PAPER

VIRTUALIZATION & Cloud Review

HOW TO MAKE YOUR APPS IN THE CLOUD SELF-OPTIMIZING?

By Greg Schulz, Server StorageIO

Densify

www.densify.com

IT managers are facing the challenge of how to match application Performance, Availability, Capacity, Economic (PACE) workload demands with the right data infrastructure including cloud resources. Without automation, manual efforts to optimize cloud resources to meet application PACE requirements results in lost time and productivity as well as increased cost and risk. This paper looks at common challenges, along with opportunities for enabling applications to be self-optimizing in both on-premises and cloud environments.

Key points include:

- Cloud resource management challenges and how to address them
- Automation and elastic compute for the on-premises virtual environment
- Match your applications demand with the right public cloud resources
- Maximize the return on your investment in cloud Reserved Instances
- How to make your applications self-optimizing while giving your credit card a rest
- Efficient, effective and productive, densification without compromise
- Business benefits of self-optimizing applications

COMMON APPLICATION CHALLENGES

Everything is not the same across different organizations, environments, application workloads, and the data infrastructure that support them. Data infrastructures are what exists inside data centers, from legacy, on-premises to public and private clouds; combining server storage, I/O network, hardware, software resources from on-premises to public clouds including containers. Data infrastructures combine hardware (server compute, memory, I/O, networks, storage), software, best practices, policies, management tools to support various application workloads.

Common application challenges spanning on-premises to the public cloud include:

- Providing adequate resources to meet application service level objectives
- All applications have PACE needs
- Application workload PACE requirements need to be aligned with resource supply
- Growing complexity of continually changing resource service offerings and costs
- Lack of visibility, insight, awareness into application PACE needs, and available resources
- Expanding focus from utilization efficiency to useful, productive enablement
- Using the wrong type or amount of resources results in increased costs and poor performance

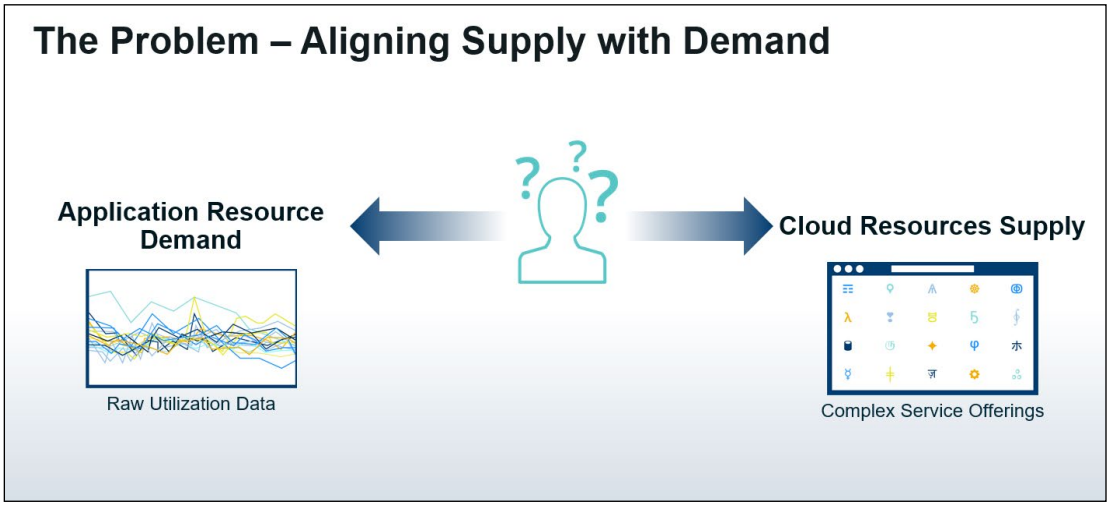

FIGURE 1: Balancing application workload demand with cloud resource supply

Mismatch between cloud provider resources and your application demands can lead to serious performance risk and massive unnecessary spend. Your credit card has to work harder in the cloud due to lack of proper alignment of cloud resources to workload needs. The business result should be increased productivity, as well as less work for your credit card as shown in **Figure 1**.

In **Figure 1**, on the demand side (left), there are applications and their workloads with various PACE resource demands (compute, memory, I/O) fluctuating all day. On the supply side (right), you have various PACE resources. These resources span legacy and software-defined virtual environments as well as public clouds.

