

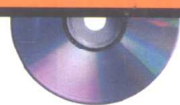


大学计算机教育国外著名教材、教参系列 (影印版)

PCI-X SYSTEM ARCHITECTURE

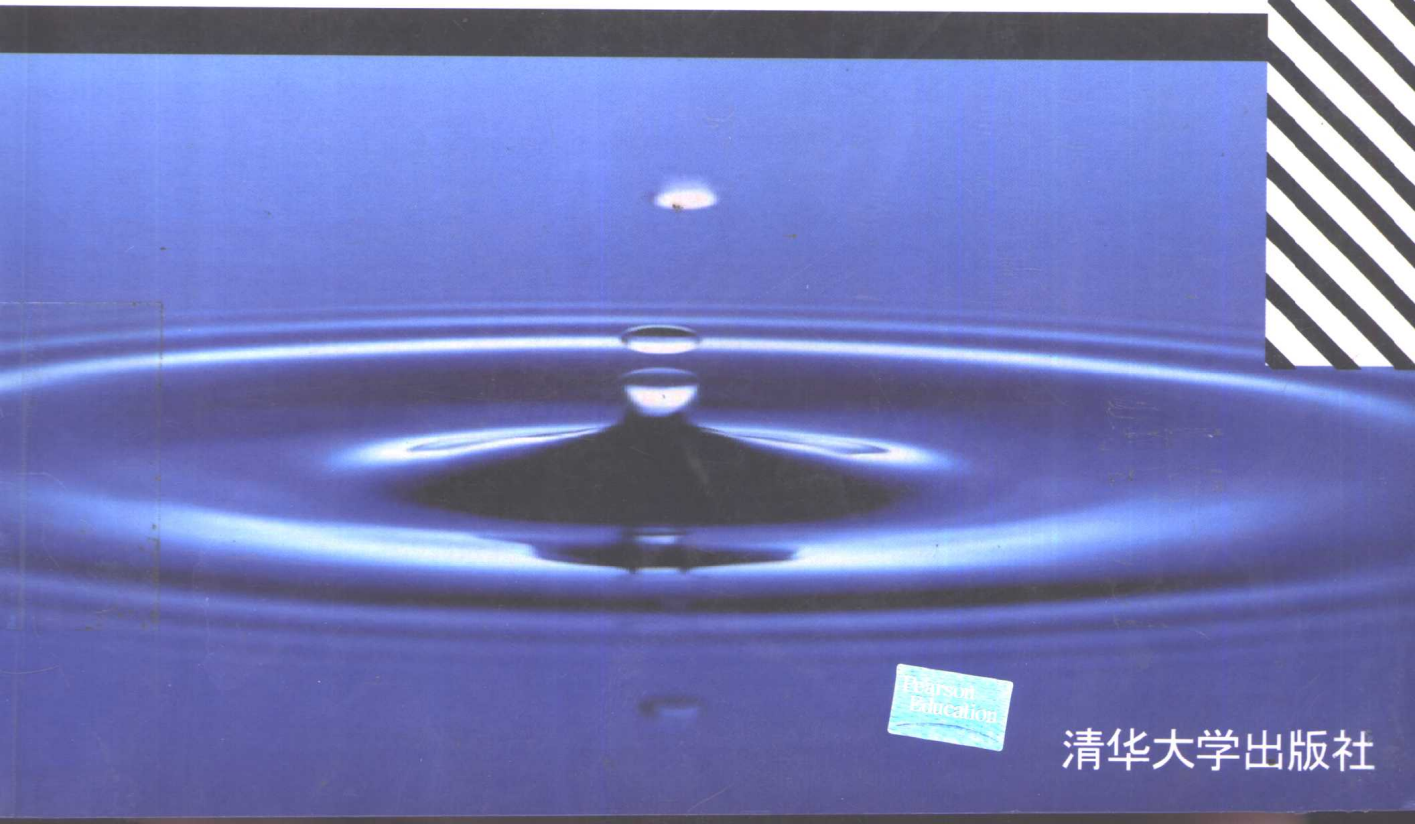
MINDSHARE, INC.

