基于流处理器的完全可编程SDN芯片

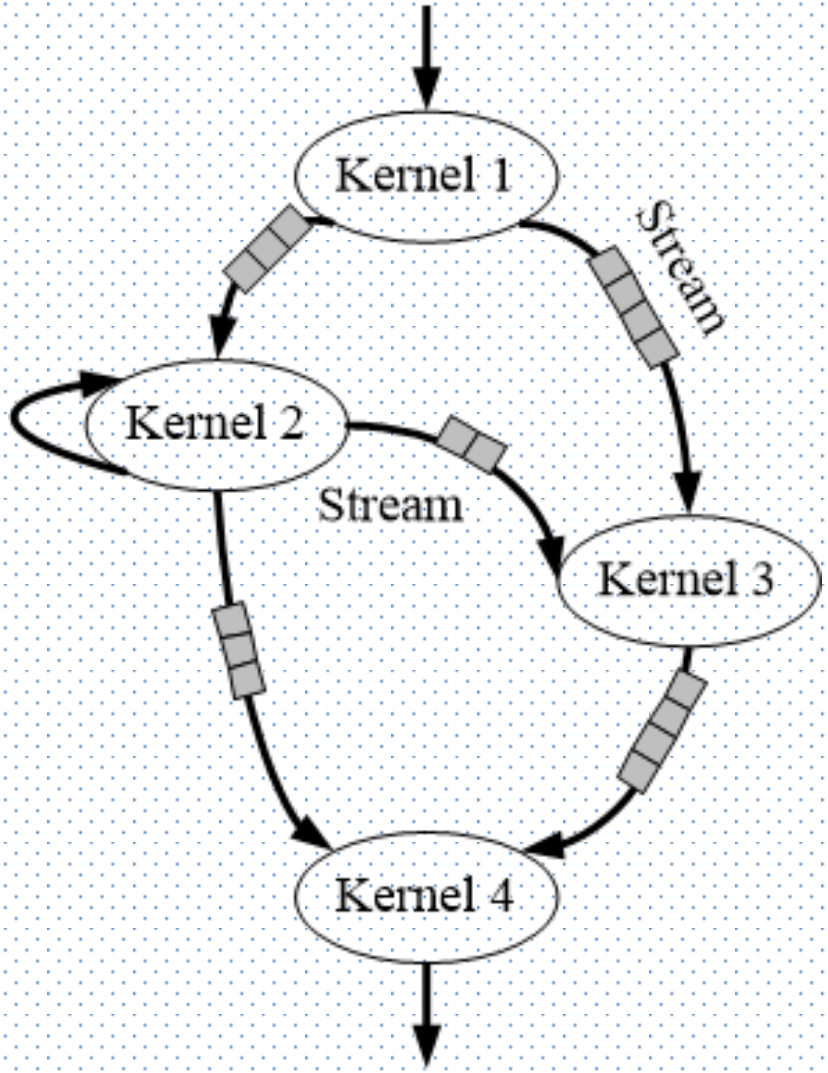
国防科技大学计算机学院
张春元
2016长沙

Agenda

- 众核流处理器内核-ET
- 基于流处理器的完全可编程SDN交换架构
- N-ET（Net-ET）SDN网络处理器

流编程模型

- 一种解耦合计算-通信的异步点火执行的程序模型
- 明确的生产者-消费者关系



我们的研究历史回顾

- 2003年开始，在image的影响下开展流处理器体系结构研究
- 第一代流处理器体系结构称为“多态自适应流体系结构MASA”，追求相当的适用性
- 获得多项NSFC和教育部博士点基金的资助
- 完成了32位系统的指令系统设计，用FPGA实现了3款原型（全集、算法定制、算法固化），均以图像和视频计算为背景
- 从模拟器、编译器、调试环境到性能分析环境，形成了比较完整的工具链
- 是FT64-2的核心技术
- 一堆论文，包括在MICRO杂志和ACM MM会议上
- 2009年出版了《流处理器研究与设计》，很多做加速器研究的研究人员的案头都有这本书

众核流处理器-ET

- 从2010年到2015年，启动ET体系结构研究
  - 该项目受到总装装备预研探索重大项目、国家自然科学基金重点项目、国家863计划项目主题项目、国家自然科学基金项目的支持
  - 具有完全自主知识产权的众核可编程处理器架构
  - 具有良好的应用适应性：面向图像视频流、网络处理器、深度学习处理应用提供处理器内核IP
- 2016年，获得国家重大专项“大数据与云计算处理系统”的支持
  - 同中科大、中科院计算所、华为、科大讯飞、银河风云合作
  - 以ET为基础构建其中的加速处理部件

