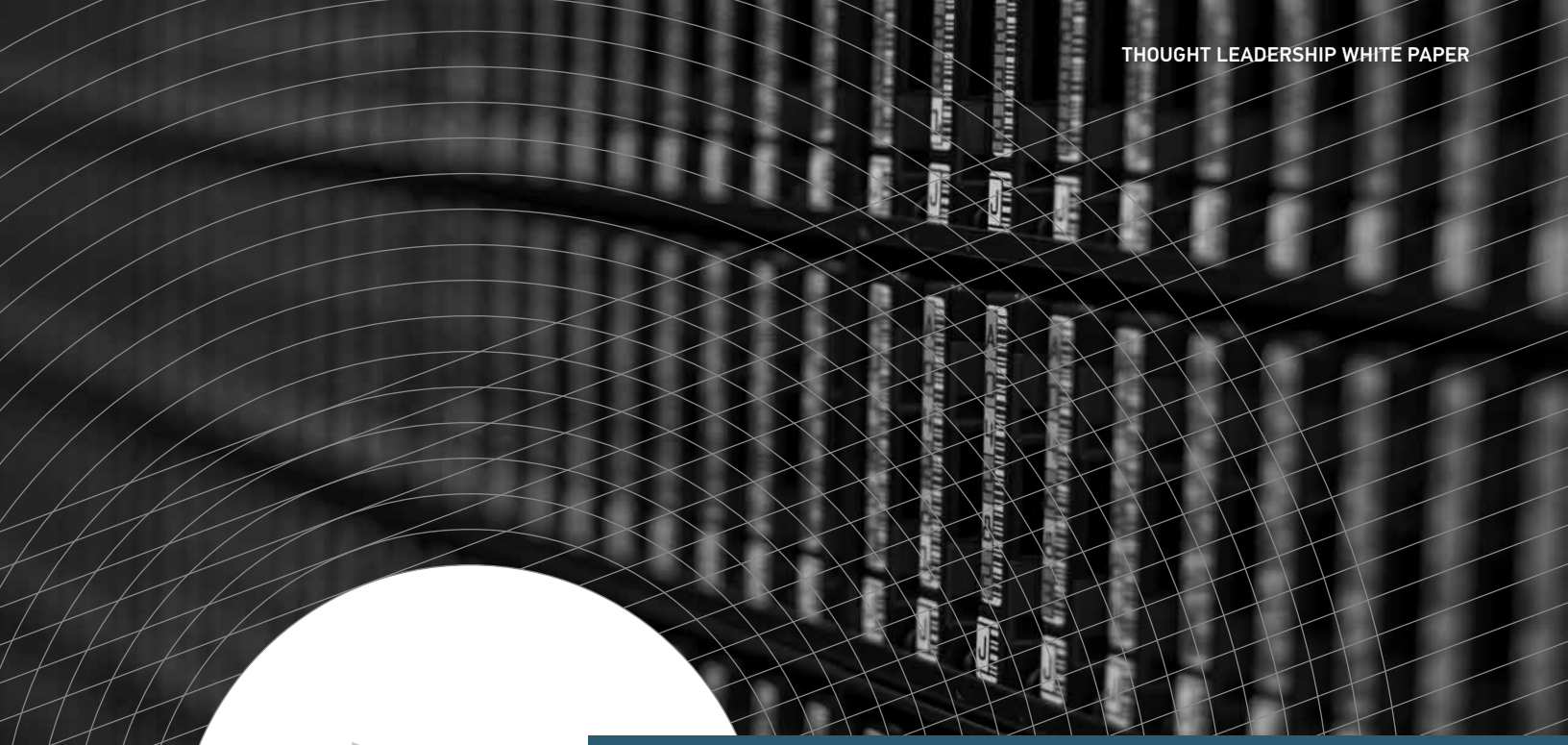
THOUGHT LEADERSHIP WHITE PAPER



Keep Your Cloud Humming: Cloud Operations Management

By Brian Singer, Lead Solution Marketing Manager for Cloud Computing, BMC Software

bmcsoftware
BUSINESS RUNS ON I.T.™

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Mark stared at his operations console, then at the report on his desk, and back again at the console. His company, a cheese-and-chocolate-of-the-month club and gift basket retailer, was facing its holiday order rush. Mark had just launched a hybrid cloud and was excited to see how it would help his company meet seasonal demand. He had worked carefully to plan his cloud to meet the ebb and flow of his business.

