



High Availability in JUNOS Platforms

Agenda

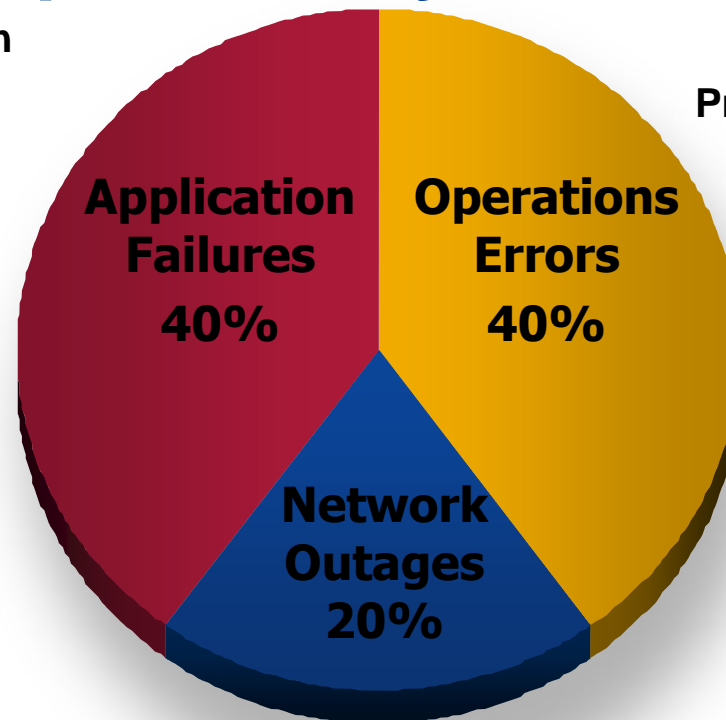
- **Introudction**
- **Reliable software architecture**
- **Reliable hardware architecture**
- **JUNOS HA features**

Threats to Dependability

Misconfiguration

DNS faults

DHCP faults



Process failures

Platform failures

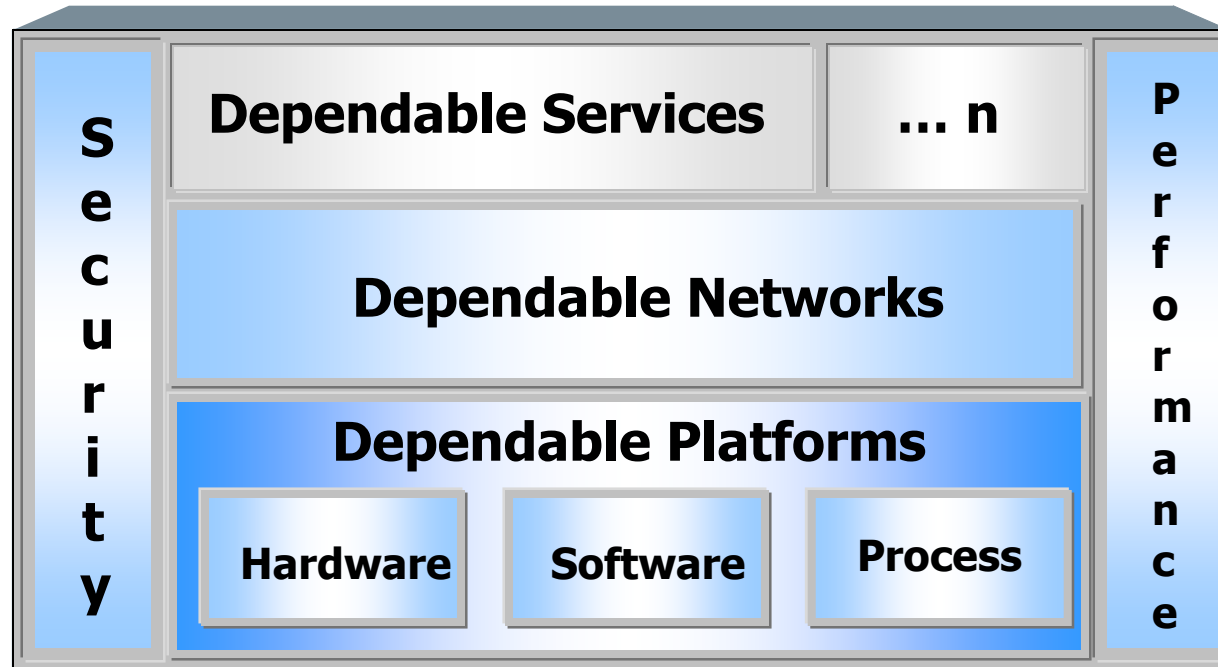
Hardware/software

Circuit failure

Source: Gartner Group, 1/01

Solution Architecture

IP Carrier-Class Availability Is a Culture, Not a Single Feature or Product



Reliable Hardware

■ Hard Fault Tolerance

- Environmental sensors
- Component redundancy
- Redundant boot devices

■ Soft Fault Tolerance

- Extensive internal diagnostics
- CRC-protected internal data paths
- ECC SDRAM

■ MTTR Reduction

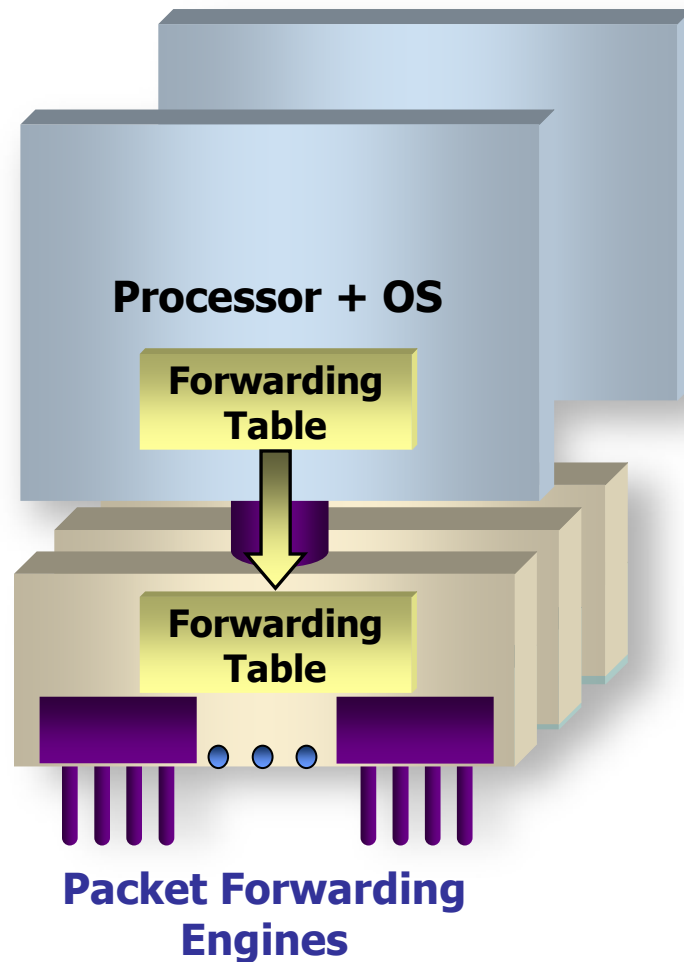
- Hot swappable components
- Field replaceable components



Why?

