O'REILLY®

Optimizing Cloud Migration

Performance Lessons for the Enterprise

Andy Still

Optimizing Cloud Migration

Performance Lessons for the Enterprise

Andy Still

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Optimizing Cloud Migration
by Andy Still

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Editor: Brian Anderson
Production Editor: Nicholas Adams
Interior Designer: David Futato
Cover Designer: Randy Comer

July 2016: First Edition

Revision History for the First Edition
2016-06-23: First Release

978-1-491-96030-1

[LSI]

Table of Contents

Optimizing Cloud Migration

Introducing the Trend: the Move to the Cloud

Cloud services are redefining how many businesses are building and hosting their applications. Flexibility, scalability, cost reduction, and reduced overheads are just some of the reasons why the case for moving to the cloud is compelling to many businesses. This is a very real trend, with a 2015 survey reporting that 72% of executives stated that the cloud was essential to their strategy, and 90% of businesses reported using the cloud in some capacity.

This move is also accompanied by a move away from server-based solutions to a world of Software as a Service-based solutions—with modern applications increasingly moving toward being jigsaw puzzles built from many different building blocks. Load balancing, file storage, databases, search, caching, authentication, data warehousing, microservices, APIs, media streaming, data processing, job queuing, and workflow are just some of the services available to build cloud-based applications. True cloud applications are fundamentally different from traditional hosted applications, not just in how they are hosted, but in the nature of how they go about solving problems to deliver resilient and flexible solutions.

The promise of the cloud, therefore, is to enable you to build a system with levels of performance and availability that wouldn't have been available to you when building an on-premise solution (at least without an investment of time and money that is beyond the scope of most companies). There are many challenges to achieving this, both practical and technological, but one area that is often overlooked is that of Internet performance.

This book will help take you on that journey—from your first foray into the cloud, to having a highly performant cloud-based system, discussing the best methods for optimizing Internet performance at each stage.

What Is Internet Performance?

Internet performance refers to the overhead of traversing the complex path of connectivity across the global Internet between the user's ISP and the entry point to your company's infrastructure. It is also sometimes referred to as the *middle mile* or *backhaul*.

Optimizing Internet performance essentially involves optimizing the route that data takes to cross the public Internet and reach your systems. This can range from understanding the routing that is in place between different locations, or serving content from different locations based on the location of the user.

Traditionally, this area of performance has been overlooked, as it is seen as being "out of our control." However, in recent years there has been a growth in understanding from organizations that this performance is a representation of their brand, and it is irrelevant to the end user whether the degradation occurs inside or outside the company's network. This has led to a growth in demand from organizations for the visibility and control necessary to improve performance of connectivity across their online infrastructure. To meet this demand, a range of tools known collectively as *Internet Performance Management (IPM)* tools have been created.

Flawed Thinking: You Can't Control Internet Performance in the Cloud

It is a mistake to think that because of the way cloud services are provided—as off-the-shelf services—you cannot take any control of Internet performance. In actual fact, the move to the cloud can potentially give you more control over the levels of Internet performance that you can deliver.

The geographically distributed nature of cloud platforms allows you more control over where you deliver content from. The possibility of using multiple clouds to dynamically serve users based on location further enhances this. However, optimizing Internet perfor-

mance requires attention, and it is easy to deliver suboptimal Internet performance if it is not addressed properly.

The following chapters will illustrate how to stay on top of this challenge when moving to the cloud and guide you through the various steps en route to delivering a highly Internet-performant cloud solution.

Phase 1: Preparing for Your Journey to the Cloud

Before you start your journey to the cloud, there are a few important mindset changes that you need to make in order to be able to take full advantage of the systems offered by cloud providers.

The Nature of Cloud Geography

Choosing a cloud provider is a difficult decision; it is a rapidly evolving industry with new offerings coming into the market on a weekly basis. There are obvious elements that should be considered when choosing a provider, such as reputation, services available, cost, and support. However, it is important that some consideration be given to the following aspects that have a major impact on Internet performance:

Geographic distribution
: Cloud providers are global platforms and portray themselves as such. In reality, however, their services are delivered from fewer than 20 geographical locations, which tend to congregate around certain areas of the world. This, of course, is more than the range of locations offered by the majority of data centers, but it is important to understand where these locations are and how that relates to the service you are wanting to provide. Performance is an important element of this alongside other considerations such as data sovereignty, redundancy, etc.

