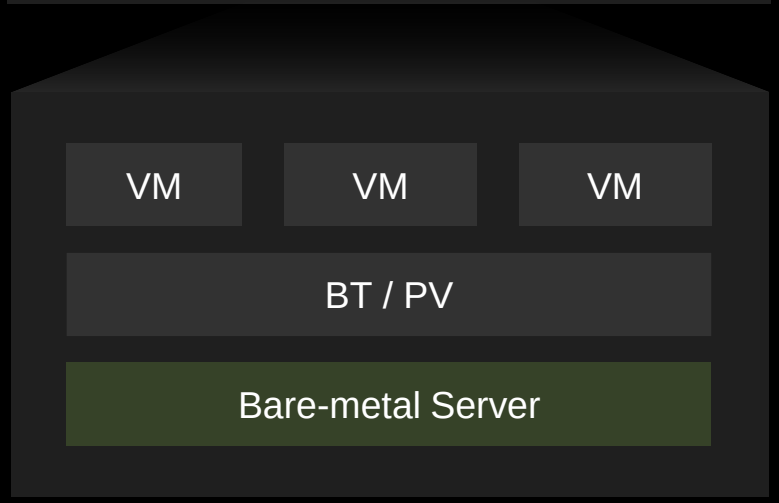
华为云擎天架构
及Zero软硬结合虚拟化3.0技术

刘劲松　华为云擎天首席架构师
黄智超　华为云Zero架构师

虚拟化技术历史变革

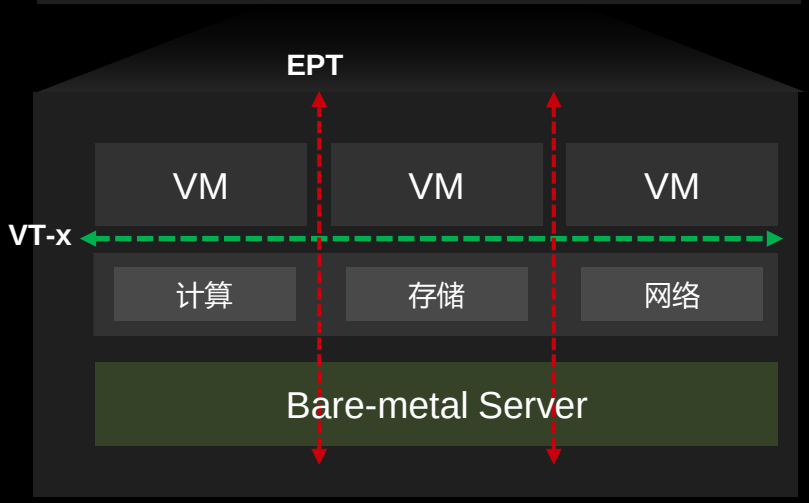
虚拟化 1.0
~ 2003



发源于实验室：纯软件虚拟化

- ✓ IBM 360
- ✓ 斯坦福 -- Binary Translation
- ✓ 剑桥 -- Para Virtualization

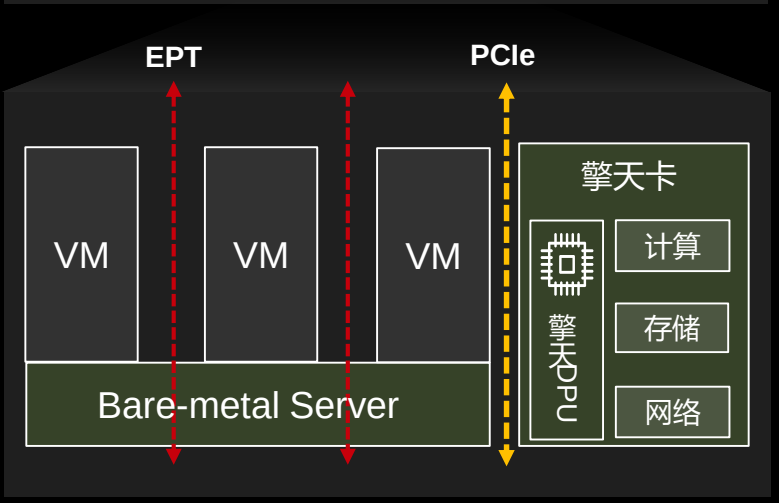
虚拟化 2.0
2004 ~ 2017



Xen/ KVM：硬件辅助虚拟化

- ✓ CPU硬件辅助虚拟化
  - CPU虚拟化：VT-x实现CPU虚拟化，开销较大
  - Mem虚拟化：EPT实现内存隔离，开销极小
  - I/O虚拟化： VT-d引入生态和弹性问题
- ✓ 3%~15%虚拟化开销
- ✓ 15%~20%数据中心税
- ✓ 虚机逃逸，数据安全
- ✓ 邻居噪音，业务抖动

