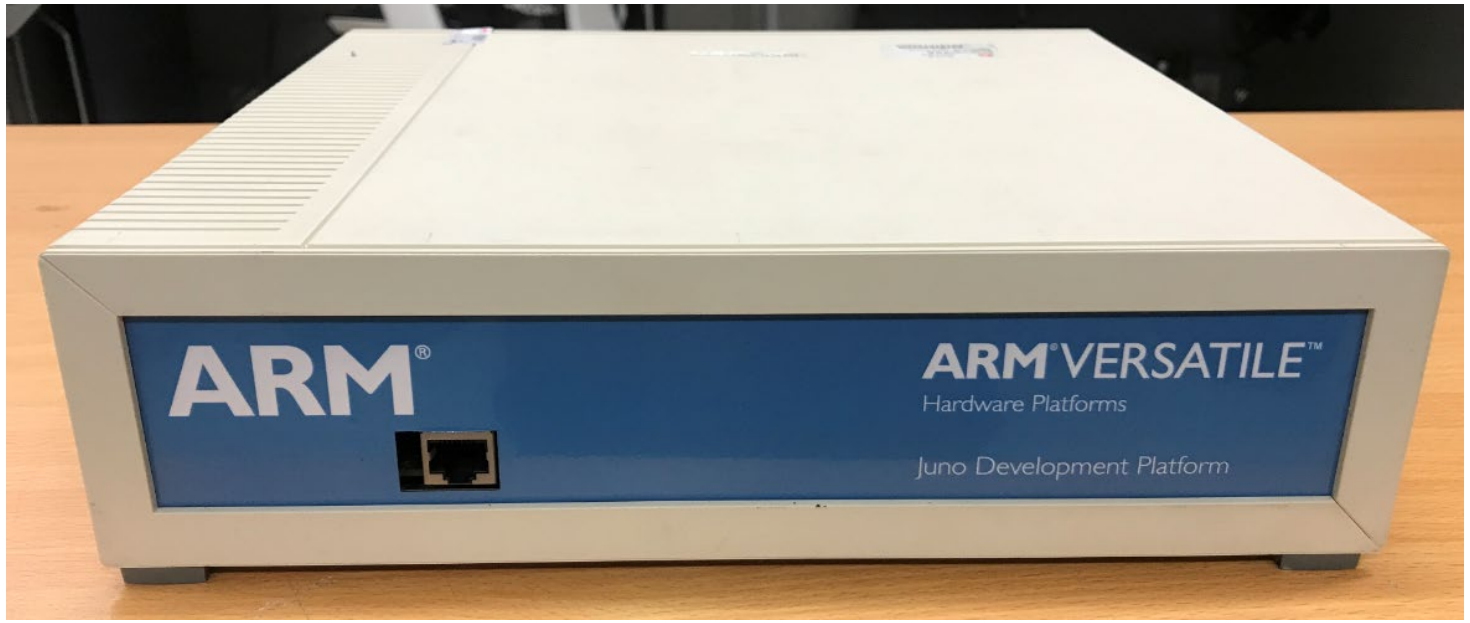
ARM64服务器Linux内核生态使能 – 历史与现状

郭寒军

Hanjun Guo <guohanjun@huawei.com>

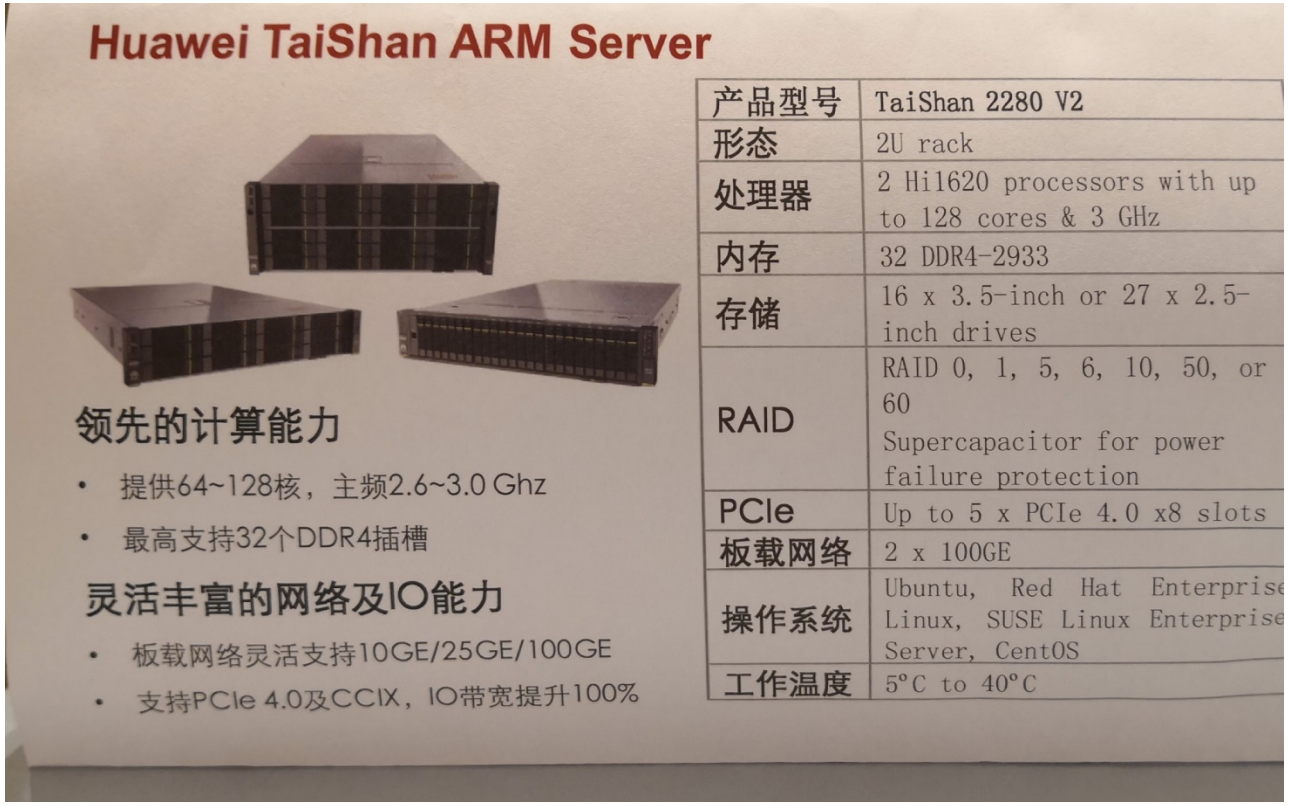
Hanjun Guo <hanjun.guo@linaro.org>

ARM64服务器芯片的发展 – 约5年前



- ARM Juno r1开发板。第一块“ARM64服务器”形态的硬件单板
- 大小核，2个A57的大核，4个A53小核；
- 4G内存，支持PCIe 2.0
- 那个时候做ARM64服务器芯片厂商都还在芯片设计中

ARM64服务器芯片的发展 – 现在



产品型号	TaiShan 2280 V2
形态	2U rack
处理器	2 Hi1620 processors with up to 128 cores & 3 GHz
内存	32 DDR4-2933
存储	16 x 3.5-inch or 27 x 2.5-inch drives
RAID	RAID 0, 1, 5, 6, 10, 50, or 60 Supercapacitor for power failure protection
PCIe	Up to 5 x PCIe 4.0 x8 slots
板载网络	2 x 100GE
操作系统	Ubuntu, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, CentOS
工作温度	5°C to 40°C

