WHITE PAPER

Meeting Application Delivery Requirements with Server Load Balancing

Sponsored by: Cisco

Lucinda Borovick
October 2011

EXECUTIVE SUMMARY

Today's enterprise network is being asked to handle more requirements than ever before — and the cost of failure is also greater than it has ever been in the past. The network lies at the heart of the IT infrastructure, linking all server, storage, and other infrastructure components in the datacenter, and any downtime or performance glitches can have significant impacts on the business as a whole.

Server load balancing has been at the heart of ensuring network robustness and scalability for over a decade. Having evolved from their early days of the first dot-com boom in which they were used primarily to provide resilience to companies' x86-based Web architectures, server load balancers now include the enterprise functionality necessary to support mission-critical enterprise applications.

Today, new initiatives such as virtualization and next-generation datacenters are placing an increased burden on the network in terms of bandwidth requirements, reliability, scalability, and resilience, and network managers are struggling to keep up. Network engineers require solutions that easily fit into their network architecture, are highly manageable, and help make them more efficient and reduce their cost of operations.

Network managers require solutions that are sufficiently robust and scalable that they can be trusted to be inserted into the network core/DMZ. They prefer a product with sufficient hardware and software performance to meet the highest performance and scalability requirements.

Server load balancing remains an important technology to support these initiatives, is a fundamental building block in datacenter design, and should be included in any end-to-end advanced network services deployment. Cisco has been one of the leaders in providing server load balancing since the technology's inception, and with its Application Control Engine (ACE) technologies, it continues to bring datacenter-ready solutions to the market.

To better understand the challenges faced by today's network managers, Cisco commissioned IDC to perform a study of U.S. and European IT executives. For this study, IDC surveyed 408 technical decision makers and datacenter managers from enterprises and service providers. This study was supported by 20 in-depth interviews of IT executives from enterprises and service providers located in the United States and Europe. These in-depth interviews consisted of both Cisco customers and non-Cisco customers.

SITUATION OVERVIEW

Key Datacenter Design Themes

The datacenter is in the midst of a transition that is reshaping requirements for the enterprise network. As applications become increasingly mission critical, requirements for network throughput, availability, and security become more important than ever, while new application initiatives push network engineers to continue evolving their network and its capabilities.

The new network, at its foundation, has a number of important trends, including the transition to virtualized IT, the increasing use of single-rack (POD) approaches for new deployments, and an increased focus on total cost of ownership (TCO) analysis.

The Transition to Virtualized IT

Virtualization is one of the most significant trends that affected IT organizations over the past five to six years. It has helped enterprises reduce physical server sprawl, drive consolidation, and increase server utilization, which allows them to reduce up-front capital and operating costs, extend the life of the datacenter, and drive greater levels of application availability.

This trend has accelerated in recent years to the point where the worldwide installed base of physical servers has largely leveled off since 2008, while the virtual server installed base continues to grow at more than 30% per year. IDC believes that in 2009, for the first time, more new application instances were deployed on virtual machines (VMs) than on dedicated physical servers, and IDC expects that the installed base of virtual servers will be equal to the installed base of physical servers by the end of 2011 (see Figure 1).

Worldwide Server Shipments and Installed Base
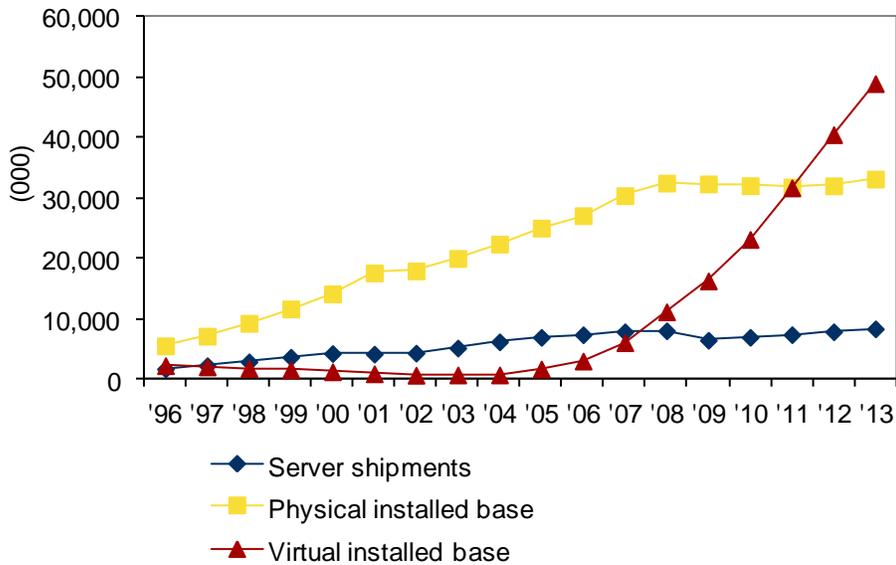
Source: IDC, 2011

