



IEEE HPSR 2002

**Building a Reliable and Scalable
Internet**

Applications, Equipment and Technology

Chris Gunner

Senior Vice President of Research and Development

May 27, 2002

Reliable Routing for the Internet



Overview

- Why is Scalability important ?
- Why is Reliability important ?
- What are the sources of Internet Downtime ?
- How can Reliability be improved ?

BTextact and the New Carrier Benchmark

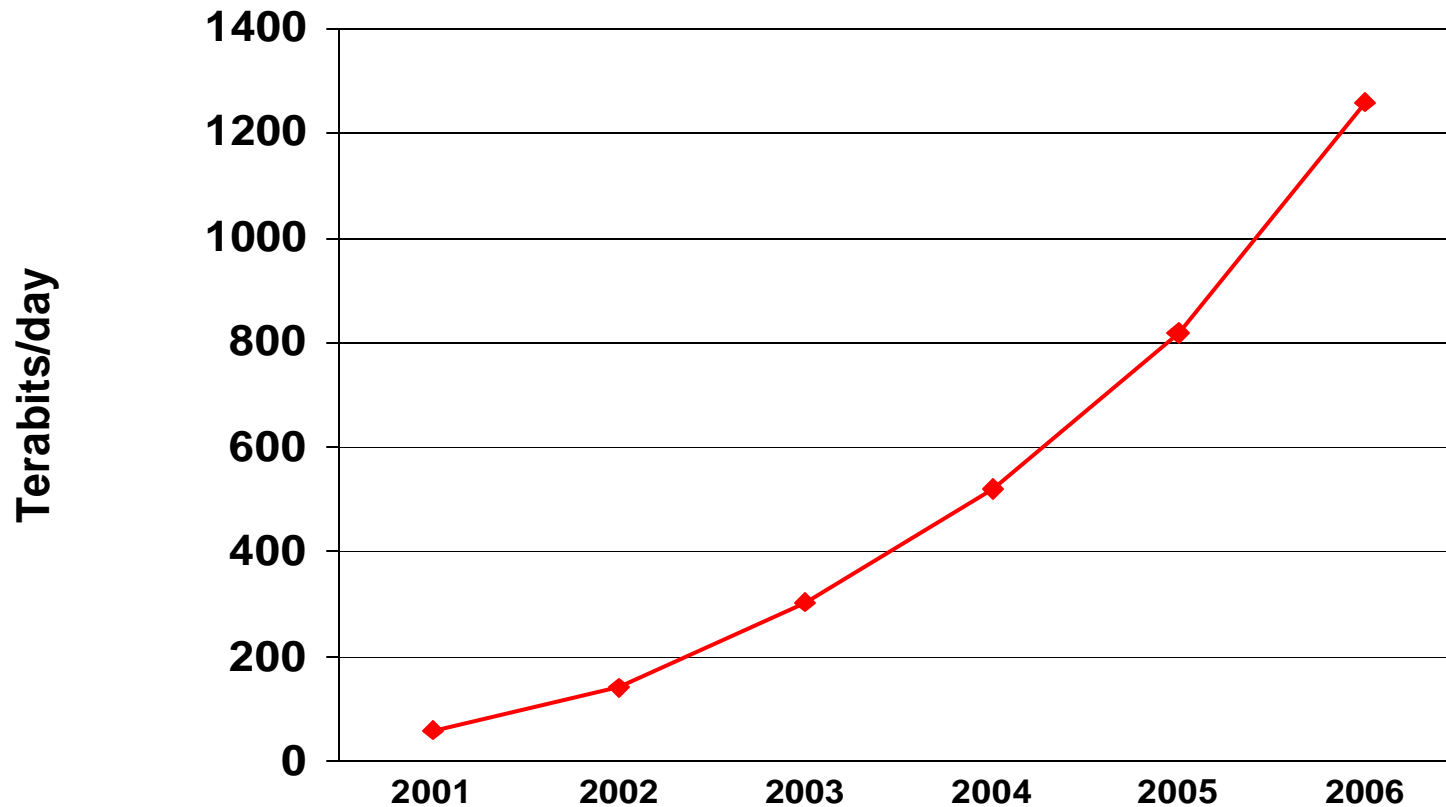
- Carrier Requirements of core IP Routers 2002¹ technology benchmark released publicly Feb. 27, 2002

Prioritised Issues For Carriers
1. Equipment Reliability and Stability
2. Scalability
3. Performance
4. Feature Support and Interoperability
5. Management
6. Total cost of Ownership
7. Environmental Considerations
8. Security

