

Network infrastructure design and implementation

(principles of data networks and storage networks)

Colin Butcher

Part 1:

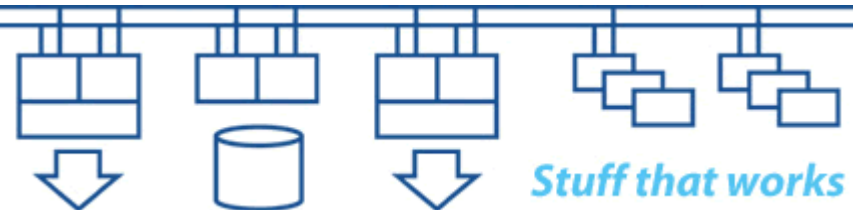
- **Basic principles of data networks and storage networks:**
 - LAN
 - WAN
 - SAN
 - Protocols

Part 2:

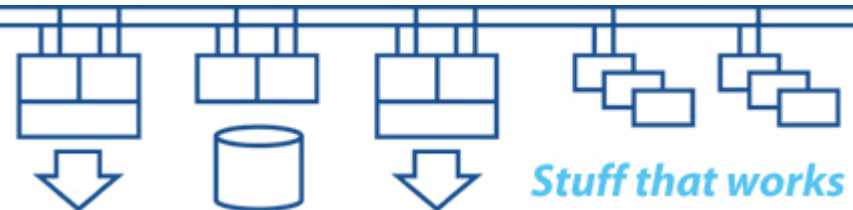
- **Building an infrastructure - putting it all together**
- **Examples**

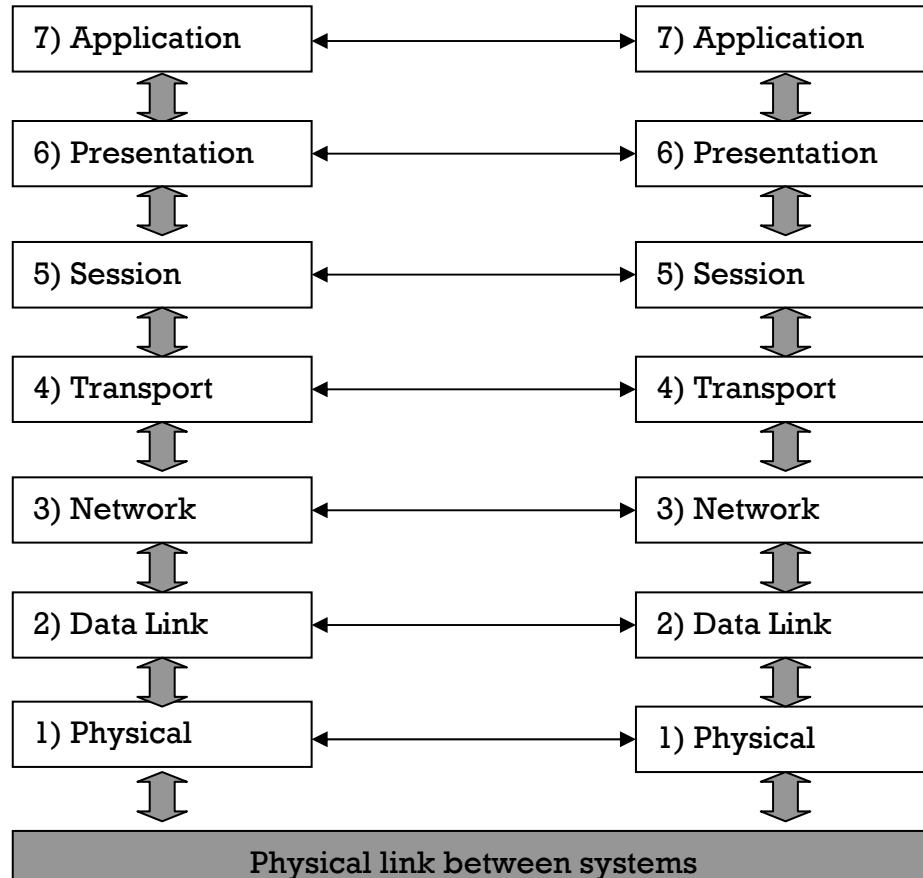
Part 1 (a):

- Data networks - LAN

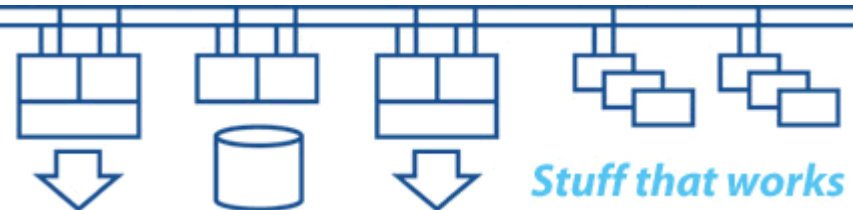


7	Application	Provides for distributed processing and access, contains application programs and supporting protocols (eg FTAM)
6	Presentation	Coordinates conversion of data and data formats to meet the needs of the individual applications
5	Session	Organises and structures the interactions between pairs of communicating applications
4	Transport	Provides reliable transparent transfer of data between end systems with error recover and flow control
3	Network	Permits communication between network entities
2	Data link	Specifies the technique for moving data along network links between defined points on the network, and how to detect and correct errors in the Physical layer (layer 1)
1	Physical	Connects systems to the physical communications media