Tom Shanley



2
PCI-X

系统的体系结构



清华大学出版社

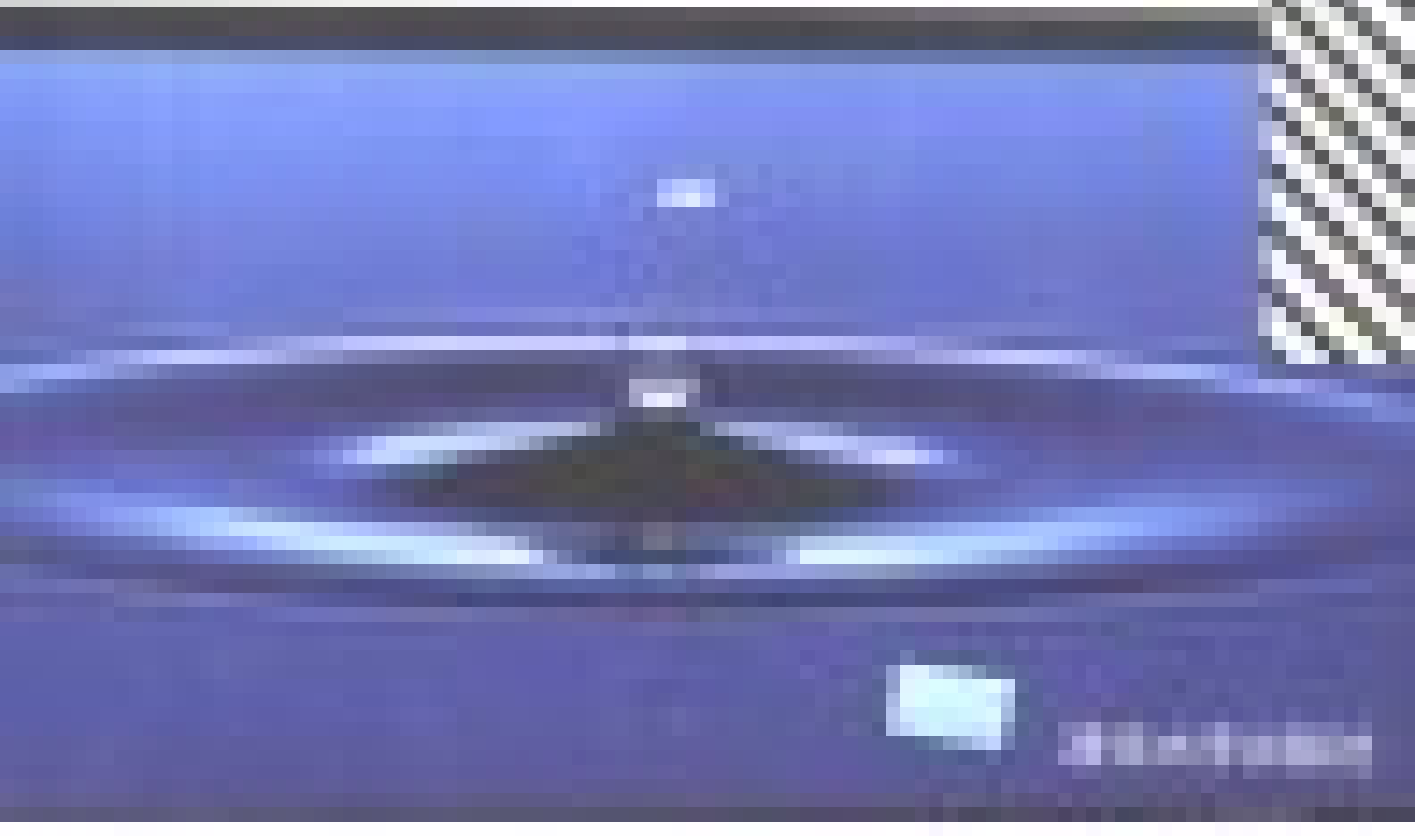
PCI-X SYSTEM ARCHITECTURE

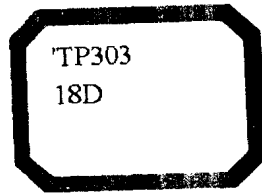
Microsoft, Inc.
Tom Chantry



PCI-X

系统的体系结构





PCI-X System Architecture

PCI-X 系统的体系结构

MindShare, Inc.

Tom Shanley

清华大学出版社

(京) 新登字 158 号

PCI-X System Architecture

Tom Shanley

Copyright © 2001 by MindShare, Inc.

Original English Language Edition Published by Addison-Wesley

All Rights Reserved.

For sale in Mainland China only.

