BCS Cheltenham & Gloucester – 16th April 2013

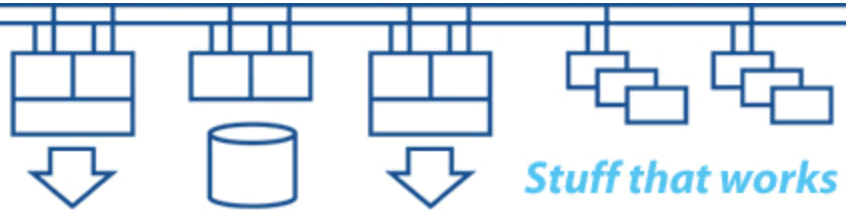
Designing and implementing network and storage infrastructures

(some of the things you always wanted to find out about data and storage networks but never got around to trying them out to see what really happens)

Colin Butcher

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Part 1:

- Basic principles of data networks and storage networks

Part 2:

- "On the wire" protocols (it's not just TCP/IP)

Part 3:

- Network infrastructures - putting it all together

Part 4:

- Examples and discussion

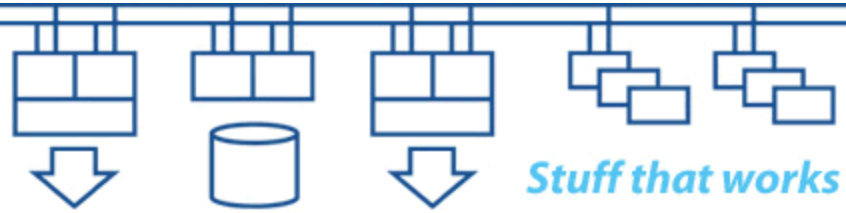
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Part 1:

- Basic principles of data networks and storage networks

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

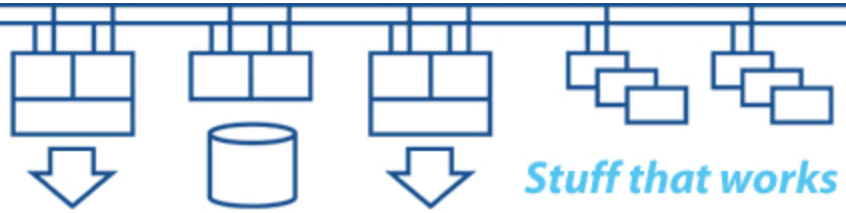


1 (a): Basic principles – data networks

- Local-Area Networks (LANs)
  - Ethernet technologies
  - Physical components and cabling
  - Protocols and addressing
  - Network Interfaces
  - Network Switches
- LAN segmentation
- LAN extension

- Wide-Area networks (WANs)

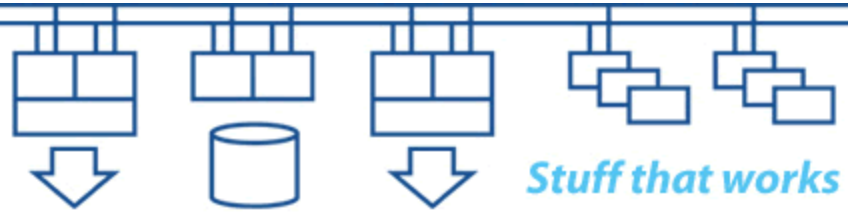
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



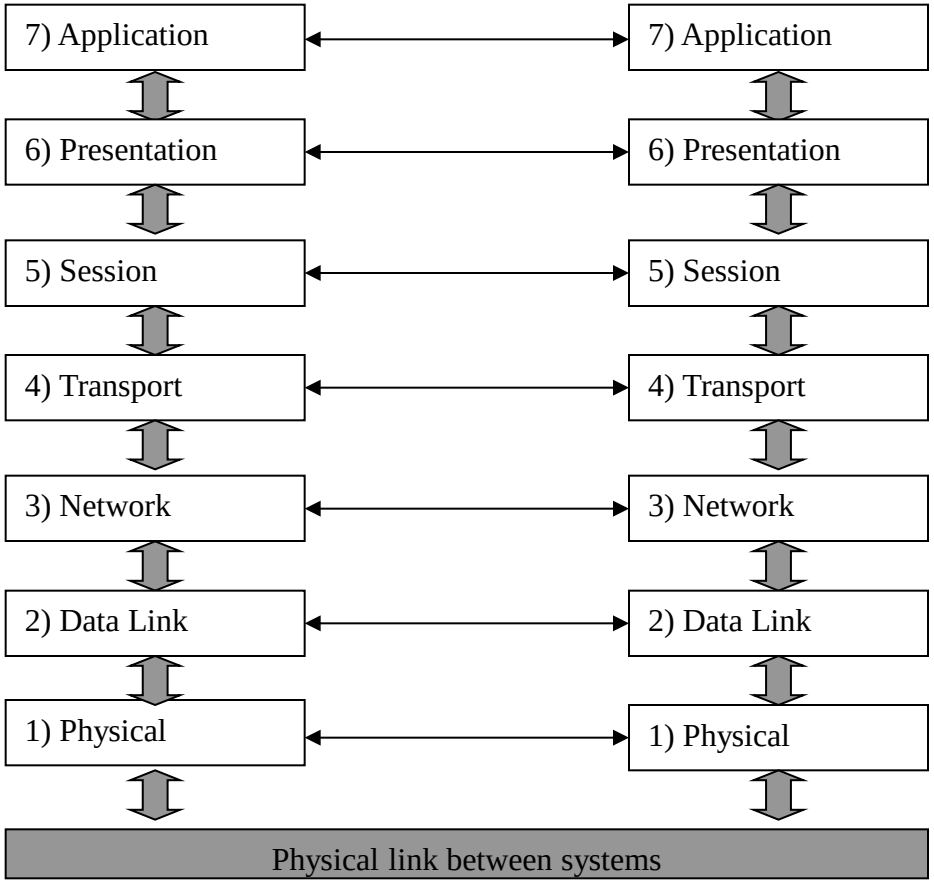
OSI Seven-layer model

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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Data flow between the layers



Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Four-layer model (predominantly TCP/IP)

- Layer 4 – Layer 3 protocol specific: TCP/IP port or socket; DECnet ‘object’ or ‘application’
- Layer 3 – Protocol specific addressing and routing layer. Needs protocol address to MAC address translation. TCP/IP ‘routing’; DECnet ‘circuit’ or ‘routing circuit’
- Layer 2 – MAC address layer, Ethernet V2 or IEE802.3 format packets. TCP/IP ‘interface’; DECnet ‘line’ or ‘csma-cd station’;
- Layer 1 – Physical layer (transmission media)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

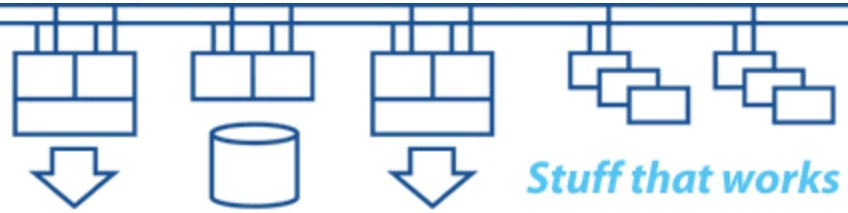
Cabling

- Transmission properties, transmitter components and receiver components are important - a square wave fed 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)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

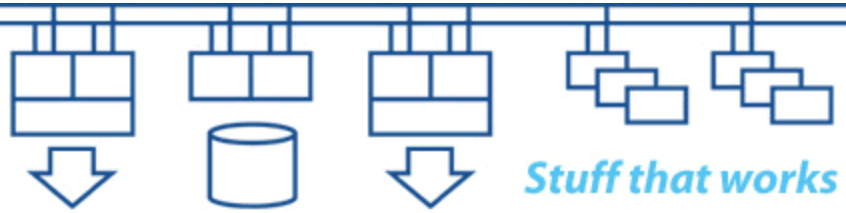


LAN - Ethernet and derivatives

- 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 (2Mbit / 11Mbit / 54Mbit / 108Mbit)
  - Note: WAP, GPRS, HSPA, UMTS, Bluetooth etc. are not wireless ethernet
  - Access control and data privacy are major issues

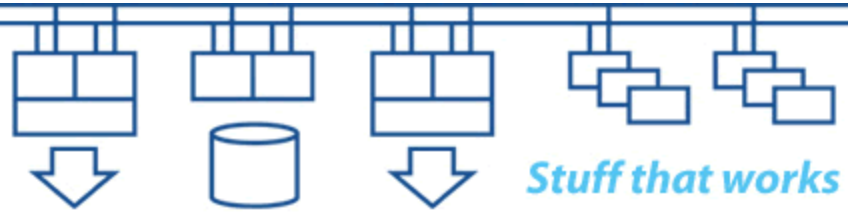
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



NICs (Network Interface Cards)

- 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)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Ethernet addressing and packets

- Hardware MAC address
- Physical MAC address
- Broadcast address
- Multicast addresses
- Point to point addresses
- Ethernet packet format v IEEE802.3 packet format
- Packet size (normal frames and jumbo frames)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

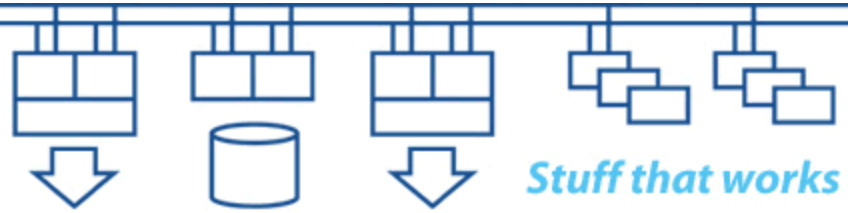
Why segment a network?

- Availability
- Performance
- Security

How can you segment a network?

- Multiple NICS in systems
- Repeaters
- Bridges
- Switches
- VLANs
- Routers

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Repeaters

- 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"*

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Bridges

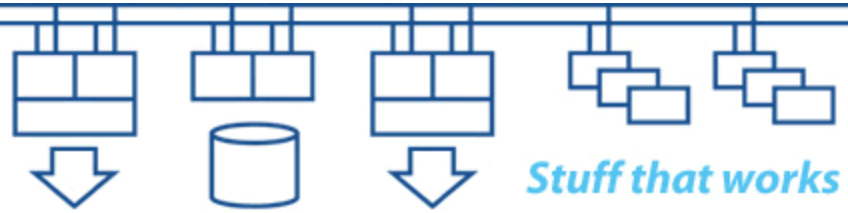
- Packet content based (Layer 2)
- Store and Forward
- Easy to use and configure
- Poor control of bandwidth (filtering)
- Spanning tree algorithm
- Provides an extended LAN
- Not all protocols can tolerate the inherent delays in working over an extended LAN
- Remote booting (MOP, BOOTP etc.) will absorb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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Switches

- Introduces 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, Layer 4 switching
  - Layer 2 is protocol independent – MAC address based
  - Layer 3 generally refers to TCP/IP routing layer
  - Layer 4 generally refers to TCP/IP ports, eg: HTTP port 80

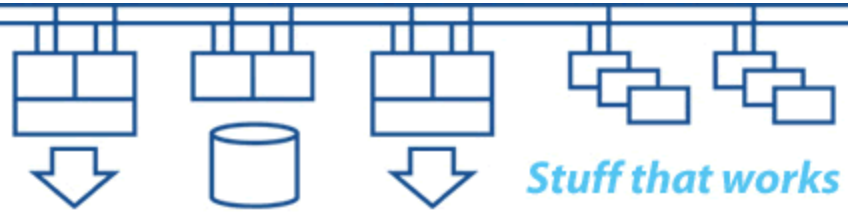
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



