浅谈多核处理器与系统软件研发

--实例研究: 思科QuantumFlow处理器与系统软件结构

陈怀临
《弯曲评论》
www.tektalk.cn

彎曲評論
科技 · 人物 · 潮流

提纲

- 《弯曲评论》
- 系统软件
- 工业动态
- 系统理解
- 实例研究：思科QuantumFlow
- 结束语

彎曲評論

科技 · 人物 · 潮流

《弯曲评论》（www.tektalk.cn）

- 目标：非盈利性智库机构
- 领域：
  - 科技跟踪
  - 专题分析
  - 人物报道
  - 学术打假

彎曲評論

科技 · 人物 · 潮流

《弯曲评论》（www.tektalk.cn）

最近工作总结：

×专题分析

《思科Quantum处理器与战略研究》

《对中国系统软件的思考与建议》

《对华为系统软件的战略思考（上）》

《对华为系统软件的战略思考（下）》

《《对国防科大麒麟操作系统研发的思考》

《中国计算机发展史略(1956-2006)》

彎曲評論

科技 · 人物 · 潮流

《弯曲评论》（www.tektalk.cn）

最近工作总结：

×科技书籍：

《PowerPC and Linux Kernel Inside》

《Linux 核 心》（The Linux Kernel)（下）

《Linux 核 心》（The Linux Kernel)（中）

《Linux 核 心》（The Linux Kernel)（上）

《MIPS CPU 体系结构概述，Linux/MIPS内核》（下）

《MIPS CPU 体系结构概述，Linux/MIPS内核》（上）

《See MIPS Run》

彎曲評論

科技 · 人物 · 潮流

《弯曲评论》（www.tektalk.cn）

最近工作总结：

×人物评述：

《邓稼先传》

《海外学人》

《计算的美丽-图灵奖的第一个四十年》（上）

《计算的美丽-图灵奖的第一个四十年》（下）

彎曲評論

科技 · 人物 · 潮流

系统软件

- 操作系统
  - 桌面操作系统
  - 服务器操作系统
  - 嵌入式操作系统
- 编译器与工具链（gcc, binutil, gdb…）
- 编程环境，中间件
  - PVM，MPI，OpenMP
  - Mapreduce，Hadoop
  - CORBA，DCOM

