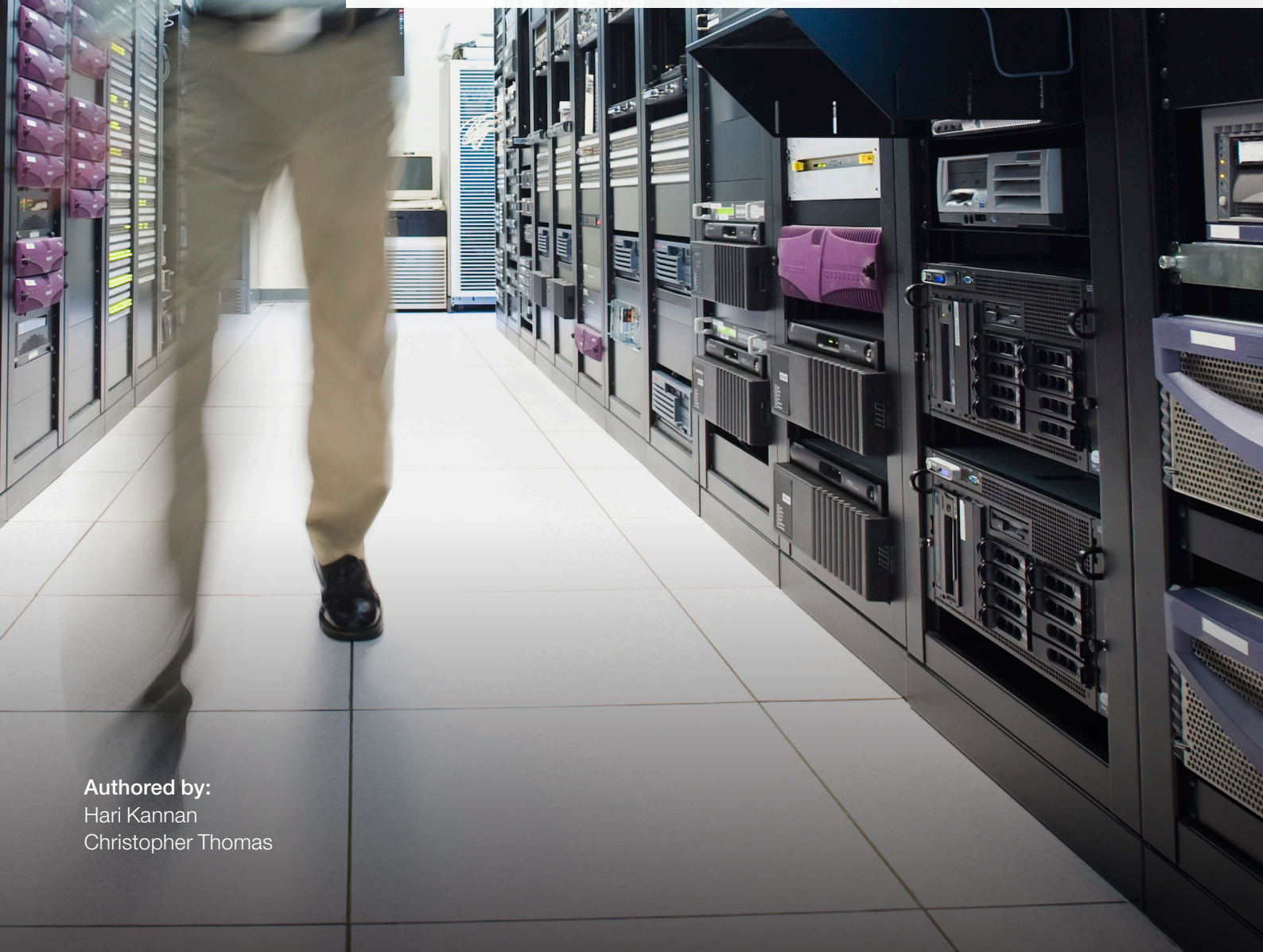
McKinsey&Company

How high-tech suppliers are responding to the hyperscaler opportunity

To win in the hyperscaler market, tech vendors must take an entirely new approach.

