PCI-E note 2.0

SYSTEM software — BIOS

operating SYSTEM — Linux

PCI Express 总线性能[2][3]

PCI Express 版本	推出	Line 编码	原始传输率[i]	带宽（带宽）[i]				
				×1	×2	×4	×8	×16
1.0	2003	8b/10b	2.5 GT/s	250 MB/s	0.50 GB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s
2.0	2007	8b/10b	5.0 GT/s	500 MB/s	1.0 GB/s	2.0 GB/s	4.0 GB/s	8.0 GB/s
3.0	2010	128b/130b	8.0 GT/s	984.6 MB/s	1.97 GB/s	3.94 GB/s	7.88 GB/s	15.8 GB/s
4.0	2017	128b/130b	16.0 GT/s	1969 MB/s	3.94 GB/s	7.88 GB/s	15.75 GB/s	31.5 GB/s
5.0[5][6]	2019[7][8]	NRZ 128b/130b	32.0 GT/s[ii]	3938 MB/s	7.88 GB/s	15.75 GB/s	31.51 GB/s	63.0 GB/s
6.0	2021	PAM4 & FEC 128b/130b	64.0 GT/s	7877 MB/s	15.75 GB/s	31.51 GB/s	63.02 GB/s	126.03 GB/s

i. ^ [1.0] [1.1] 每条通道（lane）是全双工通道。

ii. ^ 出于技术可行性，最初也考虑过25.0 GT/s

以PCIe 2.0为例，每秒5GT（Gigatransfer）原始数据传输率，编码方式为8b/10b（每10个比特只有8个有效数据），即有效带宽为4Gb/s = 500MByte/s。

1.3. PCI Express Fabric Topology

A fabric is composed of point-to-point Links that interconnect a set of components – an example fabric topology is shown in Figure 1-2. This figure illustrates a single fabric instance referred to as a hierarchy – composed of a Root Complex (RC), multiple Endpoints (I/O devices), a Switch, and a PCI Express to PCI/PCI-X Bridge, all interconnected via PCI Express Links.

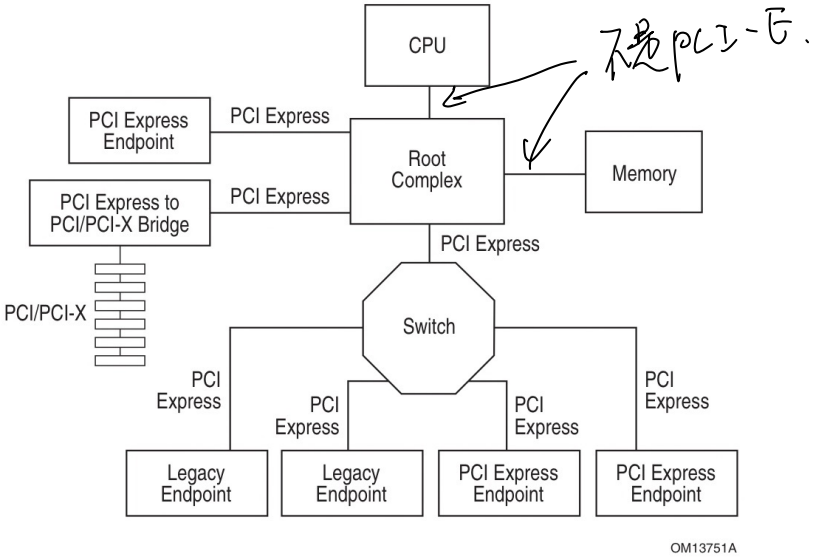


Figure 1-2: Example Topology

1. RC连cpu和memory是I/O口.

2. RL有3个PCI-E口，通下口是一个domain

3. RL可拆包.

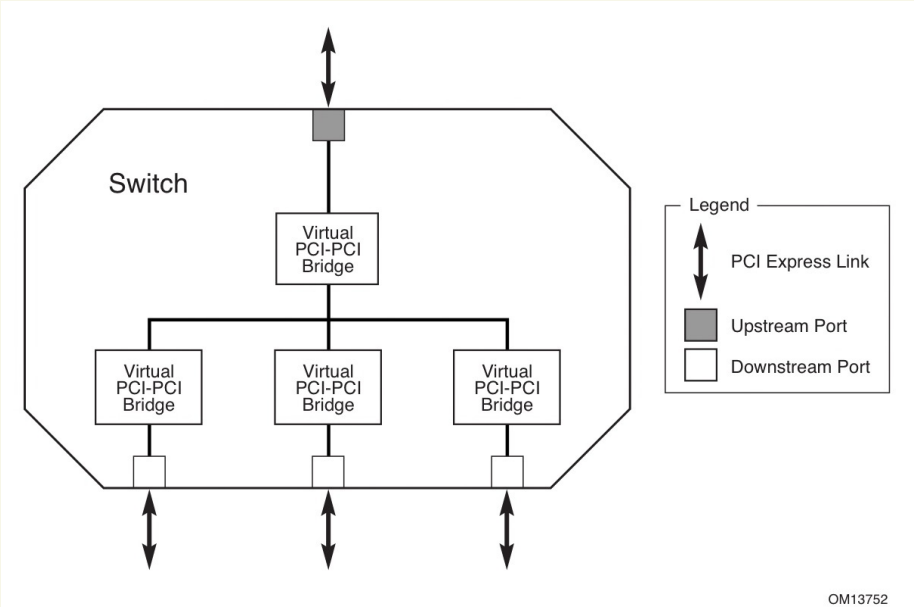


Figure 1-3: Logical Block Diagram of a Switch

1. pci-E switch 当输出端口通用 address-Based routing.

2. switch 不可拆包

3. ingress port 间仲裁可用 round-robin & weighted round-robin

Note that this layering does not imply a particular PCI Express implementation.

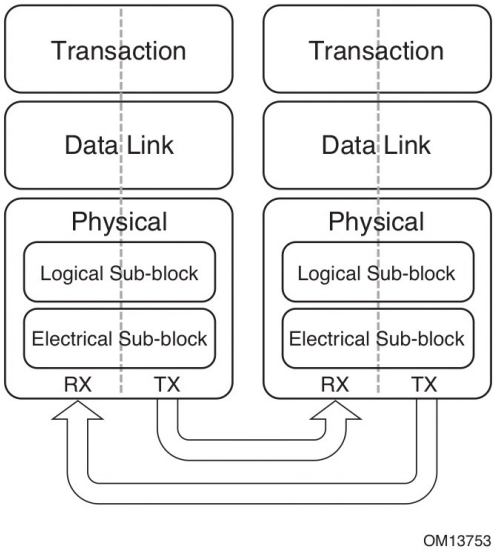


Figure 1-4: High-Level Layering Diagram

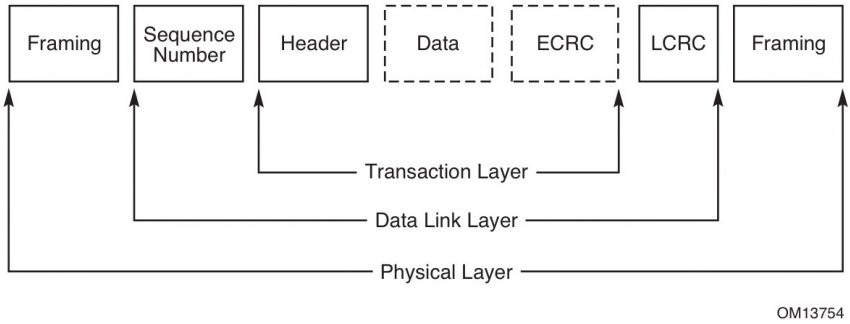


Figure 1-5: Packet Flow Through the Layers

transaction layer:

1. communicate transactions, R&W

2. 每个包有唯一标识

3. 4种地址空间

- memory
- I/O
- configuration
- address message ← 相当于带外中断，电源管理信号。

5. flow-control

Data link layer

1. 链路管理，数据查错与纠错
2. 在TLP层基础上加 data protection code 和 sequence number.
3. Rx-检测，如出错，则要求对方重传，
4. DLLP包用于 link management.

physical layer.

1. HW driver, input buffers.
2. 并串，串并转换
3. PLL
4. interface init,

At a high level, the key aspects of the Transaction Layer are:

- ❑ A pipelined full split-transaction protocol
- ❑ Mechanisms for differentiating the ordering and processing requirements of Transaction Layer Packets (TLPs)
- ❑ Credit-based flow control
- ❑ Optional support for data poisoning and end-to-end data integrity detection.

The Transaction Layer comprehends the following:

- ❑ TLP construction and processing
- ❑ Association of transaction-level mechanisms with device resources including:
  - Flow Control
  - Virtual Channel management
- ❑ Rules for ordering and management of TLPs
  - PCI/PCI-X compatible ordering
  - Including Traffic Class differentiation

This chapter specifies the behaviors associated with the Transaction Layer.

2.1.1. Address Spaces, Transaction Types, and Usage

Transactions form the basis for information transfer between a Requester and Completer. Four address spaces are defined, and different Transaction types are defined, each with its own unique intended usage, as shown in Table 2-1.

Table 2-1: Transaction Types for Different Address Spaces

Address Space	Transaction Types	Basic Usage
Memory	Read Write	Transfer data to/from a memory-mapped location.
I/O	Read Write	Transfer data to/from an I/O-mapped location
Configuration	Read Write	Device Function configuration/setup
Message	Baseline (including Vendor–defined)	From event signaling mechanism to general purpose messaging

I/O space 是为了兼容老的device，不好就使用

- ❑ Read Request/Completion
- ❑ Write Request/Completion

I/O Transactions use a single address format:

- ❑ Short Address Format: 32-bit address

configure space

Configuration Transactions are used to access configuration registers of Functions within devices.

Configuration Transactions include the following types:

- ❑ Read Request/Completion
- ❑ Write Request/Completion

memory space

Memory Transactions include the following types:

- ❑ Read Request/Completion
- ❑ Write Request

Memory Transactions use two different address formats:

- ❑ Short Address Format: 32-bit address
- ❑ Long Address Format: 64-bit address

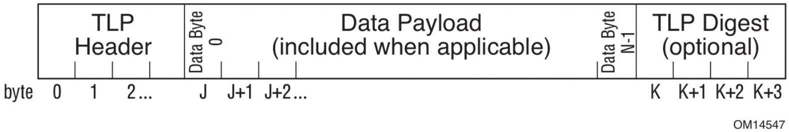


Figure 2-2: Serial View of a TLP

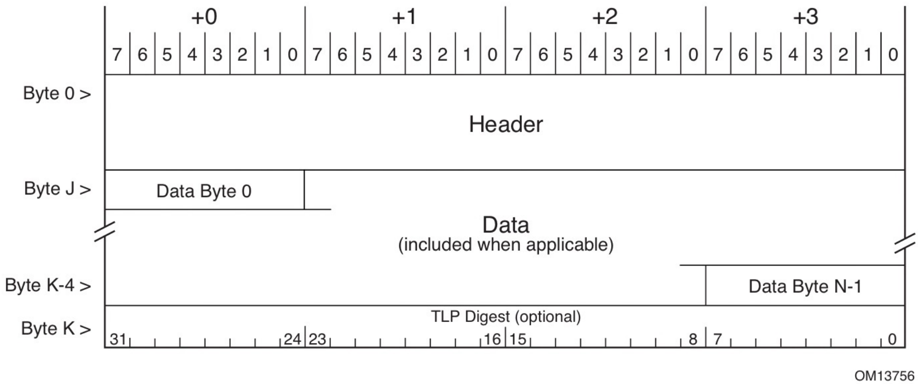


Figure 2-3: Generic TLP Format

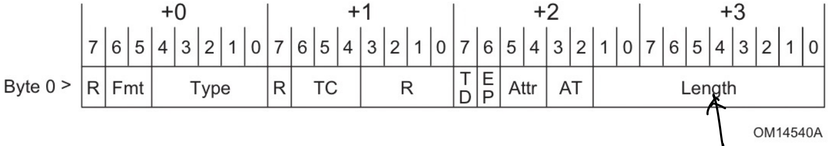
byte / bit 序和 internet 不同、字节内为大端，为的是 early address decode.