系统软件

× 嵌入式操作系统

-传统分时系统：Linux，FreeBSD

-微内核： QNX/Neutrino， L4，Mach

-大型通信操作系统：

—华为/VRP

—思科/IOS，IOS-XR，IOS-XE

通信业工业研发动态

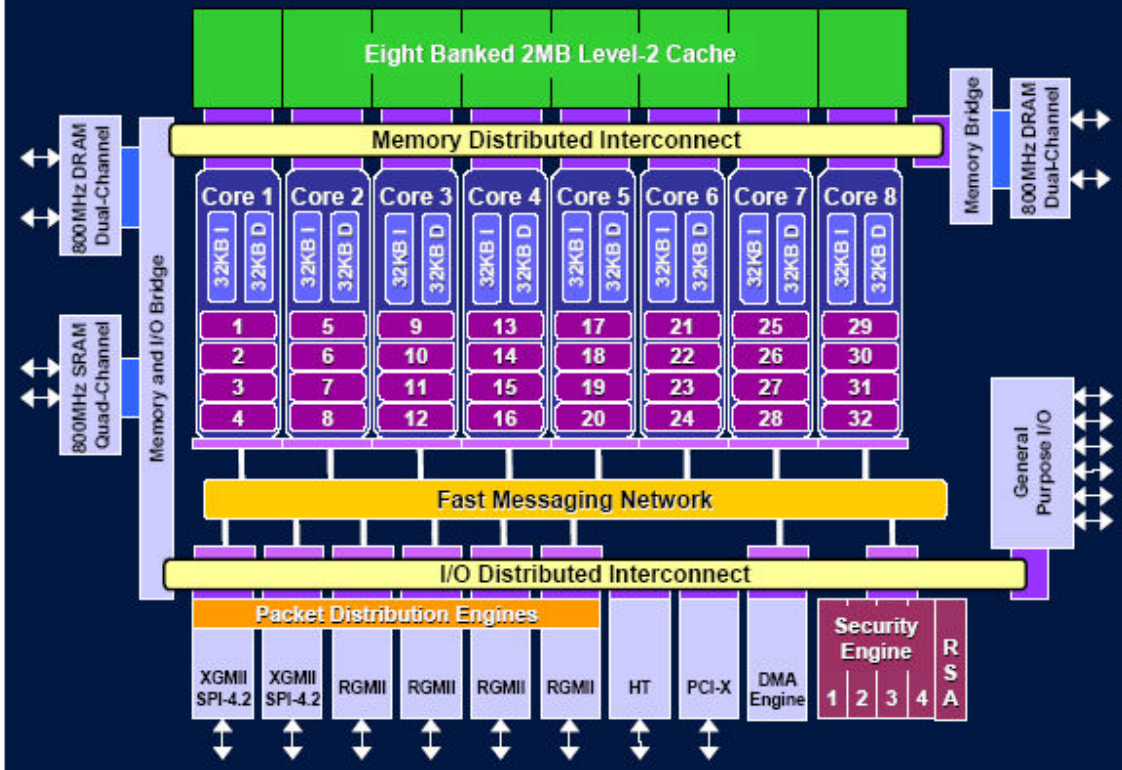
× 多核系统的持续应用

× 多核系统的多样化

×微观分布式并行计算系统

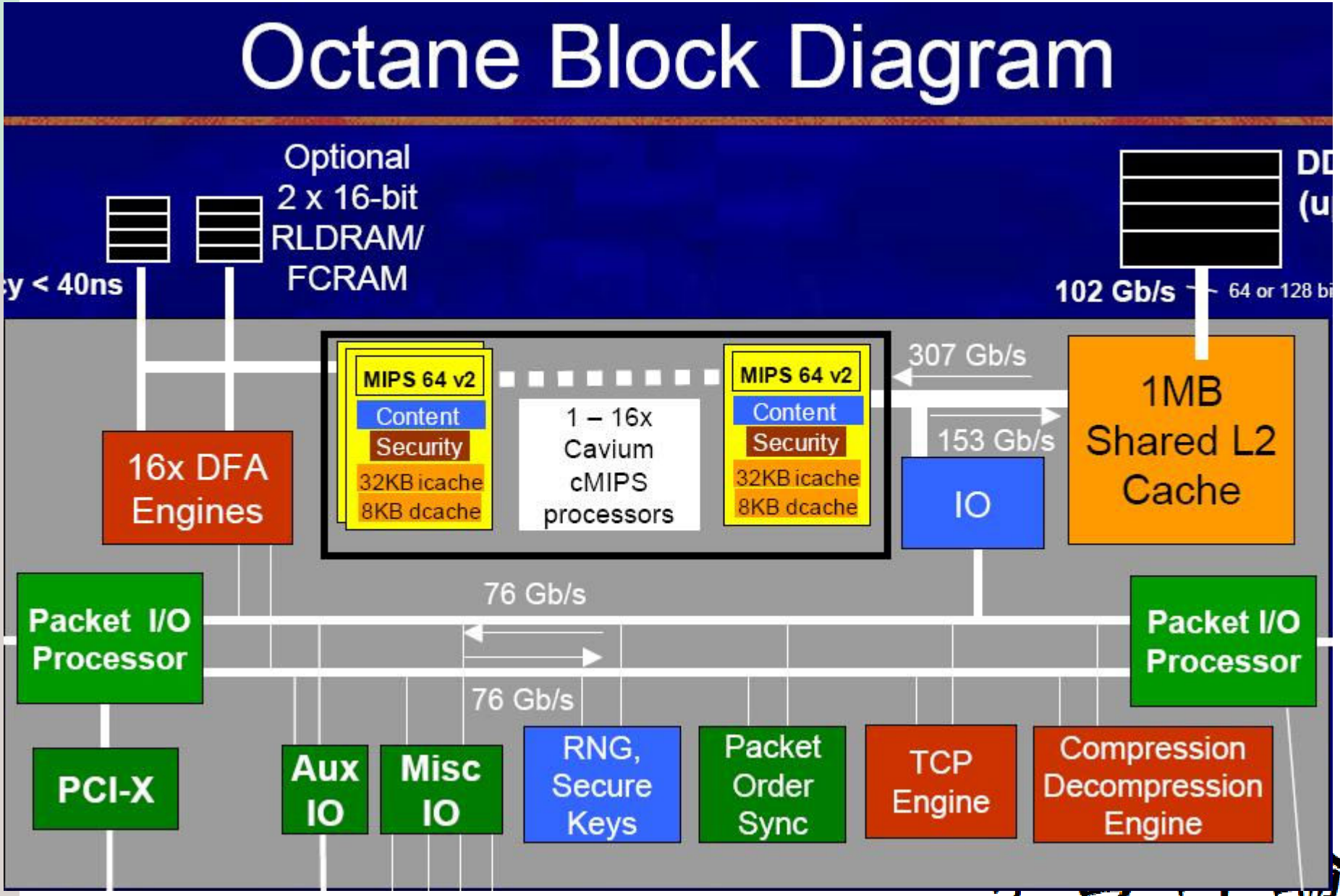
工业动态

XLR Processor Overview



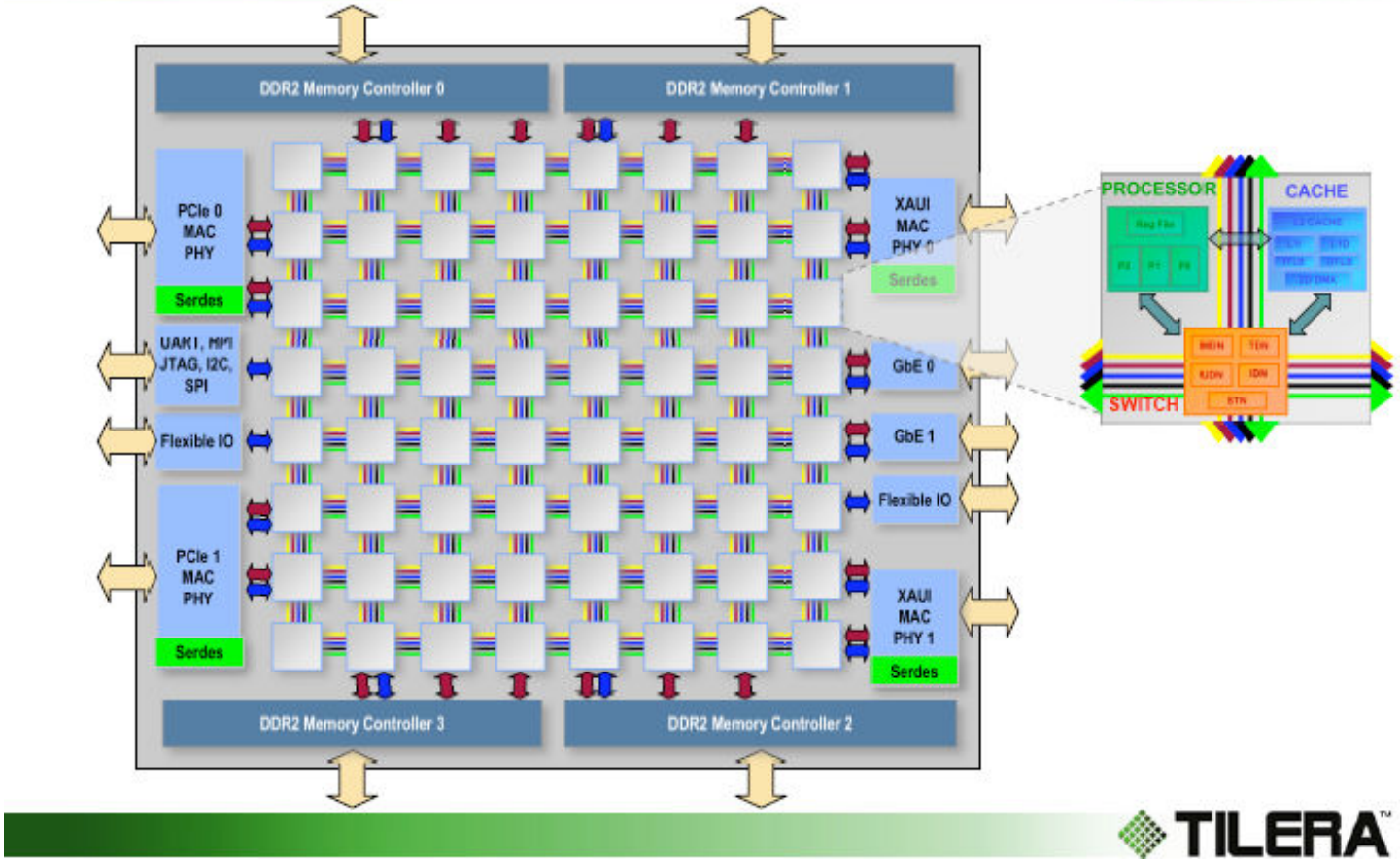
- 32 Fine Grain Threads
- 8 MIPS64™ XLR™ Cores
- 1.5GHz Operation
- 8-banked 2MB L2 Cache Fully cache-coherent
- Full speed point-to-point Interconnects (No busses)
  - Fast Messaging Network *1.5 Billion msg / sec*
  - Memory Distributed Interconnect *384Gbps Bandwidth*
  - I/O Distributed Interconnect *192Gbps Bandwidth*
- Security Acceleration Engine
  - 10Gbps Bulk Encryption
  - 4 Parallel CryptoCores

