

Contents

Preface to third edition	xiii
Abbreviations	xv
CHAPTER 1 Introduction	1
1.1 Cloud computing, an old idea whose time has come	3
1.2 Energy use and ecological impact of cloud computing	7
1.3 Ethical issues in cloud computing	8
1.4 Factors affecting cloud service availability	9
1.5 Network-centric computing and network-centric content	10
CHAPTER 2 The cloud ecosystem	13
2.1 Cloud computing delivery models and services	14
2.2 Amazon Web Services	18
2.3 Google Clouds	25
2.4 Microsoft Windows Azure and online services	28
2.5 IBM clouds	30
2.6 Cloud storage diversity and vendor lock-in	30
2.7 Cloud interoperability	31
2.8 Service-level Agreements and Compliance-level Agreements	33
2.9 Responsibility sharing between user and service provider	34
2.10 User challenges and experience	34
2.11 Software licensing	36
2.12 Challenges faced by cloud computing	37
2.13 Cloud computing as a disruptive technology	39
2.14 Exercises and problems	40
CHAPTER 3 Parallel processing and distributed computing	41
3.1 Computer architecture concepts	42
3.2 Grand architectural complications	48
3.3 ARM architecture	55
3.4 SIMD architectures	58
3.5 Graphics processing units	60
3.6 Tensor processing units	62
3.7 Systems on a chip	65
3.8 Data, thread-level, and task-level parallelism	66
3.9 Speedup, Amdahl's law, and scaled speedup	68
3.10 Multicore processor speedup	70
3.11 From supercomputers to distributed systems	72
3.12 Modularity. Soft modularity versus enforced modularity	74
3.13 Layering and hierarchy	79
3.14 Peer-to-peer systems	82

3.15	Large-scale systems	84
3.16	Composability bounds and scalability (R)	86
3.17	Distributed computing fallacies and the CAP theorem	88
3.18	Blockchain technology and applications	89
3.19	History notes and further readings	91
3.20	Exercises and problems	94
CHAPTER 4	Cloud hardware and software	95
4.1	Cloud infrastructure challenges	96
4.2	Cloud hardware; warehouse-scale computer (WSC)	98
4.3	WSC performance	100
4.4	Hypervisors	104
4.5	Execution of coarse-grained data-parallel applications	104
4.6	Fine-grained cluster resource sharing in Mesos	106
4.7	Cluster management with Borg	107
4.8	Evolution of a cluster management system	110
4.9	Shared state cluster management	111
4.10	QoS-aware cluster management	113
4.11	Resource isolation	116
4.12	In-memory cluster computing for Big Data	120
4.13	Containers; Docker containers	128
4.14	Kubernetes	130
4.15	Further readings	131
4.16	Exercises and problems	132
CHAPTER 5	Cloud resource virtualization	135
5.1	Resource virtualization	136
5.2	Performance and security isolation in computer clouds	137
5.3	Virtual machines	138
5.4	Full virtualization and paravirtualization	140
5.5	Hardware support for virtualization	141
5.6	QEMU	144
5.7	Kernel-based Virtual Machine	145
5.8	Xen—a hypervisor based on paravirtualization	148
5.9	Optimization of network virtualization in Xen 2.0	154
5.10	Nested virtualization	156
5.11	A trusted kernel-based virtual machine for ARMv8	159
5.12	Paravirtualization of Itanium architecture	161
5.13	A performance comparison of virtual machines	164
5.14	Open-source software platforms for private clouds	167
5.15	The darker side of virtualization	169
5.16	Virtualization software	170
5.17	History notes and further readings	171
5.18	Exercises and problems	172

CHAPTER 6	Cloud access and cloud interconnection networks	175
6.1	Packet-switched networks and the Internet	176
6.2	Internet evolution	181
6.3	TCP congestion control	183
6.4	Content-centric networks; named data networks (R)	185
6.5	Software-defined networks; SD-WAN	187
6.6	Interconnection networks for computer clouds	189
6.7	Multistage interconnection networks	193
6.8	InfiniBand and Myrinet	194
6.9	Storage area networks and the Fibre Channel	197
6.10	Scalable data center communication architectures	200
6.11	Network resource management algorithms (R)	204
6.12	Content delivery networks	207
6.13	Vehicular ad hoc networks	211
6.14	Further readings	212
6.15	Exercises and problems	212
CHAPTER 7	Cloud data storage	215
7.1	Dynamic random access memories and hard disk drives	216
7.2	Solid-state disks	217
7.3	Storage models, file systems, and databases	220
7.4	Distributed file systems; the precursors	223
7.5	General parallel file system	228
7.6	Google file system	231
7.7	Locks; Chubby—a locking service	233
7.8	RDBMS—cloud mismatch	238
7.9	NoSQL databases	239
7.10	Data storage for online transaction processing systems	241
7.11	BigTable	243
7.12	Megastore	245
7.13	Storage reliability at scale	246
7.14	Disk locality versus data locality in computer clouds	250
7.15	Database provenance	252
7.16	History notes and further readings	254
7.17	Exercises and problems	255
CHAPTER 8	Cloud security	257
8.1	Security—the top concern for cloud users	258
8.2	Cloud security risks	259
8.3	Security as a service (SecaaS)	264
8.4	Privacy and privacy impact assessment	264
8.5	Trust	267
8.6	Cloud data encryption	268
8.7	Security of database services	270
8.8	Operating system security	272

