Deploying Highly Resilient IP Networks

Session IPS-211

“The Janitor Pulled the Plug…”

- Why was he allowed near equipment?
- Why was problem noticed only afterward?
- Why did it take 6 weeks to determine problem?
- Why wasn’t there redundant power?
- Why wasn’t there network redundancy?
What Is High Availability?

When the Proportion of the Total Operating Time a Network Is Functional $>= 99.999%$

<table>
<thead>
<tr>
<th>Availability</th>
<th>DPM</th>
<th>Downtime Per Year (24x7x365)</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.000%</td>
<td>10K</td>
<td>3 Days 15 Hours 36 Minutes</td>
</tr>
<tr>
<td>99.500%</td>
<td>5K</td>
<td>1 Day 19 Hours 48 Minutes</td>
</tr>
<tr>
<td>99.900%</td>
<td>1000</td>
<td>8 Hours 46 Minutes</td>
</tr>
<tr>
<td>99.950%</td>
<td>500</td>
<td>4 Hours 23 Minutes</td>
</tr>
<tr>
<td>99.990%</td>
<td>100</td>
<td>53 Minutes 5 Minutes</td>
</tr>
<tr>
<td>99.999%</td>
<td>10</td>
<td>53 Minutes 5 Minutes</td>
</tr>
<tr>
<td>99.9999%</td>
<td>1</td>
<td>30 Seconds</td>
</tr>
</tbody>
</table>

Five 9's or More

How Do We Get There?

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

Joel Snyder – Network World Test Alliance 1/10/00

“Reliability: Something you build, not buy”
Building Availability

The OSI Model
- Physical
- Data Link
- Network
- Transport
- Session
- Presentation
- Application

IP Services
- Productivity Opportunity
- Investment in Infrastructure and Technology
- Stable Available Applications

Investment Meter
Convenience Meter

IP Infrastructure

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
- All of these elements are necessary, and all interact with each other
The Basics: Platform and Environment

- Redundant power
- Redundant cooling
- 1:1 or N:1 card redundancy
- Redundant route processors
- Redundant switch fabric

- Environmental controls
- Power environment
- Cabling
Platform Redundancy: Cisco 12000 GSR

- NEBS Level 3 compliant
- Automatic protection switching (APS)/Multiplex Section Protection (MSP)
- Hot swap capability
- Redundancy
  - Dual-route processors
  - Switch fabric redundancy
  - Redundant power supplies
  - Redundant cooling systems
  - Line card protection
- Maintenance bus

Effects of the Physical Plant

- MTBF
- Temperature/humidity/electronic noise
- Redundant power
- Color coding
- Device card/chassis access
- Cable labels
- Documentation
- Eliminate “The Janitor Effect”
Data Center Availability

Backbone from Hell

Avoid Highly Meshed, Non-Deterministic Large Scale L2
Typical (Better) Backbone

Client Blocks

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

Why Is Spanning Tree A Problem?

Spanning Tree State Machine

- **Blocking**: Time to Detect that Root Bridge Not Available: 20 secs
- **Listening**: Discarding Frames while Calculating New Root: 15 secs
- **Learning**: Discarding Frames while Learning Addresses: 15 secs
- **Forwarding**: Finally Forwarding Frames. To Reach this State: 50 secs
Accelerating STP Recovery

- In a highly redundant L2 network, spanning tree recalculation can cause significant delays
- Two Cisco technologies can help alleviate this delay:
  - Ether Channel for fast or gigabit ethernet
  - Spanning tree optimizations: UplinkFast, PortFast and BackboneFast

EtherChannel®

- Minimizes risk of link failure leading to spanning tree reconfiguration
- Increased availability
- Sub second recovery
- Single L2 STP link
- Single L3 subnet

Supported on the Catalyst Family as Well as Cisco IOS
**Configuring EtherChannel**

