Switch Clustering
Design Best Practices

Enterprise Networking Solutions
July 2009
NN48500-584
Version 2.2

Document Use

This document provides a quick overview of the Nortel recommended best practices for implementing Switch Clustering. Please note that the recommendations may vary between designs based on the hardware platforms being used and the feature set available on each switching product. Nortel always recommends reviewing the release notes for each product before deployment. This will help to avoid any unexpected results during network operations due to software limitations or known issues with specific releases of code.

Document Version Control

- Version 2.0 published April 2009
  - Many document updates to add more information and revise format
  - Revised VLACP timer recommendations
  - Added FDB timer change
  - Added ERS 5000 Design Requirements section
- Version 2.1 published May 2009
  - Clarified FDB timer change to apply to ERS 8600 / 8300 / 1600 only
  - Added information regarding SLPP in scaled environments
  - Added information regarding SLPP reset to clear threshold counters
  - VRRP guidelines reference ERS 5000 design requirements for additional recommendations
  - Revised recommendation on Port Rate Limiting – access ports only
  - Added additional information in ERS 5000 design requirements section – more details
- Version 2.2 published July 2009
  - Created external version for posting on the Technical Support portal
  - Corrected timer from Hold-Down to Hold-Up for RSMLT Layer 2 Edge

Please send questions, comments, or feedback to ddebacke@nortel.com

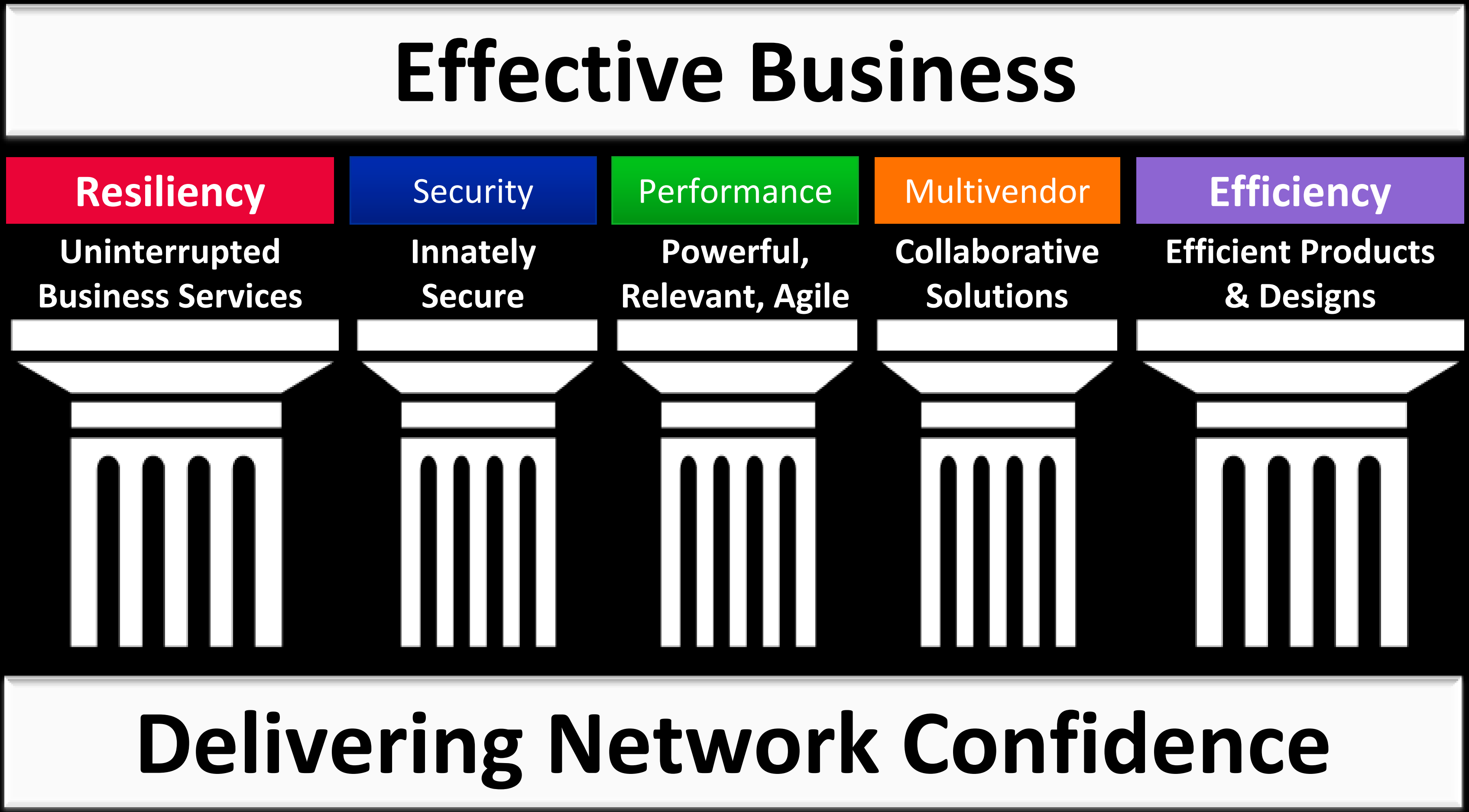
Reference Information

In addition to the guidelines provided here, the following documents are available which provide more detailed information.

- *Switch Clustering using Split Multi-Link Trunking (SMLT) with ERS 8600, 8300, 5x00 and 1600 Series Technical Configuration Guide (NN48500-518)*
- *Switch Clustering (SMLT/SLT/RSMLT/MSMLT) Supported Topologies and Interoperability with ERS 8600 / 5000 / 8300 / 1600 (NN48500-555)*
- *Resilient Multicast Routing Using Split-Multilink Trunking for the ERS 8600 Technical Configuration Guide (NN48500-544)*
- *The Small Campus Technical Solution Guide (NN48500-573)*
- *Data Center Server Access Solution Guide (NN48500-577)*

Nortel Enterprise Networking Solutions

Strategic Values



Effective Business

Resiliency

Uninterrupted Business Services

Security

Innately Secure

Performance

Powerful, Relevant, Agile

Multivendor

Collaborative Solutions

Efficiency

Efficient Products & Designs

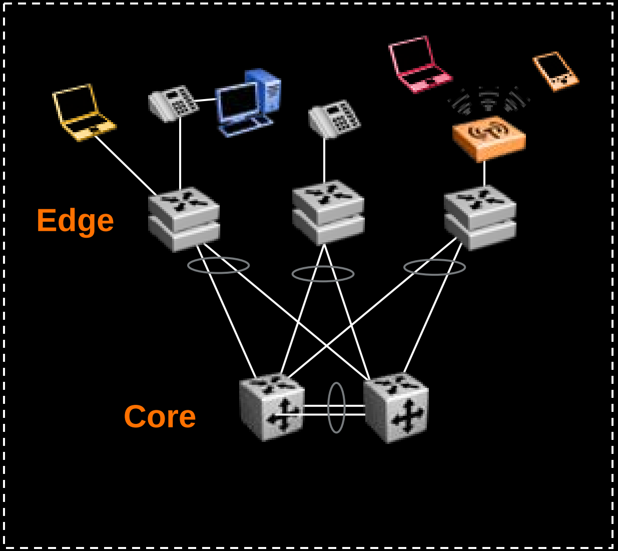
Delivering Network Confidence

Nortel’s Ethernet Switching Vision

Delivering the Foundation for Enterprise Networks

Provide an always-on Ethernet infrastructure enabling uninterrupted access to Enterprise applications and services.

- Simple and efficient network architectures
- Active / Active with fast failover and recovery
- High performance products and solutions
- Integrated security
- Technology innovation
- Energy efficiency in every product and solution
- Price / Performance leader



Resiliency – Performance – Security – Efficiency

Reliability Optimization

Choosing the Right Model

	Active–Passive	Active–Active
	n + 1	n – 1
Normal Operation	🟢 🔴	🟢 🟢
Failure Operation	🔴 🟢	🔴 🟢
	• Hot Standby • Size to constant capacity • Cost = 1/Avail x Price • Transition Timing	• Distributed Use • Size to Need • Reduce Capacity in Failure • Use Business Rules

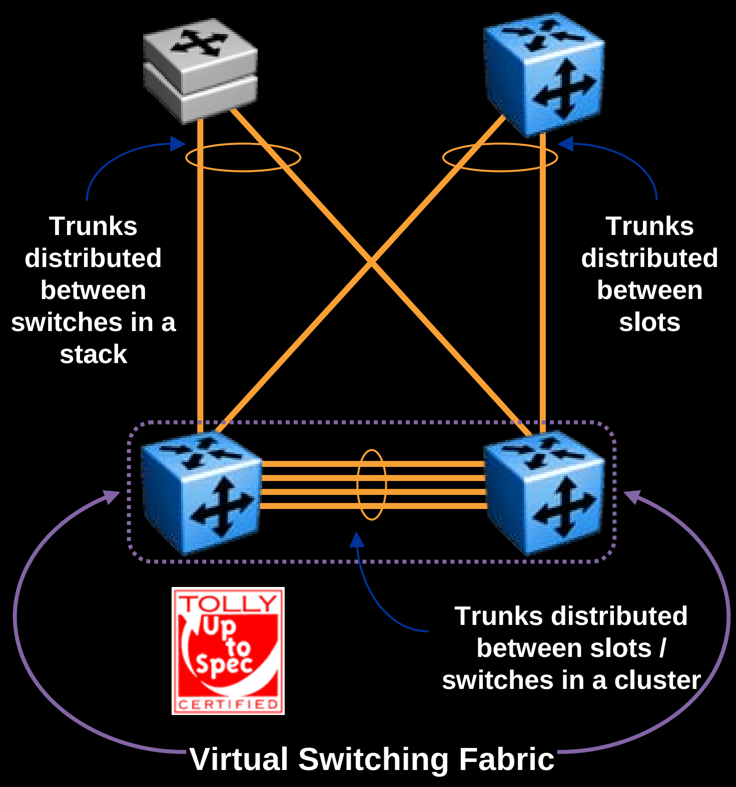
High Performance & Reliable IP Core

Active–Active

- Switch Clustering
  - Link and Nodal Redundancy (N-1)
  - Split Multilink Trunking (SMLT)
  - Routed Split Multilink Trunking (RSMLT)
- Combines Resiliency & Performance
  - All links passing traffic
  - Sub second stateful failover
  - No Spanning Tree on switch to switch links
- Virtual “Switch” Fabric
  - Centralized or distributed
  - Increased performance using full capacity
  - No single point of failure
- Scalable
  - Standalone / Stackable / Modular Chassis
  - Single pair / Square / Full Mesh
  - Mbps → Gbps → Tbps

Switch Clustering

Interoperable with ANY 3rd party edge device

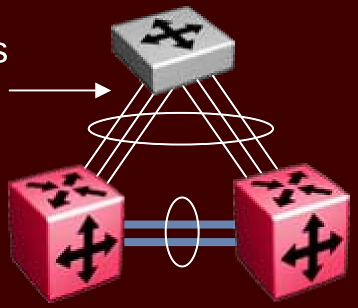

Trunks distributed between switches in a stack
Trunks distributed between slots
TOLLY Up to Spec CERTIFIED
Trunks distributed between slots / switches in a cluster
Virtual Switching Fabric

Switch Clustering

Terminology – SMLT & SLT

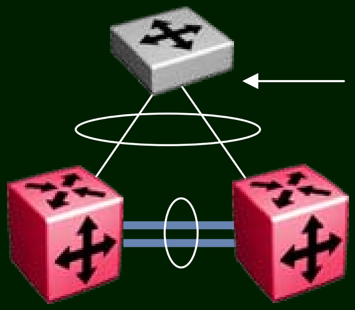
SMLT – Split MultiLink Trunking
“Standard” layer 2 design using
MLT-based connections

Scaling Uplinks Between the Edge and the Core



Maximum Number of SMLT’s supported per Cluster is number of MLT groups supported less one required for IST

SLT – Single Link Trunking
“Standard” layer 2 design using
port-based connections



Maximum Two Uplinks per Edge Connection

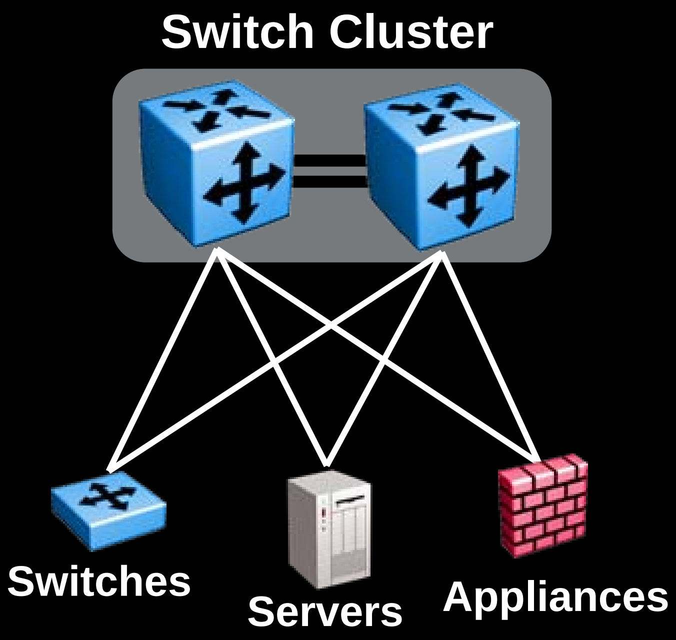
Maximum Number of SLT’s supported per Cluster is number of ports on one core switch less two required for IST

Can Configure both SMLT and SLT on a single Switch Cluster

How Do I Build a Switch Cluster ?

Physical Design

- **Create a Switch Cluster core with two like Ethernet Routing Switches**
  - ERS 8600
  - ERS 8300
  - ERS 1600
  - ERS 5000
- **Create the Inter-Switch Trunk (IST) between core switches**
  - Multilink Trunk (MLT) for resiliency
  - Responsible for forwarding/control synchronization
  - Must be the same speed: 10Mbps to 10Gbps
- **Connect edge devices**
  - SMLT, SLT, RSMLT on the Switch Cluster
  - Link aggregation configuration (802.3ad, MLT, etc.)
  - Disable Spanning Tree on link aggregation group
  - Autonegotiation enabled for Remote Fault Indication (RFI) support



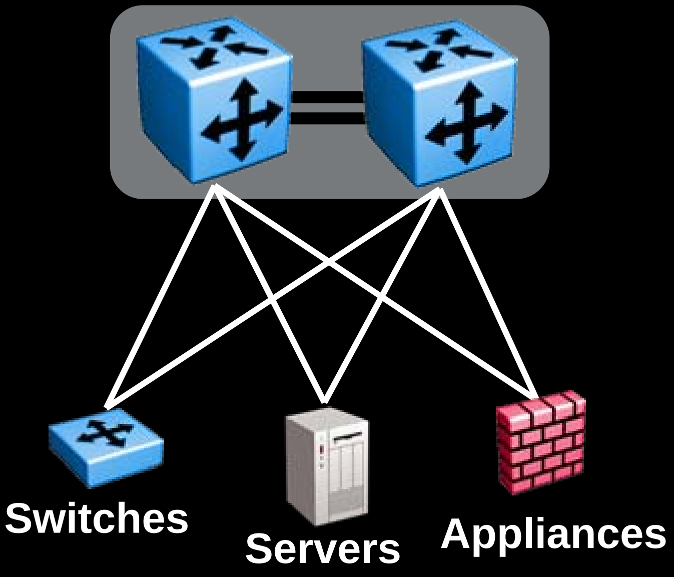
The “Simple” aspect includes the configuration

How Do I Build a Switch Cluster ?

