Delivering High Availability Routed Networks

Matt Kolon
matt@juniper.net

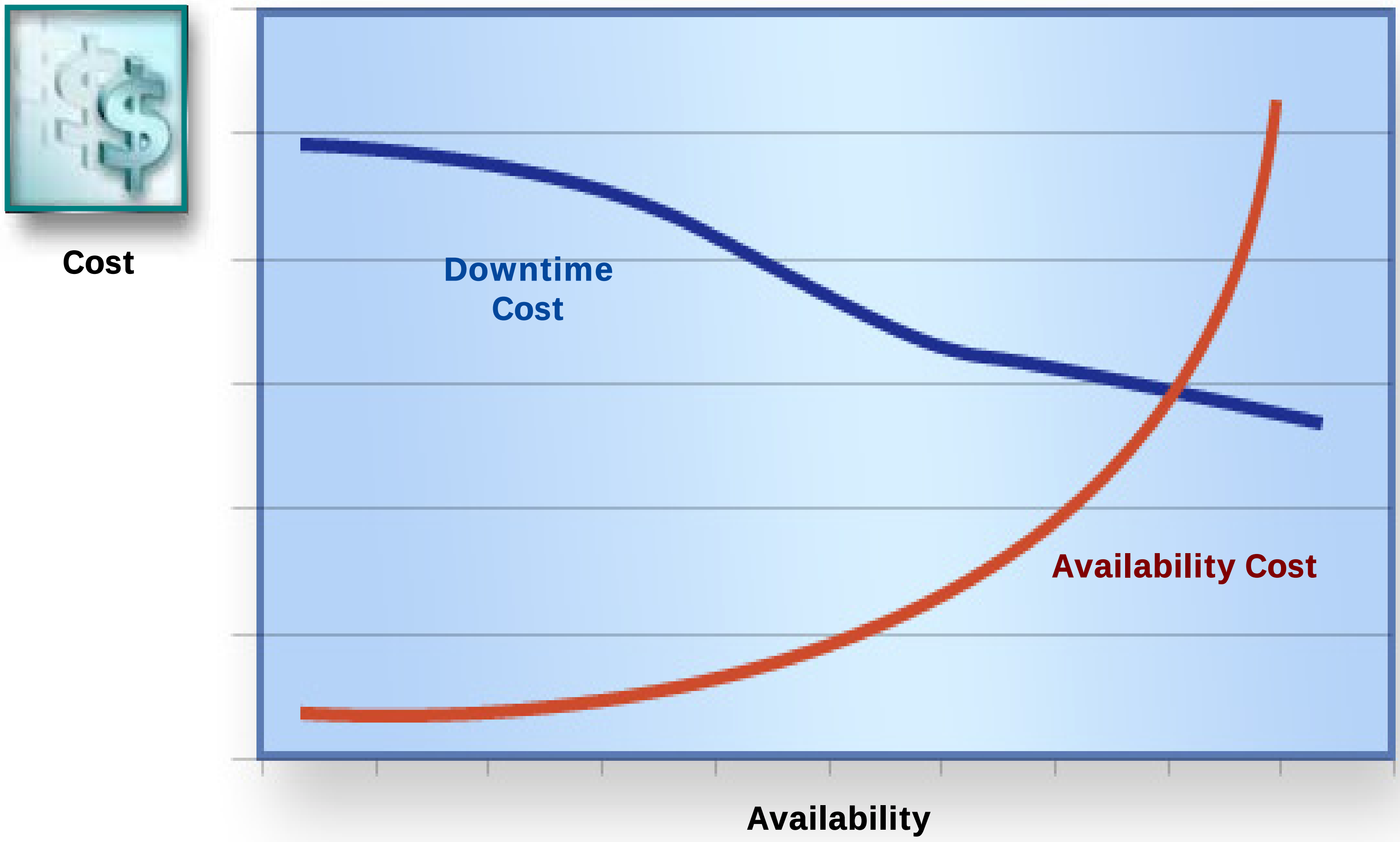
APRICOT 2005 - Kyoto

Today's IP network

- **Is an infrastructure that supports:**
  - VoIP
  - Converged data network services
  - Business VPN Services
  - And Internet access services

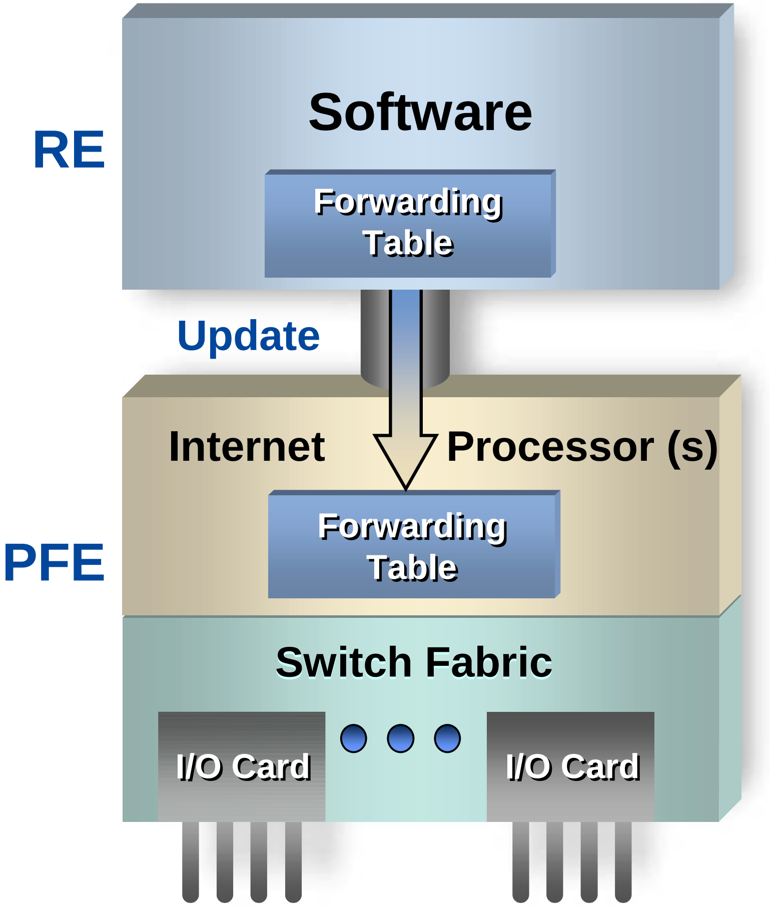
  - These carrier services typically have customer SLA's that must be supported

