IP/MPLS CORE –
HIGH AVAILABILITY DESIGN

Mars Chen <email: marschen@juniper.net>

27th January 2010

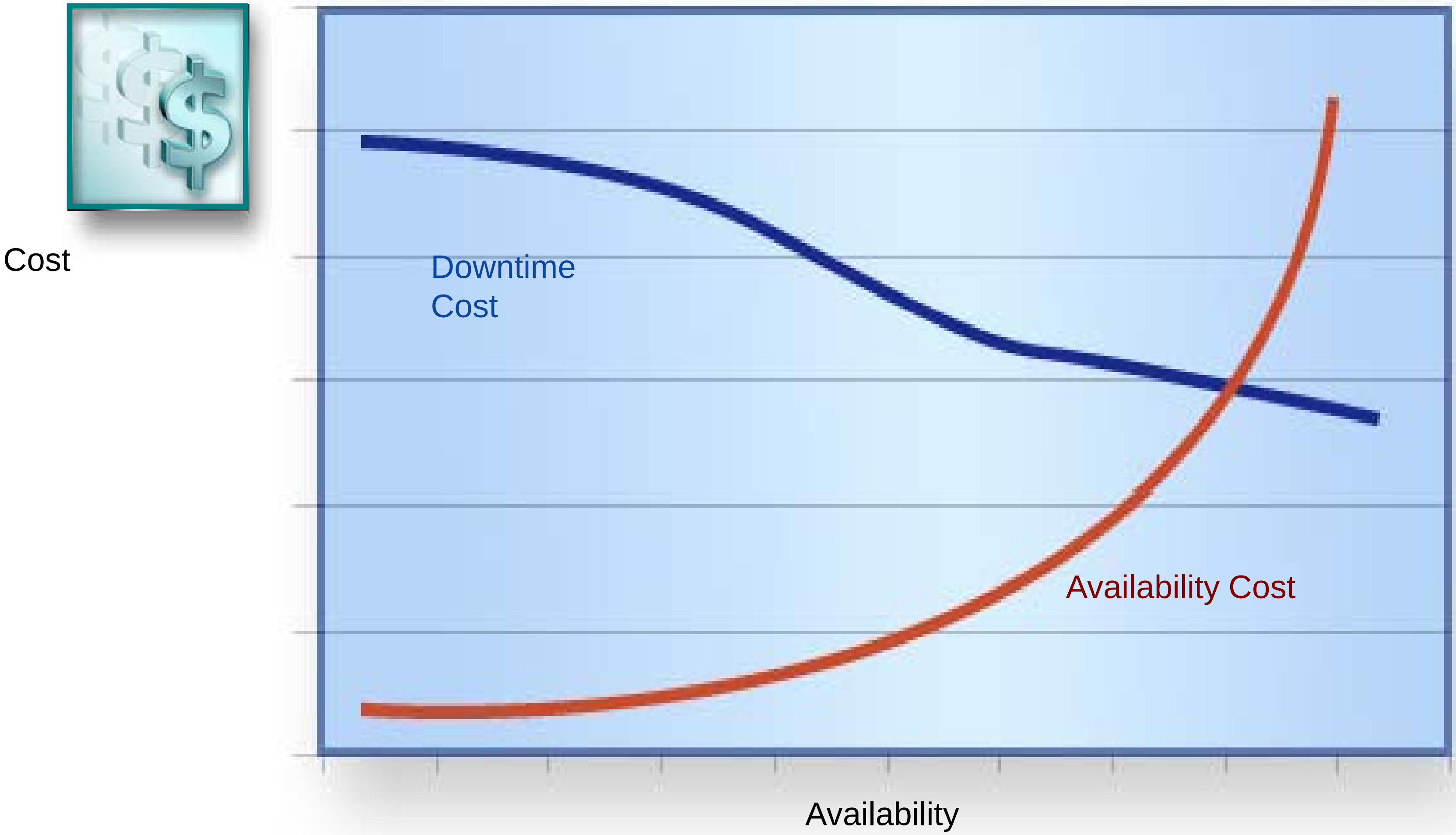
TODAY’S IP NETWORK

Is an infrastructure that supports mission critical services:

- VoIP/Mobility
- Converged data network services
- Business VPN Services
- Cloud Computing
- And Internet access services
- …………………..

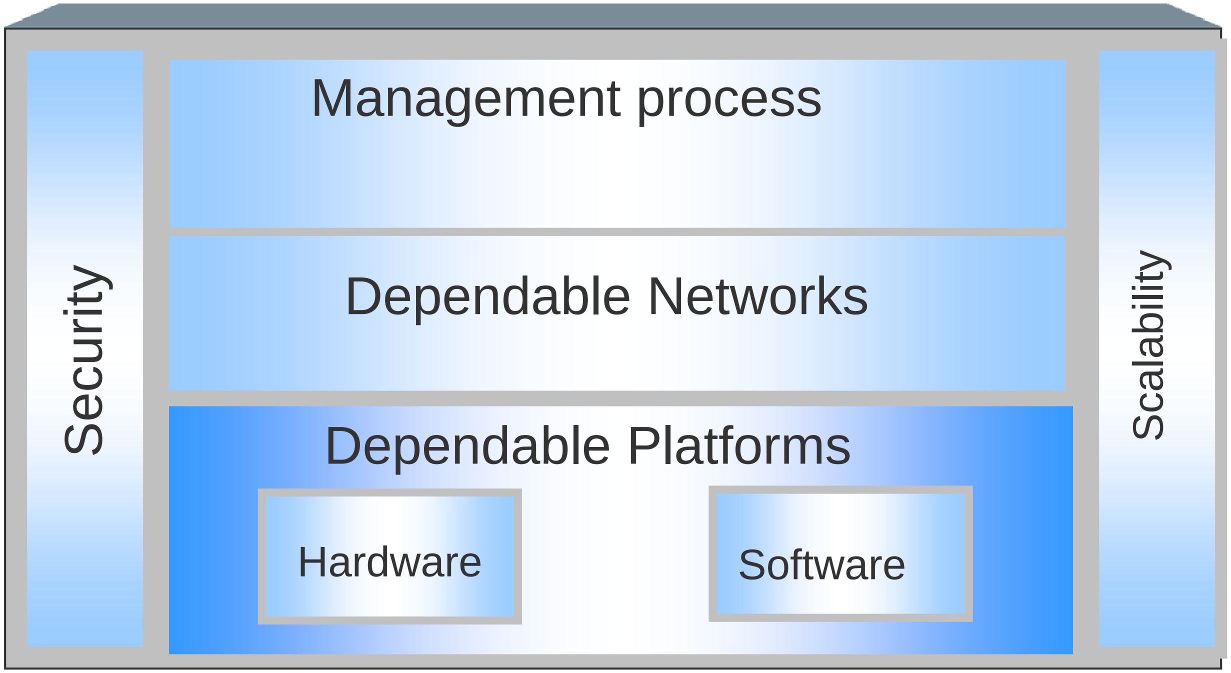
- These carrier services typically have customer SLA’s that must be supported

BUSINESS CASE FOR HIGH AVAILABILITY



HIGH AVAILABILITY SOLUTION ARCHITECTURE

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



Security

Management process

Dependable Networks

Dependable Platforms

Hardware

Software

Scalability

PLATFORM HIGH AVAILABILITY

1. Hardware
2. Software
3. Control plane
   - Nonstop forwarding (NSF) : Graceful restart + GRES
   - Nonstop routing
4. Virtualization – Virtual Chassis

LOGICAL PLATFORM VIEW OF MODERN ROUTER

Clean separation of routing and packet forwarding functions

- Consistent performance
- Stability
- Provider-class routing

Routing Engine (RE)

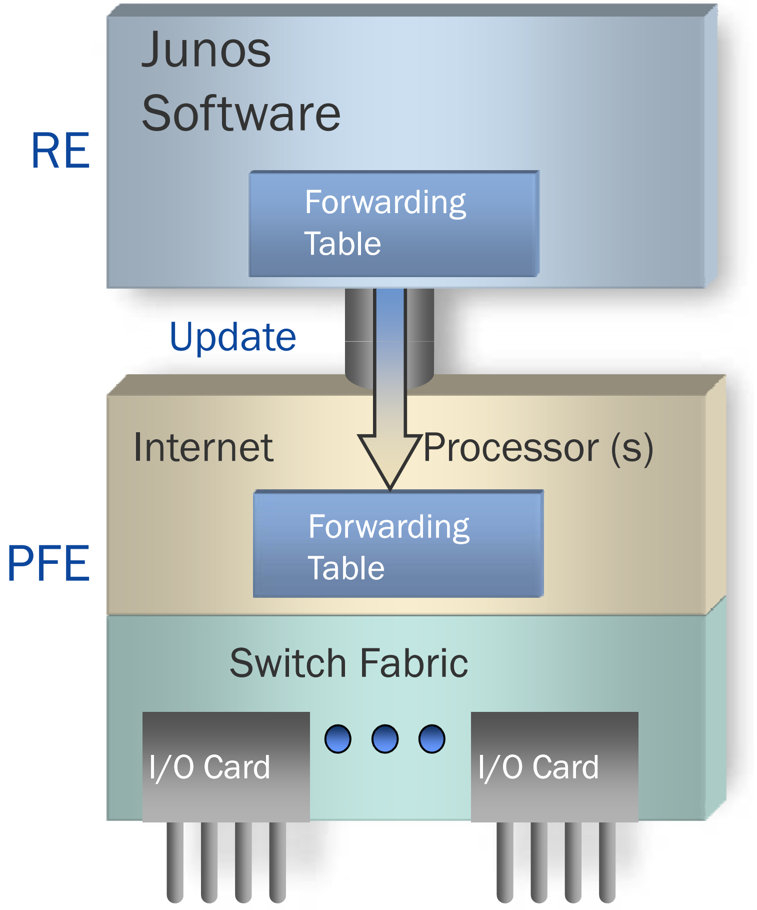
- JUNOS software

Packet Forwarding Engine (PFE)

- Processor-based design

Interfaces

- FPC/PICs



ROUTING ENGINE REDUNDANCY
GRACEFUL ROUTING ENGINE SWITCHOVER (GRES)

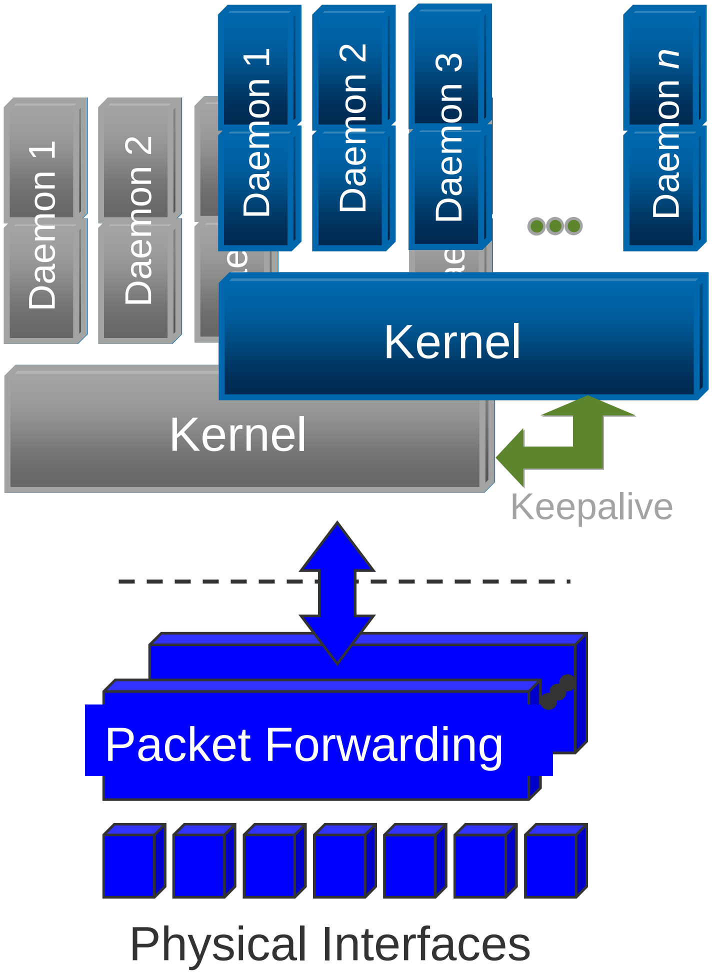
Control plane and forwarding plane separation allows continuous packet forwarding during control plane failure

GRES provides stateful replication between the master and backup REs

- Keepalives exchange info such as interfaces, and kernel.
- GRES allows fast switchover during RE failure

GRES alone is not enough

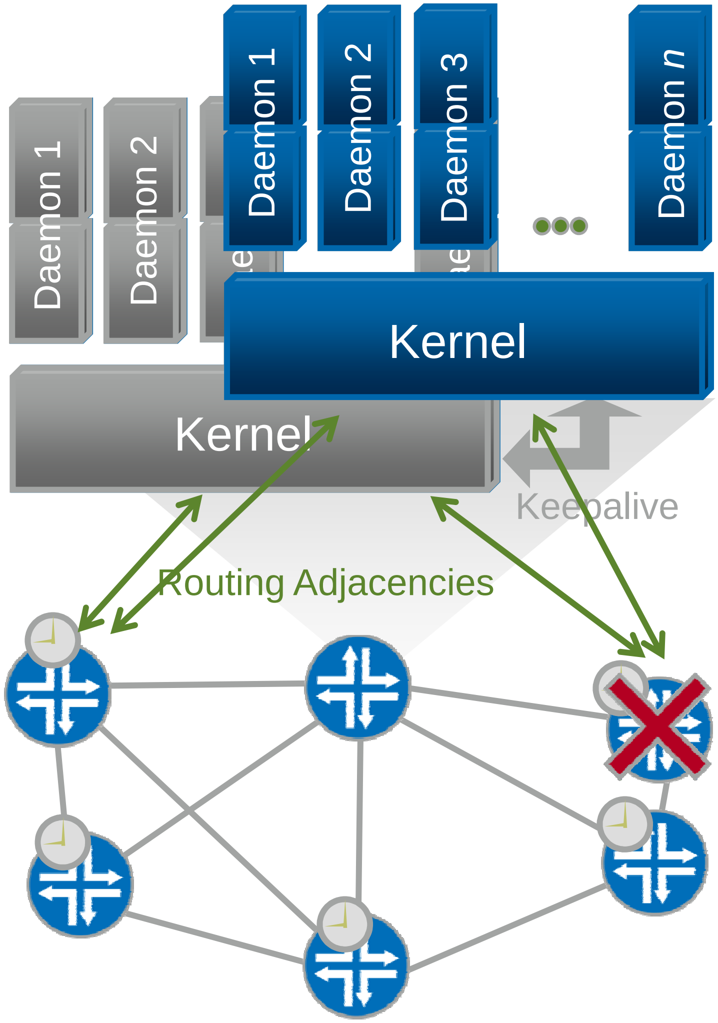
- Routing adjacencies broken during switchover
- Must be combined with either graceful restart protocol extensions or nonstop active routing



NONSTOP FORWARDING

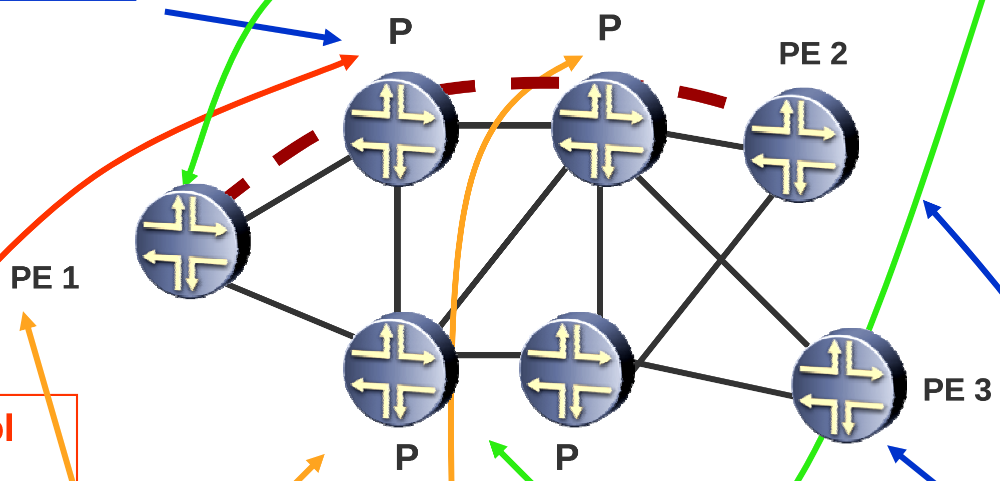
GRES + Graceful Restart

= NSF (Nonstop Forwarding)



GRACEFUL RESTART: OPERATIONS

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

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* rfc4724
OSPF	Protocol extensions New opaque-LSA type 9, "Grace-LSA"	*Graceful OSPF Restart* rfc3623
IS-IS	Protocol extensions 3 new timers New "re-start" option (TLV) in IIH PDU	*Restart Signaling for ISIS* rfc3847
MPLS	Protocol Extensions Uses signaling as described in "Graceful Restart Mechanism for BGP	Graceful Restart Mechanism for BGP with MPLS Rfc4781
RSVP	Protocol Extensions Extend rfc 3473 Recovery ERO	Extensions to GMPLS RSVP Graceful Restart Rfc5063

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 re-convergence might exceed grace period
  - For example, if there are hundreds of BGP peers
- Protocol interdependencies can slow re-convergence beyond grace period
  - For example, if BGP depends on LDP restart completion, which depends on RSVP restart completion
- Operators acceptance of GR is not widespread

NON-STOP ACTIVE ROUTING

Internal processes keep backup RE aware of protocol state and adjacency activities

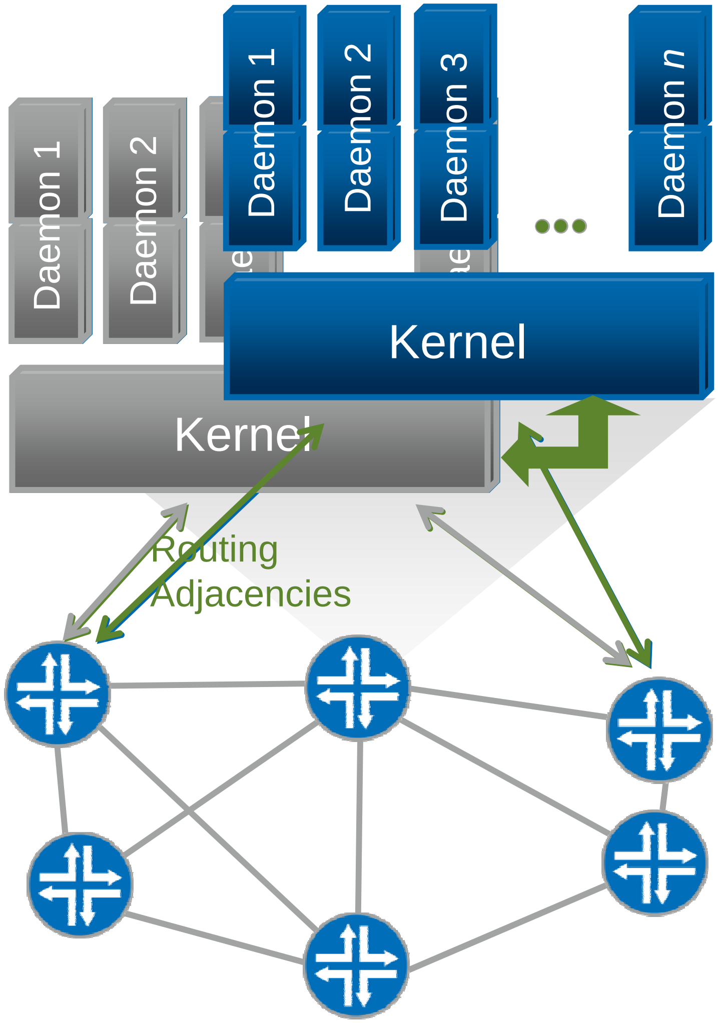
Individual routers assume sole responsibility for RE failure or switchover

- No need to run protocol extensions on neighbors

Backup routing engine becomes hot standby

- Both REs run routing protocol processes
- Relies on GRES to preserve kernel and interface info

switchover seamless to neighbors



www.juniper.net

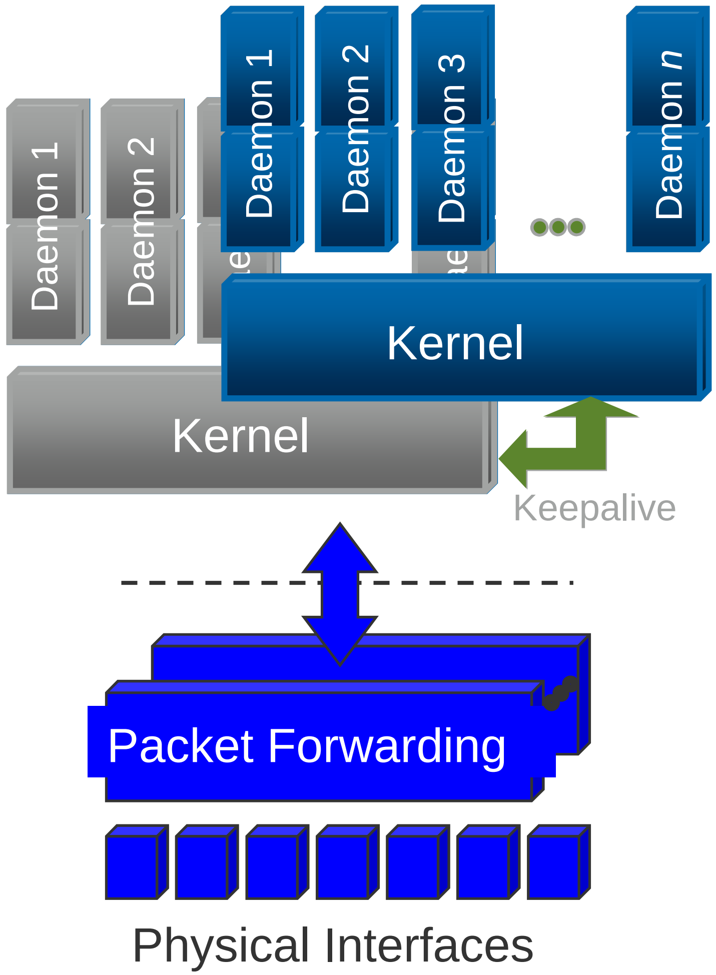
IN SERVICE SOFTWARE UPGRADE

Not every vendor extends GRES, GR and NSR support to ISSU

software upgrade isn’t merely a single point of software replacement.

Ideally, a fully redundant system can eliminate any disruption during software upgrade

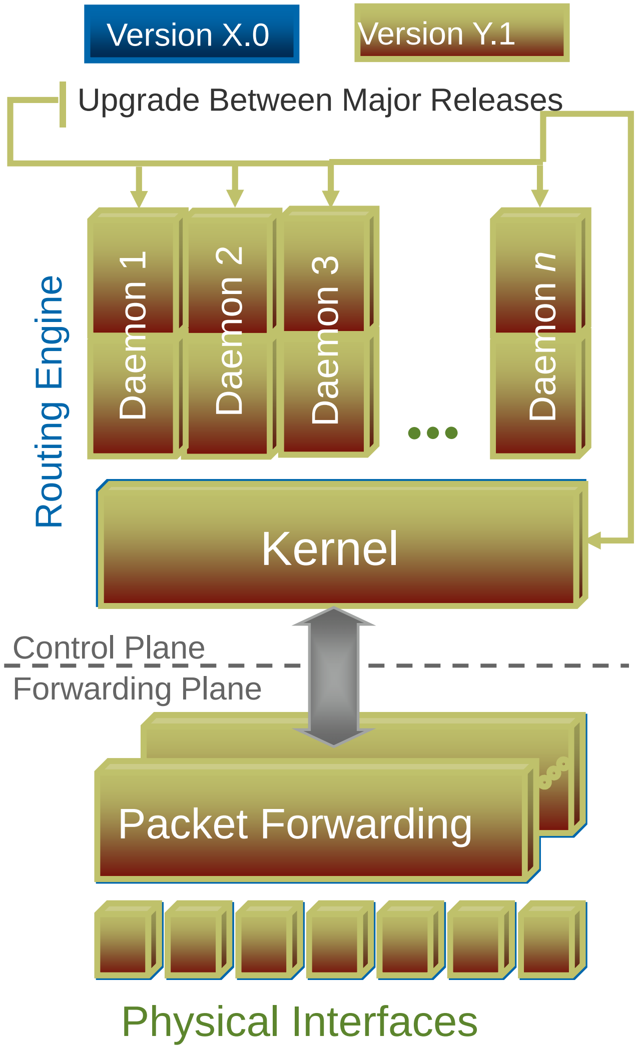
Reality is, most systems today are not fully redundant



TRUE ISSU

Upgrades entire software image

Can be done to major or minor releases



VIRTUALIZATION
VIRTUAL CHASSIS KEY ADVANTAGES

No Dependency on Dedicated Connectivity Hardware

Leverage existing Redundancy Mechanisms

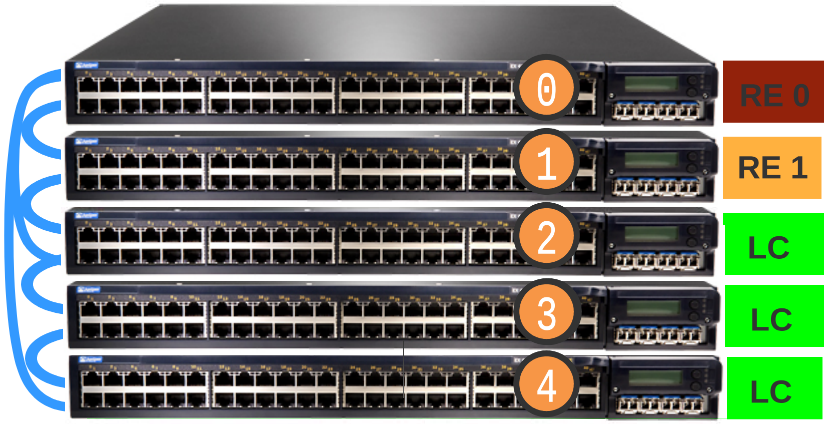
- GRES/NSR, LAG (Aggregated-Ethernet).

Operational efficiency – Single control plane as visible externally.

Failover transparent to external control entities (OSS, policy servers, AAA etc).

No routing change as visible externally: Inter-chassis link failover completely contained within VC

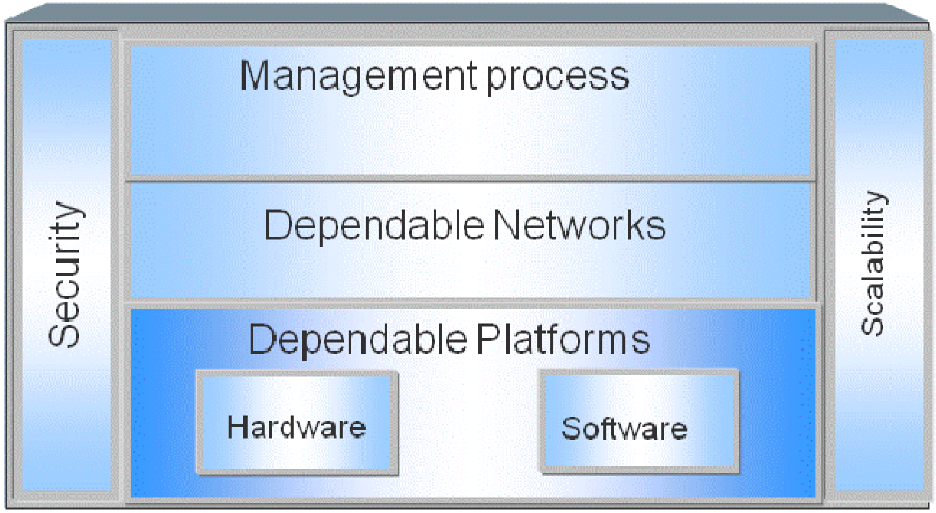
- Failover not gated by routing re-convergence.



RELIABLE NETWORKS

- SDH (Layer 1)
- Ethernet OAM (Layer 2)
- Link bundling
- VRRP
- IGP fast convergence
- BFD
- VRRP
- MPLS (RSVP) Fast reroute

- IP/LDP fast reroute



SONET/SDH PROTECTION SWITCHING

SONET APS & SDH MSP

- Redundant routers share uplink

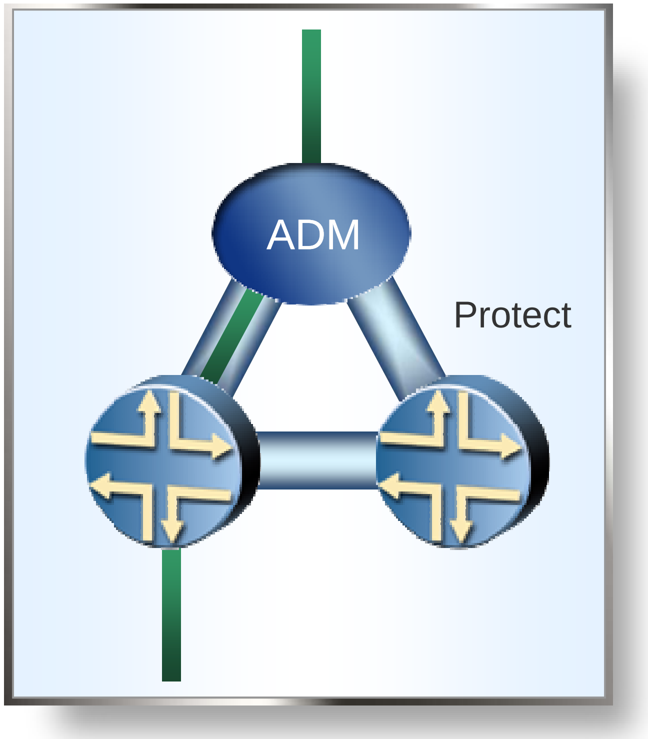
Rapid circuit failure recovery

- Used on router-to-ADM links

Interoperable with standard ADM

Working & protect circuits

- May reside on different routers
- May reside on same router



OAM LAYERS

Transport Layer

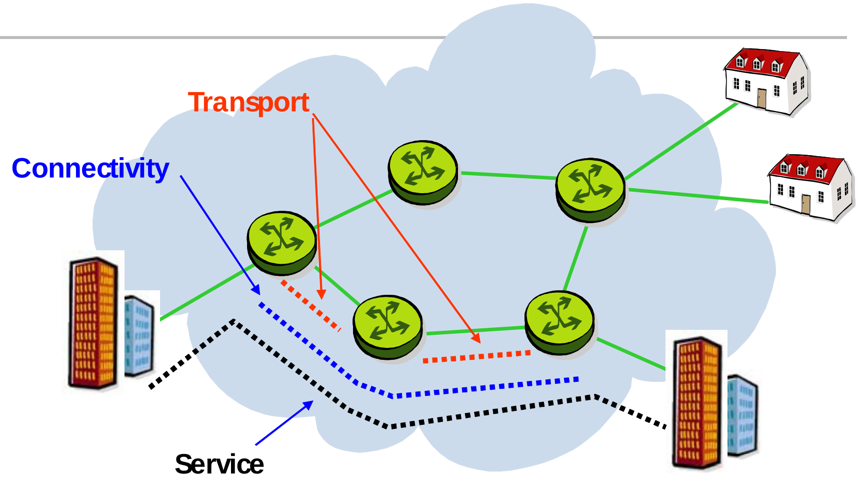
- Ensures two directly connected peers maintain bidirectional communication.
- Must detect link down failure and notify higher layer for protocol to re-route around the failure.
- Monitor link quality to ensure that performance meets an acceptable level.

Connectivity Layer

- Monitor the communication path between two non-adjacent devices.

Service Layer

- Measures and represents the status of the services as seen by the customer.
- Metrics such as throughput, round-trip delay, jitter need to be monitored in an effect to meet the Service Level Agreements (SLAs) contracted between the provider and the customer.



Services	ITU-T Y.1731	MEF Specification		
Connectivity	IEEE 802.1ag	ITU-T Y.1731	MEF Specification	
Transport Links	Ethernet link OAM	PW/MPLS OAM	EoSONET OAM	Other OAM

ETHERNET OAM FAILURE ACTION

Both 802.3ah and 802.1ag can mark the link down upon failure

MPLS FRR can be triggered by link down indication

- Ethernet Ring Protection will also link down indication

LINK BUNDLING

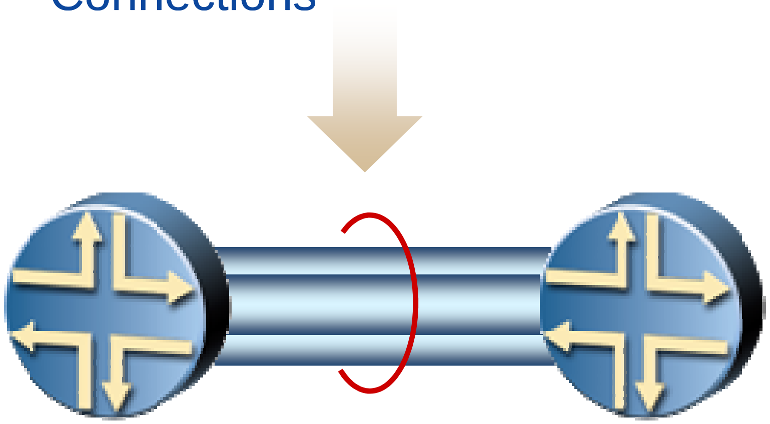
Reliable Links

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

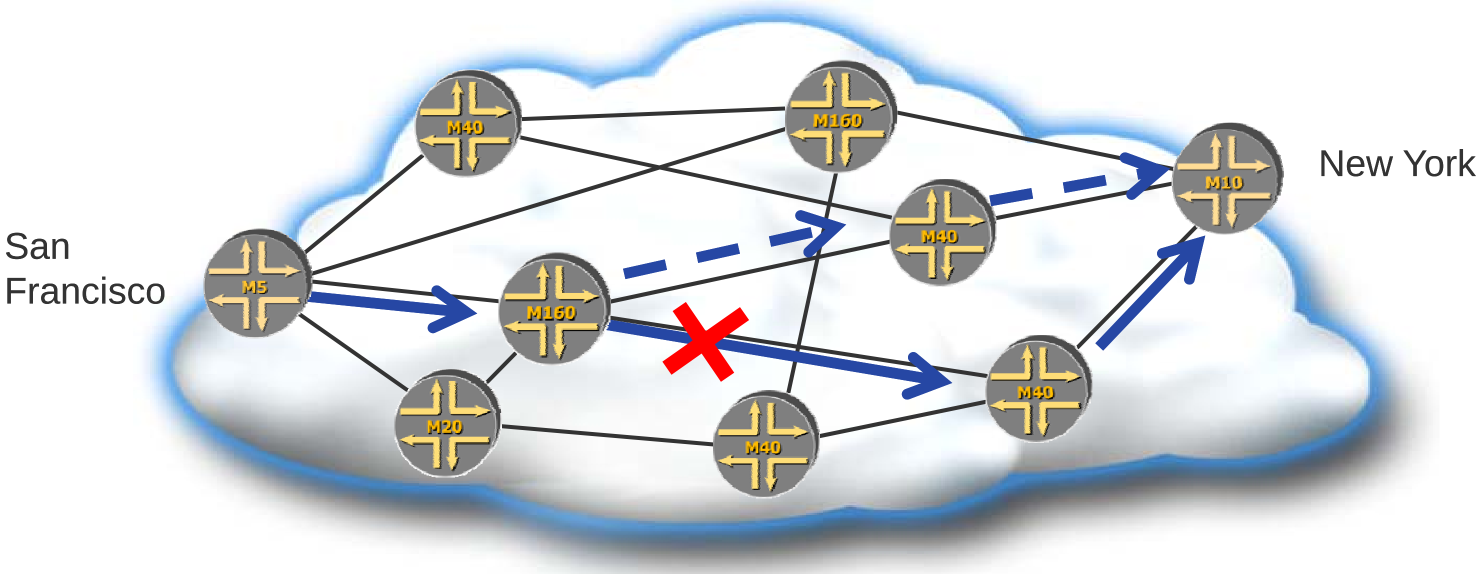
Parallel Link Technologies

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

Simultaneous Physical Connections



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
  - Separation of control & forwarding planes is key
  - Protocol expertise is key

Features	Benefits
High Priority Flooding for Interested LSPs (ISIS)	▪ Timer reduced from 100 to 20msec ▪ Faster propagation of major changes
Quick SPF Scheduling (ISIS)	▪ 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

WHAT IS BFD (BIDIRECTIONAL FORWARDING DETECTION) ?

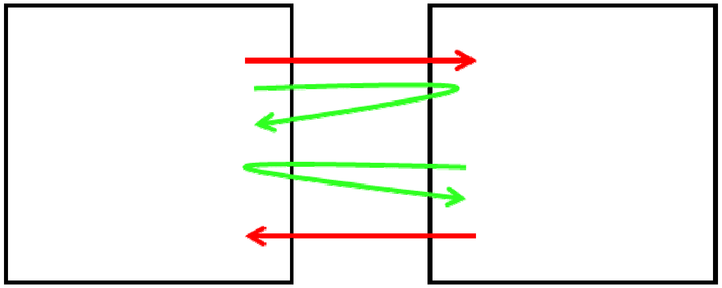
Yet Another Hello Protocol (but lightweight – compared to OSPF/IS-IS….)

Packets sent at regular intervals; neighbor failure detected when packets stop showing up

Always unicast, even on shared media

Not just for direct links; can be used over MPLS LSPs, multi-hop separated neighbors, unidirectional links.

Simple packet format, very low processing overhead. Can be implemented in forwarding plane to the extent possible.



BFD ENHANCEMENTS

- Static route
- OSPF/ISIS
- eBGP
- Multihop iBGP

- BFD for MPLS OAM
  - BFD provides LSP data plane verification. LSP-Ping verifies consistency between LSP control and data plane. BFD benefits:
    - Lightweight. Scales to large number of LSPs
    - Sub-second failure detection
    - Periodic fault detection
  - BFD over different types of LSPs:
    - Point-to-point LSP (RSVP or LDP signaled)
    - ECMP PE-PE awareness
    - P2MP LSP
    - L2 Pseudowire
  - Can use VCCV-BFD and VCCV-Ping for Pseudowires

BFD VS ETHERNET OAM 802.1AG CC COMPARISON

BFD	802.1ag CC
Layer 3 continuity check. Better for Layer 3 and MPLS	Ethernet Layer 2 continuity check. Better for Ethernet
Use ping and traceroute for loopback and trace functions	802.1ag also supports loopback and linktrace
Comparatively simpler configuration	MEPs, MIPs, MD give more flexibility but make configuration complex. Working to simplify it.
Runs on aggregate link but not on child links	Runs on aggregate and child links. Can adjust aggregate bandwidth.
Session down causes IGP, BGP to reroute	Interface down causes protocols to reroute
MPLS OAM: data plane failure detection for RSVP and LDP LSPs, LSP switch action, can trigger FRR	MPLS OAM: interface down can trigger FRR

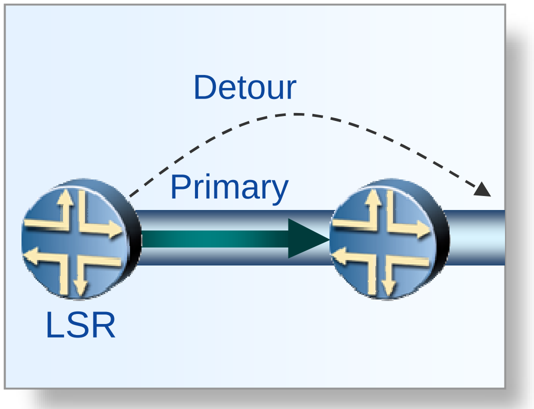
MPLS (RSVP BASED) PROTECTION

Secondary LSP

Secondary Standby LSP

FAST REROUTE

- One-to-One Backup (aka 1:1 detour) –using label swapping
- Facility Backup – using label stacking
  - Link Protection
    - Protects only against link failure
  - Node Protection
    - Protects against both link failure and node forwarding plane failure



IP/LDP FAST REROUTE

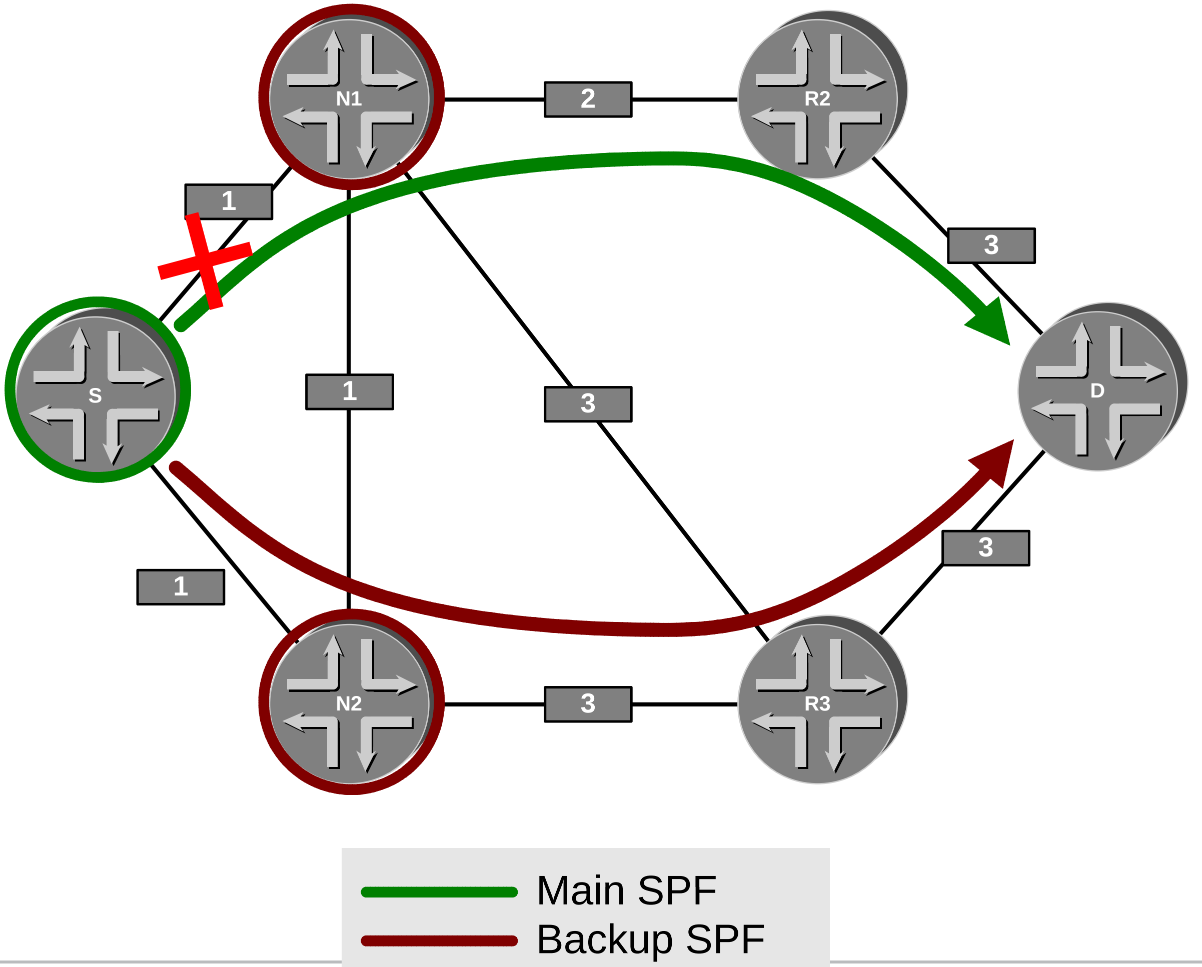
WHY?

- Some networks do not implement RSVP
- RSVP Requires PE to PE full mesh for transport plane protection
- RSVP N^2 Problem
  - 200 PEs = app. 39800 LSPs + detours & bypasses
  - Fixable using LSP hierarchy

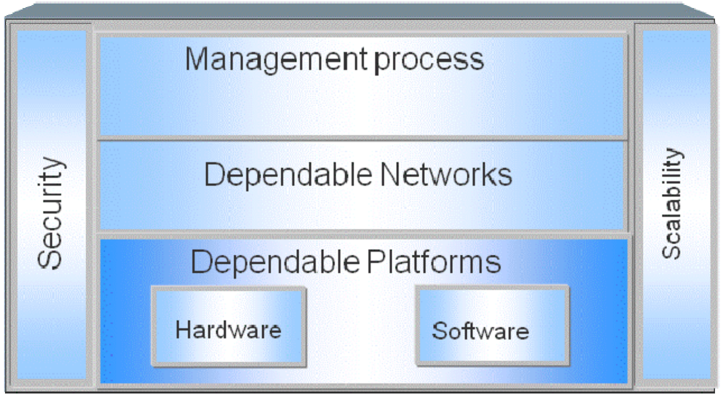
LOOP-FREE ALTERNATES (LFA)

- Adds fast-reroute (FRR) capability to IS-IS, OSPF and LDP
  - Normally only best nodal path is used for RIB walks
- Add a non-best (albeit loop free) path for backup purposes.
- How ?
  - Shared, common link state database
  - Place the SPF root at your neighbors

SPF ROOTS & LFA ILLUSTRATED



MANAGEMENT PROCESS
CONTINUOUS SYSTEMS AVAILABILITY



Security

Management process

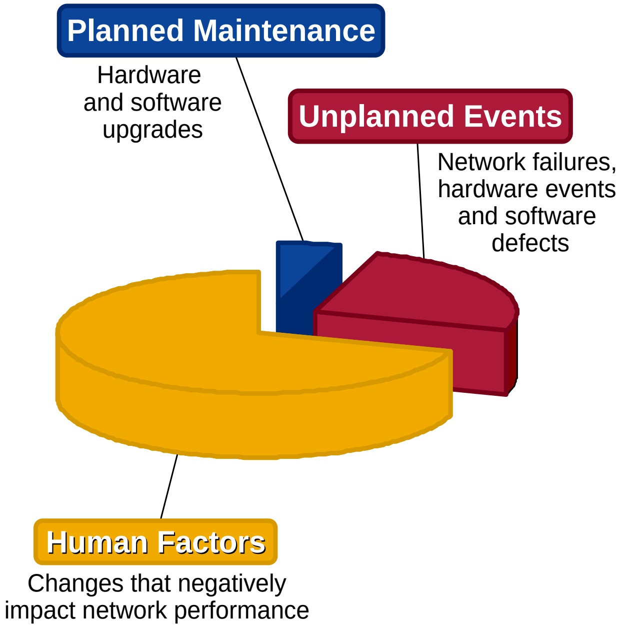
Dependable Networks

Dependable Platforms

Hardware

Software

Scalability



Planned Maintenance

Hardware and software upgrades

Unplanned Events

Network failures, hardware events and software defects

Human Factors

Changes that negatively impact network performance

Minimize impact of human factors with fail-safe mechanisms

- Prevent configuration errors and ensure compliance to configuration policies with candidate configurations, commit verifications, commit scripts

Speed response and resolution to unplanned events

- Avert downtime with transparent failover, network recovery features
- Provide proactive response and accelerate resolution with extensive instrumentation, event policies, op scripts

Reduce time of planned events

- Stable releases reduce the frequency of fixes and duration of upgrades
- In-service upgrades available in high-end routing platforms

SCRIPT AUTOMATION

- **Helps to Reduce Human Error**
- Commit Scripts - Parse Configurations upon commit
  - Generate Warnings
  - Reject the commit
  - Modify the configuration
- Op Scripts - Used to ensure compliance with company policy
  - Generate notifications
  - Automatic Diagnosis
  - Correction of Problems
- Event Policies - Monitors an event trigger
  - Works with Op Scripts
  - Wait for Notification messages

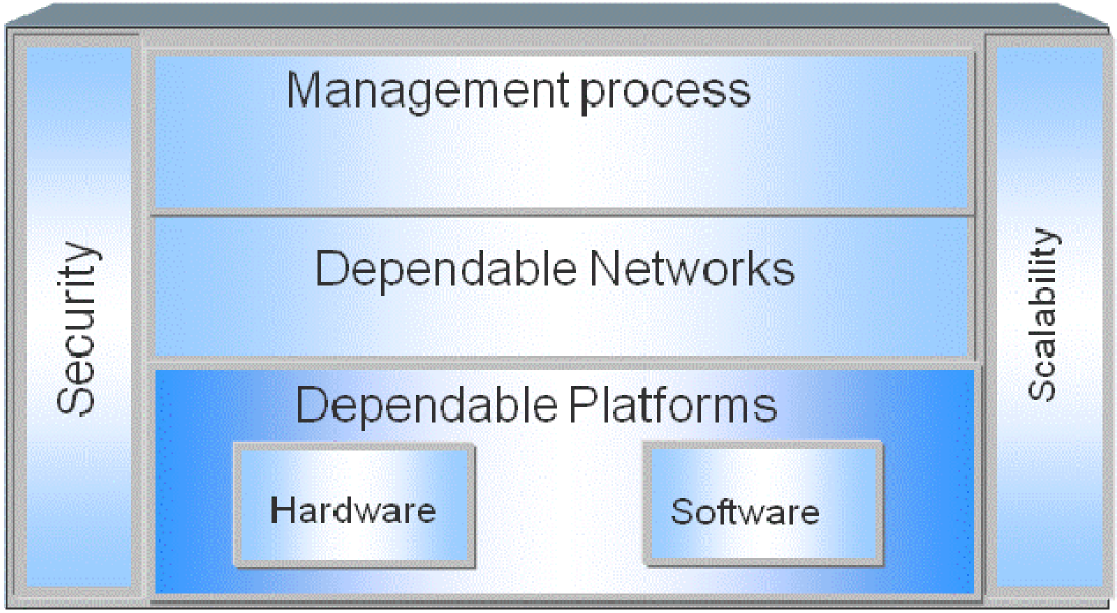
SCALABILITY/SECURITY

Scalability

- Hardware : performance/next-hops/firewall filter…
- Control plane : routes/peers/routing instances/logging….

Security

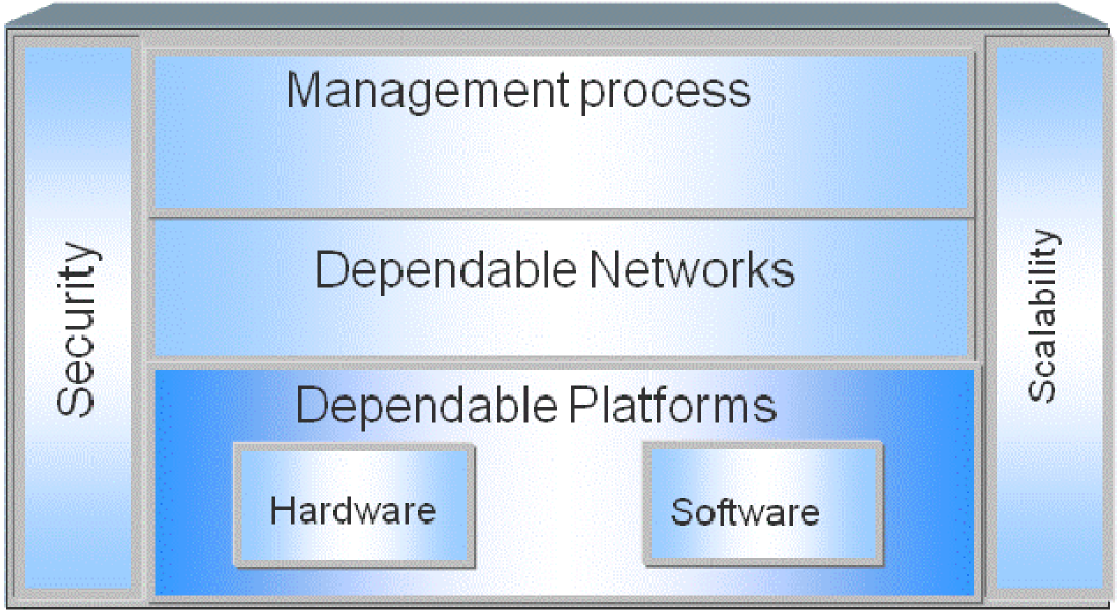
- DDOS Attacks



SUMMARY

High Availability:

- ❖ Is a culture
- ❖ Has many layers
- ❖ Is business critical



everywhere