帧格式

1. 由 requests 和 completions 组成

↑

- 返回读
- I/O ack
- 配置写



单位: 4 BYTES ↔ DW

Figure 2-4: Fields Present in All TLPs

Table 2-2: Fmt[1:0] Field Values

Fmt[1:0]	Corresponding TLP Format
00b	3 DW header, no data
01b	4 DW header, no data
10b	3 DW header, with data
11b	4 DW header, with data

Table 2-3: Fmt[1:0] and Type[4:0] Field Encodings

TLP Type	Fmt [1:0][2] (b)	Type [4:0] (b)	Description
MRd	00 01	0 0000	Memory Read Request
MRdLk	00 01	0 0001	Memory Read Request-Locked
MWr	10 11	0 0000	Memory Write Request
IORd	00	0 0010	I/O Read Request
IOWr	10	0 0010	I/O Write Request
CfgRd0	00	0 0100	Configuration Read Type 0
CfgWr0	10	0 0100	Configuration Write Type 0
CfgRd1	00	0 0101	Configuration Read Type 1
CfgWr1	10	0 0101	Configuration Write Type 1
TCfgRd	00	1 1011	Deprecated TLP Type[3]
TCfgWr	10	1 1011	Deprecated TLP Type[3]
Msg	01	1 0$r_2r_1r_0$	Message Request – The sub-field r[2:0] specifies the Message routing mechanism (see Table 2-12).
MsgD	11	1 0$r_2r_1r_0$	Message Request with data payload – The sub-field r[2:0] specifies the Message routing mechanism (see Table 2-12).
Cpl	00	0 1010	Completion without Data – Used for I/O and Configuration Write Completions and Read Completions (I/O, Configuration, or Memory) with Completion Status other than Successful Completion.
CplD	10	0 1010	Completion with Data – Used for Memory, I/O, and Configuration Read Completions.
CplLk	00	0 1011	Completion for Locked Memory Read without Data – Used only in error case.
CplDLk	10	0 1011	Completion for Locked Memory Read – otherwise like CplD.
			All encodings not shown above are Reserved.

Table 2-12: Message Routing

r[2:0] (b)	Description	Bytes 8 Through 15[5]
000	Routed to Root Complex	Reserved
001	Routed by Address[6]	Address
010	Routed by ID	See Section 2.2.4
011	Broadcast from Root Complex	Reserved
100	Local - Terminate at Receiver	Reserved
101	Gathered and routed to Root Complex[7]	Reserved
110-111	Reserved - Terminate at Receiver	Reserved

} → 指定以何种方式路由。

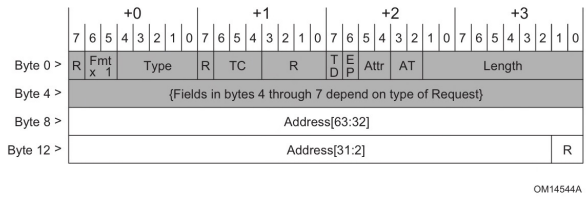
Table 2-4: Length[9:0] Field Encoding

Length[9:0]	Corresponding TLP Data Payload Size
00 0000 0001b	1 DW
00 0000 0010b	2 DW
...	...
11 1111 1111b	1023 DW
00 0000 0000b	1024 DW

读写 64/128 对齐可提高系统性能

Router - type

1. address - based



← { memory req, I/O req

Figure 2-5: 64-bit Address Routing

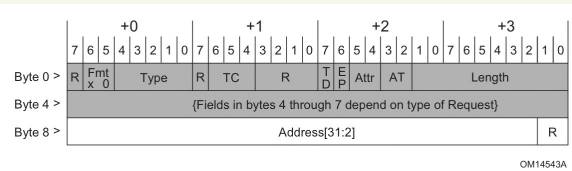


Figure 2-6: 32-bit Address Routing

2. ID - based

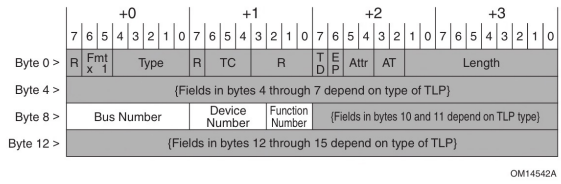


Figure 2-7: ID Routing with 4 DW Header

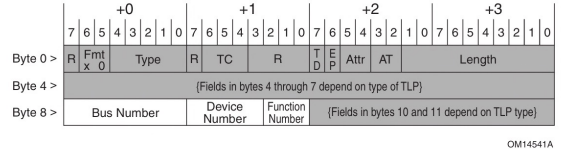


Figure 2-8: ID Routing with 3 DW Header

指定以4字节对齐，所以无效的BYTE会用BYTE_EN信号0来表示

0字节读用作flush，保证之前的写都生效。

00

描述传输信息

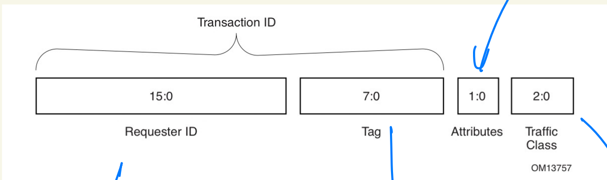


Figure 2-10: Transaction Descriptor

0优先级最高

sender的ID

描述

bus number

+ device number

+ function number.

同一时间内的唯一

1、TAG的大小决定可暂存同一个类型TLP的数量.

2、configure space 中的 capability 有的 TAG 的大小有5bit和8bit之分

3、主要用于区分多个 non-posted 消息的 cmpl

大部分的帧格式（BYTE4 – BYTE8）

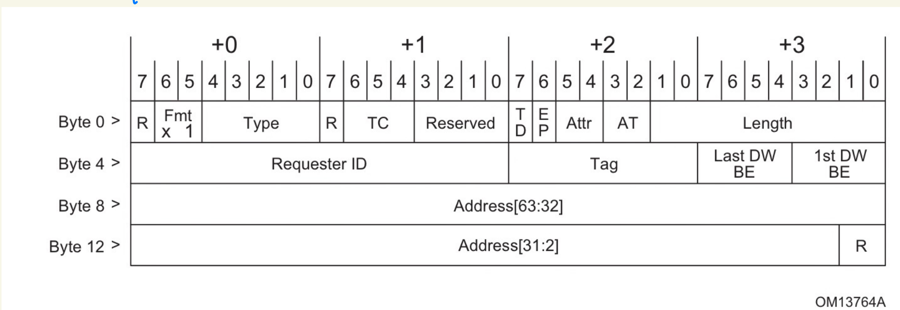


Figure 2-13: Request Header Format for 64-bit Addressing of Memory

固定格式 for I/O

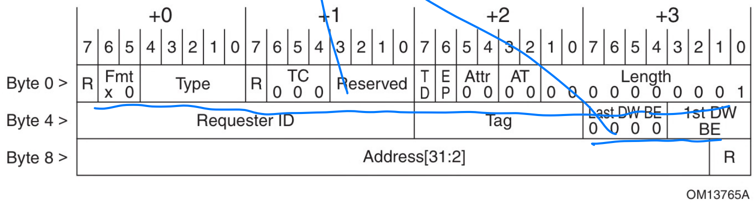


Figure 2-15: Request Header Format for I/O Transactions

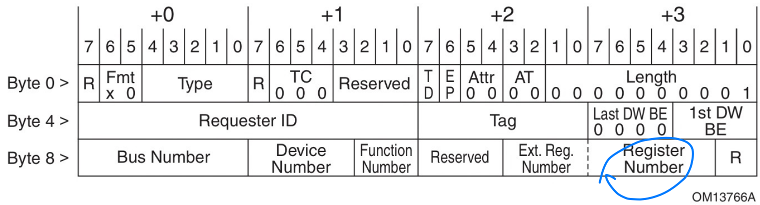


Figure 2-16: Request Header Format for Configuration Transactions

MSI / MSI-X 报文格式类似于 memory req

指定 routing 方式

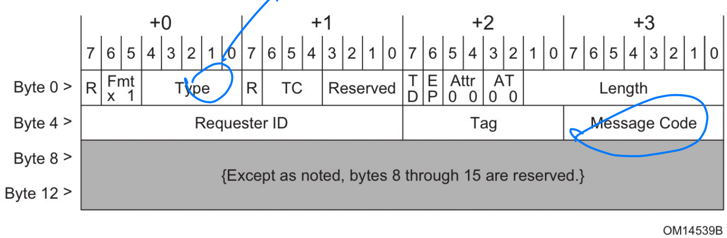


Figure 2-17: Message Request Header

不需要 completion

message code + type[2:0]

所有读请求和Non-posted写请求需要Completion.

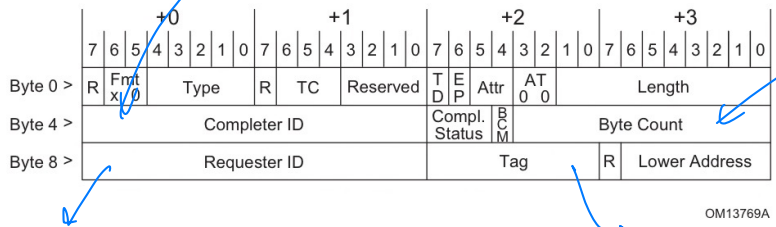


Figure 2-19: Completion Header Format

Table 2-21: Completion Status Field Values

Completion Status[2:0] Field Value (b)	Completion Status
000	Successful Completion (SC)
001	Unsupported Request (UR)
010	Configuration Request Retry Status (CRS)
100	Completer Abort (CA)
all others	Reserved

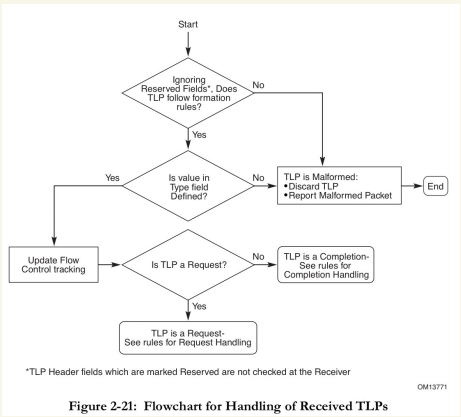


Figure 2-21: Flowchart for Handling of Received TLPs

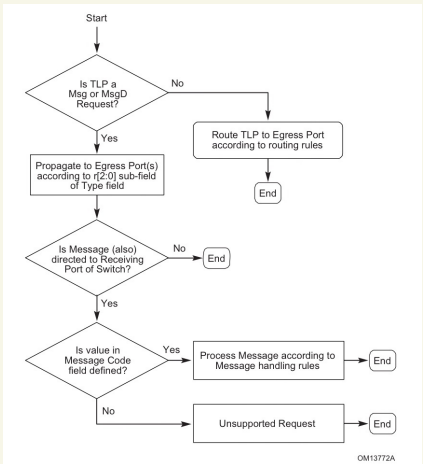


Figure 2-22: Flowchart for Switch Handling of TLPs

当 Read Req 的长度大于 RCB (RC 为 64 或 128 B，取决于 RCB 的大小) 时，有可能返回多个 completion

Table 2-22: Calculating Byte Count from Length and Byte Enables

1st DW BE[3:0] (b)	Last DW BE[3:0] (b)	Total Byte Count
1xx1	0000[11]	4
01x1	0000	3
1x10	0000	3
0011	0000	2
0110	0000	2
1100	0000	2
0001	0000	1
0010	0000	1
0100	0000	1
1000	0000	1
0000	0000	1
xxx1	1xxx	Length[12] * 4
xxx1	01xx	(Length * 4) - 1
xxx1	001x	(Length * 4) - 2
xxx1	0001	(Length * 4) - 3
xx10	1xxx	(Length * 4) - 1
xx10	01xx	(Length * 4) - 2
xx10	001x	(Length * 4) - 3
xx10	0001	(Length * 4) - 4
x100	1xxx	(Length * 4) - 2
x100	01xx	(Length * 4) - 3
x100	001x	(Length * 4) - 4
x100	0001	(Length * 4) - 5
1000	1xxx	(Length * 4) - 3
1000	01xx	(Length * 4) - 4
1000	001x	(Length * 4) - 5
1000	0001	(Length * 4) - 6

