White Paper

Freedom from Legacy Constraints: The Future of Hyperscale Data Center Performance Optimization

Freedom from Legacy Constraints: The Future of Hyperscale Data Center Performance Optimization

Optimizing performance for cloud and hyperscale data centers requires new approaches to address the unique challenges of next-generation workloads. Dramatic inflections are occurring with capacity, bandwidth, latency, power efficiency, and cost. As technology leaders invest in these infrastructures, they look to proven and purpose-built solutions to deliver low power, high performance and exceptional TCO.

What drives this shift?

Memory-intensive new cloud workloads: With the growth of memory- intensive cloud-native applications vs. traditional enterprise applications, companies increasingly turn to flexible cloud, hybrid, and edge data infrastructures to minimize hardware dependencies, cut costs, and optimize performance.

Hyperscale infrastructures: An estimated 500 hyperscale data centers will be online worldwide by 2020. These massive data centers are scaling and provisioning on-demand, and setting the standard for global computing. However, they are bringing new challenges for balancing capacity and latency, as well as the soaring costs of power management.

Workload diversity and intensity: Data processing and computation requirements vary tremendously by application, and newer workloads in areas such as big data, machine learning, business analytics, matrix math, and complex decisioning demand larger memory capacity and greater memory bandwidth. Additionally, with increased adoption of high-level programming paradigms and interpreted languages, businesses need the interoperability and freedom of open computing standards and platforms.

Power efficiency and environmental sustainability: The massive power consumption requirements of data centers have led to new opportunities for innovations in renewable and sustainable energy management. Issues of power density and power delivery are also of the utmost importance in managing both capital and operational expenditures.

Designing for Performance: Considerations and Challenges

As cloud computing professionals optimize architectures to meet their business challenges, they must prepare for the specific requirements of emerging workloads for cloud and hyperscale datacenters. These include entirely new approaches to memory, power, and features. Because there is no “one size fits all,” the following issues should be considered:

Legacy system constraints. Businesses that must migrate some or all of their existing applications and processes to the cloud face specific logistical challenges and architecture demands. Efficiency and reliability may be limited by these legacy requirements, while cloud-native applications will not require these tradeoffs to achieve performance. Additionally, architectures that are optimized to handle legacy requirements may not be as adaptable to future business requirements as those that have more flexibility by design.

Workloads are rapidly changing. Every application has different data processing and computation requirements. Some require single-process operations, while others require multi-threaded or parallel multi-process operations with a large number of high-performance cores. Data-intensive applications may require both larger memory capacity and more significant memory bandwidth, and the infrastructure must be able to scale accordingly. It’s worth considering whether you need the flexibility of architectures that support broad needs or the more specific focus of an architecture that excels at the tasks of your unique workloads.

Optimizing capacity, bandwidth, and latency. Next-generation workloads are typically far more memory-intensive in both bandwidth and capacity, operating on significant amounts of data at rapid rates and holding data in memory rather than in the cache. As memory increases, along with the additional bandwith required to transmit that data quickly for computation functions, latency becomes a challenge as well. With the growth of high-bandwidth usage models from AI, machine learning, and IoT, data movement and network latency is increasingly challenging for data centers. Expansion in emerging workloads can cripple a data center that is not properly designed to manage latency throughout the system.

There's no single best answer for balancing these competing priorities. Data center capacity is driven by density, both for processors and for memory – chip designers should add as many memory channels as the chip can physically support to maximize bandwidth and capacity, while still ensuring low-latency connections to memory by performing significant electrical signal integrity work.

Memory Selection: Generally, the most cost-effective general-purpose option is DDR memory. While high-bandwidth memory can offer better latency and bandwidth, it comes with a high cost and significant power requirements. This may be required for a niche application, but is not the best choice at scale.

I/O architectures: Using general purpose high-bandwidth I/O, data processes can scale on demand for storage, file access, computing and communication tasks. This flexibility enables retrieve-and-compute capabilities with minimal latency.

Open computing standards: With more and more processing workloads residing in the various software layers, businesses need the interoperability and freedom of open computing standards and platforms rather than hardwiring or integrating a particular proprietary solution. Open standards and open APIs offer the flexibility for companies to take advantage of next-generation workloads without designing around hardware constraints or being limited by vendor lock-in.

