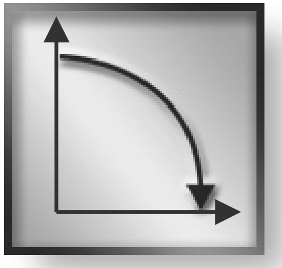
Delivering High Availability Routed Networks

Matt Kolon
matt@juniper.net
APRICOT 2005 - Kyoto

Effects of Network Outage

- **Immediate Impact**
  - Loss of Revenue
  - Repair Costs
  - SLA penalties
  - Dissatisfied customers
  - Project delays
  - Management distraction
- **Long Term Impact**
  - Damage to corporate brand
  - Customer churn, market share
  - Competition
  - Lawsuits
  - Lack of internal confidence

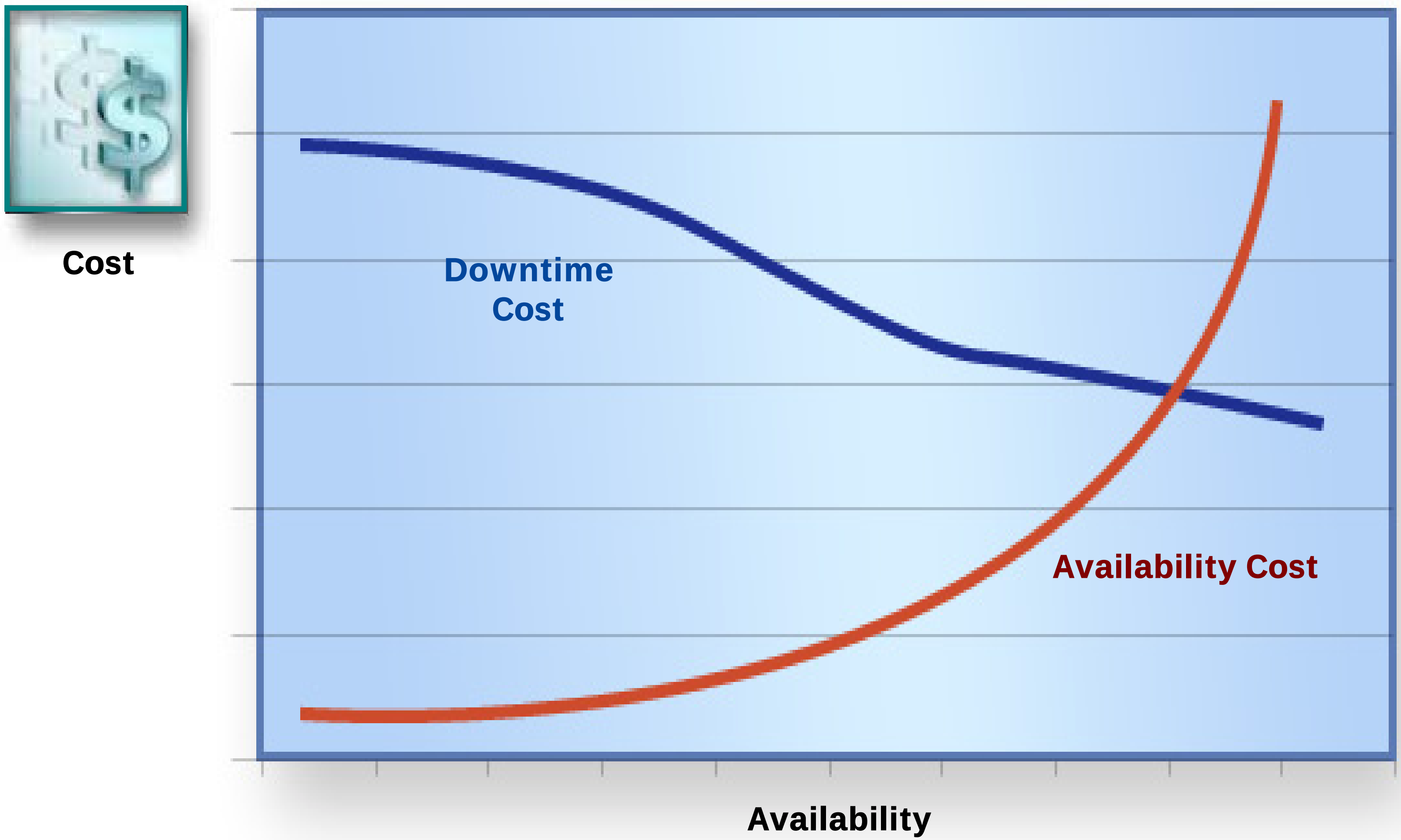
Financial



Market Share

Brand Damage

Business Case for High Availability

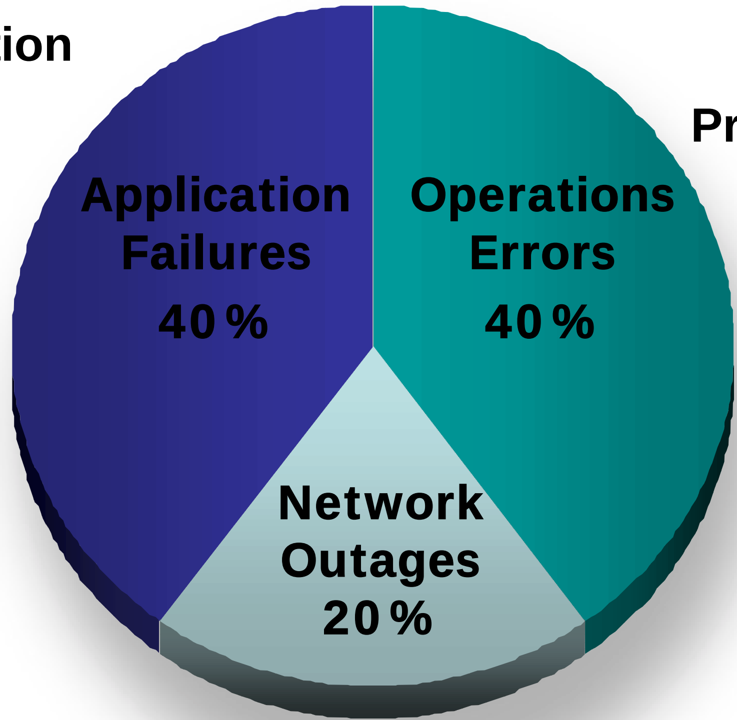


Threats to Dependability

Misconfiguration

DNS faults

DHCP faults



Process failures

Platform failures

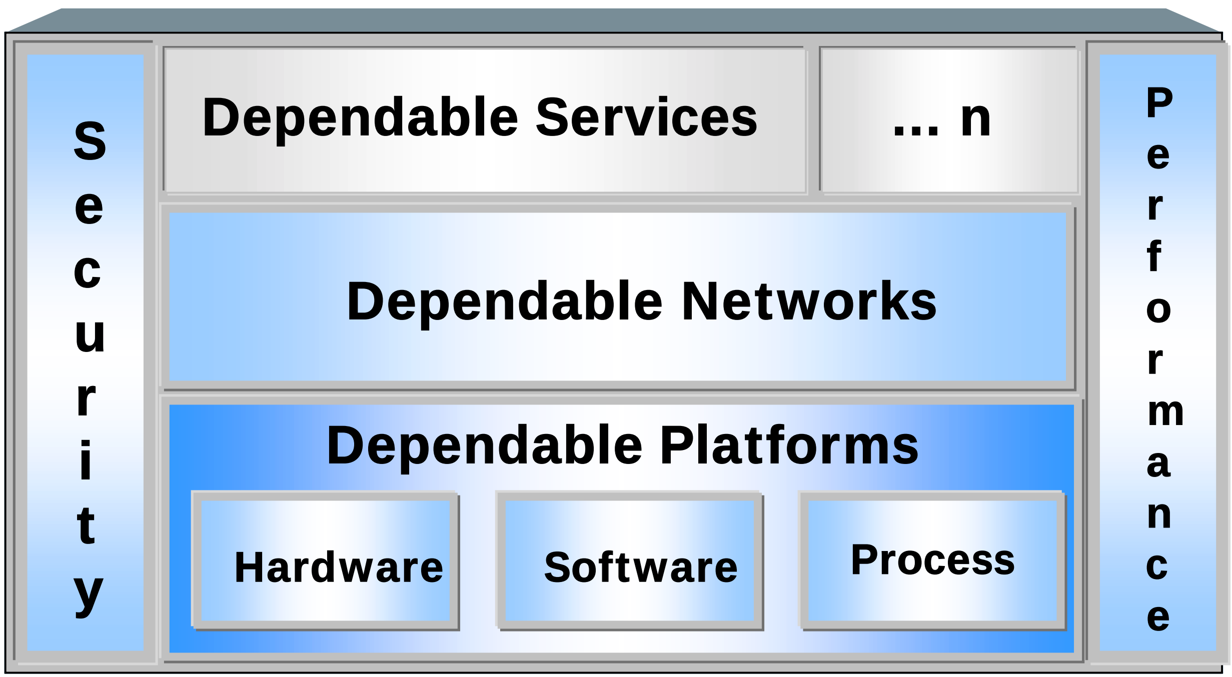
Hardware/software

Circuit failure

Source: Gartner Group, 1/01

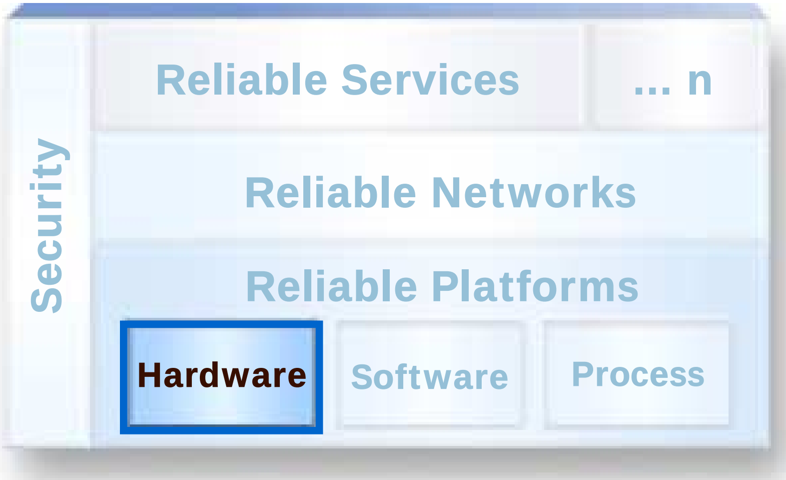
HA Solution Architecture

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



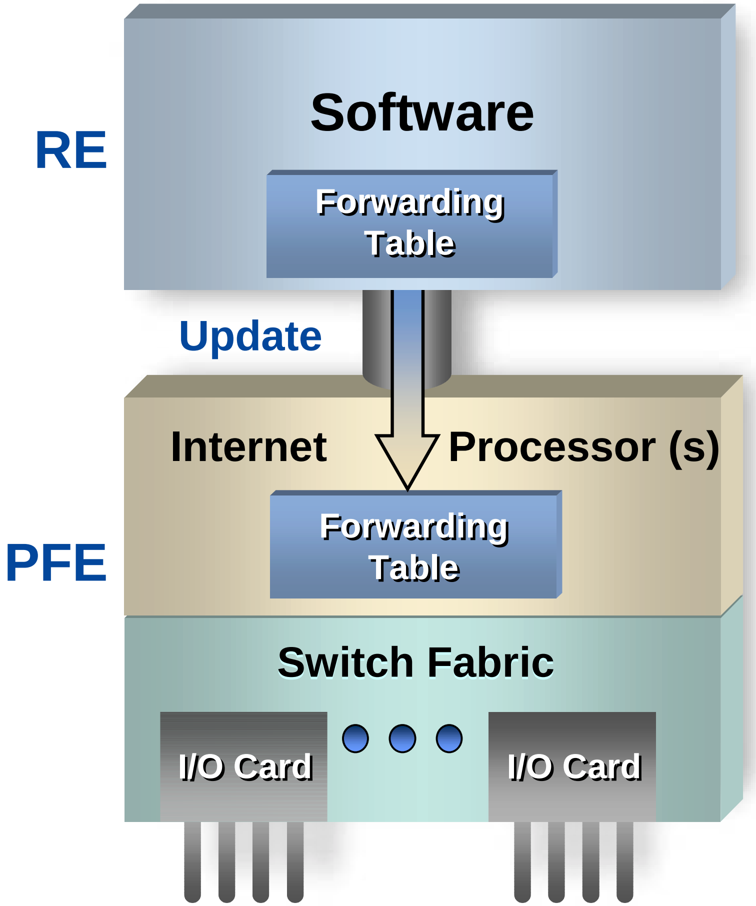
Reliable Hardware

- **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

