2021 中国智能网卡研讨会

CHINA SMARTNIC WORKSHOP

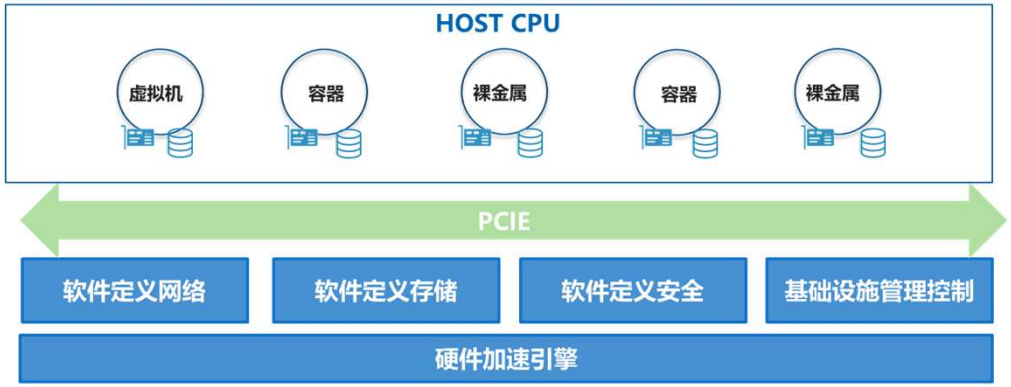
浪潮智能网卡的创新与实践

王昭峰

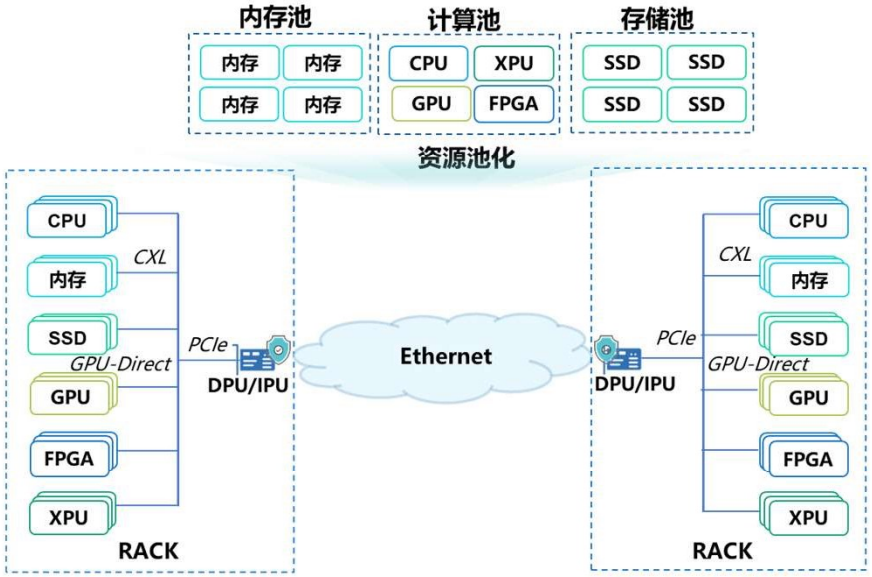
智能网卡使能云数据中心新架构

- 云计算的两大特性：**虚拟化**和**资源池化**；
- 智能网卡加速了基础设施的**虚拟化**和**资源池化**，云数据中心架构迈向**“Data-Centric”**和**“可组合”**；