Mark thought everything was going smoothly. However, the October trends report indicated that, due to slow response times, some customers were having difficulty placing orders for pumpkin-shaped chocolate bars. Other customers were able to submit their orders, but then those orders were lost. If his cloud could not keep up with Halloween, how would it manage other holidays?

Mark's company was almost ready to release a new line of their popular holiday-themed cheeses. He needed to quickly find and solve the issue that was slowing down the response times, but he was not quite sure where to start or how long it would take. He needed to know about such problems in real time so they could be fixed before anyone even realized there was an issue.

Cloud computing is transforming the IT landscape. The ability to provide on-demand, self-service, elastic computing capacity is dramatically altering the role of enterprise IT — and the roles within it. Cloud computing also presents a new set of challenges for operations professionals like Mark (in the example above) to consider and overcome. Organizations looking to execute a successful cloud strategy must put a premium on managing performance, capacity, and end-user experience to maximize their investment in cloud environments.

IT is ultimately responsible for all services running on cloud infrastructure, whether public, private, or hybrid. Though cloud technology simplifies many processes, it also introduces a new layer of abstraction, presenting IT with a host of new chal–lenges. The existing IT infrastructure still needs to be managed alongside the new cloud infrastructure. Furthermore, IT is still responsible for ensuring that service level agreements (SLAs) and budgetary goals are met, as well as discovering and fixing issues. Finally, once a cloud is up and running, it is important to focus on continually improving service delivery.

To meet these requirements, IT should proactively monitor for performance issues and maintain optimal use of available resources. Three key components are required:

1. **Service Level Enforcement —** The way to make sure a cloud service is addressing business requirements is through an actively managed SLA enforced with end-user experience management.
2. **Proactive Service Performance Management —** IT can proactively manage performance across public and private cloud infrastructures through predictive analytics for performance monitoring and identification of performance issues.
3. **Continuous Resource Optimization —** To make the best use of cloud resources, IT must balance the way that workloads are distributed throughout the cloud infrastructure and also "right-size" individual cloud services on an ongoing basis.

This white paper discusses the challenges with cloud operations management and the best practices needed to overcome them.

CLOUD VERSUS TRADITIONAL OPERATIONS

At its core, cloud operations is about running and optimizing your hybrid cloud to deliver against SLAs. Meeting these SLAs is accomplished by understanding the current state of your resources and how they will change going forward.

While sometimes perceived to be temporary, most cloud services exist for weeks, months, or even years, and the number of workloads that run in cloud infrastructure will continue to grow as shared resources support more and more traditional IT workloads. When you add up all the operating systems, middleware, tooling, applications, and hypervisors in the average IT environment, this management task can overwhelm many organizations with ongoing repair and maintenance of the cloud infrastructure.

To understand how to operate cloud environments versus traditional IT environments, it's important to first understand the unique characteristics of cloud environments. Three attributes define cloud computing: agility, elasticity, and efficiency.

These attributes create complex differences between cloud operations management and data center management, as explained below.

AGILITY — SERVICES INSTEAD OF APPLICATIONS OR MACHINES

Business buyers do not want a cloud that just delivers Linux® boxes. They want a cloud that can deliver end-to-end business services. Furthermore, they are not satisfied with a service just being available; it must be available immediately and run reliably. IT organizations are competing against external providers that provide exactly this experience.

With thousands or tens of thousands of service requests, the cloud has to ensure that services are reliably provisioned. It must also be ready to support cloud users and be monitored by administrators. Manual processes that might have sufficed in a physical world are no longer feasible in a cloud environment. The entire on-boarding (and off-boarding) process needs to be standardized and automated — and provide self-service functionality.

Service reliability is not a new requirement. It is also a very important measurement in the traditional data center. However, maintaining high levels of reliability becomes even more critical and challenging in the cloud. As a result, reliability has a huge impact on how operations must be architected in a cloud environment.

