



PCIe DMA IP设计方案

**版本说明**

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 版本 | 更新记录 | 编辑 |
| 2020-06-11 | V0.1 | 创建文档 | 郭巍 |
|  |  |  |  |
|  |  |  |  |

**目录**

[**版本说明** 2](#_Toc43117230)

[1. 项目简介 3](#_Toc43117231)

[2. F37X板卡规格介绍 4](#_Toc43117232)

[2.1 板卡概述 4](#_Toc43117233)

[2.2 参数规格 5](#_Toc43117234)

[3 DMA设计方案 7](#_Toc43117235)

[3.1 功能列表 7](#_Toc43117236)

[3.2 系统框图及模块划分 7](#_Toc43117237)

[3.3 dma\_client\_port模块设计 8](#_Toc43117238)

[3.4 dma\_psdpram模块设计 10](#_Toc43117239)

[3.5 dma\_if\_pcie\_us\_wr模块设计 12](#_Toc43117240)

[3.6 DMA IP的参数化配置说明 16](#_Toc43117241)

[3.7 DMA IP的处理时延说明 17](#_Toc43117242)

[4 DMA设计的寄存器接口 18](#_Toc43117243)

# 项目简介

本项目为PCIe DMA IP的设计开发项目，主要是考虑在低时延要求下，设计实现Card-to-Host方向，面向流的数据传输到CPU的功能。该IP在保留完整的可变长帧的多通道传输功能的基础上，接近极限的传输带宽，尽量压缩传输时延。IP提供逻辑自定义的描述符管理机制，具备实用性和一定的灵活性。本项目在F37X加速卡实现并完成测试。

# F37X板卡规格介绍

## 板卡概述

FPGA即现场可编程门阵列，它是可编程器件的基础上进一步发展的产物，作为专用集成电路（ASIC）领域中的一种半定制电路而出现，既解决了定制电路的不足，又克服了原有可编程器件门电路数有限的缺点。

F37X是Inspur浪潮基于Xilinx最新的VU37P系列 FPGA技术开发的一款基于PCI Express 3.0总线的FPGA的异构计算加速卡，具有高性能、高密度、高带宽的特点，全球首款集成HBM2的FPGA加速卡。其主控芯片VU37P是基于Xilinx 16nm工艺的Virtex UltraScale+ FPGA芯片，设计为高密度的全高半长PCIe插卡，采用板载内存颗粒，最大支持三通道24G内存。此外，VU37P支持2个100Gb光口，可以实现数据直接从网络到板卡的处理，无需经过CPU，大大减低了传输延时。同时，VU37P既支持传统的Verilog、VHDL等硬件描述语言开发，同时又支持OpenCL高级语言开发，在软件生产力上也取得了质的飞跃。

VU37P提供28.1TOPS的INT8计算性能和460GB/s的超高数据带宽，适合于机器学习推理、视频转码、图像识别、语音识别、自然语言处理、基因组测序分析、NFV（网络功能虚拟化）、大数据分析查询等各类应用场景，实现高性能、高带宽、低延迟、低功耗的AI计算加速。

F37X卡外观正面如图1-1所示。

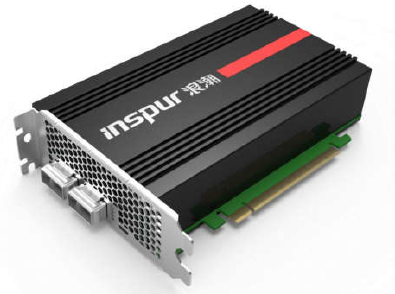


图1-1 F37X外观正面图

F37X系列加速卡产品主要功能如下：

* 硬件低耦合

F37X是标准的全高半长占用一个标准PCIe槽位，服务器适配性更强，无需额外机架空间，无需额外结构支持、无需额外散热支持、无需外接电源，对服务器的适应性强，安装简便，易于大规模部署。

* 高性能低功耗

相比于软件加速方案，F37X 的加速性能提升数倍到数百倍；F37X 的功耗仅为 75W(Average)，每瓦功耗获得的性能优势明显，使得算法应用性价比、性能功耗比相较传统服务器 CPU 有量级倍数的提升。

* 在线升级支持

F37X系列加速卡支持PCI Express在线动态重构加速算法，根据不同应用场景动态加载算法逻辑而无需重启服务器。用户可使用此应用灵活地加载更新不同的加速算法。

