

Contents

Part 1 InfiniBand Objectives	1
Chapter 1 Introduction	3
Scope.....	5
A Brief History.....	6
Chapter 2 New Server I/O Technology	9
Comparing InfiniBand to Existing Technology.....	11
Performance	15
Cluster Capability	16
Reliability and Availability	17
Efficiency	18
CPU Efficiency	18
System Efficiency.....	21
Scalability.....	23
Protection.....	23
Manageability.....	25
Summary	25
Part 2 InfiniBand Architecture	27
Chapter 3 A Switched Fabric Architecture	29
Topology.....	29
Fabric Infrastructure	30
Fabric Characteristics	34
Segmentation and Re-assembly.....	35
Virtual Lanes	36
Targeted Applications.....	38
Chapter 4 Architectural Structure	41
Queue Pairs.....	41
InfiniBand Services.....	42
Accessing InfiniBand Services - The Channel Interface	44
Transport Service Types	47
Reliable Connection and Unreliable Connection Service	48
Reliable Datagram Service	49
Unreliable Datagram Service	51
Raw Datagram Service	52

InfiniBand Layers	54
Transport Layer	55
Network Layer	55
Link Layer.....	56
Physical Layer	56
QP Operations	57
Memory Registration.....	57
Receive Queue Operations.....	58
Send Queue Operations	58
Chapter 5 Management Infrastructure	63
Management Classes	63
Management Elements.....	65
Management Messages and Methods	66
Subnet Manager	67
Subnet Administration	69
General Services	69
Communication Management.....	71
Service IDs.....	72
Channel Establishment	73
Service ID Resolution and Q_Keys.....	74
Chapter 6 Key InfiniBand Architecture Concepts	75
Specification Scope and the Full Picture	75
Nodes and End Nodes	77
Fabric, Subnet, and Partitions	78
Fabric	79
Subnet.....	79
Partitions	80
Addressing.....	83
Addressing Memory	83
Communication Addressing	83
Local Addressing.....	84
Global Addressing	86
Address Vectors	87
Channel Addressing	89
Protection Domains	90
I/O Infrastructure.....	91
Legacy Networking	92
Part 3 InfiniBand Deployment	97
Chapter 7 Deployment Strategies	99
Applying the Technology	100
Deployment Models	103
Dedicated Subnet.....	103

Shared I/O Interconnect.....	107
System Interconnect	112
Timeline of Product Capabilities.....	115
HCA Capabilities.....	115
Management Considerations:.....	117
Chapter 8 A General Service Model.....	119
Channel Interface and Verbs.....	120
Service Model Considerations.....	121
The Operating System Service Model.....	123
InfiniBand Clients	124
Client Service Interface	125
Kernel Agents.....	126
The Service Interface	129
InfiniBand Channel Access Layer.....	129
Service Interface Functions	131
User-Level Functions.....	132
Kernel-Level Functions.....	140
Chapter 9 Operating System Components.....	147
Core Components	147
InfiniBand Resource Manager	147
Installation and Configuration of an HCA	147
Initialization of HCAs	148
Allocation of Resources.....	149
Memory Registration	152
Completion Handler	153
Event Handler	154
Communication Manager	156
CM Responsibilities.....	157
Usage Models	158
Client-Server Model.....	159
IPC Model.....	160
Using a Name Server	161
CM Operation.....	163
Path Selection	166
CM Protocol.....	167
QP State Transitions	169
Queue Pair Parameters.....	173
Automatic Path Migration	178
Other Considerations	183
SMI/GSI Access Agent.....	185
SMI Access	185
GSI Access	187
Supporting GSAs	190