**On a Catalyst 6XXX:**

```
Console> (enable) set port channel 2/2-8 mode desirable
Ports 2/2-8 left admin_group 1.
Ports 2/2-8 joined admin_group 2.
Console> (enable)
```

**On a Cisco 7500:**

```
Router(config)# interface port-channel 1
Router(config)# ip address 10.0.0.1 255.255.255.0
Router(config)# ip route-cache distributed
Router(config)# interface fasteth 0/0
Router(config)# no ip address
Router(config)# channel-group 1
Router(config)# interface fasteth 0/1
Router(config)# no ip address
Router(config)# channel-group 1
FastEthernet 0/1 added as member-2 to fechannell
```

**Spanning Tree Optimizations**

**PortFast:** Allows a port on a switch to which an end station is attached to proceed directly to forwarding state

```
set spantree portfast 1/1 enable
set spantree portfast bpdu-guard enable
```

**UplinkFast:** Allows a wiring closet switch to transition a backup link directly into forwarding state

```
set spantree uplinkfast enable rate 40
```

**BackboneFast:** Allows a distribution or core switch to proactively seek out a new path to the STP root bridge

```
set spantree backbonefast enable
show spantree backbonefast
```
Reduce Unnecessary Peering

- Problem: Routers peer across each wiring closet VLAN
- Passive interfaces reduce peering overhead

4 VLANs per Wiring Closet
16 VLANs Total
16 Routed Paths between X and Y
Routing Overhead *16
Impacts Convergence Time

Solution: Make Wiring Closet VLAN Interfaces Passive on Routers X and Y

X

To Backbone

Y

Using Passive Interfaces

- Using passive interfaces at the distribution layer:
  - Reduces unnecessary peering
  - Speeds dykstra calculations
  - Speeds L3 convergence

interface Vlan31
  description Link to backbone
  ip address 10.31.0.81 255.255.0.0
  no ip directed-broadcast
  ip hello-interval eigrp 1 1
  ip hold-time eigrp 1 3
  !
  router eigrp 1
  passive-interface Vlan10
  passive-interface Vlan11
  passive-interface Vlan12
  passive-interface Vlan13
  passive-interface Vlan99
  network 10.0.0.0
Layer 3 Backbone

- Client Block
- Access L2
- Distribution L3
- Core L3
- Server Farm Block
- Distribution L3
- Access L2

Benefits of a L3 Backbone

- Multicast PIM routing control
- Load balancing
- No blocked links
- Fast convergence EIGRP/OSPF
- Greater scalability overall
- Router peering reduced
- Cisco IOS features in the backbone
Server Availability

- Campus
- Service Provider
- Internet Access
- Data Center
- Wide Area
- Telecommuters
- ISDN
- Remote Offices
- Partners/Extranet
- Customers/Electronic Commerce
- WAN Offices
- WAN Core

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

Core/Backbone
- Catalyst L3 Switch=Router

Catalyst L2 Switch
- Server Farm

Dual-homed Server—Primary NIC Recovery (Time 1–2 Seconds)
Redundant IP Servers Using Server Load Balancing

User requesting 10.1.1.1 gets directed to one of several identical DNS servers. Eliminates the server as a single point of failure. Good backup strategy for TCP/UDP based servers.

SLB—Flexible Configurations

- Monitor the status of a server in a server farm and take it out of service
- Load-balancing between servers in a server farm
- Firewall load-balancing
- Redundant local directors
- Sticky connections
- Web Cache balancing
- WAP gateway balancing
- SYNGuard against DoS attacks
- Private and public servers
- NAT session redirection
- DNS, FTP, HTTP, HTTPS, IMAP, MATIP-A, NNTP, POP2/3, RealAudio/Video via HTTP, RADIUS, SMTP, Telnet, XOT
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

Router1:
interface ethernet 0/0
bandwidth 128
ip address 169.223.10.1 255.255.255.0
standby 10 ip 169.223.10.254

Router2:
interface ethernet 0/0
bandwidth 1500
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

Putting All Of It Together

Resiliency Example

Local Director removes failed server from SLB (L4-7) (5 sec.)

