数据为中心的FPGA加速器技术
—— 可编程性

演讲嘉宾　吕高锋

二〇二一年九月

计算和网络

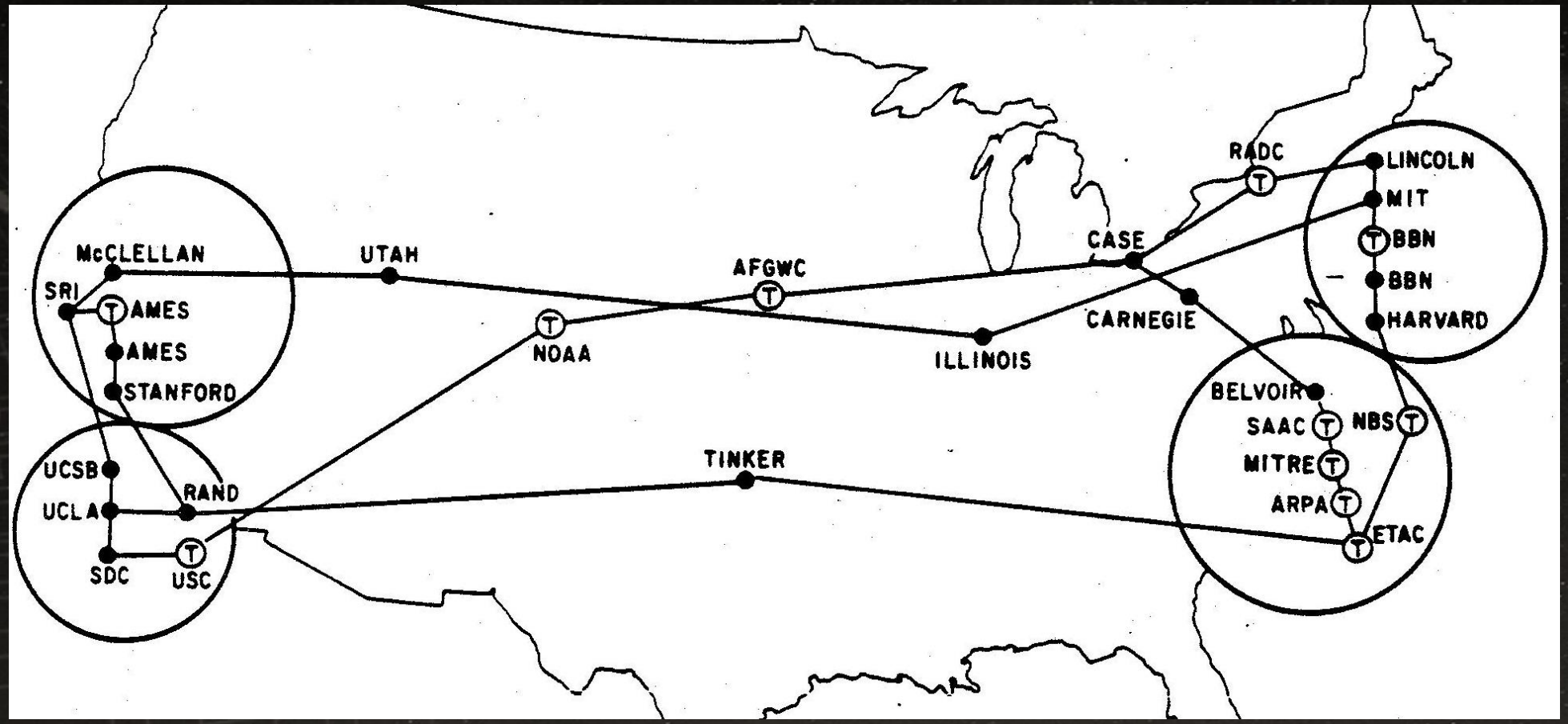
FPGA加速器

应用场景

讨论

一、计算和网络

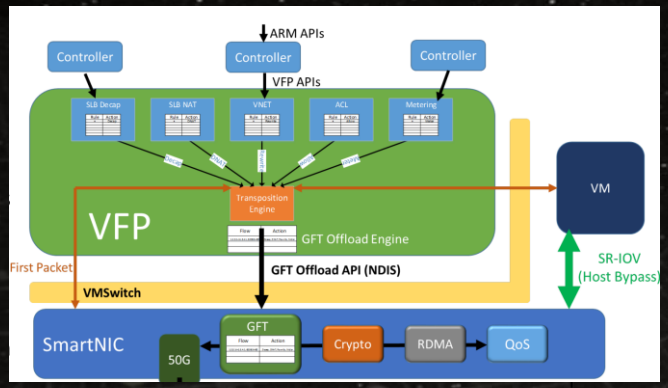
ARPANET



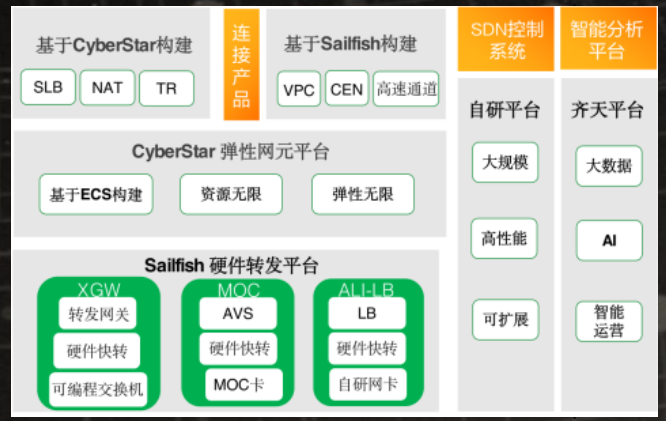
一、计算和网络

云计算和数据中心网络

Network Interface Controller

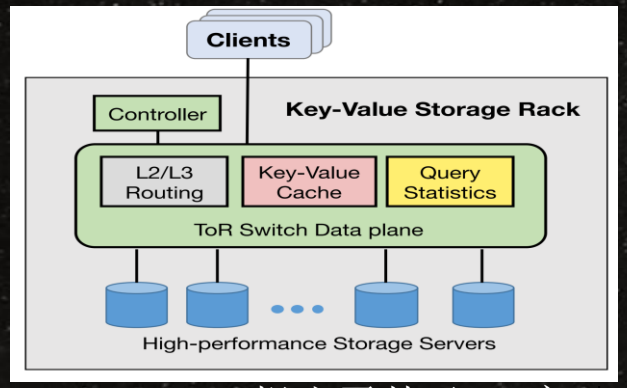


微软HostSDN设计了基于FPGA网卡的VFP

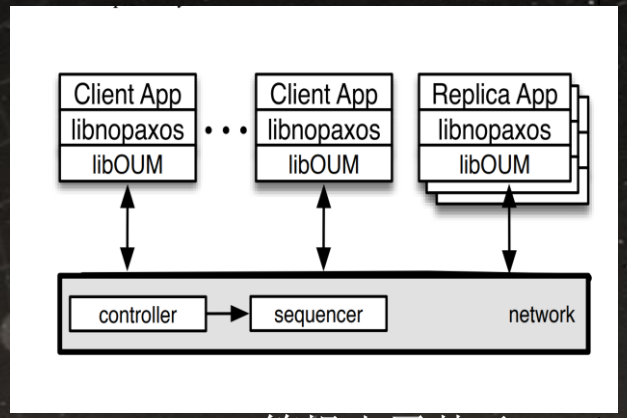