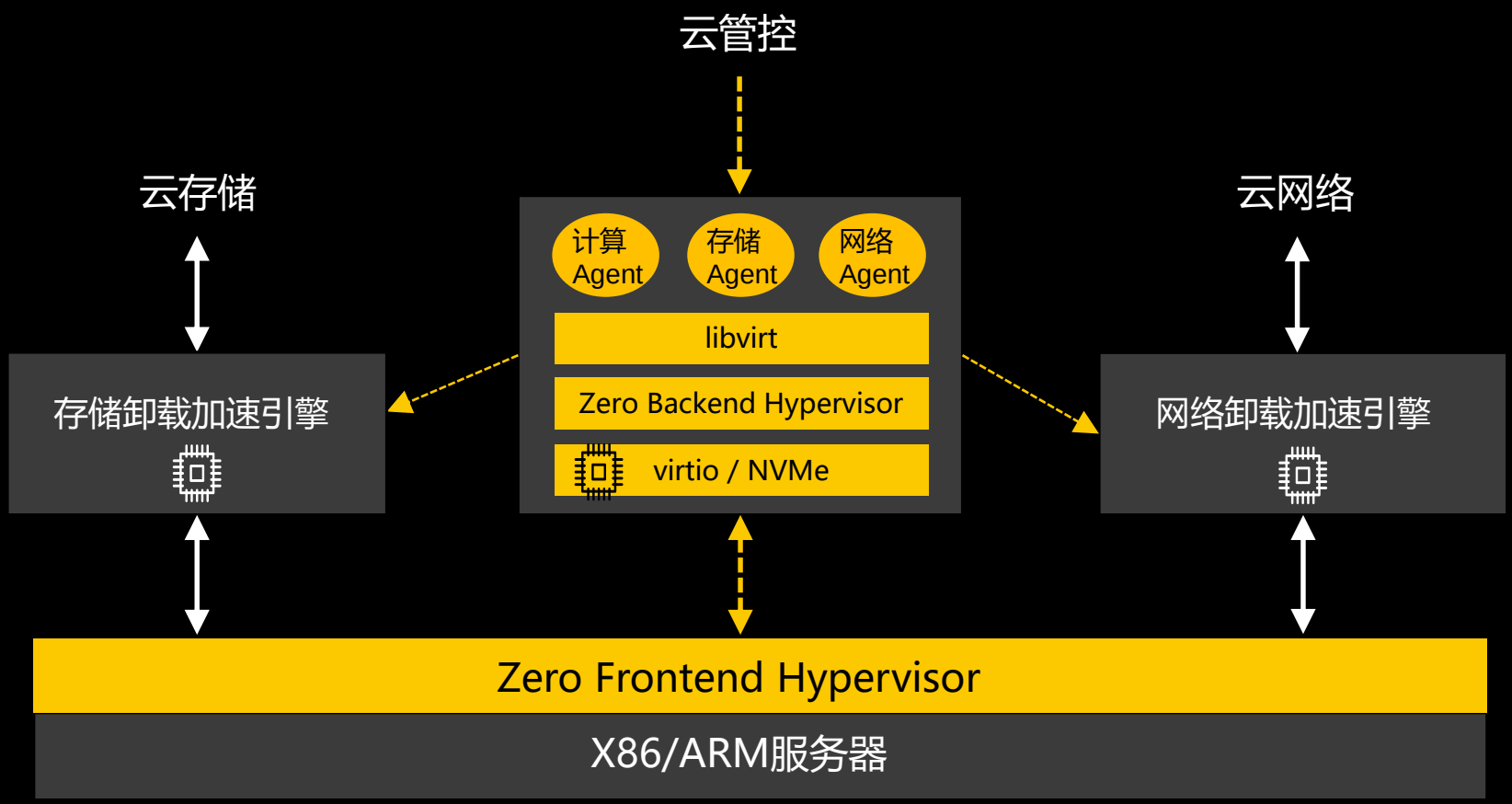
虚拟化 3.0
2017 ~



Zero软硬结合虚拟化3.0

- ✓ 前后端隔离、卸载、加速，简约不简单
- ✓ 前端自研 Zero Hypervisor
  - 资源0预留，虚拟化'0'开销，业务'0'抖动
  - 极简隔离与划分，近裸金属算力
  - 安全隔离 + 性能隔离
- ✓ 后端自研 擎天卡 + 擎天DPU
  - virtio虚拟化语义硬件实现，裸机虚机架构归一
  - libvirt / qemu卸载，无缝接入云管控、云存储、云网络
  - 网络I/O卸载加速 + 存储I/O卸载加速 + ……

擎天架构



控制流

数据流

- ✓ 单控制面，无缝接入现有云生态
- ✓ 多数据面， I/O路径隔离与扩展
- ✓ 裸机虚机架构归一
- ✓ X86/ARM多处理器架构支持

擎天计算虚拟化

近裸金属性能

资源0预留

Split-hypervisor

虚拟化'0'开销

近裸机性能

业务'0'抖动

企业级虚拟化服务

弹性 & 运维

在线灵活配置

VirtIO、SR-IOV、NVMe

极致弹性发放

K级虚机/容器 分钟级/秒级扩容

10毫秒级 ~ 亚秒级三热技术

内核&用户态热补丁、组件级热替换技术

OS内核热替换技术、虚拟机直通热迁移技术

安全 & 可信

安全芯片

企业级可信根

极简TCB

企业级可信基

安全硬隔离

防虚拟机逃逸攻击

擎天网络虚拟化

2019

virtio / OVS全卸载

- ✓ 支持VirtIO-net
- ✓ 快路径流表L1 Cache加速
- ✓ 多线程I/O无锁并发

2 x 25G / 1200万 PPS

20μs ~ 30μs时延

2020

自研GAEA网络引擎

- ✓ No openflow, 路由表业务编排
- ✓ Doorbell中断/轮询模式，DMA零拷贝转发
- ✓ 首包时延200us以内，毛刺率≈0
- ✓ Trunkport虚机容器网络归一

