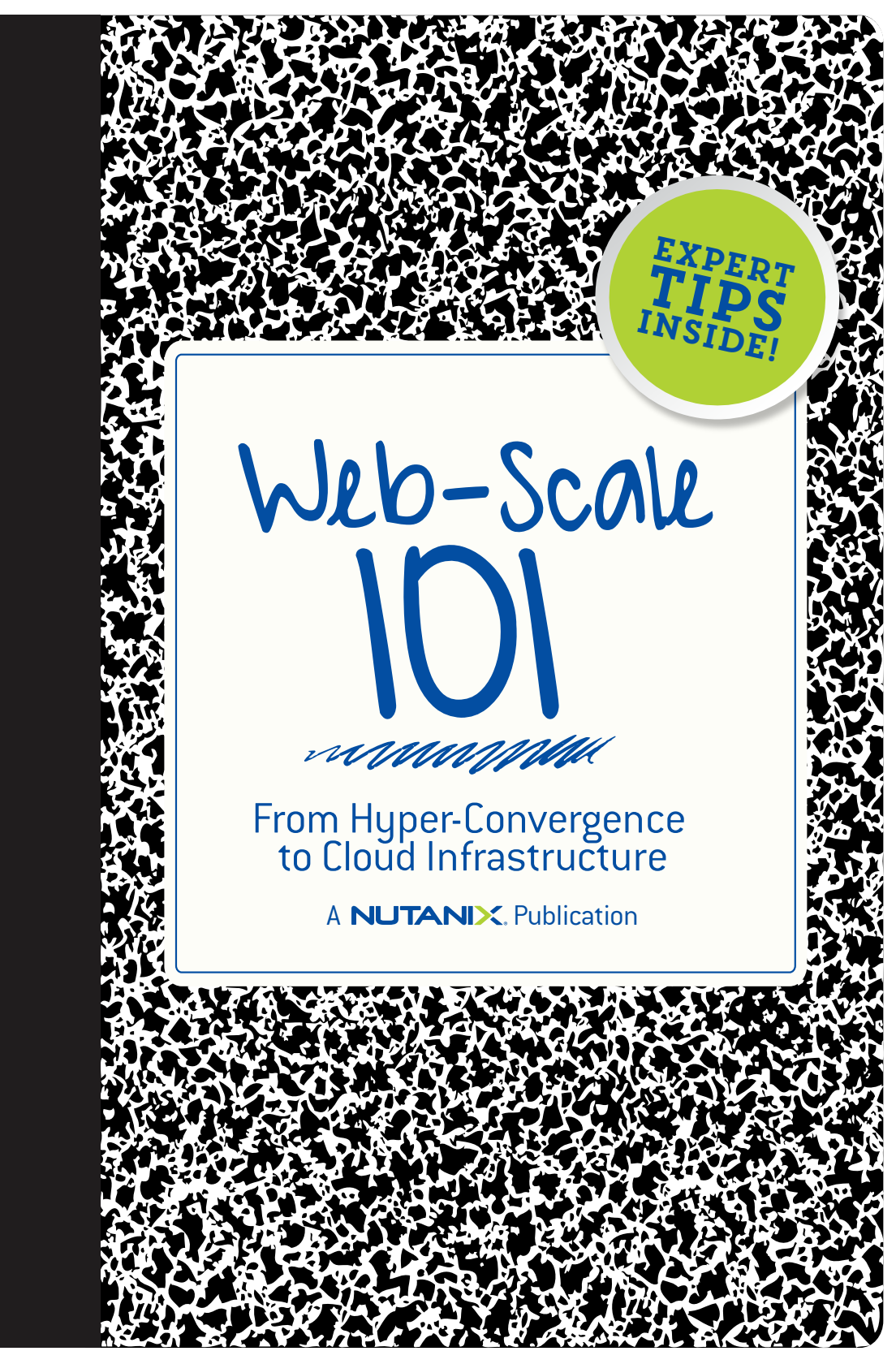

EXPERT
TIPS
INSIDE!
Web-Scale
101
From Hyper-Convergence
to Cloud Infrastructure
A NUTANIX Publication

NUTANIX

Lesson Plan:
Web-Scale 101

Lesson Objectives:

1. What is **web-scale** and what are its key characteristics?
2. Where did **web-scale principles and technologies** come from?
3. What are the potential **benefits** to enterprise IT from web-scale?
4. What are the different approaches to bringing **web-scale principles** and technologies into enterprise datacenters?
5. Why are **web-scale principles essential** to designing hyper-converged solutions?

