SYS-CON

Cloud Computing: It's the Future of Enterprise

CLOUD COMPUTING JOURNAL

CLOUDCOMPUTING.SYS-CON.COM

MAY/JUNE 2009 / VOLUME: 2 ISSUE 3

MEASURING CLOUD PERFORMANCE

CLOUD COMPUTING JOURNAL

PRESIDENT & CEO
Fuat Kircaali fuat@sys-con.com

Senior VP, Editorial & Events
Jeremy Geelan jeremy@sys-con.com

EDITORIAL
Executive Editor
Nancy Valentine nancy@sys-con.com
Associate Online Editor
Nicole Russo nicole@sys-con.com
Research Editor
Bahadir Karuv, PhD bahadir@sys-con.com

ADVERTISING
Senior VP, Sales & Marketing
Carmen Gonzalez carmen@sys-con.com
Advertising Sales Director
Megan Mussa megan@sys-con.com
Associate Sales Manager
Corinna Melcon corinna@sys-con.com
Advertising Events Associate
Krisandra Russo krisandra@sys-con.com
Susan Wechtler susan@sys-con.com

EVENTS
Events Manager
Sharmonique Shade sharmonique@sys-con.com

PRODUCTION
Lead Designer
Abraham Addo abraham@sys-con.com
Associate Art Directors
Louis F. Cuffari louis@sys-con.com
Tami Beatty tami@sys-con.com

CUSTOMER RELATIONS
Circulation Service Coordinator
Edna Earle Russell edna@sys-con.com

SYS-CON.COM
Web Designers
Stephen Kilmurray stephen@sys-con.com
Richard Walter richard@sys-con.com

ACCOUNTING
Financial Analyst
Joan LaRose joan@sys-con.com
Accounts Payable
Betty White betty@sys-con.com

SUBSCRIPTIONS
subscribe@sys-con.com
Call 1-888-303-5282
For subscriptions and requests for bulk orders, please send your letters to Subscription Department $49/yr (6 issues)

EDITORIAL OFFICES
SYS-CON MEDIA
577 Chestnut Ridge Rd.,
Woodcliff Lake, NJ 07677
Telephone: 201 802-3000 Fax: 201 782-9601
Virtualization Journal (ISSN #1549-1331)
is published bimonthly (6 times a year)

Postmaster: send address changes to:
Virtualization Journal
SYS-CON MEDIA
577 Chestnut Ridge Rd.,
Woodcliff Lake, NJ 07677

FOR LIST RENTAL INFORMATION:
Kevin Collopy: 845 731-2684 kevin.collopy@edithroman.com
Frank Cipolla: 845 731-3832 frank.cipolla@epostdirect.com

REPRINTS
For promotional reprints, contact reprint coordinator
Megan Mussa megan@sys-con.com

SYS-CON MEDIA

DOWNLOAD PRINT VERSION

Comparing Cloud Computing Providers

BY ALAN WILLIAMSON

I've just come back from hosting another packed Cloud Computing Bootcamp, at the Cloud Computing Expo in Prague, Europe. We've done a number of these bootcamps in the past six months and we're starting to see a common theme of questions bubble through. Namely, how fast is it, and what support can I expect?

A number of real alternatives are beginning to pop up on the cloud infrastructure scene now. In the past year alone, we've seen the number of providers nearly double, giving the "granddaddy" of cloud computing, Amazon, a run for its money.

Now we have real choice to build solutions that are no longer locked into a particular provider – assuming of course we don't immerse ourselves too deeply in their feature list. But how can you compare providers?

The first comparison point is performance. This is a common question asked at our bootcamps and it is important to understand that you are never going to get the same level of throughput as you would on a bare-metal system. As it's a virtualized world, you are only sharing resources. You may be lucky and find yourself the only process on a node and get excellent throughput. But it's very rare.

At the other extreme, you may be unlucky and find yourself continually competing for resources due to some "noisy neighbor" that is hogging the CPU/IO. The problem is that you never know who your neighbors are. The only way you can guarantee the exclusive resource is to move out of the cloud.

When comparing providers, it's important not to try and compare too closely the performance of say a small instance at Amazon, with the small slice at Mosso, or the small server-RAM at GoGrid. It's very tempting to do it, and you would think on paper that this would be a fair way of knowing how many instances from each provider you need to host your cloud application, but it's not. The only real way to tell is to literally try it. Cloud Computing is so inexpensive that you have the luxury of trying a number of providers at the same time to see which satisfies your own metrics.

The other big area of concern that you can use as a comparison point is support. Just how available are they when things go wrong. Can you reach out to someone to talk to? Can you email a helpdesk and get your query answered?

With Amazon You Are on Your Own

This is sadly where Amazon fails big time. They have their public forums, but other than paying for support specifically, there is no free help. You are on your own. Amazon takes little responsibility and it's quite common to arrive in the morning and discover an instance has been turned off due to some fault and you've lost all the data on those disks. At Amazon you have to build redundancy in from the start.

Just recently we had the need to move what was historically running on a 12 bare-metal server infrastructure over to the cloud. Naturally we would need more than 12 instances and, as a rule of thumb, for every real server, you need two cloud instances. It's not a hard and fast rule, more of a guideline. For this particular client, we didn't have existing accounts set up with the providers so we had to sign-up there and then and get moving. We architected the system, and decided that GoGrid was going to be the best provider for us at this time. However, we got sidelined instantly and had to ask technical support about a specific issue. On the Saturday morning the question was asked, and we have yet to get an answer to that issue. Their online accounting system was not approving new accounts over the weekend so we had to wait until the following Monday to start using GoGrid.

CONTINUED ON PAGE 5

Cloud Computing – Removing Risk and Conserving Capital

Core to the ability to preserve capital issue is the ability to upsize your IT infrastructure on demand

BY DAVID LINTHICUM

So, I'm a computer guy. What the heck is capital? As defined by Wikipedia it's:

"...any liquid medium or mechanism that represents wealth, or other styles of capital. It is, however, usually purchasing power in the form of money available for the production or purchasing of goods, etcetera. Capital can also be obtained by producing more than what is immediately required and saving the surplus." (http://en.wikipedia.org/wiki/Financial_capital)

In other words, it's money in the bank that allows the business to run. Thus, the more money we have in the bank, the more we can purchase things for the core business such as inventory that can be sold, or new plant equipment that will save the company money during production. In any event, it's good to keep as much capital as possible on hand to invest in the business, and not into infrastructure such as data centers, hardware, and software.

Considering the case study above, if we need to conserve capital, than cloud computing seems to be the way to go. We pay as we use the service, there isn't any hardware and/or software to buy, and thus we can keep the money in the bank for other purposes.

While many look at the technical value of cloud computing, the ability to preserve capital is the primary reason people leverage cloud computing from a business perspective. Startups can launch an entire company with almost no IT expense, new divisions can be created with little IT capital investment, and as the business expands, there is no need to reinvest in hardware, software, and data center resources as the business scales up. Moreover, there is no reason to keep those capital resources around as the business scales down. With cloud computing, it's just a matter of paying more, or less, for the use of the service. We can call this the value of upsizing and downsizing on-demand.

Core to the ability to preserve capital issue is the ability to upsize your IT infrastructure on demand, or simply pay more money for additional computing capacity. Many cloud computing providers call this being elastic, or the ability to grow or contract to accommodate the business. There are a few terms to consider here:

- Service Tiers
- Existing Resources
- Ability to Scale

Service tiers refers to the fact that some cloud computing providers offer services using tiers of service, meaning that while you can purchase some capacity for $1,000 a month, the next tier up is $2,500, and the next one up from that is $5,000. They are not selling the service for a more granular billing mechanism, to say it in another way. Moreover, you have to consider the contracts as well. While some are monthly, some are yearly, and some are both. Thus, many of those contracts can drive a commitment that may not be right for the enterprise. Be sure to read the service agreement carefully before selecting a cloud computing provider.

Existing resources refers to the fact that you may have some computing capacity around in the data center that's already bought and paid for, and thus using those resources should be considered. However, you also need to consider the costs of development and maintenance, which are typically where the real costs come in. The trick is to understand that your CIO will probably ask about those resources during the cloud computing discussions, and you need to have a well-thought-out answer.

Ability to scale refers to the ability for the cloud computing provider to actually provide the capacity you require to support the additional computing resources you'll need. While most can scale up to your needs, there are some that won't be able to handle the additional load no matter how much money you pay them. You need to determine that up front.

About the Author

David S. Linthicum, Editor-in-Chief of SYS-CON's Virtualization Journal, is an internationally known cloud computing and service-oriented architecture (SOA) expert. He is a sought-after consultant, speaker, and writer, and formed Blue Mountain Labs, a leading cloud computing advisory, validation, and consulting organization. In his career, Linthicum has formed or enhanced many of the ideas behind modern distributed computing including EAI, B2B Application Integration, and SOA, approaches and technologies in wide use today. For the last 10 years, he has focused on the technology and strategies around cloud computing, including working with several cloud computing startups. His industry experience includes tenure as CTO and CEO of several successful software and cloud computing companies, and upper-level management positions in Fortune 500 companies. In addition, he was an associate professor of computer science for eight years, and continues to lecture at major technical colleges and universities, including University of Virginia and Arizona State University. He keynotes at many leading technology conferences, and has several well-read columns and blogs. Dave has authored 10 books, including the ground-breaking "Enterprise Application Integration" and "B2B Application Integration."

david@bluemountainlabs.com.