Routing

The geographical location is of course only part of the story. Cloud providers don't own a worldwide network; they rely on transit providers to connect cloud locations to markets. So, it is also essential that the cloud provider has appropriate routing in and out of the geographical location. For example, if your users are in Indonesia, a cloud provider based in Singapore would seem appropriate, but that would be undermined if upon further investigation it turned out that they routed all traffic from Indonesia via Los Angeles. While it sounds absurd, this is a genuine example of the sort of routing that can exist, and similar examples are not uncommon.

Resiliency

It is important to understand the level of resiliency that is being offered in a particular cloud region. Cloud providers will provide multiple physical data centers in a region and allow for automatic distribution of services across these data centers.

Key Concept—The Nature of Buying Has Changed

Previously, the buying process was about descriptions of the capabilities of service provision backed by service-level promises. ISPs would generally be open about the nature of connectivity they had in place and would be willing to work with you to improve that in bespoke ways if necessary.

Cloud providers are typically very reticent about sharing any details of the nature and levels of resilience they have in connectivity, focusing their SLAs on the services that they provide rather than the level of connectivity to specific markets. This is partly because it is not part of their stated services and partly because they cannot own or control the entire path to every market. This leaves the responsibility for ensuring the level and quality of connectivity with you. It is essential, therefore, that you put effective monitoring in place (see "4. Build a Comprehensive Monitoring Solution" on page 15).

Understanding the nature of the geographic distribution of cloud providers enables you to start making an informed choice about which will be the best provider for your service to minimize latency (of course, latency will only be one element considered when decid-

ing the appropriate cloud provider). A good starting point for this is to look at the region that is geographically nearest; however, that region may not actually be the best option. It is essential that you also consider the peering arrangements that the cloud provider has in place, and therefore the routing that will actually occur between your users and the cloud location.

Cloud providers often don't have the most optimized routing between end users and their systems, so it is crucial to test this as much as possible up front to select the best locations. It's even more important to continue monitoring this after the systems are in use by the public. IPM tools are core to your ability to understand the impact of cloud geography and topology on your users.

Before starting your journey to the cloud, it's important to understand what the exact nature of the cloud is.

Flawed Thinking: The Cloud Is Just Another Data Center

It is easy to think of the cloud as simply a replacement data center with on-demand virtual machines. For many people, the first instinct is to just "lift and shift" their existing infrastructure to a cloud provider. This approach often results in disillusionment with the cloud, as it results in emphasis of the negative without taking advantage of the positives that the cloud has to offer.

The real benefits of the cloud are in its dynamic nature, the ability to create and destroy infrastructure on demand, the ability to use the scalable services, the ability to create geographically distributed systems, etc. If you are just creating a fixed number of servers with installed software, then you are likely building a system that is less reliable and possibly more expensive than that provided by a traditional data center.

It is often said that servers within data centers are like pets, whereas within the cloud (or other virtualized platforms) they are like cattle (a phrase that's widely used but I think was originally coined by Randy Bias).

That is, when creating a system in a data center, you can:

- Carefully craft a system to meet your exact requirements

- Investigate the physical location and the connectivity supplied
- Define the exact hardware and configuration and apply bespoke optimizations if required
- Apply your own monitoring and negotiate access to the core infrastructure monitoring from the data center

On top of all this, you can speak to the people responsible.

In the cloud, you give up control over many of these elements. You select from a range of offerings that are predefined and build your systems on top of them. The servers become throwaway; if there are any problems or if your requirements change, they are destroyed and new ones created. To those with an on-premise mindset, this can seem very limiting. However, when exercised to full advantage, the cloud can be incredibly powerful and liberating.

