Case Study:

Intel® Infrastructure Processing Unit (IPU)
CMCC Cloud

Leveraging Intel's Data Center Technology, China Mobile Independently Developed the Hypercard Hardware Accelerator Card

Contents

Overview

Under the new situation that the country has launched the east-to-west computing resource transfer project and laid out a nationwide integrated computing network, more industries are further promoting the digital transformation process by migrating businesses to the cloud and using the cloud for key services. However, the resulting dramatic increase in the size of data also poses more challenges to the efficiency of the cloud infrastructure.

It is the persistent goal of China Mobile in building its Mobile Cloud to provide users with higher-quality, more efficient, flexible and secure cloud services based on innovative technical architecture and solid infrastructure. To accelerate the effort to achieve this goal, the China Mobile Cloud Capability Center (hereinafter referred to as Mobile Cloud) joined hands with Intel to build a Panshi architecture based on its proprietary BC-Metal bare-metal server and Hypercard hardware accelerator card, providing a compute power base for Mobile Cloud.

With products such as Intel® Xeon D® processors and Intel Stratix® 10 DX FPGA as its underlying hardware, Hypercard not only delivers outstanding performance through bare-metal servers, but also enables a variety of Infrastructure as a Service (IaaS) capabilities to be offloaded. This hardware implementation of device virtualization and data forwarding in lieu of the traditional software-only virtualization architecture can not only effectively reduce the computing overhead and free up server computing resources, but also significantly improve the performance of cloud services in networking, storage, among others.

Background and Challenges

Cloud service has gradually become a must-have for the "New IT Infrastructure" in various industries, and telecommunication operators are playing an increasingly important role in it. The data shows that, in recent years, the telecommunication operators have had higher growth rates[1] and are gaining more market recognition in the cloud services market. Taking China Mobile as an example, its Mobile Cloud revenue reached CNY 23.4 billion in the first half of 2022, an increase of 103.6%[2] over the same period last year. According to the analysis report, we can see that relying on the mature network environment and business model provided by China Mobile, Mobile Cloud has been deeply rooted in the China Mobile ecosystem, helping users to be in the cloud as soon as they are in the network.

This is not there by accident. Today, with more and more recognition of the concept of cloud-network convergence, telecommunication operators such as China Mobile, which are rich in cloud network resources and have good operational experience, are obviously the seeded players in the field of cloud services in the future. Particularly against the new background that China has launched the east-to-west computing resource transfer project on a full scale, China Mobile plans to build a computing network by deeply integrating a variety of information technologies, with computing power at the center and networking as the foundation, so as to provide a powerhouse for the country's social and economic development.

However, while China Mobile continues to enhance the influence of cloud services, the efficiency of its cloud infrastructure is also facing more challenges. The networking of the computing power and the expansion of cloud service applications mean that the data centers hosting the China Mobile cloud services will use more distributed computing and storage architectures. This will undoubtedly lead to larger-scale horizontal data traffic loads, including receiving, forwarding, storing and processing data. Internal analysis from China Mobile shows that the port throughput of its data center NICs is rapidly evolving from 10GBps to 25GBps and even 100GBps and above, which poses multiple challenges:

- Input and output (I/O) bottleneck caused by heavy data load and large data flow. In the traditional virtualization architecture, VirtIO and other I/O virtualization technologies usually require the deep involvement of processors, resulting in a huge consumption of computing resources. When the volume of data increases, performance jitter will emerge due to resource contention, which leads to Service Level Agreement (SLA) inconsistencies.
- Restriction of massive data forwarding on network performance. Data forwarding is typically done using virtual switching software such as OVS (Open vSwitch), which consumes a lot of computing resources. When data scales dramatically, OVS competes with other service processes in the virtual machine for computing resources, affecting both system performance and network bandwidth utilization.
- In terms of storage, with the diversification of workload types in the cloud service environment, the interfaces and protocol stacks for high-speed data storage are also changing rapidly, such as the virtio-blk standard storage interface and Non-volatile Memory Express over Fabrics (NVMe-oF). Traditional storage models are often not as resilient and flexible as they are expected to be, when used dealing with such new interfaces and protocol stacks.
- To cope with the comsumption of computing power caused by virtualization, China Mobile uses bare-metal servers to provide users with high-performance and highly available cloud service capabilities in critical scenarios. However, in the traditional cloud service architecture, the cloud management system typically shares processor resources with virtual machines, while bare-metal servers require exclusive use of resources. Therefore, it prevents bare-metal servers from elastic management and delivery. If nested virtualization is used, the overhead incurred can result in performance loss.

