suas aplicações no planejamento de capacidade. Esse capítulo apresenta a especificação detalhada desse modelo da carga de trabalho, implementado por dois controles fuzzy, chamados de RAJIN (abreviatura para Intensidade de Rajadas) e CSWeb (abreviatura para Capacidade do Serviço Web).

5.1 Fundamentação Geral

Definiu-se para desenvolvimento dos controles fuzzy RAJIN e CSWeb diversos dados, que são ser descritos nas próximas seções. Essas informações são úteis para a implementação computacional do controle e nos ajustes necessários para os controles fuzzy representarem o comportamento da carga de trabalho.

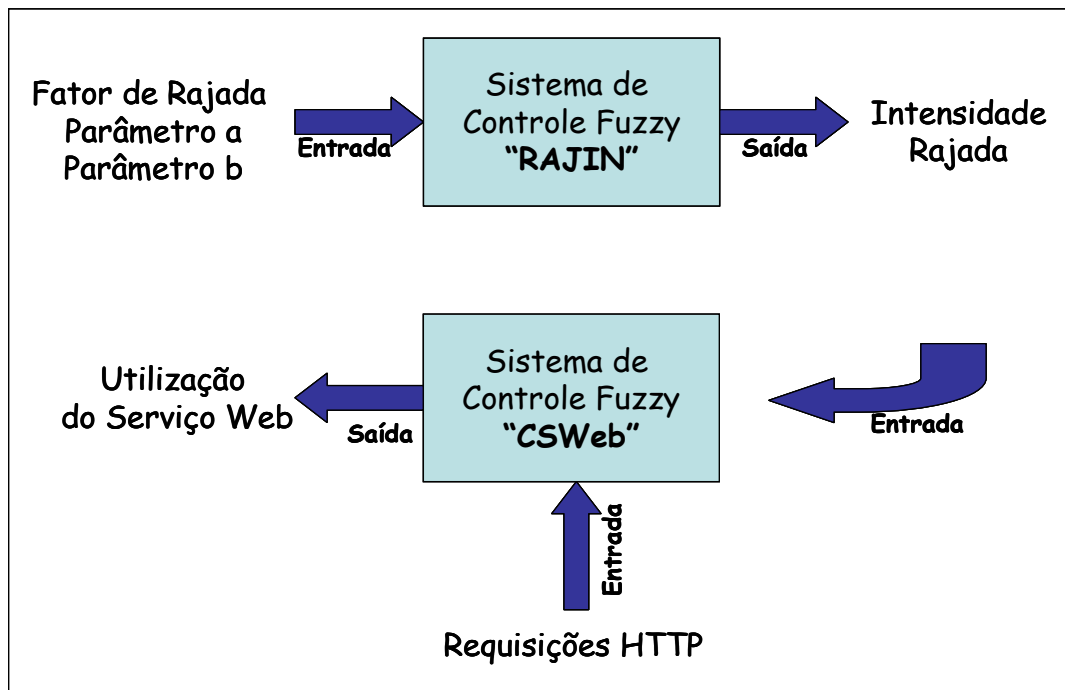


Figura 5.1 – Controle Fuzzy RAJIN e CSWeb.

A **Figura 5.1** mostra o fluxo de processamento e as variáveis de entrada e de saída dos dois controles fuzzy. O controle fuzzy RAJIN é usado para obter um número percentual que indique a intensidade da rajada ocorrida em um intervalo de tempo analisado. Para isso recebe como variáveis de entrada os parâmetros (a,b) , apresentados na Seção 3.1 como fator de rajada. O controle fuzzy CSWeb recebe como variável de entrada a intensidade da ocorrência da rajada e a taxa de chegada de requisições http. O resultado do processo de inferência realizado pelo controle é a variável de saída que quantifica a utilização do serviço Web.

As variáveis de entrada e saída dos controles fuzzy foram especificadas com as seguintes informações:

- Nome, sigla e descrição: informações para identificar as variáveis, descrevendo suas principais características.
- Quantificadores Fuzzy: identificam as opiniões recolhidas para a especificação do modelo para os estados que as variáveis podem assumir, sob o ponto de vista qualitativo (alta, muito baixa, média). Alguns estudos de Reed (1999) e Buckley (2004) abordam quantificadores específicos para o planejamento de capacidade.
- Universo de Discurso: as informações do universo de discurso podem ser fornecidas pelo especialista em planejamento de capacidade para serviços Web ou pelo arquiteto de sistemas que projetou o serviço Web. Entrevistas com o usuário também podem ser úteis nesse processo. O universo de discurso deve representar, de maneira aproximada, de acordo com a percepção do especialista, um intervalo com os valores numéricos que podem ser atribuídos às variáveis de entrada e saída.
- Formas (*shapes*): usadas no processo de fuzzificação das variáveis, são transformadas em números fuzzy. Para os dois controles fuzzy foram escolhidas formas trapezoidais e triangulares, comumente utilizadas em soluções para planejamento de capacidade que utilizem lógica fuzzy devido a sua fácil definição e implementação.

Tabela 5.1 - Operadores utilizados nos Controles Fuzzy.

Operador	Tipo
Operador de Agregação	Operador Compensatório de Zimmerman com $\lambda = 0.5$
Operador de Implicação	Mamdani (T Norma – Min)
Operador de Defuzzificação	Centro de Área

A **Tabela 5.1** lista os operadores usados na implementação dos controles fuzzy. A máquina de inferência do controle utilizou a norma T (Mamdani), para determinar as regras que foram ativadas pelas variáveis de entrada.. A agregação das regras foi implementada pelo operador de Zimmerman. Os dois operadores foram escolhidos por melhor se adequarem ao problema (DIAO et al., 2002).

O processo de defuzzificação foi implementado utilizando o método do centro de área, escolhido devido à sua grande aplicação em controles fuzzy. O método apresenta como resultado um valor de saída simples que facilita a análise por ser bastante intuitivo (SILVEIRA, 2005).

5.2 Sistema de Controle Fuzzy Rajin

O sistema de controle fuzzy Rajin é utilizado para quantificar, em valores percentuais, a intensidade da ocorrência do fenômeno de rajadas em uma carga de trabalho em um determinado intervalo de tempo. Ela utiliza valores dos parâmetros (a,b), definidos na abordagem operacional apresentada na Seção 3.1.

Nas subseções seguintes serão apresentadas as especificações utilizadas para implementar as variáveis de entrada e saída, a base de conhecimento composta de fatos e regras definidos por especialistas e os operadores utilizados na ativação, agregação e defuzzificação.

5.2.1 Definição das Variáveis de Entrada e Saída

A Tabela 5.2 apresenta a definição das variáveis de entrada e de saída. A cada variável são atribuídas uma descrição e sigla, ambas descritas nessa tabela. Os parâmetros (a,b) foram definidos como duas variáveis de entrada, representados pelas siglas FRAJA e FRAJB, fornecidas ao controle fuzzy para desencadear o processo de inferência.. Como resultado, o controle produz uma variável de saída que indica a intensidade da rajada, representada pela sigla RAJ. Essa variável foi implementada para facilitar a definição, na forma percentual, da intensidade em que ocorre o fenômeno de rajadas em uma determinada carga de trabalho. Por representar uma escala percentual, o universo de discurso da variável RAJ foi definido entre 0 e 100.

Tabela 5.2 - Variáveis de entrada e saída – Controle Fuzzy Rajin.

	Id	Variável	Objetivo	Quantificadores Fuzzy	Universo de Discurso	Formas (Shapes)
Entrada	FRAJA	Parâmetro a	Indica o número médio de requisições realizadas pelos clientes do serviço Web durante um intervalo de tempo.	Baixa (B) [0;1,5] Média (M) [1,5;4,5] Alta (A) [4,5;6]	[0, 6]	Trapezoidal Triangular
	FRAJB	Parâmetro b	Indica o percentual de tempo que a taxa de requisição recebida pelo serviço Web ficou acima da média do intervalo de tempo analisado.	Baixa (B) [0;12,5] Média (M) [12,5; 37,5] Alta (A) [37,5;5 0]	[0, 50]	Trapezoidal Triangular
Saída	RAJ	Intensidade ocorrência Fenômeno Rajadas	Indica o volume, em percentual, da intensidade da ocorrência do fenômeno de rajadas nas requisições http dos clientes de um serviço Web.	Baixa (B) [0;25] Média (M) [25;75] Alta (A) [75;100]	[0, 100]	Trapezoidal Triangular

A Figura 5.2 ilustra a variável de entrada FRAJA, com o universo de discurso definido entre 0 e 6. Esses valores foram obtidos por opinião de especialistas e encontrados em estudos de alguns autores, como Arlitt et al. (1996; 2001) e Banga (1997; 1999). A Figura 5.3 ilustra a variável de entrada FRAJB, com o universo de discurso definido entre 0 e 50, em escala percentual. Estudos de Almeida et al. mostram que valores percentuais superiores a 50 são relacionados a um aumento da demanda da carga de trabalho e não caracterizam a ocorrência do fenômeno de rajadas (MENASCÉ et al., 2004).

Cada variável possui três quantificadores fuzzy, definidos qualitativamente em alto, médio e baixo. A Tabela 5.2 registra, na coluna “Quantificadores Fuzzy”, os intervalos em que cada informação qualitativa é válida. As figuras 5.2, 5.3 e 5.4 ilustram, respectivamente, o processo de fuzzificação das variáveis FRAJA, FRAJB e RAJ. O número fuzzy obtido utiliza as formas trapezoidal e triangular. No gráfico, a linha de cor verde representa o quantificador fuzzy com o valor qualitativo “baixo”, a linha de cor azul o quantificador com o valor “médio” e a linha vermelha o quantificador com o valor “alto”.

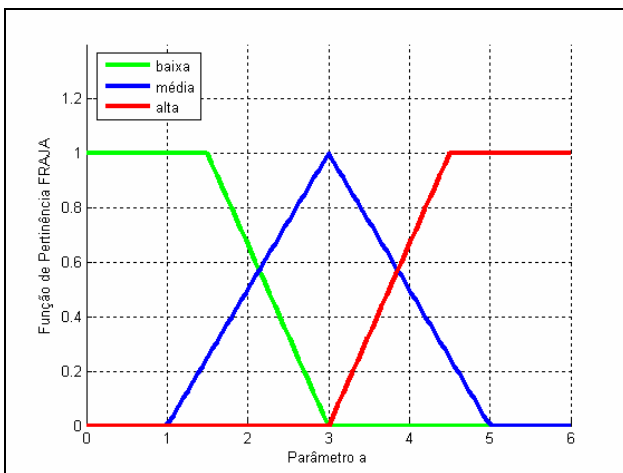


Figura 5.2 - Variável de Entrada FRAJA - Parâmetro a.

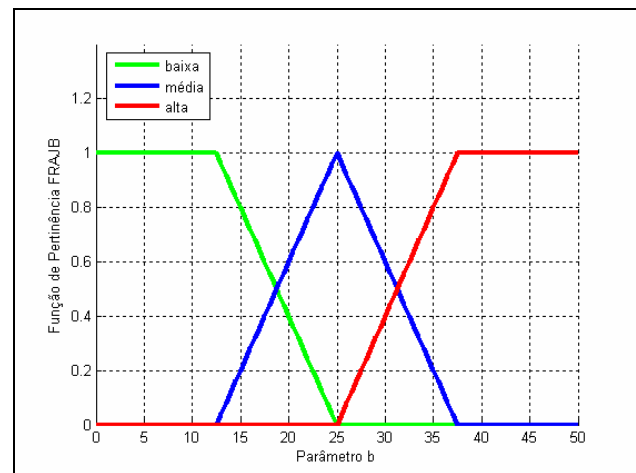


Figura 5.3 - Variável de Entrada FRAJB - Parâmetro b.

A Figura 5.4 ilustra a variável de saída RAJ. Essa variável assume um valor, resultante do processo de defuzzificação, que será pertencente ao intervalo (0,100). Esse resultado é uma representação, na forma de um número, da percepção intuitiva do usuário e do especialista em planejamento de capacidade da ocorrência do fenômeno de rajadas em uma determinada carga de trabalho.

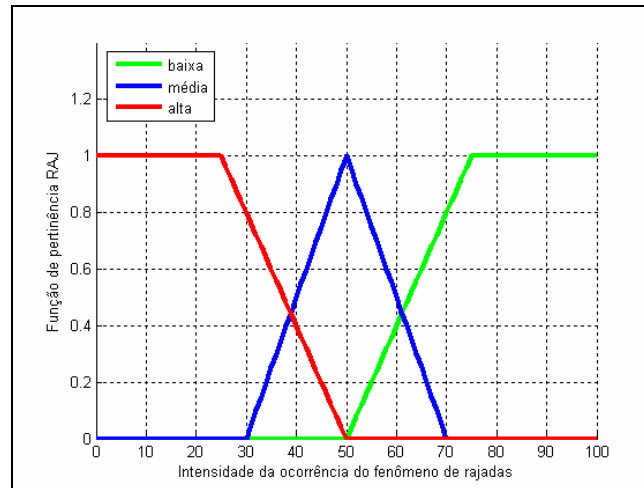


Figura 5.4 - Variável de Saída – RAJ – Intensidade do Fenômeno de Rajadas.

5.2.2 Base de Dados e de Conhecimento

A Tabela 5.3 apresenta a base de dados com as regras de inferência para o sistema de controle fuzzy. Foram definidas sete regras, a partir da opinião de especialistas em serviços Web. Todas as regras e fatos usados no processo de inferência do controle foram definidos pela combinação das variáveis FRAJA e FRAJB. A partir da obtenção dos valores dos quantificadores fuzzy de ambas variáveis, é possível confirmar ou negar um determinado fato.

Tabela 5.3 - Base de Conhecimento (Regras e Fatos) - Controle Fuzzy Rajin.