Business Case for High Availability

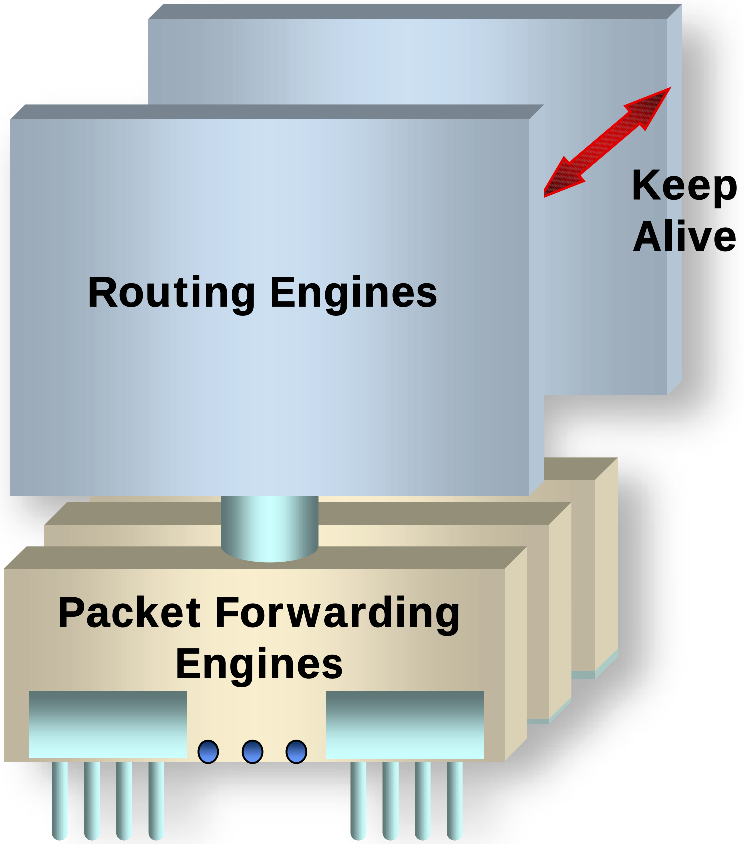


A Logical Platform View

- 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

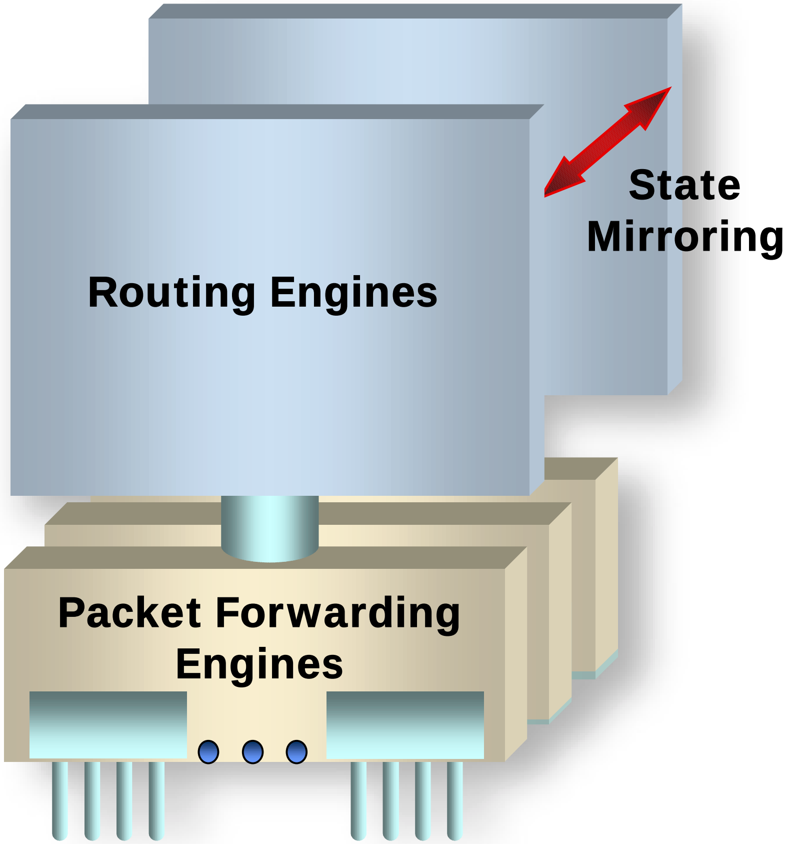


Simple RE Failover



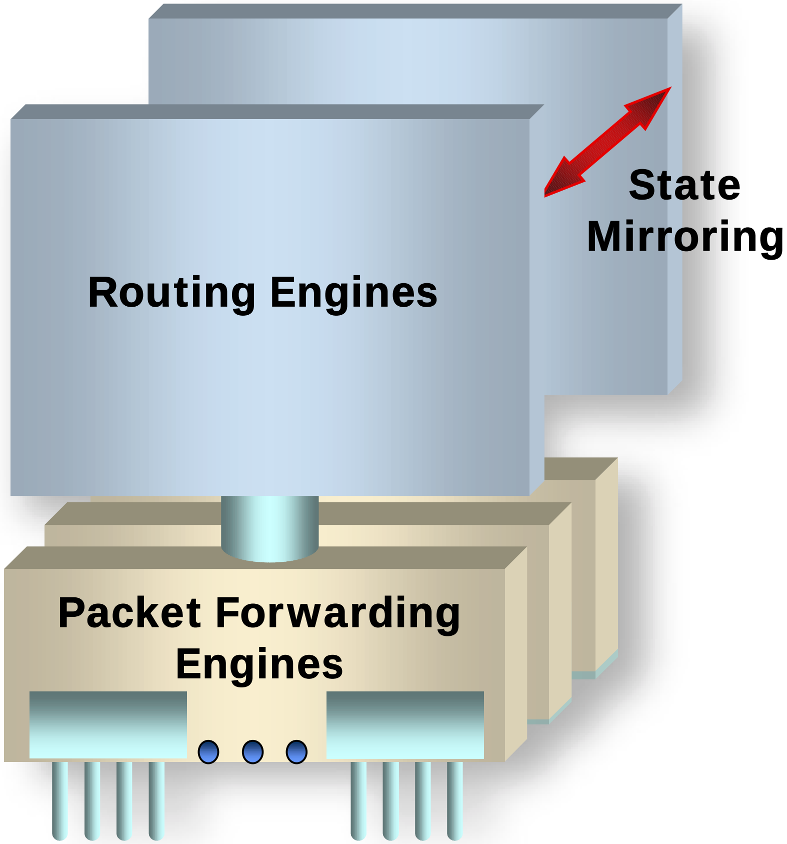
- Protects against Single Node Hardware Failure
- Redundant Routing Engines run keepalive process
- Automatic failover to secondary
