Contents

Foreword xxi
Acknowledgments xxiii
About the Authors xxv

Introduction 1

Part I: Staffing a Scalable Organization 7

Chapter 1: The Impact of People and Leadership on Scalability 9
- Introducing AllScale 9
- Why People 10
- Why Organizations 11
- Why Management and Leadership 17
- Conclusion 20
  - Key Points 20

Chapter 2: Roles for the Scalable Technology Organization 21
- The Effects of Failure 22
- Defining Roles 23
- Executive Responsibilities 25
  - CEO 25
  - CFO 26
  - Business Unit Owners, General Managers, and P&L Owners 27
  - CTO/CIO 27
- Organizational Responsibilities 29
  - Architecture Responsibilities 29
  - Engineering Responsibilities 30
  - Production Operations Responsibilities 30
  - Infrastructure Responsibilities 31
  - Quality Assurance Responsibilities 31
  - Capacity Planning Responsibilities 32
- Individual Contributor Responsibilities and Characteristics 32
  - Architect 33

Software Engineer 33
Operator 34
Infrastructure Engineer 34
QA Engineer/Analyst 35
Capacity Planner 35
An Organizational Example 35
A Tool for Defining Responsibilities 37
Conclusion 41
Key Points 41

Chapter 3: Designing Organizations 43
Organizational Influences That Affect Scalability 43
Team Size 46
Warning Signs 50
Growing or Splitting Teams 52
Organizational Structure 55
Functional Organization 55
Matrix Organization 57
Conclusion 60
Key Points 61

Chapter 4: Leadership 101 63
What Is Leadership? 64
Leadership—A Conceptual Model 66
Taking Stock of Who You Are 67
Leading from the Front 69
Checking Your Ego at the Door 71
Mission First, People Always 72
Making Timely, Sound, and Morally Correct Decisions 73
Empowering Teams and Scalability 74
Alignment with Shareholder Value 74
Vision 75
Mission 78
Goals 79
Putting Vision, Mission, and Goals Together 81
The Causal Roadmap to Success 84
Conclusion 86
Key Points 87

Chapter 5: Management 101 89
What Is Management? 90
Project and Task Management 91
Building Teams—A Sports Analogy 93
Upgrading Teams—A Garden Analogy 94
Measurement, Metrics, and Goal Evaluation 98
The Goal Tree 101
Paving the Path for Success 102
Conclusion 103
Key Points 104

Chapter 6: Making the Business Case 105
Understanding the Experiential Chasm 105
Why the Business Executive Might Be the Problem 106
Why the Technology Executive Might Be the Problem 107
Defeating the Corporate Mindset 109
Forming Relationships 111
Setting the Example 111
Educating Other Executives 111
Using the RASCI Model 112
Speaking in Business Terms 112
Getting Them Involved 113
Scaring the Executive Team with Facts 113
The Business Case for Scale 114
Conclusion 117
Key Points 117

Part II: Building Processes for Scale 119

Chapter 7: Understanding Why Processes Are Critical to Scale 121
The Purpose of Process 122
Right Time, Right Process 125
How Much Rigor 126
How Complex 128
When Good Processes Go Bad 130
Conclusion 131
Key Points 132

Chapter 8: Managing Incidents and Problems 133
What Is an Incident? 134
What Is a Problem? 135
The Components of Incident Management 136
The Components of Problem Management 139
Resolving Conflicts Between Incident and Problem Management 140
Incident and Problem Life Cycles 140
Implementing the Daily Incident Meeting 141
Implementing the Quarterly Incident Review 143
The Postmortem Process 143
Putting It All Together 146
Conclusion 148
Key Points 148

Chapter 9: Managing Crisis and Escalations 149
What Is a Crisis? 149
Why Differentiate a Crisis from Any Other Incident? 150
How Crises Can Change a Company 151
Order Out of Chaos 152
The Role of the "Problem Manager" 153
The Role of Team Managers 155
The Role of Engineering Leads 156
The Role of Individual Contributors 157
Communications and Control 157
The War Room 158
Escalations 160
Status Communications 160
Crises Postmortems 161
Crises Follow-up and Communication 162
Conclusion 163
Key Points 163

Chapter 10: Controlling Change in Production Environments 165
What Is a Change? 166
Change Identification 168
Change Management 170
Change Proposal 172
Change Approval 174
Change Scheduling 174

Change Implementation and Logging 176
Change Validation 176
Change Review 177
The Change Control Meeting 178
Continuous Process Improvement 178
Conclusion 179
Key Points 180