Flawed Thinking: The Cloud Is Not Just Another Data Center

As flawed as it is to view the cloud as just another data center, conversely, it is just as wrong to start thinking of cloud providers as being something more than data centers.

While the way cloud providers run their operations and the nature of the services they provide are very different from a traditional data center, it is important to remember that ultimately, they *are* just data centers. When you drill down to the core, they are simply buildings full of racks and servers with connectivity to the Internet. They face exactly the same challenges as those faced by traditional data centers.

When considering Internet performance, this is an essential point to remember. Cloud providers connect to the Internet in just the same manner as any other provider. Also, like any other provider, the nature of that connectivity is driven by many factors, including practical, economic, and political ones, as well as performance. The varying levels of importance given to these considerations are of course a business decision.

Currently, Internet performance is not something that cloud providers use as a selling point; they typically sell more on price and func-

tionality, which suggests that Internet performance is not a top priority when building data centers.

Flawed Thinking: Your Applications Will All Sit On *Your* Servers

The days of applications sitting on your servers in your corporate network are ending. The days of them only using systems that you install and host are ending. Creating modern applications has become a matter of using Software as a Service and third-party services, alongside more traditional server-hosted solutions, as building blocks to build your complete applications. Your finished application may even span multiple cloud providers in addition to interacting with on-premise and other third-party systems.

Obviously, the distributed nature of systems provides additional challenges for Internet performance, and it is important that you understand some core pieces of information related to your application. You must:

1. Understand the impact of performance issues caused by connectivity issues between the different elements of the system
2. Have systems in place to react to poor performance in these interactions
3. Have monitoring in place to understand what is happening and the impact it has

Because you don't control everything, your responsibility shifts to understanding when problems are happening and then mitigating them.

Phase 1: Dos and Don'ts

Do

- Embrace the benefits of the cloud-based services that are available
- Embrace the freedom of creating and destroying services on demand

- Be aware of the different regions in which cloud services are offered and choose appropriately

Don't

- Think the cloud is the same as on-premise hosting
- Think that cloud providers don't face the same challenges as traditional data centers when it comes to optimizing connectivity
- Expect the same level of control you have over hosted applications
- Assume that cloud providers' network connectivity will be foolproof

Phase 2: Beginning Your Journey to the Cloud

When starting a migration to become a cloud-focused organization, there are four rules of good practice:

1. Start small and gradually migrate systems
2. Test, test, test—prove everything before committing to the move
3. Understand your performance expectations
4. Build a comprehensive monitoring solution

These rules apply equally when thinking only about Internet performance.

1. Start Small and Gradually Migrate Systems

Any rollout to the cloud should be completed as a gradual transition, moving the lower-risk or biggest-win areas first while having systems that communicate back to your on-premise solution.

Typically, legacy applications and data migration are the highest-risk areas, so the aim should be to create cloud-based services that mitigate their risks. For example, the first phase may be to create an API in the cloud that provides access to data from an on-premise database—cloud-based data caching services can be used to deliver data returned from the API. Typically, this could be targeted at a specific region to evaluate the Internet performance. You can then gradually extend that cloud-based data provision until it eventually ends up not needing to communicate back to the source database at all. It is

also possible to use an A/B testing approach to roll out the new system to a small percentage of users and optimize it before rolling it out to the full user base.

Starting small minimizes the risk of the move to the cloud and allows you to investigate the Internet performance at a point where it is still possible to move to an alternate provider.

2. Test, Test, Test—Prove Everything Before Committing to the Move

The nature of the cloud is that everything is throwaway, you pay for what you use, and you can scale up and down at will. This allows you to try things out, see the reality of the situation, fail fast, and then move on. The systems you test on can also be completely live-like, giving the opportunity to do some full-performance testing.

The benefits of this for functional correctness are well documented, but the impact on Internet performance is in some ways more important. It allows you to fire up systems, run tests from distributed geographical locations, and monitor the Internet performance.

The important takeaway here is that if there are issues, you can raise them with the provider. It's likely that the provider won't be able to do anything about those issues, but the nature of the engagement allows you to walk away and investigate alternatives. It also allows for an A/B type release, gradually releasing the systems to subsets of users to determine whether the testing you have done is still valid with those real users.

