

Capacity Planning For

WEB PERFORMANCE

**METRICS,
MODELS, &
METHODS**

DANIEL A. MENASCÉ • VIRGILIO A.F. ALMEIDA



Capacity Planning for Web Performance

Metrics, Models, and Methods

Daniel A. Menascé

Department of Computer Science, George Mason University, Fairfax, Virginia
menasce@cs.gmu.edu

Virgilio A.F. Almeida

Department of Computer Science, Federal University of Minas Gerais, Brazil
virgilio@dcc.ufmg.br



Prentice Hall, PTR
Upper Saddle River, New Jersey 07458
<http://www.phptr.com>



CONTENTS

PREFACE	xi
1 WHEN PERFORMANCE IS A PROBLEM	1
1.1 Introduction	1
1.2 Client/Server Performance	2
1.3 The Capacity Planning Concept	6
1.4 Web Server Performance	8
1.5 Intranet Performance	10
1.6 Internet Service Provider (ISP) Performance	11
1.7 Summary	12
BIBLIOGRAPHY	15
2 WHAT ARE CLIENT/SERVER SYSTEMS?	16
2.1 Introduction	16
2.2 The World of Networks	17
2.2.1 Genesis	17
2.2.2 Types of Networks	17
2.2.3 Protocols	21
2.3 The World of Clients and Servers	28
2.3.1 The Client/Server Paradigm	28
2.3.2 Server Types	30
2.3.3 Architectural Issues	31
2.4 Concluding Remarks	32
BIBLIOGRAPHY	33

3	PERFORMANCE ISSUES IN CLIENT/SERVER ENVIRONMENTS	35
3.1	Introduction	35
3.2	Communication-Processing Delay Diagrams	35
3.3	Service Times and Service Demands	40
3.3.1	Service Times at Single Disks and Disk Arrays	41
3.3.2	Service Times in Networks	54
3.3.3	Service Times at Routers	60
3.4	Queues and Contention	61
3.5	Some Basic Performance Results	64
3.5.1	Utilization Law	64
3.5.2	Forced Flow Law	65
3.5.3	Service Demand Law	65
3.5.4	Little's Law	65
3.5.5	Summary of Basic Results	67
3.6	Performance Metrics in C/S Systems	67
3.7	Concluding Remarks	69
	BIBLIOGRAPHY	70
4	WEB SERVER AND INTRANET PERFORMANCE ISSUES	71
4.1	Introduction	71
4.2	More than Just Servers	72
4.3	Where Are the Delays?	77
4.3.1	Anatomy of a Web Transaction	77
4.3.2	Bottlenecks	80
4.4	Perception of Performance	81
4.4.1	Metrics	82
4.4.2	Quality of Service	83
4.5	Infrastructure	85
4.5.1	Basic Components	85
4.5.2	Proxy, Cache, and Mirror	85
4.6	Web Server	89
4.6.1	Architecture	89
4.6.2	Workload	90
4.7	Intranet and the Internet	94
4.7.1	Bandwidth and Latency	94
4.7.2	Traffic	95

4.8 Capacity Planning	96
4.9 Summary	97

BIBLIOGRAPHY	98
---------------------	-----------

5 A STEP-BY-STEP APPROACH TO CAPACITY PLANNING IN CLIENT/SERVER SYSTEMS	100
--	------------

5.1 Introduction	100
5.2 Adequate Capacity	101
5.3 A Methodology for Capacity Planning in C/S Environments	102
5.4 Understanding the Environment	103
5.5 Workload Characterization	106
5.5.1 Breaking Down the Global Workload	107
5.5.2 Data Collection Issues	109
5.5.3 Validating Workload Models	110
5.6 Workload Forecasting	111
5.7 Performance Modeling and Prediction	112
5.7.1 Performance Models	112
5.7.2 Performance Prediction Techniques	115
5.7.3 Performance Model Validation	115
5.8 Development of a Cost Model	117
5.9 Cost/Performance Analysis	119
5.10 Concluding Remarks	119

BIBLIOGRAPHY	120
---------------------	------------

6 UNDERSTANDING AND CHARACTERIZING THE WORKLOAD	121
--	------------

6.1 Introduction	121
6.2 Characterizing the Workload for an Intranet	122
6.2.1 A First Approach	124
6.2.2 A Simple Example	126
6.2.3 Workload Model	129
6.3 A Workload Characterization Methodology	132
6.3.1 Choice of an Analysis Standpoint	132
6.3.2 Identification of the Basic Component	132
6.3.3 Choice of the Characterizing Parameter	132
6.3.4 Data Collection	133
6.3.5 Partitioning the Workload	134

6.3.6	Calculating Class Parameters	138
6.4	Bursty Workloads	149
6.5	Conclusions	151
BIBLIOGRAPHY		153
7	USING STANDARD INDUSTRY BENCHMARKS	155
7.1	Introduction	155
7.2	The Nature of Benchmarks	156
7.2.1	Benchmark Hierarchy	157
7.2.2	Avoiding Pitfalls	158
7.2.3	Common Benchmarks	159
7.3	Component-Level Benchmarks	159
7.3.1	CPU	160
7.3.2	File Server	162
7.4	System-Level Benchmarks	164
7.4.1	Transaction Processing Systems	164
7.4.2	Web Servers	166
7.5	Conclusions	172
BIBLIOGRAPHY		173
8	SYSTEM-LEVEL PERFORMANCE MODELS	174
8.1	Introduction	174
8.2	Simple Server Model I—Infinite Population/Infinite Queue	174
8.3	Simple Server Model II—Infinite Population/Finite Queue	181
8.4	Generalized System-Level Models	184
8.5	Other System-Level Models	186
8.5.1	Infinite Population Models	187
8.5.2	Finite Population Models	191
8.6	Concluding Remarks	194
BIBLIOGRAPHY		196
9	COMPONENT-LEVEL PERFORMANCE MODELS	197
9.1	Introduction	197
9.2	Queuing Networks	197
9.3	Open Systems	199
9.3.1	Single-Class Open Queuing Networks	199