- BTextact Testing focuses on validation of Carrier Requirements in operational carrier scenarios.
 - Major Paradigm Shift in Testing

[1] http://www.btexact.com/white_papers/downloads/WP113.pdf

Continuing growth in IP traffic



Total IP Traffic

2001

2002

2003

2004

2005

2006

59.6

140.1

303.7

521.7

818

1257.8

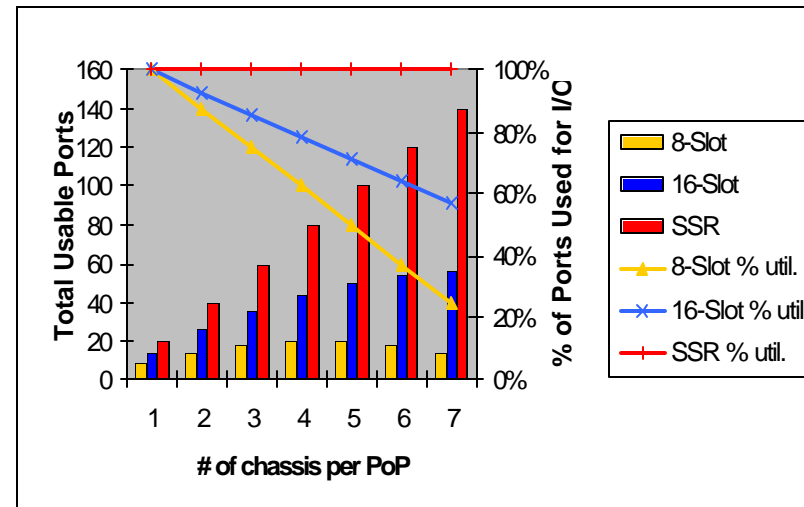
Source: Ovum

IP traffic growth 88-121% (CAGR 00'-05' source; RHK)

Scalability

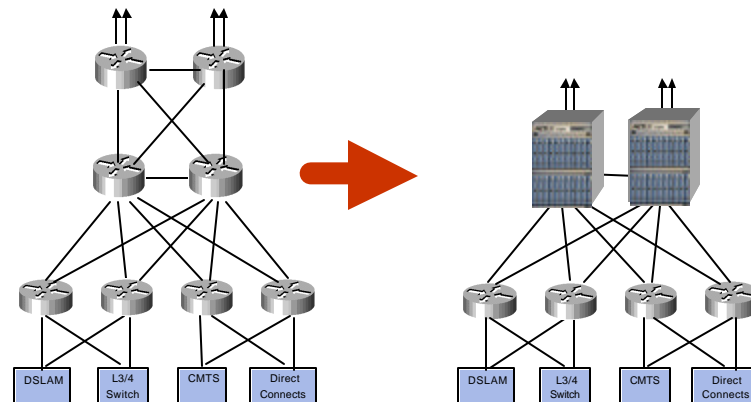
Why does Scalability matter?

