TechBrief

Leveraging Redundancy to Build Fault-Tolerant Networks

Introduction

The high demands of e-commerce and Internet applications have required networks to exhibit the same reliability as the public switched telephone network. Fault-tolerance and redundancy have become critical differentiators for networking equipment.

High availability networks must continue to operate when components fail unexpectedly and also during planned network upgrades and changes. Redundancy protects the network in both situations. By eliminating single points of failure network designers can create highly resilient networks for mission-critical applications.

But high availability networks require more than just redundant hardware. The network must also have the intelligence to optimize the use of those redundant components. Network software must take into consideration the impact of component failures on the Layer 2 and Layer 3 protocols that enable communications within a network.

Extreme Networks provides both the hardware redundancy and the intelligent software required by highly resilient networks for mission-critical applications.

Physical Redundancy

Extreme's high availability networking strategy begins inside of each BlackDiamond chassis switch and Summit stackable switch. The key components of an Extreme switch are replicated to ensure continued operation if a component fails. This redundancy includes power supplies and switch fabrics.

Dual Load-Shared Power Supplies

Extreme's BlackDiamond chassis switches are configured with dual load-sharing power supplies. Each of these power supplies has enough capacity to power the entire chassis, but in normal operation half of the required current is drawn from each power supply. A benefit of this load sharing is that the reduced current draw prolongs the life of the power supplies. If a power supply does fail, the other will support the chassis until a replacement can be fitted.

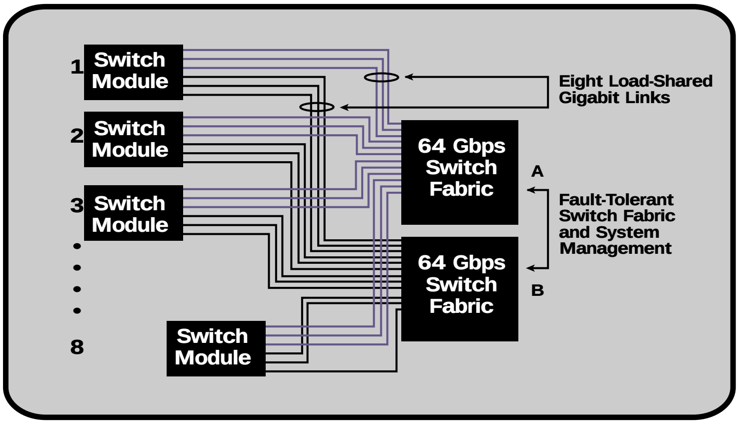
Another key benefit is that each power supply can be connected to a physically separate power source. Most data centers are wired with at least two separate circuits because power source failures are more likely than power supply failures. Some competitive products are configured with N+1 power supplies. This offers some redundancy, but the problem with this approach is that the individual power supplies cannot sustain the entire chassis. Therefore, if an input circuit fails, it causes the device to stop functioning, or stop supplying power to certain I/O blades in the device causing an unpredictable "brown out" situation.

Extreme's Summit stackable switches can optionally be connected to SummitRPS redundant power supplies that provide backup capabilities similar to the BlackDiamond chassis switch power supplies. Each SummitRPS provides redundant power supply support for up to two Summit switches.

Redundant Switch Fabrics

The switch fabric is essential to the operation of a switch because it handles the packet flows between ports. Because the switch fabric is critical to the operation of a switch, BlackDiamond switches are designed with fully redundant switch fabrics.

A unique feature of Extreme's redundant switch fabrics is that in normal operation both are fully active and the switch uses both the primary and secondary fabrics for data flow. The advantage of this design is that the switch is constantly using the secondary switch fabric. If the primary fails there is no doubt that the secondary is ready to take over.



In a BlackDiamond chassis, a single fabric provides four channels to each I/O slot and each channel is 1 Gbps, full duplex. The addition of a second switch fabric module doubles the number of channels to each slot from four to eight.

If the primary switch fabric fails, the switch will perform a soft reboot and continue operation on the other switch fabric. If the secondary switch fabric fails, the primary switch fabric will handle all of the traffic by itself, and notify management systems that the secondary switch fabric has failed. A secondary switch fabric failure does not require a soft reboot, and will result in zero downtime.