Regras e Fatos	
1	SE Parâmetro a (FRAJA) é Alto E Parâmetro b (FRAJB) é Alto, ENTÃO , Intensidade de Rajadas (RAJ) é Alta
2	SE Parâmetro a (FRAJA) é Alto E Parâmetro b (FRAJB) é Médio, ENTÃO , Intensidade de Rajadas (RAJ) é Média
3	SE Parâmetro a (FRAJA) é Alto E Parâmetro b (FRAJB) é Baixo, ENTÃO , Intensidade de Rajadas (RAJ) é Média
4	SE Parâmetro a (FRAJA) é Médio E Parâmetro b (FRAJB) é Alto, ENTÃO , Intensidade de Rajadas (RAJ) é Média
5	SE Parâmetro a (FRAJA) é Médio E Parâmetro b (FRAJB) é Médio, ENTÃO , Intensidade de Rajadas (RAJ) é Alta
6	SE Parâmetro a (FRAJA) é Médio E Parâmetro b (FRAJB) é Baixo, ENTÃO , Intensidade de Rajadas (RAJ) é Média
7	SE Parâmetro a (FRAJA) é Baixo E Parâmetro b (FRAJB) é Baixo, ENTÃO , Intensidade de Rajadas (RAJ) é Baixa

A Tabela 5.4 representa a matriz de regras de inferência. Ela é construída para facilitar a representação computacional de regras e fatos, respeitando suas características definidas na base de conhecimento pelos termos “Se-Então-Senão”. Cada linha e coluna representam uma variável de entrada e a intersecção de ambas é o valor do quantificador fuzzy assumido pela variável de saída.

Esse modelo de representação permite registrar as regras de inferência em um vetor com a mesma dimensão dos elementos da linha e coluna da matriz (COX, 1999).

Tabela 5.4 - Matriz Regras de Inferência - Controle Fuzzy Rajin.

FRAJA	FRAJB Alta	FRAJB Média	FRAJB Baixa
Baixa	NULA	NULA	RAJ Baixa
Média	RAJ Média	RAJ Alta	RAJ Média
Alta	RAJ Alta	RAJ Média	RAJ Média

Definiram-se duas posições nulas na matriz, em decorrência de informação obtida com o especialista em planejamento de capacidade. Considerou-se, nessa representação, que não seria inferida ocorrência parâmetro *b* alto ou médio com parâmetro *a* baixo.

5.3 Controle Fuzzy CSWeb

O sistema de controle fuzzy CSWeb recebe duas variáveis de entrada e infere a utilização do serviço Web. O controle CSWeb, em conjunto com o controle RAJIN, implementa o modelo de carga de trabalho do serviço Web com a ocorrência do fenômeno de rajadas. Uma das variáveis é a quantidade de requisições http efetuadas pelos clientes do serviço Web. A segunda variável é a intensidade da ocorrência do fenômeno de rajadas em uma carga de trabalho, valor determinado pelo controle fuzzy RAJIN.

A definição de utilização do serviço Web foi apresentada na Seção 3.1. Nas subseções seguintes, serão apresentadas as especificações utilizadas para implementar as variáveis de entrada e de saída, a base de conhecimento composta de fatos e regras definidos por especialistas e os operadores utilizados na ativação, agregação e defuzzificação das regras ativadas.

5.3.1 Variáveis de Entrada e de Saída

A Tabela 5.5 apresenta a definição das variáveis de entrada e de saída. A cada variável são atribuídas uma descrição e sigla, ambas descritas nessa tabela. A quantidade de requisições http e a intensidade da rajada foram definidas como duas variáveis de entrada, representadas pelas siglas NHTTP e RAJ, fornecidas ao controle fuzzy para desencadear o processo de inferência. Como resultado o controle produz uma variável de saída que indica a utilização do serviço Web, representada pela sigla USW. Essa variável foi implementada para facilitar a definição na forma percentual da utilização do serviço Web e como a mesma é afetada com a ocorrência do fenômeno de rajadas. Por representar uma escala percentual, o universo de discurso da variável USW foi definido entre 0 e 100.

A Tabela 5.5 também define o universo de discurso das variáveis de saída e entrada. As variáveis de entrada NHTTP (Número de Requisições HTTP dos clientes do Serviço Web) e RAJ (Intensidade da ocorrência do Fenômeno de Rajadas) assumirão valores entre (0,200) e (0,50). Nas

variáveis de entrada, esses intervalos definem os valores inseridos no sistema e que são utilizados no processo de fuzzificação. O universo de discurso da variável NHTTP foi definido por especialistas em planejamento de capacidade em serviços Web.

Tabela 5.5 - Variáveis de entrada e saída – Controle Fuzzy CSWeb.

	Id	Variável	Objetivo	Quantificadores Fuzzy	Universo de Discurso	Formas (Shapes)
Entrada	NHTTP	Número de Requisições Http dos Clientes do Serviço Web.	Indica o número médio de requisições realizadas pelos clientes do serviço Web durante um intervalo de tempo.	Baixa (B) [0,50] Média (M) [50-150] Alta (A) [150,200]	[0, 200]	Trapezoidal Triangular
	RAJ	Intensidade da Ocorrência do Fenômeno de Rajadas	Indica o volume, em percentual, da intensidade da ocorrência do fenômeno de rajadas nas requisições http dos clientes de um serviço Web.	Baixa (B) [0,25] Média (M) [25,75] Alta (A) [75,100]	[0, 100]	Trapezoidal Triangular
Saída	USW	Utilização do Serviço Web.	Indica o percentual de utilização da capacidade do serviço Web.	Baixa (B) [0,25] Média (M) [25,75] Alta (A) [75,100]	[0, 100]	Trapezoidal Triangular

Cada variável possui três quantificadores fuzzy, definidos qualitativamente em alto, médio e baixo. A Tabela 5.5 registra, na coluna “Quantificadores Fuzzy”, os intervalos em que cada informação qualitativa é válida. As figuras 5.5, 5.6 e 5.7 registram, respectivamente, o processo de fuzzificação das variáveis NHTTP, RAJ e USW. O número fuzzy obtido utiliza as formas trapezoidal e triangular. No gráfico, a linha de cor verde representa o quantificador fuzzy com o valor qualitativo “baixo”, a linha de cor azul o quantificador com o valor “médio” e a linha vermelha o quantificador com o valor “alto”.

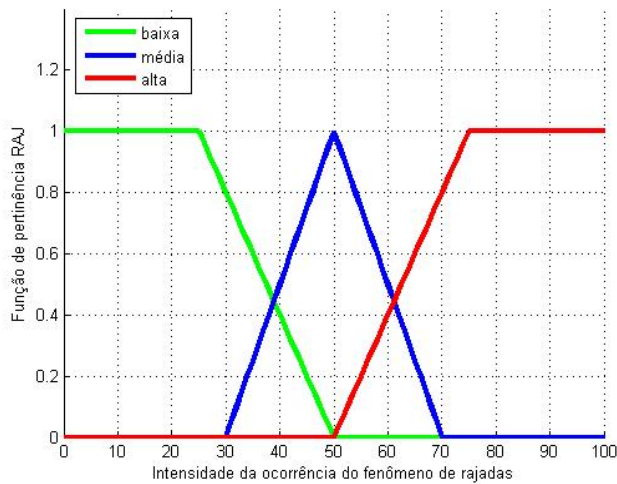


Figura 5.5 – RAJ: Intensidade do Fenômeno de Rajadas.

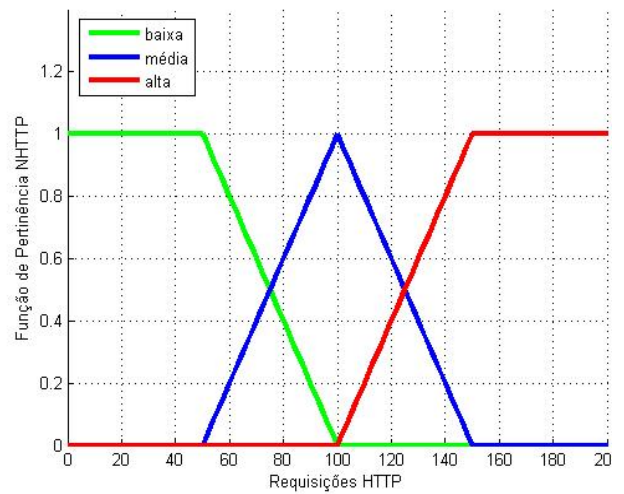


Figura 5.6 – NHTTP: Requisições http.

A Figura 5.7 ilustra a variável de saída USW. Essa variável assume um valor, resultante do processo de defuzzificação, que será pertencente ao intervalo (0,100). Essa é representação, na forma de um número fuzzy, da percepção intuitiva do usuário e do especialista em planejamento de capacidade da ocorrência do fenômeno de rajadas em uma determinada carga de trabalho.

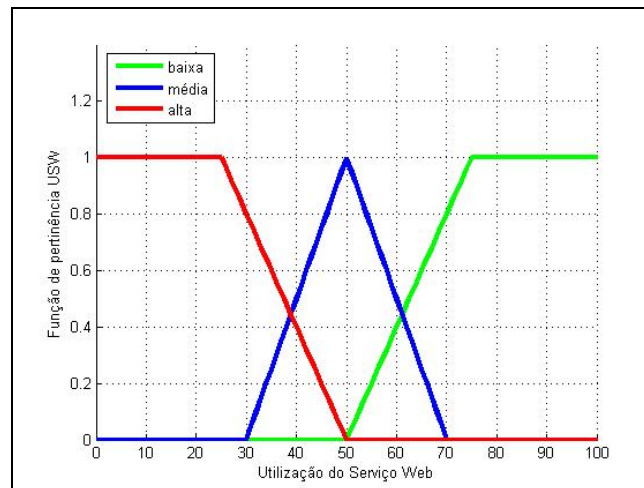


Figura 5.7 – Variável de Saída - USW – Utilização do Serviço Web.

5.3.2 Base de Dados e de Conhecimento

A Tabela 5.6 apresenta a base de dados com as regras de inferência para o sistema de controle fuzzy. Foram definidas nove regras, obtidas a partir da opinião de especialistas em planejamento de capacidade para serviços Web.

Tabela 5.6 - Base de Conhecimento – Controle Fuzzy CSWeb

Regra	
1	SE Requisição do Usuário (NHTTP) é Alta E Intensidade de Rajadas (RAJ) é Alta, ENTÃO , Utilização do Serviço (USW) é Alta
2	SE Requisição do Usuário (NHTTP) é Alta E Intensidade de Rajadas (RAJ) é Média, ENTÃO , Utilização do Serviço (USW) é Alta
3	SE Requisição do Usuário (NHTTP) é Alta E Intensidade de Rajadas (RAJ) é Baixa, ENTÃO , Utilização do Serviço (USW) é Alta
4	SE Requisição do Usuário (NHTTP) é Média E Intensidade de Rajadas (RAJ) é Alta, ENTÃO , Utilização do Serviço (USW) é Alta
5	SE Requisição do Usuário (NHTTP) é Média E Intensidade de Rajadas (RAJ) é Média, ENTÃO , Utilização do Serviço (USW) é Alta
6	SE Requisição do Usuário (NHTTP) é Média E Intensidade de Rajadas (RAJ) é Baixa, ENTÃO , Utilização do Serviço (USW) é Média
7	SE Requisição do Usuário (NHTTP) é Baixa E Intensidade de Rajadas (RAJ) é Baixa, ENTÃO , Utilização do Serviço (USW) é Baixa
8	SE Requisição do Usuário (NHTTP) é Baixa E Intensidade de Rajadas (RAJ) é Média, ENTÃO , Utilização do Serviço (USW) é Baixa
9	SE Requisição do Usuário (NHTTP) é Baixa E Intensidade de Rajadas (RAJ) é Alta, ENTÃO , Utilização do Serviço (USW) é Média

O controle fuzzy CSWeb possui duas variáveis de entrada e uma variável de saída, que fornece a solução do sistema. Essa particularidade permite pensar no controle como uma matriz de regras de inferência, representada na Tabela 5.7. A matriz de regras de inferência é construída a partir das regras e fatos da base de conhecimento. A linha da matriz representa a variável de entrada RAJ e a coluna a variável NHTTP. A intersecção da linha e coluna da matriz é um estado do quantificador fuzzy da variável de saída USW, definido como alto, médio ou baixo.

Os valores de M e N são determinados pela quantidade de quantificadores fuzzy de cada variável que forma a linha e coluna. No controle CSWeb, $M = 3$ e $N = 3$, entretanto, nem todas intersecções de linha e coluna precisam ser definidas com um estado. Na matriz de regras de inferência da Tabela 5.7 duas posições são nulas. As informações sobre o valor do quantificador fuzzy definido pela intersecção da linha e coluna da matriz são fornecidas também por especialistas do planejamento de capacidade do serviço Web e também por usuários.

Tabela 5.7 - Matriz Regras de Inferência - Controle Fuzzy CSWeb.

NHTTP	RAJ Alta	RAJ Média	RAJ Baixa
Baixa	USW Média	USW Baixa	USW Baixa
Média	USW Alta	USW Alta	USW Média
Alta	USW Alta	USW Alta	USW Alta

5.4 Implementação Computacional e Considerações Finais

Utilizou-se na implementação dos controles fuzzy RAJIN e CSWeb o software MATLAB 7.0.1 R14 (MATLAB, 2005), escolhido por apresentar uma linguagem simples e diversas bibliotecas para gerar gráficos, que foram utilizados em todos os processos do controle fuzzy: fuzzificação, agregação, defuzzificação. Porém, não foi utilizado a biblioteca Fuzzy presente nessa ferramenta. Decidiu-se programar todas as funções utilizadas para a implementação dos controles fuzzy, com o objetivo de conhecer em profundidade todo os detalhes necessários à implementação computacional de uma solução desse tipo. Descreve-se a seguir, de forma bastante superficial, os controles aqui implementados.