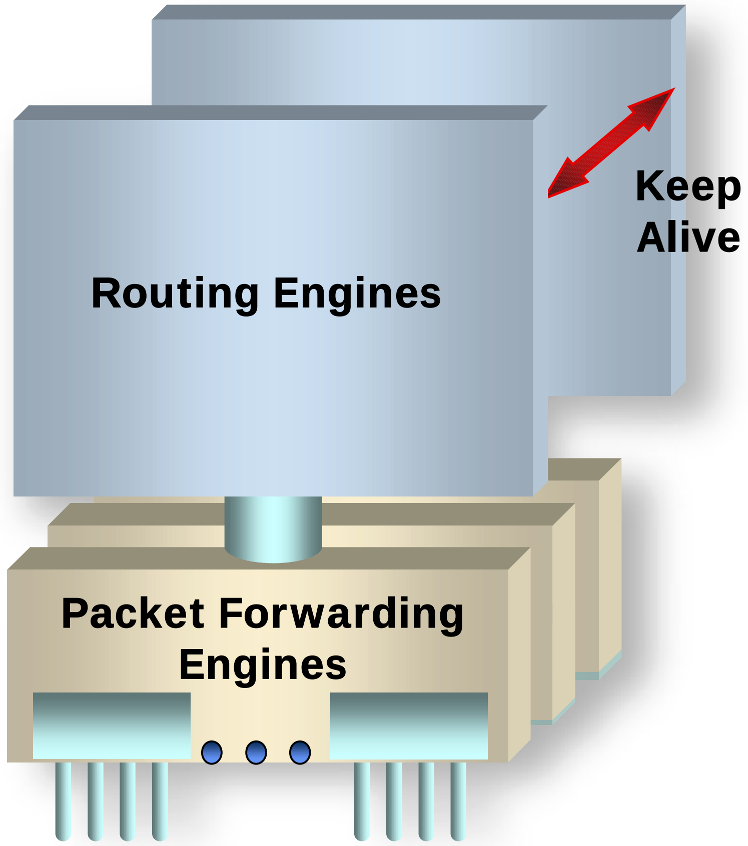


A Logical Platform View

- Hardware modularity is fundamental
- Clean separation of routing and packet forwarding functions
- Different vendors have different names, but for example:
  - **Routing Engine (RE)**
    - Routing protocol and management functions
  - **Packet Forwarding Engine (PFE)**
    - Packet forwarding and processing
- Multiples of each module allow redundancy and failover

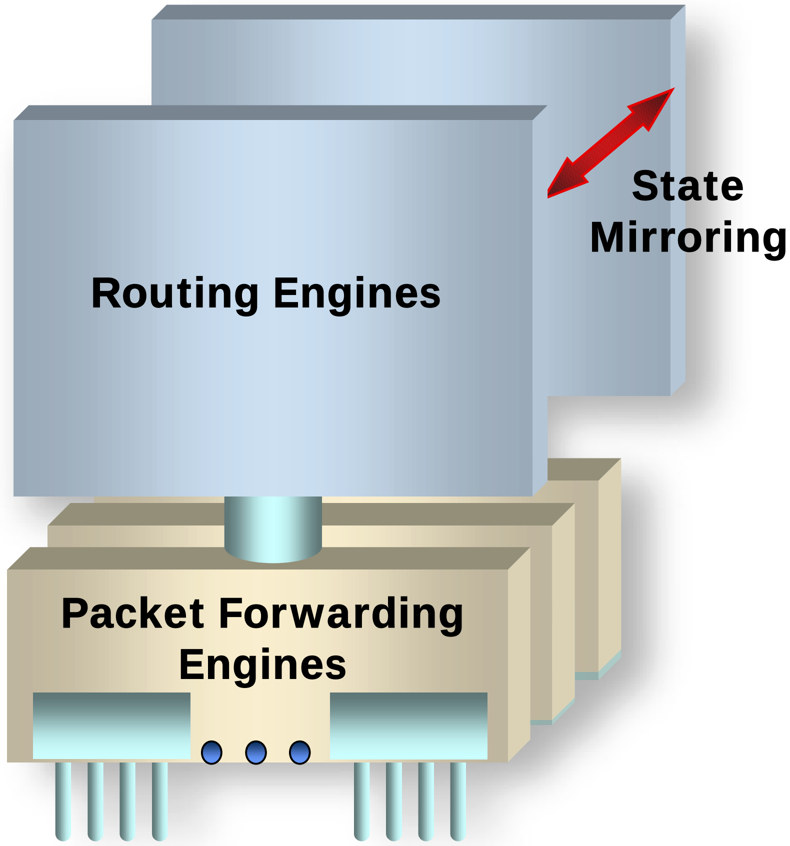


Simple RE Failover



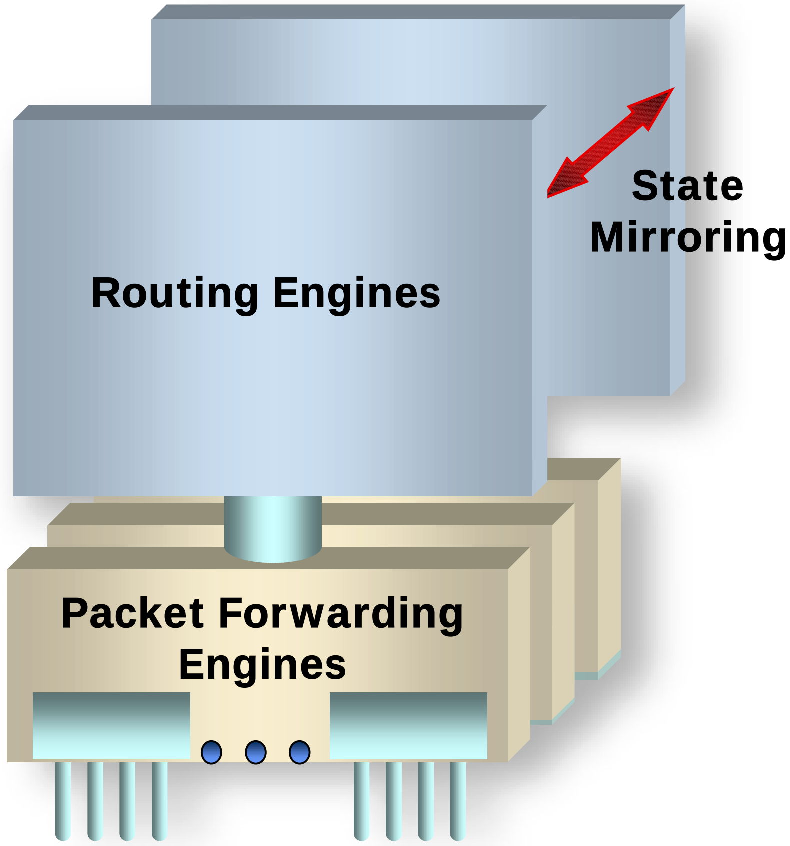
- Protects against Single Node Hardware Failure
- Redundant Routing Engines run keepalive process
- Automatic failover to secondary
- Configuration synchronized between RE's
- Configurable timer
- Routing Process restarts
- Requires PFE reset

Stateful Protocol Mirroring



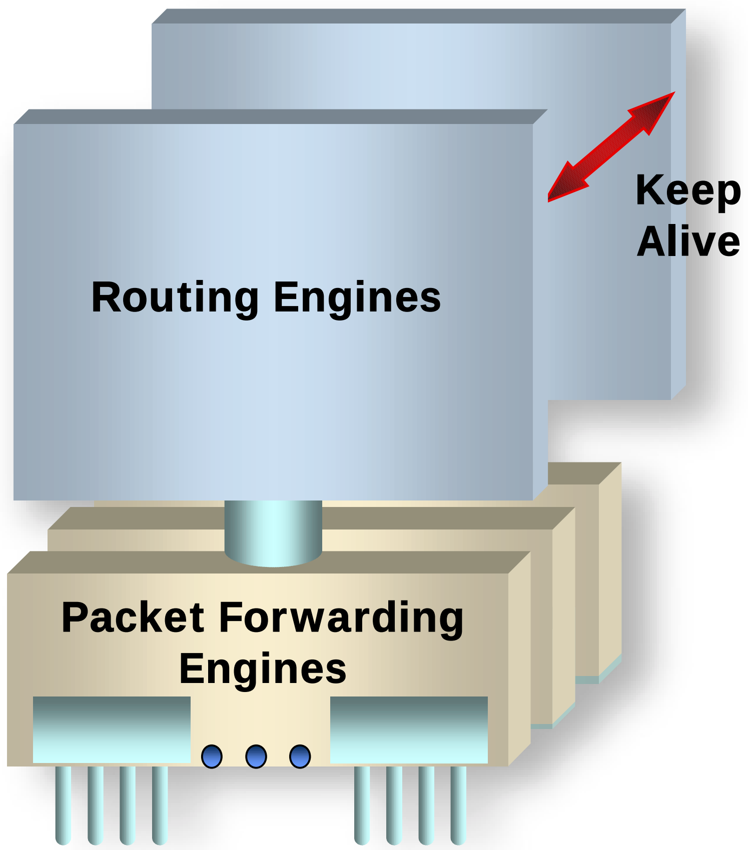
- Protects against Single Node Hardware Failure
- Redundant Routing Engines Mirror each others state
- BGP & TCP
- Theoretically ISIS & OSPF
- Automatic failover to secondary
- Advocated by some vendors, claiming Carrier-Class IP

