InformationWeek :: reports
Reports.InformationWeek.com
February 2012 $99

Next

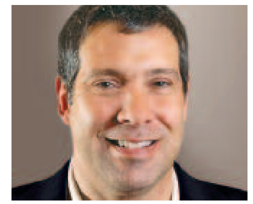
Understanding Flat Networks

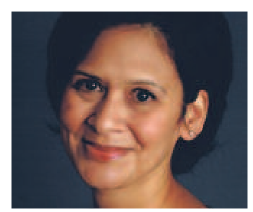
A flat network or fabric provides more paths through the network and can maximize bandwidth and better support a highly virtualized data center. We'll look at standards-based approaches to designing a flat network, including TRILL and SPB, as well as proprietary vendor implementations. We'll also discuss the implications of moving to a flat network, and provide guidance to help you decide whether this approach is the right one for your data center.

By Jeremy Littlejohn

Report ID: S4070212

TABLE OF CONTENTS





ABOUT US

InformationWeek Reports' analysts arm business technology decision-makers with real-world perspective based on qualitative and quantitative research, business and technology assessment and planning tools, and adoption best practices gleaned from experience. To contact us, write to managing director **Art Wittmann** at *awittmann@techweb.com,* content director **Lorna Garey** at *lgarey@techweb.com,* editor-at-large **Andrew Conry-Murray** at *acmurray@techweb.com,* and research managing editor **Heather Vallis** at *hvallis@techweb.com.* Find all of our reports at *reports.informationweek.com*

Jeremy Littlejohn
InformationWeek Reports

As president and chief analyst of RISC Networks, a consulting firm specializing in business technology analytics, **Jeremy Littlejohn** oversees approximately 300 engagements per year and works closely with CIOs and IT managers to optimize the reliability, scalability and performance of their entire IT infrastructures. Jeremy is a Cisco Certified Internetwork Expert (CCIE) with more than 13 years of experience in IT, focusing heavily on unified communications, LAN/WAN and data center technologies. RISC Networks services virtually every type of industry, including pharmaceuticals, healthcare, manufacturing, retail, banking, finance and technology. You can learn more about RISC Networks at www.riscnetworks.com.

EXECUTIVE SUMMARY

The traditional tiered network design is a staple of today's LAN and data center networks. However, this design has limitations caused in part by Spanning Tree Protocol (STP), which restricts the number of paths that traffic can take through the network. That restriction can affect performance and reliability requirements that are emerging because of virtualization and other factors, such as the convergence of storage and data networks.

An alternative network design, called a flat network or a fabric, is now available to IT. A flat network opens more paths and can increase available bandwidth. Flat network options include both standards-based approaches, such as TRILL (Transparent Interconnection of Lots of Links) and SPB (Shortest Path Bridging), as well as proprietary vendor approaches. All these approaches address shortcomings of STP and can make a data center network more flexible and responsive to the changing demands of highly virtualized environments.

However, a flat network also has its downsides, including the need to rearchitect the LAN. Depending on the approach that IT takes to a flat network, it may also require new gear; for instance, TRILL will likely require hardware upgrades because it uses a new frame type that has to be added to the standard Ethernet frame. Meanwhile, the proprietary approaches will lock IT to a single vendor.

This report examines the limitations of tiered networks and STP, explains the alternatives that allow IT to build a flat network, outlines potential downsides of this new approach, and offers guidance to help IT determine whether a flat network is a fit.

Getting Inside Flat Networks

The virtualization and consolidation of servers and workstations causes significant changes in network traffic, forcing IT to reconsider the traditional three-tier network design in favor of a flatter configuration. Tiered networks were designed to route traffic flows from the edge of the network through the core and back, which introduces choke points and delay while providing only rudimentary redundancy.

Enter the flat network. This approach, also called a fabric, allows for more paths through the network, and is better suited to the requirements of the data center, including the need to support virtualized networking, VM mobility, and high-priority storage traffic on the LAN such as iSCSI and FCoE. A flat network aims to minimize delay and maximize available bandwidth while providing the level of reachability demanded in a virtual world.