* 支持对第三方算法

Inspur浪潮提供免费的参考设计和参考IP，供学习和熟悉板卡应用。

## 参数规格

F37X加速卡参数规格如下表1-2所示：

表1-2 F37X参数规格表

|  |  |  |
| --- | --- | --- |
| 规格 | | 参数 |
| 产品 | | F37X |
| 物理参数 | FPGA | XCVU37P-L2FSVH2892 |
| PCIE | PCI Express Gen3 x16 |
| HBM DRAM | 8GB片上HBM2，提供460GB/s带宽 |
| DDR | 3 通道 72bits DDR4 SDRAM，最大支持24GB板载内存 |
| Ethernet接口 | 支持2路 QSFP28 100GE光口 |
| 外形（L\*H） | 167mm ⅹ 111mm |
| 电源参数 | 电源 | PCIE插槽12V@75W供电+外部Aux供电12V@75W |
| 功耗 | 150W(Peak),75W(Average) |
| 散热 | 支持双槽位被动散热 |
| 环境要求及标准 | 工作温度 | 0°C～35°C |
| 储藏温度 | -40°C～ +70°C |
| 湿度 | 93±3%RH |
| 认证标准 | 符合EMC国际电磁兼容性标准FCC Class A规范和CE规范 |
| 其他 | 板卡尺寸 | 采用PCIe标准全高半长规格：167mm（长）x 111mm（高） |
| 加载升级 | 支持通过Jtag对Flash中的FPGA版本进行在线升级  支持通过pcie对Flash中的FPGA版本进行在线升级 |

# DMA设计方案

## 3.1 功能列表

* PCIe硬核对外为PCIe Gen3 x 16接口，对内的RQ/RC的stream接口与DMA IP连接，默认为512位总线宽度
* DMA IP的用户侧接口为可配置的N路FIFO读接口，FIFO的数据位宽为1024位，N可在IP例化时配置
* 支持外部FIFO数据的PIPELINE周期可设置（代码中暂未支持），默认为1
* 支持内部分段式RAM的PIPELINE周期可设置，最短可设置为0（时延最小）
* 支持发送数据的长度可变，需在申请发送时就提供要发送的帧长
* 支持TLP帧的最大负载可配置，用户发送帧长的最大值只受描述符中长度限制，自动进行TLP分帧操作
* 支持Host主机分配多块RAM空间给PCIe RC设备进行C2H方向数据的接收
* 暂不支持TLP帧拼接（straddle）操作

## 3.2 系统框图及模块划分

该DMA IP的顶层模块名称为pcie\_dma\_c2h\_top，按照数据流的处理方向，依次分为三个主要模块。如图1的系统框图所示，从右到左分别为dma\_client\_port模块、dma\_psdpram模块和dma\_if\_pcie\_us\_wr模块。dma\_client\_port模块主要完成多个用户侧端口的选择仲裁，将选择出的端口数据送给写分段式RAM的控制逻辑，后者将其写入分段式双端口RAM。用户端口在申请发送带宽时，会提交要发送帧的帧长信息，描述符的生成电路会根据申请的帧长及软件配置的Host主机的目的存储空间地址和存储分段式RAM的地址，将和此帧相关的描述符信息写入dma\_if\_pcie\_us\_wr模块。dma\_psdpram模块是分段式双端口RAM的实现模块，每个RAM块数据端口为128比特，当RQ总线为512比特时，共8个块可拼接为1024比特的数据位宽，用于数据的读写。整个分段式双端口RAM的存储空间最小配置为1KB。dma\_if\_pcie\_us\_wr模块实现TLP帧的生成和发送功能，主要分成三个状态机实现。TLP帧的分帧状态机负责接收将要发送的数据帧的长度及地址信息，对于需要分帧的数据，依次产生每次发送的帧长信息，并存储在op\_table中，提供给另外两个状态机使用。TLP帧的读地址控制，依次产生读地址，将分段式RAM中的发送帧数据读出，同时TLP帧的发送控制状态机进行组帧并发送到RQ总线上。