本书影印版由培生教育出版集团授权清华大学出版社在中国境内（不包括香港特别行政区、澳门特别行政区和台湾地区）独家出版、发行。

未经出版者书面许可，不得以任何方式复制或抄袭本书的任何部分。

本书封面贴有培生教育出版集团激光防伪标签，无标签者不得销售。

北京市版权局著作权合同登记号：图字：01-2002-1663

书 名：PCI-X System Architecture

作 者：Tom Shanley

出版者：清华大学出版社（北京清华大学学研大厦，邮编 100084）

[http:// www.tup.tsinghua.edu.cn](http://www.tup.tsinghua.edu.cn)

印刷者：清华大学印刷厂

发行者：新华书店总店北京发行所

开 本：787×960 1/16 印张：46

版 次：2002 年 8 月第 1 版 2002 年 8 月第 1 次印刷

书 号：ISBN 7-900643-34-6

印 数：0001~3000

定 价：68.00 元（含光盘）

出版说明

进入 21 世纪, 世界各国的经济、科技以及综合国力的竞争将更加激烈。竞争的中心无疑是对人才的争夺。谁拥有大量高素质的人才, 谁就能在竞争中取得优势。高等教育, 作为培养高素质人才的事业, 必然受到高度重视。目前我国高等教育的教材更新较慢, 为了加快教材的更新频率, 教育部正在大力促进我国高校采用国外原版教材。

清华大学出版社从 1996 年开始, 与国外著名出版公司合作, 影印出版了“大学计算机教育丛书(影印版)”等一系列引进图书, 受到了国内读者的欢迎和支持。跨入 21 世纪, 我们本着为我国高等教育教材建设服务的初衷, 在已有的基础上, 进一步扩大选题内容, 改变图书开本尺寸, 一如既往地请有关专家挑选适用于我国高校本科及研究生计算机教育的国外经典教材或著名教材以及教学参考书, 组成本套“大学计算机教育国外著名教材、教参系列(影印版)”, 以飨读者。深切期盼读者及时将使用本系列教材、教参的效果和意见反馈给我们。更希望国内专家、教授积极向我们推荐国外计算机教育的优秀教材, 以利我们把“大学计算机教育国外著名教材、教参系列(影印版)”做得更好, 更适合高校师生的需要。

计算机引进版图书编辑室

2002.3

Acknowledgments

The author would like to extend special thanks to Allen Goodrum and Dwight Riley for their time and insights.

大学计算机教育国外著名教材、教参系列 (影印版) 图书目录

1. UNIX Network Programming Vol. 2: Interprocess Communications 2nd ed. 1999/*W. Richard Stevens* (UNIX 网络编程卷 2: 进程间通信 第 2 版 580 页)
2. The 80x86 IBM PC and Compatible Computers (Volumes I & II) Assembly Language, Design, and Interfacing 3rd ed. 2000/*Muhammad Ali Mazidi, Janice Gillispie Mazidi* (80x86 IBM PC 及兼容计算机卷 I 和卷 II: 汇编语言, 设计与接口技术 第 3 版 1020 页)
3. Principles of Distributed Database Systems 2nd ed. 1999/*M. Tamer Özsu, Patrick Valduriez* (分布式数据库系统原理 第 2 版 688 页)
4. Network Security Essentials: Applications and Standards 2000/ *William Stallings* (网络安全基础教程 382 页)
5. Cryptography and Network Security: Principles and Practice 2nd ed. 1999/*William Stallings* (密码学与网络安全: 原理与实践 第 2 版 590 页)
6. Data Structures and Algorithm Analysis in C++ 2nd ed. 1999/*Mark Allen Weiss* (数据结构与算法分析 C++描述 第 2 版 606 页)
7. PCI-X System Architecture 2001/*Tom Shanley* (PCI-X 系统的体系结构 734 页)
8. Introduction to Automata Theory, Languages, and Computation 2nd ed. 2001/*John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman* (自动机理论、语言和计算导论 第 2 版 535 页)
9. Operating Systems: A Systematic View 5th ed. 2001/*William S. Davis, T. M. Rajkumar*(操作系统实践与应用 第 5 版 636 页)
10. Introduction to Logic Design 2002/*Alan B. Marcovitz* (逻辑设计基础 584 页)

