Approaching Cloud Computing Performance

David S. Linthicum
Deloitte Consulting

Performance in a cloud computing environment is a complex topic. The way that IaaS and PaaS clouds operate is very different from traditional single-tenant on-premises platforms. Therefore, the way you evaluate and test IaaS and PaaS clouds is different as well.

Teradata Universe EMEA 2018 recently stated: "A majority of the largest companies in the world (83 percent) agree that the cloud is the best place to run analytics, according to a new survey by Vanson Bourne on behalf of Teradata (NYSE: TDC), the leading cloud-based data and analytics company. In the next five years, by the year 2023, most organizations want to run all of their analytics in the cloud. But, an overwhelming 91 percent say that analytics should be moving to the public cloud at a faster rate."[1]

So, what's wrong with the survey numbers? "According to the survey, some of the biggest barriers to moving analytics to the cloud are security (50 percent), *immature and low-performing available technology (49 percent),*[emphasis added] regulatory compliance (35 percent) and lack of trust (32 percent)." [1]

While performance is typically included with a list of other concerns, such as security, privacy, governance, and compliance, it's coming up more and more as enterprises move to the public cloud with mixed results.

Most enterprises that leverage public cloud platforms believe that elasticity and performance come along for the ride. However, as systems get larger and require more resources, it has become obvious that public clouds have specific performance bottlenecks that must be understood and dealt with.

As more benchmarks are published, we'll see more instances where public cloud providers have systemic performance issues that are not easy to fix. While many will toss faster hardware at the problem, the culprit is typically the way the cloud is designed. Or, more often, how the workloads themselves are designed and deployed.

EMERGING CASE STUDIES

The approaches to dealing with tenant management vary greatly from cloud-to-cloud, and have different performance and scaling characteristics. Also, some cloud providers have widely distributed datacenters to reduce the number of hops it takes to get to them, and thus reduce latency.

Copublished by the IEEE CS and IEEE ComSoc
2325-6095/18/$33.00 USD

Other providers are more centralized and may have latency issues for clients who are geographically far away from the datacenter.

Performance issues, such as the latency problems, can only truly be solved with a complete public cloud restoration project, which is unlikely to happen. That means we'll probably see a few of the public cloud providers faceplant due to a systemic performance problem that they just can't solve in time to capture the market. This is the same pattern repeating itself that appeared within other emerging technologies over the years. Cloud technology is just the latest repeat.

So, what are those in enterprise IT to do when considering cloud-computing performance? There are some basic performance concepts that you should consider as you select public cloud providers, and as you migrate data and processes to the cloud.

It's interesting that almost 50 percent of applications that migrate to the cloud are reported as performing worse or the same when on a better, more "muscled up" public cloud platform. The amount of memory you select or the number of cores that are part of your configuration have little to very random effects. This, according to my experience as a practitioner in the space for the last 10 years.

These performance issues may seem to randomly crop up, but we're actually seeing a few patterns start to emerge:

- **Failure to leverage cloud-native features.** Applications try to find their best path through the platforms hosted in the cloud. In many cases, what the application requests and what the cloud platform provides don't match up, in terms of optimizing performance. For instance, I/O systems have more layers on the cloud platform, and thus the response will lag the on-premises I/O system.
- **Unwillingness to make applications cloud-native.** Most enterprises opt for the lift-and-shift option when it comes time for application migration. It's cheap, fast, and low risk. However, the end results could be less than desirable, such as lower performing application workloads, or higher running costs in the cloud because an application wasn't optimized to leverage public cloud resources.

There are other macro pattern problems to remember as well, such as distance between servers and consumers, poor application design, and the public cloud and its approach to tenant management.

NETWORK LATENCY

Generally speaking, the distance between you and your cloud center will determine the amount of latency you'll experience. You should select a provider with a regional data center that's fairly close to the people or systems that will consume the cloud service. You should test the connection as well, understanding that shorter distances don't always guarantee better network response.

Most public cloud computing providers allow you to lock systems into regional datacenters. That means the providers typically won't move your data outside of that datacenter unless they're dealing with a business continuity situation. This should reduce the latency you'll experience over time, but will also increase costs since there is usually an up-charge for this service.

However, network latency is not the only network-related performance issue. In many instances, network latency can occur within the enterprises itself. Older networking equipment, as well as lack of firmware and software updates, can mean that the latency is self-inflicted.

