#SHAREorg

SHARE
Technology · Connections · Results

Ethernet Fabrics and the Cloud: Avoid the Fog and Smog

Dr. Steve Guendert
Brocade Communications

Dr. Casimer DeCusatis
IBM Corporation

February 7, 2013
Session 12735

SHARE in San Francisco 2013

Abstract

This session will discuss Ethernet Fabrics: what they are, what their business and technical value is, and how to implement them as part of your cloud architecture including with System z. It will also dispel misconceptions to clear the smog and fog from the cloud. The focus will be on the Open Data Center Interoperable Network (ODIN) model.

Agenda-Overview

- Introduction
- A need for progress in data center network design
- Data center network transformation
- What is an Ethernet Fabric?
- The Open Datacenter Interoperable Network (ODIN)
- System z
- Conclusion and questions.

The Need for Progress is Clear

30 percent

Energy costs alone represent about 30% of an office building's total operating costs

42 percent

Worldwide, buildings consume 42% of all electricity – up to 50% of which is wasted

85%

In distributed computing 85% of computing capacity sits idle

18%

Anticipated annual increase in energy costs

20x

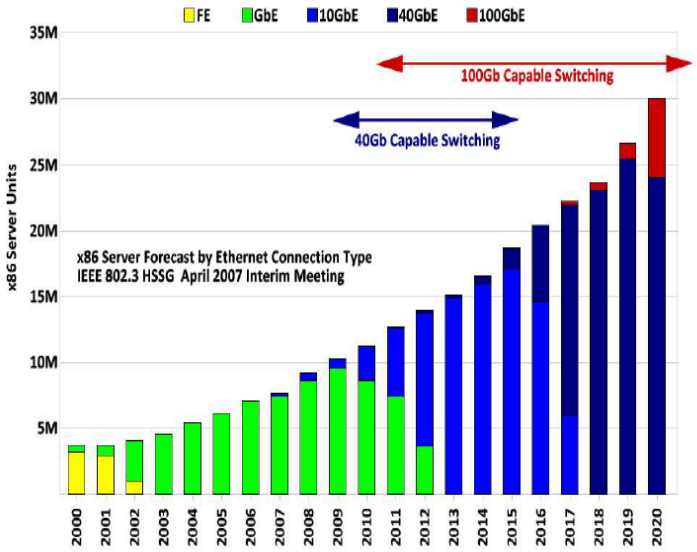
Growth in density of technology during this decade. Energy costs higher than capital outlay

50%+

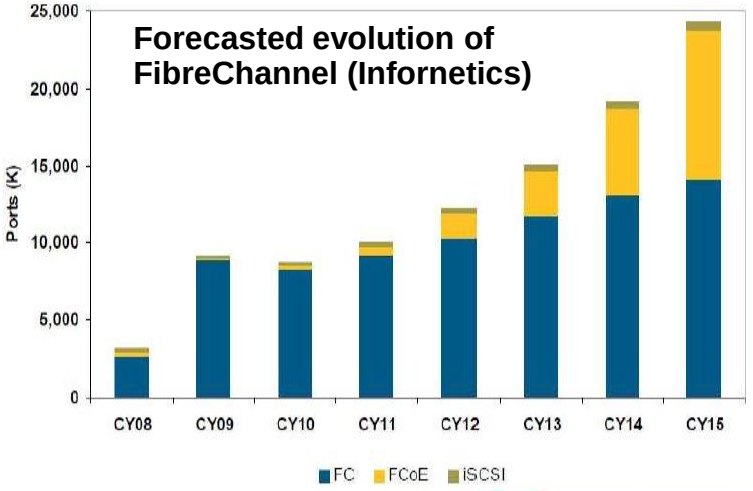
More than half of our clients have plans in place to build a new data center/network facilities as they are out of power, cooling and/or space

- Rapidly increasing demand for 10 Gbps server connections
  - Transition to 10G is happening *now and will be mainstream from 2012*
  - Broad deployment of 10GBaseT will simplify DC infrastructure
    - by easier server connectivity, while delivering bandwidth needed for heavy virtualization and IO intensive applications
- Server Virtualization
  - To stop wastage of server CPU resources
- Exploding East-West traffic volumes
  - to support multitier applications and high performance computing
- Proliferation and mobility of Virtual Machines
  - to address fluctuating workload
  - by on-demand starting, moving and (hopefully also) decommissioning VMs
- Increased complexity
  - Drives focus to maintaining the infrastructure
  - Rather than to adding business value by leveraging new infrastructure services

Forecasted evolution of Ethernet (IEEE, 2007)

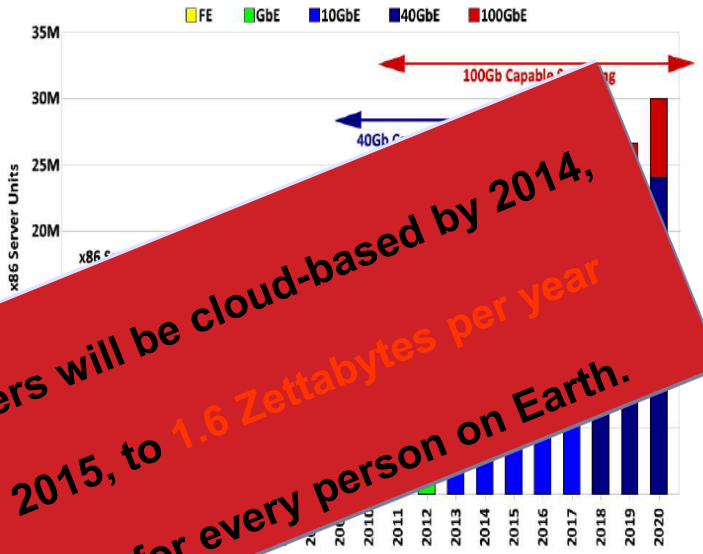


Forecasted evolution of FibreChannel (Infornetics)



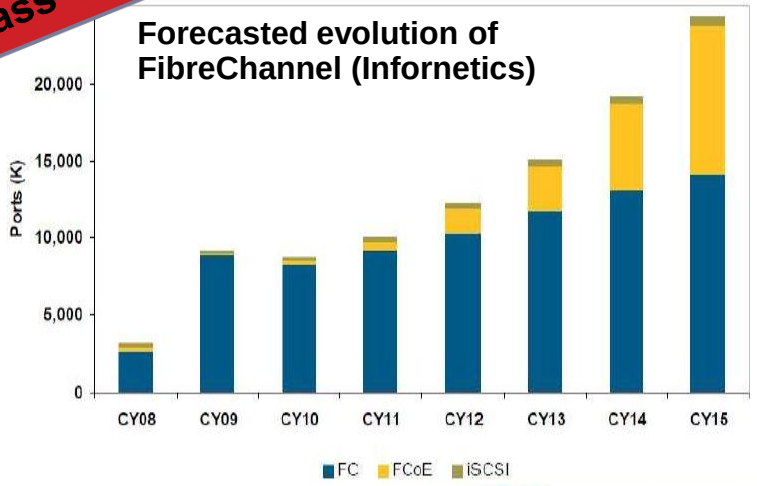
- Rapidly increasing demand for 10 Gbps server connections
  - Transition to 10G is happening *now and will be mainstream from 2012*
  - Broad deployment of 10GBaseT will simplify DC infrastructure
    - by easier server connectivity, while delivering bandwidth needed for heavy virtualization and IO intensive applications
- Server Virtualization
  - To stop wastage of server CPU resources
- Growing traction of LAN/SAN converg
  - 70% of IT decision makers believe in it (Forrester, 201
  - Barriers are organizational rather than techno
- Proliferation and
  - to address flu
  - by on
  - ue by leveraging new infrastructure

Forecasted evolution of Ethernet (IEEE, 2007)

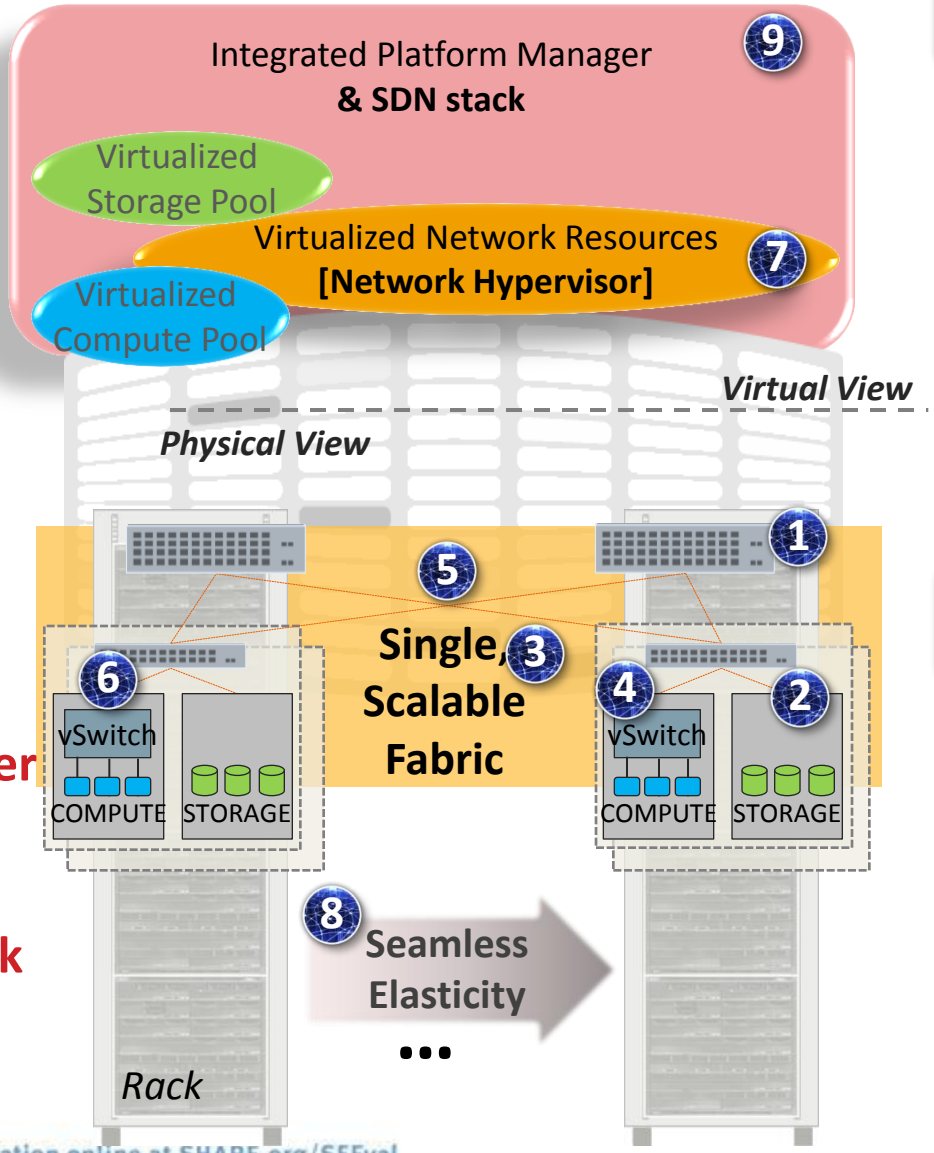


More than 50 percent of computing workloads in data centers will be cloud-based by 2014, and global cloud traffic will grow more than 12 times by 2015, to 1.6 Zettabytes per year — the equivalent of over four days of business-class video for every person on Earth.

Forecasted evolution of FibreChannel (Infornetics)



How is the data center evolving ?