But a flat network also requires some trade-offs, including the need to rearchitect your data center LAN and adopt either new standards such as TRILL (Transparent Interconnection of Lots of Links) and SPB (Shortest Path Bridging), or proprietary, vendor-specific approaches. We'll look at how a flat network differs from a traditional tiered infrastructure, examine potential shortcomings of new approaches and provide guidance on whether this new networking model is right for your environment.

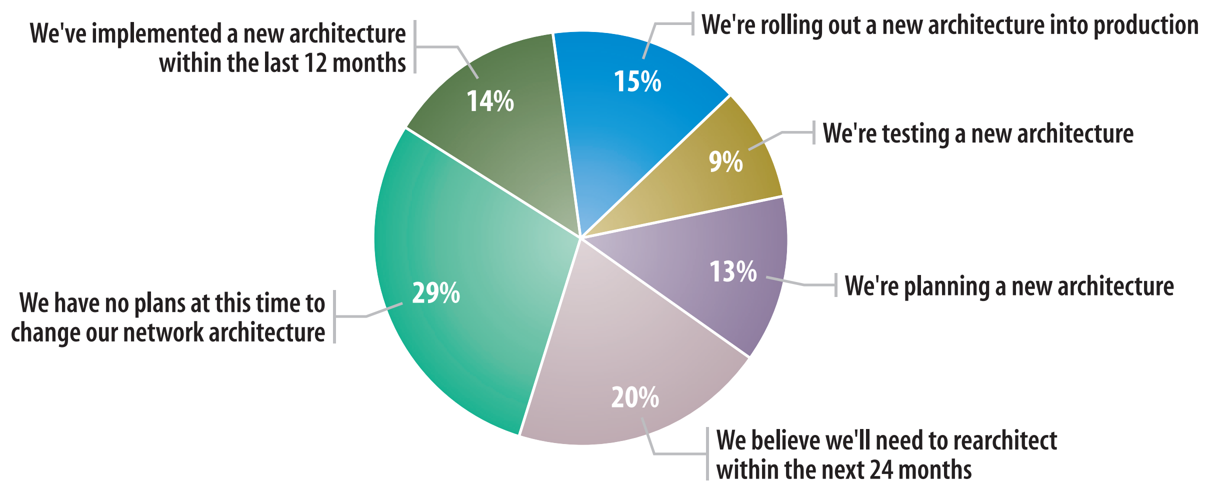
How We Got Here

Ethernet won the battle for the LAN well over a decade ago. (I remember doing a few

Figure 1

Rearchitecting Data Center Network?

Have you, or are you considering, rearchitecting your data center network?



Base: 501 respondents at organizations using data center networking products R3900112/7
Data: *InformationWeek* 2012 Data Center Networking Vendor Evaluation Survey of 510 business technology professionals, November 2011

Token Ring implementations when I began my career, but they were already on the way out). However, Ethernet suffers from two (or more, depending on who you ask) significant limitations. First is the learning mechanism for Ethernet switches. Because Ethernet has a flat address space, with no hierarchical summarization mechanisms, each switch in a LAN must learn the complete forwarding table for the entire LAN.

Switches watch Ethernet frames as they come in, and then make a corresponding entry in their forwarding tables for the source MAC address and ingress interface. The next time a frame is destined for that MAC address, the switch finds the destination interface in its forwarding table and sends the frame out that interface. This means the switch carries a large forwarding table that has lots of entries. In addition, each destination address must map to only one forwarding port. In short, everyone must know how to reach everyone else, and there is no support for multiple paths through the network to reach another station.