STP forwards quickly after primary uplink failure with UplinkFast (L2) (2 sec.)

HSRP moves gateway to alternative gateway (L3) (2 sec.)

Distributed Director routes app request to alternative site if server or access fails (L3) (5 sec.)

PortFast for immediate STP forwarding for end stations (L2) (1 sec.)

Stateful fail over of SLB with serial link (L4-7) (2 sec.)

HSRP-Track moves gateway only when both uplinks fail

Routing protocol tuning allows for quick re-route (L3) (4 sec.)

Technologies like DPT allow for very fast IP reconvergence (L2-3) (2 sec.)
Campus Availability

- Redundancy on the campus MAN varies depending on technology used
- Main choices are usually one of:
  - SONET/SDH
  - DPT/SRP
  - SRP or SONET/SDH over WDM
  - ATM
  - Gigabit Ethernet
### SONET/SDH APS

- Available on all SONET/SDH line cards
- Telcordia GR-253/ITU standards compliant
- K1/K2 link-layer control
- 3 levels of protection
  - Port failure
  - Line card failure
  - Router failure

### APS: Automatic Protection Switching

Provides Automatic Failover Protection for SONET/SDH Lines

Across a Single Card

Across Two Cards in the Same Chassis

Even across Two Separate Chassis
**APS: Automatic Protection Switching**

- Provides Protection for SONET/SDH Lines
  - Across a Single Card
  - Across 1 Cards in the Same Chassis
  - Even across Two Separate Chassis

- **APS**
  - **ADM**
  - **Active Ring**
  - **Protection Ring**

**Design**

**Dynamic Packet Transport**

- No protection bandwidth as in SONET/SDH—uses ring wrap
- Bandwidth consumed only on traversed—destination stripping
- Multiple nodes transmit concurrently
- Fairness via SRP-fa
- Self-healing through IPS
- Multicast and prioritization support
Intelligent Protection Switching

- Like SONET/SDH, DPT provides
  Proactive performance monitor and 50 ms self-healing via ring wrapping
- Unlike SONET/SDH, DPT provides
  Signaling via explicit control messages
  Multi-layer awareness and elastic cooperation
  Differentiated handling by priority
  Ring subnet, rather than point-to-point
  Fast IP service restoration on large rings

WAN Availability

- Campus
- Campus WAN
- Service Provider
- WAN Core
- Wide Area
- ISDN
- Internet Access
- ISP POP
- Remote Offices
- VPNs
- Customers/Electronic Commerce
- Partners/Extranet
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, preferably on separate routers
- Try to have it written into your SLA that your backup path terminates into separate equipment at the service provider, and that your lines are not trunked into the same paths as they traverse the network

Circuit Diversity

THIS is better than....

THIS, which is better than....

Whoops. You’ve been trunked!
Using 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, if employed with proper circuit diversity, can provide redundancy to TDM lines. Has the value-added effect of increasing your bandwidth.

MLPPP Bundle

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 fail
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

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
```

FTP Server

Frame Relay 128K

ATM OC-3
BGP Multi-path Load Sharing

- If two identical eBGP paths are learned from the same neighboring AS, and maximum-paths is greater than 1, install up to maximum-paths into the routing table.
- Otherwise, use the lowest router-id to break the tie, and install just a single route.
- Up to 6 maximum-paths routes are permitted.

