GIGAOM RESEARCH

Web-scale IT for the enterprise: the journey to modernization

Barb Goldworm

October 31, 2014

This report is underwritten by Nutanix.

Table of Contents

Executive summary

Web-scale principles, architecture, and technologies have enabled some of the largest and most successful web companies and cloud service providers such as Google, Amazon, Facebook, and Microsoft to build highly successful hyperscale datacenters. Enterprises are taking notice. They face significant challenges, as well as shortcomings in their traditional IT infrastructures, as they compete with public cloud providers and try to respond to user demands now and in the future. To compete effectively, enterprises need to learn lessons from the hyperscale datacenters and public cloud providers, and rethink the way they operate.

This report will help CIOs, IT decision makers, business strategists and decision makers, cloud platform providers, service provider executives, data managers/developers, and enterprise software and technology vendors understand the fundamentals of web-scale IT and learn how web-scale principles can be leveraged inside enterprise IT.

Key findings include:

- Web-scale principles and design can help organizations build private cloud operations that achieve scalability, elasticity, resiliency, availability, and agility.
- Hyperconvergence helps break IT silos by merging compute and storage, putting the data closer to where it's needed, minimizing data movement, and providing the ability to manage compute and storage holistically.
- Enterprises that would like to reap the benefits of web-scale architecture on a smaller scale, without customized efforts and dollars, should consider turnkey enterprise solutions that offer many of the same benefits without the same time, staff, or budgets.

Introduction

The past decade has seen significant change in IT – both in emerging technologies in infrastructure and applications and in ways of doing business. New technologies such as server and desktop virtualization, flash storage, and network virtualization, along with the rise of public cloud options like Amazon Web Services have caused IT organizations to re-evaluate, re-architect, and re-shape their infrastructures and organizations. As part of the public cloud explosion, companies like Amazon, Google, and Microsoft have scaled to a whole new level, causing them to re-architect their own infrastructures and organizations to create a new way of building and operating these hyperscale datacenters in an approach now known as web-scale IT. Leveraging the approach, architecture, tools, and technologies developed by these pioneers can bring huge benefits to all enterprises, regardless of their size.

Evolution of enterprise computing: physical to virtual to cloud

From the monolithic mainframe datacenter era of the 1970s, to the PC era of the 1980s, to the departmental server era of the 1990s, enterprises have struggled to meet user demands while trying to manage IT operations efficiently, at the lowest cost possible. The early 2000s saw the beginning of the virtualization era.

Physical to virtual

The emergence of x86 virtualization a few years ago brought huge change along with huge benefits to enterprise computing, allowing organizations to consolidate their largely underutilized servers, dramatically reducing the number of servers and thus the space, power, and cooling requirements in their datacenters. Virtualization – the process of abstracting away from the underlying layer (or layers) of hardware (or software) – allowed workloads to be moved from one server (or desktop) to another, regardless of the hardware vendor, configuration, etc. underneath. The return on investment (ROI) of consolidation was an obvious choice for most organizations. Live migration made real-time workload movement possible with no disruption to the running workloads, adding availability and disaster recovery (DR) capabilities to the benefits of consolidation. Today, server virtualization is pervasive across most IT organizations, with the remaining organizations planning the migration from physical to virtual.