Automated

6. Virtual Machine network state automation
7. Multi-tenant aware Network Hypervisor

Integrated

8. Self-contained expandable infrastructure
9. Platform Manager & Software Defined Networking Stack

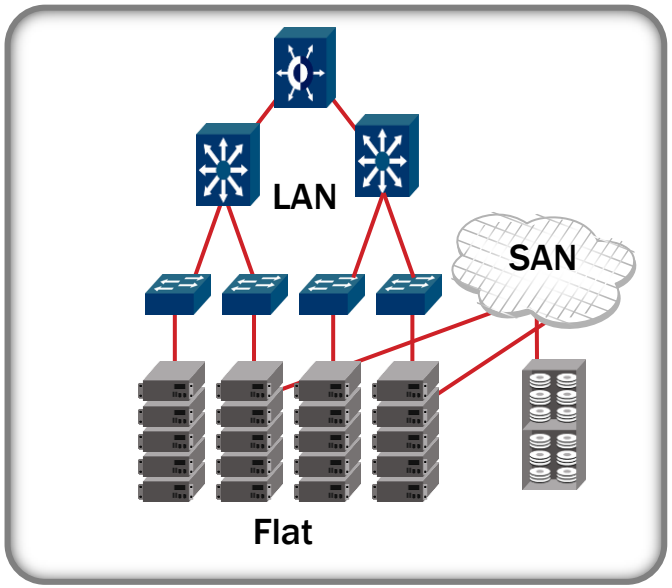
Optimized

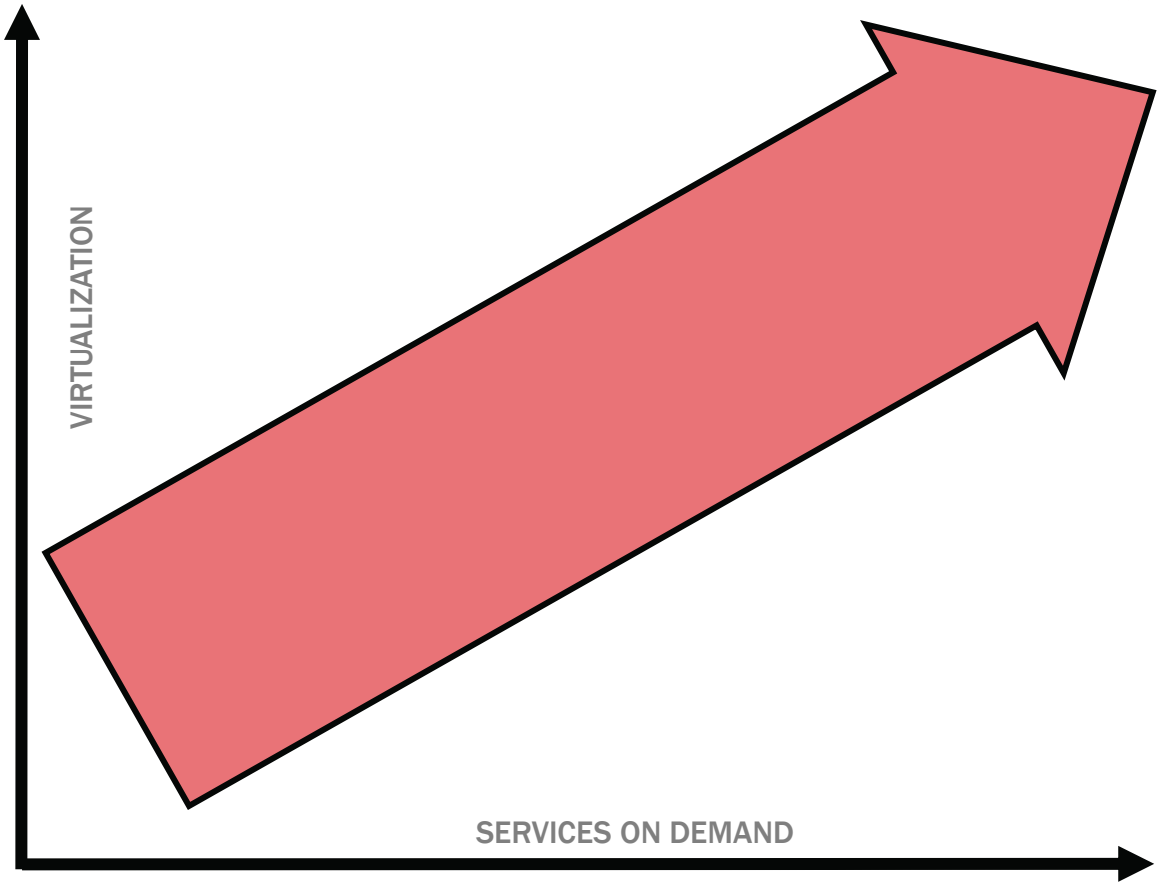
1. Fabric managed as a single switch
2. Converged fabric
3. Scalable fabric
4. Flexible Bandwidth
5. Optimized Traffic

Data Center Network Transformation

From networks to Ethernet fabrics

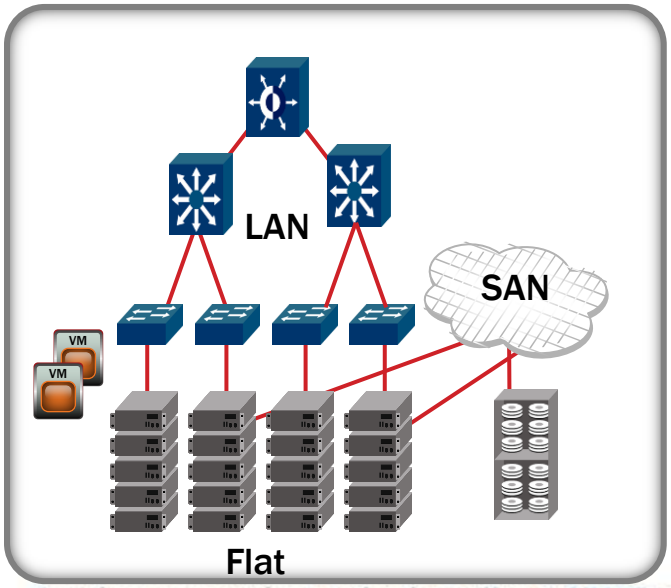
- Timeframe: 1990s
- Focus: Improve connectivity, packet delivery
- Historically 1 app:1 server; north-south traffic
- Virtualization ⇨ limited scalability
  - Traffic load strain
  - Increasing east-west traffic
- STP: one path, narrow VM mobility

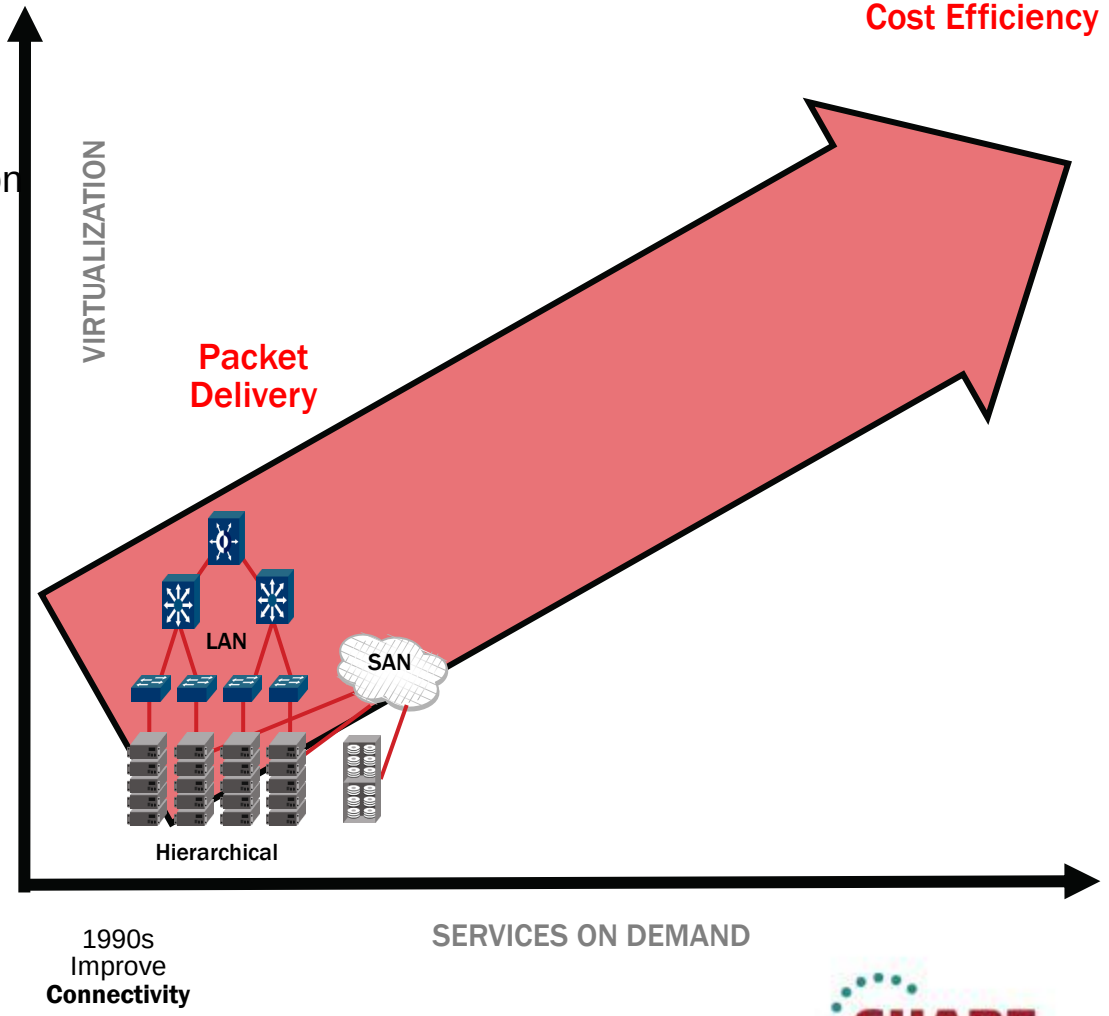




Data Center Network Transformation
From networks to Ethernet fabrics

- Timeframe: 2000s
- Focus: Improve performance/app delivery
- A more powerful, *flatter* network
  - Higher traffic, east-west, avoid congestion
  - Collapse layers to reduce complexity
- High density, high bandwidth, wire speed
- Layer 2 challenges remain…