About This Book

The MindShare Architecture Series	1
Cautionary Note	2
This Book Assumes PCI Background Knowledge	3
Specifications This Book Is Based On	3
Organization of This Book	3
Who This Book Is For	4
Prerequisite Knowledge	4
Documentation Conventions	5
Hexadecimal Notation	5
Binary Notation	5
Decimal Notation	5
Bits Versus Bytes Notation	5
Bit Fields (Logical Groups of Bits or Signals)	6
Timing Diagram Drawing Convention	6
Clock-by-Clock Timing Diagram Description	8
Signal Polarity	8
Visit Our Web Site	8
We Want Your Feedback	9

Part 1: Basic Concepts

Chapter 1: PCI Needed Improvement

Wait States Yield Poor Performance	13
Relatively Slow Clock Speed	14
Transfer Size Unknown	14
PCI Delayed Transactions Are Inefficient	15
Initiator Retries Use Up Valuable Bus Time	15
Initiator Doesn't Supply Transfer Count	15
Delayed Completion	16
Initiator's Transaction ID Is Sketchy at Best	16
Snoops Hurt Performance	16
Host/PCIX Bridge Knows AGP's Area of Memory Is Non-Cacheable	16
Memory Used by PCI Masters May or May Not Be Cached	17
Snoops Slow Down PCI Accesses to Main Memory	17
Snoop Traffic on Processor Bus Can Hurt	
Processor(s)	17
Main Memory Less Available to Processor(s)	17
Latency Timer Use Not Optimized in PCI	19
Min_Gnt Register: Timeslice Request	19

Contents

Latency Timer: "Timeslice" Register	19
Insufficient Info to Select a Good Value	20
Data Phase Parity Error Recovery Usually Not Possible	21
PCI Data Parity Error Recovery	21
Important Note Regarding Chipsets That Monitor PERR#	22
No Indication of Device Width	22
MSI Feature Optional in PCI Environment	22
Introduction	22
Advantages of MSI Interrupts	23
Power Management Optional	23
Legacy PCI Devices—No Standard PM Method	23
Device Support for PCI PM Optional	23
Discovering Function's PM Capability	23
Configuration Software Constrained by 32-Bit Memory BARs	26
Stepping Yields Poor Performance	27

Chapter 2: PCI-X Improves on PCI

PCI-X Is Backward-Compatible With PCI	29
PCI-X Is More System-Centric Than PCI	30
Higher Clock Speeds Possible	31
Wait States Eliminated	31
Data Transferred in Blocks	31
In PCI, Data Transferred as a Series of Data Items	31
In PCI-X, Data Transferred as a Series of Blocks	32
Master Cannot Delay Transfer of First Data Block	32
Target Can Only Delay Transfer of First Block	32
Master and Target Can Only Disconnect on Block Boundaries	32
But...There Are Two Exceptions	32
Disconnecting on Block Boundaries	33
Master Disconnection of a Transfer	33
Target Disconnection of a Transfer	33
Latency Timer Usage	33
Requester and Completer	34
Transfer Size Specified	34
Requester ID and Transaction ID Are Specified	34
Split Transactions Replace Delayed Transactions	35
Target Cannot Transfer Data Within 16 Clocks From FRAME# Assertion	35
Long-Latency Access Handling in PCI	35
Long-Latency Access Handling in PCI-X	35
Split Transactions More Efficient Than Delayed	36
Example: Requester and Completer on Same Bus	37
Example: Requester and Completer on Different Buses	38

Dynamic Traffic Analysis and Load Tuning	40
Data Phase Parity Error Recovery	41
General.....	41
Recovery Usually Not Possible in PCI.....	41
PCI-X Chipset Does Not Monitor PERR#.....	42
What the PCI-X Spec Says	42
Requester Hardware Doesn't Handle Error Itself.....	42
With Appropriate Master and Device Driver Design, Recovery Possible	42
Device Driver With No Recovery Capability = SERR#	43
If Bus in PCI Mode, Bridge Monitors PERR#.....	43
64-Bit Connection Indication	44
MSI Feature Mandatory	44
Power Management Mandatory for Add-In Devices.....	44
Snoops Can Be Eliminated	44
Memory BARS Must Be 64-Bit Width.....	45
Bus Masters That Access Memory Must Support DAC Command	45
Stepping Eliminated.....	45
Fast Back-to-Back Transactions Eliminated.....	46