ELASTICITY — POOLS INSTEAD OF BOXES

Many businesses look to the cloud to solve the immediate business demands that their traditional data centers cannot easily accommodate, such as seasonal peaks or rapid growth that outpaces expansion plans. The cloud offers elastic resources that can be scaled up and down to meet these types of dynamic demands.

The economical way for an internal or external provider to supply elasticity is to make sure the cloud infrastructure can support a large and varied number of resources. This is a big shift from the traditional data center to cloud. In the traditional data center, the provider supplies a box (either a physical or virtual server). In the cloud, the provider supplies pools of compute units, storage arrays, and networks devices. These resources must be pooled to provide high utilization and availability. Whereas a typical enterprise data center houses thousands of servers (with 20–35 percent utilization), a cloud could easily encompass hundreds of thousands of virtual machines (with up to 70–80 percent utilization).

EFFICIENCY — SHARING INSTEAD OF OWNING

While many business buyers look to the cloud to provide them with agility, IT buyers mainly seek a way to reduce their costs. More specifically, they want to reduce their capital expenditures (CapEx) by spreading them across more users. Cloud provides economies of scale that reduce unit costs — especially when combined with an on-demand model — and this translates into lower variable costs that get passed on to cloud users. The economies of scale translate not only to savings on CapEx through better utilization of commodity resources, but also to operating expenses (OpEx) through automated management processes. In one example, a large insurance company saved $200 million by consolidating nine data centers that were each run by separate business units down to a single, shared, cloud data center. The cost savings largely came through the efficiencies gained by sharing the infrastructure.

BEST PRACTICES FOR OPERATIONS MANAGEMENT IN THE CLOUD

Cloud operations management encompasses all the processes that are involved with ensuring the cloud runs smoothly once it is operational. It looks past simple day-one tasks, such as provisioning, to the nuts and bolts of ensuring services are properly delivered, SLAs are met, and the cloud is working the way it's supposed to. Shifting the model from boxes, applications, and ownership in a traditional data center to pools, services, and sharing in the cloud requires current operations management to manage multiple tenants and SLAs on the same shared infrastructure.

FOCUS ON THE END USER

Business clouds are about delivering services to an end user. If nothing else, IT has to ensure that the end user is receiving an experience that's consistent with his or her expectations and individual SLAs. At the same time, end users have learned from consumer-facing cloud services, such as Gmail or Facebook, that they do not necessarily need to know or consider where a service is running (though they might want some control over it). They just want it to work.

IT must be able to peel back the layers of a cloud like an onion. A problem could occur in a physical network connection at the ISP level, in an internal network, or in a virtual network managed by a virtual switch. CPU utilization could be running rampant and slowing down a service on a virtual machine or on the physical machine on which the service is actually running. Each possible cause introduces an entirely different set of remediation procedures. End-user experience management has to happen at the service or application level, but it cannot stop there if IT wants to solve problems as they occur.

MANAGE PERFORMANCE

Just as it does in traditional data centers, IT must make managing performance a key priority in the cloud. Consider that in the old (idealized) physical world, if a problem caused a service to use 80 percent of the CPU on a server, it was not very important to debug. There was probably only one service running on the server, and except for using a little bit more power, such a problem was not going to bother anyone.

In a virtual world, it becomes somewhat more important to remediate a problem like this, as virtualizing other workloads on the server provides cost-savings. However, in a cloud-based world, where usage is typically metered and pay-as-you-go, such problems quickly become expensive and a waste of resources. This is true for both private and public clouds. If you care about monitoring performance, then it follows that you should care about performance metrics. It's crucial for cloud operations to establish baselines for the metrics that cost money and impact the end-user — CPU usage, RAM usage, response times, and latency — and then to monitor performance against those baselines.

Consider a workload that is running in a public cloud. Without performance monitoring, IT's only way of knowing that a problem has occurred is when it impacts an end user. IT may look into a trouble ticket and notice that the service is running out of resources. In the public cloud, it's very simple to increase the size of the server that the service is running on, but it's also costly. Without an established baseline, IT has no way of knowing whether the application actually requires more resources, or if there is an underlying issue causing the increased resource usage that can be fixed. Even though the resources are readily available, spending more may not be the most efficient way to address the problem, and this approach can quickly wipe away any cost savings from using cloud infrastructure.

