Datacenter Rack Switch
Redundancy Models

Server Access Ethernet Switch
Connectivity Options

NANOG46
June 16, 2009
Dani Roisman
droisman ~ at ~ peakwebconsulting ~ dot ~ com
Introductions
Server Access Explained

• Enterprise and Content datacenters are home to racks full of servers, providing connectivity via Ethernet

• Typically, an Ethernet switch is located within the rack to aggregate all access connectivity for local servers, then uplinked to a network core

• There are multiple ways in which to provide redundant connectivity to this “top-of-rack” switch, each with their own benefits and drawbacks
Who May Be Interested?

- Network operators responsible for server connectivity to multiple racks – these design options scale from one rack switch to hundreds of rack switches in a datacenter
- Folks whose jobs rely upon undisrupted connectivity to servers throughout their datacenter
- Organizations battling with network redundancy versus stability trade-offs
Reference: Sample Network

Distribution layer is made up of Layer 3 switches.

If small enough, Distribution + Core may be collapsed into same chassis.

Distribution Switch #1

Core / Distribution

Distribution Switch #2

Upstream

Core Network

Server Rack

All servers connect to top-of-rack switch, which in turn has upstream connections to rest of network.

Rack Switch #1

Server(s)
Basic Definitions

- "Layer 1" - physical cabling infrastructure, physical link, patch panels, Ethernet cords etc.
- "Layer 2" - switching and bridging, VLANs etc.
- "Layer 3" - IP routing, using static or dynamic protocols such as RIP, EIGRP, OSPF, BGP, etc.
- "Distribution" - provides upstream path to rest of the network (sometimes collapsed with Core)
- "Rack Switch" a.k.a. Top-of-Rack Switch - all servers within a rack connect to this switch, and this switch in turn has uplinks to the distribution equipment (often mounted at middle of rack)
Model 0
No Redundancy
Model 0.5
Manual Activation
Manual Activation

Backup port is either physically disconnected or administratively disabled.

One NIC per server connected to switch.

Server Rack

Upstream Core Network

Core / Distribution

Distribution Switch #1

Distribution Switch #2

Server(s)

Rack Switch #1

NANOG 46 Dani Roisman
droisman~at~peakwebconsulting~dot~com
Manual Activation Explained

• Pretty simple – configure typical L2 redundant environment, say “I’m afraid of L2 loops,” and then shut down one of the uplinks
• In response to an primary link outage, requires manual intervention to enable the backup port
• Do yourself a favor, shut down at the distribution side
• Do yourself another favor, shut down the primary before enabling the backup
• Do yourself a third favor, stay away from this model
Model 1

“The Cliché” Layer2 + STP
“The Cliché” Layer2 + STP

STP Priority 8K (8192) (primary root)
STP Priority 16K (16384) (secondary root)

Intentional network loop

STP blocks towards higher cost to root

One NIC per server connected to switch
L2 + STP Explained

• Network loops are intentionally created to provide redundant paths
• Utilizes spanning-tree protocol, 802.1d or preferably RSTP 802.1w, may utilize MST (802.1s) depending on equipment vendor and network size
• STP will automatically detect and block loops during normal conditions, and will unblock to provide failover during an outage
• Relies upon receipt and processing of BPDUs to decide where to block loops
L2 + STP Redundancy

STP Priority 8K (8192) (primary root)
STP Priority 16K (16384) (secondary root)

Core / Distribution

Server Rack

Primary link is severed
Rack switch set to STP priority 40K (40960)

STP unblocks and service continues
One NIC per server connected to switch

STP unblocks and service continues
L2 + STP Benefits

- Extremely common model, simple to use and understand
- Easy to verify backup connectivity and available redundancy (via STP blocking state)
- Most flexible option, any subnet can be extended to any server in any rack via VLAN tagging, easy server mobility (e.g. VM)
- Allows for centralized “services” to be deployed in transparent mode (SLB, firewalls, etc.)
- Helps conserve public IPv4 addresses
L2 + STP Drawbacks