A Figura 5.8 mostra o código utilizado para realizar o processo de fuzzificação da variável de entrada NHTTP, pertencente ao controle fuzzy CSWeb. O algoritmo utiliza duas formas (“*shapes*”) utilizadas para representar o número fuzzy, trapezoidal e triangular, utilizando parâmetros definidos nos vetores “baix”, “med” e “alt”. O grau de pertinência, retornado pela função triangular ou trapezoidal, é armazenado nos vetores “fpNHTTP_baixa”, “fpNHTTP_media”

e “fpNHTTP_alta”. Todas as variáveis de entrada e de saída, de ambos os controles fuzzy, foram implementadas por funções com o mesmo algoritmo apresentado na Figura 5.8.

```

% Calcula os valores da função de pertinência da
% variável de entrada NHTTP - quantidade de requisições http recebidas
% dos clientes do serviço web.

% utiliza shape triangular / trapezoidal nas bordas
index = 0;
for x=NHTTpa:(NHTTPb-NHTTpa)/I:NHTTPb
    index = index + 1;
    intervalo_HTTP (index) = x;

    y = 0;

    fpNHTTP_baixa (index) = 0;
    fpNHTTP_media (index) = 0;
    fpNHTTP_alta (index) = 0;

    % Baixa
    if x >= baix(1) & x <= baix(2)
        fpNHTTP_baixa (index) = 1;
    end
    if x > baix(2) & x <= baix(3)
        fpNHTTP_baixa (index) = (baix(4) - x) / baix(5);
    end

    % Media
    if x > med(1) & x <= med(2)
        fpNHTTP_media (index) = (x - med(4)) / med(5);
    end
    if x > med(2) & x <= med(3)
        fpNHTTP_media (index) = (med(6) - x) / med(7);
    end

    % Alta
    if x > alt(1) & x <= alt(2)
        fpNHTTP_alta (index) = (x - alt(4)) / alt(5);
    end
    if x > alt(2) & x <= alt(3)
        fpNHTTP_alta (index) = 1;
    end
end
end

```

Figura 5.8 – Implementação da variável de entrada NHTTP.

A Figura 5.9 mostra um detalhe do código do controle fuzzy CSWeb realizando o processo de defuzzificação. Nessa etapa, as regras ativas estão representadas em vetores que são agregados na função “opmand”, que implementa método de Mamdani, na etapa de inferência. A agregação das regras ativas é realizada pela função “ziemmerman” e a etapa de defuzzificação é implementada pelo método do centro de massa, através das funções “defuzz” e “max”.

```
% base de conhecimento - determina se regra deve ser
% ativada ou não.
opmand = mandani(RAJ,fpraj_alta,int_raj,fpNHTTP_baixa,int_HTTP);

% agregação das regras ativas.
Agrega_AB = ziemmerman(lamb,opmand1,opmand2);

% defuzzificação - aplica método do centro de área.
Def_AB = defuzz(Agrega_AB,intervalo_USW);
MeioCentroArea = max(Agrega_AB) / 2;
```

Figura 5.9 – Algoritmos do controle fuzzy CSWeb.

Os controles fuzzy não utilizaram, em sua codificação, bibliotecas como a “Toolbox Fuzzy”, do software MATLAB. Como consequência, o tempo consumido na programação dos controles foi bastante elevado. Constatou-se que o tempo dispendido nesse esforço poderia ser melhor aplicado na análise da eficiência dos operadores utilizados na implementação do controle fuzzy.

Com a descrição dos controles fuzzy RAJIN e CSWeb, utilizados na implementação do modelo de carga de trabalho com ocorrência do fenômeno de rajadas, bem como a base de conhecimento usada no processo de inferência, as formas e o universo de discurso das variáveis de entrada e saída, é possível fazer, no próximo capítulo, as análises dos resultados obtidos pelos modelos de carga de trabalho utilizando a abordagem operacional e o controle fuzzy.

6 Análise de Resultados

O objetivo principal deste capítulo é a avaliação do comportamento dos modelos determinístico e fuzzy, quando submetidos a cargas originadas por usuários, que demandam tanto serviços comuns da Internet quanto serviços diferenciados para comunidades específicas. Descreve-se o ambiente de teste dos serviços Web utilizados, seu padrão de armazenamento de dados históricos e um estudo da evolução histórica de sua carga de trabalho. O fenômeno de rajadas é analisado em diversas amostras dos serviços Web, utilizando o modelo determinístico. Por fim, a mesma análise das amostras é realizada pelo modelo fuzzy, e comparações sobre os resultados das duas implementações são apresentadas.

6.1 Ambiente de Teste

Para implementar o modelo determinístico descrito nas seções anteriores e efetuar a análise dos resultados obtidos, escolheu-se uma comunidade composta por funcionários, alunos e professores de uma Instituição de Ensino Superior que oferece diversos cursos de graduação, com alguns serviços Web acessíveis somente ao público interno, facilitando, portanto, a restrição de acessibilidade do grupo.

Foi possível dividir os usuários da comunidade em dois grupos, um contendo apenas os alunos dos cursos de graduação e outro contendo todo o coletivo de alunos, professores e funcionários técnico-administrativos. A partir dessa divisão, foi possível determinar os serviços Web acessados pelos dois grupos de maneira distinta. Seguindo esse critério, foram selecionados dois serviços Web:

1. GA (Gestão de Alunos), que é acessado pelos discentes dos cursos da instituição para consultas de notas, freqüências, documentação e emissão de requerimentos. O GA é um sistema cliente-servidor formado por uma base de dados única com informações dos discentes, docentes e funcionários da Instituição. O serviço GA pode ser acessado pelos alunos a qualquer momento do dia ou da noite, pois é disponibilizado também na Internet.

2. Cachê de Páginas Web (Proxy), que é acessado por aproximadamente 3.000 usuários da instituição, formados por funcionários técnico-administrativos, docentes e discentes. Esse serviço fornece todo conteúdo Internet da instituição: páginas, arquivos com áudio e vídeo, comércio eletrônico, motores de busca, pesquisadores automáticos e bases bibliográficas on-line, dentre outras necessidades da comunidade.

A Figura 6.1 descreve a infra-estrutura que dá suporte a esses serviços Web. Nela observa-se que os serviços são disponibilizados por servidores distintos, instalados numa mesma rede. Essa rede está conectada à Internet através de um firewall, e os usuários dos serviços estão localizados em pontos da rede interna, pois o serviço Proxy está disponível somente na rede interna da instituição.

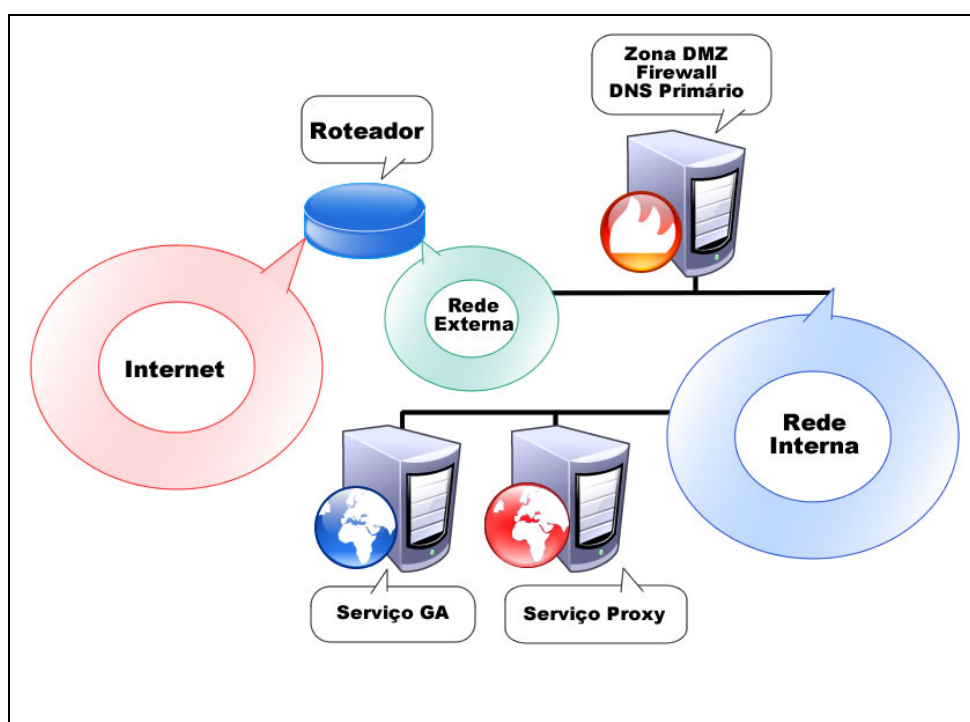


Figura 6.1 - Serviços Web analisados - GA e Proxy.

A Tabela 6.1 traz a especificação do hardware dos servidores dedicados em que estão hospedados os serviços Web. Os servidores são dedicados somente a esses serviços, separadamente. Essa característica facilita o estudo de carga, pois toda carga de trabalho recebida nesses computadores é exclusiva para cada serviço Web.

Tabela 6.1 - Configuração de hardware dos serviços Web.

GA	Proxy
AMD 1.7 GHz Memória - 512 MB Disco Rígido - 30 GB IDE Windows 2000 Advanced Server Servidor Web IIS 5.0 MS Sql Server 2000	Intel Pentium II Dual 500 MHz Memória - 512 MB Disco Rígido 18.3 GB SCSI Red Hat Linux 9.0 Squid Versão 2.6

6.1.1 Instrumentação dos Registros Históricos

Uma condição necessária ao desenvolvimento do modelo determinístico da carga de trabalho de um serviço Web é a disponibilidade dos registros históricos. Uma dificuldade encontrada no processo de planejamento de capacidade é a padronização desses registros. O pouco conhecimento das ferramentas de medição por parte dos administradores de sistemas é outro fator que dificulta a construção do modelo, pois muitas vezes não há uma série temporal disponível com os registros históricos do serviço.

A seguir estão descritos os registros históricos disponíveis para cada um dos serviços analisados:

a) Serviço GA

No trabalho foram considerados os códigos de usuário de cada discente e o “timestamp” (data, hora, minutos e segundos) do momento de login no serviço Web. Esses registros foram gerados pelo próprio serviço Web e armazenados em uma base de dados de um servidor MS SQL Server 2000. Essa forma proprietária de registros foi implantada em 2004 e, através dela, foi possível determinar os acessos únicos de cada cliente para traçar o padrão de dados históricos.

Entretanto, constatou-se que os dados registrados eram insuficientes para uma análise mais detalhada do fenômeno de rajadas. Para acabar com essa limitação, a partir de abril de 2005 alterou-

se o código que registrava as requisições do usuário, tendo como referência a metodologia CBMG (Modelo de Grafo com o Comportamento do Cliente) (MENASCÉ; ALMEIDA, 2003b). Esse novo formato, ainda que proprietário, permitiria uma análise de outras características mais complexas da carga de trabalho, como a ocorrência do fenômeno de rajadas.

A Tabela 6.2 lista as 18 funcionalidades que o serviço GA oferece. Cada requisição do cliente tem registrado em uma base de dados SQL o número da funcionalidade, o “timestamp” (data, hora, segundo) e o endereço IP. Registrar cada funcionalidade requisitada permite determinar com precisão a quantidade de requisições do cliente e a requisição de maior frequência, dentre outras informações.

Tabela 6.2 - Funcionalidades do serviço GA.

Funcionalidades Serviço GA	
1 – Login	10 - Frequência
2 - Informações Pendentes	11 - Ementa
3 - Informações Gerais	12 - Plano de Ensino
4 - Documentos	13 - Requerimentos HE
5 - Endereços	14 - Requerimentos Reg. Domiciliar
6 - Vestibular	15 - Requerimentos Gerais
7 - Notas por Período Letivo	16 - Requerimentos
8 - Notas	17 - Requerimento Gerado
9 - Frequência por Período Letivo	18 - Fale com a Secretaria

b) Serviço Proxy

O registro dos acessos ao serviço Proxy foi implementado seguindo o padrão http (FEITELSON et al. 2005). A funcionalidade para registrar no padrão http já estava disponível no software Squid (SQUID, 2005), no qual foi implementado o serviço Proxy. Foram registradas as seguintes características:

1. Endereço IP cliente;
2. Data, hora, minuto e segundo do acesso;
3. Operação realizada (GET, POST);
4. Endereço da URL completa do objeto Web acessado;
5. Versão do http suportado pelo navegador do cliente;
6. Resposta do Cachê de Disco.

Uma dificuldade encontrada na análise desse serviço foi o grande volume de dados gerados diariamente. Para realizar a carga do banco de dados, foi gerado um programa na linguagem SQL, nativa do banco de dados Oracle 9i. Esse programa analisava o log diário do serviço web Proxy, calculando os valores para intervalos de tempo pré-definidos.

Um período de três dias de requisições gera em média 500 Mb de dados de registros históricos. Por exemplo, as informações coletadas no mês de abril de 2005 geravam uma única base de dados de 8 Gb. Para processar todas as informações foi necessário utilizar um servidor de banco de dados dedicado Oracle 9i, e aproximadamente, nove horas de processamento foram gastas nessa tarefa.

6.1.2 Registro do Tempo de Uso da CPU

Um dado importante no planejamento de capacidade é a utilização do serviço Web. Assim, é possível verificar, fixada uma escala de tempo, o quanto de capacidade do serviço Web está sendo utilizado para atender a carga de trabalho recebida. Para obter essa informação são necessários registros históricos do tempo gasto no processamento das requisições http do intervalo analisado. Esse dado é identificado como tempo de uso da CPU em estudos de Feitelson (2005) e Reed (1999).