High Tech Practice October 2018



Authored by:
Hari Kannan
Christopher Thomas


27
29
31
33
35
37
39

How high-tech suppliers are responding to the hyperscaler opportunity

Companies pursuing revenue synergies can't take them for granted. Leaders need a clear grasp of where those synergies lie—and the persistence to capture them.

Over the last three years, estimates suggest, hyperscalers have spent $185 billion on data centers—about $75 billion in 2017 alone. Amazon, Microsoft, Google, Apple, and Facebook are responsible for almost 70 percent (some $50 billion) of these huge capital expenditures.[1] Hyperscalers already account for about a third of total data-center network traffic. According to Cisco, that share will jump to more than 55 percent in the next three years, when they will command more than 50 percent of all installed data-center servers.[2]

Spending on data centers rose by 20 percent in 2017. The entire supplier landscape and IT value chain have felt the effects of this explosive growth, which has been quite lucrative for suppliers of data-center components, such as storage, memory, servers, networking, power, cooling, and peripherals. But the suppliers' journey hasn't been easy—many have had serious difficulty addressing this new, concentrated, and episodic demand from a few huge customers. To meet their needs, successful suppliers have dramatically changed the way they approach R&D, manufacturing, and sales.

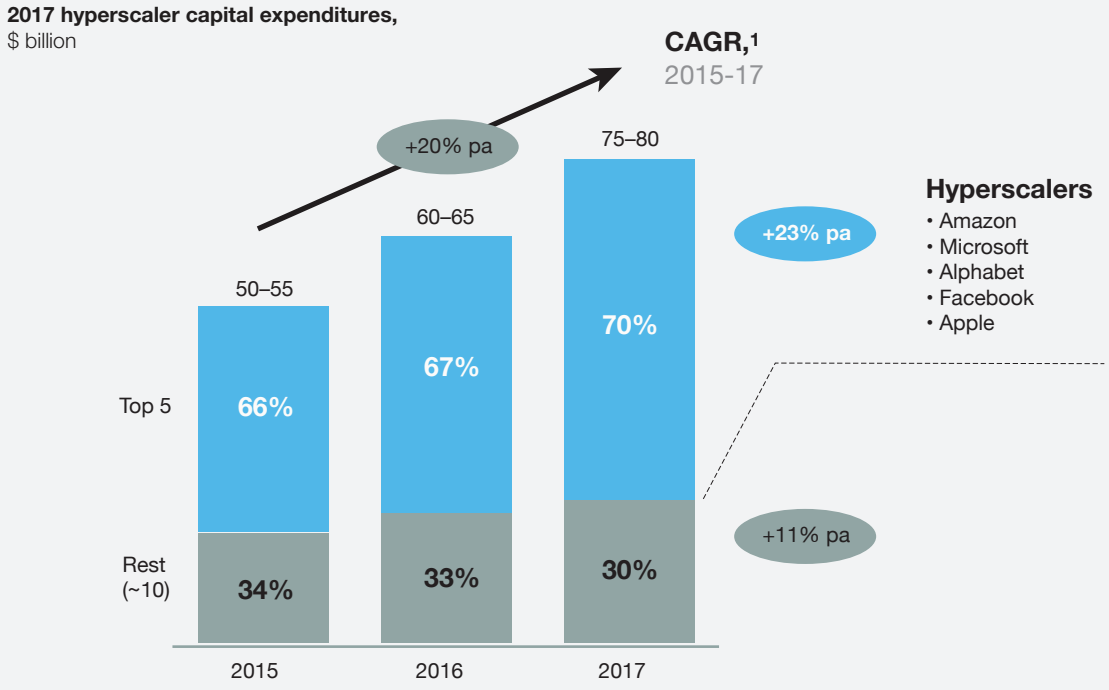
These suppliers have mastered a delicate balancing act: satisfying the needs of their existing partners—the OEMs—while building a new long-term customer base among the hyperscalers. In this article, we dig deep to understand this phenomenon, how it affects suppliers, and what companies will require to succeed in this market.

Understanding hyperscalers

Hyperscalers have recently enjoyed unprecedented growth in public-cloud and data-center revenues. The year-on-year quarterly revenues of Microsoft Azure, for example, soared by approximately 90 percent in the first quarter of fiscal year 2018,[3] while the year-on-year revenues of Amazon Web Services rose by more than 40 percent for 2018.[4] IDC believes that spending on public-cloud services will continue to increase by more than 20 percent per annum, to reach $277 billion by 2021,[5] and that this revenue growth will require the continued expansion of the hyperscalers' data-center infrastructure.