Stateful Protocol Mirroring



- Great Idea!
- Difficult to do without replicating errors as well as "good" state
- Potential for "bug mirroring"
- Much more challenging in a rich service environment than an IP-only core

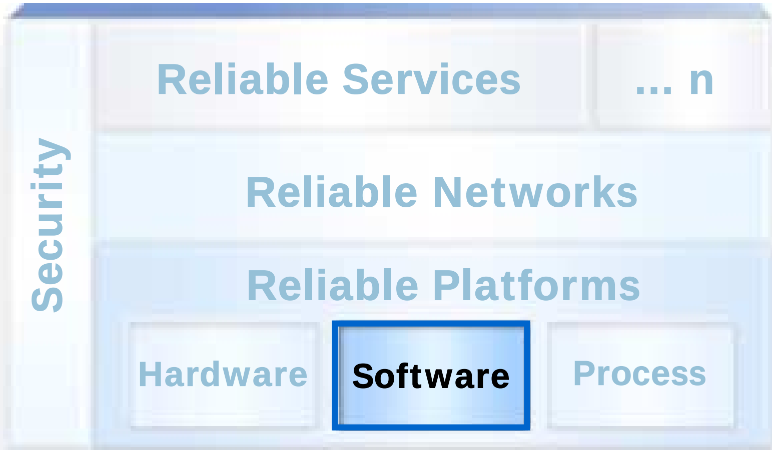
Graceful RE Switchover



- **Protects against Single Node Hardware Failure**
- **Primary (REP) and Secondary (RES) utilize keepalive process**
  - Automatic failover to RES
  - Synchronized Configuration
- **REP and RES share:**
  - Forwarding info + PFE config
- **REP failure does not reset PFE**
  - No forwarding interruption
  - Only Management sessions lost
  - Alarms, SNMP traps on failover

Reliable Software

- **Hard Fault Tolerance**
  - Redundant REs
  - Different software versions
- **Soft Fault Tolerance**
  - Separate control and forwarding
  - Modular processes can be restarted independently
  - Processes protected in own memory space
  - Individual process watchdogs
- **MTTR Reduction**
  - Incremental software upgrades
  - Modularity to speed up testing



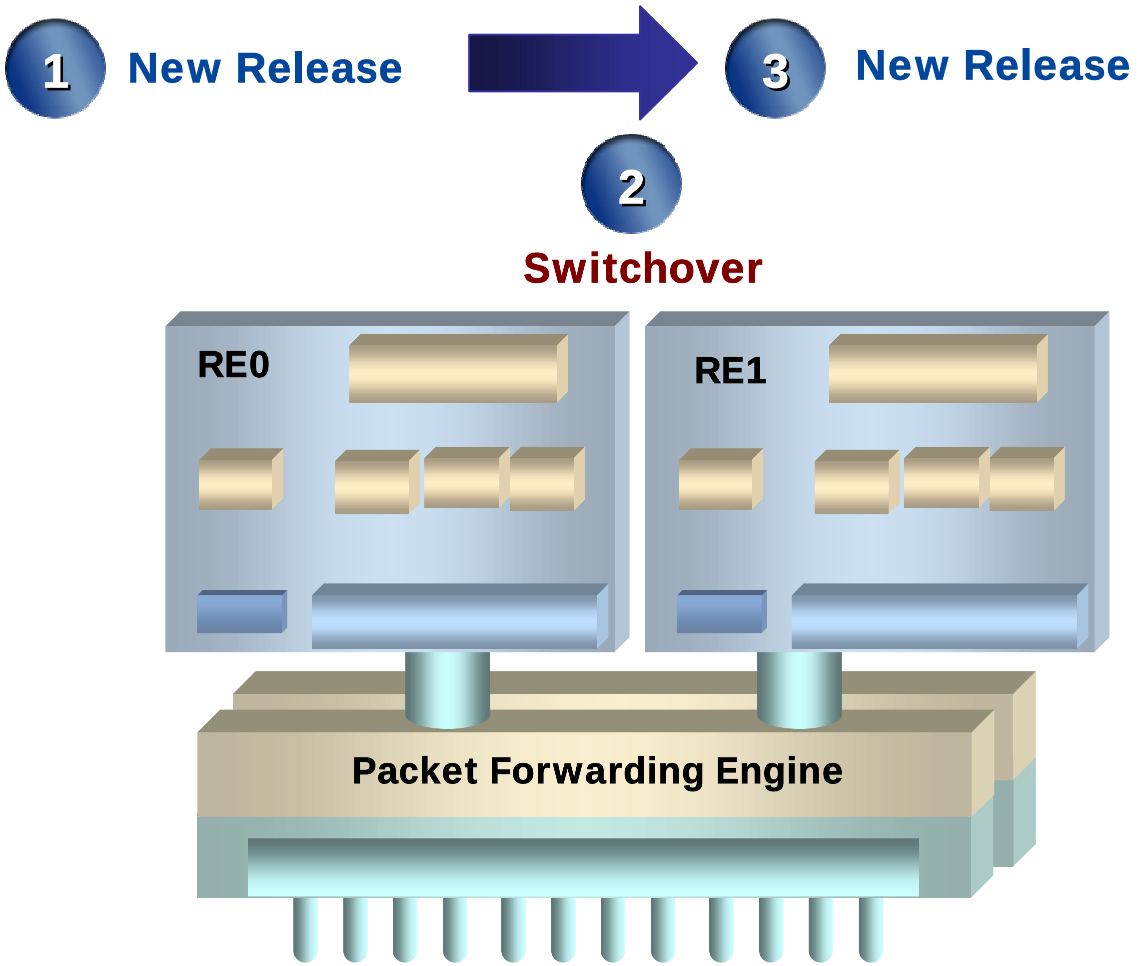
Software Reliability Principles

- **Loose coupling of modular components**
  - A single failing component will not crash the box
  - Localizes complexity
  - Creates conceptual boundaries to contain problems
  - Clean interfaces between system components (well-defined, efficient APIs)
- **Memory protection**
  - Processes cannot scribble on each others' memory
- **Adding complexity will not improve reliability**
  - If base software is not expandable, maintainable, reliable, then adding additional layers won't help
  - "Make it as simple as possible, but no simpler."

    --Albert Einstein

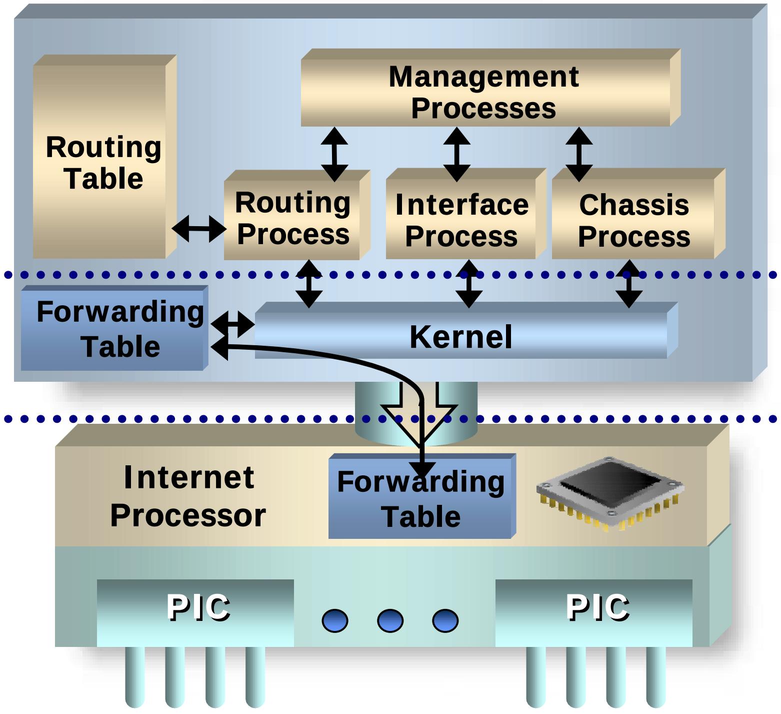
In-Service Software Upgrades

- **Leverages**
  - Graceful RE Switchover
  - Graceful Restart Protocol Extensions
- **Preserves forwarding**
  - In any RE failure
- **Delivers**
  - In-service software upgrades
- **Might also be enabled by stateful mirroring**

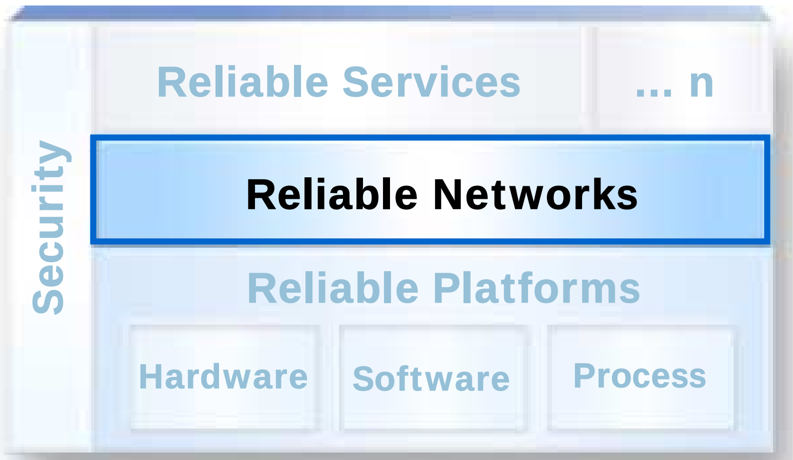


In-Service Software Upgrades

- **When Software is modular:**
- **(JUNOS, for example)**
  - "jinstall" is a complete software distribution
  - "jroute"
    - Routing protocols
  - "jkernel"
    - Operating system
  - "jpfe"
    - PFE software



