Resilient Network Design Concepts

Mark Tinka

“The Janitor Pulled the Plug…”

- Why was he allowed near the equipment?
- Why was the problem noticed only afterwards?
- Why did it take 6 weeks to determine the problem?
- Why wasn’t there redundant power?
- Why wasn’t there network redundancy?

Network Design and Architecture…

- … is of critical importance
- … contributes directly to the success of the network
- … contributes directly to the failure of the network

> **"No amount of magic knobs will save a sloppily designed network"**

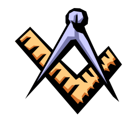
Paul Ferguson—Consulting Engineer,
Cisco Systems

What is a Well-Designed Network?

- A network that takes into consideration these important factors:
  - Physical infrastructure
  - Topological/protocol hierarchy
  - Scaling and Redundancy
  - Addressing aggregation (IGP and BGP)
  - Policy implementation (core/edge)
  - Management/maintenance/operations
  - Cost

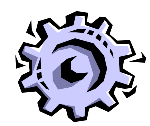
The Three-legged Stool

- Designing the network with resiliency in mind
- Using technology to identify and eliminate single points of failure
- Having processes in place to reduce the risk of human error



Design

Technology



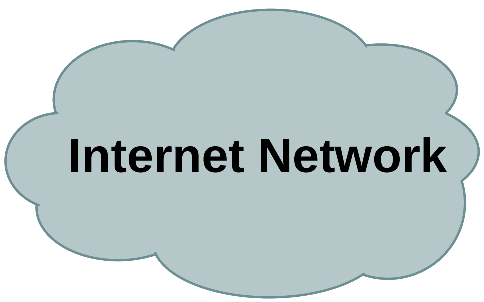
Process

- All of these elements are necessary, and all interact with each other
  - One missing leg results in a stool which will not stand

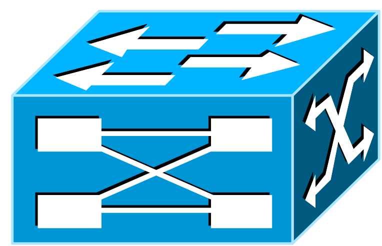
Design

New World vs. Old World

- Internet/L3 networks
  - Build the redundancy into the system
- Telco Voice and L2 networks
  - Put all the redundancy into a box

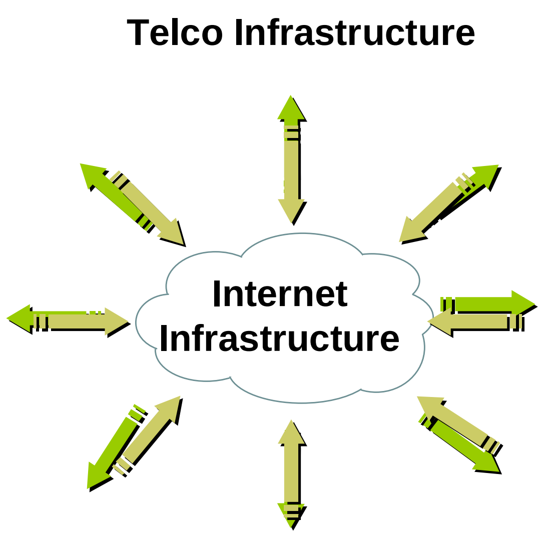


vs.



New World vs. Old World

- Despite the change in the Customer ↔ Provider dynamic, the fundamentals of building networks have not changed
- ISP Geeks can learn from Telco Bell Heads the lessons learned from 100 years of experience
- Telco Bell Heads can learn from ISP Geeks the hard experience of scaling at +100% per year



How Do We Get There?

Design

"In the Internet era, reliability is becoming something you have to build, not something you buy. That is hard work, and it requires intelligence, skills and budget. Reliability is not part of the basic package."

Joel Snyder – Network World Test Alliance 1/10/2000
"Reliability: Something you build, not buy"

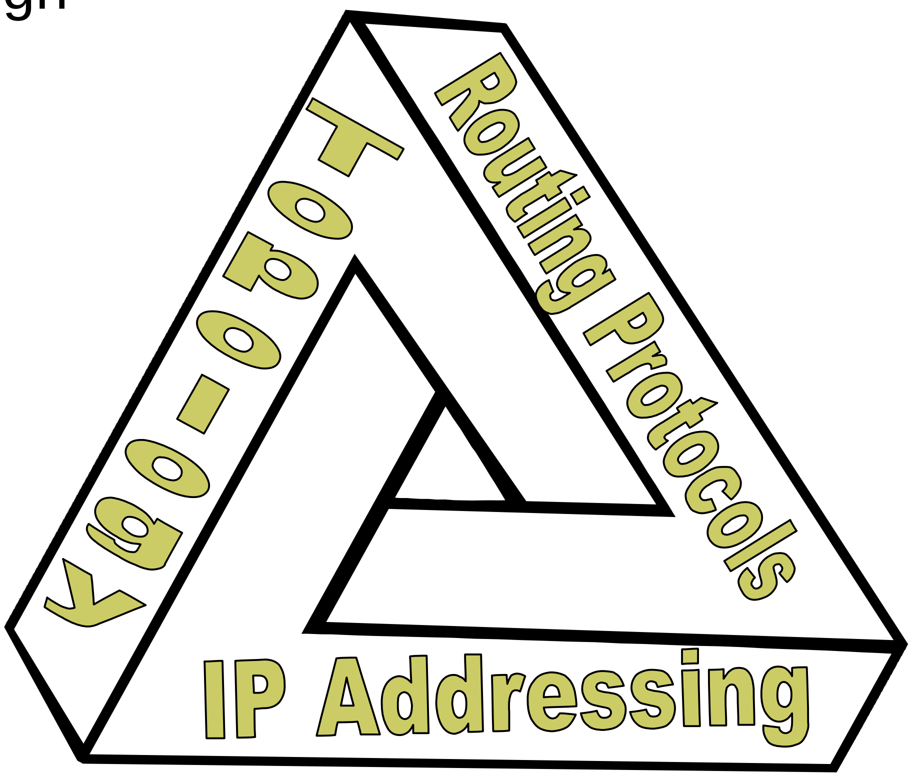
Redundant Network Design

Concepts and Techniques

Basic ISP Scaling Concepts

Design

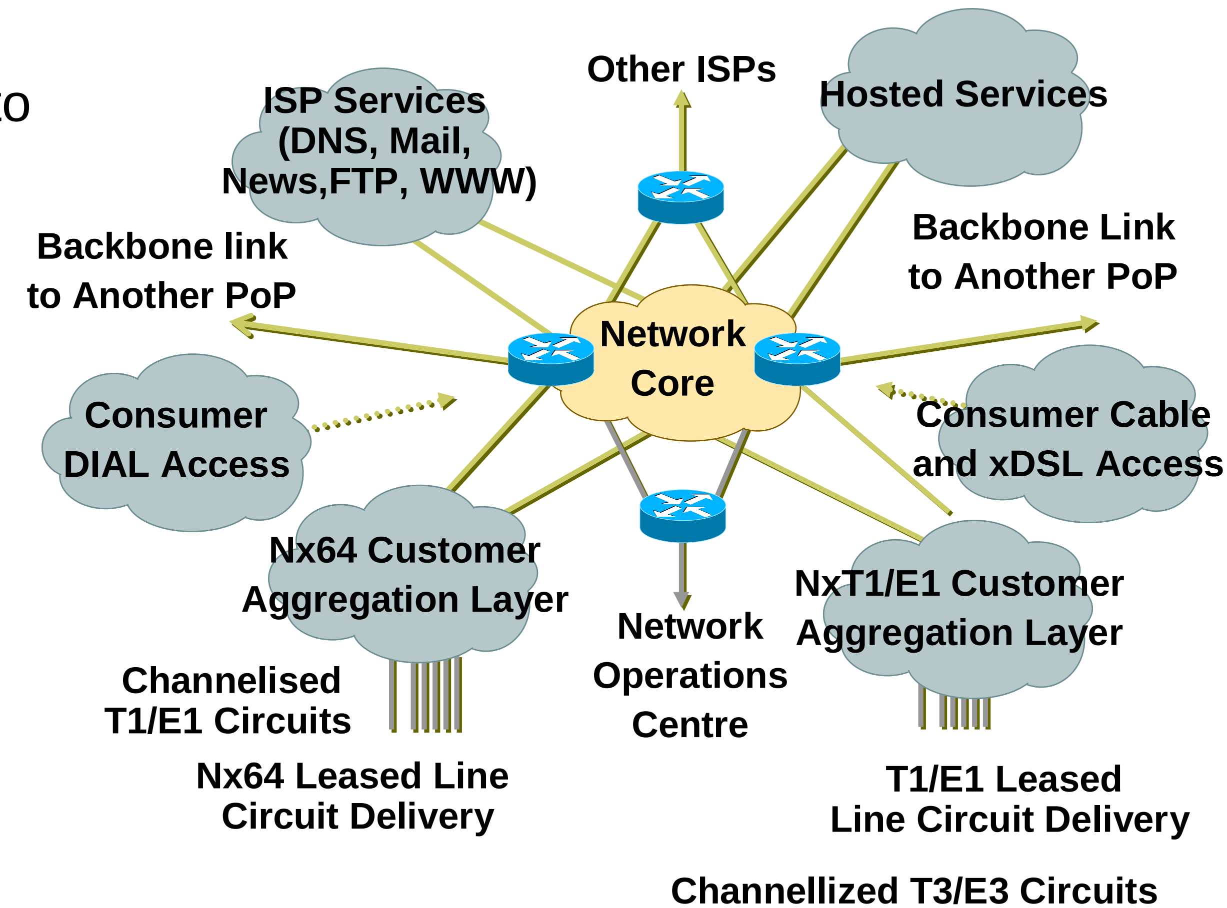
- Modular/Structured Design
- Functional Design
- Tiered/Hierarchical Design Discipline



Modular/Structured Design

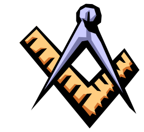
Design

- Organize the network into separate and repeatable modules
  - Backbone
  - PoP
  - Hosting services
  - ISP Services
  - Support/NOC



Modular/Structured Design

- Modularity makes it easy to scale a network
  - Design smaller units of the network that are then plugged into each other
  - Each module can be built for a specific function in the network
  - Upgrade paths are built around the modules, not the entire network