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

- Implemented within core switches
- Also implemented in NICs / device drivers (teaming)
- VLAN tagging of packets (802.1Q)
- Physical port based VLANs (at layer 2)
- Protocol based VLANs
- Connectivity between VLANs
- QoS (Quality of Service) and bandwidth reservation
- Tagged and untagged ports

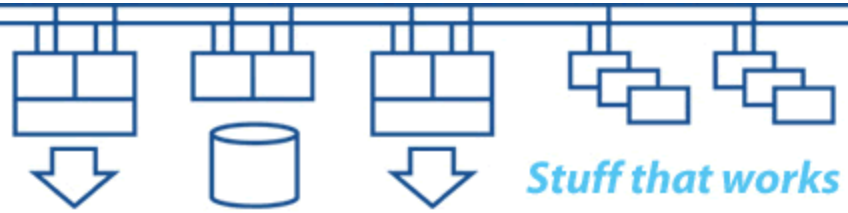
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



No single point of failure

- When there's more than one path, packets don't always arrive in the order in which they were sent
- Different manufacturers (Cisco, HP, Extreme etc.) have slightly different terminology and features (eg: Cisco 'etherchannel'; Procurve 'meshing'; Extreme 'EAPS ring'; etc.)
- Inter-switch links can be "trunked" to provide sufficient bandwidth
- NIC teaming / interface bonding / LAN failover
- Switch stacks
- Virtual routers

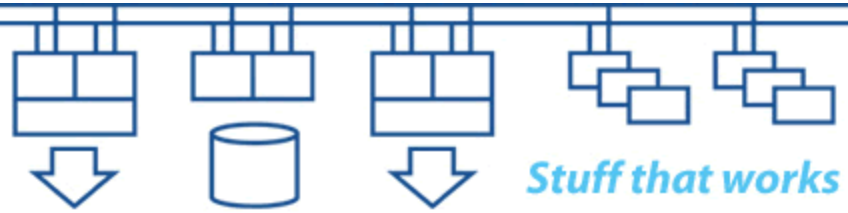
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



LAN extension between sites

- Link "glitches" will cause traffic disruption, so use routing (layer 3, not layer 2) to minimise disruption
- VLANs can extend across multiple switches
- Wave division multiplexing (DWDM, CWDM) can be used for extended distance inter-switch links

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Wireless ethernet

- Shared bandwidth ("flat" network)
- Security issues (access control, authentication, data encryption)
- Roaming issues (multiple Access Points and MAC address migration between ports)
- Management issues
- Antennas (coverage and beam patterns)
- Wireless repeaters and bridges

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Storage networks

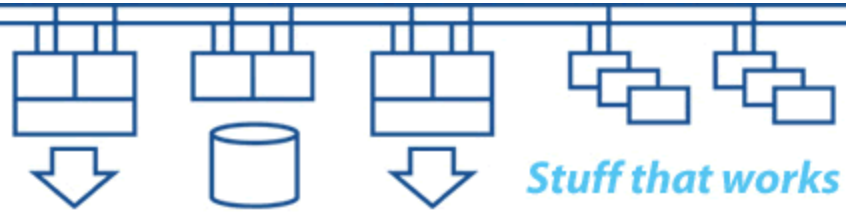
- Fibrechannel technologies (1 / 2 / 4 / 8 … Gbps)
- Storage devices (disc arrays and tape drives)
- Host Bus Adapters (HBAs) and FC switches
- WWIDs and WWNs (WWNNs and WWPNs)
- SAN Segmentation (switching, routing, zoning etc.)
- Storage subsystems and device presentation
- SAN extension (FC over IP, DWDM 'dark fibre', etc.)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Fibrechannel

- A switch based network optimised for shifting large quantities of data with high throughput and low latency
- All endpoints (HBAs, storage controllers etc.) uniquely identified with a WWID (World-Wide ID)
- Multiple switches can be interconnected
- Inter-switch links can be trunked
- The network between the storage devices and the systems is known as a fabric
- Large fabrics need to be segmented
- High availability typically uses a dual-fabric SAN

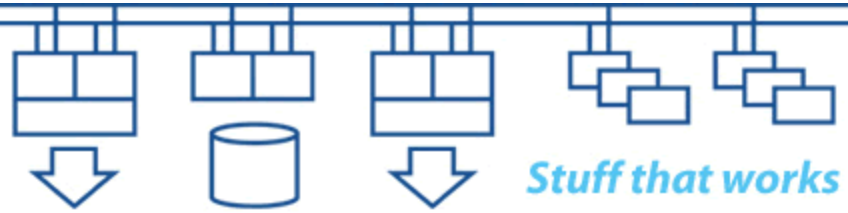
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Device addressing and presentation

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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



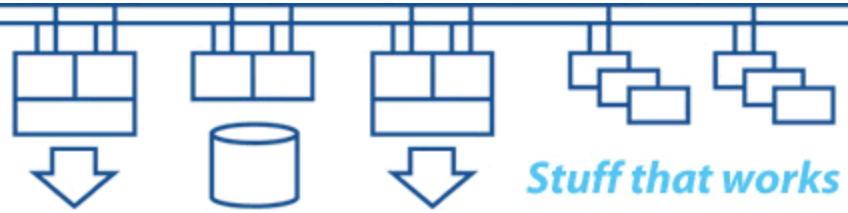
SAN segmentation

- Device paths and visibility can be controlled by zoning in the switches
- Zones can be physical port based, or WWID (WWPN or WWNN) based (known as "soft zoning")
- Zones can overlap (think Venn diagrams)
- Current zoning best practice uses the "single initiator, multiple targets" model