Having proved the concept after testing shouldn't end the testing. After migration to the cloud, ongoing testing (either explicit testing or by validating real-world performance via monitoring) should take place to ensure that the solution is still optimal. Moving a cloud system into production doesn't necessarily require a long-term commitment or mean that that platform is set in stone—if it is not meeting requirements, then it should be modified. Don't be afraid to move clouds.

3. Understand Your Performance Expectations

Unless the specific goal of the migration is to realize performance improvements, it is likely that the performance of the system after migration is only required to match the existing system's performance (though obviously any performance improvements would be gratefully received). Therefore, before migrating anything to the cloud, it is essential that you understand the nature of your application's performance as it currently exists.

The following four-stage process is good practice for identifying any application performance standards, but it applies equally when looking specifically at Internet performance:

Stage 1: Define a performance vision for the application
: This will describe the nature of good performance for your system at a conceptual level, defining which elements of performance are important to your business. If this has already been defined for the existing solution, it will remain the same for the migrated solution.

Stage 2: Understand the nature of the system
: Make sure you are aware of the nature of the system you are migrating by answering the following key questions:

  - What are the high-risk areas? What areas of the system are most prone to poor performance, or what areas are impacted the most by poor performance? For example, for an ecommerce site, a product page is an area where poor performance has a particularly large impact because it directly reduces sales. Equally, it could be that product search was a high-risk area because it was an area where performance had previously been seen to be negatively impacted by change.
  - Which areas currently have performance issues? What is the cause of those issues? Are there particular areas that have been identified as triggering poor performance? For example, it could be that high CPU usage on your database server causes poor performance, affecting the reliability of the integration to your accounts system.
  - What is the impact of poor performance on the user, the business, and on other systems? Do users stop using the system when poor performance hits, or do they get mal-

formed orders? Is the business affected by poor sales, increased complaints, or migration to competitors? Does poor performance result in other systems within the organization experiencing issues? For example, does your system running slowly impact the ability of the warehousing system to correctly process orders?

Understanding the impact allows the creation of a risk-based assessment of the importance of performance on all areas of the system.

Stage 3: Understand the characteristics of the end users

It is essential when migrating a system to the cloud that you understand the system's end users. Very important when considering Internet performance is the users' geographic location, but it is also important to understand:

- The type of people they are (as this reflects their tolerance for poor performance). For example, tech-savvy customers who have the opportunity to buy the same product elsewhere will likely be intolerant of poor performance. On the other hand, users of an internal system for non-urgent business tasks will tend to be much more tolerant of poor performance.
- The types of system and connectivity they will be using (e.g., desktop or mobile). Mobile performance is impacted by a much wider set of issues than just the server performance (e.g., network latency, poor connectivity, or device performance). It is important that you understand the varying impact of your performance on these types of users as opposed to high-speed broadband desktop users if they make up a significant element of your users.

Also remember that end users may actually be other systems (your own or third parties) that your system under migration interacts with.

Stage 4: Define acceptance criteria/KPIs for performance

Completing this analysis allows the creation of acceptance KPIs that can be used as the basis of proof-of-concept tests when evaluating cloud providers/solutions. These will then flow through into validation testing after migration is complete.

Key Concept—Think in Business Transactions, Not Individual Elements

When looking at cloud systems, you are generally looking at a complex set of interactions between different elements—some on your cloud-based servers, some using cloud-based tools, and some provided by third parties. The infrastructure, latency, and network connectivity between all these elements are generally out of your control.

When evaluating performance, it is perfectly valid to set targets for specific individual elements where possible, but it is essential that you start setting targets for the full end-to-end process to understand the impact of this dispersal of functionality to systems outside your direct control.