- **HA is a requirement in Core/Edge/Multi-service networks.**
- **Stability has been a major differentiation from our competitor. Some thought from happy customers.**
 - JUNOS routers don't require a reboot for more than a year in my network.
 - Don't know how to debug. We haven't met any problem.

JUNOS – Reliable Platform Architecture

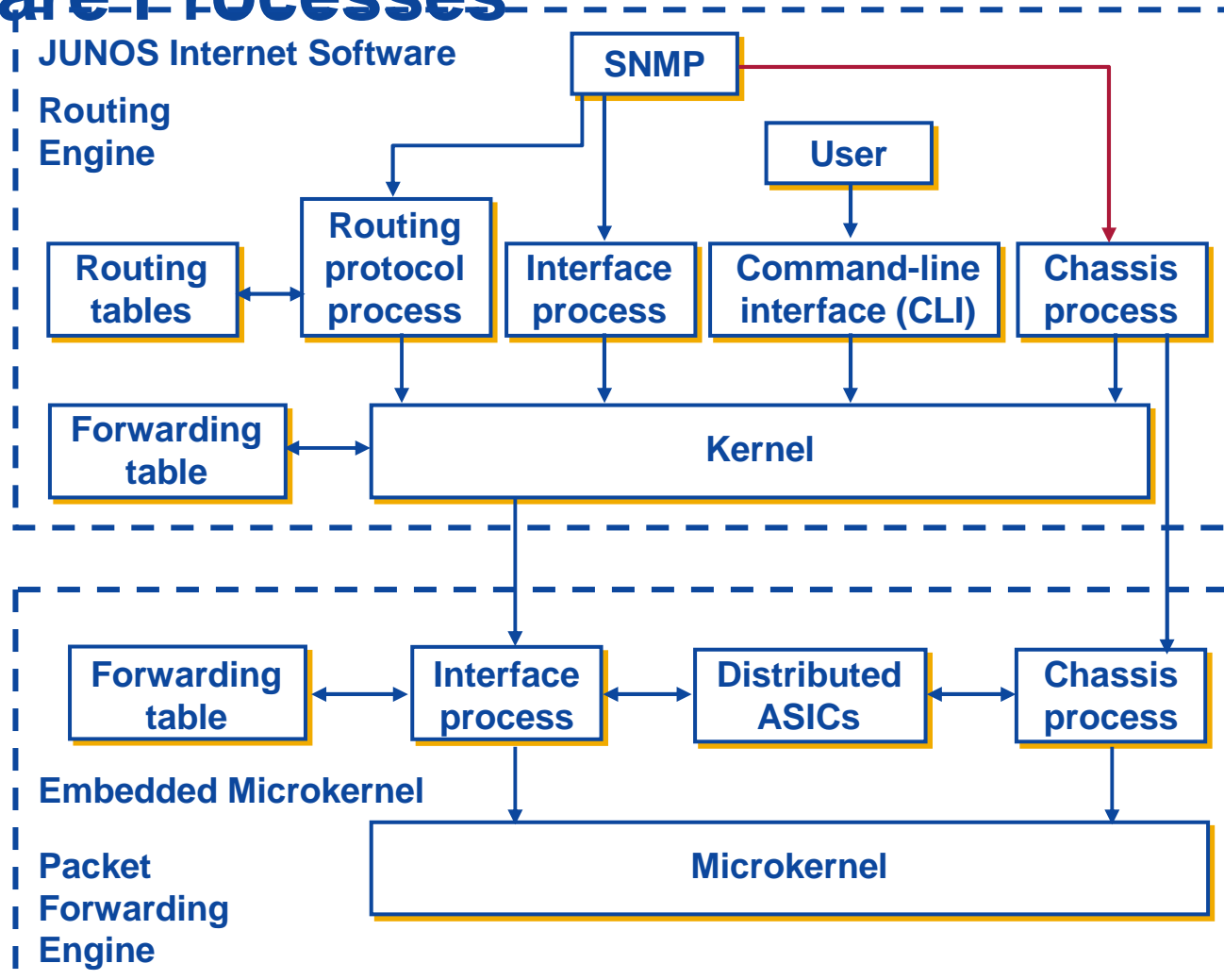


**Clean separation ensures
that processes are
protected from each
other, resulting in Higher
Reliability**

**Foundation for reliable
Software**

**Foundation for Graceful
Routing Protocol
restart operation**

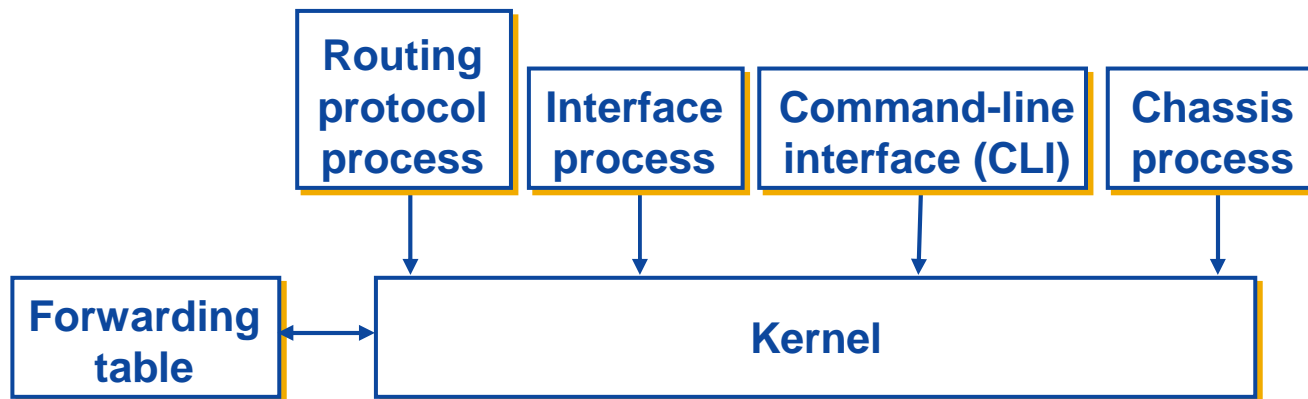
Software Processes



JUNOS Kernel

Provides the underlying infrastructure for all the JUNOS software processes

- Provides the link between the routing tables and the Routing Engine's forwarding table
- Responsible for all communication with the Packet Forwarding Engine, including keeping the Packet Forwarding Engine's copy of the forwarding table synchronized



Kernel Robustness

■ Fully Independent Software Processes

- Routing, Interface Control, Management, Chassis Management, SNMP, CLI, APS, VRRP
- Protected Memory Environment
 - Serious error in one module does not impact other modules or packet forwarding

JUNOS Advantages

- **One single binary for all platforms!**
- **No need to re-certify the software for edge and core**
- **One software to learn and manage.**
- **Individual restartable process.**

