LITE-ON Storage is now Solid State Storage Technology Corporation

Next-Generation Server CPUs

Next-Generation Server Overview

The unprecedented expansion of the Cloud in recent years, at least in part due to the ongoing pandemic, has created the need for faster and more efficient servers. A myriad of services can be hosted from a single platform but at the same time, it's necessary to have a selection of hardware to meet specific needs. For example, a data center may have different requirements depending on if it’s streaming games, engaging in high performance computing (HPC), or processing clients and data through online transaction (OLTP) or analytical processing (OLAP), respectively. Perhaps more importantly, flexibility and scalability – such as offered through a range of core counts with discrete chiplets/tiles and caching ability – help tailor the hardware to the workload to improve effective efficiency.

Recent developments in hardware, for example with DDR5 and PCIe® 5.0 (read our Alder Lake blog), have made it possible to have powerful platforms anchored by all-new central processing units (CPUs). The three major players – AMD, Intel, and ARM, the last of which is currently being courted by NVIDIA – have new hardware coming out in the next two years designed to meet the growing demand. While they share similarities in some respects, such as being able to leverage high bandwidth memory (HBM, a 3D form of SDRAM) or increased caches, there are also important distinctions that allow for delineation. As data centers must make the most of limited space there are trade-offs to be had depending on workload, but also certain bottlenecks need to be alleviated at the platform level.

AMD

AMD has two next-generation CPU families coming out for servers, based on the Zen 4 architecture with a basis in “Zen 4c” (“c” for cloud) and/or “Zen 4d” with Zen 5-based servers to follow. These are centered on the SP5 socket (LGA 6096) and are separated into two products lines: Genoa and Bergamo. Both have up to 12-channel DDR support with expected PCIe® 5.0 support as well, although there are some significant differences. Primarily, Genoa is intended to compete with Intel’s parts as a general-purpose CPU while Bergamo is oriented at the cloud. For this reason, Bergamo is designed around higher peak throughput rather than higher clocks, with for example a cap of 128 cores versus 96 for Genoa. AMD is utilizing a form of 3D cache, or V-cache, on these designs; they may compete with Intel directly using HBM in the future and will be using it with their more specialized AI chips.

There are also differences at the chiplet level with Genoa having 8 cores per core chiplet die (CCD) to

Bergamo’s 16. The intent is for Bergamo to be oriented at multi-thread performance while having better power consumption and core density, which includes a redesigned cache. Genoa, on the other hand, reaches for higher clocks – at 320W and up to 400W, which is unusually high. AMD intends to leverage TSMC’s 5nm process for the new architectures in order to achieve a significant improvement in instructions per clock (IPC). Ultimately this approach gives AMD flexibility in product offerings while supporting DDR5 and PCIe® 5.0 and remaining scalable. AMD also has the possibility to rework older architectures before committing to a Zen 5 design in the future.

Intel

AMD’s approach clearly illustrates the importance of cache and eventually HBM (and specifically, HBM2e) for upcoming servers. Intel especially has a focus on this coupled with their Optane DC persistent memory. Memory hierarchies, as such, can be useful for a variety of things, such as additional caching and even the elimination of the need for traditional DRAM. Intel has embraced these possibilities with their Sapphire Rapids Xeon (SPR) platform with optional HBM2e support. Specifically, this is through their non-universal memory architecture (NUMA) modes. Intel also has moved towards multiple dies connected with the embedded multi-die interconnect bridge (EMIB) or ultra path interconnect (UPI). Historically Intel has relied on monolithic dies while this demonstrates a new tiled architecture that allows for better scalability, particularly as core counts rise.

Intel’s SPR is utilizing Golden Cove CPUs with AVX512-FP16 and AMX instruction set support, using the same Intel 7 (10nm Enhanced SuperFin) process as Alder Lake but on the Eagle Stream server platform. PCIe® 4.0 and 5.0 are supported along with DDR5, plus of course HBM2e including as L4 cache. Intel promises a 19% IPC uptick versus Cypress Cove, through a number of architectural and production improvements, and these chips will utilize more L2 cache per core than the consumer parts. Intel has brought updates to other features, such as quick assist technology (QAT) for encryption and compression, matrix accelerators for AI, and more. Intel does have an additional advantage here with their Optane technology which can be used in both flat and HBM caching modes.

ARM

Although AMD and particularly Intel have been powerhouses in the server market, ARM may be underestimated by many who do not realize just how much is fueled by their processors in the server space. The best example would be Amazon’s Web Services (AWS), a leading cloud provider that powers much of the world’s web presence. Amazon intends to bring out the Graviton 3 CPU, but they are not alone in innovating here – there is also Ampere’s Altra, Fujitsu’s A64FX, and many more next-generation ARM-based designs. However, Amazon’s Graviton 3 provides a good example of ARM’s orientation within the industry. Amazon lists many workloads that will benefit from this architecture, from application servers to HPC and machine learning (ML). Many large customers rely on AWS services for these workloads, from ride-sharing to game streaming (see our related blogs).

Amazon’s upcoming ARM CPU will utilize a seven silicon die in a chiplet-based design with the promise of +25% performance per core over the existing Graviton 2. Floating-point performance is doubled, ML performance is tripled, and more. As with the upcoming options from AMD and Intel, ARM will have DDR5, PCIe® 5.0, and HBM support, with an eye on HBM3 specifically, promising much higher memory bandwidth. The core count will also increase, up to 64 cores, with a corresponding improvement in pipeline horsepower. In order for all of this to remain efficient, the new CPUs will be based on ARMv8.5 at 5nm with a power cap of around 100W. Clearly, these chips are meant to be more scalable and powerful than the previous generation while remaining particularly efficient compared to the competition.

Summary

In an age driven by information the amount of data collected and processed every day continues to grow. This data is becoming more useful over time as algorithms and machine learning adopt it to novel patterns. Over time more processes are becoming automated and people are interacting online in increasing numbers – whether through streaming, gaming, or virtual reality (VR). Meanwhile, the sciences require immense computing power and bandwidth for data analytics to help solve some of the greatest challenges to mankind, from decoding viruses to predicting the weather. All of this is supported in data centers and managed by server CPUs.

The leading cloud providers have seen business boom, allowing smaller companies to license server instances to improve their internal processes. Businesses both small and large can use this as a force multiplier, increasing the efficiency of everything from record-keeping to streamlining product introduction. The cost lies in server space, power efficiency and cooling, and effectiveness relative to the workload. Next-generation CPUs, therefore, have to be scalable and flexible, boosting bandwidth across the entire platform in order to mitigate potential bottlenecks. AMD, Intel, and ARM have some exciting technologies on the way that will ensure the world does not slow down.

Our SSD Solutions

PCIe™ - Our ED1 Series is a powerful, high performance SSD made for edge storage applications. It comes in M.2 and U.2 form factors.

SATA - Our ER2 SSD Series delivers affordability and performance with superior random read/write speeds of up to 90,000/45,000 IOPS. It comes in M.2 and 2.5" form factors.

Please contact our Solid State Storage Technology Corp. expert for more information.

Request Full Specs Sheets

ABOUT US

A subsidiary of KIOXIA Corporation, **Solid State Storage Technology Corporation** is a global leader in the design, development, and manufacturing of digital storage solutions. We offer a comprehensive lineup of high-performance customizable SSDs for the Enterprise, Industrial, and Business Client markets. With various form factors and interfaces, our SSD solutions help businesses simplify their storage infrastructures accelerating variable workloads, improving efficiency, and reducing total cost of ownership.

Learn more at www.ssstc.com