Para registrar o tempo gasto no processamento das requisições http dos serviços Proxy e GA foram usadas técnicas e ferramentas de software diferentes para cada um. No serviço Proxy as informações de tempo de processamento foram coletadas pelo software TOP (FOCA LINUX, 2005), presente em diversas distribuições do Sistema Operacional Linux. O serviço GA teve o tempo de processamento coletado pelo software PM (performance monitor) (WINDOWS NT, 1999) presente no Windows NT Server.

6.2 Definição dos Padrões de Análise dos Registros Históricos

A análise foi feita em diferentes períodos e escalas de tempo dos registros históricos dos serviços GA e Proxy. A Tabela 6.3 mostra cada serviço Web e os tipos de análises efetuadas, que foram ocorrência e intensidade do fenômeno de rajadas e padrão de dados históricos. Para análise de ocorrência de rajadas utilizaram-se registros históricos em escalas de tempo variando entre 1 minuto e 1 hora. Na determinação dos padrões dos dados históricos foram utilizados períodos de tempo maiores, baseados em dia, mês e ano. Essa opção foi feita para evidenciar o comportamento da carga de trabalho em períodos de tempo com maior demanda para predição de evolução da carga de trabalho.

Tabela 6.3 - Análise dos Serviços Web.

Serviço	Período	Escala de Tempo	Modelo Analisado
GA	2004, Janeiro a Dezembro	1 ano	padrão de dados históricos
Proxy	2005, Fevereiro	1 mês	padrão de dados históricos
Proxy	2005, Fevereiro, dia 18	1 dia	padrão de dados históricos
GA	2005, Abril, dia 26, 10h00	1 hora	rajadas
GA	2005, Abril, dia 26, 10h22	1 minuto	rajadas
Proxy	2005, Abril, dia 14, 9h00	10 minutos	rajadas
Proxy	2005, Abril, dia 14, 9h00	1 hora	rajadas
Proxy	2005, Abril, dia 11, 9h00	1 minuto	rajadas
Proxy	2005, Abril, dia 11, 10h00	10 minutos	rajadas

6.2.1 Padrão dos Dados Históricos

Uma análise do padrão dos dados históricos da carga de trabalho dos dois serviços foi realizada em diferentes escalas de tempo, de acordo com a disponibilidade dos registros de dados históricos.

a) Serviço GA

A Figura 6.2 mostra o registro da médias de acessos diários nos meses de janeiro a dezembro de 2004. Observa-se uma tendência de crescimento nos meses de Julho e Dezembro, correlacionada à publicação das notas e freqüências finais das disciplinas cursadas pelos alunos. Os maiores picos de acesso aconteceram na integralização das notas de disciplinas anuais, no mês de dezembro, com média diária de 83 acessos. Quase o dobro da segunda maior média do ano, justamente em Julho, que é o mês de integralização das notas e freqüências das disciplinas semestrais. Há uma significativa queda de requisições nos três primeiros meses do ano, ocasião das férias escolares.

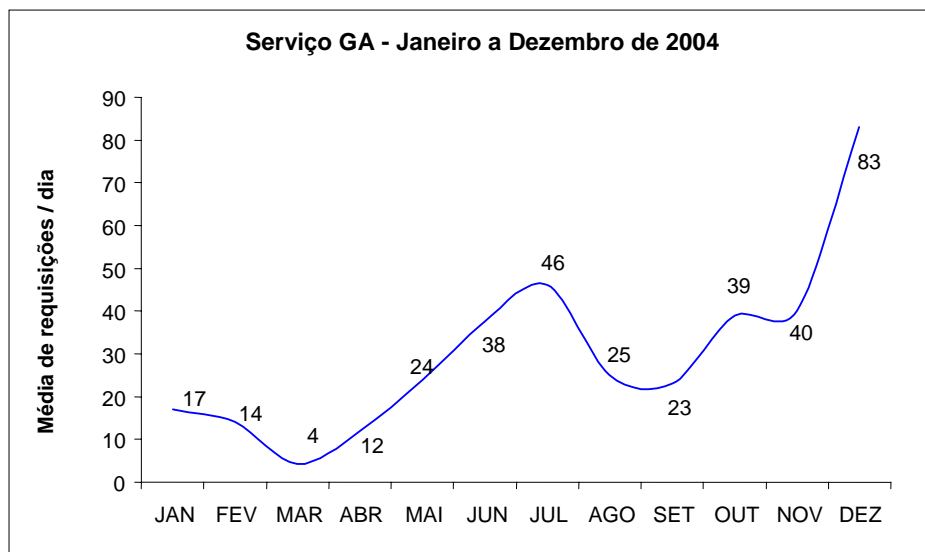


Figura 6.2 - Serviço GA - requisições anuais.

A Figura 6.4 mostra a evolução da quantidade de acessos http em 2004 e 2005 para o serviço GA. O eixo y representa o total de requisições http recebidas pelo serviço em determinado mês. Esse padrão apresentado pode ser classificado como sazonal, pois é influenciado por eventos do calendário escolar: publicação de notas, férias, dentre outros.

Uma justificativa para o aumento do número de acessos no serviço no ano de 2005 é a oferta de novas funcionalidades, a consolidação do serviço na comunidade e a mudança no padrão de registro dos dados históricos. A versão anterior do serviço GA só registrava o login do cliente, e a mudança de 2005 inclui o registro no acesso a todas as funcionalidades do serviço.

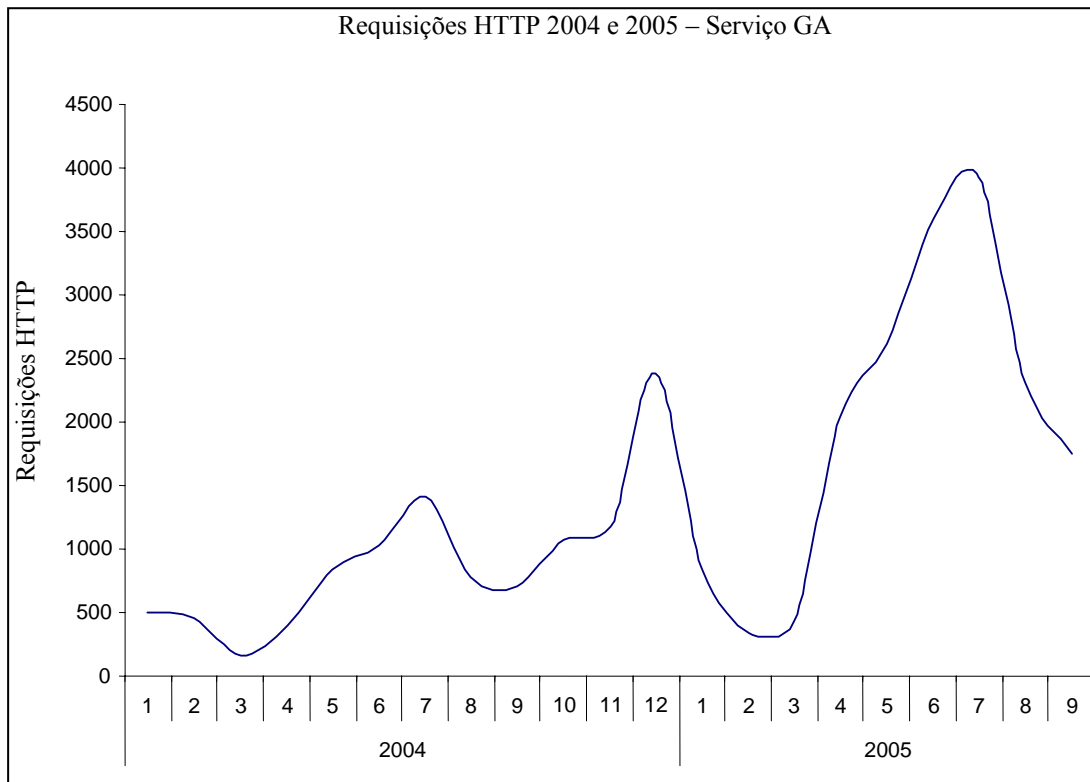


Figura 6.3 - Serviço GA - Requisições http 2004 e 2005.

b) Serviço Proxy

A Figura 6.4 mostra o total de requisições ao serviço Proxy no mês de Fevereiro de 2005. Uma média de 330 mil acessos ao dia foi registrada. Em decorrência do carnaval, nota-se uma série de 04 dias com quase nenhum acesso. No restante do mês, o serviço Web recebe grande quantidade de requisições nos dias úteis e nos finais de semana a utilização é praticamente nenhuma. Em média, foram registradas 505 mil requisições http por dia útil nesse período.

O padrão de registros históricos do tipo sazonal é o mais adequado para tipificar o comportamento dessa carga de trabalho, pois é influenciada pelo calendário escolar da instituição, sob o ponto de vista administrativo. Os acessos são estáveis durante a semana, de segunda a sexta. Nos finais de semana e feriados são registrados poucos acessos, pois não há o funcionamento da instituição.

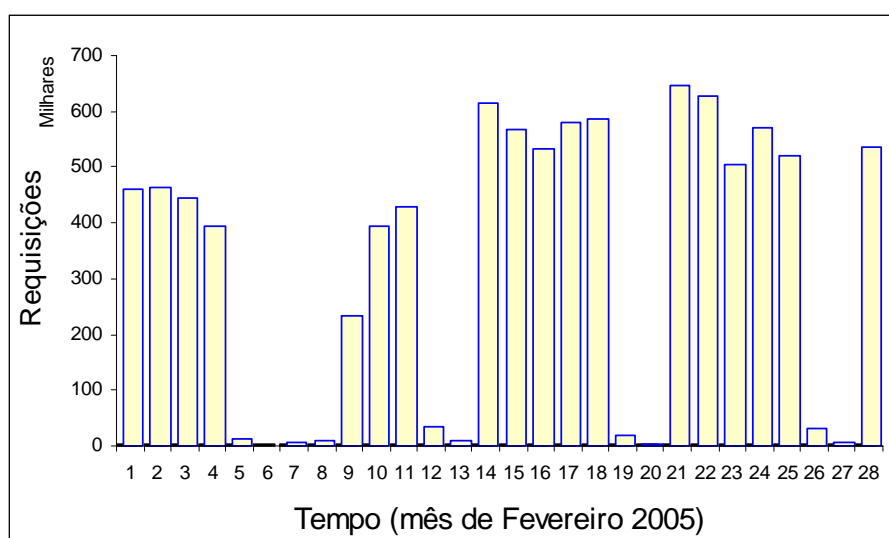


Figura 6.4 – Requisições http do serviço Proxy.

A Figura 6.5 mostra o total de requisições ao serviço Proxy, no dia 18 de fevereiro de 2005, É um dos maiores registros de requisições diárias, com 586 mil requisições. Os maiores picos de utilização são nos horários entre 9 e 16 horas.

O acesso diário analisado no dia 18 está inserido no contexto do padrão sazonal do mês de Fevereiro. O padrão dos dados históricos pode ser mais precisamente tipificado como cíclico, com uma curva crescente até as 10 horas, estável até as 16 horas e decrescente a partir desse período.

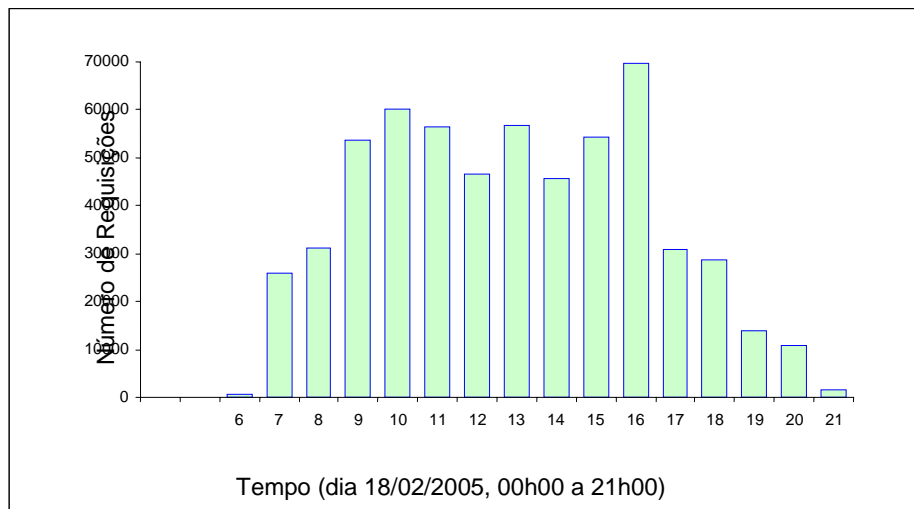


Figura 6.5 – Requisições http diárias do serviço Proxy.

Os dois horários com o maior pico de acesso são 10 e 16 horas, que representam o meio do expediente da manhã e da tarde. Esse tipo de padrão de registros históricos da carga de trabalho existia também nos ambientes de *mainframe*, em que recebia a alcunha de “10-14”, representando picos de acesso as 10 e 14 horas.

O comportamento “10-16” identifica tipicamente uma relação direta entre as requisições dos usuários e o horário administrativo de funcionamento da instituição. Pode-se notar um fato que comprova essa afirmação: os horários de início e término dos acessos, 6 e 21 horas, são os mesmos que definem o período de funcionamento diário da instituição analisada.

6.2.2 Conclusões Sobre a Análise do Padrão dos Dados Históricos

O serviço GA é oferecido a um público específico da instituição (discentes dos cursos de graduação) e tem o padrão de dados históricos definido pelo calendário escolar. Ocorrem picos de utilização em dois meses do ano: julho e dezembro. Esses meses são apontados no calendário como períodos de fechamento de notas e frequência dos alunos.