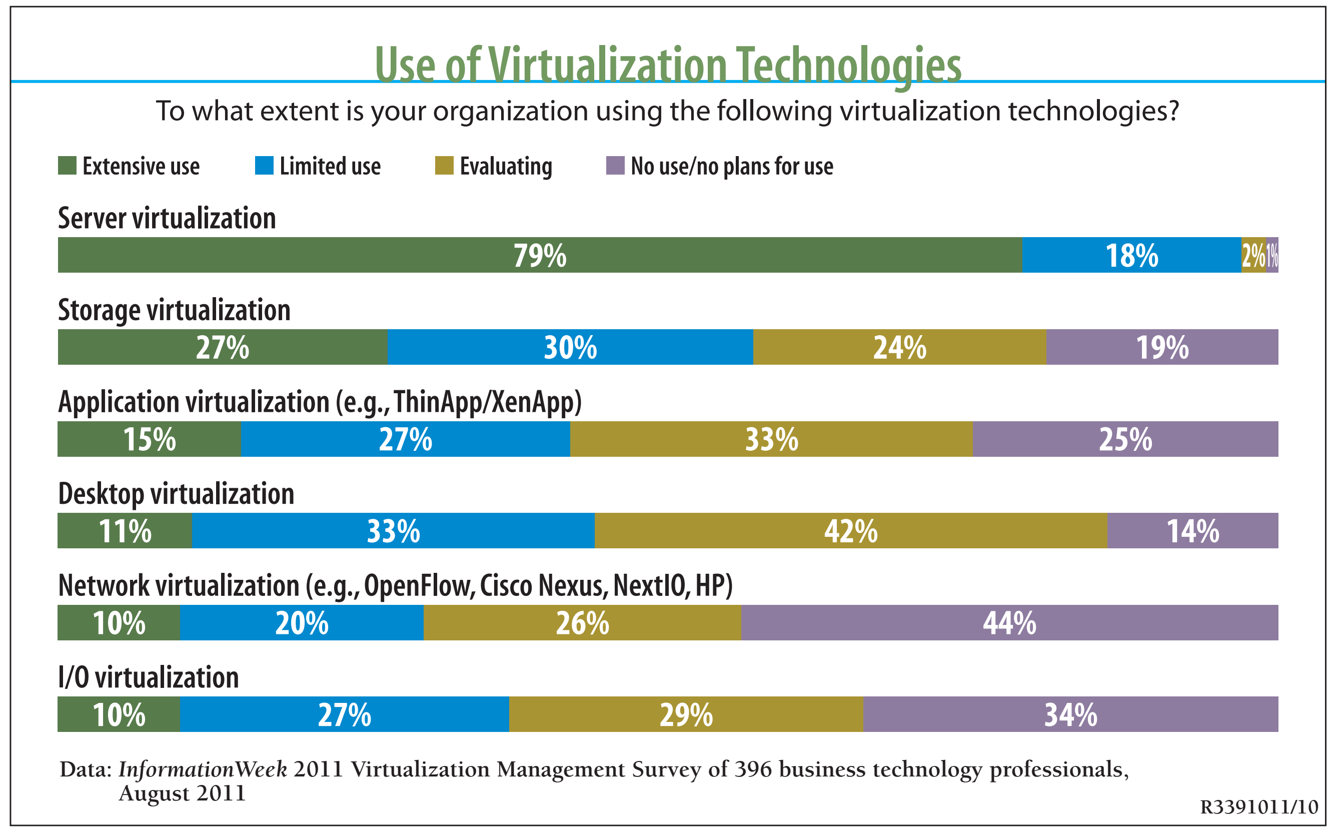
The second protocol limitation is the forwarding mechanism. When an Ethernet switch does not have a MAC and interface pair in its forwarding table, or it receives a broadcast Ethernet frame, the switch makes a copy of the frame and forwards the copy out all interfaces. Because there is no Time to Live (TTL) header field in Ethernet, if there is a physical loop in the network these frames will

FAST FACT

85%

In the author's client engagements, that's the amount of output loss on 100 Mbps interfaces that occurs on interfaces showing between 75% to 84% utilization.

Figure 2



be copied and propagated repeatedly throughout the network until it crashes.

Radia Perlman created the Spanning Tree algorithm, which became part of the Spanning Tree Protocol (STP), to solve this combination of issues. Ms. Perlman certainly doesn't need me to come to the defense of Spanning Tree–but I will. I like Spanning Tree, because it works. I would say that in at least 40% of the networks I see, Spanning Tree has never been changed from its default settings, but it keeps the network up, while at the same time providing some redundancy.