- Systems (HBAs) need to have BIOS type support for booting from SAN devices

- *TIP: HP "SAN design reference guide" is useful*

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



SAN extension between sites

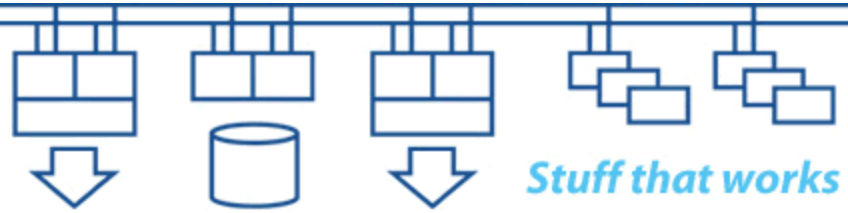
- Inter-site links can be "trunked" (as with data networks) to provide sufficient bandwidth
- Link "glitches" will cause fabric resets and rescans, so use FC routing in large extended SANs to minimise disruption
- Zones can extend across multiple switches (as with VLANs)
- Wave division multiplexing (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)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

1 (c): Wide Area Networking (WAN)

- ISDN, POTS
- Leased Line (KiloStream, MegaStream, T1 etc.)
- Frame Relay
- ATM
- MPLS
- "Dark fibre" and Wave Division Multiplexing
- SONET / SDH etc.
- ADSL / SDSL
- VPNs
- Managed services (usually TCP/IP based)
- Encapsulation and tunnelling
- FC over IP, FC over Ethernet

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Routing

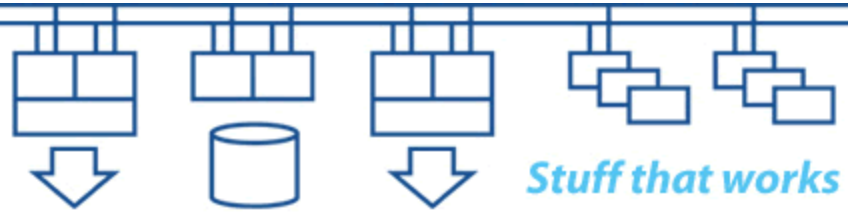
- Routers do not need to be involved in the normal inter-node traffic within a LAN, other than keeping track of who's where and making themselves known
- Routers build knowledge of address (node or interface) reachability on a per-protocol basis
- Protocol address based (Layer 3)
- Need to design addressing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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

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
- Can set up systems as dedicated multiprotocol routers if you need both DECnet and TCP/IP routing

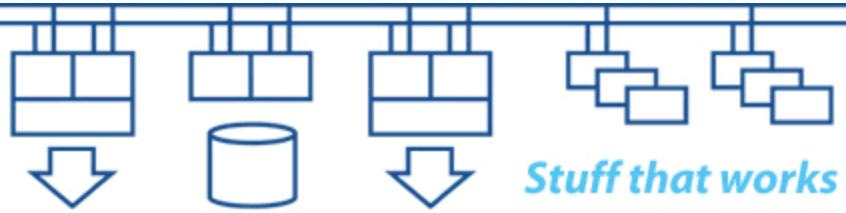
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Firewalls are used to block / allow specific traffic flows

- Separate devices or can be integrated into the router
- Need to design firewall rules
- Need to understand traffic flows and endpoints
- Good control over what talks to what
- Generally TCP/IP only
- Can add latency and limit bandwidth

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Part 2:

- "on the wire" network protocols

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Stuff that works

2: Some common data protocols

Typical network protocols "on the wire":

- SCS (clustering)
- TCP/IP (and all it's component sub-protocols)
- DECnet-Plus (NSP,OSI and "DECnet over IP" transport layers) or DECnet Phase IV (NSP transport only)
- DECdns (not to be confused with TCP/IP's DNS/BIND)
- LAT (DECserver terminal access etc.)
- MOP and Remote Console (DECserver, LAVC boot etc.)
- DTSS (can be disabled)
- LAD and LAST (Infoserver etc.)
- AMDS (quorum adjustment)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

TCP/IP V4 layers

- 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)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

TCP/IP services

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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

DECnet Phase IV - addressing

- 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 LAN adapters based on DECnet node address, so cannot connect multiple LAN adapters to the same VLAN

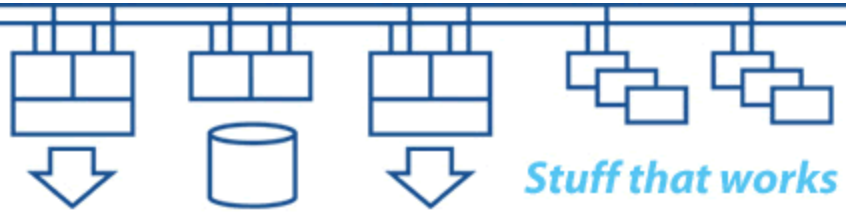
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

DECnet Phase IV – who's where?

- DECnet "hidden information":
  - End Node to Routers (end node hello packets)
  - Routers to Routers (routing updates)
  - Routers to End Nodes (router hello packets)

- DECnet Phase IV bases the MAC address on the node number, so no need for routers on LAN except for determining adjacencies

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



- Predominantly TCP/IP

- NetBIOS / NetBEUI and NetBIOS over TCP
- WINS naming service
- Windows file sharing
- Windows printer sharing
- Pathworks provides DECnet for Windows and Windows file & printer sharing, largely replaced by CIFS (based on Samba)
- "NIC teaming" for availability

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

- Predominantly TCP/IP

- "NIC teaming" for availability
- DECnet for Linux project
- SAMBA for Windows file & printer sharing

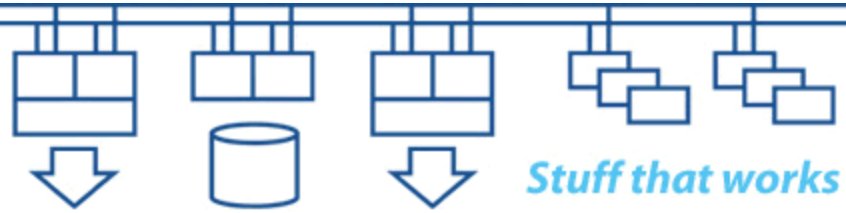
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Part 3:

- Network infrastructures - putting it all together

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Network link characteristics

- Bandwidth – determines throughput
  - Large packets shift more data with less overhead
- Signal path quality and reliability
  - Retransmits severely affect overall throughput
- 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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Traffic flow and performance

- 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 – 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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Availability (1)

- 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
- May wish to take advantage of installed bandwidth capacity to provide additional functionality when everything is working
- Document and diagram the network configurations
- Install local UPS for network hardware

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Availability (2)

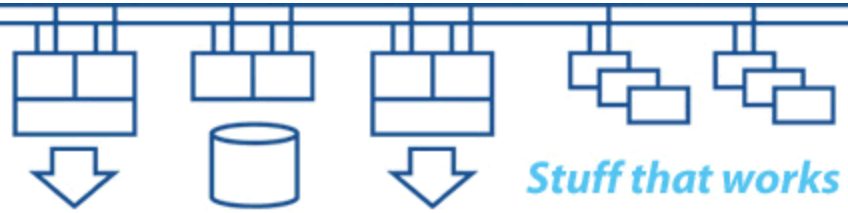
- Duplicate devices such as network printers and terminal servers on different paths
- Segment the network to provide gradual degradation rather than wholesale failure
- Have a fallback plan for getting to remote site equipment (eg: dial-up modem to console port of router)
- Network management is all about problem identification and rectification
- Learn the warning signs for common problems

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Data connectivity

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

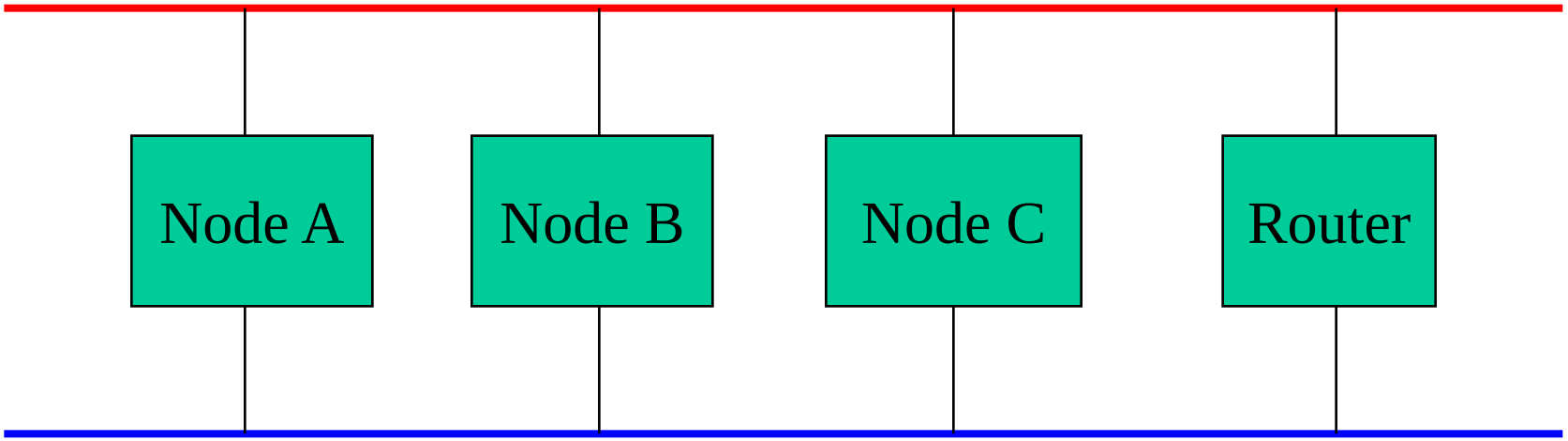
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Multiple LAN interface behaviours

- NIC teaming and equivalents
- load balancing over multiple interfaces
- network booting

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Dual LAN behaviour



- TCP/IP – multiple NICs per subnet, dynamic routing
- DECnet Phase IV– L1 routers or end-node failover
- DECnet-Plus –Multi-homed ES or IS, load balancing

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Part 4:

- Examples
- Discussion

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Stuff that works

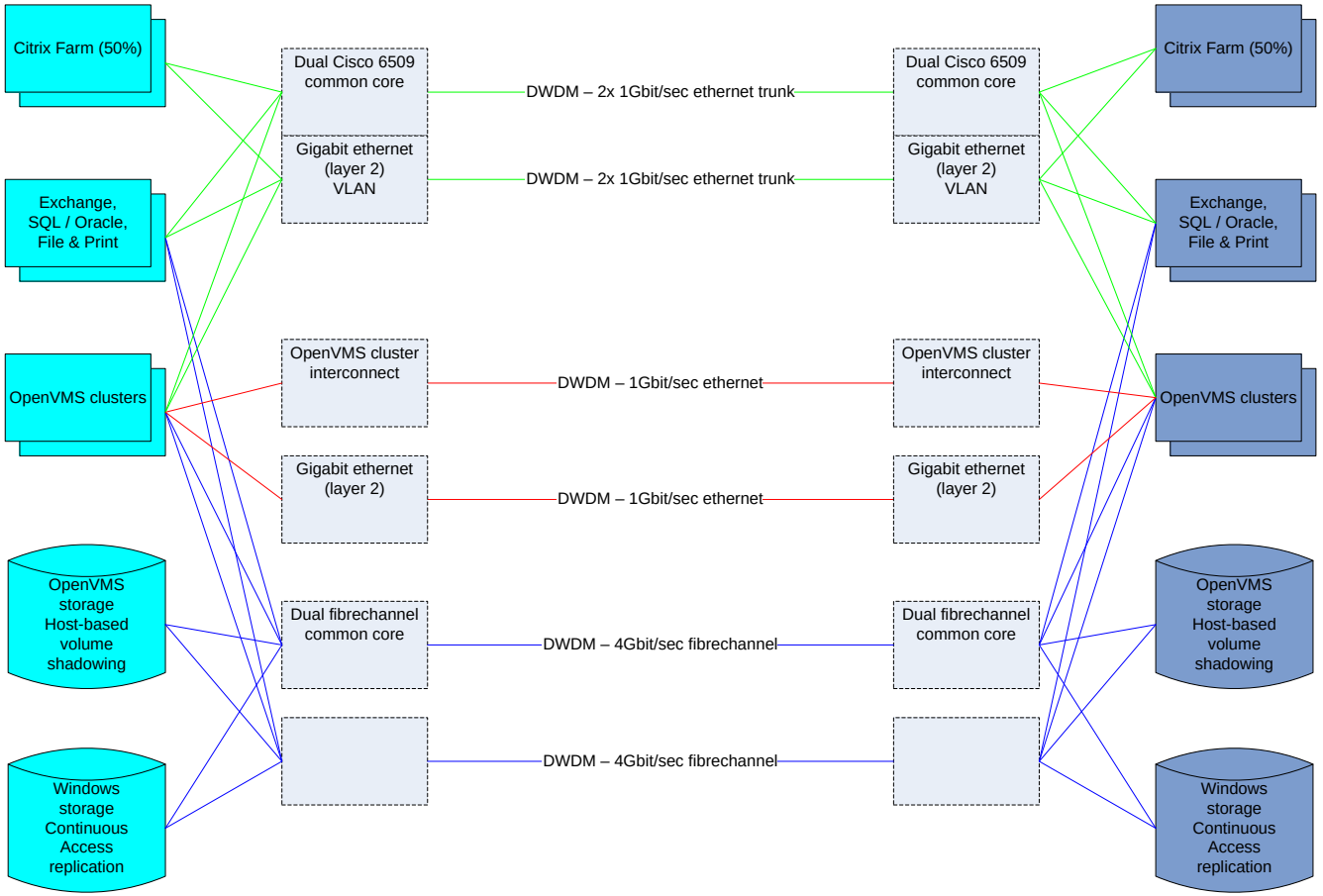
Examples and discussion

- Safety-critical and mission-critical system:
  - HP Integrity Servers
  - Merge 3x regional clusters to single national cluster
  - EVA P6350s for storage
  - Multiple NIC connectivity (NSPOF)

- Similar principles apply in many other cases

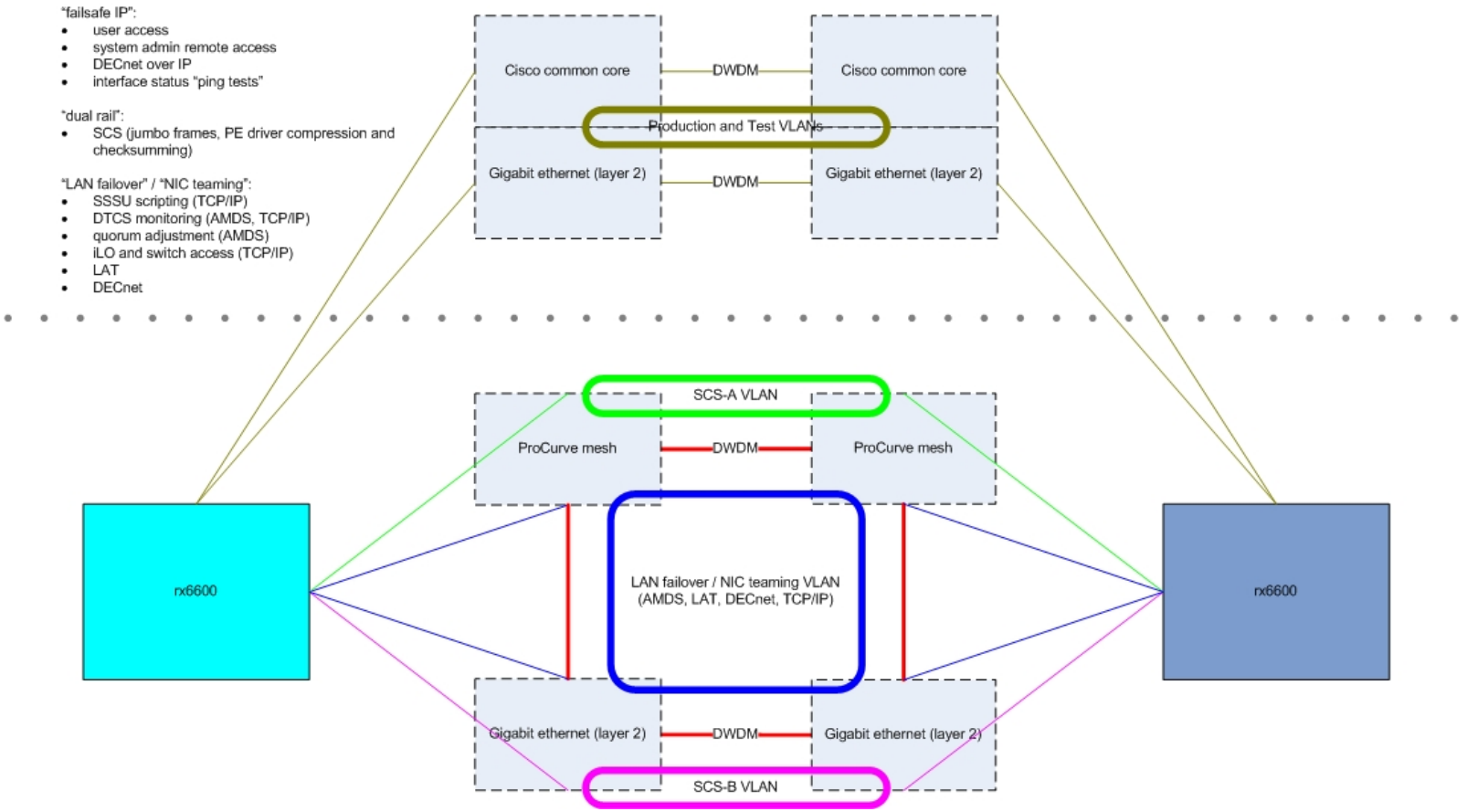
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Example overall infrastructure



