Performance Guarantees for Web Applications

Jiang, D.

2012

document version
Publisher's PDF, also known as Version of record

Contents

Acknowledgement **v**

1 Introduction **1**

2 Related work **7**
2.1 Framework 8
2.1.1 Client-side latency 8
2.1.2 Network-induced delays 10
2.1.3 Server-side latency 11
2.2 Metric determination 12
2.2.1 Performance metrics 12
2.2.2 Cost metrics 13
2.3 Scalability mechanisms 13
2.3.1 Business-logic tier 14
2.3.2 Data tier 15
2.3.3 Load balancing 18
2.4 Dynamic control 20
2.4.1 Workload characterization 21
2.4.2 Admission control and request scheduling 23
2.4.3 Dynamic resource provisioning 26
2.4.4 Hosting environment 30
2.5 Conclusion 31

3 Challenges **33**
3.1 Application scalability challenge 34
3.1.1 Methodology 35
3.1.2 Evaluation 36
3.1.3 Discussion 36
3.2 Performance modeling challenge 37
3.2.1 Methodology 39

3.2.2 Evaluation . . . 41
3.2.3 Discussion . . . 43
3.3 Resource heterogeneity challenge . . . 44
3.3.1 Methodology . . . 45
3.3.2 Evaluation . . . 47
3.3.3 Discussion . . . 53

4 Making Web applications scalable **55**
4.1 System model . . . 57
4.1.1 Goal . . . 57
4.1.2 Data denormalization constraints . . . 58
4.1.3 Scaling individual data services . . . 59
4.2 Data Denormalization . . . 59
4.2.1 Denormalization and transactions . . . 59
4.2.2 Denormalization and read queries . . . 60
4.2.3 Case studies . . . 62
4.3 Scaling Individual Data Services . . . 64
4.3.1 Scaling the financial service of TPC-W . . . 65
4.3.2 Scaling RUBiS . . . 67
4.4 Evaluation . . . 67
4.4.1 Experimental setup . . . 68
4.4.2 Costs and benefits of denormalization . . . 68
4.4.3 Scalability of individual data services . . . 70
4.4.4 Scalability of the entire TPC-W . . . 71
4.5 Discussion . . . 73
4.6 Conclusion . . . 74

5 Resource provisioning for multi-service Web applications **75**
5.1 System design . . . 77
5.1.1 System model . . . 77
5.1.2 Performance model of a single service . . . 78
5.1.3 Resource provisioning of service instances . . . 81
5.1.4 Resource provisioning of cache instances . . . 84
5.1.5 Shifting resources among services . . . 86
5.2 Evaluation . . . 87
5.2.1 Experiment setup . . . 87
5.2.2 Model validation for single service . . . 89
5.2.3 Comparison with the state of the art . . . 90
5.2.4 Provisioning of multi-service applications . . . 92
5.3 Conclusion . . . 95

6 Resource provisioning in Cloud environments 99
6.1 System design . . . 100
6.1.1 Solution outline . . . 100
6.1.2 Web application hosting . . . 101
6.1.3 Online profiling . . . 102
6.1.4 Performance prediction . . . 107
6.1.5 Resource provisioning . . . 109
6.2 Evaluation . . . 110
6.2.1 Experiment setup . . . 110
6.2.2 Importance of adaptive load balancing . . . 110
6.2.3 Effectiveness of Performance Prediction and Resource Provisioning . . . 113
6.2.4 Comparison with other provision techniques . . . 115
6.3 Conclusion . . . 117

7 Conclusion 119
7.1 Research contributions . . . 119
7.2 Lessons learned . . . 121
7.3 Future directions . . . 122

Samenvatting 125

Bibliography 131