acmqueue

Moving to the Edge: An ACM CTO Roundtable on Network Virtualization

How will virtualization technologies affect network service architectures?

The general IT community is just beginning to digest how the advent of virtual machines and cloud computing is changing their world. These new technologies promise to make applications more portable and increase the opportunity for more flexibility and efficiency in both on-premises and outsourced support infrastructures. However, virtualization can break long-standing linkages between applications and their supporting physical devices. Before data-center managers can take advantage of these new opportunities, they must have a better understanding of service infrastructure requirements and their linkages to applications.

In this ACM CTO Roundtable, leading providers and users of network virtualization technologies discuss how virtualization and clouds impact network service architectures, in their abilities both to move legacy applications to more flexible and efficient virtualized environments and to enable new types of network functionality.

PARTICIPANTS

SIMON CROSBY CTO, Virtualization and Management Division, Citrix Systems
OLIVER TAVAKOLI CTO, VP SLT Architecture and Technology Group, Juniper Networks
LIN NEASE Director of Emerging Technologies for ProCurve Networking, Hewlett-Packard
MARTIN CASADO VP, CTO, founder, Nicira Inc.
SURENDRA REDDY VP, Cloud Computing, Yahoo!
CHARLES BEELER General Partner, El Dorado Ventures
STEVE BOURNE Chairman, ACM Professions Board; CTO, El Dorado Ventures; past president, ACM
MACHE CREEGER (moderator) Principal, Emergent Technology Associates

CREEGER Our discussion will focus on how virtualization and clouds impact network service architectures, both in the ability to move legacy applications to more flexible and efficient virtualized environments and what new functionality may become available. I would like each of you to comment on the challenges and opportunities people will face in the next couple of years as the world progresses with these new platform architectures.

CROSBY Virtualization challenges the binding of infrastructural services to physical devices. One can no longer reason about the presence or the utility of a service function physically bound to a device and its relationship to a specific workload. Workloads now move around, based on demand, response time, available service capacity, resource prices, etc. While the networking industry was founded on a value proposition tied to a specific physical box, virtualization as a separation layer has introduced a profound challenge to that premise. Moreover, given the progress of Moore's law and the large number of VMs (virtual machines) we can run per server, the implicit change to networking is that the last-hop switch is necessarily a feature of the hypervisor or hardware of the server and not a traditional hardware switch in the physical network.

NEASE We've broken a paradigm. Because the core can now handle so many workloads, dedicated network devices are not being asked to solve the same problem. Networks in the past have reinforced the concept that user equals device, equals port, equals location. With virtualization, those identity relationships are now dead. Networking will need to evolve as a result.

CASADO Networks have always been built in support of some other higher-priority requirements, and as a result people have never been required to produce good stand-alone network architectures. If I'm building an operating system that people use as a platform for applications, I must have nice application platform abstractions. Networks have never had to do that.

Originally, the leverage point was in the network because it was central. Because of this, networks have always been an obvious place to put things such as configuration state. Now the leverage point is at the edge because the semantics there are very rich. I know where a VM is, I know who's on it, and I know when it joins and when it leaves. As a result, I don't require traditional service discovery and often don't need multicast. Because the leverage point is at the edge, the dynamic changes completely; and because the semantics are now more interesting at the edge, you have a clash of paradigms.

NEASE We've seen the same process take place with blade servers. When we started centralizing some of the functions that used to be distributed among many servers, we could easily and authoritatively know things that used to be difficult to obtain. Things—for example, in station state, address, user, etc.—became much easier to determine because the new blade architecture made it very convenient.

CASADO Most scalable clouds are subnetworked. They are architected to not deal with VLANs (virtual LANs) or do flat anything and are not going to try and do one big L2 domain. It's all subnet'ed upfront with the cloud overlaid on top.

Networks do not have virtualization built in, and VLANs are not virtualization in a global sense. It's easier just to treat the entire network infrastructure as a big dumb switch and hold the intelligence at the edge because that is where the semantics are.