基础设施硬件虚拟化

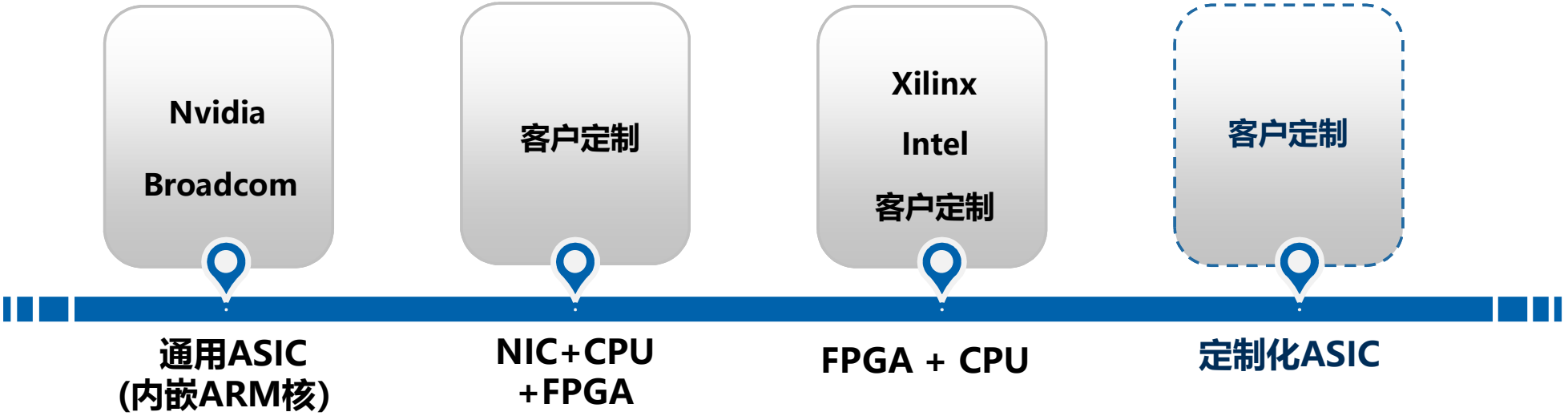


基础设施资源池化



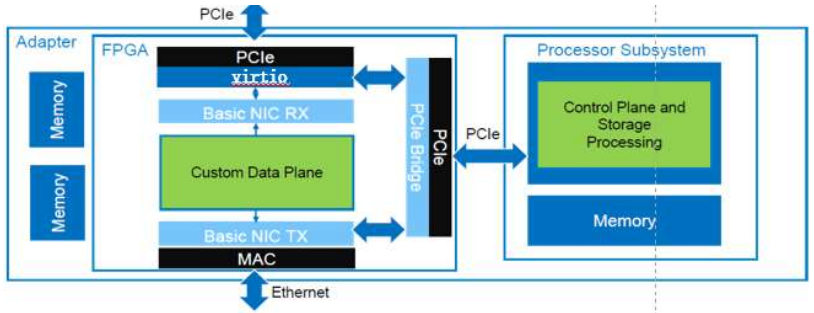
市场上主流的智能网卡硬件架构

- 硬件架构： **SoC 架构，NP架构， FPGA+CPU架构，通用ASIC架构**（内嵌ARM)和**定制化ASIC架构**；
- 产品形态：单卡、双卡，OCP卡；
- 方案选择：业务需求，基于性能，可编程，功耗和成本之间的平衡。



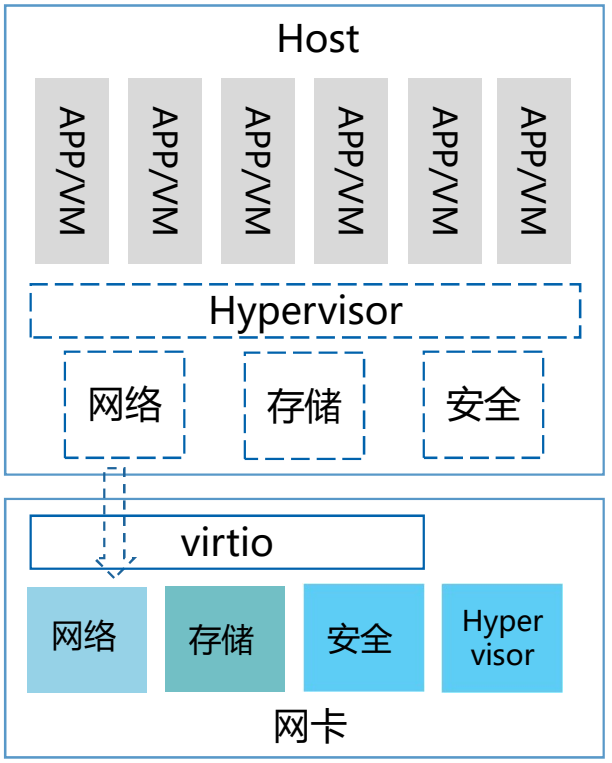
浪潮智能网卡解决方案

FPGA+CPU 架构，产品方案更灵活



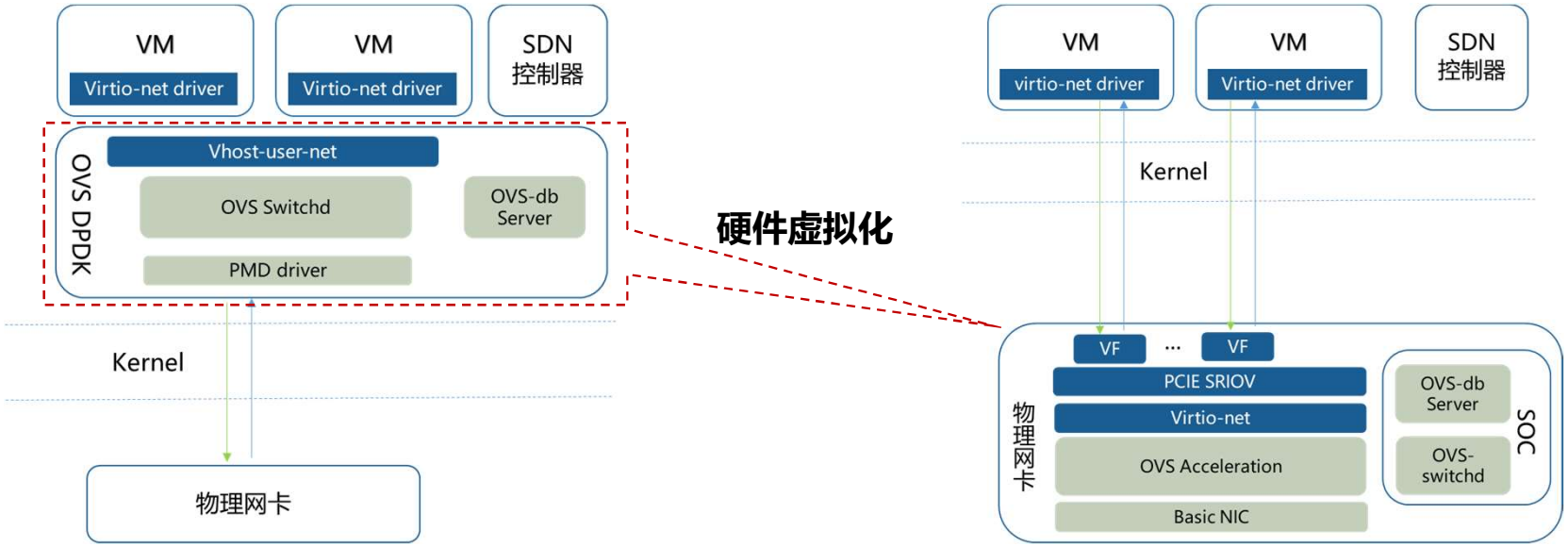
- **高性能**：FPGA提供了接近ASIC的处理能力，而X86则为异常处理、存储和安全业务提供了高速处理能力；
- **全可编程**：软、硬件全可编程，满足客户业务持续演进，保障硬件投资。

