

Drive north through Silicon Valley and it's almost impossible to miss the gargantuan hangars at NASA's Moffett Field—reminders of a time when airships were seen as the future of travel.

Today's massive data processing facilities—so-called “hyperscale” data centers that can span nearly 13 acres—are the modern equivalents of those airship hangars. Although they may not dominate the skyline in the same way, their global influence is already much more pervasive, whether they are enabling rapid Google file searches, streamlining the sharing of photos and videos, accelerating commodity futures trading, or handling any number of other data processing activities (see sidebar: “Hyperscale Data Center Defined”).

However, the data centers' bland exteriors give little hint of the design and development pace of the next generation of centers with even higher processing capacity. They reflect none of the day-to-day challenges that confront those responsible for operating the centers. And there is nothing to indicate how difficult it has become to source the hardware and supporting systems—not only in terms of the next builds and site expansions but in terms of reliably providing replacement parts and subassemblies.

These complexities are representative of what is happening across the computing equipment industry today. As a result, hyperscale data center operators are debating whether they can obtain better value propositions from the original design manufacturers (ODMs) that focus on

## Hyperscale Data Centers

What's the Value of a Server Brand?

Vinod Devan and Kevin Keegan

WHITE PAPER

high-scale, low-cost production of computers, or from original equipment manufacturers (OEMs) that design, install, and service robust, efficient, “fit for purpose” computers. The answer comes from understanding the critical decision drivers and evaluating the total costs relative to functional need over the useful life of the equipment.

Cost-effective sourcing of high-performance servers at scale has been facilitated in recent years by the rise of highly capable ODMs—contract manufacturers such as Foxconn, Wistron, Compal, and Quanta that build servers in high volumes to customers’ specifications, using their leverage with component and subassembly suppliers to hold down costs of goods sold.

In effect, ODMs offer “generic boxes” at attractive first costs, whereas OEMs—companies such as Dell and HP—furnish “branded” solutions. That is, OEM servers are typically made by ODMs but their services and value-add extend further along the value chain. OEMs deliver more of a “one-stop shop” approach to everything, from purchasing to replacement parts management and warranties to on-site service management and reverse logistics. Complicating matters is the fact that ODMs in related industries are climbing the value chain—witness mobile handset maker HTC, now investing in its own branded phones worldwide, or Acer, which successfully became an OEM after starting out in the ODM business. For enterprise servers, this level of value chain creep has not happened yet, but there is evidence of ODMs courting customers with value-added services.

The decision for business leaders in data processing environments is not a simple binary choice, and it’s plainly not a decision about server product quality. The sheer pace of growth in the hyperscale sector, and the pressures on initial purchase price in some server farm-based companies versus life cycle cost challenges in others makes it doubly difficult to make clear-eyed sourcing decisions.

This paper will examine why reflexive decisions driven largely by first cost can have negative repercussions for effective data center operation. It will point out why OEMs can no longer rely solely on the aura of a brand and how they will instead need to increase focus on the value of their services. The paper will also show how ODMs that seek direct impact on the hyperscale sector must now tailor their approaches to “steal” end users from the very OEMs they now count as customers.

### Sizing the Market for Hyperscale Data Centers

Data processing environments are becoming almost unimaginably large. In its first phase, the ground floor of Microsoft’s recently opened Chicago data center is built to hold up to 56 containers, each packed with as many as 2,500 servers.<sup>1</sup> The whole facility could have as many as 440,000 Windows®\* servers on the first floor

#### SIDEBAR

## Hyperscale Data Center Defined

There is no hard and fast definition of the hyperscale category—it is too new for that. The term, coined less than two years ago, refers to massive homogeneous “IT factories” comprising hundreds of thousands of identically or similarly specified servers that usually are powered by x86 processors. New deployments often involve more than 20,000 servers; some industry observers expect that the largest customers will soon begin ordering and deploying more than 100,000 servers at a time. Rough estimates say that 15 percent of all servers sold are now installed in hyperscale data centers.

alone—more than five times the number of servers found in today's larger data centers.<sup>2</sup>

The new supersized facilities are being driven primarily by growth in demand. Just one example: Since Facebook is expanding faster than planned—the social networking company now has more than 500 million users—it has recently announced that it will double the size of its first wholly owned data center being built in Prineville, Oregon, where it broke ground as recently as January 2010.<sup>3</sup>

Of course, the demand drivers are by no means limited to Facebook photos: Research firm IDC's ongoing Digital Universe study notes that the volume of global data—from e-mails to video clips to corporate databases to sweeping scientific analyses—will increase 44-fold from 800 billion gigabytes in 2009—a 62 percent step up from 2008—to 35 trillion gigabytes by 2020.<sup>4</sup>

Concurrently, more of the activities of corporations and institutions are reliant upon massive data processing environments—and the organizations' leaders are increasingly ready to outsource the management of their data centers or their entire data processing functions. Specialists such as Emerson, Equinix, and SunGard are typical of the IT providers that offer such services; part of their added value lies in the scale of their operations and their continuous investments to ensure that their computing platforms offer the right blends of cost, quality, reliability, and serviceability.