As we noted earlier, McKinsey estimates that the top five hyperscalers spent more than $50 billion on capital expenditures in 2017 (Exhibit 1).[6] These outlays are growing by more than 20 percent annually, and that phenomenal growth should continue during the next few years. Although data-center spending isn't public, our estimates suggest that a majority of these capital expenditures—especially the growth—goes toward data centers and the associated infrastructure. That includes servers, storage, networks, memory, accelerators such as ASICs (application-specific integrated circuits), and FPGAs (field-programmable gate arrays), facilities, power, and cooling.

Exhibit 1 **Hyperscalers are spending heavily on capital expenditures, mostly for data centers.**



[1]Compound annual growth rate

Source: Annual reports; press releases; Synergy Research Group

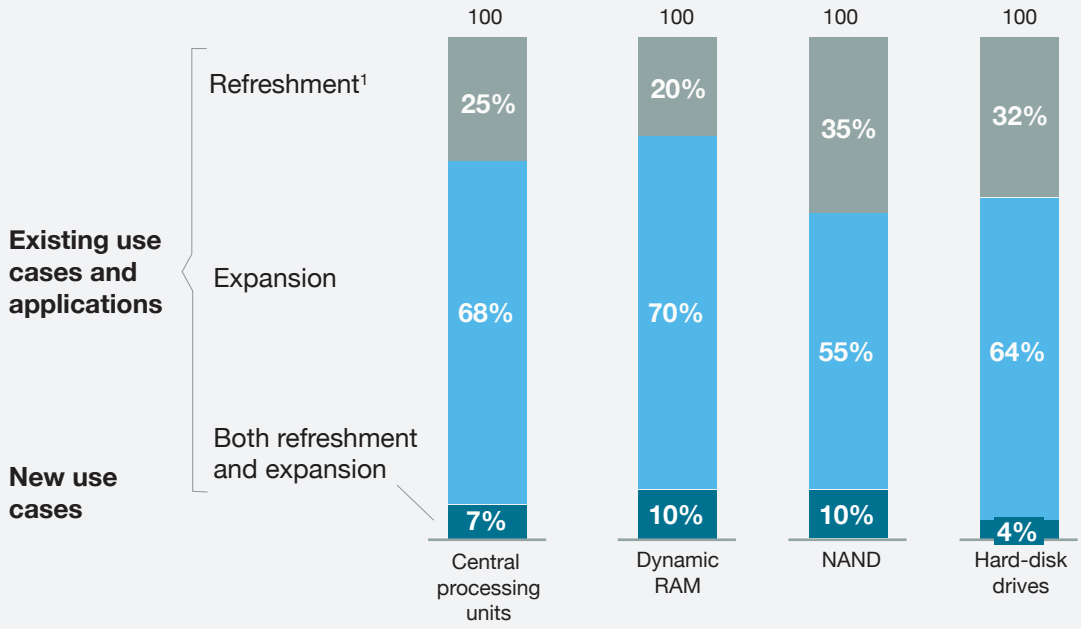
It's easy to see the data-center infrastructure in the public-cloud technology stack as part of a service for end customers. But core businesses such as search, social networking, e-commerce, and new and expanded platform capabilities—for instance, voice, vision, augmented reality and virtual reality, and machine learning—also require the continued expansion of data centers. That's especially true for investments in processing, such as graphics-processing units (GPUs) and central processing units (CPUs); silicon chips, such as DRAM and NAND; and storage drives, including magnetic memory and flash.

Planning and deploying these data centers is a complex and arduous task, as even the hyperscalers acknowledge.[7] At the highest level, their demand reflects the need to replace existing equipment, as well as expansion across geographies and capacities. We estimate that current use cases and applications account for a majority of this demand (Exhibit 2).

Less than one-fifth of total demand involves purchases to replace existing equipment, although the exact percentage varies among hyperscalers; geographies; components; and types of end applications, equipment, and infrastructure.

