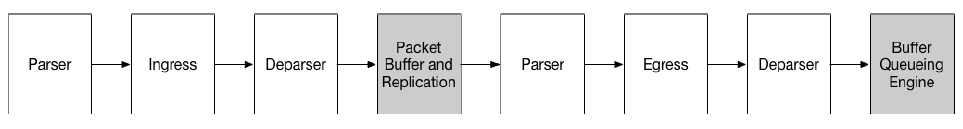
软件交换机实现文档

1. 实现结构

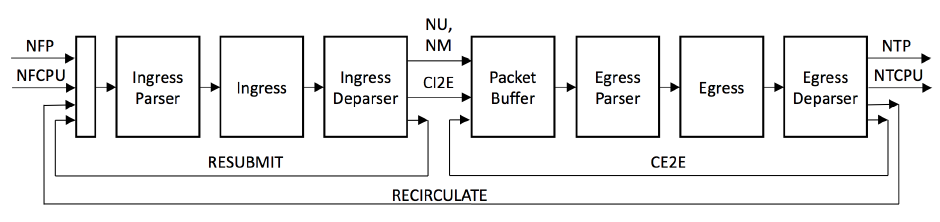
软件交换机参照PSA（Portable Switch Architecture）的结构进行实现。PSA的结构包括Ingress pipeline（parser、Ingress match-action、deparser）、Packet Replication Engine（PRE）、Egress pipeline（parser、Egress match-action、deparser）和Buffer Queueing Engine（BQE）。



需要注意的是，虽然P4Runtime提供了对parser的ValueSet，但是其只支持parser transition的modify，不支持insert和delete。所以需要将parser和deparser融入到match-action中，分模块来实现每个match-action功能。

1. Packet Path

首先需要了解一个packet在交换机中可能的路径，如下图所示。



1. 正常的从ingress pipeline到egress pipeline然后转发到输出端口；
2. Packet recirculation：recirculation是允许数据包重新进入数据包处理逻辑的技术。一个例子就是MPLS，如果一个MPLS数据包有mpls\_pop action，之后是IPv4协议，那么接下来就是要解析IPv4，但是在mpls\_pop之前是不可能知道接下来的协议是IPv4的。比如pop\_mpls(0x0800)，接续的action是将其发送到recirculate port，然后交换机将其发送到ingress中。“implementing features such as multiple levels of tunnel encapsulation or decapsulation”。
3. Packet resubmit：在一个数据包上重复进行ingress processing，其才开始不会进入deparser，而是按照原来的header和payload再次进入ingress进行处理，只不过其port变成了RESUBMIT。“deploy multiple packet processing algorithms on the same packet. For example, the original packet can be parsed and resubmitted in the first path with additional metadata to select one of the algorithms. Then, the resubmitted packet can be parsed, modified and deparsed using the selected algorithm”。
4. Packet Cloning：将数据包发送到指定的端口（除了正常流程的数据包之外）的机制，通过合适的数据平面配置可以用一个clone操作做多份clone。Clone的一个use case就是packet mirroring：除了将数据包发送到指定端口之外，还要讲packet复制一份发送到例如监控设备之类的输出端口。还有一个use case就是做arp的broadcast，其需要在ingress处clone，发送到应有的egress port中。一个是在ingress 的clone，其clone数据包是进入ingress parser的样子；一个是在egress的clone，其clone数据包是egress deparser之后的样子。
5. 运行时可重配置的软件交换机
6. 应模拟实现我们的TCAM/SRAM资源池，并模拟实现TCAM并行查找的机制以及SRAM的exact match的hash机制；要求：controller应掌握TCAM/SRAM资源池的分配情况（每种长度字段的表项数量、每片SRAM/TCAM中的资源分配情况）；表项应可以通过远程通信进行insert、modify、delete等操作；action部分应为primitive profile（group）的指针；
7. 统一exact、lpm、range、ternary四种match的查找方式，各种组合起来应该去什么地方查找；
8. 在软件实现中，多个模块应该是多个线程并行运行（或多个进程，多个进程的话需要实现进程间通信），模块之间传递数据包也应模拟出来；模拟实现我们的crossbar机制；
9. 将parser、deparser融合到match-action模块中，需实现模块的parser->matcher->executer的模板化设计；要求：实现模块的模板不能变，但是其参数需要变化，控制器可以通过远程通信来改变其值；多个模块之间可以共享的是解析出来的字段（线程通信、进程通信）；
10. 应有counter、meter这些字段，direct counter、direct meter是与entry相连的，但是indirect counter/meter不与其相关，控制器应能读取、修改这些值。