CROSBY Networking vendors sell differentiated networking value propositions to their customers. As IaaS (Infrastructure as a Service) comes into wider use, APIs will shift. If I have an investment in IT skill sets to manage Juniper equipment in my private data center, can I use those skills, configurations, and policies off premises in the cloud?

IaaS challenges the traditional vendor/customer roles for networking equipment. It may be that the cloud vendor purchased equipment from a specific vendor, but there is no way for that vendor to surface its unique value proposition to the IaaS customer. Does this necessarily force commoditization in network equipment? I think it does. Google, for example, reportedly already builds its own networking gear from industry-standard parts.

REDDY In the next two to three years our goal is to make the building of an application, its packaging, and deployment completely transparent. I want to specify SLA (service-level agreement), latency, and x-megabit-per-second throughput and receive a virtual network that satisfies the requirement. I don't care if it's Cisco, Juniper, or whatever. What I want is a service provider that constructs and delivers the network that is required. As the end user, I care about only the above-the-line result.

CROSBY Deciding whether to buy an HP or a Juniper switch is a localized problem of scale, complexity, IC integration, etc. To outsource that work, I will have to go to a large number of cloud vendors and attempt to determine for my combined networking and compute problem how to

choose the best vendor. That's way too hard if I already have a strong preference for and investment in a particular vendor's equipment, value, and management.

Alternatively, if we pursue a "genericized" feature list on the part of virtualized networks, we make it difficult for application owners to support their rich needs for control. After all, the richness of a switch/router feature set is designed specifically to address customer needs. If we genericize those features, we may not be able to support features in the cloud rich enough to meet customer needs.

NEASE That's looking at cloud computing in its infancy. How does someone decide on HP versus Juniper? Two or three vendors will come in and say, "I know the environment and I know what you're trying to run. Based on what you need, here are the things I'd recommend."

Cloud vendors are not going to become a monopoly overnight. It will evolve over some meaningful period of time, and eventually a few major winners will emerge, depending on area.

CASADO We all agree that if you take a slider bar and move it fully to the right to "future," you're going to have some massive consolidation, with a few large vendors left standing. The question is how long will that slider bar take to get to the end result?

CROSBY I talk to CIOs who are already telling their employees that that there will be no net new servers, and any new server purchases will require their sign-off. This has motivated operations teams to find ways to rent server cycles by the hour.

A key opportunity arising from virtualization and the cloud is to enable CIOs to address the labor challenges of today's owned infrastructure. CIOs will absolutely take advantage of every opportunity to outsource new workloads to hardware they do not have to purchase and that is automatically provisioned and managed without expensive labor costs, provided that key enterprise requirements—SLAs and security and regulatory compliance—can be met.

TAVAKOLI One of the things that virtualization buys you is homogeneity. When you run on top of a hypervisor, you don't really care what the drivers are; you are relying on the system. We have to get to the same degree of homogeneity on the network side. The question is both economic and technical: Who is best positioned to solve the massive network management problem?

You could take virtual switches and, as Arista has done, stitch them into your existing environment, while leaving the virtual switch in place. You can take the Cisco approach of replacing that virtual switch with your own virtual switch and pull everything back to an aggregation point. At Juniper, we want to build what is in effect a stateless, high-capacity, 100,000-port switch but without backhauling everything to the "god box" in the middle.

NEASE We are talking about taking a network function, decomposing it into constituent parts, and choosing a different place to implement those parts. One of the parts is "figure out where to send this," and that part gets separated into multiple parts depending on the specific attributes of your existing physical infrastructure. Owning your own assets is still going to make sense for some time to come because the complexity of your existing data-center infrastructure will restrict what you can actually hire someone else to do for you as a service.

TAVAKOLI There are solutions today that will work in a garden-variety 5,000-square-foot data center. They take a general approach and are not as concerned about things such as end-to-end latency. That's one approach. You can also take a more specialized approach and address customers that have very specific needs such as latency sensitivity.

REDDY There are two management perspectives in addressing this issue. One is at the infrastructure

level: the service provider who cares about the networks. The other is about the services received over the wire. They don't care about the network; they care about service availability and response time. From the service virtualization perspective, I need to see everything in a holistic way: network, storage, computing, service availability, and response time.

