案例研究

英特尔® 基础设施处理单元 (IPU)
移动云

借力英特尔数据中心技术，中国移动自主研发硬件加速卡 Hypercard

中国移动 China Mobile

目录

概述 1

背景挑战 1

解决方案：磐石服务器架构为移动云云场景提供全面坚实算力底座 2

定制化 BC-Metal 裸金属服务器，为用户关键业务提供出色性能 3

Hypercard 硬件加速卡有效卸载 IaaS 能力，实现算力优化 3

测试验证 4

总结与展望 5

概述

在国家启动“东数西算”工程，布局全国一体化算力网络的新形势下，更多行业正通过企业上云、关键业务用云来进一步推动数字化转型进程，但由此带来的数据规模急剧增大，也对云基础设施的效能提出更多挑战。

如何以创新的技术架构、坚实的基础设施来为用户提供更为优质、高效、弹性和安全的云服务能力，是中国移动在打造移动云业务持之以恒的目标。为加速实现这一目标，中国移动旗下中国移动云能力中心（以下简称“移动云”）携手英特尔，以自研 BC-Metal 裸金属服务器与 Hypercard 硬件加速卡共同构建磐石架构，为移动云提供算力基座。

Hypercard 硬件底层依托英特尔® 至强® D 处理器以及英特尔 Stratix® 10 DX FPGA 等产品，不仅能通过裸金属服务器为用户提供出色的性能，还能实现多种基础设施即服务（Infrastructure as a Service，IaaS）能力卸载。这一通过以硬件方式实现设备虚拟化和数据转发来替代传统的纯软件虚拟化架构，不仅能有效降低算力开销，释放服务器计算资源，且可大幅提升云服务在网络、存储等方面的性能。

背景挑战

云服务已逐渐成为各行各业“IT 新基建”的必选项，电信运营商也在其中扮演着越来越重要的角色。数据表明，近年来，各电信运营商正在云服务市场中展现出更高的增速[1]，并获得更多市场认可。以中国移动为例，其 2022 年上半年的移动云收入达 234 亿元，同比增长 103.6%[2]，从分析报告可以看出，背靠中国移动提供的成熟网络环境和业务模式，移动云已深深根植于中国移动生态体系，帮助用户实现“入网即入云”。

这绝非偶然，在云网融合理念获得越来越多认可的今天，中国移动等拥有丰富云网资源并具有良好运营经验的电信运营商显然是未来云服务领域的“种子选手”。尤其在中国全面启动“东数西算”工程的新形势下，中国移动计划以算力为中心、网络为根基，构建深度融合多种信息技术的算力网络来为社会经济发展提供强劲有力的引擎。

但在中国移动不断提升云服务影响力的同时，其云基础设施的效能也面临更多考验。算力的连接、云服务应用场景的扩展，意味着承载中国移动云服务的数据中心将更多采用

分布式计算、分布式存储的架构，这无疑将带来更大规模的横向数据流量负载（包括数据接收、转发、存储和处理等）。来自中国移动的内部分析表明，目前其数据中心网卡的端口吞吐量正快速从 10G 向 25G 乃至 100G 及以上演进，从而带来了多方面的挑战：

- 大负载、大流量数据带来的输入输出（Input/Output，I/O）瓶颈。在传统虚拟化架构中，VirtIO 等 I/O 虚拟化技术通常需要处理器深度参与，算力资源损耗巨大。当数据量增大时，就会因资源争抢而造成性能抖动，进而带来服务级别协议（Service Level Agreements，SLA）不一致问题。
- 海量数据转发对网络性能的制约。数据转发通常使用 OVS（Open vSwitch）等虚拟交换软件完成，这需要消耗大量计算资源。当数据规模急剧扩大时，OVS 就会与虚拟机中的其它服务进程形成算力资源竞争，既影响系统性能也无法充分利用网络带宽。
- 在存储方面，随着云服务环境工作负载类型的多样化，面向高速数据存储的接口、协议栈也日新月异，例如标准存储接口 virtio-blk、基于架构的非易失性内存（NVM Express over Fablics，NVMe-oF）等。传统存储模式在应对这些新接口、新协议栈时，弹性和灵活性往往并不尽如人意。
- 为应对虚拟化带来的算力损失问题，中国移动也提供了裸金属服务器，来为用户在关键场景中提供高性能、高可用的云服务能力。但在传统云服务架构中，云管理系统通常与虚拟机共用处理器资源，而裸金属却需要独占资源，因此会限制裸金属实现弹性管理与交付。如采用嵌套虚拟化方式，所产生的开销会带来性能损失。

