高可靠增强传输设计说明书

1. 总体设计

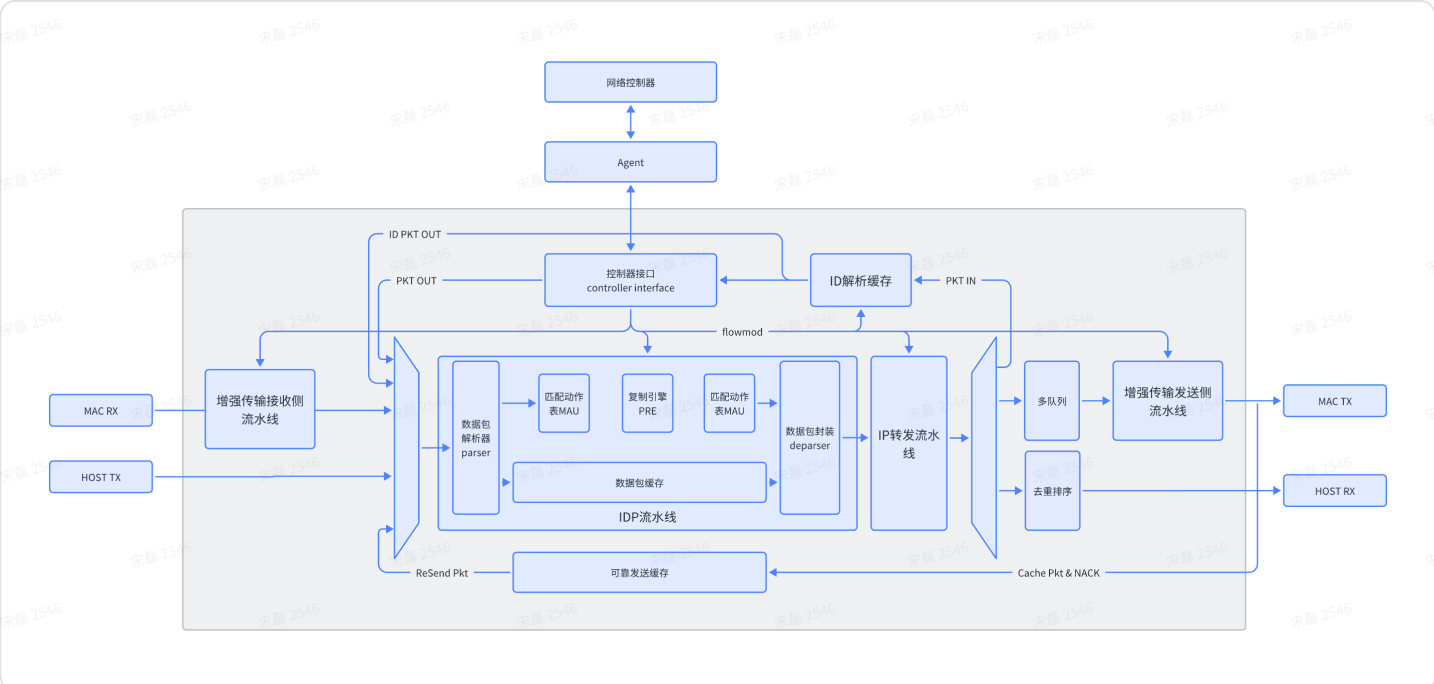
1.1 约定

a. 可靠发送缓存模块给增强传输发送流水线的ready信号始终为1，可靠发送缓存模块需自行检查数据包的完整性。

b. 可靠发送缓存需要流水线给他Reliability_Send MAU 的Index。对于发向下游的DAT报文，Reliability_Send MAU 用RSIP和DIP匹配，DAT报文同时发向MAC和可靠发送缓存模块；对于来自上游的NACK报文，Reliability_Send MAU用DstIP和SrcIP匹配，此NACK报文只发向可靠发送缓存模块。 IDP的Parser需透传所有NACK包，其中发向下游的NACK报文跳转到FIB处理，Reliability_Send MAU不匹配处理此NACK报文；来自上游的NACK包设置Outport为MAC侧，Goto Reliability_Send MAU处理。