Reliable Networks



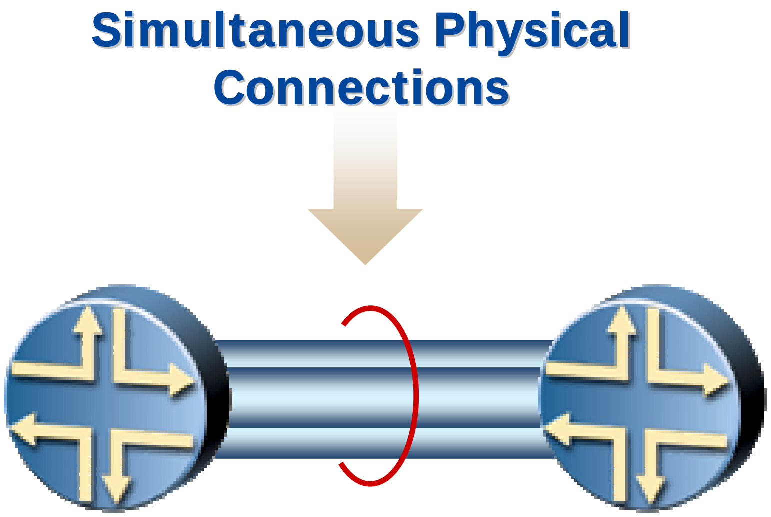
Protection and Recovery from failures

- MPLS
  - Fast reroute
  - Secondary LSPs
- VRRP

- Convergence improvements
- Graceful Restart
- Link Redundancy
- Multi-Homing
- SONET APS / SDH MSP

Link Redundancy

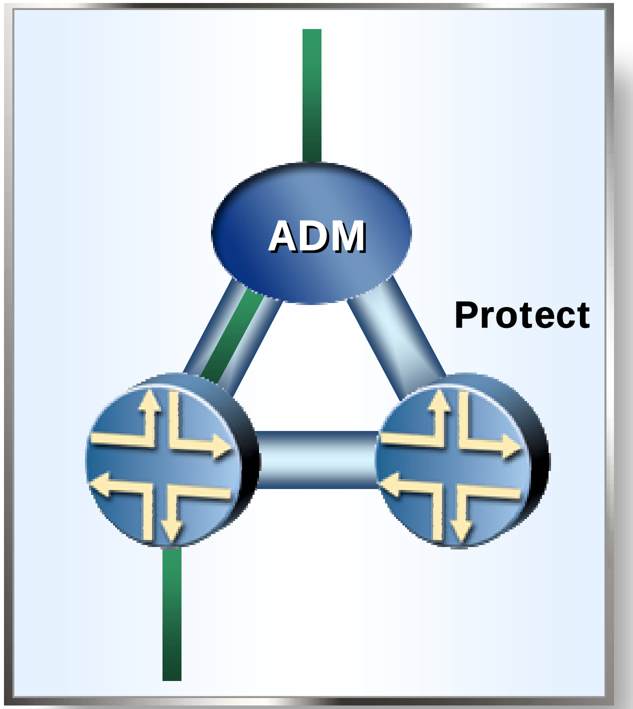
- **Reliable Links**
  - Link failure does not affect forwarding
  - Load redistributed among other members
- **Parallel Link Technologies**
  - MLPPP – T1/E1 Link aggregation
  - Multi-Link Frame Relay
  - 802.3ad – Ethernet aggregation
  - SONET/SDH aggregation



SONET/SDH Protection Switching

- **SONET APS & SDH MSP**
  - Redundant routers share uplink
- **Rapid circuit failure recovery**
  - Used on router-to-ADM links

  Layer 3 protocol convergence longer
- **Interoperable with standard ADM**
- **Working & protect circuits**
  - May reside on different routers
  - May reside on same router

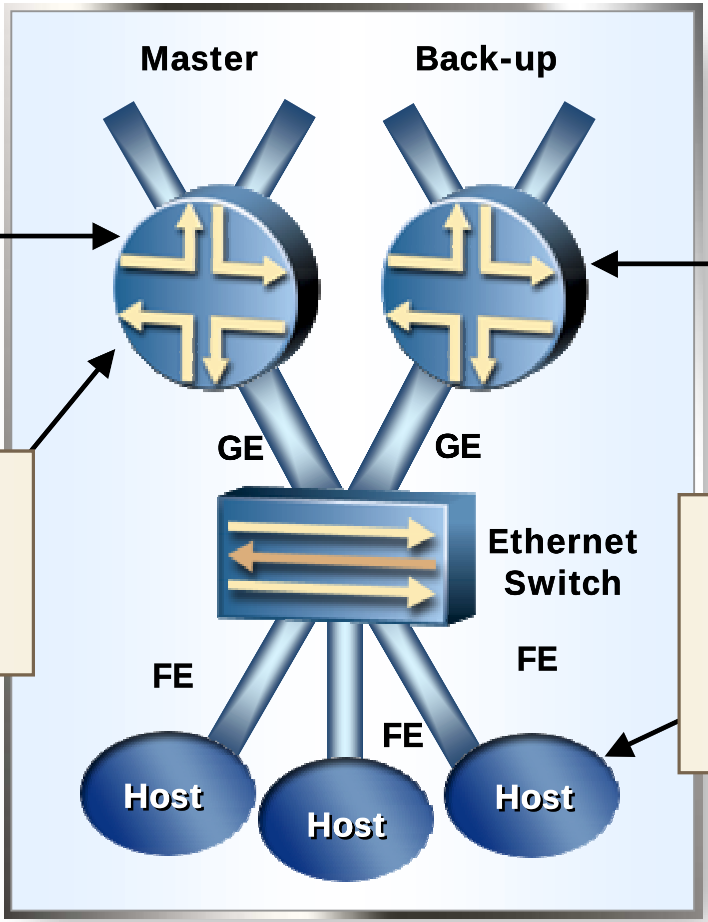


Virtual Router Redundancy Protocol

- Redundant default gateways–VRRP (RFC 2338)

Multiple routers on the subnet negotiate who will be "Master" and own the Virtual Router IP Address.

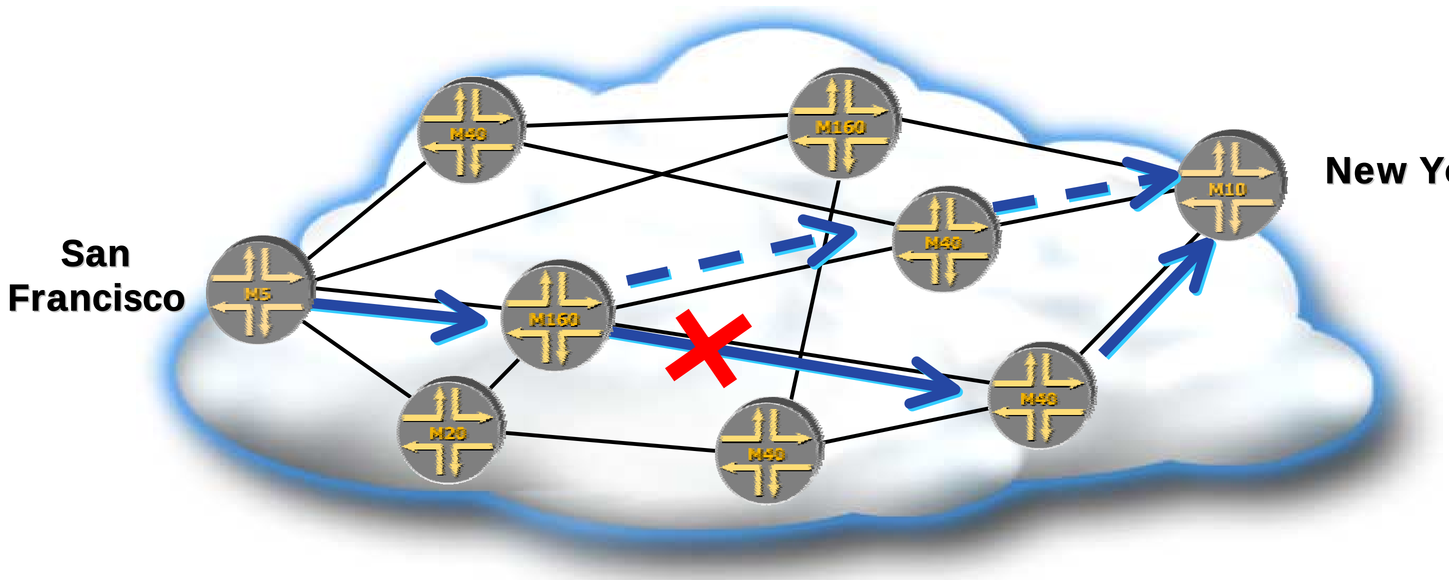
Master sends periodic hellos to communicate alive state.



All other routers are backups. Backup priority is configurable.

Hosts are preconfigured with Virtual Router IP address as default for traffic exiting the LAN.

IP Dynamic Routing



- OSPF or IS-IS computes path
- If link or node fails, New path is computed
- Response times: Typically a few seconds
- Completion time: Typically a few minutes, but very dependant on topology

