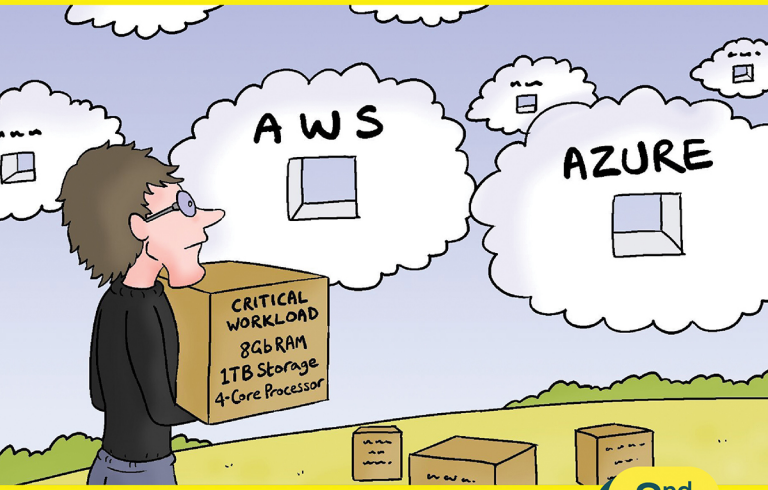
ConversationalGeek

Conversational
Workload Optimization in the Cloud

Nick Cavalancia (Microsoft MVP & Co-founder of Conversational Geek)



Learn about:

- Why the cloud doesn't guarantee workload performance
- How to go about optimizing workloads to reduce cost while improving performance

2nd MINI Edition

Sponsored by

Quest

Sponsored by Quest

Quest provides software solutions for the rapidly changing world of enterprise IT. We help simplify the challenges caused by data explosion, cloud expansion, hybrid data centers, security threats and regulatory requirements. Our portfolio includes solutions for database management, data protection, unified endpoint management, identity and access management and Microsoft platform management.

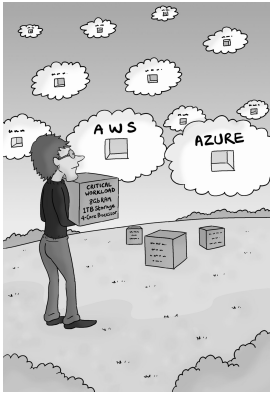
Quest's Foglight® Evolve takes a holistic and proactive approach to hybrid cloud management so you can simplify the complexity of your data center, reduce infrastructure costs, maximize system performance, and predict future costs with more accuracy.

Quest

For more information, visit
www.quest.com/Foglight-Evolve

Conversational Workload Optimization in the Cloud (Mini Edition)

by Nick Cavalancia



ConversationalGeek®

Conversational Workload Optimization in the Cloud (Mini Edition)

Published by Conversational Geek® Inc.

www.ConversationalGeek.com

Trademarks

Warning and Disclaimer

Additional Information

For general information on our other products and services, or how to create a custom Conversational Geek book for your business or organization, please visit our website at www.ConversationalGeek.com.

Publisher Acknowledgments

All of the folks responsible for the creation of this book:

Author:	Nick Cavalancia
Project Editor:	Pete Roythorne
Copy Editor:	Pete Roythorne
Content Reviewer(s):	Tim Fritz
	Gillian Ryan
	Andrea Fong

The “Conversational” Method

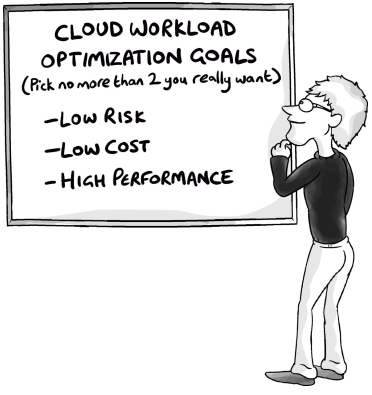
We have two objectives when we create a “Conversational” book. First, to make sure it’s written in a conversational tone so that it’s fun and easy to read. Second, to make sure you, the reader, can immediately take what you read and include it into your own conversations (personal or business-focused) with confidence.

“Geek in the Mirror” Boxes

We infuse humor and insight into our books through both cartoons and light banter from the author. When you see one of these boxes it’s the author stepping outside the dialog to speak directly to you. It might be an anecdote; it might be a personal experience.