In fact, as the data traffic processing load has always been growing faster than the computing power in large data centers in recent years, its use of computing resources has been increasing. Statistics show that 30% of computing power in data centers is use on traffic processing, which is evocatively referred to as Datacenter Tax.[3]

One of the effective strategies to deal with the above challenges is to have more computing power while reducing its use. To have more computing power, you need to incorporate, in cloud services, products which consume less computing power and bring better performance. To reduce its use, device virtualization and data forwarding features are to be implemented using hardware in a new way instead of software in the past. It will effectively offload the IaaS capability and release the computing resources from the cloud servers.

Based on this concept, with the help of Intel's comprehensive chip and hardware and software ecosystem capabilities, the proprietary BC-Metal bare-metal server and Hypercard hardware accelerator card are developed for Mobile Cloud. it creates a new Panshi architecture providing users with much better network and storage performance in real-world cloud service applications.

Solution: The Panshi Server Architecture Provides a Comprehensive and Solid Computing Power Base for Mobile Cloud Scenarios

As more applications such as IoT, Artificial Intelligence, and Big Data analytics are deployed in the cloud, users have more requirements for high performance, low latency, and security in the cloud environment. To meet these requirements, Mobile Cloud has built a new generation of Panshi server architecture based on the hardware acceleration technology. As shown in Figure 1, the Panshi server architecture consists of the BC-Metal server and Hypercard hardware accelerator card.

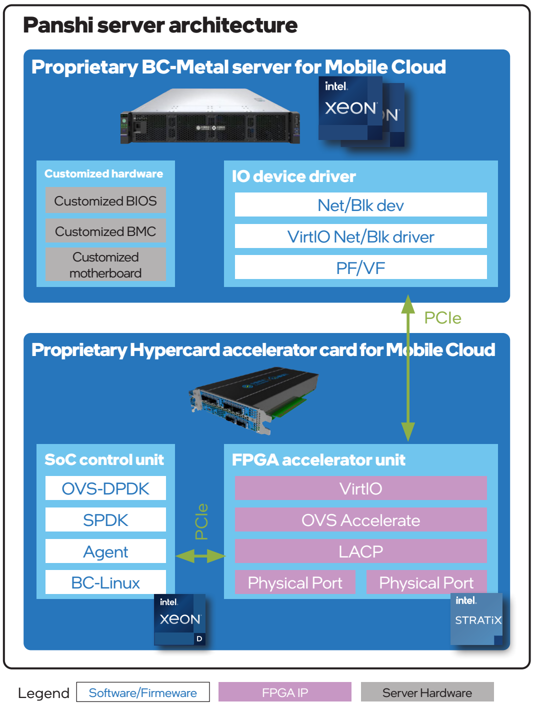

Figure 1. Mobile Cloud Panshi Server Architecture

The BC-Metal server is based on the 3rd Generation Intel® Xeon® Scalable Processor platform. Compared with general-purpose server, it is designed with various customized hardware, such as customized Basic Input Output System and customized BMC (Baseboard Manager Controller), to provide customers with more secure and efficient computing services. The Hypercard hardware accelerator card developed for Mobile Cloud is based on Intel® IPU reference design. The FPGA accelerator unit works with the SoC control unit to offload the IaaS service, providing cloud service providers with a leading IaaS service solution. The BC-Metal server is connected with the Hypercard hardware accelerator card via the Peripheral Component Interconnect Express (PCIe) bus and other custom interfaces.

■ Customized BC-Metal Bare-metal Server Provides Users with Excellent Performance for Key Services

Unlike common cloud hosting offerings, bare-metal servers allow cloud tenants to have exclusive access to resources such as processors and memory. This physical machine-like feature and benefit allows users to have high-density computing power and low-latency I/O support for their key services in a cloud environment. This design makes the choice of hardware infrastructure critical. To this end, Mobile Cloud combed through a large number of user business scenarios and added a large number of customized hardware and software products to the BC-Metal server based on business needs, including customized BIOS, customized BMC and customized motherboard. With these customized products, the BC-Metal server can provide efficient services for the user's business across its full life cycle, in terms of service monitoring, power management, system security and cooling efficiency, and operation and maintenance management.