Logical Design

- **Default Gateway Redundancy**
  - Virtual Router Redundancy (VRRP)
    - Backup Master enhancement (active/active)
  - Routed Split MultiLink Trunking (RSMLT-Edge)
    - ERS 8600 / ERS 8300
- **Simple Loop Prevention Protocol (SLPP)**
  - Prevents loops in Switch Cluster networks
  - Disables uplink port where loop is detected
  - Enabled on Access SMLT/SLT ports – disabled on IST
- **Virtual LACP (VLACP)**
  - Lightweight protocol for end-to-end health check
  - Detect end-to-end failure by propagating link status between ports that are either
    - Physically connected point-to-point
    - Logically connected point-to-point across an intermediate network
  - Does not perform link aggregation functions

Switch Cluster

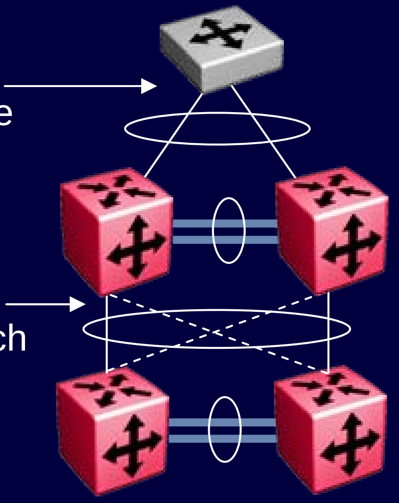


The “Simple” aspect includes the configuration

Switch Clustering

Terminology – RSMLT

RSMLT – Routed Split MultiLink Trunking "Standard" Core Routing Layer 3 Design or Routing with Layer 3 Edge



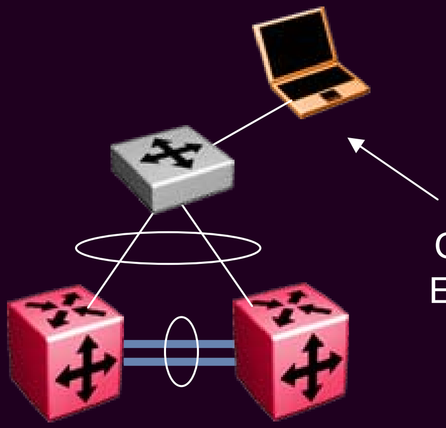
Routed Connection Between Edge and Core

Routed Connection Between Switch Clusters

Sub-second failover without modifying any layer 3 protocols or timers

Support for square or full mesh designs

RSMLT Edge – Routed Split MultiLink Trunking Replacement for VRRP with Layer 2 Edge



Provides Default Gateway for End Stations

Scales beyond maximum number of VRRP instances

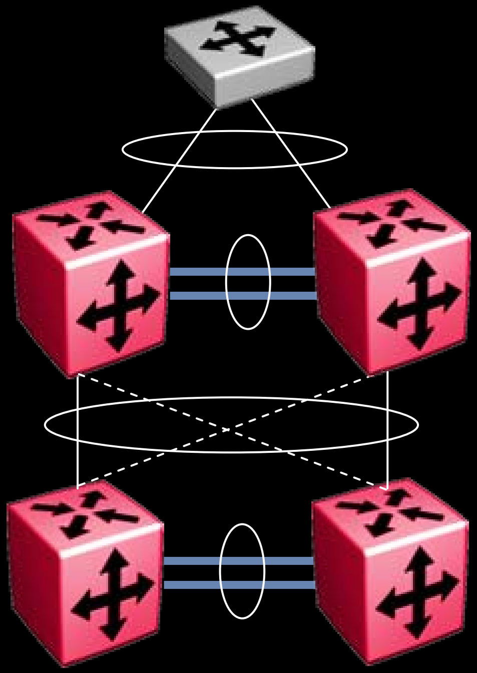
Do not run VRRP and RSMLT on the same edge VLAN simultaneously

Supported on ERS 8600 and ERS 8300 platforms

Nortel Switch Cluster Core

Routed SMLT (RSMLT) Solution

- Unparalleled Layer 3 resiliency
- End to end sub-second failover for routed VLAN traffic
- RSMLT will take care of resiliency
- Perfect complement to SMLT/SLT from the edge



- All routers in the core VLAN run standard IGP such as RIP or OSPF
- No tuning of IGP necessary
- No VRRP, ECMP required in core VLANs
- Multiple routed RSMLT Core VLANs supported per switch

- Support for square and full mesh topologies in the core
- Support for triangles to the edge – when used as a VRRP replacement or Layer 3 edge

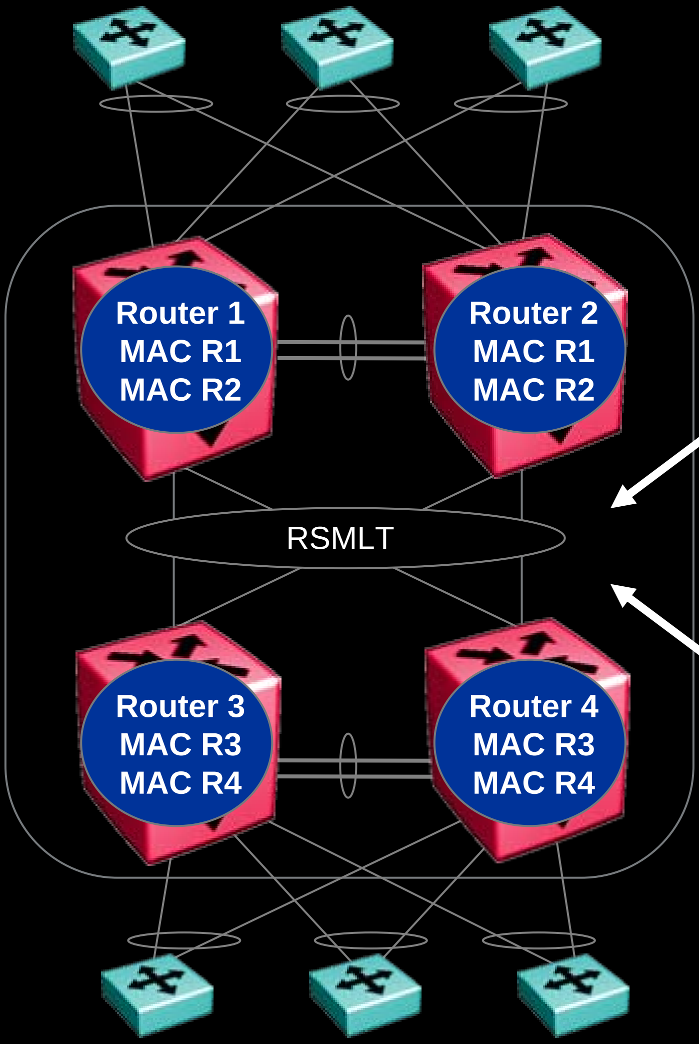
Sub-second Failover in Layer 3 Environments

LAN Solution Example

Routed SMLT (RSMLT) Operation

Normal Operation

- RSMLT aggregation switch pairs exchange local router MAC addresses
- Both router MAC addresses are routing packets on both RSMLT aggregation switches
- SMLT/RSMLT ensures that no packets are duplicated



Core VLAN between all four routers

- RSMLT enabled on core VLAN(s)
- All cluster switches in the same IP subnet
- Routing protocol enabled
- No ECMP / VRRP required

Any IGP

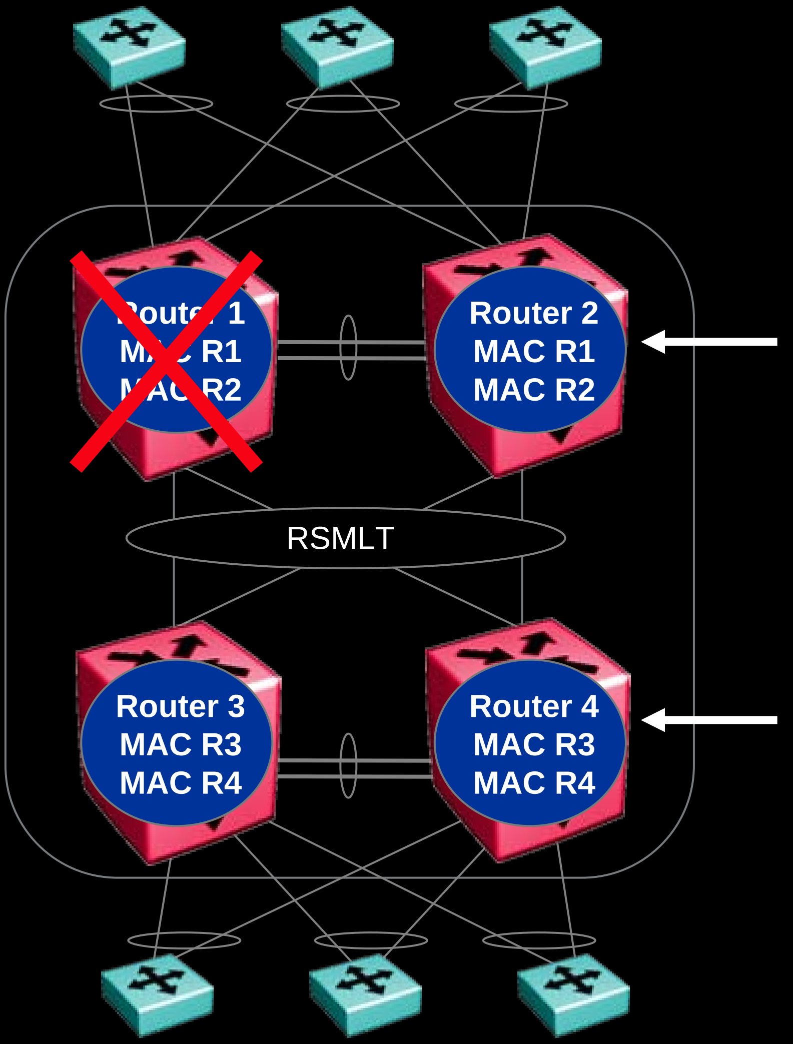
- OSPF or RIP
- IGP knows nothing of RSMLT
- Protocol timers unchanged

LAN Solution Example

Routed SMLT (RSMLT) Operation

Failure Operation

- Routing protocol converges just as normal
- Router 2 continues to forward for Router 1
- After routing protocol is converged MAC R1 is flushed



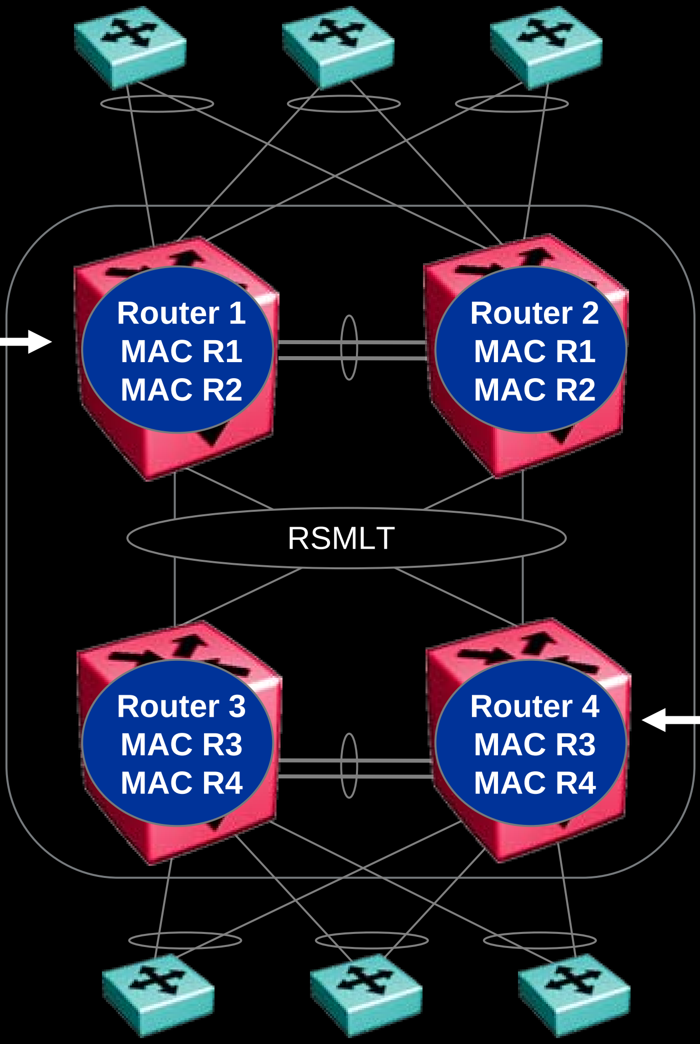
Hold-up timer kicks in

- Keeps MAC R1 alive for timer duration
- Timer should be greater than IGP convergence time

Router 3 & Router 4 continue to forward without any service disruption

LAN Solution Example

Routed SMLT (RSMLT) Operation



Hold-down timer kicks in →

- **Will not start forwarding traffic until timer expires**
- **Timer should be greater than IGP convergence time**

Recovery of Router 1

- **Routing protocol converges just as normal**
- **After routing protocol is converged Router 1 and Router 2 begins to forward for MAC R1 and MAC R2**