GSI Redirection.....	192
Resource Consolidation	192
Partition Handling Efficiency.....	192
Dedicated QPs	193
Supporting GSMs.....	193
SA Access.....	194
Ancillary Agents.....	195
I/O Resource Manager.....	195
Discovering and Tracking I/O Resources.....	195
Matching I/O Drivers.....	197
Abstracting Driver Access to I/O Information	200
Coordinating Shared Resources	202
Trivial Subnet Manager.....	202
Legacy Communication Driver.....	203
Overview of Legacy Communication.....	203
Basic Requirements	205
Raw Operation.....	206
Drawbacks.....	207
Part 4 InfiniBand-Based Applications	209
Chapter 10 Designing an Application.....	211
Selecting the Application Model	211
Selecting the Transport Service Type	212
Reliable Connection.....	214
RC Channel Characteristics	214
Application	215
Unreliable Connection.....	216
UC Channel Characteristics	216
Application	216
Unreliable Datagram	217
UD Channel Characteristics	217
Application	218
Reliable Datagram	219
RD Channel Characteristics	220
Application	222
Transport Selection Criteria	225
Completion Handling Policies.....	225
Completion Queues	225
Completion Events.....	227
Error Handling Policies.....	230
Error Classification.....	230
Error Handling.....	232
Immediate Errors.....	232
Completion Errors.....	236
Asynchronous Errors.....	242

The Process Flow	243
Initialization	244
Completion Processing	247
Error Handling	251
Programming Errors	252
Operational Errors	252
The Service Provider	253
Reliable Connection Service Provider	254
Unreliable Connection Service Provider	255
Unreliable Datagram Service Provider	255
Reliable Datagram Service Provider	258
The Service Client	259
Reliable Connection Service Client	259
Unreliable Connection Service Client	260
Unreliable Datagram Service Client	260
Reliable Datagram Service Client	261
Chapter 11 Working with the Technology	263
Channel Operation	264
Basic Principles	265
Initialization	265
Queue Pairs	266
Protection Domains	267
Completion Queues	268
Memory Registration	268
Memory Windows	269
Creating Channels	270
Summary of InfiniBand Objects	271
Using the Technology	274
Service IDs and Establishing Channels	274
I/O Operation	275
Networked Service	276
Subnet Services	276
IPC	277
Using CQs	277
Considerations for Application Developers	278
Limited number of QPs	278
Ordering and the Fence Indicator	278
Path Considerations	280
QP's Affinity to a Port	281
Chapter 12 Design Example	283
Developing a Storage Driver	283
Process Flow	283
Initialization	284
Assignment	284

Input Transactions.....	285
Output Transactions.....	286
Adding I/O Controllers.....	287
Variations.....	288
Part 4 InfiniBand-Based Applications	289
Chapter 13 Subnet Management	291
Subnet Manager Roles.....	291
General Subnet Management Duties	292
Initial SM Operation.....	294
Discovery	294
Configuring the Subnet	295
Establishing Paths.....	295
Sweeping the Subnet.....	296
SM Hand-off.....	296
SM Disable.....	297
Direct Attach.....	297
Subnet Maintenance.....	297
Segmenting and Joining Subnets	297
Stitching Subnets Together.....	298
Trivial Subnet Management.....	298
Chapter 14 InfiniBand Management Applications.....	301
Understanding InfiniBand Management Architecture.....	301
Management Classes	302
Managers	304
Management Agents	304
Class Managers	305
Privileged Operations.....	306
Role of the Subnet Manager.....	307
Other Management Classes.....	307
I/O Management.....	308
IOUnitInfo	309
IOControllerProfile.....	310
ServiceEntries	311
Using Device Management Information.....	311
Locating the I/O Controller.....	312
Matching the IOC to an I/O Driver	312
Appendix A Glossary of Acronyms and Technical Terms	315
Acronyms.....	315
Definitions.....	317