At the same time, the growing acceptance of cloud computing—both private and public clouds—is accelerating the shift toward larger and larger aggregations of servers. Gartner notes that more than a third of all IT workloads will be cloud-based by 2012.

The rapid rise in data center size means that, for many organizations, infrastructure decisions—traditionally not viewed as strategic—now

need to be reevaluated in light of a host of enterprise risk management criteria, from investment costs and large-scale system reliability to regional energy footprints. Specifically, operators are being forced to rethink fundamental decisions about sourcing the hardware backbone.

### The Many Challenges of Equipment Sourcing

It is important to state that there is really no one homogeneous hyperscale data center sector. Many organizations investing for the first time in large data processing environments—for example, new niche search engine sites supporting national salvage auto parts databases or online clearinghouses for used office equipment—will likely emphasize low hardware procurement costs until their business models are proven and operations are scaled up. By contrast, organizations with value propositions that require exceptional reliability and security—mandatory in sectors such as financial services and health care data records—will need to invest in premium hardware and software accordingly.

Experienced users such as Yahoo! and Facebook understand how to tweak costs because they know which systems and subsystems are needed and which are not. They are likely to demand more rigorous reliability testing and evidence of predictable failure rates, for instance. They will look for comprehensive parts provisioning capabilities—sometimes even on-premise facilities, not unlike the just-in-time parts warehouses next to automotive assembly plants that are stocked during suppliers' regular “milk run” deliveries. And, in some situations, data center operators will require reverse logistics capabilities so they can easily return surplus or defective components and systems.

Hardware sourcing decisions are also a function of the experience of operators' business leaders and their operational cohesiveness.

In some situations, the data centers' owners are young companies whose management teams have a keen eye for the total costs of ownership but who may not grasp all the factors driving up those costs. In other cases, the procurement teams and their colleagues in operations lack the close links necessary for optimal procurement and implementation. In many instances, there is a limited understanding of all the sourcing options available. And there are many situations where data center operators fear that providers of servers and systems will burden them with premium service fees or unnecessary extras—more memory, perhaps, or higher-rated power supplies than they need.

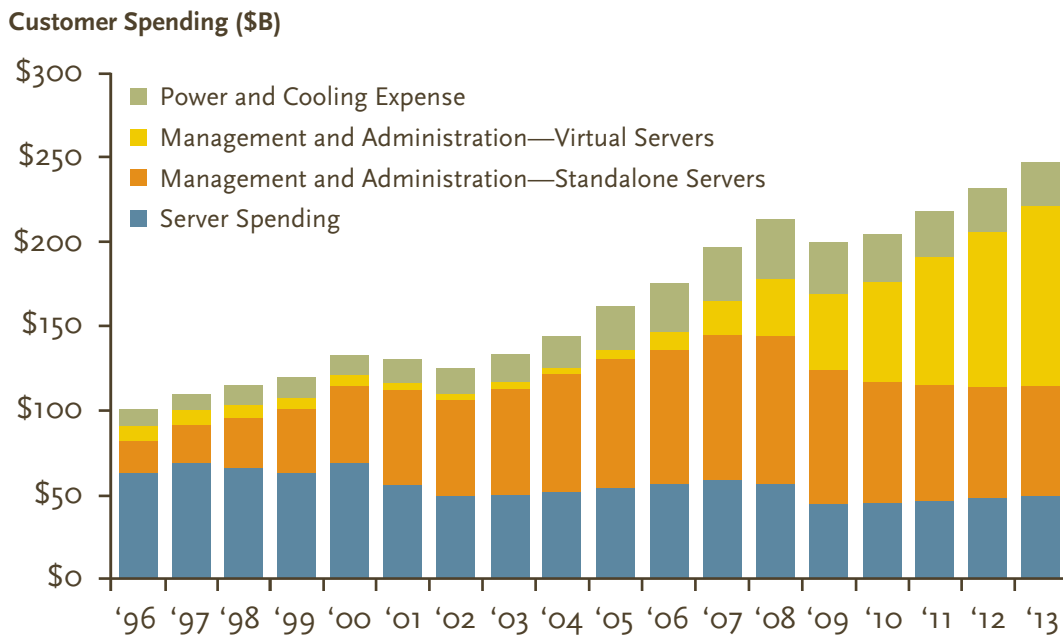
The good news is that most hyperscale center operators have a very clear idea of the basics. They are well aware that scale matters in