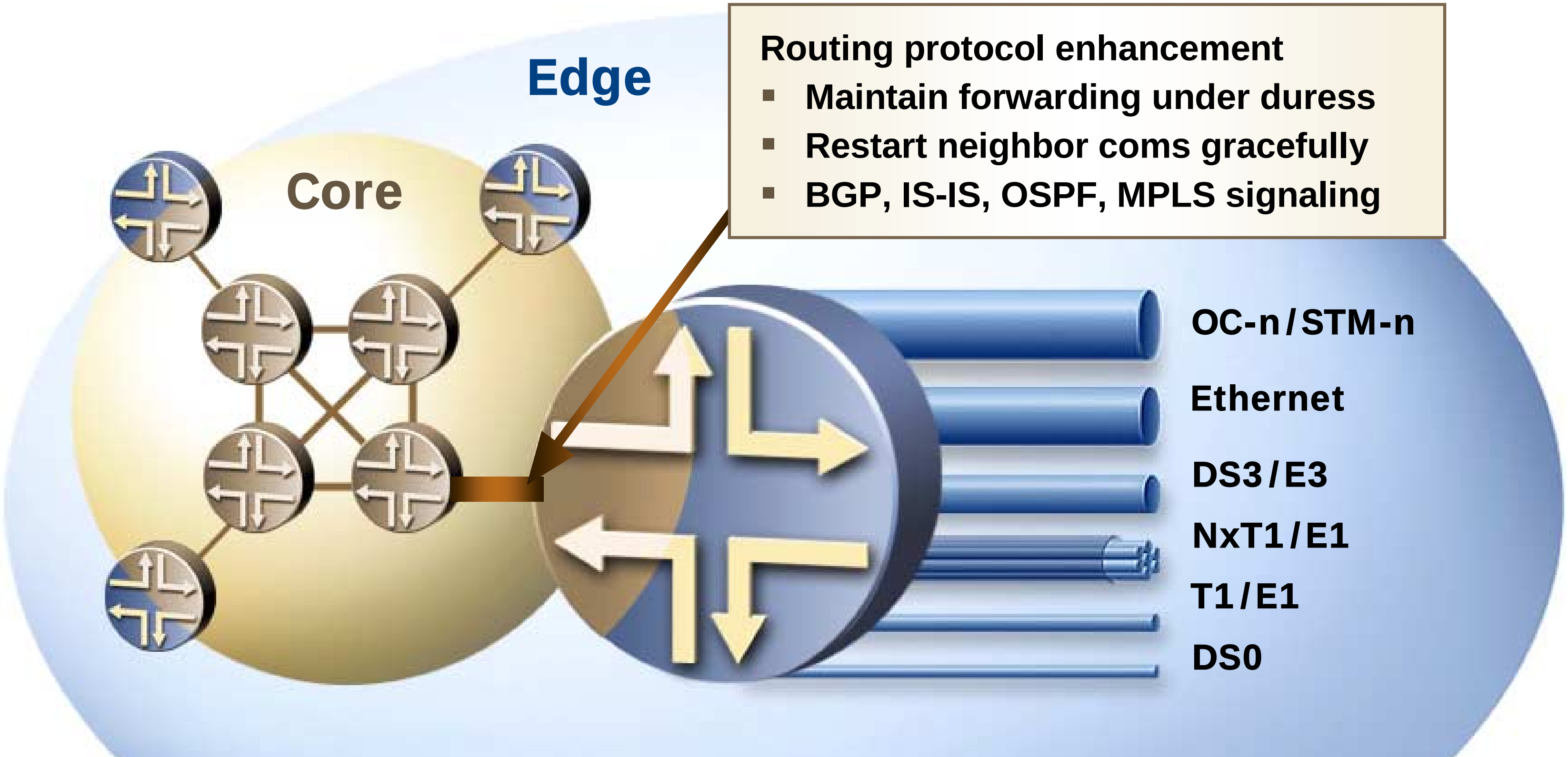
Faster Router Convergence

- Faster convergence improves Network Reliability

Features	Benefits
High Priority Flooding for Interested LSPs (ISIS / OSPF)	▪ Timer reduced from 100 to 20msec ▪ Faster propagation of major changes
Quick SPF Scheduling (ISIS / OSPF)	▪ Reduces time from 7 sec to 50 msec ▪ Speeds calculation of optimum path
Sub-second Hellos (ISIS)	▪ Lowest Hello Time possible for IS-IS, 333msec ▪ Faster Link Failure Detection
RIB and FIB Enhancements (BGP)	▪ Indirect Next Hop implies faster convergence

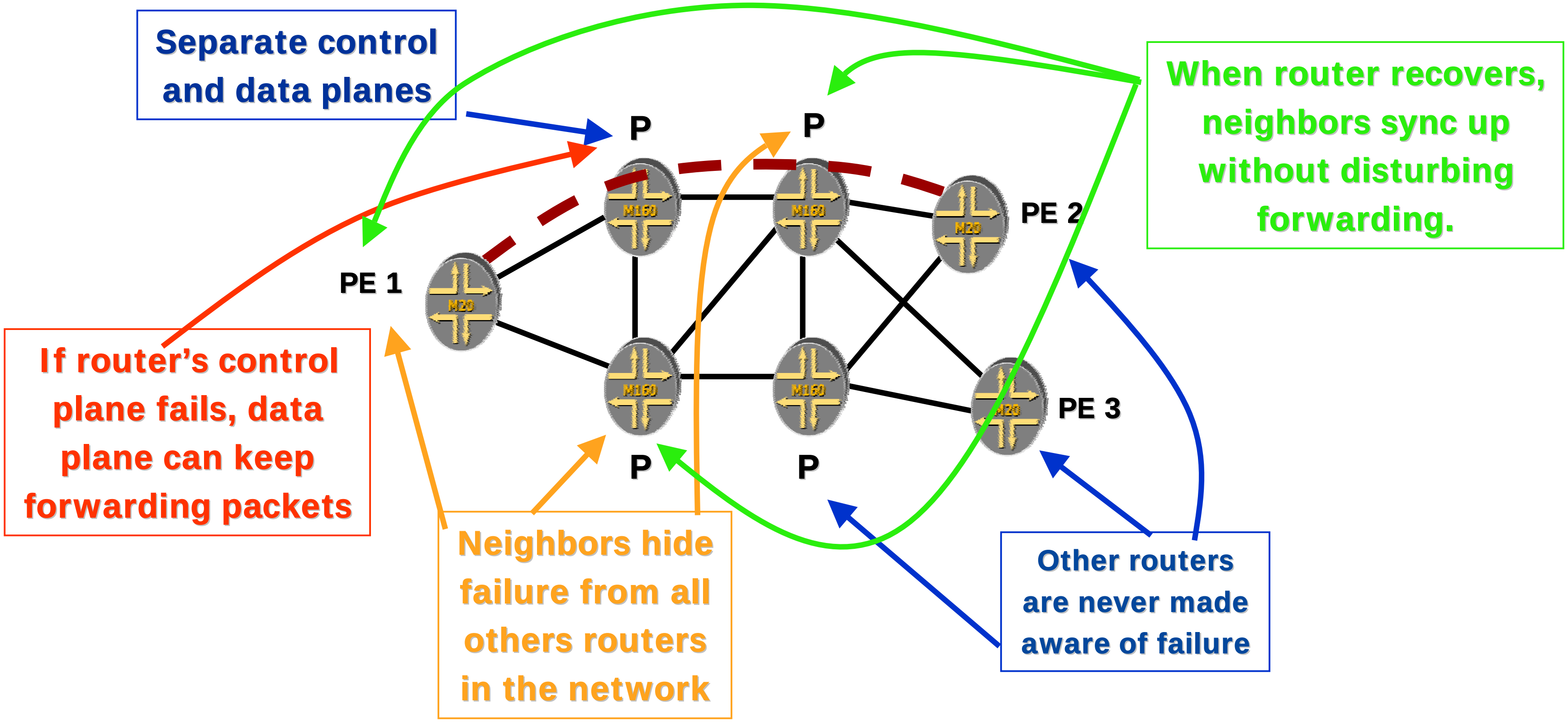
Routing Protocol Graceful Restart

Routing protocol enhancement
- Maintain forwarding under duress
- Restart neighbor coms gracefully
- BGP, IS-IS, OSPF, MPLS signaling



- **Protocol extensions distinguish between control, data plane failure**
- **Protects against Routing Protocol Module failure**
- **Failure invisible to everyone but peers**
- **Stepping stone for non-stop forwarding**

Graceful Restart - How ?



Graceful Restart Protocol Details

Purpose - Continue forwarding (PFE) during a restart of routing (RE)

	Changes	IETF
BGP	Protocol extensions Per-peer configuration Various timers with configurable defaults	*Graceful Restart Mechanism for BGP* draft-ietf-idr-restart-08.txt
OSPF	Protocol extensions New opaque-LSA type 9, "Grace-LSA"	*Hitless OSPF Restart* rfc3623
IS-IS	Protocol extensions 3 new timers New "re-start" option (TLV) in IIH PDU	*Restart Signaling for ISIS* draft-shand-isis-restart-04.txt
MPLS	Protocol Extensions Uses signaling as described in "Graceful Restart Mechanism for BGP	Graceful Restart Mechanism for BGP with MPLS draft-ietf-mpls-bgp-mpls-restart-03.txt
RSVP	Protocol Extensions Extend rfc 3473 Recovery ERO	Graceful Restart Extensions draft-rahman-rsvp-restart-extensions-00.txt

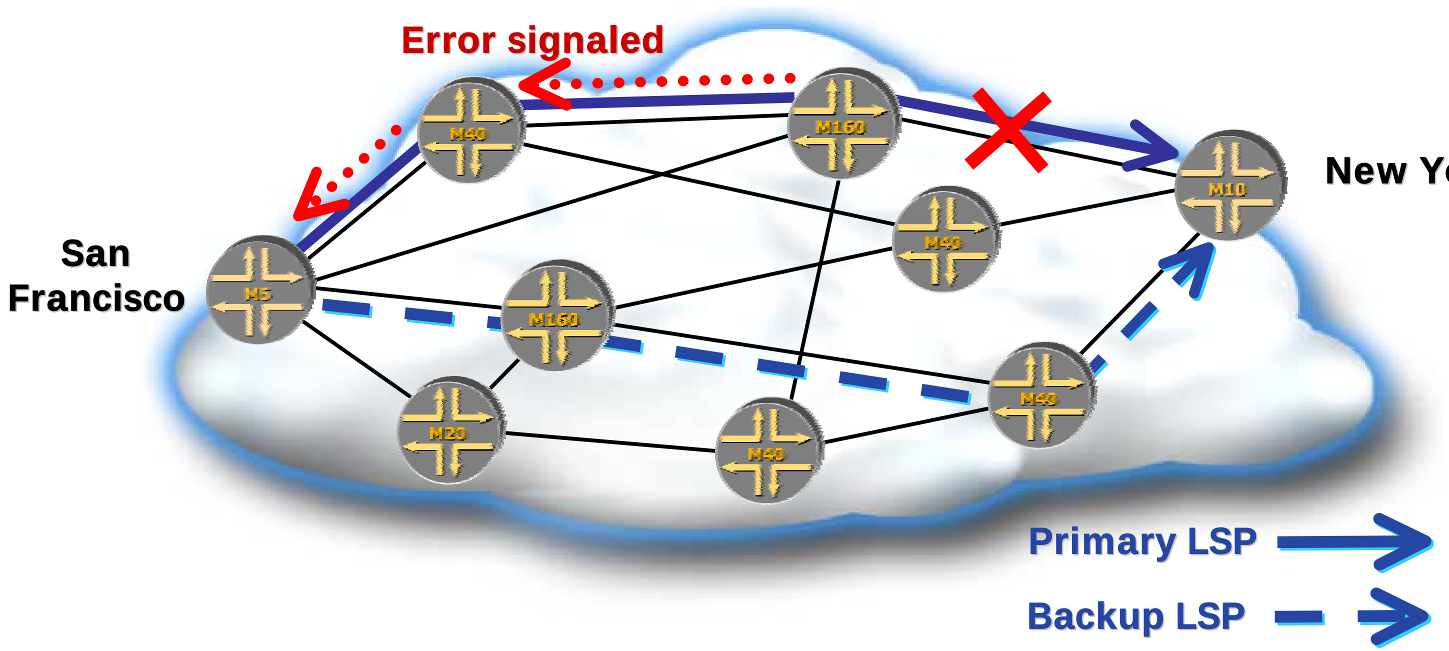
MPLS-based mechanisms

- Path protection (aka Secondary LSP)
- Local 1:1 (aka LSP/Detour Protection Fast Reroute)
  - Protects against both link failure and node failures
- Local 1:N (aka Facility-based Fast Reroute)
  - Link Protection Fast Reroute (Protects only against link failure)
  - Node Protection (Protects against both link failure and node forwarding plane failure)