Table 2-24: Ordering Rules Summary Table

Row Pass Column?		Posted Request	Non-Posted Request		Completion	
		Memory Write or Message Request (Col 2)	**Read Request** (Col 3)	**I/O or Configuration Write Request** (Col 4)	**Read Completion** (Col 5)	**I/O or Configuration Write Completion** (Col 6)
Posted Request	**Memory Write or Message Request** (Row A)	a) No b) Y/N	Yes	Yes	a) Y/N b) Yes	a) Y/N b) Yes
Non-Posted Request	**Read Request** (Row B)	No	Y/N	Y/N	Y/N	Y/N
	I/O or Configuration Write Request (Row C)	No	Y/N	Y/N	Y/N	Y/N
Completion	**Read Completion** (Row D)	a) No b)Y/N	Yes	Yes	a) Y/N b) No	Y/N
	I/O or Configuration Write Completion (Row E)	Y/N	Yes	Yes	Y/N	Y/N

IMPLEMENTATION NOTE

Large Memory Reads vs. Multiple Smaller Memory Reads

Note that the rule associated with entry D5b in Table 2-24 ensures that for a single Memory Read Request serviced with multiple Completions, the Completions will be returned in address order. However, the rule associated with entry D5a permits that different Completions associated with distinct Memory Read Requests may be returned in a different order than the issue order for the Requests. For example, if a device issues a single Memory Read Request for 256 bytes from location 1000h, and the Request is returned using two Completions (see Section 2.3.1.1) of 128 bytes each, it is guaranteed that the two Completions will return in the following order:

1st Completion returned: Data from 1000h to 107Fh.

2nd Completion returned: Data from 1080h to 10FFh.

However, if the device issues two Memory Read Requests for 128 bytes each, first to location 1000h, then to location 1080h, the two Completions may return in either order:

1st Completion returned: Data from 1000h to 107Fh.

2nd Completion returned: Data from 1080h to 10FFh.

– or –

1st Completion returned: Data from 1080h to 10FFh.

2nd Completion returned: Data from 1000h to 107Fh.

VC (virtual channel) — 用TC labels来区分不同的 traffic，防止单一个流量阻塞造成瓶颈

TC0 硬绑定 VC0

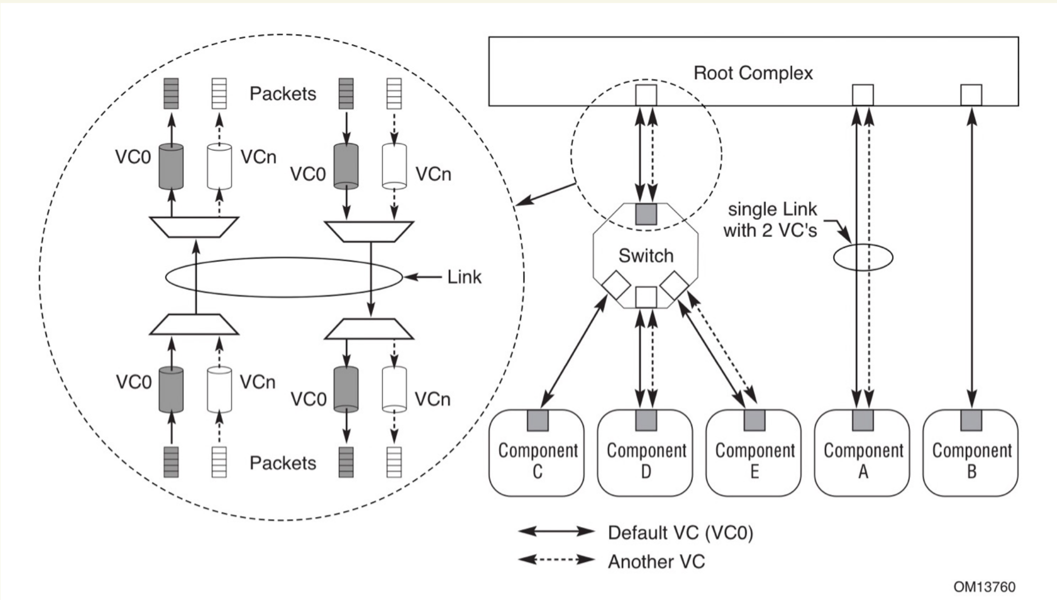


Figure 2-24: Virtual Channel Concept – An Illustration

每个VC建议有独立的queue，buffer 和控制逻辑.

TC → VC 是多对一的关系，每个 port 有独立的 mapping.

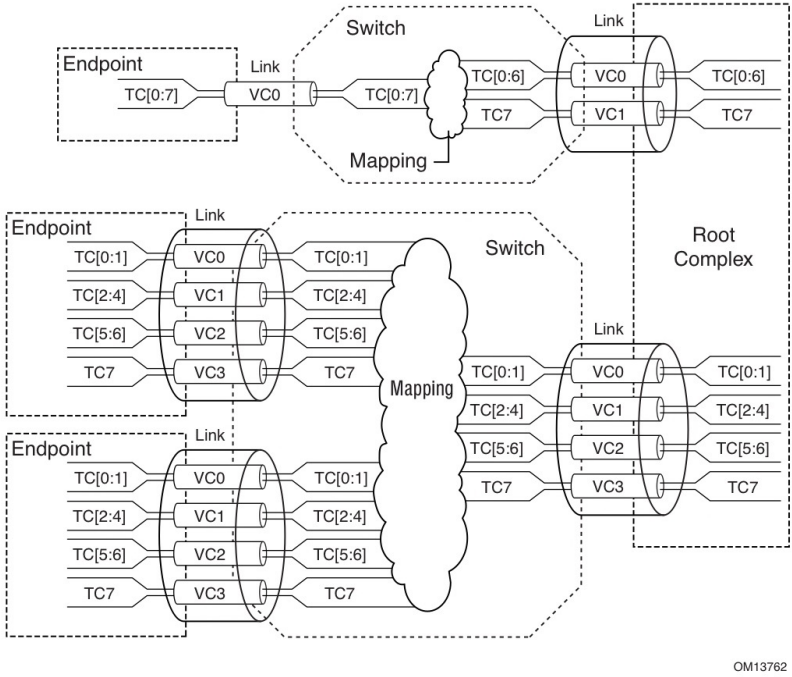


Figure 2-26: An Example of TC/VC Configurations

flow control 是 transaction layer 干的事

1. 使用 FCP，是 DLLP 的一种.
2. 以 credit 为单位，4DW 为单位
3. 不同 VC 有单独的 flow control.
4. 有 6 种不同种类

Table 2-26: Flow Control Credit Types

Credit Type	Applies to This Type of TLP Information
PH	Posted Request headers
PD	Posted Request Data payload
NPH	Non-Posted Request headers
NPD	Non-Posted Request Data payload
CPLH	Completion headers
CPLD	Completion Data payload

Table 2-27: TLP Flow Control Credit Consumption

TLP	Credit Consumed[14]
Memory, I/O, Configuration Read Request	1 NPH unit
Memory Write Request	1 PH + n PD units[15]
I/O, Configuration Write Request	1 NPH + 1 NPD Note: size of data written is never more than 1 (aligned) DW
Message Requests without data	1 PH unit
Message Requests with data	1 PH + n PD units
Memory Read Completion	1 CPLH + n CPLD units
I/O, Configuration Read Completions	1 CPLH unit + 1 CPLD unit
I/O, Configuration Write Completions	1 CPLH unit

在 DATA link layer. 有32位 CRC, 即 LCRC.

在 Transaction layer, 有32位 CRC, 即 ECRC.

{ ECRC. 于于发送端, 检验于接收端.

LCRC, 于于发送端. 每次转发会重新计算.

即使用于 error forwarding. → { Read completion DATA,
Write DATA.

DL_down 时 transaction layer 忘记所有接收报文.

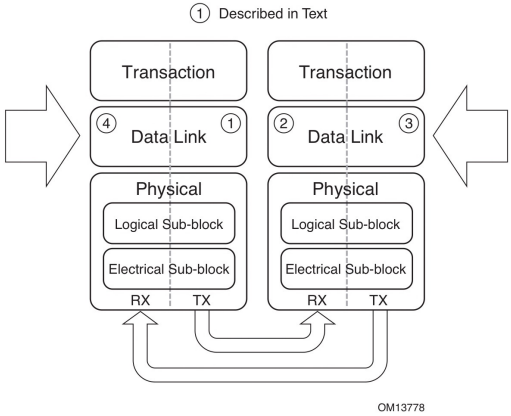


Figure 3-1: Layering Diagram Highlighting the Data Link Layer

DATA Link Layer 提供 TCP 交互的服务.

1. 接受 TCP 下发至物理层
2. 接受物理层上发至 TL.
3. 错误检测及重传.
   a. TCP 序列号.
   b. 存储 TCP 以便重传
   c. 数据完整性接受
   d. 检测 link 状态.

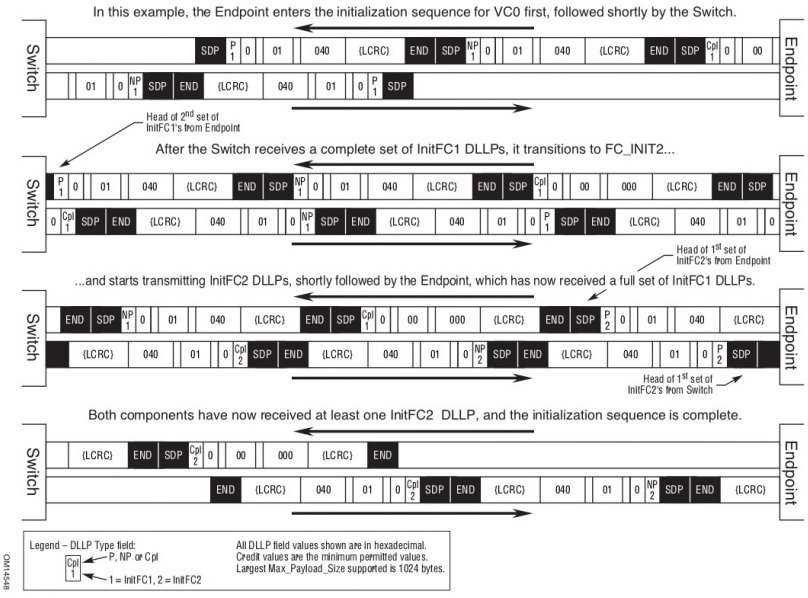


Figure 3-3: VC0 Flow Control Initialization Example

DLLPs: Ack, Nak, InitFC1&FC2 - UpdateFc, for power manage

↓

flow control

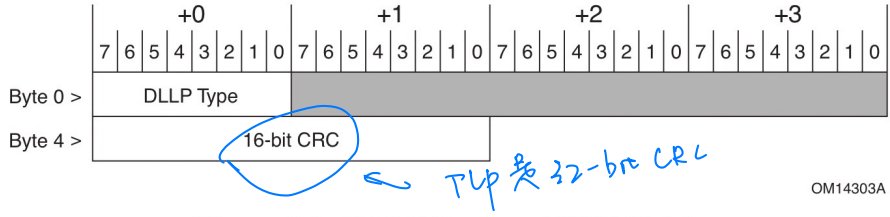


Figure 3-4: DLLP Type and CRC Fields

Table 3-1: DLLP Type Encodings

Encodings	DLLP Type
0000 0000	Ack
0001 0000	Nak
0010 0000	PM_Enter_L1
0010 0001	PM_Enter_L23
0010 0011	PM_Active_State_Request_L1
0010 0100	PM_Request_Ack
0011 0000	Vendor Specific – Not used in normal operation
0100 $0v_2v_1v_0$	InitFC1-P (v[2:0] specifies Virtual Channel)
0101 $0v_2v_1v_0$	InitFC1-NP
0110 $0v_2v_1v_0$	InitFC1-Cpl
1100 $0v_2v_1v_0$	InitFC2-P
1101 $0v_2v_1v_0$	InitFC2-NP
1110 $0v_2v_1v_0$	InitFC2-Cpl
1000 $0v_2v_1v_0$	UpdateFC-P
1001 $0v_2v_1v_0$	UpdateFC-NP
1010 $0v_2v_1v_0$	UpdateFC-Cpl
All other encodings	Reserved