The growing importance of server virtualization is creating challenges for the network, and the rapid adoption of virtualization is happening in spite of — not because of — the flexibility, or lack thereof, in customers' networks. Until recently, few network equipment vendors offered support for virtualization in a standardized, holistic way. "Virtualization will have the greatest impact in the networking environment, where you can have virtual switches, combined switches," said one European manufacturing company interviewed for this project. "This has a major impact."

Illustrating this point, IT managers surveyed for this project ranked virtualization at the top of their datacenter network priorities. Their first priority was investing to implement VM-aware network switches, followed by investing to deploy VM-specific advanced-level network services (see Figure 2).

"Virtualization will have the greatest impact in the networking environment, where you can have virtual switches, combined switches. This has a major impact."

Datacenter Network Priorities

Q.   As you think about your current datacenter network and any new major initiatives or new products you are currently planning or implementing, please indicate which of the following investments are a priority using a 1 to 7 scale, where 1 = not at all a priority and 7 = a high priority.
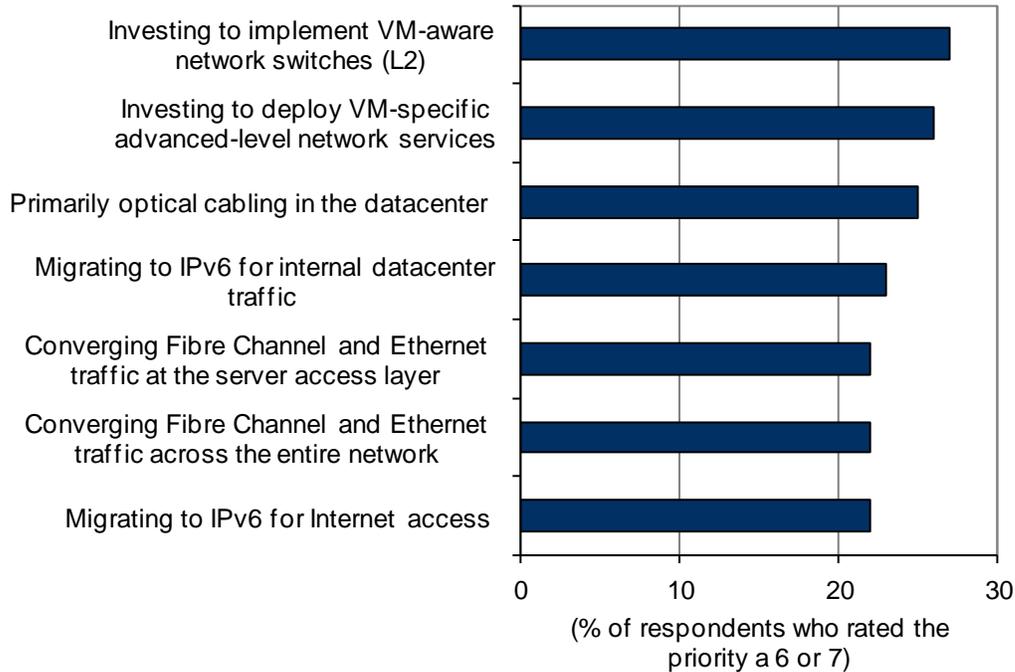
n = 408

