G00293310

Predicts 2016: Servers Changing Roles in the Data Center

Published: 4 December 2015

Analyst(s): Martin Reynolds, Joe Skorupa

Servers are about to take a broader role in data center architecture, as new memory capabilities, software network appliances and hyperconverged architectures make current hardware obsolete. Data center managers must plan new server deployment strategies for 2018, to cut costs and gain flexibility.

Key Findings

- New memory technologies will increase server memory footprints by five to 10 times, inside current cost envelopes. 5TB servers will become affordable, enabling new implementations of transaction processing, analytics and machine learning.
- Software appliances will use servers to displace networking equipment, a trend that is already displacing hardware in Web-scale environments. Simpler networking hardware will displace budgets from capital to license and maintenance.
- Hyperconverged systems will turn into software appliances, boosting servers to 70% of data center purchases and materially increasing data center transaction bandwidth.

Recommendations

- Inventory applications that can use large memory footprints, and make remediation plans to accommodate new memory technologies before 2018.
- Design budgets to shift from capital networking equipment purchases, to servers running network functional software under license and maintenance agreements.
- Identify, optimize and capture budget savings available from HCIS deployments, as HCIS software displaces stand-alone storage hardware.

Strategic Planning Assumptions

- By 2018, 15% of server memory bits will use in-metal memory technology.

- By 2018, 20% of new server capacity will be used to displace networking and security equipment, up from less than 2% today.
- By 2018, 70% of hyperconverged infrastructure sales will be license-only on Tier 1 servers.

Analysis

This document was revised on 9 December 2015. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

Servers are subject to three major forces that will change the shape of the data center. First, Web-scale technology is driving standardization, pricing and operating costs to new lows. Taking advantage of these servers opens cost reduction opportunities across data center infrastructure.

Second, new memory technologies that provide flash memory scale but DRAM-like properties will enable servers to take on complex applications at a fraction of today's cost. Large memory systems could drop to a fraction of today's price, offering 5x or 10x capability improvements. Both Intel-Micron (3D XPoint) and HP (memristor) are promising delivery of these technologies.

Third, network functional virtualization can drive out the need to buy hardware for security or application management. Instead, open source or maintained software, running on a server, can provide agility and flexibility at a fraction of the price of a rack appliance. Application delivery controllers and network security appliances are already moving to software, affecting the growth of the hardware network appliance market.

Finally, a hyperconverged integrated system (HCIS) brings new scalability and flexibility to compute and storage requirements. HCIS brings easy system scalability, and its interweaving with compute capacity improves its transaction bandwidth by reducing network chokepoints. It is important to deploy these technologies now, to take advantage of the performance and flexibility they offer as appliances. All-software versions should emerge by 2018, increasing flexibility and choice.

What You Need to Know

Data center managers can make significant gains in performance and cost by taking advantage of these emerging technologies. The key is to inventory applications and functions now, and plan how to deploy them in 2018 or sooner, against these new opportunities and requirements. In 2018, we expect Intel's Purley platform to deploy, bringing these new technologies into the mainstream.

In the interim, we will see new classes of Peripheral Component Interconnect Express (PCIe) nonvolatile storage emerge — late in 2016 or early in 2017 — that can act as caches for solid-state drive (SSD) main storage. By caching frequent reads and writes, main storage can move to less expensive multilevel cell flash, without concern of exceeding lifetime write limits.

- Audit your applications. Identify those that will benefit from large nonvolatile memory footprints, and put plans in place to take advantage of the new technology in 2018. Follow Intel's roadmap announcements and align your plans with Intel's Purley technology refresh.

- Identify networking functions that can be replaced with software components, and build roadmaps to integrate them into your data center architecture. Work with your data center providers to identify and implement these opportunities.
- Plan to implement hyperconverged systems in 2016, taking advantage of their flexibility, capacity and performance to support agile development and deployment.

Strategic Planning Assumptions

Strategic Planning Assumption: By 2018, 15% of server memory bits will use in-metal memory technology.

Analysis by: Martin Reynolds

Key Findings:

DRAM is the staple today for server main memory. However, the physical limitations of storing a measurable charge at the surface layer of a piece of silicon constrains DRAM density and cost advances. Flash memory solves the problem by building 3D structures in the metal, with 64 layers delivering 128Gb (16GB) flash memories. However, these memories are slow. First, because they require a block write, rather than a row write. A block write requires all rows in the block to be erased and rewritten. Second, they use multiple charge levels to store multiple bits per cell. This process is very sensitive, and requires continuous voltage and timing optimization.