Exhibit 2 **Current use cases and applications account for a majority of the hyperscalers' demand.**

2017 hyperscaler demand breakdown for refreshment and expansion, existing and new use cases, market estimates



[1]Includes Amazon, Alphabet, Facebook, Microsoft, Alibaba, Tencent, and Baidu.

Expansion generates the remaining demand. Annual variations in replacement and expansion rates introduce further complexities, resulting from external factors—for example, step changes in capabilities, including new CPU refresh cycles; changes in capacity, such as 3D NAND; the level of penetration of existing use cases; and the number of new ones launched in a given year. Since these factors can change continually, demand planners must understand how they are likely to pan out.

Such variations also create considerable uncertainty about infrastructure choices: hyperscalers have to make many decisions (such as the choice of CPU or GPU) relatively close to deployment. Estimating demand by component type also requires an understanding of the growth rates of different workloads and use cases—and thus demand by server type (Exhibit 3). Since new and emerging use cases with volatile market-penetration rates drive significant capacity, probabilistic scenario-based approaches are needed to predict demand. Often, the result is a wide variation between long-term forecasts and near-term needs.

The next level of complexity comes from the interplay of the top-down and bottom-up drivers of demand. Although the overall revenue streams

Exhibit 3

The variety of servers makes planning the capacity hyperscalers will need more complex.

Key server archetypes and sample use cases

Archetypes	General purpose	High-performance computing	Storage optimized	Memory optimized	GPU driven
Description	• Most commonly used; usually the starting point in public cloud	• Maxed out on processing power	• Optimized for moving large amounts of data quickly	• Optimized for more memory to access as data at memory speeds	• Fitted with high-performance graphics processing units and supporting hardware
Prominent use cases	• Development and testing workloads	• Numerical simulation	• High-throughput content	• In memory analytics • In-memory databases • Large-scale indexing and caching	• Heavy rending, simulation • Machine leaning, artficial intelligence, and deep learning • Augmented reality and virtual reality
Typical configuration	• Midpower central processing units • ~3–4 gigabytes RAM per core • Mid–high storage mix of solid-state and hard-disk drives	• High-power central processing units • Maxed-out RAM • Application-specific integrated circuits for specialized workloads	• Midpower central processing units • High dynamic RAM • Mix of fast and slow storage, depending on workload	• High-power central processing units • Maxed-out memory • High-performance storage	• Highest-performance central and graphics-processing units maxed out on cores • Maxed-out dynamic RAM • High-performance solid-state drives
Intensity of power and cooling	• Medium	• High	• Low	• Medium	• Very high
Network-throughput requirements	• Medium	• Low	• High	• High	• High

are estimated top-down by the leaders of product lines, the supporting data-center growth must occur bottom-up, or locally. Even demand planning and deployment for expanding data centers or replacing their equipment take place on the local level. Demand from each data center must then be aggregated, so that the centralized procurement function can do the job of managing a varied set of global suppliers. Supply-chain processes should ensure that the right equipment and parts reach data centers around the world. But since mismatches between the bottom-up and top-down estimates are common, additional cycles of demand-forecast refinements may be required. Suppliers thus have still less time to react to orders.