众核流处理器-ET

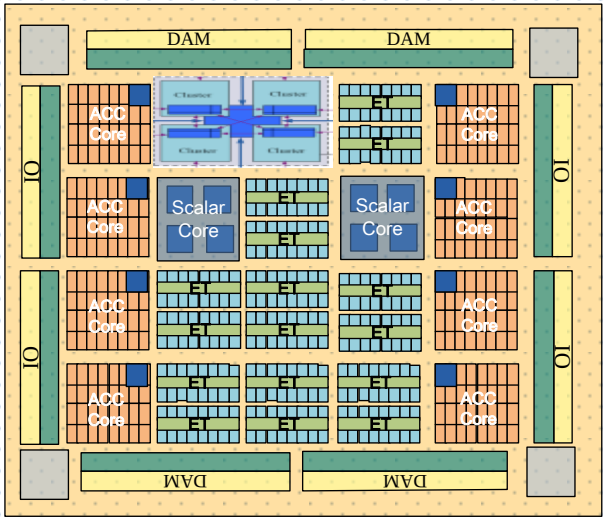
基于ET的验证原型

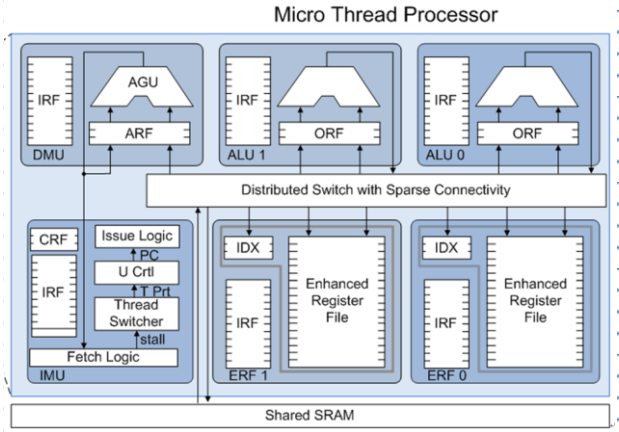
多ET的服务器原型

众核流处理器ET的技术特征

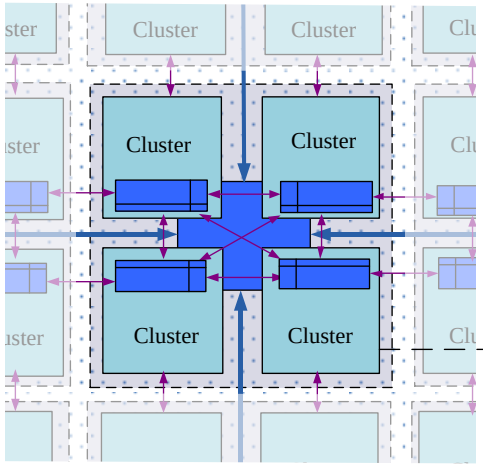
- 多级显示并行：面向应用可配置指令集，同时支持ILP、DLP、TLP模式，VLIW核心，单核心支持双线程
- 具有扩展到4K众核的能力，最大可提供万亿次计算性能，提供多核多线程保序功能
- 支持异构众核，提供对网络查找表搜索、图像处理、卷积神经网加速协处理引擎的无缝对接
- 多套片上集群网络，多通道硬件流预取缓冲、硬件管理的动态线程Dispatch和流调度，缓解存储和I/O性能压力
- 完全可编程，采用提供软件可编程管理的片上存储器和寄存器资源，提供软件可编程的协处理器通讯接口
- 简化核心，做到小巧、低功耗
- 提供微码级编程接口和Kernel C语言编译器、IDE

众核流处理器-ET

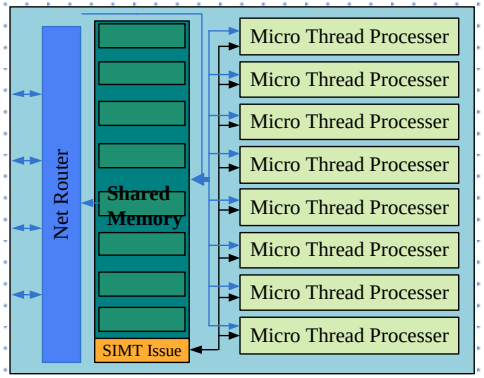




Quad-Grid Network



Cluster



ET体系结构

- 同时多粒度并行
- 三层组织形态
  - 格点（Grid）
  - 流计算簇（Stream Computing Cluster）
  - 微线程处理核心(Micro Thread Processor Core)
- 融合了GPU、通用CPU和DSP的节点组织优势，能够在多个维度开发程序的并行性
- 应用场景
  - 在1GHz下，大约每128Gbps网络流量的线速处理，需要1个格点
  - 4个格点可以满足512Gbps报文处理的需求

ET体系结构：流程序层所见

- 顶层格点是最基本执行节点
  - 由多个通用可编程流处理器内核集群或者协处理器寄存构成
  - 可以作为独立的模块单位
  - 根据芯片规模，在芯片上实现1个或多个格点
- 中间层流计算簇
  - 支持多线程多模式并行流式计算
  - 含有多个微线程处理器，共享流缓存、NoC接口、Local Scratchpad等资源,
  - 计算簇阵列等组件在不同处理核中可以是异构的
- 底层的微线程处理核心是最基础的执行节点
  - 由多功能算术部件、多模式局部寄存器、指令部件、数据加载部件和互连组成

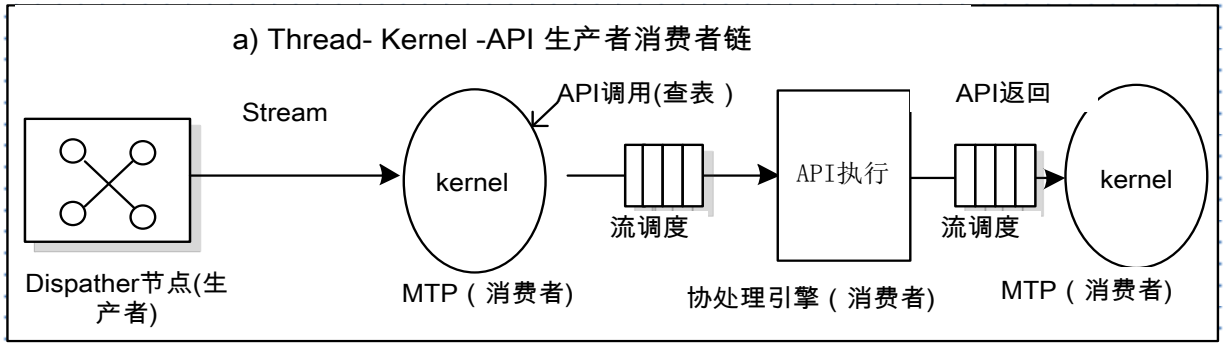
MTP微线程处理器

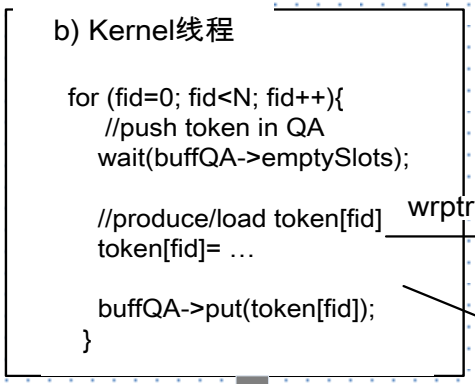
- MTP 面向功耗优化、结构简单的微线程处理器核作为ET的最底层计算单元
  - 适配轻量级线程的计算模式完成计算任务
  - 可实现针对网络应用配置指令集
- 核采用VLIW结构的处理器
  - 包括两（多）个ALU，增强型的数据寄存器文件（支持多模式访问），一组小的分布式操作数寄存器文件，一个本地的稀疏交换总线、一个独立的数据存取单元和一个指令单元
  - 使用较小容量的指令寄存器文件（IRF，Instruction Register File）来降低每次取指令的开销
- 每个MTP支持两个线程，每个线程每次可处理一个报文的一次ACTION