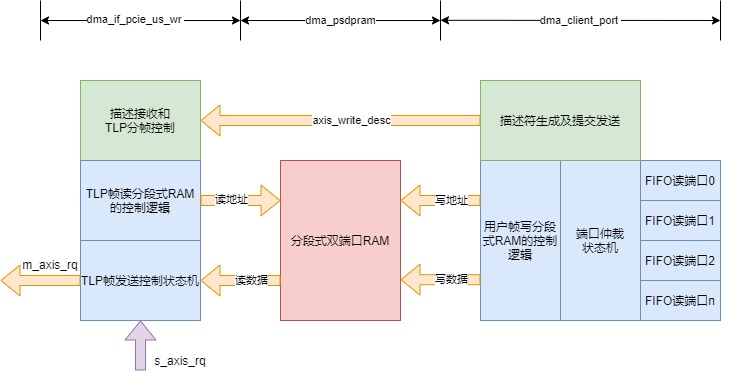


图1 系统框图

## 3.3 dma\_client\_port模块设计

* 功能描述

主要完成多个用户侧端口的选择仲裁，将选择出的端口数据送给写分段式RAM的控制逻辑，后者将其写入分段式双端口RAM。

* 接口描述

|  |  |  |  |
| --- | --- | --- | --- |
| 信号名称 | 方向 | 位宽 | 说明 |
| Clk | In | 1 | 时钟信号 |
| Rst | In | 1 | 复位信号，高复位 |
| 写描述符接口 | | | |
| m\_axis\_write\_desc\_pcie\_addr | Out | PCIE\_ADDR\_WIDTH | Host主机的内存空间地址 |
| m\_axis\_write\_desc\_ram\_addr | Out | RAM\_ADDR\_WIDTH | 分段式RAM的缓存空间地址 |
| m\_axis\_write\_desc\_len | Out | LEN\_WIDTH | 帧长信息 |
| m\_axis\_write\_desc\_tag | Out | TAG\_WIDTH | 标签信息 |
| m\_axis\_write\_desc\_valid | Out | 1 | 描述符有效标志 |
| m\_axis\_write\_desc\_ready | In | 1 | 描述符Ready信号 |
| 读多端口的帧数据接口 | | | |
| read\_frame\_enb | Out | PORTS | 帧数据读使能 |
| read\_frame\_tdata | In | PORTS\* FRAME\_DATA\_WIDTH | 帧数据 |
| read\_frame\_ready | In | PORTS | 帧数据的ready标志 |
| read\_frame\_len | In | PORTS \*LEN\_WIDTH | 帧长信息 |
| read\_frame\_tag | In | PORTS \*TAG\_WIDTH | Tag信息 |
| 分段式RAM的写数据接口 | | | |
| ram\_wr\_cmd\_be | Out | SEG\_COUNT\*  SEG\_BE\_WIDTH | RAM的字节使能标志 |
| ram\_wr\_cmd\_addr | Out | SEG\_COUNT\*  SEG\_ADDR\_WIDTH | RAM写地址 |
| ram\_wr\_cmd\_data | Out | SEG\_COUNT\*  SEG\_DATA\_WIDTH | RAM写数据 |
| ram\_wr\_cmd\_valid | Out | SEG\_COUNT | 每个RAM块的写有效标志 |
| ram\_wr\_cmd\_ready | In | SEG\_COUNT | 每个RAM块的写Ready标志 |
| 配置信息接口 | | | |
| enable\_port, | In | PORTS | 用户侧端口的使能信号 |
| enable\_tlp\_ram, | In | 3 | Host侧缓存空间的使能数量，1-4个可配 |
| pcie\_ram\_base\_addr, | In | 4\*PCIE\_ADDR\_WIDTH | 每个Host侧缓存空间的基地址 |
| pcie\_buf\_size\_kb, | In | 16 | 每个缓存空间中缓存Buf的大小，KB为单位 |
| pcie\_buf\_cnt\_max | In | 16 | 每个缓存空间中缓存Buf的数量，pcie\_buf\_size\_kb \* pcie\_buf\_cnt\_max等于每个缓存空间的大小 |

* 关键控制逻辑描述

图2展示了控制多端口数据调度的状态转移关系。状态机复位后进入STATE\_IDLE0状态。复位结束后，如果有数据请求，则进入STATE\_ARBIT状态。在STATE\_ARBIT下，根据帧长生成读使能的周期数，以及描述需要的Host内存地址和分段式RAM的地址，生成写描述符的控制信号，并且在m\_axis\_write\_desc\_ready有效的情况下，判断转移条件完成状态转移。在STATE\_READ状态下，根据读使能周期数产生读使能信号，并在读完成后生成确认信号，在确认信号下，如果请求信号有效，就返回STATE\_ARBIT状态，否则转移到STATE\_IDLE0状态。