Consumerization of IT and the emergence of public cloud

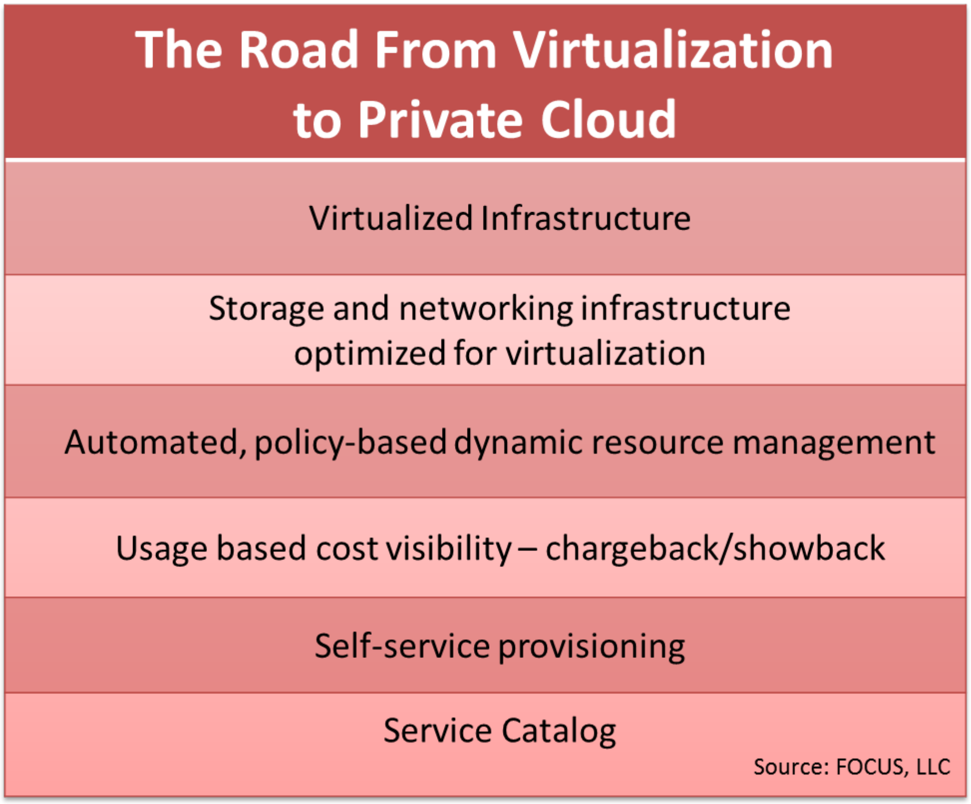
The advent of mobile phones and tablets along with the growing use of internet services and applications has created a new generation of users known as "born digital." As this generation has made its way into the corporate world, its changed expectations have created the "consumerization of IT." Users are now accustomed to instant-on, always-available, instant response-time capabilities, with broad access to real-time information, using their own device of choice, with their own user preferences. For example, technology has enabled consumers to get real-time traffic information with dynamic route suggestions and the prices of goods across hundreds of stores, all from their mobile phones. In enterprises, the newest, coolest devices often are in the hands of the execs, not just the techs. As a result of consumerization, overall user expectations are extremely high.

During this same time, there was an explosion of public clouds options such as Amazon Web Services, offering anyone with a credit card the ability to purchase pay-as-you go computing options, including primarily Software as a Service (SaaS) and Infrastructure as a Service (IaaS), and to a lesser extent Platform as a Service (PaaS). When IT departments couldn't meet corporate managers' expectations, the managers frequently circumvented IT, running a credit card swipe on Salesforce.com or Amazon, and doing it themselves, often called shadow IT or rogue IT. This movement put additional pressure on IT organizations, placing them in direct competition with the cloud service providers for their own internal customers.

The road to cloud

As a result of this competition, IT organizations have had to look at ways to operate more like a service provider to their own internal users. This has included improving their service delivery and end-to-end availability, becoming more flexible and agile, managing their datacenter resources better and more efficiently thus reducing their costs, and offering features like self-service and service-catalogs. In other words, they began delivering their own private cloud, or at least became more cloud-like in their on-premises offerings. Since most organizations have already made significant progress in virtualization, organizations must move beyond basic virtualization and consolidation, and take the necessary steps from there, starting with their own hardware and software infrastructures to deliver private cloud capabilities. In addition to optimizing the infrastructure, new types of management and orchestration are needed, for policy-based automation, cost visibility, self-service capability, and service catalogs.

The Road from virtualization to private cloud



(Source: FOCUS, LLC)

As organizations move down this road, they can learn a number of lessons from observing the public cloud providers:

- IT must learn to act like a service provider, treating its own users as service consumers. Services also must be easy to consume or users will go to the public cloud providers.
- To be competitive, automation is essential, and to automate, standardization is needed. It also will be necessary to break the IT silos and integrate across them – servers, storage, networking, security, desktops, applications, and mobile.
- Public cloud services are appealing because of the convenience and ease of use they offer. But in the long run, public clouds may not offer the cost savings that many users are expecting.
- IT needs to become a strategic advisor for all computing including traditional private, public, and hybrid cloud, and help determine what best fits where and how to integrate them together.

