TREND 4

# Harnessing hyperscale: Hardware is back (and never really went away)

Eclipsed by more than a decade of innovation in software, the hardware world is again a hotbed of new development as demand soars for bigger, faster, lower-cost data centers. Does your IT organization understand the new developments allowing companies to realize the benefits of "hyperscale" systems? In this new world, hardware matters more than ever in transforming enterprises into digital businesses with access to unlimited computing power that can be turned on and off as needed.

## Why now?

**Rising demand for scale:** Across industries, demand for processing at scale is surging. Businesses need reliable hardware to support the immense amounts of data processed for predictive analytics and real-time insights.

**Hardware and server architecture innovation surge:** From advances in storage to power consumption to processors to server architecture, infrastructure innovations such as nonvolatile memory are paving the way for faster, cheaper, and bigger hyperscale systems.

**Open source:** Facebook's Open Compute Project is accelerating the adoption of infrastructure innovations by sharing those breakthroughs freely. Founded in 2011, the Open Compute Project has already grown to more than 60 official members and thousands of participants.[1]





Who said that everything old is new again? He may have had a point—especially where computing hardware is concerned.

Not so long ago, every announcement of a new, multicore microprocessor or personal computer met with fanfare—lots of buzz about clock rates and cache memory capacity. But in recent years, the prevailing sentiment has been that hardware no longer matters—that x86 servers are nothing more than off-the-shelf commodities and all the important advances now happen in software.

Fast forward to today and looking to the future, it is becoming clear that "hardware as an afterthought" is a blinkered view. In fact, it is now a harmful view because it will make it more difficult for enterprises to evolve into digital businesses.

Every industry will be touched by the technologies being developed for the era of "hyperscale" computing systems—the supersized, super-scalable, and resilient data centers pioneered by heavily data-dependent companies like Google and Facebook. Innovations in technologies

such as low-power CPUs, solid-state data storage, and in-memory computing will benefit the performance of all enterprises' servers and data centers, enabling their next generation of infrastructure to support the digital transformation of their business.

Unilever, a consumer goods company; Pirelli, a tire manufacturer; and the NBA, a professional basketball league, are not companies that would typically be associated with one another, let alone be considered in the same context as Amazon and Google. And yet, similar to those technology giants, these and other traditional companies like them are faced with challenges that hyperscale computing can help to solve: immense amounts of data that need to be processed at speed. However, these companies had neither the scale nor the computing hardware to solve these challenges in-house. So they looked to hyperscale hardware appliances to help them perform at the speed of their nonstop businesses. For example, Unilever, Pirelli, and the NBA are all using SAP HANA to perform real-time analytics on huge datasets to gain insight and competitive advantage.[2] HANA gives them the capabilities of hyperscale that Amazon and Google enjoy in a single appliance instead of an entire data center. Others are using solutions from Oracle, IBM, and Teradata.

These data challenges exist across every industry and for companies of all sizes. This is why every CIO will again be tasked with understanding the advantages and trade-offs of hyperscale systems and big data appliances, as well as the opportunities and risks behind choosing which hardware will be used to inform and optimize their digital business.

## Inside hyperscale

Before we examine the hardware innovation that is being propelled by hyperscale developments, it is important to explain what the word *hyperscale* means. Accenture uses the term to describe not just the physical infrastructure of giant and distributed systems that support the data centers and provide companies such as Google, Amazon, and Facebook with the computing power to deal with vast volumes, variety, and velocity of data, but also the ability to scale computing tasks to achieve performance that is orders of magnitude better than the status quo. Hyperscale data centers consume storage, bandwidth, memory, and computing cycles on a scale unimaginable to most. They make pragmatic use of the latest hardware innovations, but never at the expense of scalability.