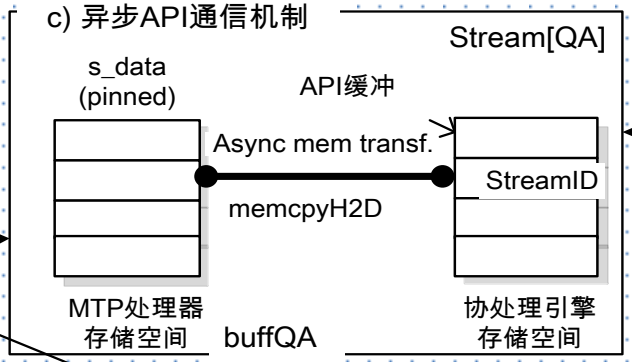
MTP微线程处理器

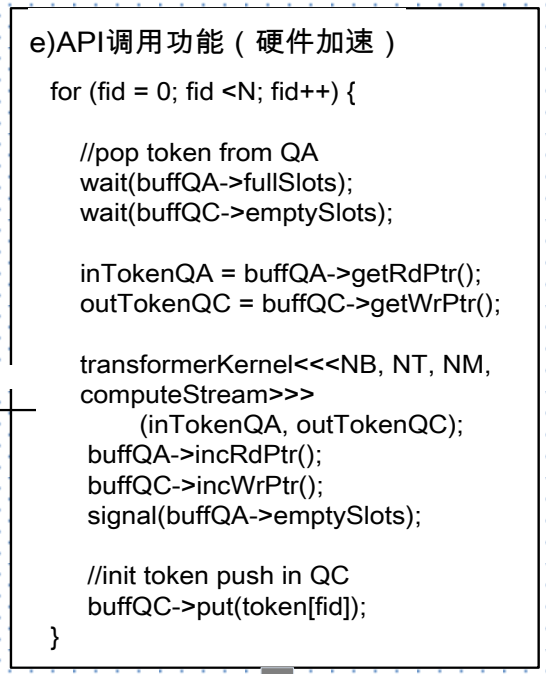
- 指标分析
  - 28nm工艺下，在25mm^{2}可集成不少于160个这样的核心处理器，执行频率不小于1.5GHz，并且完全流水化，每cycle执行两条计算指令（双VLIW指令发射槽），指令吞吐率达到480GOPS
- 执行模型
  - 网络报文预解析后分配StreamID，MTP调用查表引擎后可继续执行或切换报文处理
  - 各种调用根据StreamID再次进入流调度排队，完成MATCH后，携带StreamID的返回信息可以在其他空闲MTP上执行ACTION
  - 这种异步调用和跨节点返回机制能够大大提高报文处理的吞吐率和PPS
  - ET高效的指令缓存技术和片上集群网络数据流通信机制（包括多套通信网络的防死锁机制）

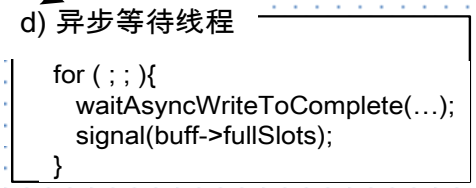
程序执行模式











wrptr

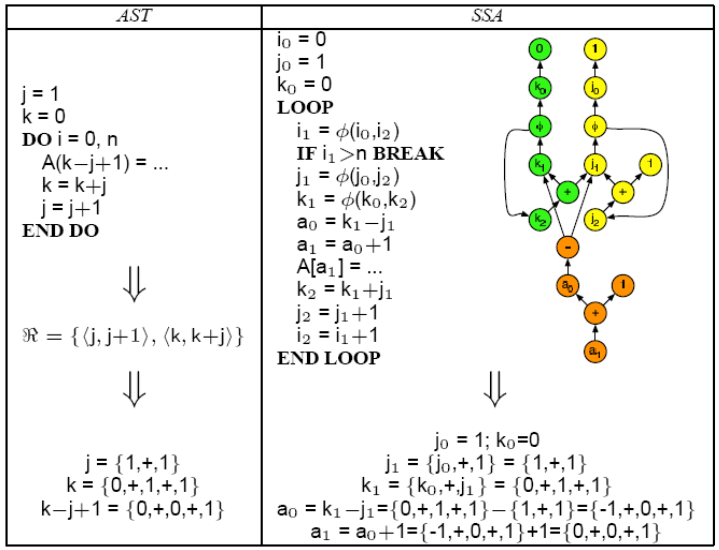
rdptr

执行在MTP

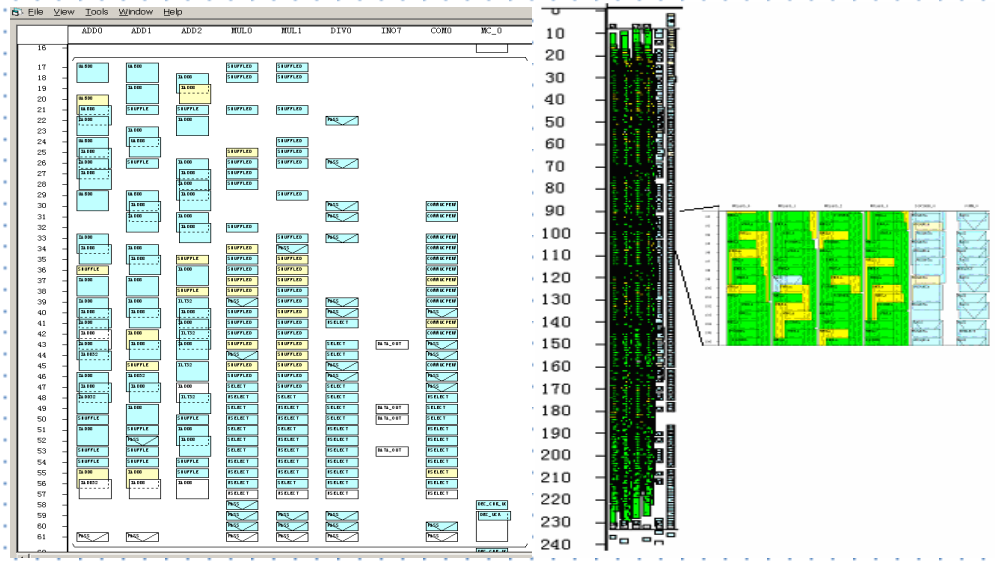
执行在协处理引擎

编程与编译

- SPC允许用户手工书写高度优化的微码，也可以通过编译器快速形成自动优化的代码。SPC提供kernelC语言和编译器。Kernel-C以C语言的子集为基础，添加支持数据流、协处理器API以及实时和自适应优化编程的扩展关键字。
- Kernel C和编译器VLIW结果的例子