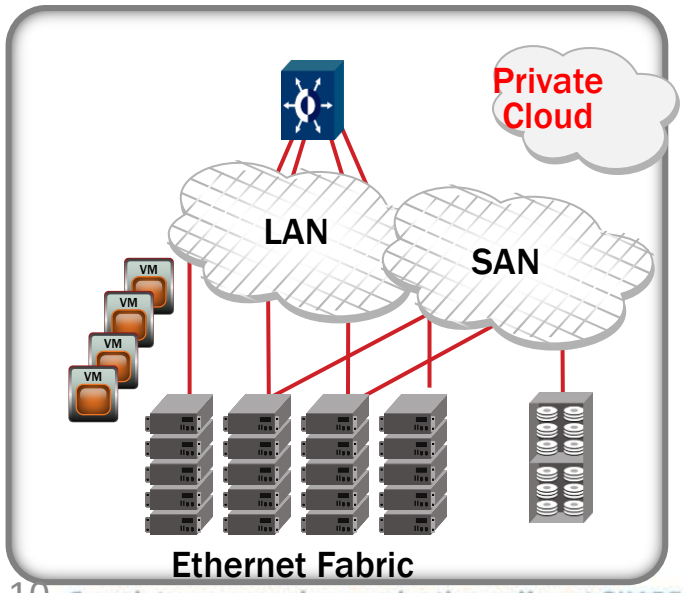


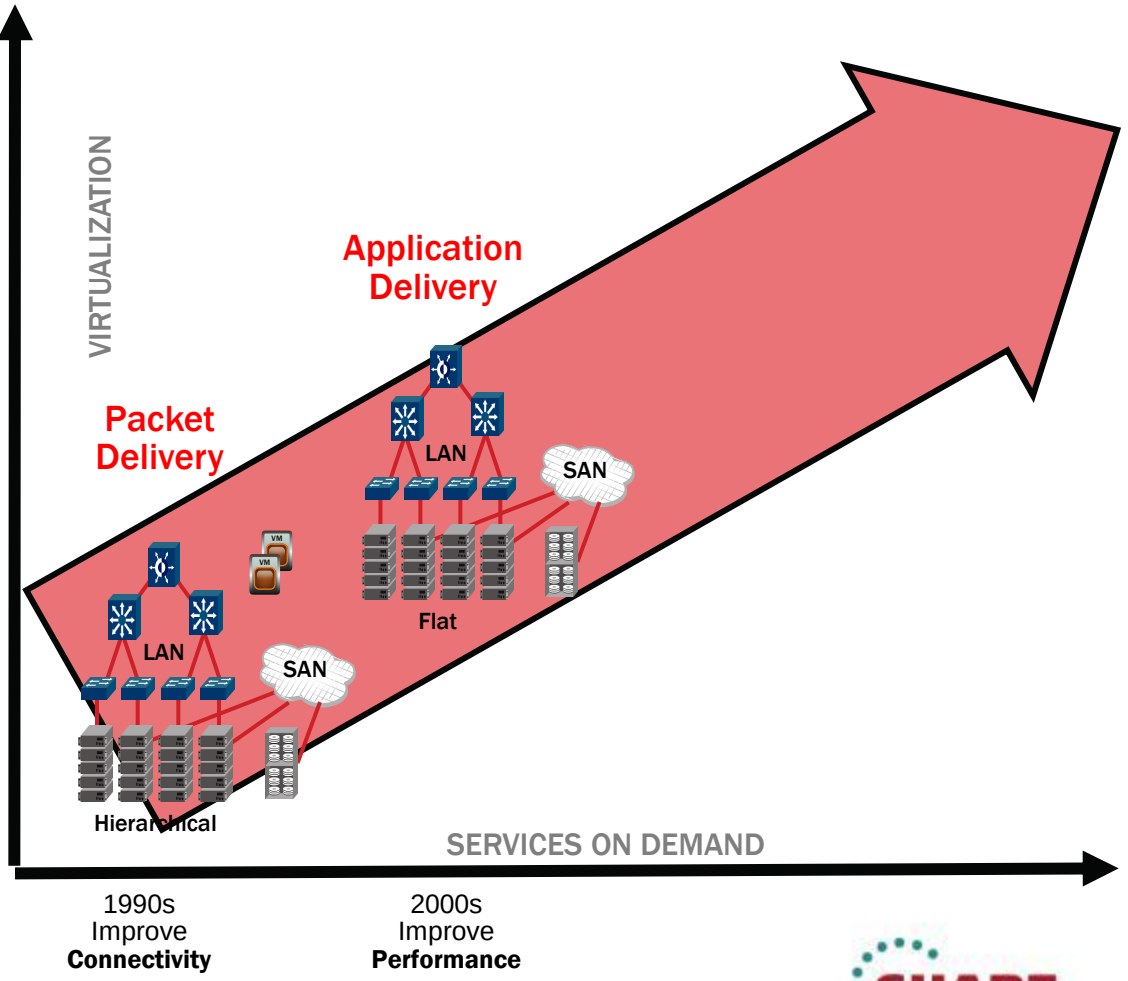


Data Center Network Transformation

From networks to Ethernet fabrics

- Timeframe: 2010s
- Focus: Improve agility
- Large, flat Layer 2, high speed, high availability
- **All** paths active—no STP
- Flexible topology
- Ability to converge IP/storage
- Wide, intelligent Virtual Machine (VM) mobility
- Manage as a single entity
- Virtualize for the cloud

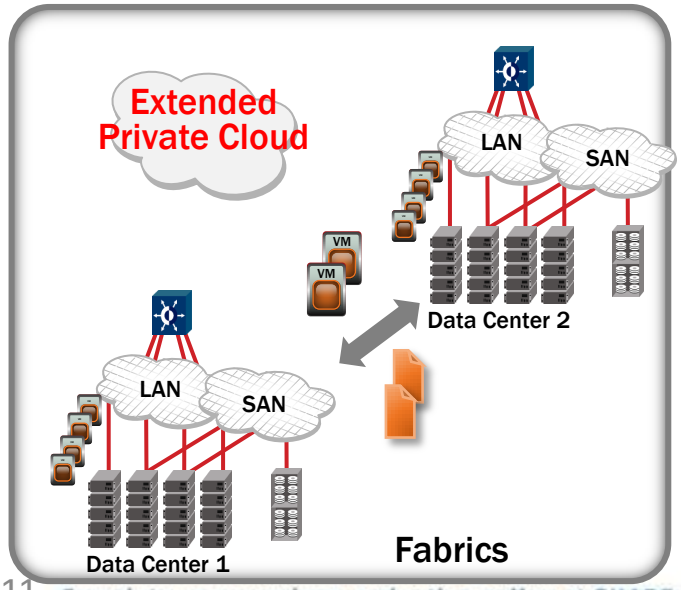


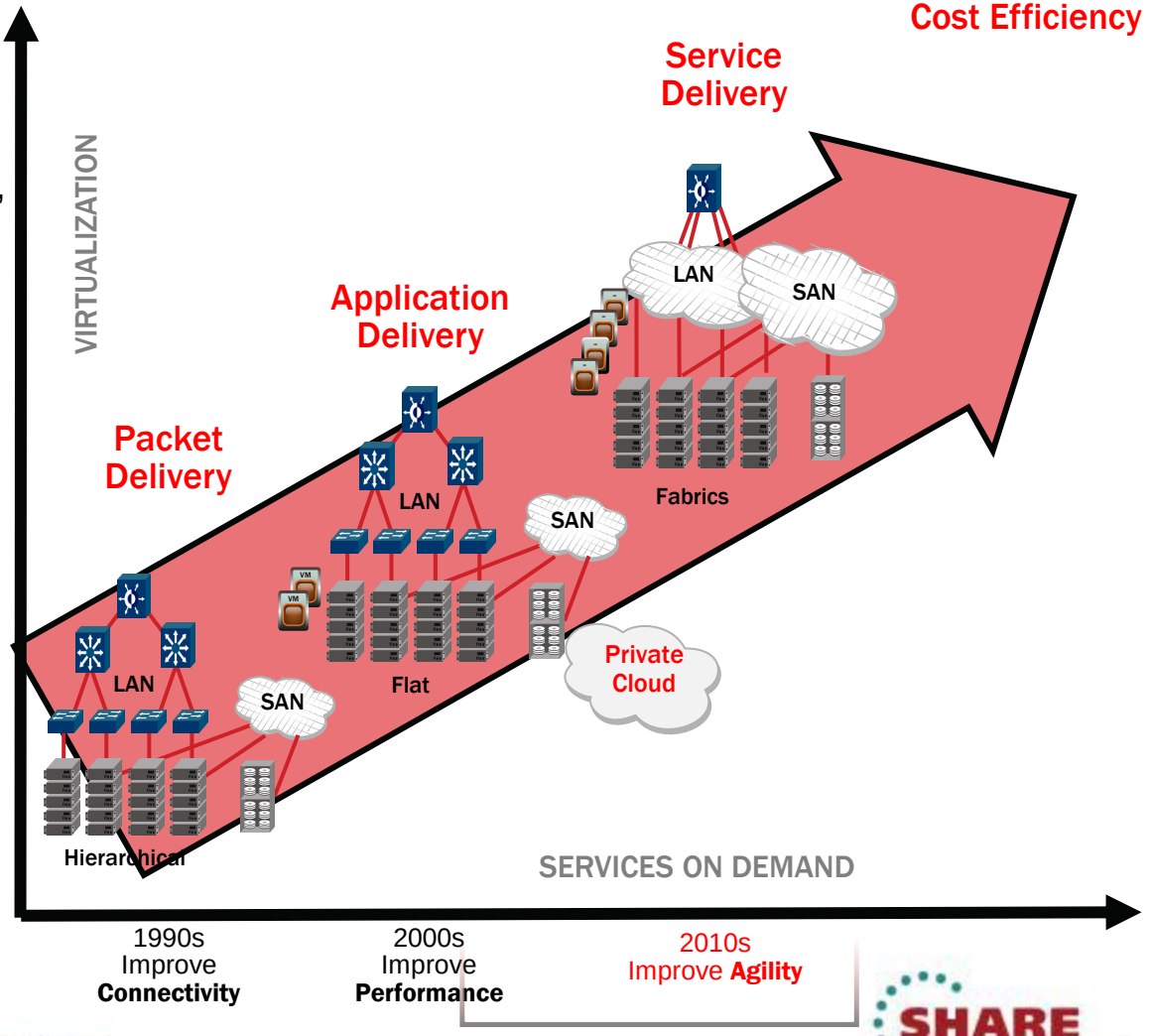


Data Center Network Transformation

From networks to Ethernet fabrics

- Timeframe: 2015+
- Focus: Improve the user experience
- Leverage resources across data centers
  - More flexibility to scale
  - Relocate applications for greater efficiency
- Layer 2 over distance, seamless mobility, rapid access
- Building on expertise to extend the private cloud

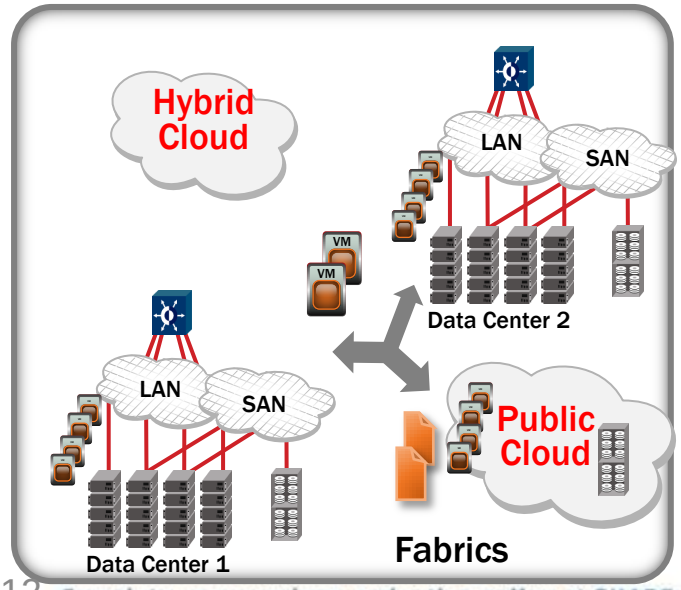


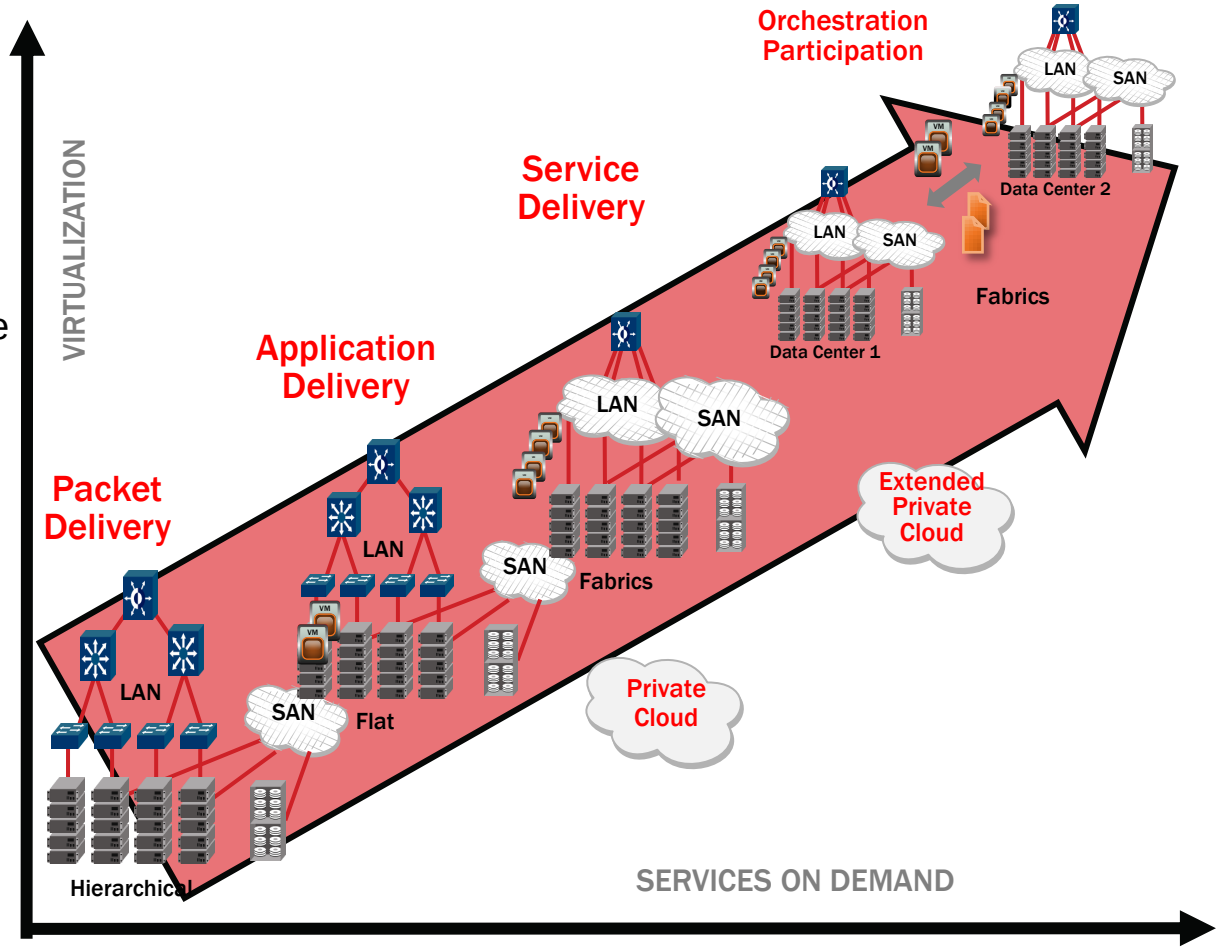


Data Center Network Transformation

From networks to Ethernet fabrics

- Timeframe: 2015+
- Focus: Orchestration
- Leverage service provider resources
  - Meet spikes/seasonal demand cost effectively
  - Accelerate application deployment
  - Resiliency in the event of a site outage
- Standards-based, open support, integrated management