NUTANIX

Lesson Structure:

1. **Walk through the history** of the enterprise datacenter, covering big trends that have shaped datacenter evolution
2. **Talk about the complexity** in the modern datacenter and its impact
3. Examine the **emergence of web-scale IT** in Internet companies like Facebook and Google
4. **Identify defining characteristics** of web-scale
5. Look at the **relevance of web-scale technologies and principles** to enterprises
6. Discuss the **business benefits** that web-scale delivers
7. Outline **different approaches** to bring web-scale to the enterprise datacenter

Waves of Innovation

The enterprise datacenter has gone through waves of disruption and transformation over the past couple of decades.

1990s Entire ecosystems of server and storage infrastructure, software, applications and tools were built around proprietary platforms from companies such as Digital Equipment Corp, Sun Microsystems, IBM, SGI and many more. Hardware silos were the norm, vendor lock-in was pervasive and the cost of switching between platforms was very high. It was during this era that Storage Area Networks emerged to provide resiliency for the data.

The Late 2000s The rise of server virtualization established x86-based servers as the de facto hardware infrastructure platform for the enterprise and created a fundamental shift in the datacenter landscape. Compute infrastructure went from scale-up to a modest scale-out architecture, and the underlying hardware got abstracted from the layers above. Customers could now deploy a shared compute platform based on commodity hardware and intelligent software that eliminated platform lock-in and could be scaled on demand. This generation also saw the rise of network storage systems, such as SAN and NAS.

The Late 2000s

While virtualization was a big step forward, it did not solve all datacenter problems. Shared storage such as SAN, while required for virtualization, was complex and time consuming to deploy and to manage.

While it was possible to deploy a virtual machine in minutes, provisioning storage, network and security resources still took days or weeks. The growing prevalence of fast flash storage exacerbated the performance problem since storage network latencies became more pronounced. Also, multiple incompatible hypervisor platforms led to hypervisor-based silos that made it difficult, if not impossible, to share infrastructure across hypervisor boundaries.

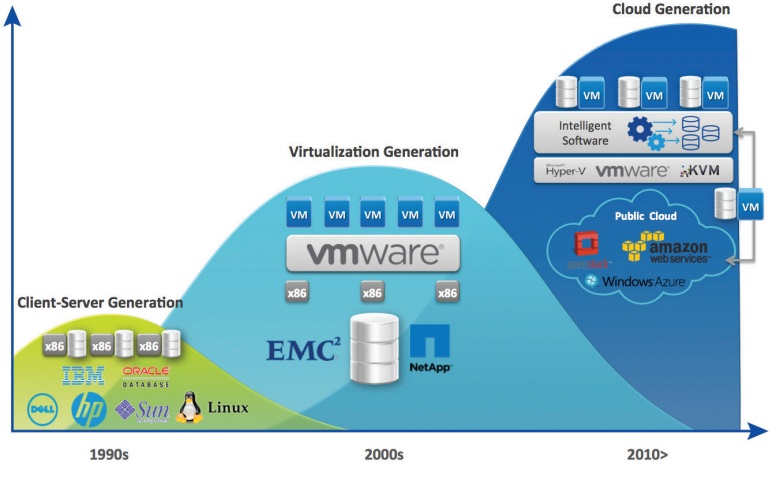


Figure 1: Waves of disruption in the enterprise data center

Now We are at the start of the cloud generation. The vision for the future is a breakdown of boundaries between private and public clouds, and between different hypervisor platforms, enabling complete mobility of virtual machines and data. Public cloud services such as those from Amazon and Google have set the bar for ease of use and rapid self-service, and enterprise IT users expect private infrastructure to step up to the challenge and provide a seamless experience.

Expert Comment
(from Joep Piscaer @jpiscaer)

Although we're only barely scratching the surface of the possibilities of the cloud generation, cloud computing is leading the way for new scale-out compute and storage architectures, breaking down technology silos and removing much of the complexity of the now outdated monolithic architectures. Simpler infrastructure architectures end up simplifying consumption models as well, enabling the consumerization of enterprise IT.

Rise of Complexity

Over the years, datacenters have become incredibly complex.

This complexity is caused in part by the clash between legacy architectures and emerging technologies. While virtualization has enabled unprecedented fluidity in compute resources, network-based storage commonly used in virtual environments is still based on legacy scale-up architectures. Every aspect of the infrastructure lifecycle is a challenge for traditional storage, from buying and deploying to configuring, managing and scaling.

Common Complexities of the Infrastructure Lifecycle:

1. **Infrastructure Planning** Storage administrators need to plan for capacity and performance several quarters in advance and then buy appropriate resources. Customers take on the risk of predicting growth incorrectly because that was the only way to procure new storage infrastructure, because they will need to rip and replace their storage arrays or have to deal with managing multiple systems.

