Table of Contents

Foreword. xiii

Preface. xv

Part I. Introduction

1. Introduction. 3
   The Sysadmin Approach to Service Management 3
   Google's Approach to Service Management: Site Reliability Engineering 5
   Tenets of SRE 7
   The End of the Beginning 12

2. The Production Environment at Google, from the Viewpoint of an SRE. 13
   Hardware 13
   System Software That "Organizes" the Hardware 15
   Other System Software 18
   Our Software Infrastructure 19
   Our Development Environment 19
   Shakespeare: A Sample Service 20

Part II. Principles

3. Embracing Risk. 25
   Managing Risk 25
   Measuring Service Risk 26
   Risk Tolerance of Services 28

Motivation for Error Budgets 33

4. Service Level Objectives. 37
Service Level Terminology 37
Indicators in Practice 40
Objectives in Practice 43
Agreements in Practice 47

5. Eliminating Toil. 49
Toil Defined 49
Why Less Toil Is Better 51
What Qualifies as Engineering? 52
Is Toil Always Bad? 52
Conclusion 54

6. Monitoring Distributed Systems. 55
Definitions 55
Why Monitor? 56
Setting Reasonable Expectations for Monitoring 57
Symptoms Versus Causes 58
Black-Box Versus White-Box 59
The Four Golden Signals 60
Worrying About Your Tail (or, Instrumentation and Performance) 61
Choosing an Appropriate Resolution for Measurements 62
As Simple as Possible, No Simpler 62
Tying These Principles Together 63
Monitoring for the Long Term 64
Conclusion 66

7. The Evolution of Automation at Google. 67
The Value of Automation 67
The Value for Google SRE 70
The Use Cases for Automation 70
Automate Yourself Out of a Job: Automate ALL the Things! 73
Soothing the Pain: Applying Automation to Cluster Turnups 75
Borg: Birth of the Warehouse-Scale Computer 81
Reliability Is the Fundamental Feature 83
Recommendations 84

8. Release Engineering. 87
The Role of a Release Engineer 87
Philosophy 88

Continuous Build and Deployment 90
Configuration Management 93
Conclusions 95

9. Simplicity.. 97
System Stability Versus Agility 97
The Virtue of Boring 98
I Won't Give Up My Code! 98
The "Negative Lines of Code" Metric 99
Minimal APIs 99
Modularity 100
Release Simplicity 100
A Simple Conclusion 101

Part III. Practices

10. Practical Alerting from Time-Series Data.............................. 107
The Rise of Borgmon 108
Instrumentation of Applications 109
Collection of Exported Data 110
Storage in the Time-Series Arena 111
Rule Evaluation 114
Alerting 118
Sharding the Monitoring Topology 119
Black-Box Monitoring 120
Maintaining the Configuration 121
Ten Years On... 122

11. Being On-Call.. 125
Introduction 125
Life of an On-Call Engineer 126
Balanced On-Call 127
Feeling Safe 128
Avoiding Inappropriate Operational Load 130
Conclusions 132

12. Effective Troubleshooting.. 133
Theory 134
In Practice 136
Negative Results Are Magic 144
Case Study 146

Making Troubleshooting Easier 150
Conclusion 150

13. Emergency Response. 151
What to Do When Systems Break 151
Test-Induced Emergency 152
Change-Induced Emergency 153
Process-Induced Emergency 155
All Problems Have Solutions 158
Learn from the Past. Don't Repeat It. 158
Conclusion 159

14. Managing Incidents. 161
Unmanaged Incidents 161
The Anatomy of an Unmanaged Incident 162
Elements of Incident Management Process 163
A Managed Incident 165
When to Declare an Incident 166
In Summary 166

15. Postmortem Culture: Learning from Failure. 169
Google's Postmortem Philosophy 169
Collaborate and Share Knowledge 171
Introducing a Postmortem Culture 172
Conclusion and Ongoing Improvements 175

16. Tracking Outages. 177
Escalator 178
Outalator 178

17. Testing for Reliability. 183
Types of Software Testing 185
Creating a Test and Build Environment 190
Testing at Scale 192
Conclusion 204

18. Software Engineering in SRE. 205
Why Is Software Engineering Within SRE Important? 205
Auxon Case Study: Project Background and Problem Space 207
Intent-Based Capacity Planning 209
Fostering Software Engineering in SRE 218
Conclusions 222

19. Load Balancing at the Frontend.. 223
Power Isn’t the Answer 223
Load Balancing Using DNS 224
Load Balancing at the Virtual IP Address 227