BlackDiamond switch fabrics are also easily replaceable and there are no active components on the backplane. This means that the failure of an active component, such as an ASIC or processor, will not cause the switch to stop functioning or function in an unpredictable manner. Competitive products utilize a crossbar switch fabric that includes active components. These devices contain very complicated ASICs that may fail under high heat or other conditions, causing the entire backplane of the device to be replaced, not to mention the complete failure of the device.

Hot Swappable Modules

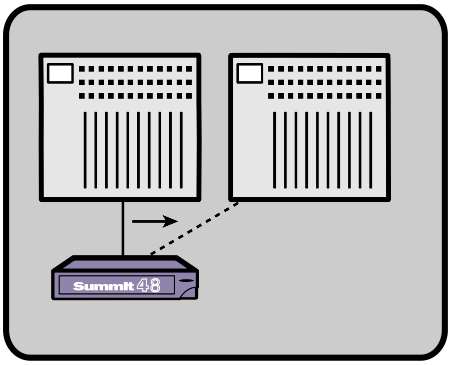
All modules in the BlackDiamond chassis, including the switch fabrics, are hot swappable. This enables modules to be replaced without rebooting or resetting the switch, and the replacement of one module does not affect the operation of any other modules. Hot swappable components enable the network to continue operation when a switch component fails or when switch configurations are changed.

Redundant PHY

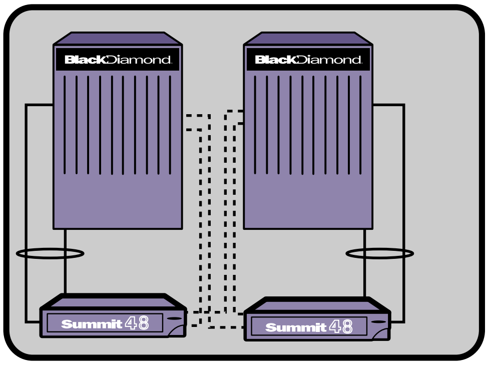
The first Layer 1 redundancy feature is the redundant PHY capability that is designed into the physical network interface of Extreme Networks switches. Redundant PHY is a protocol-independent IEEE 802.3z standards-based mechanism that enables a primary and secondary port to share a logical switch interface. Redundant PHY provides the industry's fastest failover to a redundant network path. This redundant port can be connected to the same switch as the primary, or for increased resiliency, it may be connected to a different switch, in dual-homed cores. No reconfiguration is necessary – just plug and play.

The primary and secondary or backup ports use the same logical switch interface but are presented as different physical ports on the switch. This provides a very simple, low cost way of implementing redundant links on a single switch port. Redundant PHY will make the secondary link active only if the primary link has failed. When a primary link fails the secondary link is activated almost instantly – in less than one second.

Failover is fast because failures are detected at the physical link layer and because network communication continues to use the same logical port – there is no need to find a new route or tree when a failure occurs. Additional control over failover and failback characteristics is offered by Extreme's SmartRedundancy™. This feature controls bring-up and failback characteristics of redundant PHYs by using only the primary if it is available. Fail back from the redundant link to the primary link occurs automatically once the link is restored. This feature is also used in conjunction with link aggregation as outlined below.



In this diagram, the redundant port can be connected to the same switch as the primary, or for increased resiliency, it may be connected to a different switch in dual-homed cores.



SmartRedundancy provides additional control over failover and failback characteristics. In this diagram, SmartRedundancy is used in conjunction with link aggregation.

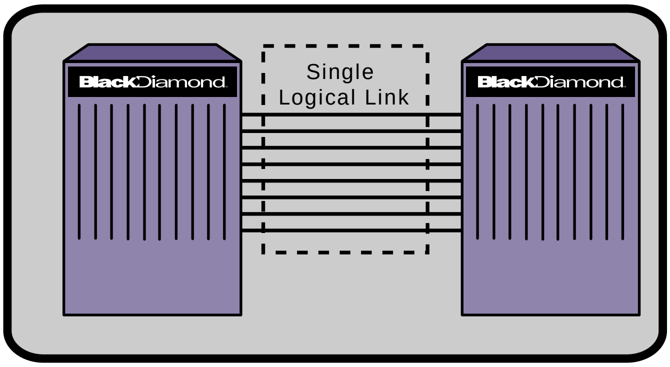
Extreme Networks Wiring Closet Solutions

Physical and mechanical redundancy is critical to building a reliable network, but hardware component reliability is just the beginning. Extreme switches also include several features that enhance Layer 2 resiliency.