- Configuration synchronized between RE's
- Configurable timer
- Routing Process restarts
- Requires PFE reset

Stateful Protocol Mirroring



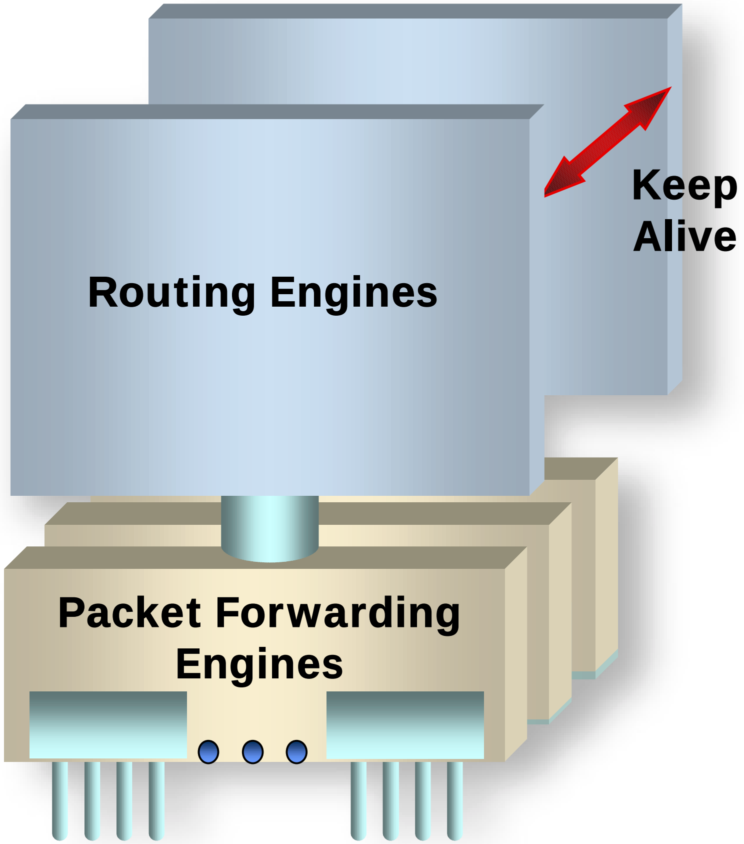
- Protects against Single Node Hardware Failure
- Redundant Routing Engines Mirror each others state
- BGP & TCP
- Theoretically ISIS & OSPF
- Automatic failover to secondary
- Advocated by some vendors, claiming Carrier-Class IP

Stateful Protocol Mirroring



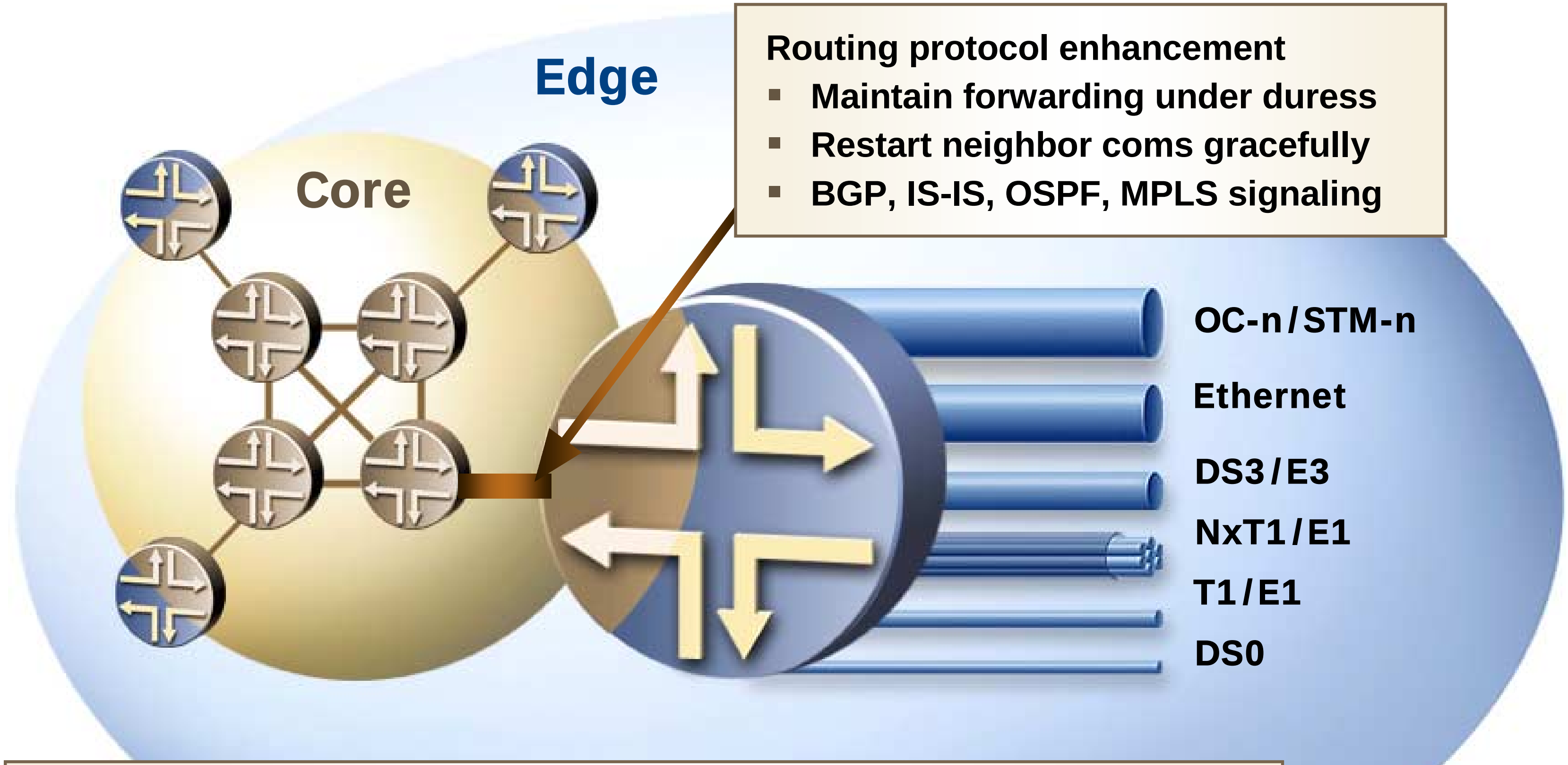
- Great Idea!
- Difficult to do without replicating errors as well as “good” state
- Potential for “bug mirroring”
- Much more challenging in a rich service environment than an IP-only core

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

Routing Protocol Graceful Restart

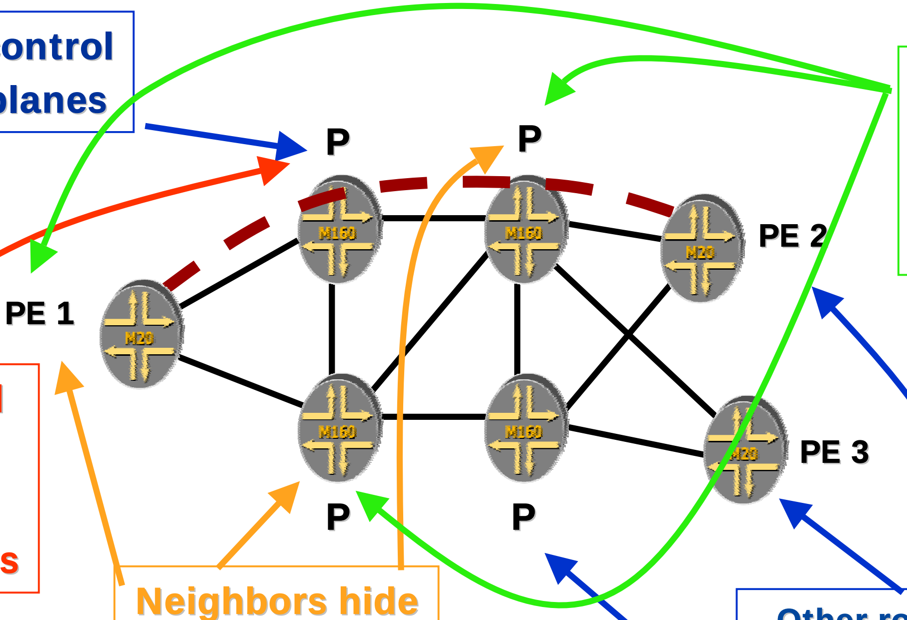


Routing protocol enhancement

- 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 ?



Separate control and data planes

When router recovers, neighbors sync up without disturbing forwarding.

If router's control plane fails, data plane can keep forwarding packets

Neighbors hide failure from all others routers in the network

Other routers are never made aware of failure

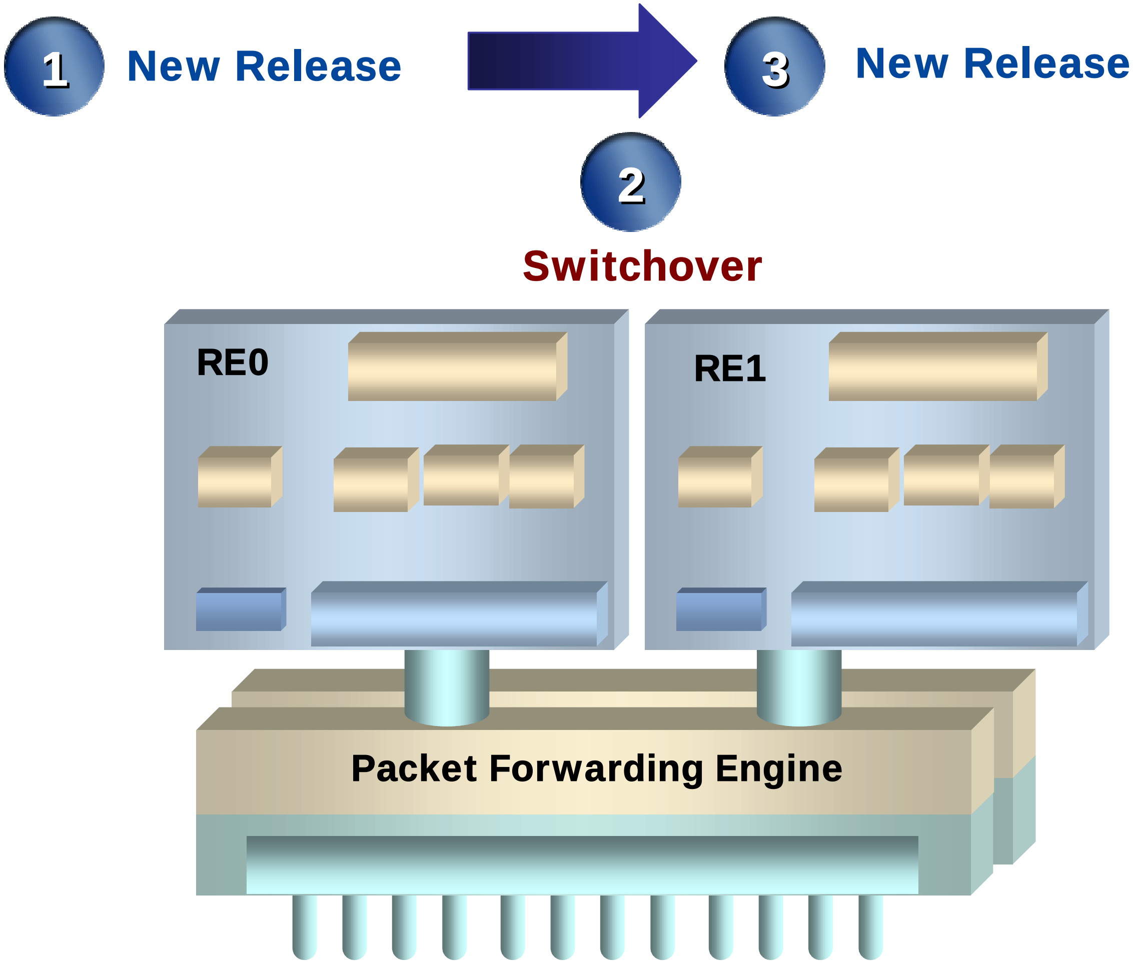
Software Reliability Principles

- **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

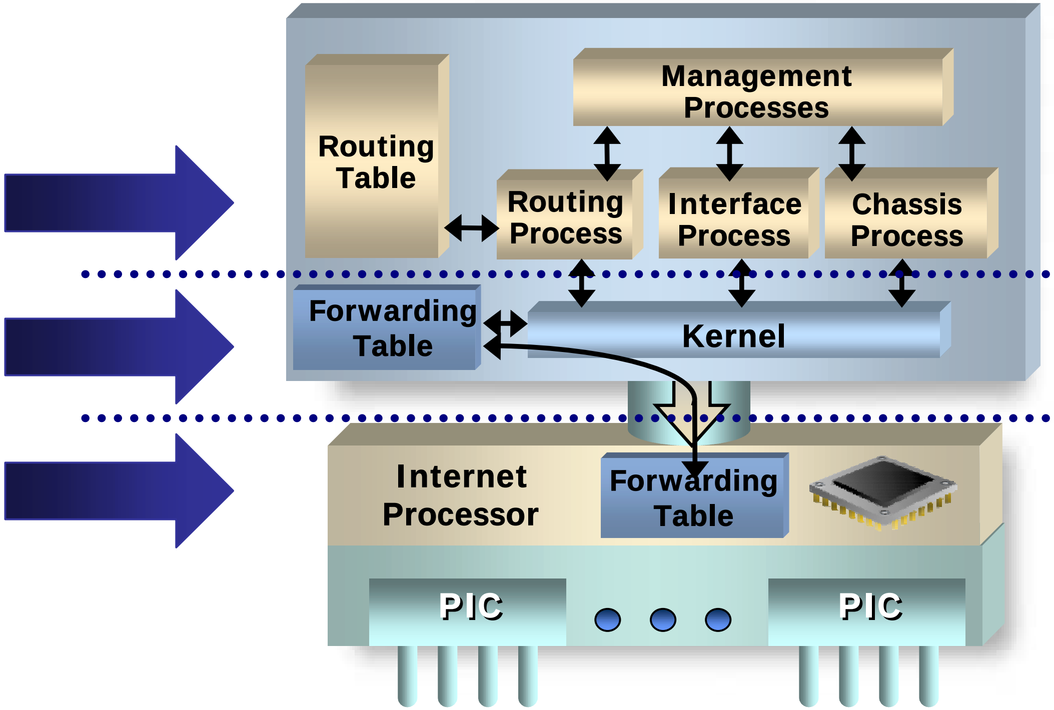
In-Service Software Upgrades

- **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**

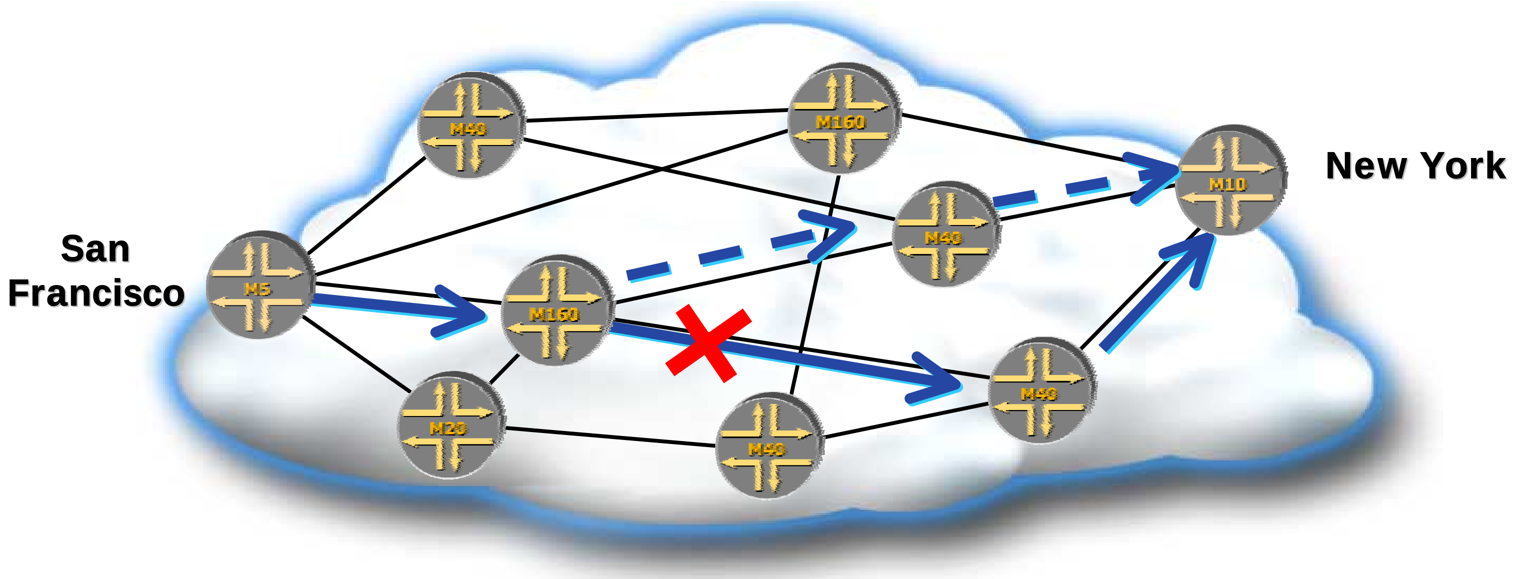


In-Service Software Upgrades

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

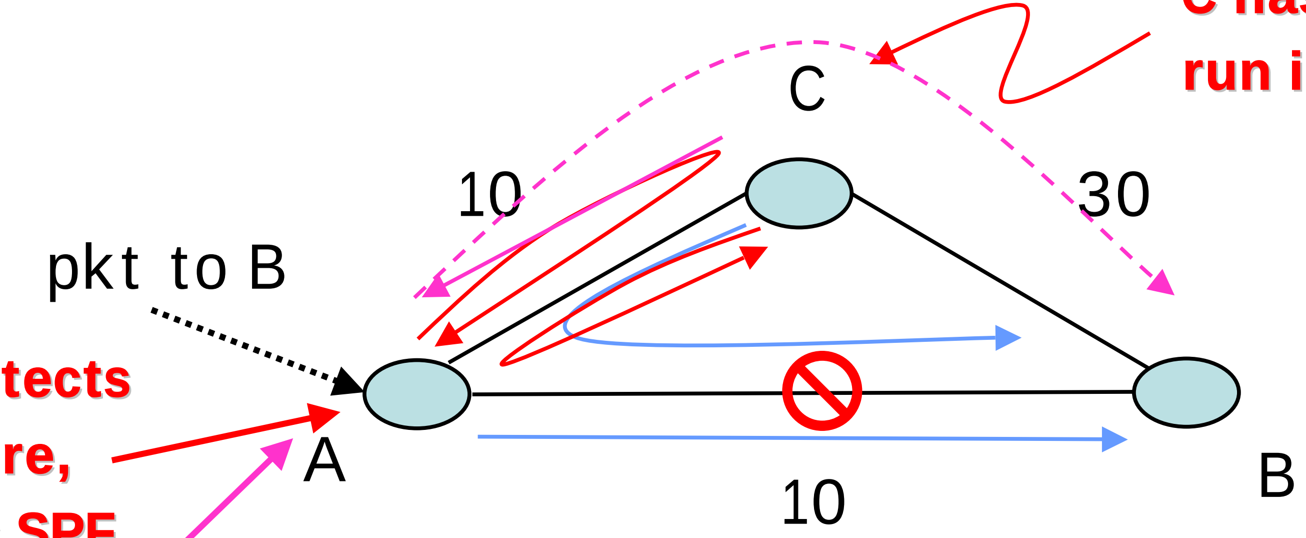


IP Dynamic Routing



- 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

MPLS Fast Reroute vs IP



C has not yet run its SPF

A detects failure, runs SPF

A detects failure, reroutes pkts to B into LSP

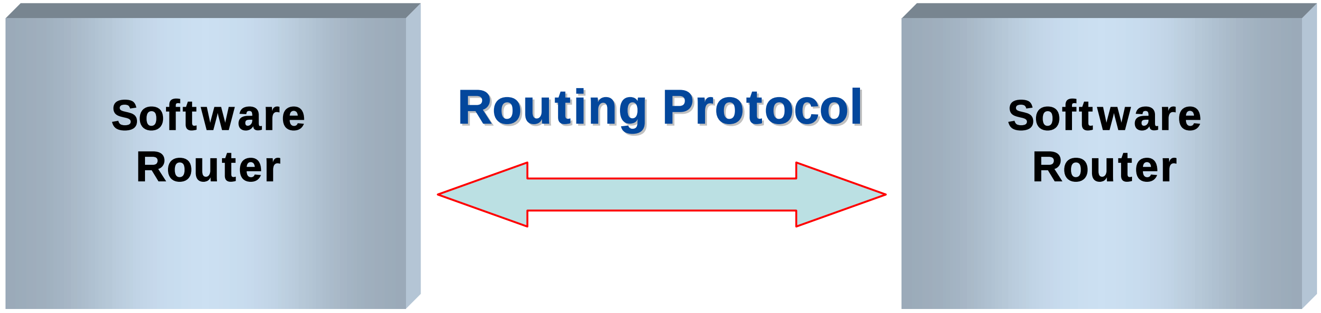
IP routing to B

New IP routing to B (transient forwarding loop)

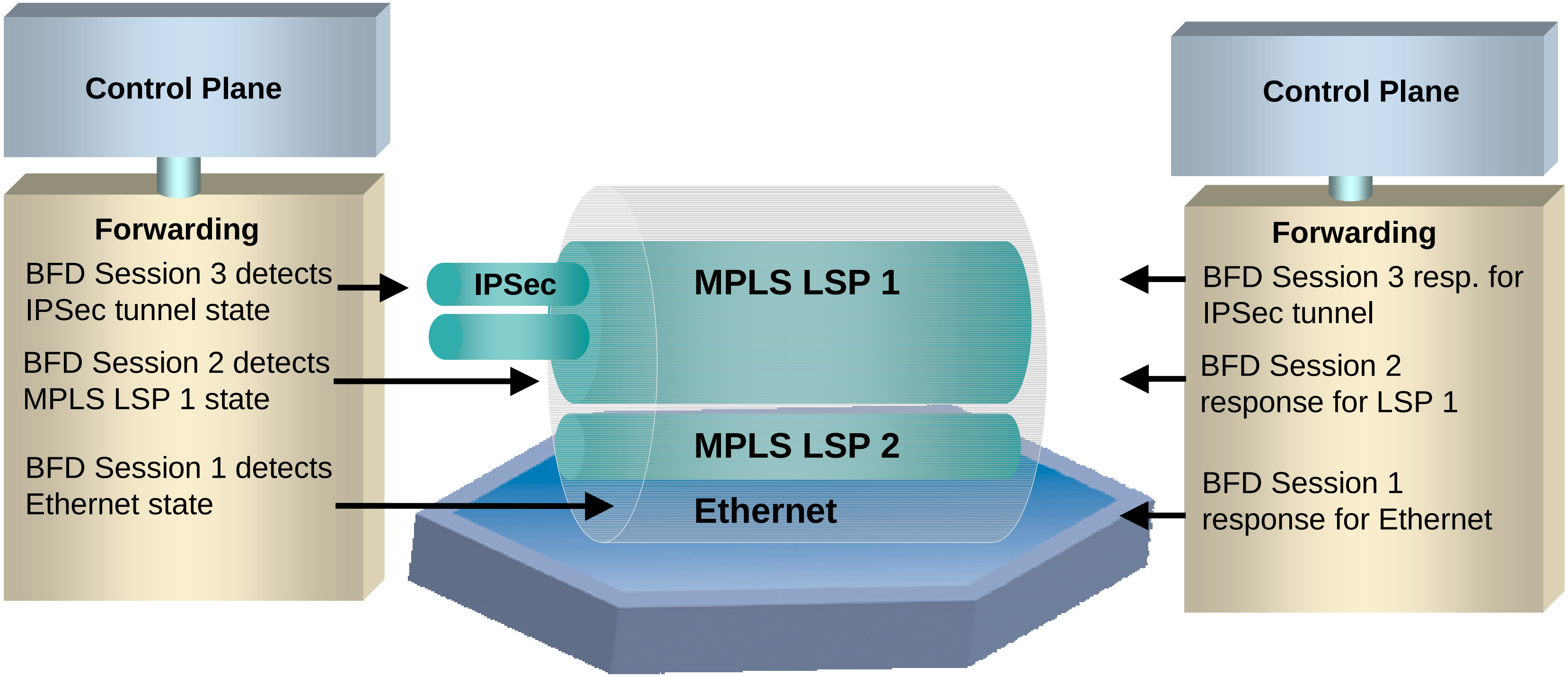
MPLS detour to B (no loops)

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!



BFD Applications

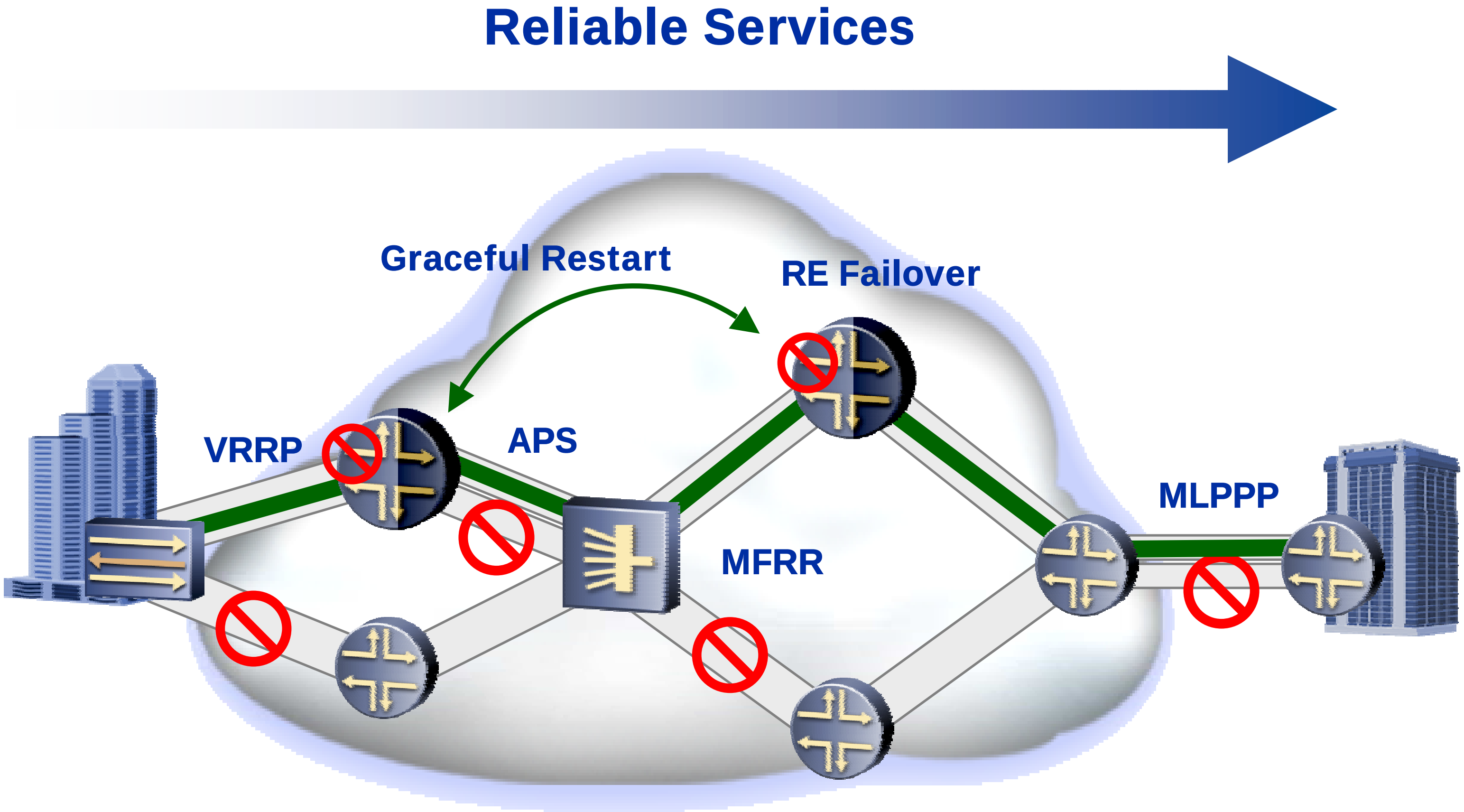

Control Plane
Forwarding
BFD Session 3 detects IPSec tunnel state
BFD Session 2 detects MPLS LSP 1 state
BFD Session 1 detects Ethernet state
IPSec
MPLS LSP 1
MPLS LSP 2
Ethernet
Control Plane
Forwarding
BFD Session 3 resp. for IPSec tunnel
BFD Session 2 response for LSP 1
BFD Session 1 response for Ethernet

BFD can detect failures within across many transports, and is therefore useful for many applications.

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**

Goal: Reliable Services



Thank you!