Kunpeng 920: The First 7-nm Chiplet-Based 64-Core ARM SoC for Cloud Services

Jing Xia, Chuanning Cheng, Xiping Zhou, Yuxing Hu, and Peter Chun, *HiSilicon Technologies Company, Ltd., Shenzhen, 518129, China*

Kunpeng 920 is the second generation server processor designed by HiSilicon based on ARM architecture. Kunpeng 920 is able to achieve cost efficiency for various workloads through using a variety of chiplets and hybrid process technologies. The unique recomposition(s) of these flexible chipsets allows new designs to be created. The Kunpeng series processors combine technology innovations from various levels to improve efficiency, eliminate bottlenecks, and deliver value and performance. Its key features are as follows: The Kunpeng 920 core is specifically designed with superscalar architecture with the support of vector extension to provide leading features for high-performance computing applications; the coherent cache subsystem is created to integrate multicores into single chiplet (e.g., 7-nm process node) with a ring design that is ultralow-latency (<15 ns), nonblocking and bufferless; a dedicated parallel small-IO block is developed to achieve high-bandwidth (e.g., 400 GB/s) interdie connection for 2-D package solutions; IO die is redesigned (e.g., 16-nm process) so that the latest standard interface (e.g., PCI4.0) can be leveraged to scale up the System on a Chip (SoCs) and connect them with other IO devices; two or four Kunpeng 920 can work together as single symmetric multiprocessor system with cache coherent nonuniform memory access fabric.

The 21st century has witnessed a big data explosion that sets up profound challenges in processing and storing data. Thus, there has been a greater demand for high-performance processors that can handle data movements more efficiently. Superscalar CPUs serves as *de facto* solutions for requirements of operating system, control-intensive applications, single-thread programs, etc.

In January 2019, when Kunpeng 920 was launched, complex instruction set computer (CISC)-based server CPUs still dominated the market. They are Intel's flagship CPU Xeon platinum 8180,[6] IBM Z14,[5] and AMD ZEN1 series.[1] However, reduced instruction set computer (RISC) still has a higher theoretical efficiency proven in practice. For instance, ARM CPUs have occupied and dominated the mobile market.[2,7–9]

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3085578
Date of publication 1 June 2021; date of current version 14 September 2021.

Due to the development of the multicore systems, the energy efficiency advantage of ARM CPUs have been highlighted. The ARM64-based server/PC CPUs are used/recognized by various market leaders within the industry. In November 2018, Amazon released cloud computing infrastructure-as-a-service and platform-as-a-service platforms mainly for web servers based on ARM CPUs. Graviton[4] [graviton] contains 16 Cortex-A72 cores and has a network bandwidth of 10 GB/s. At the same year, Marvell (Cavium), which has been working in ARM server market for several years, launched ThunderX2,[10] which is a 32-core SMT4 ARM-v8.1 CPU based on TSMC 16-nm process node. The aforementioned CPUs have specific targets. For example, Amazon's multicore CPU is low power for concurrent web services, whereas Marvell is high performance for general applications.

As a more advanced process node is used, the cost of tapeout becomes higher. Fortunately, due to the benefits provided by chiplet technology, it is possible to divide a conventional SoC into smaller functions and implement them into different dies. This method

has two advantages: 1) significantly improving the yield of a single die; and 2) integrating dies of different processes for different functions (e.g., general computing/IO/AI) into one SoC to create a better balance between cost and performance.

With such ideas in mind, Huawei in 2019 launched an ARM SoC named Kunpeng 920 to support Huawei's "Cloud, Management and Edge" services. Kunpeng 920 could meet the following three design goals:

1) great scalability for multiple market segments (e.g., servers and PCs);
2) high energy efficiency and good performance;
3) reusability of dies.

By adopting the chiplet integration, the dies (i.e., chiplets) are designed for different functions, which can easily be reused in various applications. For example, the compute die can also be used in wireless base stations (e.g., TianGang series), whereas the IO die can be used to support the communication in AI accelerators (e.g., Ascend 910).