Link Aggregation

Link aggregation (or load-sharing) is another Layer 2 redundancy feature that increases both reliability and performance. Extreme's link aggregation is based in part on the proposed IEEE 802.1ad standard. This feature allows as many as 8 physical links to function as a single logical link. A key benefit of link aggregation is that it provides very fast, sub-second failover if a physical link fails or is removed from service.

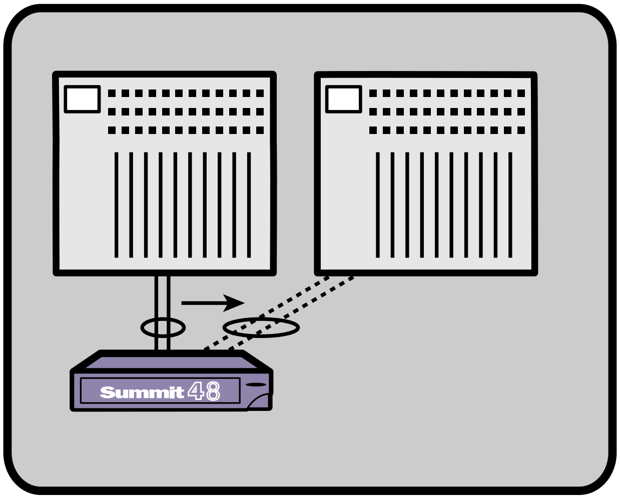
Link aggregation has the added benefit of fully utilizing the bandwidth of up to eight links because traffic is distributed across all links. This simultaneously increases performance and the reliability of the network. You may also choose from various algorithms to distribute traffic across the links. Distribution may be based on port, Layer 2 - 4 session-based address information or round-robin criteria.



Link aggregation provides very fast, sub-second failover if a physical link fails or is removed from service. Link aggregation has the benefit of fully utilizing the bandwidth of up to 8 links because traffic is distributed across all inks.

Combined Link Aggregation and Redundant PHY

Using SmartRedundancy it is possible to combine link aggregation and redundant PHY capabilities such that ports configured for link aggregation coordinate their failover capabilities with their use of redundant PHYs. For example, if ports 49 and 50 of a Summit48 form a load-shared link to another switch, a failure of one of the active ports will result in both ports utilizing the redundant PHYs in their failover thus preserving the integrity of the point-to-point load-shared link. Just as importantly, should the original link be restored, the switch will roll back to the configured primary on both ports.



If ports 49 and 50 of a Summit48 form a load-shared link to another switch, a failure of one of the active ports will result in both ports utilizing the redundant PHYs in their failover thus preserving the integrity of the point-to-point load-shared link. Just as importantly, should the original link be restored, the switch will roll back to the configured primary on both ports.

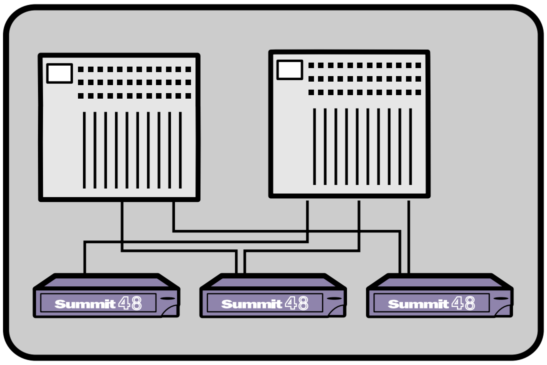
Spanning Tree

Spanning tree is the industry-standard Layer 2 protocol that provides a redundant loop-free topology. When a switch has spanning tree enabled, each port is in one of two modes, either forwarding or blocking. When a link in the spanning tree breaks, a new spanning tree is computed and the switch ports change state as needed to create a new spanning tree that re-establishes full connectivity.

Extreme switches can support up to 64 instances of spanning tree. This is far more than required by most modern networks because other techniques, such as Layer 3 routing, are typically used to build resilient networks. Another aspect of Extreme's implementation is the use of spanning tree "Domains" whereby a single spanning tree instance may be used to protect multiple VLANs simultaneously. This reduces the overhead of running a "spanning tree per VLAN" and allows far more flexible configuration.