c. 高可靠传输流表由控制器根据拓扑规划主动提前下发。如果表项未下发，高可靠数据包匹配不上的话，默认动作应该是作为普通IDP包转发，而不是PKTIN。

1.2 模块位置



1.3 特殊要求

1.3.1 高可靠功能开关

1.4 指标讨论

主要指标为：

指标	数值	讨论
高可靠流数（主要指标）	SEARouter 512条流（DPSS-200 下为1024）	高可靠以相邻节点的IP地址对标识一条流，因此此流数应该与节点的相邻IP数相当，即与FIB表的下一跳地址数相当。 ~~对于接入级L3交换机，其FIB表规模一般在4-256K条。（交换机：华为CloudEngine S5735-L-V 1K IPv6 4K IPv4。交换芯片：盛科CTC5118 12K，楠菲微SF6524 256K，春熙SW6048 32K）~~ SEARouter场景下，若为underlay组网，相连的节点数量级在32-64；若overlay组网，经过拓扑规划，每个拓扑下支持最大32-64个出度，单节点支持8个拓扑，即最大512个下一跳。 在DPSS-200场景下，小节点最大吞吐2Gbps，每条流2Mbps，支持1024条流，即全部都可以高可靠。 在CDN直播项目中，小节点流数1024条，大节点流数2048条，均在此指标覆盖范围内。
吞吐（主要指标）	10Gbps	在SEARouter场景下，接入交换机上连10Gbps，可以满足流量均为高可靠流量的需求。 在DPSS-200场景下，SEANet流量最大为2Gbps，全部为高可靠流量，可满足需求。 受限于DDR本身读写能力。
每条流的缓存包数（次要指标）	最大128K（DPSS-200下为32K）	~~在本规格设计所遵循的《SEANet报文设计v0.73》中，NACK报文中“未确认序号数量”为8bit，即最多请求256个连续数据包。~~ 在DPSS-200小节点场景下，每个拨号最大100Mbps计算，1500字节数据包时为8.3Kpps，在指标覆盖范围内。 在SEARouter场景下，最坏情况是总吞吐10G口满载一条流，1500字节数据包时为800Kpps，以国内广域公网点到点RTT 100ms计算， RTT需存储80K个包。SEARouter下每条流最大存储包数是128K，覆盖了此需求。
数据包总缓存容量（次要指标）	4GB，共用（DPSS-200下为	在DPSS-200小节点场景下，总吞吐2Gbps，1500字节数据包时为170Kpps，以国内广域公网点到点RTT 100ms计算， RTT需存储17K个包，即25MB，在指标覆盖范围内。

	1GB，独占）	在SEARouter场景下，总吞吐10Gbps，1500字节数据包时为800Kpps，以国内广域公网点到点RTT 100ms计算， RTT需存储80K个包，即128MB，在指标覆盖范围内。

经过调研，国内点到点RTT最大值80ms-100ms比较保险，跨国的话150-200ms。

2. 增强传输接收侧流水线

2.1 PHV

PHV号	有效位	说明
B	8	PKT_PROPERTY
B	8	Inport
B	8	Outport
B	8	IPIndex
B	低2位	TableMask
B	8	PKTIN_TableIndex
B	8	SEANet传输层Offset
B	8	SEANet传输层Type
B	8	SEANet传输层Net Flag
H	16	PKT_Len
H	16	Session data
W	32	protocol
W	32	DstIP [31:0]
W	32	DstIP [63:32]
W	32	DstIP [95:64]

W	32	DstIP [127:96]
W	32	RSIP [31:0]
W	32	RSIP [63:32]
W	32	RSIP [95:64]
W	32	RSIP [127:96]
W	32	PKT_RPN
W	32	FLOW_RPN

2.2 外部接口及其边带信号

1. 输入边带信号

顺序（从低位开始）	位宽	说明
1	8	Inport

该边带信号 连接自 总体输入边带信号

2. 输出边带信号

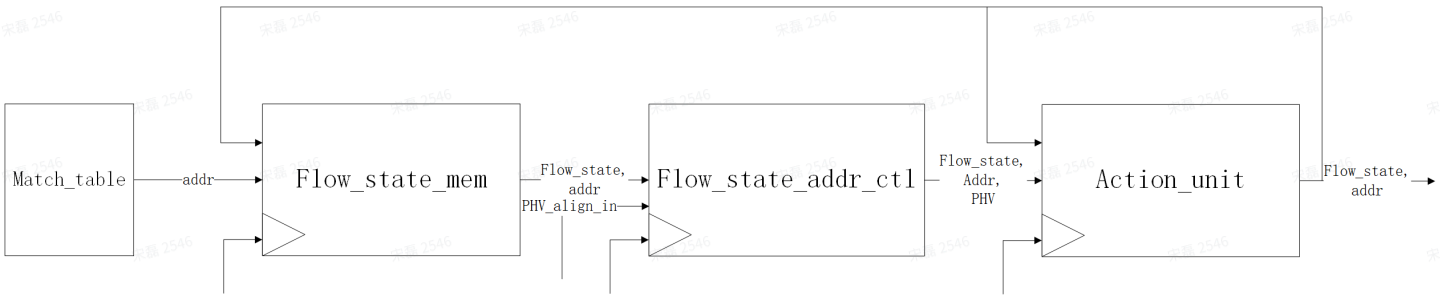
顺序（从低位开始）	位宽	说明
1	8	Inport
2	8	Outport
3	8	Tid（start table）
4	8	PKT_PROPERTY

该边带信号 连接到 mux模块输入边带信号

2.3 模块设计

利用了PHV中PKT_PROPERTY字段高两位（流水线内部私有状态）来触发是否发NACK。根据flow.rpn与pkt.rpn大小关系判断是否触发NACK，若需要触发NACK，则置pkt_property字段高2bit等于2'b11。