power management

P: post

NP: no-post

CpT: compTetion

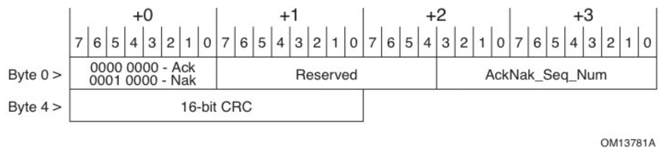


Figure 3-5: Data Link Layer Packet Format for Ack and Nak

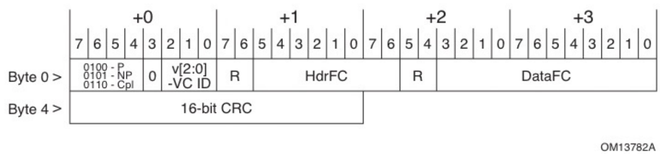


Figure 3-6: Data Link Layer Packet Format for InitFC1

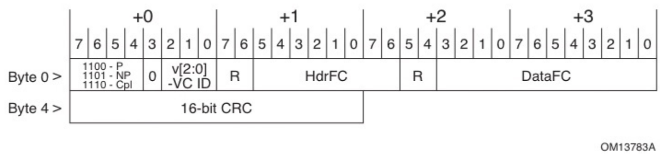


Figure 3-7: Data Link Layer Packet Format for InitFC2

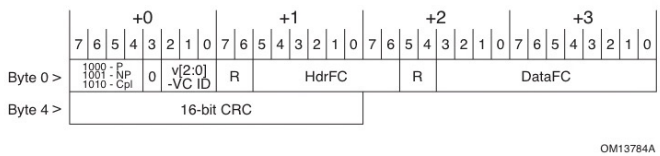


Figure 3-8: Data Link Layer Packet Format for UpdateFC

为TLP下发至DATA Link Layer与报文，会附上Sequence Number和32位的LCRC [Link CRC]

seq := (seq+1) mod 4096

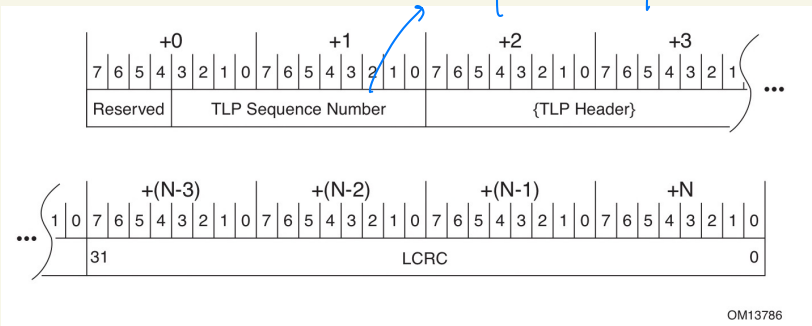


Figure 3-12: TLP with LCRC and Sequence Number Applied

通常来说，发送方会对每一个TLP在Replay buffer中做备份，直到其接收到来自接收方的ACK DLLP，确认该TLP已成功被接收，才会删除这个备份。如果接收方发现TLP存在错误，则会向发送方发送NAK DLLP，然后发送方会从Replay buffer中取出数据，重新发送该TLP。

transmission 有优先级。

1) Completion of any transmission (TLP or DLLP) currently in progress (highest priority)
2) Nak DLLP transmissions
3) Ack DLLP transmissions scheduled for transmission as soon as possible due to:
   receipt of a duplicate TLP –OR–
   expiration of the Ack latency timer (see Section 3.5.3.1)
4) FC DLLP transmissions required to satisfy Section 2.6
5) Retry Buffer re-transmissions
6) TLPs from the Transaction Layer

7) FC DLLP transmissions other than those required to satisfy Section 2.6

8) All other DLLP transmissions (lowest priority)

接收方：

1、维护下一个 expected seq。

2、记录 ACKNAK_LATENCY_TIMER.

3、物理层报错后，忽略收到的TLP

4、收到 seq 和 expected seq 不一致，

a. 如果比 expected TLP 小，则是个 duplicate 包，回 ACK.

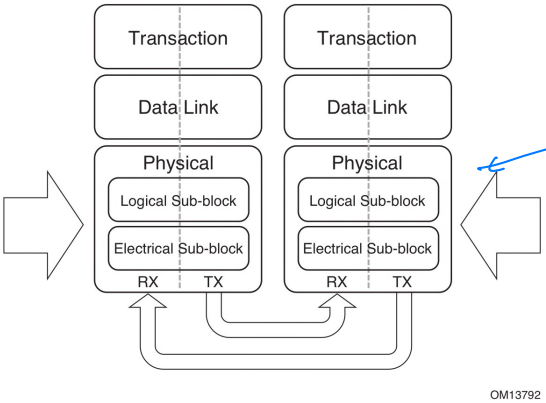


Figure 4-1: Layering Diagram Highlighting Physical Layer

Gen2 - 8B/10B encoding.

多lane时每个symbol下发

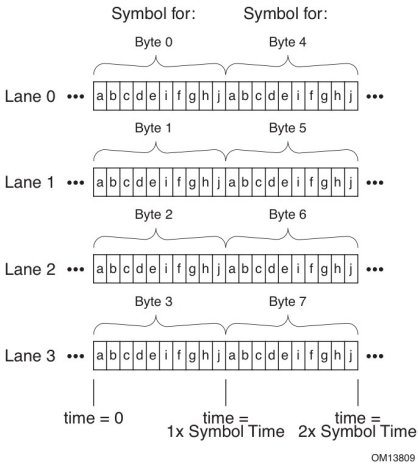


Figure 4-4: Bit Transmission Order on Physical Lanes - x4 Example

8B/10B中有些预留类symbol，用以区分不同种类的Frame或用于link management.

当链路空闲时，进入logic IDLE状态，此时持续发送0，再被解码8B/10B

Table 4-1: Special Symbols

Encoding	Symbol	Name	Description
K28.5	COM	Comma	Used for Lane and Link initialization and management
K27.7	STP	Start TLP	Marks the start of a Transaction Layer Packet
K28.2	SDP	Start DLLP	Marks the start of a Data Link Layer Packet
K29.7	END	End	Marks the end of a Transaction Layer Packet or a Data Link Layer Packet
K30.7	EDB	EnD Bad	Marks the end of a nullified TLP
K23.7	PAD	Pad	Used in Framing and Link Width and Lane ordering negotiations
K28.0	SKP	Skip	Used for compensating for different bit rates for two communicating Ports
K28.1	FTS	Fast Training Sequence	Used within an Ordered Set to exit from L0s to L0
K28.3	IDL	Idle	Used in the Electrical Idle Ordered Set (EIOS)
K28.4			Reserved
K28.6			Reserved
K28.7	EIE	Electrical Idle Exit	Reserved in 2.5 GT/s Used in the Electrical Idle Exit Ordered Set (EIEOS) and sent prior to sending FTS at speeds other than 2.5 GT/s

electrical idle是无数据传输，进入此状态需发送特别的order set，同时退出EI也是一段序列。但要以比较低速的频率。

和IEEE中的8B/10B一样存在disparity，初始化link时TX方向可任取，RX方向取symbol lock时的第一个的symbol。

实际使用中即使出现disparity不一致也可通过配置不认为是一个错误

信号symbol → lane分发

order sets：每个lane都会发

TLP & DLLP：per lane分发

多lane时 STP/SDP必须出在lane mod 4 = 0上

END & EDB不在lane width-1上时，需补上PAD

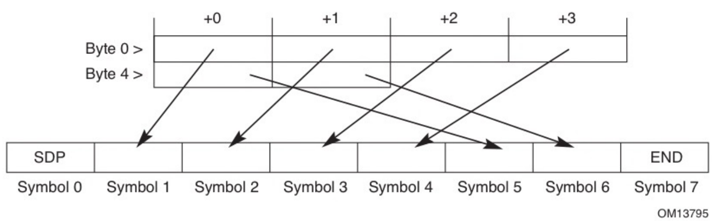


Figure 4-6: DLLP with Framing Symbols Applied

扰码，如果使能了扰码，则其在TX端8B/10B前做，RX端在8B/10B后做解码

在实现方法上是side-stream，且只对数据symbol，special symbol不会，COM容易初始化状态。(重要)

disable scrambling is implementation specific and beyond the scope of this specification.

The data scrambling rules are the following:

- ❑ The COM Symbol initializes the LFSR.
- ❑ The LFSR value is advanced eight serial shifts for each Symbol except the SKP.
- ❑ All data Symbols (D codes) except those within a Training Sequence Ordered Sets (e.g., TS1, TS2) and the Compliance Pattern (see Section 4.2.8) are scrambled.
- ❑ All special Symbols (K codes) are not scrambled.
- ❑ The initialized value of an LFSR seed (D0-D15) is FFFFh. Immediately after a COM exits the Transmit LFSR, the LFSR on the Transmit side is initialized. Every time a COM enters the Receive LFSR on any Lane of that Link, the LFSR on the Receive side is initialized.
- ❑ Scrambling can only be disabled at the end of Configuration (see Section 4.2.6.3.5).
- ❑ Scrambling does not apply to a loopback slave.
- ❑ Scrambling is always enabled in Detect by default.

Link init 和 training.

The following are discovered and determined during the training process:

- ❑ Link width
- ❑ Link data rate
- ❑ Lane reversal
- ❑ polarity inversion.