流处理器的Kernel-C语言编程



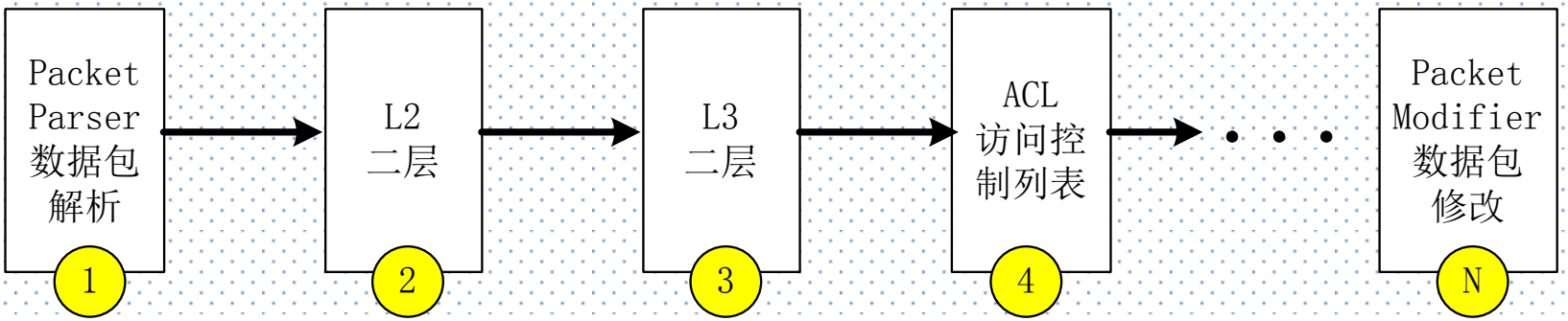
Kernel-C编译器产生的可视化微码结果

Agenda

- 众核流处理器内核-ET
- 基于流处理器的完全可编程SDN交换架构
- N-ET（Net-ET）SDN网络处理器

传统交换芯片架构

- 传统交换芯片流水线



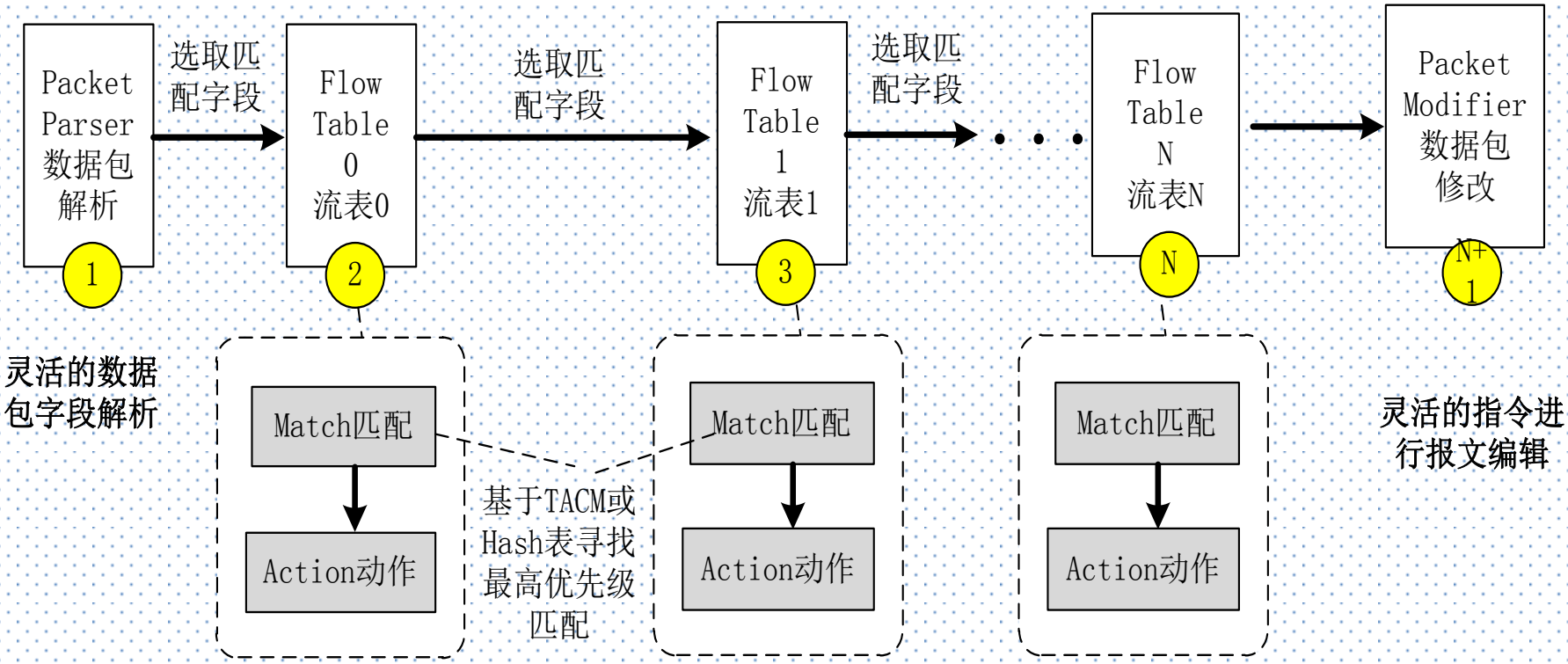
传统交换芯片架构

- 传统的ASIC网络交换芯片，Table、TCAM都是特定协议处理功能对应的，协议处理是固定的硬件逻辑，流水线深度可以到几百级，其中表与逻辑是都硬连线固定关系
- 特定的协议有自己特定的处理模式和处理过程，某个协议处理过程中要编辑什么字段，做什么动作都是确定的
- 传统交换芯片没有多级流表的概念，不可编程，但是性能和吞吐率高，资源（表和TCAM资源不浪费)、功耗和成本优势比较大

基于流水线的SDN交换芯片架构

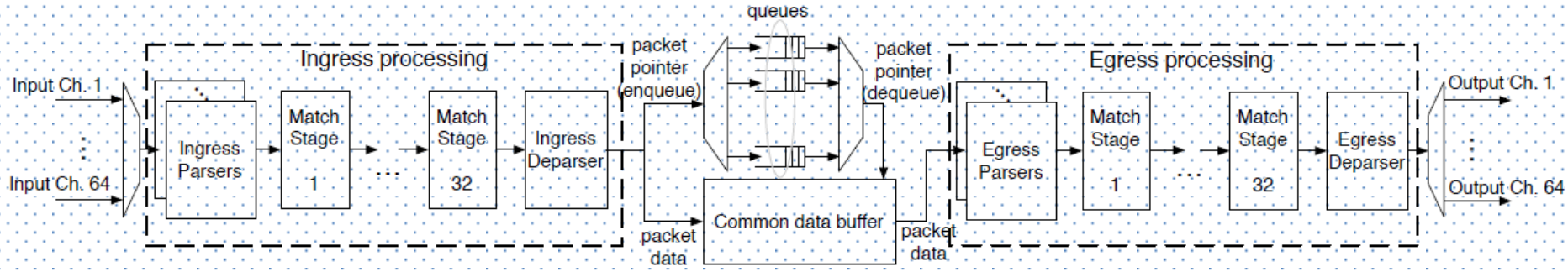
- 按照OpenFlow标准定义直接实现芯片
  - 不再区分是要做路由查找还是二层Mac查找或者MPLS/Trill/Fcoe/PBB/Nat查找
- 只关心要用报文中的哪些字段组合去做匹配查找，查找完之后出什么样的动作，而且一级流表处理后的部分结果可以作为下一级流表处理的输入参数
- 所有这一切都是中性的、协议无关的
- 在传统的商业交换芯片架构设计中，只有ACL TCAM有类似OpenFlow流表的功能，至少需要数十Mbit的TCAM，而片内TCAM在集成电路电路中是一种极其昂贵的资源，功耗、成本、面积代价极大
- OpenFlow提出了多级流表的概念，而且没有限制有多少级。这对传统交换芯片架构设计也是挑战