2.3.1 有状态的ActionUnit



数据一致性动作模块由三个子模块构成，分别为flowstate_mem、flowstate_addr_ctl、action_uint。其中三个子模块的功能为：

1) flowstate_mem：该模块输入为，匹配地址addr，更新后的流状态bcd_flowstate以及对应地址；输出为，根据输入地址查询得到的flowstate数据；

实施步骤为：根据上游模块的匹配得到的地址，从模块的Bram中读出flowstate数据。根据actionunit模块输出的地址信号以及修改后的flowstate，更新模块中存储的流状态信息。

2）flowstate_addr_ctl：该模块输入为phv，flowstate以及对应addr；输出为，phv，addr，flowstate以及对应addr信号，match_sel；

实施步骤为：该模块会进行输入flowstate信号与PHV信号的同步，同时在该模块中，会比较输入addr信号与最近三个addr信号是否相同，以检查由flowstate_ram中读出的flowstate数据是否为最新数据并产生m_phv_match_sel信号传递给action_unit模块。其工作原理为：若不存在相同地址，则flowstate本身为最新数据，若存在相同地址，则actionuint模块根据match_sel信号选出flowstate的最新结果。

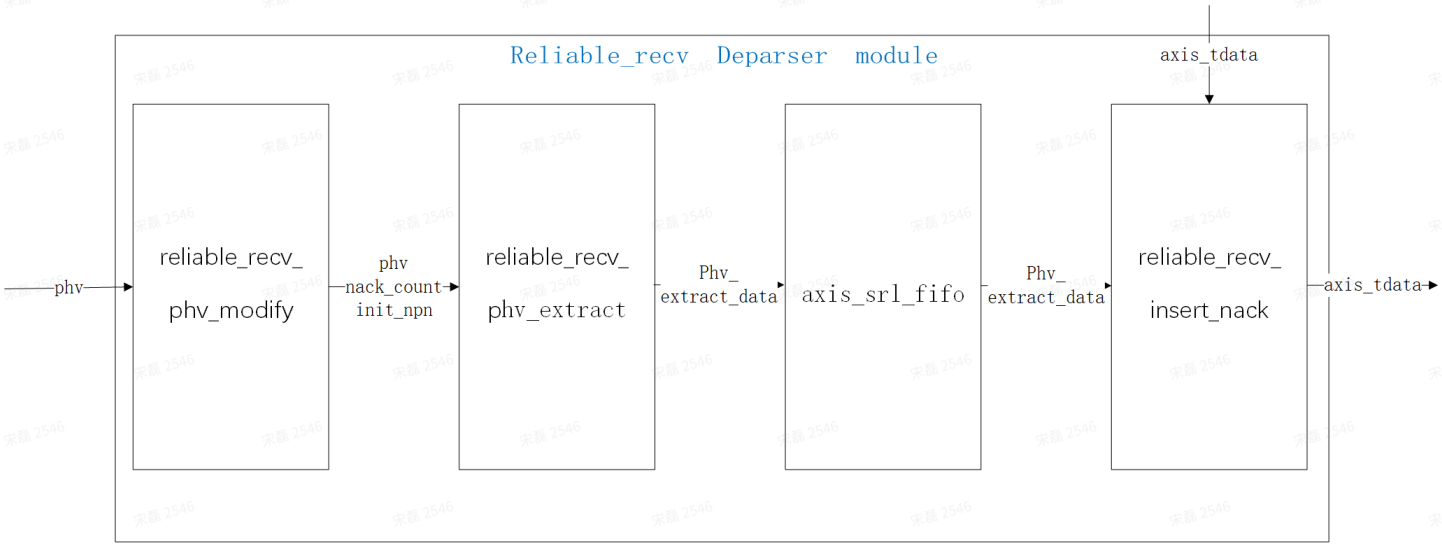
3）actionunit：该模块输入phv，flowstate，match_sel，addr信号；输出addr，以及修改后的phv、flowstate信号；

实施步骤为：该模块将缓存最新的三个flowstate结果，根据match_sel信号选择实际参与运算的输入数据为，三个缓存的最新flowstate结果，亦或者输入flowstate数据，以保证数据一致性。

2.3.2 Deparser（含NACK报文发生器）

在以往流水线deparser模块的基础上，额外增加功能:构造seanet_NACK报文并发送到axi-stream数据总线上，该NACK报文在可靠发送缓存包的phv中pkt_property字段高2bit为2‘b11时触发，并将在NACK字段填写缺失数量（npn_num）以及起始序号（initial_npn）。此外，其srcip设置为可靠发送缓存包的dstip，其dstip设置为可靠发送缓存包的rsip，并goto到fib表上。

