

Contents

Preface ix

Chapter 1 Introduction 1

- What's the Problem? 1
- How VIA Solves the Problem 8
 - Functional Partitioning 10
 - Separation of Control and Data 10
- Fundamental Concepts of VIA 11
- Performance Case Study 12

Chapter 2 Historical Background and Architectural Roots 17

- Two Parallel Architectures 17
 - Shared Memory and Multiprocessing 18
 - Distributed Memory and Message Passing 21
 - Combining SMP and Message Passing 24
- Prior Research 25
 - Cornell U-Net 25
 - Princeton SHRIMP Project 26
 - Scheduled Transfer (ST) 26
 - AM and FM 27
- Enter VIA: A New Industry Standard 27

Chapter 3 Principles of Operation 31

- The VI Stack 31
- Data Movement Operations 33
 - Send/Receive 33

RDMA-Write	34
RDMA-Read	35
Using RDMA	35
Initiating Data Movement Operations	36
Descriptors	36
Work Queues	37
The Receive Assumption	38
Doorbells	38
Completions	42
Polling for Work Queue Completion	42
Blocking On Work Queue Completion	43
Completion Queues	44
Resource Management	45
Opening the VI NIC	45
Registering Memory	46
Creating Completion Queues	47
Creating VIs	47
Connections	48
Reliability Levels	50
Formats of Descriptor Segments	52
Control Segment Fields	52
Descriptor Address Segment	57
Descriptor Data Segment	57

Chapter 4 The VIA Specification 59

The Partners	59
About RDMA-Read	61
The VIA Specification	62
The VIPL 'Example'	63
The VI Developer's Guide	64

Chapter 5 Memory Translation and Protection 67

Overview	67
TPT Format	68
Registering Memory	69
Memory Handles	69
Translating Addresses	70
Doorbell Tokens	71
Protection Tags	72
Page Sizes Greater Than 4K and 64-bit Addresses	73
TPT Limitations	74

Chapter 6 VIPL Made Simple 77

Application Overview	77
Client Program	78
Server Program	89

PuzzleCommon.h 103

Chapter 7 Industry Activity 107

Standards Activities 107
 VIDF 107
 FC-VI 107
 VI-IP 108
 DAT Collaborative 108
 DAFS 109
VI Hardware Implementations 109
 Compaq ServerNet II 109
 Emulex/GigaNet 109
 Troika Networks 110
 QLogic 110
VI Software Stacks 111
 M-VIA 111
 MVICH 111
 Microsoft Winsock Direct 112

Chapter 8 Applications 113

Scalable Databases 113
 IBM DB2 114
 Oracle Database 116
 Microsoft SQL Server 117
Storage 117
High Performance Computing 118
 Future VIA Applications 119

Chapter 9 InfiniBand Architecture 123

InfiniBand Overview 123
Differences Between VIA and IBA 124
 Scope 124
 Verbs 124
 Atomic Operations 125
 Memory Protection Model 125
 Descriptors and Work Queues 126
 Completion Model 126
 Reliability Levels 127
 Other 128

Chapter 10 Conclusions 129

Successes 129
Areas for Improvement 130
 Limited Number of VIs 130
Thread Pool Model 131
 Manager Thread Code 132

Worker Thread Code	132
Connection Orientation	134
Memory Registration	135
Receive Assumption, Lack of Flow Control	136
Insufficient Standardization	136
The Future	137

Appendix A VIPL Calls 139

Hardware Connection	139
VipOpenNic	139
VipCloseNic	140
Endpoint Creation and Destruction	141
VipCreateVi	141
VipDestroyVi	142
Connection Management	143
VipConnectWait	143
VipConnectAccept	145
VipConnectReject	146
VipConnectRequest	147
VipDisconnect	149
VipConnectPeerRequest	150
VipConnectPeerDone	152
VipConnectPeerWait	153

Memory Protection and Registration	155
VipCreatePtag	155
VipDestroyPtag	156
VipRegisterMem	157
VipDeregisterMem	158
Data Transfer and Completion Operations	159
VipPostSend	159
VipSendDone	160
VipSendWait	161
VipPostRecv	163
VipRecvDone	164
VipRecvWait	165
VipCQDone	166
VipCQWait	167
VipSendNotify	169
VipRecvNotify	171
VipCQNotify	173
Notify Semantics	175
Completion Queue Management	175
VipCreateCQ	175
VipDestroyCQ	176
VipResizeCQ	177
Querying Information	178
VipQueryNic	178
VipSetViAttributes	179
VipQueryVi	180
VipSetMemAttributes	181
VipQueryMem	182
VipQuerySystemManagementInfo	183
Error Handling	185
VipErrorCallback	185
Name Service	188
VipNSInit	188
VipNSGetHostByName	189
VipNSGetHostByAddr	191
VipNSShutdown	192

References 195

- Publications, Studies, and Projects 195
- Web Sites 196

Index 199