MARKET ANALYSIS

The Evolution of Server Architectures

Sponsored by Intel

SUMMARY

The dramatic rise in cloud computing is focusing new attention and driving important developments in the core hardware that's enabling it—servers. From their early beginnings as super-sized PCs, today's servers have evolved into a staggering array of different shapes, sizes, and configurations. New variations are designed to meet the rapidly expanding needs of modern data center and cloud computing environments. Along with these physical changes have come an even greater diversity of system architectures. These designs are intended to meet the huge array of different workload types—from traditional to virtual machine-based to containerized microservices and beyond—they now need to be able to run. As a result, whereas pure raw computing horsepower used to be one of the only metrics that was considered by server OEMs, enterprises, and cloud service providers, that line of thinking no longer makes sense. Instead, it's critical for server builders and end users to think not only about all the different hardware elements that guarantee a good performance match between different servers and different workloads, but the software tools that can be used to optimize those workloads as well.

"While raw CPU performance is important for servers, the demands of today's public, private and hybrid cloud-based solutions emphasize the need to think of servers as software-driven systems that need to be optimized for specific workloads, not just a box of CPUs."—Bob O'Donnell, Chief Analyst

INTRODUCTION

As many have observed, the COVID-19 pandemic has led to dramatic changes in how organizations are adopting cutting-edge technologies. That, in turn, is forcing them to rethink their computing infrastructure and its suitability to new business priorities. From a critical need to support distributed workforces, increasing dependence on the cloud, a growing demand for more advanced AI-driven analytics to improve the efficiency of their efforts, and substantial increases in security threats, organizations are taking on a wider range of workloads, many of which demand different hardware requirements. At the same time, there's been an explosion of different approaches to server design and an increasingly competitive array of different suppliers. The result is potential confusion about the how to best match the right solutions with the appropriate workload demands.

To better understand these issues, it's worth diving into the specific components of a server to see how they've evolved, and to understand how they can impact the suitability of different servers to different workloads.

SERVER HARDWARE EVOLUTION

Traditionally, most of the focus around servers has been on the CPU. After all, CPUs have been (and continue to be) the primary workhorses in getting critical computing workloads finished quickly. To that end, over the last few years there have been notable increases in the efficiency of individual processing cores, as well as the number of cores embedded into a single chip.

For certain types of applications, particularly traditional benchmarks, these advancements continue to be very important. However, given the wide range of workloads that servers now need to run, other aspects of server architectures also need to be considered. For example, new sets of instructions being integrated into the latest generation CPU architectures can help them be more effective or efficient for specific types of workloads.

The way different elements of chip are packaged together can also have an important impact on performance. High-speed lateral interconnects as well as stacked, 3D-like architectures are enabling both new levels of performance as well as new degrees of flexibility in piecing together the critical silicon components inside today's modern servers.

But as powerful and well-suited to a variety of tasks as CPUs may be, they aren't always the best or most efficient choice for many types of modern workloads. In fact, one of the most important trends in server evolution has been the proliferation of different standalone "xPUs"—chips specifically optimized for accelerating different tasks, such as graphics, AI/ML,

or even customized/programmable architectures (a task that FPGAs, or Field Programmable Gate Arrays, are best suited to). In these instances, new high-speed interconnect standards such as CXL (Compute Express Link) and CCIX (Cache Coherent Interconnect for Accelerators) are opening possibilities for integrating multiple components from different vendors (or even the same vendor) in a more "system-like" fashion outside the main CPU, but within the overall server architecture. Companies that offer a wide range of different XPU architectures have an advantage in being able to put together server systems that are specifically tailored to a variety of applications and workloads.

MEMORY AND STORAGE ADVANCES

In addition to compute enhancements, there's been important work done on memory and storage architectures within servers and, in some cases, even a blurring of the lines between them. One of the biggest challenges that's always existed for servers is moving enormous chunks of data into and out of the main computing elements, with various levels of caching used as a mechanism for managing and improving the efficiency of that data flow. Part of the problem has traditionally been that there's a big tradeoff between the speedy performance but extremely high cost of DRAM, and the much lower cost but significantly slower response of storage elements, such as flash-based SSDs.