The workloads (e.g., big-data, distributed storage and ARM-native applications) are expected to be the target applications of Kunpeng 920. They require high concurrency, scalability, and fast IO on the chip. The design of Kunpeng 920 takes the following characteristics of such workloads into consideration:

- Processing such big data needs more computational power. According to the forecast of Huawei GIV2025,[3] the new data volume expects to reach 180ZB by 2025, 18 times of that in 2018.
- The single-thread instructions per cycle of such workload (e.g., HBase and Redis) are usually less than 1 given the applications profiling. For example, the latency of DDR access often causes the maximum compute efficiency to be reduced.
- For large amounts of data, distributed storage service is equally important. This kind of service has multithread interactive IO demands requiring low latency.
- A large number of edge-side ARM-native workloads can be gradually migrated to the cloud or be collaborated with the cloud in the future, hence resulting in the cloud to carry a large number of edge-side-like ARM-native workloads. This kind of workload requires 1) scale-out capability; 2) virtual machine capability; and 3) quality-of-service deployments. As a result, consistent ARM instruction set architecture (ISA) should be able to support better software development and performance tuning.
- More computational intensive workloads are handled by domain-specific accelerators. For example, memory-intensive workloads (e.g., HPCG—high-performance conjugate gradients) becomes increasingly important for servers to handle. Such workloads require low memory-access latency.

From the aforementioned workloads characteristics, the following four design principles for Kunpeng 920 are applied:

1) A high-performance CPU core should effectively process single threaded tasks to eliminate the long-tail effects. The CPU core should also have the ability of processing data in parallel.
2) Most of the ARM-based CPU/SoC designed for mobile devices is small, but Kunpeng 920 designed for cloud servers is much larger in scale. Hence, an elaborated cache coherence mechanism needs to be implemented to deal with multilevel cache coherence challenges.
3) A dead-lock-free, lack of buffer, and multiring Network on Chip (NoC) is designed. The NoC is adapted to cloud service characteristics (i.e., low latency, high bandwidth, and QoS-compatibility) and achieves minimum area.
4) To realize interdie communication, a proprietary coherent parallel interconnection mechanism is developed. With Chip-on-Wafer-on-Subtract (CoWoS) integration technology, the coherent interdie communication is able to be achieve low-latency (e.g., $<$ 15 ns) and high-throughput (e.g., 400 GB/s).
5) An IO interconnection is implemented with 16-nm IO die to obtain low cost and high performance. The IO die also supports the latest IO communication protocol (i.e., PCIe 4.0).

This article is organized as follows: The "Kunpeng 920 SoC Overview" section presents the high-level overview of our SoC. Next, there are the details of the CPU-compute die and compute-IO die in the "CPU-Compute Die Design" and "Compute-IO Die Design" sections, respectively. Then, presenting multilevel interconnection methods in the "Multilevel Interconnection" section. Through experiments in the "Evaluation" section, proofs that the aforementioned SoC has high-performance, high-energy, and high-area efficiency in different scenarios are explained. Finally, the "Conclusion" section concludes this article.

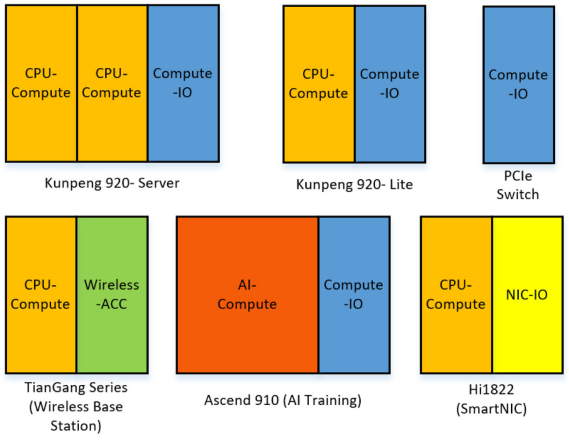

FIGURE 1. LEGO-based production.

KUNPENG 920 SoC OVERVIEW

A Kunpeng 920 SoC is designed using a LEGO-style architecture. The LEGO-style architecture can assemble multiple dies (i.e., chiplets), which can be integrated inside a single package using suitable process technologies. To achieve high integration efficiency and reduce design/manufacturing costs, several types of primitive chiplets are created. These primitive (LEGO-style) chiplets can be selected and combined together to meet various requirements. Furthermore, specialized die combination technologies and high-bandwidth coherent interconnects are proposed to facilitate the LEGO-style chiplet integration.