Within these boxes I can share just about anything on the subject at hand. Read ’em!

Your Cloud Workloads Need to be Optimized



It's practically a given today that organizations are shifting operations to the cloud; the current expectations put upon businesses demand the improved accessibility, and availability found more easily in the cloud. According to LogicMonitor's *Cloud 2025* report, three-quarters of IT decision makers believe that 95% of all workloads will run in the cloud within 5 years.

Additionally, the 2020 pandemic had a significant influence on plans to move to the cloud, as 87% of those same IT decision makers agree the pandemic has accelerated their migration to the cloud.

But, once in the cloud, workloads need to offer the same – if not better – performance as when they were on-premises. The cost-effective assumption of the cloud must be founded on getting *more performance for your money* and not just *spending less*.

So, what's standing in the way of you achieving optimized use of your cloud workloads?

The Challenge of Workloads in the Cloud

In principle, the idea of running workloads in the cloud sounds like a piece of cake – just replicate a VM there and let it run all happy, fast, and efficient. However, it's not that simple; your goal is to not just push VMs into the cloud, but to do so in a way that optimizes the workload *and* does so cost-effectively.

When thinking about placing all your workloads into the cloud, there are some very real challenges you should consider:

1. Resources vs. Rates – The first thing you'll notice is that your workloads need very specific and tangible resources (think RAM, processors/cores, storage, etc.). But when you look at your cloud options, you have rate plans with bundles of resources that you need to somehow fit into, which can lead to either underusage or wasted resources within a given plan.

2. Wasteful Consumption – You're likely not using all the resources you have on premises, which further complicates the selection of cloud plans. A good example is a storage device connected to a server or on-premises VM that is not in use.

3. Inefficient Consumption – Some workloads aren't needed 100% of the time (DevOps

and Q&A workloads are great examples). If you don't need a VM on during the night, why not spin it down? The point here is these kinds of VMs only elevate your resource numbers, pushing you to spend more than you need to on the cloud.

4. Sprawl is Inevitable – In addition to what's being moved to the cloud, there will be additional VMs spun up and forgotten, orphaned resources, etc. All this sprawl consumes precious budget, impacting your ability to continue to move on-prem workloads to the cloud.

because of these challenges, you have to goals when moving workloads into the cloud: First, they must be optimized to ensure you know they are running at peak performance. Secondly, they also need to run cost-effectively to ensure your plans to migrate can continue.

Balancing Workloads & Clouds

There's this cloud nirvana that's possible – one where you have exactly the right workloads, in exactly the right cloud, using exactly the right plans, costing exactly the right amount of money. But, getting there involves you diving into workload optimization from a few perspectives:

- Performance – What kind of performance does a given workload *really* need? Does it need to perform at peak performance 100% of the time, or can it be at less than peak (say at 90-95%) for part of the day?

- Cost – You can't decouple the cost of running a workload in the cloud the way you can in your own data center. In your data center, you've already paid for it, so money just shifts within the business. But in the cloud, it's real money that is transferred *outside* the business, so performance is always a matter of cost.

- Plans – I've mentioned this already, but how closely do the plans meet your hardware requirements? Take AWS's on-demand EC2 instances for example, a *t3a.2xlarge* instance has 8 vCPUs and 32GBs of RAM. The next lower instance is 4 vCPUs and 16GB of RAM. What happens if you need 8 vCPUs and only 8GBs of RAM? Most vendors have commoditized these plans, forcing you to "fit" your resource needs into whatever they offer.

- Storage Tiers – There are a ton of storage options. It used to just be "hot, warm, and cold", but now there are many more choices; AWS has eight currently, for example. The general rule is (and this isn't 100% accurate, so take it with a grain of salt): *the lower the cost, the longer the minimum retention, slower the retrieval, and the more costly the egress fees.*

- Cloud Placement – I saved the most important for last. When you consider the previous four optimization perspectives, it's important to realize that the answer might not be to place everything in, say, AWS. Some cloud providers may handle certain kinds of workloads better than others. So, it's important to ask the question which cloud provider addresses each of your optimization concerns best for a given workload.

- Levels of Service – Organizations are focusing more today on not just is a workload performing, but are users having a good experience. The idea of an XLA (eXperience Level Agreement) is gaining ground and has its roots firmly planted in the idea that the workload providing the service is optimized to begin with.

So, what's the right way to go about optimizing your cloud workloads?

