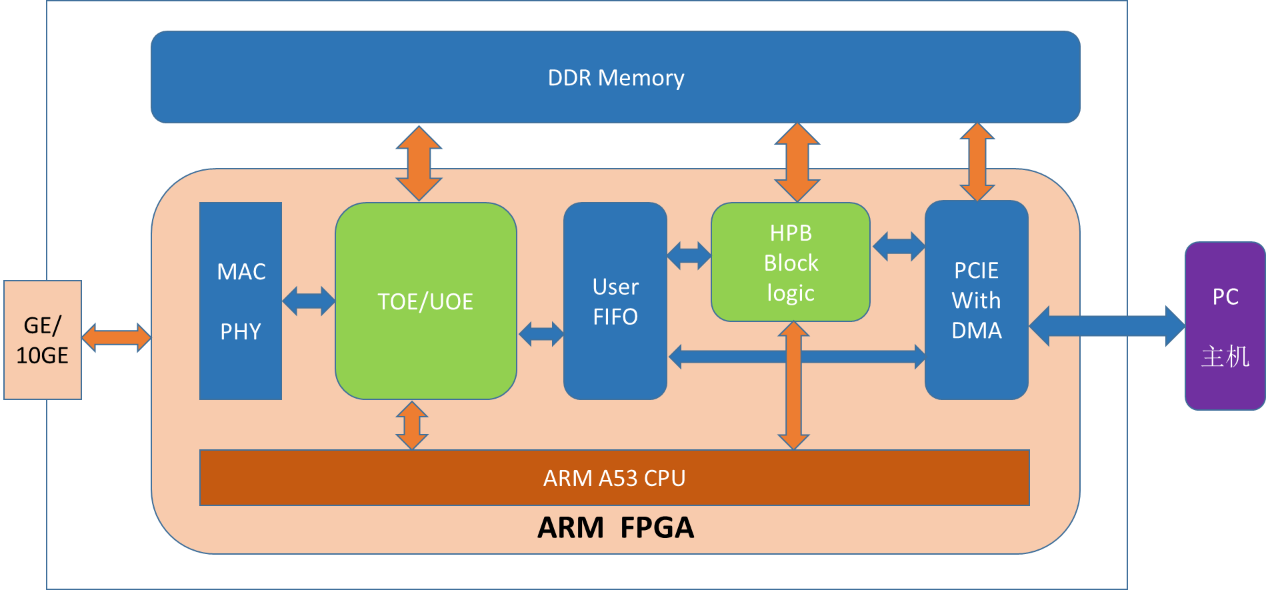
HPB（芯链）硬件加速引擎总体方案

本文对HPB硬件加速引擎的总体方案进行详细规划

HPB硬件加速引擎是一款基于ARM + FPGA架构的 FPGA定制化加速网卡，其在硬件TCP/IP 协议栈基础上，配合HPB加速单元，实现 HPB 体系架构下的网络流量卸载及高吞吐、高并发链接处理，通过修改调整缓存方式可以支持并发处理几百条流到几百万条TCP并发链接的维护处理。



HPB硬件加速引擎总体架构图

子模块描述

**该系统由五个大模块构成**

1. **MAC + PHY模块  
     
   这是**FPGA厂商提供的标准IP，支持1G/10G 自适应模式，可选 32 位低时延 10 G 以太网 MAC 或 64 位以太网 MAC，支持 10G 数据速率

* 选择 PHY 层的外部 XGMII 或内部 FPGA 接口
* 在客户端发送及接收接口上支持 AXI4-Stream 协议
* 支持缺损空闲计数以实现最大数据吞吐量；在各种条件下保持最小 IFG 并提供线路速率性能
* 针对所有设备支持包含带内 FCS 和不含带内 FCS 的缺损空闲计数
* 全面的统计收集
* 支持双向 802.3 和 802.1Qbb（基于优先级）流量控制
* 提供 MDIO STA 主接口以管理 PHY 层
* 支持 VLAN、巨型帧和 WAN 模式
* 定制前导模式
* 独立 TX 及 RX 最大传输单元 (MTU) 帧长度
* 可任意定制；针对功能性的行业资源使用