- Huawei基于自研核的Hi1620芯片，以及服务器；
- 国内：华为，华芯通，飞腾；美国：Cavium，Ampere，高通

软件生态使能先于芯片

- X86一般在新芯片出来前（约2年），相关软件会得到支持，包括Linux内核；
- ARM64服务器芯片的使能，也需要朝这个目标努力
- 软件生态使能先从OS这个底座开始
  - 软硬件解耦
  - 服务器相关特性使能
  - 性能优化
- 主要回顾Linux内核生态使能的历史，以及现状

ARM平台多样性决定了软硬件解耦的重要性

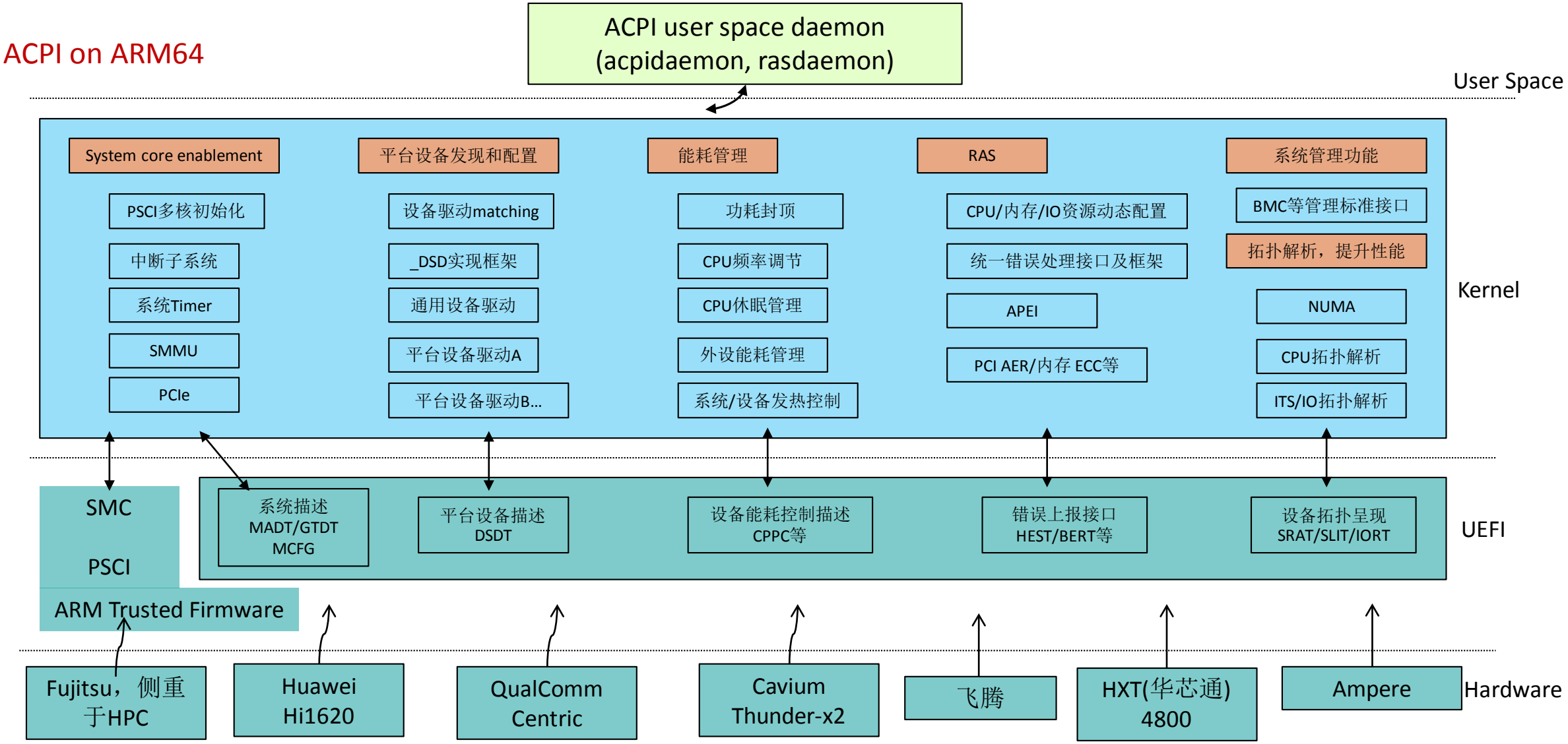
- 从客户的视角(比如阿里，腾讯，百度有着大量数据中心的公司)，希望买一台ARM服务器跟x86是一样的，包括OS安装，后续运维等。
- 不能忍受每个厂家都有自己OS，对运维和后续维护是个巨大的挑战
- ARM嵌入式的多样性在服务器上不一定是好事 (Diversity is good, but uncontrolled diversity is bad.)
- 因此软硬件解耦是第一个需要解决的事情
- 目标：Make ARM server as boring as it should be.

