
Contents

SECTION I: IN THE BEGINNING

1 History 3

- Computing 3
- Early Mechanical Devices 3
- Computing Machines 11
- The 1960s 17
- The 1970s 22
- The 1980s 26
- The 1990s 30
- The 21st Century 33

2 Definition and Components 35

- P2P 37
 - Napster 38
 - Gnutella 38
- Types 40
 - Computational Grid 40
 - Distributed Servers and Computation Sites 41
 - Remote Instrumentation 41
 - Data Archives 42
 - Networks 43
 - Portal (User Interface) 43
 - Security 44
 - Broker 45
 - User Profile 45
 - Searching for Resources 46
 - Batch Job Submittal 46
 - Credential Repository 48
 - Scheduler 48
 - Data Management 49
 - Data Grid 50

Storage Mechanism Neutrality	51
Policy Neutrality	51
Compatibility with Other Grid Infrastructure	51
Storage Systems	51
Access or Collaboration Grid	52
Large-Format Displays	52
Presentation Environments	53
Interfaces to Grid Middleware	53
Others	54
Scavenging Grid	54
Grid Scope	56
Project Grid, Departmental Grid, or Cluster Grid	56
Enterprise Grid or Campus Grid	58
Global Grid	58

3 Early Adopters 59

Computational and Experimental Scientists	59
Bioinformatics	60
Corporations	60
Academia	60
University of Houston	61
University of Ulm Germany	61
The White Rose University Consortium	62
Science	62
Particle Physics	62
Industries	63
Gaming	63
Financial	65
Wachovia	66
RBC Insurance	66
Charles Schwab	66
Life Science	67
The American Diabetes Association	67
North Carolina Genomics and Bioinformatics Consortium	69
Spain's Institute of Cancer Research	69
Petroleum	69
Royal Dutch Shell	69
Utilities	70
Kansai Electric Power Co., Inc.	70
Manufacturing	70
Ford Motor Company	70
Saab Automobile	71
Motorola	71
Government	71
NASA	72
U.S. Department of Defense	72
European Union	73

Flemish Government	74
Benefits	75
Virtualization	75

SECTION II: THE PARTS AND PIECES

4 Security	83
Security	83
Authentication	84
Reciprocity of Identification	85
Computational Efficiency	85
Communication Efficiency	86
Third-Party Real-Time Involvement	86
Nature of Security	86
Secret Storage	87
Passwords	87
Private Key	88
Block Ciphers	89
Stream Ciphers	89
Public Key	91
Digital Signature	96
Authorization	101
Delegation of Identity	102
Delegation of Authority	103
Accounting	103
Audit	103
Access Control	104
DAC	104
MAC	105
Allow and Deny	106
Satisfy	107
Role-Based Access	107
Usage Control	108
Cryptography	108
Block Cipher	109
Stream Ciphers	110
Linear Feedback Shift Register	110
One-Time Pad	111
Shift Register Cascades	111
Shrinking Generators	112
Accountability	112
Data Integrity	115
Attenuation	116
Impulse Noise	116
Cross Talk	116
Jitter	117
Delay Distortion	117

Capability Resource Management	118
Database Security	121
Inference	121
Server Security	124
Database Connections	125
Table Access Control	125
Restricting Database Access	130
DBMS Specific	131
5 The Hardware	133
Computers	133
Blade Servers	138
Storage	140
I/O Subsystems	143
Underlying Network	143
Operating Systems	144
Visualization Environments	144
People	145
6 Metadata	147
Grid Metadata	152
Data Metadata	153
Physical Metadata	154
Domain-Independent Metadata	154
Content-Dependent Metadata	154
Content-Independent Metadata	155
Domain-Specific Metadata	155
Ontology	155
User Metadata	155
Application Metadata	156
External Metadata	156
Logical Metadata	157
User	157
Data	158
Resources	158
Metadata Services	158
Context	158
Structure	158
Define the Data Granularity	159
Database	159
Access	159
Metadata Formatting	160
XML	161
What Is XML?	161
Application	168
MCAT	169
Conclusion	170