What does this scale actually mean? In the world of hyperscale data centers, there are very few players. Google is the largest with well over a million servers; Microsoft also has more than a million. Then comes Amazon with about half a million, and Facebook and Yahoo with a couple hundred thousand each.[3] In terms of power consumption, Google's global operations continuously draw 260 million megawatts of power—roughly a quarter of the energy generated by a nuclear power plant.[4]

Not only are these companies operating on a scale that defies most people's imaginations, but the services provided by these hyperscale data centers are expected to be up 100 percent of the time—not just 99.99999 percent of the time. This is the era of "always on." As a case in point, note that companies such as Amazon, Facebook, and Google are not just responsible for their own bottom lines—they are increasingly responsible for numerous other businesses' survivability as well. Today, hyperscale data centers are pivotal to an enormous ecosystem of companies and organizations that rely on hyperscale platforms as a necessary component of their business models.

But the demands on hyperscale computing do not stop with the need to be always on. It is equally critical to be "always optimized"—application-specific computing technologies are available to solve increasingly specialized tasks. For example, hyperscale data centers with the ability to horizontally scale worldwide with commodity servers are appropriate for a website. But complex calculations for applications such as 3D rendering, DNA sequencing, and cryptography are more effectively and efficiently handled by a system of servers with advanced graphics processing units (GPUs)—specialized processors optimized for parallel processing of very large floating-point number calculations. GPUs are so effective at these tasks that application-specific integrated circuits (ASICs) are starting to become available that offer orders of magnitude more processing capability than server farms full of advanced GPUs with dramatically less energy consumption and a smaller physical footprint.

There are two main reasons why this hyperscale trend is important to the enterprise. First, hyperscale is driving significant hardware innovation that will put businesses at a nexus where decisions on public or private cloud, as

well as commodity versus specialized hardware, become significant differentiators. Second, these hyperscale systems are no longer limited to large Internet companies.

Utilities, oil and gas, and automotive are only a few examples of industries that are quickly approaching the hyperscale level of interconnected devices, sensors, and data centers. Utilities now have networks of integrated sensors and smart meters that rival those of communication companies. Ford, GM, and Toyota are building vehicles with hundreds of sensors, telematics, and real-time connectivity. As all of this information begins to be collected every day, every hour, and, in extreme cases, multiple times per second, the systems required to store and analyze this data at speed will soon demand hyperscale solutions. Take gas turbines at power plants, for example. "We're almost putting a data center on a gas turbine," GE said recently in reference to the hundreds of sensors the company is placing on those machines to capture data. If those sensors, combined with robust data analysis, are able to improve efficiency by just 1 percent, that could save nearly $6 billion a year.[5]

**"Not only are these companies operating on a scale that defies most people's imaginations, but the services providzed by these hyperscale data centers are expected to be up 100 percent of the time—not just 99.99999 percent of the time. This is the era of 'always on.'"**

# The innovation behind the hyperscale push

The rapid growth of hyperscale systems has sparked a renaissance in hardware innovation from which all businesses stand to benefit. While the tech-savvy reader can jump into a bit more depth around the technology in the "Driving innovation in hardware" sidebar in this chapter, everyone should understand that innovations around processors, storage, and specialized hardware are proving to be opportunities for tangible benefits for the business.

Putting limits on design specifications, such as energy consumption, has driven innovation. Hewlett-Packard (HP) drew inspiration from mobile processors for its Moonshot servers, which operate with up to 89 percent less energy than traditional servers.[6] In a similar focus on energy efficiency, Facebook reduced energy consumption by 38 percent in its Prineville, Oregon, data center compared with existing facilities by using technologies developed as part of the Open Compute Project.[7]

Storage advances are providing enterprises with fresh ways to access and manipulate data faster. Flash storage arrays are becoming the norm for e-commerce sites and financial services firms with trading platforms.[8] E*Trade uses EMC's XtremIO storage array to maintain exceptional performance for customers while performing data maintenance tasks in parallel with no performance degradation, giving E*Trade the ability to scale operations and support new features for users.[9]