为提升重传效率，可靠发送缓存触发NACK时，在deparser模块中先出NACK包，后出可靠发送DAT包。



NACK报文格式如下：

```
//dstmac,srcmac,0x86dd(ipv6),14bytes                                        ///eth
//0x6000,0x0000,0x0043(length),0x92(idp),0x3f,    8bytes                    ///ipv6
//srcip,dstip,              32bytes                                         ///ipv6
//0x93,0x34,50*(0x00),      52bytes                                         ///idp
//0x01(version),0x02(NACK),0x000f(length),0x0000(checksum),0x0000_0000(pn) ///public hdr
//0x??(npn_num), 0x????_???? (initial npn)  5bytes                          ///NACK

//nack.srcip=rel.dstip;nack.dstip=rel.rsip
//goto fib
//total length=22+32+52+15=121 bytes
```

3. 增强传输发送侧流水线

3.1 PHV

PHV号	有效位	说明

B	8	PKT_PROPERTY
B	8	Inport
B	8	Outport
B	8	IPIndex
B	低1位	TableMask
B	8	PKTIN_TableIndex
B	8	IP Offset
B	8	SEANet传输层Offset
B	8	SEANet传输层Type
B	8	SEANet传输层Net Flag
H	16	PKT_Len
H	16	Flow Index
W	32	protocol
W	32	DstIP [31:0]
W	32	DstIP [63:32]
W	32	DstIP [95:64]
W	32	DstIP [127:96]
W	32	RSIP [31:0]
W	32	RSIP [63:32]
W	32	RSIP [95:64]
W	32	RSIP [127:96]
W	32	SrcIP [31:0]
W	32	SrcIP [63:32]
W	32	SrcIP [95:64]

W	32	SrcIP [127:96]
W	32	PKT_RPN
W	32	InitialNPN

3.2 外部接口及其边带信号

输入边带信号

输出到MAC的边带信号

输出到可靠发送缓存的边带信号

顺序（从低位开始）	位宽	说明
1	8	Inport
2	8	Outport
3	8	IP Offset (Byte)
4	8	PKT_PROPERTY
5	16	PKT Length (Byte)
6	16	Flow Index
7	8	SEANet传输层头部Offset (Byte)

3.3 模块设计

3.3.1 有状态的ActionUnit

参考2.3.1的设计。

3.3.2 Deparser

标准的基于Edit的Deparser，通过Modify模块修改Pkt.RPN。

进出此流水线的数据包，outport的type均为2b'10，流水线不改outport type，只修改PKT_PROPERTY中可靠缓存是都命中的1bit。在Deparser之后需要有数据包分流逻辑，根据PKT_PROPERTY决定转发行为，具体如下表：

包类型	PKT_PROPERTY 高可靠DAT	PKT_PROPERTY NACK	PKT_PROPERTY 可靠缓存命中	转发行为
普通包	0	0	0	转发
高可靠DAT包，本节点命中	1	0	1	转发，同时复制发给可靠发送缓存模块
高可靠DAT包，本节点未命中	1	0	0	local:PKTIN（原为转发） 非local：转发
NACK包，本节点命中	0	1	1	只发给可靠发送缓存模块
NACK包，本节点未命中	0	1	0	local：丢弃 非local：转发