20. Load Balancing in the Datacenter.. 231
The Ideal Case 232
Identifying Bad Tasks: Flow Control and Lame Ducks 233
Limiting the Connections Pool with Subsetting 235
Load Balancing Policies 240

21. Handling Overload.. 247
The Pitfalls of “Queries per Second” 248
Per-Customer Limits 248
Client-Side Throttling 249
Criticality 251
Utilization Signals 253
Handling Overload Errors 253
Load from Connections 257
Conclusions 258

22. Addressing Cascading Failures.. 259
Causes of Cascading Failures and Designing to Avoid Them 260
Preventing Server Overload 265
Slow Startup and Cold Caching 274
Triggering Conditions for Cascading Failures 276
Testing for Cascading Failures 278
Immediate Steps to Address Cascading Failures 280
Closing Remarks 283

23. Managing Critical State: Distributed Consensus for Reliability.................. 285
Motivating the Use of Consensus: Distributed Systems Coordination Failure 288
How Distributed Consensus Works 289
System Architecture Patterns for Distributed Consensus 291
Distributed Consensus Performance 296
Deploying Distributed Consensus-Based Systems 304
Monitoring Distributed Consensus Systems 312
Conclusion 313

24. Distributed Periodic Scheduling with Cron.. 315
Cron 315
Cron Jobs and Idempotency 316

Cron at Large Scale 317
Building Cron at Google 319
Summary 326

25. Data Processing Pipelines.... 327
Origin of the Pipeline Design Pattern 327
Initial Effect of Big Data on the Simple Pipeline Pattern 328
Challenges with the Periodic Pipeline Pattern 328
Trouble Caused By Uneven Work Distribution 328
Drawbacks of Periodic Pipelines in Distributed Environments 329
Introduction to Google Workflow 333
Stages of Execution in Workflow 335
Ensuring Business Continuity 337
Summary and Concluding Remarks 338

26. Data Integrity: What You Read Is What You Wrote.... 339
Data Integrity's Strict Requirements 340
Google SRE Objectives in Maintaining Data Integrity and Availability 344
How Google SRE Faces the Challenges of Data Integrity 349
Case Studies 360
General Principles of SRE as Applied to Data Integrity 367
Conclusion 368

27. Reliable Product Launches at Scale.... 369
Launch Coordination Engineering 370
Setting Up a Launch Process 372
Developing a Launch Checklist 375
Selected Techniques for Reliable Launches 380
Development of LCE 384
Conclusion 387

Part IV. Management

28. Accelerating SREs to On-Call and Beyond.... 391
You've Hired Your Next SRE(s), Now What? 391
Initial Learning Experiences: The Case for Structure Over Chaos 394
Creating Stellar Reverse Engineers and Improvisational Thinkers 397
Five Practices for Aspiring On-Callers 400
On-Call and Beyond: Rites of Passage, and Practicing Continuing Education 406
Closing Thoughts 406

29. Dealing with Interrupts. 407
Managing Operational Load 408
Factors in Determining How Interrupts Are Handled 408
Imperfect Machines 409

30. Embedding an SRE to Recover from Operational Overload. 417
Phase 1: Learn the Service and Get Context 418
Phase 2: Sharing Context 420
Phase 3: Driving Change 421
Conclusion 423

31. Communication and Collaboration in SRE. 425
Communications: Production Meetings 426
Collaboration within SRE 430
Case Study of Collaboration in SRE: Viceroy 432
Collaboration Outside SRE 437
Case Study: Migrating DFP to F1 437
Conclusion 440

32. The Evolving SRE Engagement Model. 441
SRE Engagement: What, How, and Why 441
The PRR Model 442
The SRE Engagement Model 443
Production Readiness Reviews: Simple PRR Model 444
Evolving the Simple PRR Model: Early Engagement 448
Evolving Services Development: Frameworks and SRE Platform 451
Conclusion 456

Part V. Conclusions

33. Lessons Learned from Other Industries. 459
Meet Our Industry Veterans 460
Preparedness and Disaster Testing 462
Postmortem Culture 465
Automating Away Repetitive Work and Operational Overhead 467
Structured and Rational Decision Making 469
Conclusions 470

34. Conclusion. 473

A. Availability Table.. 477

B. A Collection of Best Practices for Production Services........................... 479

C. Example Incident State Document.. 485

D. Example Postmortem.. 487

E. Launch Coordination Checklist.. 493

F. Example Production Meeting Minutes... 497

Bibliography.. 501

Index... 513