K级虚机/容器

分钟级/秒级扩容

Now

自研全栈RDMA

- ✓ 自研LDCP免PFC拥塞控制算法
- ✓ 支持RoCE、vRoCE
- ✓ 百倍组网规模提升

2 x 100G / 4000万PPS

< 10μs时延

擎天存储虚拟化

2019

VirtIO / SPDK/ EBS全卸载

支持virtio-blk/ virtio-scsi

SPDK / EBS全卸载

多线程I/O无锁并发

100μs、100W IOPS

Now

下一代存储虚拟化引擎

HUAWEI CurreNET

自研RDMA，LDCP拥塞控制算法

All-Flash存储

Append数据布局，智能FTL

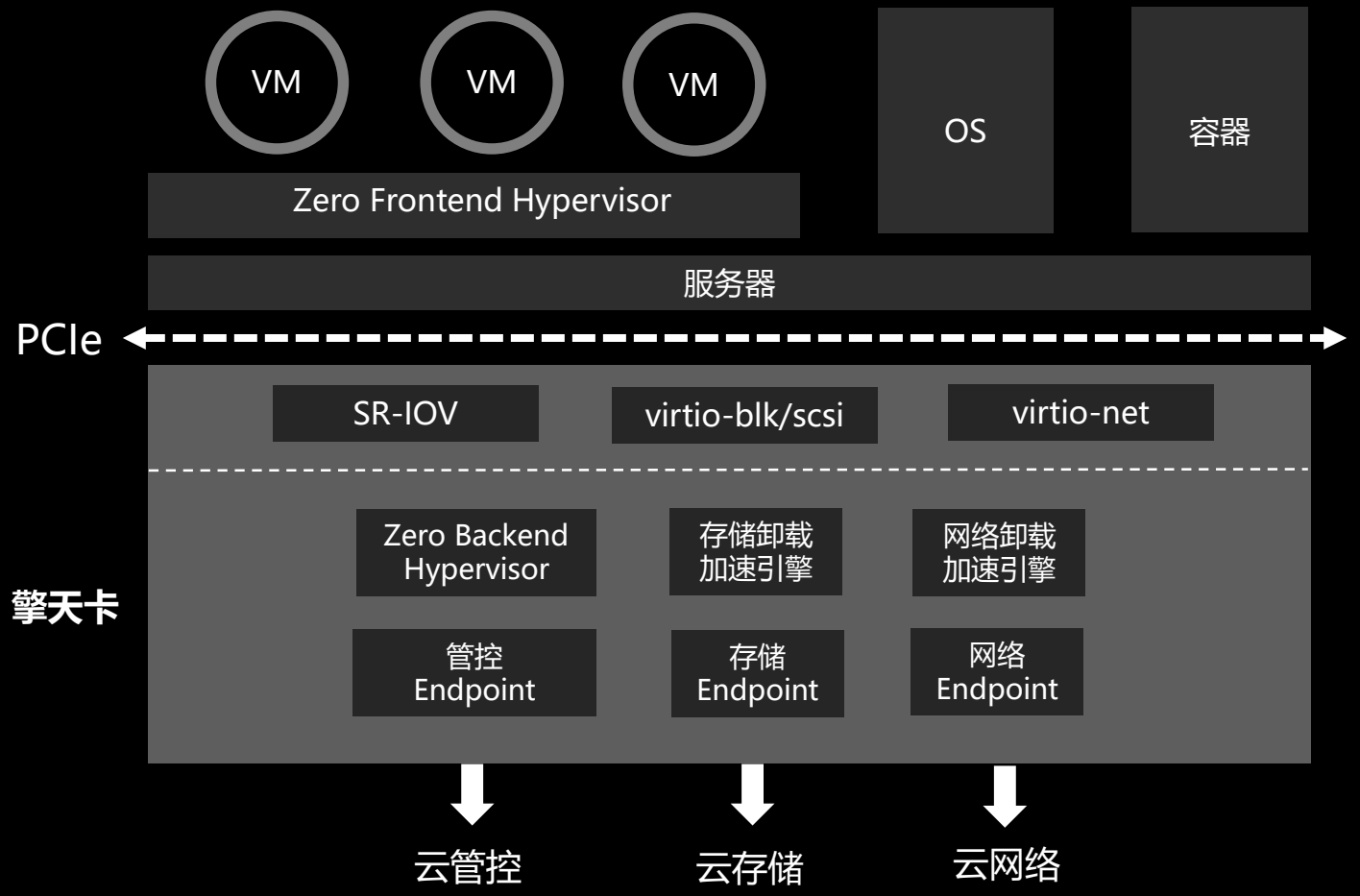
硬件加速处理

EC、DIF、I/O条带免拷贝转发

极致稳定，确定性QoS

Per-VM QoS

擎天架构统一支持裸机、虚机、容器实例



融合归一

支持裸机、虚机、容器多种实例

极简零损

近裸机的性能和稳定性

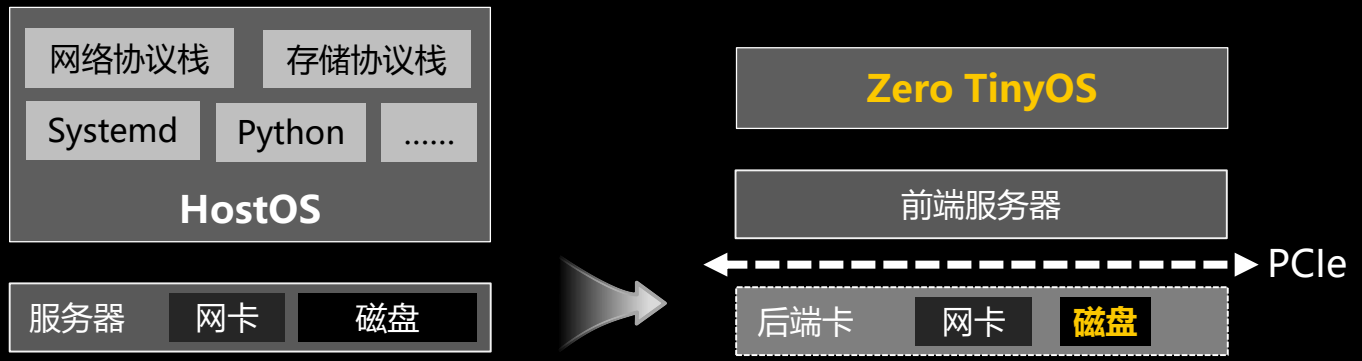
极速I/O

芯片加速

极简运维