8.9	Virtual machine security	273
8.10	Security of virtualization	275
8.11	Security risks posed by shared images	278
8.12	Security risks posed by a management OS	281
8.13	Xoar—breaking the monolithic design of the TCB	283
8.14	Mobile devices and cloud security	286
8.15	Mitigating cloud vulnerabilities in the age of ransomware	287
8.16	AWS security	289
8.17	Further readings	290
8.18	Exercises and problems	291
CHAPTER 9	Cloud resource management and scheduling	293
9.1	Policies and mechanisms for resource management	294
9.2	Scheduling algorithms for computer clouds	296
9.3	Delay scheduling (R)	298
9.4	Data-aware scheduling (R)	303
9.5	Apache capacity scheduler	306
9.6	Start-time fair queuing (R)	307
9.7	Borrowed virtual time (R)	311
9.8	Cloud scheduling subject to deadlines (R)	315
9.9	MapReduce application scheduling subject to deadlines (R)	320
9.10	Resource bundling; combinatorial auctions for cloud resources	322
9.11	Cloud resource utilization and energy efficiency	325
9.12	Resource management and dynamic application scaling	328
9.13	Control theory and optimal resource management (R)	329
9.14	Stability of two-level resource allocation strategy (R)	333
9.15	Feedback control based on dynamic thresholds (R)	334
9.16	Coordination of autonomic performance managers (R)	336
9.17	A utility model for cloud-based web services (R)	338
9.18	Cloud self-organization	342
9.19	Cloud interoperability	344
9.20	Further readings	346
9.21	Exercises and problems	346
CHAPTER 10	Concurrency and cloud computing	349
10.1	Enduring challenges	350
10.2	Communication and concurrency	353
10.3	Computational models; communicating sequential processes	358
10.4	The bulk synchronous parallel model	360
10.5	A model for multicore computing	361
10.6	Modeling concurrency with Petri nets	363
10.7	Process state; global state of a process or thread group	369
10.8	Communication protocols and process coordination	374
10.9	Communication, logical clocks, and message delivery rules	376
10.10	Runs and cuts; causal history	381

10.11	Threads and activity coordination	385
10.12	Critical sections, locks, deadlocks, and atomic actions	392
10.13	Consensus protocols	397
10.14	Load balancing	399
10.15	Multithreading in Java; FlumeJava; Apache Crunch	405
10.16	History notes and further readings	407
10.17	Exercises and problems	408
CHAPTER 11	Cloud applications	411
11.1	Cloud application development and architectural styles	412
11.2	Coordination of multiple activities	415
11.3	Workflow patterns	419
11.4	Coordination based on a state machine model—zookeeper	422
11.5	MapReduce programming model	425
11.6	Case study: the GrepTheWeb application	428
11.7	Hadoop, Yarn, and Tez	431
11.8	SQL on Hadoop: Pig, Hive, and Impala	435
11.9	Current cloud applications and new applications opportunities	440
11.10	Clouds for science and engineering	442
11.11	Cloud computing and biology research	446
11.12	Social computing, digital content, and cloud computing	448
11.13	Software fault isolation	450
11.14	Further readings	451
11.15	Exercises and problems	452
CHAPTER 12	Big Data, data streaming, and the mobile cloud	453
12.1	Big Data	454
12.2	Data warehouses and Google databases for Big Data	456
12.3	Dynamic data-driven applications	463
12.4	Data streaming	466
12.5	A dataflow model for data streaming	470
12.6	Joining multiple data streams	473
12.7	Mobile computing and applications	475
12.8	Energy efficiency of mobile computing	478
12.9	Alternative mobile cloud computing models	479
12.10	System availability at scale (R)	482
12.11	Scale and latency (R)	484
12.12	Edge computing and Markov decision processes (R)	488
12.13	Bootstrapping techniques for data analytics (R)	492
12.14	Approximate query processing (R)	495
12.15	Further readings	498
12.16	Exercises and problems	499
CHAPTER 13	Emerging clouds	501
13.1	A short-term forecast	502
13.2	Machine learning on clouds	503

13.3	Quantum computing on clouds	508
13.4	Vehicular clouds	518
13.5	Final thoughts	527
APPENDIX A	Cloud projects	529
A.1	Cloud simulation of a distributed trust algorithm	529
A.2	A trust management service	534
A.3	Simulation of traffic management in a smart city	540
A.4	A cloud service for adaptive data streaming	545
A.5	Optimal FPGA synthesis	550
A.6	Tensor network contraction on AWS	552
A.7	A simulation study of machine-learning scalability	559
A.8	Cloud-based task alert application	561
A.9	Cloud-based health-monitoring application	565
APPENDIX B	Cloud application development	571
B.1	AWS EC2 instances	572
B.2	Connecting clients to cloud instances through firewalls	575
B.3	Security rules for application- and transport-layer protocols in EC2	577
B.4	How to launch an EC2 Linux instance and connect to it	581
B.5	How to use S3 in Java	582
B.6	How to manage AWS SQS services in C#	585
B.7	How to install SNS on Ubuntu 10.04	586
B.8	How to create an EC2 placement group and use MPI	588
B.9	StarCluster—a cluster computing toolkit for EC2	590
B.10	An alternative setting of an MPI virtual cluster	590
B.11	How to install hadoop on eclipse on a windows system	592
B.12	Exercises and problems	595
Literature	597	
Glossary	621	
Index	635	