```
router bgp 109
  network 131.108.0.0
  network 192.31.7.0
  neighbor 131.108.200.1 remote-as 167
  maximum-paths 3
```

Convergence

- 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.
- For detailed comparisons of OSPF, EIGRP:
  - RST-207 deploying OSPF
  - RST-209 deploying EIGRP
Factors Determining Protocol Convergence

- 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

- 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
Methods to Improve OSPF Scaling

- Route summarization
  
  On ABR:
  ```
  router ospf 100
  area 1 range 128.213.64.0 255.255.224.0
  ```

  On ASBR:
  ```
  router ospf 100
  summary-address 128.213.96.0 255.255.224.0
  ```

- Stub or not so stubby areas (NSSA)

OSPF Scaling

OSPF Area 0.0.0.0
Backbone
2500 Routes at ABR

OSPF Area 0.0.0.4
100 Routers in Area Transit

ABR

Frame Relay

2500 Routes/
~46 Routes per Summary LSA
~54 1500-Byte Packets Required*
100 Routers
5400 Packets*
1500 Bytes
810000 Bytes*
8 Bits =
~64.8 Mbps Bandwidth Requirement
OSPF Scaling—Stub Areas

- 52 Bytes Per Default-Route Packet
  - 100 Routers
  - 5200 Bytes
  - 8 Bits = ~41.6 Kbps Bandwidth Requirement