- Transmission properties are important - a square wave at in one end needs to be recognisable as a square wave coming out at the other end
- Copper:
 - Co-axial (thick-wire, thin-wire)
 - Twisted pair (Category 5, 5E, Category 6 etc.)
- Fibre-optic:
 - Monomode (typically 9 micron)
 - Multimode (typically 50 or 62.5 micron)



- 10 Mbit/sec
- 100 Mbit/sec (Fast ethernet)
- 1,000 Mbit/sec (Gigabit ethernet)
- 10,000 Mbit/sec (10Gigabit ethernet)

- Copper / fibre (different transmission characteristics)

- Wireless ethernet (access points connect wireless and wired LANs, shared bandwidth, security issues)

- FDDI, Token ring

- Provide connection between IO subsystem and network
- Copper / fibre / wireless physical interfaces
- On-NIC processing:
 - Packet creation
 - Address filtering
 - Encryption
 - Protocol processing (TCP/IP offload - TOE)
 - Sometimes worth disabling it!

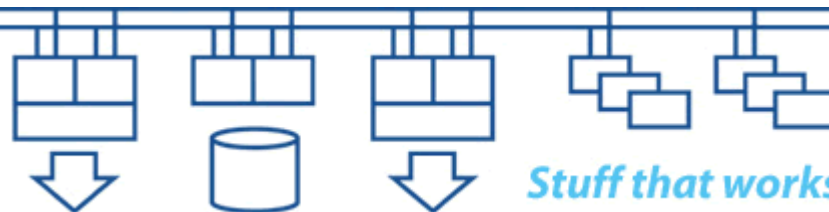
- Inherently broadcast medium, hence CSMA/CD
- Hardware & Physical MAC addresses
- Broadcast address
- Multicast addresses
- Point to point addresses
- Ethernet packet format v IEEE802.3 packet format
- Packet size (normal frames and jumbo frames)
- Jumbo frames - frame size varies with NIC type and must be supported by every device all the way through the infrastructure

Why segment a network?

- Performance
- Security
- Availability

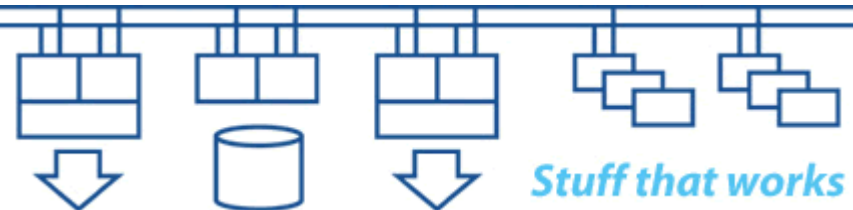
How can you segment a network?

- Multiple NICs
- Repeaters
- Bridges
- Switches
- VLANs
- Routers



- Layer 1 devices (“flat” network)
- Provide electrical fault isolation
- Simply re-time and re-transmit signal
- No control of bandwidth
- Beware of cumulative end to end delay exceeding maximum permissible frame timing – which leads to ‘folklore’ such as the “three repeater rule”

- *TIP: Beware of the generic term “hub”*



- Packet content based (Layer 2)
- Store and Forward
- Easy to use and configure
- Poor control of bandwidth (filtering)
- Spanning tree algorithm
- Provides an extended LAN
- Not all protocols can tolerate the inherent delays in working over an extended LAN
- Remote booting (MOP, BOOTP etc.) will consume bandwidth

- Introduce parallelism
- Speed of chipsets (latency & bandwidth)
- Full duplex operation on a single device per port basis
- Traffic monitoring (mirror ports)
- Link aggregation
- Bandwidth control
- “Store and forward” versus “Cut through” switching
- Layer 2, Layer 3 etc. switching
 - Layer 3 generally refers to TCP/IP routing layer
 - Layer 4 generally refers to TCP/IP port numbers, eg: HTTP port 80 traffic)

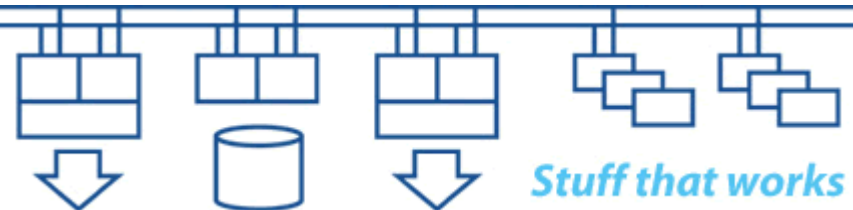
VLANs are another way to segment a network for performance and security

- Implemented within core switches
- Also implemented in some NICs / device drivers
- Port based VLANs
- Protocol based VLANs
- Connectivity between VLANs
- VLAN tagging of packets (802.1Q)
- VLAN tagging of packets out of NICs
- QoS (Quality of Service) and bandwidth reservation

- Shared bandwidth (“flat” network)
- Security issues (access control, authentication, data encryption)
- Roaming issues (multiple Access Points and MAC address migration between ports)
- Complex to configure and manage (especially in a large environment)
- Antennas (coverage and beam patterns)
- Wireless repeaters and bridges

Part 1 (b):

- Wide area data networks (WAN)