Octane Block Diagram



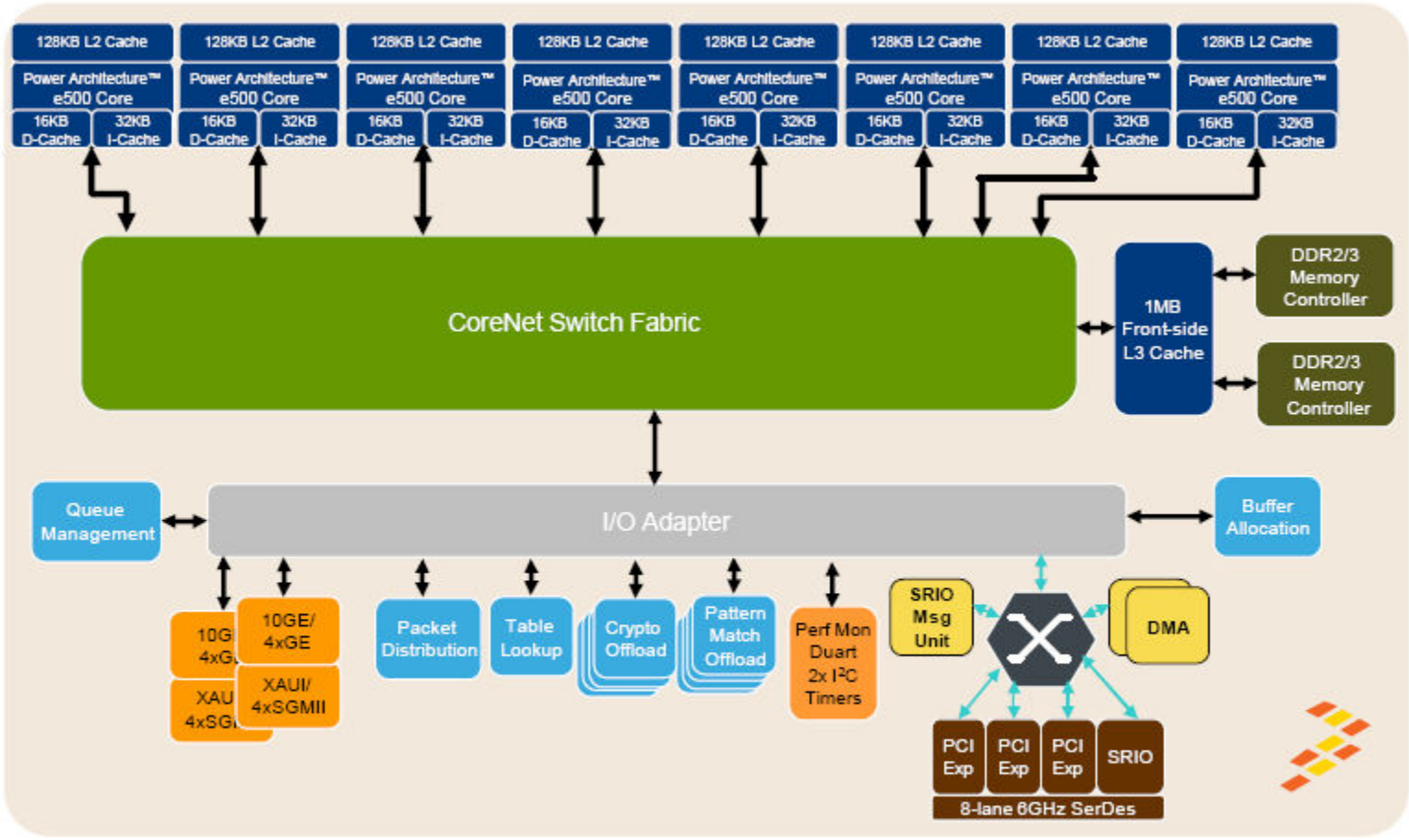
TILE64 Processor Block Diagram

A Complete System on a Chip



TILERA

MPC8578 Product Concept



freescale semiconductor

系统理解

× 系统（System）与操作系统（Operating System），核心（Kernel)的关系

×计算单元的多样化（ASIC，FPGA，CPU，NP）

×互连网络的多样化（Bus，Switch Fabric，Interconnect）

×数据报文（Packet）驱动，调度。

×中断，异常处理的简化

×性能的考量，牺牲层次性和透明性

实例研究：思科QuantumFlow

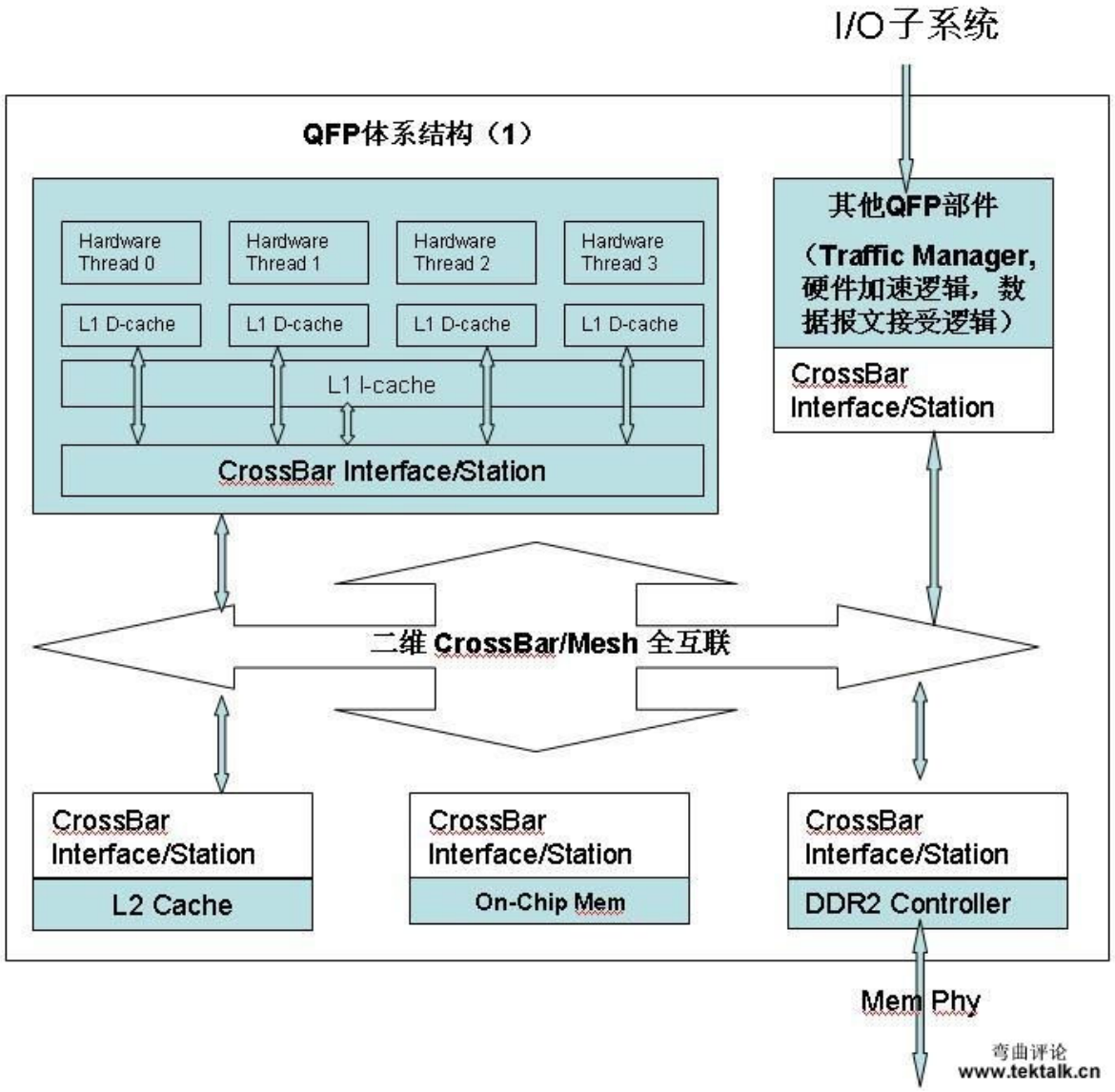
Forty custom cores are known as packet processor engines (PPEs).

Each PPE:
- Is multithreaded (4/PPE)
- Programmed in C-language
- Has no fixed-feature pipeline and can process any flow
- Offered full header and payload visibility on demand
- Consumes less than 400 mW