Secondary LSPs

- An LSP may have multiple paths
- Primary path is the preferred path to set up and use
- Secondary paths are alternatives, to be used when the primary fails
  - Usually node/link disjoined from primary
    - The level of overlap between the primary and the secondary could be controlled
- Secondary path may result in wasting resources
  - Resources reserved for secondary are reserved all the time, yet used only when the primary fails

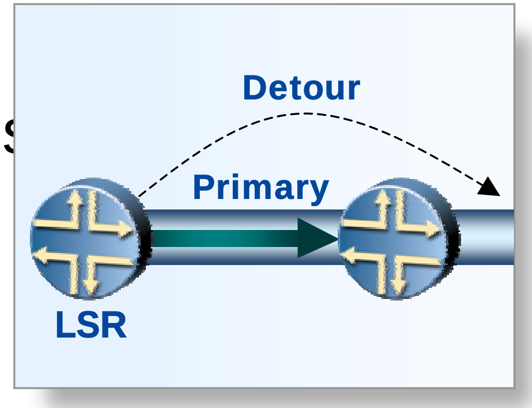
Secondary LSPs



- **Primary & secondary LSPs established a priori**
- **If primary fails**
  - Signal to ingress router to use secondary LSP
- **Faster response than routing protocol, requires wide area signaling**

MPLS Fast Reroute

- **Increasing demand for “APS/MSP-like” redundancy**
  - MPLS resilience to link/node failures
  - Control-plane protection required
    - Frequent code upgrades = instability
  - Cost of APS/MSP protection



- **Solution: MPLS Fast-reroute**
  - RSVP Extensions define Fast Reroute

Fast Reroute

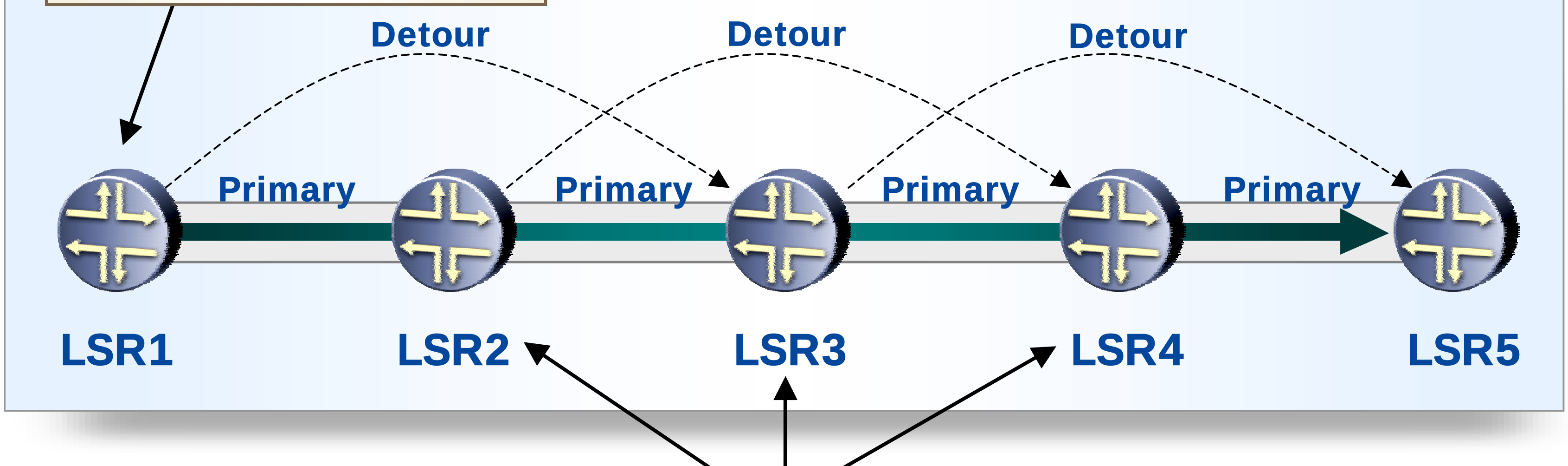
- Head-end of LSP enables fast reroute
- When signaled, each intermediate node calculates its own path to the tail-end
  - Uses CSPF and reservation
  - Doesn't duplicate reservations on a single link (but does duplicate on the network as whole)
- If any node sees the interface over which the primary LSP is routed go down, that node can instantly switch to backup
- Head-end discovers later and can reroute in a way that is more globally optimal

Complexity Comparison

- **Secondary LSPs**
  - Signaled by ingress LSR only, protects path
  - + additional constraints can be applied
  - + tries to stay away from primary path nodes and links
  - - additional management and planning
  - - switch is done at the ingress router only
  - + more scalable
- **FRR**
  - Each LSR along the path protects configured links
  - - limited path constraints
  - + no additional path definitions configuration

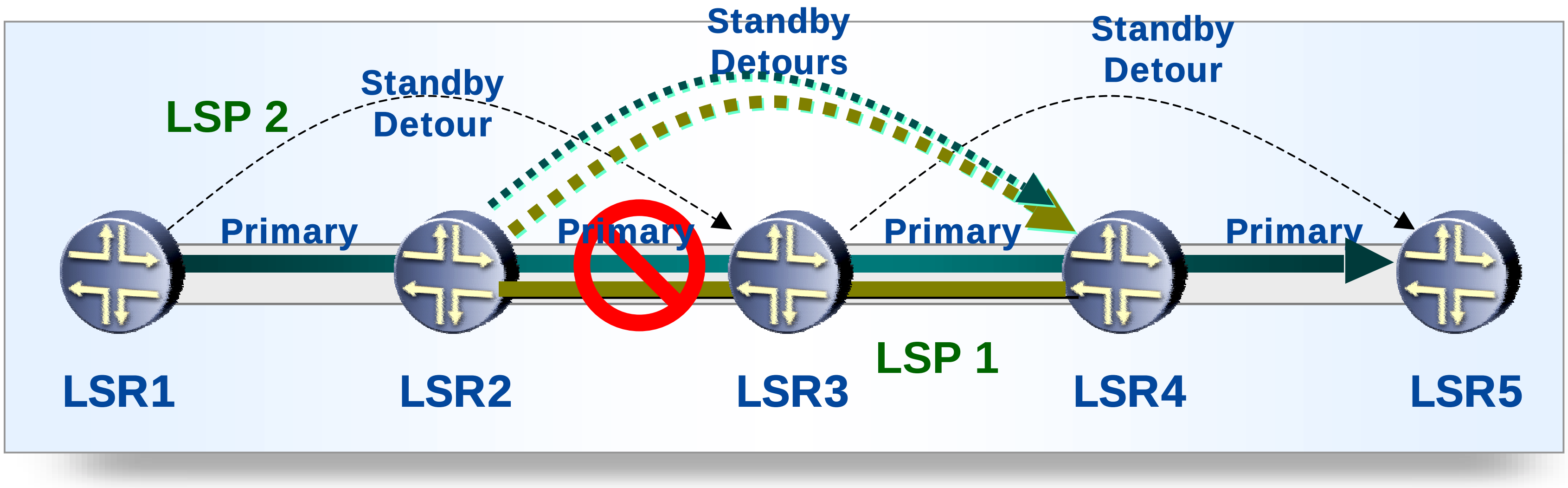
Local 1:1 Protection Operation

Single user command at head end to enable Fast Reroute.



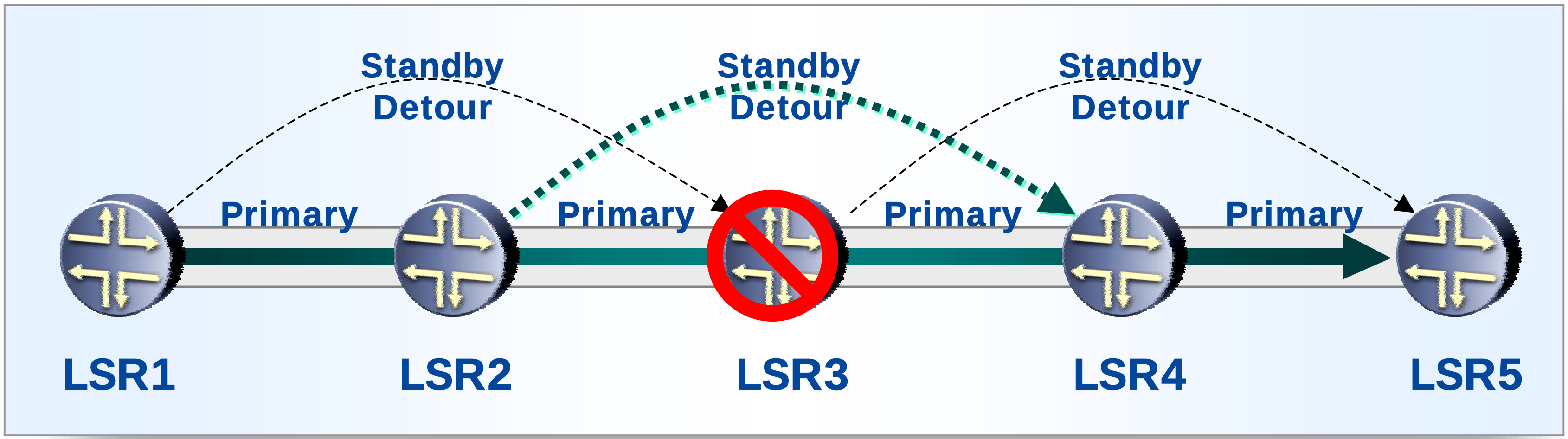
- **Fast reroute is signaled to each LSR in the path**
- **Each LSR computes and sets up a detour path that avoids the next link and next LSR**
- **Each LSR along the path uses the same route constraints used by head-end LSR**

Local 1:1 Protection Operation: Link Failure



- **LSR2 detects that an interface in an LSP has gone down and reroutes via standby detour**
  - Recovery time is limited by the time to detect the failure
    - Comparable to SONET APS
- **Packet loss is minimized to the unlucky few that were transiting at the time of failure**