4. 可靠发送缓存

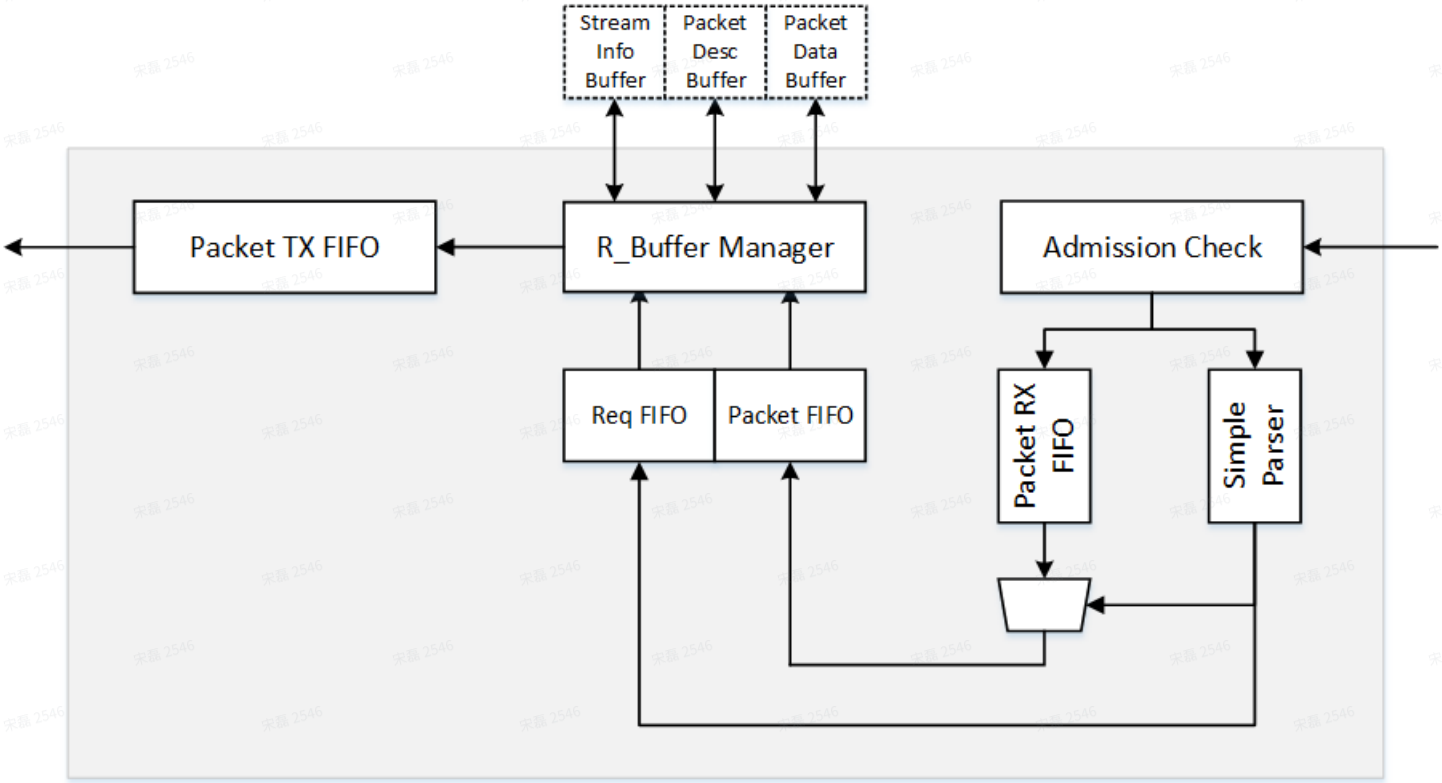
4.1 现阶段约定

1. 该模块不产生背压，当超过处理能力时直接丢包，为避免收取“半包”的情况，该模块还需做数据包准许进入校验。
2. 目前包描述符空间暂时为静态分配，支持的流数为2048，每条流存储的包数根据板卡的DDR规格配置：DDR总容量8G时，每条流128K个包；DDR总容量2G时，每条流64K个包；DDR总容量1G时，每条流32K个包。
3. 目前支持的NACK包NPN字段长度为定长4B，仅支持“第一NACK范围”。

4.2 模块概况

本模块是独立于流水线的，它接收需要缓存的可靠传输数据包（DAT）并进行缓存；同时接收重传请求（NACK）并将其指定的重传数据包从缓存中取出并发送。该模块与数据包处理流水线的输入输出交互，主要接口有：

- 输入：自流水线末尾，包括待缓存的可靠传输数据包以及下游节点发出的NACK重传请求；
- 输出：往流水线起点，输出数据为需要重传的数据包。

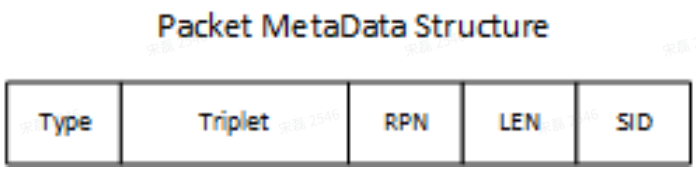


- 工作机制概述：数据包经过“准许进入”校验输入可靠发送缓存模块后，都将经过一个简单解析器模块处理，该解析器根据边带信号中附带的字段偏移信息从数据包中解析出本模块工作所需的数据包元数据，包括包类型（DAT或NACK）、包RPN、流ID（即匹配addr）、三元组（RSIP, DIP和协议类型）、包长字段。解析完成后，本模块随即根据解析出的数据包元数据信息生成相应的请求：
  - 对于DAT包的缓存请求，本模块根据请求中包含的流ID和RPN信息确定该数据包在可靠发送缓存中的存储位置并完成存储；
  - 对于NACK包的重传请求，本模块根据请求中包含的三元组和RPN信息判断重传数据包是否存在：
    - 若存在，则根据请求中包含的流ID和RPN信息确定待重传数据包在可靠发送缓存中的位置并取出发送；
    - 若不存在，则丢弃该请求。

4.3 主要模块设计

4.3.1 Simple Parser

RTB模块工作时需要的元数据包括包类型字段、流RPN、包RPN、流ID（匹配addr）、三元组（IP地址和传输层协议类型）、包长字段，包类型用于区分输入数据包为需要缓存的可靠传输数据包亦或是重传请求NACK包以确定后续的处理逻辑，三元组信息用以区分一条流，还需流ID、流RPN和包RPN信息用于判断数据包存储的位置、重传数据包的存储位置以及是否被覆盖，包长字段则需要储存在边带信号中输入给CPB模块用于数据包的实际存储。除了流RPN由RTB模块自己维护以外，其余的元数据信息都由流水线处理逻辑在边带信号中通过偏移、长度的方式提前给定。因此，RTB模块在入口处应包含一个Simple Parser模块，该模块需要能够根据边带信号中给定的偏移从数据包中提取出后续处理逻辑所需的包类型（Type）、三元组（Triplet）、包RPN、包长字段（LEN）、流ID（SID）。由于解析的类型和字段是固定的，因此该模块逻辑较通用的parser更简单，该模块的输入为标准AXIS接口，输出则是带握手的数据包元数据结构，该数据结构如下图所示。