Source: IDC, 2011

Incremental Pay-as-You-Grow Approach for New Deployments

In the past, IT organizations needed to overprovision their IT infrastructure purchases in compute, storage capacity, and network bandwidth. It was the only strategy available to hedge against unforeseen shifts in IT demand. Further, the purchases were made in separate, independent "silos." Network deployments often did not take into account server and storage needs, and vice versa, and the three components frequently did not scale in tandem. A better approach, which is followed by an increasing number of IT organizations today, is to install server, storage, and networking resources in "PODs"; each POD has "just the right amount" of server, storage, and networking resources in a dedicated rack unit, all designed to work together. This approach allows organizations to scale their IT infrastructure in a balanced manner so that the amount of compute power scales hand in hand with the amount of storage and networking resources required while ensuring that each of these infrastructure elements works in concert. This improves the IT infrastructure's flexibility and agility and enables organizations to rip and replace resources as required.

Focus on TCO Analysis

In the past, an IT organization's core charter consisted of defining and understanding the enterprise's computing needs and implementing and maintaining solutions that satisfied those needs. From a budgetary perspective, IT's primary goal was keeping up-front (mainly capital) expenses as low as possible. Ongoing operating costs were a relatively small part of the IT organization's budget. But in the past 10 years, the operating costs of managing and maintaining a server network — predominantly personnel management costs and power and cooling costs — have grown to the point where organizations now spend more on server management than on new server acquisitions. Given the condition of the current global economy, enterprises are looking to cut costs wherever they can, and IT departments must now strongly consider the entire lifetime cost of ownership for any new technologies before they integrate them into their infrastructure.

Customer Challenges

As always, network managers continue to face a number of challenges in managing their network infrastructure. Some of these are "tried and true" challenges that have been around for some time, while others are new and emerging.

Tried-and-True Challenges

Even as network managers work to evolve their infrastructures to meet today's changing needs, they must always keep in mind a number of core requirements. The datacenter is the nerve center for the enterprise and the core of IT investment. At the end of the day, the enterprise must ensure that the network can provide high performance and availability for the many mission-critical applications run in the datacenter. "Our number one guiding principle is reduced latency, with availability as number two," said one U.S. financial exchange. "That is why we have a flat architecture without any single points of failure."

"Our number one guiding principle is reduced latency, with availability as number two. That is why we have a flat architecture without any single points of failure."

The heart of the traditional challenges are reliability, availability, and security:

▱ **Reliability.** As organizations continue to rely on enterprise applications to run their core business functions, the reliability of the network is critical. Key network features must support business continuity even in the face of errors. A U.S. telecom operator interviewed for this project stated, "Consumer expectations for delivering quality and response time are just getting higher and higher, which places an additional burden on us to get it right." Further, reliability was rated the most important product attribute in a server load balancer in IDC's global survey of IT managers (see Figure 3).

"Consumer expectations for delivering quality and response time are just getting higher and higher — which places an additional burden on us to get it right."

Important Product Attributes

Q. *On a scale from 1 to 7, where 1 = not at all important and 7 = highly important, please rate*
   the importance of the following product attributes in selecting server load balancing
   solutions.
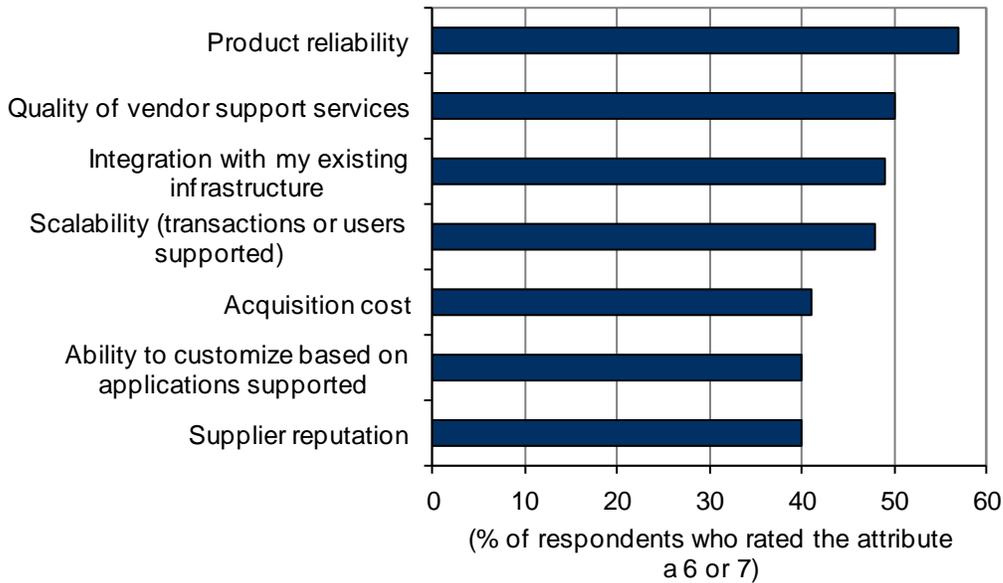
n = 408