LEGO-Style Chiplet

There are five types of primitive chiplets, which are called *CPU-Compute die, AI-Compute die, Compute-IO die, NIC-IO die, and Wireless-accelerate (ACC) die* and are shown in Figure 1. These primitive chiplets share common connections and follow the same physical design rules illustrated in Figure 1, which are listed as follows:

- Different primitive chiplets can adopt various widths but have the same height.
- Two chiplets can only be combined in the horizontal direction.
- Due to the constraints of physical pins, AI-Compute dies, Compute-IO dies, NIC-IO and Wireless-ACC dies can only be placed on either left end or right end of the chip design. In other words, only a CPU-Compute die can be placed between two other chiplets.

The LEGO-style architecture permits various combinations of dies to be integrated seamlessly with different packaging technologies (e.g., 2.5-D Interposer or MCM—multichip module) to meet with different workload requirements. In Figure 1, there are five SoC chips introduced (i.e., Kunpeng 920-Server, Kunpeng 920-Lite, TianGang Series, Ascend 910, and Hi1822-SmartNIC). The LEGO-style chiplets can work as standalone SoC. For example, the Compute-IO Die shown in Figure 1 can also function as a PCIe Switch.

For Kunpeng 920, Silicon interposer (i.e., TSMC CoWoS, 65-nm process BEOL) is adapted as the interdie connection. Due to the high density of BEOL, interdie bandwidth is able to achieve up to 400 GB/s with coherency, which is much higher than the interconnection density obtained by competing products (e.g., AMD ZEN2) based on MCM.[a] The reason of using CoWoS is twofolded. First, when multiple memory-bound applications are allocated on the same computing die, the interdie connection to I/O most-likely becomes the bottleneck for MCM solution. Second, as Kunpeng 920 supports both uniform memory access (UMA) and nonuniform memory access (NUMA) modes, the UMA model can achieve similar performance as the NUMA mode only with the help of high-speed interdie connection.

Kunpeng 920 SoC Series

The Kunpeng 920 SoC is mainly designed for servers to implement the features, such as multicore parallelism, security, connectivity, and RAS (i.e., reliability, availability, and serviceability). Kunpeng 920-Server integrates two CPU-Compute dies and a Compute-IO die. Kunpeng 920-Lite is designed for the PC market that requires high reliability and good performance. It integrates a CPU-Compute die and a Compute-IO die that is capable of providing high concurrency and high-throughput interconnections.

CPU-COMPUTE DIE DESIGN

The CPU-Compute Die contains three components: 1) TaiShan V110 cores; 2) last level cache (LLC) design; and 3) bufferless NoC. The following section introduces them.

TaiShan V110 Core Design

The TaiShan V110 core is composed of multiple types of units to support high-performance scalar/vector processing and effective load/store operations. Note that Kunpeng 920 is designed for processing cloud computing workloads (e.g., data analysis and database workloads). Thus, the design space exploration takes

[a]Kunpeng920 is also downward compatible with MCM to create cost-effective solutions.

TABLE 1. Taishan V110 performance and performance/area.

	Intel Broadwell E5-2650 v4*	Intel Skylake 8180**	AMD ZEN1 7601***	Kunpeng920
Process	Intel 14 nm	Intel 14 nm	GF 14 nm	TSMC 7 nm
Area (mm^2, with L2)	6.9#	8.5#	7#	1.5
Performance	6.73	9.22	7.16	5.28
Max Freq. (GHz)	3.2	3.8	3.2	3.0
Performance/GHz	2.3	2.4	2.2	1.77#
Perf./Area	0.333	0.282	0.314	1.13
Test Condition	ICC 18.0	ICC 18.0	AOCC1.0	GCC9.1.0

#Estimated.
*https://spec.org/cpu2017/results/res2018q1/cpu2017-20180216-03627.html
**https://spec.org/cpu2017/results/res2018q1/cpu2017-20180122-02853.html
***https://spec.org/cpu2017/results/res2018q4/cpu2017-20181126-09823.html

the characteristics of the target workloads into consideration to determine appropriate microarchitecture of TaiShan V110 core. The details of TaiShan V110 are introduced as follows.

Superscalar Pipeline

The superscalar pipeline is mainly composed of instruction fetch, instruction decode, instruction dispatch, and integer execution. The Instruction Fetch Unit fetches up to four instructions from the L1 I-cache per cycle. The Branch Prediction Unit supports both static and dynamic branch prediction policies. The dynamic policy is specifically implemented with a two-level dynamic branch predictor that uses branch target buffer for high-speed target address generation. For the target workloads, a four-width superscalar pipeline is sufficient to achieve a good balance of power performance area (PPA).