MOVE BEYOND PERFORMANCE WITH A FOCUS ON CAPACITY OPTIMIZATION

Virtualization is the precursor to cloud. It's a key enabling technology that has allowed organizations to dramatically increase hardware utilization. Managing capacity is critical to unlocking the potential of virtualization. You need to know ahead of time whether service *x* and service *y* can run on the same machine. If service *y* starts requiring too many resources, you need to know (so that you can migrate it before it slows down service *x*).

Cloud takes capacity optimization a step beyond virtualization. In the cloud, capacity optimization is not just important; it's essential. Your cloud cannot exist without it — at least not the kind of cloud that will provide lasting benefits to your business. In a cloud environment, deployment of virtual machines and services is automated. Provisioning happens without human intervention. It follows, then, that capacity optimization must be automated as well. The provisioning engine needs to know which physical servers have the resources to run a service. It also must understand when the underlying infrastructure is nearing capacity limits and provide the ability to rebalance infrastructure when capacity constraints are reached. Since the cloud is an abstracted pool of physical resources, it will not be immediately clear when those resources are close to being used up unless there is a capacity optimization solution in place.

Additionally, capacity optimization can make or break your cloud when it comes to cost efficiency. If you have resources provisioned for services that are not being used, you will be spending money needlessly. With the new resource-based pricing models for virtualization, you must optimize physical server and virtual configuration very carefully. This means that it's crucial to provision virtual machines only as they are used — and to de-provision them when they are no longer needed. Your cloud management solution should be able to handle that level of automation without requiring manual intervention. This is very difficult in practice and requires a level of workflow maturity and capability that many solutions do not yet have.

MASTERING CLOUD OPERATIONS REQUIREMENTS

Operations management needs to master six new capabilities to deliver on the promise of cloud.

1. OPERATE ON THE "POOLS" OF COMPUTE, STORAGE, AND MEMORY

Traditionally, operations management solutions have provided coverage for individual servers, storage arrays, or network devices. With the cloud, it becomes imperative to operate at the "pool" level. You have to look beyond what can be monitored at the individual device level. Operations organizations must ensure that they have immediate access to the operational status of the pool. That status could be aggregated by workload (current usage) and capacity (past usage and future projections). Perhaps more importantly, the status needs to accurately reflect the underlying health of the pool, even though individual component availability is not the same as pool availability. The operations management solution you use should understand the behavior of the pool and report the health status based on it.

2. MONITOR ELASTIC SERVICE

Elasticity is central to cloud architectures, which means that services can dynamically expand and contract based on demand. Your operations management solution must adapt to this dynamic nature. For example, when monitoring the performance of a service, monitoring coverage should expand or retract with the service — automatically. This means that a manual process cannot be used to figure out and deploy monitoring capabilities to the target. Your operations management solution needs to know the configuration of that service and automatically deploy or remove necessary agents.

Another important consideration is coverage for both cloud and non-cloud resources. This is most critical for enterprises building a private cloud. Why? Chances are that not every tier of a multitier application can be moved to the cloud. There may be static, legacy pieces, such as a database or persistent layer, which are still deployed in the physical boxes. Services must be monitored no matter where resources are located, in the cloud or on premises. In addition, a management solution should natively understand different behavior in each environment.

When resources are located in both private and public clouds, your operations solution should monitor services in each seamlessly. It should also support inter-cloud service migration. At the end of day, services must be monitored no matter where their resources are located. Your operations management solution must know their location and understand the behavior of services accordingly.

3. DETECT ISSUES BEFORE THEY HAPPEN

Compared to workloads in the traditional data center, workloads in the cloud exhibit a wider variety of behavioral issues due to their elastic nature. When service agility is important, relying on reactive alerts or events to support stringent SLAs is not an option — particularly for service providers. You need to detect and resolve issues before they happen. Yet, how do you do that? First and foremost, you should implement a monitoring solution that knows how to learn the behavior of your cloud infrastructure and cloud services.