driving efficiency. They push leading efficiency concepts like power usage effectiveness (PUE)—that is, total facility power consumption divided by the power used by IT equipment. Google, for example, publishes regular bulletins of its PUE levels compared to U.S. Environmental Protection Agency goals.<sup>5</sup>

Typically, today's data center operators will specify designs that enable applications to be optimized and power utilization to be minimized. They are well aware that power and cooling expenses must be factored in and that management and administration—especially of virtualized servers—are critical total cost components (see Figure 1). Most operators also insist that their new server systems, once installed, require minimal cost to run so their resources can be applied to the organization's core busi-

Figure 1: Spending on Data Centers Goes Far Beyond the Server Costs Themselves

*WW Spending on Servers, Power and Cooling, and Management/Administration*



Source: IDC

ness activities. And their designs allow system capacity to be stretched at only a small incremental cost. Above all, they focus on uptime and reliability as the most critical design factors.

However, the incredibly rapid rise and relative immaturity of the hyperscale market means that many operators are still uncertain about which categories of equipment providers can meet all of their performance needs most cost-effectively over the long term.

### Five Critical Drivers of Sourcing Decisions

As an “IT factory,” the hyperscale data center is under every bit as much business scrutiny as an automobile assembly plant or a retail store. Its rates of revenue generation and its total cost efficiency and effectiveness are of constant concern to the operator’s chief executive officer (CEO), chief financial officer (CFO), and shareholders. Its energy efficiency is a matter for the local community and for facilities management as well as for the CEO and CFO. The data center’s “greenness”—not limited to its energy consumption—is increasingly of interest to shareholders, not to mention employees and environmental watchdogs. The predictability of its performance is tracked by top management. And its exposure to risk is under the microscope of the board of directors as well as that of its investors.

There are five areas of focus that merit particular attention:

1. **Rightsizing reliability:** System reliability is the lifeblood of the value proposition for any massive data processing environment. The occasional downtime incurred by high-profile operators such as Twitter and Facebook creates uncomfortable headlines; similar incidents are unthinkable in the world of financial services transaction processing.

However, this does not imply that the goal is reliability at any cost. Leading providers of hyperscale data center equipment are adept at rightsizing the reliability of their installations. They understand that even the customers who are most concerned about uptime should not have to bear heavy non-recurring engineering (NRE) costs to achieve high overall levels of reliability (see Figure 2).

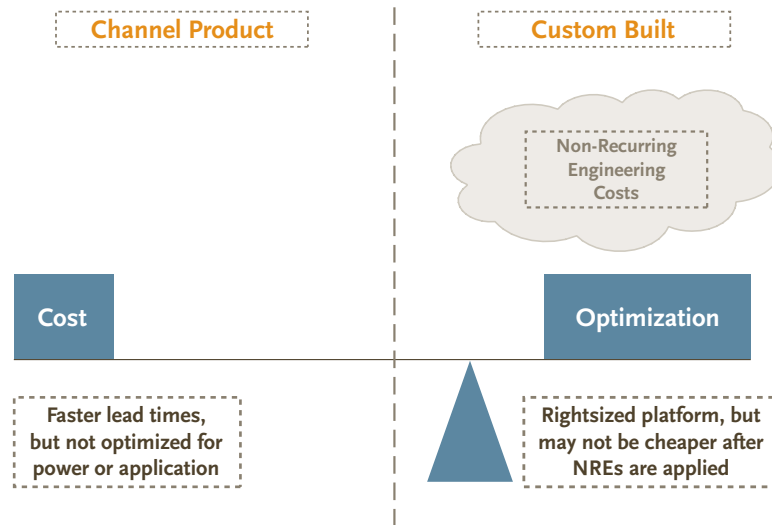
As a result, OEM providers are attuned to the distinction between *server* reliability and *system* reliability. With most large data centers running at Tier 2 reliability levels—99.741 percent availability, as defined by the Uptime Institute—the emphasis is on redundant components and subsystems rather than highly specified server boxes that might incur significant NRE costs.<sup>6</sup>

For example, OEM providers will typically use redundant hot-plug fans; when a fan fails, the server can be removed from the rack and the fan replaced while the unit is still running. They will also ensure that each rack has the optimal electrical power budget—not so light that the rack cannot support peak loads, and not so heavy that it consumes more power than is typically necessary.