Figure 2. China Mobile's Proprietary BC-Metal Server

At the same time, Mobile Cloud has partnered with Intel to incorporate the 3rd Gen Intel® Xeon® Scalable Processor (Ice Lake) as the core computing engine of the server for better server performance in multiple facets:

- More cores and better architecture bring a significant improvement in computing performance. Its single processor has up to 40 cores and 80 threads, and its max clock speed can be up to 3.1GHz, which can effectively meet the needs for high-density computing.
- More memory is supported, with max memory capacity up to 12TB. PCIe-Gen4 is supported, achieving higher I/O bandwidth per core. Moreover, up to six Intel® UltraPath Interconnect (Intel® UPI) channels effectively increase the bandwidth for I/O-intensive workloads across processors.
- A number of built-in enhancements, such as Intel® Deep Learning Boost Technology (Intel® DL Boost) and Intel® Software Guard Extensions (Intel® SGX) technology help in Artificial Intelligence (AI) scenarios and data security.

■ Hypercard Hardware Accelerator Card Effectively Offloads IaaS Services and Optimizes Computing Power

While the BC-Metal server helps to get more computing power in the Panshi server architecture, the Hypercard hardware accelerator card developed for Mobile Cloud is a powerful tool in the Panshi server architecture to offload IaaS services and reduce the use of computing power. In terms of physical specifications, it features a full-height, half-length PCIe design that connects to the server host via PCIe and custom interfaces, and has a separate power supply design. The accelerator card provides multiple SFP28 optical ports externally, which can be used for high-speed transmission of various types of service data.

Figure 3. China Mobile's Hypercard Hardware Accelerator Card

This brand new product was developed based on Intel FPGA and SoC design as a programmable network device that enables acceleration of infrastructure functions in the data center, resulting in smarter management of system-level resources. The core of FPGA accelerator unit is the Intel Stratix® 10 DX FPGA chip. With its advanced architecture design, packaging technology, more transceivers and its support on hardcore PCIe-Gen4 to achieve more bandwidth than the previous generation FPGA, and coupled with hardware full programmability, this programmable logic chip can flexibly collaborate with Intel® oneAPI to customize designs for higher performance including high throughput and low latency. It will help complete tasks such as I/O virtualization and OVS forwarding, manage infrastructure, and offload network and storage functions.

At the core of the SoC control unit is the Intel® Xeon® D Processor, which provides superior single-core performance, many security enhancements, and a host of built-in hardware acceleration capabilities to host the management functions offload required by the Panshi server architecture. The good x86 compatibility of Intel® Xeon® D Processors and the good ecosystem support with other Intel® architecture hardware can help users achieve rapid migration of system code or application capabilities, thus enhancing offloading efficiency.

- **Network Offloading**

With both chips, the Hypercard hardware accelerator card can effectively enable a wide range of IaaS offloading, covering networking, storage, and security. In terms of network offloading, first the accelerator card can "harden" the virtualization capability by offloading I/O virtualization to the FPGA chip. Cloud service typically uses I/O virtualization technologies such as VirtIO and SR-IOV (Single Root I/O Virtualization) to send and receive network data. Taking VirtIO as an example, it uses a front- and back-end separation architecture that allows virtual machines to interchange data with the VirtIO back-end in the host with the help of VirtIO drivers. Traditionally, this interchange process requires the involvement of the operating system core and consumes a lot of computing power of the processors. As I/O throughput continues to increase, performance bottleneck will emerge. In the design of the Hypercard hardware accelerator card, the I/O virtualization for VirtIO and others is offloaded to the FPGA chip. It not only significantly improves the I/O performance, but also fully frees up the system's computing power. As a result, the computing power consumed for virtualization does not exceed 10%[4].

Another type of network offloading is the "hardening" of the forwarding plane. For example, OVS is a common virtual exchange software in Mobile Cloud services, but it usually relies on the processor core polling for data forwarding processing. As a result, as data center bandwidth continues to increase, the processor resources required increase over time. According to estimates from China Mobile experts, a 25GBps-bandwidth cloud service scenario requires about 18% of processor resources reserved for it[5]. When the bandwidth further expands to 100GBps, this percentage reserved will need to increase, thus squeezing computing power needed for key services.

In the design of the Hypercard hardware accelerator card, the OVS forwarding plane is offloaded to the FPGA chip for processing, freeing up computing resources while significantly improving network forwarding performance, so that the network forwarding capacity can reach up to 31 million PPS[6].