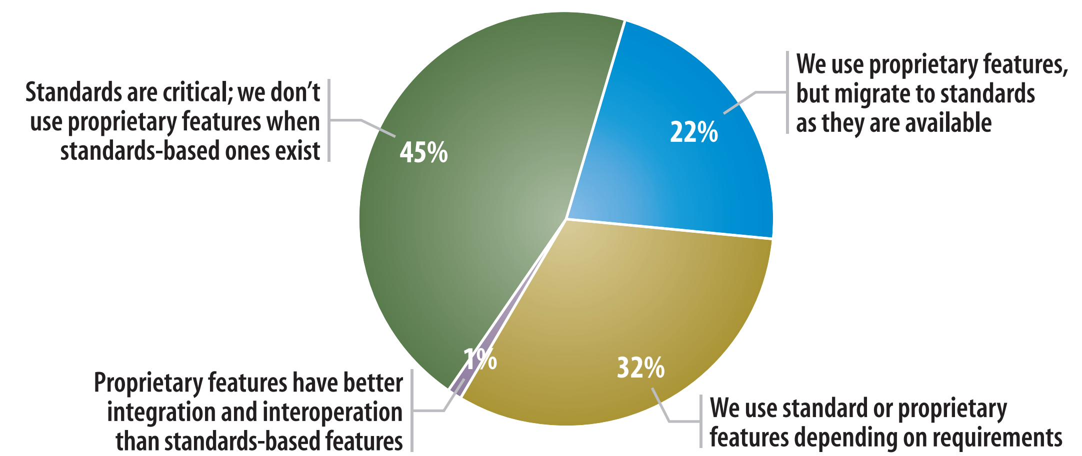
However, while STP solves significant problems, it also forces a network design that isn't optimized for many of today's data center requirements. For instance, STP paths are determined in a north-south tree, which forces traffic to flow from a top-of-rack switch out to a distribution switch and then back in again to another top-of-rack switch. By contrast, an east-west path directly between the two top-of-rack switches would be more efficient, but this type of path isn't allowed under STP.

Flat networks provide solutions. The major challenges that flat network technologies solve are poor path optimization, failover timing, limited or expensive reachability, and latency. Simply put, we need to be able to reach any machine, wherever it is in the network, while using the best path through the LAN to do so. This will lower latency, provide access to more bandwidth and provide better ROI for the network infrastructure in the data center. If a device fails, we want to recover immediately and reroute traffic to redundant links. These are all problems today with three-tier designs that use Spanning Tree. The original 802.1D Spanning Tree can take up to 52 seconds to fail to a redundant link. RSTP (802.1w) is much faster, but can still take up to 6 sec-

Figure 3

Standards-Based vs. Proprietary Features

Which statement best reflects your organization's position on standards-based features vs. proprietary features?



Data: *InformationWeek* LAN Equipment Vendor Evaluation Survey of 444 business technology professionals, July 2011 R3160811/19 onds to converge. It's an improvement, but six seconds can still be an eternity in the data center.