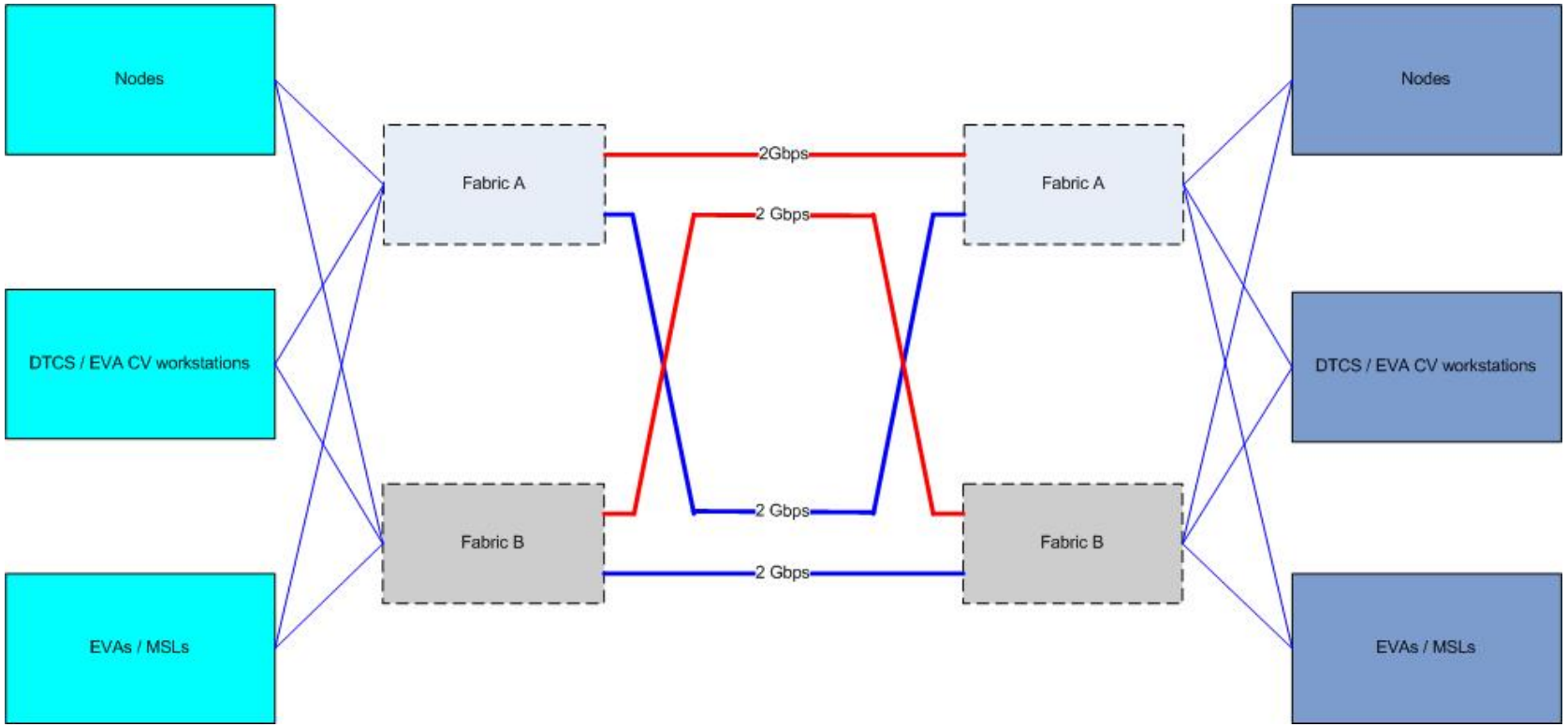
Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Common network infrastructure



Copyright © Colin Butcher, XDelta Limited, Apr. 2013

SAN fabric infrastructure

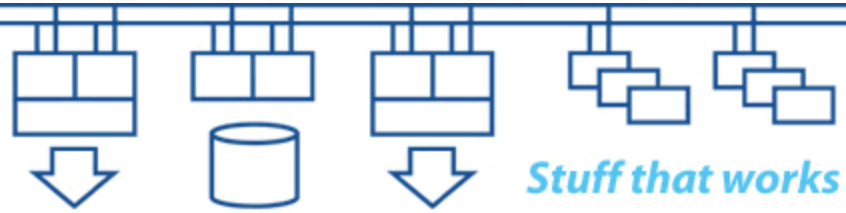


Copyright © Colin Butcher, XDelta Limited, Apr. 2013

System configurations

- System configuration – the art is to select components that work well together and which provide the bulk of what you need with minimal additional work
- Establish the minimum requirements that have to be met – and do it as well as possible
- Availability and performance have to be designed in
- Monitoring and automation are key components
- Understand the typical behaviour of your systems and be aware of changes
- Configuration control – hardware, firmware, settings, software etc.