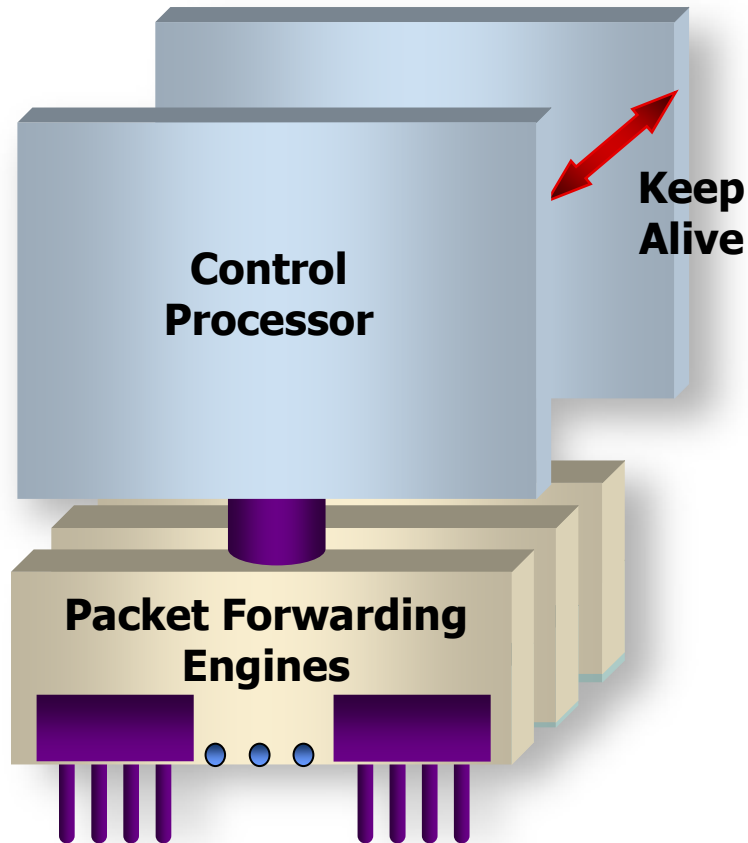
Agenda

- **Reliable software architecture**
- **Reliable hardware architecture**
- **JUNOS HA features**

Which platforms is fully redundant?

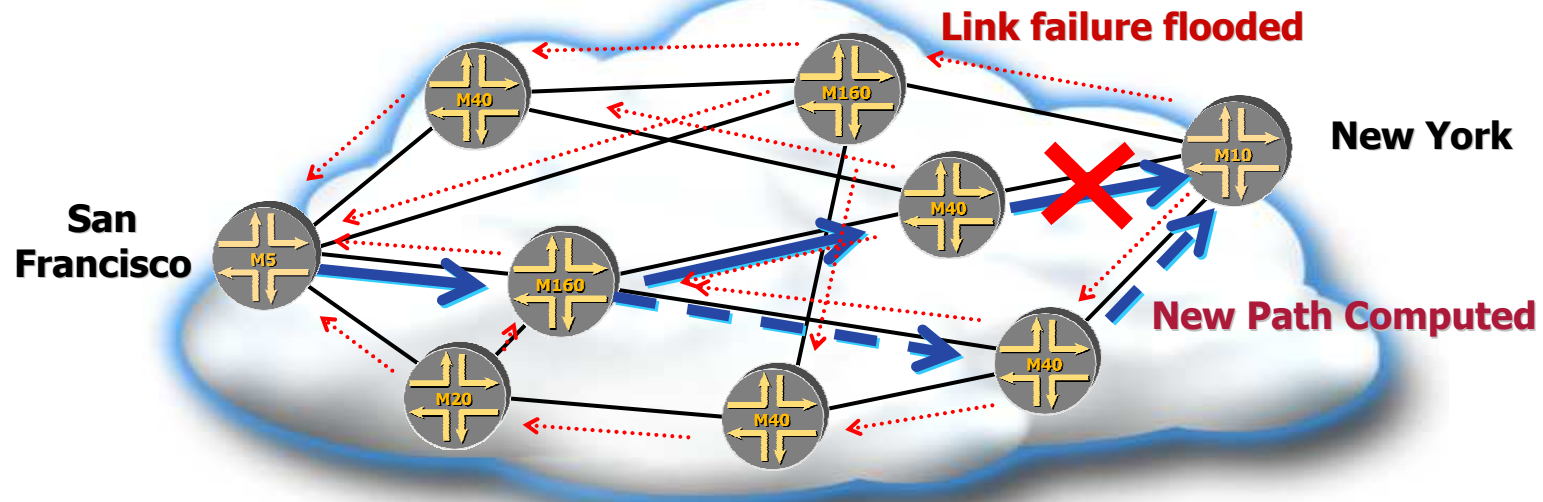
- **Including FANs, RE, Switch Fabric**
 - M20/M40e/M10i/M320/T320/T640/M120
- **Fully redundant \leftrightarrow N+1 backup.**
- **E.g. 4 SIB in M320 switch fabric. One down less 25% of throughput running at max. speed.**

Simple Processor Failover



- Protects against Single Node Hardware Failure
- Redundant control processors run keepalive process
- Automatic failover to secondary
- Configuration synchronized between processors
- Configurable timer
- Routing process restarts
- Requires PFE reset

IP Dynamic Routing



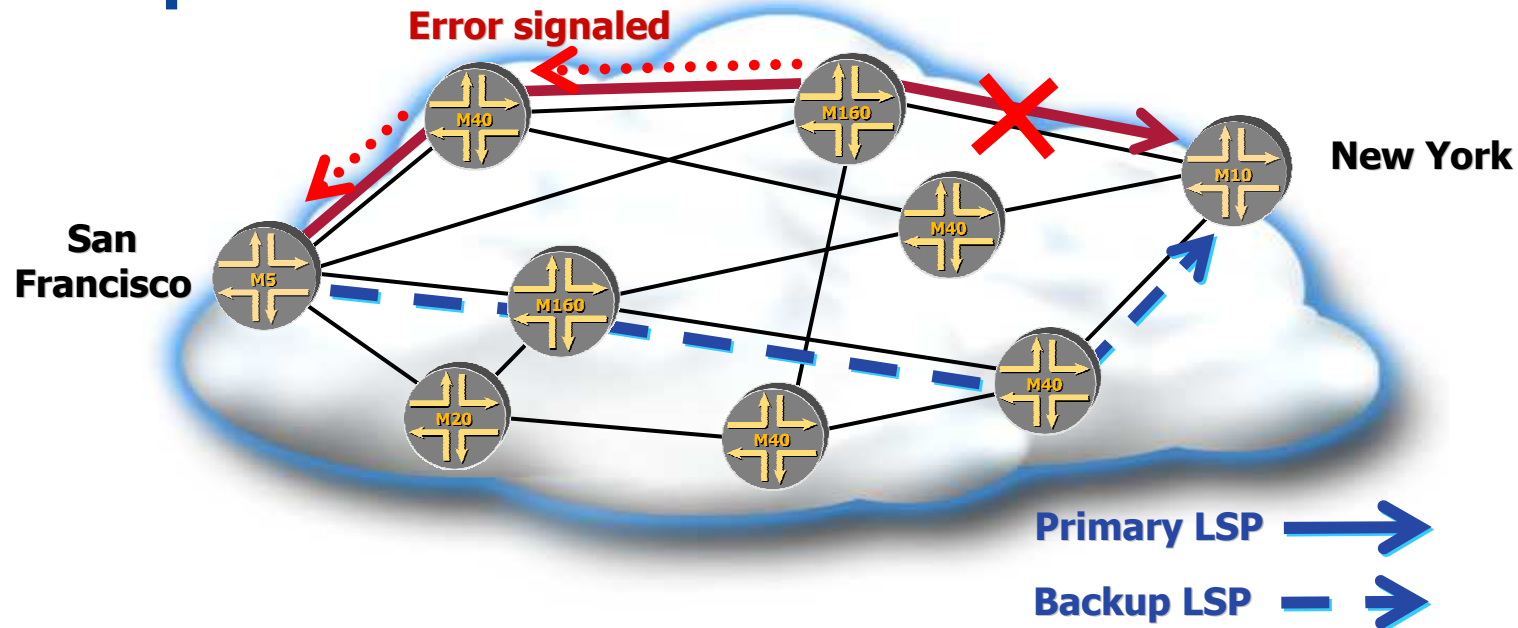
- OSPF or IS-IS computes path
- If link or node fails, New path is computed
- Response times: Typically a few seconds
 - Can be optimized to ~100 milliseconds