A common challenge is the many different options along with permutations of how cloud resources can be chosen to host various workloads. Approaching the application to resource alignment challenge manually results in lost efficiency and operational risks. Operational risks include lost productivity and availability, wasted capacity and costs from choosing the wrong resources for different workloads.

PUBLIC CLOUD PAIN POINTS

IT managers are faced with making decisions about their data infrastructures (servers, storage, I/O networking, hardware, software) to host applications across multiple public clouds as well as hybrid on-premises environments. Managing applications along with their data infrastructure resources across multiple clouds at scale also means dealing with different technology tools and environments. A key challenge is the inability to quickly and correctly manually handle the process of identifying appropriate cloud resources for different application workloads.

There are many different public cloud service providers (e.g., Amazon Web Service [AWS] Google Cloud, Microsoft Azure). Cloud service providers have numerous data infrastructure service offerings. Cloud-based data infrastructure service offerings include various compute instance options, storage, along with relational database services among others. Cloud compute

resources vary from dedicated bare metal or metal as a service (MaaS) to software-defined virtual cloud instance and container options.

Adding to the cloud complexity, there are various families of cloud instances available in different configuration sizes to meet multiple workload needs. Also adding to the public cloud complexity are the numerous options for buying (renting or subscribing) to cloud instances among other resources. The result is that IT organizations are acquiring (buying, subscribing, renting, leasing) as well as deploying the wrong data infrastructure resources to meet their workload PACE requirements.

By not using the appropriate data infrastructure resources on-premises or in public clouds is increasing cost and potentially resulting in lost productivity. Increased costs occur from over-provisioning and under-utilization of resources (hardware and software). Lost productivity can happen from selecting the wrong resources to cut cost.

SOLVING APPLICATION CHALLENGES – WHAT TO DO

With an understanding of the various application and cloud optimization challenges, lets shift focus to what to do to resolve issues. Having timely, accurate insight, and awareness is essential. Also important is being able to classify and process workload needs along with available resources. This is where in-depth learning analytics comes into play to leverage policy-based automation of both small as well as large-scale cloud (and on-premises) deployments.

In general, to enable cloud (and on-premises) application optimization:

- Instead of looking at what you have, understand what your applications need
- Move from people-based decisions on what resources to use
- Leverage automation and deep learning to make timely accurate choices
- Give the right resources, both type and quantity to meet workload needs
- Avoid over-allocating, or starving applications from resources they need
- Enable densification that combines resource utilization with improved performance
- AI/ML/DL and automation need to provide evidence of how they will help you

SOFTWARE DEFINED VIRTUAL, CONTAINER AND CLOUD OPTIMIZATION

From on-premises bare metal or MaaS to software-defined virtual and containers, along with public cloud, there is a common challenge. The common problem is selecting the appropriate data infrastructure resources to meet application PACE resource requirements.

First generation software-defined and virtual deployment focused around consolidation to boost utilization and capacity efficiency. Second generation optimization for the software-defined cloud, VM and containers expand the focus to effectiveness, performance and productivity enablement.

Part of enabling next generation workload optimization is to look beyond simple averages, peaks, means, standard deviation activity results. What this means is leveraging deep learning to look further into what is occurring on different days at various times. The result is to

