Data Center Fabric Architectures

Ivan Pepelnjak (ip@ioshints.info)
NIL Data Communications

Podatkovne komunikacije
Data Communications

NIL ASSIST

NIL HYPER center

Who is @ioshints?

- Networking engineer since 1985 (DECnet, Netware, X.25, OSI, IP ...)
- Technical director, later Chief Technology Advisor @ NIL Data Communications
- Started the first commercial ISP in Slovenia (1992)
- Developed BGP, OSPF, IS-IS, EIGRP, MPLS courses for Cisco Europe
- Architect of Cisco's Service Provider (later CCIP) curriculum
- Consultant, blogger, book author

Focus:

- Core routing/MPLS, IPv6, VPN, Data centers, Virtualization
- Rock climbing, mountain biking ;)

Agenda

- Why do we care?
- What exactly is a fabric?
- What shall I ask for?

Common fabric architectures

- Shared management plane
- Shared control plane
- Shared data plane
- Flow-based configuration

Warning: the author is known to be highly biased toward scalable L3 solutions

Why Does It Matter?

Cloud computing is the future

Regardless of personal opinions and foggy definitions

Cloud computing requires large-scale elastic data centers

Hard to build them using the old tricks

Modern applications generate lots of east-west (inter-server) traffic

Existing DC designs are focused on north-south (server-to-user) traffic

What Is a Fabric?

Juniper

- Any-to-any non-blocking connectivity
- Low latency and jitter
- No packet drops under congestion
- Linear cost and power scaling
- Support of virtual networks and services
- Modular distributed implementation
- Single logical device

The answer seems to depend on the capabilities of your gear

Cisco

- Open (standards-based)
- Secure (isolation of virtual zones)
- Resilient (fault-tolerant, stateless)
- Scalable
- Flexible (incl. auto-provisioning)
- Integrated (compute, network & storage)

Brocade

- Flatter
- Intelligent (auto-discovery)
- Scalable (multi-pathing)
- Efficient (automatic shortest path fwd)
- Simple (single logical entity)

What Should You Ask For?

Forwarding features

- Storage & network integrated over 10GE (iSCSI or FCoE)
- Lossless traffic classes (DCB)
- Massive L3 multipathing
- Optional: L2 multipathing
- Fewer hops (lower latency)
- More east-west bandwidth

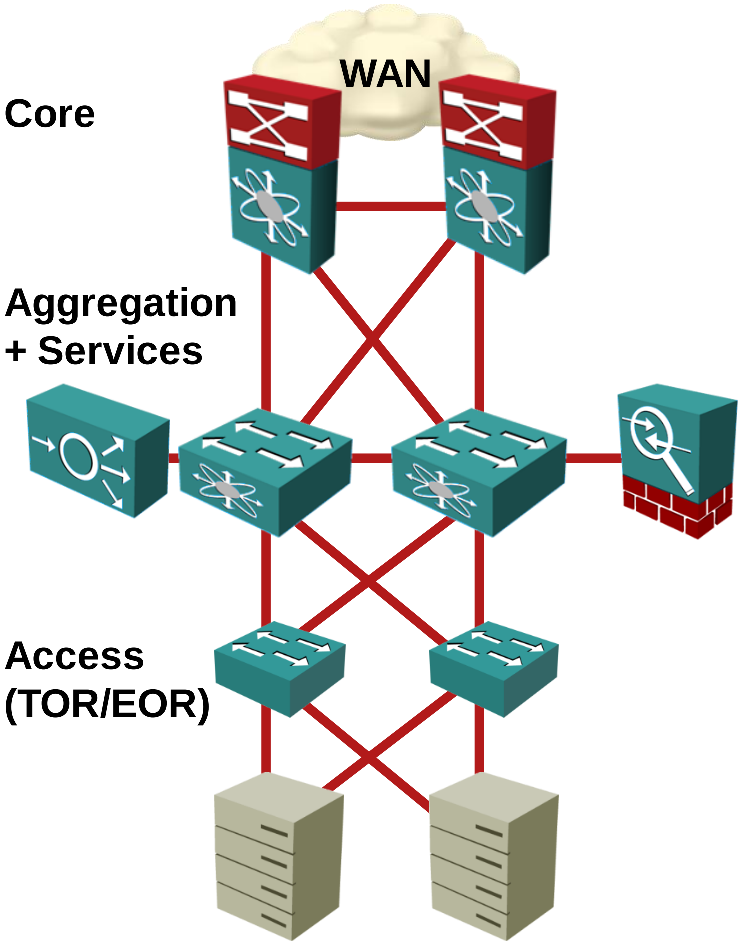
Control & management features

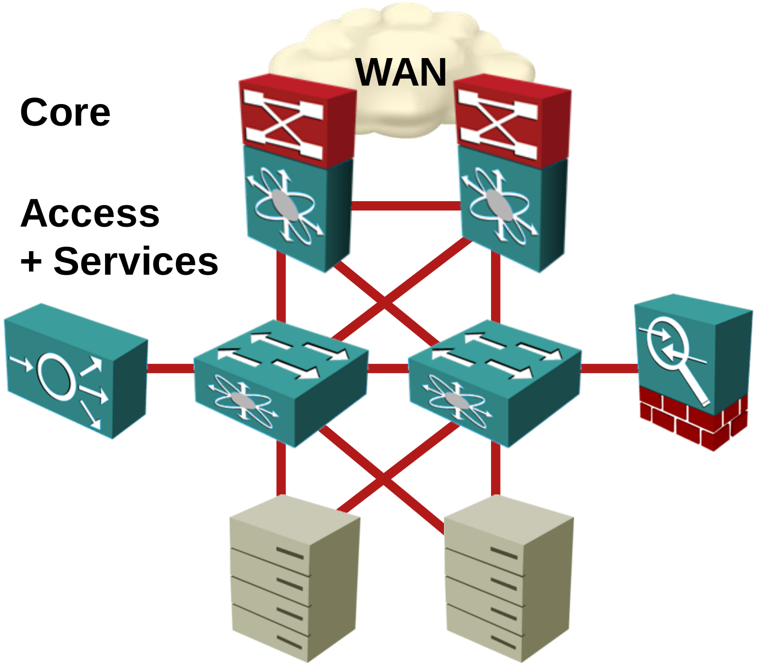
- Efficient management
- Simplified provisioning
- STP-less bridging
- Tight integration with server virtualization
- Seamless insertion of security services