9.3.2	Multiple-Class Open Queuing Networks	202
9.4	Closed Models	204
9.4.1	Single-Class Closed Models	205
9.4.2	Multiple-Class Closed Models	209
9.5	Modeling Multiprocessors	213
9.6	An Intranet Model	217
9.7	Concluding Remarks	220
BIBLIOGRAPHY		221
10 WEB PERFORMANCE MODELING		222
10.1	Introduction	222
10.2	Incorporating New Phenomena	222
10.2.1	Burstiness Modeling	222
10.2.2	Accounting for Heavy Tails in the Model	227
10.3	Client-Side Models	228
10.3.1	No Cache Proxy Server Case	228
10.3.2	Using a Cache Proxy Server	235
10.4	Server-Side Models	237
10.4.1	Single Web Server	238
10.4.2	Mirrored Web Servers	243
10.5	Concluding Remarks	246
BIBLIOGRAPHY		248
11 WORKLOAD FORECASTING		250
11.1	Introduction	250
11.2	A Forecasting Strategy	251
11.3	From Business Processes to Workload Parameters	253
11.4	Forecasting Techniques	256
11.4.1	Regression Methods	256
11.4.2	Moving Average	258
11.4.3	Exponential Smoothing	259
11.4.4	Applying Forecasting Techniques	261
11.5	Concluding Remarks	262
BIBLIOGRAPHY		263

12 MEASURING PERFORMANCE	264
12.1 Introduction	264
12.2 Performance Measurement Framework	265
12.3 Measurement Techniques	267
12.4 Data Collection Tools	270
12.4.1 Hardware Monitor	270
12.4.2 Software Monitor	271
12.5 Performance Model Parameters	274
12.5.1 Queues	275
12.5.2 Workload Classes	275
12.5.3 Workload Intensity	276
12.5.4 Service Demands	276
12.5.5 Parameter Estimation	278
12.6 Collecting Performance Data	281
12.6.1 Network	281
12.6.2 Server	282
12.7 Concluding Remarks	288
BIBLIOGRAPHY	290
13 WRAPPING UP	292
BIBLIOGRAPHY	295
A GLOSSARY OF TERMS	296
B ABOUT THE CD-ROM	309
B.1 The Workbooks	309
B.2 HTTP Log Sample and Program	310
SUBJECT INDEX	311



Capacity Planning For WEB PERFORMANCE

DANIEL A. MENASCÉ • VIRGILIO A.F. ALMEIDA

As more and more businesses rely on distributed client/server and Web-based applications, performance considerations become extremely important. **Capacity Planning for Web Performance** uses quantitative methods to analyze these systems. It leads the capacity planner, in a step-by-step fashion, through the process of determining the most cost-effective system configurations and networking architectures. The quantitative methods lead to the development of performance-predictive models for capacity planning. Instead of relying on intuition, ad hoc procedures, and rules of thumb, **Capacity Planning for Web Performance** provides a uniform and sound way for dealing with performance problems. A large number of numeric and practical examples help the reader understand the quantitative approach adopted here.

Includes a CD-ROM containing several Microsoft Excel® workbooks supported by Visual Basic® modules, samples of http logs, and programs to process them. The Excel workbooks allow the reader to immediately put into practice the methods and models discussed here.

About the Authors

DANIEL A. MENASCÉ

is a Professor of Computer Science at George Mason University, VA. He has published extensively in the area of performance modeling, client/system performance evaluation, and software performance engineering. Menascé was elected a Fellow of the Association for Computing Machinery (ACM) in recognition of outstanding contributions to information technology.

VIRGILIO A.F. ALMEIDA

is a Professor of Computer Science at the Federal University of Minas Gerais (UFMG), Brazil. He has published extensively in the area of distributed systems and World Wide Web performance. Almeida held visiting faculty and research positions at Boston University and XEROX PARC.

"This book takes the mystery out of analyzing Web performance. The authors have skillfully culled through more than 25 years of research, and selected the results most critical to Web performance, and developed important new material that deals directly with the special properties of applications that run on the Web. With everything together in a single volume, Menascé and Almeida have created a superb starting point for anyone wishing to explore the world of Web performance."

Jeffrey P. Buzen

Chief Scientist and CoFounder, BGS Systems

"Many have said that the Web is too amorphous and chaotic to permit meaningful performance forecasts. Menascé and Almeida demolish this myth. Throughput, response time, and congestion can be measured and predicted, all using familiar tools from queueing networks that you can run on your own computer. There is no other book like this. It is a first."

Peter J. Denning

Professor of Computer Science,
George Mason University
and former President of the ACM

"This excellent book presents a new way to model, analyze, and plan for these new performance problems associated with the Web's bursty and highly-skewed load characteristics. A valuable resource for students and for Web administrators."

Jim Gray

Senior Researcher, Microsoft Research

"This is a welcome approach to the performance analysis of today's Web-based Internet. It is a useful and practical treatment that is eminently accessible to the non-mathematical professional. An impressive feature the authors provide is to deal directly with the fractal nature of Web-based traffic; no simple and practical treatment has been offered before, and theirs is a timely contribution."

Leonard Kleinrock

Professor of Computer Science, UCLA

PRENTICE HALL
Upper Saddle River, NJ 07458
<http://www.phptr.com>

ISBN 0-13-693822-1



9 0000



9 780136 938224