Design

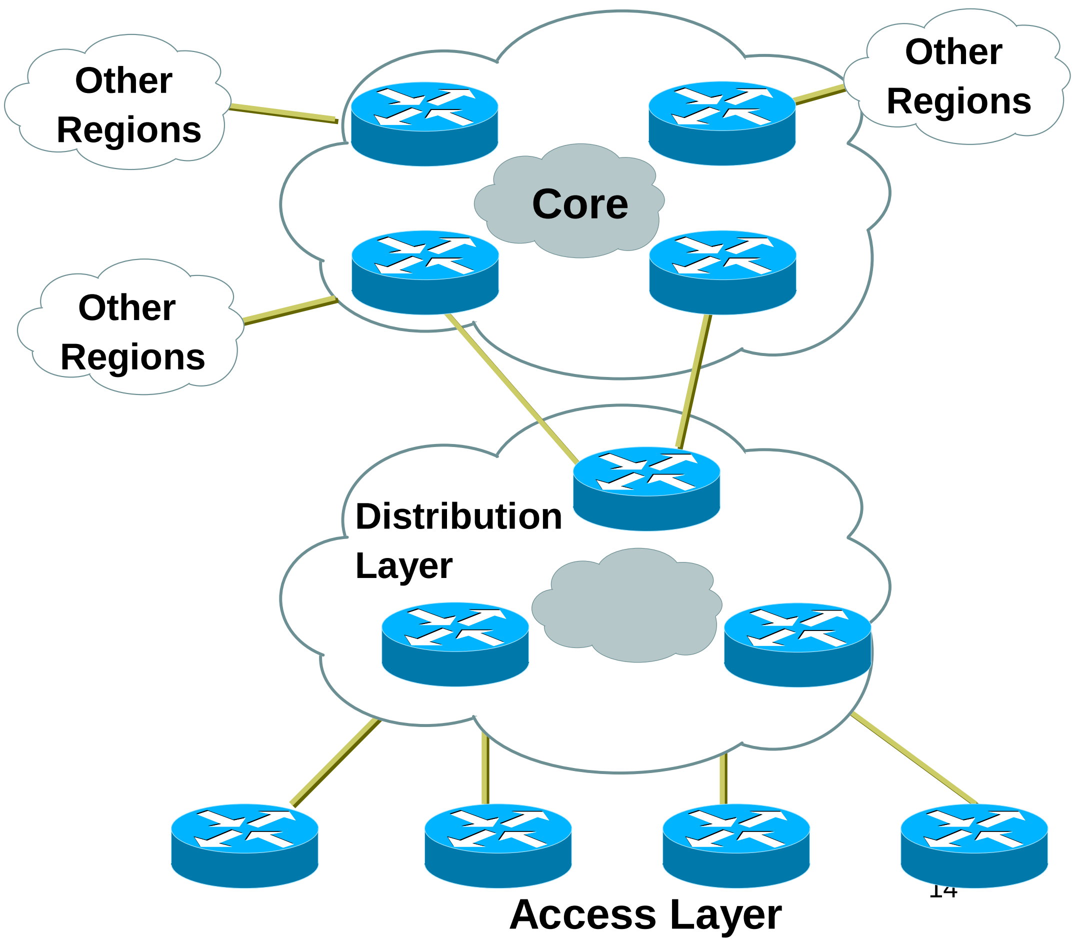
Functional Design

- One Box cannot do everything
  - (no matter how hard people have tried in the past)
- Each router/switch in a network has a well-defined set of functions
- The various boxes interact with each other
- Equipment can be selected and functionally placed in a network around its strengths
- ISP Networks are a systems approach to design
  - Functions interlink and interact to form a network solution.

Design

Tiered/Hierarchical Design

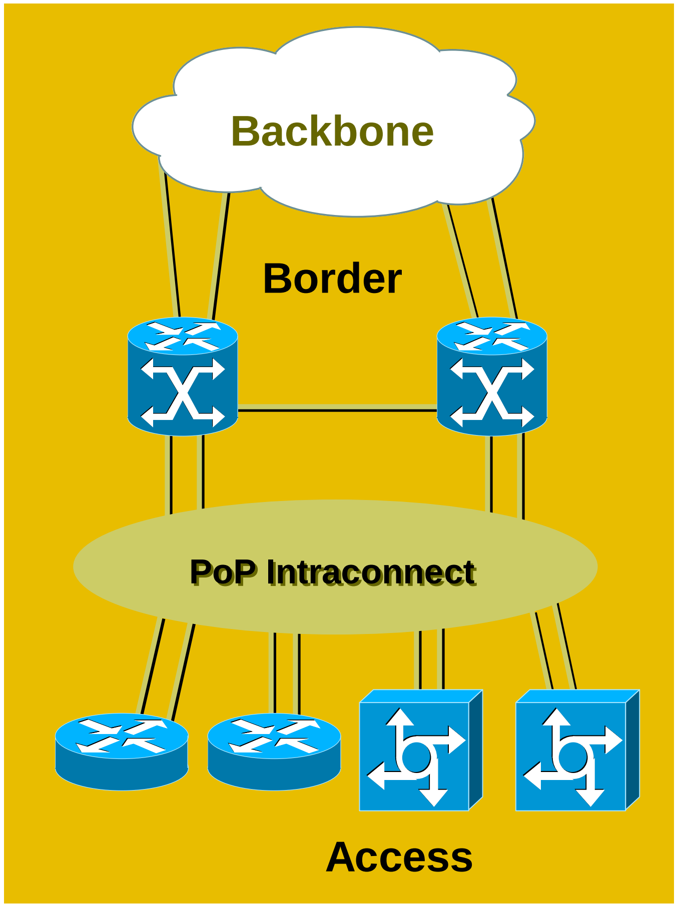
- Flat meshed topologies do not scale
- Hierarchy is used in designs to scale the network
- Good conceptual guideline, but the lines blur when it comes to implementation.



Multiple Levels of Redundancy

Design

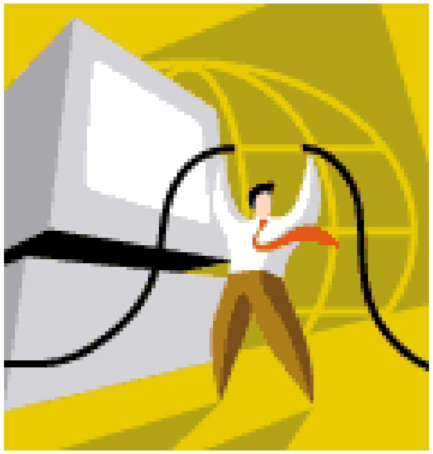
- Triple layered PoP redundancy
  - Lower-level failures are better
  - Lower-level failures may trigger higher-level failures
  - L2: Two of everything
  - L3: IGP and BGP provide redundancy and load balancing
  - L4: TCP re-transmissions recover during the fail-over



Multiple Levels of Redundancy

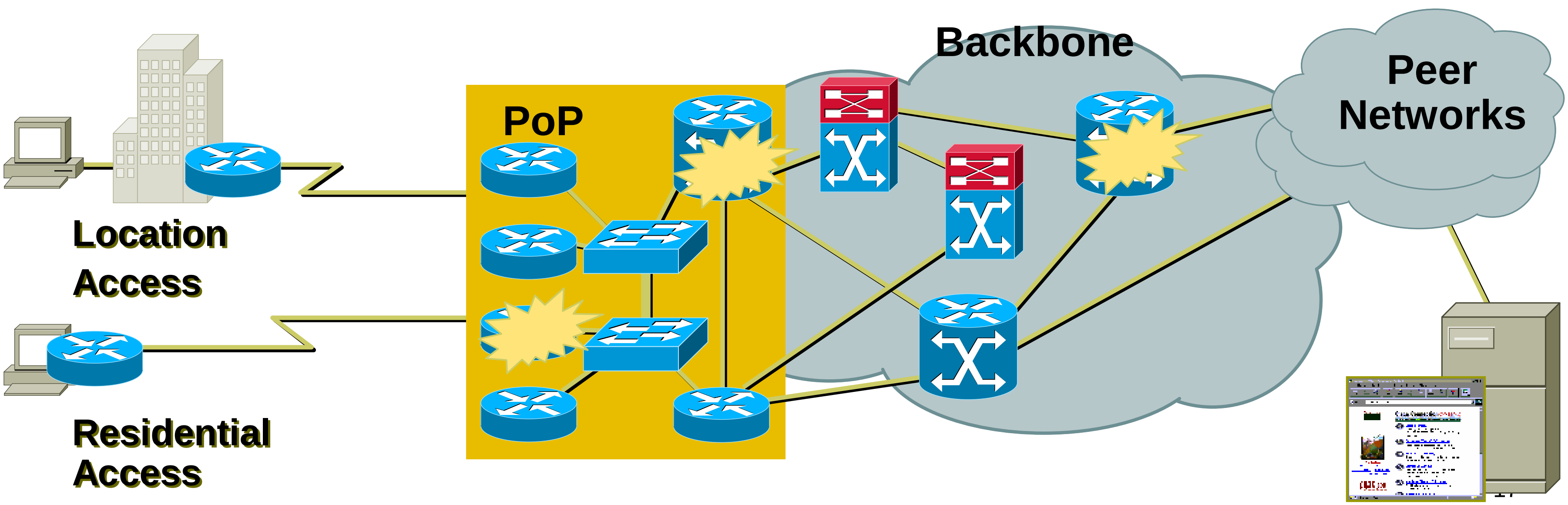
Design

- Multiple levels also mean that one must go deep – for example:
  - Outside Cable plant – circuits on the same bundle – backhoe failures
  - Redundant power to the rack – circuit over load and technician trip
- MIT (maintenance injected trouble) is one of the key causes of ISP outage.