Comparing Cloud Computing Providers

CONTINUED FROM PAGE 3

Mosso's Technical Support Is Top-Notch

Not wishing to lose the weekend and wait, we went to our second choice, Cloud Servers from Mosso. They have instant sign-up and within the hour we had our instances all spun up starting to work. We had some teething issues with Mosso. Performance isn't terribly good on their lower end instances and you have to move far up the chain before you get anywhere near the level we expected. But their technical support is excellent and top-notch. They were on hand at all times during the weekend, assisting us with some issues we had.

GoGrid's lack of responsiveness lost them what is effectively a $20k yearly account. Such is the world of instant server provisioning – live by the instance, die by the instance.

Support should never be underestimated. In our experience a good support network outstrips any kinks or issues a platform may have. Flexiscale in the UK is another cloud provider that has had us ring them at all hours, and they've been there answering and helping us out.

Google App Engine: 'Cloud It Yourself'

The big vendors have a long way to go to with their support. Amazon requires you pay. Google App Engine isn't good at answering direct questions. Their attitude seems to be, CIY: Cloud It Yourself. But this won't last.

While the large players have poor direct one-on-one support, they usually have the most online documentation and largest attended forums. But this will only go so far in diagnosing a specific issue.

If your business relies on the speedy resolution of an issue from your cloud provider, then test them before you try. Send them that email out-of-hours to see how quickly and how useful the reply actually is. Try and reach someone on the phone. Look around the forums to see how many questions are left unanswered.

Get a feel for how they treat problems, because this little piece of research could make the difference between having a great cloud experience from one that could have you running back for your bare metal systems.

About the Author

Alan Williamson is Editor-in-Chief of Cloud Computing Journal.

Cloud Computing and the Enterprise

Leveraging both on-premises and Cloud systems

BY SCOTT YOUNG

Cloud Computing – a topic that has been steadily growing in popularity and interest in the IT industry over the past decade – has the potential to significantly change how IT works by offering ready access to new capabilities, less expensive IT resources and unrivaled flexibility for businesses.

Enterprises have typically viewed Cloud Computing as a service delivery trend that applies primarily to small-to-medium businesses. But with large vendors like IBM, Dell, Microsoft, and Google adopting positions on Cloud Computing that view appears to be changing and will continue to change as larger enterprises gain a better understanding of how to match software and service delivery models to particular types of users.

Google and Microsoft provide good examples of Cloud-based services. Google is a leader in the Cloud Computing space and its Gmail and Apps provide well-known and trusted alternatives to on-premises dedicated infrastructure, yet issues around data security, service resiliency, and access remain such that most large organizations are unlikely to opt for an "all or nothing" approach to Cloud Computing.

Despite the cost-savings advantages that Cloud Computing can provide, most organizations seem unwilling to choose between their existing on-premise infrastructure and Cloud-based services. If the idea of abandoning their existing infrastructure doesn't scare them off, the thought of managing yet another system most likely will. Take the real life example of a 1,000-person software company in Silicon Valley that spends roughly $580,000 a year to maintain its on-premises messaging/collaboration infrastructure (including the Fibre Channel SAN) compared to about $80,000 a year (according to the Google Cost Calculator) to maintain Web-based service costs.

The reality is that most large organizations would prefer a scenario where they leverage both on-premises and Cloud systems based on cost, flexibility, and user needs. That said, supporting two different mechanisms doesn't make much sense unless you can manage them both without increasing IT overhead. Other challenges involved in adopting Cloud Computing include issues such as the security and privacy of business data in remote third-party data centers, fear of platform lock-in, and concerns about reliability and performance. But what if organizations could switch over some percentage of their users and leverage the best of both worlds to meet requirements for different types of users? Many enterprises are realizing that they can no longer ignore the flexibility and cost-savings potential of cloud computing and are choosing to mix their on-premises collaboration applications with it.

When a development team creates an on-premises application, most of the requirements for that application already exist. An operating system provides basic support for executing the application and interacting with storage, while other computers in the environment offer services such as remote storage. Similarly, cloud applications can be built on a cloud foundation. Both kinds of applications can access infrastructure and application services provided on-premises and in the Cloud. Just as on-premises platforms support today's applications, Cloud platforms provide services for the applications likely in the future.

Organizations want to get the best of both worlds with tools that allow them to save money by deploying software based on user requirements rather than be forced into a one-size-fits-all plan. Before IT organizations look to external cloud providers, they may opt for internal Clouds – "cloudy" environments that are implemented in a company's own data centers. Since on-premise and Cloud platforms can be used together as a "shared resource," enterprises are seeking tools and technologies that enable them to share management and avoid the complexity and IT overhead of managing two very different environments.

In some instances, large organizations are deploying their own Cloud Computing environments, using service-oriented architecture (SOA) and the Internet to allow users to access global IT resources. Organizations can also combine services and resources housed in their own data centers with those of third-party service providers, or integrate services from third-party providers into their environments. Internal environments deliver many of the benefits of Clouds and position enterprises to use external Clouds in the future, as supplier offerings improve and barriers to enterprise adoption are overcome. Software-as-a-service (SaaS) is playing a significant role in integration and building internal cloud environments, and represents a major component of the enterprise cloud computing model.

Companies, including Intel IT, are taking advantage of SaaS and niche infrastructure as a service (IaaS) implementations whenever possible to build internal Cloud Computing environments from the inside out. Using SaaS for Cloud integration allows enterprises to integrate multiple SaaS services such as those provided by Salesforce.com and eAutomate, but also offers the wherewithal to blend Cloud services into their traditional IT resources.

There are now dozens of companies scoping out this SaaS-IT integration market. The growth of SaaS applications and Cloud-based computing has been staggering – it's estimated by Gartner that worldwide SaaS revenues will grow to $19.3 billion by year-end 2011, while the market opportunity for Cloud-based computing is in excess of $160 billion. SaaS is becoming increasingly more flexible and customizable than ever before, with more and more new tools and strategies available to help IT departments work out integration issues.

Cloud Computing is no longer just for small and mid-sized companies. The scalability and cost efficiency that it provides make it appealing to much larger enterprise organizations looking to compete more efficiently in the global market. The odds are good that within the next five years the popularity of Cloud Computing in the enterprise will grow significantly. Yet Cloud Computing alone is not the answer. The Cloud doesn't yet provide the full spectrum of an on-premises environment and enterprises are not inclined to completely abandon their multimillion-dollar on-premises investment. The key to achieving the greatest success is for enterprises to use efficient software to integrate their existing on-premises infrastructure with the cloud.

About the Author

Scott Young is vice-president of marketing and product management at Ensim Corporation, Santa Clara, CA. Scott has over 17 years experience working as a senior executive in technology industries including enterprise software, internet services, and semiconductors.

syoung@ensim.com

Into the Cloud We Go.....

Have we thought about the security issues?

BY DAVID HOBSON

A new shift in computing is upon us – Cloud Computing. As our use of computing resources evolves from mainframes to PCs and networks we are now facing a major shift in the way we work. This could have dramatic effects on the way we use computers, both for work and play.

But the security issues need to be discussed, risks assessed and judgements made knowing the risks and issues. For some Cloud Computing makes a lot of business sense, for others, it may create confusion.

So what is Cloud Computing? For many it's the natural evolution of the Internet. The Internet has provided a major shift in the way we work. Less than 20 years ago, there was a comment by Ray Noorda, the CEO of Novell – "If you don't have an e-mail address on your business card, you will be considered a nobody" and most people did not believe it. Twenty years later and it seems pretty much everyone has an e-mail address, if not one at work, then a Hotmail, Gmail, or Yahoo! account. And these e-mail accounts are the first example of Cloud Computing!

Cloud Computing gets its name from network diagrams where the Internet is always shown as a cloud, as the route taken through the Internet can not normally be defined and is unknown. The route is irrelevant. The concept of Cloud Computing is that the central computer system or systems are hosted in the Internet and their actual location is irrelevant to the application and its successful deployment. The architecture is relatively simple – a data store and server are hosted on the Internet, and the client can access the server from anywhere. Normally the client will have a Web-based front-end to make access even easier. The first major examples are the e-mail services from Hotmail and the like mentioned above.

The concepts of Cloud Computing have evolved to the concept being promoted today where there will be no need to purchase software; it will be rented either on an annual basis or on a pay per use model. And now the model has added the concept of free use of software in return for receiving ads.

The major benefit of Cloud Computing for a user is financial. There's no need to invest in hardware infrastructure, or software. However there are a number of issues that need to be considered.

The old definition of security is as valid today as it ever was – CIA – Confidentiality, Integrity, and Availability. And these three areas need to be addressed by any potential user of Cloud Computing. The major issue is confidentiality. If you're giving your data to a third party, you have no control over it. So who have you given it to? What is the access to the data? Who sees it? Can it be taken and used by someone else? Who administers this? What assurance do you have that your data is confidential? Are you happy with a contractual warranty? If so, what is your recourse if the contract is breached?

Are you convinced as to the integrity of your data? Can it be tampered with? If it was tampered with, would you know – most people wouldn't.

> "The old definition of security is as valid today as it ever was – CIA – Confidentiality, Integrity, and Availability"

Are you satisfied with the segregation of data? What is the chance of "leakage" and how is this protected and tested?