Conceptually, the best way to address those differences is by creating technologies that incorporate elements of both, such as speed near that of traditional DRAM, but with the persistence of storage devices. A technology like Intel Optane PMem lives in this middle ground, offering higher capacities, as well as potentially faster performance and lower cost per GB versus the alternatives. Optane PMem doesn't replace either memory or storage, instead it works as a supplement to overcome common bottlenecks that occur in many server workloads.

Integrating these technologies into server architectures is yet another critical tool to help build servers that are optimized for specific types of workloads. In fact, for certain types of applications and environments, the performance benefits of integrating Optane memory and/or storage technologies can be significantly larger than changes in CPUs.

For many server systems, other interesting storage technologies that can have a big impact on server performance are VROC (Virtual RAID on CPU) and VMD (Volume Management Device), both of which are tied to Intel server CPUs. Together, VROC and VMD allow an NVMe-based RAID system to attach directly to an appropriate server CPU and boot on its own, without the added cost and complexity of a standalone RAID controller. This, in turn, can lead to significantly better system throughput and faster access to data in a cost-effective manner.

THE IMPORTANCE OF SOFTWARE AND TOOLS

On top of server hardware components, software tools are also critical for matching the right server to the right workload. Optimized versions of popular software development tools, for example, can not only help programmers save time, but they can also ensure that their code takes maximum advantage of all the unique hooks, APIs and instruction set extensions enabled by modern CPUs. While this sounds relatively straightforward, the real-world impact of these optimizations can make for very meaningful speed enhancements.

In the constantly evolving world of AI frameworks, having versions that are optimized for the unique capabilities of a given CPU can provide similar types of benefits, with the additional bonus of ensuring that all the most recent changes to the framework are properly mapped to the latest low-level software libraries that run on new CPUs.

Given the ever-expanding range of workload types, it's also important to think about the range of different software tools available from different vendors. While it's great to have a few general-purpose software tools, the likelihood of finding something that can help optimize very specific types of applications often remains low. In an era where servers are a key cog in delivering the complete solutions that both business and consumer customers have become so dependent upon, a wider variety of software options becomes increasingly important. As a result, not only high-level programming tools, but drivers, APIs and other types of customized software that can eke out the best possible performance from a given CPU for different types of workloads become key differentiators as well.

In addition, with the growing use of multiple types of computing components in modern servers, software tools that can help programmers fully leverage dedicated accelerators as well as core CPUs are becoming essential. One of the reasons we haven't seen a great deal of previous generation servers use multiple types of accelerators is that each kind required its own types of specialized software tools and different programmers with unique skill sets. The appearance of new APIs and other software tools that can help programmers who don't have experience with chip-specific software tools to still leverage them opens up enormous worlds of possibilities. Abstracting away the underlying complexity of these multi-chip or multi-accelerator hardware designs means that a much larger group of programmers can start to take advantage of them.

TESTING AND VALIDATION

Yet another critical, but commonly overlooked, differentiator across server component providers is the range of testing and validation tools and processes that they and their OEM

customer partners have. The more assistance a CPU vendor can provide to an OEM, the better the experience for the end customer.

Directly related are the assistance of people like support engineers who can help both system designers and software developers overcome any challenges they may run into. Again, the number of these hardware and software engineers and their range of experience can have a significant impact on how easy (or challenging) the successful completion of a project proves to be.

While it may not be immediately apparent, extensive experience in the creation and direction of key industry technology standards is also a key differentiating factor. Companies that have engineers who helped invent, develop and promulgate critical standards can leverage that knowledge into developing more robust testing and validation tests for those standards.

Conclusion

Given the increasingly vital role they play in all aspects of our lives, there's arguably never been a better time to be part of the industry that enables and deploys server technology. CPU performance certainly remains a critical differentiator in this new era, but it's only one tool in an increasingly diverse toolbox that server buyers need to consider when making purchases for specific types of applications and workloads. New chip instructions, new types of specialized accelerators, unique options for faster memory and storage access, advanced software tools to leverage the diverse and growing range of compute elements, and well-tested validation processes are all equally important factors that impact the real-world usage and effectiveness of servers. Ultimately, customers need to make decisions based on the overall value of the solution and how well it helps map to key business priorities that the organization is trying to achieve.