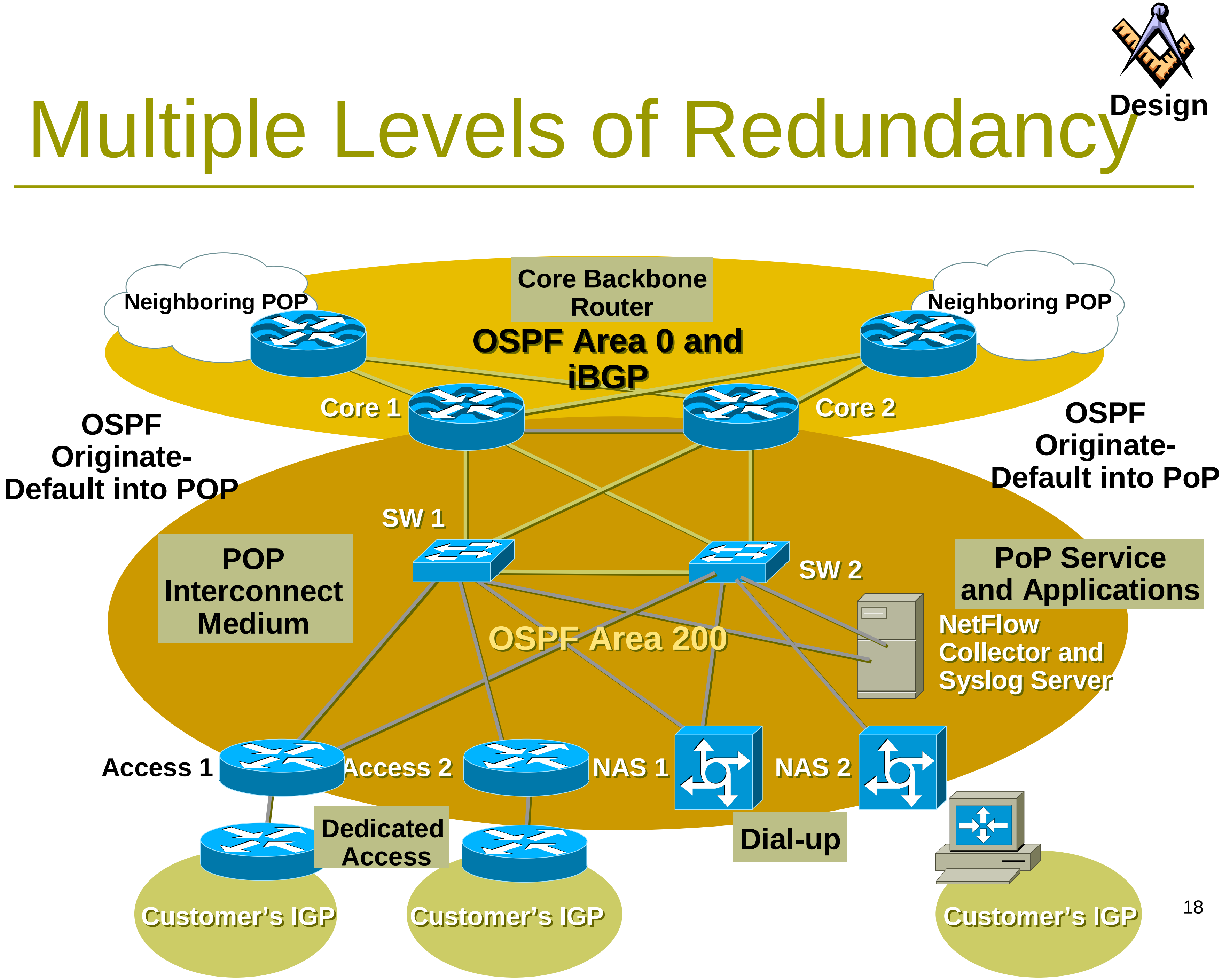


Multiple Levels of Redundancy

- Objectives –
  - As little user visibility of a fault as possible
  - Minimize the impact of any fault in any part of the network
  - Network needs to handle L2, L3, L4, and router failure



Multiple Levels of Redundancy



Redundant Network Design

The Basics

The Basics: Platform

Design

- Redundant Power
  - Two power supplies
- Redundant Cooling
  - What happens if one of the fans fail?
- Redundant route processors
  - Consideration also, but less important
  - Partner router device is better
- Redundant interfaces
  - Redundant link to partner device is better

The Basics: Environment

Design

- Redundant Power
  - UPS source – protects against grid failure
  - “Dirty” source – protects against UPS failure
- Redundant cabling
  - Cable break inside facility can be quickly patched by using “spare” cables
  - Facility should have two diversely routed external cable paths
- Redundant Cooling
  - Facility has air-conditioning backup
  - …or some other cooling system?

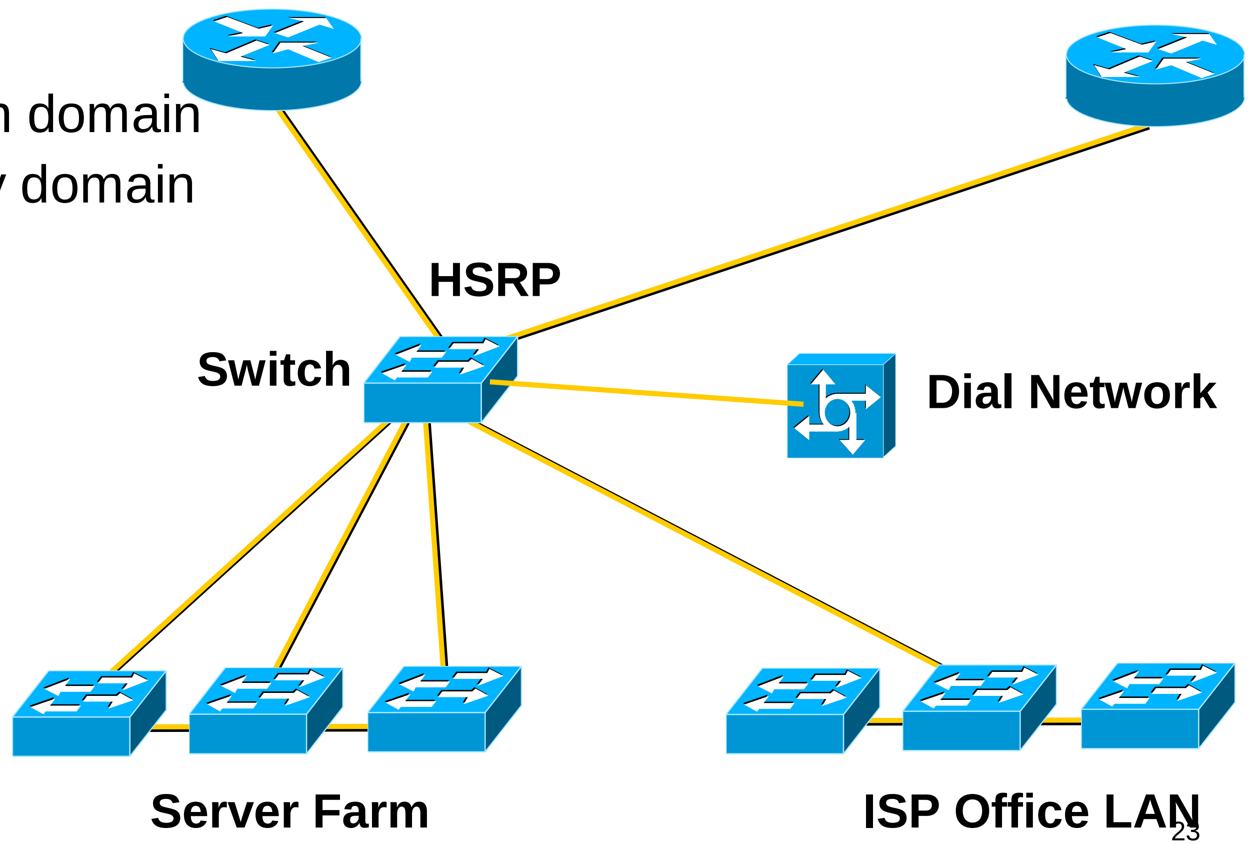
Redundant Network Design

Within the DataCentre

Design

Bad Architecture (1)

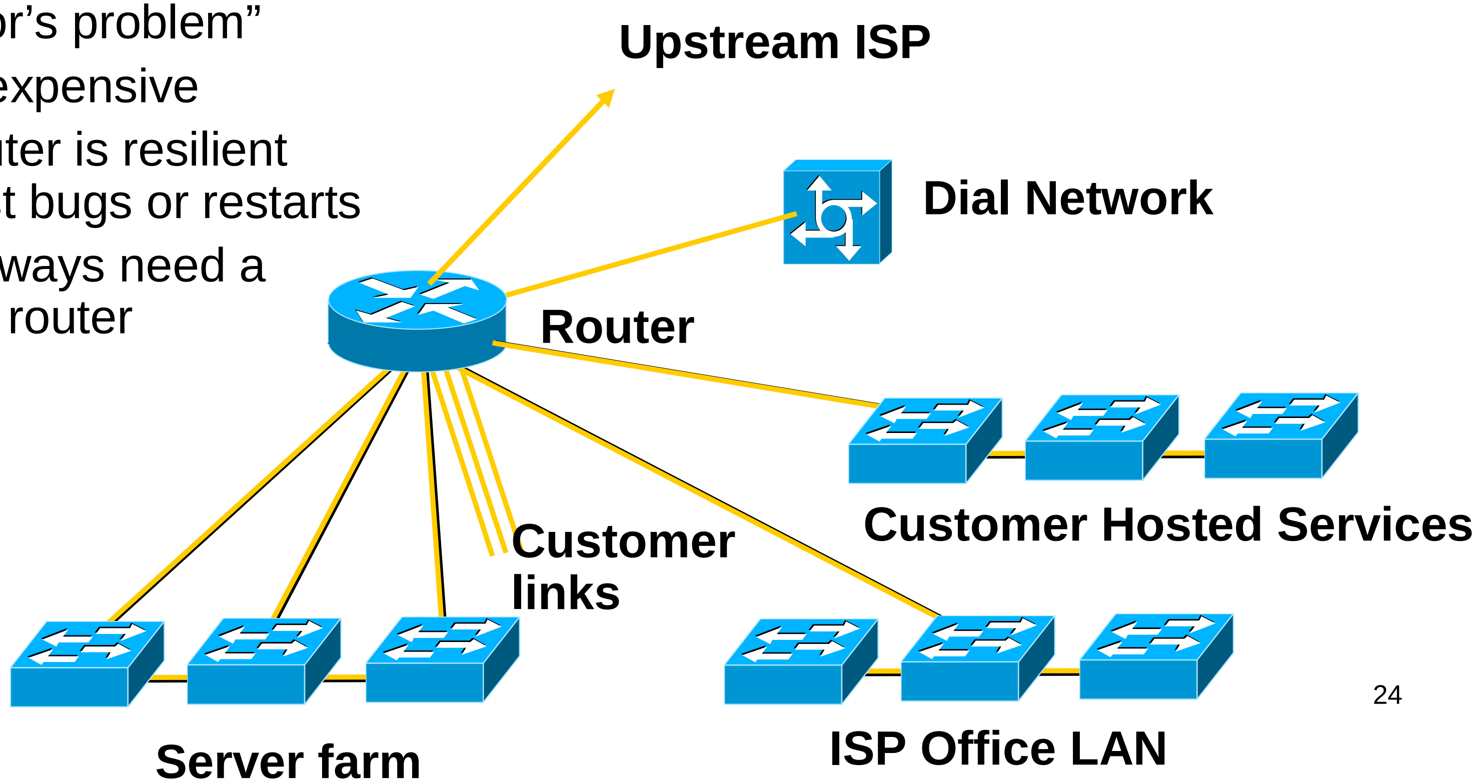
- A single point of failure
  - Single collision domain
  - Single security domain
  - Spanning tree convergence
  - No backup
  - Central switch performance