ODMs that aspire to capture more of the hyperscale data center market will benefit by shifting from a mindset in which higher NRE costs are tolerable in the quest for higher reliability to one that emphasizes the *rightsizing* of reliability.

2. **Sound planning for future capacity:** Facebook’s doubling of the capacity of a data center before construction has been completed amply demonstrates the need for a long view of data center needs. In its latest Digital Universe study, IDC illustrates the need to

Figure 2: It Is Crucial to Balance the Costs and Features of Hyperscale Data Centers



Source: IDC

plan more than one step out. If a major new storage management center takes two years to implement, the Digital Universe will be twice as big when the center goes live as it was when the project was being planned.<sup>7</sup>

It is essential to design, configure, and install any data center with future business needs in mind. Yet some operators have yet to master the system life cycle management and road mapping necessary to ensure that the data center can be updated smoothly and progressively as technology evolves—for example, as RAM memory shifts from DDR2 to DDR3 and hard drives transition to terabyte capacities.

Then there is the relatively more ambiguous evolution of market demands and preferences that disrupt even the best-laid plans. What's needed are regular “technology road map” reviews with systems providers that are constantly holding such reviews with

their suppliers and that can easily synthesize what they learn into systematic, cost-efficient, and visible upgrade paths. Such reviews are standard for most OEMs—and typically less complete in the case of ODMs.

- 3. Guaranteed supply of components and subsystems:** In some cases, hyperscale data centers must go live at a set time on a given day if end user needs are to be met—and penalties are to be avoided. So the operators have to be confident that the equipment they have ordered can be delivered, installed, tested, and brought on stream against tight deadlines. This cannot happen if their systems providers cannot guarantee constancy of supply of the components needed to build the systems—or to replace them in future.

Not long ago, many systems providers found themselves scrambling for supplies of RJ45 connectors (a component that costs less than

\$1) following a catastrophic fire at a major supplier's plant. By contrast, providers that had prequalified multiple suppliers were easily able to ensure sourcing of the connectors.

Best-in-class OEMs and ODMs have well-honed supplier management practices that help them avoid such difficulties. But farsighted data center operators will favor the providers that are willing to invest in parts depots located at or near the data center itself. (For example, Dell has built what it calls the Onsite Parts Locker—a Web-based parts and service system that provides immediate, round-the-clock visibility of and access to repair stock.) Savvy data center operators are also looking for integrators that have well-established channels and proven processes for handling spare parts and parts distribution, reverse logistics for swap-outs, and critical situation parts management. And they give higher grades to integrators that invest continually in the resources and skills needed for effective supplier management and quality controls.

- 4. Unambiguous accountability:** Given the complexity of today's giant data processing facilities, operators must be able to tap a network of services in order to confidently maintain and upgrade their equipment. Few are equipped to bear the responsibility for servicing their systems singlehandedly. In the absence of contract documentation that delineates the ODMs' or OEMs' after-sale responsibilities, there may be few ways of resolving repair and maintenance issues smoothly and at minimal cost.

This area is not typically a strong suit for ODMs. One consequence: More and more technical and engineering staff at suppliers of components and subsystems are receiving calls from operations teams at hyperscale data centers. In effect, they are seen as the

experts who can be relied upon to help when there is an unexpected shortfall in performance or system reliability.

That was the case involving a major Internet-based business in Asia which had specified unique server form factors for its hyperscale data center. But the company ran into difficulties when a flurry of system incompatibilities on site could not be resolved easily. The first phone call for help went not to the server ODM, but rather to the supplier of the servers' microprocessors; that supplier flew in a group of its engineers to troubleshoot and debug the system.

OEMs assume accountability for such on-site challenges—and plan to avoid such issues in their system designs in the first place. If ODMs are to compete directly with OEMs for hyperscale market share, they have to demonstrate comparable accountability. They will have to demonstrate true “ownership mentality”—and invest in the suite of service offerings and customer support systems, skills, and infrastructure to ensure that inevitable technical challenges are dealt with promptly and professionally.