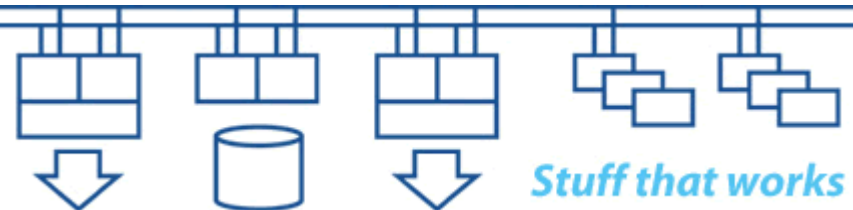


- ISDN, POTS, Leased Line
- Frame Relay
- ATM
- MPLS
- “Dark fibre” and DWDM / CWDM
- SONET / SDH etc.
- ADSL / SDSL
- VPNs
- Managed services (usually TCP/IP based)
- Encapsulation and tunnelling
- FC over IP, FC over Ethernet

- Routers do not need to be involved in the normal inter-node traffic within a LAN, other than keeping track of who's where and making themselves known
- Routers build knowledge of address (node or interface) reachability on a per-protocol basis
- Protocol address based (Layer 3)
- Need to design addressing scheme
- Bandwidth control
- Design routing paths
- Routing table updates are propagated between routers

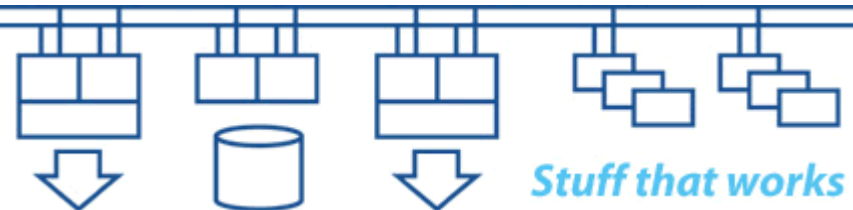
Routers are generally used to interconnect LANs over a WAN

- Separate devices or can be integrated into the core
- Need to design protocol addressing scheme and areas
- Good control over bandwidth
- Layer 3 devices – protocol address based
- IPV6 is common in big core routers
- Rare to find DECnet routing in modern routers – it's a TCP/IP dominated world in the WAN

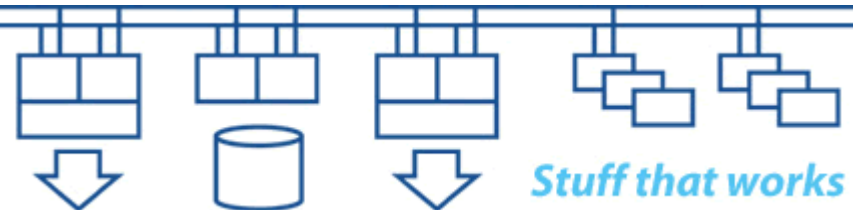


Part 1 (c):

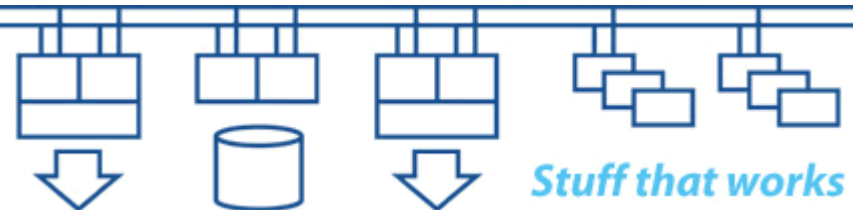
- Storage networks (Fibrechannel)



- Fibrechannel technologies (1 / 2 / 4 / 8 ... Gbps)
- A switch based network optimised for shifting large quantities of data with high throughput and low latency
- SAN fabrics connect storage devices (disk arrays and tape libraries) and Host Bus Adapters (HBAs) in systems
- All devices are uniquely identified with WWNs and WWIDs (device IDs, port IDs)
- SAN Segmentation (switching, routing, zoning etc.)
- Storage subsystems and device presentation
- SAN extension (eg: FC over IP)



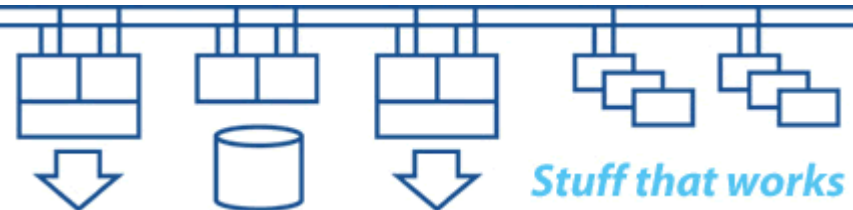
- EVA family and XP family from HP
- Array controllers interface to the SAN
- Mirrored cache in the paired array controller
- Paired array controller drives the disks
- Array controllers presents devices to the fabric once Vdisks are created in the array
- Array controllers map storage to physical disks
- Array controllers managed by SAN appliance
- Systems can interact with the controller by means of scripting through the SAN appliance (eg: snapshots, snapclones, mirrorclones)



- WWIDs are unique
- Systems and storage controllers scan the fabric to build a list of paths between devices
- Storage devices (eg: EVA Vdisks) are presented to the fabric by the array controller
- Device presentation can be controlled to limit access to specific paths (by WWID)
- Devices are presented to the fabric by the storage controller with a LUN (logical unit number) and (required by some operating systems) a device identifier (OS unit ID)