There are other approaches to moving the memory element off of the surface silicon, and into the metal array. In-metal memory scales with the number of layers, and leaves the silicon surface area open for control elements. The challenge is finding a storage element that works well in the metal array.

The leader — until recently — was HP's memristor technology, which stores data as ions that migrate across a layer of titanium dioxide. The storage elements can be tiny, as the ions are far more robust and stable than electronic charge stores. However, HP has never demonstrated working silicon.

In 2015, Intel announced the in-metal product 3D XPoint that appears to have similar characteristics to the memristor. Intel is secretive about the memory element, but we believe it to be a form of phase change memory (PCM) that Intel and Micron have scaled down to tiny dimensions. Intel has publicly demonstrated the technology in an SSD, shown technology diagrams and disclosed key parameters, and leaked roadmaps that show that 3D XPoint will be mainstream in 2018.

Our belief is that this memory, based on existing PCM products, will have a cycle time in the 100 nanoseconds (ns) read and 300ns write range. These speeds are slower than DRAM, but adequate for many large memory projects.

Indications, based on wafer images and die size, are that this memory has a density comparable with advanced NAND flash, but at two layers rather than 64 layers. Therefore, manufacturing costs

will be similar to those of flash memory. Performance will be close to DRAM, with the added benefit of nonvolatility.

If the technology succeeds, it will benefit from existing layer and multilevel storage capability, and has many years of growth ahead. Given the public demonstrations, and barring some kind of unforeseen problems, this technology will go from demonstration to mainstream over the next three years.

Market Implications:

We anticipate that Intel will drive this technology into the server market, targeting applications that require large main memory systems. There are several implications to this strategy.

Pricing will be carefully managed. At $3 per GB, we believe that the product will achieve 80% margins. However, mainstream DRAM is priced at about $4 per GB, and server memory — bought through the vendor — can be 10 or 20 times this amount. These margins are important to server profits, and suppliers will hold prices high rather than expand the memory market. Therefore, to expand the market, Intel will have to drive end-user demand through its sales engineers, and create processes that allow servers to be delivered with this new memory but without the traditional memory markup.

We believe that a 5TB system will cost tens of thousands of dollars, rather than hundreds of thousands of dollars. This price reduction will open up new markets for in-memory computing; online, complex transaction processing; large-scale analytics; "cool" virtual machines (VMs) (little used, but needing a fast response); and scaled-up applications based on DBMS that scale best with large memory systems, which cover most vendors other than Oracle.

We estimate that there is capacity — and need — for this technology to represent about 15% of all server bits shipped in 2018. As these chips are only about 10% of the size of a DRAM chip, there is plenty of manufacturing capacity available. Also, these chips will not replace DRAM in the foreseeable future (within five years). Instead, they will significantly increase the number of servers shipped with large memory footprints. They will, however, add downward pricing pressure to DRAM.

Recommendations:

Applications require work in several areas to use these large memory systems.

Nonvolatility is one of the most important aspects of this new technology. There are four implications for server users:

- **Security —** Although we expect that platform memory encryption will be available, there are always surprises. Recycled, stolen or seized servers will potentially contain recoverable data in storage formerly presumed to be lost when the power is lost. Therefore, sensitive data may still need to be encrypted, using a volatile key in DRAM loaded from an external server.
- **Consistent state —** One of the benefits of nonvolatile memory (NVM) is the ability to rapidly recover from a system failure. However, to do this, the NVM must always have a consistent,

recoverable state. Applications need to be tuned to guarantee that the NVM is always fully recoverable without state errors.

- **Storage drivers —** If the application cannot take advantage of the NVM, a memory-based storage driver can convert the space into extremely fast storage. Although not as efficient as coding for fast main memory, this approach can substantially accelerate applications with minimal effort.
- **Booting —** It should be possible to create a small boot partition in the NVM, eliminating the power and cost of hard disk or SSD boot drives.

There are other aspects that drive the need to tune applications. As this memory is somewhat slower than DRAM, it is important to keep code in DRAM space. This need will drive new options in operating systems and server management.