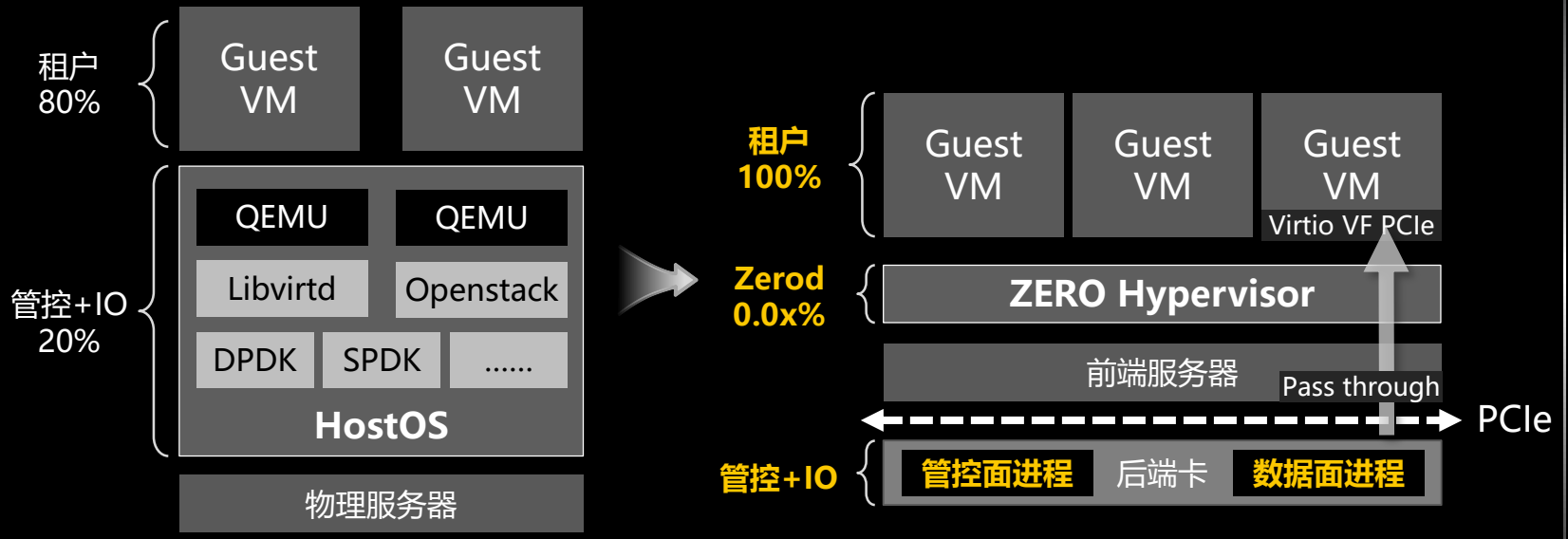
热补丁
组件热替换
OS热替换
虚拟机直通热迁移

Zero TinyOS



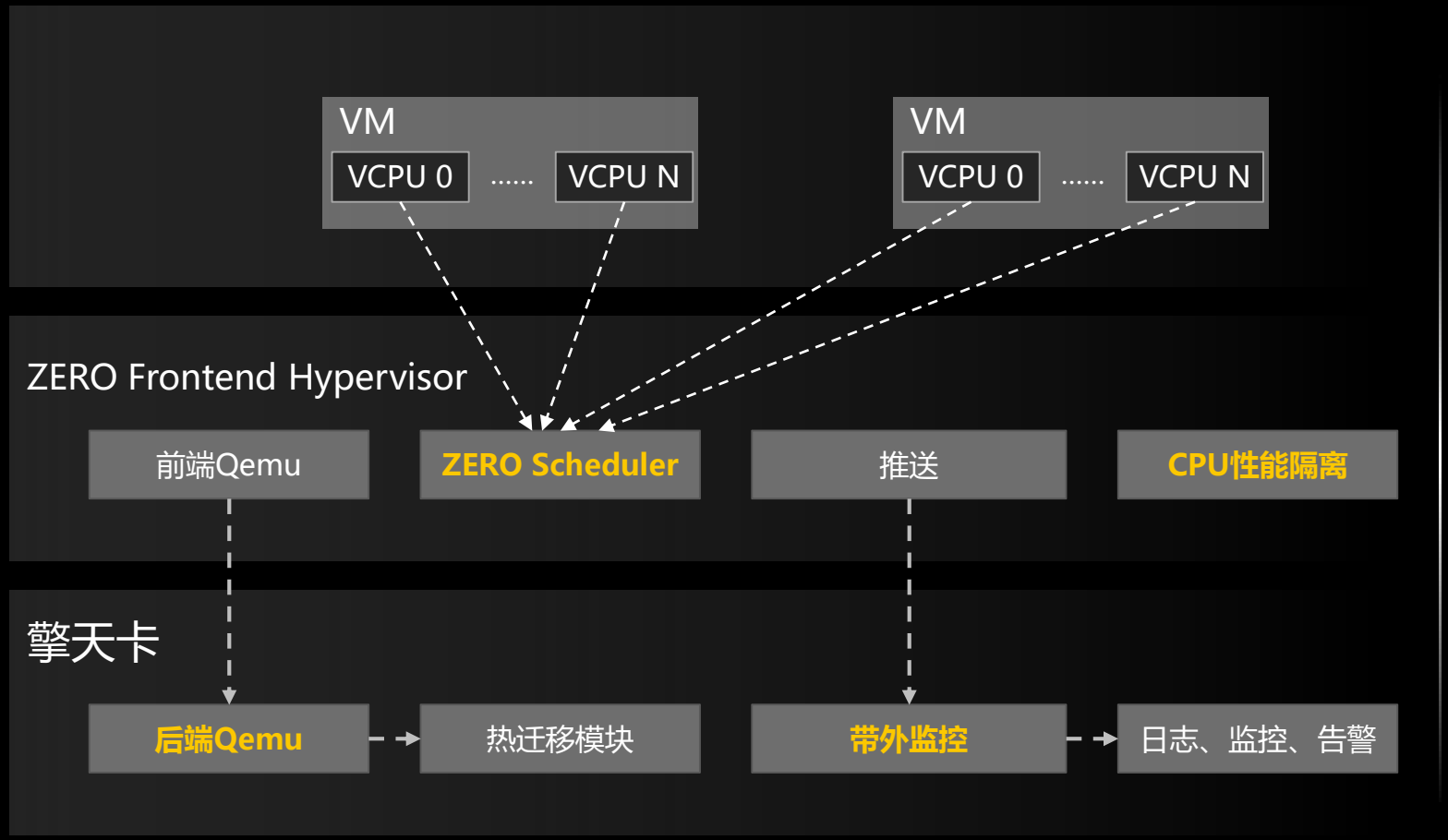
- 基于EulerOS极简定制
- 前后端Virtio-vsock通信
  - ✓ No NIC
  - ✓ No TCP/IP
  - ✓ No file system except rootfs
  - ✓ No systemd
- TinyOS最小TCB，秒级启动