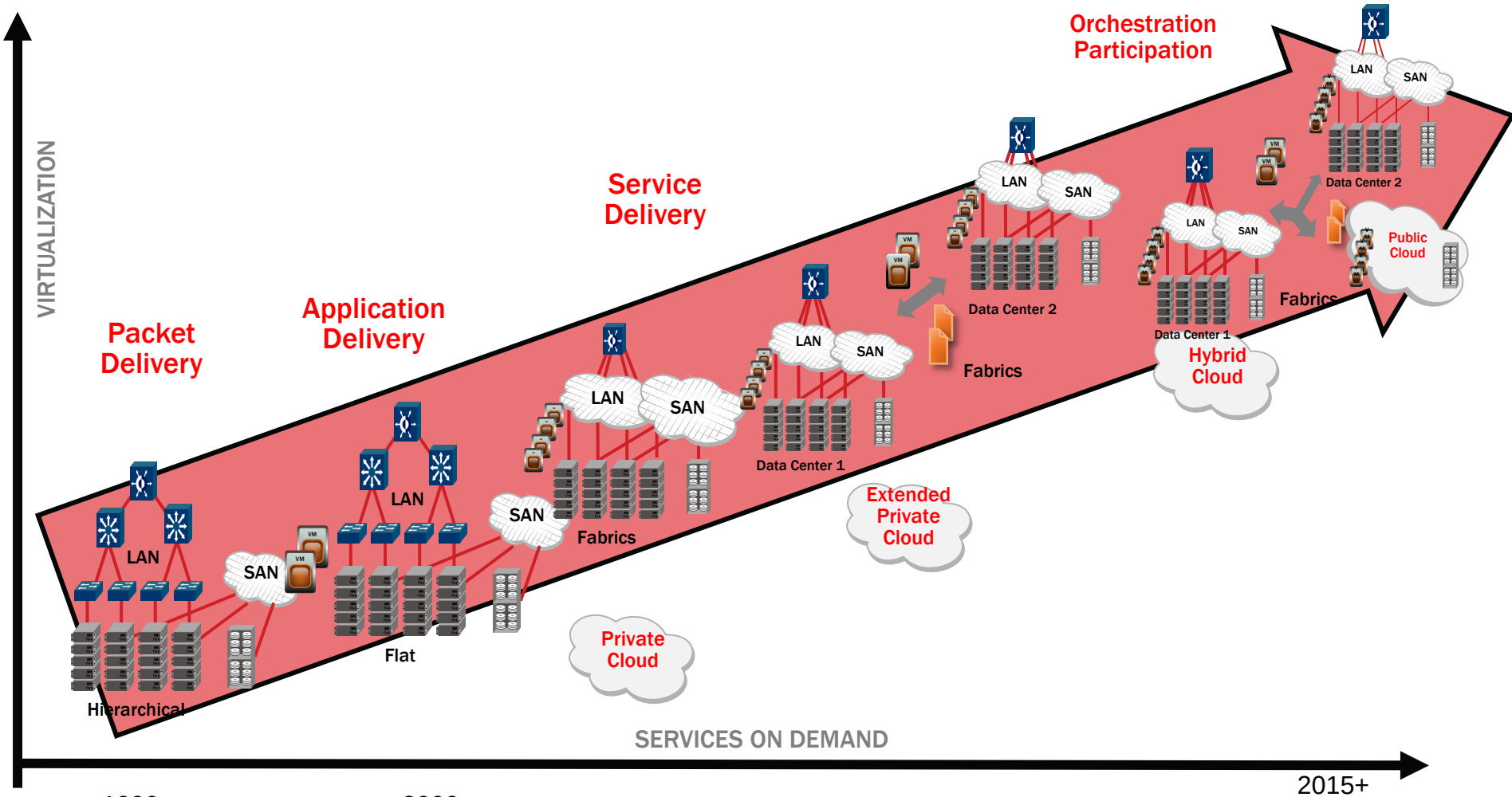




Data Center Network Transformation

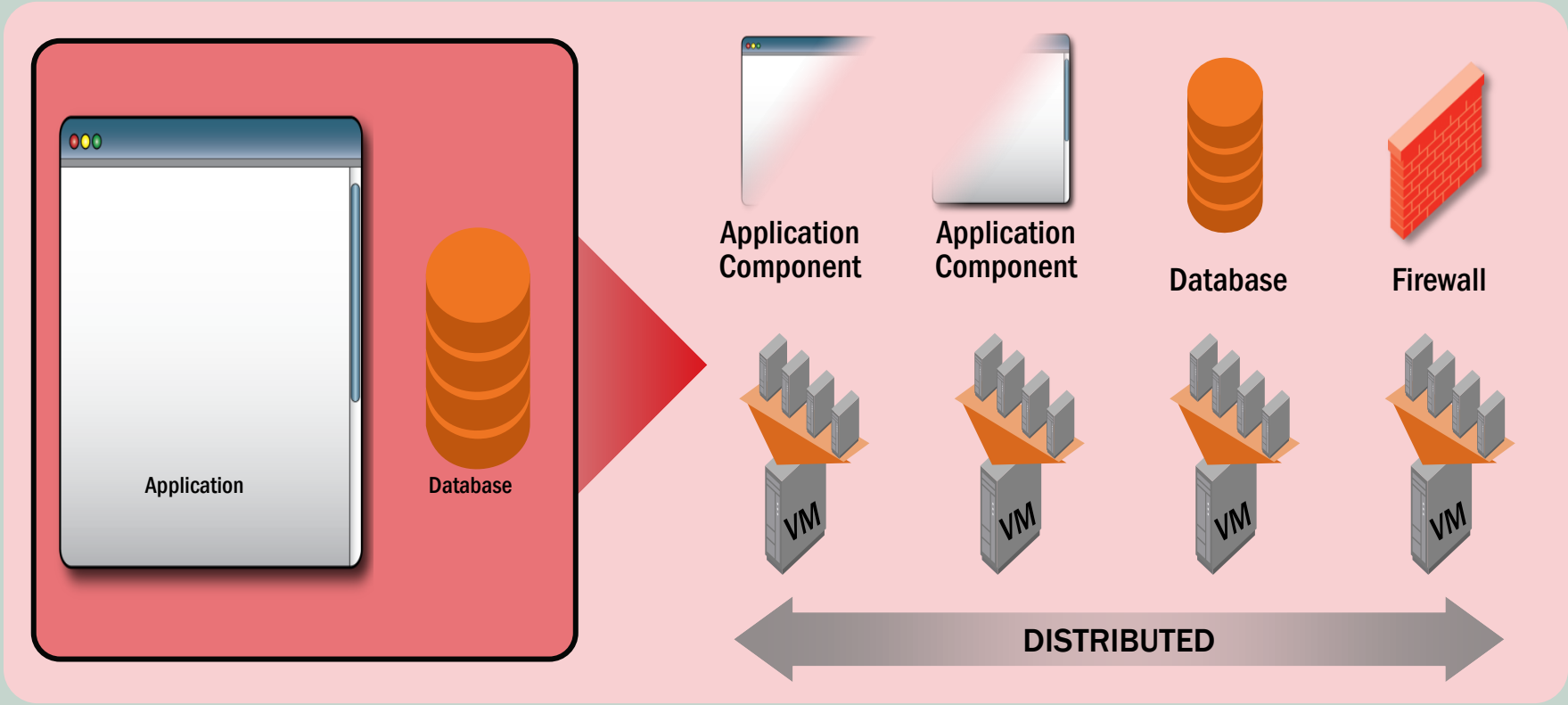
From networks to Ethernet fabrics

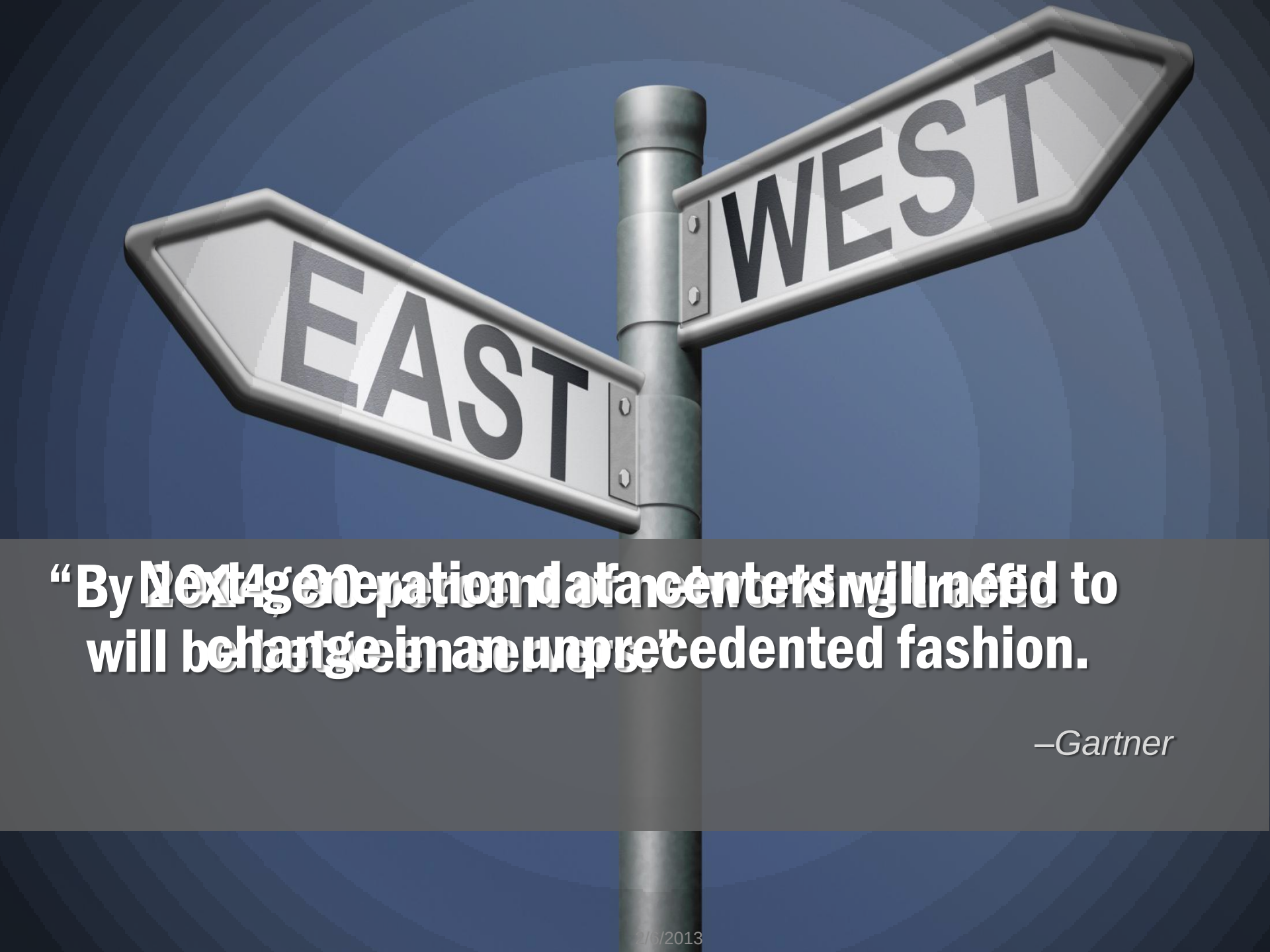
Business Agility
Cost Efficiency



What Are the Effects of This Transformation?