- **Storage Offloading**

For traditional data storage processing, every I/O between network and storage devices requires frequent exchanges of data between the user and core states in an "interrupted" manner. The whole process requires multiple processor context switches and memory copies, which is not only less efficient, but also consumes a lot of computing resources.

Now, cloud service is incorporating protocol stacks like NVMe-oF to address those weaknesses. Remote Direct Memory Access (RDMA), for example, enables data to be delivered directly to network devices without going through the operating system by using techniques such as core bypass and zero-copy, thus eliminating the overhead of data replication and process context switching and significantly freeing up the processors. However, the NVMe-oF stack, if still processed by the processors, will undoubtedly result in lack of resiliency and flexibility.

To address this issue, the Hypercard hardware accelerator card offloads the virtio-blk standard storage interface, NVMe-oF protocol stack and others to the FPGA chip, which completes the core packet encapsulation/decapsulation, congestion control, and other workloads. In addition, the accelerator card has incorporated the SPDK (Storage Performance Development Kit) framework, which provides user-state acceleration capabilities such as polling, asynchronous operation, lockless NVMe drivers, Bdev common block layer and optimized application framework to effectively improve data storage and forwarding performance.

- **Bare-Metal "Cloudification"**

While BC-Metal bare-metal servers bring high performance to the Panshi server architecture, how to continue to make it elastic and agile for cloud services was also a key consideration in the design of the Mobile Cloud product. Bare-metal servers offload these tools and capabilities to the Hypercard hardware accelerator card, instead of deploying tools for virtualization. At present, the accelerator card allows Mobile Cloud to isolate physically infrastructure management and tenants, and support elastic bare-metal services by using the hot-swappable feature of VirtIO devices. Therefore, cloudified bare-metal servers can be provided to users in the way the traditional cloud hosts were provided.

It is worth noting that in a multi-tenant virtualized environment, the Panshi server architecture, which uses hardware acceleration technology, can also offload the virtual machine manager to the accelerator card. For example, it supports seamless hot migration of vDPA-based virtual machines, allowing cloud services to be more efficient in resource allocation.

Test Validation

To validate the improvements in network forwarding performance and storage performance achieved by the Panshi server architecture, Mobile Cloud, together with Intel, conducted a series of comparative tests for verification, and the results are as follows:

- **Network Performance Testing**

First of all, in the network performance test, as shown in Figure 4, the forwarding rate of the Panshi server architecture in the single-stream scenario is 5.5 times higher than that of the ordinary server in terms of forwarding performance, while it is 3.1 times higher in the multi-stream scenario. In terms of network bandwidth performance, the bandwidth of the Panshi server architecture in the single-stream scenario is 5.5 times that of the ordinary server, while that in the multi-stream scenario is 2.1 times. The forwarding rate of the Panshi server architecture in the multi-stream scenario reaches 3100 million PPS in terms of absolute value, while the network bandwidth reaches 42 Gb/s. The result is satisfactory[7].

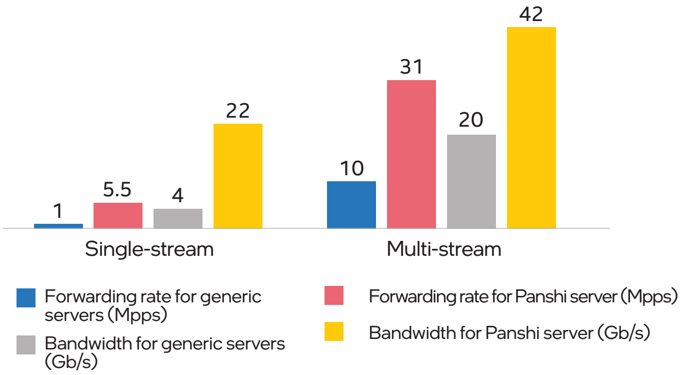

Figure 4. Comparative testing in network performance for the Panshi server architecture

- **Storage Performance Testing**

In the storage performance test, as shown in Figure 5, the IOPS of the Panshi server architecture in the single-disk scenario is 5.5 times that of the ordinary server in terms of IOPS performance. The bandwidth of the Panshi server architecture in the single-disk scenario is 1.82 times that of the ordinary server in terms of storage bandwidth performance. Excellent storage performance is achieved[8].