Os resultados obtidos no estudo da carga de trabalho do serviço GA podem ser usados para:

- Predição da evolução da carga de trabalho futura em relação a mudanças do calendário escolar. A demanda irá acompanhar os eventos de divulgação de notas e frequência.
- Contornar situações de gargalo de desempenho. Para evitar uma demanda inadequada à capacidade do serviço, pode-se sugerir uma divulgação das notas e frequências que não coincidam nos mesmos períodos para ambos os cursos.
- Interromper, nos períodos de maior demanda, as funcionalidades secundárias do serviço, para uma melhor utilização dos recursos existentes.
- Alocar um computador para realizar balanceamento de carga nos períodos de maior demanda, dividindo o processamento das requisições dos clientes.

O serviço Proxy, oferecido apenas no ambiente interno, tem a sua carga de trabalho fortemente associada ao horário de funcionamento da instituição, apresentando um padrão cíclico. Os picos de acesso são no período da manhã, às 10 horas, e à tarde, às 16 horas. Os resultados obtidos no estudo da carga de trabalho do serviço Proxy podem ser utilizados para restrição do acesso, nos horários de maior demanda, a determinados tipos de arquivos, como áudio e vídeo.

O estudo do padrão dos dados históricos da carga de trabalho dos dois serviços também fornece subsídios para a definição de acordos de nível de serviço com os clientes, planejamento da disponibilidade e de momentos adequados de interrupções para manutenção.

6.3 Ocorrência do Fenômeno de Rajadas

O capítulo 3 apresentou uma abordagem operacional para análise da carga de trabalho, com o objetivo de determinar a ocorrência do fenômeno de rajadas, sua intensidade e seu conseqüente impacto na utilização do serviço Web. Aplicou-se essa análise em diversas amostras de ambos os serviços, com escala de tempo definida a partir de 1 minuto até 1 hora. Para cada amostra, foram calculados os fatores de rajada (a,b) utilizando as equações (3.4 e (3.5. Seus resultados aparecem na Tabela 6.4:

Tabela 6.4 - Fator de rajada da carga de trabalho.

Serviço	Período	Escala	Qtde. épocas n	Média de acessos	k ⁺	Qtde. acessos Log	A+	a	b	b % raj.
GA	2005, Abril, dia 26, 10h22	1 minuto	18	6	12	108	99	1,375	0,667	66,7 %
GA	2005, Abril, dia 26, 10h00	1 hora	37	7,1	9	263	202	3,158	0,243	24,3%
Proxy	2005, Abril, dia 14, 9h00	10 minutos	658	17,94	300	11806	7439	1,382	0,455	45,5%
Proxy	2005, Abril, dia 14, 9h00	1 hora	3572	25,43	1378	90841	52135	1,488	0,386	38,60%
Proxy	2005, Abril, dia 11, 10h00	10 minutos	652	19,92	278	12989	8791	1,58	0,426	42,6%
Proxy	2005, Abril, dia 11, 10h00	1 minuto	60	13,4	18	804	517	2,14	0,3	30%

6.3.1 Análise de Rajadas no Serviço GA

A análise do serviço GA foi realizada em dados colhidos no dia 26/04/2005, escolhidos de maneira aleatória. A opção em utilizar os dados históricos do ano de 2005 deu-se em razão da implantação de um mecanismo mais eficiente no registro das requisições dos clientes a partir dessa data. Duas escalas de tempo foram consideradas: 1 hora e 1 minuto, respectivamente, para as 10h00 e 10h22.

A Figura 6.6 mostra o gráfico da carga de trabalho do serviço GA no dia 26 de Abril, as 10h22. Em 66 % do tempo a taxa de chegada de requisição (parâmetro b) está acima da média do intervalo. O parâmetro a é 1,37 vez a média do intervalo. O eixo x é a quantidade de épocas n usada na divisão da escala de tempo da carga de trabalho.

Nessa primeira análise, é pequena a variação da requisição acima da média do intervalo (1,37) e ocorrendo na maior parte do mesmo (66%). Essa característica indica que o serviço recebe uma carga de trabalho intensa, mas com rajada mínima ou inexistente.

Entretanto, apenas a análise realizada nessa escala de 1 minuto é insuficiente para afirmar que o serviço não apresenta ocorrência do fenômeno de rajadas. Distorções podem ocorrer em escalas de tempo inadequadas ao tempo médio de sessão do cliente de um determinado serviço (ALMEIDA et al., 2002); (MENASCÉ; ALMEIDA, 2003a).

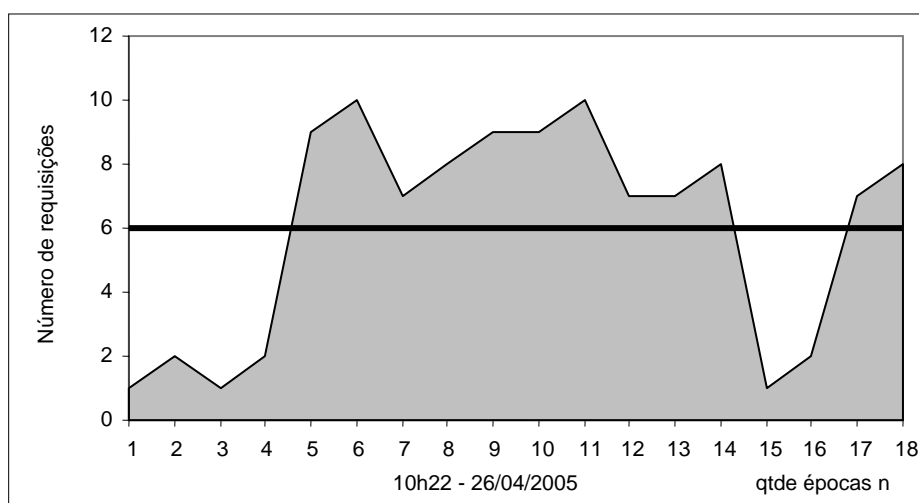


Figura 6.6 – Serviço GA: requisições http em 26/04 com escala de tempo de minuto.

A Figura 6.7 mostra o gráfico da carga de trabalho do serviço GA no dia 26 de Abril, às 10h00. Em 24,3 % do tempo a taxa de chegada de requisição (parâmetro b) está acima da média do intervalo. O parâmetro a é 3,158 vezes a média de requisições do intervalo. O eixo x é a quantidade de épocas n usada na divisão da escala de tempo da carga de trabalho.

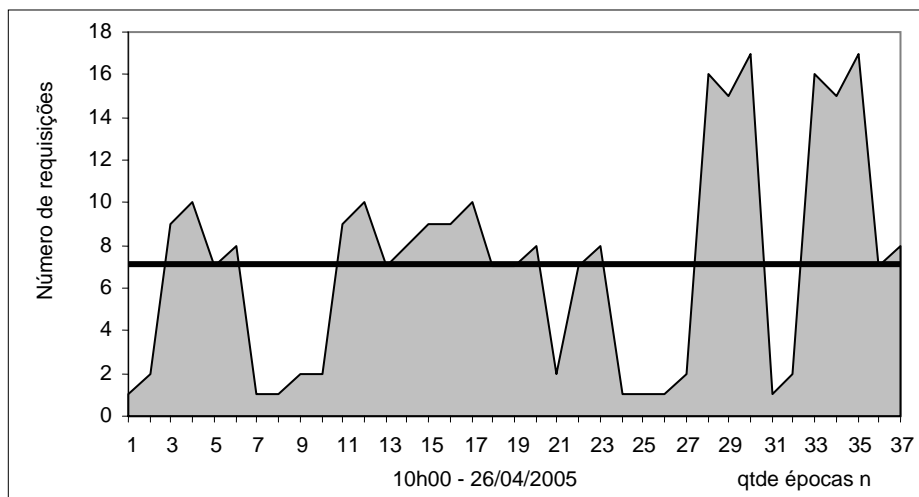


Figura 6.7- Serviço GA: requisições http em 26/04 com escala de tempo de 1 hora.

Na segunda análise utilizou-se uma escala de 1 hora. O fator de rajada (a, b) apresentou a ocorrência da suposta rajada em um percentual de tempo do intervalo bem inferior ao primeiro estudo. Entretanto, o parâmetro a é 3,158 vezes maior que a média do intervalo, o que pode causar comprometimento no desempenho do serviço de 20% a 40% (WANG et al., 2003).

Essa segunda análise permite comprovar que o serviço GA apresentou a ocorrência do fenômeno de rajadas no período analisado para a escala de tempo de 1 hora. Esse fato encontrado na construção do modelo de carga de trabalho do serviço GA indica a necessidade de analisar uma mesma amostra de dados históricos com duas ou mais escalas, definidas entre 1 minuto e 1 hora.

6.3.2 Análise de Rajadas no Serviço Proxy

O serviço Proxy foi examinado a partir de dados levantados em dois dias distintos, analisando-se em cada caso o impacto das rajadas em escalas diferentes de análise. Segue-se a análise de cada caso.

Caso A

As figuras 6.7 e 6.8 mostram o acesso ao serviço Proxy no dia 14 de Abril as 9h00 horas usando como escala de tempo 1 hora e 10 minutos, respectivamente. Esse horário foi escolhido por ser um horário de pico de utilização do serviço, conforme constatado na análise do padrão de sua carga de trabalho.

Observa-se que não há uma diferença significativa entre o fator de rajada (a,b) registrado na escala de 1 hora (1,382; 45,5%) e na escala de tempo de 10 minutos (1,488; 38,6%). Os valores, por serem próximos, mostram uma adequação entre a escala de tempo escolhida na análise da carga de trabalho.

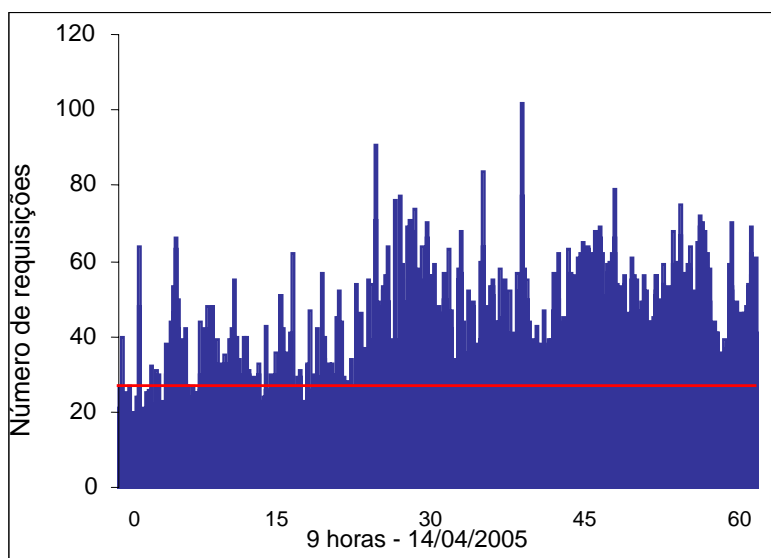


Figura 6.8 – Serviço Proxy: requisições http - 14/04 - escala de tempo - 1 hora.

Pode-se constatar que ocorre o fenômeno de rajadas nesse período, mas em uma intensidade que pode ser considerada “leve”. Essa intensidade não comprometeria em

mais que 20% de diminuição da capacidade do serviço em processar requisições para os clientes em circunstâncias normais.

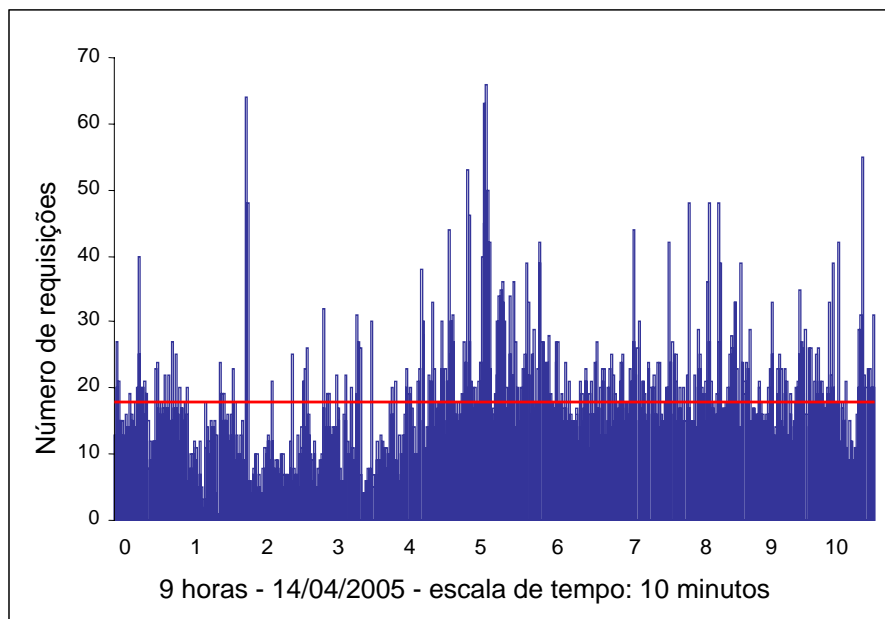


Figura 6.9 - Serviço Proxy: requisições http - 14/04 - escala de tempo: 10 minutos.

Caso B

As figuras 6.10 e 6.11 mostram o acesso ao serviço Proxy no dia 11 de Abril, às 9h00 e 10h00, usando como escala de tempo 1 minuto e 10 minutos, respectivamente. Esse horário foi escolhido por ser um horário de “pico” de utilização do serviço, conforme constatado na análise do padrão de sua carga de trabalho. Foi escolhido um horário e uma escala semelhante à realizada na seção anterior para observar os resultados obtidos pelo modelo determinístico.