Routing Protocol Convergence

- Faster convergence improves reliability

Features	Benefits
High Priority Flooding for LSPs / LSAs	■ Faster propagation of major changes
Quick SPF Scheduling	■ Speeds calculation of optimum path
Sub-second Hellos	■ Faster Link Failure Detection
RIB and FIB Enhancements (BGP)	■ Indirect Next Hop implies faster convergence

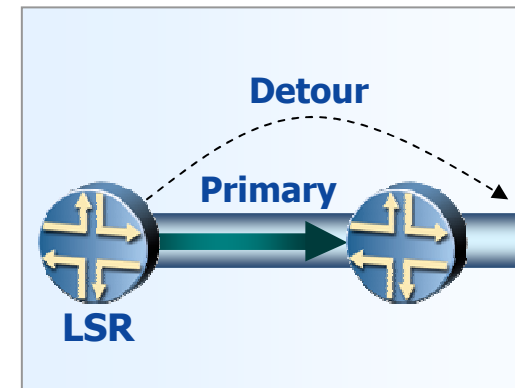
Backup Label Switched Paths



- Primary & backup LSPs established a priori
- If primary fails
 - Signal to head end, Use backup
- Faster response, requires wide area signaling

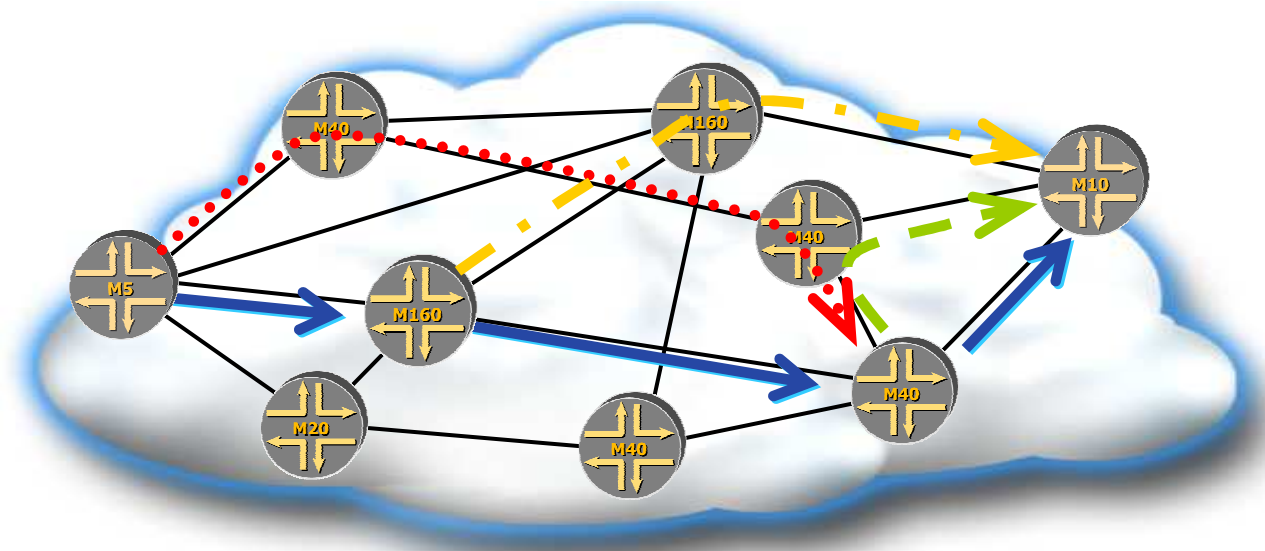
MPLS Fast Reroute

- Increasing demand for “APS-like” redundancy
 - MPLS resilience to link/node failures
 - Control-plane protection required
 - Avoid cost of SONET APS protection
- **Solution: MPLS Fast-reroute**
 - RSVP Extensions define Fast Reroute
 - LSPs can be set up, a priori, to backup:
 - One LSP across a link and optionally next node, or
 - All LSPs across a particular link



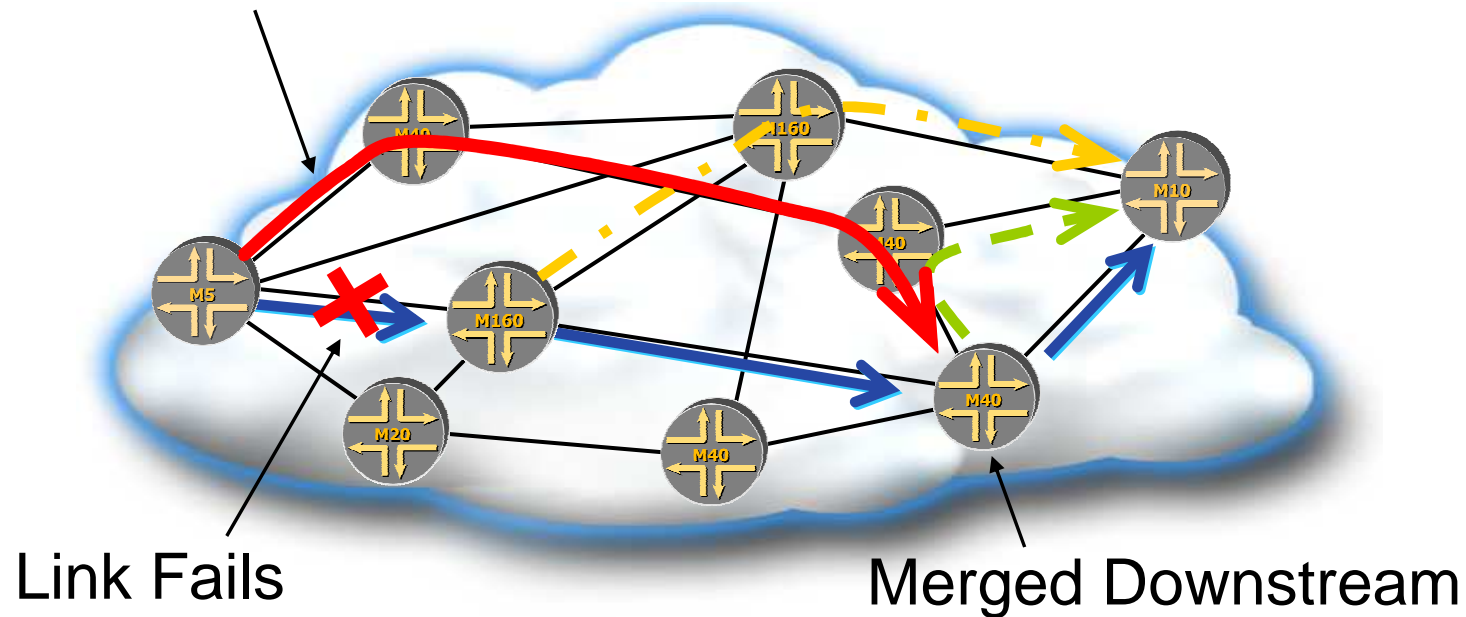
1:1 Protection

- For each LSP, for each node
 - Set up one LSP as backup
 - Merge into primary LSP further downstream
 - Backs up link and downstream node