Source: IDC, 2011

☒ **Availability.** The ability of the infrastructure to continue processing workloads, even in the face of faults or failure of individual components, is of fundamental importance. Ensuring the attainment of agreed-upon service levels, such as uptime levels for critical applications, is key to supporting core business processes for employees, customers, and partners. As one European insurance company IT executive put it, "Our guiding principles for the network are high availability, high performance, every VLAN everywhere being stable."

"Our guiding principles for the network are high availability, high performance, every VLAN everywhere being stable."

☒ **Security.** Security is one of the top issues that keep CIOs awake at night. As more data comes online, an increasing amount of sensitive information, such as customer and supplier information, costs and prices, contracts, and intellectual property, is at risk. Organizations are increasingly taking an end-to-end approach to their information security and as such are requiring "security aware" datacenter network infrastructure. One of the key security challenges specific to virtualized environments is enforcing security policies at the granular level of individual VMs or groups of VMs, and organizations should ensure that they have the appropriate security devices in place to do so.

New and Emerging Challenges

In addition to dealing with the time-honored challenges of reliability, availability, and security, network managers must consider a number of new and emerging challenges, including increasing resiliency, providing network support for siloed virtual and physical devices, reclaiming stranded capacity, and issues surrounding IT organizational structure:

⊡ **Increasing resiliency.** Not only must the network be able to heal itself in the face of failures, but now this capability must be extended to the entire network. In the past, datacenters were more likely to segregate their tier 1 applications and house them on their highest-performance, most reliable infrastructure. Today, as virtualization enables IT organizations to consolidate their applications, *every* server in the datacenter has become mission critical. It is no longer enough to provide high resiliency for a set of tier 1 servers; now datacenter managers must ensure that all applications are highly resilient in the face of failure. And the fact that today's global enterprise is more likely to have federated applications housed in multiple datacenters places further emphasis on the need to provide resiliency throughout the entire network.

⊡ **Supporting siloed virtual and physical services.** IDC has spoken with many datacenters that have "compartmentalized" the servers they use to handle virtualized workloads from their dedicated workload servers. Unfortunately, many datacenters that pursue this approach are unable to share network bandwidth and network services between the types of servers. It's not acceptable to maintain separate physical and virtual server pools and not be able to treat the whole network architecture as a single unit. The network needs to see all nodes and to be able to treat them and implement policies on them individually, with unified management and orchestration, regardless of whether it is communicating with virtual or physical dedicated servers.

⊡ **Reclaiming stranded capacity.** One of the great promises of virtualization is the ability to reduce stranded capacity; that is, to shift workloads to servers with available capacity. This allows datacenters to "rightsize" the number of servers they support and avoid having to overprovision server resources. While some ability to balance workloads across available servers is provided in most virtualization technologies, server load balancing can play a crucial role by understanding the bandwidth available in particular servers and if one server is over its capacity threshold, to automatically shift traffic to a server that is not busy. Server load balancing can also help reclaim stranded capacity even in a nonvirtualized cluster environment by shifting traffic to the servers in the cluster best able to meet the demand.