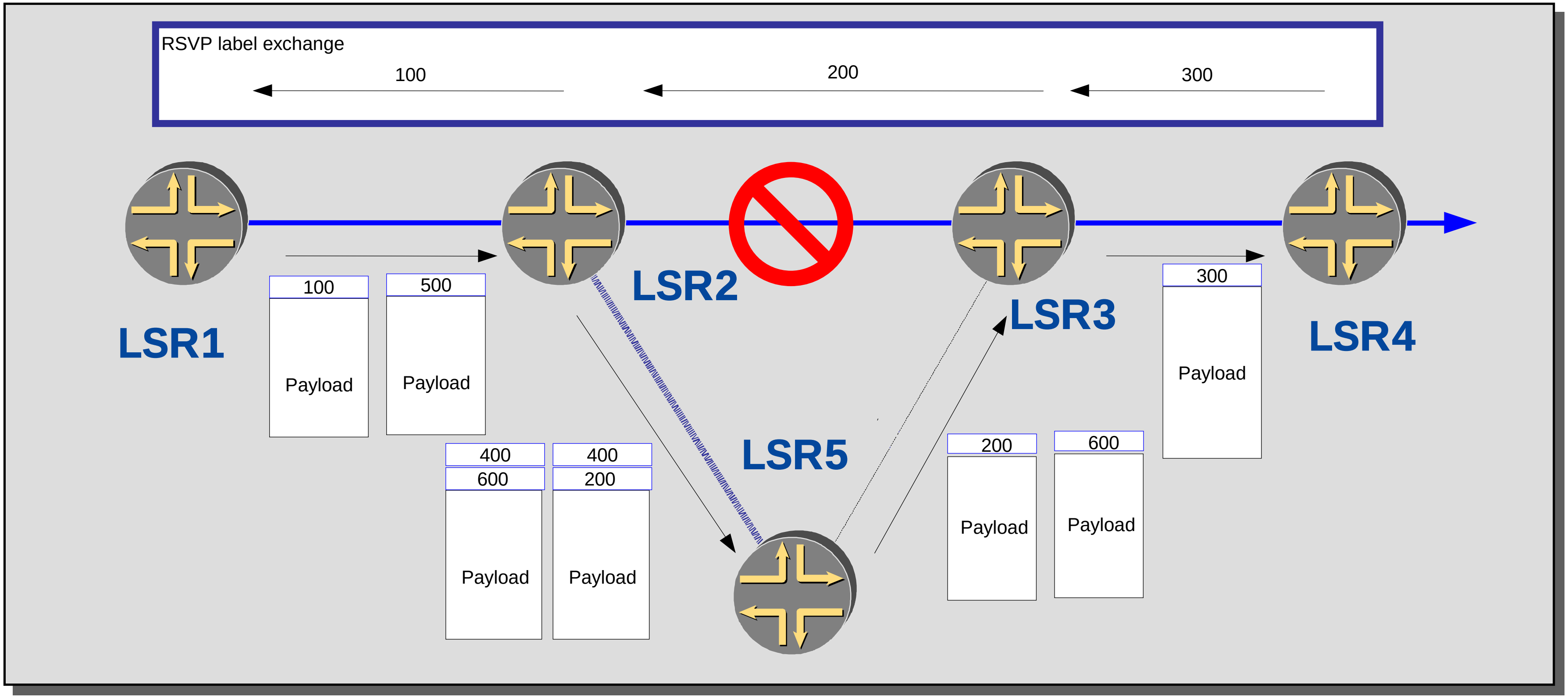
Local 1:1 Protection Operation: Node Failure



- **LSR2 detects that neighbor's (LSR3) forwarding plane has gone down and reroutes via standby detour**
  - Recovery time is limited by the time to detect the failure
- **Packet loss is minimized to the unlucky few that were transiting at the time of failure**

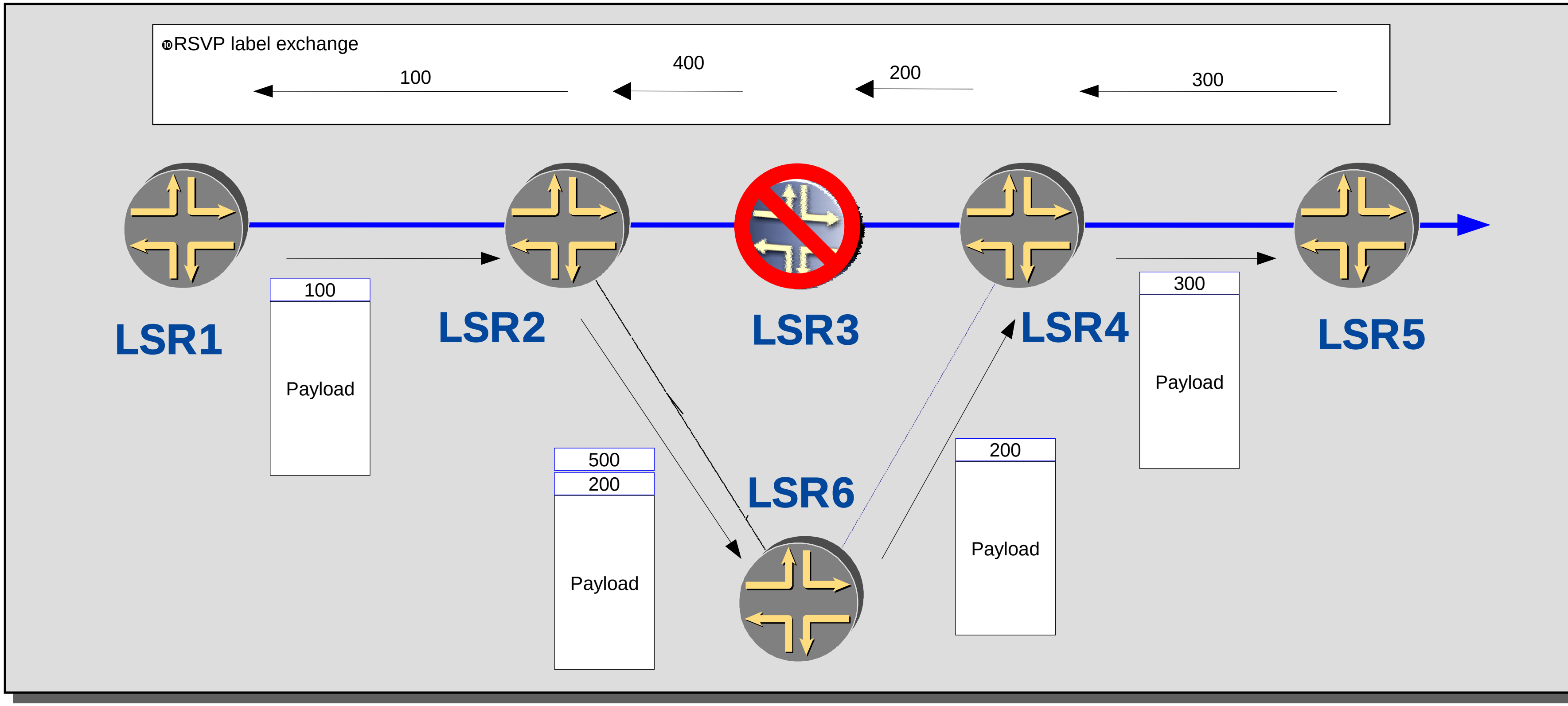
Juniper Your Net

1:N Link Protection



- Each LSR detects that an interface has gone down and reroutes all the Protected LSPs traversing the interface via the Bypass LSP
  - Recovery time is limited by the time to detect the failure
- Packet loss is minimized to the unlucky few that were transiting at the time of failure

1:N Node Protection

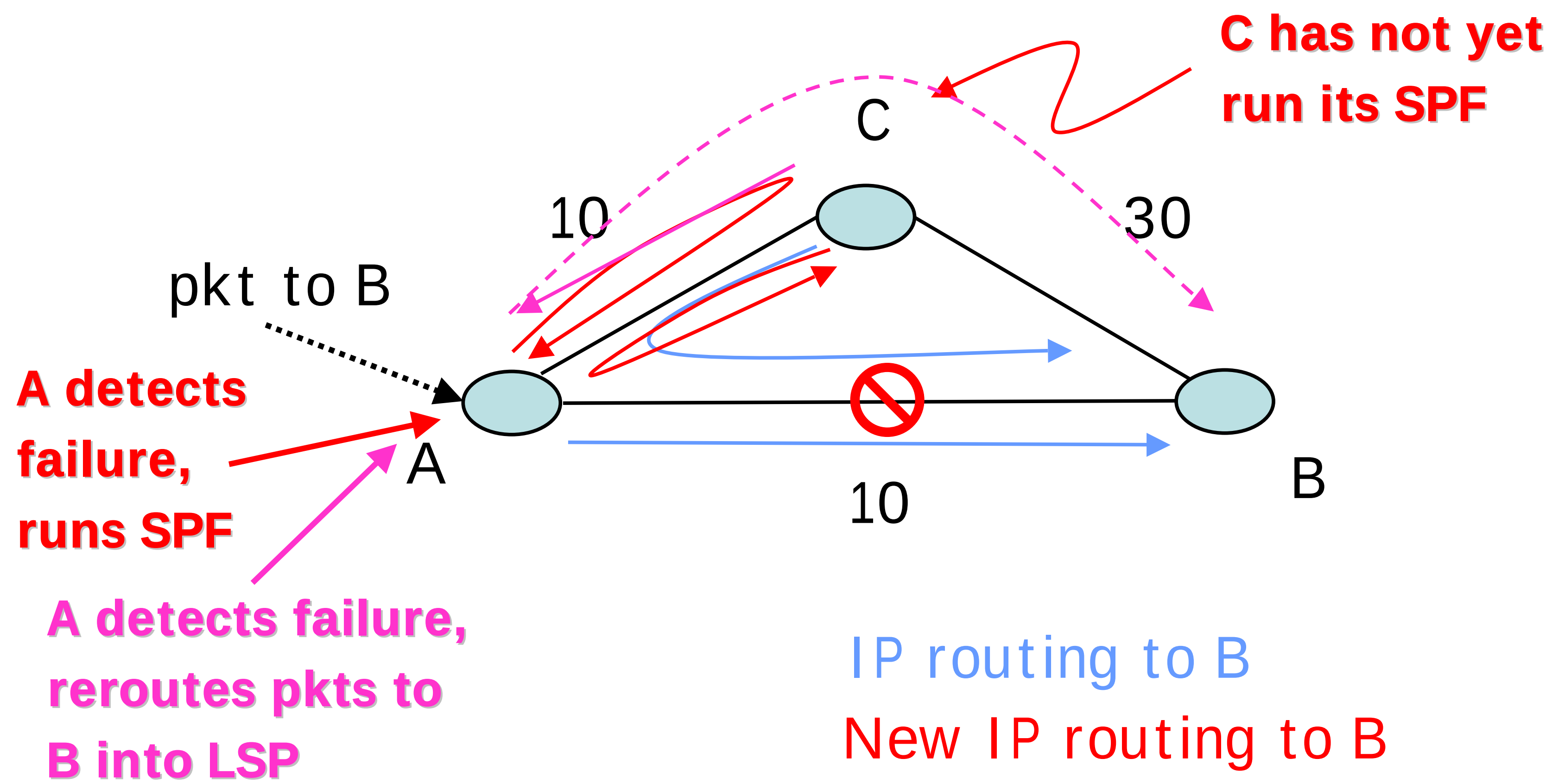


- **Each LSR detects that an interface has gone down and reroutes all the Protected LSPs traversing the interface via the Bypass LSP**
  - **Recovery time is limited by the time to detect the failure**
- **Packet loss is minimized to the unlucky few that were transiting at the time of failure**

Which one to use?