Index 329

Figures

1.1 Roots of the InfiniBand Architecture 7

2.1 CPU Speed versus I/O Speeds 9

2.2 Shared Bus Topology 12

2.3 Switched Fabric Topology 13

2.4 Shared Bus Architecture 13

2.5 InfiniBand Switched Architecture 14

2.6 Channel Isolation 18

2.7 CPU Speed Barriers 19

2.8 CPU Access Costs 20

3.1 InfiniBand Switched Topology 30

3.2 Fabric Infrastructure 31

3.3 Subnet Infrastructure 31

3.4 InfiniBand Subnet Example 33

3.5 Segmentation and Re-assembly 35

3.6 Virtual Lane Concept 37

3.7 I/O Concepts 40

4.1 Work / Completion Queue Architecture 45

4.2 Channel Interface 46

4.3 Connected Service 48

4.4 Reliable Datagram Service 50

4.5 Unreliable Datagram Service 51

4.6 Raw EtherType Service 52

4.7 Raw IPv6 Service 53

4.8 InfiniBand Communication Model 54

4.9 Protected Memory Access 61

5.1 Management Elements 65

5.2 Subnet Management Models 68

5.3 General Services Physical Model 70

5.4 General Services Logical Model 71

6.1 Typical Host Environment 76

6.2 Examples of Partition Structure 82

6.3 I/O Unit Model 91

6.4 I/O Communication Stack 92

6.5 Legacy Networking using a LAN NIC 93

6.6 Legacy Networking using InfiniBand Fabric as a LAN 93

6.7 LAN Connectivity 94

7.1 How Will You Deploy InfiniBand? 99

7.2 InfiniBand Example 100

7.3 Web Service Example 102

7.4 InfiniBand as a Private I/O Interconnect 104

7.5 Private I/O Topology 105

7.6 System with Private Subnets 106

7.7	Shared I/O Example	107
7.8	Shared I/O Topology.....	108
7.9	I/O Partitions	109
7.10	Sharing using Multiple I/O Partitions	110
7.11	Limited Membership Sharing.....	111
7.12	Data Center Example	112
7.13	Data Center Alternative	113
8.1	Channel Access Layer.....	119
8.2	Operating Environment.....	123
8.3	Channel Access Hierarchy	130
8.4	Kernel Agent Hierarchy.....	131
9.1	Communication Management Interactions	171
9.2	Alternate Paths	178
9.3	Alternate Ports	179
9.4	Automatic Path Migration States	180
9.5	Automatic Path Migration Activation and Rearm.....	182
9.6	SMI/QP0 Relationships.....	186
9.7	GSI Structure (per port)	188
9.8	Dedicated GSM QPs	190
10.1	Connection-oriented (RC and UC) Service	214
10.2	Unreliable Datagram Service.....	217
10.3	Reliable Datagram Service.....	220
10.4	Reliable Datagram Relationships	223
10.5	Completion Queue.....	226
10.6	Overall Process Flow.....	243
10.7	Generic Completion Processing	249
10.8	Dedicated Completion Queues.....	250
10.9	RC Service QP Relationships.....	255
10.10	UD Service Port/Partition Relationship.....	256
10.11	RD Service EEC/QP Relationship	259
11.1	Application Communication Model	267
11.2	Protection Domain	268
11.3	Hierarchy of InfiniBand Objects	273
11.4	InfiniBand Communication Model.....	280
13.1	Subnet Manager Behavior	293
14.1	Management Elements.....	302
14.2	Subnet Manager Role as Class Manager.....	305

Tables

2.1	Summary of Differences and Benefits	11
4.1	Service Types	47
8.1	Access Library Functions.....	132
8.2	Kernel-Level Access Library Functions	140
8.3	Access Library Parameters.....	141
9.1	Sources for QP Parameter Values	174

9.2	Sources for Address Vector Parameters	175
9.3	Sources for CM:REQ MAD Parameters	176
9.4	Sources for CM:REP MAD Parameters.....	177
9.5	IORM Functions.....	201
9.6	IORM Driver Function Parameters	201
10.1	Transport Service Type Characteristics	225
10.2	Completion Indications	229
11.1	Hierarchy of InfiniBand Objects	272
14.1	Components and Resulting String Formats.....	313
14.2	Compatible Strings	314