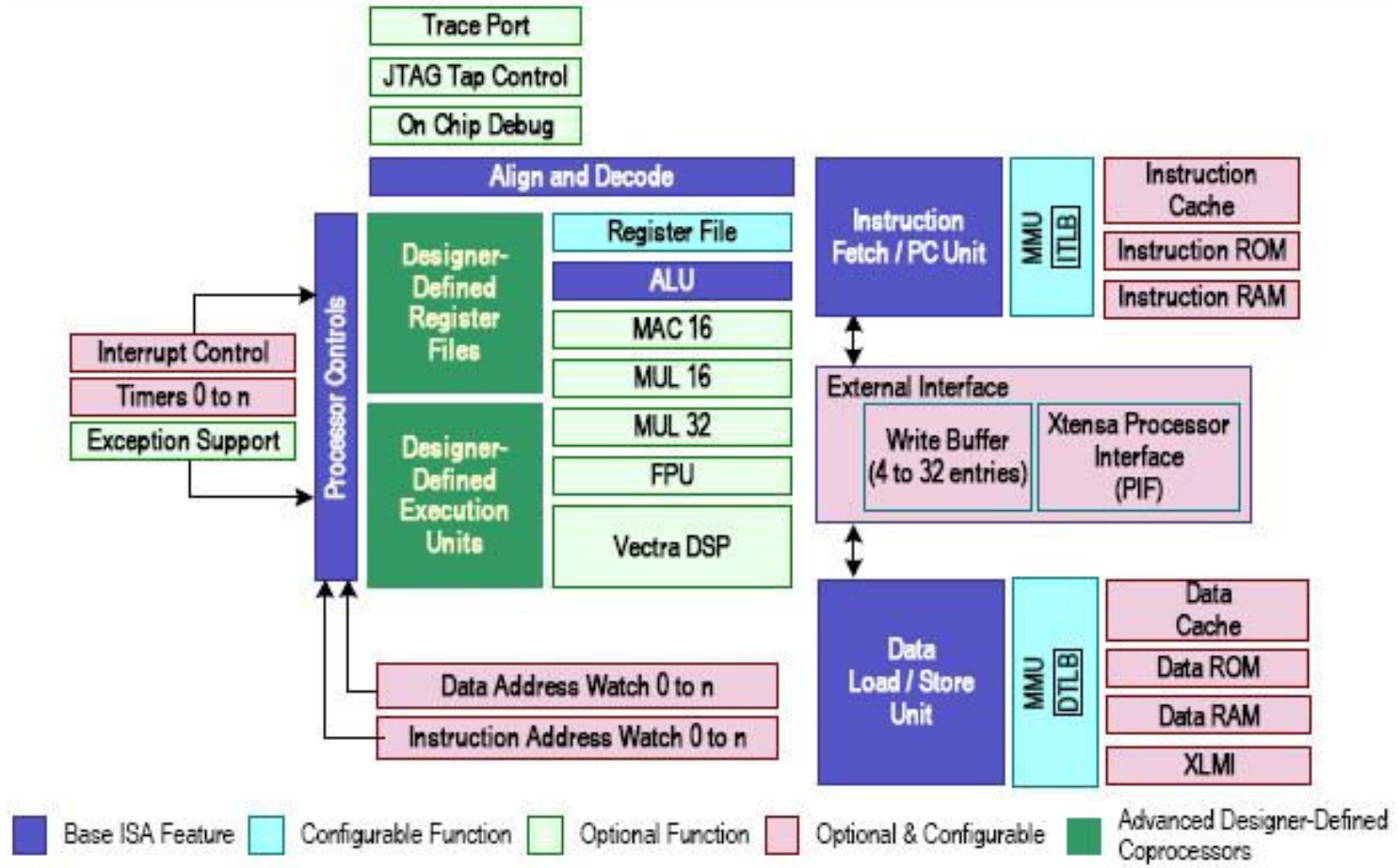
实例研究：思科QuantumFlow

× 项目启动时间：2002年Q3或Q4【笔者注：SPP是2002年流片的。CRS-1是2004年推出的。】
×研发耗资：1亿美金
×芯片主要定位：边缘（Edge）路由器，企业路由器。
×芯片解决问题：Stateful Service与转发（Forwarding)合一。
×首发系统：ASR1000
×主频：1.2GHZ【笔者注：ESP-5G:900MHz. ESP-10G:900MHz.ESP-20G:1.2GHz】
×晶体管数目：8亿
×（PFE）内存：DDR2。
×数据报文内存（Packet Buffer）：ESP-5G:64M.ESP-10G:128M.ESP-20G:256m
*CAM:外挂TCAM 【笔者注：ESP-5G:10M. ESP-10G:10M.ESP-20G:40M】
×功耗：80瓦
×多核：40，4 Way-Thread。来自Tensica的Xtensa。
×片内互联（Interconnect）：Crossbar Switch
×片外互联：ESI 【笔者注：在将来的新QFPzhong，将是Interlaken】
×数据报文接口：4个10GBPS SPI4.2 。
×工艺：90nm
×流片：德州仪器

