RSTP to MST
Spanning Tree Migration in a Live Datacenter

NANOG47
October 20, 2009
Dani Roisman
droisman ~ at ~ peakwebconsulting ~ dot ~ com
Introductions
What Is This Talk About?

• We found ourselves in the situation where we had to migrate from per-VLAN RSTP to MST due to high resource utilization at the network core
• Primary challenge: migrate with minimal impact to a live production network
• Secondary challenge: define best-practices for MST deployment that will yield maximal stability and future flexibility
• We had minimal reference material or previous experience (previous failures)
What Isn’t This Talk About?

• Not going to focus on the inner-workings of RSTP and MST
  – State machine
  – Topology change notifications
  – CST (common spanning tree)
  – IST (internal spanning tree)
  – CIST (common and internal spanning tree)

• Check vendor docs and whitepapers if you need to know about those, I’m just focusing on real-world migration experience
Who May Be Interested?

• Networks who run datacenters with a large number of VLANS
• Layer 2 connectivity to top-of-rack switches
• Folks experiencing scaling issues with large number of spanning tree instances
• Some vendor equipment runs in a default STP mode that needs to be adapted as the network grows
Our Environment
Network Overview

• Managed hosting facility needing flexibility of any VLAN to any server, in any rack
• Each customer has their own dedicated VLAN and subnet
• Layer 2 switches in racks, uplinks to Layer 3 collapsed core/distribution switches
• Need to have fault-tolerance, no SPoF, and therefore running STP to rack switches
• Supporting over 200 customers in the live datacenter, therefore over 200 live VLANs
Reference: Sample Network

- Core Layer 3 Switch #1
- Core Layer 3 Switch #2
- ISPs etc.
- L2 Rack Switch
- Servers

Rack #1
Rack #2
Rack #3
Rack #n

NANOG 47 Dani Roisman
droisman~at~peakwebconsulting~dot~com
Problem Seen

- High number of STP “logical port instances” using Cisco’s default per-VLAN RSTP (a.k.a. “rapid-pvst”) on the cores
- Multiply VLAN * Interface count = logical port instances
- In short, too many spanning tree instances for the Layer 3 core switch
- Concerns include CPU utilization, memory, other resource exhaustion at the core
Vendor Support: Per-VLAN STP

• Cisco: per-VLAN is the default configuration, actually can *not* switch to single-instance STP
• Foundry: offers per-VLAN mode to interoperate with Cisco
• Juniper MX and EX: offers VSTP to interoperate with Cisco in newer code versions
• Force10 FTOS and newer SFTOS: Offer PVST+ to interoperate with Cisco
Looking Back

• Perhaps we’re too spoiled with Cisco’s default implementation of per-VLAN STP?
• Don’t actually need per-VLAN STP, don’t want to utilize alternate path during steady-state since we want to guarantee 100% capacity during failure scenario
• Just need Layer 2 redundant path to the rack, one primary link that will fail over to a backup link
What Were Our Options?

• Simplest solution: collapse from per-VLAN to single-instance STP
• Since we were running a Cisco shop, the only way to reduce STP instances was to migrate to standards-based 802.1s MSTP
• Ironically, called “Multiple Spanning Trees” – however we were looking to collapse number of STP
MST Solution Design
MST Explained

- MST allows us to map multiple VLANs into a single instance of MST
- MST = IEEE 802.1s, developed as a standard to allow vendor interoperability for multiple spanning trees
- Typical use for non-Cisco equipment: load sharing by distributing VLANs across different path
- Driver for use on Cisco equipment: reducing the number of STP instances
Problems with MST

- MST introduces a new configuration complexity: all switches within one “region” must contain identical VLAN-to-MST instance mapping.
- This means that any VLAN or MST changes made must be updated universally throughout the datacenter.
- Issue with change control: all rack switches must be touched whenever VLAN mappings are updated.
Design Ideas Considered