CPU零预留、内存零预留



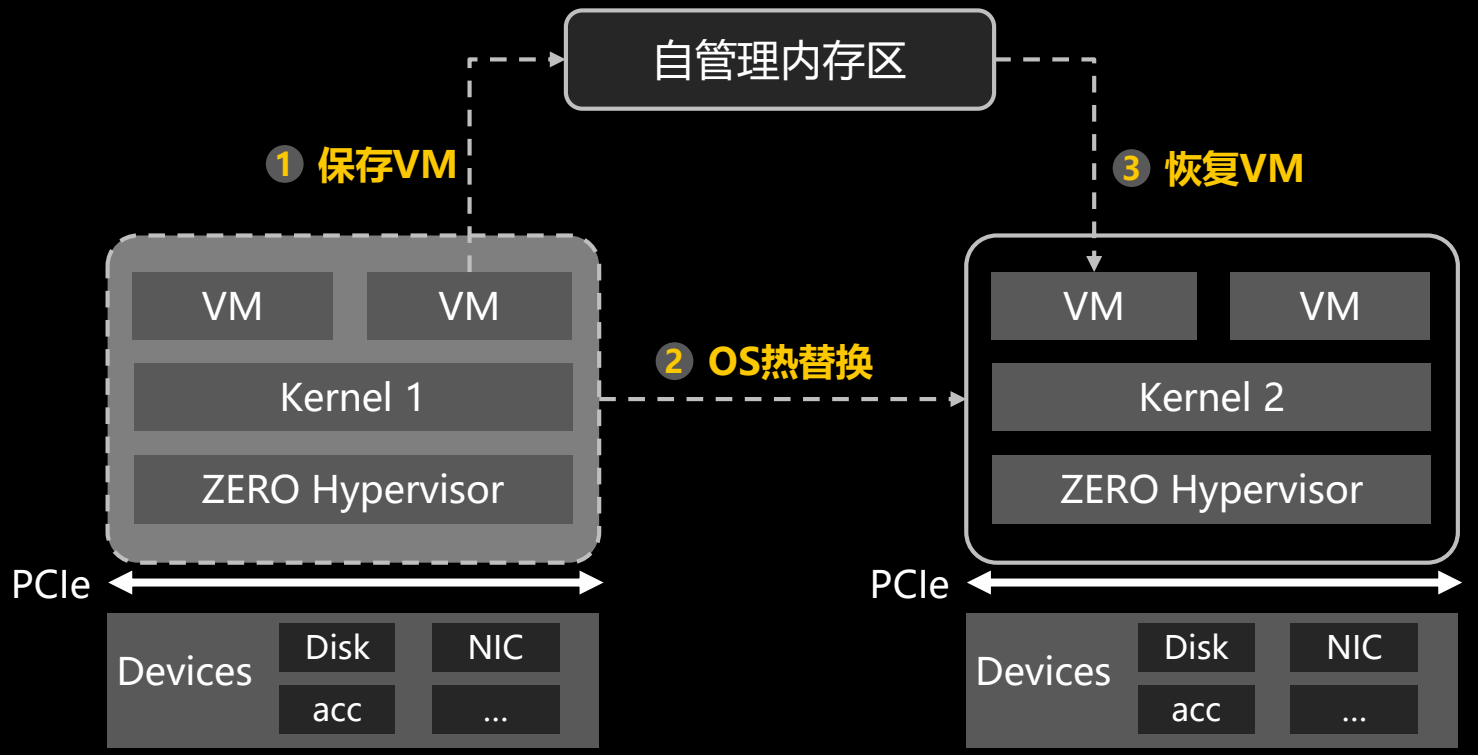
- CPU零预留
  - ✓ 计算、存储、网络组件全卸载
  - ✓ Split-qemu
  - ✓ Zerod 0.0x%开销，利用空闲CPU资源
  - ✓ 带外监控
- 内存零预留
  - ✓ Zero TinyOS
  - ✓ 用户态内存管理
  - ✓ Rmap优化
  - ✓ E820隐藏

虚拟化‘0’开销，业务‘0’抖动



- 虚拟化‘0’开销
  - ✓ 擎天架构红利
  - ✓ VF直通 + Post Interrupt
  - ✓ 1:1 VT-x极简定制
- 业务‘0’抖动 → 解决neighbor noise问题
  - ✓ 邻居搬家，计算、存储、网络全卸载
  - ✓ Zero scheduler，基于优先级的vCPU抢占调度
  - ✓ 性能隔离: RDT，PMT，bus lock, …
  - ✓ Split-qemu
  - ✓ 带外监控

OS内核热替换



- VM快速save/restore机制
- VM内存自管理、持久化
- IOMMU继承，设备无关
- 新OS内核百毫秒级快速启动

欢迎交流合作