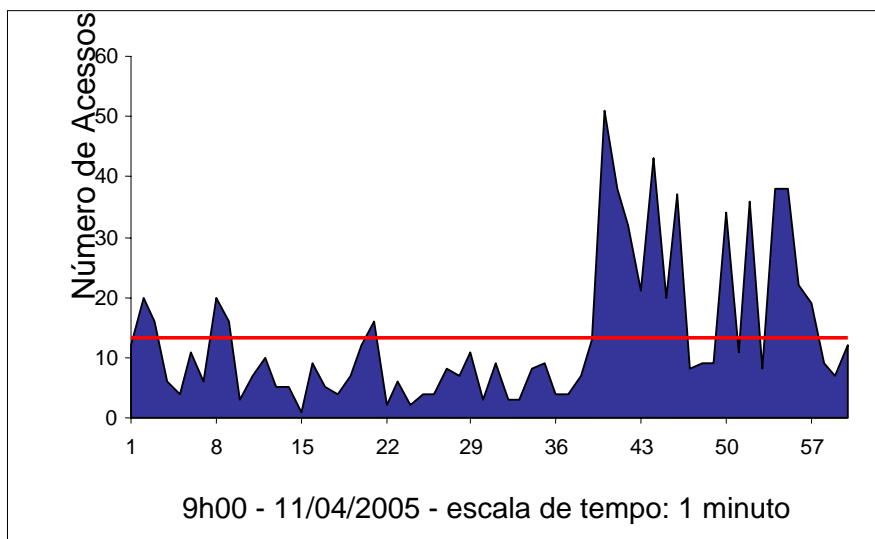


Figura 6.10 - Serviço Proxy: requisições http - 11/04 - escala de tempo: 10 minutos.

Observa-se que há uma diferença significativa entre o fator de rajada (a,b) registrado na escala de 1 minuto (1,58; 42,6%) e na escala de tempo de 10 minutos (2,14; 30%). A ocorrência do fenômeno de rajadas é registrada nas duas análises, mas novamente com mais definição na amostra com escala de tempo maior, de 10 minutos. Essa escala de tempo apresentou valores de fator de rajada (a,b) bastante próximos para os dias 25 e 11 de Abril, para o serviço Web Proxy, mesmo analisando intervalos de tempo distintos, as 9 e 10 horas.

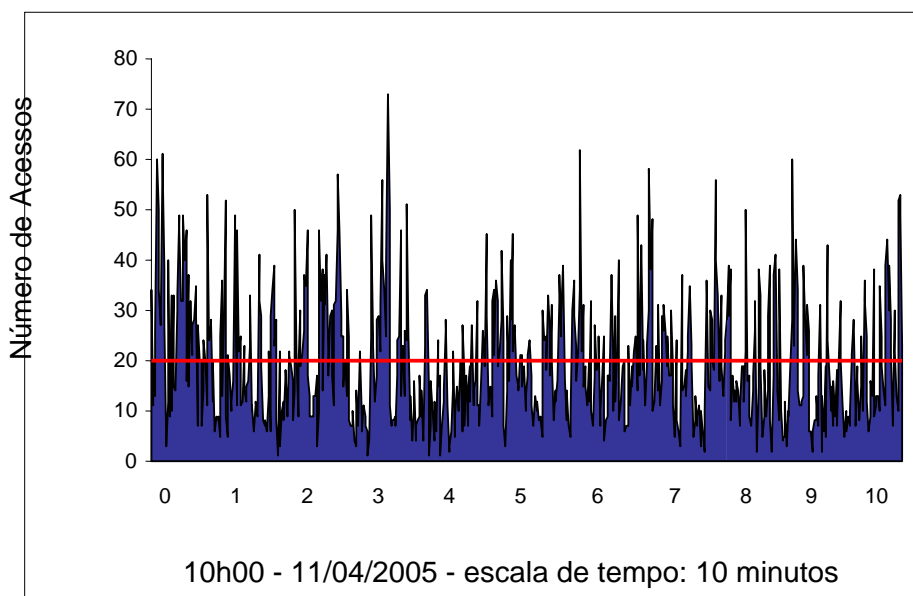


Figura 6.11 - Serviço Proxy: Requisições http - 11/04 - escala de tempo: 10 minutos.

O valor obtido utilizando a escala de tempo de 10 minutos apresenta um fenômeno de rajada que efetivamente afetaria a capacidade do serviço Web. A escala de 1 minuto, a exemplo do acontecido no serviço GA, mostra-se menos adequada à análise de serviços com as características apresentadas na carga de trabalho estudada.

6.3.3 Conclusões Sobre a Análise da Ocorrência do Fenômeno de Rajadas

O estudo através do modelo operacional obteve resultados consistentes. Algumas conclusões:

- A capacidade de processamento de ambos os serviços é afetada pelo fenômeno de rajadas.
- Mesmo em horários de pico de demanda, os dois serviços não têm sua capacidade severamente afetada.
- A escala de tempo é fundamental para uma análise mais precisa. Os valores com resultados mais consistentes foram 10 minutos e 1 hora.
- O modelo utilizado ofereceu resultados satisfatórios em resposta às amostras analisadas e pode ser considerado como representativo da carga de trabalho de ambos serviços.

6.3.4 Utilização do Serviço Web

Um dos principais objetivos do profissional que executa o planejamento de capacidade é evitar o momento de saturação do serviço Web, e um dado importante a ser observado nesse contexto é a utilização do serviço Web, cuja equação (3.7) foi definida na seção 3.1. A utilização do serviço pode ser obtida através da multiplicação da taxa média de chegada de requisições http (*throughput* - λ) pela taxa média de espera pelas requisições http recebidas (demanda - μ).

Os dados históricos coletados em ambos os serviços analisados foram utilizados no cálculo da utilização do serviço, conforme listado na Tabela 6.5. A coluna “Tempo de Processamento” (tempo de CPU) registra a quantidade, em segundos, gasta no processamento das requisições http recebidas pelo serviço. As colunas μ e λ são usadas para obter o valor da coluna U, que representa o percentual de utilização do serviço.

A carga de trabalho do serviço GA sofreu a ocorrência de rajadas em apenas uma das escalas de tempo que foram analisadas (1 minuto). Na escala de tempo de 1 minuto, chegou-se à conclusão que uma carga de trabalho intensa havia ocorrido (66,7% do tempo). Mesmo nesse cenário, a utilização do serviço GA atinge valores de máximo na faixa de 30%. Esse dado é útil para o responsável pelo planejamento de capacidade pois informa que a capacidade do serviço está adequada, não havendo saturação mesmo em situações de pico de utilização e de ocorrência de rajadas.

Para o serviço Proxy, existe a ocorrência do fenômeno de rajadas em todas as amostras analisadas. Uma constatação é a alta utilização do serviço nas quatro escalas de tempo analisadas, com o valor mínimo de 53,33% e o máximo de 86,25%. A ocorrência do fenômeno de rajadas mais intensa é na amostra do dia 14 de Abril, 9h00, com escala de 1 hora. Nesse período a utilização do serviço chega a 86,25%. Usando uma escala de tempo menor, de 10 minutos, no dia 14 de Abril, a utilização do serviço é de 70%, mesmo com um fator de rajada maior (1,58; 42,6%).

Essa variação na análise da carga de trabalho do serviço Proxy pode ser explicada pela variação da escala de tempo utilizada na análise de cada amostra da carga de trabalho. Analisando somente as cargas de trabalho com escala de tempo de 10 minutos, para o dia 11 e 14 de abril, podemos constatar que a carga de trabalho é um pouco mais intensa no dia 14 (1,38; 45,5%) do que no dia 11 (1,58; 42,6%) utilizando a capacidade do serviço 4% mais (74% no dia 14 contra 70% no dia 11). Uma conclusão é que a intensidade do fator de rajada afetou a capacidade do serviço.

Uma carga de trabalho do serviço Proxy com menor intensidade do fator de rajada (2,14; 30%) e uma escala de tempo de 1 minuto, em 11 de Abril, causa uma utilização do serviço em 53%. O parâmetro b desse registro ficou 8% abaixo do segundo menor valor, e o período no qual o parâmetro a é maior que a média do intervalo, por ser curto,

não chega a afetar a capacidade completa do serviço Web. Entretanto, tal conclusão não deve ser considerada isoladamente, em razão da influência das escalas de tempo nas medições de resultados. Outra escala, de 10 minutos, registra a utilização do serviço Proxy em 70,17%, no mesmo período de tempo.

Tabela 6.5 – Utilização do Serviço Web.

Serviço	Período	Quantidade acesso http	Intervalo (segundo)	Tempo CPU (segundo)	λ transações por segundo	μ requisições por segundo	U % abordagem operacional
GA	2005, Abril, dia 26, 10h22 1 minuto	108	60	18	1,80	0,167	30 %
GA	2005, Abril, dia 26, 10h00 1 hora	263	3600	371	0,07	1,411	10,31 %
Proxy	2005, Abril, dia 14, 9h00 10 minutos	11806	600	448	19,68	0,038	74,67 %
Proxy	2005, Abril, dia 14, 9h00 1 hora	90841	3600	3105	25,23	0,034	86,25 %
Proxy	2005, Abril, dia 11, 10h00 10 minutos	12989	600	421	21,65	0,032	70,17 %
Proxy	2005, Abril, dia 11, 10h00 1 minuto	804	60	32	13,40	0,040	53,33 %

6.4 Modelo de Carga de Trabalho Fuzzy

O capítulo anterior apresentou o modelo para análise de carga de trabalho que utiliza em sua implementação dois controles fuzzy, RAJIN e CSWeb. Esse modelo avalia a ocorrência do fenômeno de rajadas em uma carga de trabalho de um serviço Web. A análise desse modelo busca verificar a sua eficiência comparando seus resultados com os obtidos através da abordagem operacional utilizada.

A próximas seções descrevem os resultados obtidos pelo controle RAJIN no cálculo da variável de saída RAJ (intensidade de rajada) e do controle CSWeb, para a variável USW (utilização do serviço Web).

6.4.1 Intensidade de Rajada – Controle RAJIN

A primeira fase da análise é determinar os valores intensidade da rajada, variável de saída do controle fuzzy RAJIN. A Tabela 6.6 lista os resultados produzidos pelo processo de inferência do controle. Os valores de entrada foram submetidos ao controle com o parâmetro a variando de 0 a 6, e o parâmetro b , de 0 a 50. Esses valores correspondem a todos os intervalos definidos no universo de discurso de ambas as variáveis, e foram fornecidos por especialistas em planejamento de capacidade para serviços Web.

Resultados semelhantes estão presentes em estudos diversos, como Arlitt (1996) e Banga (1999). Usando a abordagem operacional, na seção 6.3 foram mostrados os resultados do parâmetro (a,b) para os serviços GA e Proxy. Foram registrados para os dados analisados valores para $a = [1,4]$ e $b = [30\%,50\%]$.

Tabela 6.6 - Resultados entrada e saída controle fuzzy Rajin.

Parâmetro b	Variável de Entrada: Parâmetro a						
	0	1	2	3	4	5	6
0	0	20	22	20	22	20	20
5	0	20	22	20	22	20	20
10	0	20	22	20	22	20	20
15	0	20	22	20	23	22	22
20	0	20	22	22	23	22	22
25	0	20	22	22	24	22	23
30	0	52	55	55	54	55	55
35	0	52	55	55	54	55	55
40	0	75	78	80	80	80	80
45	0	75	78	80	80	80	80
50	0	75	78	80	80	80	80

A análise da ocorrência do fenômeno de rajadas torna-se possível com a definição de conceito de intensidade de rajada, que é obtida por inferência do controle fuzzy RAJ. A Tabela 6.6 lista todos os possíveis valores de intensidade de rajada encontrados em uma carga de trabalho. O limite máximo da intensidade de rajada calculado pelo controle foi 80%, atingido quando o valor do parâmetro b chega em 40% e do parâmetro a atinge o valor de 3.

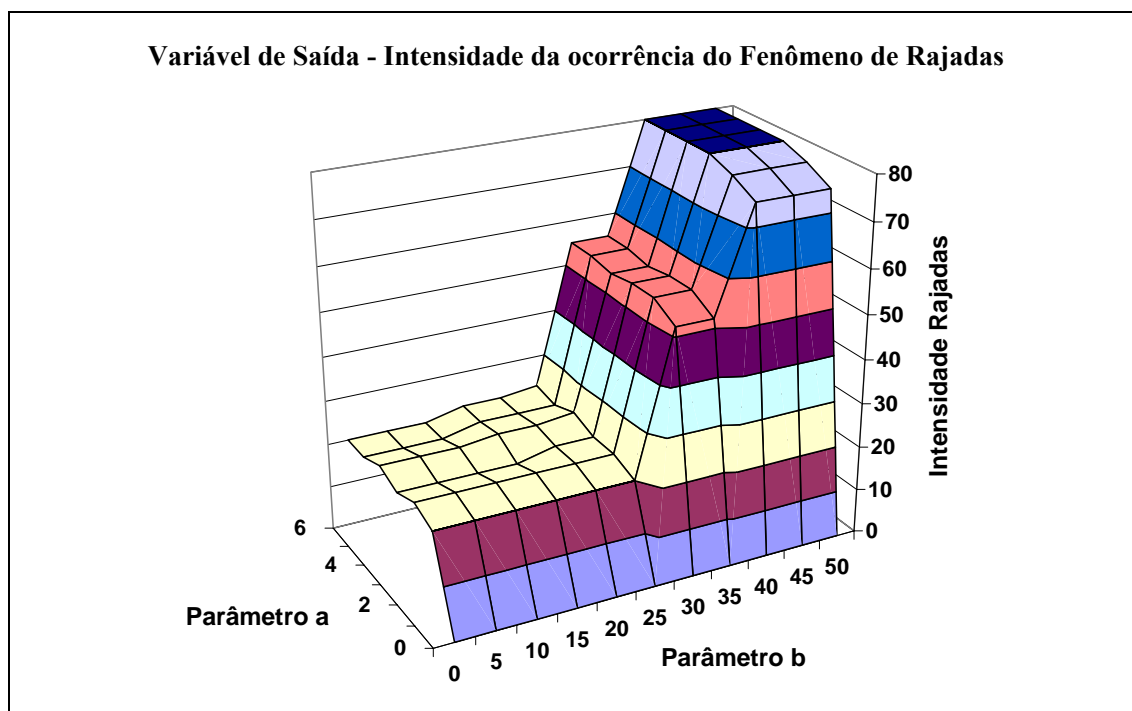


Figura 6.12 - Ocorrência do Fenômeno de Rajadas.