Applications will be disaggregated

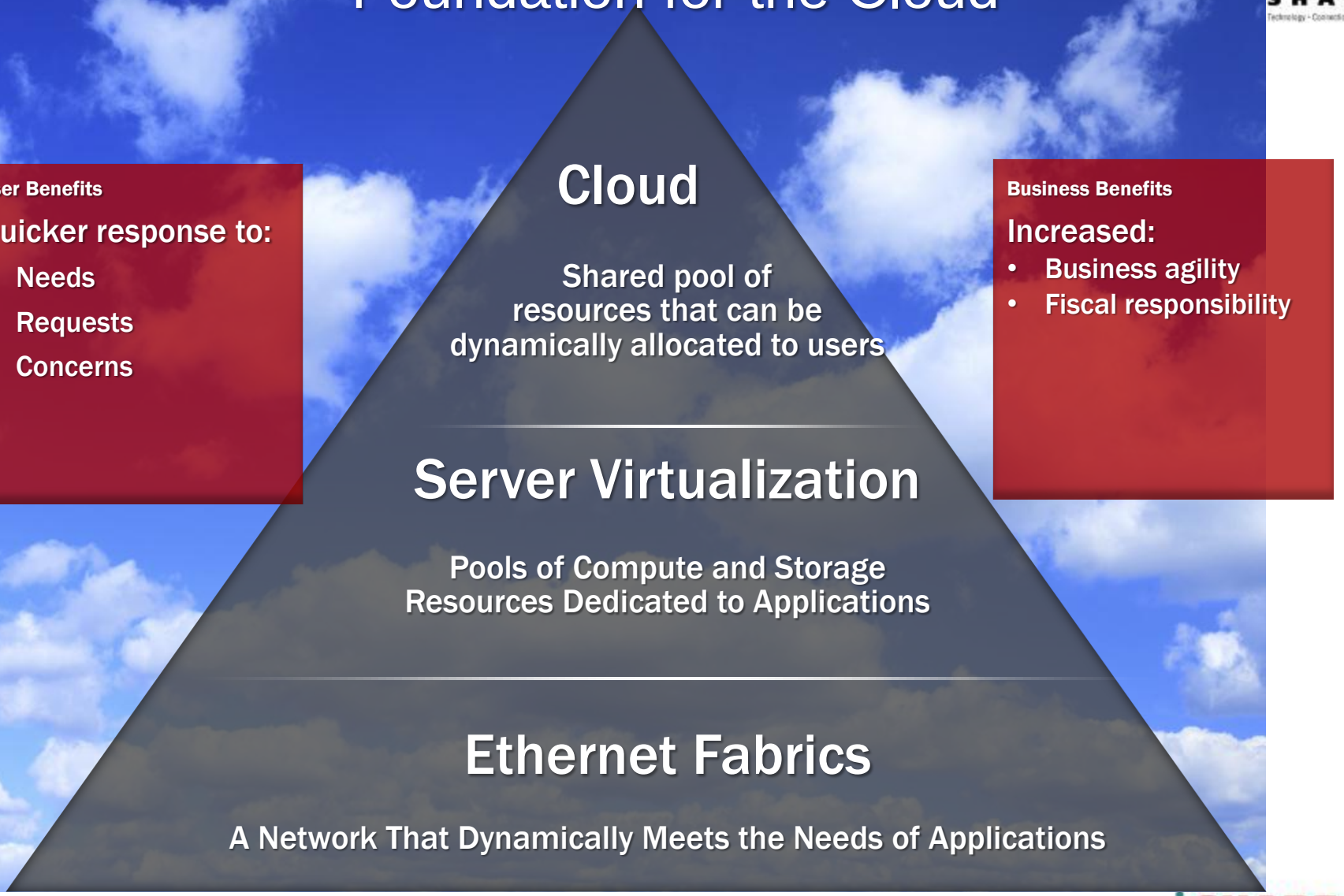




“By 2014, 20 percent of next-generation data center networking will be changed in an unprecedented fashion.

–Gartner

ETHERNET FABRICS
Foundation for the Cloud



User Benefits

Quicker response to:

- Needs
- Requests
- Concerns

Business Benefits

Increased:

- Business agility
- Fiscal responsibility

Effortless Connectivity
Better Service Delivery

- Resilient
- Flexible topology
- Scalable/elastic
- Flat architecture

Network Automation
Simpler Service Orchestration

- Logical chassis
- Automatic VM alignment
- Seamless convergence of storage, voice, and video

WHY ETHERNET FABRICS?

Future-Proof Data Center Networks

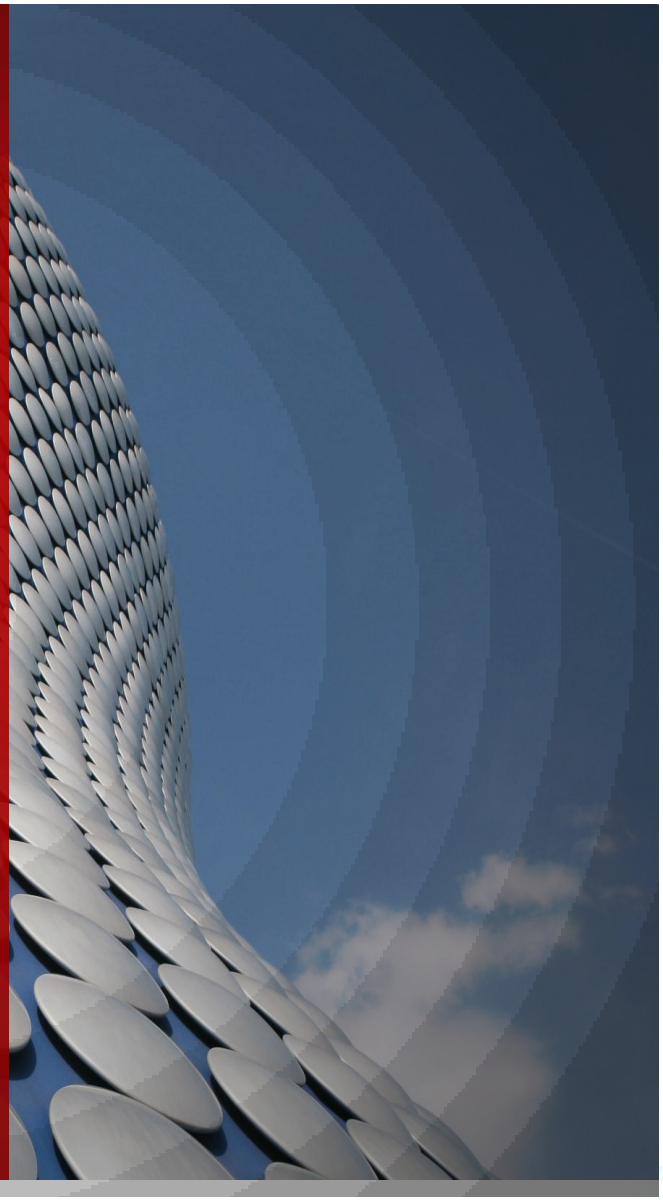
THE BUSINESS BENEFITS OF ETHERNET FABRICS

Enables organizations to:

- Leverage IT as an asset
- Reduce operational expenditures for
  data centers
- Install a data center infrastructure that is transparent to applications and users because
  it “just works” and is automated, flexible,
  and dynamic

STANDARDS, TERMS, AND TECHNOLOGIES

TRILL, SPB,
Flat Networks,
and Convergence



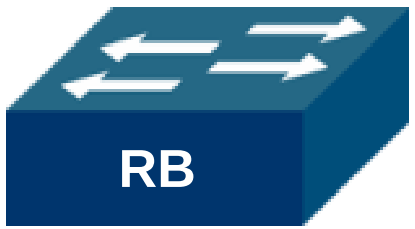
Ethernet Fabrics 101 Vernacular

Useful terms and definitions

- TRILL (Transparent Interconnect of Lots of Links) and SPB (Shortest Path Bridging)—Standards that provide multi-path, multi-hop capabilities in Ethernet fabrics
- Convergence—The ability of a single network infrastructure to support the needs of multiple technologies
- Fabric-based infrastructure versus storage fabric versus Ethernet fabric:
  - Fabric-based infrastructure—A Gartner term that refers to creating a fabric for everything
  - Storage fabric—Commonly called a Storage Area Network (SAN)
  - Ethernet fabric—A new network architecture for providing resilient, high-performance connectivity between clients, servers, and storage
- Flat network—A network in which all hosts can communicate with each other without needing a Layer 3 device

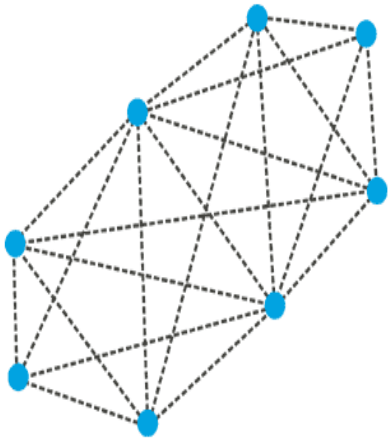
TRILL—Transparent Interconnect of Lots of Links

Overview



Devices are Routing Bridges (RBridges)

Data plane uses TRILL protocol



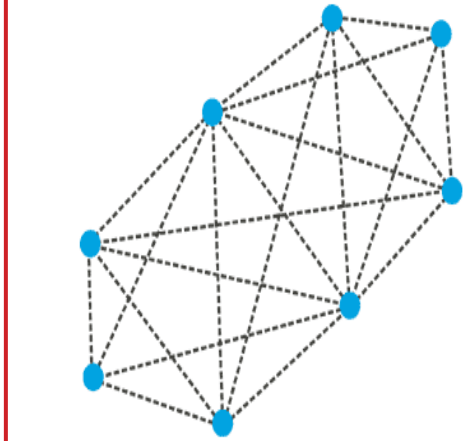
Control plane uses IS-IS Layer 2 link-state routing protocol

SPB—Shortest Path Bridging

Overview

Devices are Ethernet bridges (support 802.1ad stacking, ag OAM, and ah PBB)

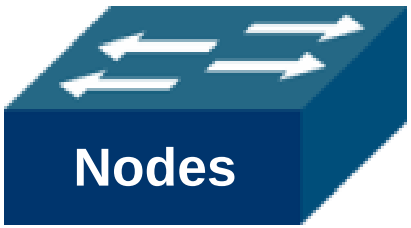
Data plane uses MAC-in-MAC



Control plane uses IS-IS Layer 2 link-state routing protocol

TRILL and SPB Use of IS-IS

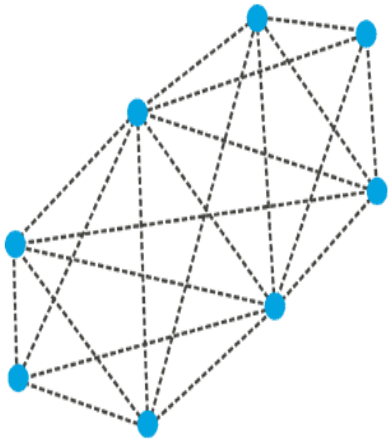
Functions



RBridges and SPB bridges:

- Use link-state Hellos to find each other
- Calculate shortest paths to all other RBridges/bridges
- Build routing tables

TRILL—Ingress RBridges encapsulate TRILL data; egress RBridges decapsulate TRILL data
SPB—Ingress bridge adds external MAC (destination); egress bridge removes external MAC

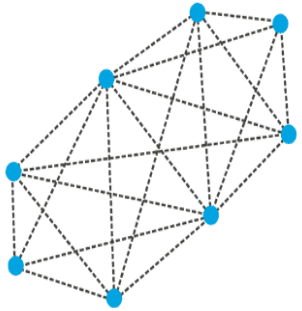


Link-state protocols

- Flood configuration information to nodes
- Used for shortest-path calculations
- Distribute configuration database

Role of Link-State Routing

Discovery and shortest path



Link-State Routing Protocols Are Used To:	Link-State Routing Neighbor Information
• Discover Ethernet fabric members • Determine Virtual LAN (VLAN) topology • Establish Layer 2 delivery using shortest-path calculations • Nodes tell every node on the network about their closest neighbor • The nodes distribute only the parts of the routing table containing their neighbors	• Gathered continuously • The list is flooded to all neighbors • Neighbors in turn send it to all of their neighbors and so on • Flooded whenever there is a (routing-significant) change • Allows nodes to calculate the best path to any other node in the network