Bad Architecture (2)

Design

- A central router
  - Simple to build
  - Resilience is the “vendor’s problem”
  - More expensive
  - No router is resilient against bugs or restarts
  - You always need a bigger router

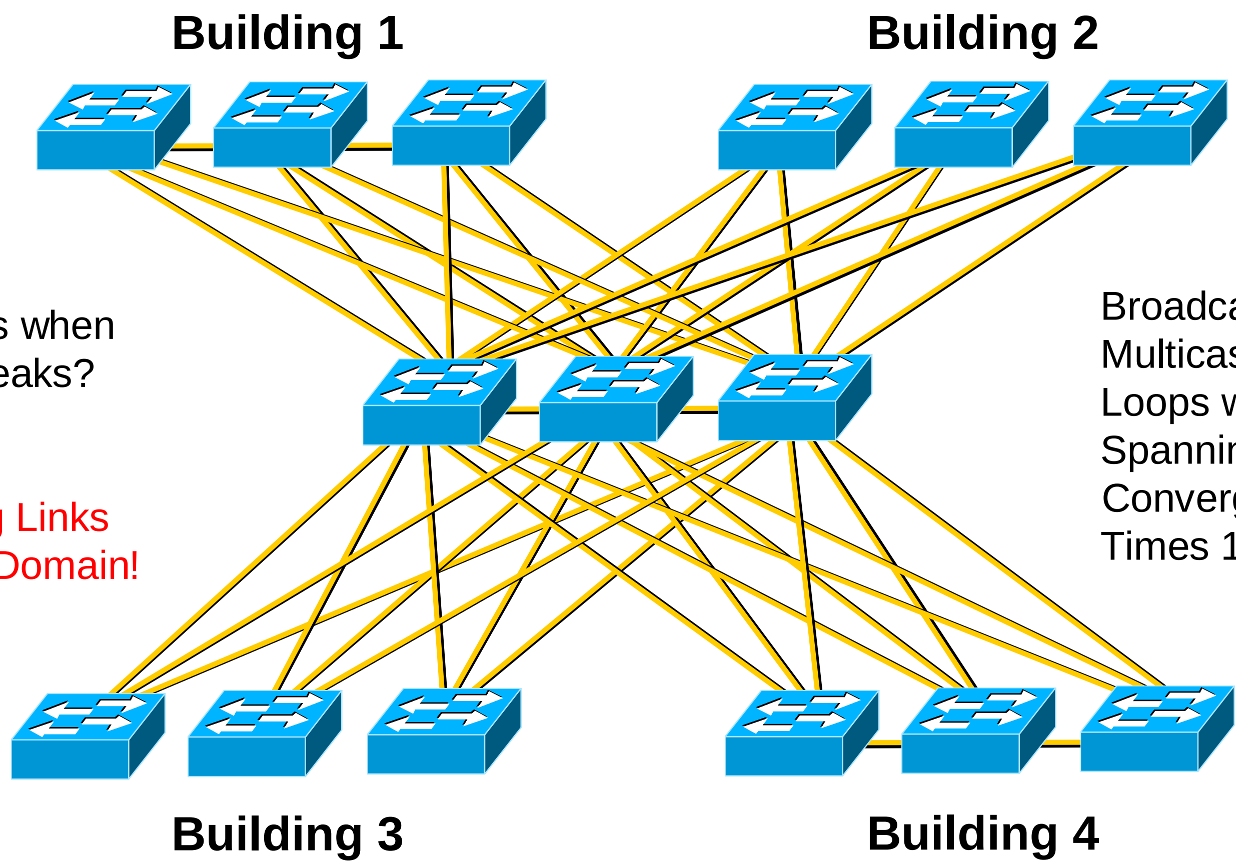


Even Worse!!

Design

- Avoid Highly Meshed, Non-Deterministic Large Scale L2

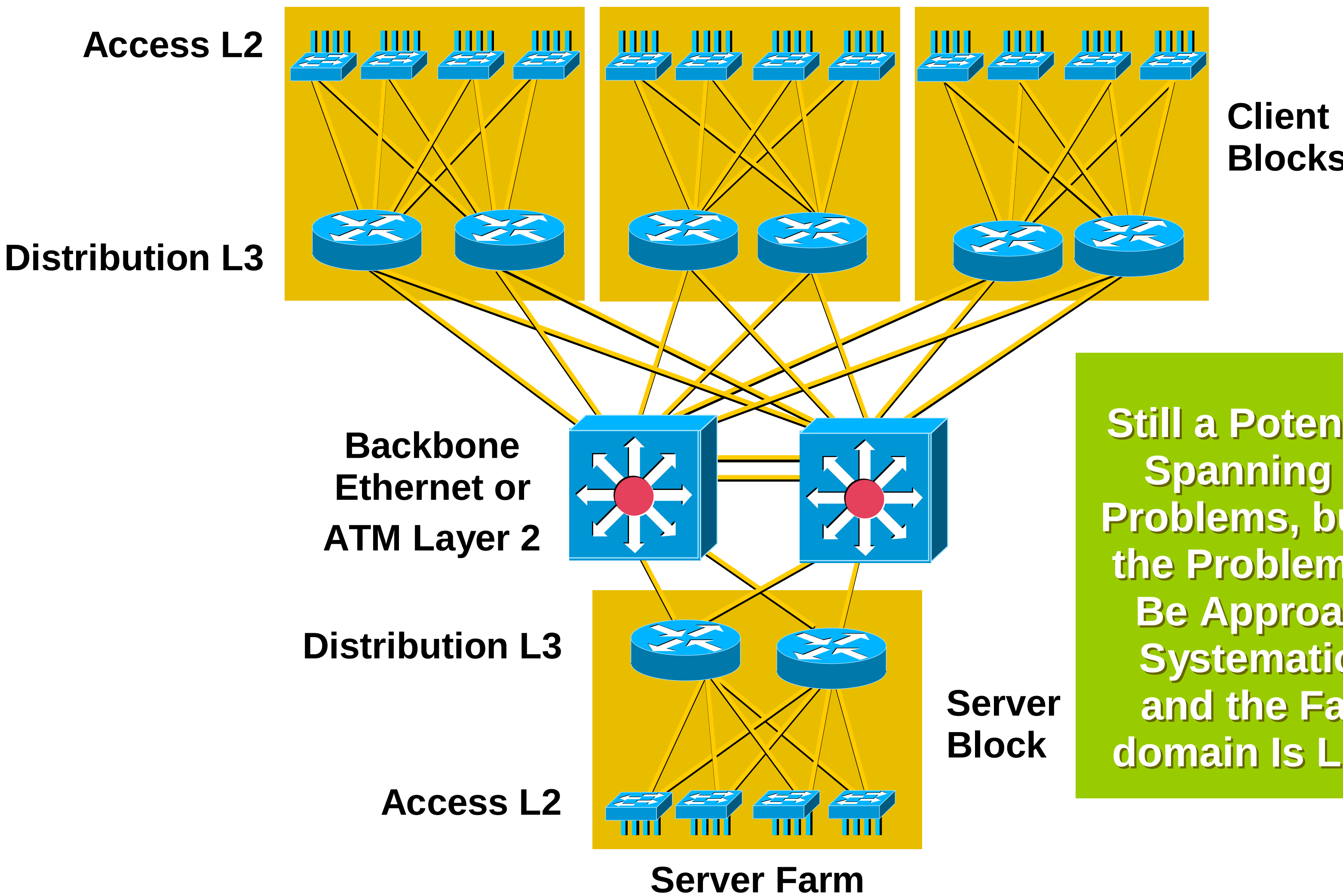
Where Should Root Go?
What Happens when Something Breaks?
How Long to Converge?
Many Blocking Links
Large Failure Domain!



Broadcast Flooding
Multicast Flooding
Loops within Loops
Spanning Tree Convergence Time
Times 100 VLANs?