多级流表 + [MATCH-ACTION]流水线



流水线的SDN交换芯片

- Case Study: Pat Bossharty, Nick McKeown et," Forwarding Metamorphosis: Fast Programmable Match-Action Processing in Hardware for SDN", SIGCOMM 2013

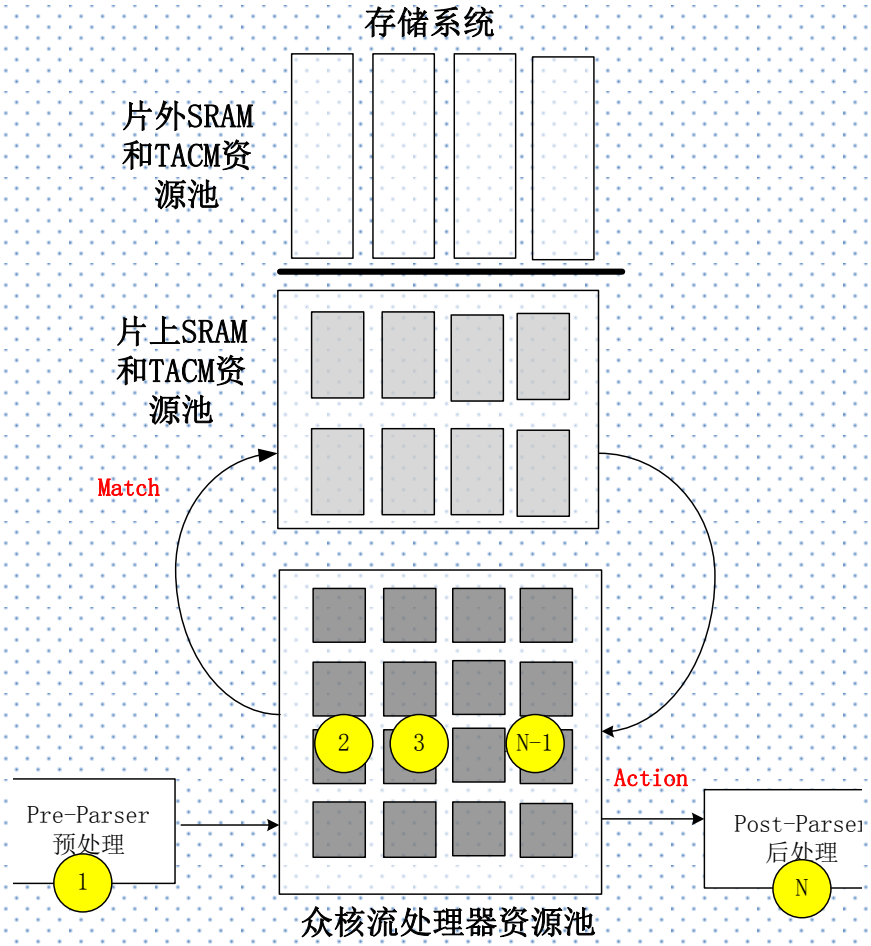


流水线的SDN交换芯片特征分析

- 多级流表构成流水线
- 大量的TCAM和SRAM（用于hash表）
  - 约传统交换芯片的4~8倍，成本很高（传统交换芯片的大部分面积已经被SRAM和TCAM占用）
- 深度流水。算上SRAM访问延迟，流水线深度仍然在数百这个量级
- 流水线（半）固化。实际是一种依赖表项配置实现的半可编程，不是完全可编程