2. **Resource Silos** Storage products are often designed for a narrow range of use cases (e.g., high-capacity iSCSI-based storage arrays for midrange/non-critical applications, all-flash arrays for high-performance applications), forcing administrators to manage silos of resources.

3. **Storage Management** Storage is provisioned and managed in terms of disk groups, storage volumes, LUNs or file systems, while compute resources are managed at the granularity of VMs and virtual disks (VMDK, VHD, etc), further adding to management challenges.

In response to technology complexity, IT organizations have become complex and specialized.

Process frameworks such as ITIL have been developed to bring discipline and manageability to IT service delivery. These frameworks, however, add further layers of cost and rigidity, slowing down service delivery and making the organization less agile.

When IT teams spend a majority of their time and resources in block-and-tackle mode dealing with infrastructure issues, they are not able to focus on initiatives that will move the needle for business.

Complexity is **toxic** because it **stifles** innovation.

Emergence of Web-Scale IT

Consumer Internet giants like Google, Facebook and Amazon started pushing the limits of existing infrastructure systems and processes in ways that traditional businesses did not.

They needed infrastructure that could scale and support their aggressive business requirements, **for example, rapid application development and deployment cycles, the need to scale on demand, and cost containment**. A small tweak to the security settings of a feature could make it go viral and see rapid growth in adoption. Infrastructure needed to be able to elastically grow to meet business demand.

These companies tried to use existing infrastructure architectures, but quickly realized that they were a poor fit for their needs. In response, they developed a better approach to datacenter infrastructure that leveraged a new set of technologies, processes and organizational structure. In effect, they eliminated centralized scale-up network storage such as SAN and NAS that were at the core of many data problems and instead built scale-out distributed systems. They hired teams of engineers, researchers and scientists to build these technologies, and contributed them to the industry as either research papers or open source projects.

This new approach to IT infrastructure, known as web-scale IT, allowed these companies to achieve better business agility and predictable scale while lowering the total cost of ownership (TCO).

Operational simplicity and efficiency became a prerogative so that operational costs did not increase linearly as infrastructure was scaled. Power and space also became important considerations as the infrastructure footprint grew. These businesses in effect handcrafted their IT infrastructure and processes to work optimally for their business needs. Infrastructure resources could be added without limits when needed and managed with limited manual intervention.

For example, the typical ratio of administrator to server in web-scale environments is around 1:20,000* and more.

Expert Comment

(from Denis Guyadeen @dguyadeen)

Google published a seminal technical paper that talked about the Google File System in 2003. A year later, they followed with research papers on MapReduce and BigTable. Not to be outdone, Amazon published papers on its key-value store called Dynamo. These papers raised the level of public awareness about what these companies were doing and gave birth to Hadoop and the Big Data and NoSQL movements.

*Data Center Knowledge, Facebook Ops: Each Staffer Manages 20,000 Servers, November 20, 2013, http://www.datacenterknowledge.com/archives/2013/11/20/facebook-ops-staffer

Did You Know?

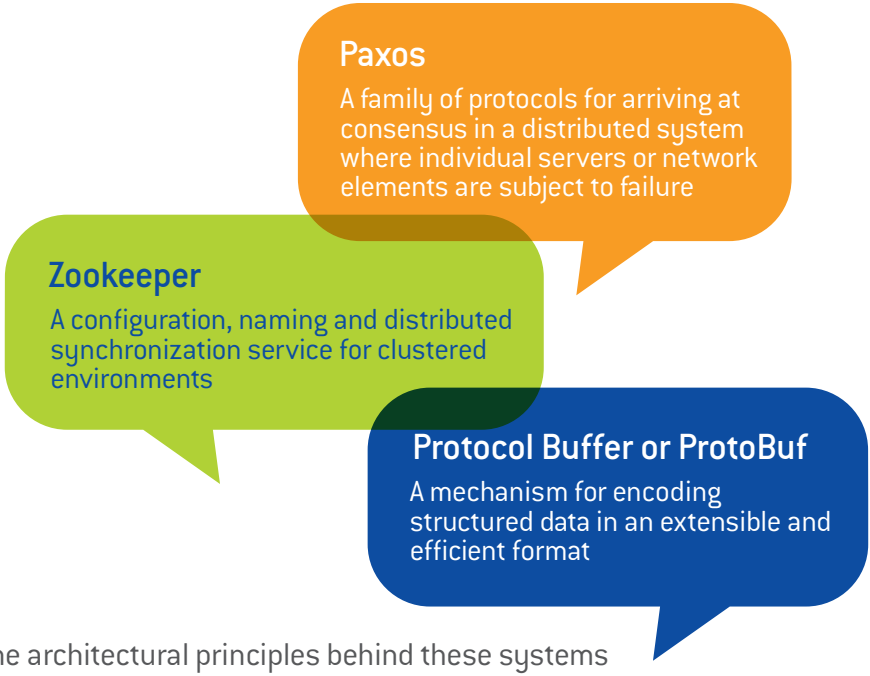
The core tools and technologies behind web-scale are built on the foundation of fundamental research around scalable distributed systems. A short list of examples is provided below.