SIMD Extension

TaiShan V110's SIMD architecture is compatible with NEON-basic instructions and also supports part of the extension instructions of ARM-v8.2. With the above acceleration framework, TaiShan V110 can enable vectorized acceleration for server applications as well as mobile ones.

Load/Store and Cache

Considering the target applications, it is important to guarantee high-efficient continuous vector accesses in addition to low-latency scalar loads/stores.

Core Parameters

TaiShan V110 is mainly designed for cloud computing (e.g., big-data and database) workloads, which tend to have high concurrency and involve extensive integer operations. Thus, the single-core score of SPEC INT 2017 benchmark is selected as a metric to evaluate the performance of CPUs, and mainstream CPU cores with respect to performance and performance/area are listed in Table 1.

According to Table 1, TaiShan V110 has much better PPA (e.g., three times better performance per area) compared to other baseline CPUs.

LLC Design

LLC Design Exploration

In recent years, the value of using shared L3 cache as LLC has been widely recognized on ARM-based SoC. With stringent physical constraints, there is a wide design space to be explored for LLC design. In the design of Kunpeng 920, the global LLC of the SoC is sliced into each CPU Cluster, so that the LLC and CPU Cluster form a NUMA relationship. Thus, how to select the proper size of each cluster needs to be carefully considered to maximize its benefits. Taking into account the multiple factors, four CPU cores per cluster are selected to obtain the best PPA score for the current process node.

Memory Strategy With Coherency

For the counterparts, LLC adopts either private mode or shared mode: Private mode is usually used when each CPU core hosts relatively independent data for tasks; shared mode is usually used when tasks within the SoC share a large amount of data. In private mode, each CPU cluster and the corresponding LLC slice form a private group, which can avoid the cluster from accessing high-latency cache slices. In shared mode, all LLC slices are combined to act as one block to increase the reuse of data inside the SoC.

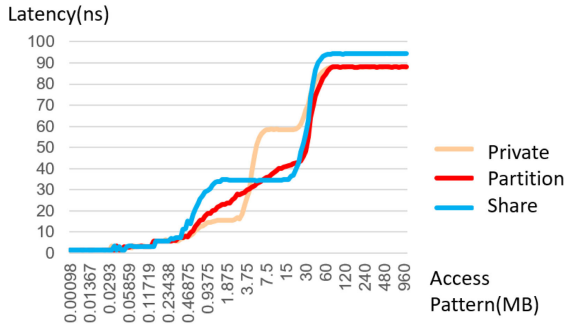

FIGURE 2. Kunpeng 920 cluster partition model merit.

In real-life applications, it is observed that a large number of applications would have different cores carrying different task slices (e.g. different tasks or different phases of the same task), leading to inefficiency for the homogeneous CPU SoC to handle these situations.

Therefore, assuming that a mechanism can dynamically allocate LLC based on runtime applications, the partition mechanism is introduced, which constitutes NUMA between the remote cache slice and the cluster's corresponding private cache. As measured in Figure 2, the current design of partition mode has a significant impact on the performance improvement of the Kunpeng 920—when the access pattern exceeds the nearby LLC slice, it is observed that the access latency rises gently instead of coarsely stepping up. Compared with other modes, partition strategy significantly reduces the average latency (e.g., more than 5%).

It is important to note that a full cache-coherence is enabled in intradie, interdie, and SoC level, which relies on the cache mechanism in conjunction with the HCCS (Huawei Cache-Coherent System) and HHA (HCCS Home Agent).

Bufferless NoC Design

For Kunpeng 920, a ring NoC is determined to be a suitable choice. In addition, a dual-ring is adapted to mitigate the shortcomings of data propagation in both directions. The bufferable NoCs consume considerable area and energy[12] and require complex logic designs, which lead to a poor scalability-limited frequency. The aforementioned limitations make the issues of "Memory Wall" more severe. Profiling real-life workloads shows that the proportion of uniform random traffic on the NoC is high, which implies that the bufferless NoC is a suitable choice for our SoC design. The results of electronic system level tests (bufferless NoCs transmit directly ratio is more than 99% in Libq. HPCG GeekBench and Bit-Data-Bench) show that the bufferless NoC saves significant buffer resources and greatly simplifies the NoC design at the cost of negligible latency jitter.