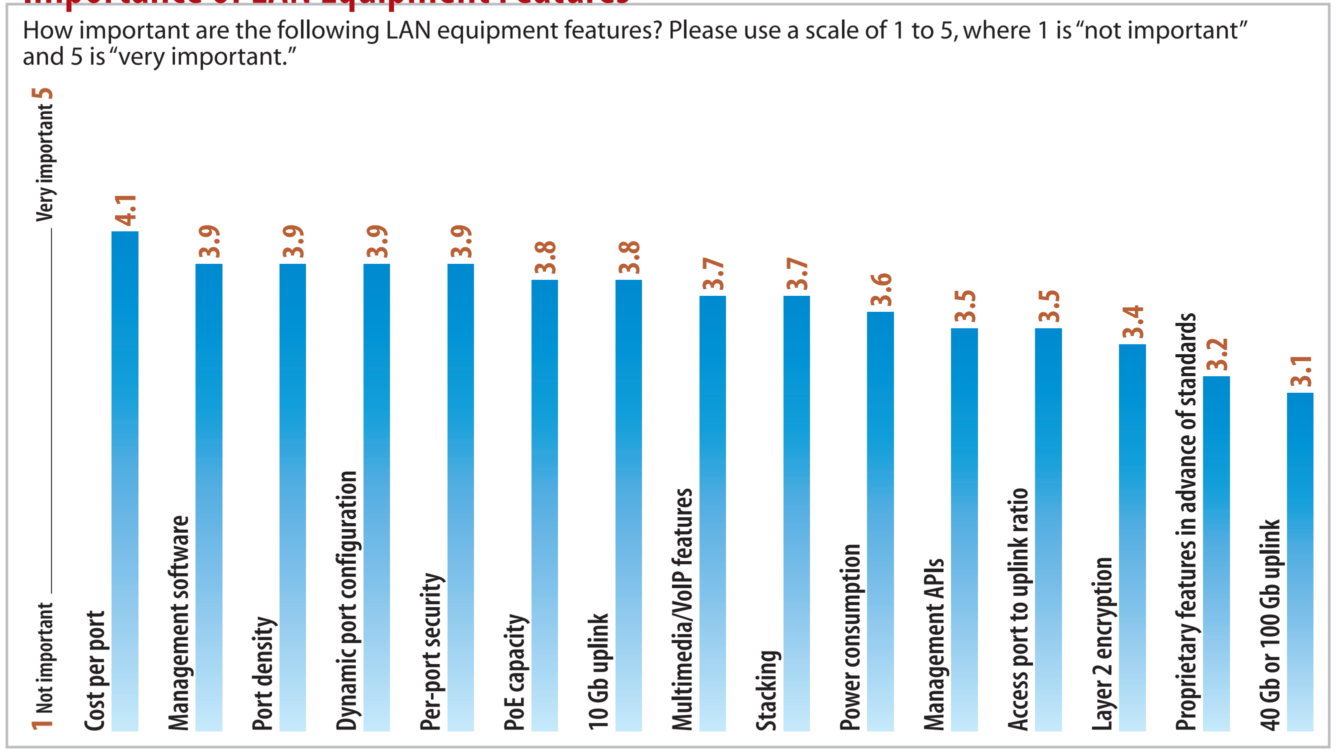
How to Get Flat

Newer protocols and technologies are emerging to address some of these problems, including TRILL, SPB, and several proprietary efforts from infrastructure vendors. In general, I classify these approaches into two major groups. Group one, consisting of TRILL (from the IETF), SPB (from the IEEE), and others like Cisco's FabricPath and Brocade Virtual Chassis Switching (VCS), aims to make LAN switches smarter. At a high level, these technologies address the learning and distribution of forwarding paths using a link state routing protocol (IS-IS) and the forwarding of traffic from Host A to Host B across multiple paths. They differ significantly in their execution.

TRILL adds a new frame type (a TRILL frame) to normal Ethernet frames and forwards them between ingress and egress RBridges. Think of RBridges as zip codes. The final street address (destination MAC of Host B) doesn't matter until you are in the right zip code (egress RBridge). The ingress RBridge maintains a mapping of final addresses (MACs) to zip codes (RBridges) and then all intermediate switches move traffic from one zip code to the other. This additional encapsulation header also includes a TTL field to help stop flooding produced by physical loops. TRILL does not have a

Figure 4

Importance of LAN Equipment Features



Note: Mean average ratings R3160811/14
Data: *InformationWeek* LAN Equipment Vendor Evaluation Survey of 444 business technology professionals, July 2011

full suite of Operations, Administration and Maintenance (OAM) standards like SPB does, but it does provide some basic network troubleshooting functionality, such as path tracing.

Unlike TRILL, SPB does not encapsulate traffic in a new frame type. Instead, SPB uses either Q-in-Q or MAC-in-MAC encapsulation, which is available today on many switch models. Once it learns its mapping of destination address to egress switch through IS-IS, it encapsulates the original frame in a new (Q-in-Q or MAC-in-MAC) frame and readdresses the new frame to the egress switch.