IO设备硬件虚拟化
网络、存储、安全硬件卸载



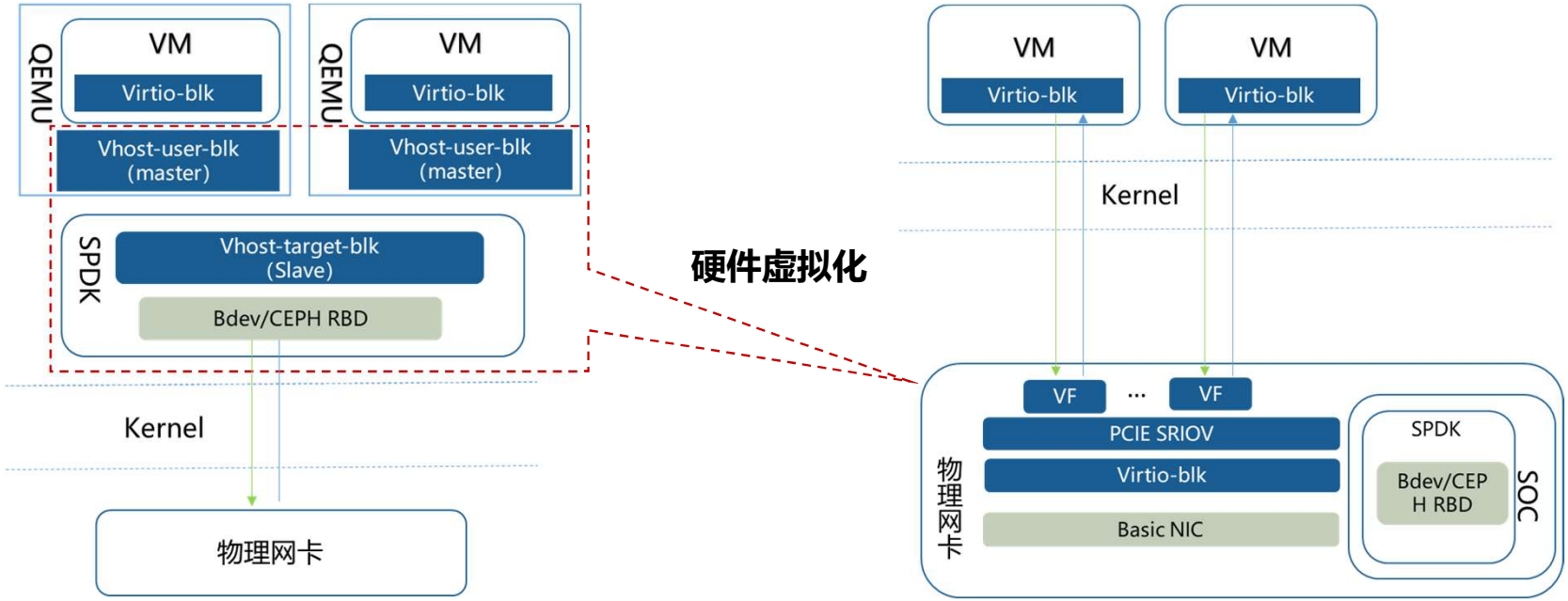
网络IO设备硬件虚拟化

- 接口卸载：硬件实现virtio-net；
- 任务卸载：数据面基于FPGA的可编程硬件网络包处理，SoC侧处理异常报文处理和控制面；
- 支持OVS，VLAN，VxLAN，Conntrack Table，以及IPSec等；
- 减少Host端CPU的负载，降低报文的延迟和抖动。