4. Build a Comprehensive Monitoring Solution

Cloud providers will give you access to a lot of information about their systems via extensive and configurable dashboards. These are an excellent resource and should form an integral part of your system health monitoring; however, they only tell part of the story. Cloud provider dashboards are only focused on the elements of the system that they are providing, which leaves some important gaps:

- The performance and behaviors within your application (APM)
- The experience actually being had by end users (EUM)
- The Internet performance of your application (IPM)

These gaps can leave you in a situation where all your cloud dashboards are green, but your system is unavailable. The dashboards are not designed for in-depth or worldwide reporting. It is important then that your monitoring is based on how your *application* is performing, not how the *server* is performing. There are several sets of tools that are essential for this.

Application Performance Monitoring

Application Performance Management/Monitoring (APM) tools (such as AppDynamics or New Relic) were originally created to give an understanding of what is going on "under the hood" of your

application, analyzing every application request down to the method call or SQL query level. APM tooling is also important for understanding the impact of third-party dependencies on the performance of your application.

User Experience Monitoring

The objective of this type of monitoring is to reflect what your user is actually seeing. There are two models for this type of monitoring: Real User Monitoring (RUM) and End User Monitoring (EUM).

RUM gathers data from all user activity and passes that data back to a central collection server (typically by injecting a snippet of JavaScript into every page), which allows for analysis of your users' exact experience. This will flag any unexpected behavior and can help you drill down to identify the cause of the problem. RUM is also useful for determining whether there is a pattern to the types of users who are experiencing a particular problem.

EUM is similar, but relies on synthetically generated, regularly repeated tests of specific functionality. EUM will quickly show if tasks are varying over time and whether key functionality is still acting as expected.

A good monitoring solution will combine elements of both models.

End-to-End Transaction Monitoring

In recent years, APM products have shifted their focus to also look at the end-to-end breakdown of a user request, giving an understanding of what the user has experienced within the browser (including the performance of client-side scripts) and allowing the tracing of that same request right through your application. This incorporates APM and EUM in a single solution.

This is a very valuable set of tooling, providing a deep view into end-to-end performance on an aggregated basis (by type, location, or technology of user). These tools are also able to drill down into specific outliers to determine the cause of issues.

Network Performance Monitoring

While RUM and EUM give you a good understanding of what the end user is experiencing and APM illustrates what's going on on

your server, network performance monitoring (NPM) looks at the areas in between (though only within your infrastructure, not the public internet).

In traditional data centers, this would involve operational management tools such as Nagios, or NPM tools such as Zabbix or SolarWinds to see details of how your network infrastructure is behaving. (It's worth noting that these two types of tools are increasingly overlapping.)

The network infrastructure is largely hidden from you in cloud environments, but NPM is still an important tool if you are using a hybrid cloud approach that combines cloud services with on-premise applications.

Internet Performance Management

Despite the standard monitoring tools described so far, there still remains a visibility gap, even when providing end-to-end monitoring, as it has limited insight into the Internet performance of your system. This is where IPM tooling can be valuable.

IPM tooling (such as Dyn Internet Intelligence) is designed to give you insight into the behavior of the connectivity between your users and cloud providers—the middle mile or backhaul. This is shown in terms of both availability and performance, which allows you to determine whether users are being impacted by connectivity or routing issues with a cloud provider and take action to mitigate it.

Key Concept—Monitoring Must Become Dynamic

In a traditional environment, adding a server to a monitoring solution was often a manual process included as part of the rollout.

In the cloud world, you are living in a dynamic, ever-changing environment where servers and other services may well be added and removed at any time. It is important, therefore, that all monitoring solutions you use reflect this and are able to dynamically pick up and remove elements, either automatically or via an API.

Phase 2: Dos and Don'ts

Do

- Stage your migration
- Test and POC everything first; take advantage of the throwaway nature of the cloud and stage the release using an A/B testing-type approach
- Consider performance part of testing
- Test on an environment that's as live-like as possible
- Take advantage of cloud systems to mitigate risk
- Continue to test and validate systems into production
- Analyze the routing seen by real users using analytics software
- Understand who and where your users are
- Define some KPIs/acceptance criteria before starting POCs or migrations
- Think in terms of end-to-end transactions, not individual elements
- Think beyond user interactions; include back-office integrations
- Have monitoring in place to identify poor performance and mitigations to handle it
- Have monitoring in place to identify Internet performance problems as they arise
- Look at how the application is performing, not the server
- Use APM and IPM tooling to get an end-to-end understanding

Don't

- Set a target of improved performance where that is not the objective of the migration
- Rely on cloud providers' dashboards
- Assume all users have the same tolerance and expectations of performance

Phase 3: Enhancing Your Cloud Solution

Having started your migration to the cloud, there are a set of considerations that will enable you to start taking your cloud-based systems to the next level.