- Stability varies based vendor and device model
- Limitations to interoperations in a multi-vendor environment
- Configuration errors, cabling errors, hardware failure can cause entire datacenter shut downs
- Some network troubles may be difficult to trace
- Heavily depends on rack switch CPU health
- Often requires extra tweaks to achieve desired performance / stability
- May not be able to utilize backup capacity
Model 2
Full Layer 3
Full Layer 3

Both Distribution devices learn subnet via IGP from rack switch

Upstream
Core
Network

Dotted line represents Routing Link

Rack switch acts as servers’ default gateway

One NIC per server connected to switch

All servers in this rack are exclusively numbered, multiple subnets are possible

NANOG 46 Dani Roisman
droisman~at~peakwebconsulting~dot~com
Full L3 Explained

- Distribution ⇔ rack switch ports are configured as Layer 3 on both sides
- Distribution ⇔ rack switch links are point-to-point routing links
- Default gateway for servers is rack switch
- Routing protocol is run between rack switch and distribution (typically IGP such as OSPF)
- Redundancy is based on standard routing protocols, e.g. link state, metric/cost
Full L3 Benefits

- Simple to use and understand
- No L2 loops, no STP or associated risks
- Works very well for managed hosting environments with many small customer subnets
- Extremely robust, configuration errors will impact only single rack in worst case
- Easy to verify backup connectivity and available redundancy (via routing adjacencies)
- Supports multi-path to take advantage of all capacity
Full L3 Drawbacks

- Significantly increases cost of rack switches
- Racks must be *exclusively* numbered, which means unique subnet(s) per rack
- Inefficient use of public IPv4 address space, one subnet must be allocated per rack
  \[ /28 = 13 \text{ servers}, \ /27 = 29 \text{ servers}, \ /26 = 61 \text{ servers} \]
- Relocating a server (or VM) to another rack means renumbering, limits NIC teaming
- Additional “services” must run in Layer3 mode (e.g. SLB, firewall), may require fancy routing tricks, PBR, etc.
Model 3
Layer 3 Dist with Layer 2 RS
Layer 3 Dist with Layer 2 RS

Both Distribution devices announce subnet via IGP.

Redundancy protocol for servers’ default gateway

HSRP or VRRP

10.10.10.1/24

10.10.10.2 10.10.10.3

All servers in this rack are exclusively numbered in 10.10.10.0/24

Can use inexpensive / unmanaged* Layer 2 rack switch

*R* managed always recommended

One NIC per server connected to switch
L3 Dist with L2 RS Explained

- Distribution ports are configured as Layer 3 (a.k.a. "routed ports")
- Default gateway for servers on distribution using a redundancy protocol such as HSRP or VRRP
- Layer 2 adjacency is formed through the rack switch (a.k.a. “V-shaped”)
- Distribution switches announce reachability via IGP, as with standard L2 model
- Multiple subnets available via 802.1q tagged sub-interfaces from distribution to rack switch
Layer 3 with Layer 2 RS Redundancy

- Distribution switch #1 withdraws subnet from IGP.
- Only Distribution switch #2 announces subnet via IGP.
- Redundancy protocol fails over.
- One link is severed.
- All servers in this rack are exclusively numbered in 10.10.10.0/24.
- One NIC per server connected to switch.
L3 Dist with L2 RS Benefits

• STP elimination as with full L3 model
• Lower cost than full L3 model, since rack switches are L2 only
• In fact, can run ultra-cheap commodity switches, e.g. not even manageable
L3 Dist with L2 Drawbacks

• As with full L3 model, requires *exclusive* numbering, which means subnet can only exist on that one rack, as well as public IPv4 inefficiencies
• Cannot use upstream backup link capacity without additional configurations (e.g. multiple HSRP/VRRP groups w/alternating priorities and alternating default gateways on servers)
• May impose multi-netting if multiple subnets are required and 802.1q sub-if is not available
• Not very common or straightforward
Model 4
Link Failover
Link Failover