← Router 3 & Router 4 continue to forward without any service disruption

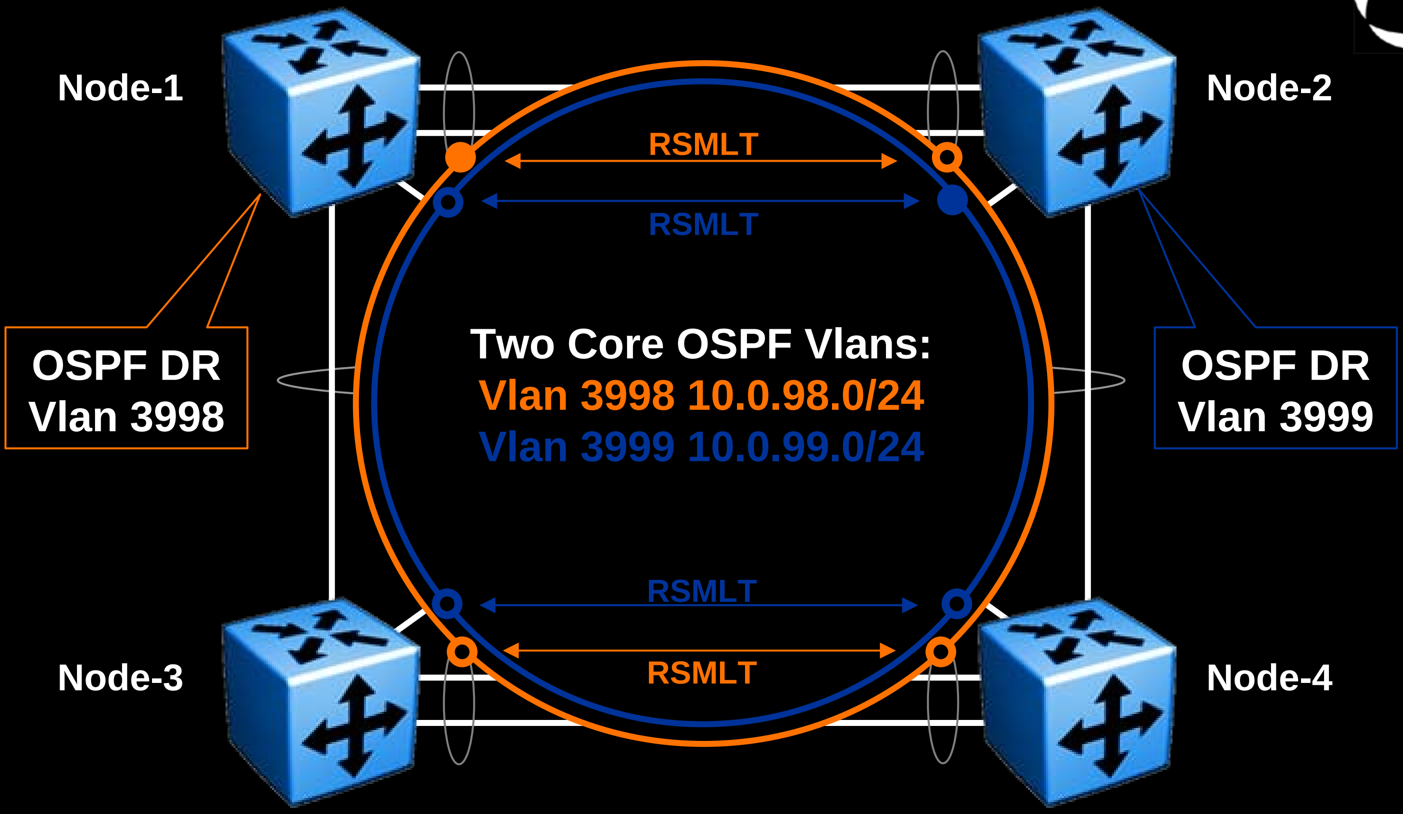
RSMLT with Dual Core OSPF VLANs

Enhancing Availability

- Issue
  - In the event of losing the DR, the BDR is immediately elected as the new DR on the broadcast segment. As soon as the new DR is elected all routers must immediately perform an SPF run and issue new LSAs for the segment. Since the DR is the only router generating the Network LSA for the segment, a new Network LSA needs to be generated by the new DR (the old one no longer being valid) and every router on that segment also needs to refresh their own router LSA.
  - The SPF run is done as soon as the router has detected a new DR. So if the router is quick, it will perform the SPF run before receiving the new LSAs for the segment; as such all OSPF routes across this particular segment are not possible after the 1st SPF run. Whereas if the router is slow, by the time it detects the new DR and does it's SPF run, it already has received most (if not all) of the new LSAs from its neighbors and it's routing table will be still be able to route most (if not all) routes across the segment.
  - Bottom line, if the DR fails, 2 SPF runs are necessary to restore routes across the segment. But the OSPF hold down timer will not allow 2 consecutive SPF runs to occur within the value of the timer. That timer defaults to 10secs on the ERS 8600 and can be reduced to minimum 3 secs. So every time the DR is lost, traffic interruption of up to 10secs (or 3secs if hold down timer is lowered; but probably it's best not to touch this timer) occurs.
  - In a classical OSPF routed design, this never constitutes a problem, since OSPF is running over multiple segments (pt-pt routed links) so even if a segment cannot be used (following loss of DR and 1st SPF run), routes are re-calculated over alternative segments (also if the adjacencies are real pt-pt, then there is no DR/BDR in the first place).
  - But with RSMLT designs, we typically only deploy a single OSPF routed vlan, which constitutes a single segment.
- Solution
  - Create a second OSPF Core VLAN and force different nodes to become the DR for each VLAN.
  - Each OSPF Core VLAN will have DR (set priority to 100) and no BDRs (set OSPF priority to 0 on all routers/switches not intended to become the DR).
  - No BDR is necessary since the two VLANs back each other up from a routing perspective.

RSMLT with Dual Core OSPF VLANs

Enhancing Availability



Node-1

Node-2

RSMLT

RSMLT

OSPF DR
Vlan 3998

Two Core OSPF Vlans:
Vlan 3998 10.0.98.0/24
Vlan 3999 10.0.99.0/24

OSPF DR
Vlan 3999

RSMLT

RSMLT

Node-3

Node-4

● ● IP interface on VLAN with **OSPF priority = 100**

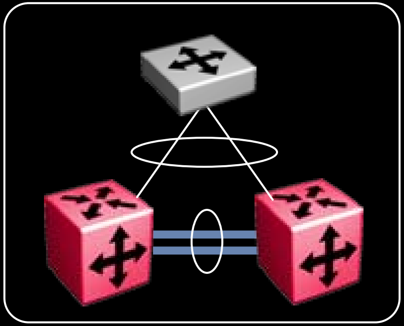
○ ○ IP interface on VLAN with **OSPF priority = 0**

Switch Clustering

Supported Topologies

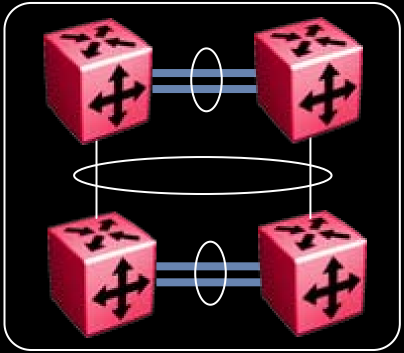
Triangle

- Single Switch Cluster at the core with edge directly connected



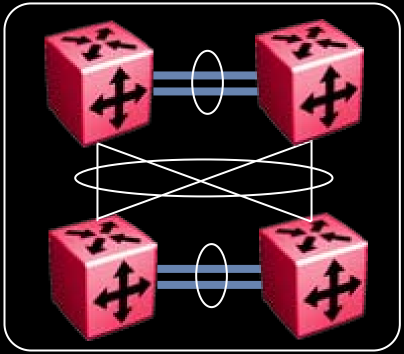
Square

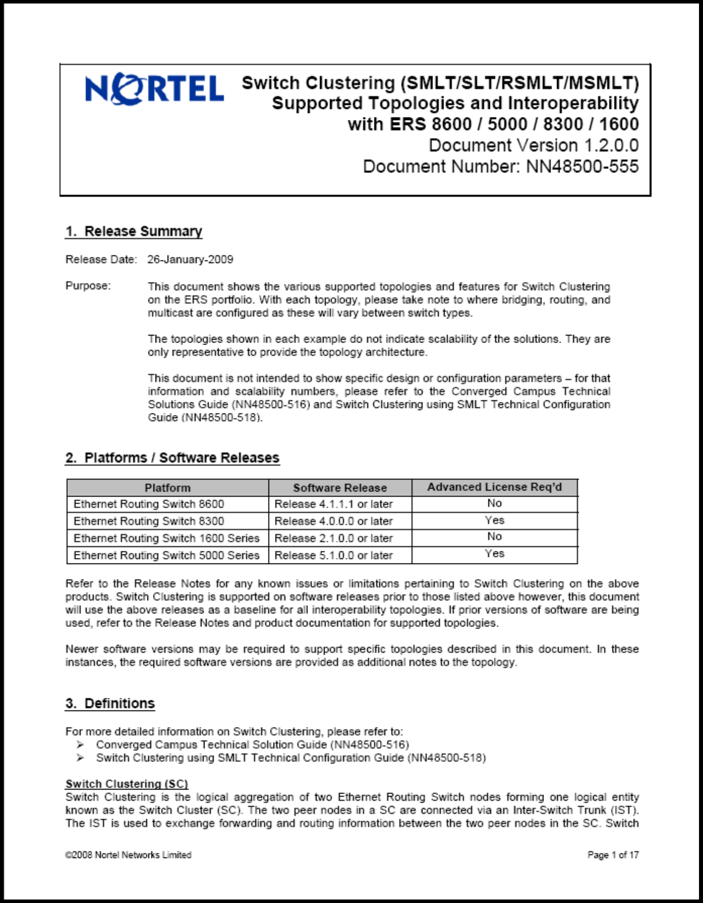
- Two pairs of Switch Clusters interconnected by SMLT. Squares can be scaled with additional pairs of Switch Clusters



Full Mesh

- Expanding on the Square topology, the full mesh adds additional connections between the pairs so that each switch has at least one connection to every other switch in the square





NORTEL **Switch Clustering (SMLT/SLT/RSMLT/MSMLT)**
Supported Topologies and Interoperability
with ERS 8600 / 5000 / 8300 / 1600
Document Version 1.2.0.0
Document Number: NN48500-555

1. Release Summary

Release Date: 26-January-2009

Purpose: This document shows the various supported topologies and features for Switch Clustering on the ERS portfolio. With each topology, please take note to where bridging, routing, and multicast are configured as these will vary between switch types.

The topologies shown in each example do not indicate scalability of the solutions. They are only representative to provide the topology architecture.

This document is not intended to show specific design or configuration parameters – for that information and scalability numbers, please refer to the Converged Campus Technical Solutions Guide (NN48500-516) and Switch Clustering using SMLT Technical Configuration Guide (NN48500-518).

2. Platforms / Software Releases

Platform	Software Release	Advanced License Req'd
Ethernet Routing Switch 8600	Release 4.1.1.1 or later	No
Ethernet Routing Switch 8300	Release 4.0.0.0 or later	Yes
Ethernet Routing Switch 1600 Series	Release 2.1.0.0 or later	No
Ethernet Routing Switch 5000 Series	Release 5.1.0.0 or later	Yes

Refer to the Release Notes for any known issues or limitations pertaining to Switch Clustering on the above products. Switch Clustering is supported on software releases prior to those listed above however, this document will use the above releases as a baseline for all interoperability topologies. If prior versions of software are being used, refer to the Release Notes and product documentation for supported topologies.

Newer software versions may be required to support specific topologies described in this document. In these instances, the required software versions are provided as additional notes to the topology.

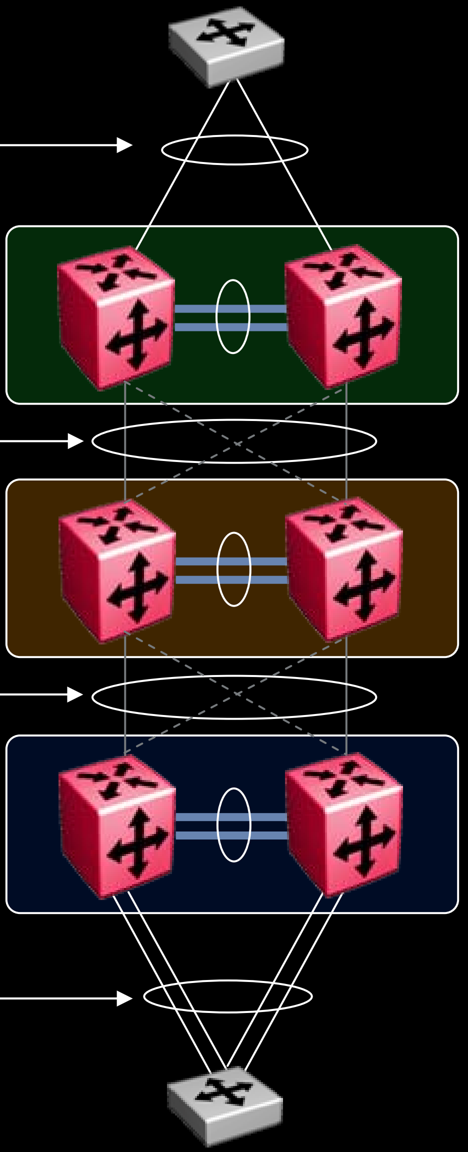
3. Definitions

For more detailed information on Switch Clustering, please refer to:

- ➢ Converged Campus Technical Solution Guide (NN48500-516)
- ➢ Switch Clustering using SMLT Technical Configuration Guide (NN48500-518)

Switch Clustering (SC)
Switch Clustering is the logical aggregation of two Ethernet Routing Switch nodes forming one logical entity known as the Switch Cluster (SC). The two peer nodes in a SC are connected via an Inter-Switch Trunk (IST). The IST is used to exchange forwarding and routing information between the two peer nodes in the SC. Switch

Switch Clustering Solution



Access SLT
(Triangle)

Core SMLT Layer 2
Core RSMLT Layer 3
(Square or Full Mesh)

Core SMLT Layer 2
Core RSMLT Layer 3
(Square or Full Mesh)

Access SMLT
(Triangle)

Switch Cluster (SC1)

Switch Cluster (SC2)

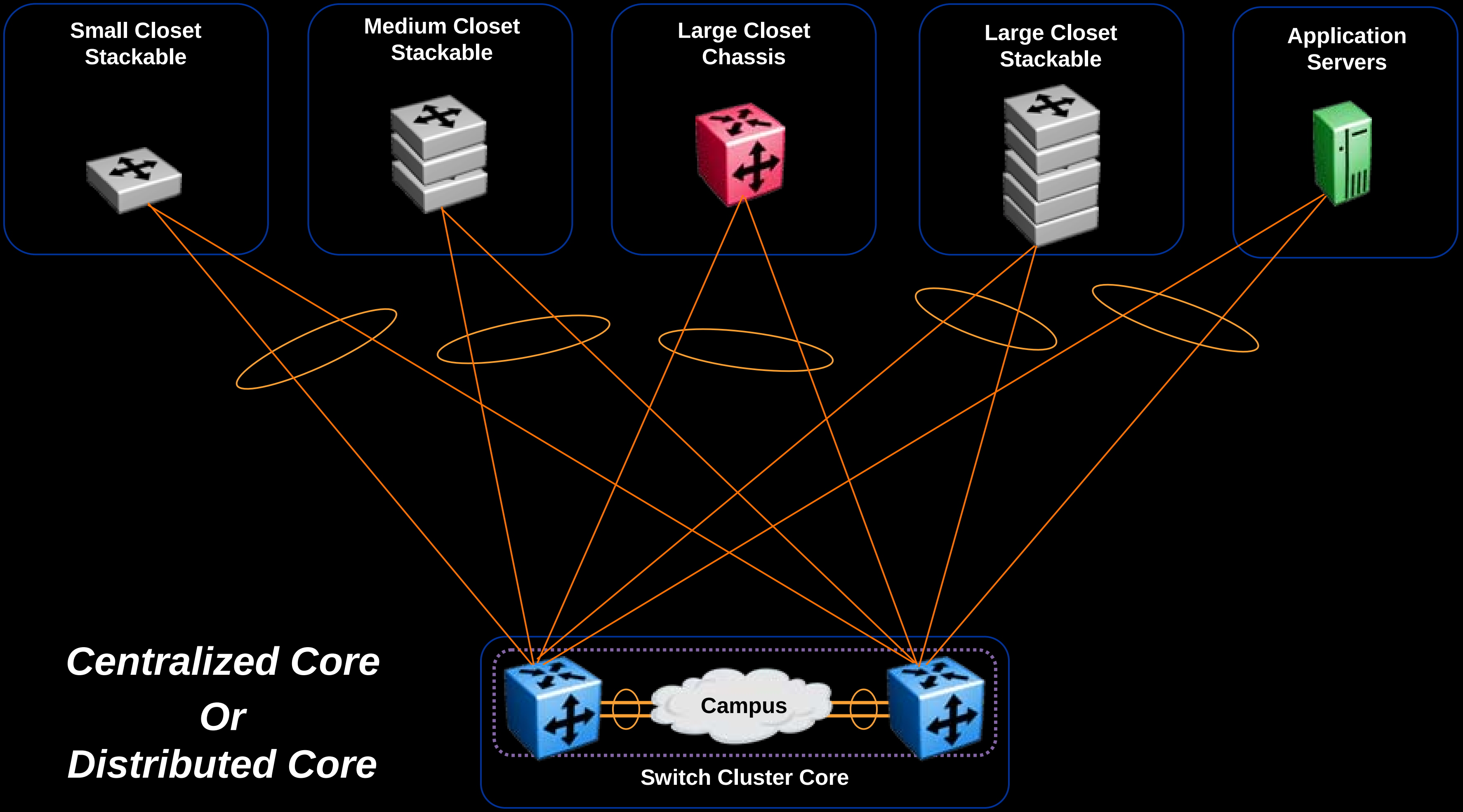
Switch Cluster (SC3)