Compare the architectures before comparing boxes & features

Software features can change, broken architecture will haunt you

The Flattening Myths





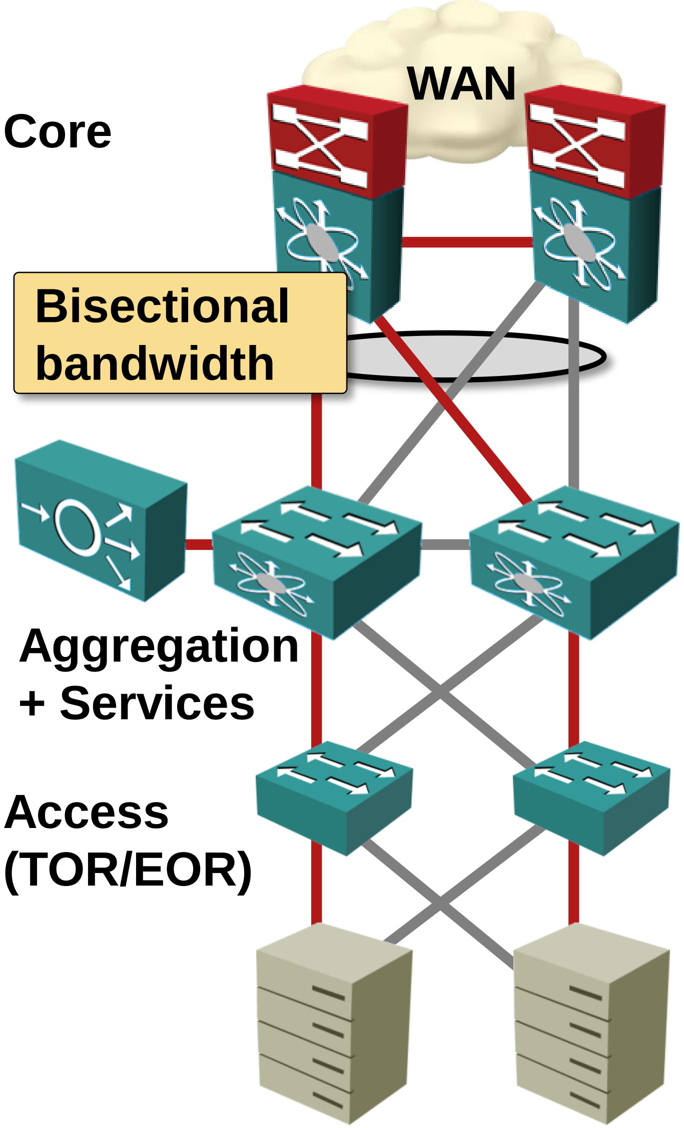
Benefits of 2-tier architecture

- Lower oversubscription
- Reduced hierarchy, fewer management points
- Enabled by high-density core switches

Crucial questions remain

- Positioning of services infrastructure (FW, LB)
- Routing or bridging (N/S and E/W)

Spanning Tree Issues

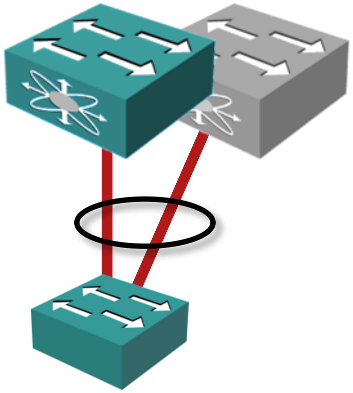


Problem: STP blocks half the links

Solutions

- **Route as much as you can**
- Multi-path bridging (TRILL/802.1aq)
- Multi-chassis link aggregation
- Server-side LACP support
- Split-horizon switching in hypervisor hosts

Multi-Chassis Link Aggregation (MLAG) Basics



Link aggregation (LAG) bundles parallel links into a virtual link

- Virtual link is not blocked by STP
- Standardized in 802.3ad/802.1ax

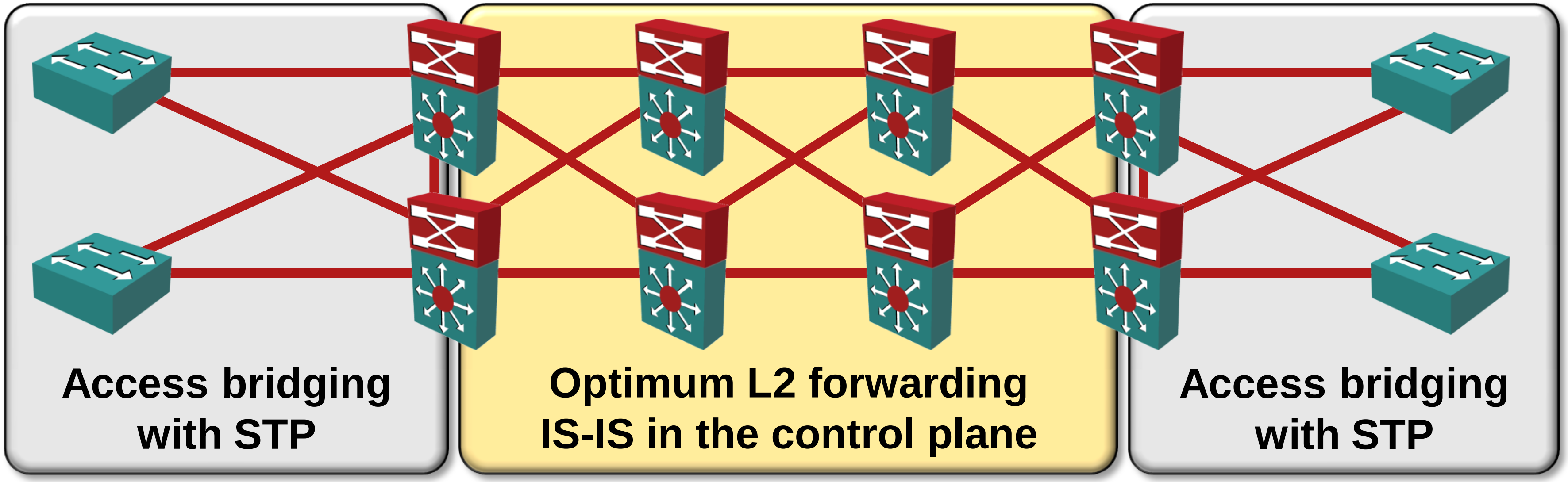
Links connected to different switches cannot be aggregated