With the same bandwidth, the area of bufferless NoC can be reduced by 50%–70%[11] compared with bufferable NoC, under the setting of Kunpeng 920. The NoC occupies less than 7% area of the whole SoC. Its frequency is up to 3 GHz, which is much better than that of buffer-able NoCs.

Summary

Based on the findings in the aforementioned sections, the Kunpeng 920 CPU-Compute Die is designed as illustrated in Figure 3(a).

The Kunpeng 920 CPU-Compute Die consists 32 TaiShan V110 cores, each with its own L1 and L2 Cache; four cores share a common LLC TAG and their corresponding LLC DATA is mounted on the NoCs nearest neighbor for more flexibility; DDR controllers are organized in pair, and an HHA is implemented for each CPU-Compute Die's coherence management; generic interrupt controllers are used to manage interrupts; super cluster link layer connector (SLLC) is designed to support interdie communication with cache coherence, which is discussed in the section "Interdie Connection"; the Peri_ICL is used for the functions of the system and is adapted to solutions other than the Kunpeng 920, fully demonstrating the advantages of the LEGO-style architecture.

COMPUTE-IO DIE DESIGN

Kunpeng 920 enables cost effective yet high-performance IO interconnection through 16-nm IO die. It supports the newest IO protocol (i.e., PCIe 4.0).

Compute-IO Die's Application and Characteristic

As illustrated in Figure 3(b), the compute-IO die is responsible for Kunpeng 920's IO functions. Such compute-IO die is also compatible to be used with other SoCs (e.g., Ascend 910 AI accelerator). Additionally, Compute-IO die can be independently used as a PCIe-switch chip. Compute-IO die is designed to meet the requirements of high throughput and relatively long-distance communication, storage accessing, and acceleration of interactive tasks. The long-distance communication can be classified into three categories:

1) long-distance coherent communication;
2) PCIe bus (on PCB weak-coherence/noncoherence interconnection);
3) Ethernet (cable interconnection).

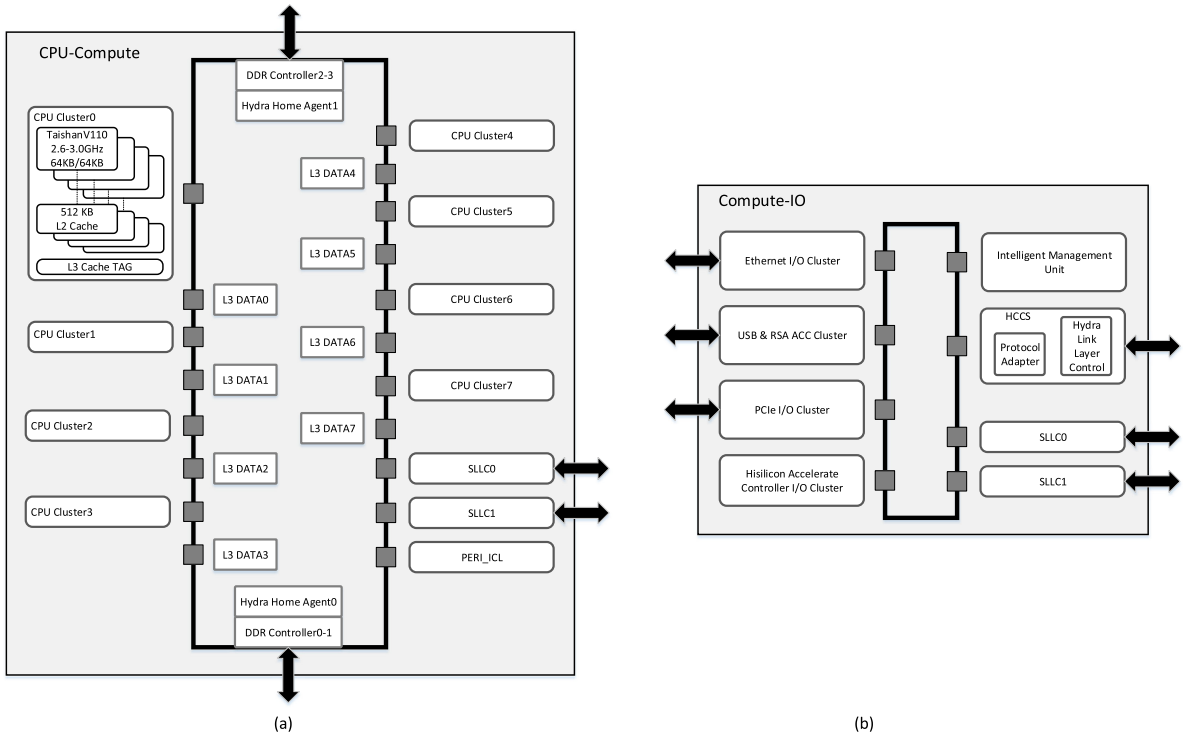

FIGURE 3. Kunpeng 920 chipset solution. (a) CPU-compute die. (b) Compute-IO die.