Cascaded
Switch
Clusters

Network Design Flexibility

Reducing OPEX

Simple Two Tier Architecture



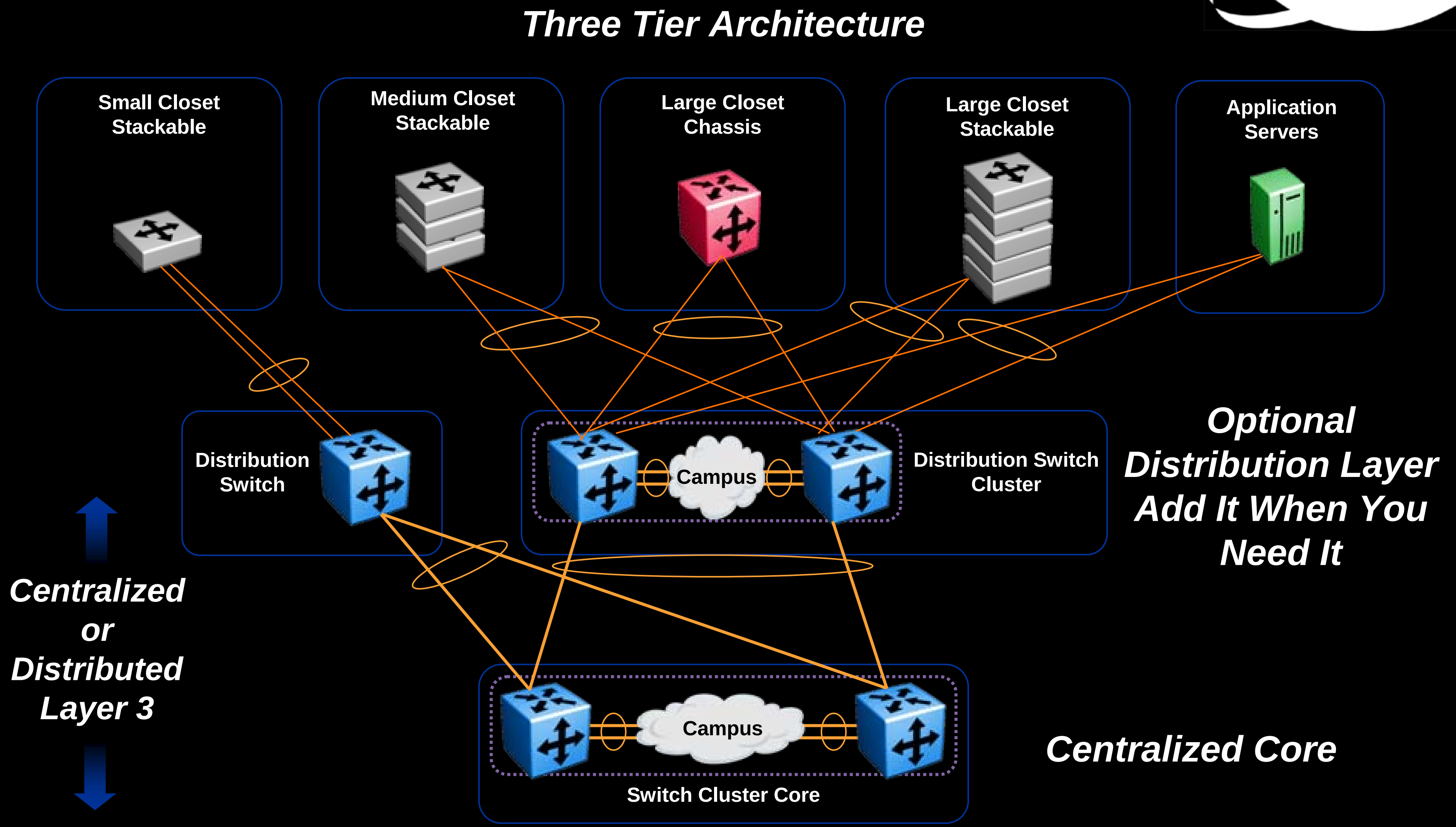
Centralized Core

Or

Distributed Core

Network Design Flexibility

Reducing OPEX



Switch Clustering Design Best Practices

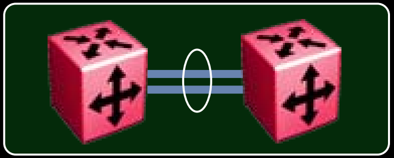
Disclaimer:

The recommendations provided here are based on large scale network testing validated by Nortel. Using values and timers outside of these recommendations are permitted at the user’s own risk based on specific network design requirements. If issues are encountered when running outside of these recommendations, the first step will be to move to recommended values before pursuing further.

Creating the Switch Cluster

Configuration Recommendations

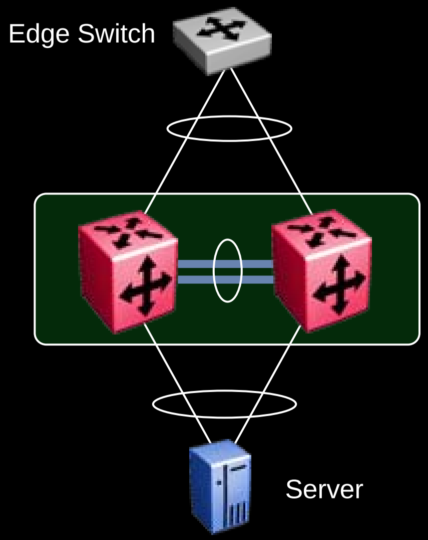
- The IST should be a distributed MLT for added resiliency between the Switch Cluster Cores
  - The DMLT links that comprise the IST must be of the same speed
  - Number and speed of links for the IST is determined by the amount of traffic that would use the IST during a failed condition – during normal operation, the majority of traffic across the IST is from devices that are not dual-homed
  - Use a private address space with 30 bit mask for IST IP’s
  - Enabling OSPF passive on the IST is supported to ease network management of the Switch Cluster
  - Do NOT use the IST IP addresses as the next hop address for any static routes
- All VLANs that span SMLT/SLT connections must be tagged across the IST



Link Aggregation

Connecting the Edge to the Switch Cluster

- Switch Cluster is agnostic of Edge devices
  - Must enable a form of link aggregation on the Edge
- ERS 8600 or ERS 8300 Switch Cluster
  - Edge devices can use 802.3ad or static link aggregation
- ERS 5000 Switch Cluster
  - Edge devices must use static link aggregation
  - 802.3ad over SMLT/SLT is not supported
  - IGMP over SMLT/SLT is not supported
- Server Connectivity
  - ERS 8600 and ERS 8300 support both 802.3ad and static link aggregation over SMLT/SLT
    - When using 802.3ad over SMLT on ERS 8600, enable:
      - LACP SMLT-SYS-ID
  - ERS 5000 Switch Cluster must use static link aggregation
  - For details on NIC teaming and Nortel Switch Clustering, refer to Data Center Server Access Solution Guide (NN48500-577)



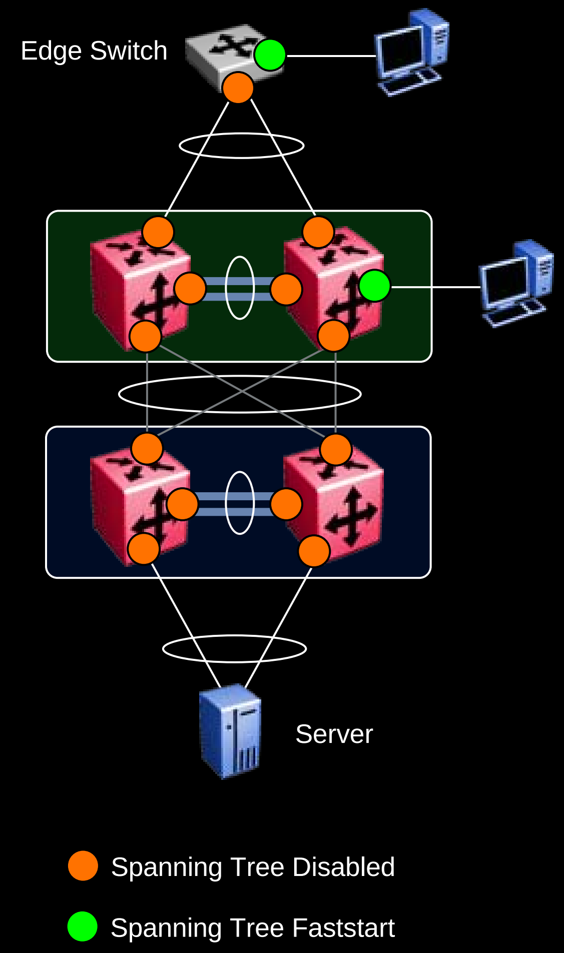
Switch Clustering Best Practices

Configuration Checklist

☑ **Spanning Tree**

Disabled on uplink and IST ports – Core
Disabled on uplink ports – Edge
Faststart enabled on all other ports

- Spanning Tree – Core
  - Globally enabled on all Switches
  - Must be disabled on IST and Uplink ports
    - STP must be manually disabled on ERS 5000 uplink and IST ports
    - STP is automatically disabled on ERS 8600/8300/1600 uplink and IST ports
- Spanning Tree – Edge
  - Globally enabled on all Switches
  - Must be disabled on the uplink ports forming the MLT
- Spanning Tree Faststart
  - Enable on all non-uplink and non-IST ports for added protection
  - Allows ports to come up in forwarding state immediately
  - Spanning Tree will still block if a loop is detected via BPDUs



BPDU Filtering

Feature Overview

- Allows the network administrator to achieve the following:
  - Block an unwanted root selection process when an edge device is added to the network. This prevents unknown devices from influencing an existing spanning tree topology.
  - Block the flooding of BPDUs from an unknown device
- When a port has BPDU-Filtering enabled and it receives an STP BPDU, the following actions take place:
  - The port is immediately put in the operational disabled state
  - A trap is generated and the following log message is written to the log:
    - BPDU received on port with BPDU-Filtering enabled Port <x> has been disabled
  - The port timer starts
  - The port stays in the operational disabled state until the port timer expires

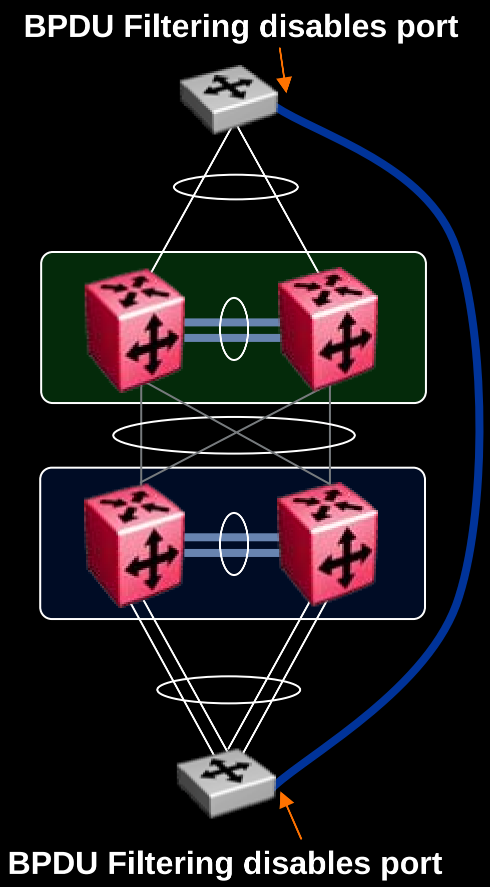
Switch Clustering Best Practices

Configuration Checklist

☑ **BPDU Filtering**

Enabled on Edge access ports

- Enable on all Edge access ports
- Set timeout to 0
  - Port remains disabled until manual intervention to re-enable it
- BPDU Filtering is not supported on MultiLink Trunk (MLT) ports
- Supported on
  - ERS 2500 with Release 4.2
  - ERS 4500 with Release 5.1
  - ERS 5500 with Release 5.1
  - ERS 5600 with Release 6.0
  - ERS 8300 with Release 4.2



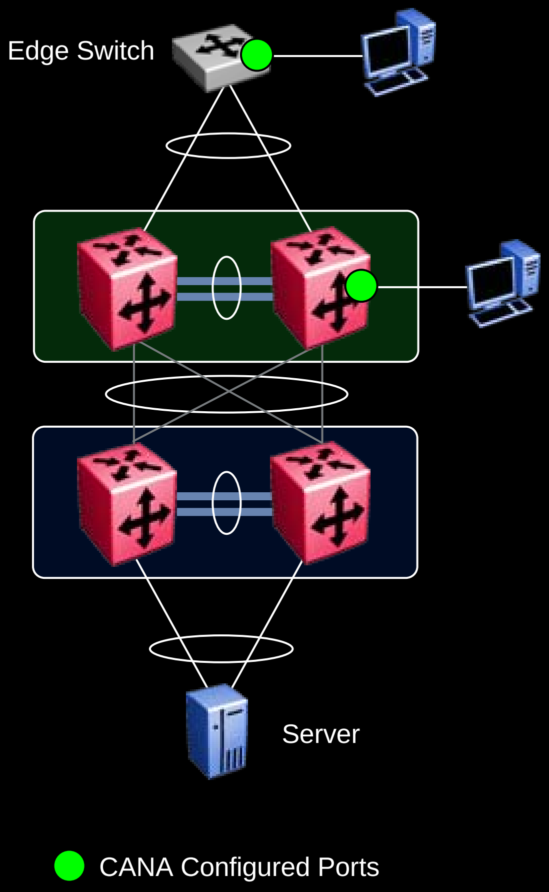
Switch Clustering Best Practices

Configuration Checklist

☑ **Autonegotiation enabled on all ports**

Can use CANA on access ports if required

- Autonegotiation
  - Leave enabled on all ports
  - Ensures RFI (remote fault identification)
  - Autonegotiation does not exist for 10Gig – RFI is already built-in
- Custom Autonegotiation Advertisements (CANA)
  - Provides capability to set autonegotiation advertisements for speed and duplex
  - Useful to limit end station connectivity bandwidth
  - Supported on:
    - ERS 2500
    - ERS 4500
    - ERS 5000
    - ERS 8300
    - ERS 8600 (RS modules only)



Switch Clustering Best Practices

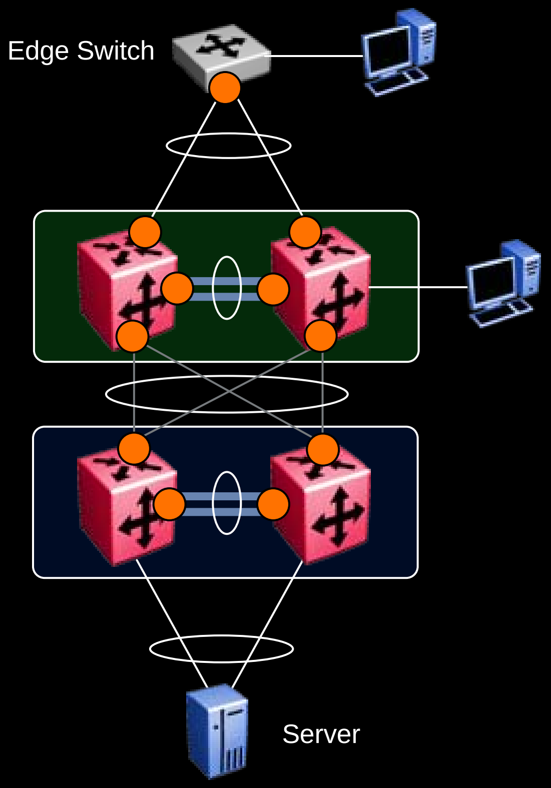
Configuration Checklist

☑ 802.1Q enabled on uplink / IST ports

☑ Discard Untagged Frames enabled on uplink / IST ports

☑ Increase FDB Timer on Switch Cluster Core VLANs

- 802.1Q VLAN Tagging
  - Enable 802.1Q on uplink ports even if only one VLAN at the Edge
  - Facilitates ease of adding VLANs to the uplinks in the future
  - Allows for use of Discard Untagged Frames on uplinks
- Discard Untagged Frames
  - Any untagged packets received on the port will be dropped at ingress
  - Protection mechanism against Edge switch reset to factory default or a new switch not configured properly being added to the network and possibly causing a loop in the network
  - Do not enable on ERS 5510 when using VLACP
- Increase FDB Timer per VLAN (ERS 8600 / ERS 8300 / ERS 1600)
  - Increase the FDB timer on all Switch Cluster Core Layer 3 VLANs from the default of 300 seconds to 21601 seconds (1 second greater than the ARP timer)
  - Reduces the amount of re-ARPs when the FDB timer for a given MAC ages out
  - Leave FDB timer at default of 300 seconds on ERS 5000 Switch Cluster