详细模块信息，参考FPGA厂商提供的用户手册

**二、TOE模块**

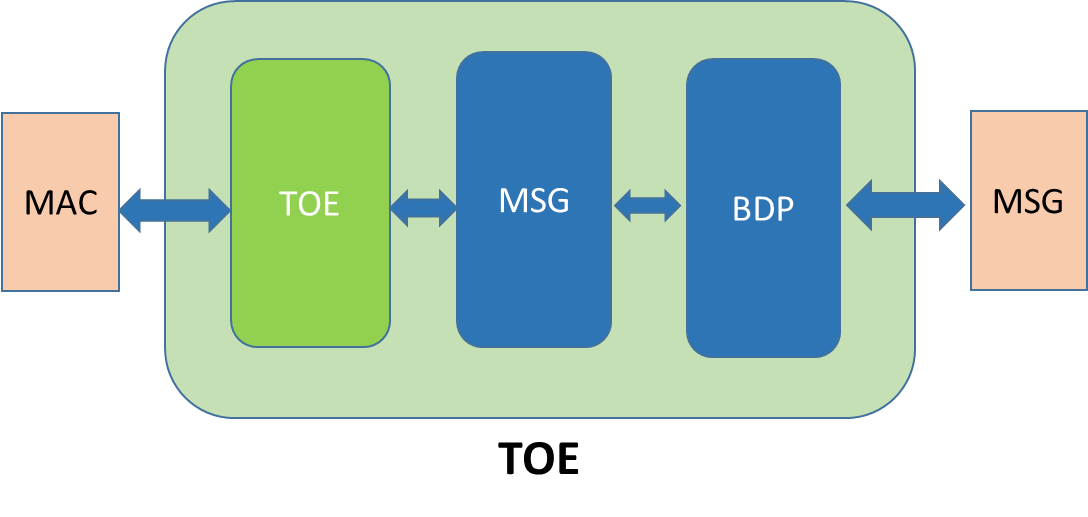
该模块为适应HPB架构的硬件TCP/IP协议栈模块，通过片上CPU及上位机PC管理，代替CPU执行网络协议打包的方案,使用特定的硬件电路来完成以太网包头的处理。 本模块的具体功能如下:

1.网络报文解析与分发，不属于本HPB加速引擎处理的报文送交CPU处理，会话报文组装发送，属于加速报文，交由硬件加速处理单元处理

2.HPB节点间通讯会话表维护

3.会话报文可靠性处理，报文错误重传，报文乱序调整等网络可靠性处理。

暂时不支持IP分片报文



TOE模块分三个子模块实现

1. TOE模块，报文发送与接收处理模块，上行方向，TOE模块主要实现网络报文提取，提取报文的核心五元组信息用于后续处理，报文分析同时将报文分块缓存至Memory中

下行方向，TOE模块根据处理结果对报文进行组装发送，发送时，接收从MSG模块来的报文描述符，提取发送缓存Memory中的数据，组装报文发送出去。

子模块框图如下：



TOE 模块子模块列表

|  |  |
| --- | --- |
| 模块名称 | 功能描述 |
| RX\_MEM | 接收报文缓冲区域，将接收到的报文存入缓冲区域，同时将关键信息送入队列管理，TCP报文最关键的信息位于报文头Cell中，核心处理均基于报文的第一个Cell进行处理，RX\_MEM将报文进行cell粒度的存储，提高ram的利用率，链表方式进行管理. |
| PKT\_iQue | 报文入队列处理模块，提取各个通路的报文的第一个cell，并对Cell队列进行管理 |
| PKT\_AN | Packet Analysis,报文解析模块，用于对提取的报文Cell进行解析，识别以太网报文的各个字段，构造描述符,以为后续模块使用 |
| TCP\_RX | Tcp报文入口方向处理，判断tcp报文是否乱序,进行调序等处理 |
| PKT\_EDIT | 报文编辑，对需要进行处理的报文根据解析结果进行处理，组合成合适的报文，或者MSG组，交送后续模块进行处理 |
| TCP\_TAB/TCP\_POLL | TCP序列表，TCP\_TAB是一个Memory,支持多个读写接口,记录TCP链接的序列表，与TCP\_POLL共同构成接收与发送TCP会话表，TCP\_POLL实现地址递增，轮询表项，发现内容过期，发出表项删除命令 |
| PKT\_oQue | 发送报文组装模块，从TX\_MEM（发送缓存区域）提取报文Cell，组合成标准以太网报文，交送报文编辑模块进行报文编辑 |
| TCP\_TX | Tcp报文出口方向处理，Tcp报文输出方向处理，读取对应会话的信息，得到输出报文的封装头信息 |
| MSG\_TA | 发送消息处理，接收来自MSG模块的发送方向的消息，进行消息解析与预处理。 |