事实上，由于近年来在大型数据中心中，数据流量处理负载的增长速度始终高于算力增长速度，其对算力资源的占用率也越来越大。一项统计数据表明，目前数据中心中 30% 的计算是在作流量处理，甚至有人将之形象地称为“数据中心税（Datacenter Tax）”[3]。

应对以上挑战的有效策略之一，是对算力资源进行“开源节流”。所谓开源，是指在云服务中引入算力损耗更小、性能更为突出的产品。而节流则是以全新的硬件方式来取代传统软件实现的设备虚拟化和数据转发功能，从而有效实现 IaaS 能力卸载，释放云服务器算力资源。

基于这一理念，借助英特尔全面的芯片及软硬件生态能力，移动云自研 BC-Metal 裸金属服务器与 Hypercard 硬件加速卡，由此打造全新磐石架构，并在云服务实践中为用户提供更具优势的网络和存储性能。

解决方案：磐石服务器架构为移动云云场景提供全面坚实算力底座

随着更多物联网、人工智能、大数据分析等应用被部署到云服务中，用户对云环境的高性能、低延迟以及安全性也有了更多的需求。针对这些需求，移动云基于硬件加速技术打造了新一代的磐石服务器架构。如图 1 所示，磐石服务器架构包括 BC-Metal 服务器与 Hypercard 硬件加速卡。

图 1 移动云磐石服务器架构

其中，BC-Metal 服务器基于第三代英特尔® 至强® 可扩展处理器平台开发，与通用服务器相比，加入了大量定制化硬件设计，如定制化 BIOS（Basic Input Output System）、定制化 BMC（Baseboard Manager Controller）等，为客户提供更安全、更高效的算力服务。而移动云自研的 Hypercard 硬件加速卡则基于英特尔® IPU 参考设计，由 FPGA 加速单元与 SoC 控制单元合作完成 IaaS 服务的卸载，为云服务提供商提供领先的 IaaS 服务解决方案。BC-Metal 服务器与 Hypercard 硬件加速卡之间通过高速串行计算机扩展总线标准（Peripheral Component Interconnect Express，PCIe）总线和其它自定义接口相连接。

■ 定制化 BC-Metal 裸金属服务器，为用户关键业务提供出色性能

不同于常见的云主机产品，裸金属服务器允许云租户独享处理器、内存等资源。这一类似物理机的特性和优势，让用户的关键业务在云环境中也能获得高密度算力和低延迟 I/O 支持。这一设计，让硬件基础设施的选择就显得至关重要。为此，移动云通过对大量用户业务场景的梳理，根据业务需求，在 BC-Metal 服务器中加入了大量定制化的软硬件产品，包括定制化 BIOS、定制化 BMC 以及定制化主板等。借助这些定制化产品，BC-Metal 服务器可以在服务监控、供电管理、系统安全散热效率以及运维管理等方面，为用户业务提供全生命周期的高效服务。

图 2 中国移动自研服务器 BC-Metal

与此同时，移动云也与英特尔开展合作，引入第三代英特尔® 至强® 可扩展处理器（Ice Lake）作为服务器的核心算力引擎，多方位提升服务器性能：

- 更多的内核、更优的架构带来算力性能的大幅提升，其单个处理器最高可配备 40 内核 80 线程，最高主频达 3.1GHz，可有效应对高密度计算所需；
- 支持更多内存，最大内存容量可达 12TB。支持 PCIe Gen4，从而实现更高的每核 I/O 带宽。同时多达六条的英特尔® 超级通道互联（英特尔® UPI）通道也有效提高了 I/O 密集型工作负载时的处理器间带宽；
- 多项内置增强技术，例如英特尔® 深度学习加速技术（英特尔® DL Boost）和英特尔® 软件防护扩展（英特尔® SGX）技术，对在人工智能（Artificial Intelligence，AI）场景及数据安全性方面提供助益。