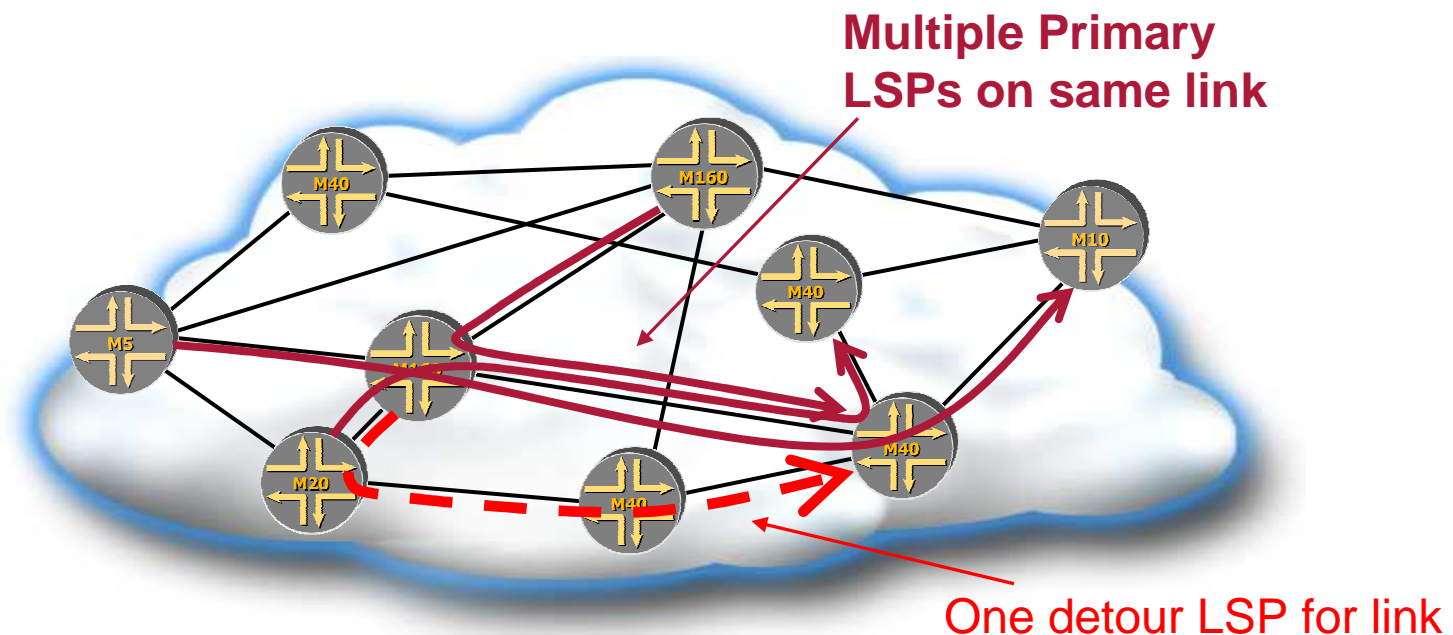
1:1 LSP Protection

Traffic uses detour LSP



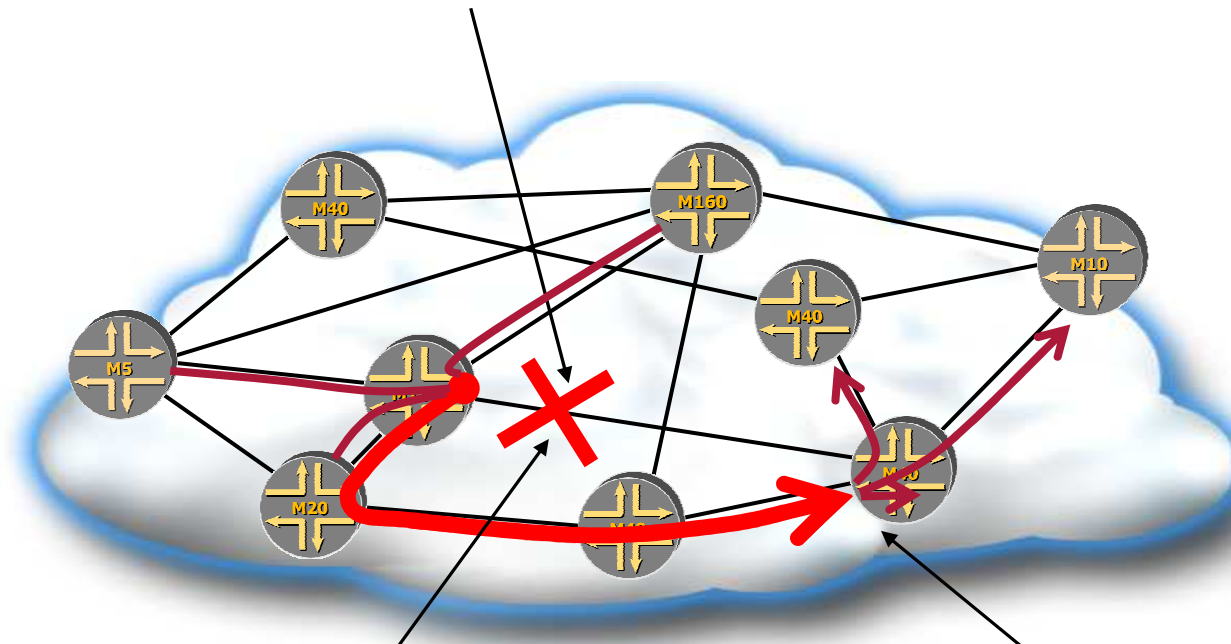
1:N Link Protection

- **For each link, for each neighbor**
 - Set up one detour LSP to backup the link
 - Uses LSP Hierarchy to backup all LSPs which were using failed link



1:N Link Protection

Link Fails



Primary LSPs multiplexed
over one detour LSP

LSPs demultiplexed
at next node

1:N Link and Node Protection

- For each link
 - For each node 2 hops away
 - Detour LSP backs up link & intermediate node
 - Uses LSP Hierarchy to backup all LSPs to that node
 - If there are two 2-hop paths to that node, setup two detour LSPs
 - For each node 1 hop away
 - Detour LSP backs up LSPs ending at that node

MPLS Fast Reroute

- **Provides fast recovery for LSP failure**
 - Based on a priori backup of detour LSPs
 - (eg, ~5 millisecond for tens of LSPs with 1:1)
- **There are significant tradeoffs between the approaches**
 - Number of LSPs required
 - Whether node failures are protected
 - Ability to reserve resources for backup LSPs
 - Optimality of routes

Summary of MPLS Methods

- **End-to-End backup LSPs**
- **MPLS Fast Re-Route**
 - 1:1 LSP protection
 - 1:N Link protection
 - 1:N Link plus node protection
- **All of these are interoperable based on IETF standards**