Cassandra

A distributed database management system designed to handle large amounts of data. Cassandra belongs to a class of DBMS called NoSQL databases in which data is stored and retrieved in models other than relational tables.

Hadoop

A software framework for storing and large-scale processing of datasets on clusters of commodity hardware. At its core, Hadoop includes a distributed file system called HDFS and MapReduce, a programming model for large-scale data processing



The architectural principles behind these systems are fundamentally different from the monolithic architecture of traditional systems.

Expert Comment
(from Stu Miniman @stu)

The cost, simplicity and operational efficiency of Amazon, Google and Facebook is the bar that IT is now managed against. Web-scale methodologies of software-led distributed architecture are entering the enterprise market in the form of solutions that do not require a team of PhDs to configure and maintain. Enterprises can stop being envious of hyperscale IT and start to look at ways to close the gap.

> "By 2017, web-scale IT will be an architectural approach found operating in 50% of global enterprises, up from less than 10% in 2013."
>
> – Gartner, Strategic Technology Trend: Web-Scale Singularity Means Goodbye to Conventional IT Wisdom, February 4, 2014

Until recently, web-scale IT was a possibility for only a handful of Internet companies that had the scale and business need to justify the cost of adopting a radically new IT approach. The IT environment in Internet companies differed from traditional enterprise datacenters. Figure 2 summarizes the difference between cloud/Internet environments and enterprise IT along three dimensions – agility, scale and cost.

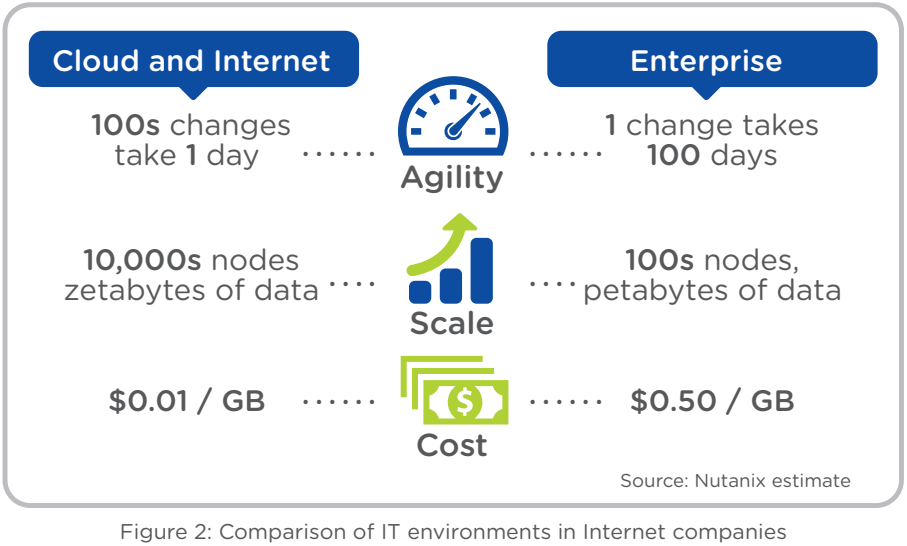


Figure 2: Comparison of IT environments in Internet companies and cloud service providers vs enterprises

Unlike most enterprise datacenters that run a wide range of standard off-the-shelf and in-house applications, these environments had a small set of services that were core to their business. They did not run applications on standard virtualization platforms, but instead ran either bare metal or on lightweight abstraction technologies. Finally, unlike traditional enterprises, Internet companies made extensive use of REST-based programmatic interfaces to extensively automate workflows and manage their infrastructure, rather than using more manually intensive graphical user interfaces.

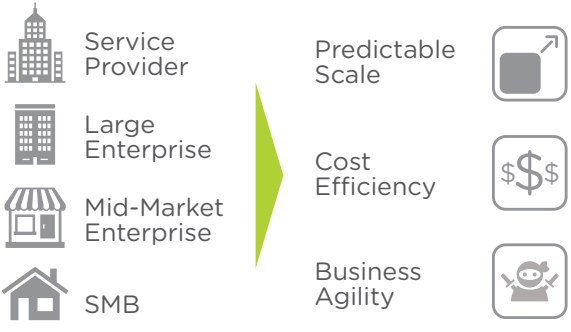


Figure 3: Web-Scale Benefits

But now, businesses are facing many of the same challenges as Internet companies. Budgetary pressures are forcing enterprise IT to find ways to do more with less and keep costs down. As technology becomes more pervasive, businesses are finding ways to use technology to accelerate business growth. This in turn is driving demand for agility in IT services. As businesses grow, IT resources need to scale predictably. While businesses may not match a Facebook or Google in absolute scale of operations, they face many of the same challenges and limitations of traditional infrastructure.

A few large companies, including big banks and global retailers **adopted web-scale IT early on** for a portion of their IT operations that were critical to their business. These companies built custom applications that could leverage web-scale technologies.

But for the most part, enterprises were blocked from embracing web-scale IT for multiple reasons.

- They relied on standard applications running in virtualized environments **that were not designed to use web-scale technologies** built for bare metal workloads.
- For these businesses, IT was a business enabler, **not a source of competitive differentiation or innovation**.
- **They could not justify the high cost** or operational risk of building custom IT environments.
- **Hiring and retaining IT talent was also a challenge**. IT organizations in web-scale environments were culturally very different from traditional enterprise IT, staffed by generalists and with few organizational silos and no technology religion. Web-scale datacenters also typically used the DevOps methodology for application development, which was radically different from traditional IT. Though they had different teams handling backup, security, etc., the underlying culture embraced collaboration and encouraged automation.

Did You Know?

DevOps is first and foremost about a shift towards a web-oriented operations (WebOps) culture. Another key ingredient of DevOps, the concept of Infrastructure as Code, is a technological shift to building and managing infrastructure programmatically. The common goal underlying DevOps is Continuous Delivery. DevOps emerged hand-in-hand with Continuous Delivery, a design and development practice that involves making small, frequent changes and testing at every step in order to reduce the risk inherent in deploying new code.

But now, a new approach has emerged to bring web-scale benefits to the enterprise. Turnkey hyper-converged infrastructure solutions are incorporating the core principles and technologies behind web-scale into their architecture while allowing customers to run the same applications and services that they do today. These solutions make the benefits of web-scale accessible to all enterprises, regardless of size.

Expert Comment

(from Darren Ashley @BEarenaDarrenA)

Web-scale technology is driving a fundamental change in the relationship between IT departments and their customers, the business. By reducing cost but improving time to market, performance and availability, IT now becomes a true enabler for the business, rather than the ball and chain that legacy-based architectures have seen them saddled with.

Ingredients of Web-Scale

Web-scale does not refer to any particular technology or system, but is rather a transformative architectural approach to deploying and managing datacenters.

The term does not refer to a specific technology. Web-scale technologies are designed with a core set of architectural considerations that differ from traditional enterprise IT infrastructure in a number of ways.

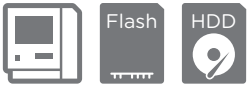
Did You Know?

Internet and SaaS companies typically manage petabytes of data and can push hundreds of software updates every month. For example, Shutterstock, the world's largest online stock photo marketplace makes more than 600 software updates a month and manages petabytes of data. These companies also need to be able to launch new services quickly and scale them on demand.

As technology becomes increasingly crucial to non-technology companies, enterprise IT is faced with similar challenges around agility and predictable scale. For example, banks are publishing mobile apps with check scanning to provide better service to their customers. Retailers are using big data to learn more about their customers and to deliver a more fulfilling shopping experience.

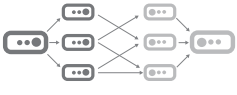
Ingredients of Web-Scale

1 **Hyper-convergence on x86 servers**
Integrated Compute and Storage

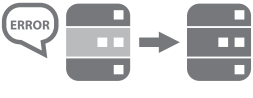


2 **All intelligence in software**
100% Software-defined

3 **Distributed everything**
Cluster-wide data and services



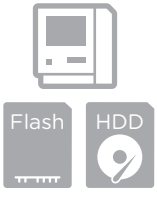
4 **Self-healing system**
Fault isolation with distributed recovery



5 **API-driven automation and rich analytics**
Data-driven efficiency

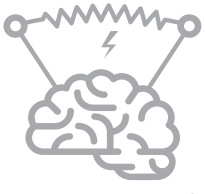
Web-Scale Characteristics

Hyper-Convergence on x86 Servers



Rather than using expensive, special-purpose big-iron hardware, web-scale systems use standard, lower cost x86-based servers. Integrated compute and storage, combined with data locality implemented in software, eliminates the complexity of storage networks and performance-degrading bottlenecks. By bringing data closer to compute, expensive data movement is minimized, resulting in significantly better performance at scale. With intelligent software, infrastructure can be scaled out at the granularity of a single x86 server, enabling fractional consumption and incremental growth. This simplifies infrastructure buying and lowers space and power requirements. By simplifying storage, hyper-convergence also helps break down organizational barriers and streamline operations.

Intelligence in Software



Web-scale infrastructure is **100% software-defined**. All the intelligence is in the software layer, and the system does not rely on special purpose hardware for resilience, performance acceleration or core functions. This allows new capabilities to be quickly added to the infrastructure without requiring hardware upgrades. Infrastructure can be configured on the fly to support different application needs without requiring rewiring or additional hardware purchase.

Web-Scale Characteristics

Distributed Everything

The software is designed in such a way that all data, metadata and operations are distributed across the entire cluster. With **no bottlenecks or chokepoints** in the data or control path, the environment can grow without limits.

Self-Healing System

Web-scale infrastructure is built from the ground up for resilience and always-on operation, **with the ability to recover quickly from individual component failures without bringing down the overall system**. The distributed software isolates faults and restores health using all the resources available to it, while keeping the cluster up and running. Self-healing and rolling upgrades are the norm in these environments

Pro Tip

In order for an infrastructure solution to be considered web-scale, it is essential for both the data fabric and control fabric to follow web-scale principles. For instance, the management software needs to be software-defined, fully distributed and resilient to failures.

Web-Scale Characteristics

API-based Automation and Rich Analytics

Web-scale IT systems are **heavily automated** and use the comprehensive REST API provided by the infrastructure to script routine tasks. Rich analytics enable data-driven efficiency. This allows companies to do more with less and save on operational expenses.

At its core, **web-scale IT is about bringing radical simplicity back to the datacenter.** Scale-out architectures and uniform building blocks simplify the infrastructure buying process by eliminating the need to predict needs in advance. Automation, system resilience and distribution abstract the operational complexity of managing growing systems.

Expert Comment

(from Andre Leibovici @andreleibovici)

Hyper-convergence opens up opportunities for infrastructure standardization and this is the catalyzer for a more agile form of IT. When hyper-convergence meet web-scale technologies mainstream enterprises start being able to benefit from heavy automation and analytics, reducing operational costs to run IT infrastructure.

Benefits of Web-Scale

Web-scale principles and architectures are a big shift in infrastructure management and operation.

By embracing web-scale architecture, businesses can realize significant benefits compared to traditional three-tier infrastructure architectures:

1 Predictable Scale:

Distributed systems that use small x86 server building blocks and have a shared-nothing, massively parallel software architecture can be scaled predictably on demand. Predictable scale is an important operational advantage for companies that anticipate growth because it eliminates the guesswork from infrastructure planning and removes the risk of incorrectly predicting demand for the future. Web-scale is not just about growing infrastructure to petabytes of data or thousands of servers, but rather about achieving just-in-time scaling and pay-as-you-grow economics.

Benefits of Web-Scale

2

Business Agility:

100% software-defined systems allow IT to quickly deploy and repurpose infrastructure to meet dynamic business requirements. Late binding of policies allows these systems to be configured at runtime for a wide range of performance, availability and resilience requirements non-disruptively. Feature velocity in software without requiring hardware upgrades is another important benefit. Automation and rich analytics also eliminate error-prone, time-consuming manual operations, while architectural simplicity makes deployment and management much easier

Benefits of Web-Scale

3

Lower TCO:

Web-scale capabilities can lower the TCO of enterprise infrastructure by 40-60% relative to legacy solutions by reducing both CapEx and OpEx. For example, hyper-convergence lowers power and space requirements of infrastructure by up to 80% relative to traditional three-tier solutions. Automation and analytics reduce the need for manual operation and enable proactive issue resolution. Scale-out architecture with fractional resource consumption makes it possible to buy infrastructure only when needed, thereby increasing resource utilization.

Benefits of Web-Scale

"I can buy based solely on my current needs, and easily scale when I need to – enabling us to rapidly adapt to new requirements"

– Serco

"96% reduction in power costs, 83% reduction in rack space, 53% lower TCO than alternate solutions"

– Langs Building Supplies

"Key business process shortened from ten days to one; technology up and running in an hour"

– Hyundai Australia

Expert Comment

(from Steve Greenberg @stevegreenberg)

In my practice I have found great value at both ends of the customer size spectrum. For the small shop, web-scale offers an "all-in-one" solution for IT generalists offering great benefits in cost, management simplicity, performance and growth potential. For the larger shops, it eliminates the risk for predictive planning for specific projects or whole areas within the datacenter, as well as nasty surprises with cost and performance, all while reducing the overhead of management.

Bringing Web-Scale to the Enterprise

Enterprises that want to get the benefits of web-scale technologies and architectures in their datacenter can do so in a couple of different ways.

One approach is to use open-source web-scale technologies such as Cassandra, Hadoop and MongoDB directly for systems of differentiation and innovation. Companies that build or use custom applications for their business or have a pressing business need (such as big data storage and analytics) can use web-scale infrastructure as the foundation. Many of the technologies developed by Internet companies are now available as commercially supported products built around open-source projects. For example, companies like **Cloudera and MapR use Apache Hadoop at the core of their offerings, while DataStax provides a commercial offering of Apache Cassandra**. Just as Linux became widely available and deployed through commercial distributions from Red Hat, Ubuntu, etc., enterprise-ready commercial versions of web-scale technologies mitigate many of the risks in adopting open-source.

> "By 2020, 25% of global enterprise CIOs will have had previous involvement in corporate Web-scale IT initiatives, up from less than 5% in 2013"
>
> – Gartner, Strategic Technology Trend: Web-Scale Singularity Means Goodbye to Conventional IT Wisdom, February 4, 2014

This approach makes the full power of web-scale IT accessible to enterprises that are able to use them. It requires enterprises to understand the benefits and limitations of specific web-scale technologies, and to learn how to use these technologies to deploy applications. Using web-scale technologies directly requires control over the application architecture, and customers need to develop the organizational know-how to safely and effectively use the tools.

For most companies, however, the risk and disruption of shifting to a fundamentally different IT approach is too high.

Companies are looking for ways to get more web-scale infrastructure capabilities in their environment without needing a complete overhaul. They use standard server virtualization from companies like VMware and Microsoft, and run a host of common off-the-shelf applications. They have little to no control over how these applications are designed, and don't have a team in place to learn about new web-scale technologies.

These customers can make use of turnkey hyper-converged infrastructure solutions built using web-scale technologies and architectural principles that deliver a familiar computing environment on which to run standard applications.

Expert Comment

(from Michael Webster @vcdxnz001)

In the coming years, traditional off-the-shelf enterprise applications will start to be re-architected to scale predictably on demand, or face becoming obsolete with customers instead migrating to cloud based services. We are already starting to see vendors like Microsoft make this transition to a scale-out world. Today's monster VM's will be tomorrow's-scale out applications run on modern web-scale architectures.

Turnkey hyper-converged infrastructure solutions deliver the benefits of web-scale IT but offer the added advantage of requiring no change on the part of the customer to get these benefits. These turnkey solutions are built to support a wide range of standard and custom enterprise applications in virtualized environments. In keeping with core web-scale principles, these solutions are:

- Agnostic to hypervisor platforms and don't depend on any specific hypervisor technologies
- Easy to buy, deploy and manage, allowing enterprise IT to focus on key initiatives to move business forward

Hyper-Convergence and Web-Scale

Hyper-convergence is a fundamentally different architecture from traditional three-tier infrastructure solutions that are prevalent in datacenters today.

Hyper-convergence is an architectural model for IT infrastructure in which compute and storage resources are physically packaged in small servers, typically x86 based. Hyper-converged systems, if designed well, can deliver significant benefits in terms of lower power and space requirements and simplified operations compared to traditional infrastructure. On the other hand, a poorly designed solution will be severely limited in scale and ease of use.

Well-designed hyper-converged systems are by their nature distributed systems. The building block of a hyper-converged solution is a physical x86 server with a processor, memory and storage (including flash). Customers deploy these solutions by buying several of these servers or nodes and connecting them to a top-of-rack Ethernet switch.

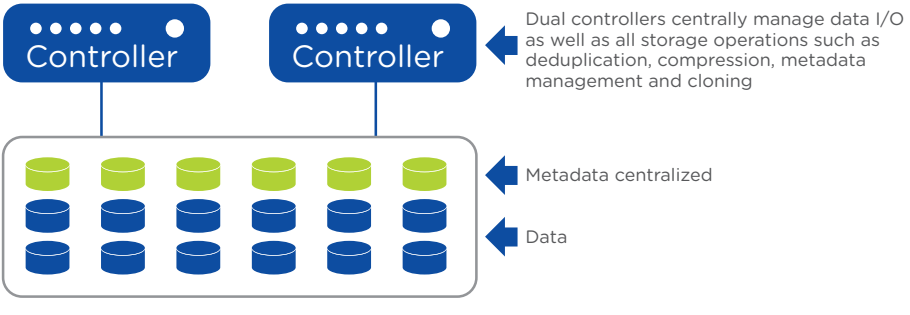


Figure 4: Scale-up storage in a traditional three-tier architecture

The distributed nature of hyper-converged systems requires a fundamental rethink of how data and metadata are stored across nodes and accessed for storage operations. In a traditional scale-up storage system, all the data is stored in drives that sit behind a pair of storage controllers. These controllers are responsible for all I/O, as well as for storage operations such as de-duplication, compression and snapshots.

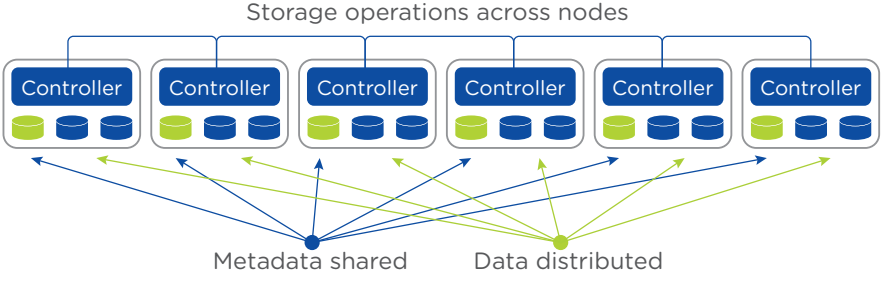


Figure 5: Data and metadata in well-designed hyper-converged systems

In contrast to this centralized approach to storage, hyper-converged systems have data spread across several nodes. Any centralized treatment of the metadata or storage operations will severely limit the scalability of the cluster and/or introduce points of failure or weakness in the system. This is where web-scale architectures come in. The principles of web-scale are tried and tested rules for building distributed systems the right way. In order for hyper-converged infrastructure to deliver the benefits of predictable scale, agility and lower TCO, they need to be designed using web-scale principles.

Expert Comment
(from Dwayne Lessner @dlink7)

Not all hyper-converged systems are created equal. In a scale-out model, distribution of all services are key: performance, metadata, data placement, data services like compression and deduplication. When additional nodes/severs are added to a cluster it needs be a non-event. System administrators should be focusing their time on applications and not worrying if their infrastructure is creating hot spots.

Conclusion

Web-scale is an architectural approach to deploying, managing and scaling datacenters that was pioneered by Internet companies and is now available to enterprises. At its core, web-scale is about minimizing operational complexity in the datacenter and enabling predictable scale, business agility and lower total cost of ownership. Enterprise IT has already started embracing web-scale principles and technologies in response to challenges and needs that traditional architectures cannot handle. **This shift to web-scale will completely transform management and operation of datacenters at any scale in the years to come.**

Test Your Knowledge

1. What is the typical ratio of admin to servers managed in web-scale datacenters?
   a. 1:65
   b. 1:450
   c. 1:1,000
   d. 1:20,000 and more

2. Which aspect of IT infrastructure does web-scale infra structure simplify?
   a. Buying
   b. Deploying
   c. Managing
   d. Scaling
   e. All of the above

3. Web-scale infrastructure is only relevant to web companies with millions of users and petabytes of data to store and manage ☐ TRUE ☐ FALSE

4. Which of the following are not benefits of web-scale infrastructure?
   a. Eliminate need for specialized IT skills
   b. Eliminate islands of storage
   c. Enable infrastructure elasticity and rapid growth
   d. Deliver overall system resiliency by building hardware that does not fail

Test Your Knowledge

5. Massive parallelism with a distributed scale-out architecture is essential to scale infrastructure without limits ☐TRUE ☐FALSE

6. Which of the following are NOT problems with traditional network-based storage?
   a. Network performance bottlenecks as server re sources are scaled and more applications come online
   b. Storage managed with different granularity than VMs (e.g., LUNs, files)
   c. Creates separately-managed resource silos be cause solutions are not versatile for a wide range of use cases
   d. So simple to manage and troubleshoot that ad mins have free time to kill

7. Cassandra is:
   a. a scalable, distributed no-SQL database built to run on commodity servers
   b. a protocol for solving consensus in a network of unreliable processors
   c. a Beatles song
   d. a system for serializing structured data

Answer Key

1. d. Web-scale companies are able to achieve an order of magnitude better leverage of administrator effort through rich automation and powerful analytics.

2. e.

3. FALSE. The benefits of web-scale infrastructure are relevant to companies at any scale. Web-scale in frastructure enables enterprises to respond to business needs quicker and scale rapidly on demand while keeping total costs down. Enterprises of all sizes care about these benefits, and can now get them without having to build infra structure from scratch.

4. d. Web-scale infrastructure is software-defined, and doesn't rely on hardware for resilience

5. TRUE

6. d.

7. a.