And finally availability. If your data is not available to you, for whatever reason, then it's no good to you. Cloud Computing may actually provide much stronger backup and provision for disaster recovery than a private enterprise. Most solutions will provide at least one backup resource, maybe more. Any subscriber should check what provisions are made. However access is required to the Internet to access your data. If for any reason an ISP failed then all access fails with it. So redundancy in Internet access is imperative. There are a number of products that offer offices small and large the ability to bind multiple ISPs to provide a virtual single access to the Internet. The other issue with availability that has to be considered is the transfer of data. There are two major areas of concern. First, one service offered in the Cloud is remote backup. If you need to get your data back from a remote data store, how long will it take to download everything in the event of an emergency? And when was this last tested. Almost certainly this will be a major issue, as the size of most people's Internet connection is relatively small compared to their LAN. The second issue is moving service providers. If you want to use a service like Salesforce.com for outsourced CRM, you may be limited to the data being stored in a proprietary format. If you are unhappy with the service and wanted to move to an alternative, how would you get your data back? And would it be useable?

In recent years, as well as CIA, three other areas have become of major concern to business – Compliance, Policy, Risk. Compliance is now a major business issue. The data being stored in the Cloud must be considered carefully. What type of data is it? Is it confidential? Are there regulations to control how and where it's stored? In the U.K. we have the Data Protection Act which is very strict on data storage. If the data is being stored in the Cloud do you know where it's being stored? Are you breaking legal requirements? Your policies on data storage must address these legal issues, and any Cloud Computing must be considered very carefully.

Finally risk. We have spoken about concerns with the data and Confidentiality, Integrity and Availability – but what if your service provider goes bust? How would you get your data back? What if the ownership changes and policies change?

One risk often not considered is that putting your data with a major provider creates a bigger target for hackers. If the service provider is hacked, or suffers some virus or security breach, how will your data be affected? Service providers have suffered already from hackers. While they will argue they can invest more in security than many people, they are without doubt a bigger prize. Some say there's much to be said for security by obscurity.

All these issues apply when outsourcing computing. Currently a lot of enterprises outsource their computing to save money. And the outsourcer provides a private Cloud to give the relevant service. But all the questions we have raised apply equally; however the answers may be easier to get from an outsourcer and contracts can be drawn up to ensure compliance with your policies.

About the Author

David Hobson founded Global Secure Systems (GSS) in 1997 and successfully merged with Peapod last year with the result of becoming one of the UKs largest IT security integrators and consultancy firms with over 2,500 customers. The company has offices in Worthing, West Sussex and Bracknell, Berkshire with additional sales offices in London and Dublin. www.gss.co.uk.

The Impact of Cloud Computing on Enterprise Architecture

Initial best practices

BY DAVID BRESSLER

It's no recent discovery that integration is IT's biggest challenge, and the difficulties associated with this obstacle have become more significant as a result of the increasing amount of information that is traveling across infrastructures.

Frequently, this difficult situation is exacerbated by non-technical managers who are not aware of all the effort necessary to integrate data properly. As a result, a potentially dangerous and costly domino effect ensues: these executives overlook or ignore the difficulties associated with integration; underlying architecture is sacrificed for speed, yet still many projects are deployed too quickly; users' first experiences are negative; and ultimately, the development team is left in a permanent state of catch-up. This approach only masks the underlying integration effort necessary for proper integration and the complexities of doing it right the first time.

Another confounding factor revolves around budget. Unfortunately, it's not uncommon for most of the money allotted to IT projects requiring significant integration efforts to be spent well in advance of revenue (or benefit) generated by an application. As a result, there's a mismatch between when spending occurs and when the offsetting benefits or revenue accrue, an accountant's nightmare. Albeit indirectly, these "cultural" and financial obstacles are among the main difficulties of integration projects.

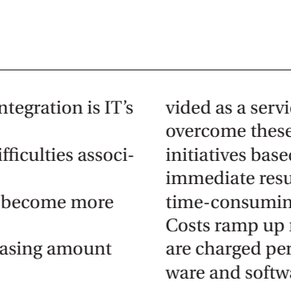
Fortunately, Cloud Computing – essentially an offering where infrastructure is provided as a service – promises to help businesses overcome these challenges. Briefly, integration initiatives based on Cloud technologies see more immediate results since they don't mandate a time-consuming infrastructure build-up process. Costs ramp up more slowly because companies are charged per user, and don't need major hardware and software investments upfront.

Businesses that leverage Cloud Computing will be able to capitalize on bringing data together in ways not previously possible, because their data is available through standardized interfaces. This, in turn, will enable them to create more innovative applications around the use of their data. Business users with better access to relevant information can make more informed and faster business decisions. This new method of uniting data and teasing out new relationships between information will become the focus of enterprise application development, and revolutionize the way corporate knowledge, which is currently hidden inside systems, is used to assist in making important business decisions.

With an understanding of the potential of Cloud Computing established, the question arises: what are best practices to successfully impact enterprise integration?

Mediation Is a Secret Weapon

The importance of a mediation layer is not unique to Cloud Computing. Creation of a mediation layer protects the enterprise from provider changes. It's the single most critical architecture enhancement a company can make when using the Cloud, because it enables the enterprise to change on their own terms and not be dictated to by the external provider.



Continue to Consider Service Level Management

Mediation also serves a second

key benefit, that of control and policy enforcement. Adding a mediation layer provides the enterprise with visibility into how the Cloud is being used and to assert enterprise standards. Visibility is critical in determining risk and optimizing business relationships.

Finally, mediation enables monitoring the vendor's service level agreements to validate that they're delivering as promised in ways that are meaningful to the business.

Keep a Focus on Security

Understandably, security is the IT executive's biggest objection associated with Cloud Computing. All the obvious components of security with an outside vendor apply, but Cloud Computing magnifies the problem, and the fear.

To address security in the Cloud, data security must move up the stack, the same way IT value does. IT needs to focus on data-layer security, since the lower layers are beyond its control.

Companies that start moving to the Cloud will find that security has been tightly integrated with their infrastructure stack (primarily at the network layers) to the point that it becomes difficult to tease security and infrastructure apart. However, by moving to the Cloud, the fine level of control over the network is lost and security at the higher layers must take over.

It goes without saying that Cloud Computing is still immature. As a result, organizations should avoid putting information in the Cloud without having a copy elsewhere. Just like everything else, investment in a contingency plan, including testing the plan, is critical.

Be Cautious But Don't Fear Mistakes

Mistakes are important learning experiences. Since there's less upfront investment in Cloud Computing, mistakes are easily forgiven. Businesses leveraging Cloud Computing can change course because they don't have as much to "throw away." This also gives businesses some good leverage over their Cloud vendor. The situation certainly becomes more complicated once an investment is made heavily, but no more complicated than switching software vendors after a COTS purchase.

The Hidden Disease: IT Culture

If infrastructure is moved and IT becomes about application integration and delivery instead of "infrastructure babysitting," the IT organization itself will have to undergo a metamorphosis to meet this new model of enterprise computing. This transformation will require leadership, clarity, and education; without that, the results will be less than optimal.

Map Out a Strategy for Success

Having a plan is necessary to aim investments at the right target. Envision Cloud-based computing as enabling the "network" to be a single application delivery platform delineated by service interfaces between components. Data is freed from application silos and made accessible through open and well-defined interfaces.

IT can then focus on delivering an innovative synthesis of data to the end users in ways that enable them to make better decisions and have a greater impact on their business.

This ability is important because it reduces costs, and because it improves both alignment with the business and agility. Ultimately, IT will be able to contribute real and measurable results by collaborating with the business. In turn, this should help the business better understand the value of IT and how to effectively use it as a competitive weapon.

Summary

We've discussed some best practices that can be used as the foundation to a Cloud Computing strategy. To ensure the success of Cloud Computing initiatives, it's important to discriminate between the hype and the reality but also realize the barrier to entry is low. Don't worry about mistakes, but make small ones first! With this strategy as a foundation, IT executives are able to revolutionize the consumption of enterprise applications in their business, while at the same time, significantly reducing the cost barriers to entry of new ideas. In short, Cloud Computing will manifest a streamlined business, focused on meaningful results at a lower cost.

About the Author

As SOA Evangelist for the Actional products, David Bressler provides Progress' customers and field teams with the expertise and experience required to deliver an enterprise service-oriented architecture (SOA). He also helps steer technology partnerships to provide pragmatic value to customer deployments. From his 15+ year career in enterprise software and data networking, David brings to customer engagements a deep knowledge of application integration and data networking, blended with an ability to leverage emerging technology to create business value. He received his MBA in International Business from New York University Stern School of Business.

Benefit Now with Cloud-Hosted Desktops

Don't wait for stateless virtual desktops

BY JEFF FISHER

The excitement around virtual desktop infrastructure (VDI) is being tempered by the concern that VDI can't be cost-justified if it doesn't completely transform traditional desktop management – and by the belief that the only way to achieve this is with stateless desktops.

The bad news is that truly stateless desktops won't be commonly available for at least another five years. But there is good news and that is, contrary to conventional wisdom, deploying virtual desktops and achieving state separation in client operating environments are not one and the same. In fact, although they will likely cross paths, they are fundamentally different objectives and should be viewed as separate endeavors. The reality is that enterprises can benefit from VDI today without capital investment or complexity, and without a completely stateless desktop, by leveraging cloud-hosted desktops delivered via desktops as a service® (DaaS®).

The State of State Separation

When you boil it all down, the argument for stateless VDI is that if an enterprise centralizes all of its virtual desktops onto data center virtual machines, it will end up with bloated images in its data center, much higher storage costs, and IT will have to manage those images just like it manages its physical desktops. Therefore, the thinking goes, if IT doesn't transform how it manages those images, i.e., via state separation technologies, the cost of VDI will be prohibitive.