The majority of modules in Compute-IO Die is designed to support long-distance interchip communication, demanding relatively strong driving capability. The PHY occupies a large chip area (approximately 30%), which can hardly shrink with more advanced process node. Therefore, it is proper and effective to adopt 16-nm process node for Compute-IO Die. The relative longer channel length of the transistor at 16-nm process node also helps reduce the leakage.

KUNPENG 920 IMPLEMENTS A MULTILEVEL MEMORY HIERARCHY. THE PRIMARY PURPOSE IS TO PLACE A SUBSET OF DATA THAT ARE FREQUENTLY USED CLOSER TO THE CPU CORE SO THAT IT MANAGES LOAD/STORE MORE EFFECTIVELY.

Compute-IO Die's Solution and Evaluation

The Kunpeng 920 Compute-IO Die is the first in the world to implement PCIe 4.0 and supports up to 40 lanes. The overall bandwidth reaches 640 Gb/s. The PCIe interface supports thousands of channels in virtual machine scenarios, further enhancing the scale-out feature of Kunpeng 920 in cloud computing workloads. By supporting share virtual memory (i.e., the virtual address space shared between the accelerator and the user state), Kunpeng 920 offers better usability with respect to IO.

Kunpeng 920 16-nm Compute-IO Die solution offers a good performance and versatility with a low cost. As shown in the following sections, the efficient Inter-Chip coherence HCCS protocol hosted by HCCS on the Compute-IO Die provides a good support for the scale-up feature for cloud computing workloads.

MULTILEVEL INTERCONNECTION

Kunpeng 920 Memory Hierarchy

Kunpeng 920 implements a multilevel memory hierarchy. The primary purpose is to place a subset of data that are frequently used closer to the CPU core so that it manages load/store more effectively. Table 2 lists the expected bandwidth and latency parameters and illustrates how smoothly they change between each memory level. The two-level interconnections, including interchip and interdie connections are designed carefully to minimize unnecessary data movements in all levels of memory hierarchy. Details are introduced as follows.

TABLE 2. Kunpeng 920 memory hierarchy: Bandwidth and latency expected.

	Bandwidth expected (GB/s)	Latency expected (ns)	Characteristic (Coherency / Noncoherency)
Core(with L1)	6000	<2	Coherency
L2 Memory	6000	<4	Coherency
L3 Memory	3000	<15	Coherency
Interdie - Access L3 Cache	400	<30	Coherency
Main Memory	200	<90 (Interdie<110)	Coherency
Inter-SoC (4 SoCs/Pod)	50	<230	Coherency / Weak coherency

Interconnections

Interconnections consist of interdie and interchip connections. A series of solutions is applied to tackle the problems with respect to the two-level connections.

Interdie Connection

SLLC is designed to deal with interdie connection with coherency.

LEGO-style interdie connections are explained in the detail as follows:

- Physical Layer: It targets CoWoS and is compatible with MCM. Small IO (i.e., Communication PHY) is customized to adapt for interdie connection.
- Link Layer: It uses virtual channel to guarantee the effectiveness of interdie NoC.
- Flow Control: It adopts credit-based flow control to enable effective interaction between sender and receiver.

Finally, as listed in Table 2, low-latency and high-bandwidth interdie connection is implemented.

Interchip Connection

- Physical Layer: A customized larger PHY is designed to provide sufficient driving capability.
- Protocol Layer: A number of request iteration of coherence protocol is optimized to reduce the latency of a single-coherence-based load/store as the transmission latency increase (due to the longer transfer distance).

As a result, the interchip connection latency is maintained within twice the latency of interdie connection, as shown in Table 2. In Figure 4, an example of a multichip system architecture is shown consisting four Kunpeng 920 SoCs. Four Kunpeng 920 SoCs are connected with one another through HCCS interconnections.