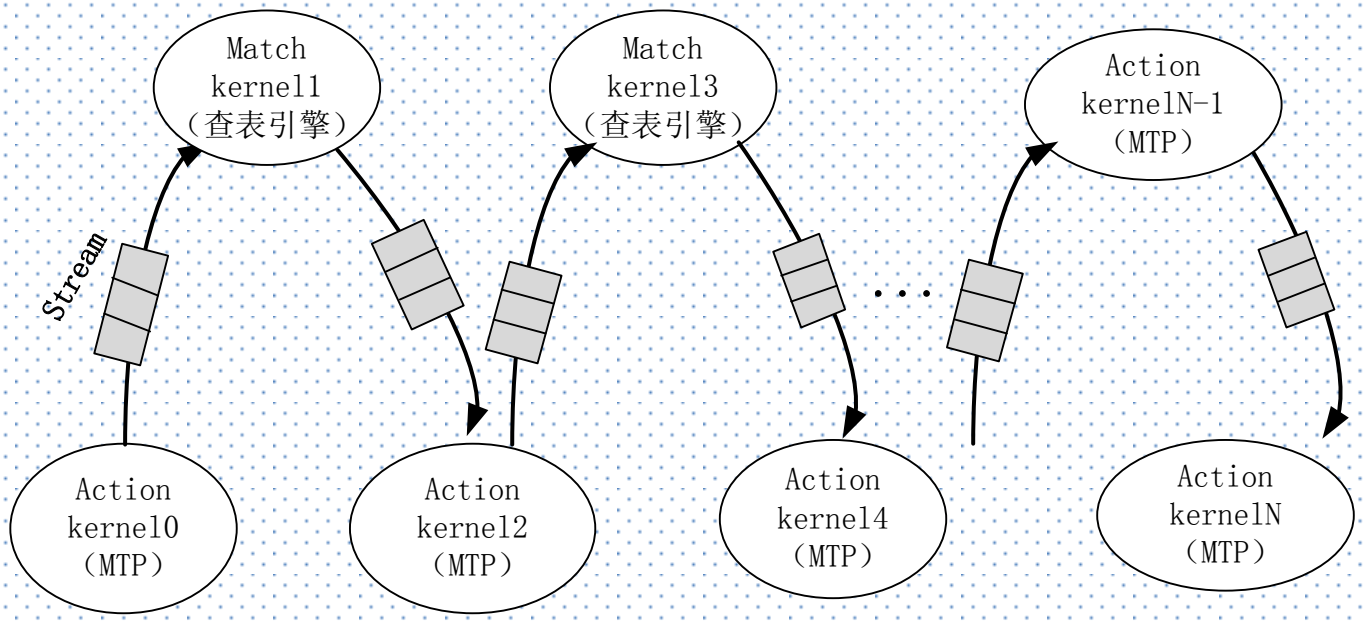
基于流处理器的完全可编程SDN架构

- 从芯片的角度，MACTH-ACTION变成处理器资源池和存储资源池之间的访问和通信
- 流水线折叠
  - 计算、存储（SRAM和TCAM）资源解锁



基于流处理器的完全可编程SDN架构

- 从流编程的角度，流处理器上执行的kernel-stream构成逻辑上的MACTH-ACTION 流水线



基于流处理器的完全可编程SDN架构

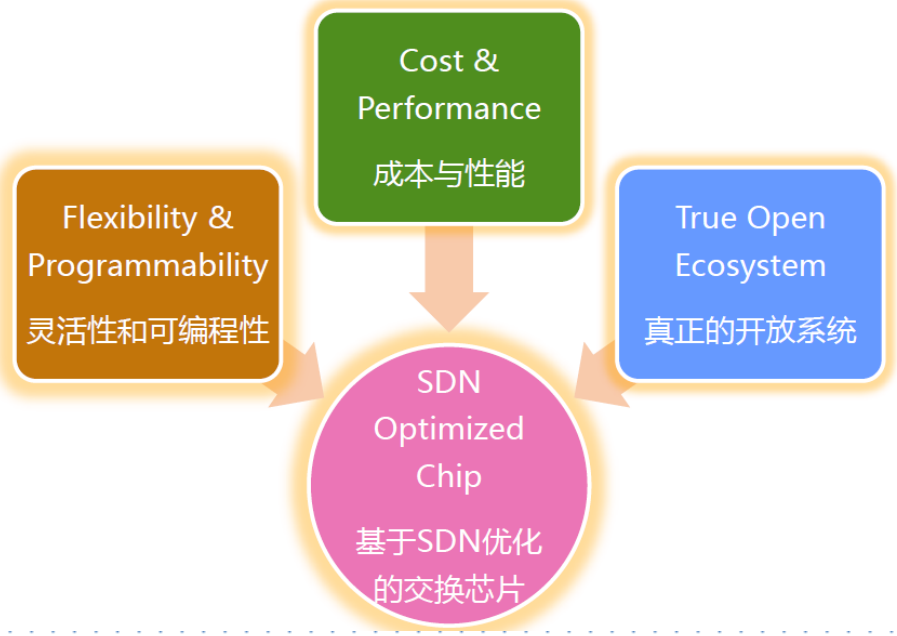
- 把深度多级流水线结构变成处理器结构
- 资源池化，共享提高利用率、吞吐率、性能弹性和灵活性
  - 片上和片外存储资源池化
  - 处理器核资源池化
- 完全可编程，处理器除了支持ACTION工作，还可以承担多种任务，例如报文解析（硬件不识别协议可通过编程支持），字段提取、报文编辑等
- 可按报文任意支持1-N级流表，N不受流水线深度限制，报文延迟同处理的流表级数线性正比，对于简单处理可节省延迟
- 流队列调度、异步调用提高处理器报文吞吐率
- 可扩展性好，处理性能主要与集成的流处理器核数正相关，表项可扩展片外存储资源
- 采用处理器IP架构，芯片设计难度比ASIC完全固化低，快速推出产品，可提供丰富的产品线

Agenda

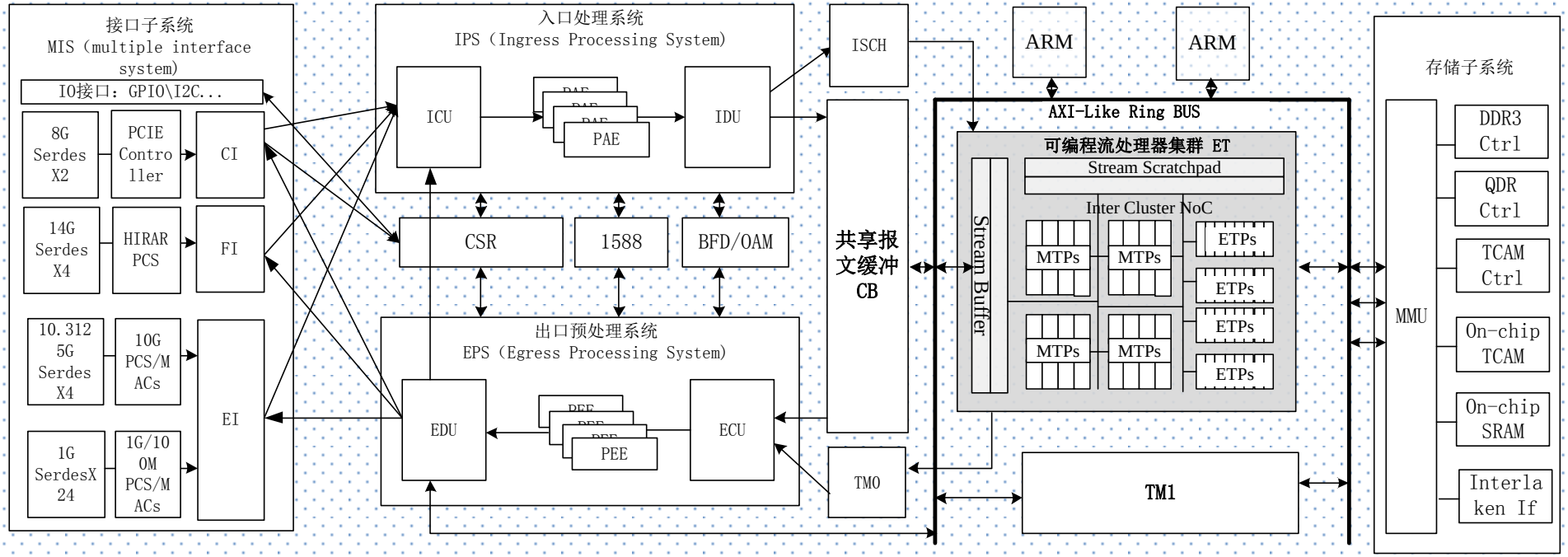
- 众核流处理器内核-ET
- 基于流处理器的完全可编程SDN交换架构
- N-ET（Net-ET）SDN网络处理器