A **Figura 6.12** apresenta o gráfico de área intensidade da ocorrência de rajadas. Pode-se notar que o aumento dos valores dos parâmetros (a,b) reflete o aumento da intensidade das rajadas, que varia de 0% a 80%, quando atinge seu valor máximo. Esse resultado é esperado pois está em conformidade com a definição das regras e fatos registradas na matriz de regras de inferência.

Valores acima de 80% não aparecem registrados pela programação das regras e fatos da base de conhecimento do controle fuzzy pois refletem a definição do conceito de rajada: a partir de um determinado ponto, a carga de trabalho passa de uma característica de ocorrência de rajada para um estado de alto volume de requisições http durante todo o intervalo de medição.

6.4.2 Utilização do Serviço Web – Controle CSWeb

A segunda etapa da análise da carga de trabalho foi efetuada utilizando o controle fuzzy CSWeb. Esse controle, conforme descrito em seções anteriores, fornece como variável de saída o valor da utilização do serviço Web. Submetendo como variáveis de entrada os valores das colunas RAJ e NHTTP, listados na **Tabela 6.7**, o resultado foi uma série quase completa de valores nulos, presentes na coluna “U Fuzzy CSWeb [0,200]”.

Analisando o resultado obtido, constatou-se que o motivo dos valores nulos era o universo de discurso da variável de entrada NHTTP (entre 0 e 200). Os valores foram definidos por especialistas em planejamento de capacidade, e mostraram-se adequados para serviços Web de comércio eletrônico e portais corporativos e governamentais, com alta utilização e crescimento. Entretanto, para uma instituição pequena, com até 2.000 clientes, com média de acessos não superior a 25 requisições http na escala de tempo analisada, os resultados obtidos foram pouco significativos (utilização do serviço web inferior a 5%).

Um ajuste nos valores do universo de discurso foi realizado para adaptar o controle à carga de trabalho analisada. Através desse processo os valores do universo de discurso da variável de entrada NHTTP foram redefinidos em um intervalo de [0,30]. Foi possível definir esse intervalo com base nos dados históricos analisados no modelo operacional. A Tabela 6.7 lista na coluna “U Fuzzy CSWeb [0,30]” os resultados obtidos no processo de inferência do controle CSWeb.

Na mesma tabela, há a coluna “U % - Abordagem Operacional” com os resultados obtidos pelo modelo analítico, para fins comparativos. A coluna a e b representam o fator de rajada (a,b) e a coluna RAJ a intensidade da ocorrência do fenômeno de rajadas presente na carga de trabalho analisada.

Tabela 6.7 – Cálculo de Utilização do Serviço – Abordagem Operacional e Fuzzy

Ser- viço	Período	a	b	RAJ	NHTTP médio intervalo	U % abord. operac.	U Fuzzy CSWeb [0,200]	U Fuzzy CSWeb [0,30]
GA	2005, Abril, dia 26, 10h22 1 minuto	1,375	66,7 %	75%	6	30 %	Nulo	20%
GA	2005, Abril, dia 26, 10h00 1 hora	3,158	24,3%	22%	7,1	10,31 %	20%	15%
Proxy	2005, Abril, dia 14, 9h00 10 minutos	1,382	45,5%	75%	17,94	74,67 %	Nulo	66%
Proxy	2005, Abril, dia 14, 9h00 1 hora	1,488	38,60%	52%	25,43	86,25 %	Nulo	80%
Proxy	2005, Abril, dia 11, 10h00 10 minutos	1,58	42,6%	75%	19,92	70,17 %	Nulo	68%
Proxy	2005, Abril, dia 11, 10h00 1 minuto	2,14	30%	55%	13,4	53,33 %	Nulo	30%

A Figura 6.13 traz os gráficos gerados pelo controle fuzzy CSWeb representando os processos de agregação das regras ativadas pelos dados recebidos das variáveis de entrada. O operador de Zimmerman (com $\lambda = 0,5$) foi usado para agregar as regras, e o operador de centro de área para realizar a defuzzificação. No gráfico, o centro da área é marcado com um sinal de positivo.

A carga de trabalho analisada na Figura 6.13 é a do serviço Web Proxy no dia 11 de Abril de 2005, com as respectivas escalas de tempo de 10 e 1 minuto. A escala de 1 minuto tem uma intensidade de rajada de 55%, e a quantidade média de requisições http é considerada de média para baixa. Como resultado, a área do gráfico traz agregadas as regras relacionadas a NHTTP e RAJ baixas e médias, representando a USW baixa.

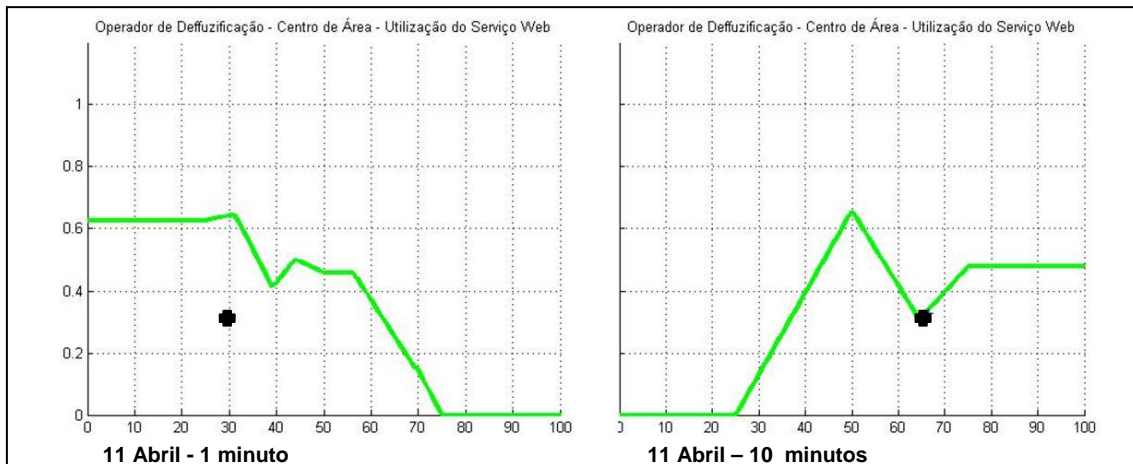


Figura 6.13 – Processo de defuzzificação - Serviço Web Proxy (11 de Abril).

A Figura 6.14 traz o gráfico do processo de defuzzificação da carga de trabalho Proxy no dia 14 de Abril, com escala de tempo de 1 hora. A intensidade da rajada (RAJ) é alta, com valor de 52%, e a média de requisições http no intervalo também é alta, 25,43. Como resultado o controle retorna a variável USW com valor de 80%. Pode-se notar no resultado do processo de agregação a ativação das regras com valores dos quantificadores fuzzy definidos como altos.

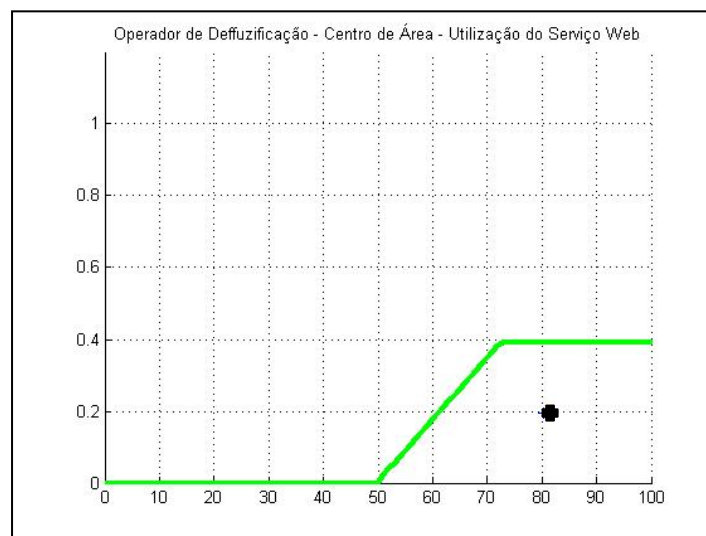


Figura 6.14 - Processo de defuzzificação - Serviço Web Proxy - 14 de Abril.

A Figura 6.15 mostra o gráfico com a comparação dos resultados obtidos na análise da utilização do serviço Web com o modelo operacional e o controle fuzzy. A coluna amarela representa os resultados da análise produzida pela abordagem operacional. A coluna verde representa os resultados do modelo com base no controle fuzzy. O eixo y indica o valor percentual da utilização do serviço Web analisado.

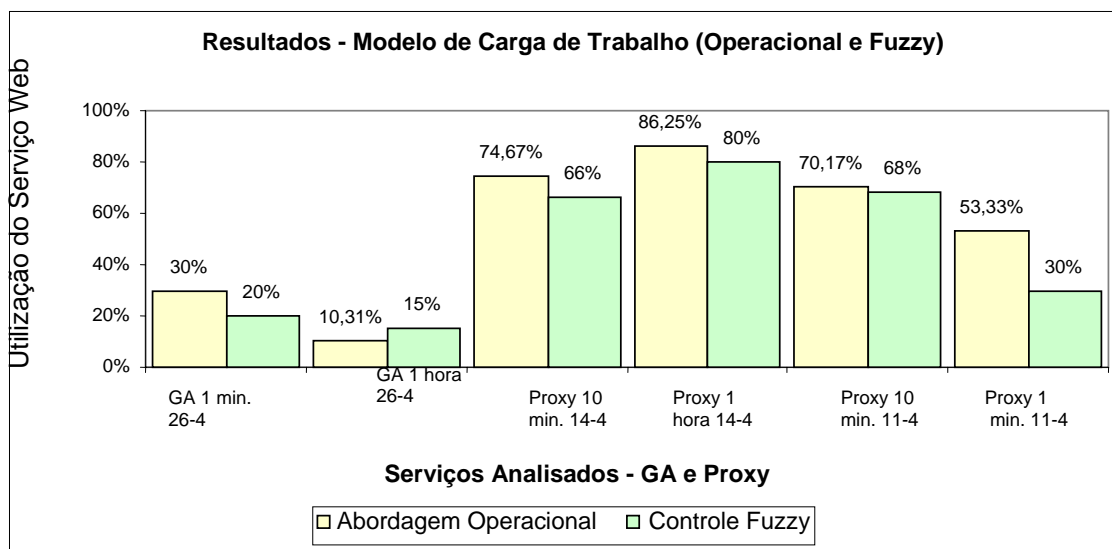


Figura 6.15 – Comparação de resultados entre o modelo operacional e fuzzy.

Analisado pelo controle fuzzy CSWeb, o serviço GA apresentou uma diferença de 33% em média em relação ao resultado da utilização do serviço calculado no modelo operacional. A mesma análise foi realizada no serviço Proxy e gerou um resultado melhor, com uma diferença média de 16% comparado ao modelo operacional. O melhor resultado obtido na análise do serviço Proxy tem relação com o processo de sintonia do universo de discurso da variável NHTTP, pois foram utilizados, como referência, os valores dos dados históricos da carga de trabalho desse serviço.

Em algumas amostras da carga de trabalho o resultado obtido pelo controle fuzzy foi próximo ao resultado do modelo operacional. Na Figura 6.15 pode-se visualizar uma diferença mínima de resultados da utilização do serviço Proxy em 11 de Abril na escala de tempo de 10 minutos: 70,17% obtido no modelo operacional e 68% no controle fuzzy.

6.5 Considerações Finais

Neste capítulo foram apresentadas as análises realizadas para várias cargas de trabalho submetidas a dois serviços Web. As análises utilizaram os modelos operacional e fuzzy para o tratamento dos dados experimentais, com bons resultados.

No próximo capítulo aparecem as conclusões finais sobre essas avaliações, bem como sugestões para trabalhos futuros nessa área.

7 Conclusão

A principal finalidade desse trabalho é o estudo das características do modelo de carga de trabalho de serviços Web adaptado ao fenômeno de rajadas, baseado em sistemas de controle fuzzy, descrito no capítulo 4. A motivação para sua realização surgiu da constatação da complexidade de implementação da solução baseada em modelo determinístico, representada pelo modelo operacional.

A partir do estudo das técnicas, modelos e *frameworks* usados em planejamento de capacidade, conclui-se que duas abordagens são usadas na sua implementação: a primeira utiliza métodos e ferramentas mais elaborados, em busca de precisão, em contraponto a uma segunda, que adota uma postura pragmática e utiliza técnicas que produzam resultados rápidos, em busca de um senso de direção.

Os principais resultados do estudo realizado sobre a abordagem operacional são:

1. Constatou-se que a necessidade de utilizar os dados históricos com as requisições http dos clientes do serviço Web para a implementação do modelo determinístico traz desvantagens:
 - Exigência de investimento em recursos computacionais para processamento dos dados históricos de um serviço Web com grande volume de acesso e muitos clientes.
 - O armazenamento inadequado dos dados históricos, por desconhecimento dos padrões da indústria, dificulta ou inviabiliza a posterior análise da carga de trabalho.
2. O conhecimento detalhado do contexto no qual o serviço está sendo oferecido é fundamental no estudo de sua carga de trabalho e está relacionado ao padrão dos dados históricos. Em ambos os serviços Web analisados há uma correlação entre o calendário administrativo da instituição estudada e o padrão da respectiva carga de trabalho.

