Are Fabrics Faster or Better Than Networks? Are they the right choice for your next generation Data Center?

Inside

Subject

Discusses the requirements for next generation networks to support large VM farms and contrasts the approaches between two different fabric models and a more traditional L2/L3 multi-layer network

Why

There is a lot of debate and vendor marketing around fabrics. The purpose of this paper is to compare some of the attributes between fabrics and networks and help customers reach a decision as to which path is the best for their IT requirements

Who Should Care

Anyone building out a VM farm, Hadoop Cluster, or looking at fabric architectures for their next-generation of IT infrastructure

Introduction

Over the past year interest in building 'fabrics' has grown throughout the networking community. This has been led from the front by many vendors with one of two approaches to building these next-generation infrastructures:

In one genesis Fabric was derived from the use of a switched fabric architecture to connect multiple linecards together in a modular chassis so that each had the ability to talk to any other linecard, at any time. To accomplish this fabrics were often either arbitrated on ingress or the fabric was over-clocked. Thus any point in time congestion could be handled by a small but reasonable amount of buffering.

In the second iteration Fabric was derived from the storage area networks built out, starting in the late 90s, where the intent was to have a large number of initiators connecting to a welldefined number of storage targets. The traffic patterns were well understood and almost always from initiator to target and back. These fabrics were designed to be lossless by implementing a credit-based fabric arbitration mechanism that guaranteed the SAN would not drop traffic under congestion.

Arista has a different and unique approach to building network architectures that achieves greater scale, lower latency, and increased power efficiency while building a network architecture based on as many open standards as possible and ensuring multi-vendor interoperability. By sticking with open standards and well established/proven architectural models the overall operating cost of the infrastructure and ease of troubleshooting and management are far enhanced over proprietary and prestandard systems.

This note explores some of the characteristics of fabric architectures and how Arista achieves equal or better results with more proven technologies.

Fabrics are Faster

One statement often made is that fabrics are faster than networks, in many cases because they require only one hop or because all links are active.

As far as requiring one lookup goes the most proprietary of the fabric architectures on the market requires five header lookups (one full lookup at the ingress switch, three lookups in the Clos tree within the Fabric Interconnect, and a final lookup at the egress switch.) Only the first lookup is a L2/ L3 lookup, subsequent are port-of-exit header lookups that tell the transit nodes in the fabric which port to forward the frame to.

The statement that all uplinks are active in a fabric implies that they are not all active in a network. This was true in flat L2 networks of the past based on STP. But since the late 1990s vendors have delivered proprietary solutions that made all links active and since 2008 blended proprietary/openstandard solutions such as MLAG have become available. MLAG requires that two switches be the same make/model from a vendor, they share learning and state information via a proprietary protocol between the two switches, and then express standards-based LACP to all adjacencies. With MLAG the size of the proprietary system is two switches.

For those to whom latency directly correlates to application performance financial trading, cluster computing/ modeling, deep analytics you can build a 2-tier network with a latency as low as 1.5 microseconds using fixed-configuration rack switches, you can build a very scalable 2-tier network with latency as low as 5 microseconds using a mix of rack and modular switches. The fastest fabrics are a bit more than that.

You can build a 2-tier L2 network based on Multi-chassis Link Aggregation and have all uplinks active. The main constraint in MLAG scaling is the density and performance of the spine switch - the denser and higher performance the systems are the larger a L2 domain you can build - currently these are as large or larger than the biggest fabrics.

The other factor that is very important to note here is the need for effective congestion management and buffering in the spine tier. In a multi-stage network with east-west traffic patterns it is quite common for some large number of ports to need to access the resources on another port - the larger the buffers are on the spine switches the more effectively they will perform before they have to drop traffic or issue a PAUSE frame to slow down traffic.

Summary: Networks offer the flexibility to be designed at the absolute lowest latencies if required (trading off the use of modular large-buffer systems for fixed-configuration low latency systems in the spine is the most effective way of achieving absolute lowest

latency/fastest forwarding) Current network solutions scale as well or better than the current fabric solutions - at L2, and certainly as you introduce routing at L3.

Simply put, from a 'Fabrics are Faster' perspective the claim doesn't hold up. The fastest fabric is a bit slower than an equally scalable network, the fastest network is 2-3x lower latency than the fabric alternative.

Enabling a large and flat Layer-2 broadcast domain

It feels like all vendors have rallied around a common shift: that FC (Fiber Channel) is not going to be the primary driver of next-generation network architectures, optimizing network support of virtual machines is going to be, especially virtual machine mobility, or vMotion in VMware parlance.

This is the main issue that the Fabric proponents are rallying around, regardless of the genesis of their strategy: fabric can move VMs anywhere because it will give you a stable large flat layer-2 network.