Chapter 3: Lowest Common Denominator Defines Mode

Bus Protocol/Speed = Lowest Common Denominator	48
Discovering a PCI-X Bus.....	48
Discovering PCI Devices on a PCI-X Bus.....	49
Some Example Systems.....	49
System With No Connectors on PCI-X Bus.....	49
Single Bus System With Connectors	49
Systems Supporting Both PCI-X and PCI Environments.....	52
Introduction.....	52
System With Single Host/PCIX Bridge	52
Dual Host/PCIX Bridge System	54

Chapter 4: Device Types and Bus Initialization

All Devices Support 33MHz PCI	56
General.....	56
133MHz PCI-X Device Must Support 33MHz and 66MHz PCI	56
When M66EN Is Grounded on a Card.....	56
Effect of Grounded M66EN on a PCI Source Bridge.....	56
Effect of Grounded M66EN on a PCI-X Source Bridge.....	56
M66EN Usage on a 66MHz PCI Card or a PCI-X Card.....	57
All PCI-X Devices Support 66MHz PCI-X Mode.....	57

Contents

PCIXCAP Indicates Protocol/Frequency Required.....	58
General.....	58
PCIXCAP Indicates Capabilities of Card's First Device.....	61
Device Not Permitted to Use PCIXCAP as Input or Output.....	61
Bridge's Interpretation of M66EN and PCIXCAP.....	61
Maximum Reliable Speed Verified by Design and Testing	63
Supplying PCI-X Devices With Protocol and Speed.....	64
General.....	64
Host/PCIX Bridge Pattern Delivery Sequence.....	64
PCIX-to-PCIX Bridge Pattern Delivery Sequence	65
The Sequence.....	65
But First.....	65
Why Bus Must Be Idle When Init Pattern Driven	67
Upon Receipt of Pattern, Device Initializes Itself	67
Init Pattern Setup and Hold Time	67
General.....	67
Relationship With Trhff	68
Reassertion of RST# Necessitates Redelivery of Pattern	68
General.....	68
Secondary Bus RST# Follows RST# on Primary Side	68
Secondary Bus RST# Under Software Control	69
Bridge Must Support Interfaces in Different Modes.....	70
Hot-Plug PCI-X Bus Initialization	71
Some Systems Permit Hot-Plug of PCI-X Cards	71
For Background on PCI Hot-Plug.....	71
Problems Associated With PCI-X Hot-Plug.....	71
Determination of Card Capabilities	72
General	72
Determining the Presence of a PCI Card.....	73
Unacceptable Method	73
Acceptable Method.....	73
Determining the Presence of a 66MHz PCI Card.....	74
Determining the Presence of a 66MHz-Capable PCI-X Card.....	74
Determining the Presence of a 133MHz-Capable PCI-X Card.....	75
Hot-Install of Card With PCIX-to-PCIX Bridge on It	75
Description	75
How to Avoid Bus Renumbering During Run-Time	76
Hot-Install of Card With PCI-to-PCI Bridge on It.....	77
Checking PCIXCAP Without Power Applied to Card.....	78
Changing Secondary Bus Number Configuration Register	78

Conservation of Address Space	78
General	78
Requesting Memory Space	79
Requesting IO Space	80
Early Configuration Access to Newly-Installed Device	80
Possible Adjustment of Max Memory Read Byte Count Values	80
Placing Device in Low-Power Mode Requires Quiesce	81

Chapter 5: PCI-X Is a Registered Bus

PCI-X Is a Low-Voltage Swing (LVS) Bus.....	83
Introduction to the Registered Nature of the Bus.....	84
Address/Command Decode Example.....	85
Data Read Example	86

Chapter 6: Intro to Commands

Commands Fall Into Three Categories	89
Command Encoding	90
Dword Commands	92
General.....	92
IO Read and Write Commands.....	92
Memory Read Dword Command	92
Configuration Read and Write Commands	93
Interrupt Acknowledge Command	93
Special Cycle Command	95
Burst Commands	95
All Burst-Oriented Commands Are Memory Transfers.....	95
Start Address Is Byte-Aligned.....	95
Transfer Length	95
Linear Addressing Is Implied	96
Memory Read Block Command.....	96
Alias To Memory Read Block Command	96
Memory Write Block Command.....	97
Alias To Memory Write Block Command	97
Memory Write Command	97
Split Completion Command	97
Dual-Address Cycle (DAC) Command.....	98

Chapter 7: Intro to Transaction Phases

The PCI Transaction Phases.....	100
The PCI-X Transaction Phases.....	100
Address Phase	100