Training does:

- ❑ Link data rate negotiation.
- ❑ Bit lock per Lane
- ❑ Lane polarity
- ❑ Symbol lock per Lane
- ❑ Lane ordering within a Link
- ❑ Link width negotiation
- ❑ Lane-to-Lane de-skew within a multi-Lane Link.

training 时会发 TS1 or TS2, 8B/10B 是不是用来的

Table 4-2: TS1 Ordered Set

Symbol Number	Encoded Values	Description
0	K28.5	COM for Symbol alignment
1	D0.0 - D31.7, K23.7	Link Number within component
2	D0.0 - D31.0, K23.7	Lane Number within Port
3	D0.0 - D31.7	N_FTS. This is the number of Fast Training Sequences required by the Receiver to obtain reliable bit and Symbol lock.
4	D2.0, D2.2, D2.4, D2.6, D6.0, D6.2, D6.4, D6.6	Data Rate Identifier Bit 0 – Reserved, set to 0b. Bit 1 – When set to 1b, indicates 2.5 GT/s data rate supported. Bit 2 – When set to 1b, indicates 5.0 GT/s data rate supported. Devices that advertise the 5 GT/s data rate must also advertise support for the 2.5 GT/s data rate (i.e., set Bit 1 to 1b). Bit 3:5 – Reserved, must be set to 0b. Bit 6 (Autonomous Change) – Downstream component: Autonomous Change/Selectable De-emphasis: When set to 1b in Configuration state and LinkUp = 1b, indicates that the speed or Link width change initiated by the Downstream component is not caused by a Link reliability issue. In Recovery state, this bit indicates the de-emphasis preference of the Downstream component. In Polling.Active substate, this bit specifies the de-emphasis level the Upstream component must operate in if it enters Polling.Compliance and operates in 5.0 GT/s data rate. In Configuration.Linkwidth.Start substate with LinkUp = 0b and in the Loopback.Entry substate, this bit specifies the de-emphasis level the Upstream component must operate in if it enters Loopback state (from Configuration) and operates in 5.0 GT/s data rate. For de-emphasis, a value of 1b indicates -3.5 dB de-emphasis and a value of 0b indicates -6 dB de-emphasis. This bit is reserved in all other states for a Downstream component. Upstream component: In Polling.Active, Configuration.Linkwidth.Start, and Loopback.Entry substates, this bit specifies the de-emphasis level the Downstream component must operate in 5.0 GT/s data rate if it enters Polling.Compliance and Loopback states, respectively. A value of 1b indicates -3.5 dB de-emphasis and a value of 0b indicates -6 dB de-emphasis. This bit is reserved for all other states. Bit 7 (speed_change) – When set to 1b, indicates a request to change the speed of operation. This bit can be set to 1b only during Recovery.RcvrLock state. All Lanes under the control of a common LTSSM must transmit the same value in this Symbol. Transmitters must advertise all supported data rates in Polling.Active and Configuration.LinkWidth.Start substates, including data rates they do not intend to operate on.
5	D0.0, D1.0, D2.0, D4.0, D8.0, D16.0, D20.0	Training Control Bit 0 – Hot Reset Bit 0 = 0b, De-assert Bit 0 = 1b, Assert Bit 1 – Disable Link Bit 1 = 0b, De-assert Bit 1 = 1b, Assert Bit 2 – Loopback Bit 2 = 0b, De-assert Bit 2 = 1b, Assert Bit 3 – Disable Scrambling Bit 3 = 0b, De-assert Bit 3 = 1b, Assert Bit 4 – Compliance Receive Bit 4 = 0b, De-assert Bit 4 = 1b, Assert Components that support 5 GT/s data rate must implement this bit as specified. Components that support only 2.5 GT/s data rate may optionally implement this bit as a Receiver. If not implemented for components that support only 2.5 GT/s data rate, this bit will be reserved and must behave as if the component received a 0b in this bit position. Bit 5:7 – Reserved Set to 0b
6 – 15	D10.2	TS1 Identifier

Table 4-3: TS2 Ordered Set

Symbol Number	Encoded Values	Description
0	K28.5	COM for Symbol alignment
1	D0.0 - D31.7, K23.7	Link Number within component
2	D0.0 - D31.0, K23.7	Lane Number within Port
3	D0.0 - D31.7	N_FTS. This is the number of Fast Training Sequences required by the Receiver to obtain reliable bit and Symbol lock.
4	D2.0, D2.2, D2.4, D2.6, D6.0, D6.2, D6.4, D6.6	Data Rate Identifier Bit 0 – Reserved, set to 0b Bit 1 – When set to 1b, indicates 2.5 GT/s data rate supported. Bit 2 – When set to 1b, indicates 5.0 GT/s data rate supported. Devices that advertise the 5 GT/s data rate must also advertise support for the 2.5 GT/s data rate (i.e., set Bit 1 to 1b). Bit 3:5 – Reserved, must be set to 0b. Bit 6 (Autonomous Change/Link Upconfigure Capability/Selectable De-emphasis) – This bit is used for Link upconfigure capability notification, as well as Selectable De-emphasis by an Upstream component, and for Link upconfigure capability notification as well as autonomous bandwidth change notification by a Downstream component. In Configuration Complete substate (both for Upstream and Downstream components): When set to 1b, indicates the capability of the component to upconfigure the Link to a previously negotiated Link width during the current LinkUp = 1b state. The device advertising this optional capability must be capable of supporting at least a x1 Link in the Lane 0 assigned in the Configuration state. In Recovery state for a Downstream component: When set to 1b by the Downstream component, indicates that the speed or Link width change initiated by the Downstream component is not caused by a Link reliability issue. In Recovery state for an Upstream component: This bit must be set to 1b if the Upstream component wants the Link to operate in -3.5 dB in 5.0 GT/s speed and reset to 0b if the Upstream component wants the Link to operate in -6 dB in 5.0 GT/s speed if it is advertising 5.0 GT/s data rates (i.e., bit 2 of Symbol 4 is 1b). In Polling.Configuration substate both by Upstream and Downstream components: This indicates the de-emphasis level of transmission in 5.0 GT/s data rate if the receiving device enters Loopback from Configuration. A value of 1b indicates -3.5 dB de-emphasis and a value of 0b indicates -6 dB de-emphasis. This bit is reserved in all the other states not covered above for an Upstream or Downstream component. Bit 7 (speed_change) – When set to 1b, indicates a request to change the speed of operation. This bit can be set to 1b only during Recovery.RcvrCfg state. All Lanes under the control of a common LTSSM must transmit the same value in this Symbol. Transmitters must advertise all supported data rates in Polling.Configuration substate, including data rates they do not intend to operate on.
5	D0.0, D1.0, D2.0, D4.0, D8.0	Training Control Bit 0 – Hot Reset Bit 0 = 0b, De-assert Bit 0 = 1b, Assert Bit 1 – Disable Link Bit 1 = 0b, De-assert Bit 1 = 1b, Assert Bit 2 – Loopback Bit 2 = 0b, De-assert Bit 2 = 1b, Assert Bit 3 – Disable Scrambling Bit 3 = 0b, De-assert Bit 3 = 1b, Assert Bit 4:7 – Reserved Set to 0b
6 – 15	D5.2	TS2 Identifier

1. 每1538个symbol中要有1个SKP order.

2、1280 UI (128 symbol) 中无TS1 & TS2，则认为在Electrical IDLE.

3. 判断有无极性翻转，靠的是TS1 & TS2中的symbol D6-15. 是D10.2还是D21.5 (D10.2的反转) 或 D5.2还是D26.5

4、FTS (fast training) 用于RX方向接收

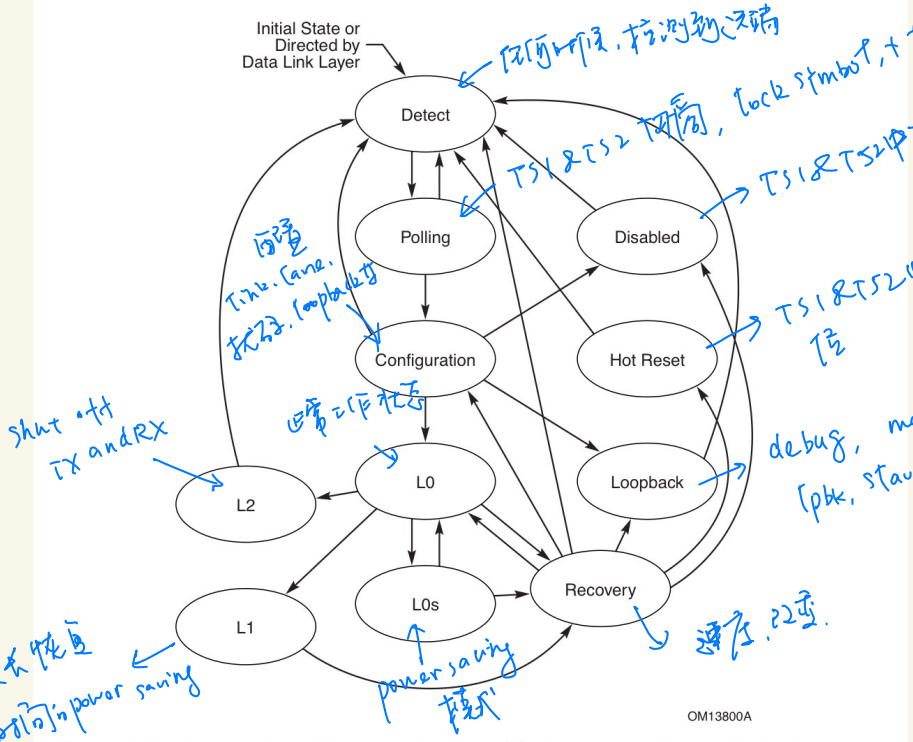


Figure 4-11: Main State Diagram for Link Training and Status State Machine

1. training 的时候，要求 lane 都以 2.5GT/s 的速度协商
2. pcie link 包含一个或多个 lane
3. TS2 中的选项和 TS1 中不同

Table 4-7: Link Status Mapped to the LTSSM

LTSSM State	Link Width	Link Speed	LinkUp	Link Training	Receiver Error	In-Band Presence[26]
Detect	Undefined	Undefined	0b	0b	No action	0b
Polling	Undefined	Set to 2.5 GT/s on entry from Detect. Link speed may change on entry to Polling.Compliance.	0b	0b	No action	1b
Configuration	Set	No action	0b/1b[27]	1b	Set on 8b/10b Error	1b
Recovery	No action	Set to new speed when speed changes	1b	1b	No action	1b
L0	No action	No action	1b	0b	Set on 8b/10b Error or optionally on Framing Violation, Loss of Symbol Lock, Lane Deskew Error, or Elasticity Buffer Overflow/ Underflow	1b
L0s	No action	No action	1b	0b	No action	1b
L1	No action	No action	1b	0b	No action	1b
L2	No action	No action	1b	0b	No action	1b
Disabled	Undefined	Undefined	0b	0b	Optional: Set on 8b/10b Error	1b
Loopback	No action	Link speed may change on entry to Loopback from Configuration.	0b	0b	No action	1b
Hot Reset	No action	No action	0b	0b	Optional: Set on 8b/10b Error	1b

[26] In-band refers to the fact that no sideband signals are used to calculate the presence of a powered up device on the other end of a Link.

[27] LinkUp will always be 0 if coming into Configuration via Detect -> Polling -> Configuration and LinkUp will always be 1 if coming into Configuration from any other state.