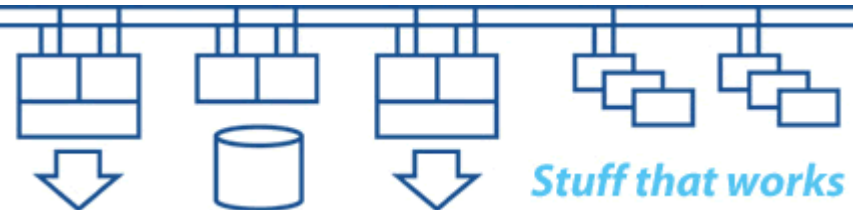
- Zones can be port based, or WWID based (“soft zoning”) – think about maintenance issues
- Zones can overlap (think Venn diagrams)
- Current zoning best practice uses the “single initiator, multiple targets” model
- VSANs can be used to segment a switch fabric
- Systems (HBAs) need to have BIOS type support for booting from SAN devices
- High availability typically uses a dual-fabric SAN
- See the HP SAN design reference guide

- Inter-site links can be “trunked” (as with data networks) to provide sufficient bandwidth
- Link “glitches” will cause fabric resets and rescans, so use FC routing in large extended SANs to minimise disruption
- Zones can extend across multiple switches (as with VLANs)
- Wave division multiplexing (DWDM, CWDM) can be used for extended distance inter-switch links
- “FC over IP” can be used to link SANs over an IP data network (beware latency issues – use QoS techniques)



Part 1 (d):

- Data network protocols



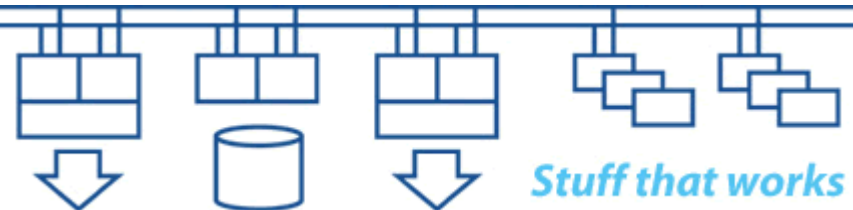
Typical network protocols:

- TCP/IP (and all its components – DHCP, BOOTP, TFTP, FTP, Telnet, HTTP, SSH, NFS etc.)
- DECnet-Plus (DECnet over IP) or DECnet Phase IV
- SCS (OpenVMS clustering)
- LAT (DECserver terminal access etc.)

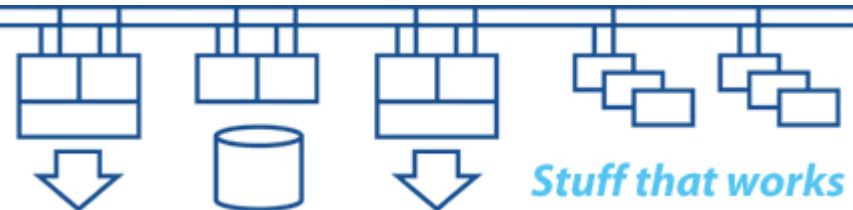
- *TIP: “Wireshark” - capture and examine packets*

- Predominantly TCP/IP
- NetBIOS / NetBEUI and NetBIOS over TCP
- WINS naming service
- Windows file & printer sharing
- Pathworks provides DECnet for Windows and Windows file & printer sharing, superseded by CIFS (Common Internet File System - based on Samba)
- “NIC teaming” for availability (eg: ProLiant servers)

- Predominantly TCP/IP
- “NIC teaming” for availability
- DECnet for Linux project
- LAT / MOP for DECservers
- SAMBA for Windows file & printer sharing



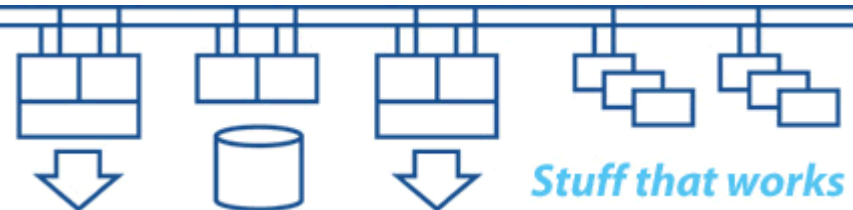
- SCS for clustering (IP clusters coming soon)
- LAD/LAST for disk serving (Infoserver)
- Pathworks (Advanced Server) - DECnet for Microsoft Windows and LANmanager functionality for OpenVMS systems. Replaced by CIFS (based on SAMBA)
- LAT / MOP for DECservers
- “LAN failover” for improved LAN availability (all protocols). Equivalent to NIC teaming
- “failsafe IP” for improved TCP/IP availability within a cluster



- Layer 4 – port or socket layer (eg: HTTP = port 80, “well known” ports allocated by convention)
- Layer 3 – IP addressing and routing layer (eg: 192.168.0.n/24, DNS/BIND resolver user to convert IP hostnames to interface IP addresses)
- Layer 2 – MAC address layer (ARP used to convert IP interface addresses to MAC addresses, cached locally)
- Layer 1 – Physical layer (transmission media)

- DNS and the BIND resolver
- DHCP address provision
- BOOTP services
- TELNET access
- FTP / TFTP file transfer
- NFS file serving
- Monitoring with SNMP
- SMTP / POP / IMAP
- Secure extensions: SSH, SSL, IPSEC
- Printing (LPR / LPD)