- MLAG: two (or more) chassis are represented as a single LACP entity
- Removes STP-induced link blocking while retaining redundancy
- Works only in dual-tree hierarchies

Ask these questions:

- Are all links in the bundle active? Example: Cat6500 w/o VSS
- Can you run STP on the LAG?

Large Scale Bridging Architecture



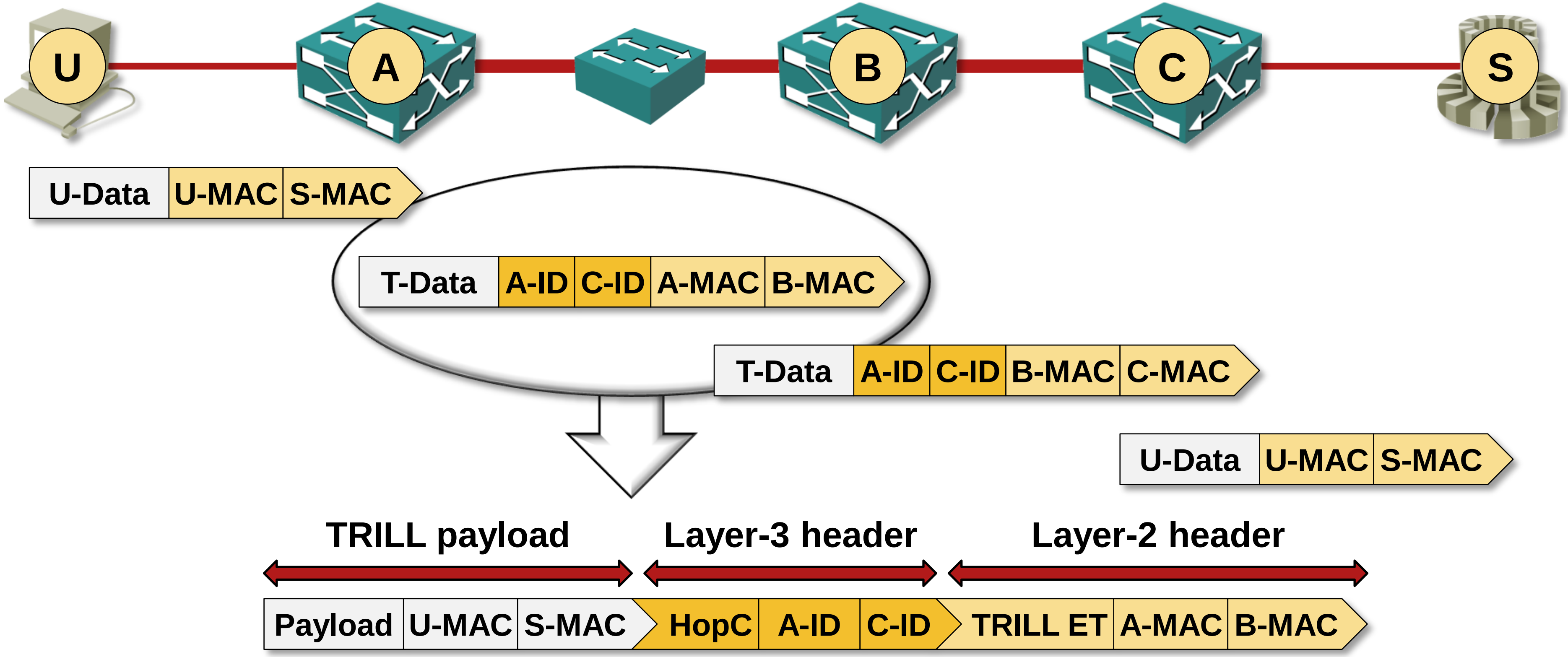
Core architecture:

- Network core implements optimum multi-path L2 forwarding
- IS-IS is run between core devices (BRouters / RBridges)
- Information gained with IS-IS SPF populates core bridging tables

Edge architecture:

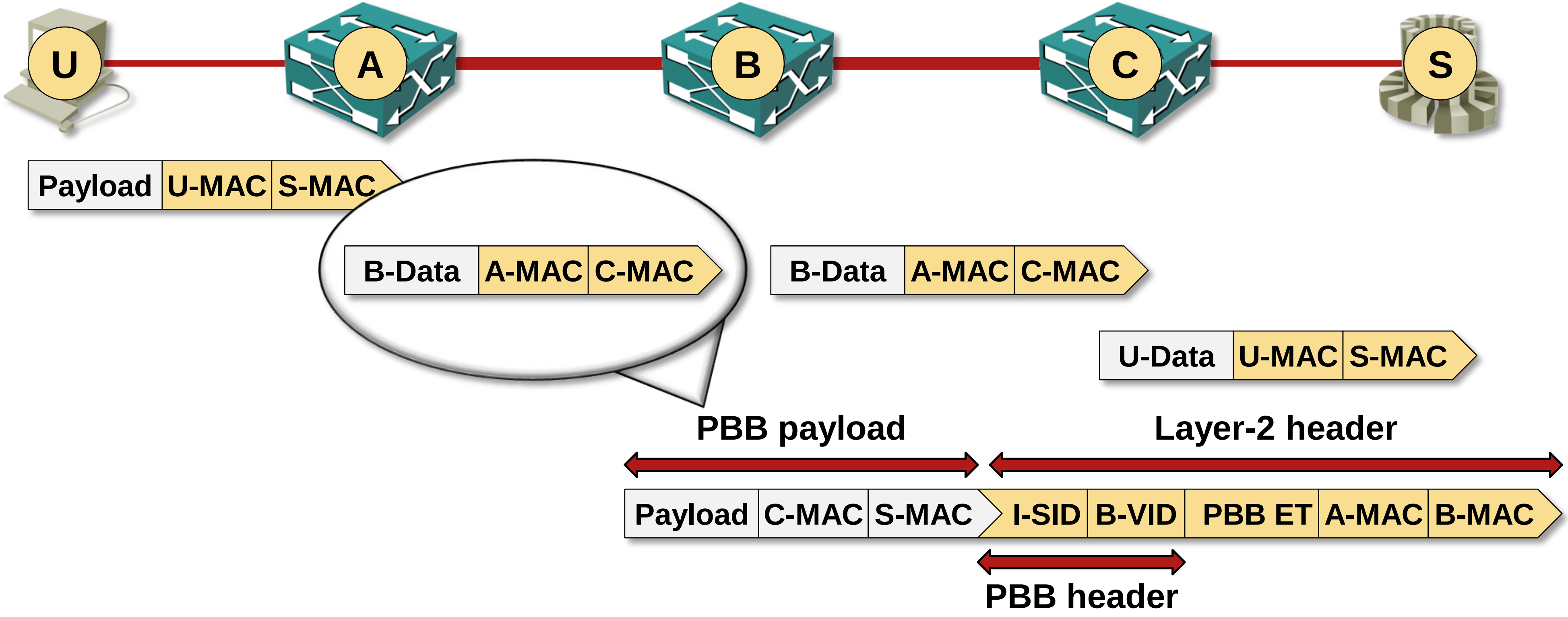
- End-to end forwarding paradigm is still bridging with dynamic MAC learning
- Edge RBridges don’t have to participate in access STP
- Dedicated forwarder is elected for each access VLAN

TRILL: Forwarding Paradigm


U
A
B
C
S
U-Data | U-MAC | S-MAC
T-Data | A-ID | C-ID | A-MAC | B-MAC
T-Data | A-ID | C-ID | B-MAC | C-MAC
U-Data | U-MAC | S-MAC
TRILL payload
Layer-3 header
Layer-2 header
Payload | U-MAC | S-MAC | HopC | A-ID | C-ID | TRILL ET | A-MAC | B-MAC

- Almost routing in the TRILL core (no router-to-host communication)
- Supports classic bridging and VLANs on inter-RBridge hops
- Requires new chipsets

802.1aq: Forwarding Paradigm



- MAC-in-MAC (802.1ah; SPBM) or Q-in-Q (802.1ad; SPBV) with a new control plane
- Not a true routing solution (bridging-over-smarter-bridging)
- 802.1aq core must be contiguous
- Reuses existing chipsets

Current L2 Multipath Implementations

Cisco – FabricPath on Nexus 7000

- TRILL-like control plane (IS-IS)
- Proprietary data plane
- Active-Active forwarding (vPC+)

Brocade – VCS Fabric on VDX switches

- Trill-compliant data plane
- Proprietary control plane (FSPF)
- No Appointed Forwarders / STP interaction