Optimizing Your Workloads

The work of optimizing a workload comes down to three simple factors – of which you can usually only choose two: *risk*, *cost*, and *performance*. You need to find a balance among these three factors, both within the context of a specific workload and across all your workloads.

What to Optimize

Let's do this by looking at things from the perspective of what you'd normally optimize if you were running workloads on-premises. In an on-premises VMware environment, for example, you can optimize on the following:

- Individual Resources – Not to beat a dead horse here, but, yeah, things like CPU and memory storage need to be balanced out given the host system has a limited amount of these resources.

- Powered off VMs – While these aren't taking up resources actively, they do need to be taken into account as part of your storage calculations when looking at what you need in the cloud.
- Snapshots – Like powered off VMs, snapshots add to your storage tally.
- VM "Zombies" – These VMs are running somewhere but aren't really in use. They look like VMs that are "alive" but are just wandering aimlessly consuming resources.
- Inaccessible VMs – A misconfigured network interface can cause a VM to be seen on a disk with a connection state, etc., but there's no means of connection. Again, more resource wastage.

These same factors will need to be a part of your workload optimization strategy.

Remember, in the cloud, optimization works differently. Since the cloud vendor owns the infrastructure, you're limited to changing rate plans instead of actual resources, compromising between the performance you need and plan you can afford.

Which Workloads Should You Optimize First?

Some of you may have so many workloads that it is difficult to decide where to start the work of optimizing. So, there are actually a few methods you can use to determine where to begin:

Most Constrained Workload

The *Theory of Constraints* holds that a system is most constrained by its most constrained resource (in our case, CPU, memory, storage, memory, or network). In essence, it's a take on the old adage "you're only as strong as your weakest link."

When looking for where to optimize first, you need to create an index of which VMs and cloud instances

are most resource constrained, and start with the one that is lacking the resources it needs the most.

Min/Avg/Max Workloads

If you have far too many workloads to manage, this approach says to leverage the index of VMs and workloads, and identify the workload that consumes the least, the most, and the average consumption. These three become the representative workloads for corresponding workload groups. You'd then attempt to find the appropriate mix of cloud providers and plans that best fit those three workload groupings and place the remaining workloads into the appropriate cloud and plan.

Priority Workloads

How priority is determined depends on the organization. In some businesses, the disaster recovery planning team identifies which workload is the most critical to maintain. In others, priority is given to the workload with the most serious performance issues (e.g., the one that is triggering the most performance alarms).

Deciding When to Optimize

There are two times when you should definitely be thinking about optimizing your workloads. The first is (if possible) *before you move*. Having the benefit of time on your side gives you an advantage when trying to optimize. However, if you wait too long, it's like trying to purchase a gift for someone just hours before their birthday celebration; you're going to make some sacrifices on what you choose and will likely over spend on your choice.

A last optimization possibility is *containerization.* You can think about putting workloads into containers, allowing you to scale up and down more quickly than you can do with VMs. While this works best when building a workload from scratch, remember you can optimize containers as easily as VM hosts.

The second time to optimize is *after you move* (didn't see that coming, did you?). The workloads you put into the cloud have a lifecycle of their own; they grow in use, they plateau, and they diminish in importance. Throughout their lifecycle, the ability to

compromise on performance also shifts, giving you options around how you can optimize the workload's place in the cloud.

The Big Takeaways

Blindly pushing a workload into the cloud most definitely *isn't* the answer; the performance characteristics achieved on-premises are due to the resources made available to the workload. The cloud needs to meet those very same needs, despite being defined by plans and not resources.

Optimization of workloads finds a balance between the performance needed, the cost of a plan, and the operational risk that is created when putting a specific workload into a chosen plan.

Plan to optimize both before you move a workload to the cloud, and then again as the business needs around that workload change over time.

When thinking about optimizing all your workloads, the use of third-party tools may be necessary to ensure you give the same attention to each and every workload, while considering every possible cloud/plan combination to meet your need.

Placing workloads in the cloud requires more work than just a migration. Ensuring acceptable levels of workload performance while lowering costs and operational risk is a balance every IT organization must deliver. In this book, I'll cover the why, what, and how around optimizing cloud workloads.

About Nick Cavalancia

Nick Cavalancia is Microsoft Cloud and Datacenter MVP, a Technical Evangelist by trade, and is a 25+ year IT veteran who regularly speaks and writes for some of today's most recognizable companies.