CASADO As you consume the network into the virtualization layer, you lose visibility and control of those components. We are just figuring out how to get it back. Networks are being consumed into the host and they're losing control. We have a set of practices and tools that we use to monitor things, provide security, and do trending, and that information is now more accessible to the guy who runs the host than to the guy who runs the network.

CROSBY Let's make it real. In a medium-size enterprise I have a LAN segment called LAN A with an IDS (intrusion detection system), and a different LAN B segment with no IDS. If I have an application running in a VM and move it from a server residing on LAN A to a server on LAN B, then I have a problem.

NEASE No, you move it only to a segment that supports the physical service. That's how networks are rendered.

CROSBY The key point is that you don't have the luxury of being asked when a VM moves; you are told. The argument that Lin (Nease) makes is that we would never move a thing to a LAN segment that is not protected. People usually don't understand the infrastructure at that level of detail. When the IT guy sees a load not being adequately serviced and sees spare capacity, the service gets moved so the load is adequately resourced. End of story: it will move to the edge. You are not asked if the move is OK, you are told about it after it happens. The challenge is to incorporate the constraints that Lin mentions in the automation logic that relates to how/when/where workloads may execute. This in turn requires substantial management change in IT processes.

TAVAKOLI You can have a proxy at the edge that speaks to all of the functionality available in that segment.

CROSBY The last-hop switch is right there on the server, and that's the best place to have all of those functions. Moving current in-network functions to the edge (i.e., onto the server) gives us a way to ensure that services are present on all servers, and when a VM executes on a particular server, its specific policies can be enforced on that server.

CASADO This conversation gets much clearer if we all realize that there are two networks here: a physical network and one or more logical networks. These networks are decoupled. If you do service interposition, then you have to do it at the logical level. Otherwise you are servicing the wrong network. Where you can interpose into a logical network, at some point in the future these services will become distributed. When they become a service, they're part of a logical network and they can be logically sized, partitioned, etc.

Today, services are tethered to a physical box because of the sales cycle, because someone comes in and says, "I've got a very expensive box and I need to have high margins to exist." As soon as you decouple these things, you have to put them into the logical topology or they don't work. Once you do that, you're untethered.

NEASE But once you distribute them, you have to make sure that you haven't created 25 things to manage instead of one.

CASADO You already have the model of slicing, so you already have virtualization; thus, nothing changes in complexity. You have the exact same complexity model, the exact same management

model.

NEASE No, if I can get problems from more than one place, something has changed. Think of virtual switching as a distributed policy enforcement point. It is not true, however, that distributed stuff is equal in cost to centralized stuff. If distributed stuff involves more than one way that a problem could occur, then it will cost more.

CASADO It would have to be distributed on the box. If you're going to inject it into one or more logical topologies, then you will have the same amount of complexity. You've got logically isolated components, which are in different default domains.

If people want the dynamics and cost structure of the cloud, they should either (a) not invest in anything now and wait a little while; or (b) invest in a scale-out commodity and make it look like Amazon. If they do not take one of these two paths, then they will be locked into a vertically integrated stack and the world will pass them by.

CROSBY The mandate to IT is to virtualize. It's the only way you get back the value inherent in Moore's law. You're buying a server that has incredible capacity—about 120 VMs per server—that includes a hypervisor-based virtual switch. You typically have more than one server, and that virtual switch is the last-hop point that touches your packets. The virtual switch allows systems administrators to be in charge of an environment that can move workloads on the fly from A to B and requires the network to ensure that packets show up where they can be consumed by the right VMs.

NEASE The people who will be left out in the cold are the folks in IT who have built their careers tuning switches. As the edge moves into the server where enforcement is significantly improved, there will be new interfaces that we've not yet seen. It will not be a world of discover, learn, and snoop; it will be a world of know and cause.

CROSBY The challenge for networking vendors is to define their point of presence at the edge. They need to show what they are doing on that last-hop switch and how they participate in the value chain. Cisco, for example, via its Nexus 1000V virtual switch, is already staking a claim at the edge and protecting its customers' investments in skill sets and management tools.