阿里Sailfish设计了基于FPGA加速器的云网关

In-Network Computing

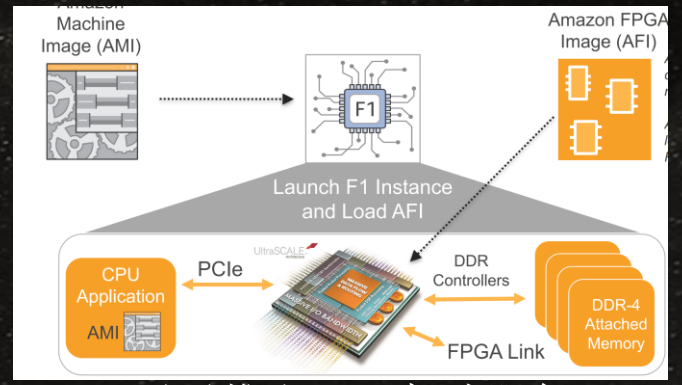


NetCache提出了基于ToR交换机的key-value加速

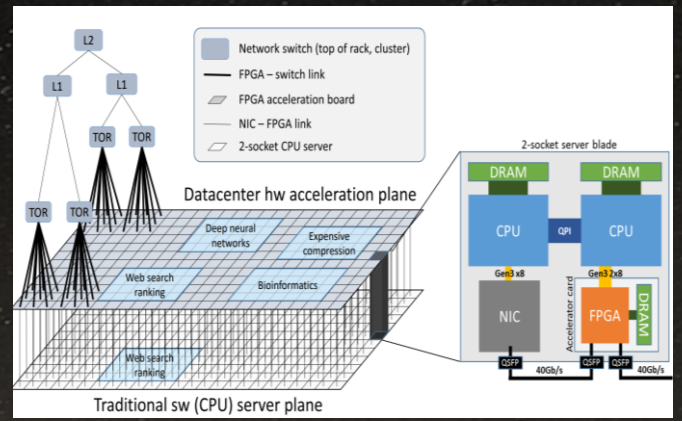


NetPaxos等提出了基于ToR交换机的同步操作加速

Distributed Applications



亚马逊推出FPGA加速服务F1，作为EC2实例提供给第三方

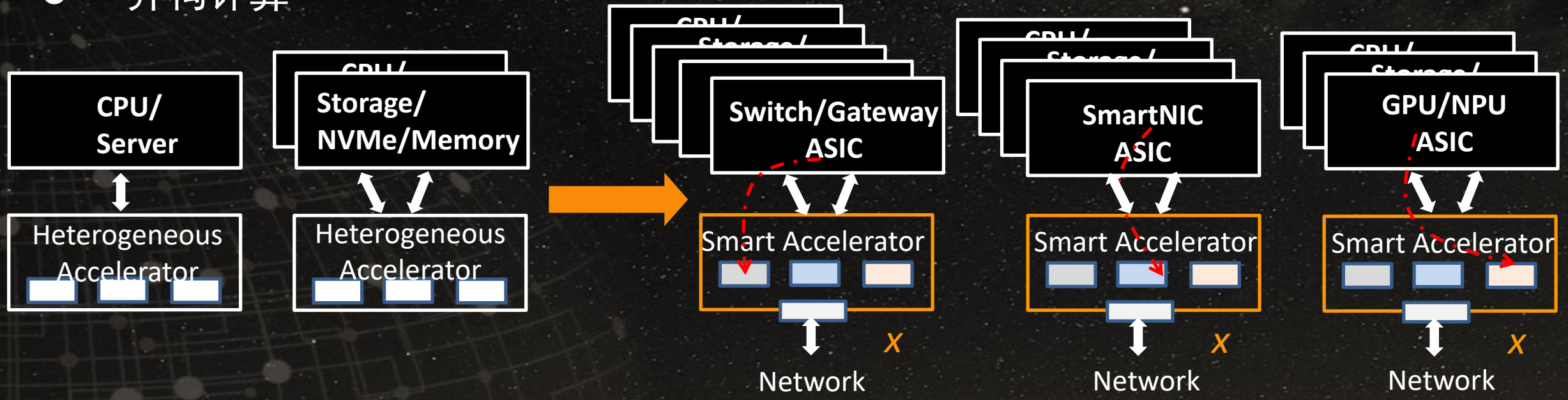


微软Azure设计了基于FPGA的Catapult对Brainwave加速

一、计算和网络

计算模型与网络模型

- 冯诺伊曼体系架构瓶颈
- “异构计算”