• Should we just go ahead and create a single MST instance that included all VLANs 1 – 4096
• Should we pre-create instances such as:
  – Instance 1: VLANs 1 – 99
  – Instance 100: VLANs 100 – 199
  – Instance 200: VLANs 200 – 299
  – Instance 300: VLANs 300 – 399
• Wait a moment, do all vendor equipment support large MST instance numbering?
• No! Some only support numbers 1 through 16
Migration Challenges

• Had to migrate production network from per-VLAN RSTP to MST with zero down time if possible
• Used a Lab environment made up of core simulated with two Layer 3 switches and three rack Layer 2 switches
• We could not find a way with zero down time, so we had to find absolute minimal downtime
• We got this down to 45 seconds (one STP blocking → learning → forwarding cycle)
Migration Lab Key Notes

• Know your roots – set cores to “highest” STP priority (lowest numeric priority value)
• Set consistent root for all VLANs
• Set priority of rack switches to lower-than-default do ensure they do not become root
• Start from roots, then work your way “down”
• MSTP runs RSTP for backwards compatibility, making migration less painful if you’re already running RSTP (802.1w)
• Choose your VLAN ⇔ Instance mapping carefully
MSTP Gotchas

• Instance numbering
  – Some devices support arbitrarily numbered MSTP instances, e.g. 1, 100, 105, 3000
  – Some only support 1 through $<\text{maxN}>$, e.g. 1 through 16, 1 through 32, etc..

• MST config has got to match on all devices
  – Region name
  – Revision number
  – VLAN $\leftrightarrow$ Instance mapping for all VLANs, even if there are no active ports in that VLAN on a given switch
Migration Example:
Cisco Equipment
Starting Point

- All devices running 802.1w (actually per-VLAN RSTP)
- Core 1: root with priority 8192 for all VLANs
- Core 2: secondary root with priority 16384 for all VLANs
- Rack Switches: blocking towards Core 2 with priority 40960 for all VLANs
Reference: Sample Network
Migration Walk-Through

• Base MSTP Configuration:

  spanning-tree mst configuration
  name DC1-MST
  revision 1
  instance 1 vlan 2-49 ! we don’t use VLAN 1
  instance 2 vlan 50-99
  instance 3 vlan 100-199
  instance 4 vlan 200-299
  instance 5 vlan 300-399
  instance 6 vlan 400-499
  instance 7 vlan 500-1000
  instance 8 vlan 2000 - 3999

• Specific to Core 1:

  spanning-tree mst 0-8 priority 8192 ! don’t forget MSTI 0

• Specific to Core 2:

  spanning-tree mst 0-8 priority 16384 ! don’t forget MSTI 0

• Rack Switches:

  spanning-tree mst 0-8 priority 40960 ! don’t forget MSTI 0
Migration Step 1: Network Hit

- Core 1: Activate MSTP
  
  spanning-tree mode mst
45 Sec. Later: Network Restored

- Wait patiently…
Migration Step 2: Core 2

- Core 2: Activate MSTP
  spanning-tree mode mst
Migration Step 3: Rack Switches

- All Rack Switches: Activate MSTP
  ```
  spanning-tree mode mst
  ```

Rack switches have brief blocking→forwarding cycle
Ongoing Maintenance
Ongoing Maintenance

• Once MST is deployed, a few things must be kept in mind:
  – All new devices should be pre-configured with identical MST parameters before being deployed on the network
  – Any VLAN $\leftrightarrow$ Instance mapping changes should be made on the root, then pushed to secondary root, then to rack switches

• Currently there is no widespread protocol for automatic propagation of MST configuration
  – Maybe VTPv3? (though Cisco only)
  – Anyone else?
Summary
Summary

• MST adds configuration complexity, so stay on your toes
• While not covered in this talk, MST allows for great multi-vendor interoperability in a Layer 2 datacenter network
• We’ve only deployed this solution a few times, I’m interested in hearing feedback and experience from others in similar situations
• Anybody know how to configure Cisco Layer 3 switches for single-instance RSTP?
Any Questions?

Thank you for listening

Peak Web Consulting is available to assist

Dani Roisman
droisman ~ at ~ peakwebconsulting ~ dot ~ com