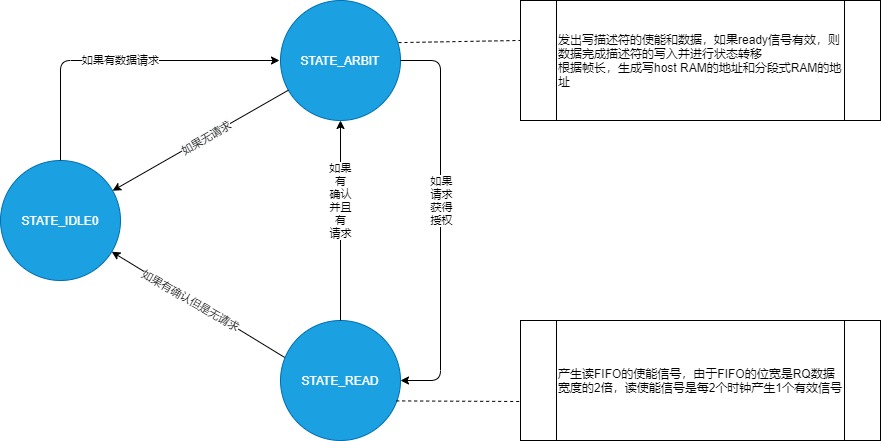


图2 多端口调度的状态机

## 3.4 dma\_psdpram模块设计

* 功能描述

分段式双端口RAM的实现模块，默认情况是PIPELINE=2。在此设计中增加了PIPELINE=0的情况，当总的缓存资源为1KB时，FPGA综合时会生成分布式RAM。在本设计中，分布式RAM的个数为8。

* 接口描述

|  |  |  |  |
| --- | --- | --- | --- |
| 信号名称 | 方向 | 位宽 | 说明 |
| 分段式RAM的写数据接口 | | | |
| Clk\_wr | In | 1 | 时钟信号 |
| Rst\_wr | In | 1 | 复位信号 |
| ram\_wr\_cmd\_be | In | SEG\_COUNT\*  SEG\_BE\_WIDTH | RAM的字节使能标志 |
| ram\_wr\_cmd\_addr | In | SEG\_COUNT\*  SEG\_ADDR\_WIDTH | RAM写地址 |
| ram\_wr\_cmd\_data | In | SEG\_COUNT\*  SEG\_DATA\_WIDTH | RAM写数据 |
| ram\_wr\_cmd\_valid | In | SEG\_COUNT | 每个RAM块的写有效标志 |
| ram\_wr\_cmd\_ready | Out | SEG\_COUNT | 每个RAM块的写Ready标志 |
| 分段式RAM的读数据接口 | | | |
| Clk\_rd | In | 1 | 时钟信号 |
| Rst\_rd | In | 1 | 复位信号 |
| rd\_cmd\_addr | In | SEG\_COUNT\*  SEG\_ADDR\_WIDTH | RAM的读地址信号 |
| rd\_cmd\_valid | In | SEG\_COUNT | RAM的读命令信号 |
| rd\_cmd\_ready | Out | SEG\_COUNT | RAM的读命令Ready信号 |
| rd\_resp\_data | Out | SEG\_COUNT\*  SEG\_DATA\_WIDTH | RAM的读响应数据 |
| rd\_resp\_valid | Out | SEG\_COUNT | RAM的读响应有效信号 |
| rd\_resp\_ready | In | SEG\_COUNT | RAM的读响应Ready信号 |