Web-scale IT

Just as enterprise IT organizations learned important lessons from the early public cloud providers, they can leverage the experience and money spent by hyperscale pioneers such as Google and Facebook and apply many of the web-scale principles to enterprise IT. These principles, processes, architectures, and potentially turnkey solutions can bring many of the benefits of web-scale IT to enterprises of varying sizes.

Benefits of web-scale IT

Scale on-demand. Web-scale, which is designed for elasticity and predictability in adding or subtracting resources as needed, is highly scalable and extremely agile.

Always-on operation, resiliency to failures. Downtime is unacceptable in hyperscale environments. A key part of web-scale design is high resilience so that it can recover from failures. Component failure is a given at large scale, so web-scale architecture and processes are designed with automated recovery as opposed to traditional IT, which is designed to prevent failure.

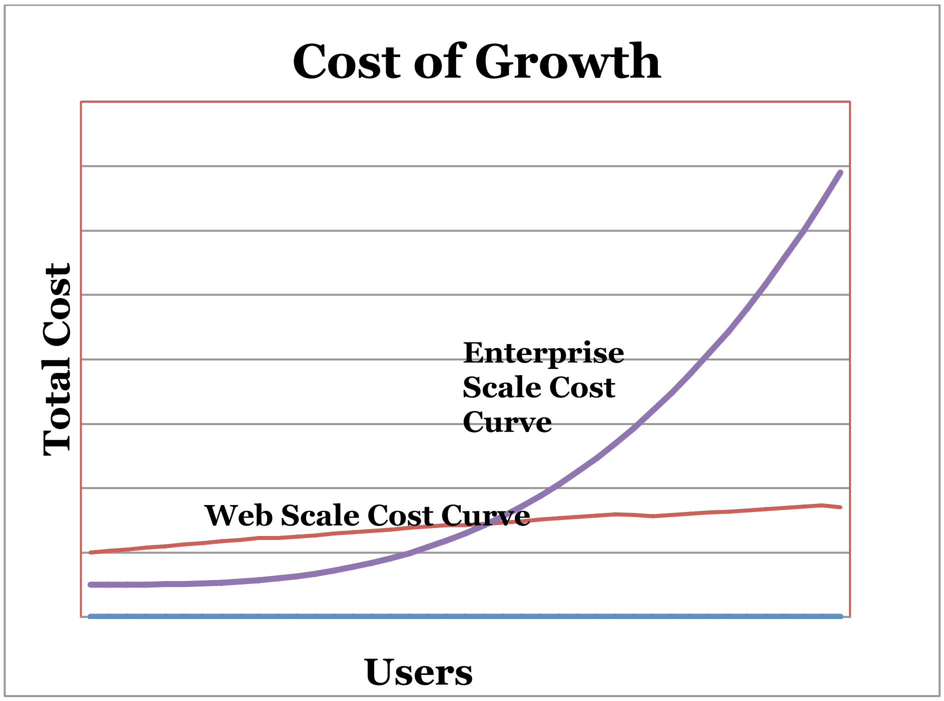
Cost-efficient performance and availability through standardization and automation. Always-on, highly available operation requires infrastructure standardization so that issues are resolved automatically. Web-scale design, with its standardized, scalable, highly available, and automated environment, is optimized to deliver performance at the least possible cost (i.e., the least amount of hardware for the most work).

Web-scale was developed for hyperscale. Spending time and money up-front was necessary to optimize every aspect of the infrastructure – particularly in components that occur in volume, such as efficiencies in servers, storage, networking, energy, and even people. Some of the largest hyperscale datacenters have been built with stripped down, commodity, off-the shelf hardware, scaling to millions of servers and providing elastic computing, at the least cost and least amount of space, power, and cooling possible. Open-source software is commonly used in web-scale design to reduce software-licensing costs.