Design for Failure at the Network as well as Application Layers

It is a mantra that is as old as the cloud itself: the cloud doesn't guarantee success, but it does give you the tools to deal with failure. The ability to dynamically create infrastructure on demand removes the dependency on hardware that is characteristic of data centers.

Everything you do in the cloud should assume failure will happen.

This is now a common practice for server infrastructure. The most famous example is Netflix's Chaos Monkey: a tool that goes around intentionally disabling elements of the infrastructure to ensure that their resiliency systems can cope. This is rarely done at the network level, but the same rules can apply.

If your system suffers Internet performance problems, such as routing issues that add overhead onto every request, then your system should be aware of this and be able to easily switch to another location. For example, if you normally serve content from Virginia because the majority of your requests come from users in Chicago, then in the event of Internet performance issues, it should be easy to switch to serving content from another relevant location, such as San Francisco. This allows you to respond not only to Internet per-

formance issues but also to outages seen at specific data centers. Alternatively, rather than switching location—as this is not always possible due to practical issues (e.g., data)—the system could also be configured to move into a lower-bandwidth/minimized-service interaction state to minimize the impact.

Understand the Cost of Performance and Monitoring as a Core Part of Capacity Planning

The cloud allows you to provide all the systems needed to deliver a scalable system, but those systems do not come for free. Anyone who has used cloud-based services will tell you that it is very easy to run up much higher bills than expected. However, this can be mitigated by intelligent system design.

Key Concept—Capacity Planning Has Changed

Capacity planning used to be about understanding the capacity of your systems and ensuring that there was always sufficient headroom to allow for anticipated short/medium-term growth. In this model, a server hitting capacity was a negative position that indicated that capacity planning was failing and that the system may soon fail.

In the cloud world, this is reversed, and the objective should be to have a system that is always operating close to capacity. Resources are so easy to scale that scaling them ahead of time is generally a waste of money.

Adding complexity will add cost—not only in terms of cloud costs, but also in terms of development and maintenance overhead. It is essential that you consider the following:

Level of usage
: Scale systems only to the level of usage that you anticipate. There is no need to future-proof systems. You build systems that can scale, not that are at a capacity to meet any anticipated future demand. Good system architecture is essential here and, like other things, cloud-based system architecture is different from on-premise system architecture. As a general rule, the aim should be to use cloud-based services where possible, as they

are prebuilt to be scalable with no input from you, and some are also built to be region-independent. Where you are building upon virtual machines, the aim should be for them to be horizontally scalable, meaning you can add and remove servers when desired with no impact on users.

Where your users are coming from
: Only scale systems to meet demand in areas where you have a user base that warrants the additional cost and effort. Building and maintaining a multiregion system is a complex task, particularly when it comes to data management, so it is not something that should be entered into lightly. Before committing, use your monitoring to determine if there is sufficient demand from the region and, more importantly, what the impact is on users of the configuration that you have in place.

When your users are coming
: The nature of cloud systems, with their "pay as you use" charging method and on-demand creation and destruction of resources, means that you can scale your system up and down as needed. It is therefore best practice to analyze when your systems are busy and scale up to meet demand and back down again afterwards. This can be on a daily, hourly, or even minute-by-minute basis.