Avaya – pre-standard 802.1aq (SPBM) on ERS 8600/8800

Juniper – completely proprietary QFabric

Plane-Based Data Center Solutions Classification

Data plane

- Packet forwarding

Control/data plane

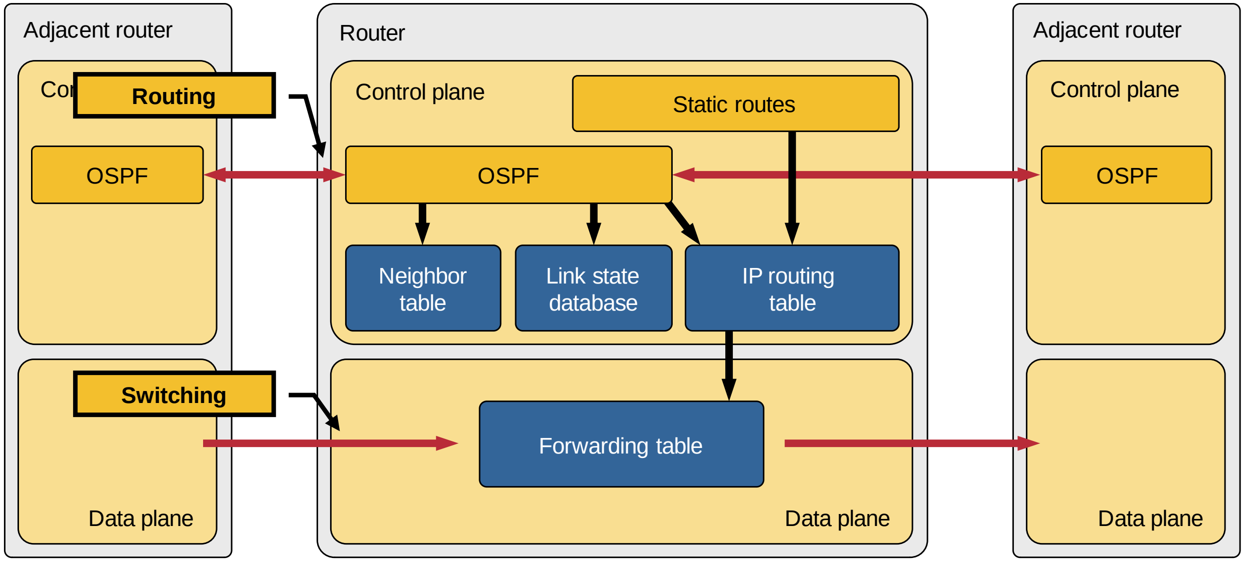
- Dynamic MAC learning
- ICMP replies

Control plane

- STP, routing protocols

Management plane

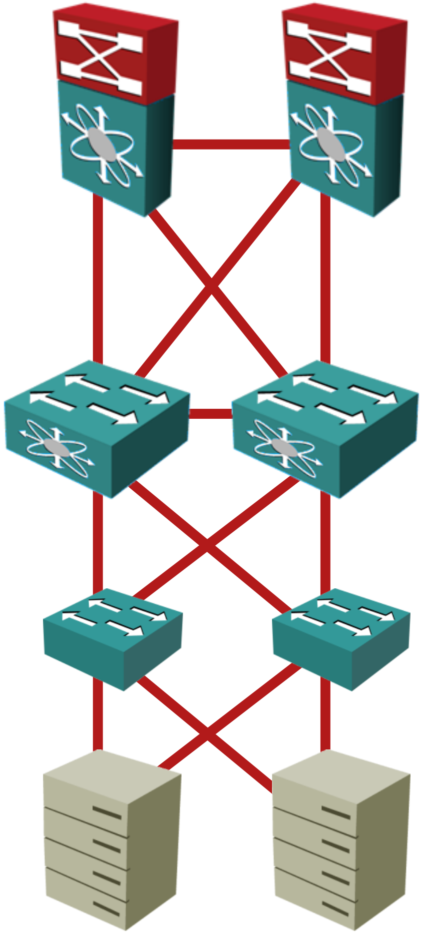
- Configuration, monitoring



Questions to ask

- What is centralized, what is distributed?
- How well does it scale?
- What are the limitations?

Independent Devices (Business-as-Usual)



Each device remains independent

- Standalone configuration
- IP addresses and L3 routing protocols
- STP device ID/priority

Examples

- Cisco Nexus 5000/7000
- Brocade VCS Fabric

Benefits

- Well-known designs, well-known challenges

Major issues

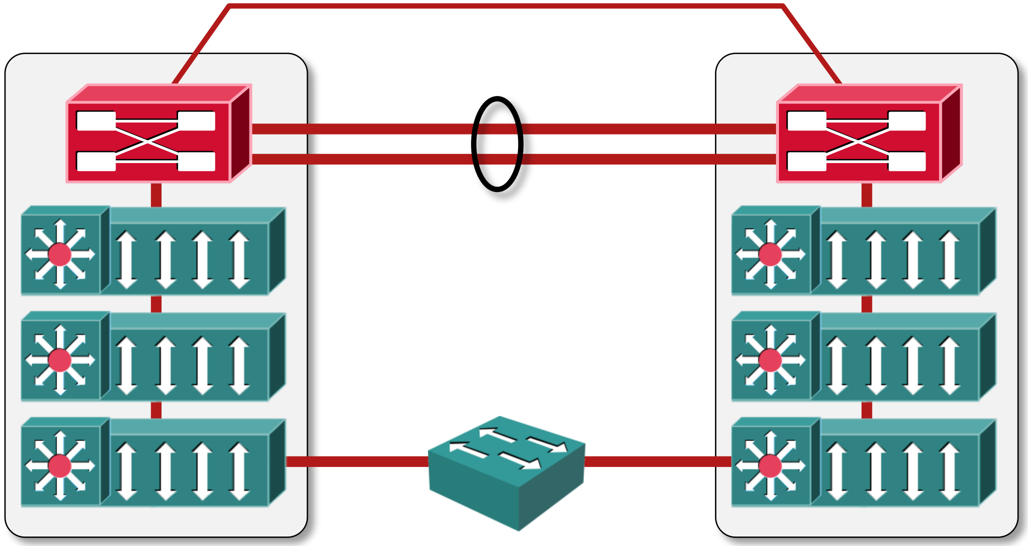
- Scales no better than what we have today
- L2 bisectional bandwidth (requires MLAG)

? L2 multipathing (requires large-scale bridging)

Example: Virtual Port Channel (vPC) on NX-OS

Supervisor engine

I/O module



- Each Nexus switch is an independent management/configuration entity
- Both supervisor engines are active
- LAG reset/split after vPC link or box failure
- LACP/STP packets are forwarded to the primary vPC member
- vPC members exchange MAC reachability information
- Off-VLAN functions (HSRP, PIM, FabricPath) work in active-active mode

One of the few solutions with full active/active LACP and full STP support

Other 2-chassis MLAG Solutions

MLAG in business-as-usual architecture offered by many vendors:

- Alcatel Lucent
- Arista Networks
- Avaya
- Cisco
- Force 10

Cisco and Avaya support L2 multipathing

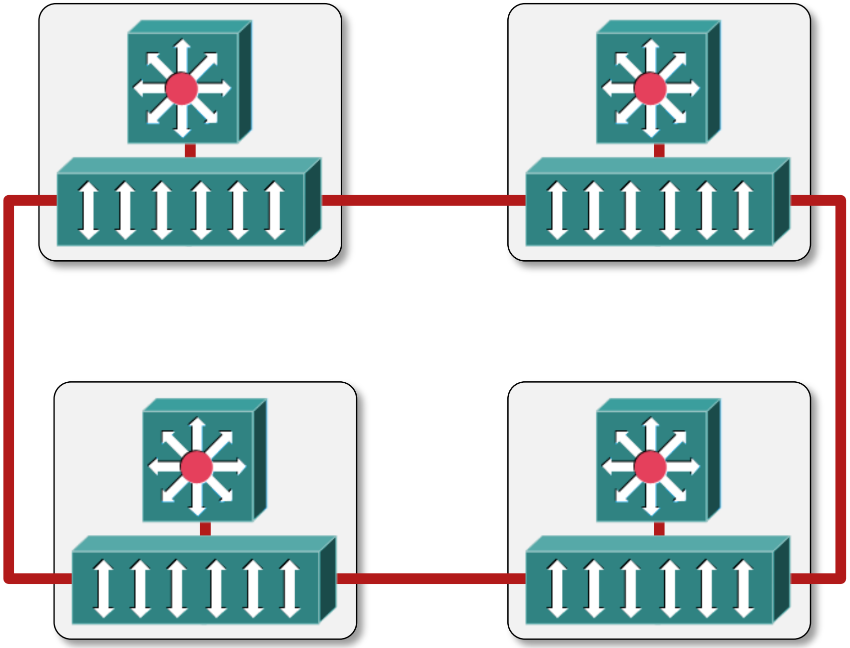
Brocade goes a step further with VLAG

Check the following features

- Active/Active or Active/Passive links in a LAG
- Standard STP/RSTP/MSTP over MLAG bundle
- Active-Active off-VLAN functions (example: VRRP gateway)