TOE关键寄存器定义

|  |  |
| --- | --- |
| 寄存器名 | 含义 |
| cfg\_toe\_mode\_loop | 环回模式模式寄存器 |
| cfg\_toe\_mode\_reorder | 是否使能乱序缓存 |
| cfg\_toe\_mode\_ins\_err | 是否插入间隔的错误 |
| cfg\_toe\_mode\_sess\_ne | 会话不存在时是否丢弃该报文 |
| cfg\_toe\_mode\_dupack\_en | 是否使能dupack发送 |
| cfg\_toe\_mode\_ins\_err1 | 是否插入连续错误 |
| cfg\_toe\_mode\_ins\_bp | 是否插入反压模式 |
| cfg\_toe\_mode\_cell\_gap | 包间隔控制 |

TOE子模块核心表项定义

Tcp\_table\_item[ TAB\_DWID\*1+127 +: 001 ] 表项有效位

Tcp\_table\_item[ TAB\_DWID\*1+126 +: 001 ] 收到ack报文的标记

Tcp\_table\_item[ TAB\_DWID\*1+125 +: 001 ] 收到fin报文的标记

Tcp\_table\_item[ TAB\_DWID\*1+124 +: 001 ] 收到rst报文的标记

Tcp\_table\_item[ TAB\_DWID\*1+120 +: 004 ] 记录收到的win scale，控制发送速率

Tcp\_table\_item[ TAB\_DWID\*1+112 +: 008 ] 统计收到的乱序报文个数

Tcp\_table\_item[ TAB\_DWID\*1+096 +: 016 ] 接收到有效数据的时间戳

Tcp\_table\_item[ TAB\_DWID\*1+064 +: 032 ] 接收到报文的源IP

Tcp\_table\_item[ TAB\_DWID\*1+048 +: 016 ] 接收到报文的源PORT

Tcp\_table\_item[ TAB\_DWID\*1+000 +: 048 ] 接收到报文的源MAC

Tcp\_table\_item[ TAB\_DWID\*0+096 +: 032 ] TCP会话的初始seqn

Tcp\_table\_item[ TAB\_DWID\*0+064 +: 032 ] 当期预期的seqn，判断是否乱序

Tcp\_table\_item[ TAB\_DWID\*0+032 +: 032 ] 收到的ackn$，用于释放发送方向的数据

Tcp\_table\_item[ TAB\_DWID\*0+016+:016] 收到的window，用于控制发送数据速率

Tcp\_table\_item[ TAB\_DWID\*0+012 +: 004 ] 发送的dupack报文统计，控制发送dupack的速率

Tcp\_table\_item[ TAB\_DWID\*0+008 +: 004 ] 接收的dupack报文统计，触发发送方向的数据快速重传

Tcp\_table\_item[ TAB\_DWID\*0+000 +: 008 ] 本地的接收window，准备发给对端的

1. MSG模块，报文TCP头信息处理模块，对输入的tcp按流记录，区分message的起始，申请message\_id等一系列流消息处理功能

完善中。

1. BDP模块，TCP/IP链接表维护，对链接表描述符进行扫描，更新，处理超时，处理重传，慢启动，拥塞避免等功能，所有操作均基于message描述符进行。

完善中。

**三、用户FIFO**

此模块是一个特殊的双端口FIFO，属于HPB加速处理报文，在经过TOE模块后，被送往该FIFO的加速缓存区，不属于HPB的其他普通网络报文，将通过PCIE DMA通道，直接上送上位机PC，交由PC及处理，该模块由一个通用的多端口FIFO与一部分空调逻辑共同构成。

**四、HPB Block Logic**

HPB 专用加速逻辑，用于HPB特定的通讯报文接受处理与特殊通讯报文发送处理，比如共识广播报文、block确认报文等HPB自定义区块通讯报文处理

**五、PCI with DMA**

该模块为FPGA厂商提供的支持SG-DMA特性的PCIE Hard IP，具体功能请参考FPGA厂商IP用户手册。

如上就是HPB硬件加速引擎的总体方案，下周我们会开源加速引擎芯片核心模块的代码。

官方网站：http://gxn.io/