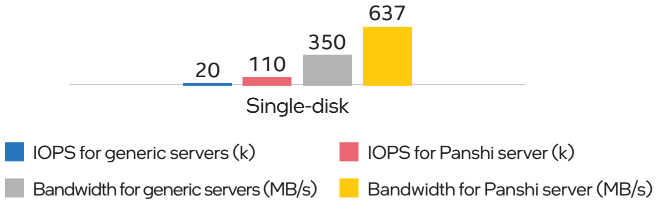

Figure 5. Comparative testing in storage performance for the Panshi server architecture

Summary and Outlook

For a long time, China Mobile has been committed to building an efficient base for Mobile Cloud through its software and hardware conergence technology researches, incorporation of a large number of core proprietary designs and continuous product performance tuning. After decades of development, Intel's server product line also has extensive platform experience, complete technical functions and mature industry ecosystem.

The successful cooperation between the two giants is facilitated by the advantages China Mobile has had in technology and ecosystem as a result of its active exploration in the field of cloud services on the one hand. On the other hand, it cannot happen without Intel's mature technology ecosystem. While optimizing the Panshi server architecture, the two giants worked together to address the needs of users in terms of network performance and storage performance. As a result, they provided optimized solutions, by introducing the community version of SPDK framework, and enriching the ecosystem for vDPA/virtio-net/virtio-blk live migration, among others. These joint tuning efforts effectively solved the performance and stability challenges of the Panshi server architecture in the deployment process, and meet the new requirements of the top-level services for underlying hardware.

At present, the Panshi server architecture has successfully gone live in the Mobile Cloud resource pool, and has become a key infrastructure in the IaaS layer of the Mobile Cloud. In the future, China Mobile will cooperate further with Intel in the broader development of the cloud resource pool. For example, the two sides plan to continue to work together to create the next-generation Hypercard hardware accelerator card for 2 x 100G and 2 x 200G throughput based on Intel® IPU reference design to provide higher network, storage, and security performance for the next-generation Panshi server architecture, thereby helping Mobile Cloud accelerate its effort to be listed among China's first public cloud service providers.

[1] This viewpont is quoted from interpretation of the IDC report *China Public Cloud Services Market (2H 2021) Tracker* published in media: https://baijiahao.baidu.com/s?id=1735285320288343583.

[2] This data is quoted from China Mobile's 2022 interim results report: http://www.chinamobile.com/aboutus/news/groupnews/index_detail_42804.html

[3] This viewpoint is quoted from the article *Talking about the Computing System Change from the Rise of DPU* published in media: https://aijishu.com/a/1060000000228825.

[4,6] Data from China Mobile. Test group configurations, processors: 2-slot Intel® Xeon® Platinum 8352Y processor, 32 cores/64 threads, 2.2GHz, 3.4GHz max turbo frequency, hyper-threading on, turbo on; memory: 384GB (12*32GB 3200MHz DDR4); NIC: HyperCard2.0; OS: BCLinux8.1, kernel version 4.19.0-193;

[5] As quoted from *China Mobile DPU Technology White Paper (2022)* published by China Mobile: http://finance.sina.com.cn/tech/roll/2022-08-05/doc-imizirav6801902.shtml.

[7,8] Data from China Mobile. Test group configurations, processors: 2-slot Intel® Xeon® Platinum 8352Y processor, 32 cores/64 threads, 2.2GHz, 3.4GHz max turbo frequency, hyper-threading on, turbo on; memory: 384GB (12*32GB 3200MHz DDR4); NIC: HyperCard2.0; OS: BCLinux8.1, kernel version 4.19.0-193. Control group configurations, processors: 2-slot Intel® Xeon® Platinum 8352Y processor, 32 cores/64 threads, 2.2GHz, 3.4GHz max turbo frequency, hyper-threading on, turbo on; memory: 384GB (12*32GB 3200MHz DDR4); NIC: 2*25GE Ethernet cards; OS: BCLinux 8.1, kernel version 4.19.0-193.

Performance varies by use,configurationand other factors. For more information, see www.Intel.com/PerformanceIndex.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel shall bear no liability for any purchase made in connection with the forecasts published in this document. All information provided here is subject to change without prior notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. For more information, please check with your OEM or retailer, or visit intel.com.

Tests document performance of components on a particular test, in specific systems. Any differences in your system hardware, software, or configuration may affect your actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more information on performance and results of benchmark test procedure, please visit www.intel.com/benchmarks

Intel does not control or audit third party data. You should review this content, consult other sources, and confirm whether referenced data is accurate.