■Hypercard 硬件加速卡有效卸载 IaaS 能力，实现算力优化

如果说 BC-Metal 服务器助力磐石服务器架构实现了算力的“开源”，那么由移动云自研的 Hypercard 硬件加速卡就是磐石服务器架构进行 IaaS 服务卸载，获得算力“节流”的有力武器。从物理规格上来看，其采用了全高半长的 PCIe 设计，可通过 PCIe 及自定义接口与服务器主机相连接，并具有独立供电设计。加速卡对外提供了多个 SFP28 光口，可用于各类业务数据的高速传输。

图 3 中国移动 Hypercard 硬件加速卡

这一全新的产品是基于英尔特 FPGA 及 SoC 设计开发的，作为可编程网络设备，能够对数据中心内的基础设施功能进行加速，从而使系统级资源的管理更加智能。其中，FPGA 加速单元的核心是英特尔 Stratix® 10 DX FPGA 芯片，这一可编程逻辑芯片通过先进的架构设计、封装技术、比前一代 FPGA 拥有更多的收发器并支持硬核 PCIe Gen4 接口从而实现更高的带宽。配合硬件全可编程特性，能灵活地实现英特尔® OneAPI 定制化设计，从而实现高吞吐和低延时的性能表现，完成 I/O 虚拟化、OVS 转发等任务，实现基础设施管理、网络和存储功能卸载。

而 SoC 控制单元的核心是英特尔® 至强® D 处理器，其提供了卓越的单核性能、诸多安全功能加强以及内置的一系列硬件加速能力，来承载磐石服务器架构各项管理功能卸载。而英特尔® 至强® D 处理器良好的 x86 兼容性，以及与其它英特尔® 架构硬件形成的良好生态支持，能帮助用户实现系统代码或应用能力的快速迁移，从而提升卸载效率。

- **网络能力卸载**

借助以上两种芯片，Hypercard 硬件加速卡能够有效地实现对多种 IaaS 能力卸载，包括网络、存储以及安全等的卸载。在网络卸载上，首先加速卡可通过将 I/O 虚拟化卸载至 FPGA 芯片来实现虚拟化能力的“硬化”。云服务通常采用 VirtIO、SR-IOV（Single Root I/O Virtualization）等 I/O 虚拟化技术来完成网络数据收发，以 VirtIO 为例，其采用了前后端分离的架构，虚拟机可借助 VirtIO 驱动与宿主机中的 VirtIO 后端进行数据交互。传

统上，这一交互过程需要操作系统内核参与并消耗大量的处理器算力。随着 I/O 吞吐量的不断提升，其也逐渐遭遇性能瓶颈。在 Hypercard 硬件加速卡的设计中，将 VirtIO 等 I/O 虚拟化卸载到 FPGA 芯片中完成，这不仅能显著提升 I/O 性能，还可充分释放系统算力，使虚拟化损耗不超过10%[4]。

另一项网络能力的卸载是转发面的“硬化”。如 OVS 是移动云服务中常用的虚拟交换软件，但其通常需要依赖处理器核轮询的方式来完成数据的转发处理。因此，随着数据中心带宽的不断提升，所需的处理器资源也与日俱增。根据中国移动专家的估算，25G 带宽的云服务场景中需要预留 18% 左右的处理器资源[5]，而当带宽进一步扩大至 100GB 时，这一比例将更为巨大，从而挤占关键业务所需的算力。

在 Hypercard 硬件加速卡的设计中，将 OVS 的转发面卸载到 FPGA 芯片中进行处理，可在释放算力资源的同时，也显著提高网络转发性能，使得网络转发能力可达 3100 万 PPS[6]。