VCS Fabric (Brocade VDX Switches in Fabric Mode)

- Each device is an independent management/ configuration entity
- Automatic ISL trunk negotiation
- Optimal trunk load balancing
- TRILL-like data plane (FSPF routing)
- External LAG can be terminated on any box in the fabric (virtual LAG)
- L2 only, no STP support in fabric mode



Brocade NOS 2.1 enhancements

- Scalability: 24 switches in fabric, VLAG termination on up to 4 switches
- vCenter integration: ESX host autodiscovery & increased VM awareness
- FC support and inter-fabric routing between FC/FCoE fabrics
- Distributed configuration

Shared Management Plane (Quilt)

- Independent control/data planes
- Shared configuration and monitoring

Examples

- Cisco UCS
- Juniper Virtual Chassis (IS-IS-like internal routing)

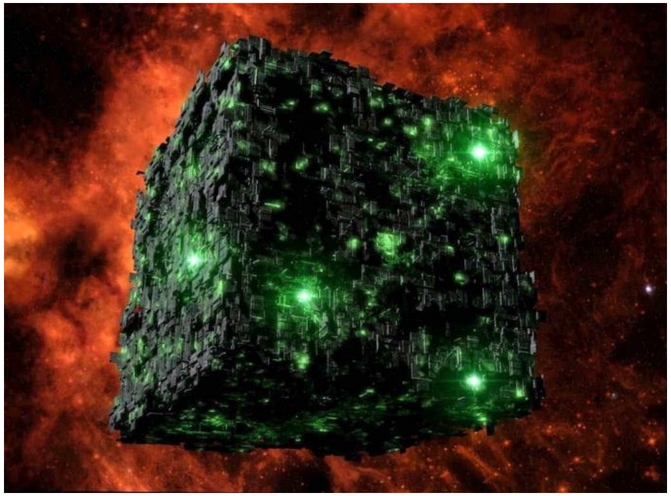
Benefits

- Single management entity
- Single-box failure does not result in fabric-wide resets

Major issues

- Most existing implementations are L2 only
  L2 is simple, L3 would be interesting

Shared Control Plane (Borg)



- Shared configuration and monitoring
- Single control plane, distributed data planes
- One set of IP addresses, one set of routing adjacencies

Examples

- Cisco’s VSS, HP’s IRF, Juniper’s XRE
- Nexus 1000V
- Most stackable switches

Benefits

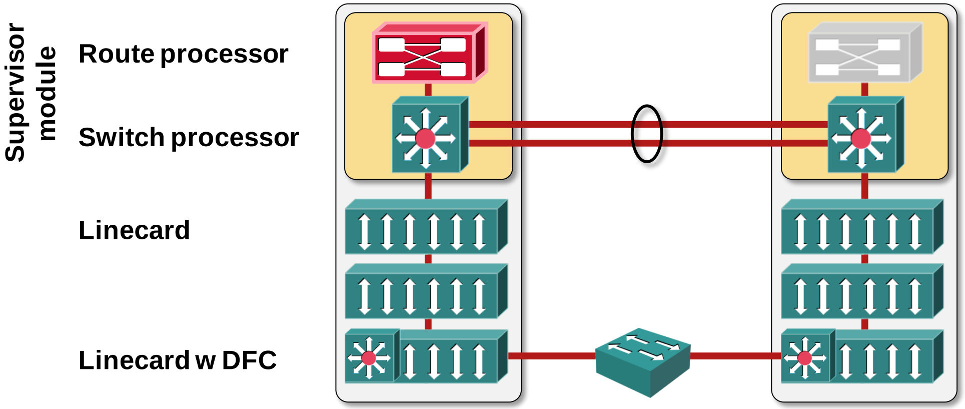
- Single management and control-plane entity

Major issues

- Loss of “master” node might result in resets
- Partitioning results in L3 split brain and/or loss of the minority part
- Does not scale as well as architectures with distributed control planes

VSS (Cisco Catalyst 6500) and IRF (HP)

- Active RP controls all switching fabrics
- Backup RP synchronized to the primary RP, takes over after failure
- All control packets sent to the primary RP (including LACP and STP)
- No need for HSRP/VRRP (use MLAG)
- Partitioning is fatal for L3 forwarding
- You lose half the system after split-brain discovery