⊡ **IT organizational structure.** The reality of many IT organizations today is that departmental hierarchies have been built up around technology expertise. It is common for IDC to speak with IT professionals who refer to themselves as part of the "network", "server," or "storage" group. While this approach makes sense from a troubleshooting and operational standpoint, it limits the IT department's ability to respond quickly to the business requirements. Similarly, it points to the need for products to have built-in "shared" management functionality, enabling the different groups to utilize specific components in the datacenter.

Benefits of a Comprehensive Application Delivery Strategy

The goal is to create a holistic datacenter in which server load balancing is used to migrate workloads across the datacenter and provide a comprehensive application delivery strategy. There are a number of benefits to implementing such a comprehensive application delivery strategy, including:

- ☑ **Higher resiliency, security, and control.** This will allow the datacenter to be more secure as control is centralized from an application traffic standpoint. A European retailer emphasized this point, stating that "server load balancers offer an extra level of security, scalability, availability, and help with maintenance," while a European financial services firm said that "the business benefit of a server load balancer is high availability and ease of server maintenance."

- ☑ **Ability to deploy new services applications faster.** IT organizations want to be able to respond quickly when a request comes from the business for a new application or to support new regulatory or emerging security threats. By having an overall standardized approach, IT organizations have the ability to provision a new application very quickly. This benefit was emphasized by an IT executive at a U.S. energy company: "In our environment, server load balancing is used by the network team to figure out what the application guys want and then make it work for them."

- ☑ **Reduced operational costs.** Datacenters that implement server load balancing can facilitate consolidation and virtualization. For many customers, this can lead to fewer physical appliances, lower power and cooling requirements, and fewer staff needed for administration.

Interestingly, organizations interviewed for this study stated that they used their server load balancers in a variety of ways, and several even referred to it as the "Swiss Army knife" of their network infrastructure. While most used them to scale their network, increase resiliency, and support virtualization, others also used them for more exotic uses such as diagnosing and troubleshooting network problems. Figure 4 demonstrates the breadth of features (and use of those features) respondents are looking for in server load balancers. Performance and scale is by far the most important feature for customers. The server load balancer brings performance and scale to the overall datacenter and enables the network team to adjust to rapidly changing demand in the most efficient and effective manner. Interestingly, as IT organizations begin to come together to form centralized consolidated teams across server, network, and storage, the need for role-based management is high, with over one-third of respondents rating it a top feature.

> "Server load balancers offer an extra level of security, scalability, availability, and help with maintenance."

> "In our environment, server load balancing is used by the network team to figure out what the application guys want and then make it work for them."

FIGURE 4

Feature Importance in Server Load Balancers

Q. Please rate the value of the following server load balancing features on a scale of 1 to 7, where 1 = not at all valuable and 7 = extremely valuable.

(% of respondents who rated the feature a 6 or 7)
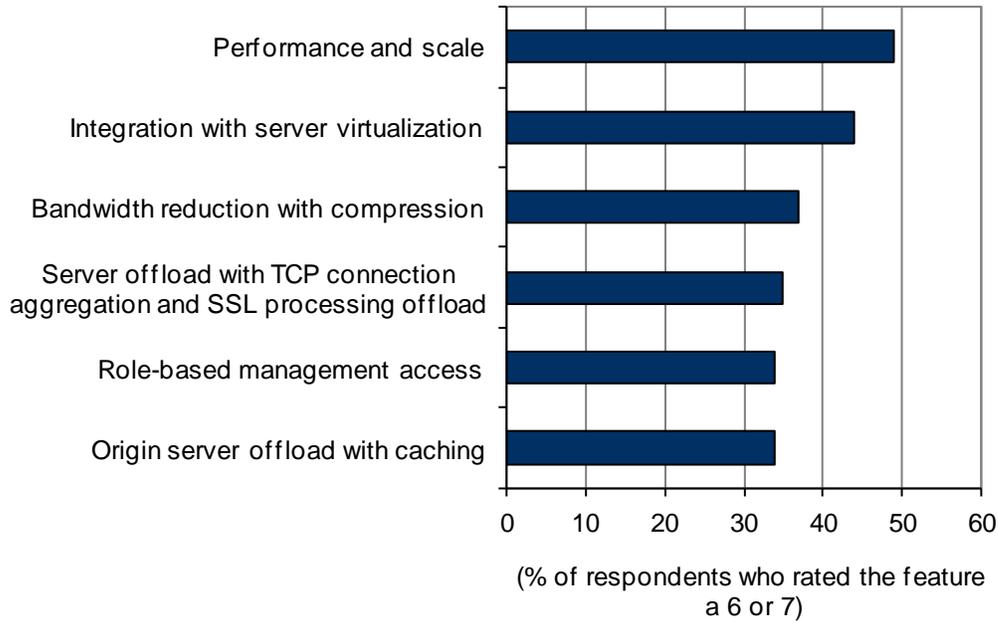
n = 408