7 Drivers	171
Business 174	
Accelerated Time to Results 174	
Operational Flexibility 174	
Leverage Existing Capital Investments 175	
Better Resource Utilization 176	
Enhanced Productivity 176	
Better Collaboration 178	
Scalability 178	
ROI 179	
Reallocation of Resources 180	
TCO 181	
Technology 183	
Infrastructure Optimization 183	
Increase Access to Data and Collaboration 183	
Resilient, Highly Available Infrastructure 183	
Make Most Efficient Use of Resources 184	
Services Oriented 185	
Batch Oriented 186	
Object Oriented 186	
Supply and Demand 186	
Open Standards 187	
Corporate IT Spending Budgets 187	
Cost, Complexity, and Opportunity 188	
Better, Stronger, Faster 190	
Efficiency Initiatives 191	

SECTION III: DATABASES IN THE GRID

8 Intr oducing Databases	195
Databases 195	
Relational Database 196	
Tuples 197	
Attributes 198	
Entities 198	
Relationship 198	
Relational Algebra 198	
Union 198	
Intersection 198	
Difference 199	
Cartesian Product 199	
Select 199	
Project 200	
Join 200	
Relational Calculus 200	
Object Database 202	
Architecture Differences between Relational and Object Databases 203	

Object Relational Database 203

SQL 205

 Select 206

 Where 206

 And/Or 206

 In 207

 Between 207

 Like 207

 Insert 207

 Update 208

 Delete 208

Database 209

Data Model 209

Schema 209

Relational Model 209

Anomalies 209

 Insert Anomaly 210

 Deletion Anomaly 210

 Update Anomaly 210

9 Parallel Database 213

Data Independence 213

Parallel Databases 214

 Start-Up 216

 Interference 216

 Skew 217

 Attribute Data Skew 217

 Tuple Placement Skew 217

 Selectivity Skew 217

 Redistribution Skew 217

 Join Product Skew 218

Multiprocessor Architecture Alternatives 218

 Shared Everything 218

 Shared Disk 219

 Shared Nothing (Message Passing) 220

 Hybrid Architecture 221

 Hierarchical Cluster 221

 NUMA 222

Disadvantages of Parallelism 222

Database Parallelization Techniques 224

 Data Placement 224

 Parallel Data Processing 224

 Parallel Query Optimization 224

 Transaction Management 224

 Parallelism Versus Fragmentation 224

 Round-Robin 225

 Hash Partitioning 225

- Range Partitioning 225
- Horizontal Data Partitioning 226
- Replicated Data Partitioning 226
- Chained Partitioning 227
- Placement Directory 227
- Index Partitioning 228
- Partitioning Data 228
- Data-Based Parallelism 228
 - Interoperation 228
 - Intraoperation 229
 - Pipeline Parallelism 229
 - Partitioned Parallelism 230
- Parallel Data Flow Approach 231
 - Retrieval 232
 - Point Query 232
 - Range Query 232
 - Inverse Range Query 232
- Parallelizing Relational Operators 233
 - Operator Replication 233
 - Merge Operators 233
 - Parallel Sorting 233
 - Parallel Aggregation 234
 - Parallel Joins 234
- Data Skew 237
 - Load Balancing Algorithm 237
 - Dynamic Load Balancing 238

10 Distributing Databases 241

- Advantages 245
- Disadvantages 245
- Rules for Distributed Databases 246
- Fragmentation 248
 - Completeness 249
 - Reconstruction 249
 - Disjointedness 249
- Transparency 249
 - Distribution Transparency 250
 - Fragmentation Transparency 250
 - Location Transparency 250
 - Replication Transparency 250
 - Local Mapping Transparency 251
 - Naming Transparency 251
 - Transaction Transparency 251
 - Performance Transparency 252
- Vertical Fragmentation 252
- Horizontal Fragmentation 254
- Hybrid 255

Replication	255
Metadata	256
Distributed Database Failures	257
Failure of a Site	257
Loss of Messages	257
Failure of a Communication Link	257
Network Partition	257
Data Access	258
11 Data Synchronization	261
Concurrency Control	262
Distributed Deadlock	262
Database Deadlocks	264
Multiple-Copy Consistency	265
Pessimistic Concurrency Control	266
Two-Phase Commit Protocol	267
Time Stamp Ordering	267
Optimistic Concurrency Control	268
Heterogeneous Concurrency Control	270
Distributed Serializability	271
Query Processing	271
Query Transformations	271
Transaction Processing	271
Heterogeneity	272
12 Conclusion	275
Index	277