存储IO设备硬件虚拟化

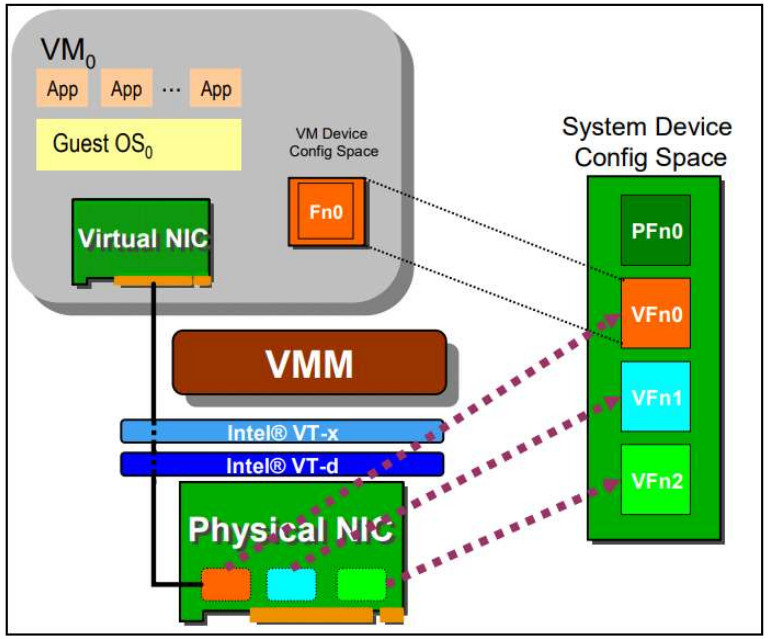
- 接口卸载：硬件实现virtio-blk；
- 任务卸载：SoC侧使能SPDK，代理控制指令和存储数据，并通过网络和远端盘交互；
- 支持块存储、NVMeoF，支持TOE加速；
- 减少Host端CPU的负载，支持裸金属、远端盘启动和云盘业务。



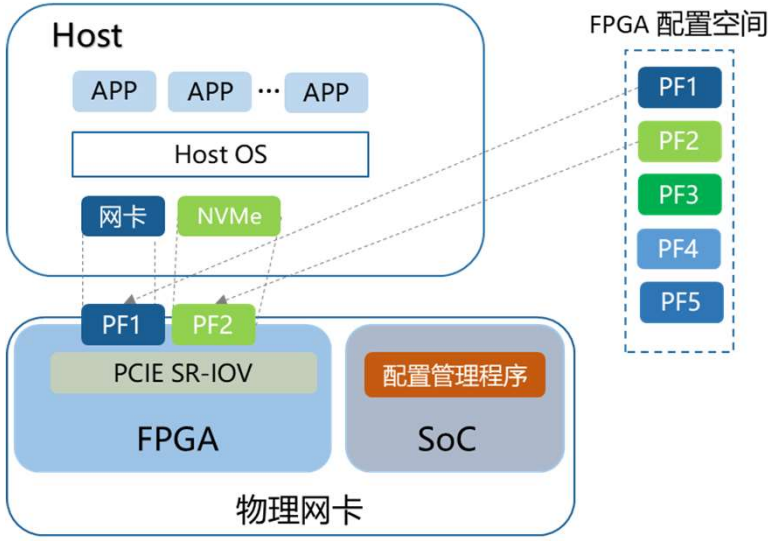
虚拟设备的管理和动态热插拔

- SR-IOV引入了两种PCIe的Function：PF和VF，通常对应着裸金属和虚拟机（VM）的应用场景；
- 在虚拟机场景下，VF的配置和管理由VMM完成，Guest OS需要支持VF的动态热插拔；
- 在裸金属场景下，PF的配置和管理由网卡SoC上管理程序负责，Host OS需要支持PF的动态热插拔。

VF的配置和管理



PF的配置和管理



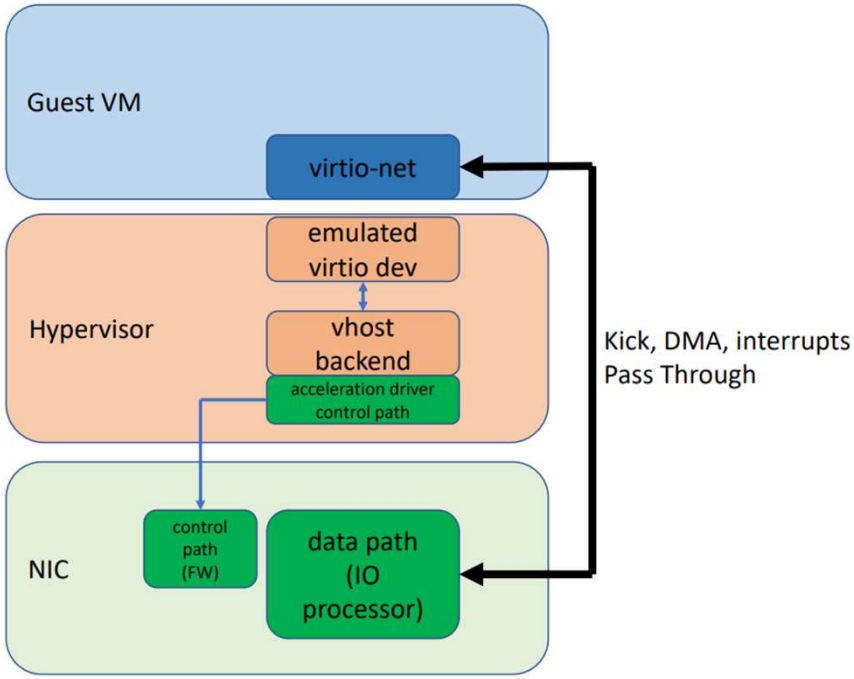
IO硬件虚拟化后的热迁移方案

热迁移的两个挑战：

- Hypervisor能够感知硬件设备状态
  - ✓ VDPA控制和数据平面分离，在监控设备状态同时，避免降低转发性能；
- 网卡能够跟踪迁移过中的脏页
  - ✓ 网卡硬件监控DMA页的跟踪，避免切换到软件处理引发迁移过程中的性能下降；