The state machine rules for configuring and operating a PCI Express Link are defined in the following sections.

clock compensation

TX:

1. 所有 lane 同时发送数据.
2. skp order sets: com + 3个连续 skp.
3. 每 1180 和 1538 间加入 skp

RX:

1. 计数到 skp 信号. com + 1~5 skp.
   为 lane 时不管在不同 lane 发的不同数量 skp.
2. 1180 ~ 1538 时间内接收
3. tolerance 每 2 个 max_payload_size.

full swing (mandatory)	low swing
de-emphasis	power sensitive — shorter channel

TX 要支持 voltage margining，per link 的设定 必须要支持. per lane 的是 optional 的

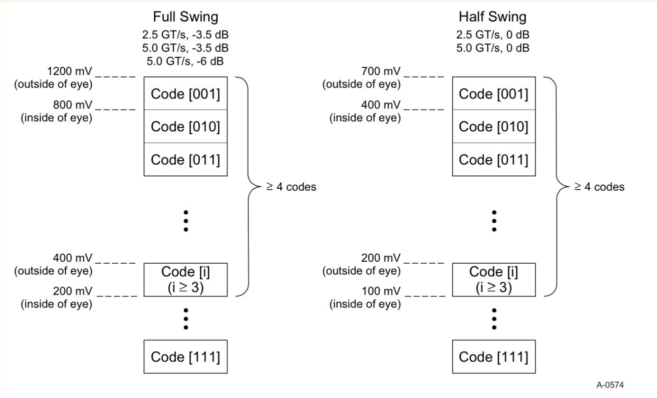


Figure 4-22: Transmitter Margining Voltage Levels and Codes

pin：引脚.　　pad：通常无法直接测量. 是 asic 焊接 pin 的地方

指 30kHz 以下信号叫之为 DC. 高于 30kHz 叫之 AC

TX:

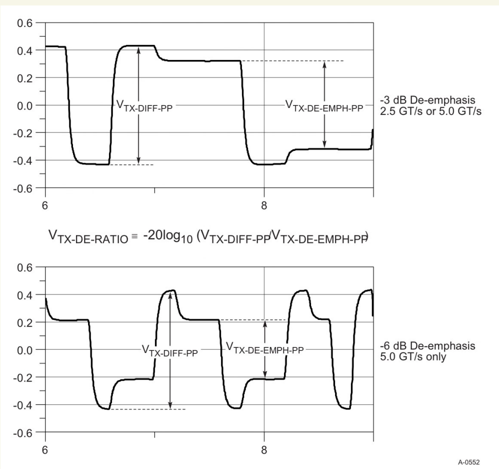


Figure 4-30: Full Swing Tx Parameters Showing De-emphasis

← TX de-emphasis

算法：

$$-20\log_{10}\frac{V_{TX-DIFF-PP}}{V_{TX-DE-EMPH-PP}}$$

RX

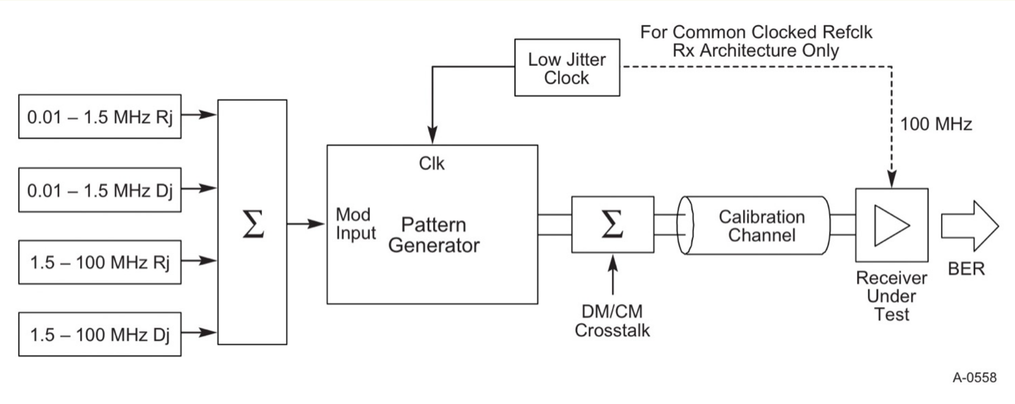


Figure 4-36: Setup for Testing Receiver

RX侧通过接收到EIE Symbol来退出electrical idle.

PCI-E link里的所有lane都是AC-coupled.

BEACON是让PCI-E从L2回到正常状态的信号 Bridge & switch 在 downstream 收到时要转发给 upstream

PLL架构

1. common refclk

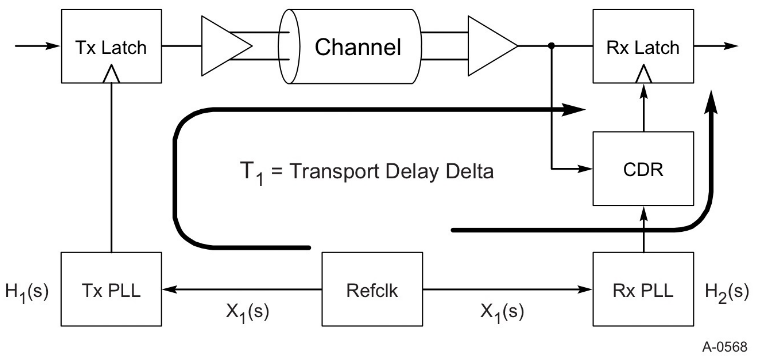


Figure 4-49: Common Refclk Rx Architecture

2. data clocked

收端：时钟使用 CDR 恢复出来的时钟

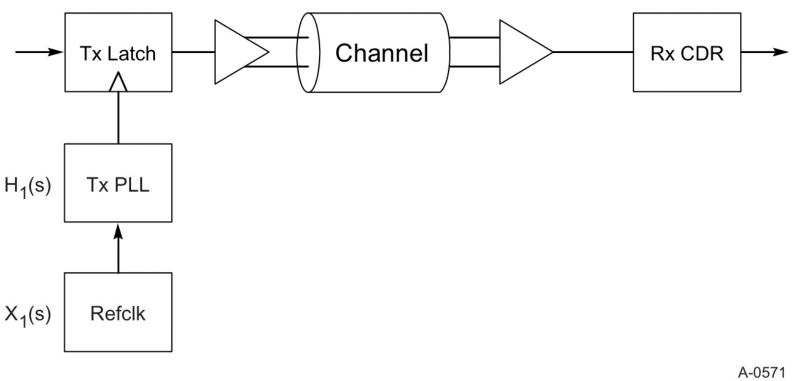


Figure 4-51: Data Driving Architecture

3. separate refclk , Tx & Rx 따로 입력 refclk

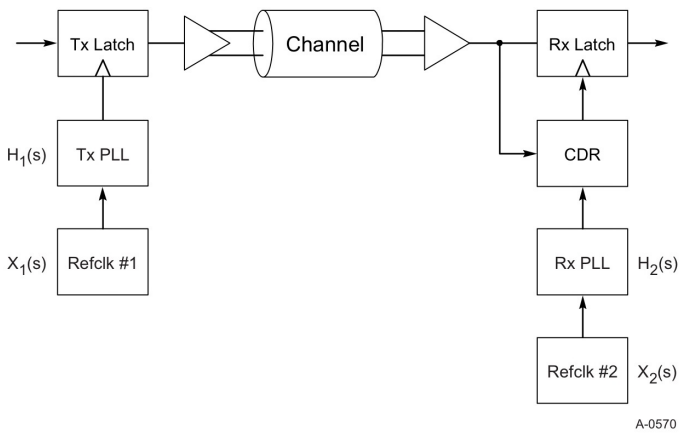


Figure 4-53: Separate Refclk Architecture

PCI Express-PM provides the following services:

- ❑ A mechanism to identify power management capabilities of a given Function
- ❑ The ability to transition a Function into a certain power management state
- ❑ Notification of the current power management state of a Function
- ❑ The option to wakeup the system on a specific event

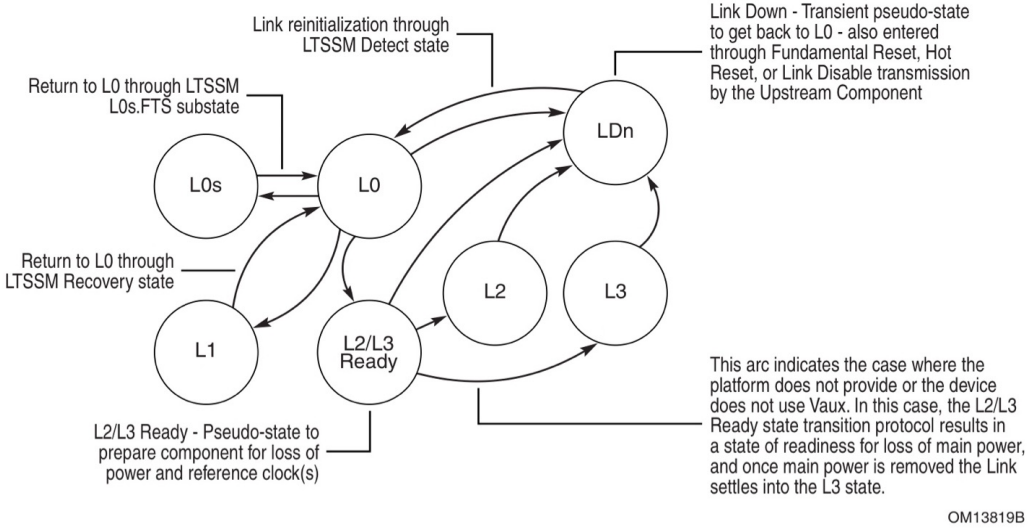


Figure 5-1: Link Power Management State Flow Diagram

1. 软件可以设置 PCI-E power state. 通过 configure write.
2. 同时 PME (WOL) 中断 system power

中断

1. Intx 模拟
   为了向后兼容的设备

2. msi 和 msi-x
   pci-e 设备必须支持，实现必须能够是通过 memory write transaction，同时基于 edge-triggered.

> If the Root Port is enabled for edge-triggered interrupt signaling using MSI or MSI-X, an interrupt message must be sent every time the logical AND of the following conditions transitions from FALSE to TRUE:
>
> - ❑ The associated vector is unmasked (not applicable if MSI does not support PVM).
> - ❑ The PME Interrupt Enable bit in the Root Control register is set to 1b.
> - ❑ The PME Status bit in the Root Status register is set.
>
> Note that PME and Hot-Plug Event interrupts (when both are implemented) always share the same MSI or MSI-X vector, as indicated by the Interrupt Message Number field in the PCI Express Capabilities register.

错误报告和日志

Base line 和 AER

1. 错误分为 fatal, nofata, Uncorrectable, Correctable

2. correctable errors
   无需 software 参与

3. uncorrectable errors
   - fatal → 必须要 reset
   - no-fatal → 需要软件

错误信号实现

1. completion status
通过回 unsupported request completion status

2. error message — 一种 TLP. message request

Table 6-1: Error Messages

Error Message	Description
ERR_COR	This Message is issued when the Function or Device detects a correctable error on the PCI Express interface. Refer to Section 6.2.2.1 for the definition of a correctable error.
ERR_NONFATAL	This Message is issued when the Function or Device detects a Non-fatal, uncorrectable error on the PCI Express interface. Refer to Section 6.2.2.2.2 for the definition of a Non-fatal, uncorrectable error.
ERR_FATAL	This Message is issued when the Function or Device detects a Fatal, uncorrectable error on the PCI Express interface. Refer to Section 6.2.2.2.1 for the definition of a Fatal, uncorrectable error.

root complex 通过 message header 中的 Request ID 来确认发生错误.

error pollution

For errors detected in the Transaction layer, it is permitted and recommended that no more than one error be reported for a single received TLP, and that the following precedence (from highest to lowest) be used:

- ❑ Receiver Overflow
- ❑ Flow Control Protocol Error
- ❑ ECRC Check Failed
- ❑ Malformed TLP
- ❑ Unsupported Request (UR), Completer Abort (CA), or Unexpected Completion[44]
- ❑ Poisoned TLP Received

Error logging

和 function 无关的错误，会在此 device 的所有 function 的 registers 中出现.

Interrupt generation

1. level-triggered interrupt — IntX message

2. edge-triggered interrupt — MSI & MSI-X

(更细) vector 表明 AER interrupt message.

