Aligning IT with the Business by Decreasing Efficiency

Lori MacVittie, 2010-22-11

Here's the conundrum: utilizing every last drop of network, storage, and compute resources can impede performance and, through it, the business' bottom line. So which do you choose?

There are a few vertical industries for which performance is absolutely critical. A delay of even a micro-second can mean a huge differential in revenue or lost opportunities. A delay of seconds is a disaster, and more than that? Might as call yourself unavailable. While most organizations do not have such stringent "do or die" performance requirements, performance is always top of mind because users, well, users and customers are increasingly demanding with regards to the performance of their hyper-connected, online-driven lives.

So along comes "cloud" and introduces the myth of 100% efficiency. Like a Maxwell House ad, providers and pundits alike tout full utilization – to the last drop – of compute, network, and storage infrastructure. Stop wasting resources! Put those idle resources to work for you! Save money! Do it now or the business will outsource you!

It's enough to make you want to do it and the performance of their applications be damned, isn't it?

So you do, and now you have to listen to complaints and watch the help desk tickets pile up regarding performance of applications – yours and everyone else's, for that matter. You didn't realize you were going to be responsible for timely responses from Facebook and Twitter, did you?

See, there are some technical reasons ***why*** operations never ran network and server infrastructure components at 100% utilization. In fact, the rule of thumb was always 60% for most organizations and a harsher 30% or so for those with more performance-sensitive business needs.

LET'S TALK ABOUT QUEUES

IMAGE CREDIT: FREEFOTO.COM

Not, not the UK English version of what many IT administrators and operators have hanging down their backs, the ***technical*** queues, the ones that handle input and output for network stacks and applications in every corner of the data center.

See, any device that processes packets (which means everything network-capable) utilizes some sort of queue to manage the reality that packets will eventually come in faster than they can be processed. As packet processing "backs up", the queue fills up. The longer it takes for packets to get through the queue and be processed, the longer it takes the overall exchange of data to occur.

One of the reasons packets might "back up" in the queue is that the time it takes to process the packet – apply security, route it to the correct VLAN/network/etc…, apply quality of service policies – is related to the utilization on the device. The more packets the device is trying to process the more it consumes CPU and RAM and associated hardware resources which translates into less available resources being spread around all the different functions that must be performed on the device. The more the resources are consumed, the slower the device can process packets.

This also happens on web/application servers, by the way, when a client is trying to receive data but is doing so over a relatively slow connection. The client can only pull data so fast, and so the send/receive queues on the web/application server remain filled with data until the client can complete the transfer. There are only so many send/receive queues available for use on a web/application server, so emptying those queues as quickly as possible is a primary focus for application delivery infrastructure as a means to improve capacity and overall performance.

In any case, there is a fixed amount of compute resources available for each device/server and it must be shared across all the queues it is managing. As the utilization of devices increases, it means that the time-slices each queue receives to process data ***decreases,*** which means fewer packets are processed with every processing time-slice. That means packets "back up" in the queue, waiting their turn to be processed. Waiting = latency, and latency = delay in service delivery.

The higher the utilization, the longer the queues. The longer the queues, the higher the latency.

It's actually a pretty simple equation when you get down to it.

A BALANCING ACT

This is where the rubber meets the road – balancing the need for speed with the need to be efficient. Idle resources are the devil's playground, right? It's a waste to leave resources unused when they could be doing ***something.*** At least that's the message being sent by cloud computing advocates, even though efficiency of IT ***processes*** is much more a realizable benefit from cloud computing than efficiency of resources.

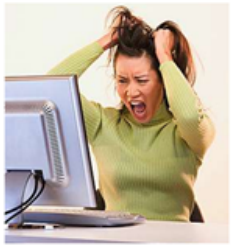
For many organizations that's absolutely true. Why not leverage idle resources to fill additional capacity needs? It just makes financial and technical sense. Unless you're an organization whose livelihood (and revenue stream) depends on speed. If even a microsecond of latency may cost the business money, then utilization is important to you only in the sense that you want to keep it low enough on every network component that touches the data flow such that near-zero latency is introduced. If that means imposing a "no more than 30%" utilization on any component policy, that's what it means – efficiency be damned.

