To Take Advantage of Cloud Computing You Must Unlearn, Luke.

Lori MacVittie, 2009-28-10

Carrying over the provisioning and capacity planning techniques used in a traditional data center to cloud computing negates the full power of ~~the Force~~ cloud computing.

One of the benefits of cloud computing is supposed to be efficiency, particularly in the utilization of compute resources. Over-provisioning of compute resources has long been one way in which IT combats the need for scalability and availability of applications but this often leaves a large percentage of compute resources unused. The utilization rule once employed as a means to ensure availability and performance of applications, i.e. no device or server should utilize more than X% of its resources at any time, is no longer acceptable as it wastes resources which in turn eats away at the bottom line.

Cloud computing, through virtualization, is supposed to address this waste by providing an on-demand model of resource provisioning. The theory is that as an application requires additional resources to scale that those resources are allocated to it. The reality is that virtualization doesn't really work that way. Resources are allocated to a virtual machine in which an application is deployed. The resources assigned are fairly static. While some models provide for bursts of resource utilization, others do not and in both cases there is a ceiling to the amount of resources that can be provisioned to any given container.

Most thus offer options, much like server hardware vendors, on the "size" of the virtual containers available. These containers vary in RAM and CPU capabilities, offering users the ability to choose a virtual container that best fits the specific needs of their applications. If an application grows beyond those capabilities users have a choice: they can vertically scale by "upgrading" to a higher resource container or they can deploy multiple instances of the existing container and load balance requests across them. From an operational perspective this is exactly the choice IT has always had.

We know that application usage patterns are, for most applications, rarely constant. There are peaks in usage; sometimes daily, sometimes weekly, sometimes seasonal. If an application normally needs X resources, then to meet demand during a peak the application needs X+Y. It's a simple equation, and one with which IT operations is very familiar. The problem in meeting that demand in a traditional architecture with either vertical or horizontal scalability methods is that most of the time it wastes resources. The additional capacity – whether from a bigger server or a new server – is not being used most of the time. It's only pulled out of the "break in case of emergency" glass case during peak demand; the rest of the time it's just sitting there, wasting power and space.

START SMALL, NOT BIG

This is what cloud computing and virtualization is supposed to address, but IT operations has to first trust the ability of cloud computing models to scale up, on-demand, as per the literature. In order to maximize the benefits of cloud computing IT actually has to provision resources based on the lowest common denominator rather than trying to provision for highest peak demand, which runs contrary to everything IT operations knows as truth about provisioning a data center to ensure availability of applications around the clock.

Consider there are two sizes of containers offered by a provider: small and large. Most of the time a small will serve your applications admirably and meets defined SLAs without issue. But during peak demand you really need a large instance to maintain both availability and performance. The answer, in a cloud computing environment, is to take advantage of scaling services and let the infrastructure handle the situation: launching new instances as is necessary to meet demand and releasing those instances as demand decreases. This keeps costs at the minimum necessary to support the application and still meets demand when necessary, but it is ***efficient***. It wastes no resources because it provisions only the bare minimum necessary for base demand and dynamically allocates additional resources as necessary.

But there is a danger that IT operations, having been burned in the past by under-provisioning, will continue to over-provision resources in the cloud to ensure that peak demand can be met by the same instance that handles normal demand. This defeats the purpose of on-demand provisioning of compute resources and wastes both the extra resources and money. The difference between cloud computing and traditional environments in this situation is that the money wasted in the former is ***tangible.*** It's clearly shown on an invoice, listed for all to see and tally up at the end of the year.

Assume a small instance is $0.10 per hour and a large is $0.40 per hour. Further assume you need peak demand for 10 hours per day that requires the compute resources of a large instance. Assume also that four "small" instances provides the equivalent compute resources of a single large instance. The difference in costs between an on-demand provisioning model in which instances are launched to meet demand versus an over-provisioned environment is certainly significant:

	On-Demand	Over-provisioned
Cost per year	$1971	$3504

Imagine, if you will, how the difference in costs will become more significant as the number of applications and amount of computing resources you need increase.

TRUST IN THE TECHNOLOGY AND YOUR PROVIDER

Virtualization is an enabler of cloud computing, but it's not perfect. The optimal solution to the most efficient use of compute resources requires a technology that will be much more granular in its ability to provision those resources on-demand. But at the moment it's a damned sight better that what we've done in the past. In order to really leverage the benefits of reduced operating expenses associated with cloud computing we have to use the technology available to achieve the end-goal and in this case it means provisioning compute resources in steps based on the the least amount necessary, not the most.

That means IT operations and users of cloud computing must unlearn what they know now about capacity planning and provisioning and embrace the concept of on-demand and automated scalability. They must learn to ***trust*** the technology and the environments, or run the risk of continuing of not benefitting as much as they could from a move to cloud computing.

- Vertical Scalability Cloud Computing Style
- Load balancing is key to successful cloud-based (dynamic) architectures
- Virtual Machine Density as the New Measure of IT Efficiency
- Trends in Enterprise Virtualization Technologies [PDF]
- Putting a Price on Uptime
- Cloud Computing: Vertical Scalability is Still Your Problem
- The Myth of 100% IT Efficiency
- How do you get the benefits of shared resources in a private cloud?

F5 Networks, Inc. | 401 Elliot Avenue West, Seattle, WA 98119 | 888-882-4447 | f5.com

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com