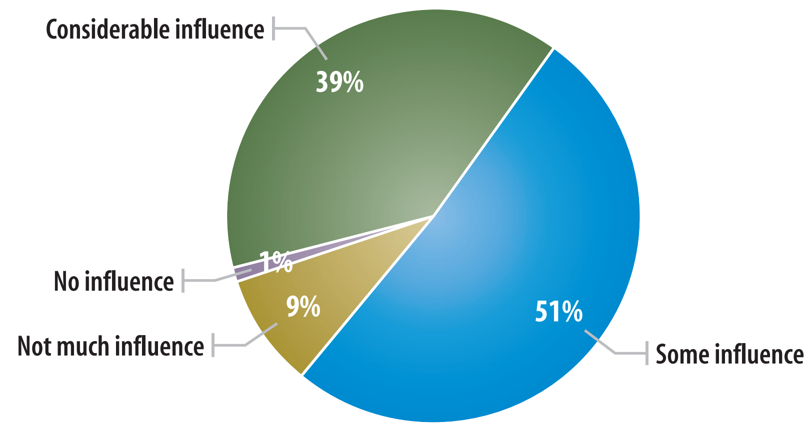
While STP solves significant problems, it also forces a network design that isn't optimized for many of today's data center requirements.

SPB and TRILL both support multiple paths, faster failover than Spanning Tree, and increased reachability. SPB proponents point to its support of legacy ASICs because no new frame format has to be added to the frame. In other words, you may only need a software upgrade to support SPB, while TRILL will likely require new hardware to handle the new TRILL Frame format.

Figure 5

Impact of Incumbent Vendor Status on LAN Equipment Purchases

When buying LAN gear, how much is your purchase influenced by having other equipment from the same vendor?



Data: *InformationWeek* LAN Equipment Vendor Evaluation Survey of 444 business technology professionals, July 2011 R3160811/10

Note that while TRILL and SPB solve some of the limitations that are introduced by STP, these technologies also have potential downsides. For example, SPB and TRILL topologies may require significant security configuration. Distributed intelligence means distributed security requirements. Take a protocol like OSPF for example, which is a link state protocol like IS-IS. It is a best practice that all OSPF neighbor relationships be based on hashed authentication to avoid the introduction of rogue routers and possible route poisoning. This security step requires configuration of keys on each OSPF router in a domain. Now think about how many LAN switches you have, even if you only count the core ones. Will you need to configure this security control on each one? This and other implications must be part of your evaluation of these approaches.

Cisco FabricPath and Brocade VCS also belong to the first group of "smart switch" technologies. Both are similar to TRILL in that they encapsulate traffic for transport across a fabric backbone. However, these are proprietary technologies that do not interoperate with TRILL. In fact, they both work differently with existing Spanning Tree networks. Brocade VCS either passes your Spanning Tree info through its fabric or drops it entirely. Cisco FabricPath terminates Spanning Tree domains at the edge of its fabric.

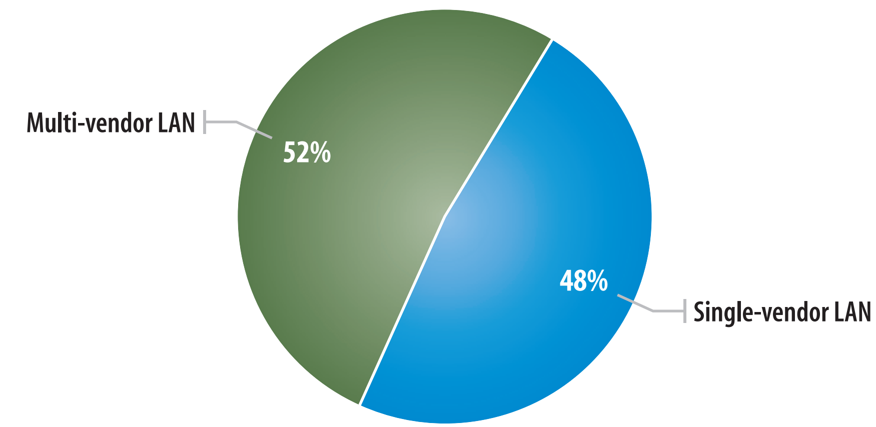
The second group of technologies looks to eliminate the inefficiencies of Spanning Tree and tiered design by eliminating the tiers altogether. This approach treats all switches as one giant switch, or fabric. Juniper's QFabric is an example. Essentially, the physical switches act as blades within a giant chassis. In Juniper's model, the control plane is handled by QFDirector, which coordinates the individual switches. The entire fabric not only appears as one giant switch to connected hosts, as with TRILL or SPB, but is also managed as one giant switch. This reduces complexity, but make sure you understand the implications of failure of critical components. Just as distributed intelligence means distributed security and management overhead, centralized management means centralized risk.