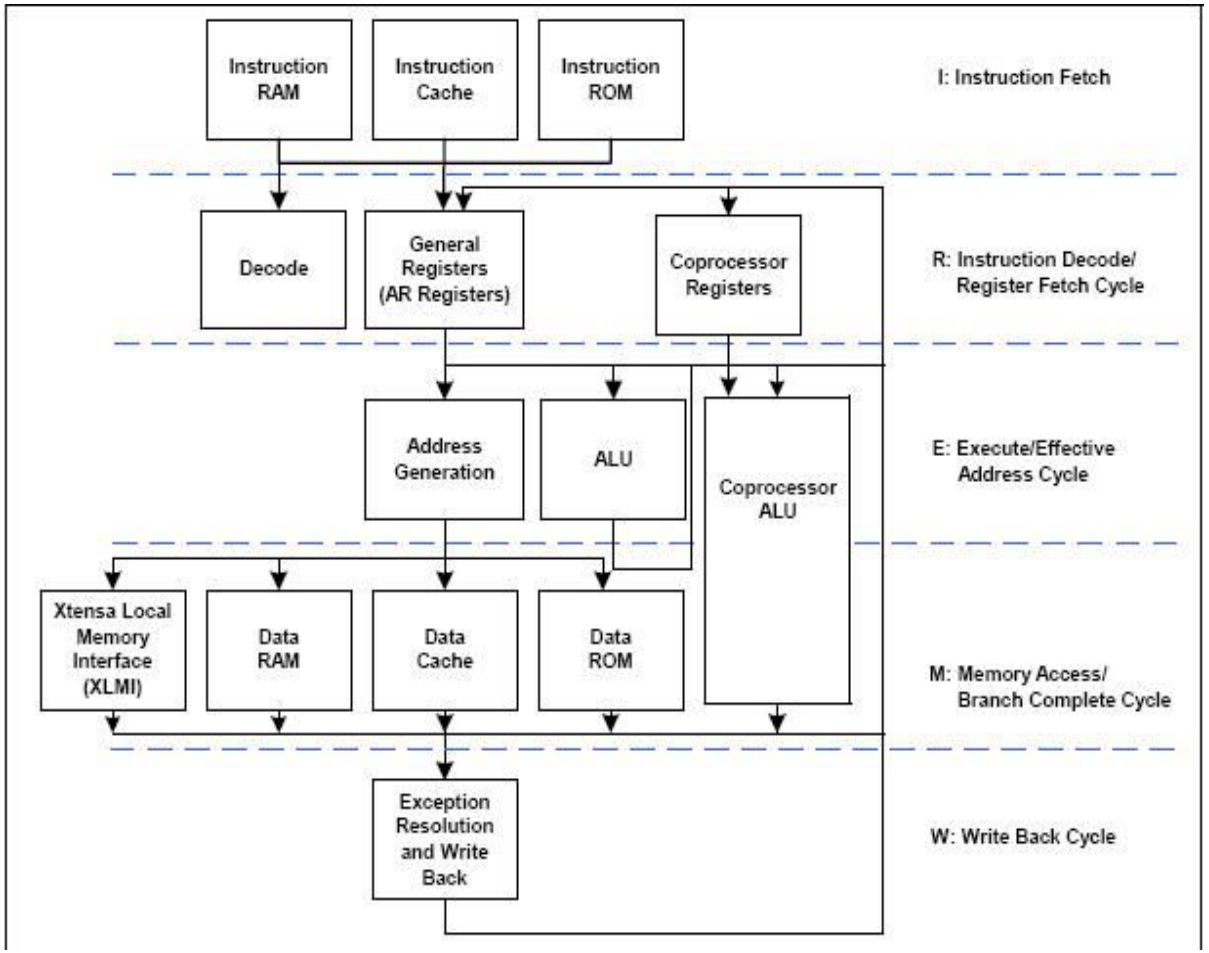
实例研究：思科QuantumFlow



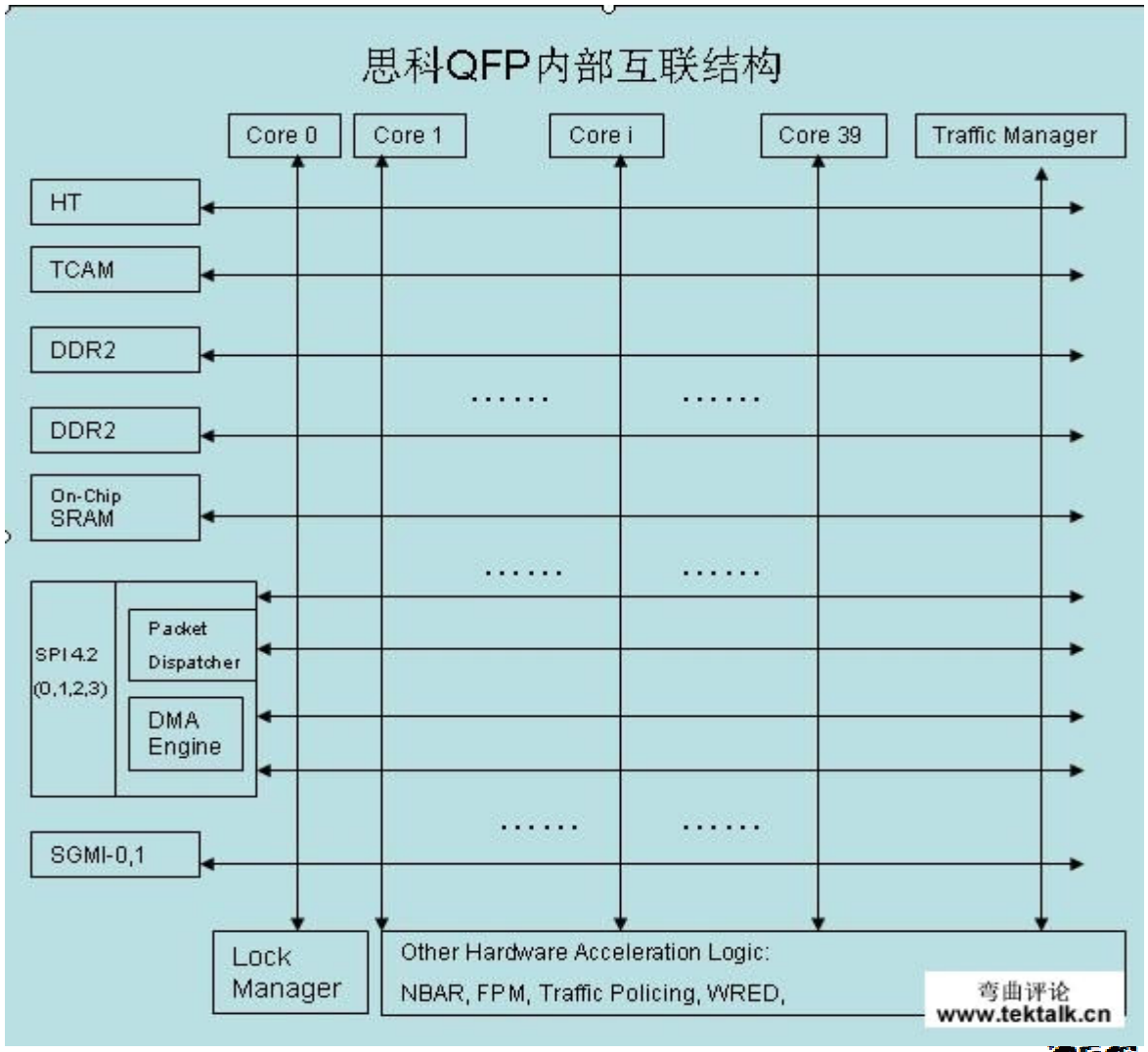
实例研究：思科QuantumFlow



实例研究：思科QuantumFlow

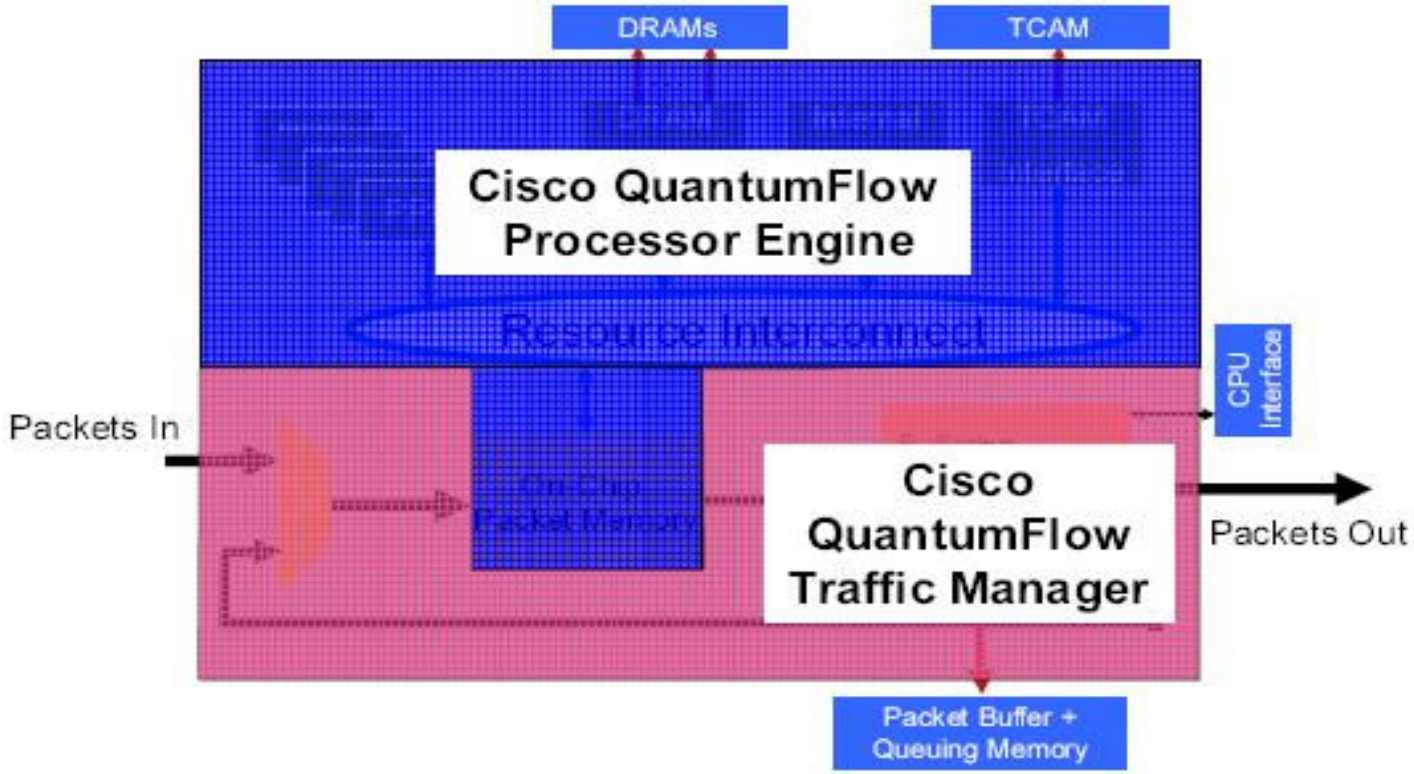


实例研究：思科QuantumFlow

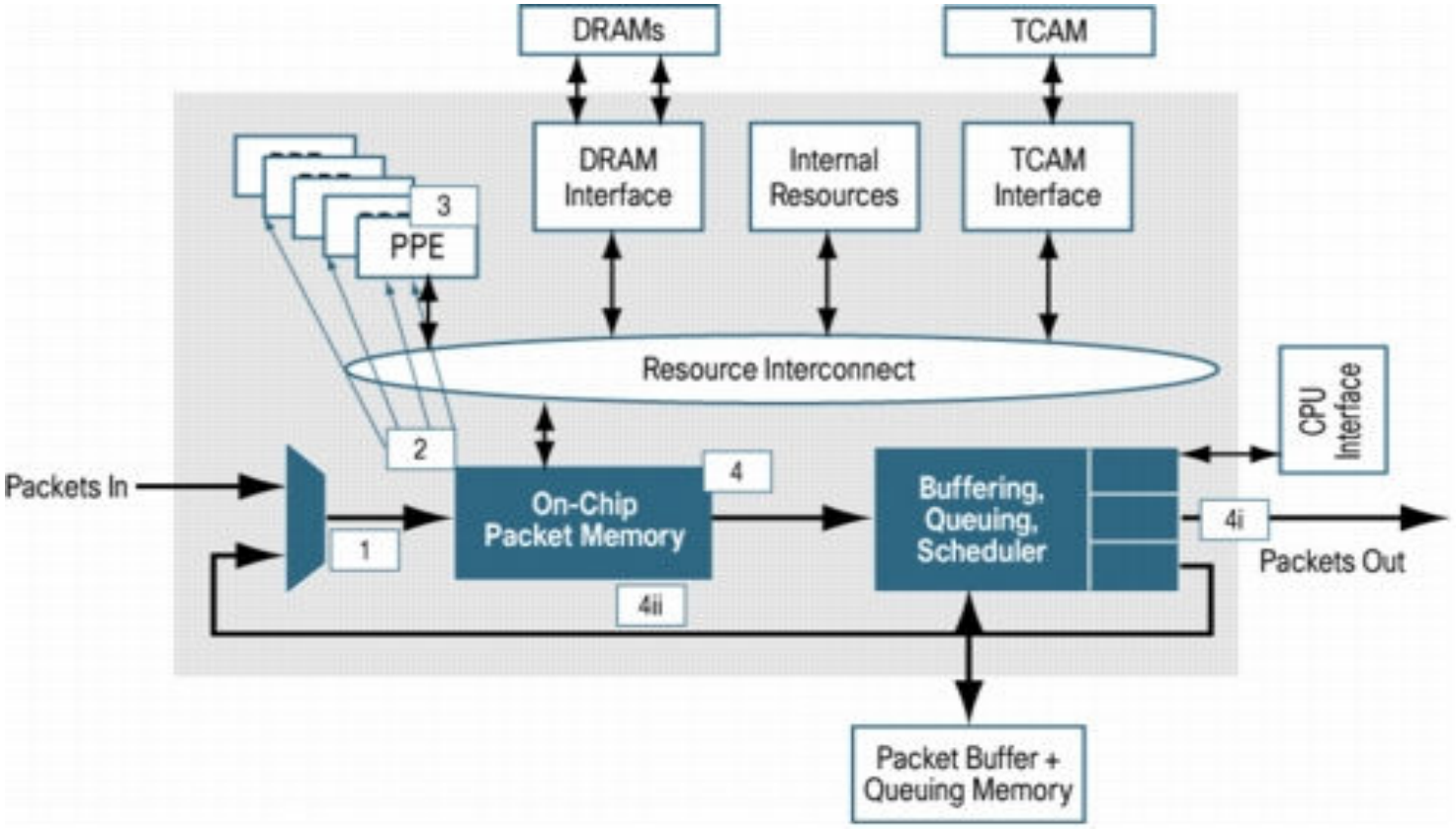


实例研究：思科QuantumFlow

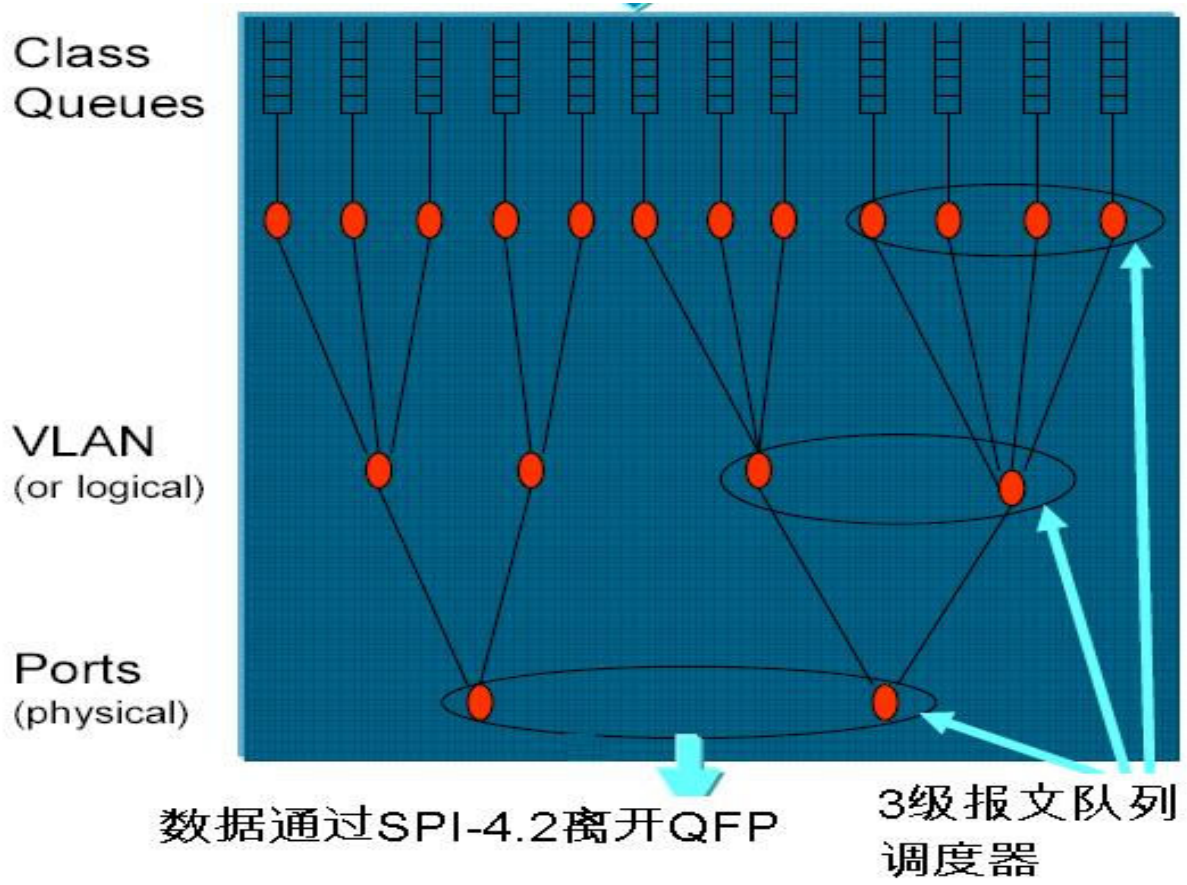
思科QFP数据报文流程



实例研究：思科QuantumFlow



实例研究：思科QuantumFlow



思科QFP的Traffic Manager结构图

实例研究：思科QuantumFlow

思科QFP Thread 软件模型

思科QFP 控制 Thread 程序逻辑结构

While（1）

{

Process 来自ESP板控制CPU的控制消息。

更新QFP的数据，如FIB或其他。

处理各种统计数据；并汇报给控制CPU

监测QFP其他Data Thread的工作异常，并做出相应处理。

其他管理工作。。。

}

思科QFP Data Thread 程序逻辑结构