- *Enables non-disruptive, cost-effective growth*

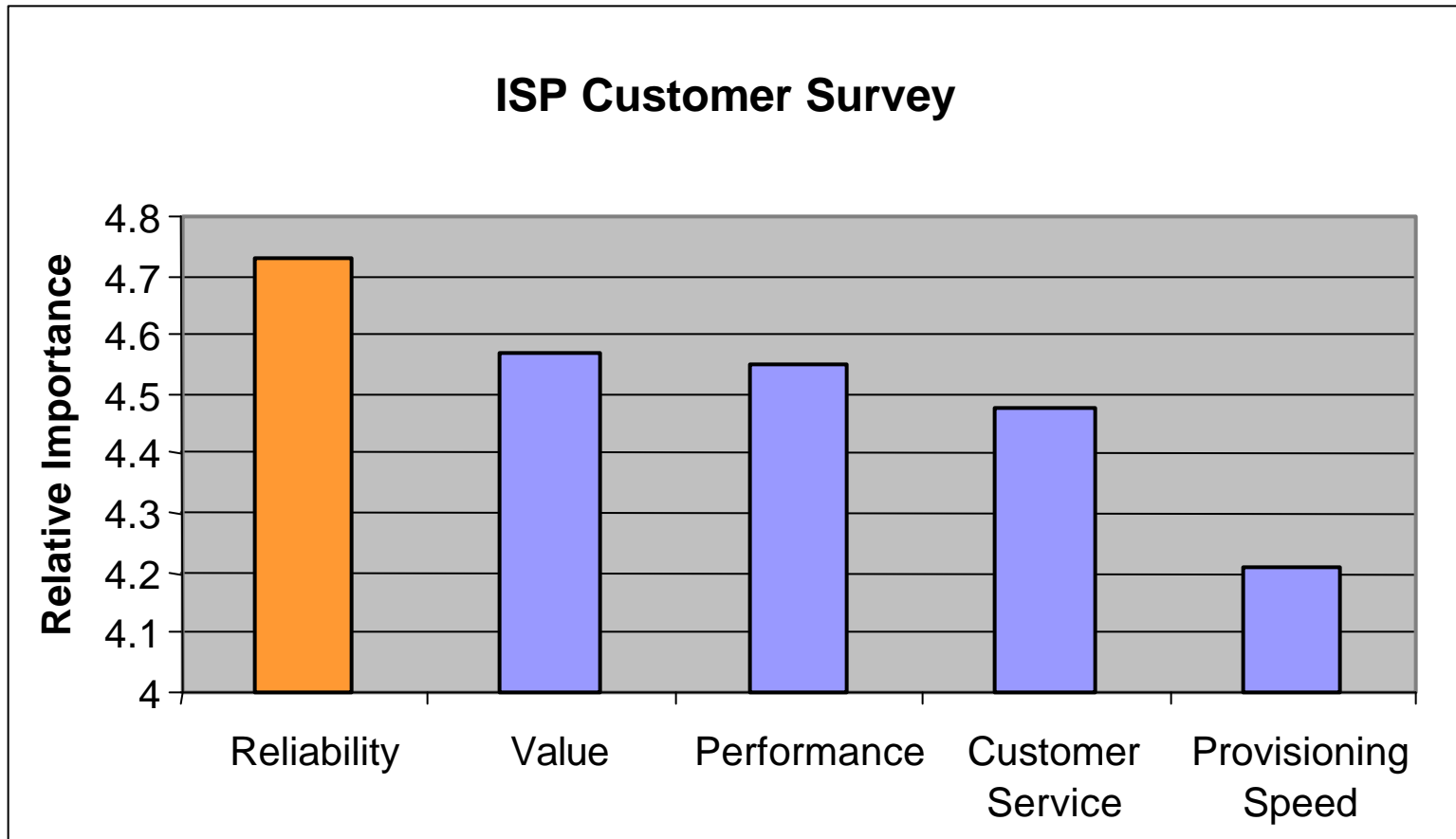


- *Enables “vertical” POP simplification*

Reduces CapEx
Increases Network Stability



Why is Reliability Important?

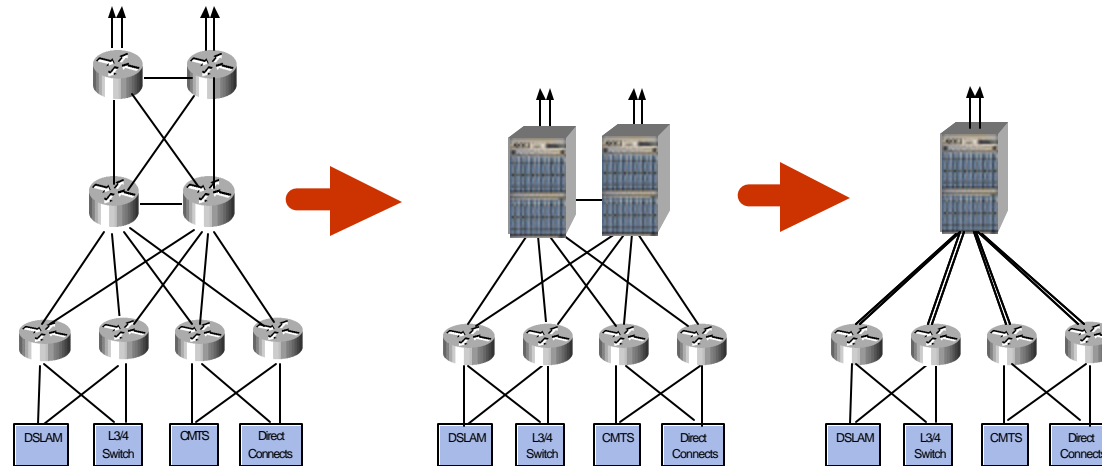


Because users care !