Chapter 11: Determining Headroom for Applications 183

Purpose of the Process 184
Structure of the Process 185
Ideal Usage Percentage 189
Conclusion 192
Key Points 193

Chapter 12: Exploring Architectural Principles 195

Principles and Goals 196
Principle Selection 199
AKF's Twelve Architectural Principles 200
N+1 Design 200
Design for Rollback 201
Design to Be Disabled 201
Design to Be Monitored 202
Design for Multiple Live Sites 202
Use Mature Technologies 202
Asynchronous Design 202
Stateless Systems 202
Scale Out Not Up 203
Design for at Least Two Axes of Scale 203
Buy When Non Core 203
Use Commodity Hardware 203
Scalability Principles In Depth 204
Design to Be Monitored 204
Design for Multiple Live Sites 205
Asynchronous Design 205
Stateless Systems 206
Scale Out Not Up 207
Design for at Least Two Axes of Scale 207
Conclusion 208
Key Points 209

Chapter 13: Joint Architecture Design 211
Fixing Organizational Dysfunction 211
Designing for Scale Cross Functionally 214
Entry and Exit Criteria 217
Conclusion 219
Key Points 220

Chapter 14: Architecture Review Board 221
Ensuring Scale Through Review 221
Board Constituency 223
Conducting the Meeting 225
Entry and Exit Criteria 228
Conclusion 230
Key Points 231

Chapter 15: Focus on Core Competencies: Build Versus Buy 233
Building Versus Buying, and Scalability 233
Focusing on Cost 234
Focusing on Strategy 235
"Not Built Here" Phenomenon 236
Merging Cost and Strategy 237
Does This Component Create Strategic Competitive Differentiation? 238
Are We the Best Owners of This Component or Asset? 238
What Is the Competition to This Component? 239
Can We Build This Component Cost Effectively? 239
AllScale's Build or Buy Dilemma 240
Conclusion 242
Key Points 242

Chapter 16: Determining Risk 243
Importance of Risk Management to Scale 244
Measuring Risk 245
Managing Risk 252
Conclusion 255
Key Points 256

Chapter 17: Performance and Stress Testing 257
Performing Performance Testing 257
Criteria 258

Environment 259
Define Tests 260
Execute Tests 261
Analyze Data 261
Report to Engineers 262
Repeat Tests and Analysis 262
Don't Stress Over Stress Testing 264
Identify Objectives 264
Identify Key Services 265
Determine Load 266
Environment 266
Identify Monitors 267
Create Load 267
Execute Tests 267
Analyze Data 268
Performance and Stress Testing for Scalability 270
Conclusion 271
Key Points 272

Chapter 18: Barrier Conditions and Rollback 273

Barrier Conditions 274
Barrier Conditions and Agile Development 275
Barrier Conditions and Waterfall Development 277
Barrier Conditions and Hybrid Models 278
Rollback Capabilities 278
Rollback Window Requirements 279
Rollback Technology Considerations 281
Cost Considerations of Rollback 281
Markdown Functionality—Design to Be Disabled 282
Conclusion 283
Key Points 284

Chapter 19: Fast or Right? 285

Tradeoffs in Business 285
Relation to Scalability 289
How to Think About the Decision 290
Conclusion 295
Key Points 296

Part III: Architecting Scalable Solutions 297

Chapter 20: Designing for Any Technology 299

An Implementation Is Not an Architecture 300
Technology Agnostic Design 300
TAD and Cost 301
TAD and Risk 302
TAD and Scalability 303
TAD and Availability 306
The TAD Approach 306
Conclusion 308
Key Points 308

Chapter 21: Creating Fault Isolative Architectural Structures 309

Fault Isolative Architecture Terms 310
Benefits of Fault Isolation 312
Fault Isolation and Availability—Limiting Impact 312
Fault Isolation and Availability—Incident Detection and Resolution 315
Fault Isolation and Scalability 315
Fault Isolation and Time to Market 315
Fault Isolation and Cost 316
How to Approach Fault Isolation 317
Principle 1: Nothing Is Shared 318
Principle 2: Nothing Crosses a Swim Lane Boundary 319
Principle 3: Transactions Occur Along Swim Lanes 319
When to Implement Fault Isolation 319
Approach 1: Swim Lane the Money-Maker 320
Approach 2: Swim Lane the Biggest Sources of Incidents 320
Approach 3: Swim Lane Along Natural Barriers 320
How to Test Fault Isolative Designs 321
Conclusion 322
Key Points 322

Chapter 22: Introduction to the AKF Scale Cube 325

Concepts Versus Rules and Tools 325
Introducing the AKF Scale Cube 326
Meaning of the Cube 328
The X-Axis of the Cube 328