```
router ospf 1
network 10.3.1.0 area 0
network 10.3.2.0 area 4
area 4 stub no-summary
```

EIGRP—Convergence

- EIGRP will track successor routes (alternate routes to be used in case of failure)
- If a successor route is found, switchover occurs with no interaction with other routers; immediate convergence
- If no feasible successor routes exist, router sends a query to neighbors to find a route
- Because queries can stretch to the very edge of the network, makes sense to limit their scope
Decreasing Query Scope

- Summarization (manual or auto)
- Distribute-lists
  Particularly on dual-homed remotes
- Stub routers (release 12.0S)
  Signals HUB router not to send queries

Decreasing Query Scope with Summarization

Queries
Replies

10.1.8.0/24

No Knowledge of 10.1.8.0/24

IP summary-address eigrp 1 10.0.0.0 255.0.0.0 on Outbound Interfaces toward Remote Routers
map-class frame-relay vcgrp router1
frame-relay end-to-end keepalive mode bidirectional

-- Dial Backup Technology

!--- password for other router defined
username ROUTER2 password 0 letmein
!
interface Serial0.1
 ip address 172.16.2.1 255.255.255.128
 frame-relay interface-dlci 102
!
interface BRI0
description ISDN for frame relay
 ip address 172.16.3.1 255.255.255.0
 encapsulation ppp
dialer map ip 172.16.3.2 name ROUTER2 broadcast 6234020
dialer-group 1
!--- Floating static route defined below
 ip route 172.16.4.0 255.255.255.0 172.16.3.2 200
!
!--- Dialer list defines interesting traffic
dialer-list 1 protocol ip list 101
!
VPN Backups

- If WAN and internet access are separate, the internet can be leveraged as a backup path.
- Since a tunnel will traverse the internet, some level of encryption is needed.
- Many variations: L2TP, GRE, IPSec.
- If you want to encrypt (you do...), need encryption—enabled version of IOS.

Site-to-Site VPN Example

The Internet

San Jose

Bangalore

interface Tunnel0
ip address 172.16.12.2 255.255.255.0

tunnel source 128.107.241.185

tunnel destination 192.135.250.69

crypto-map gre

interface Tunnel0
ip address 172.16.12.1 255.255.255.0

tunnel source 192.135.250.69

tunnel destination 128.107.241.185

crypto-map gre
Adding Encryption

crypto isakmp policy 10
authentication pre-share
! crypto isakmp key cisco123 address 192.135.250.69
! crypto ipsec transform-set one esp-des esp-md5-hmac
mode transport
! crypto map gre 10 ipsec-isakmp
set peer 192.135.250.69
set transform-set one
match address grel
! ip access-list extended grel
permit gre host 192.135.250.69 host 128.107.241.185

Multi-chassis MLPPP

To Internal Network

ISDN PRIs

sgbp group stackq
sgbp member systemb 1.1.1.2
sgbp member systems 1.1.1.3
! username stackgroup password cisco
! isdn switch-type primary-4ess
controller t1 0
framing esf
linecode b8zs
pri-group timeslots 1-23
! interface Serial0:23
ip unnumbered e0
dialer map .....
capsulation ppp
ppp authentication chap
dialer-group 1
dialer rot 1
! ppp multilink
Many of the Same Rules Apply

- Routing to the internet is not significantly different than routing to any other WAN site
- Ensure proper circuit diversity
- Protect dual paths with HSRP and track interface
- Optimize routing via load-sharing and fast convergence
Do I Need BGP?

Questions to Ask:

- Do I have more than one co-located paths to the Internet
- And
- Do I, for cost, security or other administrative reasons, need to selectively route (or have traffic routed to me) over one path rather than the other?

When Not to Use BGP

- When you have a single path to the Internet. Use a static default route instead
- When you have two paths to the internet but you don’t care which way your traffic goes. Use two default routes (and possibly load-balance)
- "My ISP says I need to use BGP so he can get the routes from my AS"
- Run BGP, send your routes, but request that the ISP send you no routes. Use a default route

BGP Route Refresh Capability RFC2918

- Facilitates non-disruptive policy changes—i.e. don’t have to clear bgp session
- No configuration is needed
- No additional memory is used
- Clear ip bgp x.x.x.x in | out
- "in" => send route-refresh (new BGP message—type 5) request to neighbor
- "out" => withdraw and resend all routes to peer, via new policy
Operational Excellence

User Error/Process

- Change management
- Process consistency
- Capacity management
- Configuration management
- Network security
- Change management

Source: Gartner Group
Process Goals

- Consistent speedy repair times
  - Configuration management
  - Fault management
  - Performance management
- Quality improvement
  - Availability metrics
  - Fault management metrics
  - Root-cause analysis
  - Performance indicators

Process Goals

- Problem avoidance
  - Network design and resiliency
  - Security
  - Proactive fault management
  - Capacity and performance management
- Successful network evolution
  - Change planning management
  - Testing and validation
Network Life-Cycle Management

- Planning/Design
- Implementation
- Operation
- Fault Management
- Performance and Capacity Management
- Security Management
- Configuration Management

New Solution Planning and Deployment

New Solution Deployment

- Design review with vendor
- Test plan (to reflect your app/network scenarios)
- Lab validation
- Solution pilot
- Solution templates
- Staffing
- Training
- Operational support handoff
Configuration Management

- Maintaining configuration consistency
- Inventory management
- IP address management
- Software version control
- Password management
- Wiring and naming conventions
- Documentation

Configuration Management

- Change Management
  - Change management procedures
  - Risk analysis
  - Testing/validation for high risk change
  - Backout plan
  - Network management and documentation update
  - Change management metrics
Performance and Capacity Management

• Base-lining
• What-if analysis (network and application)
• QoS management
• Periodic review plan and upgrade criteria
• Exception management

Fault Management

• Fault management
  24 x 7 detection, notification, escalation, resolution for link/hardware/network failures
  Proactive fault analysis plan (MIB variables, threshold violations, syslog events, review plan)
  Infrastructure (TFTP, syslog, NTP, time-stamps, out-of-band management, vendor access)
  Help desk systems (metrics, accountability)
Security Management

- Internet access
- Dial-in access
- Partner access
- Security operations
- Internet/partner monitoring
- CERT/vendor advisory review
- Security configuration practices
- Termination practices

Process Tools

- CiscoWorks Resource Manager Essentials
  - Configuration/Change/Inventory Management

- CiscoWorks Device Fault Manager
  - Fault Mgmt

- CiscoWorks Secure Policy Manager
  - Security Mgmt

- Cisco NetFlow Accounting
  - Accounting and Capacity Mgmt

- CiscoWorks Internetwork Performance Monitor
  - Performance Mgmt

- Interfaces to Oracle/Remedy
  - Problem Mgmt
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”
Deploying Highly Resilient IP Networks

Session IPS-211

Please Complete Your Evaluation Form

Session IPS-211