Table 6-2: Physical Layer Error List

Error Name	Error Type	Detecting Agent Action[54]	References
Receiver Error	Correctable	*Receiver:* Send ERR_COR to Root Complex.	Section 4.2.1.3 Section 4.2.2 Section 4.2.4.6 Section 4.2.6

Table 6-3: Data Link Layer Error List

Error Name	Error Type (Default Severity)	Detecting Agent Action[54]	References
Bad TLP	Correctable	*Receiver:* Send ERR_COR to Root Complex.	Section 3.5.3.1
Bad DLLP		*Receiver:* Send ERR_COR to Root Complex.	Section 3.5.2.1
Replay Timeout		*Transmitter:* Send ERR_COR to Root Complex.	Section 3.5.2.1
REPLAY NUM Rollover		*Transmitter:* Send ERR_COR to Root Complex.	Section 3.5.2.1
Data Link Layer Protocol Error	Uncorrectable (Fatal)	If checking, send ERR_FATAL to Root Complex.	Section 3.5.2.1

Error Name	Error Type (Default Severity)	Detecting Agent Action[54]	References
Surprise Down		If checking, send ERR_FATAL to Root Complex.	Section 3.5.2.1

Table 6-4: Transaction Layer Error List

Error Name	Error Type (Default Severity)	Detecting Agent Action[54]	References
Poisoned TLP Received	Uncorrectable (Non-Fatal)	*Receiver:* Send ERR_NONFATAL to Root Complex or ERR_COR for the Advisory Non-Fatal Error cases described in Sections 6.2.3.2.4.2 and 6.2.3.2.4.3. Log the header of the Poisoned TLP.[55]	Section 2.7.2.2
ECRC Check Failed		*Receiver (if ECRC checking is supported):* Send ERR_NONFATAL to Root Complex or ERR_COR for the Advisory Non-Fatal Error case described in Section 6.2.3.2.4.2. Log the header of the TLP that encountered the ECRC error.	Section 2.7.1
Unsupported Request (UR)		*Request Receiver:* Send ERR_NONFATAL to Root Complex or ERR_COR for the Advisory Non-Fatal Error case described in Section 6.2.3.2.4.1. Log the header of the TLP that caused the error.	Section 2.2.8.6, Section 2.3.1, Section 2.3.2, Section 2.7.2.2., Section 2.9.1, Section 5.3.1, Section 6.2.3.1, Section 6.2.6, Section 6.2.8.1, Section 6.5.7, Section 7.3.1, Section 7.3.3, Section 7.5.1.1, Section 7.5.1.2

Error Name	Error Type (Default Severity)	Detecting Agent Action[54]	References
Completion Timeout		*Requester:* Send ERR_NONFATAL to Root Complex or ERR_COR for the Advisory Non-Fatal Error case described in Section 6.2.3.2.4.4.	Section 2.8
Completer Abort		*Completer:* Send ERR_NONFATAL to Root Complex or ERR_COR for the Advisory Non-Fatal Error case described in Section 6.2.3.2.4.1. Log the header of the Request that encountered the error.	Section 2.3.1
Unexpected Completion		*Receiver:* Send ERR_COR to Root Complex. This is an Advisory Non-Fatal Error case described in Section 6.2.3.2.4.5. Log the header of the Completion that encountered the error.	Section 2.3.2
ACS Violation		*Receiver (if checking):* Send ERR_NONFATAL to Root Complex. Log the header of the Request TLP that encountered the error.	
Receiver Overflow	Uncorrectable (Fatal)	*Receiver (if checking):* Send ERR_FATAL to Root Complex.	Section 2.6.1.2
Flow Control Protocol Error		*Receiver (if checking):* Send ERR_FATAL to Root Complex.	Section 2.6.1

Error Name	Error Type (Default Severity)	Detecting Agent Action[54]	References
Malformed TLP		*Receiver:* Send ERR_FATAL to Root Complex. Log the header of the TLP that encountered the error.	Section 2.2.2, Section 2.2.3, Section 2.2.5, Section 2.2.7, Section 2.2.8.1, Section 2.2.8.2, Section 2.2.8.3, Section 2.2.8.4, Section 2.2.8.5, Section 2.2.9, Section 2.3, Section 2.3.1, Section 2.3.1.1, Section 2.3.2, Section 2.5, Section 2.5.3, Section 2.6.1, Section 2.6.1.2, Section 6.3.2

For all errors listed above, the appropriate status bit(s) must be set upon detection of the error. For Unsupported Request (UR), additional detection and reporting enable bits apply (see Section 6.2.5).

VC support — vc & port & function - based

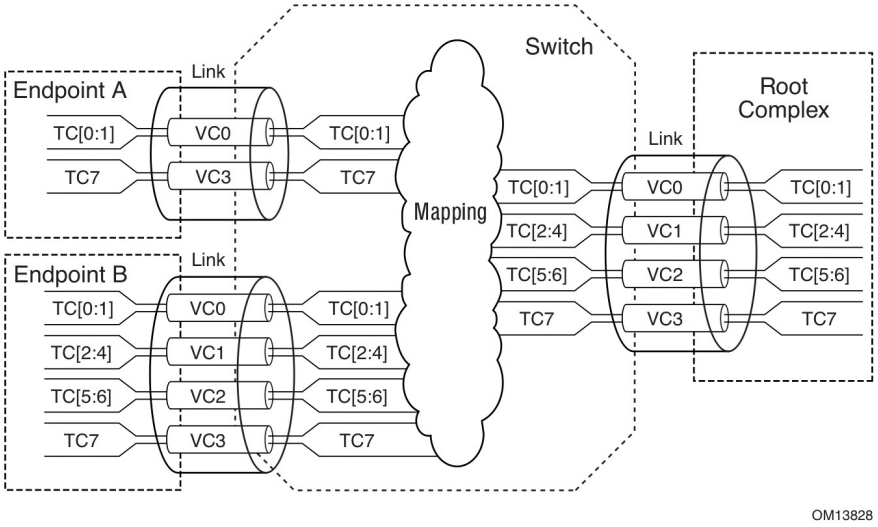


Figure 6-4: TC Filtering Example

TLP 在传输过程中不变的是 TC 值

转发

1. TLP 中 address & router 信息决定 egress port.

2. TC & VC mapping 由 TLP 中的 TC 决定

同一对 不同源 的 traffic 竞争资源为

1. flow control 是否满足

ordering rule (不同 TLP 类型 有前后顺序

2.

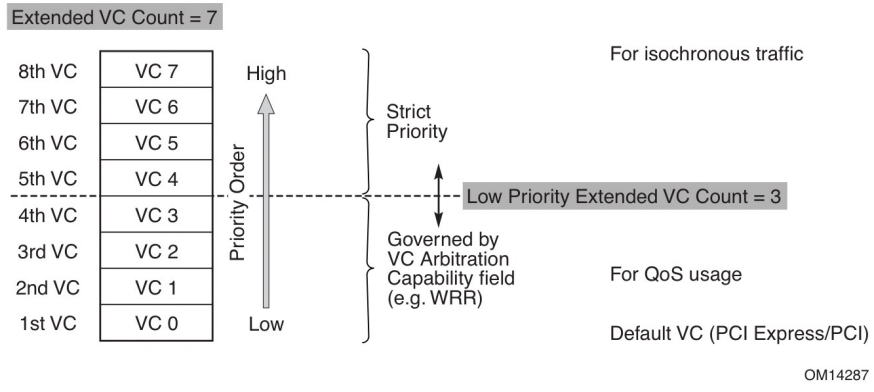


Figure 6-9: VC ID and Priority Order – An Example

VC 间仲裁:

1. Strict priority – 严格: 绝对排序
2. Round Robin – 轮询.
3. Weighted RR – 带 weight 的轮询.

多 function 的设备有多个 VC, 需要 MFVC 机制

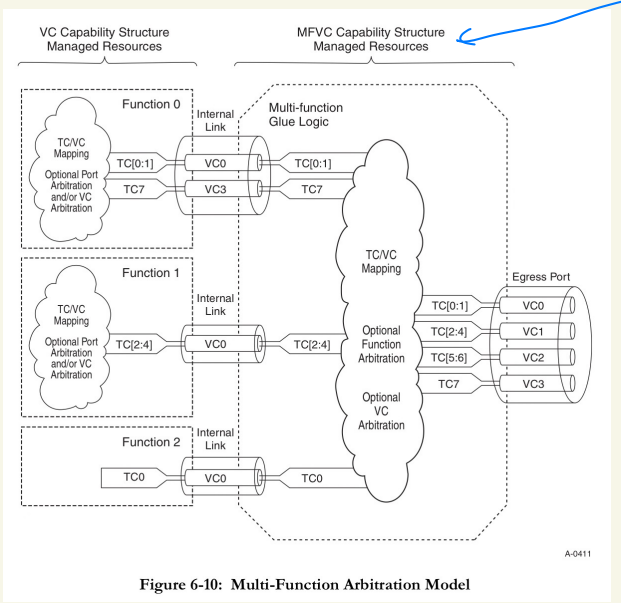


Figure 6-10: Multi-Function Arbitration Model

lock transaction

1、为了兼容老的PCI设备
2、只有RC可发lock access.
3、lock后，其他的RC ↔ legacy PCI路径做traffic阻隔.
4、TLP类型为 MRdLk & CplDLk.
5、只使用默认VC & TC.

Conventional reset. — cold. warm. cold.

1. port寄存器&状态机重置.
2、unReset后，PHY层 bring up link 做link training. 初始化 VCO
3. 一些设备需要额外的delay

除此之外，还可以有 functional Reset ← 由软件驱动

Hot plug

Table 6-5: Elements of Hot-Plug

Element	Purpose
Indicators	Show the power and attention state of the slot
Manually-operated Retention Latch (MRL)	Holds adapter in place
MRL Sensor	Allows the Port and system software to detect the MRL being opened
Electromechanical Interlock	Prevents removal of adapter from slot
Attention Button	Allows user to request hot-plug operations
Software User Interface	Allows user to request hot-plug operations
Slot Numbering	Provides visual identification of slots
Power Controller	Software-controlled electronic component or components that control power to a slot or adapter and monitor that power for fault conditions

Hot plug 相关的寄存器存在于下 downstream port 里的 Device capabilities, slot capabilities, slot control, slot status 中

Speed management.

软件可以控制链路速度 trigger retrain

ACS 是 PCI-E 的 ACL

- ❑ ACS Source Validation (V)
- ❑ ACS Translation Blocking (B)
- ❑ ACS P2P Request Redirect (R)
- ❑ ACS P2P Completion Redirect (C)
- ❑ ACS Upstream Forwarding (U)

- ❑ ACS P2P Egress Control (E)
- ❑ ACS Direct Translated P2P (T)

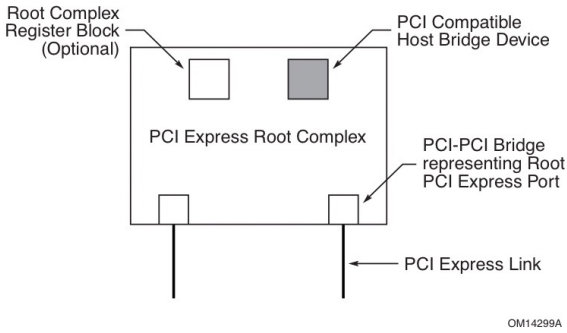


Figure 7-1: PCI Express Root Complex Device Mapping

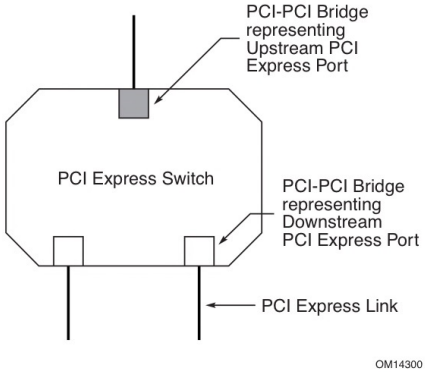


Figure 7-2: PCI Express Switch Device Mapping [67]

RC 和 switch 内部为总线型结构 PCI.