* 关键控制逻辑描述

分段式双端口RAM的默认大小为4Kbyte，PIPELINE=2，综合器根据配置自动选择综合后使用的资源。

## 3.5 dma\_if\_pcie\_us\_wr模块设计

* 功能描述

dma\_if\_pcie\_us\_wr模块实现TLP帧的生成和发送功能，主要分成三个状态机实现。

* 接口描述

|  |  |  |  |
| --- | --- | --- | --- |
| 信号名称 | 方向 | 位宽 | 说明 |
| Clk | In | 1 | 时钟信号 |
| Rst | In | 1 | 时钟复位 |
| 写描述符接口 | | | |
| m\_axis\_write\_desc\_pcie\_addr | In | PCIE\_ADDR\_WIDTH | Host主机的内存空间地址 |
| m\_axis\_write\_desc\_ram\_addr | In | RAM\_ADDR\_WIDTH | 分段式RAM的缓存空间地址 |
| m\_axis\_write\_desc\_len | In | LEN\_WIDTH | 帧长信息 |
| m\_axis\_write\_desc\_tag | In | TAG\_WIDTH | 标签信息 |
| m\_axis\_write\_desc\_valid | In | 1 | 描述符有效标志 |
| m\_axis\_write\_desc\_ready | Out | 1 | 描述符Ready信号 |
| m\_axis\_write\_desc\_status\_tag | Out | TAG\_WIDTH | 描述符处理完成的TAG |
| m\_axis\_write\_desc\_status\_valid | Out | 1 | 描述符处理完成的有效标志 |
| 分段式RAM的读数据接口 | | | |
| rd\_cmd\_addr | Out | SEG\_COUNT\*  SEG\_ADDR\_WIDTH | RAM的读地址信号 |
| rd\_cmd\_valid | Out | SEG\_COUNT | RAM的读命令信号 |
| rd\_cmd\_ready | In | SEG\_COUNT | RAM的读命令Ready信号 |
| rd\_resp\_data | In | SEG\_COUNT\*  SEG\_DATA\_WIDTH | RAM的读响应数据 |
| rd\_resp\_valid | In | SEG\_COUNT | RAM的读响应有效信号 |
| rd\_resp\_ready | Out | SEG\_COUNT | RAM的读响应Ready信号 |
| 从AXIS\_RQ接口：连接TLP读指令的发送数据接口 | | | |
| s\_axis\_rq\_tdata | In | AXIS\_PCIE\_DATA\_WIDTH | RQ通道数据 |
| s\_axis\_rq\_tkeep | In | AXIS\_PCIE\_KEEP\_WIDTH | RQ通道KEEP |
| s\_axis\_rq\_tvalid | In | 1 | RQ通道有效标志 |
| s\_axis\_rq\_tready | Out | 1 | RQ通道Ready标志 |
| s\_axis\_rq\_tlast | In | 1 | RQ通道Last标志 |
| s\_axis\_rq\_tuser | In | AXIS\_PCIE\_RQ\_USER  \_WIDTH | RQ通道User信号 |
| 主AXIS\_RQ接口：连接PCIe硬核的RQ通道 | | | |
| m\_axis\_rq\_tdata | Out | AXIS\_PCIE\_DATA\_WIDTH | RQ通道数据 |
| m\_axis\_rq\_tkeep | Out | AXIS\_PCIE\_KEEP\_WIDTH | RQ通道KEEP |
| m\_axis\_rq\_tvalid | Out | 1 | RQ通道有效标志 |
| m\_axis\_rq\_tready | In | 1 | RQ通道Ready标志 |
| m\_axis\_rq\_tlast | Out | 1 | RQ通道Last标志 |
| m\_axis\_rq\_tuser | Out | AXIS\_PCIE\_RQ\_USER  \_WIDTH | RQ通道User信号 |
| 发送序列号的输入端口 | | | |
| s\_axis\_rq\_seq\_num\_0 | In | RQ\_SEQ\_NUM\_WIDTH | 序列号 |
| s\_axis\_rq\_seq\_num\_valid\_0 | In | 1 | 序列号有效标志 |
| s\_axis\_rq\_seq\_num\_1 | In | RQ\_SEQ\_NUM\_WIDTH | 序列号 |
| s\_axis\_rq\_seq\_num\_valid\_1 | In | 1 | 序列号有效标志 |
| 发送序列号的输出端口 | | | |
| m\_axis\_rq\_seq\_num\_0 | Out | RQ\_SEQ\_NUM\_WIDTH | 序列号 |
| m\_axis\_rq\_seq\_num\_valid\_0 | Out | 1 | 序列号有效标志 |
| m\_axis\_rq\_seq\_num\_1 | Out | RQ\_SEQ\_NUM\_WIDTH | 序列号 |
| m\_axis\_rq\_seq\_num\_valid\_1 | Out | 1 | 序列号有效标志 |
| 流量控制 | | | |
| pcie\_tx\_fc\_ph\_av | In | 8 | TLP帧头的信用额度 |
| pcie\_tx\_fc\_pd\_av | In | 12 | TLP负载的信用额度 |
| 配置端口 | | | |
| enable | In | 1 | 使能信号 |
| requester\_id | In | 16 | 请求ID |
| requester\_id\_enable | In | 1 | 请求ID的使能信号 |
| max\_payload\_size | In | 3 | TLP的最大负载指示 |

* 关键控制逻辑描述

如图3所示，TLP的分段状态机状态转移图。状态机初始在REQ\_STATE\_IDLE状态，当收到上游模块发送过来的描述符时，产生第一个要发送的TLP帧长度，然后转移到REQ\_STATE\_START状态，在此状态下继续生成下一个TLP帧长，待发送的TLP帧长被写入op\_table中，供TLP发送状态机使用。当前发送的数据帧都被分割为合适的TLP帧后，状态机返回IDLE状态，再次接受新的发送描述符。