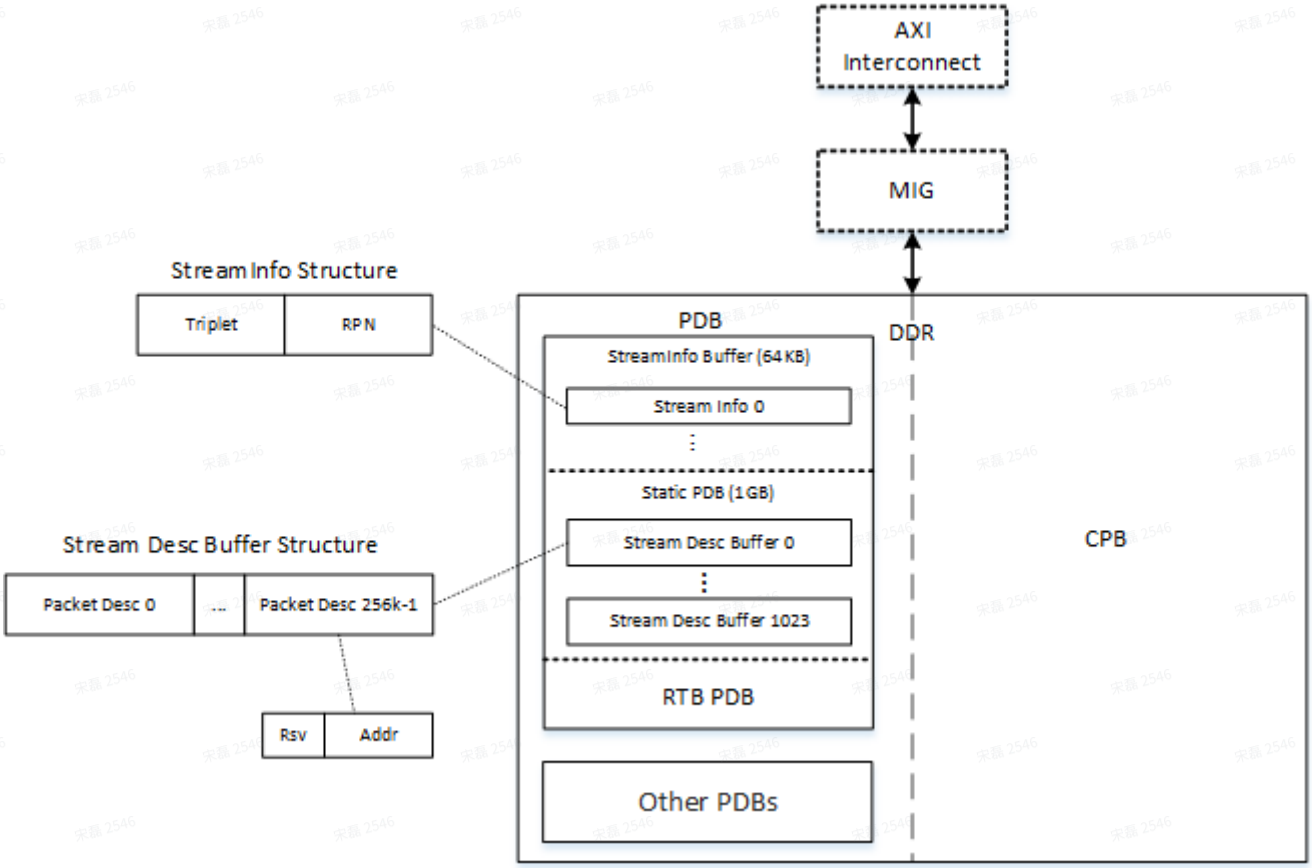


4.3.2 R_Buffer Manager

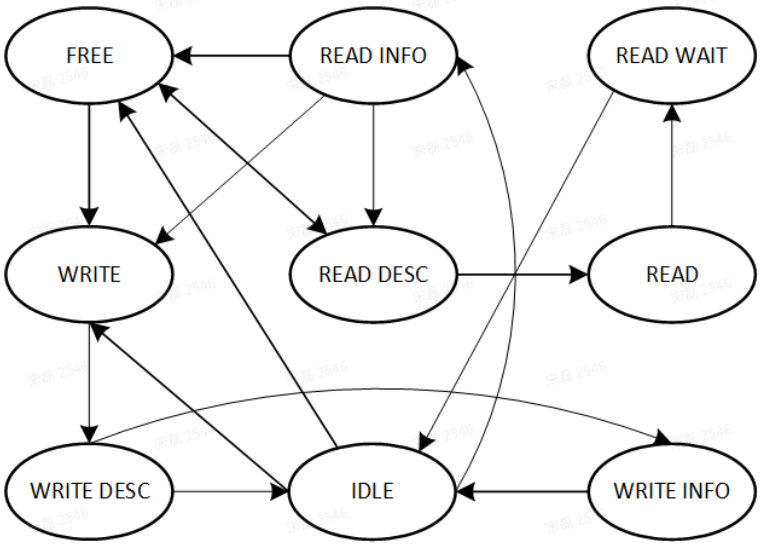
该模块与多个FIFO和基于DDR的StreamInfo Buffer (SIB) 、Packet Descriptor Buffer (PDB) 和Common Packet Buffer (CPB) 接口交互。该模块根据Request FIFO中的请求，向三个Buffer发起读、写请求，同时还需要验证重传包的正确性以及管理包描述符空间，该模块由一个主状态机、DDR读写模块和一个验证模块构成。

约定：由于本模块的特殊性，读取一个数据包时并不伴有释放需求，在数据包覆写时才需要释放，因此需要Common Packet Buffer提供一个释放Packet Buffer接口。这个需求暂时通过分离CPB模块中Read Engine中的读功能和释放功能实现。

- 描述符数据结构：包描述符由地址Addr和包大小Size构成，总位宽24bit（来源于CPB设计），Addr指向的是该数据包在Common Packet Buffer中存储的Block ID，包大小表示的是该数据包占用的Block数量。下图为存储空间分配示意图。



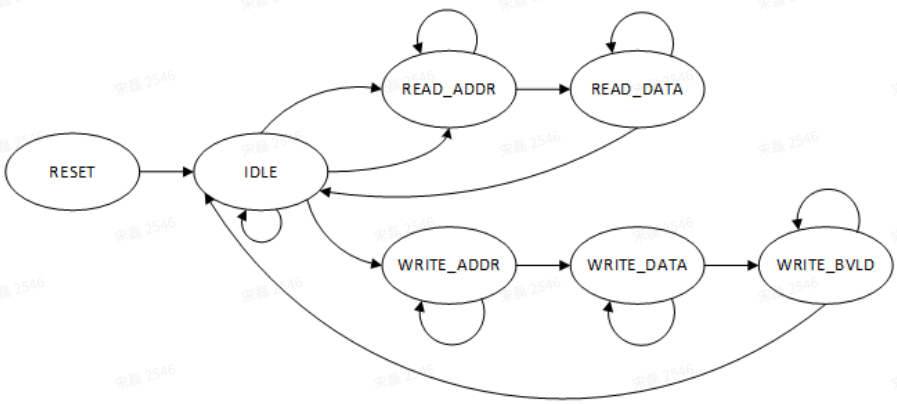
- R_Buffer Manager主状态机：使用有限状态机协调FIFO和DDR中的数据包及其描述符输入与输出；其状态跳转图如下图所示。



  - IDLE状态：检测Request FIFO，当该FIFO中有请求存在时，读取并解析一个请求，根据请求的类型和RPN决定状态的跳转；若为读请求，则下一状态跳转至READ INFO状态；若为写请求，RPN为零（新流）则下一状态跳转至READ INFO状态，RPN大于零小于PC（隶属于旧流的非覆写数据包）跳转至WRITE状态，RPN大于PC（隶属于旧流的覆写数据包）跳转至FREE状态；

- READ INFO状态：提取出请求中的buffer index和packet index信息，从StreamInfo Buffer中读出对应流的信息，包括三元组和RPN信息，输入判断模块，等待判断结果，对于读请求，若判断结果为存在，则跳转至READ DESC状态，否则则停止本次请求处理，跳转至IDLE状态；对于写请求，若判断结果为隶属于新流且写入新流空间的数据包，则跳转至WRITE状态，若判断结果为隶属于新流且写入旧流空间的数据包，则跳转至FREE状态；