Balancing usable density and avoiding stranded capacity: Designing for high-density can be challenging. If blades are not designed with the right power, they can become hot and the cooling costs become prohibitive, forcing the racks to be spaced further apart. These challenges of supplying power and cooling can limit the actual density of each rack, row, and data center, making it important to calculate not only density, but usable density as part of the planning process. Nearly half of the power costs of a data center are dedicated to cooling functions. The platform must therefore be designed not for a theoretical density, but for the actual usable density to deliver power and remove heat without having empty rows.

Additionally, different areas of the world have different power provisioning for racks. If the servers are running hot, and must be spaced for cooling, they will not be able to use the provisioned power supply, creating stranded capacity and leaving empty racks. Low-power chips, properly provisioned power supply, and improvements in cooling technologies are vital in improving usable density for hyperscale environments.

Power efficiency: Gartner estimates that ongoing power costs are increasing at least 10 percent per year, especially for high-power density servers which are requisite for hyperscale computing. Even small gains in efficiency for low-power chip design makes it possible for companies to save millions of dollars in operational expenses across thousands of servers.

Reliability, Availability, and Serviceability (RAS): It is vital to consider RAS best practices in systems infrastructures. With next-generation processors, built-in checking mechanisms can identify and fix issues in individual blocks before they can cause catastrophic system damage.

Security: Any data infrastructure design needs to have security assurances, such as putting functionality in the base of the platform, but this is especially important for cloud and hyperscale infrastructures. Innovations in security architectures in cloud computing enable vulnerabilities to be constrained, and machines that may be compromised in part can be immediately and automatically isolated from both the rest of the machine and the rest of the network. These newer architectures are also less at risk from the security issues of legacy software and older code bases.

Redefining TCO and ROI for the future of cloud computing

Total cost of ownership (TCO) for processors is quantified as performance per watt and performance per dollar. Purpose-built high performance, low-power solutions deliver unmatched TCO in both areas for public, private, and hybrid cloud architectures.

As technology leaders and data center managers make choices to optimize performance across workloads and applications, they must consider both upfront capital expenditures (CapEx) and ongoing operational expenditures (OpEx). Because of the unique requirements and opportunities inherent in cloud and hyperscale computing environments, TCO comes with a different set of considerations than traditional on-premises systems, leading to new ways to look at returns on investment (ROI).

Popular wisdom suggests that moving to cloud computing shifts expenditures from capital to operational, in that the upfront costs of servers and other hardware are distributed into ongoing platform-as-a-service expenditures, and that this is beneficial primarily to avoid issues of depreciation and obsolescence and distribute costs over time.

Though this may have been a useful framework for companies considering cloud deployments 10-15 years ago, it is no longer the way most technology leaders consider the ROI of cloud computing. The new ROI of cloud computing is predicated on more abstract – but still mission-critical – benefits such as business agility, time-to-market, and responsiveness to changing technical requirements. While these are all distinct advantages of cloud and hyperscale deployments, they can be nearly impossible to quantify on a spreadsheet.

Issues of density, scale, and power efficiency, on other hand, are all important parts of calculating TCO. The higher-density servers that comprise hyperscale infrastructures help manage capital expenses and reduce operational costs. Significant gains in power efficiency are also driving down operational costs, resulting in new possibilities for managing data center TCO.

Summary

Managing next-generation workloads in the cloud requires a sophisticated understanding of both the technical considerations and the business implications. Over the past decade architectures such as Arm® have taken the lead in a more energy efficient, lightweight, high performance approach to hyperscale datacenters. Building on those strengths, Ampere was founded to give customers the freedom from legacy constraints to accelerate the next generation of cloud computing for memory-intensive applications with the Ampere eMAG processors.

As the industry shifts to take advantage of all that cloud and hyperscale has to offer, the performance benefits of emerging cloud workloads consistently need to deliver the lowest power and TCO for private and public clouds.

AMPERE™

Ampere Computing
4555 Great America Parkway, Santa Clara, CA 95054
Phone: (669) 700-3700
https://www.amperecomputing.com

Ampere Computing reserves the right to make changes to its products, its datasheets, or related documentation, without notice and warrants its products solely pursuant to its terms and conditions of sale, only to substantially comply with the latest available datasheet.