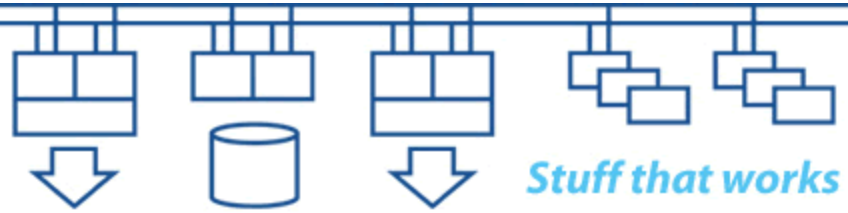
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Systems infrastructure

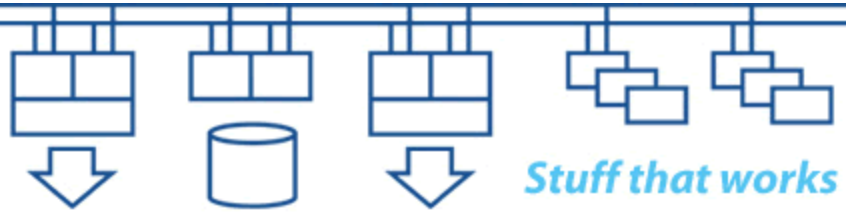
- Use multiple systems within a site – and have appropriate physical separation between them
- Use multiple sites with appropriate geographical separation
- Understand what happens when a failure occurs and how the overall system configuration is likely to respond
- Avoid the risk of more than one system thinking that it's in charge when a failure has happened
- Ensure that the surrounding infrastructure and environment is appropriate to the needs of your systems
- Configure the individual components of the systems in a manner that maximises interchangeability

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



- 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)

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Network infrastructure

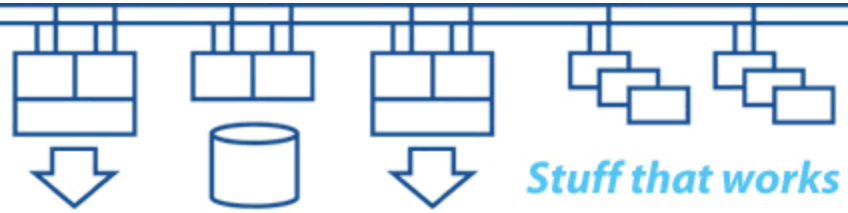
- Scale network so that overall performance is based on minimum essential number of paths and maximum estimated traffic
- May wish to take advantage of installed bandwidth capacity to provide additional functionality when everything is working
- There is no such thing as a single protocol network
- Understand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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Design and implementation

- Think big, implement small, expand as needed
- Documentation
- Modularity
- Reliability
- Scalability
- Configuration control
- Testing (development, pre-deployment, post-deployment)
- Installation
- Monitoring
- Management and firmware updates

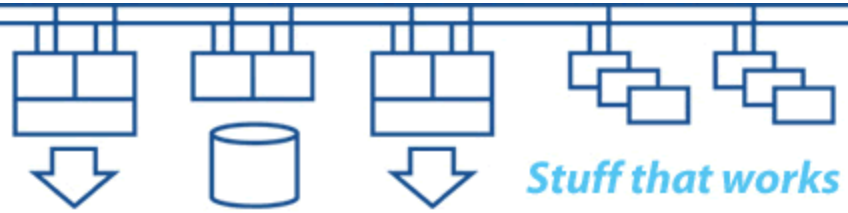
Copyright © Colin Butcher, XDelta Limited, Apr. 2013



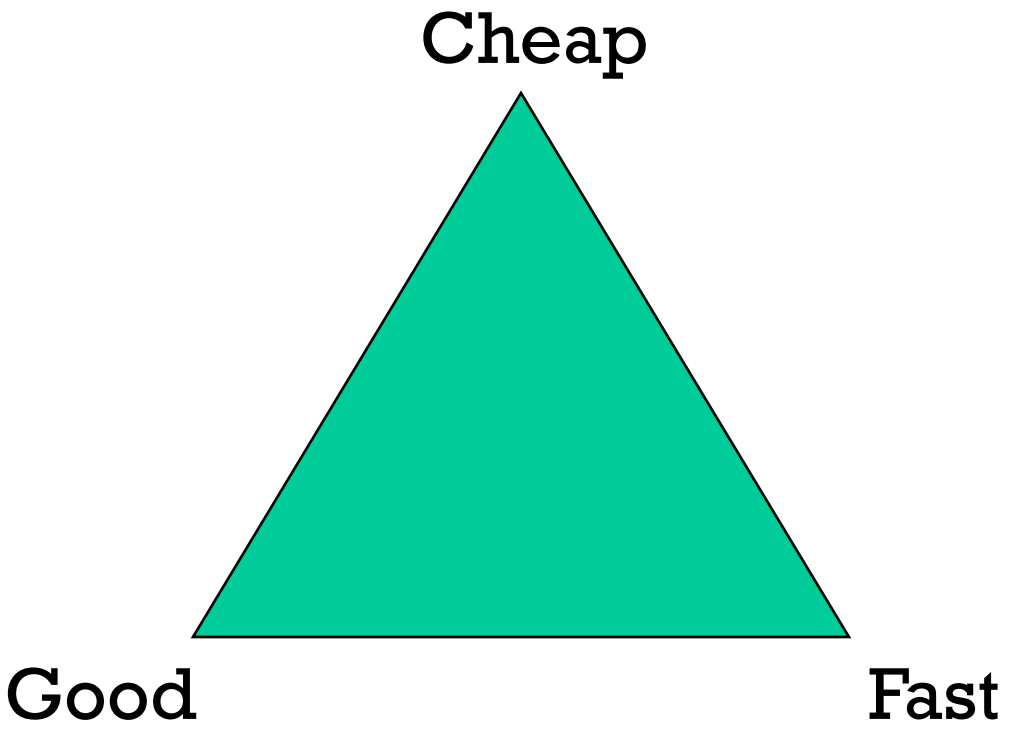
Summary

- Understand the basic principles by which networks work
- Network hardware tends to have a long life-cycle
- Minimise the chance of making mistakes
- Avoid complexity where possible
- Design for change over time
- 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

Copyright © Colin Butcher, XDelta Limited, Apr. 2013



Compromises – pick any two!



Copyright © Colin Butcher, XDelta Limited, Apr. 2013

Thank you for your participation.

Discussion!

The Chartered Institute for IT
Enabling the information society

Copyright © Colin Butcher, XDelta Limited, Apr. 2013