N-ET（Net-ET）芯片架构

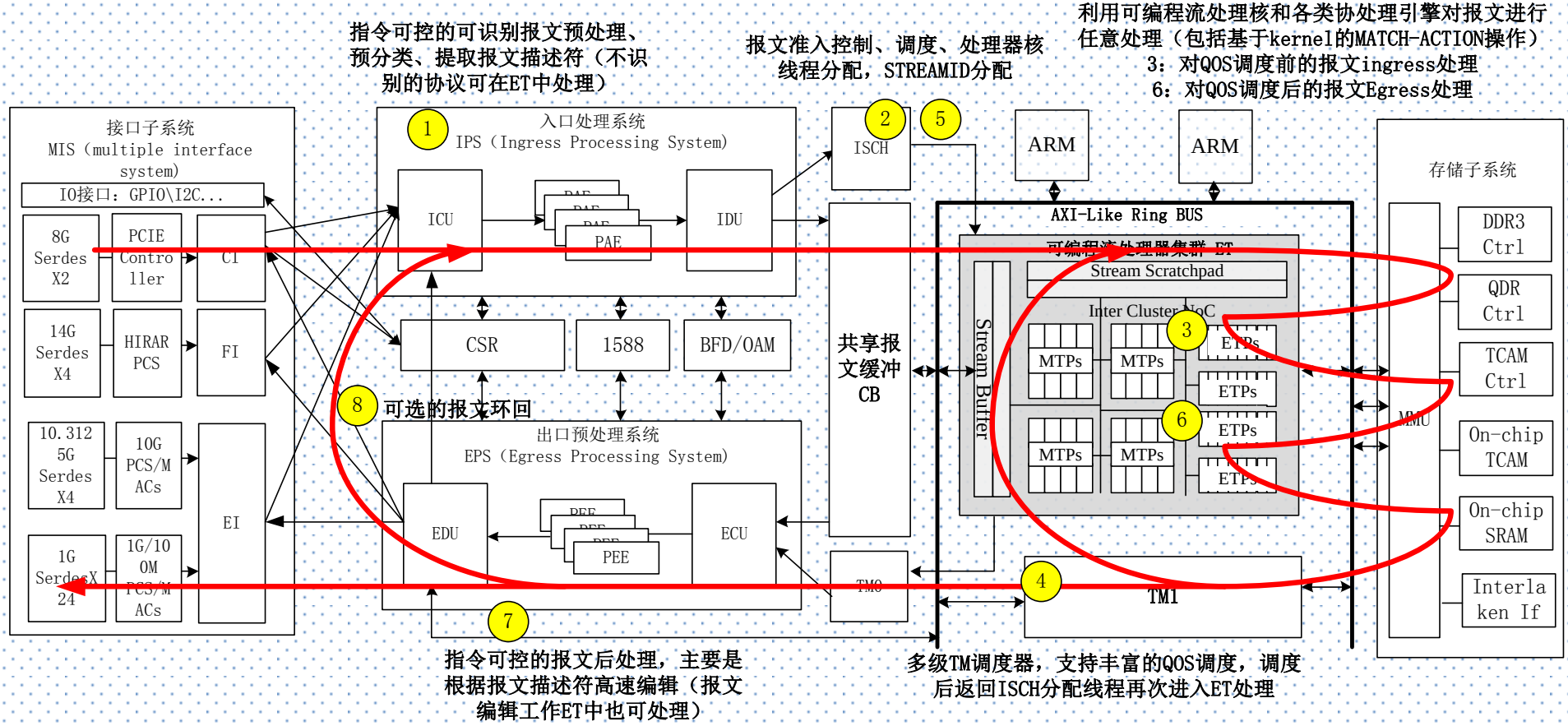
- 基于流处理器自主设计的SDN处理器芯片
- 单N-ET芯片400Gbps处理能力
- 多N-Et芯片+NR 扩展到TGbps以上