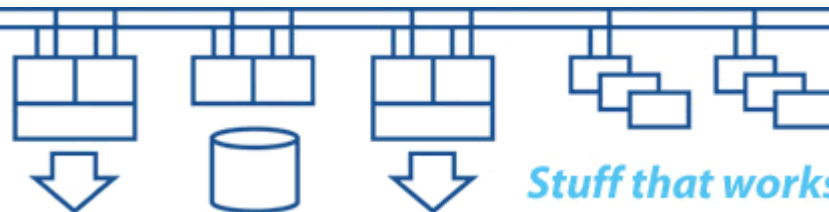
- **MAC Address formed from Node address:**
 - Area 1 - 63, Node: 1 - 1023
 - 16 bit address = (Area x 1024) + Node number
 - SCSSYSTEMID = same 16 bit value
 - AA-00-04-00-nn-mm
 - nn-mm = byte reversed hexadecimal 16 bit address
- Sets MAC address on all LAN adapters (that DECnet starts on) based on DECnet node address, so cannot connect multiple LAN adapters to the same LAN (or extended LAN). Note: Phase V lets you control which NICs change their MAC address.



- Phase IV compatible addressing on a per NIC basis – can thus select which NICs change their MAC address
- Multiple path behaviour (multi-homed End System)
- “DECnet over IP”
 - Preserves DECnet APIs for existing applications
 - Performance and availability are determined by underlying IP network infrastructure

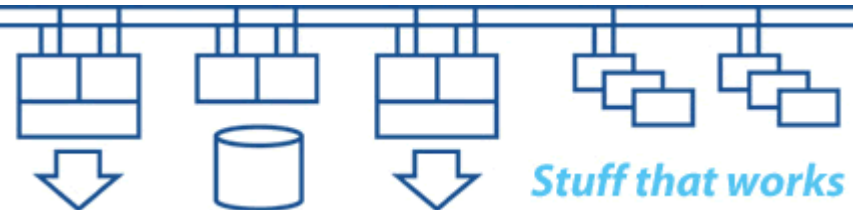
Part 2 (a):

- Network infrastructures - putting it all together



- **Signal path quality and reliability**
 - Retransmits severely affect overall throughput
- **Bandwidth – determines throughput**
 - Large packets shift more data with less overhead
- **Latency – determines round trip delay**
 - Determines how much data is in transit at any given instant
 - Data in transit is at risk if there is a failure
- **Jitter (“div latency” or variation of latency with time) – determines predictability of round trip delay**
 - Understanding jitter is important for establishing timeout values
 - Severe latency fluctuations can cause system failures

- Traffic flow, end-to-end packet delivery, delivery failure notification and performance are key parts of the design of any network protocol, as are the addressing scheme and the naming scheme
- Multicast packets are inherently “fire and forget”
- Multiple paths through a network bring additional complications – packets may no longer arrive in the order in which they were sent
- What happens when paths fail or are intermittent?
- How do we cope with bad latency or jitter?
- Time synchronisation across the infrastructure

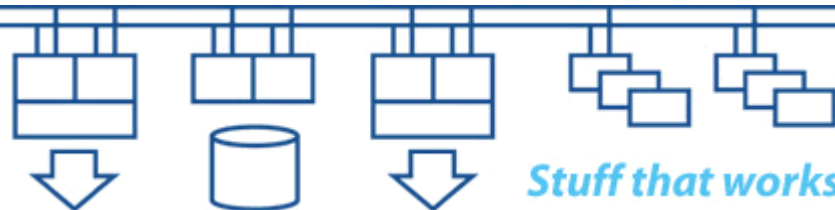


- Node naming, addressing schemes and routing mechanisms
- Multiple NICs and multiple LANs
- Map functions to NICs:
 - Management (ILO, SAN appliance, etc.)
 - Clustering
 - Network backups
 - Data transfers (eg: FTP, NFS etc.)
 - Interactive users

Relative time is more useful than absolute time

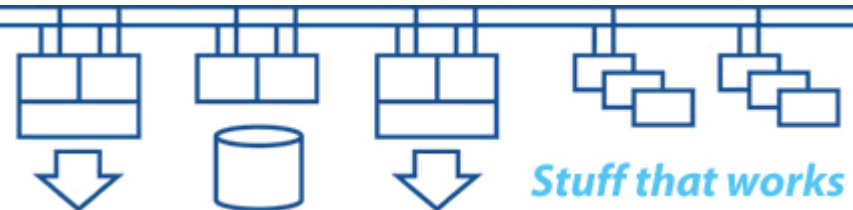
- Need to be able to order events across the network based on timestamps
- UTC Timestamp format
 - Time value
 - Inaccuracy component
- External reference clocks
- NTP
- DTSS

- Scale network so that overall performance is based on minimum essential number of paths and maximum estimated traffic
- May wish to take advantage of installed bandwidth capacity to provide additional functionality when everything is working
- There is no such thing as a single protocol network
- Understand the behaviours of the different protocols under failure conditions
- Segment the network to provide gradual degradation rather than wholesale failure



- Losing data is a disaster
- Availability is more important than performance
- Size storage subsystem based on minimum components and maximum estimated throughput
- Segment storage subsystem to provide gradual degradation rather than wholesale failure
- Need adequate backup capacity and throughput in order to meet permissible backup windows
- Understand application behaviour and storage performance requirements (bandwidth and latency)