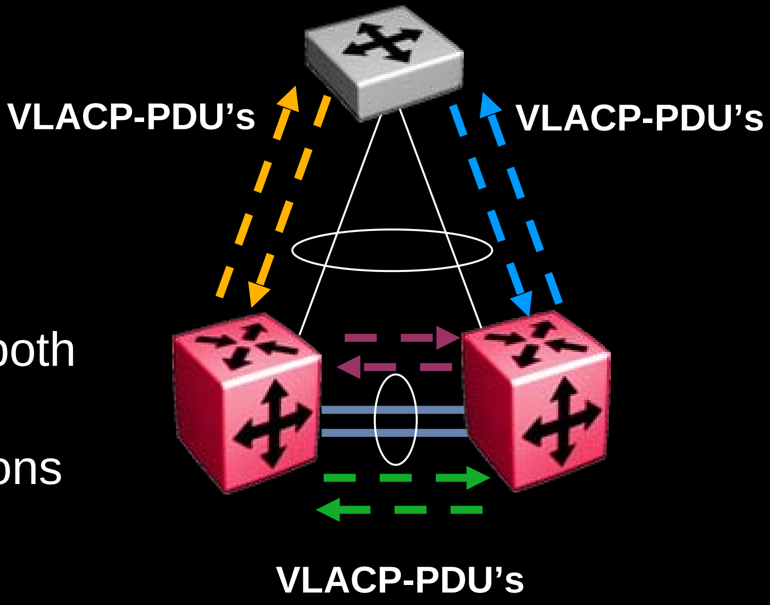


● 802.1Q VLAN Tagging and Discard Untagged Frames

Virtual LACP (VLACP)

Feature Overview

- Virtual LACP (VLACP) = Lightweight LACP
  - Detects end-to-end failure by propagating link status between ports that are:
    - Directly connected point-to-point
    - Logically connected point-to-point across an intermediate network
  - Can detect
    - Complete link failure
    - Receive or transmit link disruptions only
  - Transmits VLACPDU every “x” milliseconds so both ends of the link maintain state
  - VLACP doesn’t perform Link Aggregation functions
  - Based on LACP but is Intellectual Property of Nortel
  - Supported on:
    - ERS 2500
    - ERS 4500
    - ERS 5000
    - ERS 8300
    - ERS 8600



Switch Clustering Best Practices

Configuration Checklist

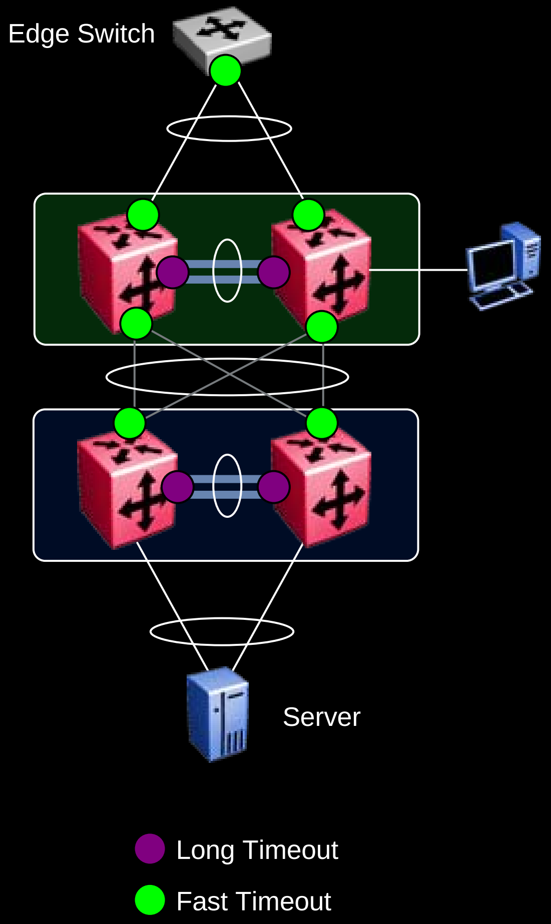
☑ **VLACP enabled on uplink and IST ports**

- Enable VLACP
  - Globally and on each individual uplink and IST port
  - Both ends must have matching Multicast MAC, Ethertype, and Timers
  - Do not enable VLACP and LACP on the same links
- For directly connected point-to-point links
  - Use reserved multicast MAC 01-80-c2-00-00-0f
  - Ensures packet is not flooded across a defaulted switch
- For end to end connections traversing intermediate networks
  - Use default MAC 01:80:c2:00:11:00

Short Timeout = Timeout Scale * Fast Periodic Timer

Long Timeout = Timeout Scale * Slow Periodic Timer

Connection Type	Fast Timer	Slow Timer	Timeout	Timeout Scale
Uplink	500ms	N/A	Short	5
IST	N/A	10000	Long	3



Simple Loop Prevention Protocol (SLPP)

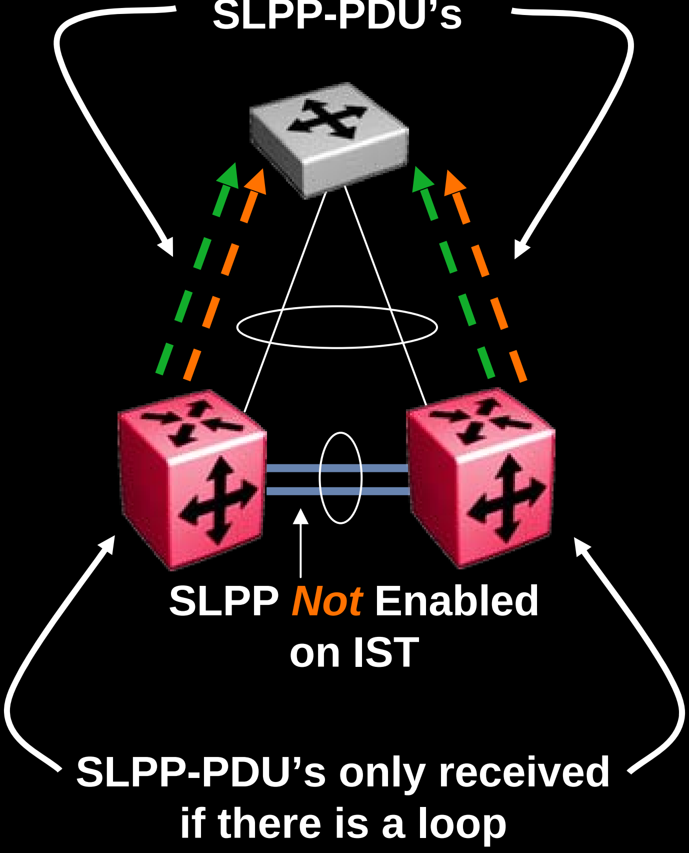
Feature Overview

- Prevents loops in a Switch Cluster network
  - Loops can occur when:
    - MLT at the edge is mis-configured
    - MLT not created at the edge but links are plugged in anyway
    - MLT configuration is lost (switch set back to factory default)
- SLPP uses an SLPP-PDU which is generated by the Switch Cluster cores
  - Loop detection is achieved by detecting whether the SLPP-PDU is received on the IST peer switch port or on the same switch where it originated
  - If the packet is received
    - The port is taken down
    - A log file entry is generated
    - An SNMP trap is sent
  - Once the port is down, it will stay in the down state and need manual intervention to be enabled
  - SLPP is enabled on a per VLAN basis and per port basis

Simple Loop Prevention Protocol (SLPP)

Feature Overview

- Enabling SLPP on a VLAN causes the switch to transmit the multicast SLPP-PDU – the packet is constrained to the VLAN on which it was sent
- The SLPP-PDU receiving and processing works only on ports where SLPP-Rx is enabled
- When SLPP-PDU receiving process works on the port which is a member of an MLT, all port members in that MLT will be taken down
- The SLPP-PDU can be received by the originated Switch or the IST peer Switch. All other switches treat the SLPP-PDU as normal multicast packet and will forward it on the VLAN
- SLPP threshold based on the sum of all packets received
- Port-based VLANs only
- Supported on:
  - ERS 8600, ERS 8300, ERS 5000



SLPP-PDU's

SLPP *Not* Enabled on IST

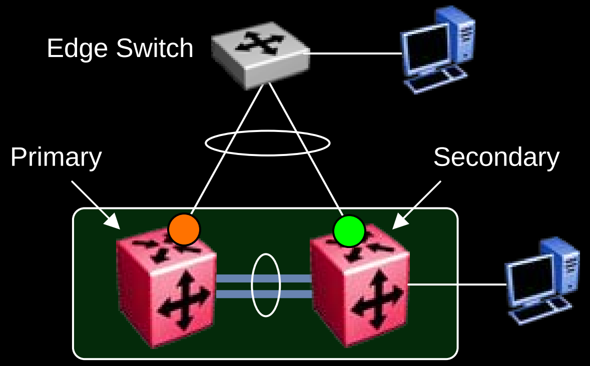
SLPP-PDU's only received if there is a loop

Switch Clustering Best Practices

Configuration Checklist

☑ SLPP enabled per VLAN and on uplink and access ports

- Enable SLPP
  - Per VLAN
  - Per port by setting Rx Threshold
- Identify one IST peer as Primary and the other as Secondary
  - Not a configurable option, strictly from a design standpoint
  - Enable RxThreshold per table below on uplink ports
- Do not enable auto recovery – Once the port is down, it will stay in the down state and need manual intervention to be enabled
- Do not enable SLPP-Rx on IST ports
  - Never want to take these ports down



● RxThreshold = 5

● RxThreshold = 50

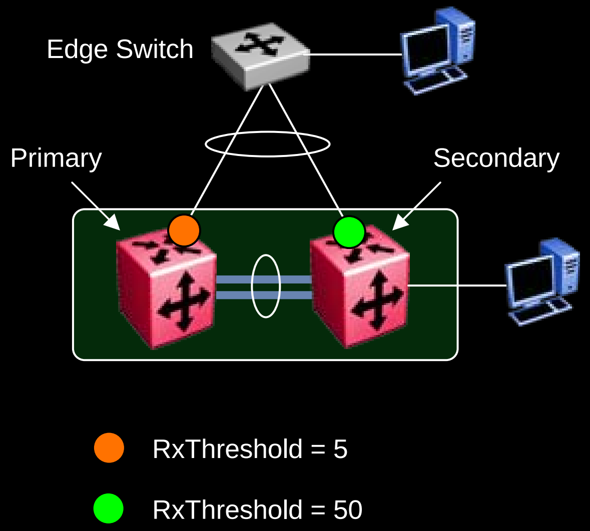
ERS 8600 Switch	Ethertype	Packet Rx Threshold	Transmission Interval
Primary	Default	5	Default (.5 seconds)
Secondary		50	

Switch Clustering Best Practices

Configuration Checklist

☑ SLPP enabled per VLAN and on uplink and access ports

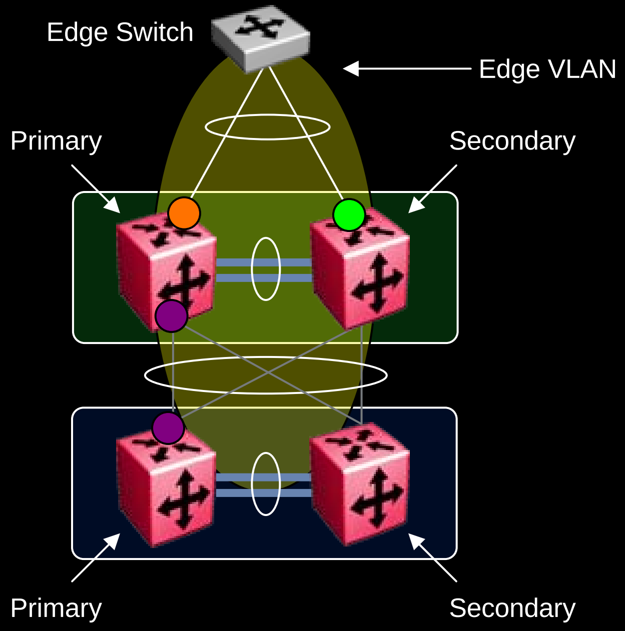
SLPP Rx-threshold is NOT reset upon any activity, but is a cumulative count. This can cause a situation in which multiple different loop events, can lead to an event where both primary and secondary links have their threshold reached and both links bring their ports down, and edge isolation could occur. A disable/enable of SLPP, which does not impact the network, should be performed after any SLPP event to clear the counter



As the number of VLANs running SLPP scale off of a specific uplink port, the Rx-threshold value may need to be increased to prevent complete isolation of the offending edge. Critical to note is that the primary goal of SLPP is to protect the core at all costs. In certain loop conditions, what may occur is the secondary switch also detecting the loop and its SLPP Rx-threshold is reached before the primary can stop the loop by taking its port down. Therefore, both switches eventually take their ports down and the edge becomes isolated. The larger the number of VLANs associated with the port, the more likely this could occur, especially for loop conditions that affect all VLANs. The recommended step here is to increase the Rx-threshold on the secondary only.
As a guideline, when the number of edge VLANs off of a specific uplink exceed 10, increase the secondary Rx-threshold to 100.

Switch Clustering Best Practices

Bridged Core



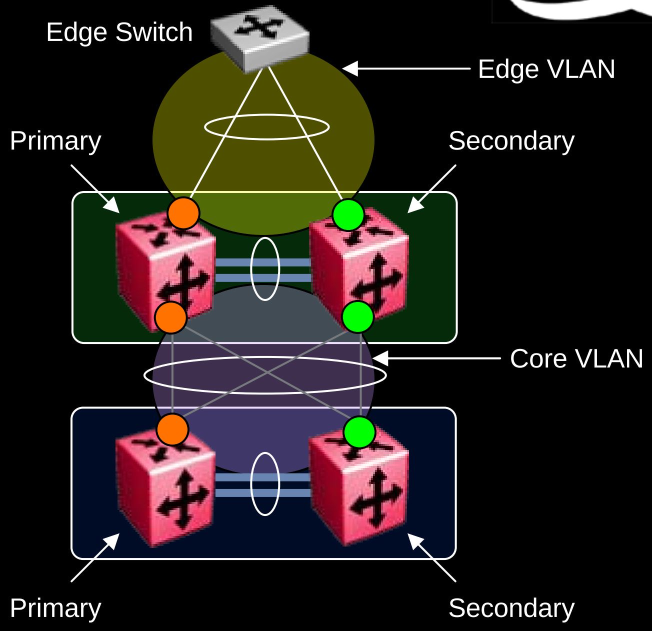
🟠 RxThreshold = 5

🟢 RxThreshold = 50

🟣 RxThreshold = 300

- Increase RxThreshold on the Switch Cluster core ports
- Loops at the edge should be caught at the edge and not shut down core ports

Routed Core



🟠 RxThreshold = 5

🟢 RxThreshold = 50

- Use/Create a VLAN that spans ONLY the routed core between Switch Clusters
- Enable SLPP on that VLAN to catch loops that occur within the core
- Loops at the edge will not cause ports to be shut down in the core