- READ DESC状态：提取出请求中的buffer index和packet index信息，从Packet Descriptor Buffer中取出相应的数据包描述符，并取出Addr和Size信息，若Type为读，则下一状态跳转至READ，否则跳转FREE；
- READ状态：等待Common Packet Buffer读ready信号拉高，置位读valid信号，并输出从描述符中取出的Addr信息，下一状态跳转至READ_WAIT；
- READ WAIT状态：等待Common Packet Buffer将请求的数据返回，并将返回的数据输出至发送端Async Packet FIFO中；本次数据传输结束后跳转至IDLE状态；
- FREE状态：根据判断结果做不同的处理；若判断结果为隶属于旧流的覆写数据包，则代表只需要释放一个数据包，否则则需要释放该条流中已存储的所有旧数据包（RPNs>PC?PC:RPNs），完成释放操作后再跳转至WRITE状态；一次释放的流程为——先跳转至READ DESC状态，提取待释放数据包的描述符，再跳转回本状态，在本状态中等待Common Packet Buffer释放ready信号拉高，置位释放valid信号，并往对应的释放接口输入取得的包描述符信息，等待释放完成信号置位，重复本流程直到所有待释放数据包被释放；
- WRITE状态：等待Common Packet Buffer写ready信号拉高，置位写valid信号，并往写AXIS接口输出数据信息，期间符合AXIS数据传输时序，直到本次写数据完成且收到返回的地址信息；下一状态为WRITE DESC。
- WRITE DESC状态：将Common Packet Buffer返回的addr信息拼接成一个packet descriptor并根据写请求中的buffer index和packet index信息写回Packet Descriptor Buffer中的相应位置；下一状态为WRITE INFO状态；
- WRITE INFO状态：根据写请求中的buffer index将流信息（三元组和RPN）更新至StreamInfo Buffer中的对应位置，下一状态为IDLE。