Automated provisioning and deployment of new versions of software is also critical for rolling upgrades and key to reducing costs. It changes the span of control from an average of 1:300-700 servers (in a virtual infrastructure) to the 1:10,000 range seen at hyperscale datacenters (or at 1:24,000 at Facebook), resulting in dramatic reductions in people costs.

Applying web-scale architecture, principles, and technologies can minimize waste and reduce cost or spread costs across many commodity-priced servers, allowing scaling to occur exponentially, with costs that grow linearly or sub-linearly. As demand increases, commodity servers are added, and with variable costs being linear, fixed costs are diluted over a growing number of users (see following figure).

Exponential user growth at sub-linear cost growth



(Source: FOCUS, LLC)

Agility, flexibility. Automated provisioning is a key to agility: new users are provisioned rapidly and changes can be quick and across large groups of users. Web-scale architecture integrates and automates storage and networking provisioning in addition to compute, allowing full-service provisioning in seconds.

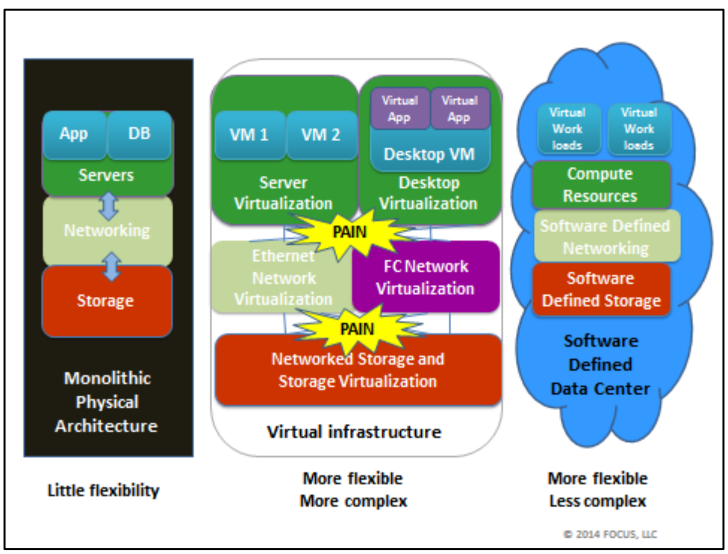
Automating software updates is also important for agility. Web-scale offers an environment that is highly adaptable, with changes made as quickly and often as needed (e.g., Google self-updating nodes). Web-scale is built on loosely coupled components that can be updated independently, facilitating changes, so the infrastructure is highly responsive to changing user/business needs. Web-scale helps organizations move towards DevOps principles, offering continuous development/upgrades, and rapid deployment, with less risk, and at lower cost. This is especially important for leading edge, innovative applications, or services needing high rates of change.

Principles and characteristics of web-scale IT

Web-scale principles offer a new approach to designing and operating datacenters that are highly scalable, offering the benefits already discussed. Building these principles into the infrastructure/environment allows developers to build applications and services that will scale, without the developers having to write the low-level software to make it happen. These web-scale principles and characteristics are baked into the underlying infrastructure.

Software-defined. Since software is more flexible than hardware, web-scale means new features should be added solely in software and run on commodity hardware, with no special-purpose hardware for any task. A hardware abstraction layer can be used to abstract the software development away from the server hardware (web-scale design can be either for a specific hypervisor or be hypervisor-agnostic). Abstraction can also leverage a shared-nothing approach to storage, rather than using more expensive, complex networked storage appliances and storage area networks as in enterprise IT. Software-defined storage (SDS) as well as software-defined networking (SDN) can help most IT shops move beyond basic virtualization towards cloud. A full software-defined data center (SDDC), running on commodity hardware increases flexibility and reduces complexity (see following figure).

Goal: Increased flexibility with reduced complexity



(Source: FOCUS, LLC)

Open-source software and the open-source community play a role in a software-defined enterprise by lowering software costs while accelerating innovation and rapid development.