To make sure we're all on the same page, let me define stateless desktops in the way that most people in our industry view it. A stateless desktop is a client operating environment (normally Windows XP or Vista) that can be built completely on the fly. That is, when a user logs in to such an environment, all of the components necessary to present a fully functioning desktop are brought together from different systems in real time. In addition, any changes a user makes to each of these components is stored somewhere on the network so that the user could theoretically move to a different system and receive the same exact experience the next time he logs on.

The four critical components or layers that make up stateless desktops (or any desktop environment for that matter) are:

1. Client OS
2. Applications
3. User personalization
4. User data

Unfortunately, in a traditional Windows desktop environment, these elements are all tightly coupled. In fact, the only layer that most IT departments have successfully managed to isolate is the user data layer. This is accomplished fairly easily by redirecting the ubiquitous Windows "My Documents" folder to the network. However, all of the three remaining layers are bound together by the Windows registry and are much more difficult to isolate. So although state separation is a great goal and there are some third-party solutions that attack parts of the problem, they don't bring us to the level of statelessness that is required for dynamically composed desktops and, in and of themselves, they are more like band-aids, each with its own integration costs and limitations.

Why You Shouldn't Wait

We all know the economic situation is putting incredible pressure on IT to cut costs and operate with much greater agility. This is why VDI is so attractive, but also why many organizations feel they must wait for statelessness to adopt it. In fact, one could argue that in the case of enterprise VDI, since the organization is building and funding the implementation, state separation technologies such as application and profile virtualization are required

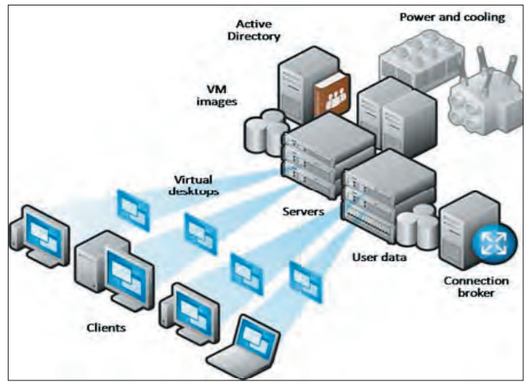


Figure 1: Enterprise VDI beckons a serious look at state separation technologies because the enterprise bears the cost burden of the desktop hosting infrastructure, including storage

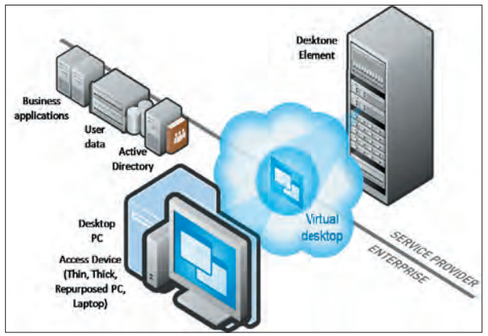


Figure 2: Cloud-hosted desktops (also known as desktops as a service or DaaS) alleviate infrastructure CAPEX for the enterprise and the associated pressure to seriously consider state separation technologies as a prerequisite for VDI

to achieve a justifiable level of CAPEX, especially with respect to storage costs.

But, why would you wait when you can get these benefits now with cloud-hosted virtual desktops delivered by service providers who have the necessary infrastructure in place and can deliver the service levels enterprises need? DaaS decouples the entire OS environment from the device layer and moves it into a data center owned and operated by a service provider. By moving instances of Windows clients into a service provider's cloud, users are able to access them from anywhere they have a network connection with just about any device, and get their own personalized desktop down to the application level.

More important, with DaaS (as opposed to enterprise VDI), the economic model changes and CAPEX for the hosting infrastructure shifts from the enterprise to the service provider. Now, the cost of the infrastructure becomes an operating expense that enterprises normally pay as a monthly subscription. Organizations can onboard their first 100 or 1,000 users, leveraging their current desktop practices while offloading the CAPEX burden to the service provider. Therefore, holding off for stateless VDI because of costs concerns is no longer necessary – you can gain significant cost savings today by outsourcing cloud-hosted desktops.

How DaaS Meets Today's Needs

DaaS allows organizations to truly leverage a model I call "VCO" for "virtualize, centralize, optimize." VCO is based on the premise that, for most IT shops (perhaps with the exception of those with extremely mature desktop management practices), simply transforming physical PCs into server-hosted virtual desktops will not only lower overall desktop TCO but will also instantly bring about other benefits (i.e., security, agility, etc.) that have been previously unachievable. To put this in the context of Microsoft's Infrastructure Optimization model, a simple server-hosted desktop virtualization implementation can immediately take the targeted desktop infrastructure from a "Basic" state to a "Standardized" one.

Service providers are well equipped to deliver DaaS in the way enterprises need and at the cost structures they require. In order to address the storage footprint challenge inherent with hosted virtual desktops, these providers are implementing highly optimized storage infrastructure, which includes advanced thin provisioning and de-duplication technologies, specifically for their DaaS offerings. This, combined with economies of scale and operating efficiencies, allows most service providers to run hosting infrastructure at a much cost lower than the average enterprise. And that is what enables organizations to get started with DaaS while realizing compelling virtual desktop economics. A case in point is Pike County School District, which has just seven IT professionals supporting 25 schools and 9,720 students and faculty that are spread across 700 square miles. Pike couldn't have afforded to implement VDI itself. A cost-savings analysis comparing what it would have taken for Pike to build and deploy 1,400 centralized virtual desktops versus what it's costing with DaaS reveals 64% savings for the TCO of each workstation.

CONTINUED ON PAGE 23

Taking .NET Development to the Cloud

The leap of faith

BY ROBERT LE MOINE

Cloud platforms, such as Microsoft Azure, offer compelling advantages for building new scalable .NET applications. But can the Cloud be used for developing existing .NET applications?

In this article, I'll explain how we've made the leap to Cloud-based development for our internal applications and the lessons we've learned along the way. Specifically, I'll describe our checklist for selecting a Cloud vendor and how we've used the virtualization capabilities of the Cloud to improve our agile development process. I'll also outline the quantifiable benefits we've seen, including saving $100,000 in capital expenditure and reducing our iteration cycle times by 25%.

All Clouds Are Not the Same

As the development team lead for Buildingi, a corporate real estate consultancy that specializes in back-office technology solutions to manage large portfolios, I'm responsible for building Web-based applications using Visual Studio, the Microsoft .NET Framework, and Silverlight. Last year we started looking at Cloud Computing to gain the advantages of a scalable, virtualized platform for software development and testing.

Most of our developers were using virtual machines (VMs) on their individual development workstations, which limited the ability to share environments across the team. In addition, giving access to customers for user acceptance testing was still a very manual process that involved taking new builds and installing them on under-utilized physical servers.

We had ruled out an onsite virtual lab solution because we estimated even a modest lab would cost well over $100,000 in hardware, storage, lab management software, and labor, which was well beyond our budget of $40,000. By moving to the cloud for a new development and test lab we hoped to avoid this large capital expense and gain the ability to:

- Quickly set up and tear down multi-machine environments
- Snapshot entire systems and store them in a central library to enable more efficient bug replication by our regionally distributed development team
- Run multiple isolated copies of the same application stack side-by-side for testing
- Enable customers to access new software builds over the Web

Microsoft Azure is a Platform-as-a-Service (similar to Google App Engine) and won't support existing multi-tier .NET applications unchanged, so we explored using Amazon EC2. EC2 is classified as Infrastructure-as-a-Service and offers Xen-based virtual machines and Cloud-based services, including storage, database, and queuing services. Essentially, this means you get Xen-based virtual machines in the cloud, can scale these machines on-demand, and only pay for what you use.

We loved the potential of EC2, but the fact that it only offers Windows 2003 and no other Windows clients (such as XP and Vista) was a non-starter for our client/server solutions. We found the user interface very limited and not suitable for use by an end customer. It also lacked workflow functionality to easily share entire application development stacks with customers or across the team.

Another requirement to consider was the ability to import VMware images of an application stack to a Cloud and run them unchanged. EC2 will only import its AMI format and this is only possible for Linux-based images. After some more research, we discovered Skytap's Virtual Lab, an emerging start-up that offers a virtual lab as a Cloud-based service and solved these requirements.

A Virtual Lab in the Cloud

Virtual lab management is a term coined to describe a new set of tools and test practices to automate labs using virtualization technology. There are vendors (such as VMware) that offer virtual lab management as an onsite package and Skytap offers these capabilities in the Cloud. Virtual labs offer some important capabilities over individual workstation virtualization solutions:

Multi-Machine Configurations

Virtual machine images are the containers that enable operating systems and applications to be isolated from physical resources. A group of virtual images that defines a complete system, including network and storage characteristics, is called a configuration. For instance, a configuration could consist of multiple Vista client machines, an Oracle database server, and a .NET application server. A configuration is a very useful concept for QA teams because it allows a whole system to be defined and isolated in a test environment. Virtual networking enables copies of the same environment to be run in parallel and the emulation of production environments during the test process. Many Cloud vendors only support single machines, so finding a solution that supports configurations is important.

Configuration Library

A configuration library allows a team to manage and organize virtual images and configurations. Standard builds and images can be created and made available to development and QA teams to save manual installation and set-up time. Additionally, the library is used to store new configurations that are cloned or created as part of a test.

Suspend, Snapshot, and Restore

The ability to suspend a complete state of a multi-machine configuration and make a snapshot (a copy at a point in time) is a major benefit of Cloud environments. This is especially useful for application development teams because when a bug is found a configuration snapshot to be taken at the point of failure and a link to the configuration added to a defect report. Instead of a developer spending hours to reproduce the defect, he or she can restore the configuration and start debugging the issue in minutes.