Cisco VSS

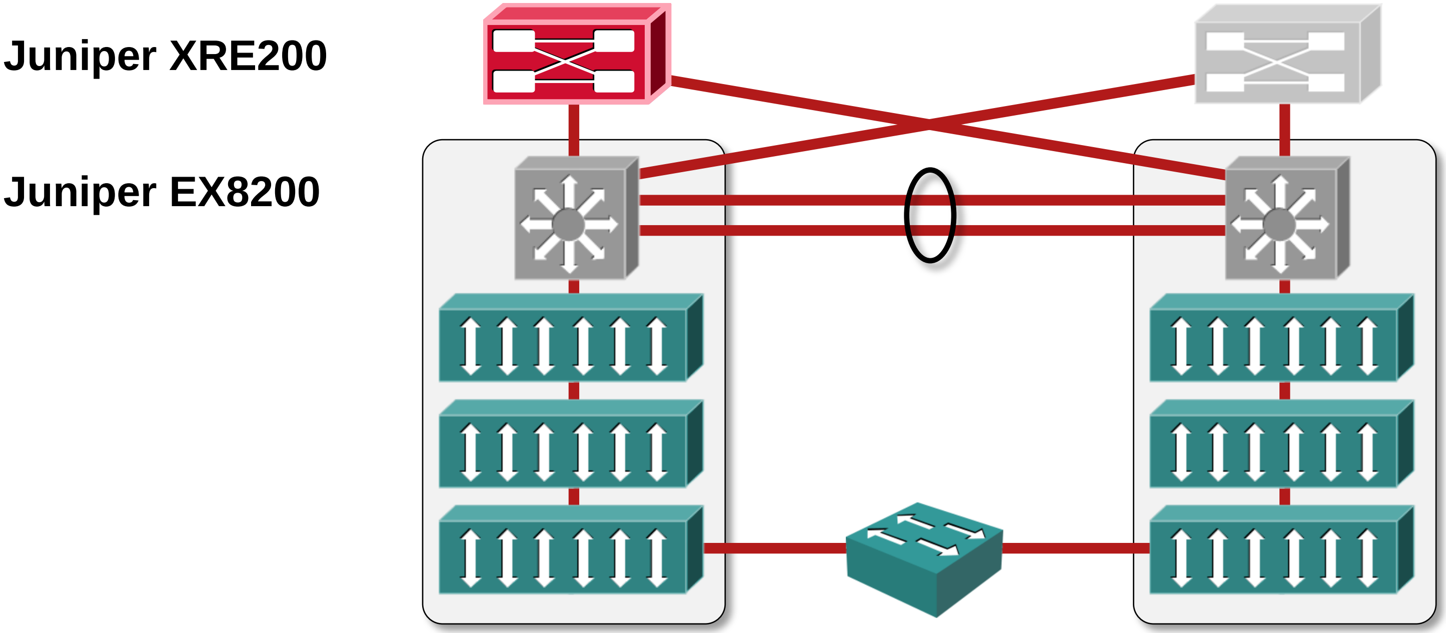
- Two Catalyst 6500 switches (one or two SUPs each)
- Split-brain detection with BFD or PAgP

HP IRF

- Two high-end switches
- Four stackable switches
- Split-brain detection with BFD, modified LACP or gratuitous ARP

Similar technologies, plenty of room for nitpicking

Virtual Chassis with External Routing Engine



- External routing engine takes over the control plane
- Supervisory modules in core switches perform maintenance functions and download data to TCAM
- All control packets are sent to primary XRE
- Backup XRE takes over after primary XRE failure

Centralized Data Plane (Tendrils)

Single control plane, centralized data plane

Examples

- Nexus 2000 port extenders
- 802.1Qbh
- WLAN controllers

Benefits

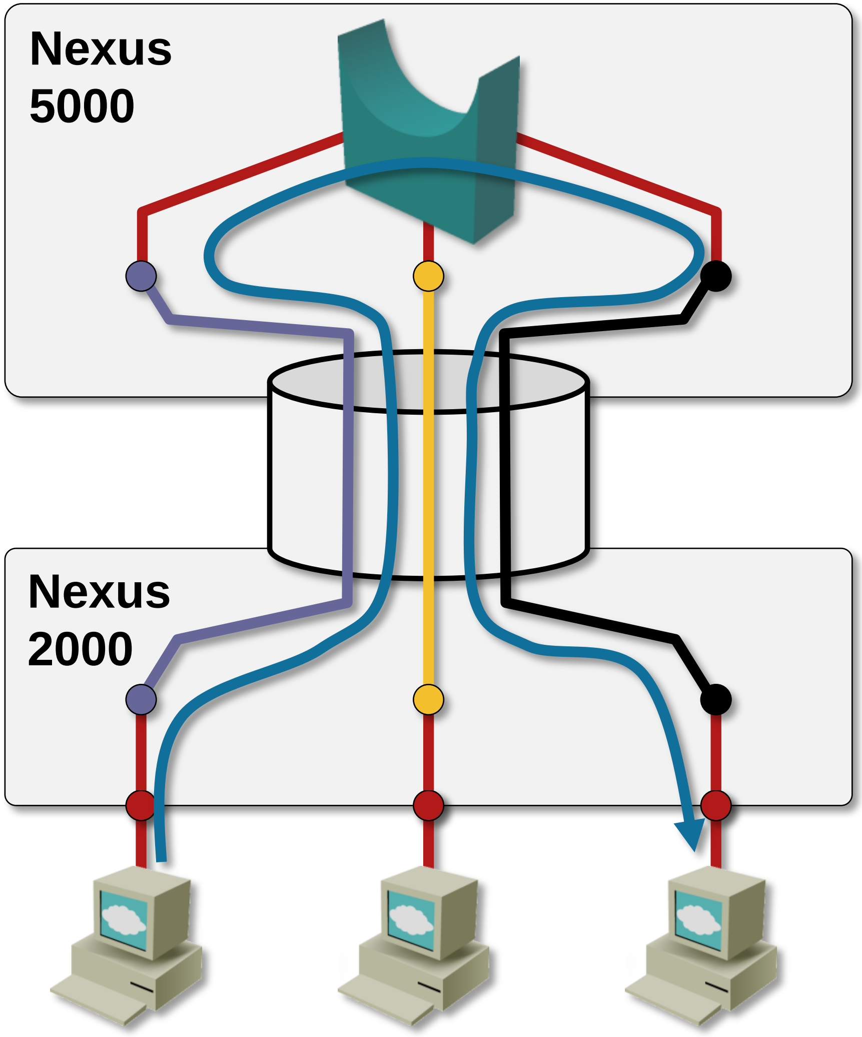
- Simple deployment and management

Major issues

- Loss of central node might results in resets or loss of the whole complex
- Suboptimal handling of east-west traffic

Port (Fabric) Extender Architecture

- Controlling bridge “owns” and configures the extenders
- Extra non-VLAN tagging (802.1Qbh) is used on the fabric links
- Port extender interfaces are configured as physical interfaces on the controlling bridge
- All traffic goes through the controlling bridge



Per-Flow Data Plane Setup (Big Brother)