A key limitation of spanning tree is its slow convergence. Networks can take up to 30 seconds to converge after a link failure. Also, network bandwidth is not used efficiently because the blocked links are not available to carry traffic. In comparison, Layer 3 switching, which is true routing, makes full use of network bandwidth. For these reasons, Extreme Networks recommends that customers use other techniques to provide resilience, such as Layer 3 routing protocols, link aggregation, or the Extreme Standby Router Protocol™ (ESRP™) whenever possible.

Extreme Standby Router Protocol



ESRP offers integrated and coordinated Layer 2 and Layer 3 redundancy, eliminating the need for separate solutions such aspanning tree and VRRP.

ESRP is a unique feature of Extreme's switches that offers integrated and coordinated Layer 2 and Layer 3 redundancy. This eliminates the most common problem of redundant switch/router implementations that require separate, independent solutions such as spanning tree and VRRP. ESRP's Layer 2 redundancy can be used without routing as an attractive alternative to spanning tree because its failover time is in the range of 2-6 seconds – much faster than spanning tree. ESRP, which is explained in more detail later in this paper, can be used in Layer 2-only environments or with Layer 3 switching.

Virtual Chassis Mode or Layer 2 Clustering This feature allows a cluster of switches to act as a single "chassis" while using multiple active dual-homed gigabit links between them. The "core" allows use of multiple VLANs with intelligent switching and loop prevention.

Layer 3 Redundancy

One of the key advantages of Layer 3 networks is that packets are forwarded along multiple routes that are dynamically updated to reflect the current network topology. If a failure occurs the network automatically re-routes traffic around the failure. The dynamic topology updates are distributed by routing protocols that automatically update routers and switches as the state of the network changes. Extreme supports the key industry-standard unicast routing protocols, Routing Information Protocol (RIP), Open Shortest Path First (OSPF), Border Gateway Protocol v4 (BGP4) for both Exterior Border Gateway Protocol (EBGP) and Interior Border Gateway Protocol (IBGP). For Layer 3 IP multicast traffic redundancy, DVMRP, PIM Dense Mode and PIM Sparse Mode are also supported.

When using Extreme switches there is no performance penalty for using this Layer 3 functionality. Extreme Networks switches can route as fast as they can switch at Layer 2 – at wire speed.

Routing Information Protocol

RIP is an early industry-standard routing protocol that can be used to build small IP networks. It also provides compatibility with legacy equipment that supports RIP. A key limitation of RIP is that it is slow to converge after a change to the network topology occurs. It sometimes requires up to two minutes for a network that uses RIP to converge.

RIP support is standard on all Extreme Networks switches – from BlackDiamond chassis switches to Summit stackable switches. This gives customers a very cost-effective option for implementing Layer 3 routing and compatibility with legacy equipment.

Open Shortest Path First (OSPF)

OSPF is a more modern link-state routing protocol that distributes routing information between routers within a single IP domain, also known as an autonomous system. When using a link-state routing protocol, each router maintains a database describing the topology of the autonomous system. Each participating router has an identical database maintained from the perspective of that router.

From the link-state database, each router constructs a tree of shortest paths, using itself as the root. The shortest path tree provides the route to each destination in the autonomous system. When several equal-cost routing paths to a destination exist, traffic can be distributed among them. The relative cost of each route is described by a single metric.