How tolerant your users are
: With an intelligent set of monitoring tools, you can determine how tolerant your users are of performance issues. For example, you may determine that users in Australia see performance that is notably worse than that seen by users in other areas of the world, which could trigger a need to invest in expanding to cloud providers with better Internet performance for Australian users. However, before making such an investment, it is a good idea to understand the impact that poor performance is having on those users. There are a couple of ways to investigate this: you could analyze the performance of your competitors to see how well you compare in that area, or alternatively, you could change performance and assess the impact. Improving performance is typically a complex task, so one option is to consciously reduce performance on your system to see the business impact. This may seem like an unusual suggestion, and it may be hard to sell within your business, but, while obviously not foolproof, it can be a quick and effective method of determining

the value of investing a lot of time and effort in performance improvements.

Combining all these factors, you can construct a system that is scaled to meet the optimal delivery to users while minimizing cost and complexity. However, like everything else, the cost of building and maintaining this system must be included when considering the cost optimization. In other words, don't spend six months of time building a system that will save the equivalent of one month of time in reduced cloud costs.

Key Concept—How Cloud Providers Sell Networking

When dealing with data centers, network provision is usually sold as a pipe into your environment. This pipe typically has a limit but with an element of bursting available. Beyond this level, throughput is usually throttled. Each machine in the environment will then have a networking card, which will have a limit to the amount of throughput it can handle. The level of this throughput will be defined as part of the hardware definition of the machine.

Cloud providers, however, sell networking differently. They usually offer unlimited throughput into your environment, charged by the byte (or GB). Therefore, the limiting factor is now the infrastructure that you run your application on, rather than the pipe into your environment.

However, cloud providers are often limited in the information they provide about the networking capabilities of their machines, and will give high-level views such as S, M, or L, with the size determining the size of the elements within the machine (memory, CPU, etc.), as well as the level of networking that it provides. Therefore, high-throughput but low-CPU/memory systems such as load balancers can often end up having to be run on much beefier machines than expected to stop hitting the networking limits of the device. This can have a sizable impact on cost when sizing a new system.

However, as you look to optimize your cloud service, there are a few common mistakes that can undermine your efforts.

Flawed Thinking: Moving to the Cloud Means You Don't Need an Ops Team

A common misconception is that a move to the cloud can be accompanied by a reduction in the levels of ops support that is needed for production systems. This is completely untrue. Cloud-based systems need as much in-house expertise as any other hosted systems. Instead, it's the nature of the job that is changing. Managing a cloud-based system is as complex as managing an on-premise system. It requires a high level of specialized knowledge and understanding of the implementation, as well as industry knowledge in both traditional networking and cloud systems. Cloud systems do not look after themselves, but they do provide a new paradigm in how to build, monitor, and maintain these systems. This requires as much expertise and management as on-premise systems.

Because little pre-emptive optimization of the core infrastructure can be done, the focus becomes more on building fault tolerance into systems, building comprehensive monitoring solutions, and the ability to react quickly to situations to take advantage of the scaling and geographical options offered by cloud-based services. The skill-sets of your ops team will need to evolve to meet these new challenges.

Flawed Thinking: Third Parties are Optimized for You

Modern systems are not only dealing with incoming systems, but they are also routing requests out to other remote systems, often over the public Internet. This could be to third-party services or other services within the organization. When assessing the Internet performance of systems after migration to the cloud, it is essential that you also consider the performance of communications with these dependencies.

Key Concept—Direct Connections Are Available

Although the default nature of cloud providers is to communicate over the public Internet, many of them offer the option to introduce a direct point-to-point connection into their infrastructure from an external point, typically your back office or other data cen-

ter. In essence, this is a leased-line connection into your cloud environment. This has several advantages over traversing the public Internet:

- More consistent network connection
- Reduced bandwidth costs
- Increased security

Phase 3: Dos and Don'ts

Do

- Assume that components could fail at any point—this includes network connectivity
- Have contingency plans in place to deal with networking issues
- Consider costs when building any solution
- Build a system that can scale, not one that is already scaled
- Understand your users when planning where and how to scale your system
- Aim to build a system that is always operating close to capacity
- Realize that cloud systems require specialist knowledge to manage them
- Realize that the nature of the work will change
- Ensure that you understand the impact of poor performance of third-party systems
- Remember to assess the performance of dependent applications
- Consider installing dedicated connections to external systems