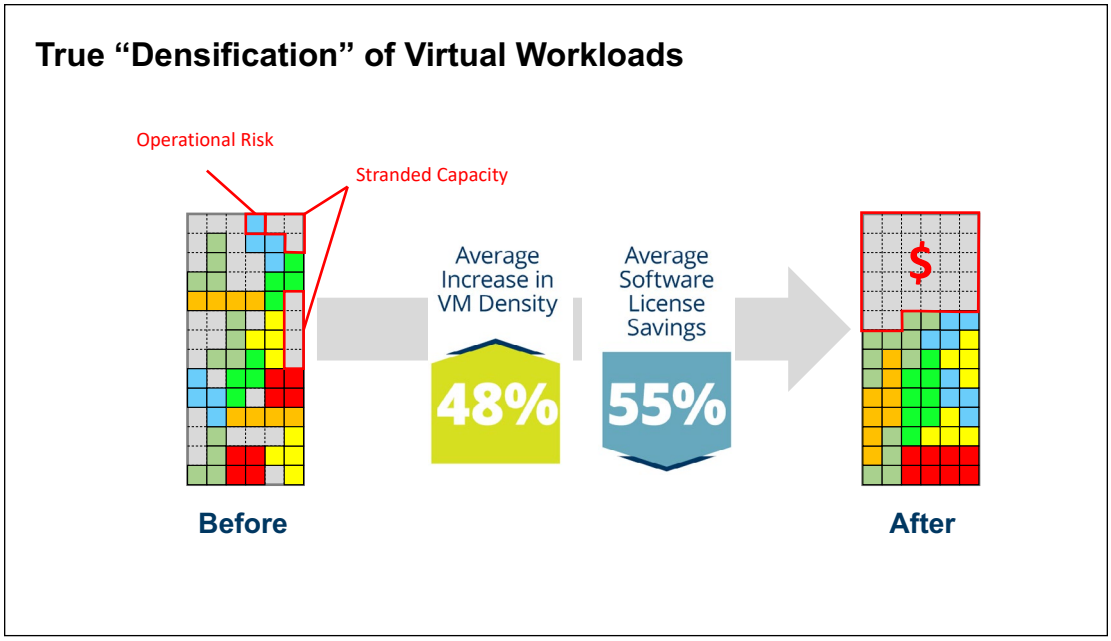

FIGURE 2: Software Defined private cloud VM optimization

correlate and gain insight into what resources are needed to boost productivity while reducing operational risk and stranded capacities.

Figure 2 shows on the left VM based applications with stranded (not usable) resources (server compute, memory, I/O, storage). On the right of **Figure 2** shows freed up resources available for allocation to other workloads along with subsequent cost savings (hardware, software, tools).

In addition to enabling application optimization for software-defined VM and cloud instances, containers can also be improved. **Figure 2** shows as an example how optimization resulting in enhanced densification balances workload PACE demand with appropriate resources. Densification results in reduced costs along with improved performance to boost productivity.

Many container deployments result in excess cloud spend due to misconfiguration of resources hard-coded in manifest configuration files. Proper resource alignment for containers is essential in that even if they only run for short durations, given a large number of workloads the resulting cloud cost is managed. For example, optimizing and aligning appropriate cloud resources to meet workload needs results in an 82% net savings, while boosting productivity shown in **Figure 3**.

Public cloud application self-optimization applies to different service providers, for example, AWS, Azure, and Google among others, as well as resource types. Self-optimization requires that the enabling software and service solution understand both the application workload PACE requirements, as well as those of the available resources. Besides knowing the available resources, other considerations include those defined by policies such as geographic location among others.