Test Types	Cloud Support
Unit Testing	Yes
Functional Testing	Yes
System Testing	Yes
Integration Testing	Yes
End-User Acceptance Testing	Yes
Performance Testing	Partial
Load Testing	Yes
Hardware-Specific Testing	No

Figure 1 Test types supported

Automation API

Automating a test lab almost always involves integrating tools and test processes. An automation API enables teams to automatically create test environments as part of the build process and initiate automated test runs once a new build has been deployed. An automation API is typically made available through a Web Services interface.

Administration, Reporting, and Security

Administration and security features include user and quota management, project creation, permissions, and authentication. Remote access to the system (for instance, for an outsourced vendor) is usually managed through secure connections via encrypted protocols and virtual private networking. Reporting lets users and administrators manage usage and quotas to keep project costs in budget.

Cloud Limitations

Before getting into detail on how we used Skytap to improve our agile development process, it's worth noting what types of testing the Cloud is not suitable for. Most Infrastructure-as-a-Service vendors offer standard infrastructure; however, some limit the configurability of processors, amount of memory, and network settings of machines (for instance, Amazon only offers Small, Large, and Extra Large specifications of machines). This can be an issue for specific test cases

where low-spec machines need to be created. Other types of test scenarios include unit testing, functional testing, system testing, integration testing, and load testing of applications and the Cloud is suitable for these tests assuming your Cloud provider supports your application platform.

There are a few use cases where the Cloud is not recommended for testing. These include tests that require specific x86 hardware (e.g., BIOS driver tests) and some types of performance and stress testing. If an application requires an onsite Web Service behind a firewall, this can usually be accessed using a VPN connection.

Extending Agile Methods to the Cloud

We follow an agile development philosophy and were keen to explore ways to improve our development process and quality using the Cloud. Since we made the move, we've seen numerous ways to improve our agile approach. Here's what we've learned:

Faster Iterations

The cloud is inherently self-service, which allows developers to get immediate access to development resources whenever they're needed. Activities such as buying, installing, and configuring physical machines are a thing of the past and we can now deploy a new development environment in a matter of minutes versus several hours (or days) previously.

Developers are rarely blocked because an IT operations request is taking too long and they can scale virtual machines up and down as needed. We've found that by removing these friction points in our development process we've reduced our iteration time by about 25%.

Customer Feedback

We've always tried to get customer feedback early and often in our development process. The challenge we had was providing customer access to applications easily. Typically we'd have to find a test stack (several machines), install a new build, and give a customer detailed instructions on how to RDP into these machines using a VPN connection. Usually, these machines couldn't be used for anything else while a customer was testing a build, which tied up machines and was inefficient.

Using a Cloud-based virtual lab, we can publish an entire development stack and give a customer a URL to access these machines. We only pay for the hours a customer uses the machines and can set auto-suspend rules after a period of inactivity. This not only improves the ease of getting customers involved earlier in the development process, but also cuts costs and improves flexibility.

Test Contexts

Many of our projects involve very specific test contexts, such as named domains, active directories with specific user accounts, and SQL server installations with users and data that closely mirror production. Our test cases include changing active directory entries, updating database entries, and changing configuration settings and we found setting up an isolated test environment very challenging.

Spinning up a test environment in the Cloud solves many the issues involved with replicating a complex multi-machine environment. Network fencing ensures environments are isolated, but that they still retain domain names, subnets, IP addresses, etc. We can run multiple copies of the same test environment side-by-side and roll back changes to a clean state quickly and easily. This capability has saved many tedious hours of trying to replicate test contexts and ensure test cases are adequately covered.

Team Interaction

We have a regionally distributed development team and this poses challenges with team communication. We've found many of these challenges have been solved with a Cloud-based virtual lab. We can now easily instant-message a URL to each other to collaborate in real-time on an application stack or to get advice. These "configuration URLs" point to entire multi-machine stacks and can only be accessed by approved team members.

In addition, our QA team can now snapshot entire multi-machine configurations and save them in the Cloud. This practically eliminates the "it works on my machine" complaint developers often pass back to the test team. Just like a TiVo can pause and replay TV programs, a tester can pause an application at the point of failure using the virtualization capabilities of the Cloud and shoot this over to the development team. This saves hours of trying to replicate issues for complex bugs.

Benefits

Overall, moving our development and test environments to the Cloud has been very successful for Buildingi. I'm often asked how to quantify the benefits of adopting a Cloud-based development environment. Every team is different, but analyzing total cost of ownership (TCO), time savings, and quality of code are good areas to focus on. Here are some of the specific benefits we've measured:

- Deployed the complete virtual lab solution in two days (including developer training)
- Saved $100,000 in capital expenses
- Reduced time to reproduce complex bugs by 80%
- Cut development iteration cycle time by 25%
- Improved set-up and tear-down of new test environments by 75% (two hours versus eight+ hours)
- Reduced time to deploy a customer user acceptance testing (UAT) environment from six+ hours to a few minutes

In summary, I'd definitely recommend looking into a Cloud-based lab environment to augment or replace your existing development and test infrastructure. We found Skytap worked best for us, but it's always important to complete a thorough evaluation based on your company's requirements. I'd recommend looking carefully at the capabilities of the Clouds you evaluate, especially with respect to supported operating systems, Web-based UIs, virtual lab automation, and configurable networking. It's a leap of faith moving to the cloud, but I'm confident you'll find significant benefits for typical .NET application development. Given Cloud vendors don't have any upfront charges and can be cancelled at any time, there's very little to lose and a lot to gain!

About the Author

Robert Le Moine is a Senior Consultant with Buildingi, helping people transform ideas into actual tools. With a background in both computer science and engineering, he has designed and implemented solutions ranging from complete IT Infrastructure designs to customized software applications. Fifteen years of experience in leading teams through the ever-changing IT and development worlds has provided Robert with the skills to guide his team through delivering solutions for small business as well as corporate environments.

IAAS

Building a Real-World IaaS Cloud Foundation

The perfect complement to software virtualization

BY KEN OESTREICH

I recently gave a talk at the Cloud Computing Expo in New York about where to begin if you're building a cloud infrastructure or "Infrastructure-as-a-Service." The response was great, so I'll try to summarize the high points here for others who are interested.

What Is IaaS anyhow?

Whenever we see a "stack diagram" of cloud architectures, most conversation centers on higher-level layers like defining what "Platform-as-a-Service" or "Software-as-a-Service" is. But underneath all of these diagrams is always a foundation layer called "Infrastructure-as-a-Service" (IaaS) or sometimes Hardware-as-a-Service (HaaS). IaaS doesn't get much attention, but it's really the critical factor to being able to provide reliability and basic services to all of the other layers (see Figure 1).

Why? Without a dependable, scalable, and expandable hardware (CPU, storage, networking) foundation, building virtualization layers and higher-level services are for naught. Plus, it should be noted, that certain specialized "cloud" infrastructures might not have a virtualization layer at all. So a flexible physical layer is even more crucial.

This IaaS layer is assumed to provide a number of idealized properties: It should "present" a pool of highly available CPUs (and maybe even differentiated CPU types), capacity/utilization data, chargeback data, and data needed by CMDB/compliance systems. In return, it should "consume" requirements for real-time server needs, storage needs, network needs, and SLA requirements. In this way the infrastructure's control policy can be set to provide the necessary SLAs and performance (see Figure 2).

Note that what we're talking about is very different from a virtual infrastructure – we're talking about the highly reliable and elastic physical infrastructure that's underlying any software infrastructure – whether or not it's physical or virtual.

In summary, such IaaS architecture ought to act like an idealized CPU pool with a simple API – not unlike what you see when using Amazon's EC2 or other hosted CPU services. But unless you're an Amazon, Salesforce, or another mega-provider, how do you go about building one of these?

Where the Industry "Went Wrong"

To understand how to construct a flexible/reliable IaaS, it helps to know why it's been so hard to build in the past.

It all starts with server hardware…motherboards to be exact. When the computer industry was just getting started, motherboards harbored a CPU and remedial I/O. But as processors got more sophisticated, they were integrated with complex I/O (e.g., Network Interface Cards or NICs) as well as with storage connectivity (e.g., Host Bus Adaptors or HBAs). Plus, there was usually a local disk, of course. These components all added-up to giving the motherboard the concept of "state."

This "state" meant that the server retained static data, specifically things like I/O addressing and storage connectivity naming, not to mention data on the local disk. Usually the local network had state too – ensuring that the IP and MAC address of the motherboard were attached to switches and LANs in a particular way. Add to this the fact that for critical applications, all of these components (plus naming/addressing) were frequently duplicated for redundancy.

This meant that if you had to replace (or clone) this server, say because of a failure, you had to reconfigure all of these addresses, names, storage connections, and networks – and sometimes in duplicate. This resulted in lots of things to administer to, and lots of room for error. And frankly, this is where fundamental "data center complexity" probably arose.

In response to dealing with failures and complexity, vendors developed special-purpose clustering and failover software – often necessarily closely coupled to specific software and hardware – to provide the reassignment of state to the new hardware and networking. This software often required hand-crafted integration and frequent testing to ensure that all of the addressing, I/O, and connectivity operations worked properly. And many of these special-purpose systems are what are in use today.

Similarly, there are equally complicated software packages for scale-out and grid computing, that perform similar operations – not for the purpose of failure correction, but for "cloning" hardware to scale-out systems for parallel computing, databases, etc. But these systems are equally complex and usually application-specific, again having to deal with replicating stateful computing resources.