Source: IDC, 2011

CISCO APPLICATION CONTROL ENGINE (ACE)

Cisco Application Control Engine (ACE) technologies are designed to increase availability, acceleration, and security of datacenter applications. The Cisco ACE family of products includes the ACE Service Module for the Cisco Catalyst 6500, Cisco 7600 routers, and the ACE 4710 standalone appliance. These switches are designed to enhance application availability, accelerate application performance, and help secure the datacenter and mission-critical application from attacks.

The Cisco ACE portfolio is designed to allow datacenters to accomplish a number of goals, including:

☐ **Increase application availability.** Cisco ACE includes a failover system with application health probes to ensure that traffic is forwarded to the most available server. ACE is also integrated with the Cisco Global Site Selector (GSS) to provide connection failover between datacenters and help ensure business continuity.

- ☑ **Secure datacenter and applications.** Cisco ACE is designed to provide the last line of application defense against application threats and denial of service (DoS) attacks with deep packet inspection; normalization of application traffic for specific protocols including TCP, HTTP, DNS, and LDP; and network security.

- ☑ **Facilitate datacenter consolidation through virtualization.** Support for virtualized architecture is a primary design element of Cisco ACE. IT managers can configure virtual devices that are isolated from one another, with the ACE Service Module supporting 250 virtual devices and the ACE 4710 appliance supporting up to 20 virtual devices. By implementing virtualization, organizations can reduce the number of servers, load balancers, and datacenter firewalls they must support, which in turn reduces operating costs, power and cooling needs, and management complexity.

- ☑ **Accelerate application performance.** Cisco ACE is designed to accelerate the end-user experience for all users, whether they are working from the office or remotely. ACE includes a range of acceleration capabilities to improve application response time, reduce bandwidth volume, and improve efficiency of protocols. Included technologies are hardware-based compression, data encoding, caching, SSL and TCP processing, and Flash Forward.

- ☑ **Enable dynamic workload scaling.** A key part of Cisco ACE is the Application Networking Manager (ANM), a centralized management tool that improves VM awareness and provides integration capabilities with VMware vCenter. Topology mapping provides continuity of operations between application and server administrators and enables real-time monitoring, configuration, and operation of virtual machines.

Cisco ACE integrates a broad set of intelligent Layer 4 load balancing and Layer 7 switching technologies with leading-edge virtualization and security capabilities. With its support of virtualized architecture and role-based administration, it helps streamline the effort involved in implementing, scaling, accelerating, and protecting applications.

Cisco ACE provides role-based system administration so that multiple departments or stakeholders can manage individual portions of the network. This also enables IT organizations to implement a range of applications from a single appliance, streamlining the process of application provisioning throughout the datacenter. The ACE 4710 comes in a one-rack unit (1RU) form factor, was built using highly available system software and hardware, and supports traffic loads of up to 4Gbps.

Cisco ACE ANM software integrates with the VMware virtual datacenter environment to support VMware operations and management. It provides service visualization with a GUI-based tool that administrators can use to visually navigate maps of the server topology and find, view, and zoom in on items of interest. It supports delegation of SSL services, provides statistical data that can be used for planning baselines, and enables automatic backup of Cisco ACE configurations according to user-defined schedules.