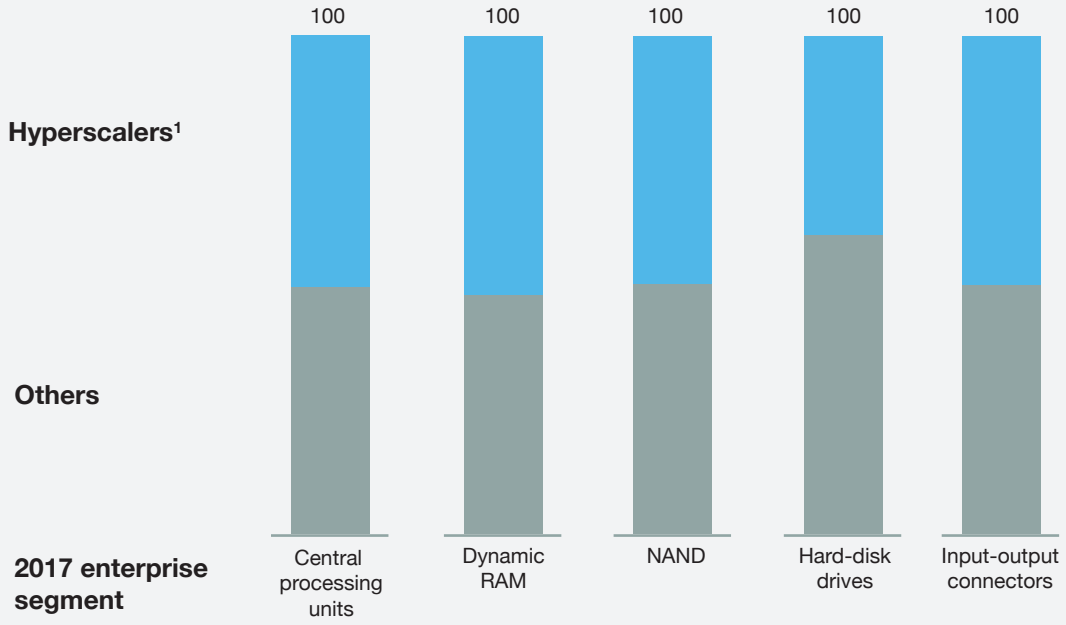
The final element of complexity lies in the innovation cycle. Hyperscalers continually develop new use cases for their customers and make existing services increasingly efficient. Often, they are also early adopters of new technologies and, in many cases, introduce their own customized versions of the components required to implement them. Hyperscalers can therefore meet an unknown proportion of their demand internally, further complicating planning, deployment, and operations in their data centers.

Understanding the changing customer landscape and its effect on suppliers

Suppliers were accustomed to working with OEM vendors, which in turn served IT shops. In fact, OEM vendors often handled end-customer requirements and demand forecasting. But hyperscalers have disintermediated the OEMs for much of their demand and engaged directly with commodity suppliers. Since hyperscalers are rapidly emerging as lucrative end customers expanding at a high and fast rate (Exhibit 4), serving them can be a source of tremendous growth and success. Yet to navigate this

Exhibit 4 **Hyperscalers, commanding a growing share of the market, are emerging as significant customers for many components.**

2017 share of hyperscalers in component markets, market estimates



[1]Includes Amazon, Alphabet, Facebook, Microsoft, Alibaba, Tencent, and Baidu.

new customer environment successfully, suppliers must adapt to the new market dynamics and invest in the right capabilities. In the quest to adapt, they face challenges in common.

Margin compression. Hyperscalers boast of their tremendously large deals and therefore command significant price discounts. Their business model relies on efficiency: maximizing the life and use of hardware, deploying components specifically designed for their own use cases, and procuring components at the lowest possible cost. These requirements pose a pricing challenge for suppliers, strongly compressing their margins and creating cost pressures on their internal supply chains and operations. Our research indicates that, on average, the server vendors' gross margins on hyperscaler deals are eight to ten points lower than their margins on other enterprise deals.

Inaccurate demand forecasting. Demand from the hyperscalers' data centers fluctuates year on year and quarter on quarter, given the seasonality of their revenue streams, the launch of new products, and changes in the customer base, among other things. Although hyperscalers can estimate their annual demand with fair accuracy, quarterly fluctuations are significant, so demand forecasting and related business planning are a challenge for hyperscalers and can make it hard for suppliers to meet and keep up with order deadlines.

Varying levels of factory utilization. Variations in quarterly demand also affect the capacity, utilization, and manufacturing plans of factories. Suppliers must rethink and realign their end-to-end operations, from procurement to manufacturing to inventory management to logistics, so that all of them suit the hyperscalers' needs. Component suppliers must also balance demand from the hyperscalers against demand from other parts of their businesses—for instance, enterprise OEMs and manufacturers of consumer devices. Such customers continue to experience relatively high volatility and frequent mismatches in their supply-and-demand patterns. The result is highly variable and unpredictable price behavior.

High demand for customization. Hyperscalers want components customized to their specific use cases, and they themselves are rapidly innovating to make new capabilities and use cases available to their own customers. This level of innovation places a strain on the suppliers, which struggle to keep up with the hyperscalers' technology road maps. What's more, suppliers can meet the required standards only by overhauling their engineering, quality-assurance, and testing capabilities for increased agility, shorter turnaround cycles, and larger test-bit quantities.