X Accelerator as a microService plaTform

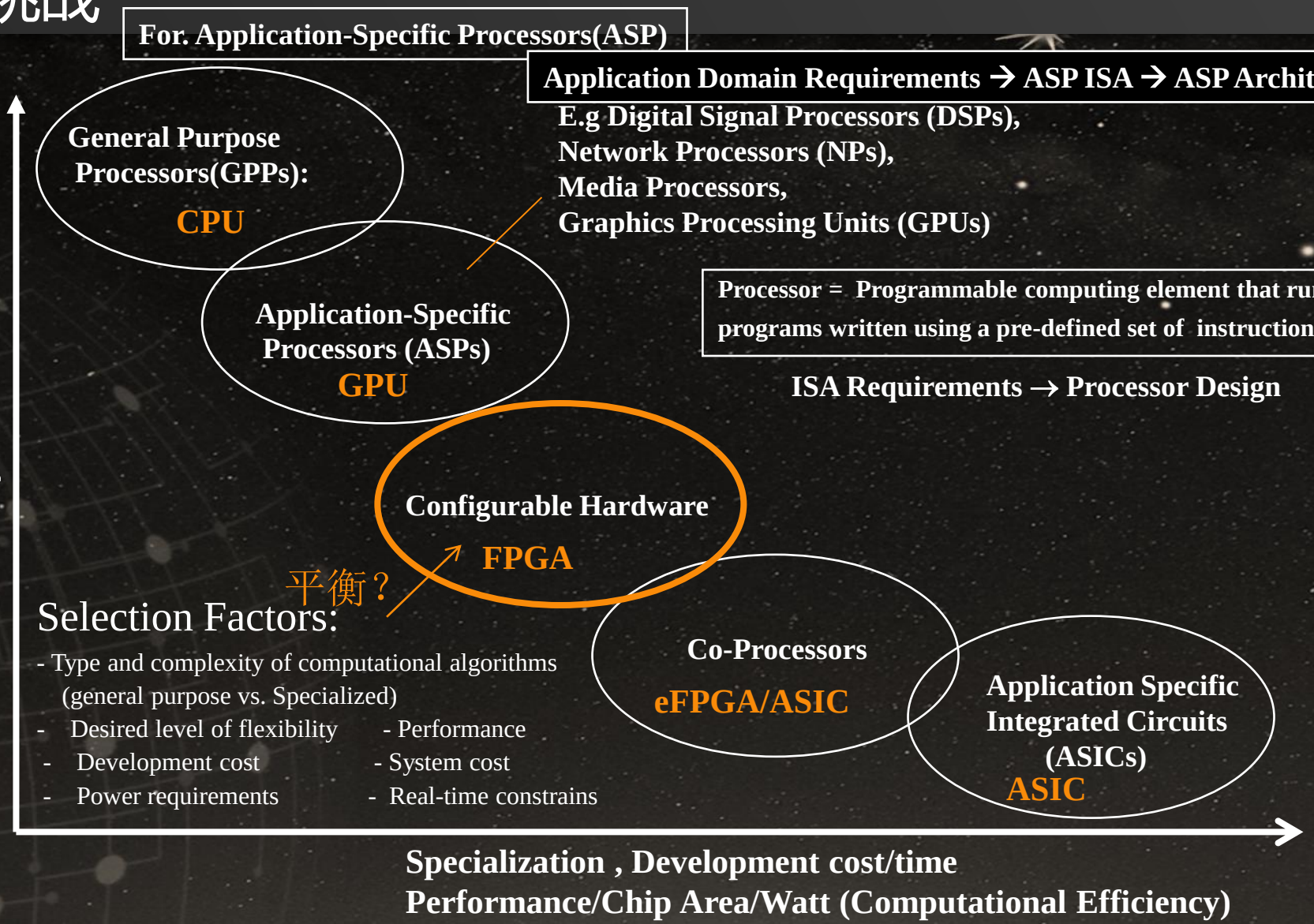
泛在网络加速器*x*

Fpga Accelerator as a microService plaTform

一、计算和网络

网络加速器x挑战

- 分组数据处理
  - 无统一模型
- 衡量标准
  - 可编程
  - 性能

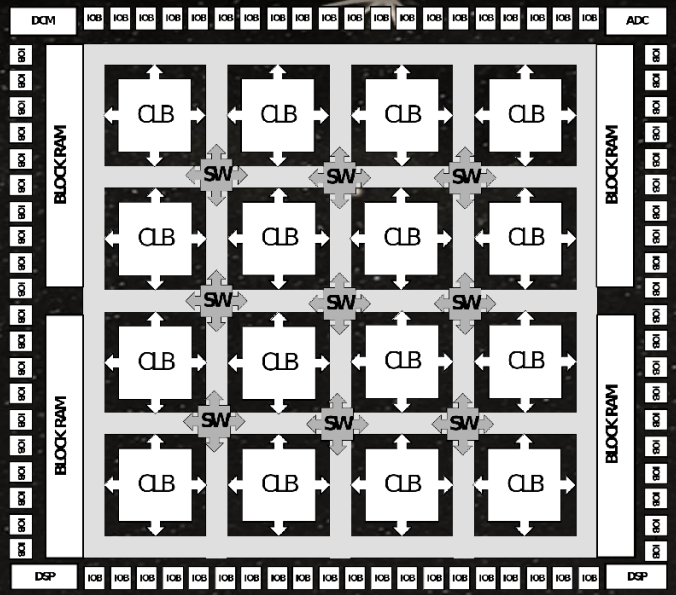


一、计算和网络

FPGA效能分析

	Zynq SoC Z-7020	Zynq SoC Z-7100	Artix-7 200T	Kintex-7 480T	Virtex-7 690T
Price (approx.)	100 €	3,000 €	190 €	2,500 €	11,200 €
Dual ARM® Cortex™-A9 MPCore	Yes	Yes	-	-	-
Programmable Logic Cells	85,000	444,000	215,360	477,000	693,120
Programmable DSP Slices	220	2,020	740	1,920	3,600
Peak DSP Performance	276 GMACs	2,622 GMACs	929 GMACs	2,845 GMACs	5,335 GMACs
Processing Power - single	180 GFLOPS	1,560 GFLOPS	648 GFLOPS	1,800 GFLOPS	3,120 GFLOPS
Technology	28 nm	28 nm	28 nm	28 nm	28nm
PCIe Interface	-	x8 Gen2	x4 Gen2	x8 Gen2	x8 Gen3
Power Consumption	2.5 W	20 W	9 W	25 W	40 W
Price Efficiency - single	0.56 €/GFLOPS	1.92 €/GFLOPS	0.29 €/GFLOPS	1.39 €/GFLOPS	3.59 €/GFLOPS
Power Efficiency - single	72 GFLOPS/W	78 GFLOPS/W	72 GFLOPS/W	72 GFLOPS/W	78 GFLOPS/W

FPGA加速性能评估:(定点MAC+单精度浮点性能）/功耗



General Structure of FPGA

		Nvidia GeForce GT 730	AMD Radeon R7 360	Nvidia GeForce GTX 970	Sapphire Radeon R9 390	Radeon R9 390X	Sapphire Radeon R9 Fury X	Nvidia GeForce GTX 980 Ti
Price (approx.)		80 €	120 €	250 €	400 €	420 €	600 €	700 €
Processing Power	Single	693 GFLOPS	1,612 GFLOPS	3,494 GFLOPS	5,120 GFLOPS	5,913 GFLOPS	7,168 GFLOPS	5,632 GFLOPS
	Double	32 GFLOPS	100 GFLOPS	109 GFLOPS	640 GFLOPS	739 GFLOPS	448 GFLOPS	176 GFLOPS
Technology		28 nm	28 nm	28nm	28nm	28nm	28nm	28nm
GPU		GK208 (Kepler)	Tobago (GCN 1.1)	GM204 (Maxwell)	Grenada (GCN 1.1)	Grenada (GCN 1.1)	Fiji (GCN 1.2)	GM200
Core Clock		902 MHz	1050 MHz	1050 MHz	1000 MHz	1050 MHz	1050 MHz	1000MHz
Power Consumption *Stress Test*		93 W	100 W	242 W	323 W	363 W	358 W	250 W
Price Efficiency		0.10 €/GFLOPS	0.07 €/GFLOPS	0.07 €/GFLOPS	0.08 €/GFLOPS	0.07 €/GFLOPS	0.08 €/GFLOPS	0.12 €/GFLOPS
Power Efficiency		7 GFLOPS/W	16 GFLOPS/W	14 GFLOPS/W	16 GFLOPS/W	16 GFLOPS/W	20 GFLOPS/W	23 GFLOPS/W

GPU加速性能评估:依据手册 @2016

	Titan Xp	BW_S10
Numerical Type	Float32	BFP (1s.5e.2m)
Peak TFLOPS	12.1	48.0
TDP (W)	250	125
Process	TSMC 16nm	Intel 14nm