IDC believes that one of the most compelling attributes of ACE is the ability it provides to virtualize and segment traffic in the datacenter, all from one physical appliance. With Cisco ACE, instead of having to buy multiple server load balancers (one for email, one for database traffic, etc.), organizations need to purchase only a single server load balancer. Cisco lets you have one appliance that can act as 10 appliances — like server virtualization — carving up the appliance itself.

Several of the Cisco customers interviewed for this project were very happy with ACE. A large U.S. financial services company stated, "The ACE has really been this dynamic fix-it-all. Two things make this possible. One thing is it plugs right into the core…. The other thing is the horsepower that ACE provides is ridiculous." A U.S. utility company added, "The ACE really allowed us to add value to where the network team has gotten visibility…all the way up [in the organization]."

"The ACE has really been this dynamic fix-it-all. Two things make this possible. One thing is it plugs right into the core…. The other thing is the horsepower that ACE provides is ridiculous."

USE CASE SCENARIOS

Cisco has had experience with customer use case scenarios demonstrating how ACE has been deployed in enterprise environments to enable virtualization and improve application efficiency and control. Two scenarios that emerged during IDC customer interviews involve datacenter consolidation and resiliency.

Datacenter Consolidation

Several of the companies interviewed for this project had datacenter consolidation initiatives under way, in planning, or recently completed. One European service provider was in the midst of a multiyear effort to consolidate applications onto virtual server farms to reduce the number of servers in the datacenter, while an Asia/Pacific service provider cited as one of its guiding principles reducing the number of datacenters (or at least limiting their size and growth). "About two years ago, the majority of customer environments had dedicated stacks of hardware," said an Asia/Pacific hosting provider. "We didn't have any headroom we could use…. We said why don't we add an element of consolidation to the network, a layer of shared function to bring [greater] agility."

"About two years ago, the majority of customer environments had dedicated stacks of hardware. We didn't have any headroom we could use…. We said why don't we add an element of consolidation to the network…."

Server load balancing played a key role in many of these engagements: By deploying this functionality, businesses are able to more easily provide connectivity to multiple virtual machines. Further, consolidation places additional spotlight on the need for the type of resiliency that can be enabled by server load balancers; having multiple eggs in a single basket (i.e., by having many virtual machines housed on a single physical server) makes the need to protect that basket more paramount.

Datacenter consolidation is not limited to servers; respondents' visions include consolidation of the number of switches and connections in the network as well. Finally, server consolidation is driving the need for increased bandwidth into the datacenter, with the rollout of 10GbE, both at the core and, increasingly, on the network perimeter.

Resiliency

Similarly, server load balancing played a role in several respondents' resiliency initiatives. "My goal is to make sure we have a rock-solid infrastructure," stated a U.S. hosting provider. "The server load balancer supports the resiliency of the whole datacenter." Other respondents strove for georesilient infrastructures and platforms and implemented server load balancing with the express goal of increasing resilience and improving scalability.

Network complexity in many enterprises has increased exponentially in the past several years, with IT organizations supporting more end users, access points, end-user device types, and applications (such as video and voice). While enterprise networks may be sufficiently robust, they are not as resilient as they need to be, and organizations are understanding the need to make investments to improve resiliency.

One U.S. healthcare provider uses server load balancing for a radiology application that runs on several servers in a cluster. Obviously, this system is critical to the radiology department's ability to provide patient care. The load balancer makes intelligent traffic decisions on where to route traffic and to which server in the datacenter. This makes the application work better and improves the end-user experience while providing a path to continue to scale the application in the future.

Other examples of front-office business applications include the use of server load balancing in public Web sites, mobile telecom services, and Web hosting applications. Organizations looking to improve resiliency need solutions at their network core; thus, any devices they install must have appropriate throughput, reliability, availability, and security functionality. Further, organizations must trust that the core technology and the vendor behind it are up to the task.

FUTURE OUTLOOK

IDC observes that the best practices of previous-generation network architectures are being rewritten to support the needs of today's virtualized datacenter with required levels of flexibility. Application functionality will span the entire datacenter, but a key challenge to the IT organization will be how to treat separate workloads individually. The products that will achieve the most market success will enable IT managers to manage the network as a whole across the entire IT infrastructure but still provide control over individual network elements.