虚拟机热迁移的过程：

- 开启VM热迁；
- 阶段一，同步设备状态到目的VM；将所有内存页传送到目的VM；
- 阶段二，使能dirty page跟踪，网卡监控到内存被修改时，标记所在页为dirty page；传送dirty page到目的VM；
- 阶段三，停止源VM，同步网卡设备的vring状态到目的VM，如vring 的读写指针；
- 启动目的VM，热迁移结束。

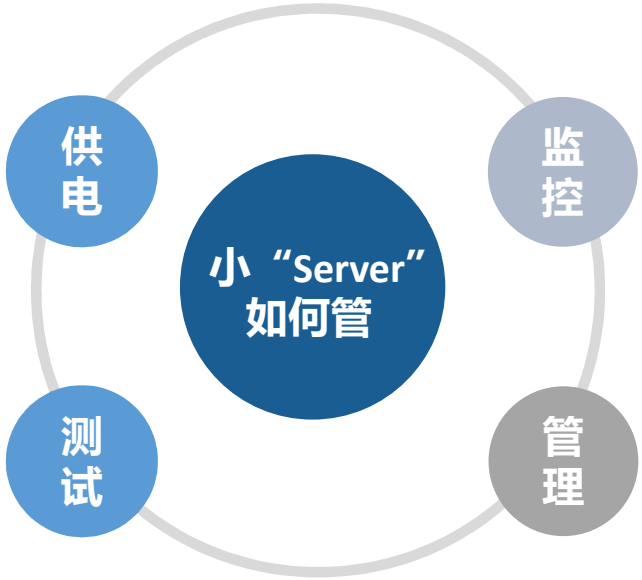


智能网卡与服务器的适配

智能网卡是大Server中的“小Server”

智能网卡如何供电，上电流程是什么，能否独立运行？

大“server”的部件，适配测试应该怎么做？



智能网卡如何监控，是否需要和服务器分开监控？

两个“Server”放在一起如何管理，相互之间如何配合？

智能网卡的供电和上电需求

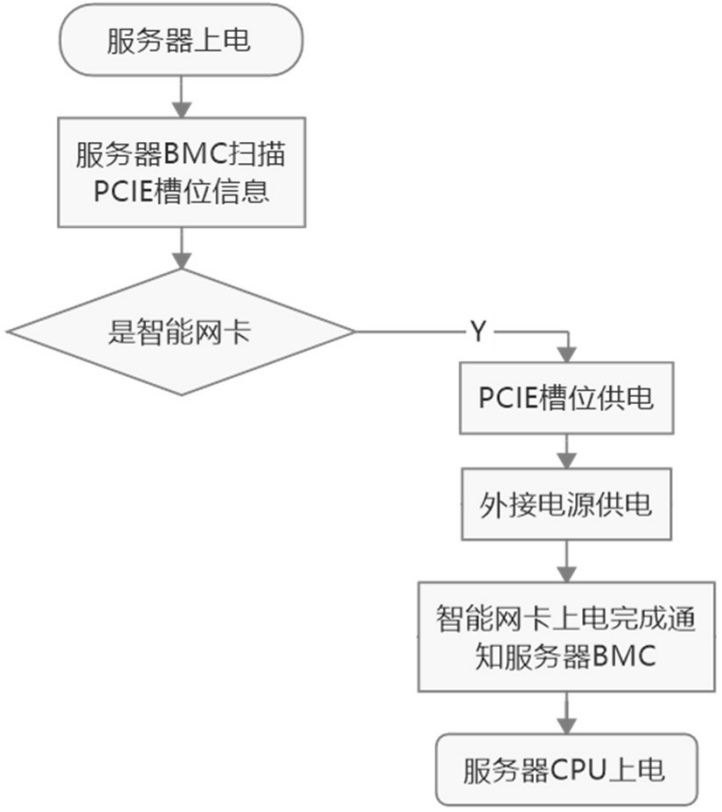
服务器运行智能网卡时，需要满足以下三个需求：

- ✓ 提供大于75W的供电满足高功率智能网卡
- ✓ 智能网卡需要在服务器CPU掉电时保持运行
- ✓ 智能网卡需要在服务器CPU上电前完成启动

供电设计

功耗	供电方式	独立供电方案
<=75W	金手指	通过智能网卡在位信号，可控制单独给智能网卡供电
>75w	金手指+外接电源	金手指与外接电源共同为网卡供电，同样根据智能网卡在位信号，可控制单独给智能网卡供电