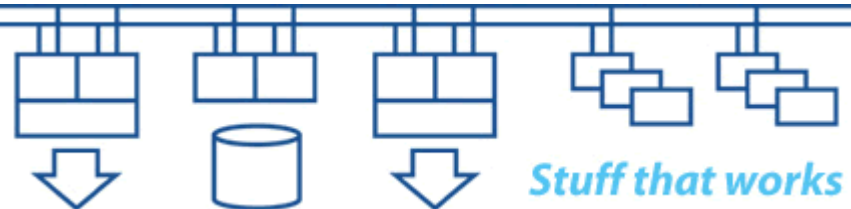
- Availability is more important than performance
- Scale network so that desired overall system performance is based on minimum essential number of paths and maximum estimated traffic
- Segment the network to provide gradual degradation rather than wholesale failure
- Have a fallback plan for getting to remote site equipment (eg: dial-up modem to console port)
- Network management is all about problem identification and rectification
- Learn the warning signs for common problems



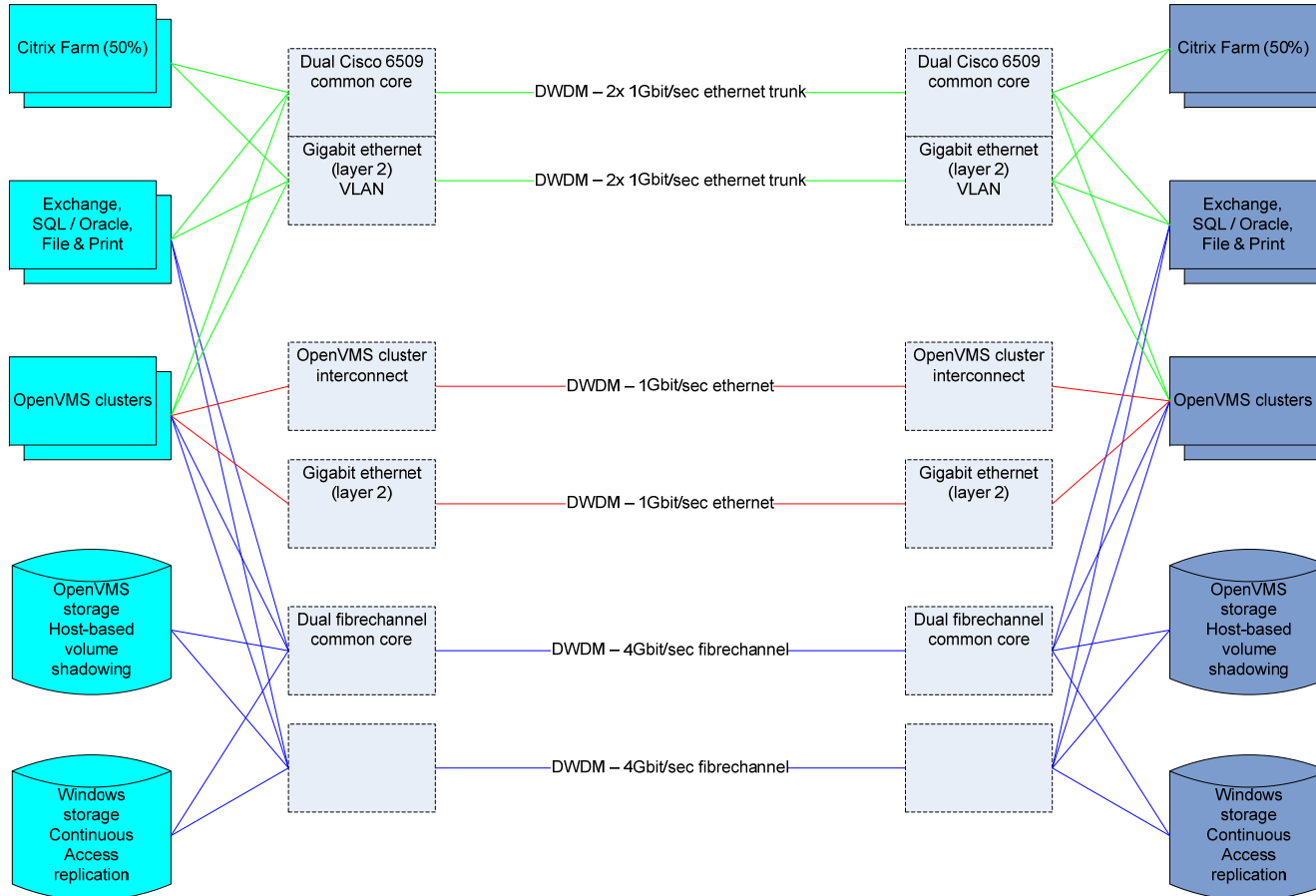
- “Converged ethernet” – fibrechannel and ethernet protocols on the same physical carrier with common interfaces and switching infrastructure
- 10GigE (and faster)
- Fibre, not copper (transmission characteristics matter)
- RNIC – offloading the bulk of the protocol handling to the NIC and minimising both CPU overhead and moving data around between the memory subsystem and the NICs

Part 2 (b):

- **Examples**



- Safety-critical and mission-critical system:
 - Migrate from Alpha to Integrity
 - Move from 3x regional clusters to single national cluster
 - Move from HSG80s to EVA4100s
 - Move to multiple NIC connectivity
- Similar principles apply in many other cases