Also, the very large memory footprint will require multiple processor threads to remain active to use the memory space. Processors will need to keep the DRAM busy. As such, there is a trade-off between the number of cores, core performance features, cache efficiency and memory transaction bandwidth to manage.

Related Research:

"The Spectrum of IMC Styles Meets the Spectrum of Business Needs"

"Market Guide for In-Memory Computing Technologies"

"Market Guide for In-Memory DBMS, 2015"

Strategic Planning Assumption: By 2018, 20% of new server capacity will be used to displace networking and security equipment, up from less than 2% today.

Analysis by: Joe Skorupa

Key Findings:

IT organizations from mid-market to Web-scale service providers are adopting new ways to build and service networks. Instead of deploying dedicated hardware appliances, they are deploying software instances in off-the-shelf servers. This approach is often called network function virtualization (NFV). For example, Amazon Web Services (AWS) incorporates an application delivery controller (ADC) as one of its per-use services. This ADC is not a physical device: rather, it is a customized software instance, written and tuned by Amazon that runs on server capacity on an as-needed basis. It can be sold on a per-use schedule, or as structured payments. The key is, these ADCs work just like a physical ADC, but the associated server revenue does not form part of the ADC market revenue. Similar trends are forming around firewalls (e.g., VMware's NSX firewall), and intrusion detection system/intrusion prevention system/data loss prevention (IDS/IPS/DLP) security appliances.

In the WAN, many service providers are looking to virtualize existing customer premises equipment (CPE), including branch office routers and firewalls, with software instances running on server infrastructure (NFV) located within service provider data centers.

Market Implications:

Buyers will see broader choices in terms of deployment options — physical appliance, virtualized physical appliance, virtual appliance offered as a service by their cloud provider, and embedded within the WAN and offered as a service. While fit-for-purpose choices will be a positive trend, it can lead to management complexity since no single vendor offers a complete solution under a single management framework, and multivendor management platforms are limited at this point.

Traditional Layer 4-7 (L4-7) vendors will be torn between their need to preserve exiting hardware-centric revenue streams while attempting to satisfy customer demand for more flexible deployment software-centric options. Additional pressure will come from Web-scale providers such as Amazon, Google and Microsoft, which developed its own ADC software as well as supported open-source vendors. Discounting virtual editions of traditional dedicated appliances to compensate for the cost of the associated servers will diminish vendor revenue further.

These trends align with those in the hyperconverged systems market, where networking, virtualization and storage systems are integrated into a single, scalable module. Just as with the large Web-scale providers, enterprise data centers will bend toward more servers and fewer dedicated networking and storage subsystems. Such a data center could easily spend more than 70% of its hardware budget on servers, up from about 40% today (based on average data center spend numbers).

Recommendations:

End users should:

- Not assume hardware is required for network and security functions that have previously been delivered in an appliance form factor.
- Pilot new virtualized deployment options to reduce costs and gain flexibility for network and security buyers, but to also ensure that adequate performance remains. Subscription pricing, total capacity pricing (instead of instance pricing) and site licensing/volume discounts can reduce costs.
- Calculate the costs/savings of managing this new mixed deployment model before committing to server-based deployments. Include integration and the cost of additional x86 and virtualization licenses as well as operational concerns.
- Build implementation cost cases for physical and virtual implementations, and use them to decide when to implement a new deployment strategy.
- Prefer software options that align with existing management tools.

Related Research:

"Magic Quadrant for Application Delivery Controllers"

"Market Guide for Communications Service Provider Software-Defined Network and Network Function Virtualization Solutions"

"How to Determine the Right Mix of Network and Security Appliances (Physical, Virtualized, Virtual) for Your Data Center"

"Tech Go-to-Market: Breaking the Bond Between Software Value and Underlying Hardware Disrupts Long-Standing Sales Motions"

Strategic Planning Assumption: By 2018, 70% of hyperconverged infrastructure sales will be license-only on Tier 1 servers.

Analysis by: Martin Reynolds

Key Findings:

Hyperconverged infrastructure is a $1.1 billion market in 2015, growing at 68% a year to $4 billion in 2018. The fundamental premise is powerful. By distributing storage across the network using storage management software integrated with a hypervisor, applications can be close to their data. Furthermore, the distributed computing capacity enables sophisticated storage management features.