While（1）

{

Block for wakenup

Fetch a new packet descriptor.

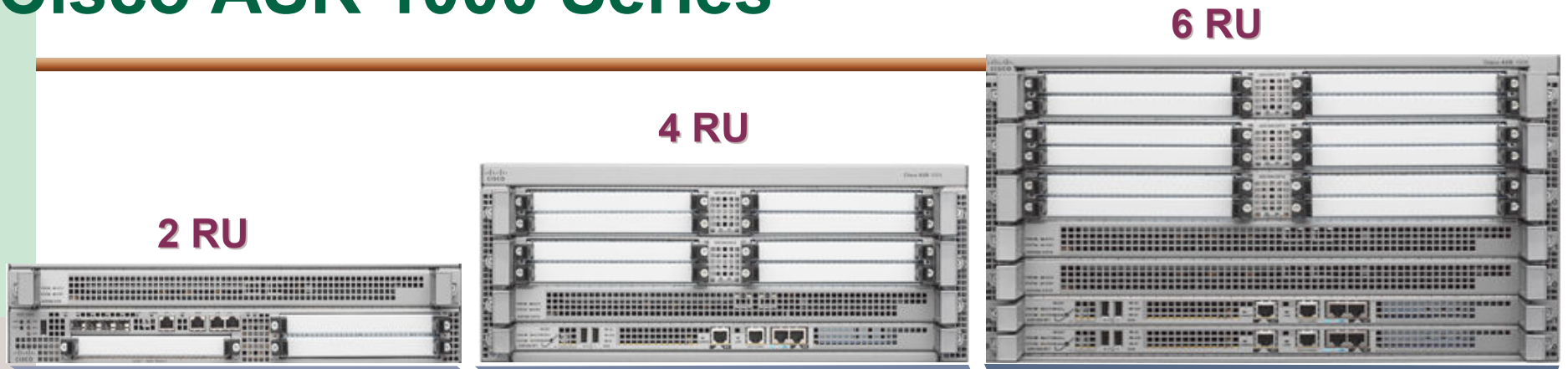
Process this packet

Dispatch the packet to QFP traffic manager queue;

Notify Traffic manager

}

Cisco ASR 1000 Series



SPA Slots	3-slot	8-slot	12-slot
# of ESP Slots	1	1	2
# of RP Slots	Integrated (RP1)	1	2
# of SIP Slots	Integrated (SIP10)	2	3
IOS Redundancy	S/W	S/W	H/W
Built in GigE	4	n/a	n/a
Height	3.5” (2RU)	7” (4RU)	10.5” (6RU)
Bandwidth	5-10 Gbps	10-40+ Gbps	10-40+ Gbps
Performance	4-8 Mpps	8-16+ Mpps	8-16+ Mpps
Air Flow	Front to Back	Front to Back	Front to Back
Power Supply (Watts)	470	765	1275

Aggregated Services & Scale

ASR1000 ILT – Naming

CPP = Cisco Packet Processor now known as the QuantumFlow Processor (QFP)

CC = CarrierCard now known as the SPA Interface Processor (SIP)

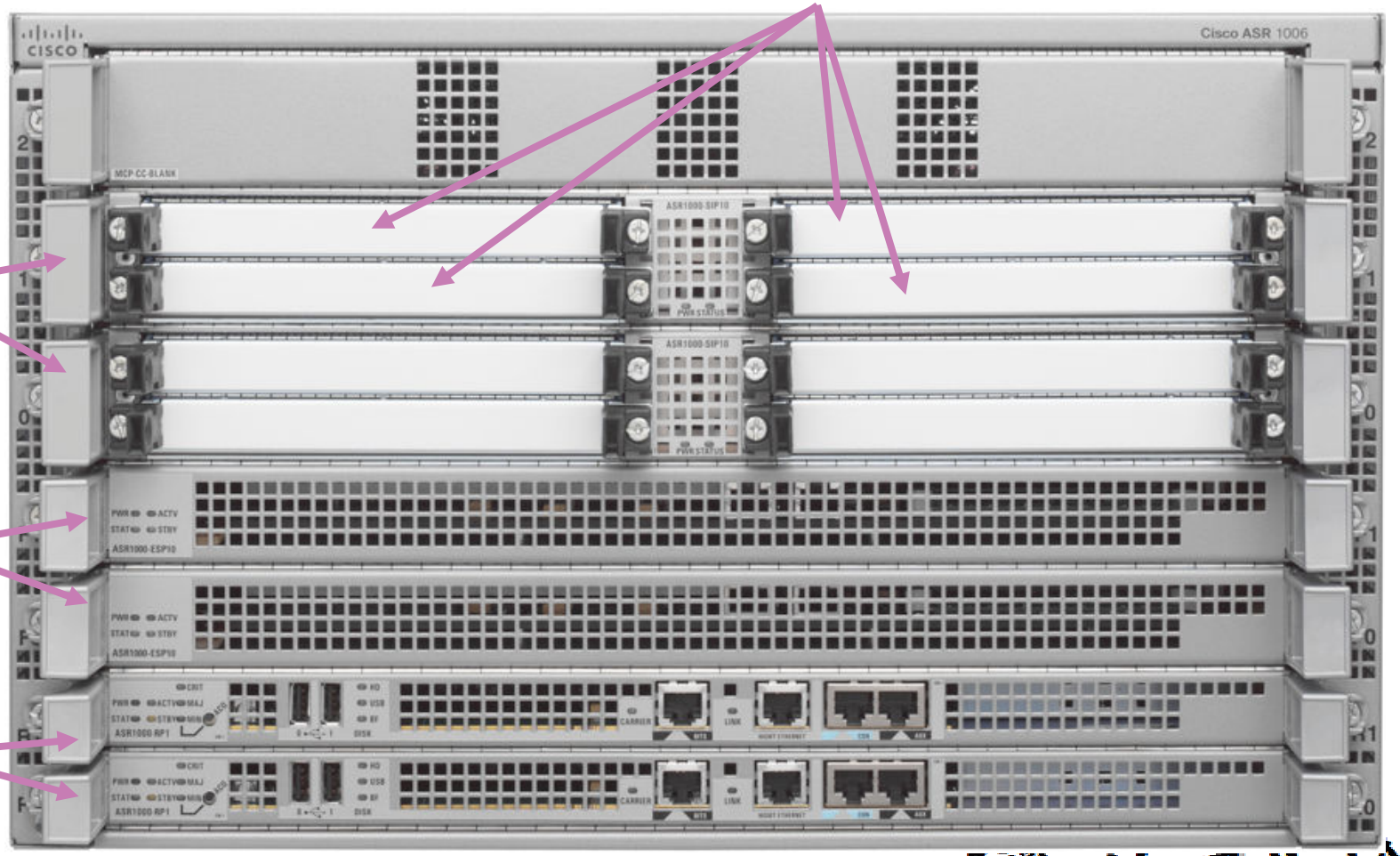
FP = Forwarding Processor now known as the Embedded Services Processor (ESP)

FRU = Field Replaceable Unit

RP = Route Processor

IOSd = IOS daemon, IOS process running on RP

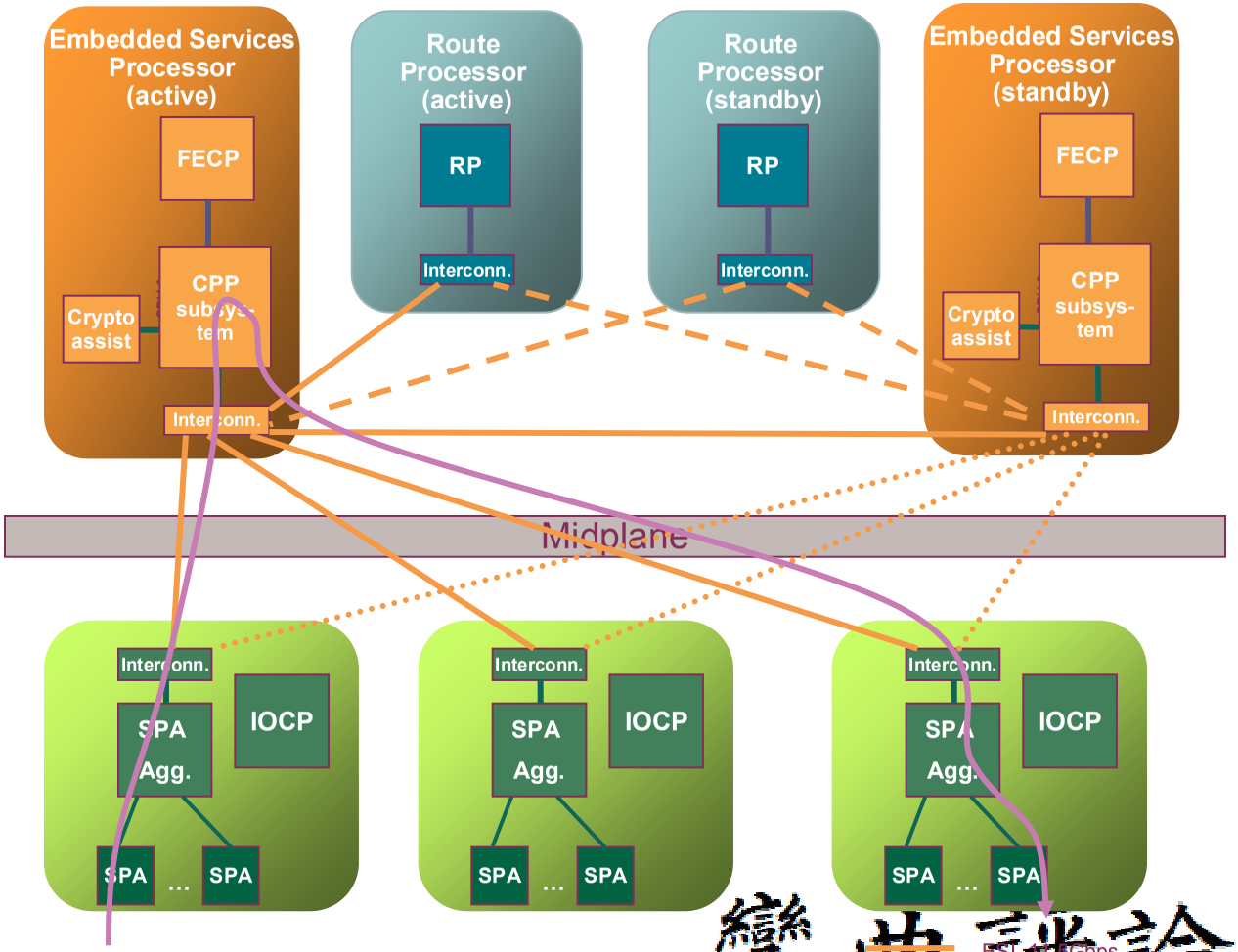
Chassis Options: ASR1006



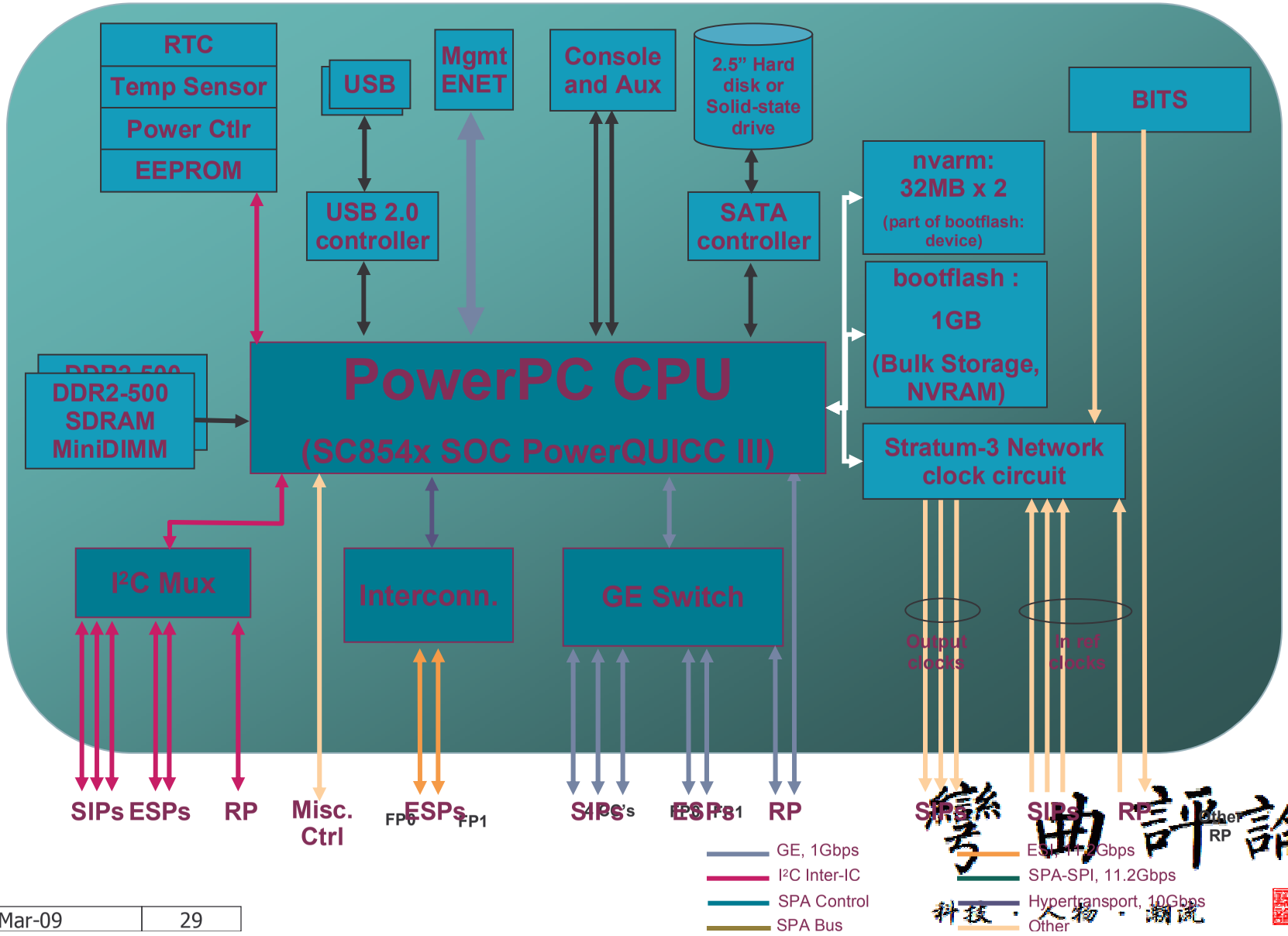
Rack Mounts and Cable Mgt not shown

System Architecture - Dataplane

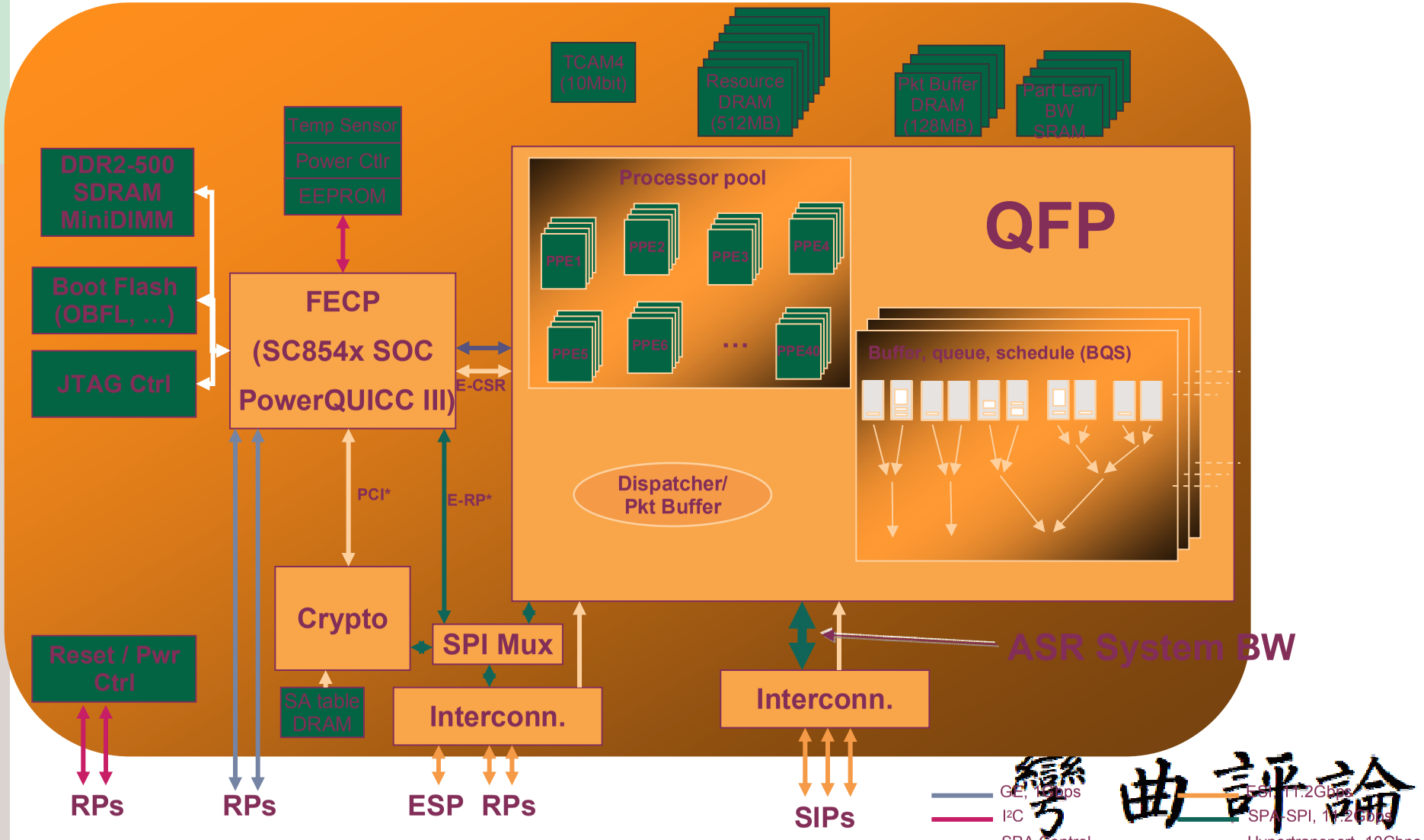
- All data forwarding is through FP
- Exception: Punt path for Legacy protocols – handled by the RP
- Interconnect ASIC in each of the functional elements provides the backplane connection through ESI links
- ESI (Enhanced Services Interface) links are used for Data forwarding
- SPA-SPI links connect to the backplane through the SPA-Agg ASIC



RP1 Block Diagram



ESP10 Block Diagram



SIP10 Block Diagram