So the industry, in an effort to add "smarts" and sophistication to the server – to enable it to failover or scale – has instead created complexity and inflexibility for itself. The question is, what could the industry have done differently, and what can we do now?

A More Elegant Approach to Infrastructure Availability & Elasticity

The logical follow-on question is: What if servers hadn't originally been designed as "stateful" entities? What if they were truly only CPUs and memory, with I/O and storage connections that could be reconfigured in software? And, what if the network "fabric" interconnecting these CPUs could be infinitely reconfigurable too? That way, network interconnections, multi-paths, virtual LANs, even things like load balancing, could also all be defined in software. That would get us closer to the ideal IaaS.

If this were possible, then the ability to re-purpose and re-assign CPUs would be massively simplified. CPUs, their compute loads, and their connections to the rest of the data center, could be easily cloned. Why? Local cloning would be the equivalent of HA, and environment cloning would be the equivalent of DR. The way that much of today's clustering and DR activities are administered would be greatly simplified with an elegant approach.

Part of the secret here is figuring a way to eliminate (or, at least, abstract away) the traditional I/O and storage naming/addressing, plus the static networking, that causes so many headaches and complications. And, there are essentially two solutions:

The first approach is a ground-up design in which the server motherboards are designed without the traditional NIC and HBA adapters at all, and where the board interconnections consist of a converged high-speed "fabric" that conveys both I/O and storage data. No Ethernet, no iSCSI, no Fibre Channel (except between the system and the "outside world"). Drivers are written for the OS so that it "thinks" it's talking to ordinary NICs and HBAs, but in reality, those entities are logical, not physical; hence they can be configured at will. Within the fabric's converged control are logical

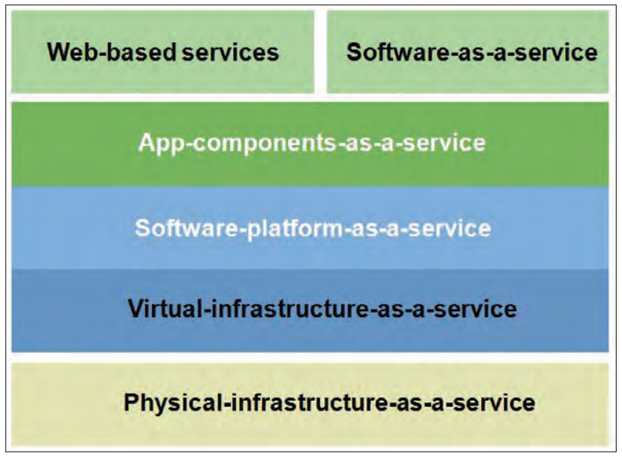


Figure 1

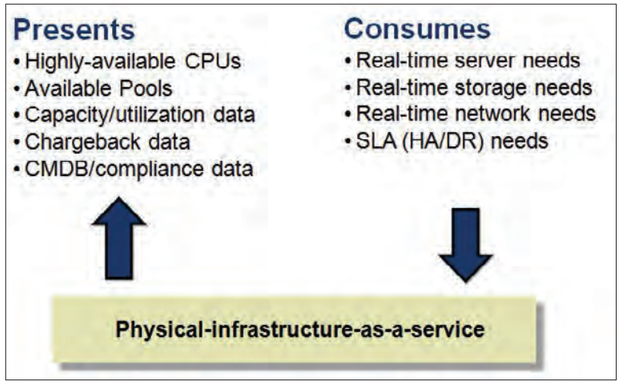


Figure 2

switches and load balancers, plus the ability to create secure single or redundant interconnections. This "purist" architecture was first brought to market by Egenera in 2001 with its BladeFrame product, which is still available today. This architecture has also more recently been brought to market by Cisco with their Unified Computing System.

The second approach leverages existing standard hardware. The same replacement device drivers are written for I/O and storage connections. However, the existing networking hardware is repurposed to provide converged I/O and storage data along a single (or dual-redundant) standard networking wire. Standard switches can be used to rout traffic in the fabric, and standard NIC and HBA cards can be used for communication between the fabric and the rest of the data center. This approach is used, for example, in a joint Dell/Egenera solution, and as a slight variant in solutions from HP and IBM.

Regardless of the approach used, it's now possible to "orchestrate" physical infrastructure in software (logically) rather than in hardware (physically). This allows IT OPs to provide those idealized IaaS services which are expected: The ability to define hardware infrastructure configurations in software, the ability to "clone" failed machines (whether they're virtual hosts or physical hosts), and the ability to "clone" entire computing/networking infrastructures (again, regardless of software, OS or VM host technology). And it's truly elegant – all this without having to worry about the minutia of special-purpose clustering, HA, or DR systems.

Engineers know that an "elegant" design is one that accomplishes a goal with the minimum of complexity and resources. And indeed, this "infrastructure orchestration" meets these criteria. Besides providing HA and DR for essentially any form of software (P or V), it also reduces the overall complexity by removing the need for multiple cables, I/O cards, and application-specific clustering. Plus, it does so in a fully flexible way. In many aspects, infrastructure orchestration simplifies for IT "plumbing" what hypervisors have accomplished in the software and OS world.

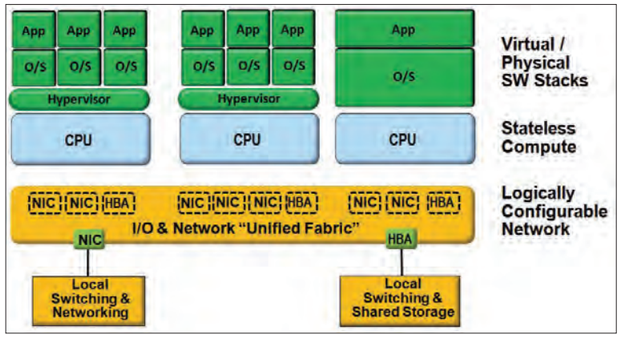


Figure 3

Operating a Real-World Infrastructure-as-a-Service

So the final question is, how does this idealized approach work as an IaaS in the real world, such as for a hosting provider, a cloud provider, or just in an "internal" or "private" utility computing environment? To do so, infrastructure orchestration mechanisms are outfitted with the following management facilities to provide the necessary foundational services:

- **Monitoring –** Which checks server and application "heartbeats," available CPUs and networks, plus other asset states. This input is fed-back both for dynamic control of the infrastructure, as well as for real-time reporting such as for compliance, CMDBs, and chargeback systems.
- **SLA Management –** Once a service (application) is provisioned – complete with an SLA – this facility monitors heartbeats, user-defined performance metrics (e.g., CPU load) and other metrics. If a user-defined threshold is breached (or, when a manual intervention occurs) this control will take over and provide appropriate advance action. This action might be simply to migrate the service to a larger (or smaller) server. Or, should it detect failure, it would provide transparent failover. And in the worst case of complete environment-wide SLA failures, it would provide a disaster-recovery style of migration to an entirely different location…once again "cloning" the original configurations.
- **Security –** In a dense, shared, and pooled environment, the security of CPUs, data, and network is paramount. So the control has to enforce tiered administrative access as well as secure multi-tenancy across all resources. For example, even "bare metal" CPU pools might remain globally unallocated, or smaller free pools might be allocated to specific users. Similarly, network security has to ensure that cross-fabric traffic is segregated appropriately.
- **Account Management & Provisioning –** The final key to the operation becomes how different users are assigned free resources, and how those users provision software to those resources. There are a number of self-serve approaches that can be taken here, and a number of commercially-available products in the space.

As mentioned earlier, this simplified form of infrastructure management for IaaS is agnostic to whether it underpins physical or virtual environments. In physical scenarios, multiple scale-out physical instances (such as Web server farms or large-scale databases) can easily be provisioned on top of the IaaS foundation – which can be managed as "elastic" if needed, cloning instances as demand warrants, or retiring instances when demand ebbs.

Similarly in virtual environments, each CPU can be automatically pre-provisioned with a virtual host – presenting an "elastic" pool of VMs when/as demand warrants. Traditional VM management tools can then transparently operate on top of the physical IaaS foundation. If new hosts are needed, the IaaS can provision as needed. And the beauty of infrastructure orchestration is that both physical and virtual environments can be simultaneously built on top of the same IaaS.

Summary

While the "mega" cloud providers such as Amazon and Salesforce.com have designed their own form of infrastructure management, this form of IaaS is available commercially from a number of vendors. Today, some of the largest commercial hosting firms, government agencies, banks, and healthcare providers are using this architecture to support both physical and virtual infrastructures – although many still may not refer to their implementation as a "cloud."

As virtual infrastructures, cloud computing, and automation begin to permeate the market, this form of infrastructure management will continue to be seen as the perfect complement to software virtualization…flexible software environment, flexible infrastructure architecture. And, as more higher-level PaaS and SaaS models establish themselves in the market, building the foundational IaaS architecture will be more prevalent than ever.

Consider that the entire IT industry will have advanced quite a bit once our complex hard-wired plumbing finally evolves into a simple and elegant IaaS foundation.

About the Author

Ken Oestreich is VP of Product Marketing with Egenera, and has spent over 20 years developing and introducing new products into new markets. Recently, he's been involved in bringing utility- and cloud-computing technologies to market. Previously, Ken was with Cassatt, and held a number of developer, alliance and strategy positions with Sun Microsystems.

koestreich@egenera.com and @Fountnhead onTwitter.