It's a good promise. The fabric trade-off is that the network team must adopt an architecture that generally requires extreme vendor lock-in: kind of like standardizing your multi-protocol routing on EIGRP was demanding in the mid 1990s (except at least in EIGRPs defense it was arguably a better RP than IPX-RIP and AppleTalk)

Do you need a proprietary fabric to build an infrastructure to support vMotion at scale? This is literally the 'million dollar question' so a few facts that should be disclosed and discussed that will help in making an informed decision:

1. The scale of the Network that can be built using L2 based on multi-chassis implementations of LACP (accepting this is still proprietary but only between 2 boxes and not the span of the entire network) is a few thousand ports. If you design at L3 its 10s of thousands of ports.
2. VMware has one of the more scalable virtualization controllers available today - and it supports 1000 hosts per vSphere instance. Each host can have 20 or so VMs based on today's hardware capabilities with Intel Westmere and the forthcoming SandyBridge.
3. The maximum number of hosts that can participate in an automated vMotion domain thatwould require L2 adjacency is 32.

4. The maximum distance you can vMotion a workload is when the network latency has a RTT of less than 10msec.

So what does this mean? Well, in short, whether you go with a proprietary fabric or a more open protocol-based network either choice can easily support the scale needed to support the maximum density of VMs and physical hosts that can participate in a stateful vMotion.

It also means that future protocols like NVGRE and VXLAN will solve the main Virtualization care-about without requiring fundamental network change, re-architecture, or capital outlay.

It means that anyone who is selling you the "move your VMs around the world and follow the sun with your workloads" is up against some hard physics lessons on how much data can be moved and how much state can be synchronized between two locations.

Summary on the specific question of "how to support large and flat L2 domains"

Both of the Fabric architectures, regardless of genesis, support decent scale of L2 domains. Arista open-standard protocol-based networks support the same or larger scales, yet induce no changes to the operating model most network administrators are familiar with.

TRILL lets you have a wider spine than multi-chassis LACP models do, but the LACP models extend the network reliability down to the host and are not limited to a total of 100 bridges.

Fabric sounds a bit sexier and does give a raison d'etre for some companies as well as a justification for asset churn; open-standard protocol based networking sounds a bit more boring but at least you can hire people who know how to manage and operate critical infrastructure.

Should I use L2 or L3 for my data center?(whether to call it a fabric or not is at your discretion)

Arista's point of view and general recommendation is that this is a function of scale and size of both the network and the number of hosts and whether the environment supports virtual machines or not as one of the major elements of scaling an architecture up is managing the size of the IP hosts tables in the switching equipment at the L2/L3 boundary (default gateway). A flat Layer-2 network works for a few hundred hosts, as the network gets close to around 1000 hosts we usually recommend moving to an IGP such as OSPF, and as the network gets closer to 3000-5000 hosts we recommend using BGP as the routing protocol as it scales more effectively, converges on link or nodal failure more quickly, and is generally more stable with less flapping.

Technologies that are in development such as VXLAN and NVGRE both promise to simplify the table size management by masking the addresses of the virtual machines so we only have to support a few MAC/IP pairs per virtualized host. This will greatly simplify table management while allowing stateful virtual machine mobility on top of a Layer-3 network through the use of UDP or TCP based tunneling from vSwitch to vSwitch. Essentially VXLAN and NVGRE invalidate the main reason for most of the Large, Flat, L2 network architectures.

The combination of network virtualization/addressing abstraction such as VXLAN/ NVGRE, ubiquitous L3 forwarding in hardware in most/all credible switching platforms, and the availability of trained network operators leads us to the most likely architecture for scaling a data center moving forward: Dual-Attached L3 Default Gateways at the Top of Rack or End of Row depending on your cabling preference.

In this architecture a pair of switches as the top of each rack will provide both L2 and L3 services to the attached hosts and function as active-active default gateways usually using an Anycast Default Gateway technology such as VARP to eliminate protocol-based default gateway redundancy that often causes flooding and congestion on MLAG/spine links. The pair of switches at the top of the rack will route into anywhere from 2-16 spine switches based on the scale of the network and data center.

This architecture when using 7050S-64s at the leaf and Arista 7508s at the spine scales to 17,664 host ports in a single two-tier network. (384 64-port leaf switches delivering 2 ports for MLAG ports, 46 host ports, and using 16 Q-SFP ports for uplinks to 16x 7508 spine switches). By contrast a similar architecture with identical oversubscription ratios based on a flat L2 network with TRILL can scale to 4,232 ports before you hit a limitation in the number of bridges supported, and a MLAG based network with only two spine switches can scale to 2,116 ports.

Summary: L3 scales to multi-tier architectures simply and meets or exceeds the requirements for scaling within the largest data centers in the world. The demand for large and flat L2 networks is driven primarily by the need for stateful virtualization which is getting met with VXLAN/NVGRE technologies on top of any topology - L2 or L3. L3 can scale to at least 4x the size of the largest L2 networks without requiring new protocols and staff retraining.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062