And innovations in server architectures are offering more options to match specialized infrastructure to the applications being run, helping CIOs identify the important requirements of each application, whether that be performance, scale, or cost. Illustrating the advantages that specialized hardware can provide, SAP has 28 customers that now run analytics jobs in HANA 10,000 times faster than they did previously. Three of these customers have experienced 100,000 times improvement. One of those three, Yodobashi, a Japanese electronics retailer, took a three-day batch process to analyze loyalty customers down to two seconds, enabling real-time couponing at the point of sale.[10]

Although the details of these innovations may not need to be completely understood outside of the IT department, it is important to recognize their potential to transform business processes and strategies. In doing so, it's paramount for companies to understand the advantages and trade-offs of these new tools as they look to procure, build, and use their next generation of IT systems.

## Adopting an open-source view of hardware

Facebook's Open Compute Project is another development that is causing significant discussion within CIO and other technology circles. The initiative involves openly sharing hardware innovations, following the model associated with open-source software projects. Members participating in this exchange of technology innovations include Goldman Sachs and Fidelity, which, in conjunction with AMD and Intel, have been reviewing new motherboard designs to incorporate in their own data centers.[11] The payoff for enterprises can be significant. For instance, Facebook designed and built its own servers and software, claiming that it can build its data centers at one-fifth of the cost of a traditional data center.[12]

## Asking hard questions about hardware

The relationship between hardware innovation and rising demand for cloud services is raising some new questions for IT leaders. First and foremost, over the next five years, every IT department will consistently be faced with the choice between leveraging external clouds and building computing infrastructure on premise. To make matters more complex, on-premise choices are capital intensive and increasingly specialized. There are computing appliances that give hyperscale capabilities from a bevy of manufacturers—SAP HANA, Oracle Exadata (now enhanced with Exalytics), IBM PureData Systems (an evolution from Netezza), and Teradata are all top-tier options. Choosing the cloud does not remove this complexity either. Not all clouds are equal; there are many varieties from which to choose. How CIOs make this choice will depend significantly on the systems they are looking to build and their demands.

One company may need highly resilient services that fluctuate little during the workday, but it may demand more of those services at the end of every month. Another may require real-time decisions on vast quantities of data.

The array of decisions now required is reflected in the sheer number of choices to be made when configuring an instance on Amazon Web Services (AWS). Amazon provides a wide range of configuration options to suit all sorts of business needs—ranging from "compute optimized" (with a high ratio of CPU to memory and lowest cost per virtual CPU of all Amazon's instance types) to "storage optimized" (suited to applications with specific input/output and storage capacity requirements).

So the decisions that IT leaders make about their data centers and cloud services must consider the underlying hardware. How will this provider's use of low-power CPUs in its data centers affect operational costs over the contract term? Can our applications run on low-power CPUs, or would GPUs be more efficient for computation? Does the code for our critical business insights need to be rewritten to take advantage of these new technologies?

To what extent can the data center scale up for a pharmaceutical client, say, if the client starts putting most of its clinical trial simulations in the cloud? Do our business requirements mean that we are better off with flash, in-memory, or hard-disk storage? The choice of hardware depends very much on what the specific application needs are, and what the usage patterns will look like.

On this digital journey, solutions must be tailored for individual use cases. Every organization is going to face a set of heterogeneous requirements that will be best served using the correct recipe of commodity versus specialized hardware and private versus public cloud architecture. This means that for the foreseeable future, enterprise infrastructures will be, by necessity, hybrid solutions—weaving hyperscale cloud, on premise, specialized hardware, and an enterprise's existing systems into a computing fabric that serves more parts of the business in more demanding, reliable, and scalable ways than before.