Design

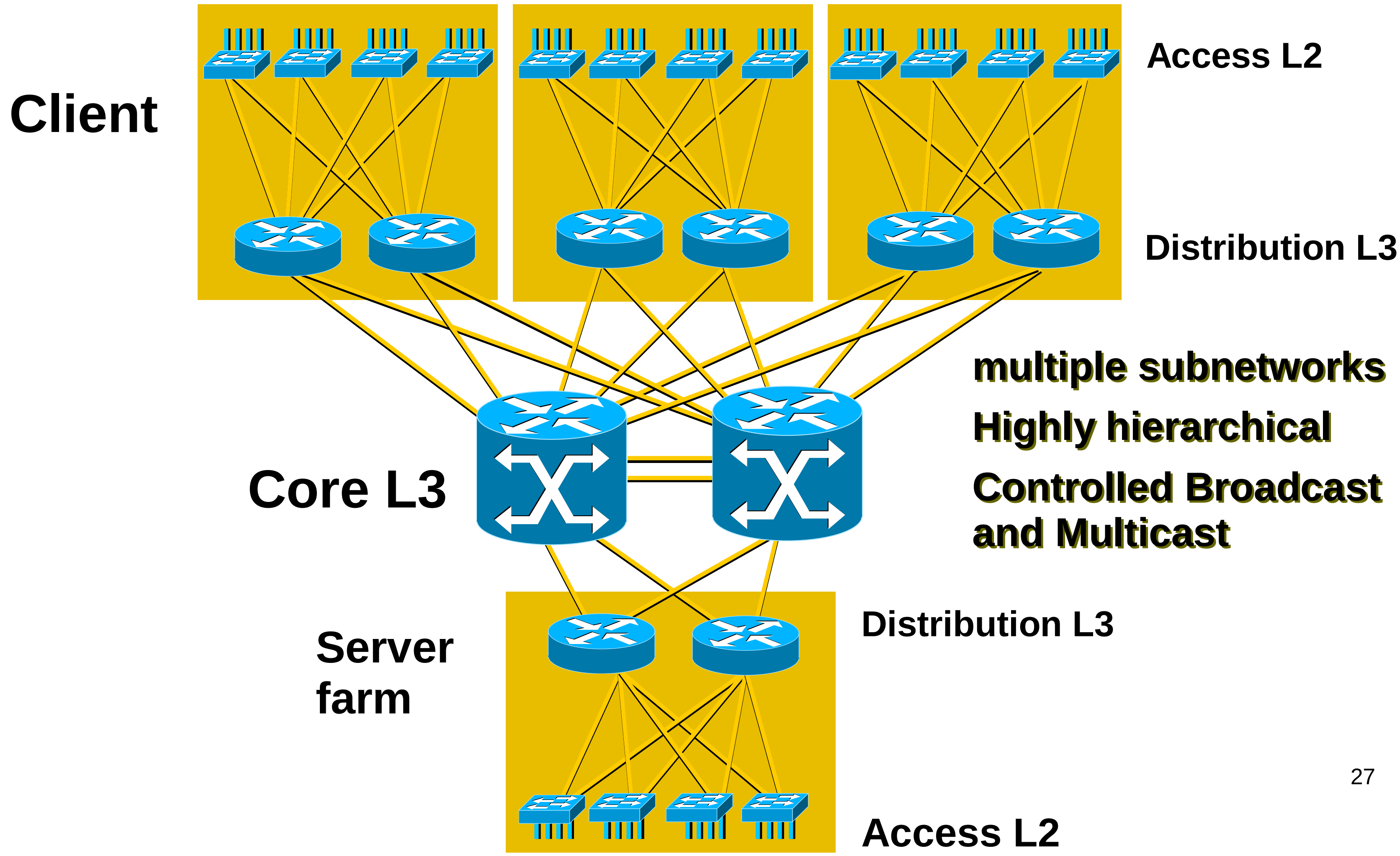
Typical (Better) Backbone



Still a Potential for Spanning Tree Problems, but Now the Problems Can Be Approached Systematically, and the Failure domain Is Limited

The best architecture

Design



multiple subnetworks

Highly hierarchical

Controlled Broadcast and Multicast


Technology

Benefits of Layer 3 backbone

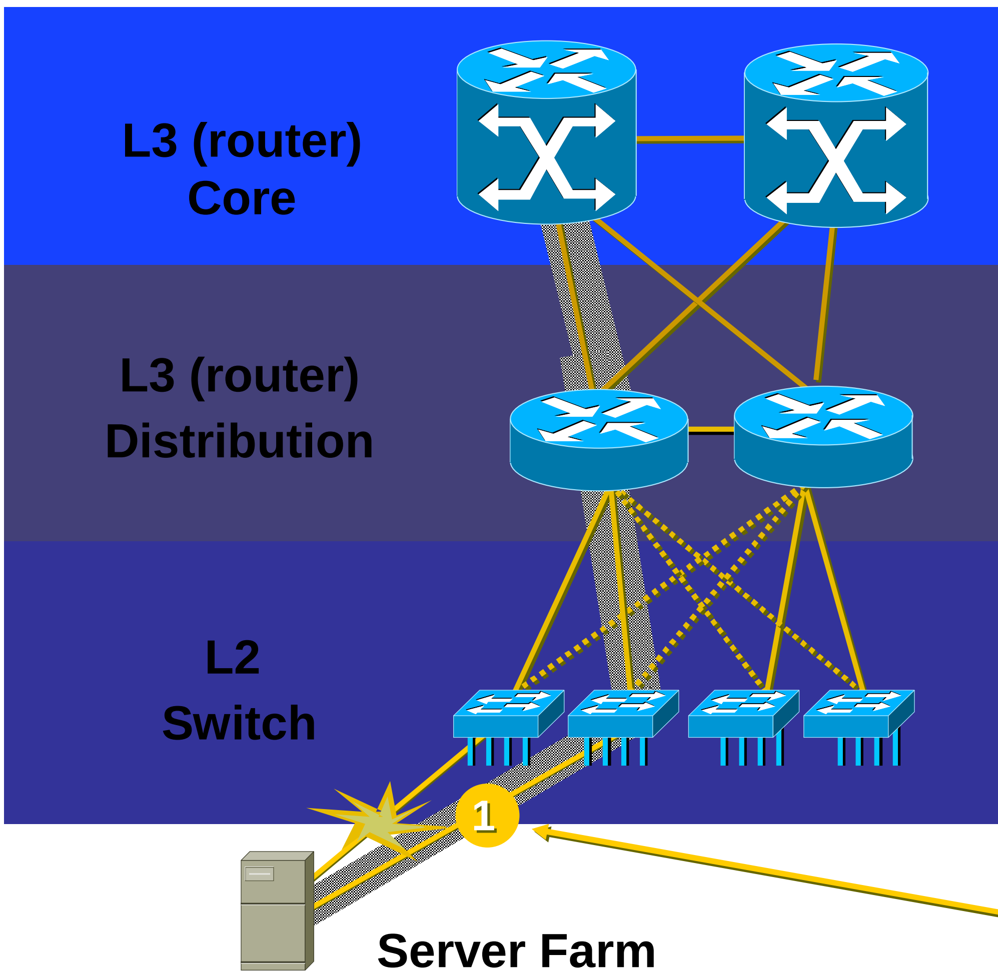
- Multicast PIM routing control
- Load balancing
- No blocked links
- Fast convergence OSPF/ISIS/EIGRP
- Greater scalability overall
- Router peering reduced

Redundant Network Design

Server Availability

Technology

Multi-homed Servers



Using Adaptive Fault Tolerant" Drivers and NICs

NIC Has a Single IP/MAC Address (Active on one NIC at a Time)

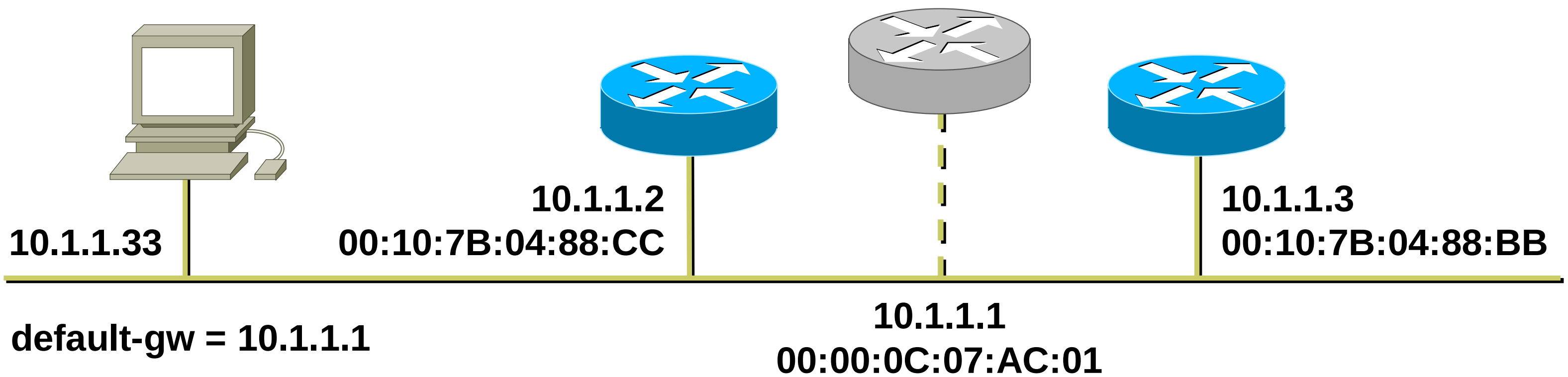
When Faulty Link Repaired, Does Not Fail Back to Avoid Flapping

Fault-tolerant Drivers Available from Many Vendors: Intel, Compaq, HP, Sun

Many Vendors also Have Drivers that also Support etherchannel

Dual-homed Server—Primary NIC Recovery (Time 1–2 Seconds)

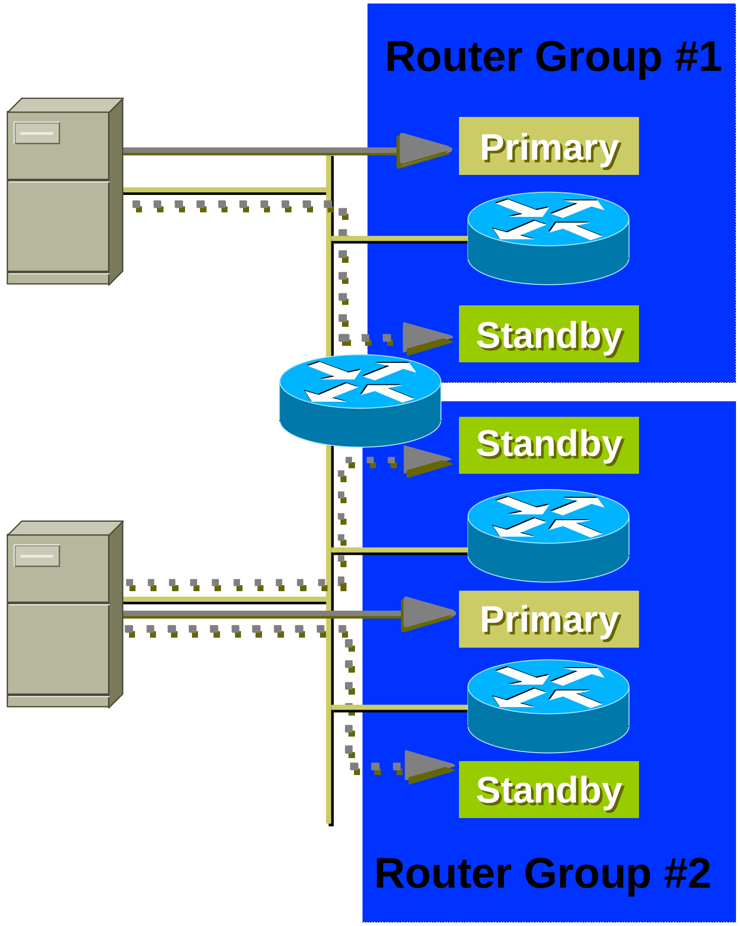
HSRP – Hot Standby Router Protocol



- Transparent failover of default router
- “Phantom” router created
- One router is active, responds to phantom L2 and L3 addresses
- Others monitor and take over phantom addresses