BEELER If I manage the network within an enterprise and am told that we just virtualized all our servers and are going to be moving VMs around the network to the best host platforms, then as network manager, since I do not have a virtualized network, this causes me problems. How do I address that? How do I take the IDS that I have on my network today, not of the future, and address this problem?

CASADO You either take advantage of the dynamics of the cloud, which means you can move it and you do scale out, or you don't. In this case you can't do these VM moves without breaking the IDS. The technology simply dictates whether you can have a dynamic infrastructure or not.

REDDY My server utilization is less than 10 percent. That number is not just CPU utilization—memory and I/O bandwidth are also limited because there are only two network cards on the each server box. All my applications are very network-bandwidth intensive and saturate the NICs (network interface cards). Moreover, we also make a lot of I/O calls to disk to cache content. Though I have eight cores on a box, I can use only one, and that leaves seven cores unutilized.

NEASE It seems like you would benefit from an affinity architecture where the communicating peers were in the same socket, but that sometimes requires gutting the existing architecture to pull off.

TAVAKOLI From our perspective, you want a switch without those affinity characteristics so you don't have to worry about "same rack, same row" to achieve latency targets. You really want a huge

switch with latency characteristics independent of row and rack.

REDDY It is more of a scheduling issue. It's about getting all the data into the right place and making best use of the available resources. We need to schedule a variety of resources: network, storage, computational, and memory. No algorithms exist to optimally schedule these media to maximize utilization.

TAVAKOLI You want an extremely holistic view. You are asking where you put a VM based on the current runtime context of the hypervisor so you can maximize utilization and minimize contention for all aspects of data-center operations, such as CPU, network, storage, etc.

NEASE You have to understand that the arrival of demand is a critical component to achieving this optimization, and it is not under your control.

REDDY At Yahoo! we built a traffic server proxy that is open sourced and has knowledge and intelligence regarding incoming traffic from the network edge. The proxy characterizes and shapes incoming traffic, and routes it appropriately.

CASADO This approach works best when linked with pretty dumb commodity switches, high fan-out, and multipath for load balancing. Then they build the intelligence at the edge. This is the state of the art.

It does not matter where the edge is in this case. Whether you enforce it at the vSwitch, the NIC, or the first-hop switch, the only thing that matters is whose toes you step on when you exercise control. The definition for the edge is the last piece of network intelligence. How that translates to a physical device— an x86, a NIC—depends on how you want to set up your architecture.

REDDY When a Yahoo! user sends a request from Jordan, a domain address maps to his IP address. Once he lands on the server, DNS (Domain Name System) is used to determine that he is coming from Jordan. I can then map this traffic to the edge server located in the data center serving the Middle East. That's your entry point and that is your edge router.

The traffic then goes to the Apache proxy, which is a layer seven router that we built. It determines that since the user is coming from Jordan, we should route service to our data center in Singapore, or in Switzerland. The traffic never comes to the U.S.

Depending on how I want to do traffic shaping, this architecture allows me to change my routing policies dynamically and route traffic to the UK, Switzerland, or Taiwan as needed. I can do all this through layer seven routing (Apache proxy layer).

TAVAKOLI This is a different solution to the problem of large data centers and virtualization in the cloud. If routing intelligence needs to move up the stack to layer seven, then by definition you're going to disenfranchise layers two and three from a bunch of policy decisions. As you move up the stack it becomes more of a general-purpose kind of application, with general-purpose processors being better suited for that type of work.

If a decision point requires that you pick something specific out of an XML schema or a REST (representational state transfer) body, then the intelligence needs to be there. The distribution of enforcement needs to be closer to the edge for it to scale.

Where precisely that edge is, whether it's on the last-hop physical switch, the NIC, or the vSwitch, is almost beside the point. With something like VEPA (Virtual Ethernet Port Aggregator), you could aggregate that up one hop and it would not significantly change the argument. The issue is about what you can ascertain from layers two and three versus what you need to ascertain from a much higher context at the application level.