Commodity hardware. Using commodity hardware reduces cost and vendor lock-in, so changing hardware vendors when necessary is easy. With all the intelligence in the software, leveraging the least expensive, commodity x86 servers that will do the job becomes possible. These commodity servers can make use of x86 hardware, flash, and HDD storage technology as appropriate for the needs of the environment. Many commodity server platforms have been purpose-built for virtualization, enabling a software-defined architecture layered onto a hardware abstraction layer. These scale-out servers should be standards-based, standardized platforms, allowing for automation through APIs.

Web-scale operation also changes the perspective towards hardware in another way – often referred to as “cattle versus pets.”

- Traditional IT treats servers as pets; when one gets “sick,” anything and everything will be done to make it better.
- Web-scale design treats servers as cattle; when one gets “sick,” it is scrapped, and another one takes its place.

Many IT organizations will use a hybrid approach: commodity hardware will become the standard, but with the intent to failover quickly and repair if possible.

Hyperconvergence: compute storage integration. Hyperconvergence, which is optimized for scale-out, goes a step beyond converged infrastructure and blurs the lines between compute, storage, and storage services (e.g., deduplication, compression, replication), creating an infrastructure-in-a-box solution. This generally involves integrated compute (processors and memory) and storage (HDD and flash) on commodity x86 servers running a hardware abstraction layer or hypervisor, networked together and managed as one entity. Hyperconverged nodes use a modular design, scalable through the addition of more hyperconverged nodes.

VMware’s 2014 announcement of EVO:RAIL allows hardware vendors to build hyperconverged appliances running VMware Virtual SAN (VSAN) Software. Nutanix was first to market in hyperconvergence, building a clean slate architecture from scratch on off-the-shelf hardware. The company started with VMware, then added KVM and Hyper-V support, as well as deduplication, global management, and a recent OEM deal with Dell. SimpliVity came largely from a storage deduplication background and targets mid-size organizations, designing around deduplication through the lifecycle of the data, and is now also available on Cisco UCS. Scale Computing (based on KVM) targets smaller organizations.

Hyperconvergence is attractive as a next-generation solution for cloud-like, elastic, agile (i.e., web-scale) requirements (particularly during hardware refresh cycles or new growth) and especially for mid-size and large IT organizations with limited IT staff and budget. It offers a flexible, efficient, and agile platform for organizations that have not implemented a SAN and for shops building out a VDI solution. At the same time, it eliminates the complexity, especially storage complexity, that has slowed VDI acceptance.

Distributed, resilient, no single point-of-failure. A distributed-everything system design, with a parallel software architecture, is critical for achieving web-scale resiliency and availability. Distributed everything includes distributed server/storage nodes, data, metadata, intelligence, administration, management, and operations control.

Distributed storage technologies such as distributed databases, file systems, and services (e.g., NoSQL, Cassandra, Hadoop, etc.) are also key for web-scale's handling massive amounts of data across a fully distributed, scalable, and resilient environment.

Elastic growth. Web-scale supports elastic growth, including number of users and transactions, amount of data, etc., and is designed for predictable scale through scale-out design. As part of cloud-like behavior, web-scale design must be able to add and subtract resources on the fly.

Automated operations and management. Automation is critical for achieving web-scale scalability, agility, responsiveness, and low-cost aspects. Advanced automation of operations and management requires use of intelligent analytics for optimized automation that must be available through a programmatic interface, API-driven, that can be used for asynchronous operations. This interface can be used by everything from scripts to standard management software to advanced analytics and correlation engines.

Automated recovery/ availability. Because the web-scale approach expects and tolerates failures without losing overall availability, automated recovery is critical. Self-healing, automated failover, and remediation are required at every level and 24x7 availability requires versioning and automated deployment across all nodes. Rolling upgrades are important for allowing multiple versions of software while maintaining consistency of data. Automated deployment must be fast, often, and reliable.

DevOps. The principle breaks human silos and eliminates disconnect/tension/lag between developers and operations to streamline putting new software out to users. Organizations should consider DevOps concepts with web-scale design for delivering an always available, agile infrastructure. Transitioning to DevOps can begin with small steps that can be integrated with the use of web-scale principles.

Bringing web-scale to the enterprise

Most enterprise organizations are either completing virtualization efforts or are working towards private cloud. As they move to private cloud, they must evaluate legacy work, which still runs on a physical infrastructure and work that has been virtualized.

- Some organizations have achieved success using enterprise virtualization with cloud orchestration tools and techniques.
- Others have had success hiring staff to translate web-scale design and tools in mega data centers for their own enterprises, using tools like Cassandra, MongoDB, MapReduce, etc. This approach requires significant investment in people skills and changes to both the infrastructure and the organization. The transition can be long and difficult, and the requisite skills can be hard to hire.
- Others have leveraged web-scale turnkey solutions from vendors such as Nutanix that allow enterprises to take advantage of web-scale principles and benefits, without requiring the same level of people resources or overall change. Using turnkey solutions that embed various web-scale principles and big-data capabilities, IT doesn't need to work directly with the tools.

Considerations and best practices

Successfully moving to a private cloud requires both strategic and tactical thinking. Learning from early adopters can make the road less bumpy.

- IT must think and act like a service provider. A well-run private cloud is the best option for certain applications and the public cloud is best for others. IT must be an advisor and a broker for all cloud services, including on-premise private clouds, public clouds, and the resulting hybrid cloud.
- Evaluate the required infrastructure for size and to overcome the bottlenecks and shortcoming of existing infrastructure. Then, standardize into a small set of building blocks. Measure, optimize, and automate anything done at scale.
- Web-scale principles and design require an assumption that hardware and software components will fail. Design for it, including automation, rapid provisioning, and orchestration (moving workloads to the working resources).
- Build redundancy at all levels with no single point of failure. Include servers and their components, power, network links into the data center.
- Consider open-source components, particularly in areas of changing needs and where progress is being made (e.g., SDN).
- Consider DevOps principles as part of the shift to more agile development. IT as a strategic competitive advantage requires delivering updates fast and frequently, with no downtime.
- Consider solutions with web-scale principles embedded in them versus using native tools with exposed APIs that require significant development.

Key takeaways

- Cloud computing (private, public, and hybrid) is an important tool for gaining a competitive advantage. Strategic planning for a transition from virtualization to private cloud is a multi-step process. Infrastructure optimization, standardization, and automation are critical success factors.
- Building on virtualization, breaking IT silos – technology, management, and organizational – is necessary to achieve the scalability and agility needed for cloud-like operation. Hyperconvergence helps by merging compute and storage, putting the data closer to where it's needed, minimizing data movement, and providing the ability to manage compute and storage holistically.
- Web-scale principles and design can help organizations build private cloud operations that benefit from the lessons of hyperscale public cloud providers, and achieve the full benefits of cloud – scalability, elasticity, resiliency, availability, and agility.
- The long-term results can be significant for enterprises with the time, staff, and money to design their own web-scale architecture from scratch. Turnkey enterprise solutions have emerged that can provide many of the benefits for enterprises that would like to reap the benefits on a smaller scale, without the customized efforts and dollars – and without all the custom work. These turnkey solutions embed the concepts of web-scale in the solution, allowing many of the same benefits without the same time, staff, or budgets.

About Barb Goldworm

Barb Goldworm is founder, president, and chief analyst of FOCUS, LLC, and is a well-recognized industry expert focused on emerging and game-changing technologies including cloud computing, virtualization, software-defined data centers, systems management and storage, with over thirty years in software architecture, development, marketing, sales, and industry analysis. She has spent the last fifteen years keynoting, presenting, teaching, and writing for numerous industry events, webinars, publications and other industry channels. Her experience includes IBM, StorageTek, Novell, and numerous startups, in engineering, marketing, and executive management roles, as well as chairing and keynoting industry events such as Interop (Virtualization), Cloud Connect, Data Decisions, Storage Decisions, Server Blade Summit, Storage Networking World, and others. In addition to publishing hundreds of research reports, white papers, and regular trade magazine columns, Goldworm is the author of the Wiley book, "Blade Servers and Virtualization." In addition to authoring Gigaom Research reports, she routinely acts as moderator for Gigaom Research webinars.

About Gigaom Research

Gigaom Research gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, Gigaom Research addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: research.gigaom.com.