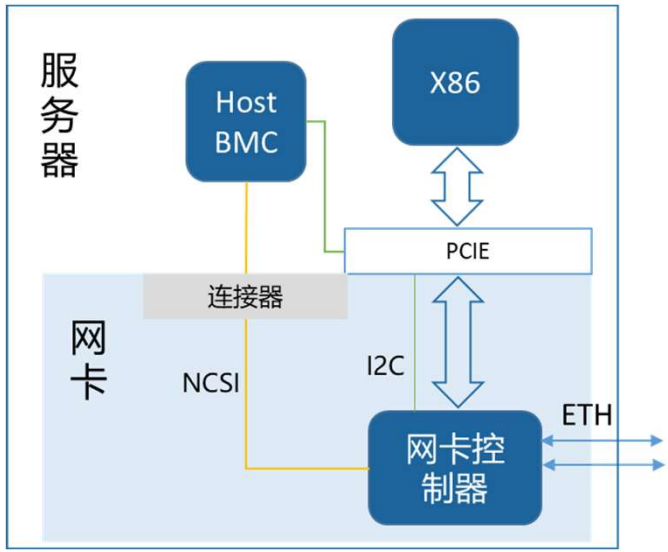
上电时序参考设计



智能网卡的监控需求

- 智能网卡是个独立运行的小系统，需要像管理服务器一样，监控整个网卡的硬件状态，记录异常日志、诊断分析故障、以及远程固件升级等；
- 传统服务器管理网卡的设计无法满足对智能网卡的管理，Host BMC的软/硬件都需要做较大修改，且非标准；
- 在网卡上采用独立的BMC监管设计，既可以解决监控管理需求，又可以避免服务器侧的软硬件修改。

传统网卡的管理控制设计



智能网卡BMC基本功能需求

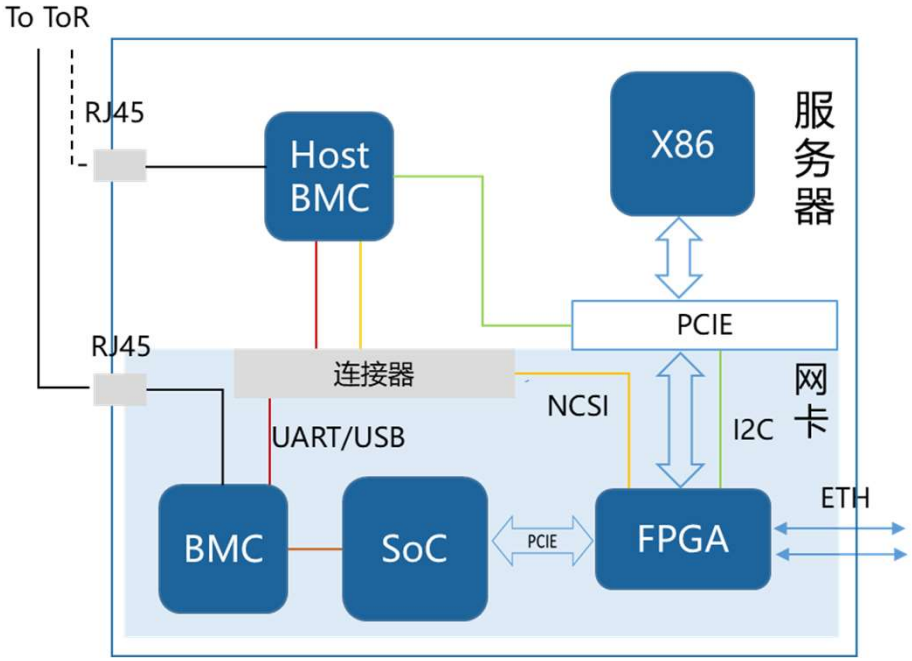
网络	Host MAC1支持static/dhcp混合模式
固件刷新	支持带外刷新 BMC ， BIOS和FPGA等固件
IPMI	IPMI命令
	电源管理
FRU	FRU
传感器类型要求	传感器类型要求
传感器定义	传感器定义
稳定性要求	S-BMC保活机制
	Self-Test状态
故障诊断	诊断日志收集
	诊断解析
	诊断日志下载
Host交互	散热相关的Sensor交互
	散热策略实现
	Host-Nic协同开关机
	Host Fake-S5状态实现
	Host Fake-S5状态下KVM、VNC显示