# Adopting hyperscale

In a dramatic shift of IT strategy, companies are leveraging hardware innovation on a piecemeal basis, with some companies going so far as to build their own hyperscale systems as a competitive differentiator. GM is one such company that is embracing the hyperscale concept and building out its own mega data centers. This is not just another private cloud, according to GM: "IT isn't here to manage daily operations. It's a strategic tool to drive business forward."[13]

With that in mind, the automaker is setting up its own hyperscale environment starting with the first of two data centers located in Michigan. Moreover, GM has cancelled its multibillion-dollar outsourcing agreements, believing that it needed to bring these strategic capabilities in-house to be closer to the business decisions needing to be made in a rapidly changing industry.[14]

GM's goal is to accelerate new product introductions, especially when it comes to keeping up with the growth of on-board vehicle electronics. Imagine fleets of vehicles with thousands of sensors sending gigabytes of data back to the data centers every second of every day from every corner of the world—all while maintaining high reliability levels in its data centers. Downtime at one of its factories can cost GM $1 million per minute, given current just-in-time inventory practices.[15]

And GM is not alone. By 2018, some 10 million BMWs will be connected, asking for, and receiving more than 1 terabyte of data every day.[16] More large organizations are interested in building their own high-efficiency data centers. Gartner estimates "large data centers"—those with more than 500 racks—will account for 29 percent of all data hardware revenue by 2017, up from 22 percent in 2011.[17]

Even if CIOs do not have an appetite for building their own infrastructure, companies still need to examine their path to becoming digital enterprises and determine whether they have the infrastructure and skills needed to support it. There is no set path on how to grow to meet the hyperscale challenges that businesses are going to face. But forward-looking companies are seeing the advantages of hyperscale technology.

At a minimum, they are seeing an opportunity to drive down the operational cost of running their data centers. As such, high-performance companies are increasingly recognizing that hyperscale systems are a vital part of becoming a digital business. To get started, technology leaders need to ask themselves, "What could our business do with unlimited compute power that can be turned on and off as needed?"

## Your 100-day plan

In 100 days, make sure your organization is informed about the available hyperscale technology options.

- Ensure that your IT organization is aware of consortiums and/or is testing the benefits of the latest hardware innovations. Identify those that are most important to your business.
- Identify your data storage needs and the magnitude of devices producing data in your network (including sensors, smart meters, devices, and data centers). Forecast their expected usage based on one-year and three-year business growth strategies.
- Create a plan that allows key data assets to be portable across architectures.
- Explore participation in open-source communities such as the Open Compute Project to leverage emerging hardware innovations.

## This time next year

In 365 days, be prepared to have your IT road map include hyperscale technologies. Have the knowledge to make the right investments and act.

- Update models of your digital business's most demanding compute processes to understand the advantages, tradeoffs, opportunities, and risks of hyperscale hardware choices.
- Create a hyperscale task force. Have it include hyperscale hardware during infrastructure planning.
- Build the reference architectures for hyperscale workloads and evaluate the applications that are deployed to hyperscale. Maintain hybrid hyperscale deployment strategies that account for improvements in offerings across SaaS, PaaS, and hardware to support varying enterprise activities.
- Investigate how edge-device computing can make applications more efficient and provide a vehicle for machine-to-machine interaction.

SIDEBAR

# Driving innovation in hardware

Innovation in three corners of the hardware world is driving hyperscale:

## Processors

Historically, complex instruction set computing (CISC)–based servers ran on high-performance x86 processors. This helped to rein in costs and maximize compatibility for mass-market applications. Over the last decade, however, the massive consumer demand for mobile phones has driven radical progress in the capabilities of low-power processors—largely to the credit, and benefit, of ARM Holdings. Low-power processors give data center designers an opportunity to address data center power costs which, in many cases, have risen to be 30 percent or more of operational expenditures. Calculations show that the cost of power to run a server over its lifetime will very often eclipse the cost of the server itself.[18]