- **存储能力卸载**

在传统的数据存储处理中，应用在网络设备与存储设备间的每一次 I/O 都需以“中断”的方式将数据在用户态和内核态之间进行频繁交换。整个过程需要进行多次处理器上下文切换以及内存拷贝，不仅效率较低且需要消耗大量的算力资源。

现在，云服务正引入 NVMe-oF 这样的协议栈来应对这些短板。以远程直接数据存取（Remote Direct Memory Access，RDMA）为例，其能通过内核旁路、零拷贝等技术，使数据直接面向网络设备传递而不经由操作系统，从而消除数据复制和进程上下文切换带来的开销，大幅度解放处理器。但 NVMe-oF 协议栈如果仍然由处理器进行处理，无疑会带来弹性和灵活性方面的不足。

针对这一问题，Hypercard 硬件加速卡一方面将 virtio-blk 标准存储接口、NVMe-oF 协议栈等卸载到 FPGA 芯片中，由其完成核心的数据包封装 / 解封装、拥塞控制等工作负载。同时，加速卡也引入 SPDK（Storage Performance Development Kit）框架，借助其提供的轮询、异步、无锁的 NVMe 驱动程序、Bdev 通用块层和优化的应用框架等用户态加速能力，来有效提升数据存储转发性能。

- **裸金属“云化”**

在 BC-Metal 裸金属服务器为磐石服务器架构带来高性能的同时，如何使之继续保持云服务的弹性和敏捷也是移动云在进行产品设计时着重思考的问题。借助 Hypercard 硬件加速卡，裸金属服务器无需部署虚拟化的工具，而是将这些工具和能力卸载到加速卡中。目前，借助该加速卡，移动云已经可以实现基础设施管理面与租户之间的物理隔离，利用 VirtIO 设备热插拔特性支持弹性裸金属服务，令裸金属在“云化”后，能以传统云主机的方式提供给用户。

值得一提的是，在多租户虚拟化环境中，引入硬件加速技术的磐石服务器架构同样也能将虚拟机管理程序卸载到加速卡中，例如对基于 vDPA 的虚拟机无缝热迁移的支持，使得云服务在资源分配时能够更为高效。

测试验证

为验证磐石服务器架构在网络转发性能和存储性能上获得的提升，移动云与英特尔一起，对其进行了一系列对比测试验证，验证结果如下：

- **网络性能测试**

首先在网络性能测试中，如图 4 所示，在转发性能上，单路数据流场景中的磐石服务器架构转发率是普通服务器的 5.5 倍，而多路数据流场景中则为 3.1 倍；在网络带宽性能上，单路数据流场景中的磐石服务器架构带宽是普通服务器的 5.5 倍，而多路数据流场景中则为 2.1 倍；在绝对数值上，多路数据流场景的磐石服务器架构的转发率达到了 3100 万 PPS，同时网络带宽也达到 42Gb/s，结果令人满意[7]。

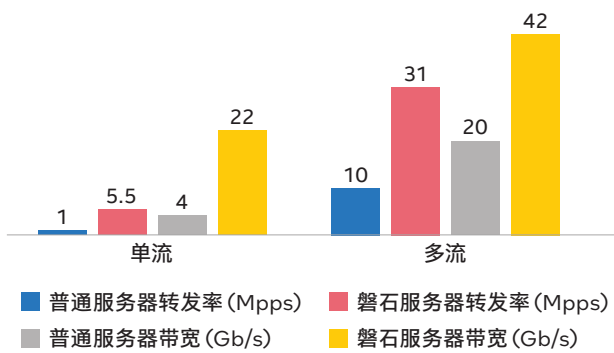

图 4 磐石服务器架构网络性能对比测试

- **存储性能测试**

而在存储性能测试中，如图 5 所示，在 IOPS 性能上，单盘场景中的磐石服务器架构 IOPS 是普通服务器的 5.5 倍；在存储带宽性能上，单盘场景中的磐石服务器架构带宽是普通服务器的 1.82 倍，实现了卓越的存储性能[8]。