智能网卡和服务器的管理拓扑

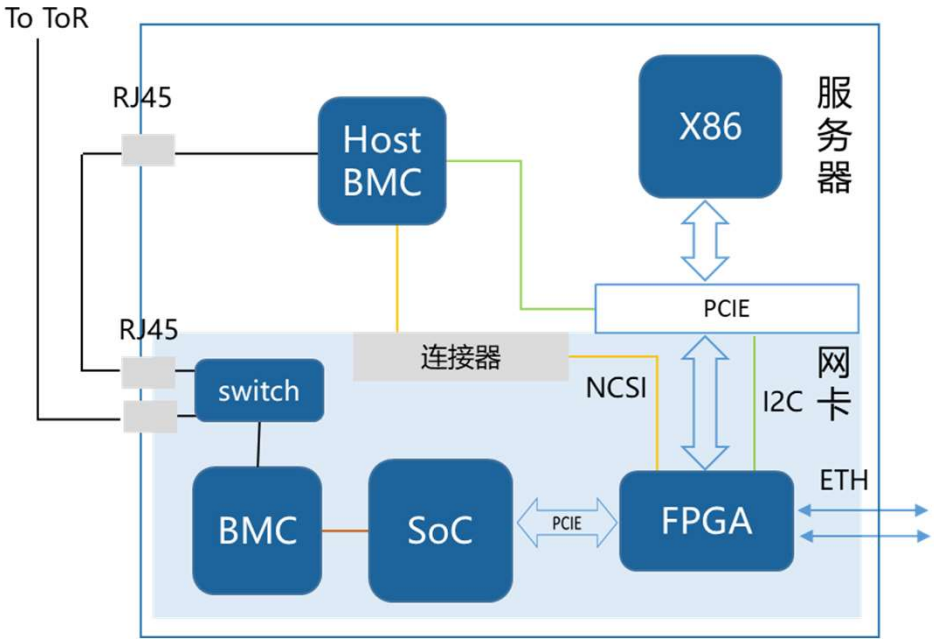
智能网卡和服务器的管理拓扑分为两种：

- ✓ 内部互联：通过UART，金手指的I2C以及NCSI，Host BMC与网卡BMC互联，两者为主从关系；
- ✓ 外部互联：通过网卡和服务器的网口互联，Host BMC与网卡BMC相互独立，分开管理。

内部互联



外部互联



服务器适配智能网卡的测试规范

硬件功能测试

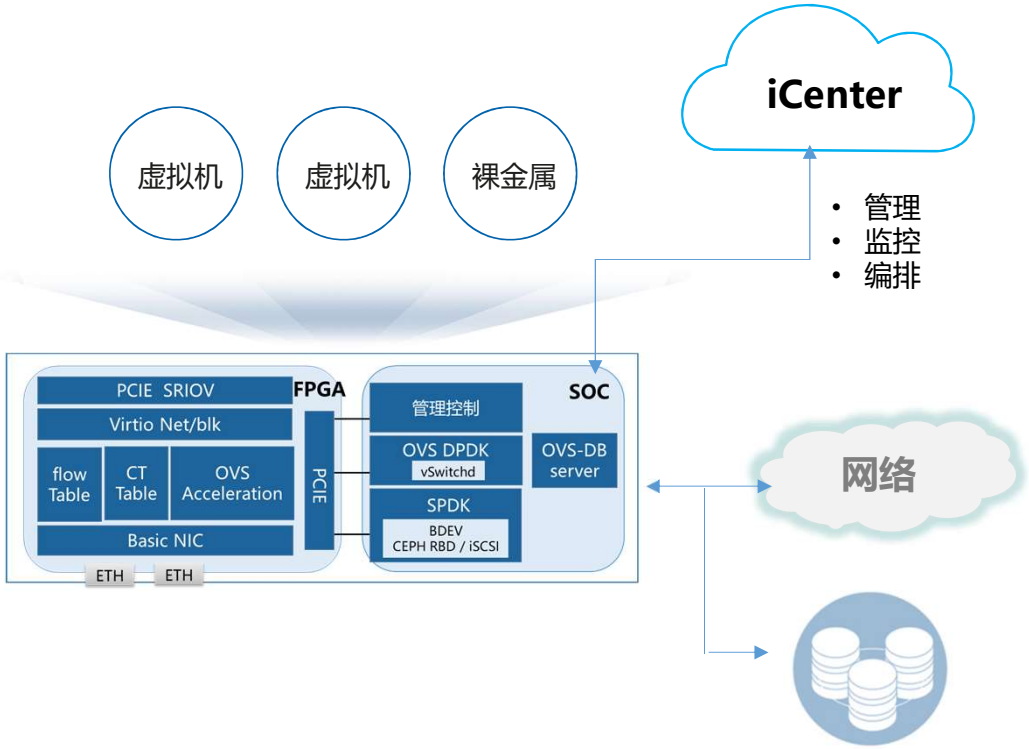
- HW - 硬件设计验证
- SIV - 信号完整性验证
- PIV/ACV- 供电及功耗验证
- Thermal - 散热设计验证
- EMCV - 电磁类验证
- REL - 整机可靠性验证
- BIOS/BMC

软件功能测试

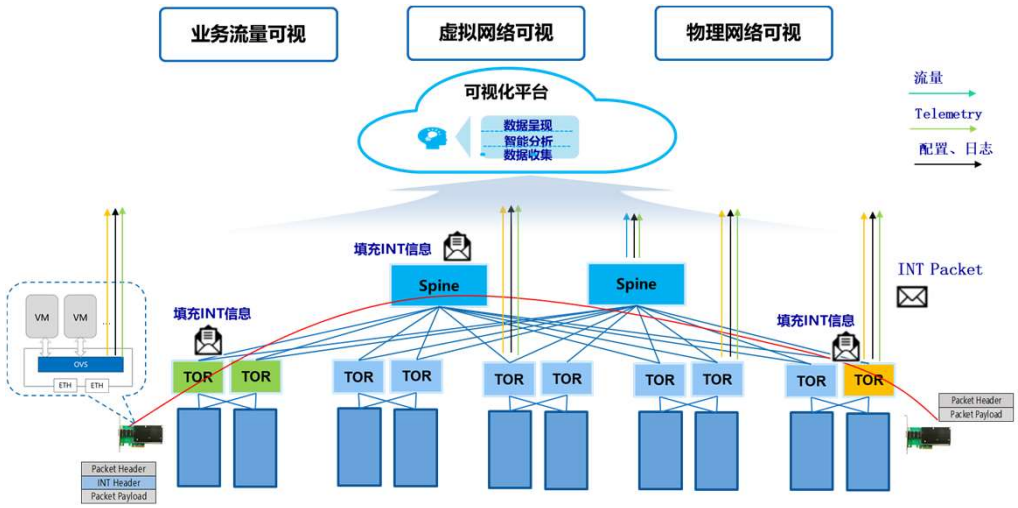
- BasicNIC - 基本网卡功能
- Networking - 网络卸载验证
- Storage - 存储卸载验证
- Secure - 安全功能验证
- Mgmt - 管理功能验证
- Perf/Press - 性能及压力验证
- SI - 系统集成及业务场景验证