EVALUATION

Experiment Setup

To sum up, Kunpeng 920-Server consists of 16 CPU clusters, 64-MB L3 cache, hardware ACC, HCCS and a series of IO. Each CPU cluster includes four Taishan V110 core. Each core has a 64-kB instruction cache, a 64-kB data cache and a 512-kB L2 cache. The IO consists of SAS/SATA 3.0 interfaces (up to 16 channels), PCIe 4.0 interfaces (up to 40 channels), low-speed IO and eight DDR4 controllers (up to 2933 MHz).

Results and Comparison

When the Kunpeng 920-Server was released in January 2019, the top performance server CPUs at the time were used for comparison including Intel Xeon 8180 and AMD EPCY 7610. They are from Intel Skylake architecture and AMD ZEN1 architecture, respectively. The performance of Kunpeng 920-Server is compared

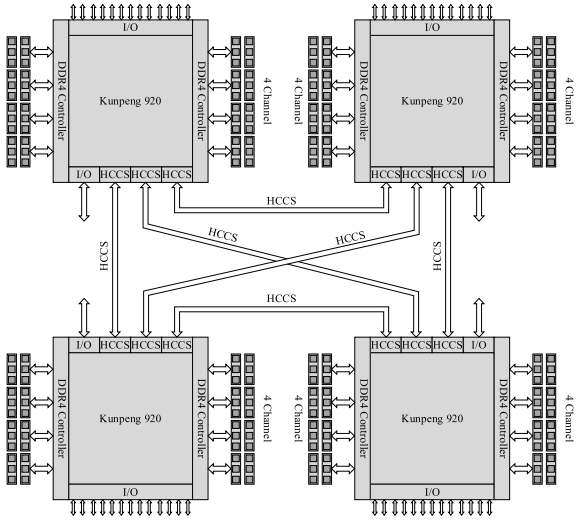

FIGURE 4. Kunpeng 920 multichip system architecture.

TABLE 3. CPU SoC performance evaluation.

	Intel Xeon 8180	AMD EPYC 7601	AMD EPYC 7742	Kunpeng 920
Process	Intel 14 nm	GF 14 nm	TSMC 7 nm + GF 14 nm	TSMC 7 nm+16 nm
Core Quantity	28	32	64	64
Thread Quantity	56	64	128	64
Area (mm^2)	698	756	1152	452
Performance (SpecINT2017)	134*	135**	323***	159****
Thermal design power (W,TDP)	205	180	225	200
Perf./Area	0.192	0.179	0.280	0.352
Test Condition	ICC 18.0	AOCC1.0	AOCC2.0	GCC9.1.0

*https://spec.org/cpu2017/results/res2018q1/cpu2017-20180122-02798.html
**https://spec.org/cpu2017/results/res2018q2/cpu2017-20180319-04037.html
***https://spec.org/cpu2017/results/res2020q1/cpu2017-20191218-20419.html
****Released by Peng Cheng Laboratory: https://spec.org/cpu2017/results/res2020q2/cpu2017-20200529-22566.html

with the aforementioned products in Table 3. We also put the AMD EPCY 7742 (AMD-ZEN2) in Table 3, which released six-months later than Kunpeng 920.

As is shown in Table 3, Kunpeng 920-Server has better performance than previously released products with a smaller chip size, which demonstrates the efficiency of Kunpeng 920. Kunpeng 920 still has better Perf./Area than later released AMD EPCY 7742 (AMD-ZEN2). In addition, Hisilicon's agile RTL-Fabrication co-design mechanism ensures that Kunpeng 920 is ahead of production on 7 nm, which ensures that Kunpeng 920 is the first 7-nm Server CPU to enter the market, which also contributes to the performance of the chip.

Behind Figure 5, it is also observed that there are greater performance advantage for Kunpeng920 (more than 20%) when Kunpeng 920's middle-end 5250 compare with Intel 6148/6248 under typical cloud workloads Hive-SQL.

Furthermore, the architecture of server CPU is also driven by the cost-effectiveness of mass production, also known as performance-per-dollar. Thanks to the benefits of LEGO-style architecture on yield and cost, Kunpeng 920 can achieve a high performance-per-dollar when the yield of 7-nm process is not so high, creating additional benefits to customers and driving up industrial progress with high performance.