- DDR读写模块：特别地，由于流信息、包描述符存储于DDR中，因此对于流信息和描述符的读写，另需要单独的DDR读写模块来执行，其FSM状态转移图如下：



- 判断模块：本模块根据输入的流信息（三元组和RPN）和包信息（IP和RPN），分别对读请求和写请求进行判断。对于读请求，，判断请求的重传数据包是否仍然存在于缓存中。首先判断该数据包隶属的流是否被关闭，判断条件为包三元组是否与流三元组一致，一致则进行下一步判断，否则丢弃该请求。第二步判断该数据包是否已被循环覆写，具体判断条件为包RPN（记为RPNp）是否落在以下范围：

$$\begin{cases} [1, RPN_s] & R = 0 \\ [R * PC + 1, RPN_s] \cup [RPN_s - R * PC + 1, R * PC] & R > 0 \end{cases}$$

其中，RPNs为数据包边带信号中携带的流表中存储的RPN，其代表着该路流存储的上一个数据包的RPN，R为循环覆写的次数，R=RPNs/PC。

对于写请求，需要判断该数据包是否隶属于一条新流（判断条件为包RPN是否为0），若是，则还需要判断是否存储于旧流缓存中（判断条件为流RPN是否不为0）。

4.4 外部接口及其边带信号

4.4.1 外部接口

名称	输入/输出	位宽	说明
clk	输入	1	时钟信号
rst	输入	1	重置信号
数据包：			
s_axis_rtb_tdata	输入	DATA_WIDTH	数据包输入
s_axis_rtb_tkeep	输入	KEEP_WIDTH	
s_axis_rtb_tvalid	输入	1	
s_axis_rtb_tready	输出	1	
s_axis_rtb_tlast	输入	1	
s_axis_rtb_tuser	输入	USER_WIDTH	
m_axis_rtb_tdata	输出	DATA_WIDTH	数据包输出
m_axis_rtb_tkeep	输出	KEEP_WIDTH	
m_axis_rtb_tvalid	输出	1	
m_axis_rtb_tready	输入	1	
m_axis_rtb_tlast	输出	1	
m_axis_rtb_tuser	输出	USER_WIDTH	
控制面：			
RSV			
Common Packet Buffer：			
m_axis_cpb_wr	I/O		写CPB通道，输出AXIS接口
s_desc_cpb_wr	I/O		写CPB通道，输入描述符接口
s_axis_cpb_rd	I/O		读CPB通道，输入AXIS接口
m_desc_cpb_rd	I/O		读CPB通道，输出描述符接口
m_desc_cpb_free	I/O		释放PB通道，输出描述符接口
s_avaliable_bd	输入	22	CPB剩余block数量
Packet Descriptor Buffer：			
m_axi_pdb	I/O		一组标准的DDR AXI接口
Stream Information Buffer：			
m_axi_sib	I/O		一组标准的DDR AXI接口

4.4.2 边带信号

顺序（从低位开始）	位宽	说明
1	8	Inport
2	8	Outport

3	8	NextTable（去Reliability_Send MAU）

5. 附录

5.1 分工

人员	任务	注意事项
宋磊	设计，测试例	
李逸飞	集成，测试例	
卢睿	高可靠传输接收侧流水线 和 高可靠传输发送侧流水线	（1）MAU中 RPN一致性问题 （2）简化的Parser （3）定制能力的Deparser：接收侧支持造NACK报文；发送侧支持同时发给可靠发送缓存
黄逍颖	可靠发送缓存模块	（1）校验数据包完整性

5.2 进展

模块	子模块	开发	仿真	集成仿真
接收侧流水线	Top	完成	未完成	
	Parser	完成	完成	
	Local MAU	完成	未完成	
	Reliable_recv MAU	完成	未完成	
	Deparser	完成（缺插入NACK，后补）	完成	
发送侧流水线	Top	完成	未完成	
	Parser	完成	完成	
	Reliable_send MAU	完成	完成	

	Deparser	完成	完成	
可靠发送缓存		完成	完成	