Innovation is not limited to power consumption. ARM licenses its chip designs, creating a rich ecosystem of chip manufacturers and software developers able to pursue niche opportunities. Benefiting from this ecosystem, HP recently announced that an upcoming module for its Moonshot server line will use a new version of Calxeda's ARM-based EnergyCore chips. HP claims that this new module will offer nearly double the performance and four times the amount of memory per module compared with the previous generation of low-power processors.[19]

Complementing these low-power CPUs, the massively parallel architectures of graphics processing units (GPUs) are ideal for big data analytics. As a consequence, x86 servers are being replaced by servers with low-power ARM processors, using GPUs as accelerators or coprocessors. This allows massively parallel database (MapD) technology to realize tremendous speed gains by storing the data in the onboard memory of GPUs instead of CPUs, as is typical. Use of a single high-performance GPU can make data processing as much as 70 times faster than a conventional CPU.[20] That is why 53 of the top 500 supercomputers now use GPU accelerators or coprocessors.[21]

## Storage

Organizations must now store more data and get more insights more quickly from the data. With data processing capacity radically improved, the bottleneck becomes the ability to store and retrieve large volumes of data as rapidly as needed.

Traditionally, data has been stored either in limited volatile memory, which provides the fastest access, or on disk, which provides much more storage but is far slower. Companies could have their data fast or cheap, but not both. Now, though, innovations in flash memory are making this trade-off unnecessary.

Vendors such as Fusion-io and Violin Memory are providing flash memory products that address hyperscale issues such as the need for low-latency, low-power consumption; a smaller data center footprint; and a high degree of resilience. These companies are providing case studies and performance numbers that indicate the potential for application performance to increase by at least 25 times, and sometimes by as much as 40 times.[22]

**"Innovation at the server and data center levels is causing a bifurcation between the types of processing capabilities being offered at hyperscale: those optimized for general-purpose, Web-scale computing and those optimized for specialized tasks."**

That means that CIOs now have new options for solving their scalability challenges. They can start asking different questions: Is tuning application performance a good use of expensive developers? Will future applications run most efficiently on commodity hardware that scales horizontally? Does specialized hardware better serve the business needs? What are the trade-offs? What is the total cost of ownership?

## Servers

Innovation is also occurring at the server and data center levels. In fact, this innovation is causing a bifurcation between the types of processing capabilities being offered at hyperscale: those optimized for general-purpose, Web-scale computing and those optimized for specialized tasks.

For general-purpose computing, data center designers now have two options to build at hyperscale: low-power systems such as HP's ARM-based Moonshot or commodity x86 servers. In the most extreme cases of the latter, Google, Amazon, and Facebook are all sourcing custom server designs from OEMs, optimizing for their specific operational objectives and computing and storage needs. Companies looking to build or expand their own commodity x86 data center architectures may find benefit in joining the Open Compute Project. For those looking to radically improve operational costs, low-power systems make sense. While these low-power systems can be configured to run standard Web or big data applications such as Hadoop, they are not compatible with software written for x86 architectures. However, the operational savings may be worth the development investment; HP claims that typical data center operational costs can be reduced by as much as 77 percent with Moonshot.[23]

With specialized tasks such as analytics on large databases and computationally intense tasks such as weather forecasting and DNA sequencing, general-purpose hyperscale systems are insufficient to deliver results at speed and scale. Database software companies such as SAP and Oracle have risen to the challenge by building in-memory computing appliances that give orders-of-magnitude performance improvements on big data analytic challenges. Existing customers of SAP

and Oracle will find streamlined transition paths with HANA and Exalytics, respectively. These platforms and others offer the benefits of hyperscale in a configuration optimized for on-premise installations.

Moving to a greater variety of complex data processing workloads, however, calls for supercomputing solutions. Similar to the in-memory-computing appliances, these individual systems do not scale, but they do provide capabilities that go beyond what hyperscale can deliver. Amazon understands this, which is why its Elastic Compute Cloud (EC2) includes two of the world's top 500 supercomputing clusters (numbers 64 and 165).[24]