3. Os resultados obtidos na análise do padrão dos dados históricos dos serviços podem ser usados em diversos contextos: predição de carga de trabalho futura, melhoria da qualidade da capacidade do serviço oferecido, seja através de investimentos para expansão da infra-estrutura, como também do ajuste das funcionalidades oferecidas.
4. A escala de tempo é um detalhe importante a ser observado, pois uma análise feita com uma escala inadequada pode trazer resultados não representativos. O ideal é trabalhar ao menos com duas escalas de tempo quando utilizar o modelo determinístico para representar a carga de trabalho, pois possibilita comparações e avalia os resultados encontrados.
5. O fenômeno de rajadas ocorre em todas cargas de trabalho analisadas e causa alterações no desempenho dos serviços Web, com maior ou menor grau de intensidade.

O segundo modelo estudado foi implementado utilizando dois sistemas de controle fuzzy. Os principais resultados e contribuições desse estudo são:

- O sistema de controle fuzzy adaptado ao fenômeno de rajadas é uma alternativa para a análise de serviços Web que não tenham registros históricos disponíveis ou padronizados.
- O modelo de carga de trabalho implementado pelo sistema de controle fuzzy calculou em menor tempo a utilização do serviço, comparado ao modelo determinístico. O controle fuzzy produziu em 1 segundo o resultado da amostra da carga de trabalho do serviço fuzzy, para o dia 11 de Abril, na escala de 10 minutos. Para esses mesmos dados o modelo determinístico produziu os resultados em 20 minutos.
- Uma contribuição interessante desse estudo foi a definição e o cálculo da intensidade de rajadas para o fator de rajada (a,b) . Esse parâmetro não existia e

utilizando o controle fuzzy foi possível inferir um valor único para a dupla de parâmetros (a,b).

- Ajustes no universo de discurso da variável de entrada da quantidade de requisições http, do controle CSWeb, trouxeram resultados melhores e mais próximos dos obtidos por modelos determinísticos. O processo de sintonia fina do controle fuzzy é necessário.
- O modelo de carga de trabalho implementado pelo controle fuzzy pode ser utilizado para predição do comportamento da carga de trabalho. Essa solução é ideal para predições do comportamento da carga de trabalho e pode ser usada em diversas simulações do impacto do fenômeno das rajadas na utilização do serviço, mesmo que ele não exista.

Assim, conclui-se que o objetivo do trabalho foi atingido, mostrando que o modelo fuzzy é representativo e simples em relação à carga de trabalho e produz resultados aproximados aos obtidos por modelos determinísticos, amplamente utilizados.

7.1 Trabalhos Futuros

A partir dos resultados obtidos com esse trabalho é possível fazer algumas recomendações para trabalhos futuros. Tais sugestões são apresentadas a seguir:

- Realizar um estudo sobre modelos de carga de trabalho adaptados à ocorrência de outros invariantes, como o fenômeno de cauda longa (lei de Zipf). As implementações do modelo da carga de trabalho podem ser feitas usando controles fuzzy e modelos determinísticos.
- Utilizar mais quantificadores fuzzy na especificação das variáveis de entrada e de saída dos controles RAJIN e CSWeb. Podem ser utilizados, por exemplo, quantificadores com as denominações: “baixo”, “meio baixo”, “médio”, “meio alta”, “alta”. Verificar, analisando os resultados obtidos, se há melhoria na precisão do resultado fornecido pelo controle.

- Utilizar outros operadores e métodos para o processo de fuzzificação e defuzzificação dos controles fuzzy RAJIN e CSWeb. Comparar se os resultados obtidos com os novos operadores estão mais próximos dos resultados dos modelos determinísticos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, V. A. F.; ARLITT M.; ROLIA J. Analyzing a web-based system's performance measures at multiple time scales, **SIGMETRICS Performance Evaluation Review**, v. 30, n. 2, p. 3-9, 2002.

ANDERBERG, M. R. **Cluster analysis for applications**. Nova York: Academic Press, 1973. 359 p. ISBN 01-205-7650-3

ARLITT M.; KRISHNAMURTHY, D.; ROLIA J. Characterizing the scalability of a large web-based shopping system. **ACM Transactions on Internet Technology**, p. 44-69, 2001.

ARLITT M.; WILLIAMSON C. Web server workload characterization: the search for invariants. In: ACM 1996 SIGMETRICS CONF. MEASUREMENT COMPUT. SYST., 1996, Philadelphia. **Proceedings...** Philadelphia: Pennsylvania, 1996. p. 126-137.

BANGA, G.; DRUSCHEL, P. Measuring the capacity of a web server under realistic loads. **World Wide Web Journal**, v. 2, n. 1-2, p. 69-83, 1999.

BANGA, G.; DRUSCHEL, P. Measuring the capacity of a web server. In: USENIX SYMPOSIUM ON INTERNET TECHNOLOGIES AND SYSTEMS (USITS), 1997, Rice University. **Proceedings...** Rice University: Department of Computer Science, 1997.

BARFORD P.; CROVELLA M. Generating representative web workloads for network and server performance evaluation. **Sigmetrics ACM**, 1998.

BRETON, P. **História da informática**. São Paulo: Unesp, 1991. 260 p. ISBN: 8571390215.

BUCKLEY J. J.; REILLY, K.; ZHENG, X. **Fuzzy probabilities and fuzzy sets for web planning**. Studies in fuzziness and soft computing. Berlin: Springer-Verlag, 2004, 190 p. ISBN 3540004734.

BUCKLEY J. J.; REILLY, K.; ZHENG, X. **Fuzzy probabilities for web planning**. Soft computing - a fusion of foundations, methodologies and applications. v. 8, n. 7, Berlin: Springer-Verlag, 2004, p. 464 – 476. ISSN 1432-7643.

BUZEN, J. P. Operational analysis: an alternative to stochastic modeling. In: INTERNATIONAL CONFERENCE PERFORMANCE COMPUTER INSTALLATIONS, 1978, Amsterdam. **Proceedings...** Amsterdam: North-Holland, p. 175-194. 1978.

CALZAROSSA M.; MASSARI L.; TESSERA D. Workload characterization - issues and methodologies. **Lecture Notes in Computer Science: Performance Evaluation: Origins and Directions**. v. 1769, p. 459-484. Springer-Verlag, 2000.

COCKCROFT A., WALKER B. **Capacity planning for internet services**. Sun Blue Prints. New Jersey: Prentice Hall PTR, 2001. 256 p. ISBN 0130894028.

COX, E. **The fuzzy systems handbook**: a practitioner's guide to building, using, and maintaining fuzzy system. Academic Press, 1999. 716 p. ISBN 0-12-194455-7.

DENNING, P.; BUZEN J. The operational analysis of queueing network models. **CMG Transactions**. p. 29-60, 1994.

DIAO, Y.; HELLERSTEIN, J. L.; PAREKH, S. Using fuzzy control to maximize profits in service level management, **Artificial Intelligence IBM Systems Journal**, v. 41, n. 3, p 403-420, 2002.

DILLEY J.; FRIEDRICH R.; JIN T.; ROLIA J. Web server performance: measurement and modeling techniques. **Performance Evaluation Journal**. v.33, p. 5-26, 1998.

FEITELSON, D.; TALBY D.; JONES J. P. **Standard workload format**. Disponível em: <<http://www.cs.huji.ac.il/labs/parallel/workload/swf.html>>. Acesso em: 10 Ago 2005.

FOCA LINUX: tipos de execução de comandos e programas. **Guia Foca GNU/Linux**, São Paulo, 31 Jul 2005. Disponível em: <<http://focalinux.cipsga.org.br/guia/intermediario/ch-run.htm>>. Acesso em: 31 Jul 2005.

FORTES, D. Grid computing. **INFO Exame**, n. 219, p. 60-67, 2004.

GUNTHER, N. J. Benchmarking blunders and things that go bump in the night. In: WORKSHOP ON SOFTWARE PERFORMANCE AND RELIABILITY (WOPR2), 2004, California. **Proceedings...** California: Menlo Park, 2004.

GUNTHER N. J., Hit-and-run tactics enable guerrilla part I and II. **Measure IT**. abr/jun 2003.

GUNTHER, N. J. Hit-and-run tactics enable guerrilla capacity planning. **IEEE IT Professional**, p. 40-46, jul/ago, 2002.

GUNTHER Neil J. Performance and scalability models for a hyper-growth e-commerce web site. **Springer Lecture Notes in Computer Science - Performance Engineering: State of the Art and Current Trends**. v. 2047, p. 267-282, 2001.

LARSEN K. R. T.; BLONJARZ P. A. A cost and performance model for web service investment, **Communications of the ACM**, v. 43, n.2, 2000.

MAMDANI, E. H.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. **International Journal of Human-Computer Studies**. V. 51, n. 2, p. 135-147, 1999.

MATLAB. **The Mathworks**, 21 Out 2005. Disponível em: <<http://www.mathworks.com/>>. Acesso em: 25 Nov 2005.

MENASCÉ D. A.; ALMEIDA, V. A. F.; DOWDY L. W. **Performance by design:** computer capacity planning by example. Prentice Hall, 2004. 552 p. ISBN: 0-13-090673-5.

MENASCÉ D. A.; ALMEIDA V. A. F. A hierarchical and multiscale approach to analyze e-business workloads Source. **Performance Evaluation**. v. 54, n.1, p. 33-57, 2003.

MENASCÉ D. A.; ALMEIDA V. A. F. **Planejamento de capacidade para serviços na web:** métricas, modelos e métodos. Rio de Janeiro: Campus, 2003. 472 p. ISBN 85-352-1102-0.

MENASCÉ D. A.; ALMEIDA V. A. F. Capacity planning: an essential tool for managing web services. **IEEE IT Professional**. p. 33-38, jul./ago. 2002.

MENASCÉ D. A.; BARBARA D.; DODGE R. Preserving QoS of ecommerce sites through selftuning: a performance model approach. In: EC'01, 2001, Florida. **Proceedings...** New York: ACM Press, 2001. p. 224-234. ISBN:1-58113-387-1

MENASCÉ, D. A.; BARBARA, D.; DODGE, R. Preserving QoS of e-commerce sites through self-tuning: a performance model approach. In: ACM CONFERENCE ON E-COMMERCE, 2001, Florida. **Proceedings...** Florida, 2001. p. 224-234.

MENASCÉ D. A. Web performance modeling issues. **International Journal of High Performance Computing Applications**. v. 14, 2000.

MENASCÉ D. A.; ALMEIDA V. A. F. **Scaling for e-business:** technologies, models, performance and capacity planning. Prentice Hall, 2000. 449 p. ISBN 0-13-086328-9.

MENASCÉ D. A.; ALMEIDA V. A. F. In search of invariants for e-business workloads. In: 2° ACM CONFERENCE ON ELECTRONIC COMMERCE, 2000, Minneapolis. **Proceedings...** Minneapolis, 2000. p. 56-65.

MENASCÉ D. A.; PERAINO B.; DINH N. Planning the capacity of a web server: an experience report. In: COMPUTER MEASUREMENT GROUP CONFERENCE, Reno, 1999. **Proceedings...** Reno, Dez. 1999.

MOLLOY, C. Using ITIL best practices to create a capacity management process **Measure IT**, v. 8, 2003.

MUNAKATA, T; YASHVANT J. Fuzzy systems: an overview. **Communications of the ACM**, v. 37, n. 3, p. 68-76, 1994.

RECEITA Federal suspende serviço. **Folha Online**, São Paulo, 09 jun. 2004. Disponível em: <<http://www1.folha.uol.com.br/folha/dinheiro/ult91u85377.shtml>>. Acesso em: 01 Ago 2004.

REED D.; AYDT R. Performance contracts: a fuzzy logic perspective. **GrADS meeting**, 1999.

SEYBOLD P.B.; MARSHAK R.T. **Cientes.com**. São Paulo: Makron Books, 2000. 382 p. ISBN: 8534611556

SILVEIRA, A. M. et. al. Identificação de abordagens administrativas: um ensaio com lógica fuzzy, **INFOCOMP Journal of Computer Science**, v. 4, n. 1, p. 36-45, 2005.

SQUID Web Proxy Cache. **Squid Web Proxy Cache**, 21 Out 2005. Disponível em: <<http://www.squid-cache.org/>>. Acesso em: 25 Nov 2005.

THING L. **Dicionário de tecnologia**. São Paulo: Futura, 2003. 1005 p. ISBN 85-7413-138-5.

WANG, Q.; MAKAROFF, D.; EDWARDS, H. K.; THOMPSON, R. Workload characterization for an e-commerce web site. In: 2003 CONFERENCE OF THE CENTRE FOR ADVANCED STUDIES ON COLLABORATIVE RESEARCH, Canada, 2003. **Proceedings...** Canada: IBM Centre for Advanced Studies Conference, Out. 2003, p. 313-327.

WANG, L.X. **A course in fuzzy systems and control**. Prentice-Hall, New Jersey, 1996. 424 p. ISBN:0-13-540882-2

WEB application deployment: a practical approach to capacity planning. **IBM Global Services**. Disponível em: <<http://www-1.ibm.com/services/us/its/pdf/g563-0339-00.pdf>>. Acesso em: 01 Jan 2004.

WINDOWS NT performance monitor in depth. **Windows IT library**. 10 Abr 1999. Disponível em: <<http://www.windowsitlibrary.com/Content/113/02/toc.html>>. Acesso em: 30 Jun 2005.

ZADEH, L.A. **Fuzzy sets and applications**: selected papers by L.A. Zadeh. New York: Wiley and Sons, 1987. 684 p. ISBN: 0471857106

ZADEH, L.A. The role of fuzzy logic in the management of uncertainty in expert systems. **Fuzzy sets and applications**: selected papers by L.A. Zadeh, Wiley and Sons, p. 413-437. 1987.

ZADEH, L.A. Outline of a new approach to the analysis of complex systems and decision processes. **IEEE Transactions on system, man and cybernetics**, v. SMC-3, n.1, 1973.