CONTINUED FROM PAGE 15

Benefit Now with Cloud-Hosted Desktops

As I stated earlier, although stateless desktops aren't a pre-requisite for cloud-hosted virtual desktop deployments, they will eventually cross paths. In keeping with Microsoft's Infrastructure Optimization model, DaaS deployments will begin in what is comparably a "Standardized" state (1:1 stateful replicas of physical desktop environments that are managed using current desktop practices).

Once enterprises get comfortable with their initial production cloud deployments, they can move along the maturity curve with state separation technologies – as they become available – to help optimize those virtual desktops, until they reach a more "Rationalized" or "Dynamic" state, where desktops can be completely composed on the fly. This will allow organizations to better take advantage of capabilities like dynamic pooling, where a single, static image is used as the source for deploying large numbers of virtual desktops. More important, it will allow enterprises to decrease the storage consumption of their cloud-hosted desktop environment, which will lower the service provider's hosting costs and ultimately result in a lower monthly cost for their virtual desktops.

The bottom line is this: the stateless desktop is a great goal, but it is not needed for VDI to be feasible or successful. If you want to begin realizing VDI's tremendous cost, security and flexibility benefits now, cloud-hosted virtual desktops as a service can get you there today..

About the Author

Jeff Fisher has 17 years of business development, marketing and technical experience in virtualization, including at Microsoft, Softricity and Citrix. He guides many elements of Desktone's corporate strategy, technology alliances, competitive positioning and market development. He worked at Wired Business and NETLAN, and has an MBA from Columbia Business School.

jeff.fisher@desktone.com

Why Assumptions about Cloud Performance Can Be Dangerous to Your Business

The response time at which services are delivered to an end user is one of the most important risks to your business

BY IMAD MOULINE

"I probably wouldn't put anything mission-critical in the cloud now." That's a recent quote from the CIO of a major consumer brand talking about his enterprise IT infrastructure. The CIO of a consumer e-commerce Web site would be equally nervous about moving its front- or back-end resources to the cloud.

As tempting as its cost savings and scalability may be, the risks of the cloud are now also coming to light. Availability and security concerns have dominated this discussion, but the performance of the cloud – the response time or speed at which services are delivered to an end user – may, in practical terms, be one of the most important risks to your business.

In this article we'll describe the performance risk inherent in the cloud. We'll explain how this applies to the various cloud services, offer data and analysis that may surprise you, and suggest a series of tough questions to ask when evaluating the performance of a cloud provider and its impact on your bottom line. Our overriding question: are you getting what you're paying for?

With broadband at critical mass, and with connection speeds continually increasing, all customers, B2C or B2B, now expect Google.com- or Amazon.com-like response times. But should we expect that Amazon's EC2 service delivers the same speedy performance one experiences when shopping at Amazon.com?

The assumption that Amazon EC2 or Google App Engine can provide levels of performance commensurate with their brand is convincing enough for the many small businesses now jumping on cloud services. However the CIO or CTO of a large complex IT infrastructure has a much longer checklist of performance considerations. Small or large, you require assurances that your cloud provider will deliver the service your business needs. But do the service-level agreements being offered today by cloud providers meet your expectations for accountability and guarantees?

Levels of Cloud Services

The cloud promises computing as utility: off-premises, on-demand, easily scalable (elastic), and paying only for what you use. The common classifications of cloud services are:

- **Infrastructure as a Service (IaaS)** – This is typically a server and storage connected to the Internet, a blank page on which to build the underlying platform and every element and application required in your infrastructure. Examples include Amazon EC2, Mosso, and 3Tera.
- **Platform as a Service (PaaS)** – Here the underlying platform is abstracted out and you're given an on-demand solution stack, a development environment on which to build your necessary applications. As the underlying platform isn't changeable, there are fewer configuration options, and therefore less flexibility than with IaaS. Examples of PaaS services include Google App Engine and Force.com from Salesforce.com (also SaaS).
- **Software as a Service (SaaS)** – This cloud option is the most mature. It provides turnkey applications on-demand, usually accessible via a Web browser. All else is hidden from view and maintained by the provider. Examples include SAP, Zoho, and Gomez.

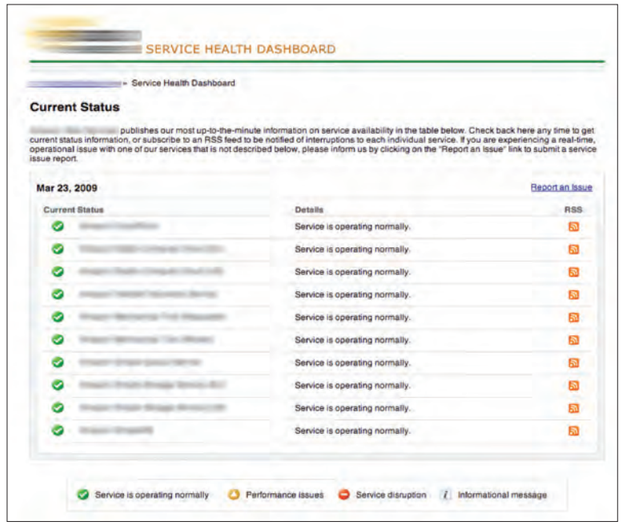


Figure 1: Example of IaaS service status dashboard

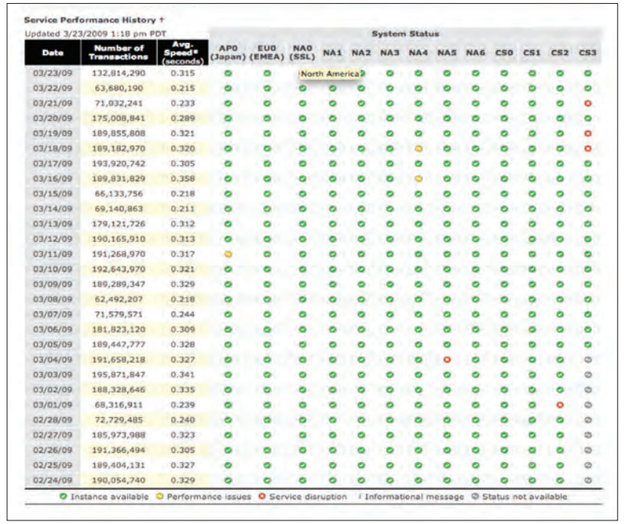


Figure 2: Example of SaaS system status dashboard

So what are your implicit and explicit expectations for each of these services?

Many cloud SLAs promise 99.99% uptime, but what does that mean to your business? Does availability mean the server is running or that applications are performing to your specifications? Is availability measured from within the cloud's

firewall or from end users' actual devices? A look at the status dashboard of a major cloud provider (Figure 1) doesn't offer this detail.

CIOs need to ensure that a cloud SLA addresses the company's specific business needs. This means every service in the delivery chain has to have someone responsible and accountable for it – just as they would in a non-cloud infrastructure with its detailed service level objectives (SLOs) from internal teams and SLAs from outside vendors. But if you're outsourcing vital portions of your infrastructure to the cloud, many of those elements are handled outside your direct control. So who is accountable if one aspect of that service performs below expectations? Watching for these potential cloud disconnects is an important part of your due diligence in evaluating Cloud services.

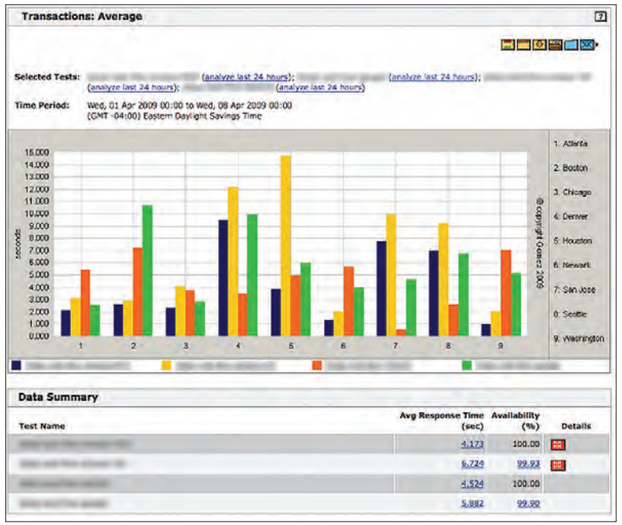


Figure 3: Geographic performance test: four major IaaS and PaaS providers (Source: Gomez, Inc.)

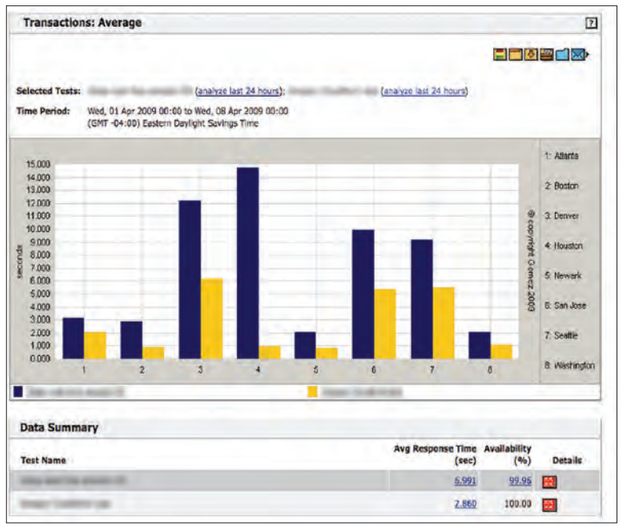


Figure 4: Geographic performance test: IaaS provider, with and without CDN (Source: Gomez, Inc.)