The Y-Axis of the Cube....331
The Z-Axis of the Cube....333
Putting It All Together....334
When and Where to Use the Cube....336
Conclusion....337
Key Points....338

Chapter 23: Splitting Applications for Scale....339

The AKF Scale Cube for Applications....339
The X-Axis of the AKF Application Scale Cube....341
The Y-Axis of the AKF Application Scale Cube....343
The Z-Axis of the AKF Application Scale Cube....344
Putting It All Together....347
Practical Use of the Application Cube....349
Ecommerce Implementation....350
Human Resources ERP Implementation....351
Back Office IT System....352
Observations....353
Conclusion....354
Key Points....355

Chapter 24: Splitting Databases for Scale....357

The AKF Scale Cube for Databases....357
The X-Axis of the AKF Database Scale Cube....358
The Y-Axis of the AKF Database Scale Cube....362
The Z-Axis of the AKF Database Scale Cube....365
Putting It All Together....367
Practical Use of the Database Cube....370
Ecommerce Implementation....370
Human Resources ERP Implementation....372
Back Office IT System....372
Observations....373
Timeline Considerations....373
Conclusion....374
Key Points....375

Chapter 25: Caching for Performance and Scale....377

Caching Defined....378
Object Caches....381

Application Caches 384
Proxy Caches 384
Reverse Proxy Cache 386
Caching Software 388
Content Delivery Networks 389
Conclusion 390
Key Points 391

Chapter 26: Asynchronous Design for Scale 393
Synching Up on Synchronization 393
Synchronous Versus Asynchronous Calls 395
Scaling Synchronously or Asynchronously 396
Example Asynchronous Systems 398
Defining State 401
Conclusion 405
Key Points 406

Part IV: Solving Other Issues and Challenges 409

Chapter 27: Too Much Data 411
The Cost of Data 412
The Value of Data and the Cost-Value Dilemma 414
Making Data Profitable 416
Option Value 416
Strategic Competitive Differentiation 416
Cost Justify the Solution (Tiered Storage Solutions) 417
Transform the Data 419
Handling Large Amounts of Data 420
Conclusion 423
Key Points 424

Chapter 28: Clouds and Grids 425
History and Definitions 426
Grid Computing 428
Public Versus Private Clouds 430
Characteristics and Architecture of Clouds 430
Pay By Usage 431
Scale On Demand 431

Multiple Tenants 432
Virtualization 433
Differences Between Clouds and Grids 434
Types of Clouds 435
Conclusion 436
Key Points 437

Chapter 29: Soaring in the Clouds 439

Pros and Cons of Cloud Computing 440
Pros of Cloud Computing 440
Cons of Cloud Computing 443
Where Clouds Fit in Different Companies 448
Environments 448
Skill Sets 449
Decision Process 450
Conclusion 453
Key Points 454

Chapter 30: Plugging in the Grid 455

Pros and Cons of Grids 456
Pros of Grids 456
Cons of Grids 458
Different Uses for Grid Computing 461
Production Grid 461
Build Grid 462
Data Warehouse Grid 463
Back Office Grid 464
Decision Process 465
Conclusion 467
Key Points 468

Chapter 31: Monitoring Applications 469

"How Come We Didn't Catch That Earlier?" 469
A Framework for Monitoring 472
User Experience and Business Metrics 476
Systems Monitoring 477
Application Monitoring 477
Measuring Monitoring: What Is and Isn't Valuable? 478
Monitoring and Processes 480

Conclusion 481
Key Points 482

Chapter 32: Planning Data Centers 483
Data Center Costs and Constraints 483
Location, Location, Location 485
Data Centers and Incremental Growth 488
Three Magic Rules of Three 490
The First Rule of Three: Three Magic Drivers of Data Center Costs 491
The Second Rule of Three: Three Is the Magic Number for Servers 491
The Third Rule of Three: Three Is the Magic Number for Data Centers 492
Multiple Active Data Center Considerations 496
Conclusion 498
Key Points 499

Chapter 33: Putting It All Together 501
What to Do Now? 502
Case Studies 505
eBay: Incredible Success and a Scalability Implosion 505
Quigo: A Young Product with a Scalability Problem 506
ShareThis: A Startup Story 507
References 509

Appendices 511

Appendix A: Calculating Availability 513
Hardware Uptime 514
Customer Complaints 515
Portion of Site Down 516
Third-Party Monitoring Service 517
Traffic Graph 518

Appendix B: Capacity Planning Calculations 521

Appendix C: Load and Performance Calculations 527

Index 535