Contents

Attribute Phase	100
Response Phase	101
Data Phase(s)	101

Chapter 8: Intro to Transaction Termination

Introduction.....	103
Initiator Termination of Transaction	104
Byte Count Satisfaction	104
Initiator Approaching Buffer Full or Dry Condition	104
Connection Timeout	105
Early Target Termination of Transaction	106
General.....	106
Target Abort.....	107
Reasons for a Target Abort.....	107
Target Abort Always Ends a Transaction	107
Retry	107
Reasons for Issuing a Retry	107
Initiator's Response to a Retry	108
Response to a Retry in PCI.....	108
Response to a Retry in PCI-X.....	108
Detailed Description of Retry	108
Single Data Phase Disconnect	108
Disconnect At Next ADB	109
Split Response.....	109
Definition	109
Commands That Will Not Receive a Split Response.....	110
Bridge Handling of Memory Writes.....	110
Bridge Handling of Transactions Other Than Memory Writes	110
Addressing a Location Within the Bridge	111
Addressing a Location on the Opposite Side of Bridge	111

Chapter 9: Intro to Split and Immediate Transactions

Definition of Requester and Completer.....	114
Definition of a Sequence	114
Definition of Requester ID, Tag, and Sequence ID	114
How Does a Device Know Its Requester ID?.....	114
Immediate Transaction	116
Definition of an Immediate Transaction.....	116
Immediate Completion Completes the Sequence... ..	116
...Unless It's a Memory Write or Memory Write Block	116
...Or Unless First Data Phase of Memory Write Receives Retry	117

Memory Writes Are Posted	117
General.....	117
Can Be Initiated Before All Write Data Is Ready.....	117
Split Transactions	118
What Problem Do Split Transactions Solve?.....	118
Split Completion Uses Sequence ID as the Address.....	118
Requester and Completer May Reside on the Same Bus	119
Example.....	119
Split Completion May Contain a Message.....	119
Memory Read May Result in Multiple Split Completion Transactions	120
Completer Has All Data Buffered Up: One Split Completion.....	120
Some Data Buffered Up: Multiple Split Completions.....	120
Requester and Completer May Reside on Different Buses.....	121
Example.....	121
Case One: Completer Issues a Split Response.....	122
Case Two: Completer Transfers Data Immediately	122
Requester Buffer Space Management	124
Allocate Buffer to Hold All of the Data, Then Arbitrate.....	124
On Split Response, Requester Must Commit Buffer Space	125
On Immediate Response, Requester Can Release Buffer Space.....	125
Bridge Can Retry or Disconnect Split Completion	125
Requesters Must Implement a Split Request Timeout	125

Part 2: Transaction Protocol

Chapter 10: Bus Arbitration

Stepping Not Permitted	129
Request and Grant Signals Are Registered	129
Example Arbitration	130
Device Design Rules	133
No Fast Back-to-Back.....	133
REQ# Can Be Asserted or Deasserted at Any Time.....	134
Issuing REQ# and Then Changing Your Mind.....	134
After REQ# Is Asserted, Assert FRAME# or Deassert REQ#.....	134
Initiator Can Start Transaction Two Clocks After GNT# Is Asserted	134
GNT# Removal in Clock Prior to FRAME# Too Late to Stop Initiation	135
Single Idle Clock Between Transactions.....	135
Multiple Idle Clocks Between Transactions.....	137
Back-to-Back Transactions by Same Initiator.....	138
Device Containing Multiple Initiators	138

Contents

Arbiter Design Rules	138
No GNT# Asserted and REQ# Detected.....	138
Must Allow Opportunity to Start Configuration Transaction	139
When Bus Not Idle, GNT# Can Be Deasserted Sooner	140
If Transaction Not Started, Arbiter May Ignore Master.....	140
Removing GNT# From One Initiator and Giving It to Another	141
Parking Recommendation	141
Bus Parking	141
Floating Bus Causes Power Drain	141
Bus Parking Prevents Bus Float	141
Rules Associated with Bus Parking.....	142
How the Initiator Deals With Preemption	143
The Basics	143
After Timeslice Exhaustion and Preemption, How Soon Must It Yield?	144
What Is the Recommended LT Value?.....	144
How Much Data Can Be Transferred During Default Timeslice?	144