- 1:1 Detour Backup
  - The number of LSPs to be protected is small
  - Finer control (at the granularity of individual LSPs) with respect to LSP priority, bandwidth, link coloring for detour/bypass LSPs is important
  - Simpler configuration is desired
  - Suitable if LSP's have divergent paths
- 1:n Facility Backup
  - Ability to protect all the LSP's on a link with a single LSP with stacking

MPLS Fast Reroute vs IP



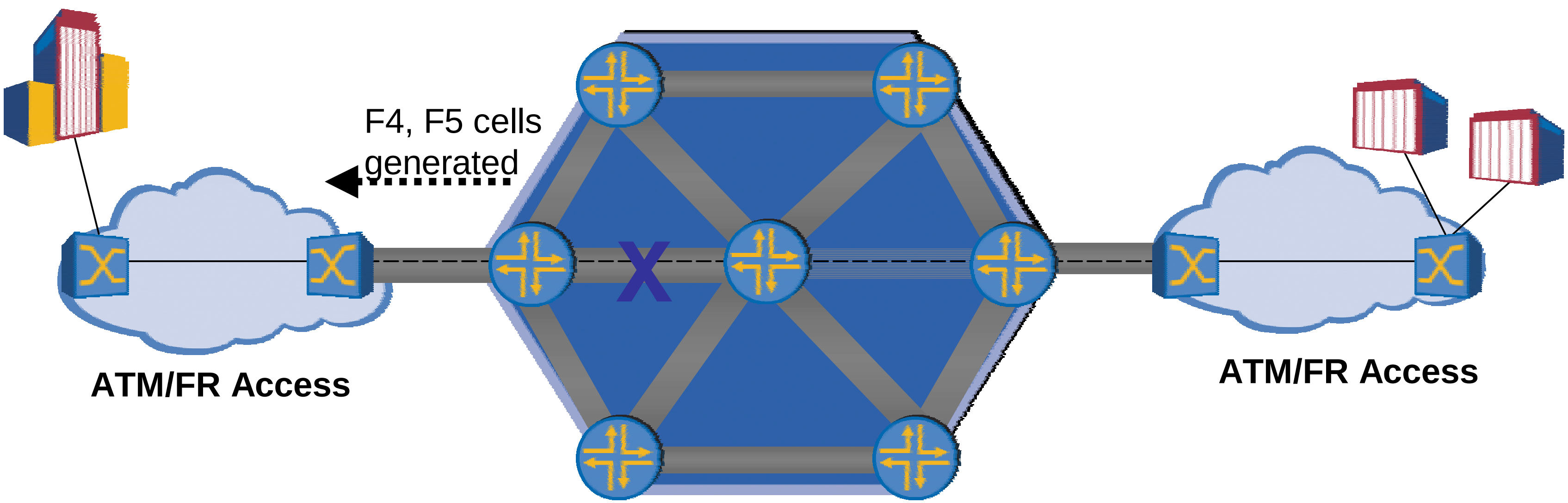
IP routing to B

New IP routing to B
(transient forwarding loop)

MPLS detour to B (no loops)

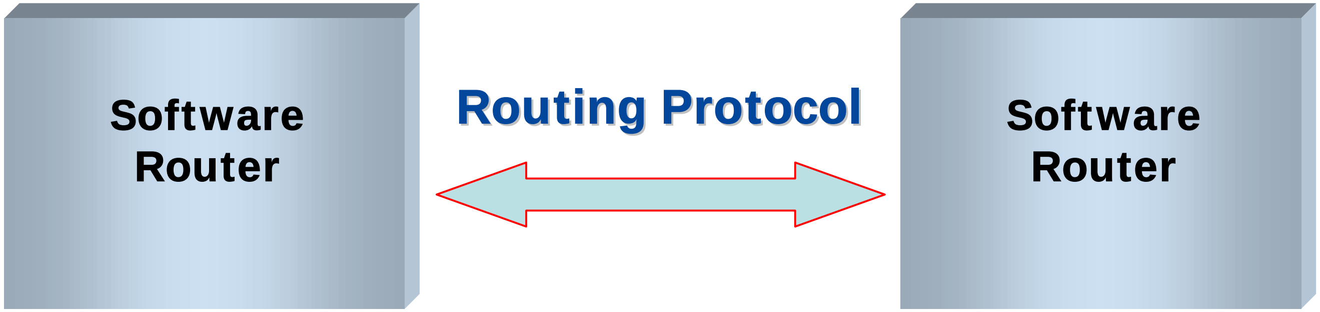
Extending to Legacy Networks

- **MPLS OAM features**
  - Use BFD and FRR, along with other mechanisms
  - Provides notification to external networks if LSP fails



BFD:Forwarding Liveliness
(Bidirectional Forwarding Detection)

- **In IP, historically a function of the routing protocol**
  - Because formerly, routing = forwarding
  - Fault resolution in perhaps tens of seconds
  - This is too slow for anything but best-effort IP
  - Sometimes there is no routing protocol!



Juniper your Net

Goals of BFD

- **Faster** convergence **of routing protocols, particularly on shared media (Ethernet)**
- **Semantic** separation **of forwarding plane connectivity and control plane connectivity**
- Detection **of forwarding plane-to-forwarding plane connectivity (including links, interfaces, tunnels etc.)**
- **A** single **mechanism that is independent of media, routing protocol, and data protocol**
- **Requiring** no changes **to existing protocols**

BFD Protocol Overview

- At its heart, Yet Another Hello Protocol
- Packets sent at intervals; neighbor failure detected when packets stop arriving
- Intended to be implemented in the forwarding plane where possible
- Context defined by encapsulating protocol
- Always unicast, even on shared media

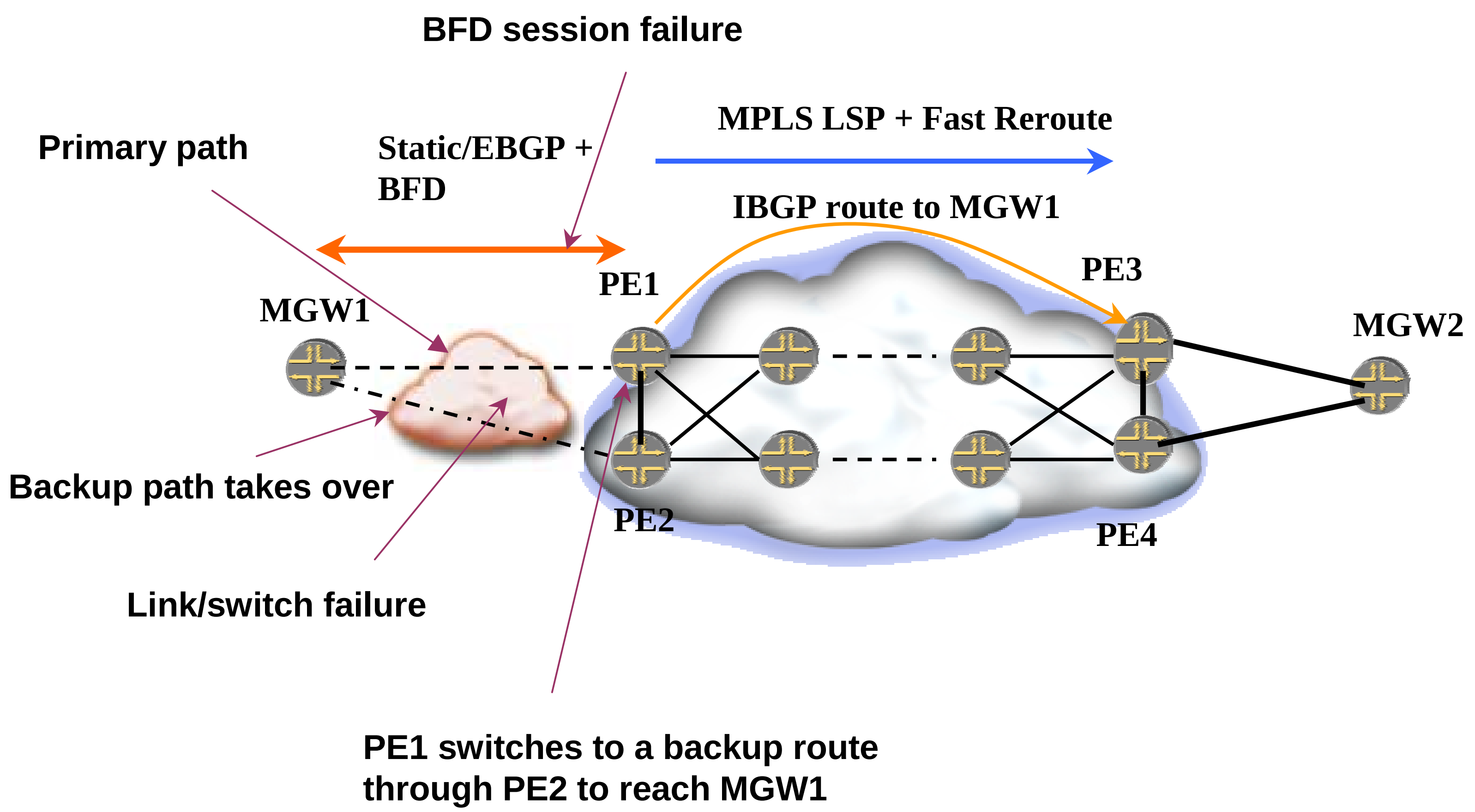
BFD Applications

- **IGP liveliness detection**
- **Tunnel liveliness detection**
  - MPLS LSPs
  - IP-in-IP/GRE tunnels
- **Edge network availability**
- **Liveness of static routes**
- **Host reachability (e.g media gateways)**
- **Switched Ethernet integrity**

BFD for IGP Liveliness Detection

- **One of the first motivations for BFD**
- **Faster convergence particularly on shared media**
  - Sub-second IGP adjacency failure detection
- **IGP hellos can be set to higher intervals**
  - Can improve IGP adjacency scaling

BFD for Edge Availability Voice over IP



Juniper your Net

Summary

Dependability:

- Is a culture
- Has many layers
- Is business critical
- Must be designed into networks from the start

- **Luckily:**
  - Vendors are providing tools for reliability
  - Many architectural options from which to choose
  - Also many protocols and mechanisms

Thank you!