Another alternative is Multi-Chassis Link Aggregation (MLAG), which is offered in some variation by most major switch vendors. MLAG switches act as a single switch for downstream STP bridges to eliminate Spanning Tree blocking of redundant paths. This allows for better utilization of links and simplifies management. Most MLAG implementations are limited to

Figure 6

Number of LAN Vendors

Do you have a single- or multi-vendor LAN?



Data: *InformationWeek* LAN Equipment Vendor Evaluation Survey of 444 business technology professionals, July 2011 R3160811/12

FAST FACT

92%

In the author's client engagements, that's the percentage of packet loss that occurs on interfaces that never show over 49% bandwidth utilization.

two aggregated switches.

Outside of TRILL and SPB, these technologies are proprietary and will tie you to a single vendor. IT professionals vigorously debate the benefits of single-vendor vs. multi-vendor networks. It's a reasonable point of contention. I come down on the side of a single-vendor LAN. I have several reasons, but one is of particular interest to this report: Spanning Tree. Plug an HP switch into a Cisco switch and see how Spanning Tree works out for you; hopefully your network won't melt down. The fact is, implementation of standards can differ among vendors, which can make operations tricky in a multi-vendor environment. Heck, getting trunks to negotiate properly with only a single vendor can sometimes be a challenge. I recommend minimizing interoperability issues by adopting one vendor for your LAN.

The original 802.1D Spanning Tree can take up to 52 seconds to fail to a redundant link.

First Understand the Problem

The decision to move from a traditional tiered network to a flat network is a significant one. Before you make that decision, you should understand the specific problem or problems you're trying to solve. A good approach is to take the following steps:

• Ensure that Spanning Tree is running correctly first. If you need immediate bandwidth relief on ISLs (Inter Switch Links), consider running MSTP (Multiple Spanning Tree Protocol) and splitting your root bridges on an even/odd VLAN distribution. This will let you forward up paths that might be blocked if you were running a single STP instance.

• Map out the blocking ports and understand whether your environment even has a problem. See how many east-west links you are blocking. Make sure you don't just look at ISLs. Understand where your hypervisors and network based storage (iSCSI for example) units are in the topology. In addition, be sure to identify security devices. The path taken by traffic today is very easy to understand because Spanning Tree limits the choices. If you flatten the network and allow multiple paths, you will need to consider how to direct flows to your security devices for inspection and logging.

• For a detailed analysis, you can capture traffic and evaluate the amount of communication from access switch to access switch. Once you have mapped your Spanning Tree, you can target specific ISLs and access switches that you have suspicions about.

• Check your ISLs for utilization and overutilization. Make sure to look for input and output discards, not just bandwidth. The discards (packet drops) are what we call "ghosts of microbursts." They show you how much you have dropped due to either an oversubscribed architecture or to utilization. If you see high bandwidth utilization and packet drops, this is a sign that flattening may be a good fit for you. In addition, 10 Gbps Ethernet may be a good fit.

• If VM mobility is a problem, trunk VLANs to your virtual switches and then log how much virtual machine movement actually occurs before you rearchitect the network. Your VMs may not be moving as often as you think. This introduces the classic chicken-and-egg conundrum: Are VMs not moving because they can't? Or do you not need to move VMs, and thus don't need the feature? Flat networks can help with VM mobility issues, but so can network automation, so don't feel that you have to go flat.