REDDY This was the best we could do given our current level of virtualization infrastructure. How do I get from where I am to take full advantage of the current state of virtualization? I want to move my distribution function from where it currently resides at the edge of the network to its core. I want to build a huge switch fabric on the hypervisor so I can localize the routing process within that hypervisor.

If you look at the enterprise applications, there is a relationship between multiple tiers. We built a deployment language that characterizes the relationship between each tier: the application tier, which is latency sensitive; the application-to-database tier, which is throughput sensitive; and the database-storage tier, which is again throughput sensitive. We then internally built a modeling architecture to characterize this information so that it can be used effectively during operations.

CASADO Where you enforce is a complete red herring. What this says to me is that because Surendra's (Reddy) data-path problems are being driven by applications, he really has to open them up. To solve this problem, you want to be able to dictate how packets move. To implement this, you will need to control your flow table.

NEASE The issue here is that a network is a single shared system, and it won't work if an individual constituent tries to program it. It has to be the center that is programmed. Effectively, it comes down to replacing the vendor's view of the protocols of importance.

CROSBY Hang on. You intend to sell a switch to cloud vendors? If that is true, every single tenant has a reasonable expectation that they can program their own networks—to implement policies to make their applications work properly and to protect them.

NEASE No, it's the service provider that programs the network on behalf of the tenants.

CROSBY If I'm a tenant of a virtual private data center, I have a reasonable right to inspect every single packet that crosses my network. Indeed, I might have a regulatory obligation to do precisely that or to run compliance checks, network fuzzing, etc. to check that my systems are secure.

CASADO This gets to become a red herring. People who are building efficient data centers today are overlaying on top of existing networks, because they can't program them. That is just a fact.

TAVAKOLI We have done an implementation of Open Flow on our MX routers that allows for precisely what you're talking about. Though not a supported product, it does provide a proof of concept of an SDK (software development kit) approach to programming networks.

NEASE There's a contention over who's providing the network edge inside the server. It's clearly going inside the server and is forever gone from a dedicated network device. A server-based architecture will eventually emerge providing network-management edge control that will have an API for edge functionality, as well as an enforcement point. The only question in my mind is what will shake out with NICs, I/O virtualization, virtual bridges, etc. Soft switches are here to stay, and I believe the whole NIC thing is going to be an option in which only a few will partake. The services provided by software are what is of value here, and Moore's law has cheapened CPU cycles enough to make it worthwhile to burn switching cycles inside the server.

If I'm a network guy in IT, I better much more intensely learn the concept of port groups, how VMware, Xen, etc. work, and then figure out how to get control of the password and get on the edge. Those folks now have options that they have never had before.

The guys managing the servers are not qualified to lead on this because they don't understand the concept of a single shared network. They think in terms of bandwidth and VPLS (virtual private LAN service) instead of thinking about the network as one system that everybody shares and is way oversubscribed.

REDDY We are moving to Xen and building a new data-center architecture with flat networks. We tried to use VLANs, but we have taken a different approach and are going to a flat layer 2 network. On top of this we are building an open vSwitch model placing everything in the fabric on the server.

My problem is in responding to the service requirements of my applications and addressing things such as latency and throughput. The data needed to address these issues is not available from either a network virtualization solution or the hypervisor.

Also, my uplink from the switches is 10 gigabits per second or multiple 10 gigabits per second, but my NICs are only one gig. If I run 10 VMs on a box, then all of the bandwidth aggregates on one or two NICs.

NEASE You guys are cheap. If you went to a backplane, then you would get a lot more bandwidth out of those servers. A KR signal on a backplane is how you get a cheap copper trace for 10-gigabit service.

CASADO Going forward, a new network layer is opening up so you can take advantage of virtualization. Traditional networking vendors certainly do not control it today and may not control it in the future. The implications are that it may not matter what networking hardware you purchase, but it may matter much more what network virtualization software you choose.

If you like the cost point and the service and operations model of the cloud, then look at Eucalyptus, Amazon, Rackspace, etc. and see how they build out their infrastructure. Today that is the only way you can get these types of flexibility and per-port costs.

It would be interesting to compare a vertically integrated enterprise with something like Amazon EC2 (Elastic Compute Cloud) in terms of cost per port and efficiency.