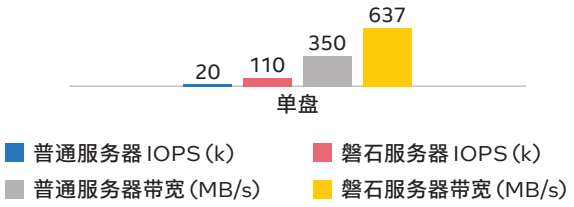

图 5 磐石服务器架构存储性能对比测试

总结与展望

一直以来，中国移动都致力于通过软硬融合技术的研究、引入大量核心自主设计以及持续进行产品性能调优，来为移动云构建高效基座。而英特尔服务器产品线经过数十年的发展与沉淀，同样也有着丰富的平台经验、完备的技术功能和成熟的行业生态。

此次强强联手的成功协作，一方面得益于中国移动在云服务领域积极探索形成的技术和生态优势，另一方面也离不开英特尔的成熟技术生态。在磐石服务器架构的优化过程中，双方共同针对用户在网络性能、存储性能上的需求开展技术攻关，并给出了相应的优化方案，例如引入社区版 SPDK 框架，对 vDPA/virtio-net/virtio-blk live migration 等的生态进行丰富等。这些联合调优工作有效地解决了磐石服务器架构在部署过程中的性能和稳定性等方面的挑战，满足了上层业务对底层硬件提出的新要求。

目前，磐石服务器架构已在移动云资源池成功落地，并成为移动云 IaaS 层的关键基础设施之一。未来，中国移动也将与英特尔一起，在更广阔的云资源池建设领域开展合作。例如，双方计划继续协同打造基于英特尔® IPU 参考设计的下一代面向 2 x 100G 和 2 x 200G 吞吐量的 Hypercard 硬件加速卡，来为新一代磐石服务器架构提供更有力的网络、存储和安全效能，从而助力移动云加速进入中国公有云行业第一梯队。

[1] 该观点援引自公开媒体发布的 IDC《中国公有云服务市场（2021 下半年）跟踪》报告解读：https://baijiahao.baidu.com/s?id=1735285320288343583。
[2] 该数据援引自中国移动 2022 年中期业绩报告：http://www.chinamobile.com/aboutus/news/groupnews/index_detail_42804.html
[3] 该观点援引自公开媒体发布的《从 DPU 的崛起谈谈计算体系变革》一文：https://aijishu.com/a/1060000000228825。
[4, 6] 数据来源于中国移动。测试组配置，处理器：双路英特尔® 至强® 铂金 8352Y 处理器，32 内核 / 64 线程，主频 2.2GHz，最大睿频 3.4GHz，超线程开启，睿频开启；内存：384GB (12*32GB 3200MHz DDR4)；网络适配器：HyperCard2.0；操作系统：BCLinux8.1，内核版本 4.19.0-193；
[5] 据援引自中国移动发布的《中国移动 DPU 技术白皮书（2022 年）》：http://finance.sina.com.cn/tech/roll/2022-08-05/doc-imizirav6801902.shtml。
[7, 8] 数据来源于中国移动。测试组配置，处理器：双路英特尔® 至强® 铂金 8352Y 处理器，32 内核 / 64 线程，主频 2.2GHz，最大睿频 3.4GHz，超线程开启，睿频开启；内存：384GB (12*32GB 3200MHz DDR4)；网络适配器：HyperCard2.0；操作系统：BCLinux8.1，内核版本 4.19.0-193。对比组配置，处理器：双路英特尔® 至强® 铂金 8352Y 处理器，32 内核 / 64 线程，主频 2.2GHz，最大睿频 3.4GHz，超线程开启，睿频开启；内存：384GB (12*32GB 3200MHz DDR4)；网络适配器：2*25GE 以太网卡；操作系统：BCLinux8.1，内核版本 4.19.0-193。