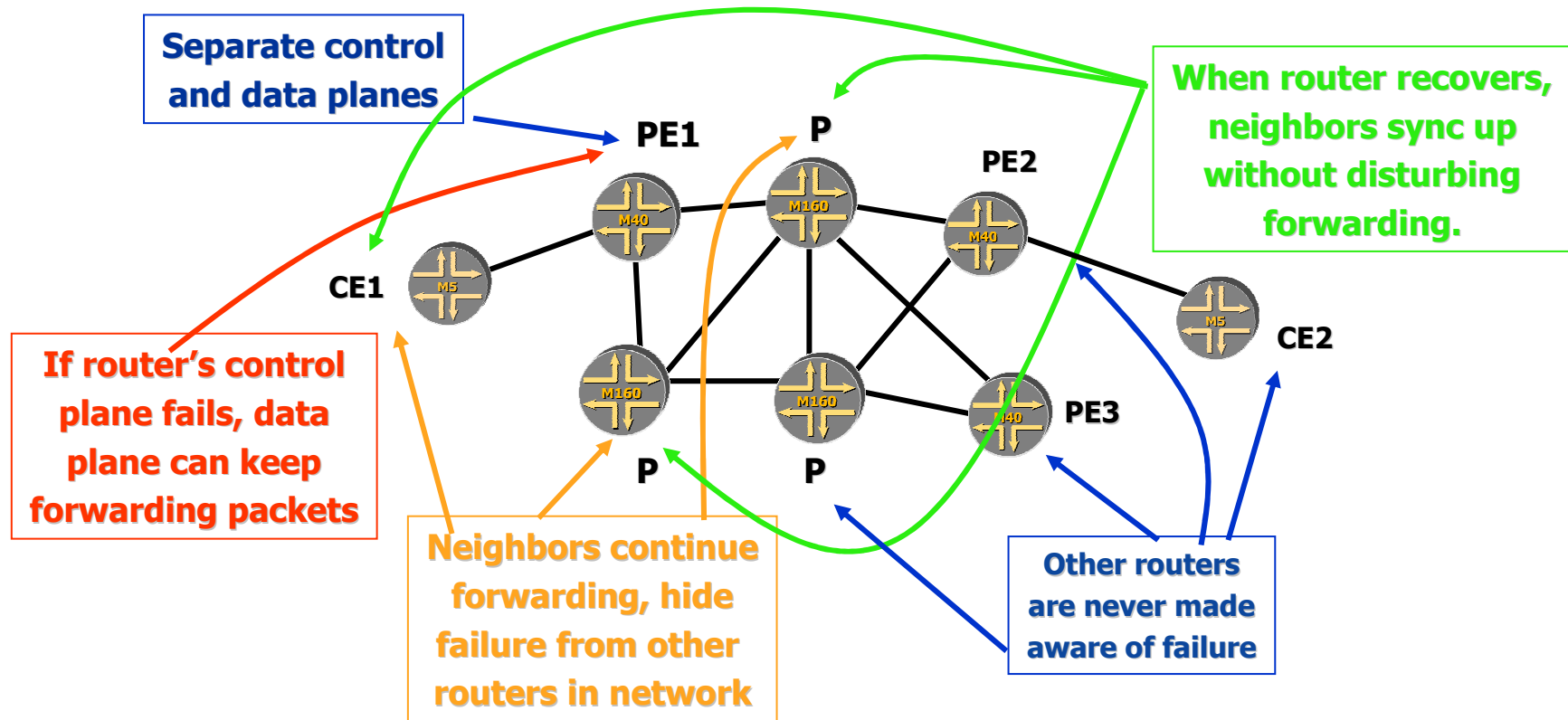
Control Plane Restart

- Control plane restarts due to planned upgrade, bugs, or hardware failure
- Hitless restart allows data plane to forward while control plane restarts
- But: If control plane fails
 - Peer routers will route around the router with the failed processor
 - Data plane won't have anything to forward
 - Routing will respond network-wide
- Restart on PE is especially disruptive

Graceful Restart

- Router agrees with its neighbors to forward data while control plane resets
- Restarting router re-establishes routing state with immediate neighbors
- Minimizes disruption
- Improves service availability
- Handles planned or unplanned outages

Graceful Restart - How ?



Graceful Restart - How ?

- **Graceful restart mechanisms are protocol specific:**
 - BGP – draft-ietf-idr-restart-05.txt
 - ISIS – draft-ietf-isis-restart-01.txt
 - OSPF – draft-ietf-ospf-hitless-restart-02.txt
 - LDP – draft-ietf-mpls-ldp-restart-06.txt
 - BGP/MPLS – draft-ietf-mpls-bgp-mpls-restart-02.txt
 - RSVP - draft-ietf-mpls-generalized-rsvp-te-09.txt
 - RIP – already build in !!!
 - PIM-SM in 6.4

What about routers that can't preserve forwarding state ?

- **Implementing a subset of graceful restart is very useful**
- **Enables such routers to take advantage of graceful restart on neighbors that preserve forwarding state**
- **CE routers can take advantage of graceful restart on service provider routers. Because traffic is sent from the edge to the restartable core routers.**

Graceful Restart at the Edge

- **Edge router typically**
 - Has one link each to many customers
 - Has a few uplinks to core routers
 - Routing is very simple, relatively static
 - Frequently no redundancy
- **Graceful restart in edge (CE and PE) routers is frequently desirable**

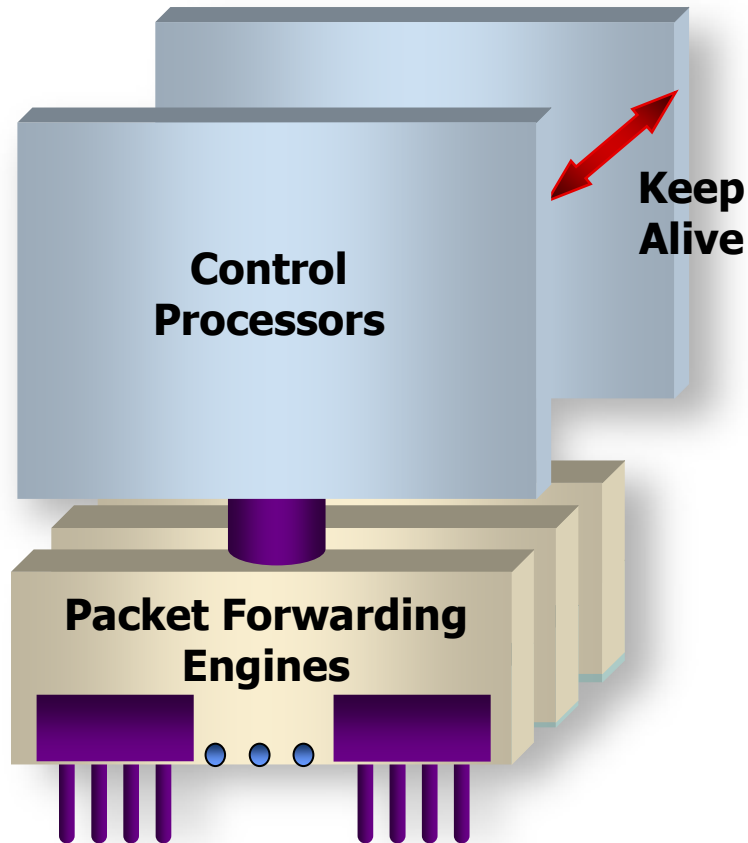
Graceful Restart in the Core

- **Core router typically**
 - Has multiple possible paths to any destination
 - MPLS Fast ReRoute is available
 - In large network, best path to some destinations may change rapidly
 - Forwarding while processor restarts may cause loops
- **For heavily meshed cores, Graceful restart in the core might not be recommended**
- **Details may vary**
 - EG: Forwarding on existing LSPs may be safe in core