Upstream Core Network

Core / Distribution

Distribution Switch #1

Distribution Switch #2

Server Rack

Identify primary port and associate with backup port

Uplink failover feature disables traffic on (or shuts) designated “backup” port as long as primary port is operational

One NIC per server connected to switch

Server(s)
Link Failover Explained

• This model uses switches connected similarly to the L2 model, but with an L1 redundancy
• Instead of creating a loop for backup and blocking using STP, the RS automatically disables all forwarding on (or shuts down) the backup link
• Many vendors support this:
  – Cisco calls this “Flex Link”
  – Foundry calls this “Protected Link Groups”
  – Juniper (EX) calls this “Redundant Trunk Links”
  – Force10 calls this “Redundant Pairs”
Link Failover Redundancy

- **Upstream Core Network**
- **Distribution Switch #1**
- **Distribution Switch #2**

**Core / Distribution**

**Server Rack**

- **Primary link is severed**
- **Uplink failover feature enables traffic on designated “backup” port**

- **One NIC per server connected to switch**

- **Designation of Switch #1**
- **Server(s)**
Link Failover Benefits

- Not based on health of distribution and rack switch CPUs, may be more reliable
- No complex “state machines” as with STP
- No loops during steady-state, backup link never passes production traffic while primary is up
- May provide rapid failover (depends on certain conditions)
- Reduced interoperability concerns, if the RS provides this feature, there is nothing needed from the distribution equipment
- Same server mobility & subnet flexibility as L2
Link Failover Drawbacks

• Some vendors place the backup port in a “down” state:
  – link is always shown as “down” therefore cannot confirm interface health or guarantee redundancy without actually testing failover
  – cannot map backup interfaces using LLDP/CDP/FDP
  – may cause delays in failover due to STP discovery as link in distribution needs to transition to forwarding

• Lack of protocol (STP / IGP) limits knowledge, redundancy may not be complete – surprise!

• Cannot load-balance traffic over redundant link

• STP usually disabled, may cause loops!
Link Failover Questions

• Is anyone using this?
• Really though – using it widely (all racks in a datacenter)?
• Why not?
• Sure, it’s not as flexible as STP, but doesn’t it provide exactly what we need at the dual-uplinked rack switch?
Model 5
Layer 2 Dist with Layer 3 RS
Layer 2 Dist with Layer 3 RS

Distribution devices are two separate switch fabrics, RSs form IGP adjacencies with other RSs and Cores.

Can use inexpensive / unmanaged* Distribution switch.

*Managed always recommended.

Server Rack

Rack switch acts as servers’ default gateway.

All servers in this rack are exclusively numbered, multiple subnets are possible.

One NIC per server connected to switch.
L2 Dist With L3 RS Explained

- Each distribution switch acts as a switch fabric for a unique multi-access routing subnet, dumb switch with no routing functionality
- Layer 3 RSs form IGP adjacencies with all other RSs and Layer 3 cores over two diverse subnets
- RS uplink ports are configured as Layer 3 ports
- Default gateway for servers is rack switch
- Redundancy is based on standard routing protocols, e.g. link state, metric/cost
Expanded L2 Dist with L3 RS

- ISPs etc.
- Layer 3 Router #1
- Layer 3 Router #2
- L2 Dist Switch #1
- L2 Dist Switch #2
- L2 Dist Switch #3
- L2 Dist Switch #4
- Core / Border
- Dist #1 Routing – 10.100.1.0/24
- Dist #2 Routing – 10.100.2.0/24
- 192.168.1.0/24
- 192.168.2.0/24
- 192.168.3.0/24
- 192.168.4.0/24
L2 Dist With L3 RS Benefits

• All the benefits of full L3, such as no loops or STP, ability to utilize redundant uplinks for additional capacity during steady-state

• If application/product requirements require Layer3 access devices, will allow for some economy in the distribution hardware (don’t even need VLAN support in the L2 distribution)

• Introduces a novel any-to-any routing over L2 switching fabric model into the datacenter
L2 Dist With L3 RS Drawbacks