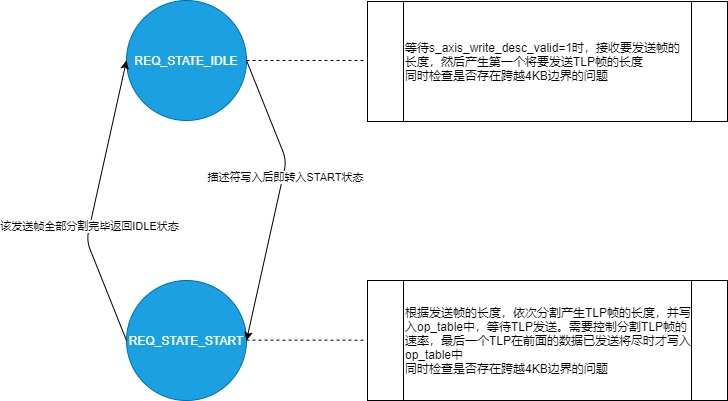


图3 TLP分段状态机

如图4所示，REQ状态完成一个TLP帧的分割，即触发读状态机的启动，在每个周期更新读地址的信息，如果没有TLP帧需要发送，则返回IDLE状态。

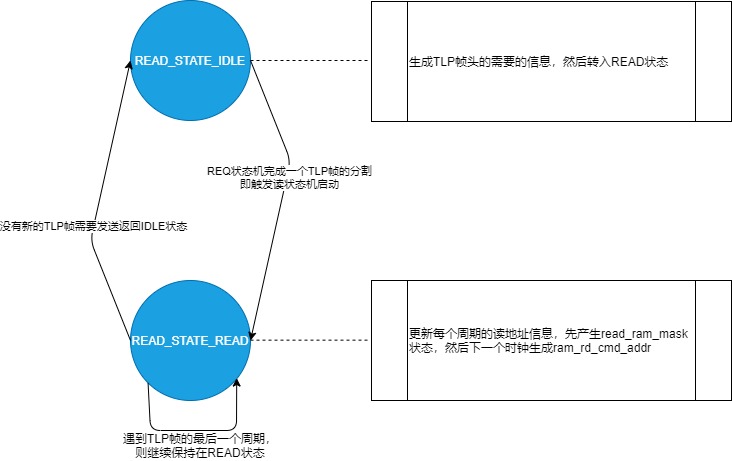


图4 读分段式RAM控制状态机

如图5所示，TLP的发送状态机在初始状态下处于IDLE状态，如果s\_axis\_rq通道有数据发送，则优先发送该通道数据，并转移到PASSTHROUGH状态运行，否则如果有本模块的TLP帧发送，则转移到HEADER\_1状态开始发送帧头。如果RQ通道使用128位及以上位宽的总线，则从HEADER\_1状态直接转移到TRANSFER进行后续数据的发送，否则还要经过HEADER\_2的状态转移。在TRANSFER状态，如果没有数据发送，则进入IDLE状态，否则直接接入HEADER\_1状态进行下一个帧头的发送。