Summary of Graceful Restart

- **Allows forwarding (data plane) to continue when control plane restarts**
- **Is useful for routers that can separate data plane and control plane operation**
 - Also useful in *neighbors* of such routers
- **Most useful where topology is limited (such as PE routers & their neighbors)**
- **Standards are being progressed for all appropriate IETF protocols**

Using Graceful Restart to allow Hitless RE Switchover



- **Protects against Single Node Hardware Failure**
- **Primary and Secondary CPUs utilize keepalive process**
 - Automatic failover
 - Synchronized Configuration
- **Processors share**
 - Forwarding info + PFE config
 - Interface state
- **Failure does not reset PFE**
 - No forwarding interruption
 - Routing and management are reset relatively rapidly
 - Alarms, SNMP traps on failover

Not the Same As Protocol Stateful Mirroring

Graceful Restart - Benefits

- **No disruption to routing and forwarding on:**
 - The restarting node, the peers or any other routers in the entire network
- **No change in the traffic patterns**
 - Preserves a steady state of traffic patterns
 - No impact on jitter, latency, packet ordering, route optimality
- **By limiting the scope of the nodes that are aware of a control plane restart to only the routing peers of the restarting node we improve control plane:**
 - Scalability across the network
 - Computation capacity and adaptability

■ Graceful Restart

- Extensions made to each protocol
- IETF standards-based solution
- Low risk and proven protocol-based solution

<u>Protocol</u>	<u>IETF Draft</u>
BGP	draft-ietf-idr-restart-05.txt
ISIS	draft-ietf-isis-restart-01.txt
OSPF	draft-ietf-ospf-hitless-restart-02.txt
LDP	draft-ietf-mpls-ldp-restart-06.txt
BGP/MPLS	draft-ietf-mpls-bgp-mpls-restart-02.txt
RSVP	draft-ietf-mpls-generalized-rsvp-te-09.txt

Graceful Restart - Limitations

- **Neighboring routers must understand GR procedures and messages**
 - Inhibits full GR implementation
 - Particularly significant on PEs
- **GR must stop if topology changes during grace period**
- **In some cases, cannot distinguish between link failure and control plane failure**
- **In some cases, routing reconvergence might exceed grace period**
 - For example, if there are hundreds of BGP peers
- **Protocol interdependencies can slow reconvergence beyond grace period**
 - For example, if BGP depends on LDP restart completion, which depends on RSVP restart completion
- **Industry acceptance of GR is not widespread**

A Better Solution is Required

- **Transparent to network peers (adjacencies and sessions should not drop).**
- **Convergence impact should be minimal**
 - Measured by the delay introduced to a “down” event that happens simultaneously or after the switchover.
- **Minimize performance impact of HA on the master RE.**
- **Switchover can happen at any point (no matter how much routing may be in flux).**

NSR Goals

- **Switchover from primary to secondary RE transparent to neighboring routers**
 - No interruption of protocol peering
 - Therefore 100% availability during switchover
- **Self-contained mechanism**
 - No required protocol extensions
 - No required cooperation from remote peers
- **Replacement for GR protocol extensions**
- **Key enabler for ISSU**

Stateful Replication

- **Warm standby**
 - For forwarding features
 - For control plane
- **Synchronization state is known before switchover**
- **No control-plane convergence impact**
- **Applicability:**
 - Resilience tool
 - Against hardware failures
 - Against software failures
 - Software upgrades

NSR Software Features

■ Supported

- Static Routes
- IGP (IS-IS, OSPF, RIP)
- BGP
- MPLS Signaling (LDP and RSVP)
- MPLS Applications (L3VPNs, VPLS, L2VPNs, L2 circuits)
- Multicast (IGMP, PIM, MSDP)

The Challenges of Software Upgrades

- **Network globalization means no off-peak traffic periods**
 - Maintenance windows become difficult or impossible to schedule
- **Five-nines reliability requirements mean ~5 minutes of downtime per year**
- **Therefore software upgrades must be performed *in-service***
 - No disruption to control plane
 - No disruption to forwarding plane
- **Traditional voice and ATM switches set the expectation for IP/MPLS switches**

ISSU Goals

- **Upgrade complete JUNOS package**
 - While maintaining sessions
 - L3 routing
 - L2 keepalives
 - Link management
 - With minimal disruption of packet forwarding
 - < 10 seconds of packet drops
- **Provide ISSU to existing hardware platforms**
- **“Hitless” is *not* a goal**
 - Goal is to make SW upgrade unnoticed by end users

Bidirectional Forwarding Detection

- **Q: Do we want every routing / signaling protocol to have sub-second Hellos?**
- **Q: How do we detect forwarding plane failures vs control plane failures?**
- **A: Bidirectional Forwarding Detection**
 - Simple rapid Hello protocol
 - 'Echo' allows test of forwarding plane
 - Does not replace OSPF/IS-IS/RSVP Hellos
 - See draft-katz-ward-bfd-01.txt

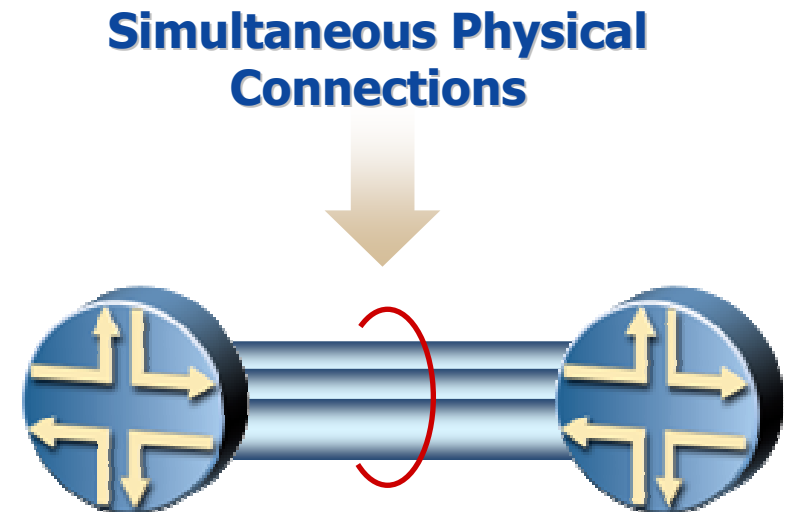
Link Redundancy

■ Link Bundling

- Link failure does not affect forwarding
- Load redistributed among other members