HSRP – RFC 2281

- HSR multicasts hellos every 3 sec with a default priority of 100
- HSR will assume control if it has the highest priority and preempt configured after delay (default=0) seconds
- HSR will deduct 10 from its priority if the tracked interface goes down

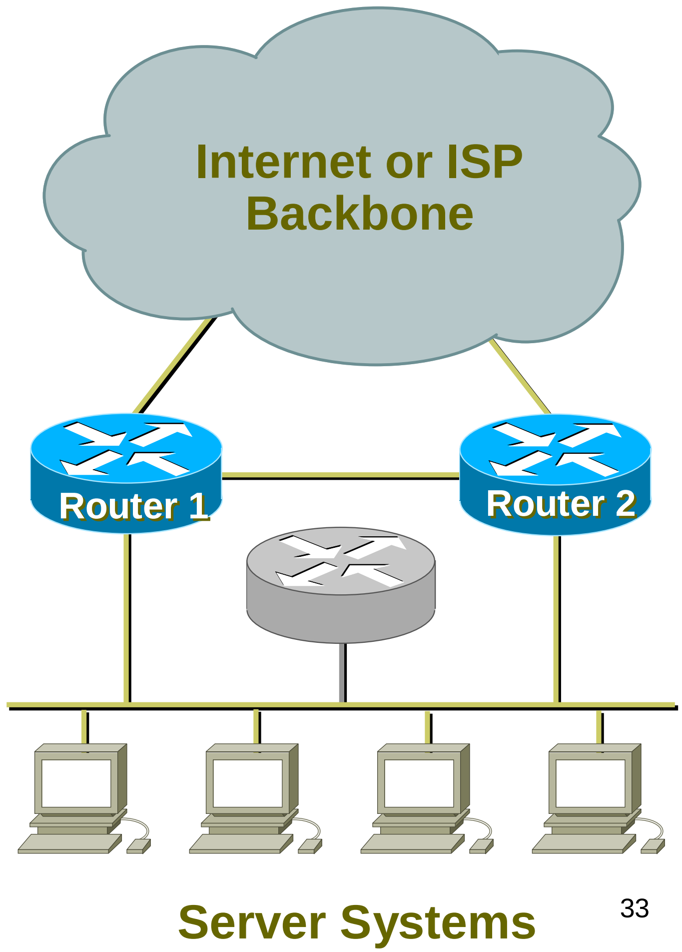


HSRP

Technology

```
Router1:
interface ethernet 0/0
ip address 169.223.10.1 255.255.255.0
standby 10 ip 169.223.10.254


Router2:
interface ethernet 0/0
ip address 169.223.10.2 255.255.255.0
standby 10 priority 150 pre-empt delay 10
standby 10 ip 169.223.10.254
standby 10 track serial 0 60
```



Redundant Network Design

WAN Availability

Design

Circuit Diversity

- Having backup PVCs through the same physical port accomplishes little or nothing
  - Port is more likely to fail than any individual PVC
  - Use separate ports
- Having backup connections on the same router doesn't give router independence
  - Use separate routers
- Use different circuit provider (if available)
  - Problems in one provider network won't mean a problem for your network

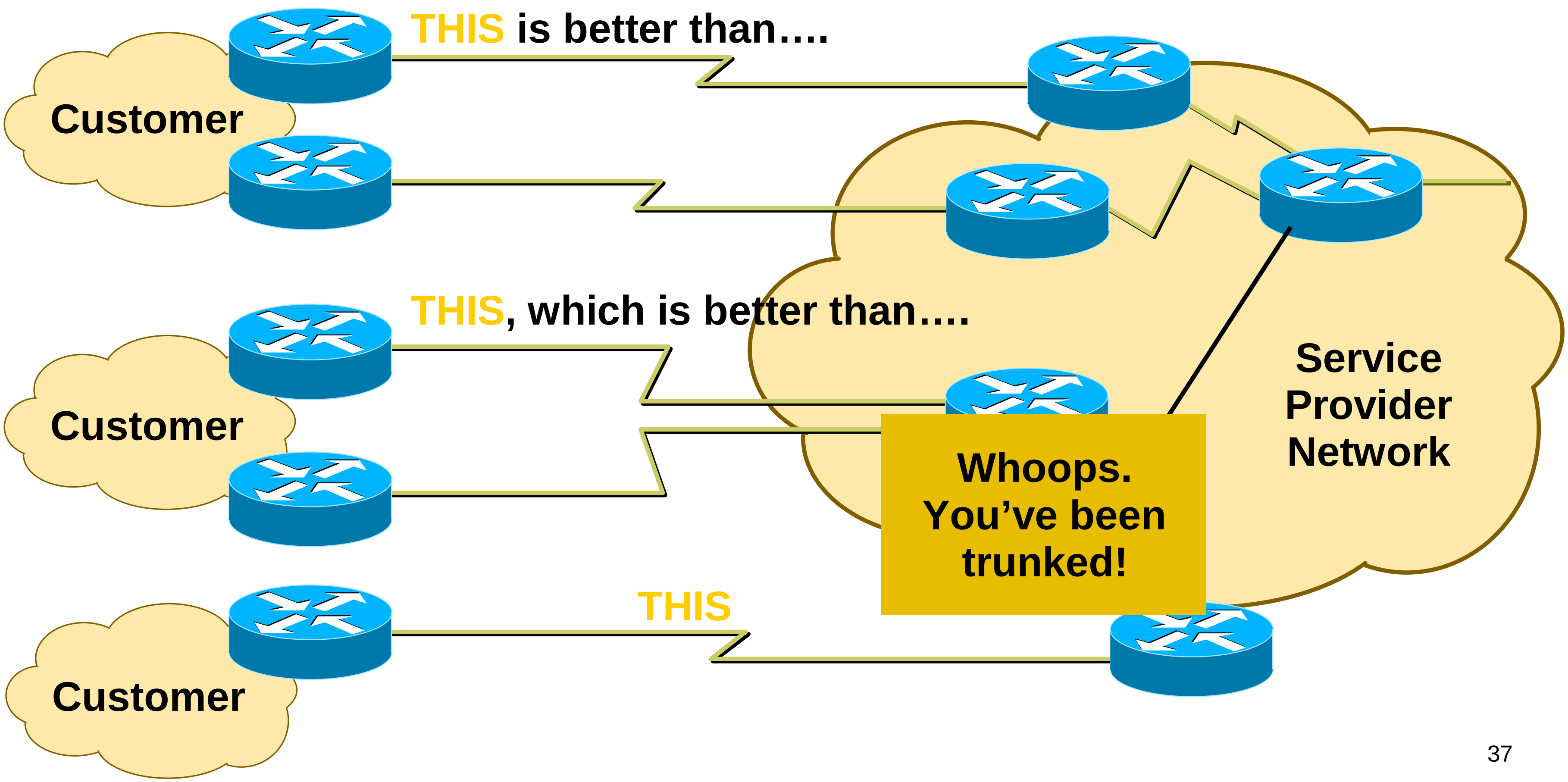
Design

Circuit Diversity

- Ensure that facility has diverse circuit paths to telco provider or providers
- Make sure your backup path terminates into separate equipment at the service provider
- Make sure that your lines are not trunked into the same paths as they traverse the network
- Try and write this into your Service Level Agreement with providers

Circuit Diversity

Technology

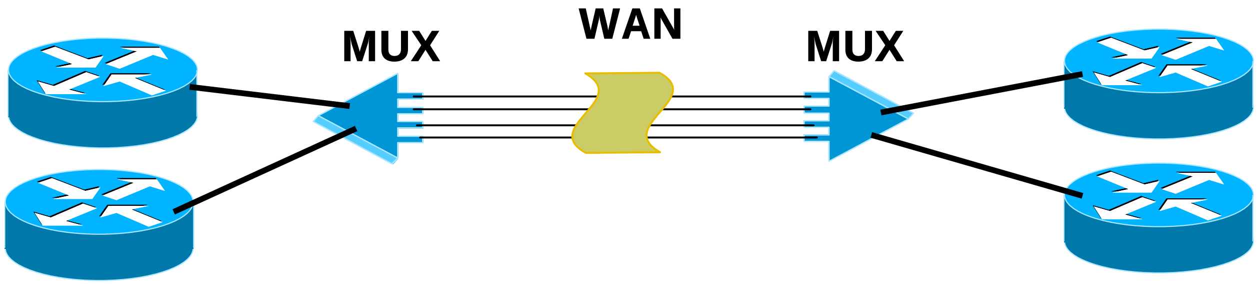


Circuit Bundling – MUX

Technology

- Use hardware MUX
  - Hardware MUXes can bundle multiple circuits, providing L1 redundancy
  - Need a similar MUX on other end of link
  - Router sees circuits as one link
    - Failures are taken care of by the MUX

Using redundant routers helps



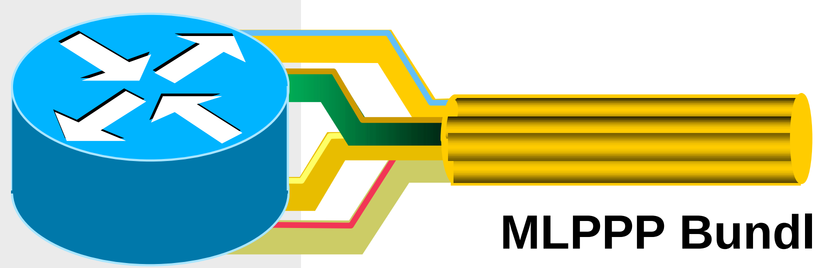
Technology

Circuit Bundling – MLPPP

```
interface Multilink1
 ip address 172.16.11.1 255.255.255.0
 ppp multilink
 multilink-group 1
!
interface Serial1/0
 no ip address
 encapsulation ppp
 ppp multilink
 multilink-group 1
!
interface Serial1/1
 no ip address
 encapsulation ppp
  ppp multilink
 multilink-group 1
```

Multi-link PPP with proper circuit diversity, can provide redundancy.