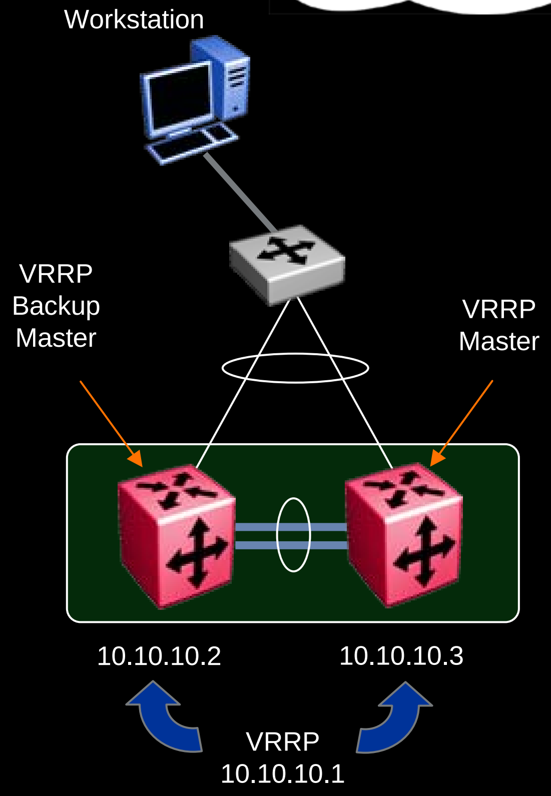
Switch Clustering Best Practices

Configuration Checklist

☑ Default Gateway Redundancy

VRRP with Backup Master

- Always use three addresses, two physical VLAN addresses and one virtual address – Do not use the physical IP address of the VLAN as the VRRP address
- Enable Backup Master on core switches
- Define VRRP Master by increasing VRRP Priority to 200 (100 is default – any value greater than 100 is acceptable)
- Balance VRRP Master between core switches across VLANs
- Set Advertise Interval to 10 seconds
- Set Hold-Down Timer to 60 seconds
- Supports up to 255 instances
- Do not use the same VRID across multiple VLANs
- See ERS 5000 Design Requirements section for other details on VRRP recommendations with ERS 5000 Switch Cluster



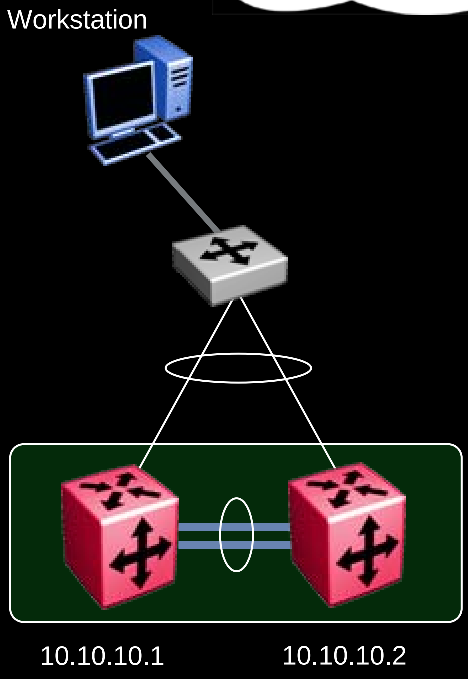
Switch Clustering Best Practices

Configuration Checklist

☑ Default Gateway Redundancy

RSMLT Layer 2 Edge

- Both IST peers can forward on behalf of each other
- Much less overhead than VRRP
- Scales beyond 255 instances
- Set Hold-Up Timer to 9999 seconds (infinity)
- Once both IST peers are up and running, must save the configuration file on each switch to ensure the peer’s MAC address is saved
- Do not use VRRP and RSMLT Layer 2 Edge on the same VLAN simultaneously
- Supported on:
  - ERS 8600
  - ERS 8300



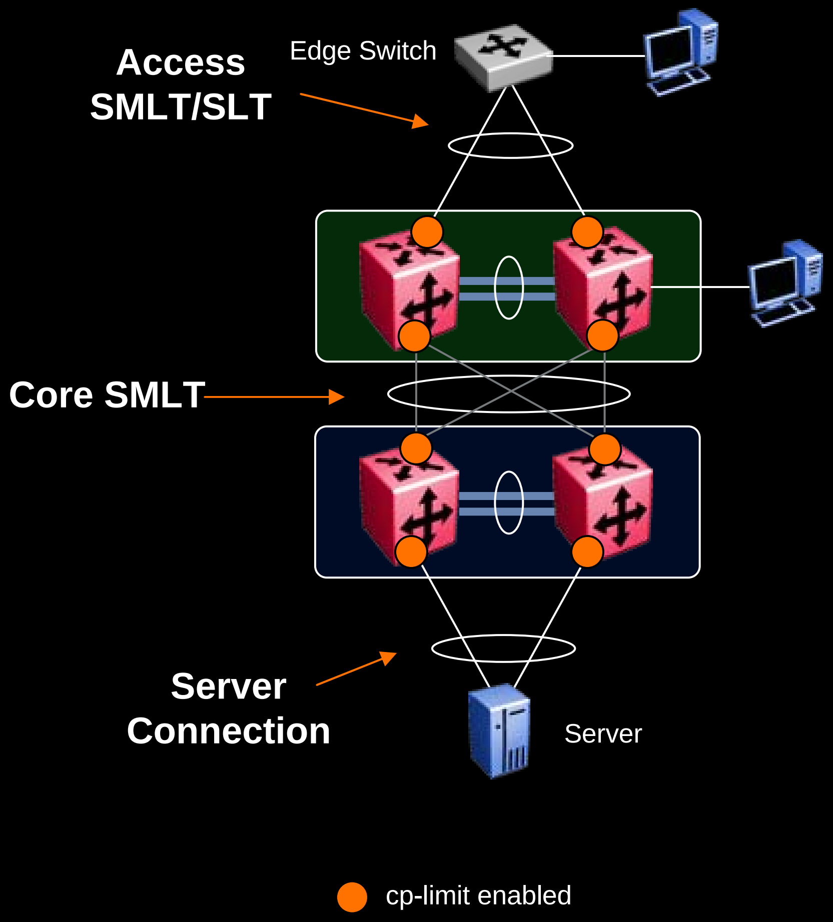
Control Plane Rate Limit (CP-Limit)

Feature Overview

- Protects CPU from broadcast and multicast storms
  - Looks at "**Control**" multicast traffic and broadcast traffic
    - Only packets destined to the CPU
  - If the defined packet rate per second is exceeded, the port is shut down
    - Need to disable/enable port to recover
  - Does NOT look at data packets (session/user traffic)
  - Does NOT protect against traffic exception traffic such as: SNMP, telnet, ICMP, IP with TTL1, Unknown SA, etc.
  - Enabled on all ports by default
    - Automatically disabled on IST ports during IST creation
- Supported on:
  - ERS 8600
  - ERS 8300
- Please note that with Release 6.0 for the ERS 5000 series, a CPU limiting feature was implemented, however, this is not a user configurable feature

CP-Limit Guidelines

	CP-Limit Values	
	Broadcast	Multicast
Aggressive		
Access SMLT/SLT	1000	1000
Server	2500	2500
Core SMLT	7500	7500
Moderate		
Access SMLT/SLT	2500	2500
Server	5000	5000
Core SMLT	9000	9000
Relaxed		
Access SMLT/SLT	4000	4000
Server	7000	7000
Core SMLT	10000	10000



Access SMLT/SLT

Edge Switch

Core SMLT

Server Connection

Server

● cp-limit enabled

Extended CP-Limit

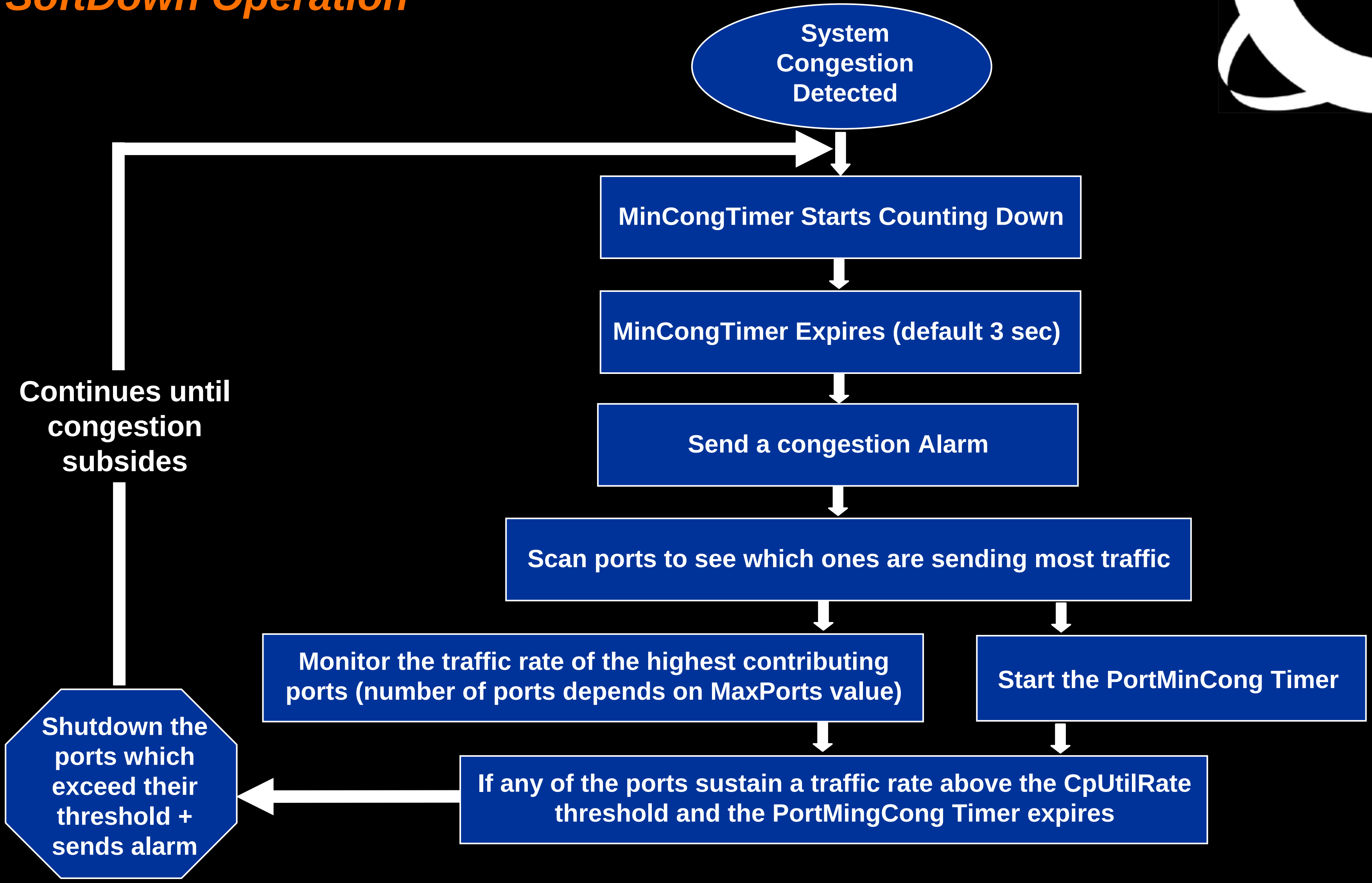
Feature Overview – ERS 8600 Only

- Can be used in conjunction with cp-limit and expands on the ability of cp-limit by monitoring
  - Buffer congestion on the CPU
  - Port level congestion on the I/O modules
- Does NOT look at data packets (session/user traffic)
- SoftDown monitors port for “x” duration, if congestion remains, port is disabled
- Can be enabled on all ports of the 8600 – the max ports value indicates the number of ports monitored during a time of congestion (ie. With maxports = 5, the 5 highest ports in terms of utilization are monitored)
- Enable SoftDown with the following values:
  - Maxports = 5
  - MinCongTime = 3 seconds (default)
  - PortCongTime = 5 seconds (default)
  - CPLimitUtilRate = Dependent on network traffic *

* Network must be baselined to understand the average utilization rate. Once the average rate is known, the CPLimitUtilRate should be set to a value of (3 * average utilization rate), but not higher than 70% under normal, average network conditions. This is a basic guideline for the “normal” enterprise network – for networks with normally high utilization, these values may differ.

Extended CP-Limit

SoftDown Operation



System Congestion Detected

MinCongTimer Starts Counting Down

MinCongTimer Expires (default 3 sec)

Send a congestion Alarm

Scan ports to see which ones are sending most traffic

Monitor the traffic rate of the highest contributing ports (number of ports depends on MaxPorts value)

Start the PortMinCong Timer

If any of the ports sustain a traffic rate above the CpUtilRate threshold and the PortMingCong Timer expires

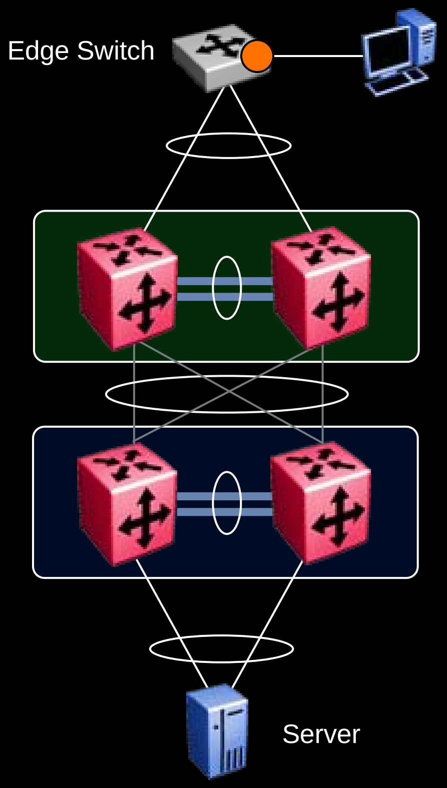
Shutdown the ports which exceed their threshold + sends alarm

Continues until congestion subsides

Broadcast / Multicast Port Rate Limiting

Feature Overview

- Enable Rate Limiting on the edge access ports to protect from broadcast/multicast storms
  - Protects against non-CPU bound traffic
  - Must understand multicast and broadcast traffic in the network before enabling rate limiting
- ERS Implementation
  - Broadcast / Multicast Rate limiting allows the user to configure the allowed amount of bcast/mcast traffic on a port. When traffic exceeds this threshold, it is dropped.
  - ERS 2500 / 4500 / 5000
    - 1 – 10% of port speed
    - Recommendation → 10%
  - ERS 8300
    - 1-100% of port speed
    - Recommendation → 10%
  - ERS 8600 (legacy, E-series, M-series modules)
    - Broadcast / multicast rate limiting
    - Allowed rate is in packets per second (pps)
    - Recommendation → 3 times normal pps
  - ERS 8600 (R-series, RS-series modules)
    - Broadcast / multicast bandwidth limiting
    - Allowed rate is in kbps
    - Recommendation → 3 times normal kbps



● Rate Limiting Enabled

Multicast and Switch Clustering

Supported Configurations & Features

- PIM-SM with Switch Clustering is supported on:
  - ERS 8600 (SMLT/SLT/RSMLT)
  - ERS 8300 (SMLT/SLT/RSMLT)
  - ERS 1600 (SMLT/SLT)
- PIM-SM and IGMP are NOT supported on the ERS 5000 Switch Cluster
- Enable PIM-SM on the IST VLAN (no unicast routing protocol is required) for fast recovery of multicast
- Enable IGMP Snooping and Proxy on the Edge switches when running multicast on the network
- When running PIM-SM over ERS 8600 Square or Full Mesh, enable mcast-smlt square-smlt flag
- For details on all supported topologies refer to:
  - Switch Clustering Supported Topologies and Interoperability (NN48500-555)
- For details on Multicast and Switch Clustering configurations refer to:
  - Resilient Multicast Routing Using Split-Multilink Trunking for the ERS 8600 (NN48500-544)

>BUSINESS MADE **SIMPLE**

NORTEL

Ethernet Routing Switching
8600
Engineering

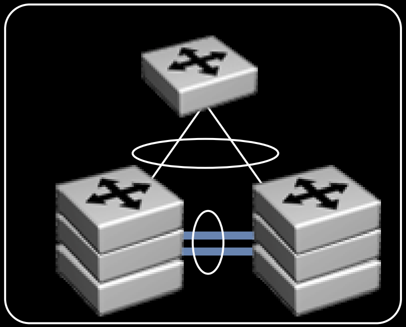
> **Resilient Multicast Routing Using Split-MultiLink Trunking for the ERS 8600 Technical Configuration Guide**