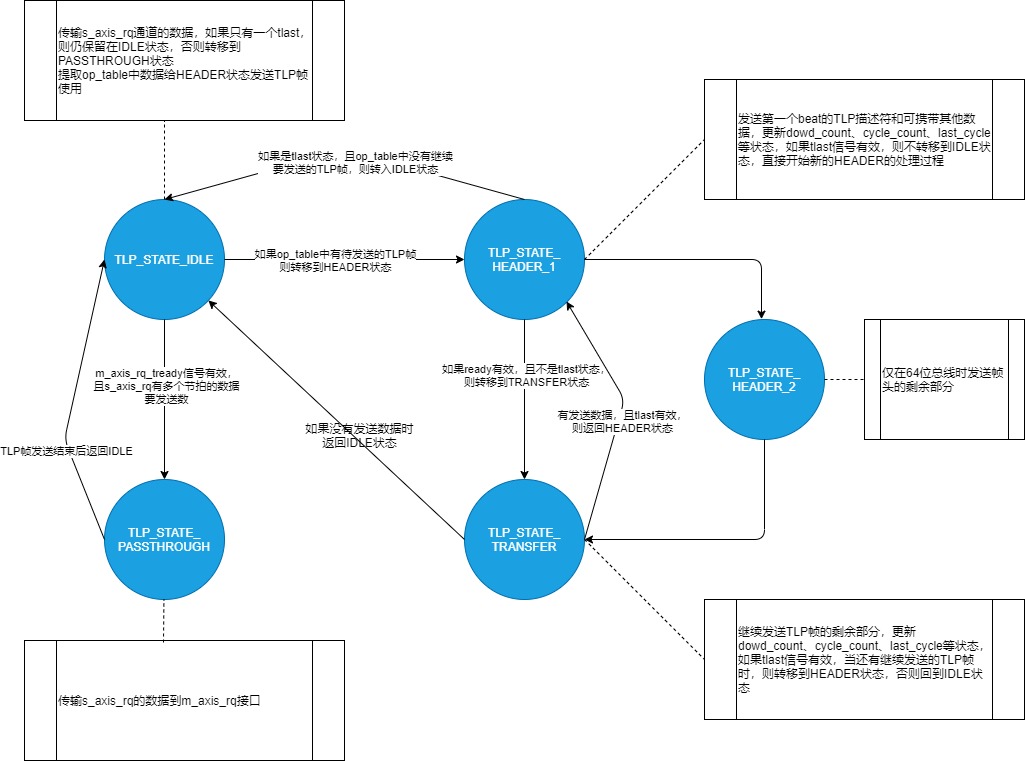


图5 TLP帧的发送状态机

## 3.6 DMA IP的参数化配置说明

|  |  |  |  |
| --- | --- | --- | --- |
| 参数名称 | 默认值 | 可选范围 | 注释 |
| AXIS\_PCIE\_DATA\_WIDTH | 512 | 64,128,  256,512 | RQ通道使用位宽，默认使用512位的最大位宽 |
| AXIS\_PCIE\_KEEP\_WIDTH | 16 | 2,4,8,16 | AXIS\_PCIE\_DATA\_WIDTH/32 |
| AXIS\_PCIE\_RQ\_USER\_WIDTH | 137 | 60,137 | RQ采用512位宽度时，使用137位的USER位宽，否则使用60位的位宽 |
| RQ\_SEQ\_NUM\_WIDTH | 6 | 4,6 | RQ采用512位宽度时，选用6，否则选4 |
| RQ\_SEQ\_NUM\_ENABLE | 0 | 0,1 | 序列号的使能位，默认为0 |
| SEG\_COUNT | 8 | 2,4,8 | RAM的分段数量，512位宽对应8段 |
| SEG\_DATA\_WIDTH | 128 | 64,128 | 每段RAM的位宽为128，如果RQ总线为64位，则采用64位 |
| SEG\_ADDR\_WIDTH | 8 | >=3 | 每段RAM的地址深度 |
| SEG\_BE\_WIDTH | 16 | 8,16 | 每段RAM的字节使能 |
| RAM\_SEL\_WIDTH | 1 | >=1 | 如果有多个分段RAM，则进行扩展 |
| RAM\_ADDR\_WIDTH | 15 |  | SEG\_ADDR\_WIDTH+$clog2(SEG\_COUNT)+  $clog2(SEG\_BE\_WIDTH) |
| PCIE\_ADDR\_WIDTH | 64 | 32,64 | Host主机内存空间的地址 |
| LEN\_WIDTH | 16 | [11,32] | 帧长信息的位宽 |
| TAG\_WIDTH | 8 | [4,16] | 标签信息的位宽 |
| OP\_TABLE\_SIZE | 32 |  | 2\*\*(RQ\_SEQ\_NUM\_WIDTH-1) |
| TX\_LIMIT | 32 |  | 2\*\*(RQ\_SEQ\_NUM\_WIDTH-1) |
| TX\_FC\_ENABLE | 0 | 0,1 | 默认关闭使能 |
| PORTS | 4 | >=1 | 用户侧端口的数量 |
| FRAME\_DATA\_WIDTH | 1024 | 512,  1024 | 用户帧数据位宽，如RQ通道的数据位宽为512，则应选1024作为数据帧位宽 |
| FRAME\_PIPELINE | 1 | 1,2 | 数据帧的PIPELINE的选择，默认为1 |
|  |  |  |  |

## 3.7 DMA IP的处理时延说明

如图6所示，在FRAME\_PIPELINE=1，内部分段式RAM的PIPELINE=0的条件下，整个数据帧的处理时延从识别到帧ready信号到RQ数据输出为5个时钟周期。在图6所示的时隙1中，用户发送数据的端口给出Ready信号，本模块选择出将要发送的端口。在时隙2中，将描述符信息写入到下级模块中，同时发出读使能信号，从上级模块读取数据。时隙3中，REQ状态机进入工作状态，生成第一个要发送的TLP帧长，并存入op\_table中，此时从用户数据的FIFO中读出第一组数据。时隙4中，产生写入到分段RAM的数据，REQ状态机根据帧长信息继续进行TLP帧的分割，TLP发送状态机看到op\_table中有待发数据，则进入到HEADER\_1状态。时隙5中，读分段式RAM的地址分2步产生，先生成读数据的掩码，后生成每个RAM上地址。时隙6中，输出RAM地址，读出的数据同时有效，本周内采样，并准备好帧头数据。时隙7中，TLP发送状态机转入TRANSFER状态，并发出第一个RQ帧数据。