TRILL vs. SPB

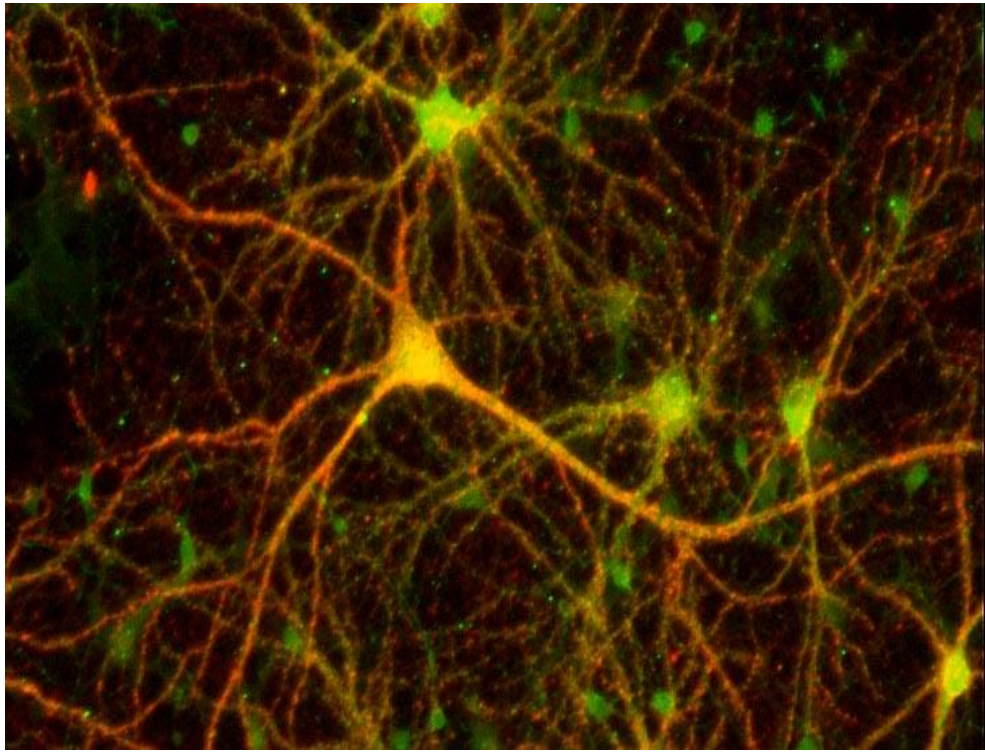
Different approaches to the same problem

Characteristic	TRILL	SPB
Standards Body	IETF	IEEE 802.1aq
Link-State Protocol	IS-IS (new PDUs)	IS-IS (new PDUs)
Encapsulation	TRILL Header	MAC-in-MAC
Multi-Path Support	Yes	Yes
Loop Mitigation	TTL	RPFC
Packet Flow	Hop by Hop	Symmetric
Configuration Complexity	Easy	Moderate
Troubleshooting	Moderate	Easy (OAM)

Flat Networking

TRILL and/or SPB allow for large Layer 2-based networks

- Hosts can directly communicate with each other without routers
- Highly interconnected, all paths available, and all links active
- Flat is synonymous with low latency
- Low latency is a fundamental building block for meeting user expectations



What Is Data Center Bridging (DCB)?

DCB is a collection of protocols that make Ethernet lossless

DCB-Related Protocols

Data Center Bridging Capabilities Exchange Protocol (DCBX)

- **Purpose**: Provides discovery and capability exchange protocol—extensions to LLDP
- **Benefit**: Enables the conveying of capabilities and configuration between neighbors

802.1Qbb: Priority-based Flow Control (PFC)

- **Purpose**: Enables control of individual data flows on shared lossless links
- **Benefit**: Allows frames to receive lossless service from a link that is shared with traditional LAN traffic, which is loss-tolerant

802.1Qaz: Enhanced Transmission Selection (ETS)

- **Purpose**: Permits organizations to manage bandwidth on the Ethernet link by allocating portions (percentages)
  of the available bandwidth to each of the groups
- **Benefit**: Bandwidth allocation allows traffic from the different groups to receive their target service rate (for example, 8 Gbps for storage and 2 Gbps for LAN). Bandwidth allocation provides Quality of Service (QoS)
  to applications

802.1Qau: Quantitized Congestion Notification (QCN)

- **Purpose**: Enables end-to-end congestion management.
- **Benefit**: Allows for throttling of traffic at the edge nodes of the network in the event of traffic congestion

Summary of Standards, Terms, and Technologies

Foundational components of an Ethernet fabric

- Flat networks—Allow any-to-any communication with routers
- TRILL—Transparent Interconnect of Lots of Links
- SPB—Shortest Path Bridging
- DCB—Data Center Bridging, provides for lossless Ethernet

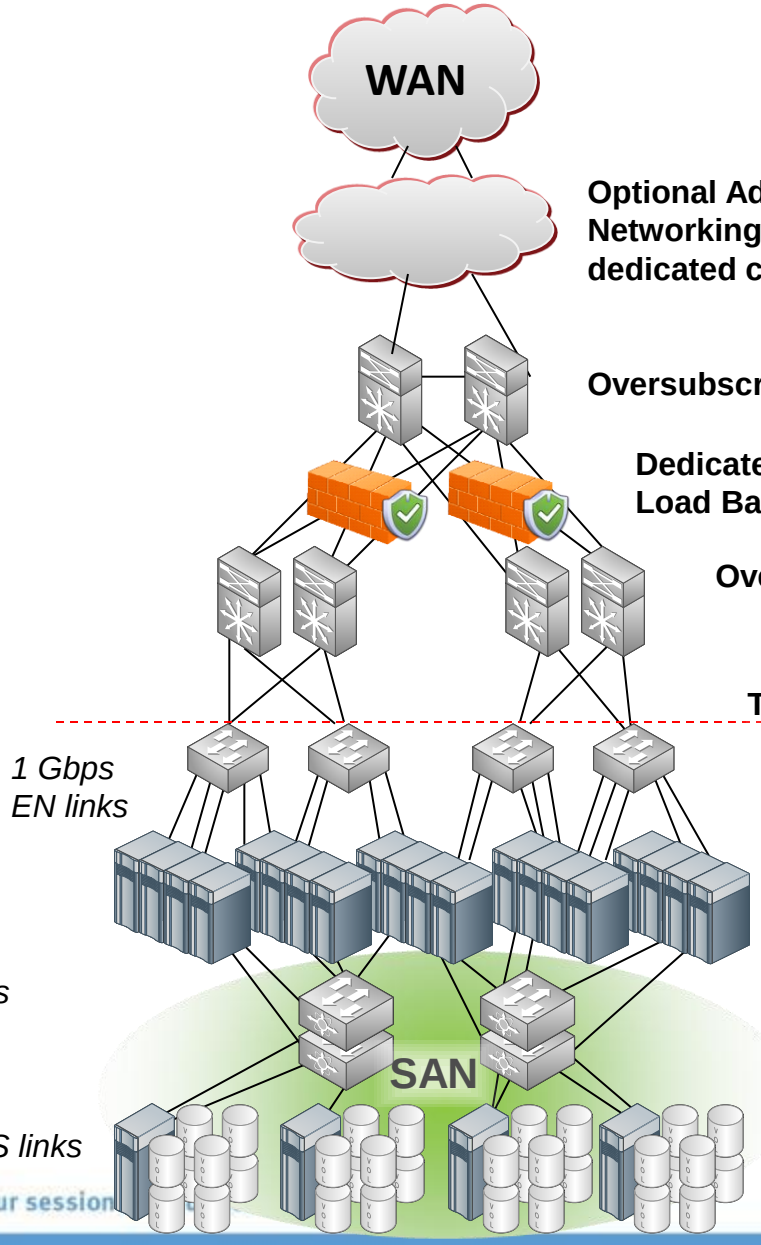
ODIN

The Open Datacenter Interoperable Network (ODIN)

- Standards and best practices for data center networking
  - Announced May 8 as part of InterOp 2012
    Five technical briefs (8-10 pages each), 2 page white paper, Q&A
    http://www-03.ibm.com/systems/networking/solutions/odin.html
  - Standards-based approach to data center network design, including descriptions of the standards that IBM and our partners agree upon

- IBM System Networking will publish additional marketing assets describing how our products support the ODIN recommendations
  - Technical white papers and conference presentations describing how IBM products can be used in these reference architectures
  - See IBM's Data Center Networking blog: https://www-304.ibm.com/connections/blogs/DCN/entry/odin_sets_the_standard_for_open_networking21?lang=en_us
  - And Twitter feed: https://twitter.com/#!/IBMCasimer

Traditional Closed, Mostly Proprietary Data Center Network



Optional Additional Networking Tiers, dedicated connectivity for server clustering

Oversubscribed Core Layer

Dedicated Firewall, Security, Load Balancing, L4-7 Appliances

Oversubscribed Aggregation Layer

Traditional Layer 2/3 Boundary

1 Gbps EN links

Oversubscribed Access Layer

Various Types of Application Servers, Blade Chassis

2, 4, 8 Gbps FC links

1 Gbps iSCSI / NAS links

Traditional Data Center Networks: B.O. (Before ODIN)

- Historically, Ethernet was used to interconnect “stations” (dumb terminals), first through repeaters and hubs, eventually through switched topologies
  - Not knowing better, we designed our data centers the same way
- The Ethernet campus network evolved into a structured network characterized by access, aggregation, services, and core layers, which could have 3, 4, or more tiers
- These networks are characterized by:
  - Mostly north-south traffic patterns
  - Oversubscription at all tiers
  - Low virtualization, static network state
  - Use of spanning tree protocol (STP) to prevent loops
  - Layer 2 and 3 functions separated at the access layer
  - Services (firewalls, load balancers, etc.) dedicated to each application in a silo structure
  - Network management centered in the switch operating system
  - Complex, often proprietary features and functions

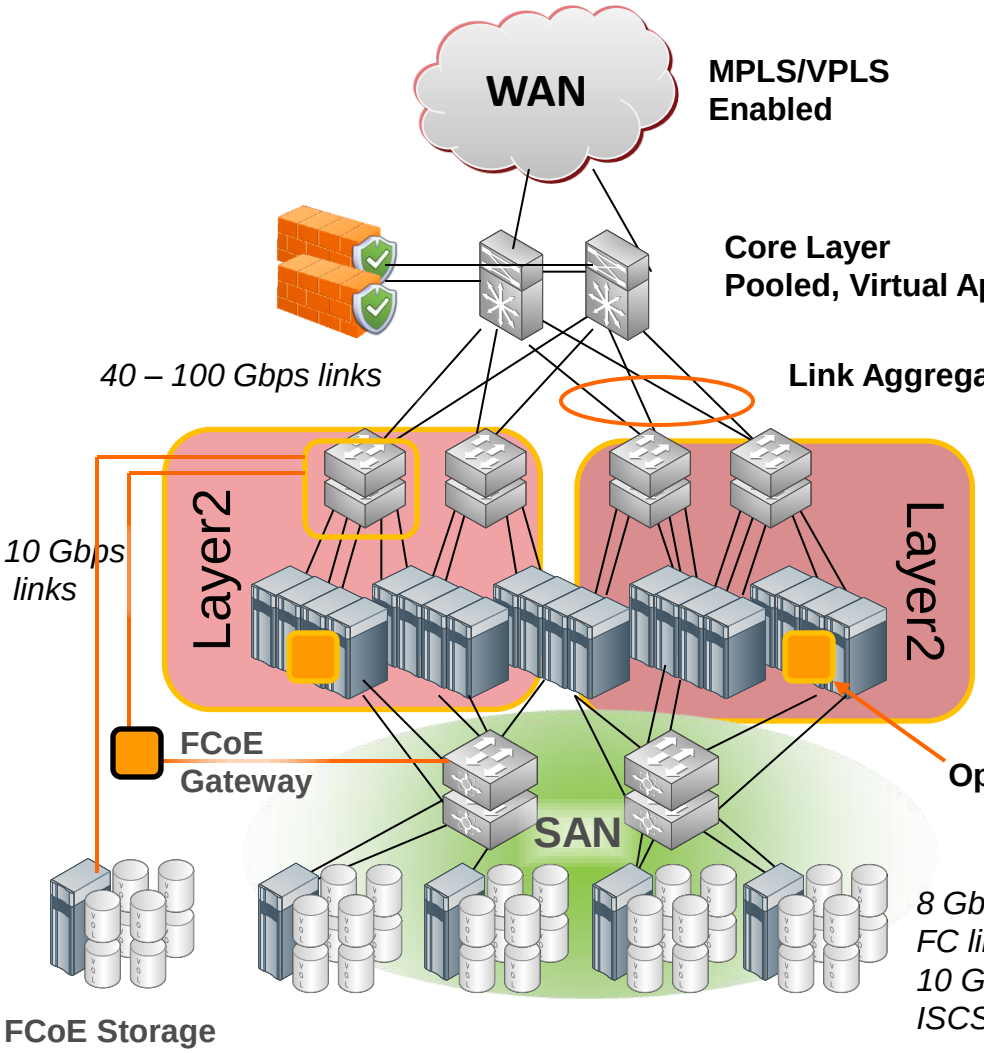
Problems with Traditional Networks

- Too many tiers
  - Each tier adds latency (10-20 us or more); cumulative effect degrades performance
  - Oversubscription (in an effort to reduce tiers) can result in lost packets
- Does not scale in a cost effective or performance effective manner
  - Scaling requires adding more tiers, more physical switches, and more physical service appliances
  - Management functions do not scale well
  - STP restricts topologies and prevents full utilization of available bandwidth
  - Physical network must be rewired to handle changes in application workload
  - Manually configured SLAs and security prone to errors
  - Potential shortages of IP Addresses