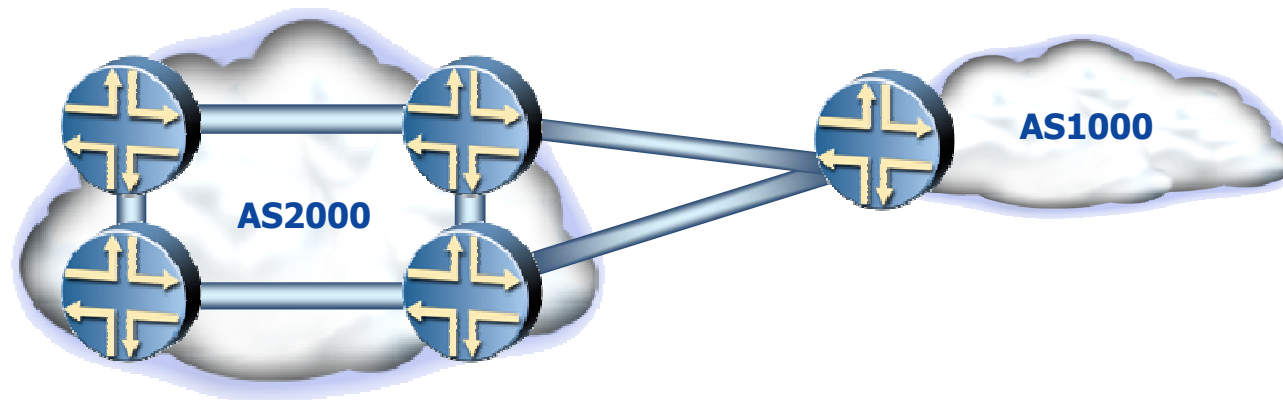
■ Parallel Link Technologies

- MLPPP – T1/E1 Link aggregation
- 802.3ad – Ethernet aggregation
- SONET/SDH aggregation
- Multi-Link Frame Relay



Resilient Edge Connectivity

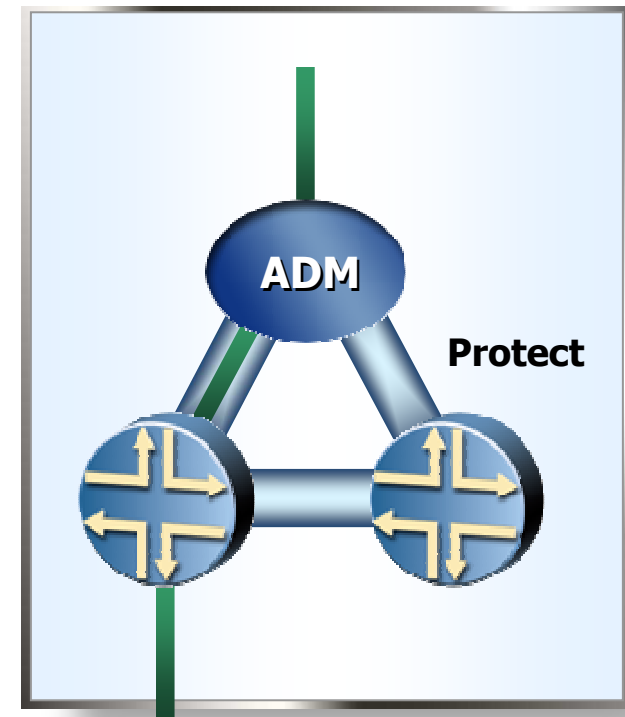
- **BGP Multi-Homing for resilient Internet and IP-VPN connectivity**
 - Stub network (static not BGP)
 - Multi-Homed Stub Network (Using BGP)
 - Multi-Homed Network (Same ISP)



Multi-Homed Stub Network (Using BGP)

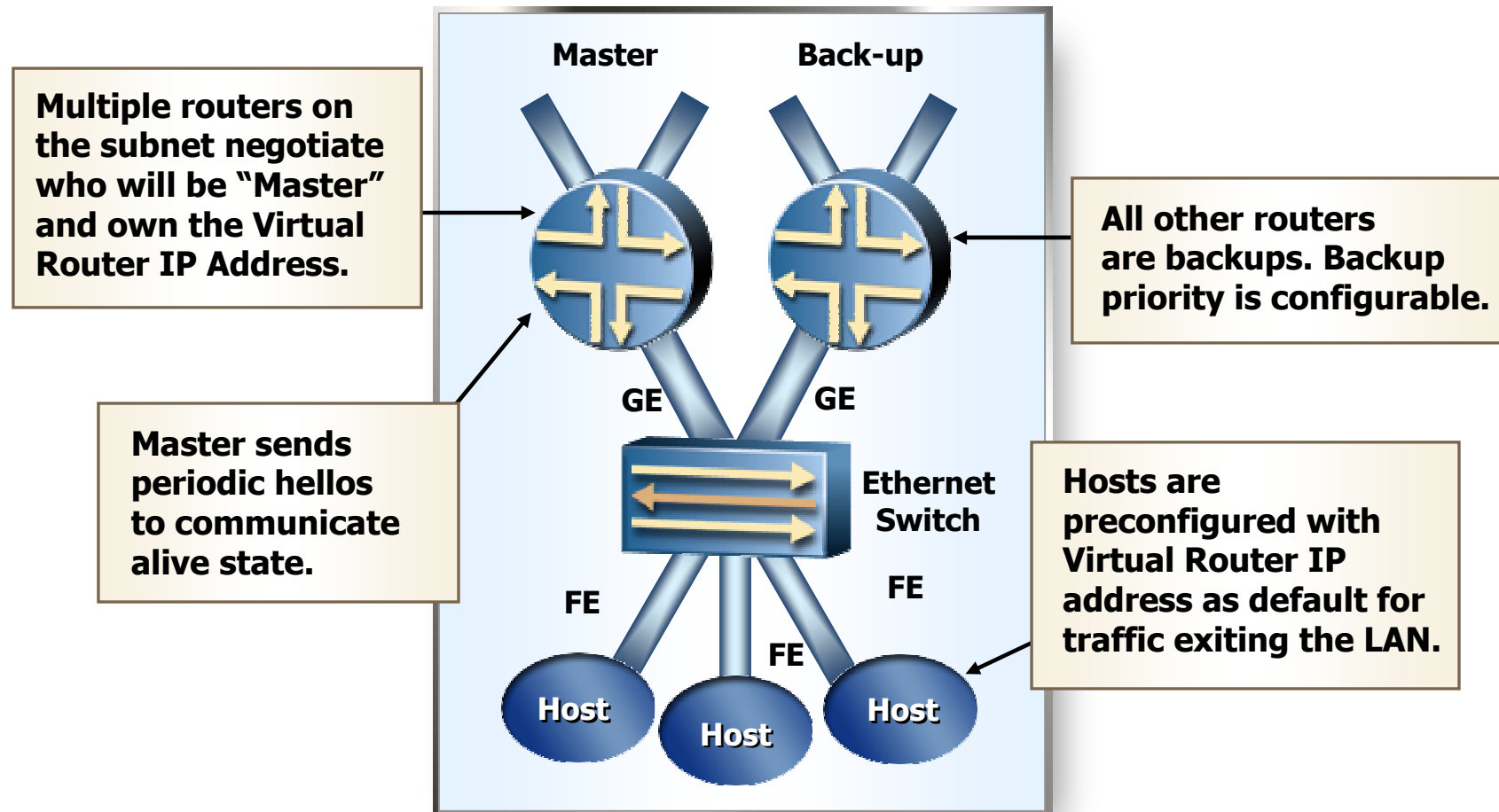
SONET/SDH Protection Switching

- **SONET APS & SDH MSP**
 - Redundant routers share uplink
- **Rapid circuit failure recovery**
 - Used on router-to-ADM links
 - 50 ms at physical layer
 - Faster than layer 3 routing protocol convergence
- **Interoperable with standard ADM**
- **Working & protect circuits**
 - May reside on different routers
 - May reside on same router



Virtual Router Redundancy Protocol

- Redundant default gateways–VRRP (RFC 2338)



Goal: Reliable Services

Reliable Services