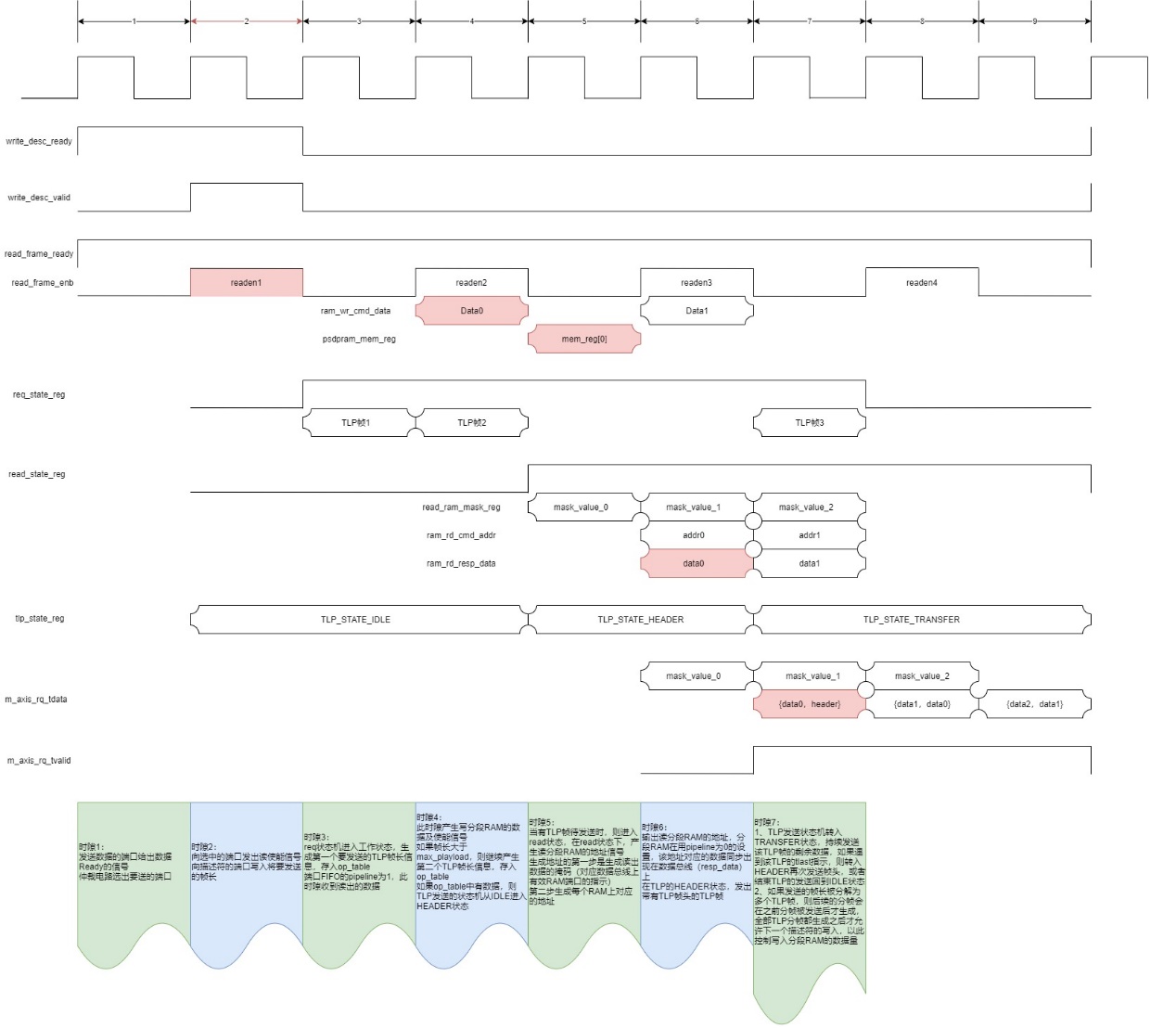


图6 处理时延的说明

# DMA设计的寄存器接口

无寄存器接口，需要配置的内容都在顶层端口上。