Router based rather than dedicated hardware MUX



MLPPP Bundle

Design

Load Sharing

- Load sharing occurs when a router has two (or more) equal cost paths to the same destination
- EIGRP also allows unequal-cost load sharing
- Load sharing can be on a per-packet or per-destination basis (default: per-destination)
- Load sharing can be a powerful redundancy technique, since it provides an alternate path should a router/path fail


Technology

Load Sharing

- OSPF will load share on equal-cost paths by default
- EIGRP will load share on equal-cost paths by default, and can be configured to load share on unequal-cost paths:

```
router eigrp 111
 network 10.1.1.0
 variance 2
```

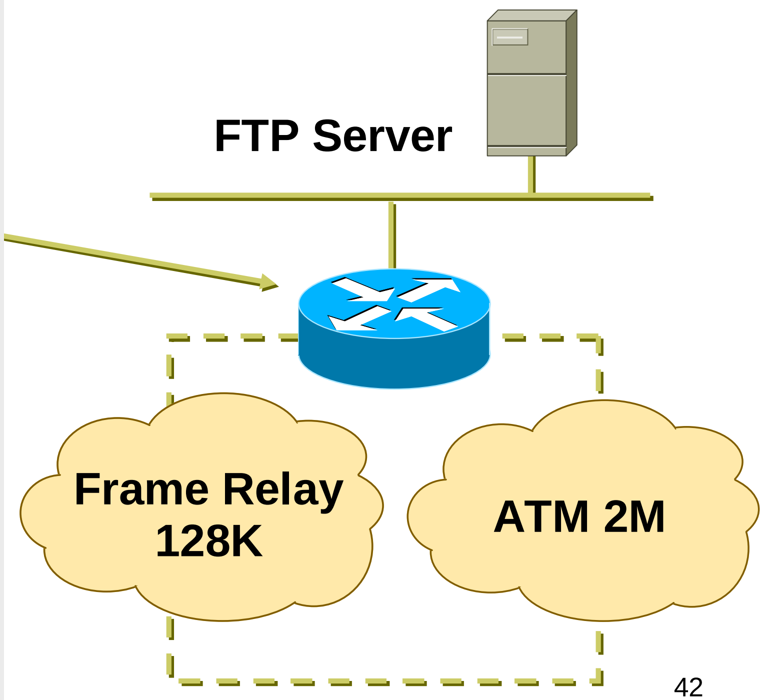
  - Unequal-cost load-sharing is discouraged; Can create too many obscure timing problems and retransmissions

Technology

Policy-based Routing

- If you have unequal cost paths, and you don't want to use unequal-cost load sharing (you don't!), you can use PBR to send lower priority traffic down the slower path

```
! Policy map that directs FTP-Data
! out the Frame Relay port. Could
! use set ip next-hop instead
route-map FTP_POLICY permit 10
   match ip address 6
   set interface Serial1.1
!
! Identify FTP-Data traffic
access-list 6 permit tcp any eq 20 any
!
! Policy maps are applied against
! inbound interfaces
interface ethernet 0
   ip policy route-map FTP_POLICY
```



Convergence

Design

- The convergence time of the routing protocol chosen will affect overall availability of your WAN
- Main area to examine is L2 design impact on L3 efficiency

BFD

- BFD - Bidirectional Forwarding Detection
  - Used to QUICKLY detect local/remote link failure
  - Between 50ms and 300ms
  - Signals upper-layer routing protocols to converge
    - OSPF
    - BGP
    - EIGRP
    - IS-IS
    - HSRP
    - Static routes
  - Especially useful on Ethernet links - where remote failure detection may not be easily identifiable.

IETF Graceful Restart

- Graceful Restart
  - Allows a router's control plane to restart without signaling a failure of the routing protocol to its neighbors.
  - Forwarding continues while switchover to the backup control plane is initiated.
  - Supports several routing protocols
    - OSPF (OSPFv2 & OSPFv3)
    - BGP
    - IS-IS
    - RIP & RIPng
    - PIM-SM
    - LDP
    - RSVP

NSR

- NSR - Non-Stop Routing
  - A little similar to IETF Graceful Restart, but…
  - Rather than depend on neighbors to maintain routing and forwarding state during control plane switchovers…
  - The router maintains 2 identical copies of the routing state on both control planes.
  - Failure of the primary control plane causes forwarding to use the routing table on the backup control plane.
  - Switchover and recovery is independent of neighbor routers, unlike IETF Graceful Restart.

VRRP

- VRRP - Virtual Router Redundancy Protocol
  - Similar to HSRP or GLBP
  - But is an open standard
  - Can be used between multiple router vendors, e.g., between Cisco and Juniper

ISSU

- ISSU - In-Service Software Upgrade
  - Implementation may be unique to each router vendor
  - Basic premise is to modularly upgrade software features and/or components without having to reboot the router
  - Support from vendors still growing, and not supported on all platforms
  - Initial support is on high-end platforms that support either modular or microkernel-based operating systems

MPLS-TE

- MPLS Traffic Engineering
  - Allows for equal-cost load balancing
  - Allows for unequal cost load balancing
- Makes room for MPLS FRR (Fast Reroute)
  - FRR provides SONET-like recovery of 50ms
  - Ideal for so-called “converged” networks carrying voice, video and data

Control Plane QoS

- QoS - Quality of Service (Control Plane)
  - Useful for control plane protection
  - Ensures network congestion do not cause network control traffic drops
  - Keeps routing protocols up and running
  - Guarantees network stability
  - Cisco features:
    - CoPP (Control Plane Policing)
    - CPPr (Control Plane Protection)

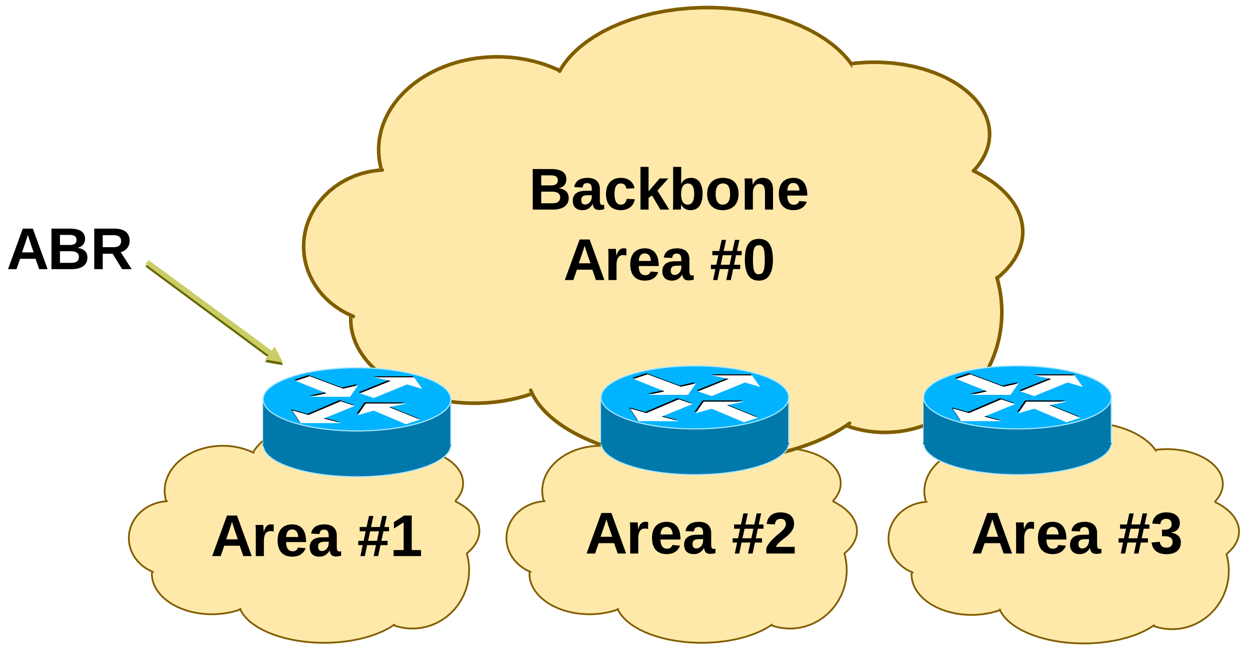
Factors Determining Protocol Convergence

Design

- Network size
- Hop count limitations
- Peering arrangements (edge, core)
- Speed of change detection
- Propagation of change information
- Network design: hierarchy, summarization, redundancy

OSPF – Hierarchical Structure

Design



- Topology of an area is invisible from outside of the area
  - LSA flooding is bounded by area
  - SPF calculation is performed separately for each area

Factors Assisting Protocol Convergence

Design

- Keep number of routing devices in each topology area small (15 – 20 or so)
  - Reduces convergence time required
- Avoid complex meshing between devices in an area
  - Two links are usually all that are necessary
- Keep prefix count in interior routing protocols small
  - Large numbers means longer time to compute shortest path
- Use vendor defaults for routing protocol unless you understand the impact of “twiddling the knobs”
  - Knobs are there to improve performance in certain conditions only

Redundant Network Design

Internet Availability

PoP Design

Design

- One router cannot do it all
- Redundancy redundancy redundancy
- Most successful ISPs build two of everything
- Two smaller devices in place of one larger device:
  - Two routers for one function
  - Two switches for one function
  - Two links for one function