软硬件解耦-规范先行

- 需要从规范层面来约束和保证
  - 芯片设计规范SBSA(Server base system Architecture)
    - 规定芯片需要支持的特性，包括CPU核的特性，中断，时钟以及PCIe特性等；
  - OS和firmware的解耦方案采用了UEFI/ACPI
    - 要同时支持Linux以及Windows，Windows只支持ACPI；
    - ACPI的一些特性支持包括RAS等比device tree有优势；
    - Device tree和ACPI启动二选一；
    - 规范最早是微软支持windows phone的，因此很多规范不满足服务器需求
  - SBBR （system boot base requirement），定义了
    - UEFI的基础要求；
    - ACPI特性支持的基础要求；
    - SMBIOS基础特性要求；
- 规范发布历史
  - 芯片/BIOS/OS/虚拟化/服务器厂商以及云服务提供商广泛参与；华为/高通/AMD/AMI/微软/ARM/redhat/HP/google/Linaro等；
  - 从ACPI 5.1开始逐渐支持ARM服务器平台；6个月发了一个新的ACPI规范，创造了历史，感谢Intel☺
  - 最新规范版本，SBSA5.0，ACPI 6.3(预计19年1月份发布)
- 华为是积极贡献者，超过20个提案被ACPI/UEFI/SBSA规范小组接收，比如NUMA/CPU拓扑/异构支持等多个关键特性；SBBR的初稿是华为和Redhat等撰写的。
- 期待更多的国内同行参与相关规范的讨论和制定，会议时间不一定是半夜☺