***failsafe IP*:**

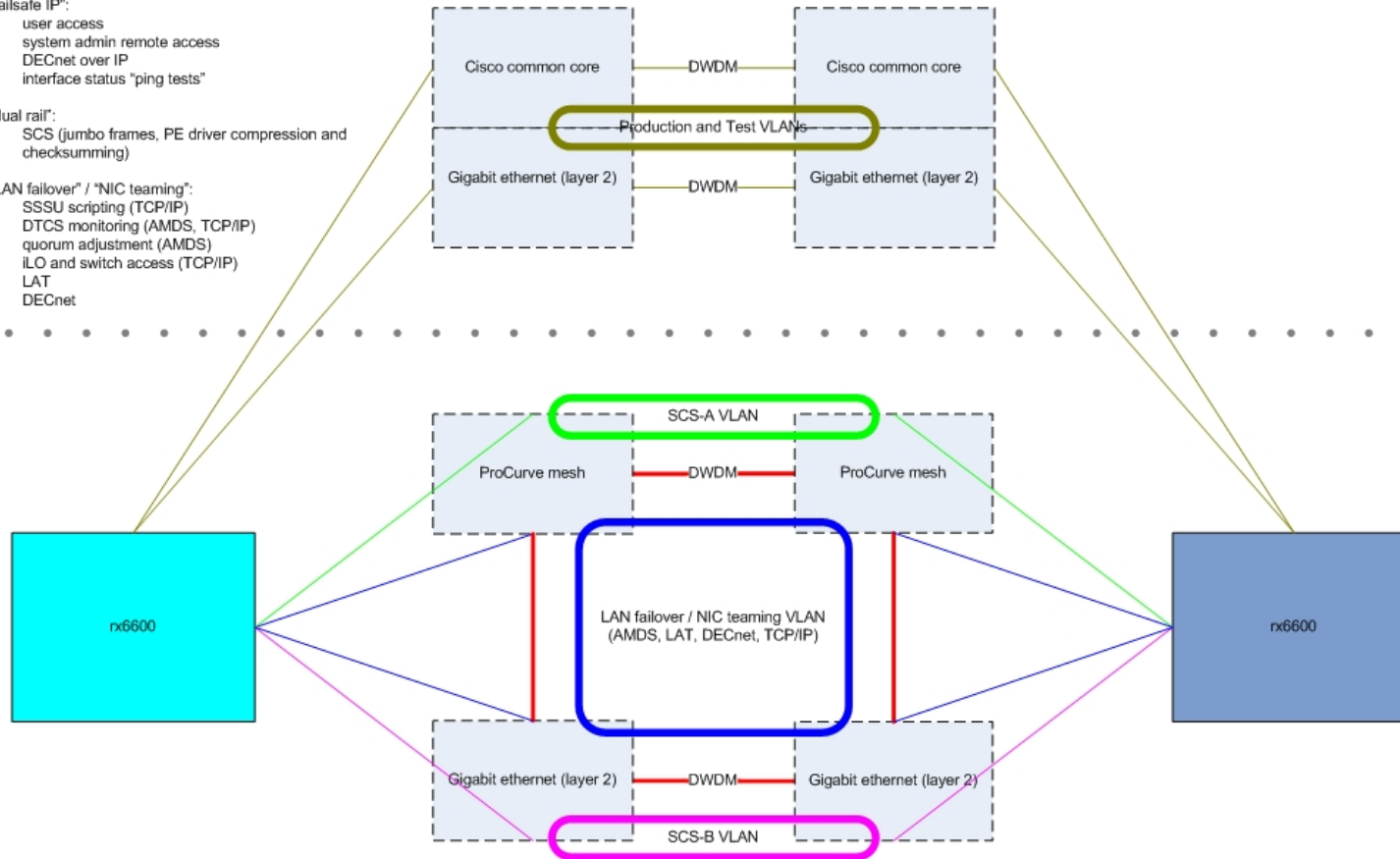
- user access
- system admin remote access
- DECnet over IP
- interface status "ping tests"