Problems with Traditional Networks

- Not optimized for new functions
  - Most modern data center traffic is east-west
  - Oversubscription / lossy network requires separate storage infrastructure
  - Increasing use of virtualization means significantly more servers which can be dynamically created, modified, or destroyed
  - Desire to migrate VMs for high availability and better utilization
  - Multi-tenancy for cloud computing and other applications
- High Operating and Capital Expense
  - Too many protocol specific network types
  - Too many network, service, and storage managers
  - Too many discrete components lowers reliability, poorly integrated
  - Too much energy consumption / high cooling costs
  - Sprawl of lightly utilized servers and storage
  - Redundant networks required to insure disjoint multi-pathing for high availability
  - Moving VMs to increase utilization limited by Layer 2 domain boundaries, low bandwidth links, & manual management issues
  - Significant expense just to maintain current network, without deploying new resources

Open Datacenter with an Interoperable Network (ODIN)



Modern Data Center Networks: A.O. (After ODIN)

- Modern data centers are characterized by:
  - 2 tier designs (with embedded Blade switches and virtual switches within the servers)
    - Lower latency and better performance
    - Cost effective scale-out to 1000s of physical ports, 10,000 VMs (with lower TCO)
    - Scaling without massive oversubscription
    - Less moving parts → higher availability and lower energy costs
    - Simplified cabling within and between racks
    - Enabled as an on-ramp for cloud computing, integrated PoDs, and end-to-end solutions
  - Optimized for east-west traffic flow with efficient traffic forwarding
  - Large Layer 2 domains and networks enable VM mobility across different physical servers
    - "VM Aware" fabric; network state resides in vSwitch, automated configuration & migration of port profiles
    - Options to move VMs either through hypervisor Vswitch or external switch

- ODIN Provides a Data Center Network Reference Design based on Open Standards

Modern Data Center Networks: A.O. (After ODIN)

- Modern data centers are characterized by:
  - “Wire once” topologies with virtual, software-defined overlay networks
  - Pools of service appliances shared across multi-tenant environments
  - Arbitrary topologies (not constrained by STP) with numerous redundant paths, higher bandwidth utilization, switch stacking, and link aggregation
  - Options to converge SAN (and other RDMA networks) into a common fabric with gateways to existing SAN, multi-hop FCoE, disjoint fabric paths, and other features
  - Management functions are centralized, moving into the server, and require fewer instances with less manual intervention and more automation
    - Less opportunity for human error in security and other configurations
  - Based on open industry standards from the IEEE, IETF, ONF, and other groups, which are implemented by multiple vendors (lower TCO per Gartner Group report)

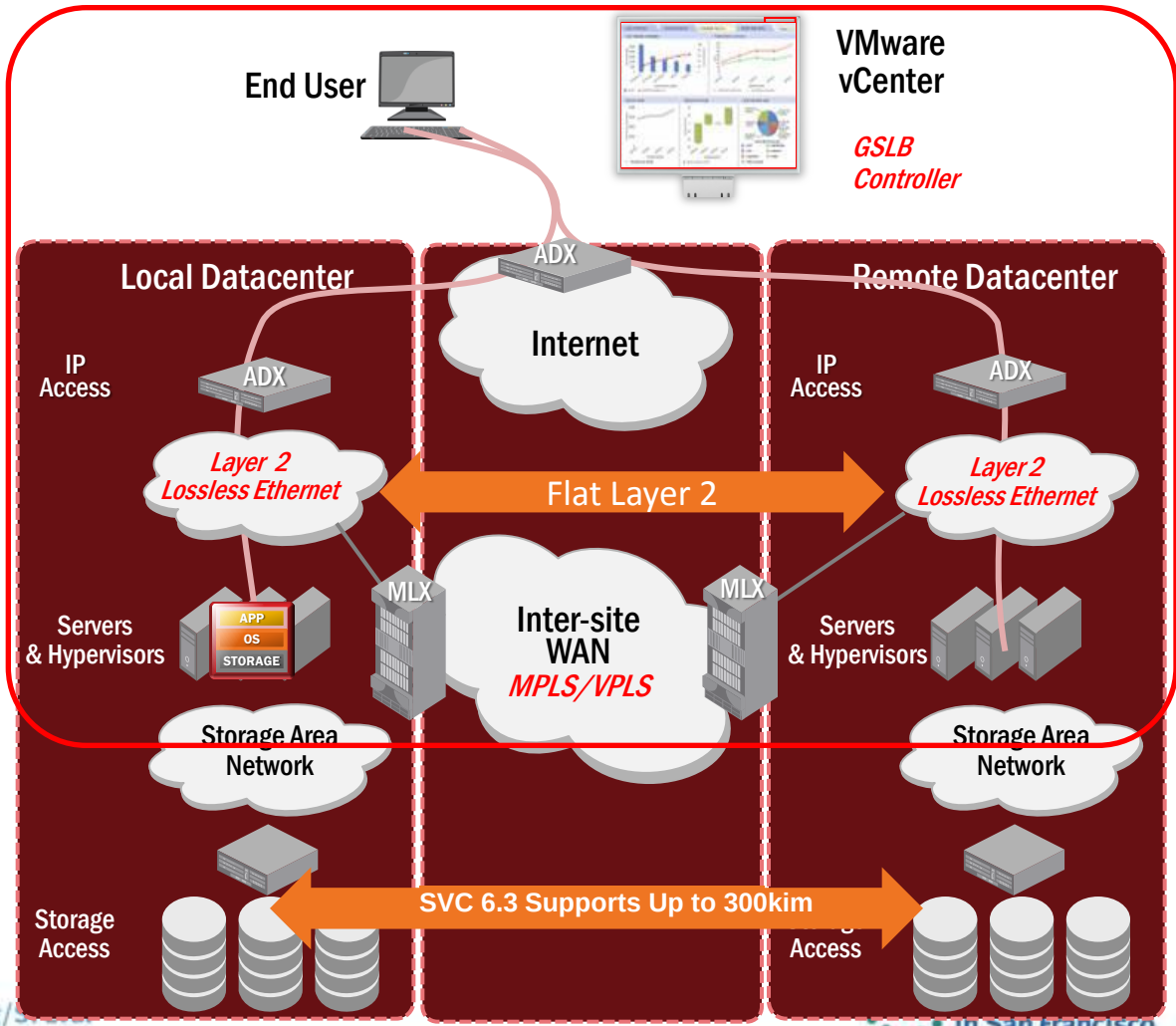
- ODIN Provides a Data Center Network Reference Design based on Open Standards

An ODIN Example: VM mobility and Multi-site Deployment

- VM mobility improves resource efficiency and application availability
- Multi-site deployment involves moving workload between two physical locations
- VM Hypervisors and Storage Virtualization provide continued access independent of physical location
- Infrastructure provides the foundation that is required for VM Hypervisors and Storage Virtualization to work in tandem and transparently
- Drivers include disaster backup and zero down time, global enterprises (follow the sun), optimization for power cost (follow the moon)

An ODIN Example: Infrastructure to support VM Mobility

- Server L2 VLAN Connectivity with Lossless Ethernet, Flat L2 fabric
- IBM Storage Volume Controller (SVC) using Stretch Clustering provides Read/Write Access to volumes across sites & provides data replication
  - **Third site for quorum disk not shown**
- Brocade Fibre Channel switches with ADX option
- VMware vMotion enables transparent migration of virtual machines, their corresponding applications and data over distance with intelligent IP load balancing
- Intersite 10G Layer 2 VLAN with MPLS/VPLS via WDM, 16G ISLs, or FC-IP option on Brocade
  - **Optional In-Flight 2:1 Compression increases link utilization (per AES-GCM-256)**
  - **Optional In-Flight switch-switch Encryption, 64 GB per IS, per AES-GCM ECB mode, 256 bit key**



Broad Ecosystem of Support for ODIN

"In order to contain both capital and operating expense, this network transformation should be based on open industry standards."

"ODIN…facilitates the deployment of new technologies"

InterOp Webinar: "How to Prepare Your Infrastructure for the Cloud Using Open Standards"

"ODIN is a great example of how we need to maintain openness and interoperability"

preferred approach to solving Big Data and network bottleneck issues

National Science Foundation interop lab & Wall St. client engagement

"…the missing piece in the cloud computing puzzle"

"…one of the fundamental "change agents" in the networking industry…associated with encouraging creativity… a nearly ideal approach…is on its way to becoming industry best-practice for transforming data-centers"

"We are proud to work with industry leaders like IBM"

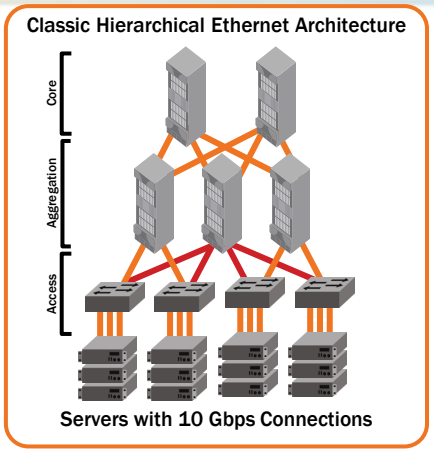
TRANSITIONING TO AN ETHERNET FABRIC

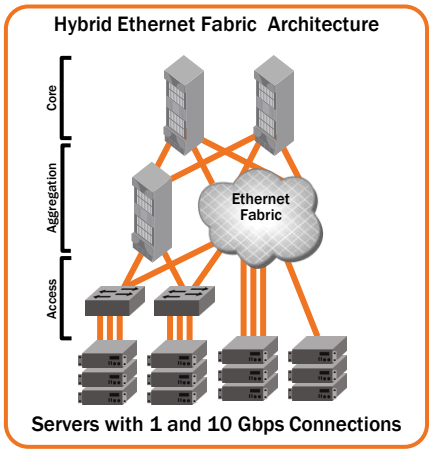
Traditional Multilayer vs. Fabric-Based Networks

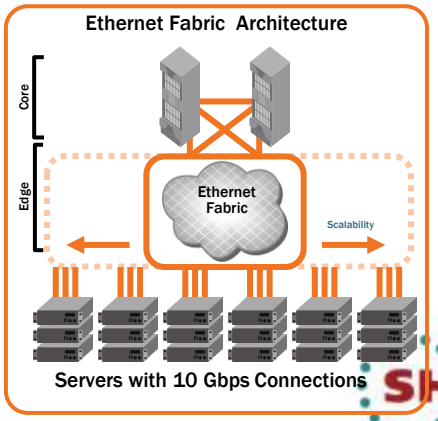
Flat Networking

Migrating to a flat network

- Migrate in stages
  - Identify applications and/or projects that can benefit from an Ethernet fabric
- Leverage current Layer 3 devices
  - All inter-VLAN communication and security boundaries handled in the same fashion
- Ethernet fabrics use updated broadcast mechanisms
  - Reduced flooding within intelligent fabric
  - BPDU drop capabilities