通过ACPI架构解耦，实现多平台OS共享BIOS与驱动架构

ACPI on ARM64



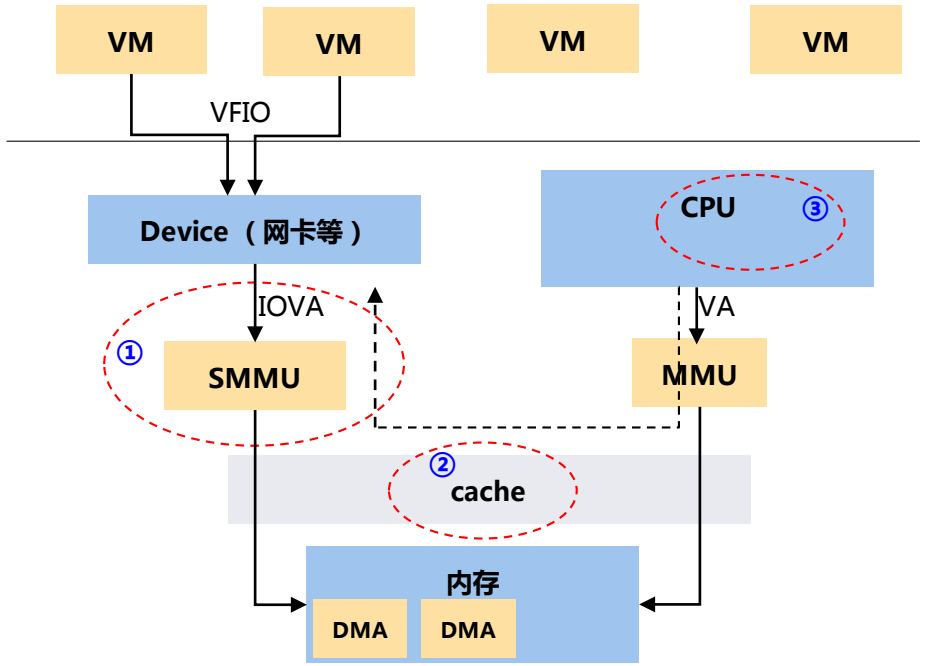
特性使能 – Linux kernel upstream

目前OS使能的状态

- Linux内核主线直接使能各个厂家ARM64服务器的芯片；
- SUSE/Redhat/Ubuntu/Oracle/Centos正式商用发行版直接支持；
- Windows支持Huawei, cavium,高通等厂家芯片。

性能优化

性能优化点很多，用一个优化I/O路径上的性能作为例子来讲一下我们做的事情。
遇到的问题: 在我们开启SMMU（IOMMU）后，I/O性能直线下降，但在虚拟化(SMMU保证虚拟机IO隔离性)以及异构(比如AI加速器)场景下，SMMU是绕不开的点。



1. IOVA查找效率低，页表释放导致TLBi以及SYNC过长，拉低整个IO性能；
2. IO设备与CPU进程cache冲突，关键DMA操作从内存读取；
3. CPU端多核多进程读写设备，未充分利用IO的并行能力。

- 针对SMMU瓶颈点
  - IOVA优化页表查找，提升未分配页查找速度；
  - TLBi+SYNC耗时时间长，通过优化SYNC命令数量，缩短等待耗时；
  - 引入Non-strict模式，延迟页表释放，极大提升性能；
- 针对cache瓶颈点 – 讨论点
  - 或许类似Intel的RDT技术可以解决，ARM类似的技术MPAM；
  - 或者DDIO(Data Direct I/O)也有效果？
- CPU侧
  - 将spinlock保护的对象修改为RCU保护，提升并行能力
  - Cache false sharing冲突检测，qspinlock

优化效果，使用iperf测试100G网卡带宽：从优化前的17Gbps提升到约94Gbps，提升约4.5倍。

优化的补丁，我们都已经推到主线内核，比如SMMU Non-strict模式的支持，在最新的4.20进入主线。

当前正在进行中的一些特性

- 芯片
  - ARM先后发布了ARMv8.1/2/3/4/5的规范，新增了不少特性
    - CPU漏洞的修复，包括spectre, meltdown等；
    - Memory tagging, VA的4个bit作为tag，来防止溢出，use after free等漏洞；
    - MPAM（Memory partition and monitor），类似于Intel的RDT，但支持虚拟机内再分配cache；
    - RAS增强
- Kernel
  - MPAM支持
  - 异构系统支持
  - ARMv8.x特性支持等等
- 规范
  - SBSA: PCIE的设计基本要求；
  - ACPI: CCIX或者其它一致性（设备内存与CPU内存cache coherent）互联协议的支持（异构系统支持）

最后来一个广告

欢迎加入华为OS内核实验室

华为操作系统部：华为端、管、云核心的OS软件基础设施

OS Kernel Lab

- Linux内核（ARM/x86/异构等）的技术研发与创新
- 低时延、高安全、高可靠、高智能的下一代OS内核技术的研究和成果转化

招聘岗位

下一代操作系统研究员/高级工程师

形式化技术研究员/高级工程师

Linux内核架构师/高级工程师

Linux内核测试专家/高级工程师

工作地

杭州、北京、上海

简历投递

Tel：王先生/18658102676

Email：hr.kernel@huawei.com