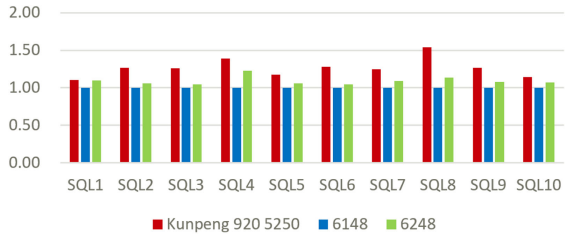

FIGURE 5. Big data Hive Perf. evaluation.

KUNPENG 920 ADOPTS A LEGO-STYLE CHIPLET-BASED DESIGN WITH HYBRID PROCESS TECHNOLOGIES TO ACHIEVE BOTH HIGH PERFORMANCE AND LOW COSTS.

CONCLUSION

In this article, Kunpeng 920 targeting for high-performance cloud computing workloads has been introduced. Kunpeng 920 adopts a LEGO-style chiplet-based design with hybrid process technologies to achieve both high performance and low costs. Kunpeng 920 is implemented with systematic solutions (e.g., bufferless Ring NoC and partition-based LLC) to address memory problems, such as coherency, latency, and throughput at multiple levels (e.g., inter-core, interdie, and interchip). With the combination of the aforementioned innovations, Kunpeng 920 demonstrates that LEGO-style chiplet-based design can be scalable and cost-effective to achieve high-performance data processing solutions. For example, the sales of Kunpeng series have been strong in Cloud and PC markets and the current results are ≈1,000 K PCS/Year. In summary, the LEGO-style chiplet-based design approach of Kunpeng 920 is proven to be effective because it can meet various PPA requirements and reduce design/manufacturing costs by the reusage of chiplets in various scenarios.

REFERENCES

1. "AMD ZEN1," 2017. [Online]. Available: https://en.wikipedia.org/wiki/Zen_(first_generation_microarchitecture)
2. "Apple-A14," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Apple_A14
3. "GIV2025," 2018. [Online]. Available: https://www.huawei.com/minisite/giv/Files/whitepaper_en_2018.pdf
4. "graviton," 2018. [Online]. Available: https://aws.amazon.com/ec2/graviton/
5. "IBM Z14," 2017. [Online]. Available: https://www.ibm.com/products/z14
6. "Intel-xeon-platinum-8180," 2017. [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/120496/intel-xeon-platinum-8180-processor-38-5m-cache-2-50-ghz.html
7. "kirin-9000," 2020. [Online]. Available: https://www.hisilicon.com/en/products/Kirin/Kirin-flagship-chips/Kirin%209000
8. "kirin-990," 2019. [Online]. Available: https://consumer.huawei.com/en/campaign/kirin-990-series/
9. "snapdragon-888," 2020. [Online]. Available: https://www.qualcomm.com/products/snapdragon-888-5g-mobile-platform
10. "thunderx2," 2018. [Online]. Available: https://en.wikichip.org/wiki/cavium/thunderx2
11. C. Fallin, C. Craik, and O. Mutlu, "Chipper: A low-complexity bufferless deflection router," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Architecture*, 2011, pp. 144–155, doi: 10.1109/HPCA.2011.5749724.
12. T. Moscibroda and O. Mutlu, "A case for bufferless routing in on-chip networks," in *Proc. 36th Annu. Int. Symp. Comput. Architecture*, 2009, pp. 196–207, doi: 10.1145/1555754.1555781.

JING XIA is a Chief Architect of KunPeng (CPU) with Turing Business Department, HiSilicon Technologies Company, Ltd., Shenzhen, China. He is the corresponding author of this article. Contact him at dio.xia@hisilicon.com.

CHUANNING CHENG is a Chief Architect of IO/Interconnect System with Turing Business Department, HiSilicon Technologies Company, Ltd., Shenzhen, China. Contact him at chengchuanning@hisilicon.com.

XIPING ZHOU is a Chief Architect of Computing Chip System with Turing Business Department, HiSilicon Technologies Company, Ltd., Shenzhen, China. Contact him at zhouxiping@hisilicon.com.

YUXING HU is a Researcher of Kunpeng (CPU)/Ascend (AI) with Linx Lab, Turing Business Department, HiSilicon Technologies Company, Ltd., Shenzhen, China. Contact him at huyuxing1@huawei.com.

PETER CHUN is a Sr. Manager of Technical Planning and Collaboration with Huawei, Shenzhen, China. He is a Member of the IEEE. Contact him at peter.chun@huawei.com.