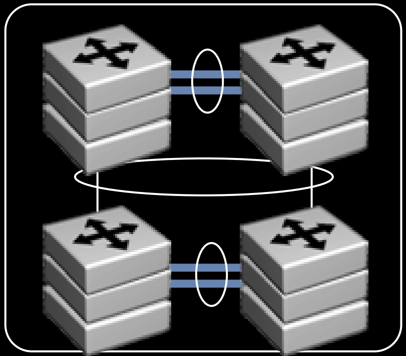
Core Systems Engineering
Document Date: January, 2008
Document Number : NN48500-544
Document Version: 1.0

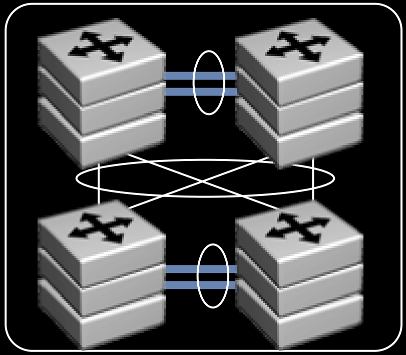
ERS 5000 Switch Clustering Design Requirements

ERS 5000 Switch Cluster

- No single point of failure in the core network
- Fast recovery when a link or a switch goes down
  - Sub-second recovery for L2 traffic in most cases
  - Sub-second recovery can be achieved in some cases for L3 traffic
- All redundant links are active – no Spanning Tree
- Switch Cluster supports
  - 1 IST
  - 31 SMLT
  - 512 SLT
- In 6.0 release, triangle, square, and full mesh are supported on both stand-alone and stack



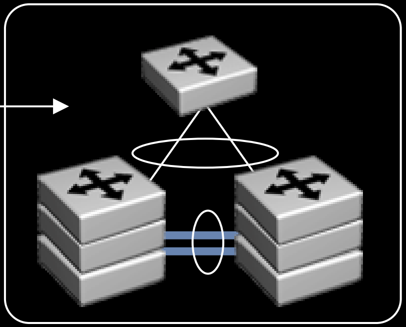




ERS 5000 Switch Cluster

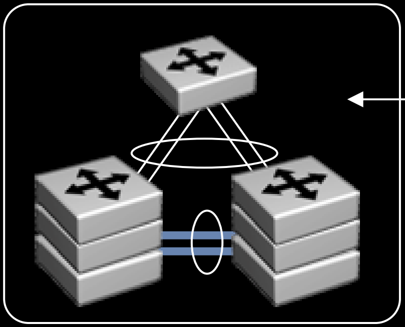
- IP Forwarding must be enabled
- STP must be manually disabled on IST / SMLT / SLT ports
- IST should be identical hardware, *but not an absolute requirement*
- LACP is not supported over SMLT/SLT
- IGMP is not supported over SMLT/SLT
- PIM-SM is not supported with SMLT/SLT or stacking
- SMLT must have at least two port members on the switch/stack, thus, you cannot create an SMLT with just one port on each pair, must be SLT

SLT



Maximum Two Uplinks per Edge Connection

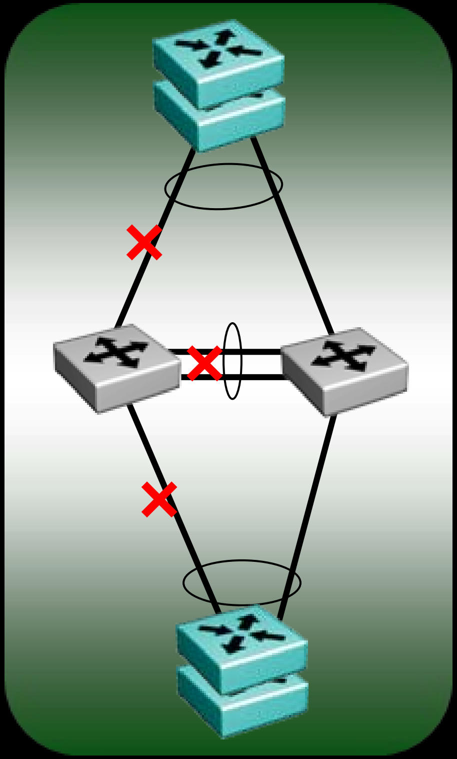
SMLT



Scaling Uplinks Between the Edge and the Core

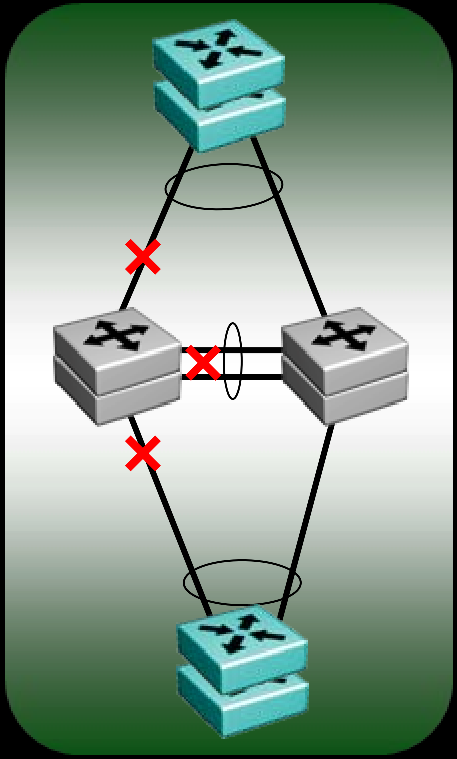
ERS 5000 Switch Cluster

- For the best resiliency, use at least three units when stacking the Switch Cluster Core



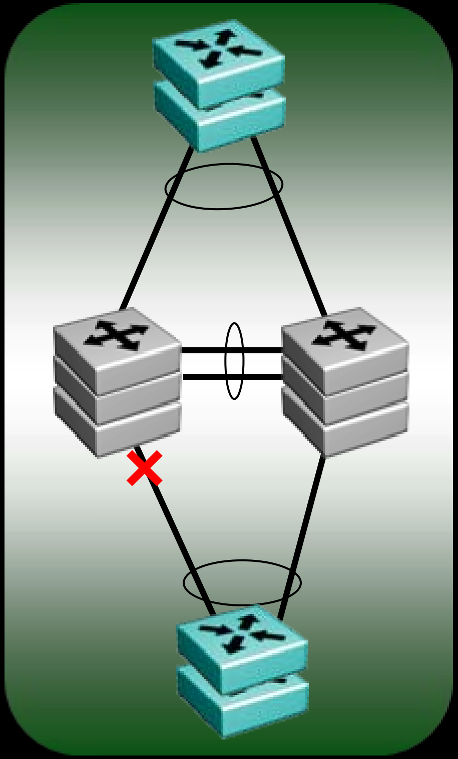
When a unit fails

- All SLT/SMLT/IST -> Down



When a unit fails

- All SLT/SMLT/IST -> Down

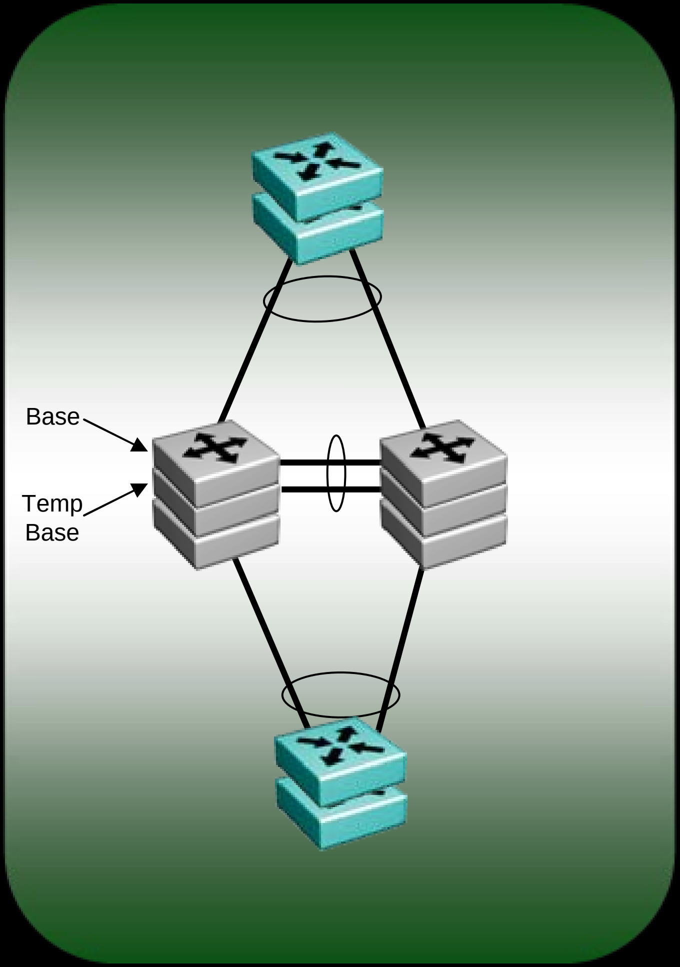


When a unit fails

- Only SLT/SMLT/IST on the failed unit -> Down

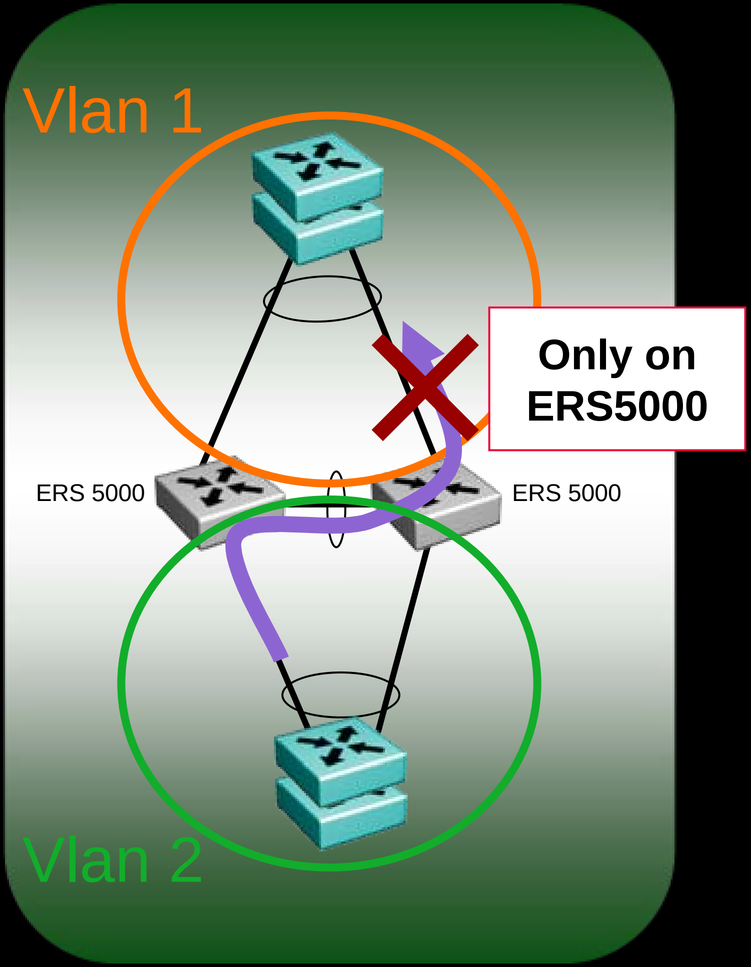
ERS 5000 Switch Cluster

- For the best resiliency, use at least three units when stacking the Switch Cluster Core
- When in stack configuration, at least one IST link should be on base unit (unit #1) and temporary base unit (unit #2). If more links are used in the IST, they can be spread across remaining units in the stack
- IST VLAN cannot be assigned as the management VLAN
- Cannot ping VRRP IP address from local CPU
- In a square or full mesh, aggregation pairs cannot use same VRID on the same VLAN



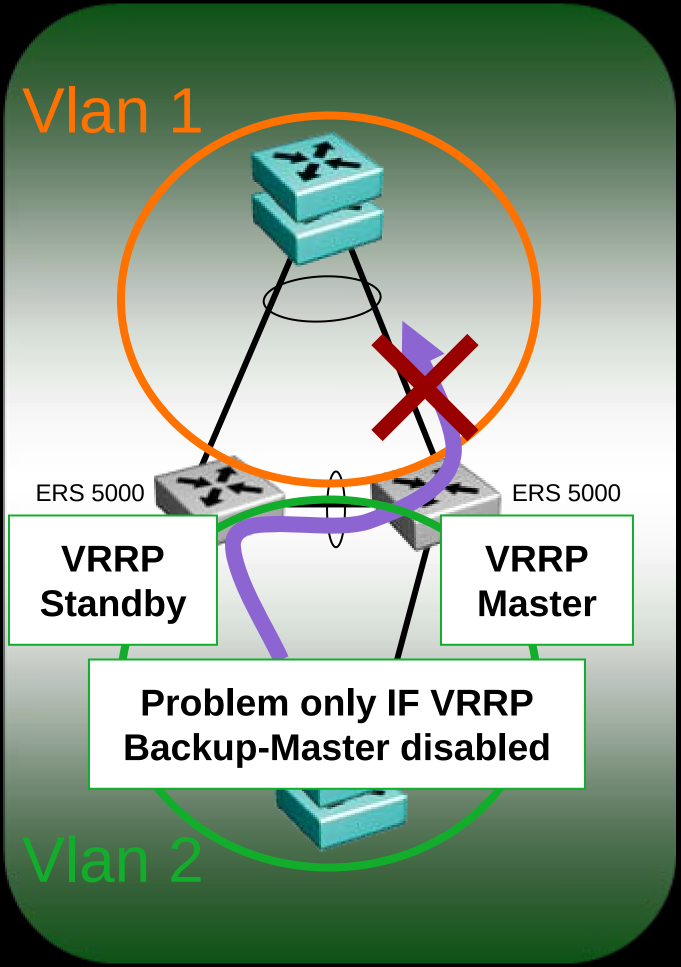
ERS 5000 Switch Cluster

- SMLT Loop prevention
  - The simple rule of SMLT is that any packet received on the IST CANNOT be L2 switched out of an SMLT or SLT port which is active; where active means that the corresponding SMLT link on the IST-peer switch is up and running
  - The ERS 5000 will also not forward L3 packets that traverse the IST out an SMLT or SLT port



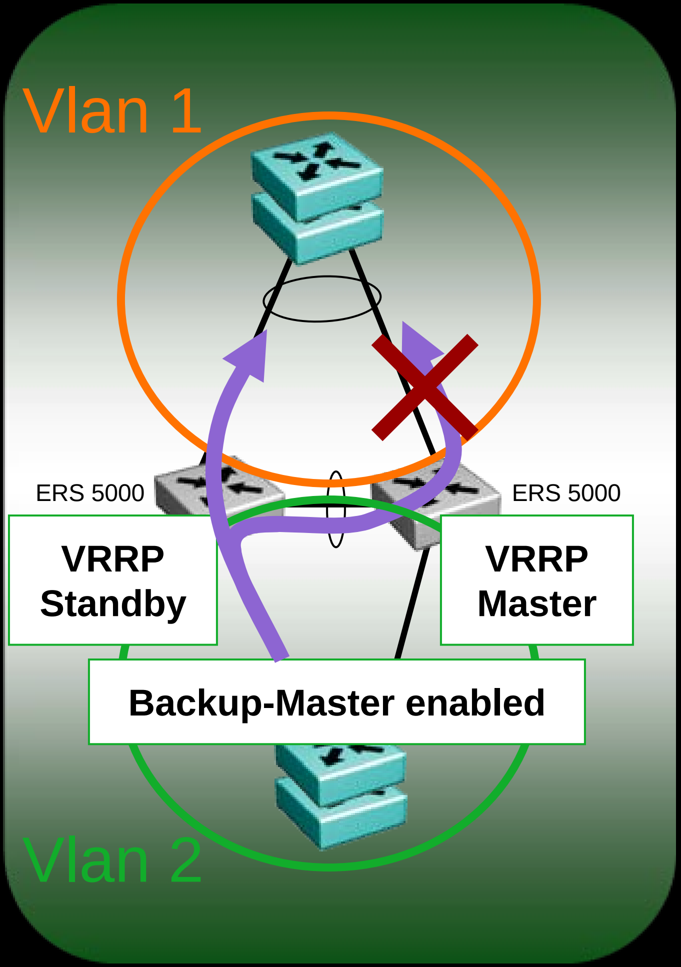
ERS 5000 Switch Cluster

- Where is this a problem ?
- VRRP
  - If VRRP Backup-Master is not enabled (or VRRP is disabled on one of the switches)