With IaaS, elasticity is the most promoted benefit. Elasticity equals velocity plus capacity. That means a quick ramp-up during peak customer usage periods, and only during those times. How efficiently can this occur based on your needs? If you ramp up early, cost benefit is diminished; too late and your system performance deteriorates under the increased load. Will this ramp-up be fast at all times of the day and across all geographies? And just how much capacity can you get? Will an additional hundred or 300 instances be there when you need them? If you're using the cloud for as-needed, behind-the-scenes data crunching, this isn't a concern, but it certainly is when you're serving a worldwide base of users.

Although not an explicit benefit, connectivity is certainly implied. You assume you're getting fast servers with redundant Internet connectivity in multi-homed data centers with good peering relationships to major networks nationally and internationally.

Similarly with PaaS, what are the implied performance guarantees? With Google App Engine, you assume the underlying service is performing at adequate speeds for your business. Velocity and capacity are a given. With PaaS, this happens transparently based on the number of customers using the system. But are all APIs functioning at mission-critical levels at all times or will a spike in usage slow down the underlying performance?

With SaaS, many of the same performance considerations apply. Can you be sure a transaction made in your Paris bureau is available minutes later for use by the San Francisco sales team trying to close an end-of-quarter deal? In other words, how long will it take the end user to complete his multi-step workflow, regardless of the time of day, point in the quarter, or geographic location? Under certain conditions, performance issues can quickly become availability issues as the increased load weighs on the system. A sample SaaS status dashboard (Figure 2) does provide average speed data, but these numbers may not reflect the experience of your end users from their desktop, laptop, or mobile device.

Putting the Cloud to the Test

A CIO who has optimized his existing IT infrastructure wants answers to these questions and others, as he considers migrating to the cloud. To illustrate some of these important points, we conducted a few basic Web application performance tests of well-known cloud services using Gomez's worldwide testing and monitoring network.

As the basis of our tests, we mocked up a small static Web application, a series of click-throughs and transactions a user might typically perform in the process of navigating a Web site. Any problem along the way triggers an availability error, since the goal is to complete the transaction successfully. In our nomenclature, availability is defined as the success rate of the transaction, not just the reachability of the site.

In our first test (Figure 3), we compared the performance of four major IaaS and PaaS services over a one-week timeframe. We've obscured the names of the vendors tested. Identical tests were conducted in nine large U.S. cities. Note that one major cloud provider (dark blue bar) fared very well in East Coast cities, but as you move farther west you see slower performance, especially in Denver and San Jose. This detailed performance monitoring isn't available in the cloud provider dashboards shown above.

In fact, no current cloud provider offers this level of performance monitoring, nor does it guarantee performance levels. This underscores the need to conduct your own performance measurements and, as this test shows, track them across all geographies. FYI, most cloud providers won't guarantee geographic placement of servers.

One might consider closing the geographic performance gap by engaging content delivery (CDN) services from your cloud provider. That's what we did in our next test of a major IaaS provider that also offers CDN. Measuring only that provider, we conducted tests with and without its CDN service in the same nine cities. The yellow bars (Figure 4) indicate the CDN did its job, significantly improving performance times. However the pattern of slower performance in western cities reappeared, with Denver, San Jose, and Seattle showing CDN-enabled response times of up to six seconds compared to Boston and Newark, both under one second.

We also performed measurements to compare only CDNs, testing two clouds and one traditional CDN vendor. We hosted our web pages on each service and ran tests from network backbones in every region of the U.S. The results (Figure 5) are average performance numbers, that is, the average of all

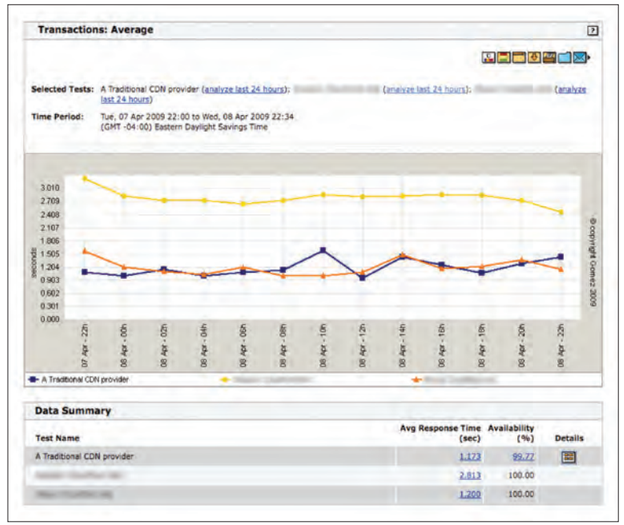


Figure 5 Performance comparison: CDN services, as tested from backbone (Source: Gomez, Inc.)

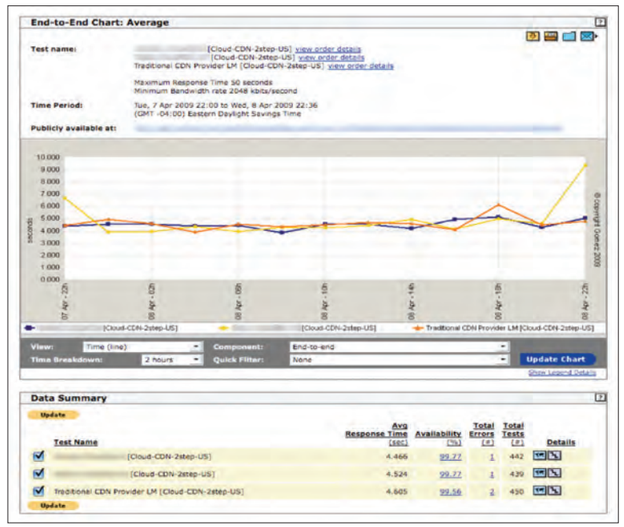


Figure 6 Performance comparison: CDN services, as tested from last mile (Source: Gomez, Inc.)

tests taken throughout a one-day test period. They showed one of the cloud CDN providers had an average response time of almost three seconds; the other two were both under one second.

Before we draw any conclusions, let's look at a parallel test of the same CDN providers, this one taken from the last mile, a network of actual home and business users with consumer-level devices and home Internet connectivity (minimum 2 mbps connection). The last mile test (Figure 6) shows almost identical performance for the three CDN providers, with response times in a tight range of 4.466 seconds to 4.605 seconds. Barely 1/15 of a second separates them. These results are quite different from the test performed from the backbone, with its data-grade connectivity and server-grade machines.

There are several lessons to be learned from this exercise. First, your customers do not live in data centers. Testing from the lab-like conditions of a backbone alone doesn't provide meaningful performance metrics. Testing from the end-user perspective is a much better approach as it mirrors the real world – your customers. Second, as accurate as these tests are, they are still average performance figures taken over time, similar to charts used to market such services. Detailed performance measurements, tuned to the specific needs and user patterns of your business, should be the basis of your assessment of Cloud services.

Demanding More from Cloud Services

In its 2008 report "Is Cloud Computing Ready for the Enterprise," Forrester Research said that cloud platforms are maturing but would not be enterprise-ready for two or three more years. Part of this maturing, we hope, will be the inclusion of detailed and relevant performance guarantees in cloud SLAs along with real-time performance monitoring by providers.

Remember too that whether your installation is simple or complex, all cloud services have one thing in common: they rely on the Internet to satisfy the needs of end users. So regardless of which provider is engaged, cloud services do not relieve IT managers of the responsibility of conducting their own ongoing performance monitoring of all Web applications delivered by the cloud.

In closing, your guidelines for measuring the performance of the cloud should always include:

1. Getting clear on what your business needs from the cloud, then testing based on those expectations, implicit and explicit,

2. Testing before deployment and continually in production, due to the constantly fluctuating nature of the cloud,

3. Knowing your end users – how they connect to the Internet, their location, times of day they logon, even which browser they use – and delivering on their expectations.

Once you've established your performance benchmarks, demand an SLA from your cloud provider that addresses all your performance needs. cloud SLAs are a work in progress and will evolve only when IT professionals demand it. Right now the client is king, as cloud providers look to fulfill the promise of utility computing, but without the risks.

About the Author

Imad Mouline is CTO of Gomez, Inc. He is a veteran of software architecture and is a recognized expert in web application design, testing and performance management. As CTO, he works with customers, analysts, industry groups and internal resources to evolve Gomez's web application experience management solutions. Prior to Gomez, Mouline was CTO at S1 Corp.

imouline@gomez.com

CommunityOne

An open developer conference

june 1–3

the moscone center

san francisco

Lightning Talks

Groovy

MySQL™

PHP

Deep Dive Tutorials

Pavilion Demos

FOSS

Cloud

Java™

Apache

RIAs

NetBeans™

Linux

Hands-on Labs

Eclipse

Virtualization

Java™

Android

OpenSolaris™

Web 2.0

GlassFish™

CloudCamp Unconference

> **"One heavyweight geek-fest."**
> – Tim Bray, 2009 speaker

Choose from more than 70 sessions on free and open-source software (FOSS) technologies and tools. Whether you're focused on creating robust Web apps, building a scalable infrastructure, or thinking about cloud computing, you'll find something to get excited about.

Register Today!

Monday is Free*

*Main conference on Monday, June 1 is free; fee required for Deep Dive tutorials on Tuesday, June 2 and Wednesday, June 3.

developers.sun.com/events/communityone

CommunityOne, 999 Skyway Road, Suite 300, San Carlos, CA 94070

Sun microsystems