Chapter 11: Detailed Command Description

Dword Commands	147
General.....	147
Command Encoding.....	148
Illegal to Assert REQ64# in Dword Transactions.....	148
IO Read and Write Commands.....	149
Basic Description.....	149
Start Address and Byte Enable Format.....	149
Start Address Is Byte-Aligned	149
Byte Enable Usage	149
IO Transaction With No Byte Enables Asserted	149
Target Response to an IO Access.....	151
General	151
Completer Handling of an IO Access	151
When Bridge Acts as Target of IO Access.....	151
Memory Read Dword Command.....	152
Start Address and Byte Enable Format.....	152
Start Address Is Byte-Aligned	152
Byte Enable Usage	152
Transaction With No Byte Enables Asserted.....	152
Target Response to a Memory Read Dword Access.....	152
General	152
Completer Handling of a Memory Read Dword Access	152
Bridge Acts as Target of Memory Read Dword Access.....	153
Configuration Read and Write Commands	153

Interrupt Acknowledge Command	154
General	154
Background.....	154
Host/PCIX Bridge Handling of Interrupt Acknowledge	155
System Using APIC Bus to Deliver Interrupts To Processors.....	157
Non-Intel System Implementation.....	159
Special Cycle Command	159
General	159
Basic Description.....	159
Address Phase	160
Attribute Phase.....	160
Data Phase.....	160
In PCI-X Special Cycle Transaction, No Wait States Permitted	160
No Target Is Permitted to Respond.....	160
Ending the Transaction.....	161
The Message Types.....	162
Devices That May Initiate a Special Cycle Transaction.....	163
Software-Initiated Special Cycle Transactions	163
Burst Commands	163
All Burst-Oriented Commands Are Memory Transfers.....	163
Burst Command Encoding	163
Start Address Is Byte Aligned	165
PCI Memory Addressing.....	165
PCI-X Memory Addressing.....	165
Linear Addressing Is Implied	166
Transfer Length.....	166
In PCI, Transfer Length Is Unknown.....	166
In PCI-X, Byte Transfer Count Defines End Address.....	167
Memory Read Block Command.....	167
Basic Description.....	167
Can Be Purely Speculative.....	167
Target Response to Memory Read Block Access	167
Must Commit Buffer Space Before Initiating.....	169
Alias To Memory Read Block Command	169
Memory Write Block Command.....	169
Basic Description.....	169
If Disconnected, Must Resume With Same Sequence ID	170
If Requester Cannot Complete Memory Write.....	170
Technically, a Write Can Be Purely Speculative, But... ..	171
Target Never Permitted to Split a Memory Write	171
Target Response to Memory Write Block Transaction.....	171
Doesn't Need All of the Write Data Ready Before Initiating	172

Contents

Emulating the PCI Memory Write and Invalidate Command.....	172
Alias To Memory Write Block Command.....	174
Memory Write Command	174
Basic Description.....	174
Can Be Purely Speculative.....	175
Target Never Permitted to Split a Memory Write	175
Target Response to Memory Write Transaction.....	175
Doesn't Need All of the Write Data Ready Before Initiating	175
Split Completion Command	176
Basic Description.....	176
At Minimum, Completer Must Handle One Split Transaction	176
Can Be Initiated Even If Bus Master Bit Is Off.....	176
Split Transactions Never Associated With Memory Writes.....	177
One or More Split Completions to Fulfill One Request	177
Separate Requests May Be Fulfilled out of Order	177
The Address Phase	178
The Attribute Phase.....	178
Claiming a Split Completion Transaction	180
Requester and Completer Are on the Same Bus	180
Requester and Completer Are on Different Buses.....	181
Master Abort or Target Abort on a Split Completion	182
Treatment of Byte Enables.....	182
Requester Must Accept the Split Completion Data	182
When Bridge Is Target of Completion, Can Retry or Disconnect.....	182
Initiator of Split Completion Can Disconnect.....	183
After Disconnect, Pick Up Where You Left Off.....	183
Split Completion Messages	183
To Disconnect First Split Completion at Imminent ADB.....	184
Disconnecting First Split Completion Is Problematic.....	184
The Problem	184
The Solution.....	185
Bridge Buffer Space Problem Due to Corrupted Split Completion.....	186
Dual-Address Cycle (DAC) Command.....	187

Chapter 12: Latency Rules

Initiator Latency Rules.....	189
Don't Start Transfer If You're Not Ready.....	189
No Initiator Wait States Permitted...Ever	190
Behavior When Preempted.....	191
Target Latency Rules	191
Target Response Time During Initialization Period	191
What's Going on During Initialization Time?	191