Principle of operations

- Unknown packets (first packets in a flow) are sent to the controller
- Controller might forward the packets to egress device (or block the flow)
- Controller installs per-flow TCAM entries in all forwarding entities in the path

Examples

- Multi-Layer Switching (remember Catalyst 5000?)
- OpenFlow (can also support all other architectures)

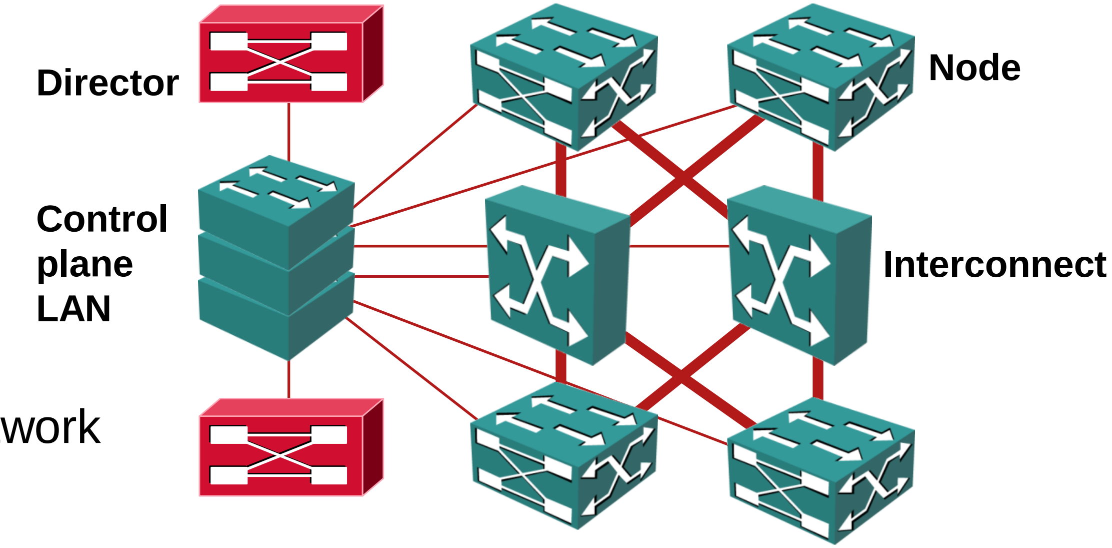
Benefits

- Can be used to implement any forwarding/routing policy

Major issues

- Per-flow forwarding architectures have never scaled
- For other issues, talk to someone who had to support MLS (even better: MLSP)

QFabric: Hardware Architecture



Director

- compute resources, runs multiple *Routing Engines*
- Redundant scalable architecture
- Add more directors when needed

Interconnect

- High-speed 3-stage 10Tbps Clos network
- Up to four interconnects per QFabric

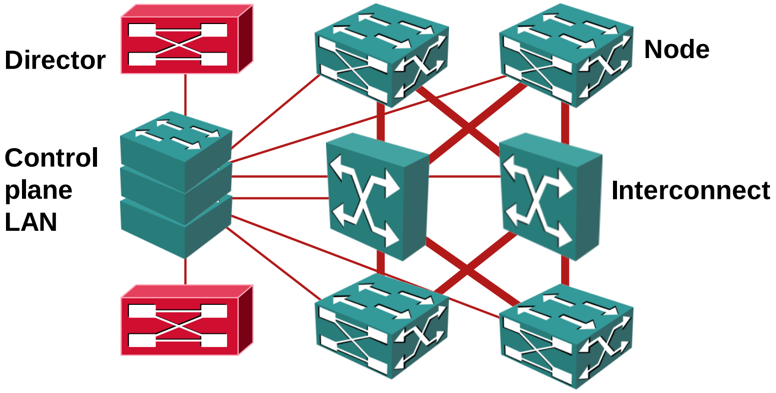
Node

- Layer2/3 packet forwarding (QFX3500)
- Single (ingress node) packet lookup (sounds like MPLS/VPN) – 5 μs across the QFabric
- 40 Gbps to the interconnects

Control plane LAN

- Out-of-band redundant GE LAN (EX4200 switches in a virtual chassis)

QFabric Control Plane



Central management plane

- Single CLI, one configuration file
- Distributed configuration updates
- Scatter/Gather monitoring (**show** outputs, SNMP)

Independent control-plane *node groups*

- Single node with local *Routing Engine*
- Two nodes in a *server group* (for MLAG)
- Up to eight nodes in a *network group*
  Processing offloaded to redundant *Routing Engines* running in Directors
- Only the *network group* provides routing protocols (OSPF, BGP) and STP support

Distributed data plane

- Each node performs full L2/L3 lookup
- Forwarding tables distributed by *Fabric Control Routing Engines*

QFabric is equivalent to a Quilt of Borgs

Conclusions

Age-old wisdom

- Don’t rush
- Evaluate what you actually need (listen to the business people, not server admins)
- Buy according to your business needs (not the nerdiness factor)
- Evolution is usually better than revolution
- Bleeding edge usually hurts

Specific to Data Center fabrics

- Large-scale bridging might be dead (even Gartner agrees with me)
- FCoE is a must-have if you have FC storage (but I would use iSCSI)
- DCB (lossless Ethernet) is a must (iSCSI will thank you)
- Revisit old designs (Clos networks)

More information

Blogs & Podcasts

- Packet Pushers Podcast & blog (packetpushers.net – Greg Ferro, Ethan Banks & co)
- BradHedlund.com (Brad Hedlund, Cisco)
- NetworkJanitor.net (Kurt Bales)
- LoneSysAdmin.net (Bob Plankers)
- The Data Center Overlords (Tony Bourke)
- StorageMojo.com (Robin Harris)
- blog.fosketts.net (Stephen Foskett, Pack Rat)
- Brass Tacks (Eric Smith)
- The Networking Nerd (Tom Hollingsworth)
- ioshints.info (yours truly)

Webinars (@ www.ioshints.info/webinars)

- Data Center Fabric Architectures (upcoming)
- Data Center 3.0 for Networking Engineers

Questions?