- As with full L3 model, more expensive RSs
- As with full L3 model, requires *exclusive* numbering, which means subnet can only exist on that one rack, as well as public IPv4 inefficiencies
- Since uplinks are not point-to-point, far-end outages are only detected after hold or keepalive timer expiration, may experience short-term blackhole during link failover
- Not very common or straightforward
Model 6
Server-Based (Multi NIC)
Server-Based (Multi NIC)

- **Upstream Core Network**
- **Core / Distribution**
  - No loops or redundancy protocols in Distribution
  - Two switches per rack, singly connected to diverse Distribution
  - Two NICs per server connected to diverse switches, using “teaming” or “bonding” *
  - *recommended active-backup mode
Server-Based Explained

• This model assigns redundancy burden to the servers
• Supply two rack switches for server access, each with diverse uplinks
• May be the best bet when using blade chassis with integrated switches
• Different methods available for servers to determine NIC usability:
  – Link state
  – ARP queries for default gateway
  – Ping default gateway
Server-Based Redundancy 1

**Core / Distribution**
- Distribution Switch #1
- Distribution Switch #2
- No loops or redundancy protocols in Distribution

**Server Rack**
- Two switches per rack, singly connected to diverse Distribution
- Primary link is severed on server detects primary NIC failure and transitions service to backup NIC

**Upstream Core Network**
Server-Based Redundancy 2

No loops or redundancy protocols in Distribution

Uplink becomes severed

Should have method to shut downlink ports

NIC redundancy protocol on server detects primary NIC failure and transitions service to backup NIC
Server-Based Benefits

• Will work with pretty much any switch – if your budget is tight you can shift some $$ from the rack access layer back to the dist / core / border layer where you need it more

• Same IP/subnet/VLAN flexibility as L2 models

• Network requires minimal attention, good for organizations that have more sysadmin skill, requiring less network engineering

• Network has less risk of Layer 2 or Layer 3 troubles, no worry of STP / IGP meltdowns due to misconfiguration or operator error
Server-Based Drawbacks

• Increases complexity of server configuration
• No good centralized way of assuring 100% redundancy (have to check at each server)
• Uses twice the number of NICs on each server
• Uses twice the number of RSs (but, you may want this anyway, to eliminate RS SPOF)
• Should have way to signal server “downlink” ports of RS “uplink” failure, otherwise may blackhole server traffic
• Redundant capacity is not utilized (if configured as recommended)
Extra Redundancy Helpers

- Most failover schemes rely upon discovery of link status changes (port up / down), which sometimes may be delayed or not properly reported
- Remote fault signaling (link negotiation)
- Unidirectional Dead Link Detection (UDLD) is useful especially with fiber links, but beware buggy implementations
- Bidirectional Forwarding Detection (BFD) is more promising than UDLD, maybe even performed in port ASIC, but stable/reliable implementations are still few
Extra Models

- There are some new models surfacing that are based on link aggregation at the rack server side, running to diverse equipment in the distribution layer:
  - Link aggregation across different members of a switch stack (multi-vendor)
  - Multi-chassis EtherChannel (“MEC” - Cisco VSS), kind of like different members of a switch stack
  - Virtual Port Channel (“vPC” - Cisco Nexus 7000)
- Ink is still drying for some of these, the models included in this talk have been well-baked
Summary
Layer 3 versus Layer 2

- Great quote from Cisco design document which summarizes the difference between Layer 3 and Layer 2:
  - A routing protocol identifies where to send packets
  - STP identifies where not to send frames

- Source:
  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html
Summary

• We have deployed and operated the preceding models over the past 12 years
• Different datacenter environments, application and security requirements, and equipment capabilities will lead to different solutions
• This list may not be exhaustive, come find me if you have other ways you’ve done this – both successes and failures!
• Is the L2/STP model seeing a comeback due to server virtualization?
Any Questions?

Thank you for listening
Peak Web Consulting is available to assist

Dani Roisman
droisman ~ at ~ peakwebconsulting ~ dot ~ com