智能网卡的应用与实践

虚拟化、裸金属云解决方案



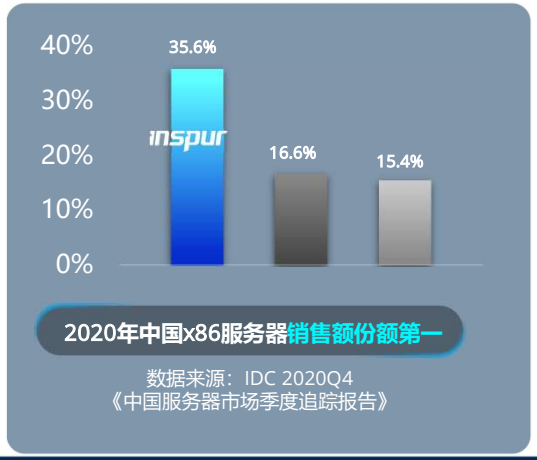
数据中心可视化方案



浪潮：全球领先的数据中心IT基础架构产品及方案提供商

- 服务器领域的全球领先企业&中国第一品牌，中国市占有率超40%，具备从芯片、板卡、部件到整机、平台软件的全栈自研能力，提供从通用服务器到关键小机的各类架构服务器产品
- 人工智能计算领域的全球领导品牌，全球市占率第一，中国市占率连续四年超过50%，拥有业界最强最全的AI计算阵列，支持从训练、推理、边缘的全AI场景
- 聚焦数据中心IT基础架构完整产品及方案，业务覆盖服务器、存储、网络等领域，覆盖中心数据中心到边缘数据中心场景应用，业务覆盖全球120个国家和地区，8个全球研发中心，6个全球生产中心，2个全球服务中心

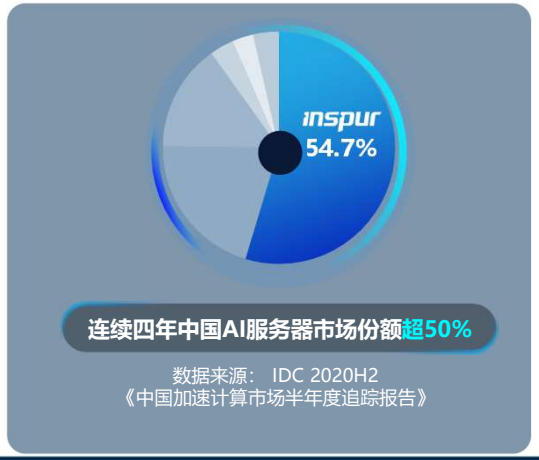
服务器 全球第三 中国第一



2020年中国x86服务器销售额份额第一

数据来源：IDC 2020Q4
《中国服务器市场季度追踪报告》

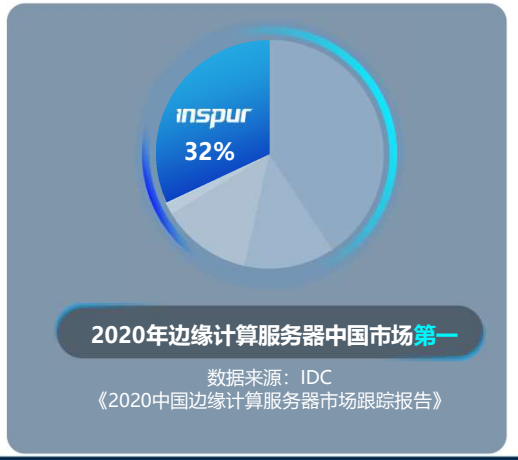
AI服务器 中国市占率超50%



连续四年中国AI服务器市场份额超50%

数据来源： IDC 2020H2
《中国加速计算市场半年度追踪报告》

边缘服务器 中国第一



2020年边缘计算服务器中国市场第一

数据来源：IDC
《2020中国边缘计算服务器市场跟踪报告》

浪潮网络—拥抱开源、开放，聚焦数据中心

开放网络解决方案

可视化控制器

- 3rd Controller
- Collector
- Analyzer
- Insight

软件NOS

- Inspur-NOS
- 3rd NOS

驱动

- SAI
- DIAG
- OpenBMC

基础硬件

- 交换机硬件平台
- 智能网卡

数据中心开放网络产品

INSPUR NOS

SC8661SL
(128 ×100G)

SC6630EL
(32 ×100G)

SC5630EL
(48 ×25G+8×100G)

CN3620EL
(48×10G + 2X40G+4X100G)

CN2610EL-48T4X2Q

CN2610EL-48T(S)4X

THANKS