- 5. Managed deployment on a large scale:** In one actual scenario, a hyperscale user had to be able to install, test, and start up an entire data center's worth of equipment at one time to meet its customer's need to stream live video feeds from the Olympics—a fixed deadline, to be sure. This is not simply a question of delivering a few FedEx boxes several mornings in a row. In the scenario described, it involved 40,000 servers—1,000 racks—showing up on site, being installed, powered up, validated, and tested, all inside a six-week window.

Of course, such mass mobilization calls for rigorous supply chain discipline to fill large

orders for servers, but that's only part of a successful hyperscale implementation. Top-notch project management skills are essential as well. OEMs will use integration facilities to which the server boxes are shipped after assembly; there, they will be crated up and packed onto fleets of large trucks. Skilled resources must be ready at the site to receive the equipment, unpack it safely, validate that nothing has been damaged, and ensure that it is still working. Meanwhile, other professionals have been handling long-lead time tasks such as obtaining the permits necessary for delivering electrical power to the facilities.

Large-scale managed deployments also call for exceptional customer-facing skills, from setting expectations for the hyperscale operator to reporting progress at key milestones.

As things stand now, OEMs provide the more compelling total value proposition for more sophisticated and more demanding hyperscale data center operators. But OEMs cannot take their edge for granted; the ODMs are demonstrating their ability to move up the value chain.

For their part, ODMs that want to capture more share of profit in the sector will have to raise their game in at least the five areas of focus described. For now, their raw cost advantage is ebbing: IDC notes that, partway through 2010, business priorities have quickly returned to prerecession status.<sup>8</sup> There is a renewed focus on “doing the right thing,” not the least expensive thing, for customers and shareholders—and for competitive advantage.

#### Sources:

<sup>1</sup> “Photos: Inside a Microsoft Data Center,” CNET.com, Nov 2, 2009, [http://news.cnet.com/2300-10805\\_3-10001679.html](http://news.cnet.com/2300-10805_3-10001679.html).

<sup>2</sup> Lai, Eric, “Walking the talk: Microsoft builds first major container-based data center,” Computerworld, April 7, 2008, [http://www.computerworld.com/s/article/9075519/Walking\\_the\\_talk\\_Microsoft\\_builds\\_first\\_major\\_container\\_based\\_data\\_center](http://www.computerworld.com/s/article/9075519/Walking_the_talk_Microsoft_builds_first_major_container_based_data_center).

<sup>3</sup> Blair, Scott, “Facebook Adds Expansion to Massive Data Center Under Construction in Prineville,” Architectural Record, August 4, 2010, [http://archrecord.construction.com/news/daily/archives/2010/08/100804facebook\\_construction\\_prineville.asp](http://archrecord.construction.com/news/daily/archives/2010/08/100804facebook_construction_prineville.asp).

<sup>4</sup> “The Digital Universe Decade: Are You Ready?” <http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm>.

<sup>5</sup> “Data Center Efficiency Measurements,” Google report 2010, <http://www.google.com/corporate/green/datacenters/measuring.html>.

<sup>6</sup> “Tier Standards,” Uptime Institute, <http://www.uptimeinstitute.org/>.

<sup>7</sup> “The Digital Universe Decade: Are You Ready?” Op. cit.

<sup>8</sup> “Three Data Centers—One Vision?” Directions 2010, IDC, March 2010.

For more information, please contact:

**VINOD DEVAN**, PRTM Principal  
vdevan@prtm.com, +1 847.430.9000

**KEVIN KEEGAN**, PRTM Director  
kkeegan@prtm.com, +1 781.434.1200

**Beijing**

+86 10.6808.0296

**Bengaluru**

+91 80.4010.0900

**Boston**

+1 781.434.1200

**Chicago**

+1 847.430.9000

**Dallas**

+1 972.980.8200

**Detroit**

+1 248.327.2500

**Dubai**

+971 4.434.5977

**Frankfurt**

+49 69.219.940

**Glasgow**

+44 141.616.2616

**London**

44 207.010.8585

**Munich**

+49 89.516.175.5

**New York**

+1 212.915.2600

**Orange County**

+1 949.752.0100

**Oxford**

+44 1235.555500

**Paris**

+33 0.1.56.68.30.30

**Shanghai**

+86 21.3860.7888

**Silicon Valley**

+1 650.967.2900

**Tokyo**

+81 3.5326.9090

**Washington, DC**

+1 202.756.1700

[www.prtm.com](http://www.prtm.com)