The best path here is to install and test networking performance management tools, which should show you where the issues are, inside and outside of the firewall. In many instances, the use of cloud just puts a spotlight on existing networking problems, and does not cause the problems themselves.

POOR DESIGN

As we discussed above, many of those tasked with migrating systems from the local datacenter to public clouds consider this an A-to-A port with few modifications required. While the system may run fine, if it's not localized to leverage the performance features of the host cloud, you're not getting the best performance bang for your cloud buck.

Moreover, public cloud-hosted performance should be considered in the design of your system, including how to deal with resource provisioning and deprovisioning, platform optimization, etc. Most public cloud providers get paid regardless of whether or not your system is optimized. However, you'll pay larger bills and suffer poor performance. Those issues will diminish with a bit of good system design that leverages the native features of the public cloud.

Some emerging best practices include:

- Understand the best approaches to leverage platforms on your public cloud provider. In many instances, they have special tricks that you can add to your code that will provide noticeable savings in performance, stability, and cloud usage costs.
- Performance is not just about use of the cloud-native APIs, but also the design of the application as a whole. In many instances, legacy applications are poorly designed. While they got by on-premises operating on single tenant platforms, in the cloud they are outed for being poorly designed. You need to consider the best approach to cloud application design. As you remediate the existing on-premises application designs into cloud-native versions, you need to consider a better holistic design as well.

UNDERSTAND THE APPROACH TO TENANT MANAGEMENT

Finally, you should understand the tenant management features of the public cloud provider, and work around any quirks. All clouds do not approach tenant management in the same way. It's helpful to understand how the cloud provider deals with multitenancy at an architecture level so you can optimize your use of the public cloud resource.

For instance, consider the approach to multiplexing I/O, such as reading and writing to storage. Or, the approach to dealing with queueing, database reads and writes, and prioritization for the CPU, memory, and storage.

Guess what? The public cloud provider is unlikely to provide you with this information. Instead you have to depend upon other users to help you understand what the tenant management approach is, and how to change applications to take full advantage.

The idea is to alter your approach to when, how, and the size of the chunks of resources you provision to support your system processing. For example, you may find that provisioning smaller amounts of resources at a time does not task the tenant management system as much as allocating one large chunk. The idea is to understand the quirks of the cloud, and work around them to optimize performance.

PERFORMANCE PLANNING

As we become more experienced with public clouds, the ability to manage performance should improve. Cloud providers should also learn and evolve their cloud computing offerings over time to provide better performance and scalability. However, for the next few years, cloud computing performance management will be a necessary skill that most enterprises will need.

The bottom line is that if you're not willing to plan for performance, then you'll probably have poorly performing public cloud workloads. More likely, you'll have huge cloud bills that get bigger every month.

Creating a performance plan requires a few core steps, including:

- Defining the performance expectations of the business, include response time, uptime, recovery time, and costs that the users are willing to pay.
- Defining the multitenant approach that your pubic cloud provider leverages, which includes approaches to I/O management, etc. Create a sub-plan that focuses on how to leverage the cloud multitenant approach so the workloads are optimized.
- Defining how the applications and data stores should be modified to take advantage of the public cloud you picked. This will typically look like T-shirt sizes—small, medium, large, and extra-large—in terms of the level of effort and cost.
- Defining a performance management plan, including tools, talent, and processes for checking that cloud performance is meeting expectations.

While this seems complex, it's really something we go through each time we change platform technology, which we've done many time in the past. This time, however, there are many more moving parts. These parts need to align properly, or performance will be an ongoing issue.

REFERENCES

1. "Survey: Companies are Bullish on Cloud Analytics, But Need to Speed Up the Pace," Teradata, 2018; https://www.teradata.com/Press-Releases/2018/Survey-Companies-are-Bullish-on-Cloud-Analyt.

ABOUT THE AUTHOR

David S. Linthicum is the Chief Cloud Strategy Officer at Deloitte Consulting, and was just named the #1 cloud influencer via a recent major report by Apollo Research. He is a cloud computing thought leader, executive, consultant, author, and speaker. Linthicum has been a CTO five times for both public and private companies, and a CEO two times in the last 25 years. Contact him at david@davidlinthicum.com.