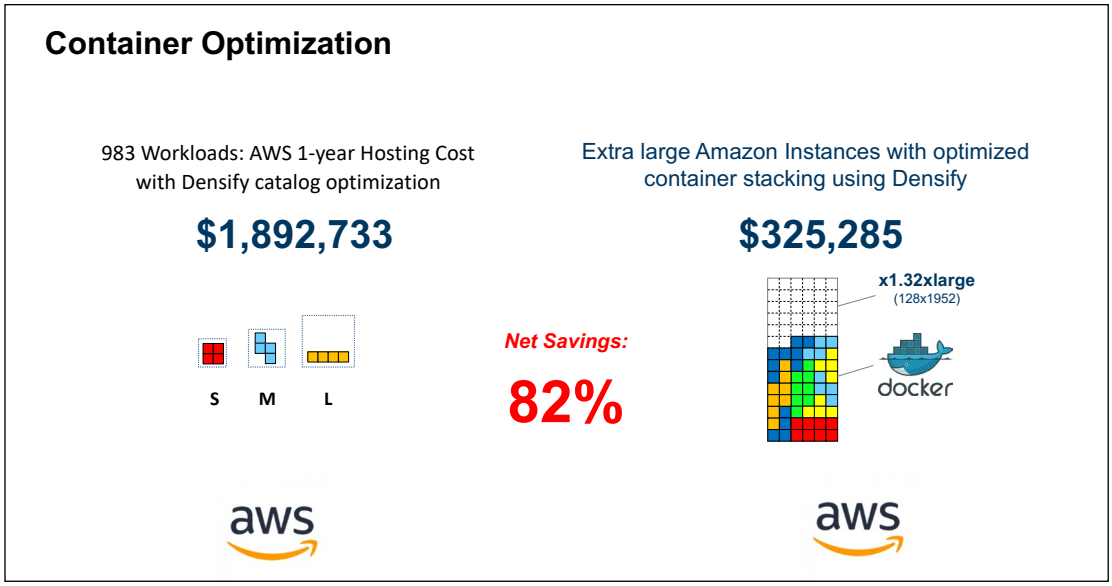

FIGURE 3: Software Defined container optimization

Cloud optimization also includes gaining insight and awareness into how pre-paid resources such as AWS Reserved Instances (RI) type and size. The problem is pre-buying the wrong instance or resource type which results in lost money in the long run. On the other hand, the solution is to optimize your applications and determine the correct type of RI or resource to use taking into consideration the desired optimal state. It's also very important to decide if it makes sense to reserve or purchase on-demand according to predicted uptime of the instances, avoiding to be locked into unnecessary financial commitments on Reserved Instances.

NEXT GENERATION AUTOMATION – CLOUD LEARNING OPTIMIZATION ENGINE

An example of the next generation automation solution in cloud optimization is Densify's Cloe (Cloud Learning Optimization Engine). Cloe is powered by machine learning technology and it makes applications running in the cloud self-optimizing, enabling them to automatically match their needs with perfectly fit cloud resources. Cloe learns and performs in-depth analysis of application workloads and their resource consumption including CPU, memory, I/O and storage.

In addition to application resource usage, Cloe looks at resource demand patterns beyond basic peaks, averages, percentile or standard deviations. Besides resource demands, Cloe also looks at and analyzes available cloud (and on-premises) resources as well as normalizes those to shared characteristics using industry-standard benchmarks.

Key to next-generation automation is leveraging deep learning that provides results and evidence of how those recommendations will help, vs. hurt application delivery. Part of providing an optimized solution, Cloe (**Figure 4**) looks for workloads that can be combined that are also complementary as opposed to those that would be competing for resources. The result is a solution that optimizes resource usage, while boosting productivity, eliminating bottlenecks, and enabling densification.

FIGURE 4: Cloud learning optimization engine (Cloe)

Figure 4 shows how Cloe leverages business policies combined with deep learning-based insight awareness to understand application demand along with PACE resource needs, as well as available cloud resource capabilities. The result is the right amount and type of cloud resources are allocated dynamically at runtime to meet application workload needs.

Cloe deployment involves changing for example in **Figure 5**, a single line of code in your applications definition template, manifest or resource configuration script. The change consists of replacing a hard-coded resource instance definition with a call to Cloe to obtain and configure the applicable resource appropriate for the given workload. Based on deep and continuous cloud learning, Cloe can adjust resource definition needs on a dynamic basis at runtime vs. static requiring a developer, DevOps, technical or operations support person to make the code change (**Figure 5**).

```
1   provider "aws" {
2     region = "${var.aws_region}"
3   }
4
5   resource "aws_instance" "web" {
6       name = "Web Server“
7
8       instance_type = "m4.large"
9       instance_type =["${aws_instance.tags:Densify-optimal-instance-type}"]
10
11      ami = "${lookup(var.aws_amis, var.aws_region)}"
12  }
```

FIGURE 5: Cloe dynamic resource call replacing static hard-coded resource definition