Don't

- Feel that any failover process has to be a complex automated process—a tested and documented manual process can be equally valid
- Assume that after moving to the cloud the ops overhead will be reduced
- Try to build a system that is sized to be future proof

Phase 4: Maximizing Your Internet Performance: Building a Multicloud Solution

As we have detailed previously, not all cloud vendors are created equal and, despite what they may like you to think, it is not necessary to tie yourself to a single provider. It is a perfectly valid solution to build your system as a jigsaw puzzle of components from multiple cloud providers. Certain cloud providers may provide services that other cloud providers do not—cloud systems are generally designed to be modular and use standard, open formats for communication. In this case, the same precautions should be taken as defined previously.

However, when thinking about Internet performance, the primary advantage of using multiple cloud providers is the locations each one offers relative to your users. It is generally good practice to move your systems as close to users as possible. (As mentioned previously, this does not always directly correlate to physical location but to closest in network hops.)

The monitoring systems defined previously, combined with the discussed performance and cost-optimization metrics, will allow you to determine when it is appropriate to consider moving to multiple cloud providers.

Resilience

Using multiple cloud providers also allows you to have a failover system in place in the event of a major availability issue affecting one of your cloud providers. This solution is dependent on having a fully dynamic DNS provision that allows for very low TTLs on their domain names. (TTL—Time To Live—is the amount of time that a domain name resolution will be cached before it is requeried. Essentially, this will reflect the amount of time until a change in a DNS record will take effect for a user.) This allows for any change made at DNS level to be very quickly propagated to the wider Internet. The solution can be implemented manually when the situation is observed, though some DNS providers will provide systems as part of their solution to automate this failover.

Key Concept—Uncouple Your DNS from Your Provider

To take advantage of using multiple cloud providers, you need a good dynamic DNS provider. Many cloud providers provide DNS services, including anycast facilities; however, this only relates to their own services. If you want to span multiple cloud providers, it would be better to use an independent DNS provider. Good dynamic DNS providers will also allow for failover (either manually or automatically) at a DNS level.

Flawed Thinking: Multicloud Has to Be Complex and Expensive

While a multicloud approach does add additional complexity into your systems from a technical point of view (different technology stacks, different services to support, separate deployment methodologies, multiple test processes, etc.) and also from a practical point of view (multiple management systems, financial arrangements, support processes, etc.), it does not mean that you can't design an intelligent system to minimize this complexity and optimize your infrastructure spending, thereby ensuring that your cost/performance ratio is in line with your business objectives.

Phase 4: Dos and Don'ts

Do

- Ensure your DNS provider allows reducing of TTLs to very low values and that you validate the performance of your DNS provision when TTLs are very low
- Consider using multiple cloud providers
- Keep your DNS provision independent of your cloud provider if you want to support multiple providers
- Ensure your DNS provider provides an anycast network to allow geographic optimization of traffic

Don't

- Overlook or overestimate the additional complexity and overhead in managing multiple cloud providers

Conclusion

The journey to the cloud is not an easy one, but if done well, it can have many benefits. However, without the correct precautions and visibility, you can easily end up with suboptimal Internet performance or infrastructure spend.

To maximize the Internet performance of your cloud-based systems, it is essential that the following elements be considered:

- Understand the nature of the cloud and how it differs from traditional data centers, and also how the two are similar
- Start small and gradually migrate systems
- Complete extensive testing, including for performance on live-like systems, then continue that testing after systems move into production
- Build a comprehensive end-to-end monitoring system, including an element of IPM (Internet Performance Management)
- Understand the nature of your users and their performance expectations
- Only scale when the cost/performance ratio meets your business objectives
- Build system mitigations for poor performance into your systems
- Consider using a multiple-cloud solution to optimize Internet performance

About the Author

Andy Still has worked in the web industry since 1998, leading development on some of the highest traffic sites in the UK. He cofounded Intechnica, a vendor-independent IT performance consultancy, to focus on helping companies improve performance on their IT systems, particularly websites. Andy is one of the organizers of the Web Performance Group North UK and Amazon Web Services NW UK User Group.