1. PCI-E link 北边为 PCI bridge to second bus

2. 只有 RC, switch 才有 PCI-bridge 结构，印没有

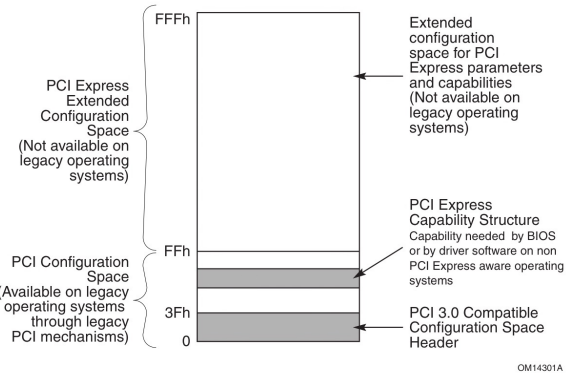


Figure 7-3: PCI Express Configuration Space Layout

1. 前256字节兼容老的PCI.

2. 0x100 - 0xFFF 为 PCI-E 扩展 config space.

3. BDF 最多有 2^{16} 个设备，每个设备有 2^{12} 大小的空间. 2. 总大小有 2^{28} BYTES. 为 256 MB.

Table 7-1: Enhanced Configuration Address Mapping

Memory Address[70]	PCI Express Configuration Space
A[(20 + *n* – 1):20]	Bus Number 1 ≤ *n* ≤ 8
A[19:15]	Device Number
A[14:12]	Function Number
A[11:8]	Extended Register Number
A[7:2]	Register Number
A[1:0]	Along with size of the access, used to generate Byte Enables

memory 映射 config 空间

1. A[63:(20+n)] 指定 memory base address

2. 其他域如方图所示

3. 对于 posted-write (通常为 memory write), software 可以通过 read 同一个地址来保证写结果

4. Configure write 和 memory Read 可能重排序. 所以需要有一种机制来保证先用 config write 配置 BASE BAR memory addr. 然后去 Read BAR 范围地址 ← 重排序导致读取到内容

和在 vendor-specific 中的寄存器当做用来在 memory-map allocate 前需要的空间配置.

1. PCI-E 设备的 up-stream port. 只能有一个 Device number, 但可有8个 function number.

2. 从 Downstream & root port 连下的 bus 另外一侧的 Device-number 为0 (除了PCI)

3. Downstream ports 可 assign 任何 device.

4. configuration-request Address:

Bus + Dev + function + exteneded & normal register number.

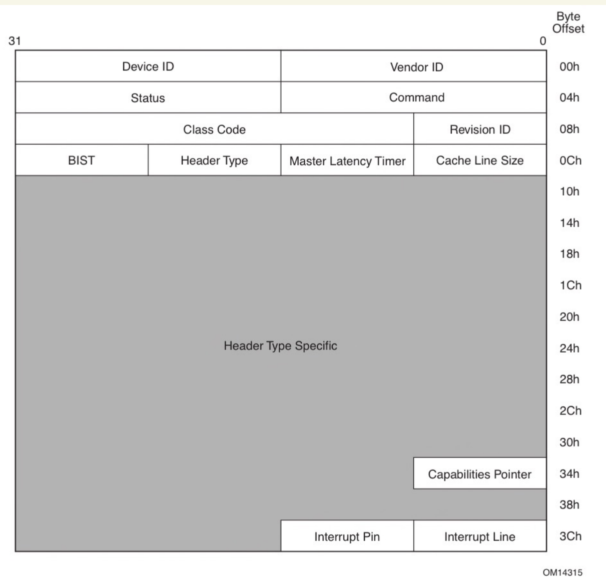


Figure 7-4: Common Configuration Space Header

← 在 config register 里不同域有不同的读写权限.

Table 7-2: Register and Register Bit-Field Types

Register Attribute	Description
HwInit	**Hardware Initialized** – Register bits are initialized by firmware or hardware mechanisms such as pin strapping or serial EEPROM. (System firmware hardware initialization is only allowed for system integrated devices.) Bits are read-only after initialization and can only be reset (for write-once by firmware) with Fundamental Reset (see Section 6.6.1). HwInit register bits are not modified by an FLR.
RO	**Read-only** – Register bits are read-only and cannot be altered by software. Register bits are permitted to be initialized by hardware mechanisms such as pin strapping or serial EEPROM.
RW	**Read-Write** – Register bits are read-write and are permitted to be either Set or Cleared by software to the desired state.
RW1C	**Write-1-to-clear status** – Register bits indicate status when read. A Set bit indicates a status event which is Cleared by writing a 1b. Writing a 0b to RW1C bits has no effect.
ROS	**Sticky - Read-only** – Register bits are read-only and cannot be altered by software. Bits are neither initialized nor modified by hot reset or FLR. Where noted, devices that consume AUX power must preserve sticky register values when AUX power consumption (via either AUX power or PME Enable) is enabled. In these cases, registers are neither initialized nor modified by hot, warm, or cold reset (see Section 6.6).
RWS	**Sticky - Read-Write** – Register bits are read-write and are Set or Cleared by software to the desired state. Bits are neither initialized nor modified by hot reset or FLR. Where noted, devices that consume AUX power must preserve sticky register values when AUX power consumption (via either AUX power or PME Enable) is enabled. In these cases, registers are neither initialized nor modified by hot, warm, or cold reset (see Section 6.6).

Register Attribute	Description
RW1CS	**Sticky - Write-1-to-clear status** – Registers indicate status when read. A Set bit indicates a status event which is Cleared by writing a 1b. Writing a 0b to RW1CS bits has no effect. Bits are neither initialized nor modified by hot reset or FLR. Where noted, devices that consume AUX power must preserve sticky register values when AUX power consumption (via either AUX power or PME Enable) is enabled. In these cases, registers are neither initialized nor modified by hot, warm, or cold reset (see Section 6.6).
RsvdP	**Reserved and Preserved** – Reserved for future RW implementations. Registers are read-only and must return zero when read. Software must preserve the value read for writes to bits.
RsvdZ	**Reserved and Zero** – Reserved for future RW1C implementations. Registers are read-only and must return zero when read. Software must use 0b for writes to bits.

0x04 中关键字域 (control)

1. Bus master. 是否可以发起 IO & mem 请求
2. parity error, 当奇校验错误报告
3. SERR enable, 系统错误报告
4. int disable, 是否关闭 Intx 中断
   ↑
   (通过 message code 传输)

0x06 关键字域 (status)

1. Int status — 是否有 pending 中断
2. parity error — 是否有 parity error
3. xxB 方式发出 error.

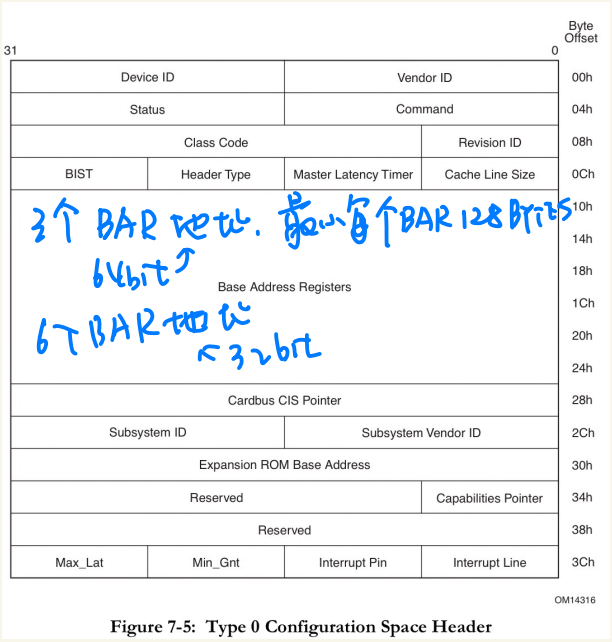


Figure 7-5: Type 0 Configuration Space Header

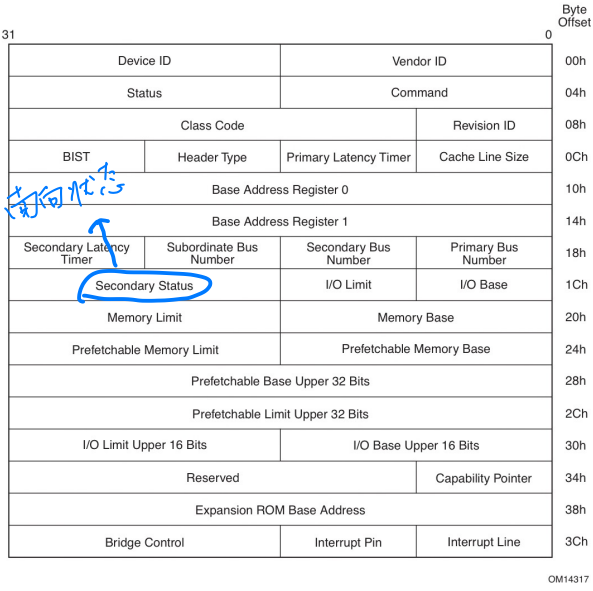


Figure 7-6: Type 1 Configuration Space Header

Capabilities

设备不支持 capability：就填 0.

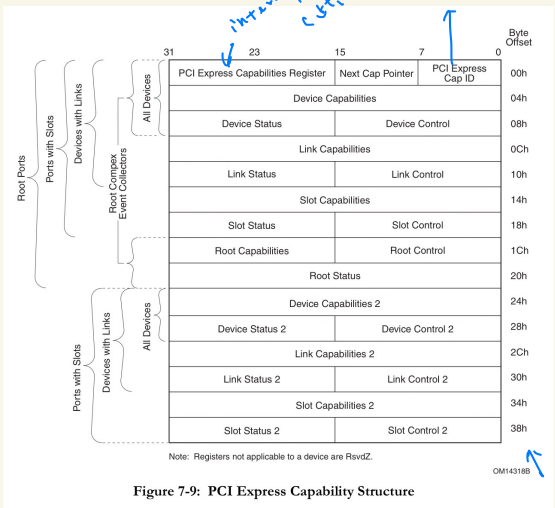


Figure 7-9: PCI Express Capability Structure

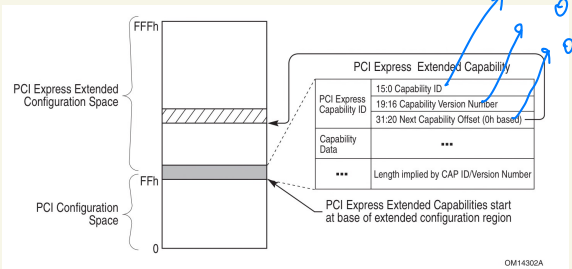


Figure 7-28: PCI Express Extended Configuration Space Layout

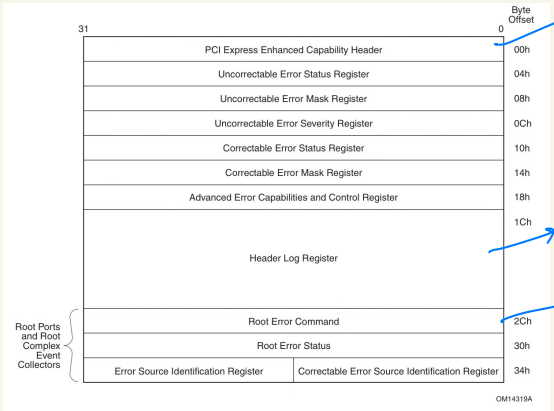


Figure 7-30: PCI Express Advanced Error Reporting Extended Capability Structure