BEELER The guys who run infrastructure for Google told us that the difference between running their own infrastructure and running their stuff on Amazon was small enough that it really made them think about whether they wanted to continue to do it themselves.

CASADO We have seen close to two orders of magnitude difference between a vertically integrated solution and something like EC2.

BEELER The relevance here is that while these issues may not affect you today as a practitioner, you should understand them because they will affect you tomorrow. In this way you can make intelligent investments that will not preclude you from taking advantage of these kinds of benefits in the future.

CASADO The leverage point for vertical integration has always come from the networking vendors. It was lost a long time ago on the servers. Someone who offers a full solution is going to be a networking vendor. If you're making a purchasing decision, then you don't have to blindly follow the legacy architectures.

I do not believe that owners of existing network infrastructure need to worry about the hardware they already have in place. Chances are your existing network infrastructure provides adequate bandwidth. Longer term, networking functions are being pulled into software, and you can probably keep your infrastructure. The reason you buy hardware the next time will be because you need more bandwidth or less latency. It will not be because you need some virtualization function.

TAVAKOLI We get caught up on whether one is implementing a new data center from scratch or is incrementally adding to an existing one. For a new data center, there are several things to keep in mind. Number one, if you're planning for the next five years, understand how you are going to avoid focusing on "rack versus row versus data center." Architect the data center to minimize location-dependent constraints but still be able to take advantage of location opportunities as they arise.

Also, have a strategy for how you can still obtain a top-level view from all the independent edge-based instances. This is especially critical in areas such as security, where you need a global view of a multiple point attack. If there's an attack that consists of multiple events that are all below individual thresholds, then there's still some correlation required up top to be able to recognize it as an attack. You cannot get away with saying that these are distributed, independent problems at the edge and that no correlation is required.

NEASE You will never remove the concept of location from networking. It will always be part and parcel of the value proposition. Bandwidth consumption will be rarer the farther you span, and latency will be shorter the closer you are. Physical location of networking resources never completely goes away. The network is never location independent and always has a component of geography, location, and physics. You cannot separate them.

CREEGER The issues involved in network virtualization are moving quickly. Mass interest in virtualizing the data center is breaking a lot of traditional physical versus logical bounds. You need to concentrate on what you're trying to achieve in your data center and what key properties you're trying to preserve. If you do decide to virtualize, do an internal cloud, or subcontract out to an external cloud vendor, then you need to parallel your architecture closely to industry leaders such as Amazon so you keep close to current accepted practices. Additionally, to avoid breakage between physical and virtual devices, you want to minimize functionality and performance investments that require device-specific configuration of the physical infrastructure. As virtual devices become more prevalent, those device-specific configurations will become more of a burden.

Some network vendors are offering products under the banner of network virtualization that provide virtual implementations of their physical devices. I believe they are being offered to preserve a customer's investments in legacy infrastructure. By preserving the status quo, however, it will be that much more difficult to take advantage of new, more efficient and functional architectures as they gain broader acceptance. The advice here suggests that you keep things simple. Avoid investing in vendor-proprietary functions, and wait to see what new architectures emerge. Once you identify these new architectures, invest conservatively as they gain acceptance.

CROSBY The key point is to be aware that your networking competence investments are going to shift radically. The new network will be automated, aware of the current locus of the workload, and dynamically reconfigure the infrastructure as the workload migrates or scales elastically.

TAVAKOLI An opportunity exists to implement very fine-grained, high-quality enforcement at the very edge of the network, including on the host. That will have to be stitched into your service model. You can scale and distribute your control to the very edge of the network, which now is on the hosts. The question is, who ends up driving the overall policy decision?

NEASE If you're a network person and you're not touching VMware and find yourself not needed, you have to ask yourself whether or not your skill set is not needed as well. The network edge has moved, and if you are not architecting the network inside the server, then your skill set may not matter.

BEELER The good news is that some systems administrators don't have a clue about networking. This is an opportunity for network engineers still to add value in the new virtualized world. Q

LOVE IT, HATE IT? LET US KNOW

feedback@queue.acm.org