While this technology exists in the traditional data center, device-level behavior evolves more rapidly and with less conformity in the cloud. That's why your solution should have the ability to learn the behavior of abstracted resources, such as pools, as well as service levels that are based on business key performance indicators (KPIs). Based on those metrics, the solution should give predictive warnings to isolate problems before they affect your customer.

To further pinpoint problems, operations should conduct a proper root cause analysis. This becomes even more critical in the cloud, where large numbers of scattered resources are involved. This information might manifest itself as a sea of red alerts suddenly appearing in a monitoring dashboard. Even though one may be a critical network alert, chances are you are not going to notice it. Your operations management solution should intelligently detect the root cause of an issue with the cloud infrastructure and highlight that network event in your dashboard, while also invoking your remediation process.

4. MAKE HOLISTIC OPERATIONS DECISIONS

In the cloud, you have to manage more types of constructs in your environment than in the traditional IT environment. In addition to servers, operating systems, and applications, you will have compute pools, storage pools, network containers, services, and tenants (for service providers). These new constructs are tightly coupled. You cannot view their performance and capacity data in silos; they have to be managed holistically.

It is important to know who your most crucial customers are — and to identify their services so you can focus on recovering them in order of priority. In addition, you may want to send out alerts to affected customers to proactively let them know there is an issue. Your operations management solution should give you a panoramic view of all these aspects and their relationships. Not only will it let you quickly isolate the problem, but it will also save you money if you know which SLAs cost more to breach and therefore should be addressed first.

5. ENABLE SELF-SERVICE FOR OPERATIONS

To give your cloud users their desired experience while also saving on support costs, it's important to provide constant feedback. Traditionally, performance data has not been available to the end user. In the cloud, however, there is a larger number of users or service requests with a relatively lower ratio of administrators. For that reason, it's important to minimize the "false alarms" or manual routine requests. The best way is to let your end users see the performance and capacity data surrounding their services.

You can also let your users define key performance indicators (KPIs) to monitor, the threshold levels they want to set, and some routine remediation processes they want to trigger (such as auto-scaling). The operations management solution should allow you to easily plug this data into your end-user portal.

6. MAKE CLOUD SERVICES RESILIENT

Resiliency is the ultimate goal of proper cloud operations management. If a solution is able to understand the behavior of cloud services and proactively pinpoint potential issues, it's natural for that solution to automatically isolate and eliminate problems. First, the solution must have accurate behavior learning and analytics capabilities. Second, a human must create well-defined policies with an automated policy engine or a human interactive process. Lastly, the solution must plug seamlessly into other lifecycle management solutions, such as provisioning, change management, and service request management.

Operations management in a silo cannot make your cloud resilient. You should plan the right architectural design as a foundation and implement a good management process that reflects the paradigm shift to ensure your success.

CLOUD OPERATIONS: MAKING CUSTOMERS HAPPY DURING THE HOLIDAYS

Yes, the cloud is here to stay. The benefits of a properly managed cloud are impossible to ignore, and as Mark found out earlier, critical to the success of your business. To properly manage a cloud — and move it from a lab to a production environment — you must focus on operations in the domains of end user experience management, performance management, and capacity optimization.

The cloud is a chance for IT to reboot itself, to prove that it is a key enabler of the business and that it can help the business achieve its goals. To learn more about cloud operations, go to **www.bmc.com/cloud**.

ABOUT THE AUTHOR

Brian Singer is the lead solution marketing manager for cloud computing at BMC Software. Singer has broad experience with building and marketing software and hardware infrastructure products spanning cloud, security management, virtualization, and networking. He has worked for Novell and QLogic and holds an MBA from MIT Sloan School of Management and a computer engineering degree from Brown University.

BUSINESS RUNS ON I.T.
I.T. RUNS ON BMC SOFTWARE

Business thrives when IT runs smarter, faster and stronger. That's why the most demanding IT organizations in the world rely on BMC Software across distributed, mainframe, virtual and cloud environments. Recognized as the leader in Business Service Management, BMC offers a comprehensive approach and unified platform that helps IT organizations cut cost, reduce risk and drive business profit. For the four fiscal quarters ended September 30, 2011, BMC revenue was approximately $2.2 billion. For more information, visit **www.bmc.com**.