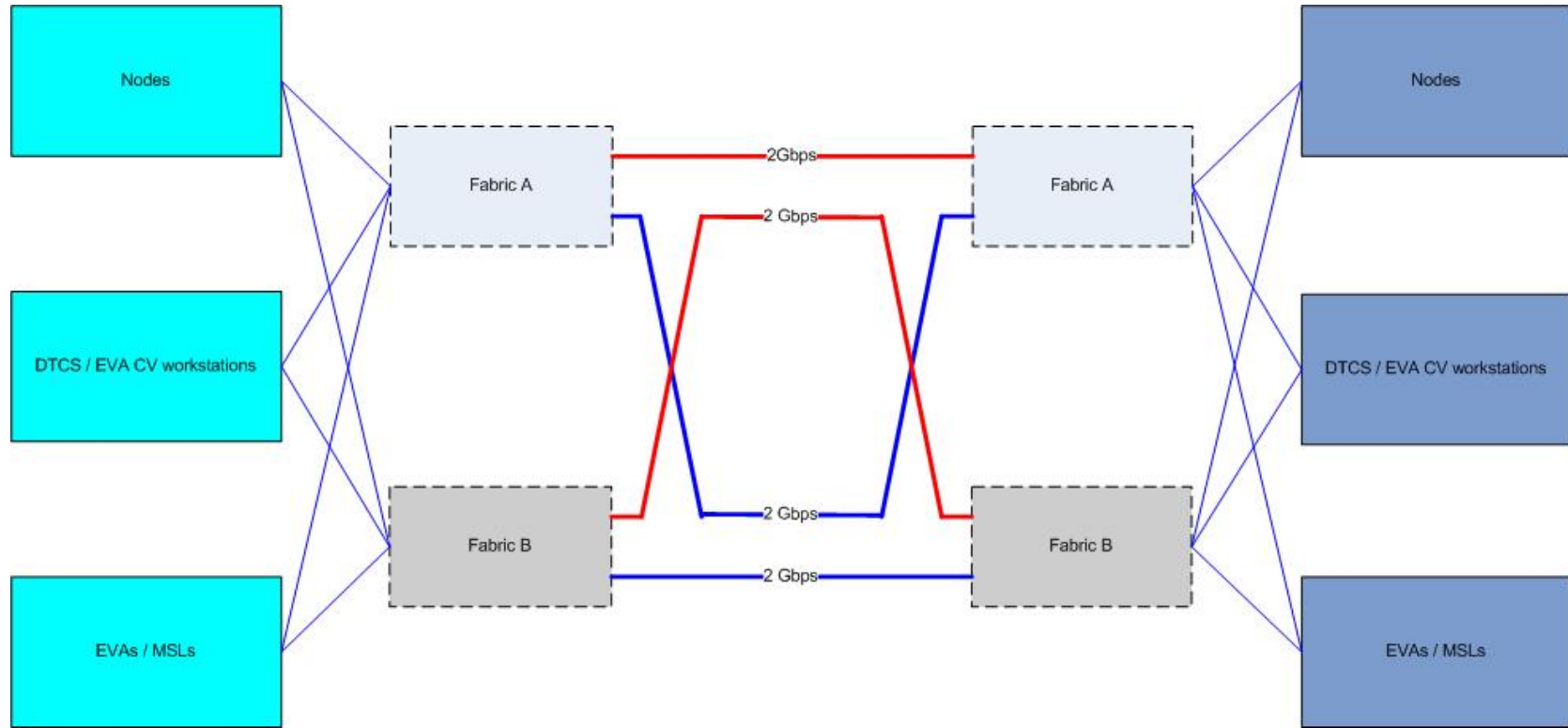
***dual rail*:**

- SCS (jumbo frames, PE driver compression and checksumming)

***LAN failover* / *NIC teaming*:**

- SSSU scripting (TCP/IP)
- DTCS monitoring (AMDS, TCP/IP)
- quorum adjustment (AMDS)
- iLO and switch access (TCP/IP)
- LAT
- DECnet





- Think big – plan for future expansion
- Understand the basic principles
- Network hardware tends to have a long life-cycle
- Avoid complexity where possible
- Design for change over time
- Minimise risk of mistakes when working on equipment
- Ensure that all cabling is tested and labelled
- Aim to minimise disruption when failures occur
- Use managed equipment that you can monitor easily
- Use the highest quality equipment that you can afford

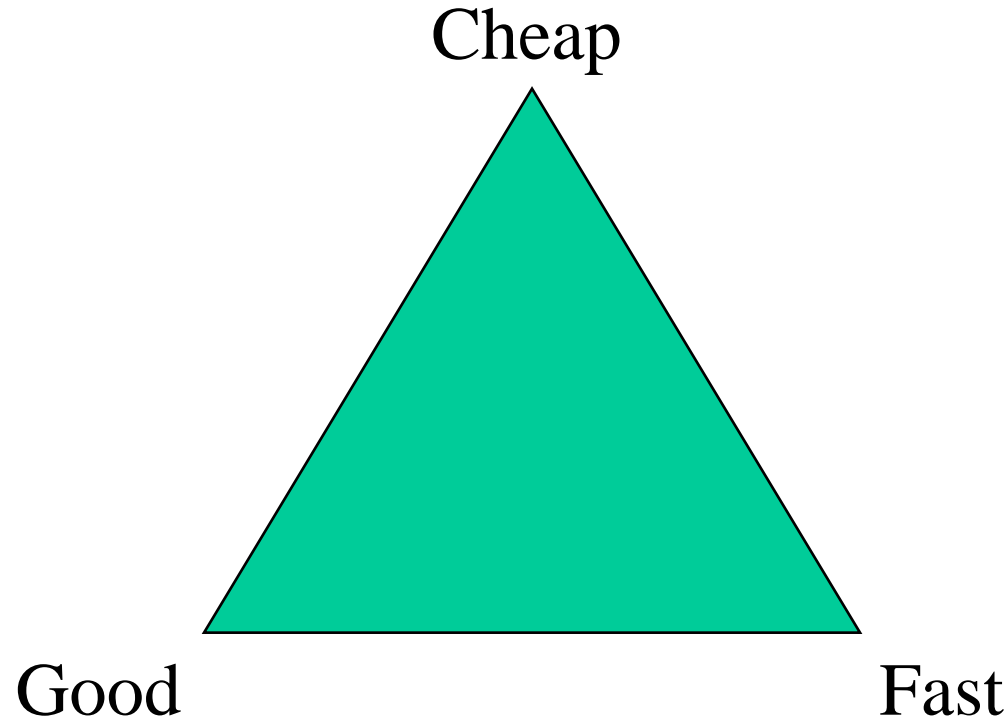
SAN information:

- HP SAN design reference guide (AA-RW86N-TE)
- HP storage subsystems web site
- Brocade web site

Data networks information:

- HP ProCurve networking web site
- Cisco web site

Standards bodies



Thank you for your participation

Q & A