The risk of disintermediation. Suppliers also face a strong risk of disintermediation or even the possibility that segments of their businesses will simply evaporate. Since Hyperscalers lead the technology road map from the front, they are hardly shy about building their own products if they think that suppliers aren't meeting their vision, pace of improvement, or quality requirements. Hyperscalers can also attract the best talent, and that too exacerbates the pressure on suppliers to deliver the products required by advancing technology. If suppliers can't meet these challenges, they could become price takers, fighting for market share through low-cost leadership.

To win in the hyperscaler segment, suppliers must revamp their operations

As suppliers gear up for these changes, a few critical capabilities have emerged as top requirements for winning in the hyperscaler segment.

Technology leadership. Suppliers must lead the way, proactively shaping the next breakthrough product or technology to meet future needs and applications. They can do so, for example, by forming partnerships and joint ventures across the ecosystem to design

and deliver next-generation technology for key components.

Balanced investments in R&D and manufacturing. Another consideration for suppliers is the balancing act between investing in R&D, on the one hand, and in manufacturing and operations, on the other. Overall, suppliers are starting to move at a faster clip, which requires higher investments in research and development. Volatile demand calls for more agile manufacturing operations, and that in turn requires investments in tooling and manufacturing-line operations.

Design to value. Developing strong design-to-value (DTV) and design-for-manufacturing (DFM) functions is essential for creating the cost structure needed to serve hyperscalers—as we have seen, they might well build parts themselves if they believe that the suppliers' offerings are designed or manufactured inefficiently. In contrast to the enterprise segment, which deploys general-purpose products in a wide variety of use cases, the hyperscalers want specific features and products designed and manufactured to serve only particular use cases. A comprehensive DTV function, from design to manufacturing, can therefore help build the suppliers' long-term value propositions.

From vendor to strategic partner. Suppliers should consider building a multidimensional strategic relationship with each hyperscaler, although this strategy is significantly harder to execute. It requires, first, a modified go-to-market approach, with direct account coverage: the supplier's representatives should engage with multiple stakeholders across the hyperscalers' functions. Such a supplier must go beyond procurement coordinators by engaging with product and data-center teams to understand the customer's needs and requirements. Second, this strategy calls for joint investments, with shared risks and rewards, which must be negotiated between the hyperscalers and the suppliers. A rich relationship with hyperscalers, combined with an analytical model for demand forecasting, can help suppliers to understand the volatility and to address the challenges proactively.

Suppliers will have to overhaul their approach, above all by transforming themselves from commodity vendors into technology partners—the critical transition for suppliers hoping to win in the hyperscaler segment.

[1] Yevgeniy Sverdlik, "Hyper-scale data center spend was up 20 percent in 2017, analysts say," Data Center Knowledge, March 2, 2018, www.datacenterknowledge.com.

[2] "Cisco global cloud index: Forecast and methodology, 2016–2021," Cisco Systems, 2018, cisco.com.

[3] "Earnings release FY18 Q1: Intelligent cloud," Microsoft, microsoft.com.

[4] Quarterly results, Amazon, amazon.com; Jordan Novet, "Amazon cloud revenue jumps 45 percent in fourth quarter," February 1, 2018, cnbc.com.

[5] "Worldwide public cloud services spending forecast to reach $160 billion this year, according to IDC," IDC, January 18, 2018, idc.com.

[6] Yevgeniy Sverdlik, "Hyper-scale data center spend was up 20 percent in 2017, analysts say," Data Center Knowledge, March 2, 2018, www.datacenterknowledge.com.

[7] Yousef Khalidi, "The network is a living organism," Microsoft, April 13, 2017, azure.microsoft.com; Daniel Terdiman, "How Facebook engineers plan to make your experience faster and more efficient in 2018," *Fast Company*, December 18, 2017, fastcompany.com.

Hari Kannan is an associate partner in McKinsey's Silicon Valley office. **Christopher Thomas** is a partner in the Beijing office.

The authors would like to thank Aykut Atali, Varanjot Kaur, Yuriy Kuklyev, Chhavi Sharma, Yu Ueda, Bill Wiseman, and Andrew Yoo for their contributions to this article.

High Tech Practice
October 2018

www.mckinsey.com