2018 微软BrainWave @Deepbench

一、计算和网络

FPGA可编程性

- 硬件描述语言HDL
  - 寄存器级操作RTL的状态机描述
  - 实现效率高、较为复杂
- 高级编程语言
  - SystemC，C等
  - 开发简单，效率尚可
- 领域定制语言DSL
  - P4（Programming Protocol-independent Packet Processors）
  - Chisel等语言
  - 开发简单，学习过程长
- 粗粒度可重构体系架构CGRA
  - 面向FPGA资源，研究任务分解和映射
  - 缺乏网络分组处理基础库
- ETH FPGA OS
  - 提出了基于microkernel的FPGA资源虚拟化方法
  - 设计了模块交互和数据交换核心
  - 缺乏网络分组处理基础库
- Microsoft ClickNP
  - 提供了丰富了网络分组处理基础库
  - 没有设计应用加速

出发点

计算和网络

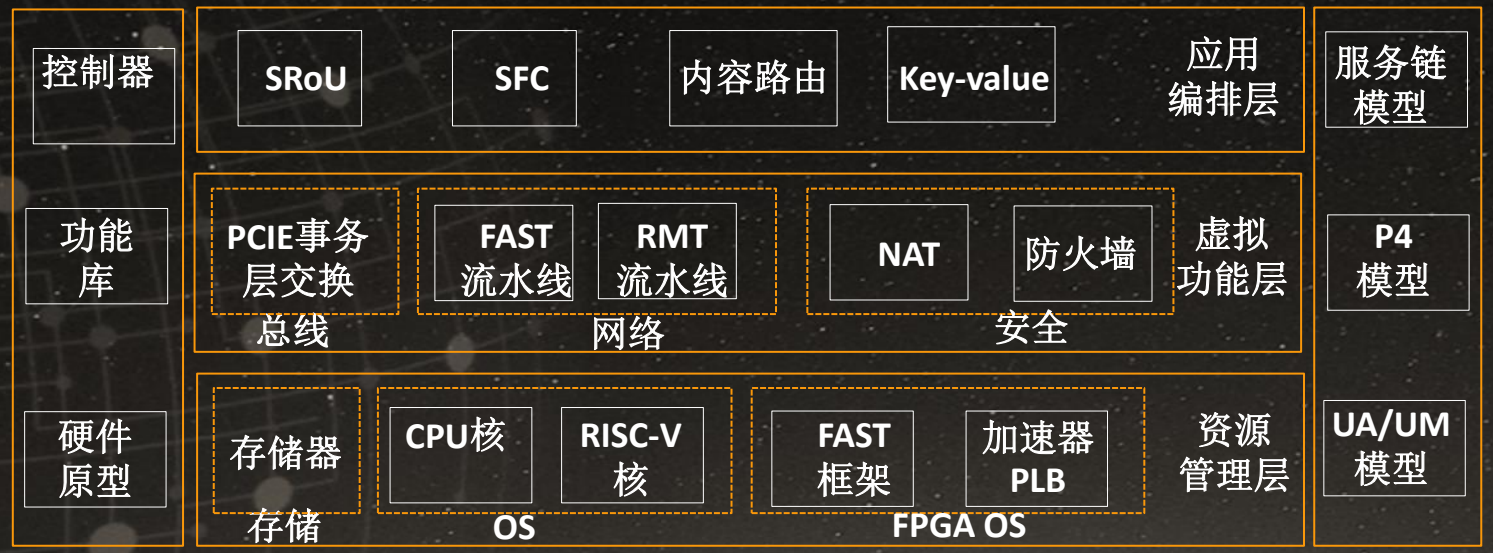
FPGA加速器

应用场景

讨论

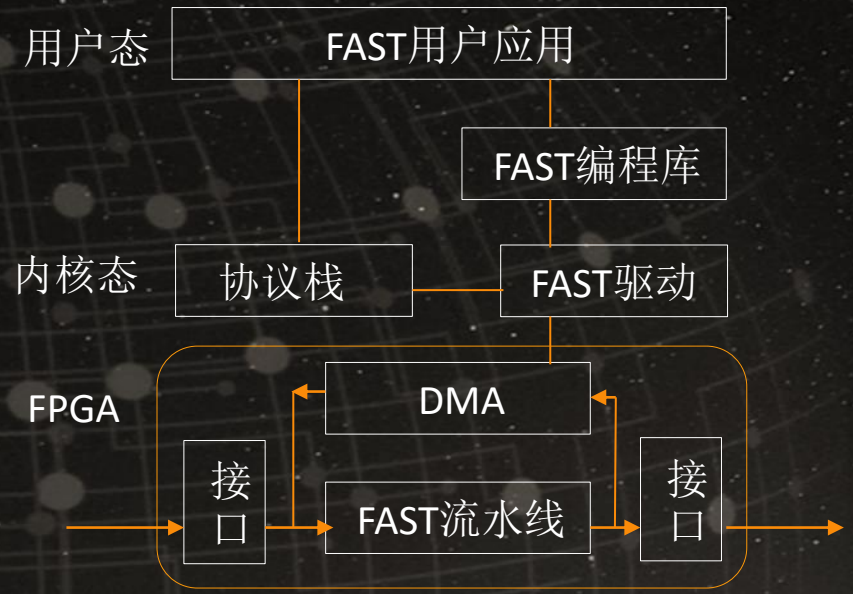
泛在网络加速器架构FIA

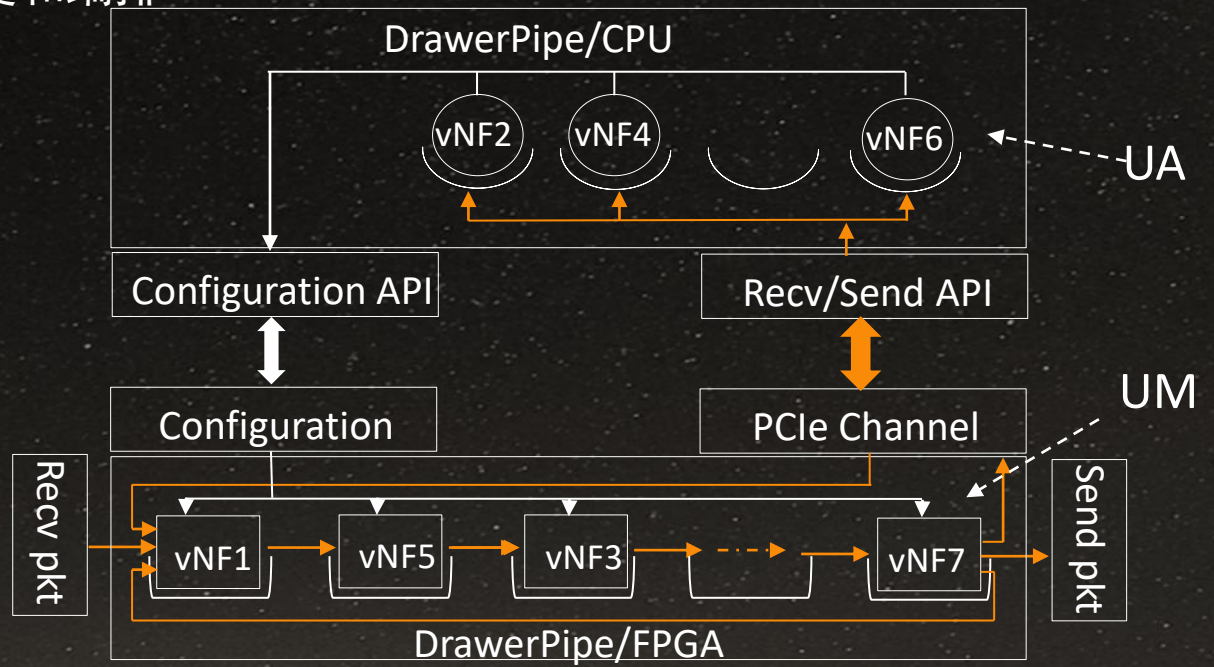
- 云-网-边协同的算力网络、在网计算的Dis-aggregation数据中心网络、以及异构计算系统总线中计算和网络协同和融合，提出了泛在网络加速器架构FIA
  - 资源层：包括FPGA和CPU等可编程资源，异构计算，软硬件协同的网络处理框架
  - 功能层：可重构通用分组处理流水线RDP与可编程数据深度处理器PDP架构，包括基础库和架构
  - 应用层：用户面功能UPF开发库和控制模型
- 解决算网融合下网络编程及性能扩展问题，为网络功能卸载、用户功能加速提供运行环境
- 赋能端/智能网卡、网/交换机、边/网关的功能，构成泛在的网络加速器服务



模型级：软硬件协同的开发模型

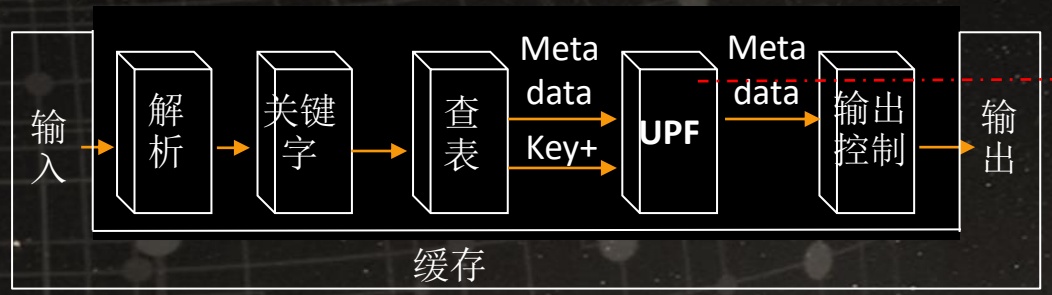
- FAST（ *F*pga *A*ccelerating micro*S*ervice pla*T*form ）架构
  - 基于进程隔离的轻量级虚拟化技术，运行软件化vNF
  - 基于FPGA OS的FPGA虚拟化技术，构建硬件流水线化vNF
  - 基于分组元数据（数据）和统一虚拟地址空间（控制），支持vNF协同
  - 基于统一编号与路由，支持软硬件vNF一体化调度和编排



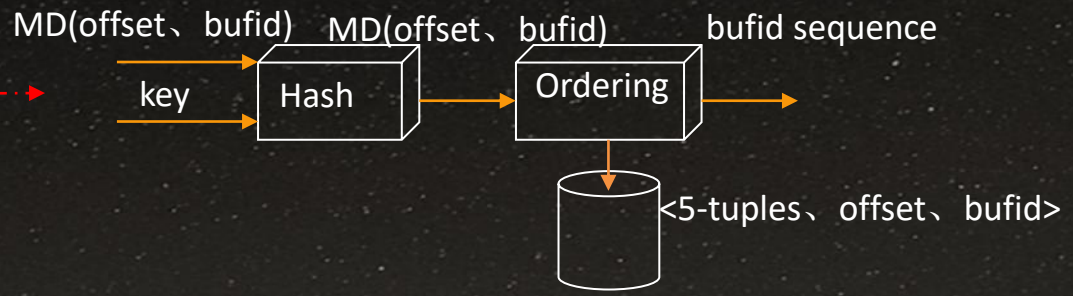


寄存器级：FPGA OS及开发框架

- 针对网络协议栈卸载和应用加速，提出了泛在网络加速器抽象FPGA OS，支持用户面功能UPF运行Runtime
  - 泛在网络加速器抽象FPGA OS：提供了DMA、网络接口、数据缓存、控制总线等外围基本模块
  - 用户面功能模块开发框架：提供关键字、时间戳、Buffer索引等Metadata信息，支持对报文内容的操作
  - Metadata：标识vNF，实现FPGA OS与UPF参数、中间结果交互
- 为实现近数据计算提供计算、网络和存储等资源，将智能网卡功能加速从网络协议栈扩展到了特定的用户面功能



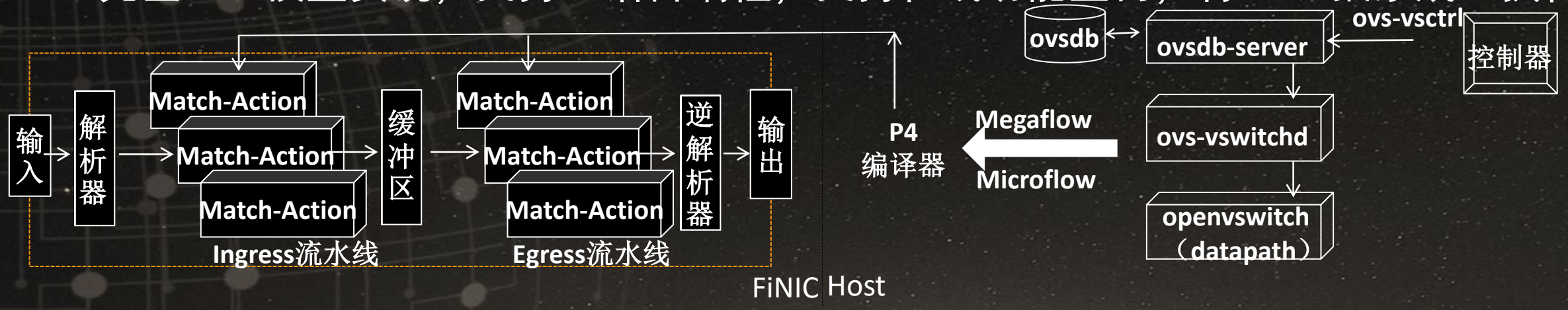
泛在网络加速器
FPGA OS



QUIC数据还原

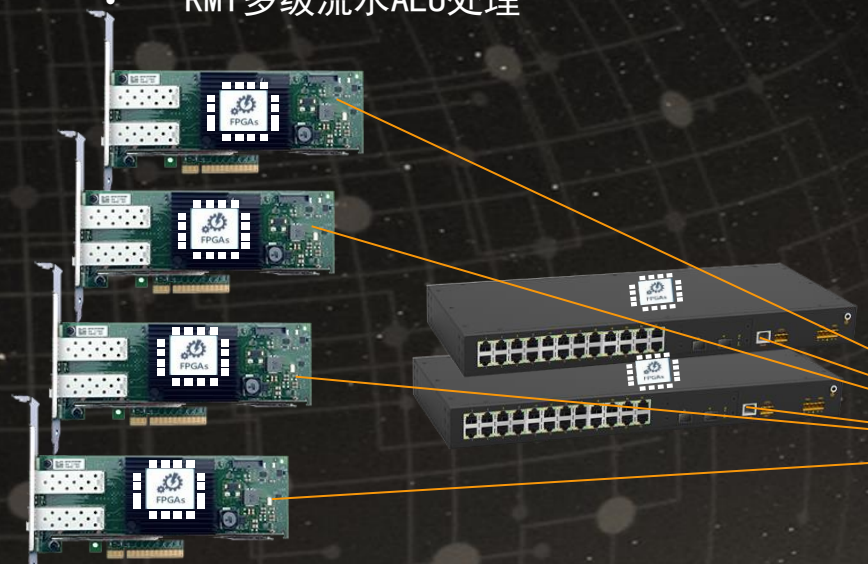
指令级：可编程数据深度处理器RMT+/P4

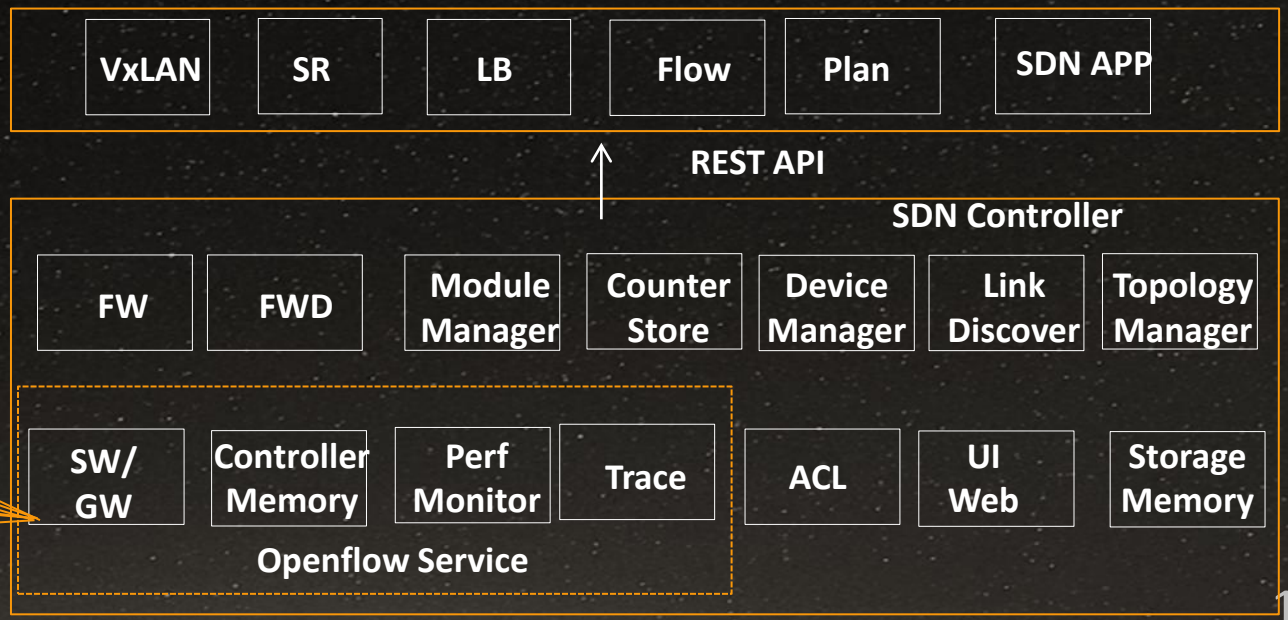
- 针对无状态和有状态的协议处理差异，协议无关的分组处理方式成为基础，设计了可编程的硬件处理逻辑RMT，支持在线功能重构，线速处理
  - 协议无关的解析引擎，包括TLV表示的协议状态转换表，关键字提取
  - 查表匹配引擎，基于CAM的带掩码的查找
  - 交叉开关，关键字等Metadata与ALU的通路
  - ALU，支持加、减、移位等基本运算
  - 逆解析器，报文头选项的编辑，报文头与报文体的合并等
- 完全RMT模型实现，支持P4语言编程，支持在线功能重构，将FPGA抽象成P4执行器



模块级：控制和编排器

- 泛在网络加速器控制和功能编排
  - 异构网络设备统一管理，支持端到端的管理
    - 基于开源控制器，设计了软件定义可重构智能网卡、SDN交换机、超融合网关的统一的抽象层，支持软件定义可重构智能网卡、SDN交换机和网关统一管理，构成泛在的网络加速服务层
    - 统一的设备状态数据库、网络拓扑显示界面
    - 支持二层、三层路由转发，支持ACL，支持负载均衡等功能应用
  - 多元网络功能统一调度，支持动态赋能
    - 基于FAST的软硬件功能协同
    - RMT多级流水ALU处理





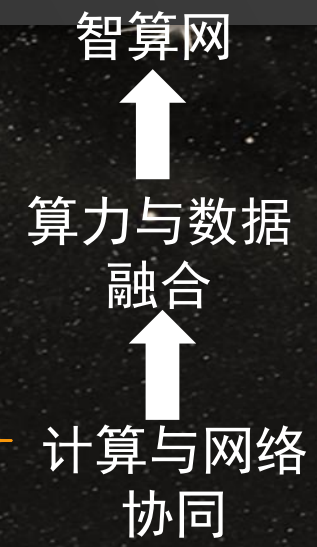
应用开发：基于加速器的程序开发接口/C/C++

- 控制抽象/OF
  - 泛在加速器中RMT实现了软件交换机OVS卸载，提供了Openflow流表的抽象
- 应用接口/DPDK
  - 泛在加速器实现了系统网络协议栈的卸载和加速，采用旁路内核的方式，将数据直接上送到用户态，实现零拷贝数据收发
    - 内核态实现报文rx/tx缓冲区的申请和映射
    - 基于dpdk框架实现rx/tx缓冲区的管理
    - 兼容dpdk网络数据处理库
  - 用户态数据零拷贝收发，减少中断处理等对服务器影响

二、 FPGA加速器

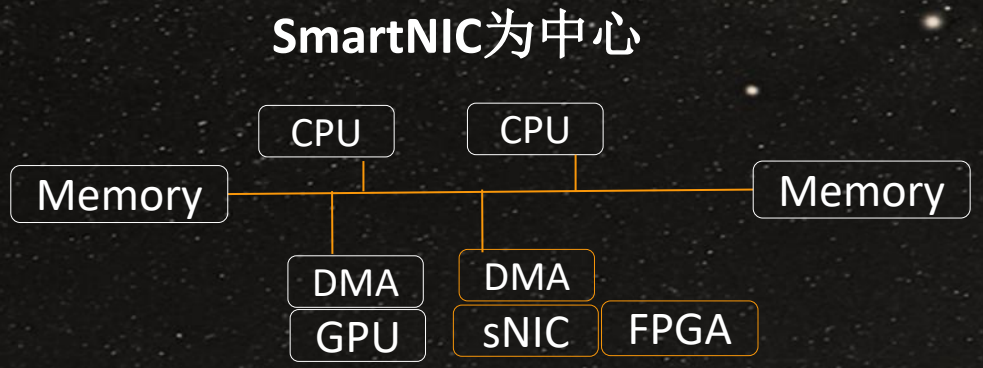
特点

- 多层次抽象和编程库、多领域开发模型
  - 支持寄存器传输级RTL、以及指令级ALU多种层次的硬件资源封装
  - 提供硬件加速框架FAST、可重构RMT/P4等开发模型
- CPU融合，软硬件协同
  - 加速器FPGA内部可编程硬件与多核RISC-V协同
  - 服务器中加速器FPGA与宿主机CPU协同
  - 支持异构加速，支撑功能软件化虚拟化，支持动态赋能
- 数据驱动，泛在计算
  - 网络流驱动数据流，分组操作支撑数据处理
  - 支持数据就近处理，实现应用在网计算，支持泛在计算



远景：以数据为中心

云计算、边缘计算、雾计算是数据驱动，按需部署



异构计算：NIC进行调度，其他为协处理器

计算和网络

FPGA加速器

应用场景

讨论

三、应用场景

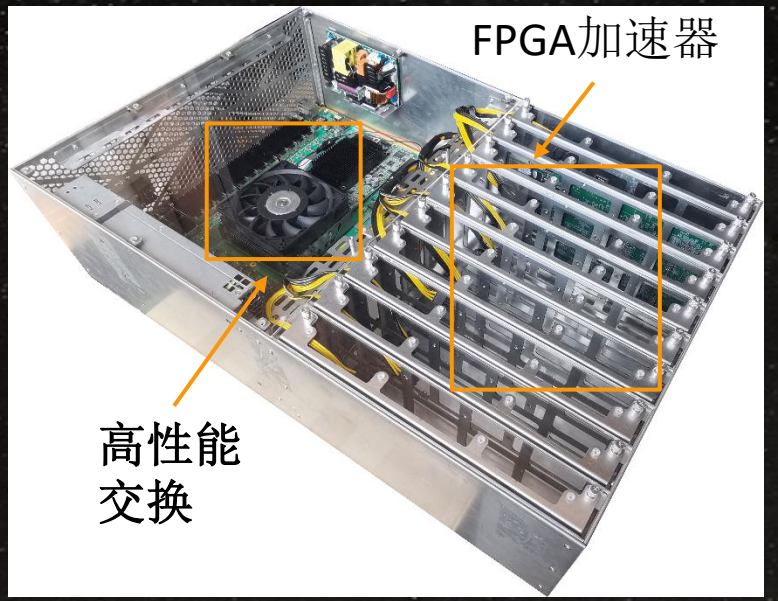
1、网卡

- 软件定义智能网卡原型
  - 云数据中心应用种类繁多，针对云-端通信，适配了主流的网络传输层协议QUIC
    - 对多种QUIC实现版本进行移植，和性能测试
    - 分析QUIC性能瓶颈，分片重组和加解密
  - 面向云数据中心多租户应用，基于智能网卡原型，部署了k8s虚拟化环境
    - 容器虚拟网络二层交换、三层转发功能
  - 与系统协议栈、云计算平台等具有良好的兼容性
  - 支持软件定义边界扩展到服务器第一跳

三、应用场景

2、数据中心网关

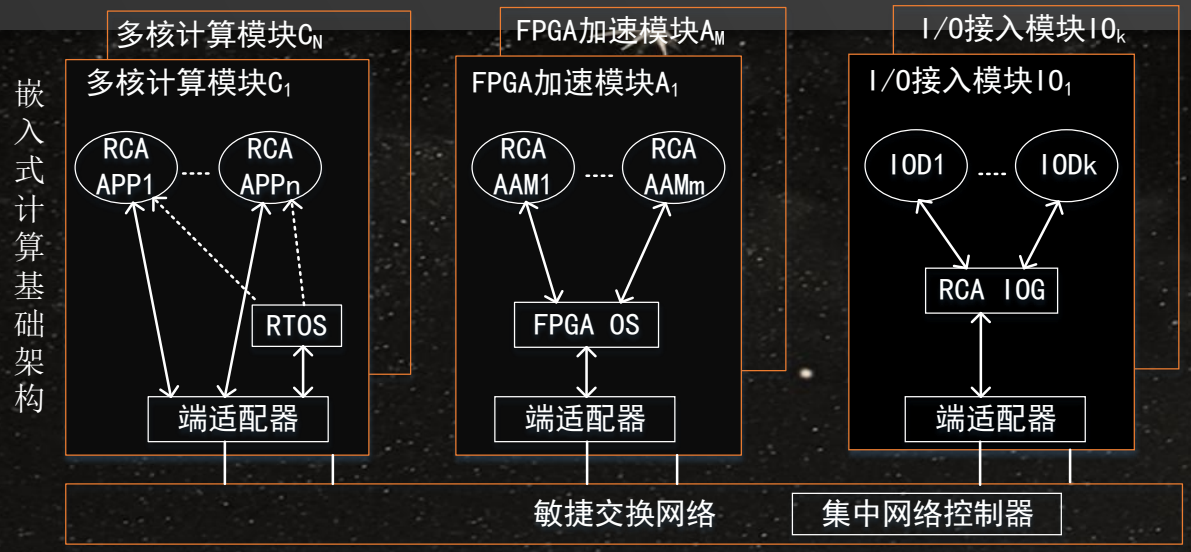
- 超融合安全网关
  - 以自主可控高性能交换芯片为前端分流器，泛在网络加速器FPGA作为业务运行平台，支持网关功能加速
    - 数据中心门户业务负载均衡、NAT等
    - 流量压缩解压缩、加解密、清洗等
    - 应用防火墙等功能卸载
  - 基于加速器FPGA开发框架对网关进行重构和功能扩展
    - 基于加速器FPGA开发新型网关功能
      - 隐蔽信道检测、地址跳变等
  - 与智能网卡共同支撑数据中心网络纵深安全

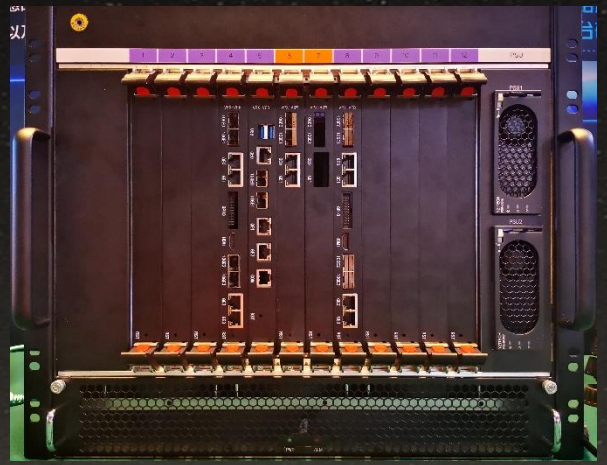


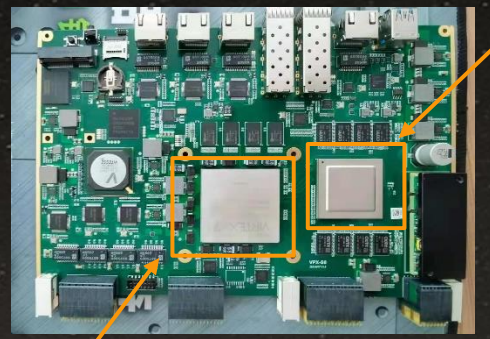
三、应用场景

3、异构计算

- 异构可重构计算架构
  - 计算、加速、I/O基于敏捷交换解耦
    - 易于资源扩展、统一管理、开发集成
- 计算：轻量级虚拟化FAST UA
  - 类Docker的虚拟化技术
  - 支持自定义功能的卸载和负载均衡
- 加速：泛在网络加速器FPGA即服务
  - 泛在网络加速器FPGA OS
  - 加速与卸载等算力提升的重要手段
- 互连：低延迟时间触发通信
  - ItRDMA和ItDMA



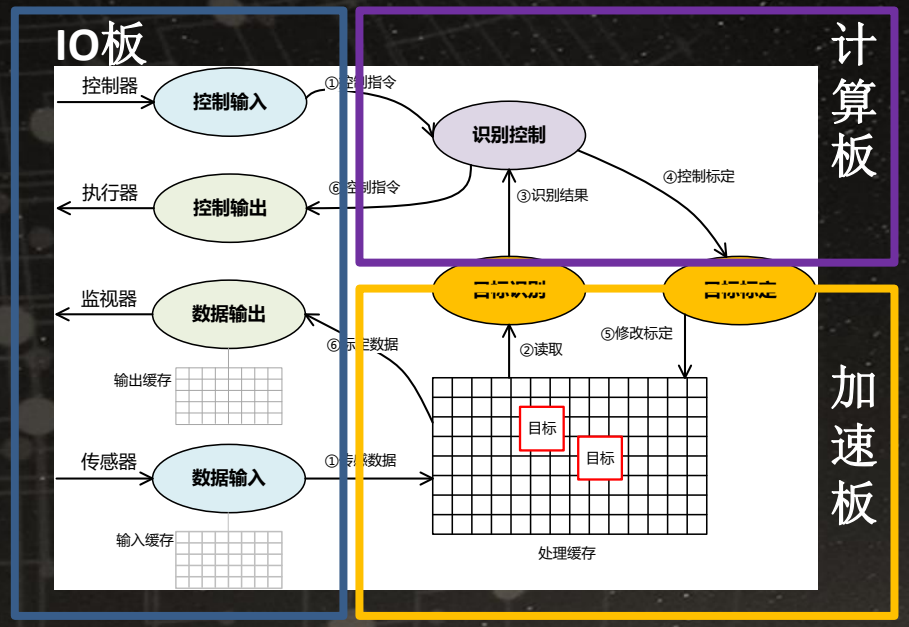




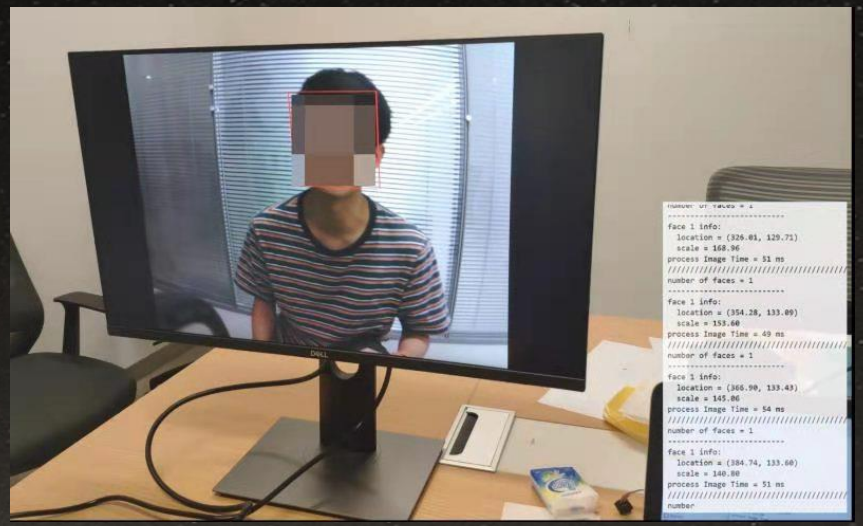
三、应用场景

3、异构计算

- 目标识别
  - 输入处理：数据到缓存的导入以及控制指令的输入
  - 计算处理：目标识别过程控制、目标识别计算以及标定
  - 输出处理：数据输出以及控制指令输出



- 人像标定应用验证
  - 输入：摄像头
  - 输出：显示器
  - 计算：目标标定PICO算法



计算和网络

FPGA加速器

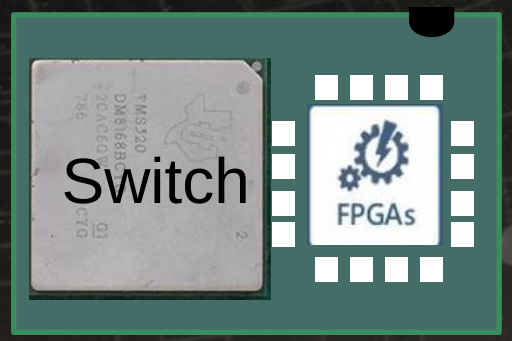
应用场景

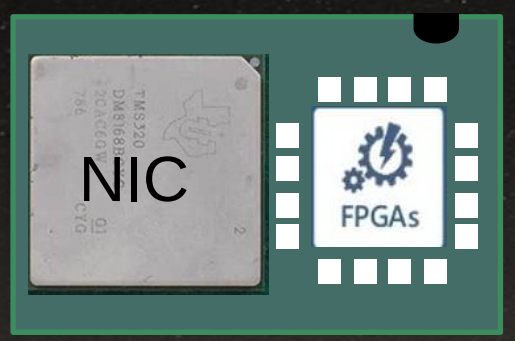
讨论

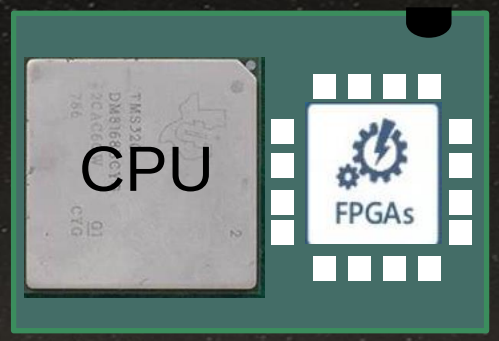
四、讨论

下一步

- 开放泛在网络加速器FPGA平台
  - 开发基于RMT的SRoUDP，以及基于RMT的RNN加速等项目
  - 打造CPU + FPGA + Switch/NIC自主可控生态
  - 作为粘合剂和鲶鱼，促进芯片发展
- 基于Chiplets的演进式DPU实现路线
  - FPGA功能验证，并加速DPU流片
  - 支持DPU等ASIC芯片部分功能重构，迭代







四、讨论

欢迎参与，敬请批评指正！