N-ET（Net-ET）实验芯片



N-ET（Net-ET）报文处理流程



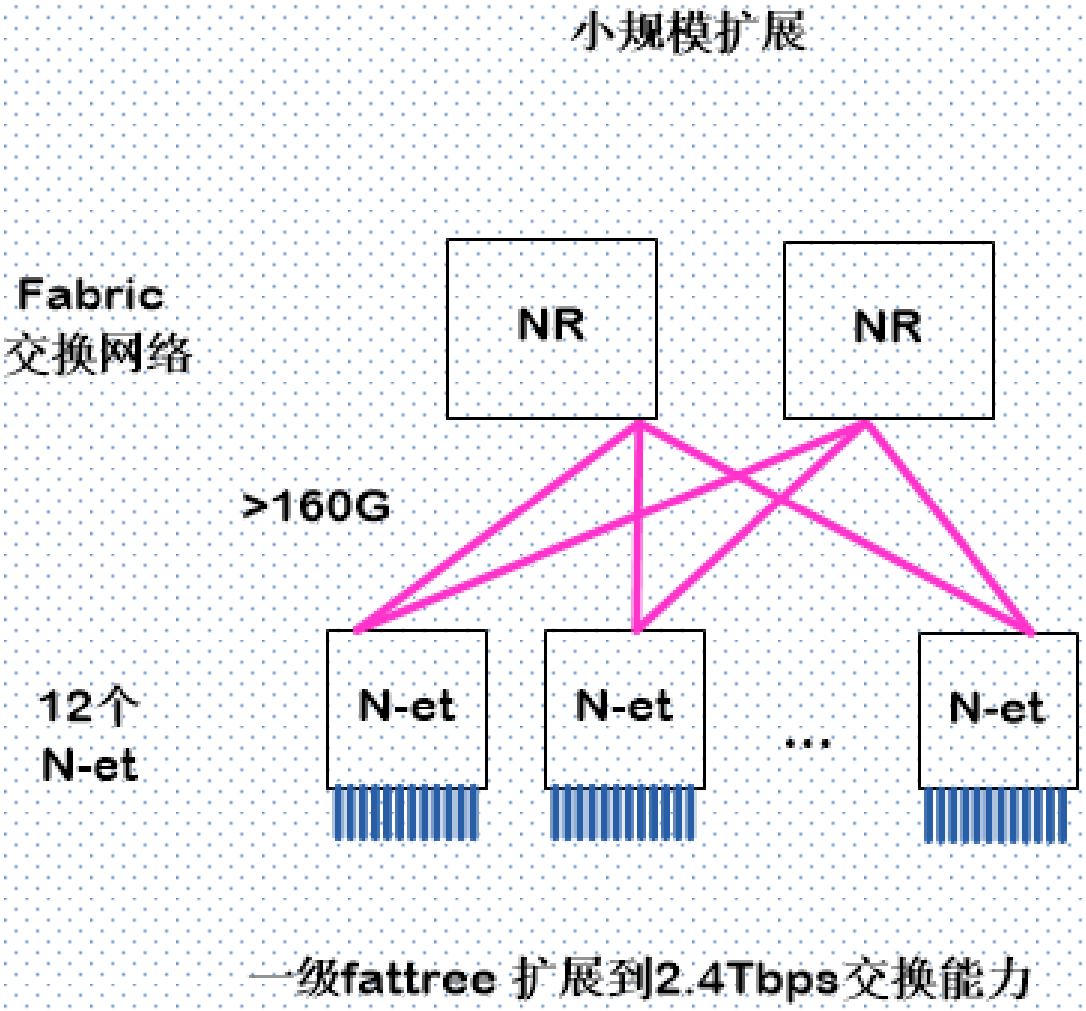
主要设计指标

- 内核工作频率：800MHz-1GHz
- 单片 >400Gbps全线速处理带宽, 800Mpps处理能力
- 支持多路片外扩展的DDR3 SDRAM控制器、QDR SRAM控制器、TCAM控制器
- 集成自主研发的流处理器
- 支持192K VLIW指令空间
- 支持200个MTP内核（400个线程），64个增强线程处理器ETP
- 64个增强线程处理器ETP分为如下几类：
  - 32个表搜索协处理器，提供片上、片外表项资源流水化Hash查找功能
  - 支持基于流的匹配
  - 16个TCAM搜索协处理器，提供片上或者片外TCAM表查找功能
  - 4个灵活计数协处理器，提供灵活的流水化计数功能
  - 4个灵活METER协处理器，提供灵活的令牌桶和CAR功能
  - 1个CSR访问协处理器，为CPU提供寄存器和表的访问接口
  - 1个锁同步协处理器
  - 2个ARM调用协处理器
- 16个报文预解析（Parser）引擎，16个报文后编辑（Edit、Modify）引擎，采用全流水架构

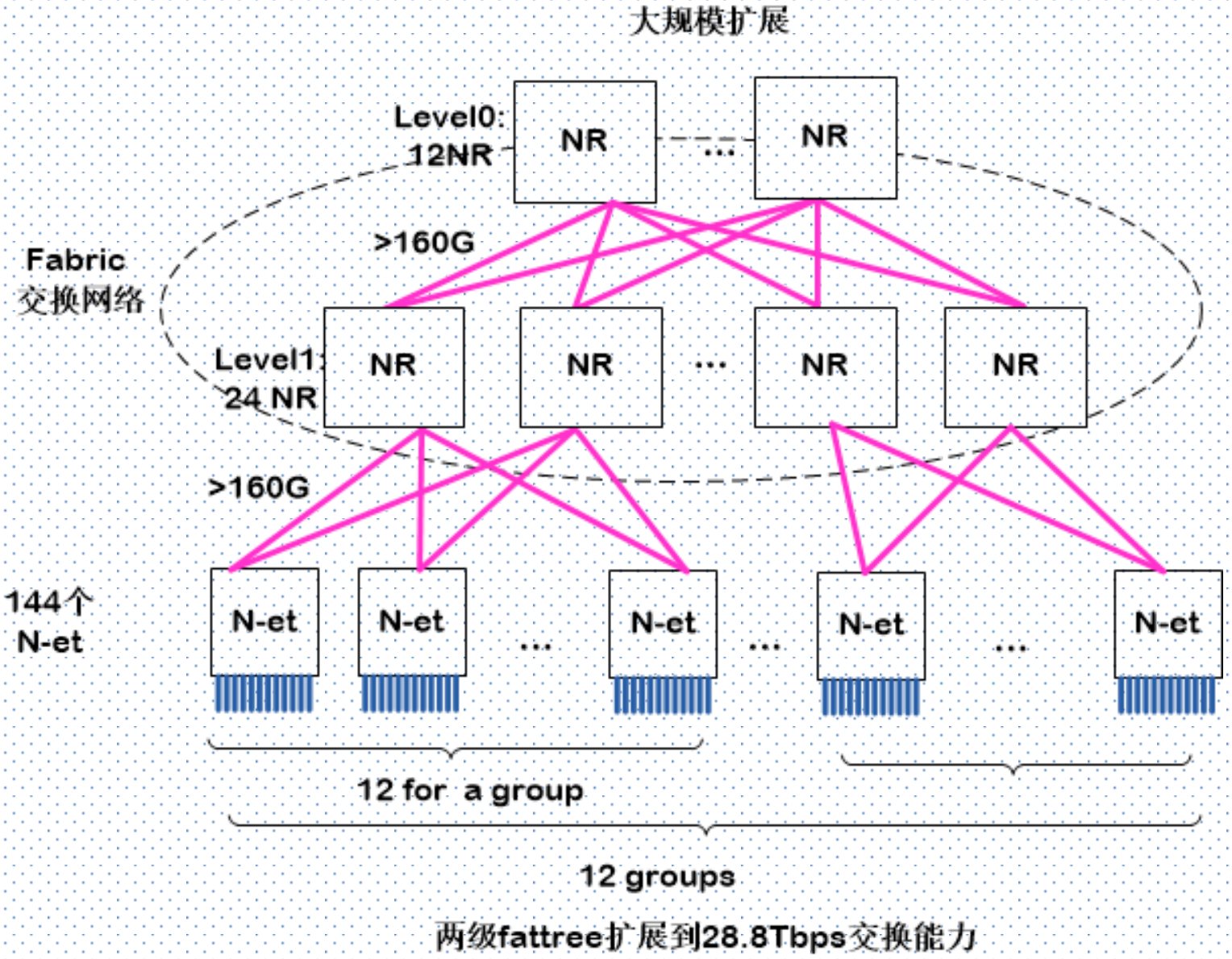
主要设计指标

- 内部集成2个流量管理部件（TM）
- 不小于16K队列、3级层次化调度
- 内部集成统计计数模块，统计计数与内部查找表复用RAM资源
- 支持不小于512K 64bit内部计数器和48M 64bit外部计数器
- 内部集成以太网OAM模块
- 支持 BFD、以太 OAM
- 内部集成IEEE1588v2硬件加速模块、支持1588协议报文识别和打时间戳功能
- 内置大容量报文缓存和查找表资源
  - 不小于6MB内部报文集中式缓存，支持报文链表插入、修改、删除
  - 不小于16MB各类内部查找表资源
- 支持1G、10G、40G、100G以太网接口，内置MAC，SERDES
- 集成4路INTERLAKEN接口
- 集成4路HIRAR接口，支持与HNR高速互联芯片对接
- PCIE3.0接口

级联扩展设计方案


小规模扩展
Fabric
交换网络
NR
NR
>160G
12个
N-et
N-et
N-et
...
N-et
一级fattree 扩展到2.4Tbps交换能力

级联扩展设计方案



两级fattree扩展到28.8Tbps交换能力

Thanks