Different business models have different performance needs and while the majority of organizations do not have such stringent requirements regarding performance, those that do will never buy into the Maxwell House theory of resource utilization. They can't, because doing so makes it impossible to meet performance requirements which are, for them, a much higher priority than utilization. Basically, the cost of failing to perform is much higher than the cost of acquiring and managing resources.

This doesn't mean that cloud computing isn't a fit for such performance-focused organizations. In fact, cloud computing can be an asset for those organizations in the same way it is for organizations trying to achieve a "good to the last drop" resource utilization policy. It's just that performance-minded organizations will set their thresholds for provisioning additional resources extremely low, to ensure optimal performance on each and every network and server component. Where most organizations may provision additional capacity when a component reaches 70-80% utilization, performance-minded organizations will likely try to remain below the 50% threshold – at around 30%. Or more accurately, they'll use a context-aware network of application delivery components that can assist in maintaining performance levels by actually watching the real-time performance of applications and feeding that data into the appropriate systems to ensure additional capacity is provisioned before performance is impacted by increased utilization. Because load on a server – virtual or iron – is a direct input to the performance equation, utilization is an important key performance metric that should be monitored and leveraged as part of the automated provisioning process.

ENGAGE the ENTIRE INFRASTRUCTURE

Performance-minded organizations aren't just financial and banking, as you might assume. Organizations running call centers of any kind should be, if they aren't already, performance-focused for at least their call center applications. Why?



OMG! This thing is so slow I can't stand it anymore! I'm going to miss my quota and lose my bonus. Someone is going to pay for this...

Because milliseconds add up to seconds add up to minutes add up to reduced utilization of agents. It means a less efficient call center that costs more per customer to run. Most call centers leverage web-based applications and delays in moving through that system mean increased call duration and lowered agent utilization. That all translates into increased costs – hard and soft – that must be balanced elsewhere in the business' financial ledger.

While certainly not as laser-focused on performance as perhaps a financial institution, organizations for whom customer costs are an important key performance indicator should be concerned about the utilization of components across their entire data center. That means balancing the costs of "idle" resources versus costs incurred by delays caused by latency and a decision: where is the tipping point for utilization? At which point does the cost of latency exceed the costs of those idle resources? That's your maximum utilization point, and it may be well below the touted 100% (or nigh unto that) utilization of cloud computing.

Don't forget, however, that this is essentially an exercise in tuning your entire data center infrastructure. You may baseline your infrastructure with a tipping point of 60% utilization, but by leveraging the appropriate application delivery technologies – caching, compression, network and application optimization, offload capabilities – the tipping point may be increased to 70% or higher. This is an iterative process requiring an agile infrastructure and operational culture; one that is able to tune, tweak, and refine the application delivery process until it's running like a finely honed race car engine. Optimally burning the right amount of resources to provide just the right amount of performance such that the entire data center is perfectly balanced with the business.

This is the process that is often overlooked and rarely discussed: that the data center is not – or should not be – simply a bunch of interconnected devices through which packets pass. It should be a collaborative, integrated ecosystem of components working in concert to enable the balance needed in the data center to ensure maximum utilization without compromising performance and through it, business requirements.

So though it sounds counterintuitive, it may actually be necessary to decrease efficiency within IT as a means to align with business needs and ensure the right balance of performance, utilization, and costs. Because "costs" aren't just about IT costs, they're about ***business*** costs, too. When a decrease in IT costs increases business costs, nobody wins.

Related blogs & articles:

- The Myth of 100% IT Efficiency
- IT Myths and Legends: Sharing Servers
- Cloud + BPM = Business Process Scalability
- WILS: What Does It Mean to Align IT with the Business
- Business-Layer Load Balancing
- Infrastructure 2.0: Aligning the network with the business (and the rest of IT)
- Caveat Emptor: Be sure to align your goals for cloud computing with provider models before you sign up
- Like Garth, We Fear Change

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