Virtualization introduces new challenges for network managers, especially in organizations that have implemented traditional, "fixed' network architectures, but server load balancing can help address this issue by bringing greater virtualization awareness to the network. It is a time-tested approach to providing greater flexibility to networks, and the technology itself has demonstrated flexibility to evolve to address new market demands. From the earliest days in which server load balancing was used to distribute traffic between Web servers, the technology has incorporated greater degrees of intelligence and functionality required to handle enterprise application traffic. IDC believes that server load balancing — currently the Swiss Army knife of the network operations team — will continue to evolve and incorporate

#230643 ©2011 IDC

greater degrees of intelligence and flexibility moving forward and will enable greater numbers of use cases.

Vendors must continue to evolve server load balancers as the foundation to support future network intelligence plays and new applications coming down the pike. This technology is applicable across a broad range of network-based applications from virtualized applications to streaming audio and video to virtual desktops, and the adaptability of the platform to handle a wide range of needs is critical to its success in adding network intelligence to create a resilient datacenter.

OPPORTUNITIES AND CHALLENGES

IDC sees both opportunities and challenges for customers as they look to adopt server load balancing to support virtualization and other application network initiatives in their datacenter, as well as for vendors such as Cisco as they evolve their product offerings.

Opportunities include:

☑ **Incorporation into the core of the network architecture.** To date, many enterprises have implemented server load balancing for specific applications, servers, or workloads but are not taking advantage of it across their entire datacenter. To fully realize the benefits of application resiliency, consolidation, and reduced operational costs, organizations need an underlying network infrastructure that is virtualization aware. To achieve this goal, organizations should look to integrate server load balancing into the core of their network across their global datacenters.

☑ **Consolidating network purchases from the same vendor: "better together."** At the end of the day, customers have the choice of standardizing their network equipment purchases from one of the leading providers or piecing together a solution using point products. One of the advantages of consolidating purchases from a single vendor is that products are designed and tested to work together, and the burden to make sure everything works together is shifted from the customer to the vendor.

Challenges include:

☑ **Limitations in scripting environment.** IDC was surprised to see that survey respondents rated support for open APIs and scripting lower than many of the other features tested for server load balancing. IDC believes that the reason is that unlike devices that focus on lower levels of the OSI stack such as WAN optimization, in which customization consists of simply identifying which protocols are broken and implementing a quick fix, for server load balancers, the customization is made at the application level, which requires a great deal more time and effort to implement. Further, there is the risk that changes introduced at this level can be "breaking changes" so that when new firmware and upgrades are included, customizations either won't work or will need to be recreated.

- **For Cisco: lack of a blade-based solution for the Nexus 7000.** While ACE comes in a variety of form factors applicable to a broad range of datacenter needs, it does not currently come in a blade form factor for Cisco's popular Nexus 7000. While this may not be an issue for a broad range of customers, many network managers prefer the simplicity and modularity of the blade paradigm and may want to see a blade-based solution.

- **For customers: understanding and taking advantage of the full range of functionality available.** Customers interviewed for this project, referring to server load balancers as the Swiss Army knife of their network, implied that there are a wide variety of use cases. One of the challenges for vendors such as Cisco is helping educate customers on the range of solutions provided by these devices and enabling customers to take advantage of the wealth of functionality available.

CONCLUSION

Server load balancing technology has evolved from its earliest days supporting Web traffic during the initial dot-com boom to become a vital component of today's enterprise datacenter. With new demands on the network operations management team brought about by virtualization and increased requirements for application reliability, availability, security, and resilience, the need for server load balancing is greater than ever.

IT executives interviewed for this project confirmed the importance of virtualization, rating it at the top datacenter network priorities, and the role of server load balancing to support it. They also demonstrated that any server load balancer introduced into the network core must be highly robust and have the highest levels of performance and scalability, rating product reliability as the most important characteristic of a server load balancer.

The Cisco Application Control Engine load balancing technology was designed to meet these challenges and to enhance application availability, accelerate application performance, and help secure the datacenter and mission-critical applications from attacks.