I'd also like to offer a word of advice about bandwidth upgrades. Our customers are often told they need to upgrade their 30% utilized 1 Gbps link to 10 Gbps to get better performance. It doesn't work like that. Only two network-induced variables have a significant impact on TCP traffic: packet loss and delay. If you aren't suffering packet loss due to overutilization or oversubscription, the additional bandwidth won't help. In addition, the performance difference between 5 milliseconds and 5 microseconds of latency in your LAN is not going to show up in your Outlook client performance.

In our customers, we see the major limiting factor in LANs being the switching architecture more so than path selection. Packet loss is the top performance problem. On one customer LAN, the 100 Mbps links often see packet loss due to overutilization. For instance, 85% of output packet loss on 100 Mbps interfaces occurs on interfaces showing between 75% to 84% utilization. Based on a 30- second interface average, these interfaces are likely reaching 100% saturization during a small window (say 500 milliseconds) and dropping traffic. This points to utilization being the culprit. Flat networks would definitely help performance for the 100 Mbps links if they were ISLs and had additional paths that could be used to distribute traffic.

Now take a customer with 1 Gbps ports. In their environment, 92% of outbound packet loss occurs on interfaces that never show over 49% bandwidth utilization. What gives? In many cases, 1 Gbps interfaces are not actually able to handle 1 Gbps of line rate traffic. Their connections to the backplane of the switch are oversubscribed, in some cases up to an 8:1 ratio. From a performance perspective, this architectures is a much larger issue than north-south Spanning Tree topologies, and should be addressed first. How prevalent is this issue? If you take Cisco 6500s, which constitute the majority of the data center switches we see deployed today, a full 33% of line cards we have assessed are oversubscribed. This is not to pick on Cisco. Cisco clearly tells customers and partners that these line cards are designed for access layer deployment, but sometimes IT folks see Gigabit and think "This belongs in the data center." These links are likely to continue to have issues, even if they have a flat network topology. The solution here is first to get line rate 1 Gbps performance, and then re-evaluate your situation by continuing to track the performance of these links. In the end, 10 Gbps Ethernet may be your answer, or you may want to spread the traffic over multiple 1 Gbps links.

Ethernet multipathing and link optimization are great if you have lots of flows that are similar in size, and you have a wide distribution of hosts. But if you are trying to address one flow or a small set of them, multipathing may not help. For instance, a single large database replication or system backup is not going to be split over multiple paths, so if that is your main issue you don't need to go flat, you need to go with more bandwidth, like 10 Gbps Ethernet.

A flat LAN architecture holds a lot of promise, but we don't recommend blind adoption. If engineers don't have time to set up a Spanning Tree root bridge priority today in order for the Spanning Tree algorithm to choose the best possible tree, can you really expect that TRILL or SPB or any other more complicated setup will be tuned properly? The LAN stability provided by Spanning Tree should not be taken for granted; it is important to clearly understand the potential ramifications of new flattening technologies. There's a reason that tiered network architectures are so prevalent: they work. If you decide to go flat, apply these new approaches judiciously.

MORE LIKE THIS

Want More Like This?

InformationWeek creates more than 150 reports like this each year, and they're all free to registered users. We'll help you sort through vendor claims, justify IT projects and implement new systems by providing analysis and advice from IT professionals. Right now on our site you'll find:

Strategy: Networking in a Virtualized World: Virtualization is rapidly evolving into a core element of next-generation data centers. This expanded role places new strains on the network. This report explore the technical issues exposed by virtualized infrastructure and looks at standards, technologies and best practices that can make your network ready to support virtualization.

Research: IT Pro Ranking: LAN Equipment Vendors: Cisco leads on the LAN, with an overall score of 77% based on product reliability, product performance and other key criteria, but HP, Dell and others are finding ways to compete. Find out how 444 IT pros who use and evaluate LAN equipment compare vendors.

Strategy: Inside OpenFlow: New software-defined networking technologies in general and OpenFlow in particular are poised to disrupt the way we manage load in highly virtualized data centers. In a world where, for decades, we've relied on Ethernet and TCP/IP standards—and where powerful vendors like Cisco and Juniper have made their fortunes based on intelligence in switches—that's a big deal.

PLUS: Find signature reports, such as the *InformationWeek* Salary Survey, *InformationWeek* 500 and the annual State of Security report; full issues; and much more.