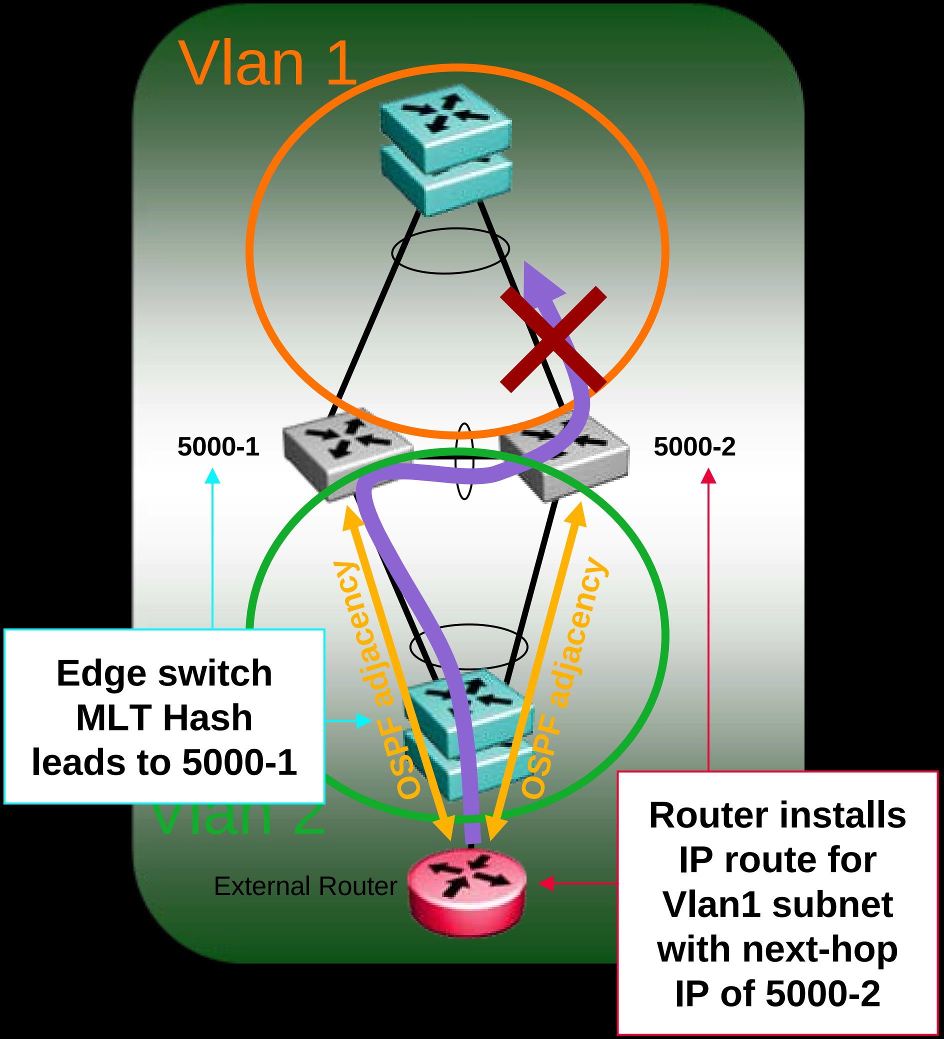
ERS 5000 Switch Cluster

- Where is this a problem ?
- VRRP
  - If VRRP Backup-Master is not enabled (or VRRP is disabled on one of the switches)
- Solution
  - Enable Backup-Master
  - Do not activate any VRRP functionality which prevents Backup-Master from being active
    - Set VRRP-Hold-down to 0
    - No Critical VRRP interfaces



ERS 5000 Switch Cluster

- Where is this a problem ?
- OSPF (also RIP)
  - If the edge router installs 5000-2 as next-hop in it’s routing tables and the edge switch MLT hashes routed traffic to 5000-1

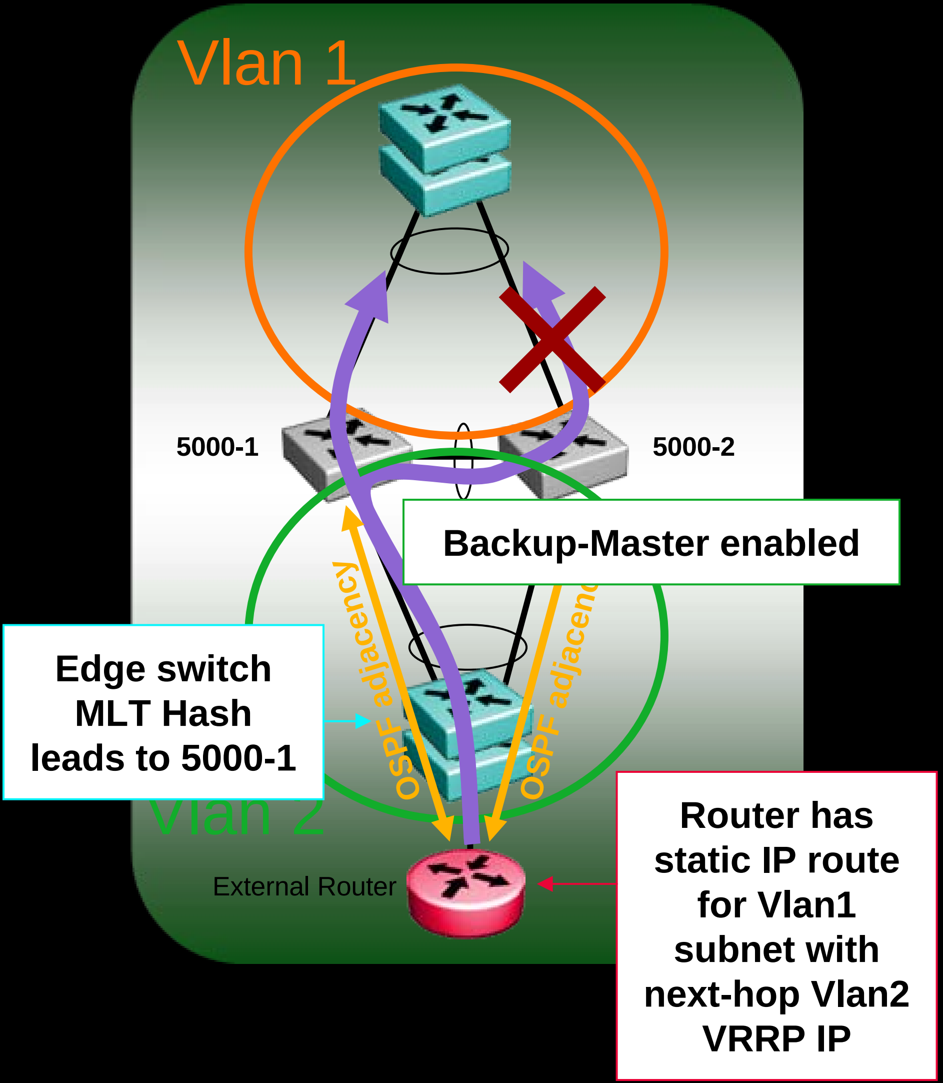


Edge switch MLT Hash leads to 5000-1

Router installs IP route for Vlan1 subnet with next-hop IP of 5000-2

ERS 5000 Switch Cluster

- Where is this a problem ?
- OSPF (also RIP)
  - If the edge router installs 5000-2 as next-hop in it’s routing tables and the edge switch MLT hashes routed traffic to 5000-1
- Solution Today - 1
  - Use Static Routes on External Router and point them to ERS5000 VRRP IP
  - VRRP Backup-Master required



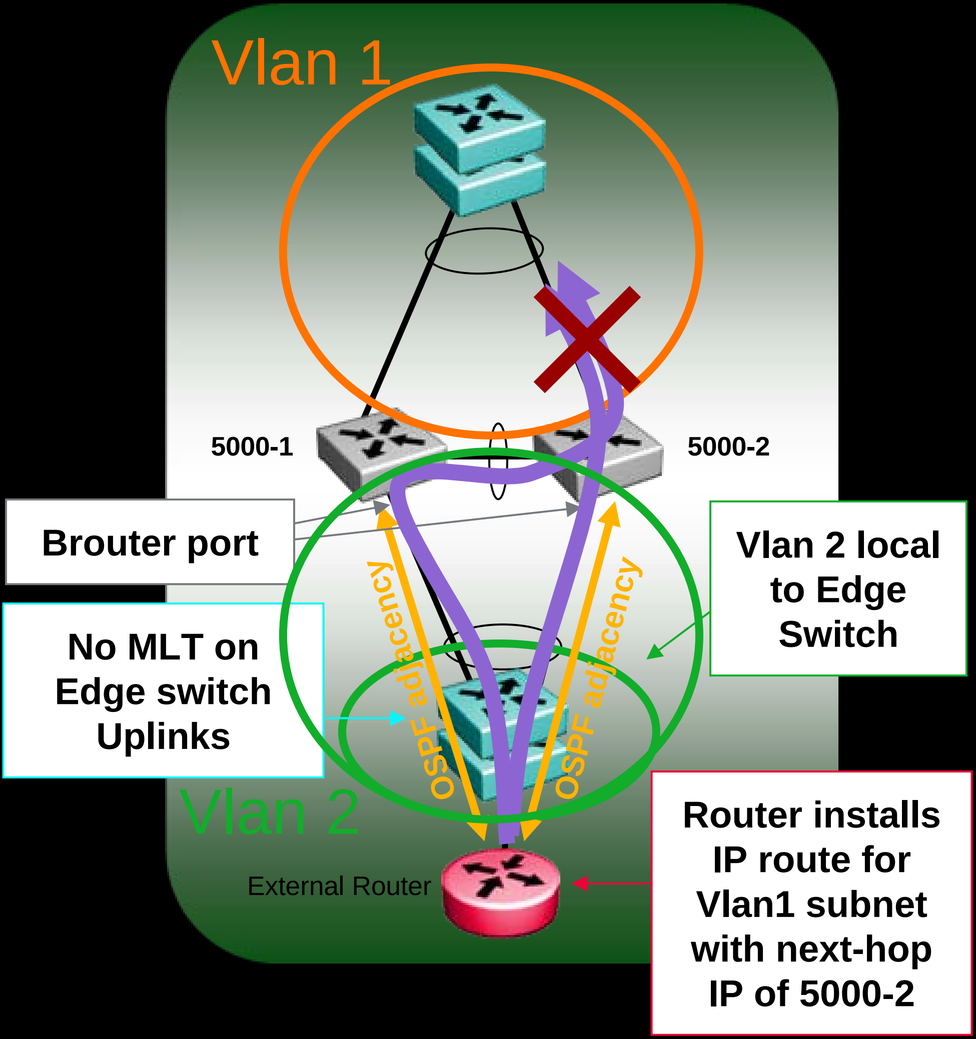
Backup-Master enabled

Edge switch MLT Hash leads to 5000-1

Router has static IP route for Vlan1 subnet with next-hop Vlan2 VRRP IP

ERS 5000 Switch Cluster

- Where is this a problem ?
- OSPF (also RIP)
  - If the edge router installs 5000-2 as next-hop in it’s routing tables and the edge switch MLT hashes routed traffic to 5500-1
- Solution Today - 2
  - If Static Routes not desired...
  - Connect edge switch with dedicated vlan (no MLT/SMLT connection)



Brouter port

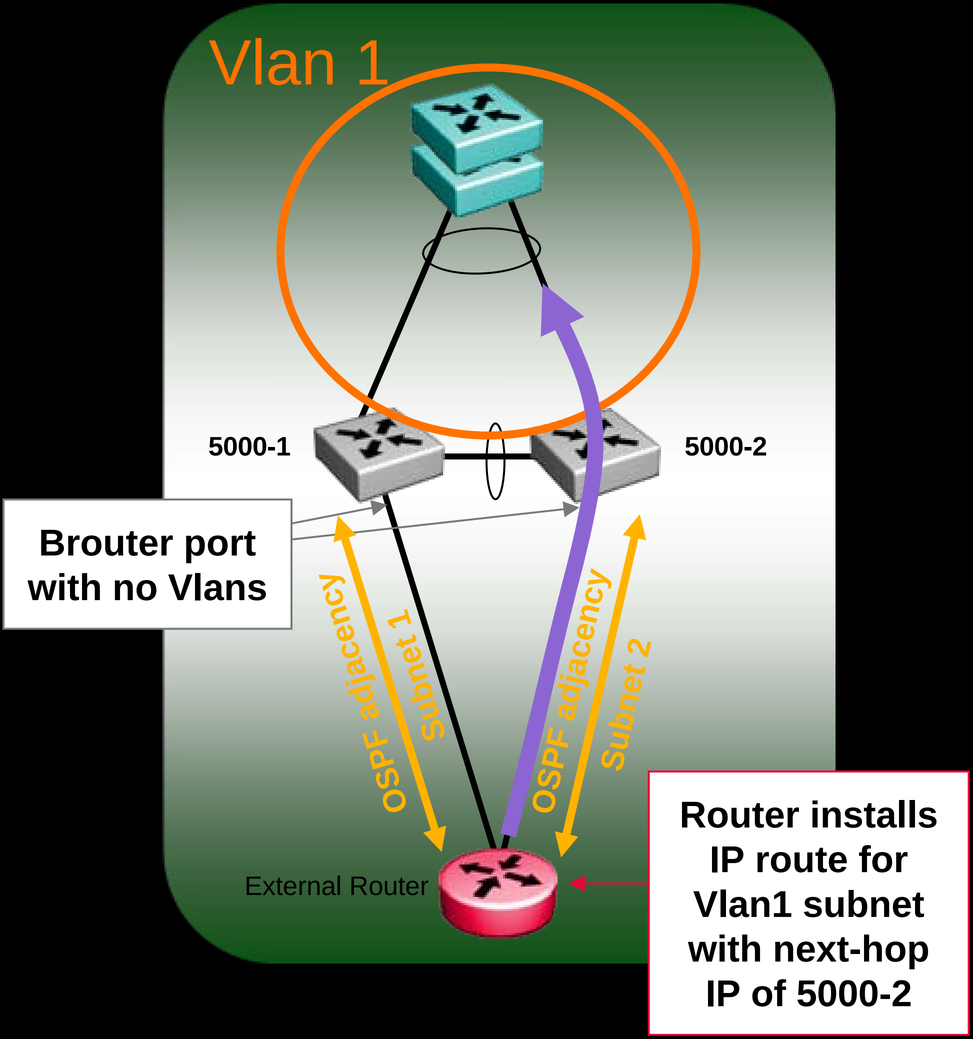
No MLT on Edge switch Uplinks

Vlan 2 local to Edge Switch

Router installs IP route for Vlan1 subnet with next-hop IP of 5000-2

ERS 5000 Switch Cluster

- Where is this a problem ?
- OSPF (also RIP)
  - If the edge router installs 5000-2 as next-hop in it’s routing tables and the edge switch MLT hashes routed traffic to 5000-1
- Solution Today - 3
  - If Static Routes not desired..
  - Connect External Router via dedicated routed links



Brouter port with no Vlans

Router installs IP route for Vlan1 subnet with next-hop IP of 5000-2

NORTEL

Business made simple