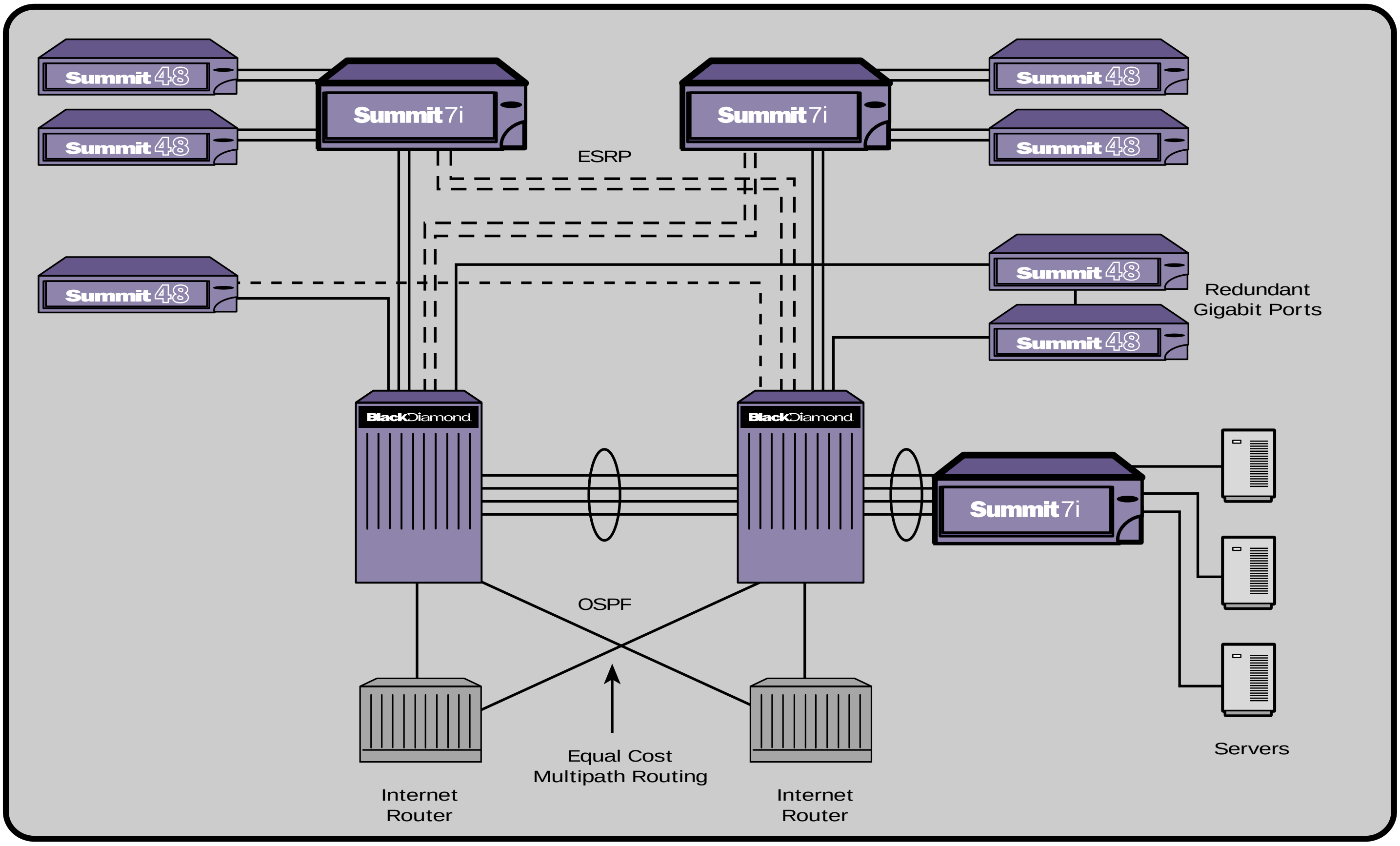
In addition to multivendor interoperability, a few of the benefits of using OSPF for building resilient networks are fast convergence, Equal Cost Multipath (ECMP) routing and a hierarchical design approach.

OSPF converges quickly after a failure. The routing table of each router is calculated from the OSPF link-state database. In a stable network, updates are sent at predetermined times but when there is a change in the network, the link-state tables are updated immediately through flooding. This ensures that each router has an accurate picture of its area. Events that would trigger an update include changes to the state of a router or an adjacent link.

If a network is configured correctly, OSPF convergence can take place in less than a second.

Equal Cost Multipath Routing

Another advantage of OSPF is that it can be used to enable Equal Cost Multipath (ECMP) routing, which distributes traffic across multiple paths with equal costs. This efficient routing mechanism increases network performance by using multiple paths simultaneously. If a path fails, traffic is redirected to an alternate path.



OSPF can be used to enable Equal Cost Multipath (ECMP) routing which distributes traffic across multiple paths simultaneously. If a path fails, traffic is redirected to an alternate path.

Finally, OSPF's hierarchical design allows parts of the network to be grouped together into areas. Link-state databases and topology updates are kept local. This significantly reduces link-state advertisement traffic and reduces the computations required to maintain the routing tables. In addition, since topology information is kept local to an area, a network problem in one area will not affect the other areas.

BGP4 and the additional extensions Extreme offers contain several redundancy and route advertisement control options for use by service providers and customers wishing to "peer" with single or multiple service providers for redundancy purposes. Some of these features are:

- Multi-exit discriminator (MED)
- Import/export filtering on peer, route, AS-Path
- Route mapping
- Use of "communities"
- EBGP Multihop capabilities
- Route damping

Intelligent Management of Redundant Routers

In Layer 3 networks with redundant routers, host systems must be kept aware of the status of the routers with which they interface with. IP hosts use default gateway routers to communicate with hosts in other networks, subnets or VLANs. Because the failure of such a gateway isolates hosts from the rest of the network, mission-critical networks should be configured with redundant gateways. But simply adding another gateway is not enough. The network must have the intelligence to use those gateways to maximum advantage. Network software must first decide which of the routers will provide the highest level of service and then it must dynamically connect hosts to that gateway router.

When redundant gateways exist, hosts must be able to determine the IP address of the currently active gateway. One approach is for hosts to use a protocol such as the ICMP Router Discovery Protocol (IRDP), Routing Information Protocol (RIP) or Proxy ARP to dynamically discover routers. Though all these methods are supported by Extreme, these approaches have some key limitations. Most hosts do not implement these protocols and even when they are implemented they may be slow to converge from primary to backup gateway.

The majority of hosts are configured statically or through the use of the Dynamic Host Configuration Protocol (DHCP) with the IP address of their default routers. To extend the benefits of router redundancy to these hosts, the routers themselves must transparently handle the switch from primary to backup when a router fails. Several solutions exist that coordinate router failovers, including the Extreme Standby Router Protocol (ESRP), Virtual Router Redundancy Protocol (VRRP) and Hot Standby Routing Protocol (HSRP).

The general approach is to make two or more routers function as a single virtual router with a single IP and MAC address. Hosts are then configured with the IP address of the virtual router rather than a physical router. This virtual router address is always assigned to the currently active physical router.

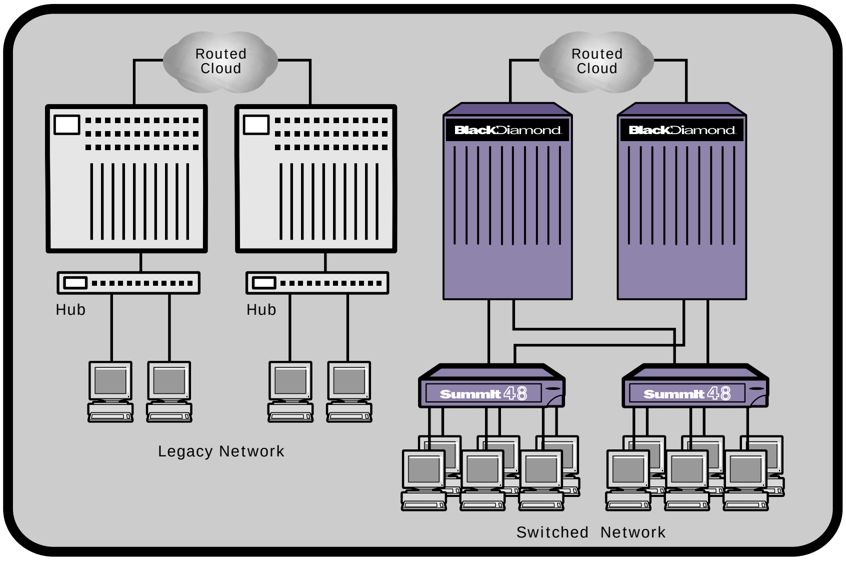
ESRP vs. VRRP and HSRP

VRRP is a proposed IETF standard that enables a group of routers to function as a single, virtual default gateway. The key limitation of VRRP and HSRP, which are functionally similar, is that they were designed to support older shared-LAN environments, not today's switched LAN infrastructures.

HSRP also includes an extension that is designed to compensate for the performance limitations of software routers. It allows routers to be configured as both primary and secondary. Some clients are configured with one router as their default gateway and the others point to the other router. This balances the load between the two routers but it also complicates client administration. This manual load balancing and its administrative burden are unnecessary with Extreme switches because they route at wire speed.

The diagram below compares a shared legacy LAN with a switched network.

VRRP and HSRP were designed for networks made up of hubs and legacy routers. Today's networks use intelligent switches and routers, which may require multiple router ports to connect with a single subnet. Dual-homed configurations are commonly used to provide high availability by eliminating single points of failure.



Legacy networks commonly use complex redundancy solutions such as VRRP and HSRP. Today's switched networks use intelligent switches and routers, which may require multiple ports to connect with a single subnet. Dual-homed configurations provide high availability by eliminating single points of failure.

VRRP assumes that if any link to a subnet is active, the router has access to the entire subnet. While this assumption may have been valid for legacy shared LANs, it does not result in a good decision in a switched LAN environment. In this example the hosts connected to the switch on the left are isolated from the network. Half of the hosts in this subnet are now inaccessible, yet VRRP does not initiate a switch to the backup router that has connectivity to the entire subnet.

Extreme's ESRP makes a better decision when a link failure occurs. ESRP is based on VRRP but it includes some extensions that enable it to make more intelligent failover decisions in today's switched networks.

An important ESRP extension counts the number of links active in an ESRP VLAN and determines primary/backup status based on this metric, always keeping the maximum number of connections as the master device. In this scenario, ESRP would determine that the backup router has more active links to the VLAN and would therefore initiate a failover to the backup router. This decision preserves connectivity to the entire VLAN. The failover time for ESRP is very short, in the range of 2-6 seconds.

ESRP can also use the status of a set of learned routes or a router uplink port as a metric when making a failover decision. When an uplink failure occurs on an active router, ESRP initiates a failover to the backup router that has an active uplink.

VRRP routers advertise the existence of subnets even when they are in backup mode. In this example, a link failure has isolated some hosts from the backup router. However, because the router continues to advertise the subnet, it may receive traffic that it will not be able to forward to those hosts.

Displaying far more intelligence, ESRP "understands" that the reason that a router is not active may be due to the fact that it has only partial subnet connectivity. Only active ESRP routers advertise their subnets to maximize the probability that inbound traffic will be delivered.

Since redundancy requirements often vary, flexibility is offered with ESRP by configuring which factors and the precedence of factors that are used in determining an ESRP failover.

	ESRP	VRRP	HSRP
Layer 3 Redundant Router	Yes	Yes	Yes
Layer 2 Redundancy / Loop Prevention	Yes	No	No
Automatically Suppress Standby Router Advertisement	Yes	No	No
Track VLAN Uplinks	Yes	No	Yes
Track Layer 3 Learned Routes	Yes	No	No
Downstream Switches are "Aware" of Failover	Yes	No	No
Configurable Content and Precedence of Failover Criteria	Yes	No	No

Extreme's ESRP offers superior redundancy and failover protection.

Another key feature that differentiates ESRP from VRRP and HSRP is that ESRP can also can be used without routing. Because ESRP has built-in Layer 2 redundancy, it can be used as a spanning tree alternative with a convergence time of 5-8 seconds using default parameters. ESRP can be utilized in a Layer 2-only environment to provide a standby Layer 2 switch in case the primary Layer 2 switch fails. This failover time is very short and is much faster than spanning tree.

Another key requirement of redundant configurations is correct handling of intra-subnet traffic of downstream Layer 2 switches in the event of failover. An all-Extreme implementation enables downstream Layer 2 switches to be "aware" that an upstream ESRP failover has occurred. Without this, downstream switches would continue to forward intra-subnet traffic to the wrong destination port until the Layer 2 forwarding database timers expire (typically 5 minutes) or are refreshed. The "ESRP Awareness" built into all Extreme switches will automatically detect an upstream ESRP failover and correctly direct all subsequent inter- and intra-subnet traffic.

Device Management Redundancy

Equally important in building redundant networks are the management and configuration aspects of the devices within the network and their interaction. All Extreme switches may contain two separate configurations and two separate software images for smooth transitions during moves, adds and changes in a network. This allows quick rollback in the event of a problem in either configuration or upgrades. Configuration management is also aided by:

- readable ASCII-based config files
- automatic timed configuration uploading
- download incremental configuration changes without rebooting
- client support for multiple/redundant Syslog, RADIUS, SNTP and DNS servers
- local and remote logging (Syslog) of all configuration changes

Summary

Extreme Networks provides the redundancy needed to build mission-critical networks. This redundancy extends to all layers of a network. Redundant components increase the reliability of each switch while the ExtremeWare software suite provides the routing protocols and standards needed to intelligently use redundant switches in both Layer 2 and Layer 3 networks.

Networks based on Extreme switches not only heal themselves, they can converge so quickly that users are often unaware that an outage has occurred.