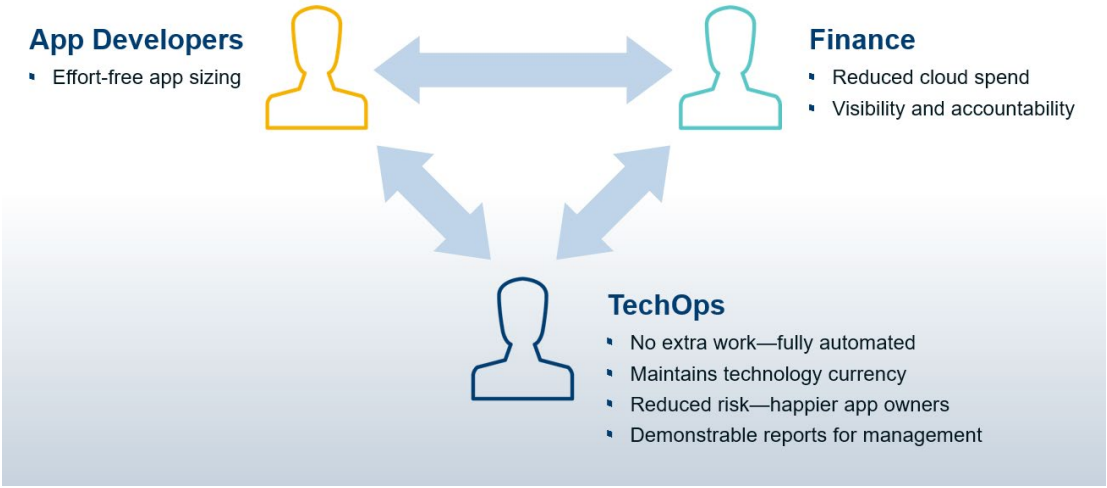

Figure 6: Self Optimizing Applications and Workloads – Everybody Wins

In **Figure 5**, a "hard-coded" resource definition (e.g., AWS EC2 m4.large) is replaced with a call to Densify Cloe which selects the optimum cloud compute instance to meet application PACE workload requirements.

BUSINESS ENABLEMENT OF SELF-OPTIMIZING CLOUD APPLICATIONS

Everybody wins (**Figure 6**) with self-optimizing cloud applications including developers, technical operations, and infrastructure support, along with finance as well as the user of applications. Regardless of if yours is a traditional IT environment, or a modern DevOps centric, startup to legacy, large or small, self-optimizing cloud applications benefits can be realized across different applications, public, private clouds along with other software-defined data infrastructure environments.

- **Infrastructure and Operations team** gets better alignment of cloud supply and application demands in a highly-automated fashion, and they don't have to beg the application owners to make the changes.
- **Developers** never have to do the guesswork on app sizing again when deploying their apps in the cloud, and their applications will be running efficiently in the cloud on the exact right resources.
- **Finance** team gets much higher cloud cost effectiveness, with increased predictability and accountability.

EVERYBODY WINS WITH SELF-OPTIMIZING CLOUD APPLICATIONS INCLUDING DEVELOPERS, TECHNICAL OPERATIONS, AND INFRASTRUCTURE SUPPORT.

ABOUT DENSIFY

Densify is the developer of Cloe (Cloud learning optimization engine), a machine-learning cloud optimization engine which enables your applications to become self-optimizing enabling them to automatically match their needs with perfectly fit cloud resources, 24/7. Densify continuously learns your applications' usage patterns and automates the function of finding optimal cloud resources. Delivered as a service, Densify customers achieve 60-80% improvement in application performance and cloud cost reduction.

SUMMARY, CALL TO ACTION AND NEXT STEPS

Instead of looking to optimize what you have, start looking at what your applications and environments need. Work backward from application PACE needs and then align applicable cloud or on-premises data infrastructure resources.

Additional summary points and things to keep in mind include:

- Align your cloud application workload demands with available resources
- Gain insight into your application PACE needs along with resource capabilities
- Drive efficiency and effectiveness from your cloud Reserved Instances (RI)
- Applies to legacy on-premises IT as well as new DevOps centric organizations
- Densification improves utilization as well as productivity without compromise
- Improved effectiveness boosts application and user productivity
- Give your credit card a rest, or, use it for other productive tasks

Take the next step to boost your cloud (and on-premises) applications to give your credit card a break by giving Densify Cloe enabled self-optimization a try. Sign up for a free-trial, and give your apps the power to self-optimize: **https://www.densify.com/try**

FIND OUT MORE AT:
WWW.DENSIFY.COM

Densify

Greg Schulz is the founder of independent IT Analyst consultancy firm Server StorageIO. He has worked as the customer in various IT organizations in different roles, as well as a vendor, consulting analyst and author of several books including "Software-Defined Data Infrastructure Essentials" (CRC Press). Greg brings a diverse background with hands-on, real-world perspective across applications, data infrastructures, hardware, software, and clouds. He is a Microsoft MVP Cloud Data Center Management and VMware vExpert.