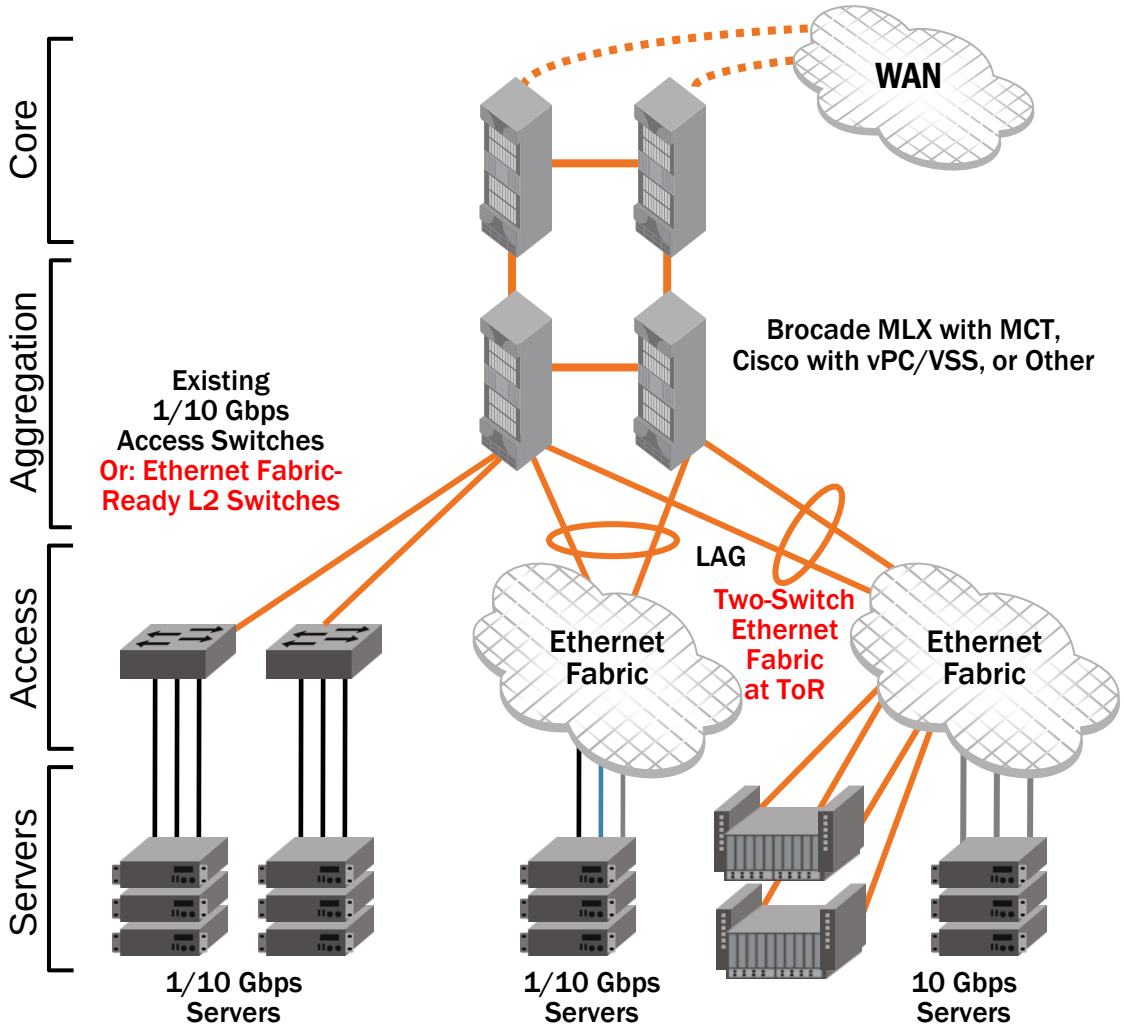






Ethernet Fabric Transition: Use Case 1

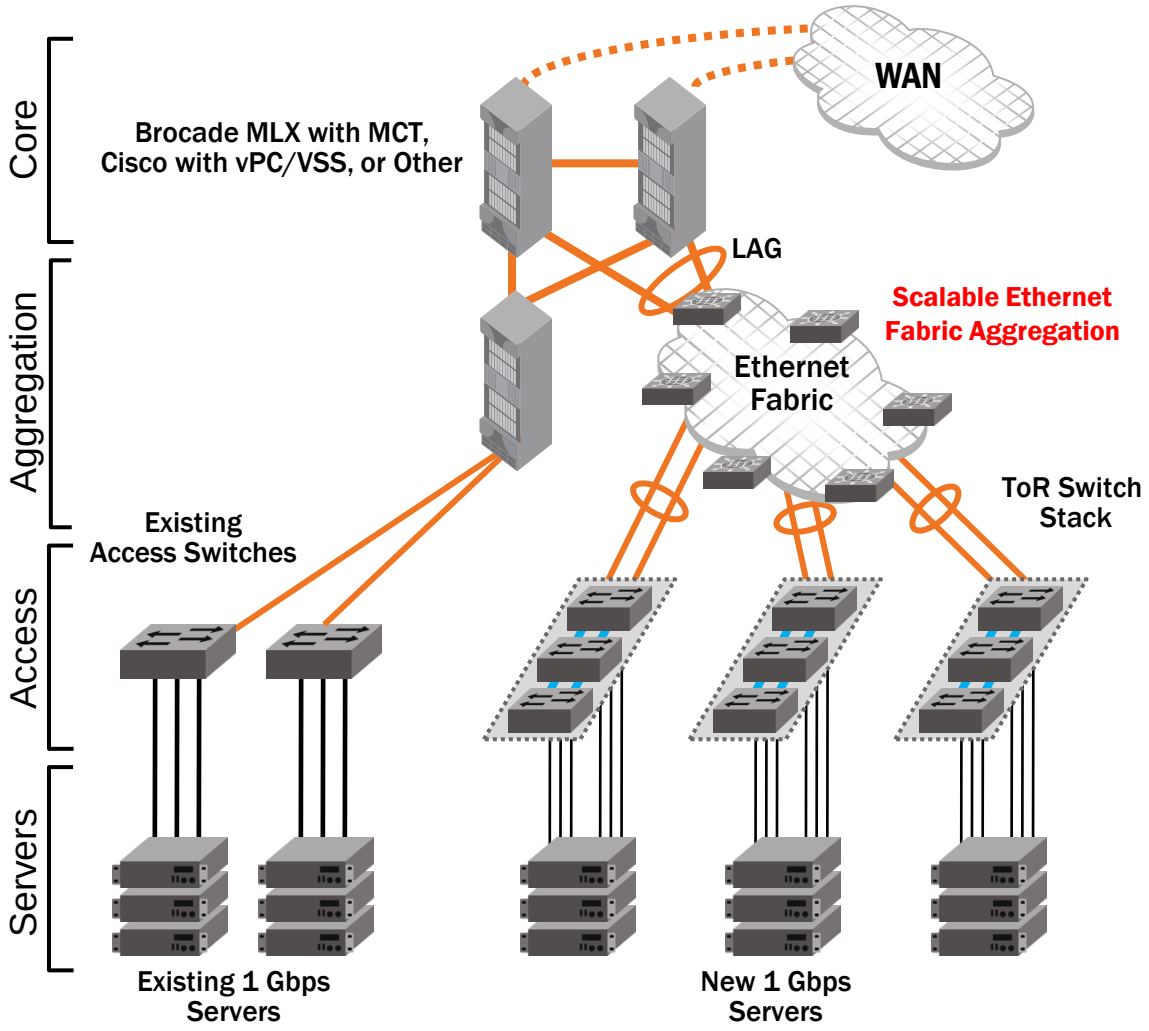
1/10 Gbps Top-of-Rack (ToR) access architecture



- Preserves existing architecture
  - Leverages existing core/ aggregation
  - Coexists with existing ToR switches
- Supports 1 Gbps and 10 Gbps server connectivity
- Active-active network
  - Load splits across connections
- No single point of failure
  - Self healing
- Fast link re-convergence
  - < 250 milliseconds
- High-density access with flexible subscription ratios

Ethernet Fabric Transition: Use Case 2

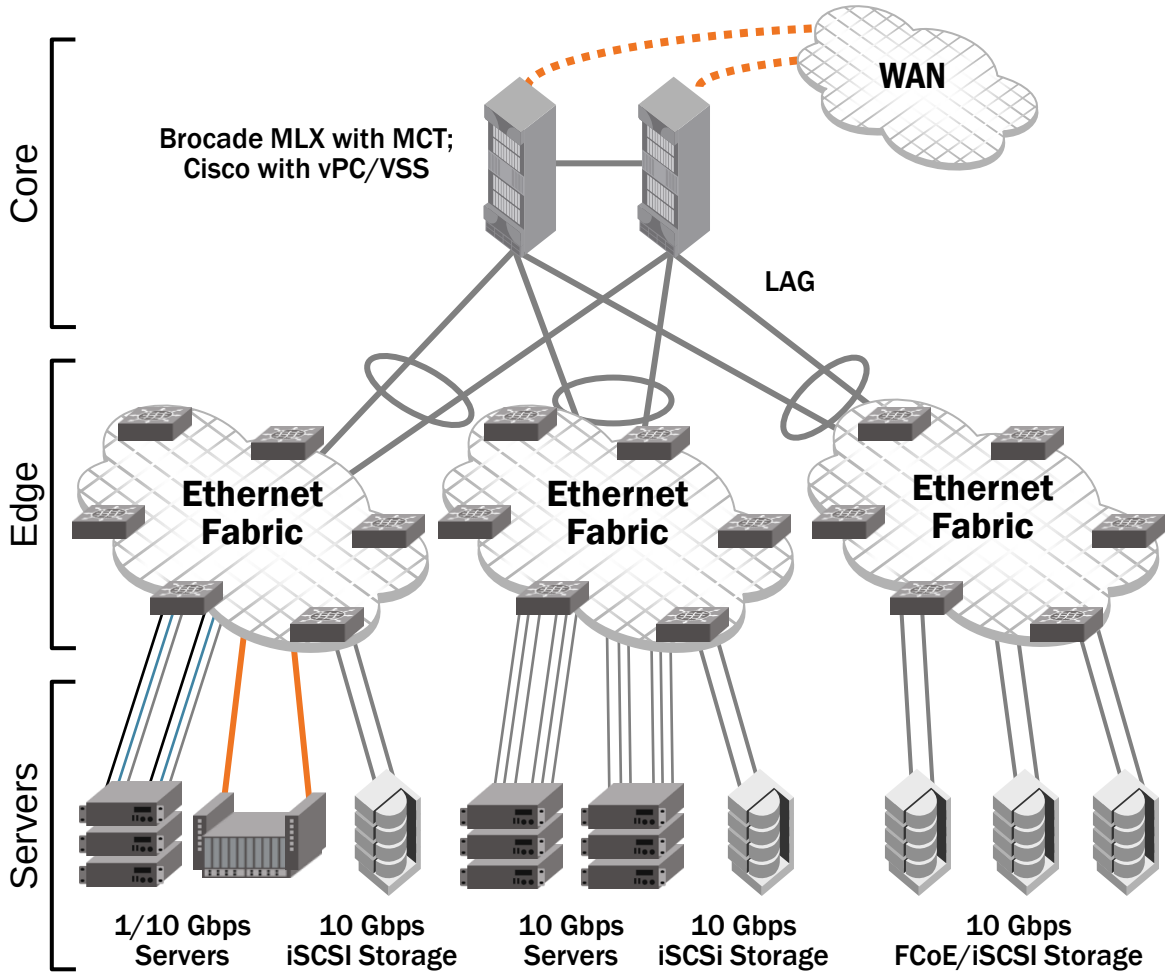
10 Gbps aggregation; 1 Gbps Top-of-Rack (ToR) access architecture



- Low-cost, highly flexible logical chassis at aggregation layer
  - Building block scalability
  - Per-port price similar to a ToR switch
  - Availability, reliability, manageability of a chassis
  - Flexible subscription ratios
- Ideal aggregator for 1 Gbps ToR switches
- Optimized multi-path network
  - No single point of failure
  - STP not necessary

Ethernet Fabric Transition: Use Case 3

1/10 Gbps access; network convergence architecture



- Flatter, simpler network design
  - Logical two-tier architecture
- Greater Layer 2 scalability/ flexibility
  - Increased sphere of VM mobility
  - Seamless network expansion
- Optimized multi-path network
  - All paths are active
  - No single point of failure
  - STP not necessary
- Convergence-ready
  - DCB support within Ethernet fabrics
  - Multi-hop FCoE support within Ethernet fabrics
  - Lossless iSCSI

Conclusions

- Accelerating change in enterprise data center networks
  - Under-utilized servers, rising energy costs, limited scalability, dynamic workload management
  - Need to automate, integrate, and optimize data center networks
- Market forces and cloud adoption are causing a deconstruction of IT models
- Classic network architectures are too complex, rigid
- Scalable, flexible, and high-performance Ethernet fabrics provide greater virtualization ROI and lay the foundation for cloud-based data centers
- Learn more about Ethernet fabrics: www.ethernetfabric.com

- The Path Towards an Open Datacenter with an Interoperable Network (ODIN)
  - Announced as part of InterOp 2012
    Five technical briefs (8-10 pages each), 2 page white paper, Q&A
    http://www-03.ibm.com/systems/networking/solutions/odin.html
  - Standards-based approach to data center network design, including descriptions of the standards that IBM and partners agree upon
  - Support from 9 major participants (including Marist College, Brocade and the IBM SDN/OpenFlow Lab)