Stopping senseless sprawl of private clouds

Even in a cloud world, reducing server count is a lofty goal in itself. After all, you would be paying for all those servers anyway. And sticking all your servers in a private cloud makes them more flexible, but not necessarily more efficient or cheaper.

So today I bring you an interview with a guy who has a full time job in keeping those server counts down. Ron Kaminski is a capacity planner at Kimberly Clark Corporation (KCC). This interview is a composite of several conversations I had with Ron at the annual CMG conferences, and a number of messages Ron wrote.

Peter van Eijk: Ron, when you joined KCC you helped stop their trend of building a new computer room every 6 months, down to no new computer rooms at all. Is that right?

Ron Kaminski: Yes Peter that was basically what happened. As a result of the development of the business, there was a big growth in applications. Each of these applications brings in a lots of servers, because of the multi-tier architecture, and because you need test and development servers.

Peter: Could you tell us a bit about the way your firm stopped the server growth? What are the most important tactics?

Ron: Invisible machines proliferate. To begin with, automate the measurement and web delivered graphical displays of the resource consumption of each and every server. This will allow you to identify unused and under used servers, which you can then repurpose to projects and applications that have a real need for them. A lot of times, a detailed performance analysis suggests changes in the application that will drastically reduce its resources needs. Think bad/missing database indexes, memory leaks, and runaway/looping processes.

Peter: Everybody loves clouds these days, and if they can't get any public clouds, they will settle for private clouds. What are the most relevant lessons for those adopters?

Ron: Basically two lessons: You still need to do capacity planning, and you have to automate or you will fail.

Many folks say "Why bother with capacity planning? The cloud is infinite!" and that is only true as long as your wallet is infinite too. We need to be sure that we have tracking and review of intended business volumes and costs, in order to make decisions about when we need to use or not use cloud resources intelligently and economically.

Peter: Yes, computing is still not free. How far do you need to go with automation?

Ron: Well, what we have learned is basically "Automate or die!" both for creation and removal. If we expect to use either internal or external clouds, we need to stop doing setups that involve humans in any way. This means that we need to have standardized services that are engineered and/or architected to be complete environments, with any database or other services in place. There should also be administrative checking/post creation review, with removal if determined to be wasteful.

A key point that we need to have is the notion of "server lifespan" and automatic deletion at end of life. Folks will fight this until they get comfortable with how easy it is to get a new one. If we don't have "delete at end of lifespan" all this automation will lead to massive server sprawl and OS image proliferation, and all cloud vendors are counting on this for their profits. They know that firms left to their own devices will "create and forget" so they will continue to bill for unused servers. We need to get very good at noticing and removing dormant servers or risk large wasteful bills.

We also need automatic "backup/ archive and restore" of cloud resources, so that they can be completely restored swiftly in a totally automated manner.

Peter: So what about performance analysis?

Ron: The universally accepted best practice is to contractually demand collectors deployed in the clouds to allow performance analysis.

Naive business people will have quite a shock when they still have slowness issues when running in giant clouds, when their applications are not engineered to succeed in the method they are deployed. Examples of these include single-threaded applications, chatty applications over long network distances, and architectures with locking/latching issues. These will still have poor performance in the biggest cloud in the world.

The truth is we will still need development and testing in the cloud world, and check up on performance.

Peter: The CMG community has a lot of technical deep dive experts, and they have taken to analyzing server virtualization in great detail. What where the most important lessons that you picked up from these presentations?

Ron: The depth and complexity of issues covered in many sessions on VMware really made me wish that we sent more VMware resources to the CMG conference. In a nutshell:

We should be at the most recent versions, as the performance and scheduling losses on older ones sometimes exceed 30% or more of the physical machine's resources.

One particularly surprising item was that when current VMware has a 4 VCPU "guest" that is in a computable state, it waits to schedule it until all 4 CPUs are simultaneously free. More scarily, 8 VCPU guests wait to be scheduled until all 8 are free simultaneously. That means that there can be lots of waiting on relatively lightly loaded machines, say if there is a CPU loop in another guest slice. Since we have so many 8 VCPU mail systems, this really gave me the willies. We need to dig deeper on these and determine the performance losses this may be causing for us.

Peter: What other policies come out of these experiences that you want to suggest to your IT management?

Ron: VMware still takes some measurable amount of resources/overhead, and we really should avoid VMware for servers of any significant continuous consumption. We need to set a corporate policy for maximum usage allowed to be virtualized.

Peter: Thanks ever so much Ron, is there any place online where people can go to get to know more about this stuff?

Ron: Join CMG and get to the conferences. Be sure to get to and bring drink tickets, especially on Thursday nights! ☺