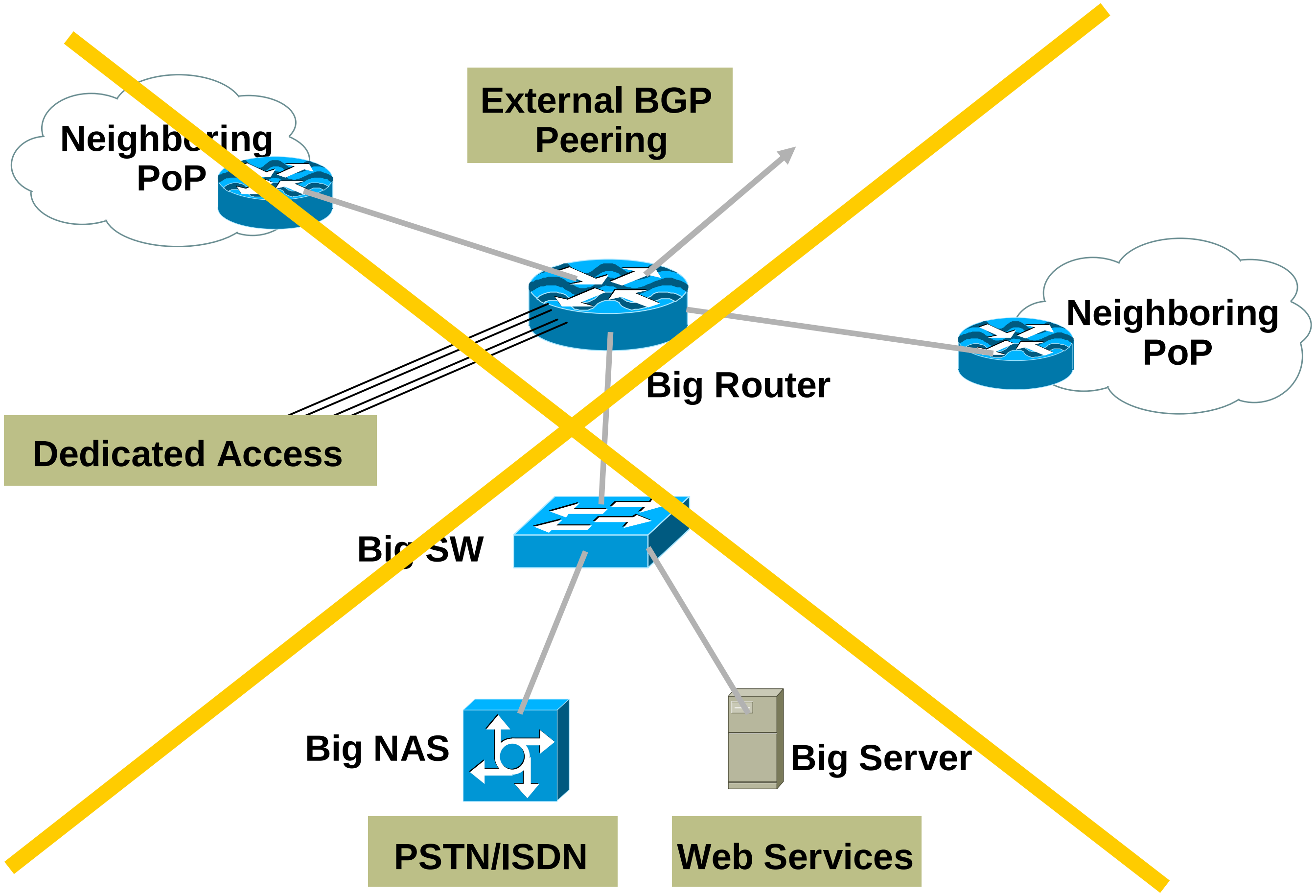
Design

PoP Design

- Two of everything does not mean complexity
- Avoid complex highly meshed network designs
  - Hard to run
  - Hard to debug
  - Hard to scale
  - Usually demonstrate poor performance

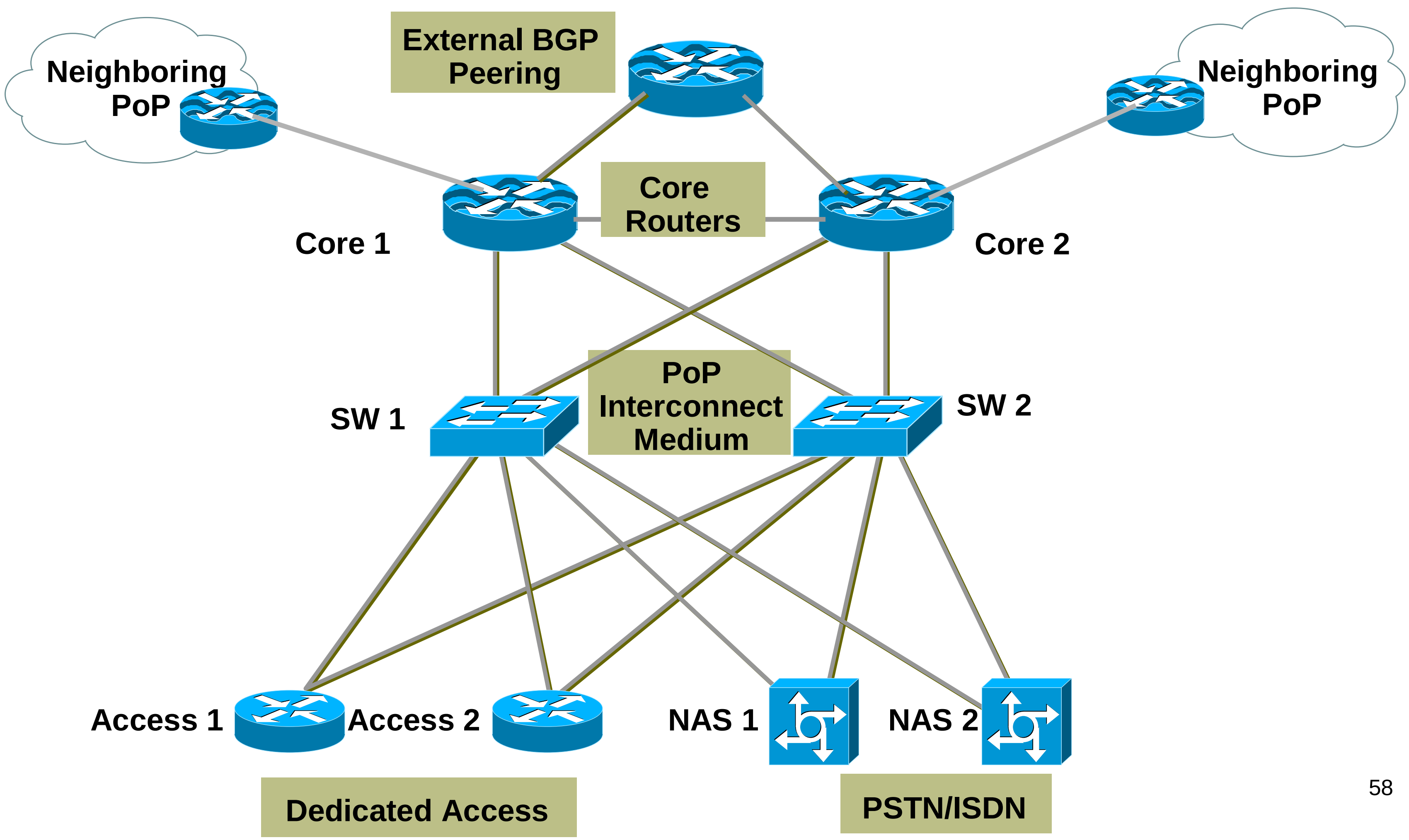
PoP Design – Wrong

Design



PoP Design – Correct

Design



Technology

Hubs vs. Switches

- Hubs
  - These are obsolete
    - Switches cost little more
  - Traffic on hub is visible on all ports
    - It’s really a replacement for coax ethernet
    - Security!?
  - Performance is very low
    - 10Mbps shared between all devices on LAN
    - High traffic from one device impacts all the others
  - Usually non-existent management

Technology

Hubs vs. Switches

- Switches
  - Each port is masked from the other
  - High performance
    - 10/100/1000Mbps per port
    - Traffic load on one port does not impact other ports
  - 10/100/1000 switches are commonplace and cheap
  - Choose non-blocking switches in core
    - Packet doesn’t have to wait for switch
  - Management capability (SNMP via IP, CLI)
  - Redundant power supplies are useful to have

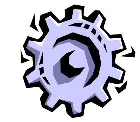
Beware Static IP Dial

Design

- Problems
  - Does NOT scale
  - Customer /32 routes in IGP – IGP won't scale
  - More customers, slower IGP convergence
  - Support becomes expensive
- Solutions
  - Route "Static Dial" customers to same RAS or RAS group behind distribution router
  - Use contiguous address block
  - Make it very expensive – it costs you money to implement and support

Redundant Network Design

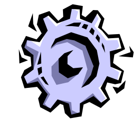
Operations!



Process

Network Operations Centre

- NOC is necessary for a small ISP
  - It may be just a PC called NOC, on UPS, in equipment room.
  - Provides last resort access to the network
  - Captures log information from the network
  - Has remote access from outside
    - Dialup, SSH,…
  - Train staff to operate it
  - Scale up the PC and support as the business grows



Process

Operations

- A NOC is essential for all ISPs
- Operational Procedures are necessary
  - Monitor fixed circuits, access devices, servers
  - If something fails, someone has to be told
- Escalation path is necessary
  - Ignoring a problem won’t help fixing it.
  - Decide on time-to-fix, escalate up reporting chain until someone can fix it

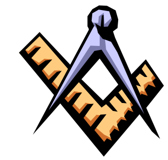
Process

Operations

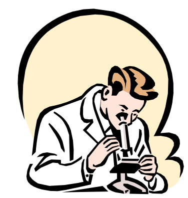
- Modifications to network
  - A well designed network only runs as well as those who operate it
  - Decide and publish maintenance schedules
  - And then STICK TO THEM
  - Don’t make changes outside the maintenance period, no matter how trivial they may appear

In Summary

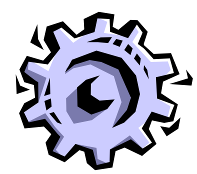
- Implementing a highly resilient IP network requires a combination of the proper process, design and technology
- "and now abideth design, technology and process, these three; but the greatest of these is process"
- And don't forget to KISS!
  - Keep It Simple & Stupid!



Design



Technology



Process

Acknowledgements

- The materials and Illustrations are based on the Cisco Networkers' Presentations
- Philip Smith of Cisco Systems
- Brian Longwe of Inhand .Ke