The performance optimizations needed by storage area network (SAN) architectures are eliminated, so the overall cost drops. Networking becomes more efficient as hot spots are distributed across multiple links. And finally, the virtualization and management capabilities provide for agility in application deployment and provisioning.

Nutanix, the largest HCIS provider, is already on track to deliver a software version of its product, as evidenced by its hardware partnerships with Lenovo and Dell. SimpliVity brings another angle: an accelerator card that performs compression for each server and provides nonvolatile memory for transaction integrity. However, the card blocks a software-only implementation, but we believe that the benefits of the card are outweighed by a software implementation and anticipate that SimpliVity has to shift to software-only to capture growth.

The combination of performance improvement, flexibility and agility, and lower costs, are a compelling reason to move to HCIS. Furthermore, the rich margins associated with enterprise disk storage fuel investment and growth for these emerging companies.

Market Implications:

To capture the storage profits, the provider has to sell the drives. Selling the drives means that they have to be in a server, or at least directly attached in the rack. Consequently, the HCIS providers today are biased toward hardware sales. Their hardware products are not in the same league as

those from the major providers, with immature support networks, limited global reach, and the need to design servers as a secondary skill.

Therefore, a licensed software model becomes an imperative for sustained growth. Users will purchase the licenses as part of a packaged system from a major provider, or add the software to existing or new hardware.

Nutanix is already moving to a software model. Its partnerships with Dell and Lenovo give the company's products global reach in terms of sales, support, distribution and configuration centers. The breadth of configurations and form factors available from these suppliers makes the HCIS product easier to assimilate. Therefore, buyers can choose between a configured system from a major vendor, self-branded hardware from the HCIS provider, or software for installation on a platform of choice. Buyers focused on agility and simplicity will select the appliance approach; buyers focused on manageability and cost will tend toward a software solution.

There are two forces at play. The first is the reduction of profits associated with enterprise drives as HCIS makes the components part of the Web-scale movement. The second is the need of the HCIS vendors to expand their markets by shifting to sales on top of established data center server providers.

The net result is a shift of profits from hardware, where the greater margin contribution comes from incremental storage drives, to licensed software, which could, for example, be based on the drives and capacity attached to a server.

This shift will be material. Our forecast shows HCIS systems at $5 billion by 2020, compared with $24.5 billion of external storage. Therefore, HCIS systems will be 15% of total storage spend, or 5% of the total budget. With the virtual machine capacity included, HCIS could foreseeably become 20% of data center spend — where it works well — in 2018.

As it matures, HCIS might also provide a compelling platform for networking components. Its distributed nature, high connectivity and available processing resources allow software networking components to readily scale with demand. Therefore, the shift to HCIS software is part of a larger move in data center infrastructure that will result in most of the hardware budget moving to servers.

Recommendations:

- Seriously consider software-based HCIS to leverage existing infrastructure and reduce capital expenditure (capex) as well as operating expenditure (opex).
- Plan to shift up to 20% of your infrastructure deployment to HCIS by 2018, if it works for you.
- Identify, optimize and capture budget savings available from HCIS deployments.
- Leverage HCIS agility to capture business-unit led projects.

Related Research:

"Deploying Hyperconverged Integrated Systems: Eight Great Use Cases"

"Competitive Landscape: Hyperconverged Integrated Systems — Single-Vendor Solutions"

"Forecast Analysis: Integrated Systems, Worldwide, 1Q15 Update"

"Forecast Analysis: External Controller-Based Storage, Worldwide, 2Q15 Update"

A Look Back

This topic area is too new to have on-target or missed predictions.

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Deploying Hyperconverged Integrated Systems: Eight Great Use Cases"

"Competitive Landscape: Hyperconverged Integrated Systems — Single-Vendor Solutions"

Evidence

The analysis and insight presented here are based on the contributing analysts' regular tracking of the industry. This is achieved through frequent engagements and discussions with market participants, including communications service providers (CSPs), network equipment providers and other market participants. This deep industry knowledge/insight is then used by the contributing analysts to anticipate the future of the industry. The predictions also had to undergo a rigorous internal peer review process.

More on This Topic

This is part of an in-depth collection of research. See the collection:

- Predicts 2016: Algorithms Take Digital Business to the Next Level

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp