

Performance and Scale in Cloud Computing

A Joyent White Paper

Executive Summary

Poor application performance causes companies to lose customers, reduce employee productivity, and reduce bottom line revenue.

Because application performance can vary significantly based on delivery environment, businesses must make certain that application performance is optimized when written for deployment on the cloud or moved from a data center to a cloud computing infrastructure.

Applications can be tested in cloud and non-cloud environments for base-level performance comparisons. Aspects of an application, such as disk I/O and RAM access, may cause intermittent spikes in performance. However, as with traditional software architectures, overall traffic patterns and peaks in system use account for the majority of performance issues in cloud computing.

Capacity planning and expansion based on multiples of past acceptable performance solves many performance issues when companies grow their cloud environments. However, planning cannot always cover sudden spikes in traffic, and manual provisioning might be required. A more cost-effective pursuit of greater scalability and performance is the use of more efficient application development; this technique breaks code execution into silos serviced by more easily scaled and provisioned resources.

In response to the need for greater performance and scalability in cloud computing environments, Joyent Smart Technologies offer scalability features and options that aid application performance, including lightweight virtualization, flexible resource provisioning, dynamic load

balancing and storage caching, and CPU bursting. The Joyent SmartPlatform development environment allows businesses to develop more efficient applications that are easily ported to virtually any open standards environment.

Contents

Introduction	4
Why Worry about Cloud Computing Performance?	4
Isolating the Causes of Performance Problems	5
Misperceptions Concerning Cloud Computing Performance	6
Understanding Performance, Scale, and Throughput	8
Horizontal and Vertical Scalability	9
Administrative and Geographical Scalability	10
Practical and Theoretical Limits of Scale	11
Addressing Application Scalability	11
Application Development to Improve Scalability	12
Joyent Solutions to Performance and Scaling	13
Conclusion	15
References	16

Introduction

As companies move computing resources from premises-based data centers to private and public cloud computing facilities, they should make certain their applications and data make a safe and smooth transition to the cloud. In particular, businesses should ensure that cloud-based facilities will deliver necessary application and transaction performance—now, and in the future. Much depends on this migration and preparation for the transition and final cutoff. Rather than simply moving applications from the traditional data center servers to a cloud computing environment and flick the “on” switch, companies should examine performance issues, potential reprogramming of applications, and capacity planning for the new cloud target to completely optimize application performance.

Applications that performed one way in the data center may not perform identically on a cloud platform. Companies need to isolate the areas of an application or its deployment that may cause performance changes and address each separately to guarantee optimal transition. In many cases, however, the underlying infrastructure of the cloud platform may directly affect application performance.

Businesses should also thoroughly test applications developed and deployed specifically for cloud computing platforms. Ideally, businesses should test the scalability of the application under a variety of network and application conditions to make sure the new application handles not only the current business demands but also is able to seamlessly scale to handle planned or unplanned spikes in demand.

Why Worry about Cloud Computing Performance?

Sluggish access to data, applications, and Web pages frustrates employees and customers alike, and some performance problems and

bottlenecks can even cause application crashes and data losses. In these instances, performance—or lack of performance—is a direct reflection of the company's competency and reliability. Customers are unlikely to put their trust in a company whose applications crash and are reluctant to return to business sites that are frustrating to use due to poor performance and sluggish response times.

Intranet cloud-based applications should also maintain peak performance. Positive employee productivity relies on solid and reliable application performance to complete work accurately and quickly. Application crashes due to poor performance cost money and impact morale. Poor performance hampers business expansion as well. If applications cannot adequately perform during an increase in traffic, businesses lose customers and revenue. Adequate current performance does not guarantee future behavior. An application that adequately serves 100 customers per hour may suddenly nose-dive in responsiveness when attempting to serve 125 users. Capacity planning based on predicted traffic and system stress testing can help businesses make informed decisions concerning cost-effective and optimum provisioning of their cloud platforms. In some cases, businesses intentionally over-provision their cloud systems to ensure that no outages or slowdowns occur. Regardless, companies should have plans in place for addressing increased demand on their systems to guarantee they do not lose customers, decrease employee productivity, or diminish their business reputation. ¹

Isolating the Causes of Performance Problems

Before companies can address performance issues on their cloud infrastructure, they should rule out non-cloud factors in the performance equation. First, and most obvious, is the access time from user to application over the Internet or WAN. Simple Internet speed testing programs, taken from a number of geographically disbursed locations,

or even network ping measurements when testing WAN connections can give companies the average network access overhead.

Next, businesses can also measure the performance of the application on a specific platform configuration to calculate how much the platform affects overall performance. Comparisons of the application speed on its “native” server in the data center compared to its response on the cloud, for example, provide an indication of any significant platform differential, but may not necessarily isolate exactly what platform factor is causing the difference.² If an application’s performance suffers once it is moved to a cloud infrastructure and network access time is not the culprit, Joyent has found from experience that the application’s architecture is likely to be at fault. Inefficiently written applications can result in poor performance on different platforms, as the application is not optimized for a particular architecture.

Optimizing application code for use on cloud platforms is discussed later in this paper, because it directly relates to the more complex topic of enabling cloud computing scalability. Optimum cloud computing performance is not a simple problem to solve nor is it addressed satisfactorily by many current cloud computing vendors.

Misperceptions Concerning Cloud Computing Performance

In a typical corporate network data center, servers, storage, and network switches perform together to deliver data and applications to network users. Under this IT scenario, applications must have adequate CPU and memory to perform, data must have sufficient disk space, and users must have appropriate bandwidth to access the data and applications.

Whenever IT administrators experience performance issues under this scenario, they usually resolve issues in the following way:

- **Poor application performance or application hang-ups.** Usually the application is starved for RAM or CPU cycles, and faster processors or more RAM is added.
- **Slow access to applications and data.** Bandwidth is usually the cause, and the most common solution is to add faster network connections to the mix, increasing desktops from 10 Mbps NICs to 100 Mbps NICs, for example, or faster disk drives, such as SCSI over fibre channel.

While these solutions may solve data center performance issues, they may do nothing for cloud-based application optimization. Furthermore, even under data center application performance enhancement, adding CPU and memory may be an expensive over-provisioning to handle simple bursts in demand. The bottom line is that cloud computing, based on virtual provisioning of resources on an Internet-based platform, does not conform to standard brute-force data center solutions to performance. When companies or cloud vendors take the simplistic “more hardware solves the problem” approach to cloud performance, they waste money and do not completely resolve all application issues. Unfortunately, that is precisely how many cloud vendors attempt to solve performance issues. While they may not always add more physical servers or memory, they frequently provision more virtual machines. Adding virtual machines may be a short-term solution to the problem, but adding machines is a manual task. If a company experiences a sudden spike in traffic, how quickly will the vendor notice the spike and assign a technician to provision more resources to the account? How much does this cost the customer?

Besides the misperception that more hardware, either physical or virtual, effectively solves all performance issues, companies may have a fundamental misunderstanding of how performance, scalability, and network throughput are interdependent yet affect applications and data access in separate ways.

Understanding Performance, Scale, and Throughput

Because the terms *performance*, *scale*, and *throughput* are used in a variety of ways when discussing computing, it is useful to examine their typical meanings in the context of cloud computing infrastructures.

Performance. Performance is generally tied to an application's capabilities within the cloud infrastructure itself. Limited bandwidth, disk space, memory, CPU cycles, and network connections can all cause poor performance. Often, a combination of lack of resources causes poor application performance. Sometimes poor performance is the result of an application architecture that does not properly distribute its processes across available cloud resources.

Throughput. The effective rate at which data is transferred from point A to point B on the cloud is throughput. In other words, throughput is a measurement of raw speed. While speed of moving or processing data can certainly improve system performance, the system is only as fast as its slowest element. A system that deploys ten gigabit Ethernet yet its server storage can access data at only one gigabit effectively has a one gigabit system.

Scalability. The search for continually improving system performance through hardware and software throughput gains is defeated when a system is swamped by multiple, simultaneous demands. That 10 gigabit pipe slows considerably when it serves hundreds of requests rather than a dozen. The only way to restore higher effective throughput (and performance) in such a "swamped resources" scenario is to scale — add more of the resource that is overloaded.

For this reason, the ability of a system to easily scale when under stress in a cloud environment is vastly more useful than the overall throughput or aggregate performance of individual components. In cloud

environments, this scalability is usually handled through either horizontal or vertical scaling.

Horizontal and Vertical Scalability

When increasing resources on the cloud to restore or improve application performance, administrators can scale either horizontally (out) or vertically (up), depending on the nature of the resource constraint. Vertical scaling (up) entails adding more resources to the same computing pool—for example, adding more RAM, disk, or virtual CPU to handle an increased application load. Horizontal scaling (out) requires the addition of more machines or devices to the computing platform to handle the increased demand. This is represented in the transition from Figure 1 to Figure 2, below.

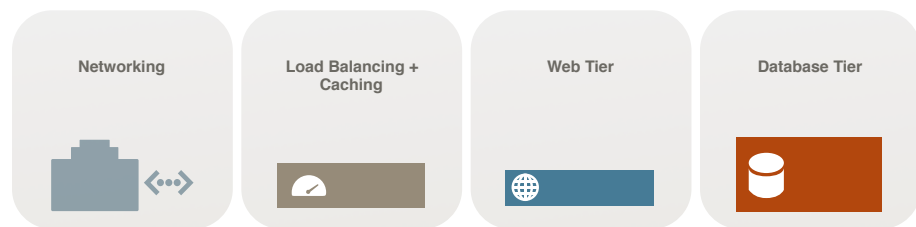


Figure 1: Basic, Single Silo, n-tier architecture

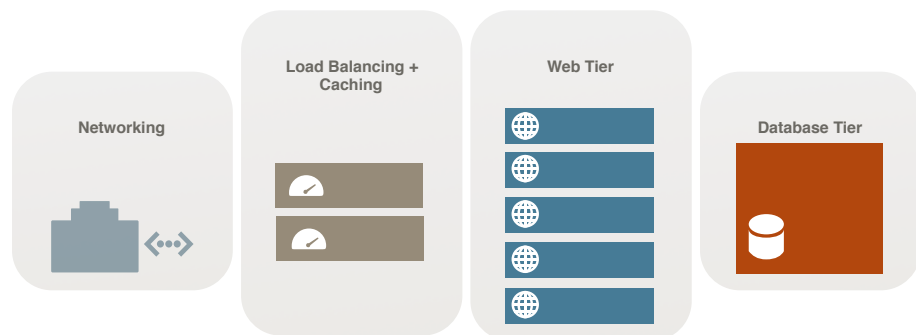
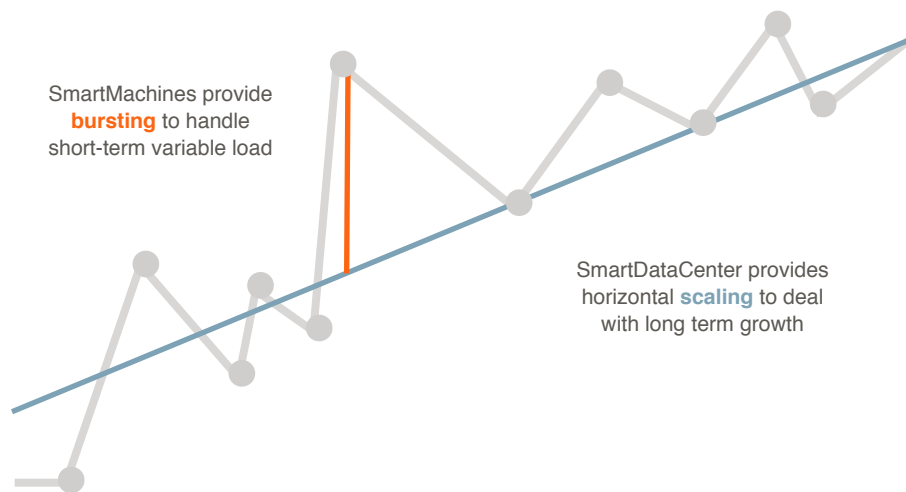


Figure 2: Horizontally scaled load balancing and web-tier. Vertically scaled database tier.

Vertical scaling can handle most sudden, temporary peaks in application demand on cloud infrastructures since they are not typically CPU-intensive tasks. Sustained increases in demand, however, require horizontal scaling and load balancing to restore and maintain peak performance. Horizontal scaling is also manually intensive and time

consuming, requiring a technician to add machinery to the customer's cloud configuration. Manually scaling to meet a sudden peak in traffic may not be productive—traffic may settle to its pre-peak levels before new provisioning can come on line.



Businesses may also find themselves experiencing more gradual increases in traffic. Here, provisioning extra resources provides only temporary relief as resource demands continue to rise and exceed the newly provisioned resources.

Administrative and Geographical Scalability

While adding computing components or virtual resources is a logical means to scale and improve performance, few companies realize that the increase in resources may also necessitate an increase in administration, particularly when deploying horizontal scaling. In essence, a scaled increase in hard or virtual resources often requires a corresponding increase in administrative time and expenses. This administrative increase may not be a one-time configuration demand as more resources require continual monitoring, backup, and maintenance.

Companies with critical cloud applications may also consider geographical scaling as a means to more widely distribute application load demands or as a way to move application access closer to dispersed communities of users or customers. Geographical scaling of resources in conjunction with synchronous replication of data pools is another means of adding fault tolerance and disaster recovery to cloud-based data and applications. Geographical scaling may also be necessary in environments where it is impractical to host all data or applications in one central location.

Practical and Theoretical Limits of Scale

While scalability is the most effective strategy for solving performance issues in cloud infrastructures, practical and theoretical limits prevent it from ever becoming an exponential, infinite solution. Practically speaking, most companies cannot commit an infinite amount of money, people, or time to improving performance. Cloud vendors also may have a limited amount of experience, personnel, or bandwidth to address customer application performance. Every computing infrastructure is bound by a certain level of complexity and scale, not the least of which is power, administration, and bandwidth, necessitating geographical dispersal.

Addressing Application Scalability

For a cloud computing platform to effectively host business data and applications, however, it must accommodate a wide range of performance characteristics and network demands. Storage, CPU, memory, and network bandwidth all come into play at various times during typical application use. Application switching, for example, places demands on the CPU as one application is closed, flushed from the registers, and another application is loaded. If these applications are large and complex, they put a greater demand on the CPU.

Serving files from the cloud to connected users stresses a number of resources, including disk drives, drive controllers, and network connections when transferring the data from the cloud to the user. File storage itself consumes resources not only in the form of physical disk space, but also disk directories and metafile systems that consume RAM and CPU cycles when users either access or upload files into the storage system.

As these examples illustrate, applications can benefit from both horizontal and vertical scaling of resources on demand, yet truly dynamic scaling is not possible on most cloud computing infrastructures. Therefore, one of the most common and costly responses to scaling issues by vendors is to over-provision customer installations to accommodate a wide range of performance issues.

Application Development to Improve Scalability

One practical means for addressing application scalability and to reduce performance bottlenecks is to segment applications into separate silos. Web-based applications are theoretically stateless, and therefore theoretically easy to scale—all that is needed is more memory, CPU, storage, and bandwidth to accommodate them, as was depicted in Figure 2. However, in practice Web-based applications are not stateless. They are accessed through a network connection(s) that requires an IP address(es) that is fixed and therefore stateful, and they connect to data storage (either disk or database) which maintains logical state as well as requiring hardware resources to execute. Balancing the interaction between stateless and stateful elements of a Web application requires careful architectural consideration and the use of tiers and silos to allow some form of horizontal resource scaling. To leverage the most from resources, application developers can break applications into discrete tiers—state or stateless processes—that are executed in various resource silos. Figure 3 depicts breaking an application into two

silos identified by their DNS name. By segregating state and stateless operations and provisioning accordingly, applications and systems can run more efficiently and with higher resource utilization than under a more common scenario.



Figure 3: Multi-silo, n-tier, scaled architecture

Joyent Solutions to Performance and Scaling

Joyent has applied its extensive experience in providing cloud computing infrastructure and services and developed its Smart Technologies range of products, from its SmartOS operating system, SmartMachine virtualization technology, SmartDataCenter infrastructure management system, to its SmartPlatform development environment. This Joyent architecture provides a highly elastic cloud infrastructure that accommodates bursts in traffic that other cloud infrastructures cannot. The Joyent Smart Technologies architecture has the following key performance and scaling advantages over traditional cloud computing infrastructures:

SmartMachine lightweight virtualization. Joyent SmartMachines have been designed to provide best possible performance with limited overhead. The SmartOS operating system combines the operating

system and virtualization to eliminate redundancy and maximize available RAM for applications.

SmartCache. Joyent makes use of all of the unused DDR3 memory in the cloud by providing a large ARC Cache pool delivering unparalleled Disk I/O. Both reads and writes are greatly improved as content that would traditionally be served from disk are cached in high speed memory without any customer interaction.

CPU bursting. The Joyent implementation of its CPU engine allows on-demand processing cycles from a resource pool of available CPUs, enabling instantaneous vertical scaling to meet bursts of application demand without costly and time-consuming manual provisioning of resources.

Choice of virtualization. While SmartMachines provide ideal performance, Joyent recognizes that legacy operating systems and development environments are required for many applications, and Joyent SmartOS provides XVM virtualization technology as an integral component of the OS allowing for other operating systems such as Windows and Linux. These operating systems are still able to take advantage of SmartOS capabilities such as SmartCache to achieve improved performance, and management by SmartDataCenter.

Build clouds on architecture not rent-a-machine. The Joyent SmartDataCenter architecture is built with performance and scale of applications in mind versus a simplistic concept of adding more and more virtual machines to solve application performance issues. Joyent understands that application architecture is supported by several tiers of servers that need low latency interconnect. Our patent pending Honeycomb design ensures that servers (Web, App, DB, Cache) while completely distributed and fully redundant, are provisioned in the highest performance and lowest latency manner possible. Rent-a-machine

cloud solutions merely move physical data center inefficiencies to virtual, cloud-based inefficiencies.

Joyent's Smart Technology infrastructure provides additional performance and scalability enhancements over traditional cloud computing platforms. Joyent's use of SmartDataCenter network provisioning include automatic caching of distributed file systems and load balancing of network traffic so that the network can dynamically conform to demands made by applications with no administrator input required; in many cases this can eliminate the need for adding physical or virtual resources.

Joyent's SmartPlatform is an application development environment that allows businesses to write more efficient application code for cloud computing environments. SmartPlatform is flexible and supports a number of standard programming tools and languages, making SmartPlatform applications portable to nearly any cloud computing environment.

In addition to these core technology advancements, Joyent professional services provide a complete methodology that helps build applications with architectures that will operate at peak performance, scale to handle demand, and maximize return on investment. Joyent's multiphase methodology, the Joyent Smart Architecture Method, helps deploy the right architecture for every application. This approach is suitable for existing applications that are transitioning to the cloud as well as new applications that are being developed from scratch.

Conclusion

Scalability is the best solution to increasing and maintaining application performance in cloud computing environments. Cloud computing vendors often resort to brute-force horizontal scaling by adding more physical or virtual machines, but this approach may not only waste

resources but also not entirely solve performance issues, especially those related to disk and network I/O. In addition, customers and vendors alike have practical limits on their ability to scale, primarily constrained by costs and human resources. Smart application development can alleviate performance issues in many cases by isolating resource-intensive processes and assigning the appropriate assets to handle the load. However, for the most part scaling to meet performance demands remains a manual process and requires vigilant monitoring and real-time response.

Joyent's Smart Technologies address many issues of scalability and performance in cloud computing, including dynamic vertical scalability, more efficient allocation of virtual resources, and efficient I/O load balancing.

References

1. Mouline, Imad. "Why Assumptions About Cloud Performance Can Be Dangerous." *Cloud Computing Journal*. May, 2009.
www.cloudcomputing.sys-con.com/node/957492
2. Nolle, Tom. "Meeting performance standards and SLAs in the cloud." *SearchCloudComputing*. April, 2010.
http://searchcloudcomputing.techtarget.com/tip/0,289483,sid201_gci1357087_mem1,00.html