Interactive Week / Telechoice ISP Survey
Sept. 10, 2001

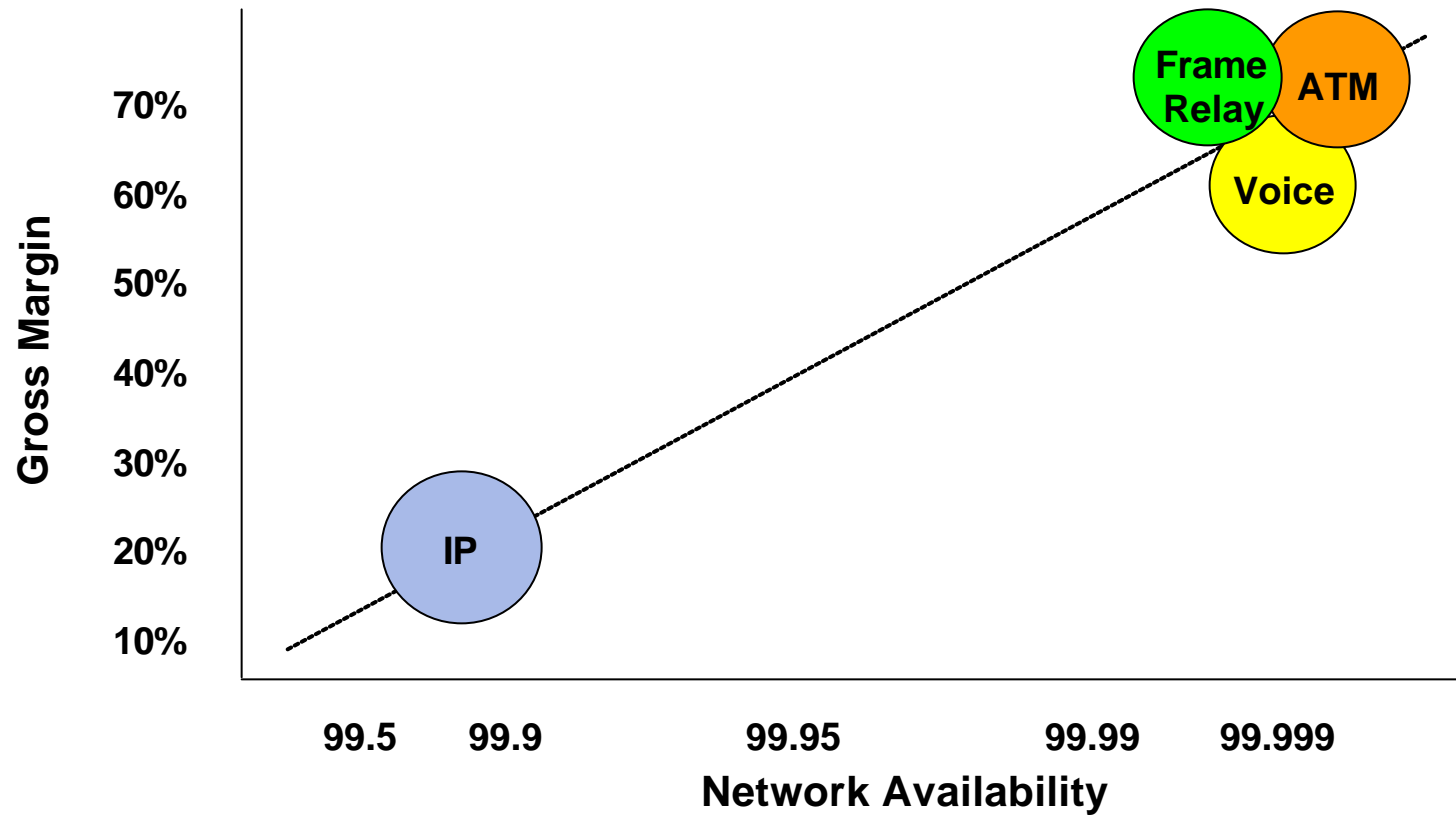
Reliability

- High Reliability allows “horizontal” POP simplification



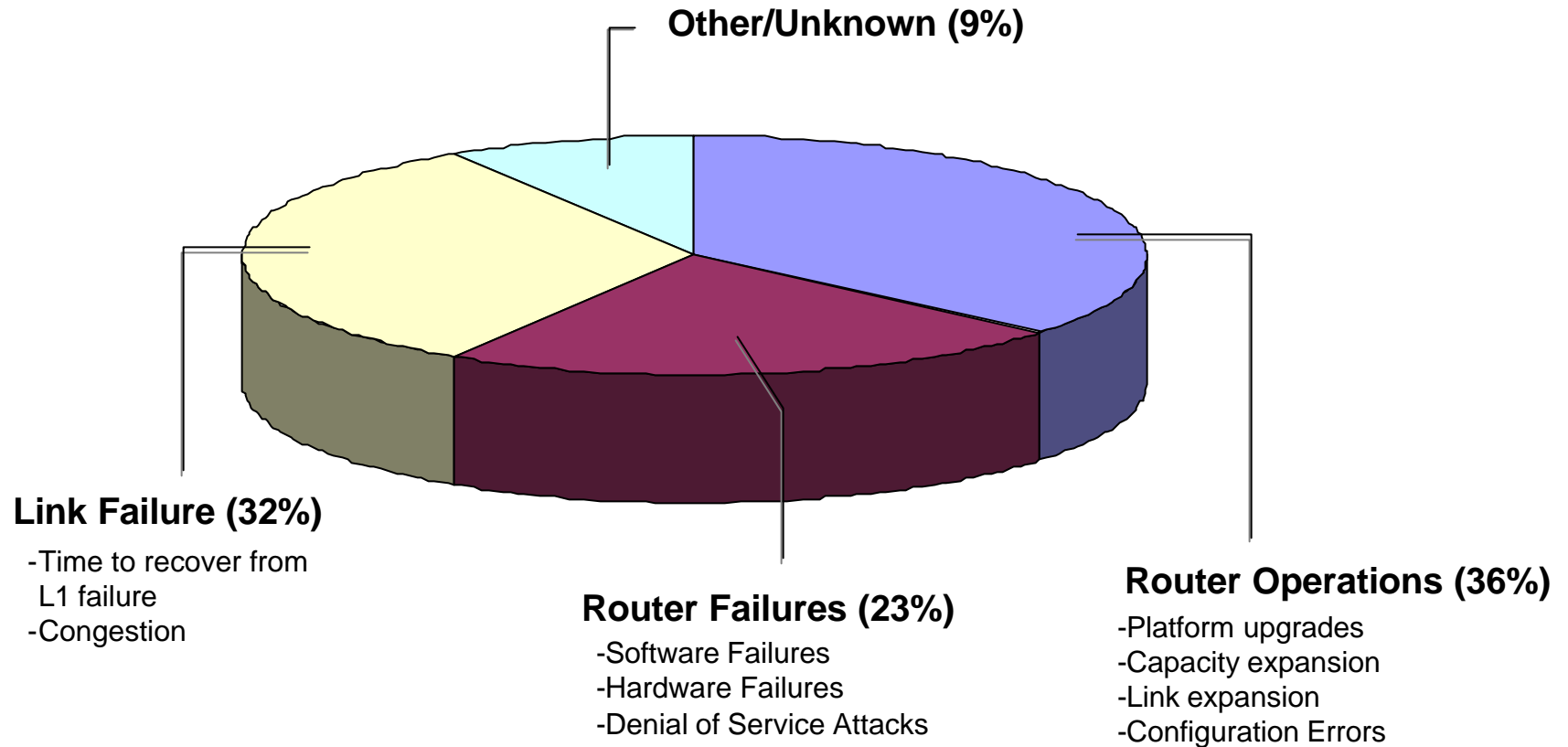
***Further Reduces CapEx, Operational cost
Further increases network stability***

Today's Network Reliability Gap



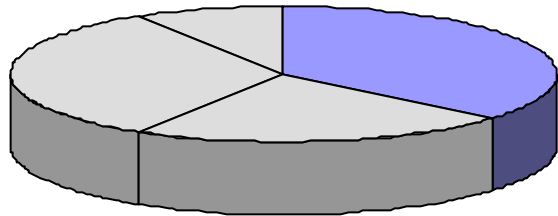
Causes of Internet Downtime

Causes of Downtime in Router-Based Wide-Area Networks



Source: University of Michigan, Avici Systems

Router Operations: (36% of downtime)

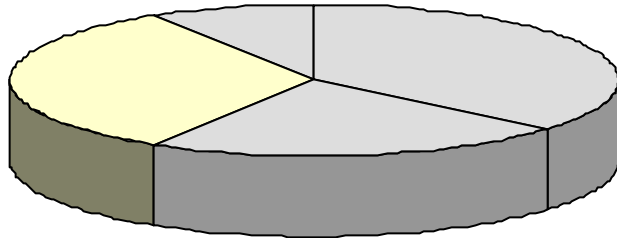


Primary Operations activities resulting in Network Downtime:

- Router platform upgrades
- PoP expansion
- Link expansion
- Configuration changes

Core routers must provide an inherently stable operational platform which significantly reduces outages and downtime

Link Failure & Congestion: (32% of downtime)



Root causes of Downtime from Link Failure and Congestion:

- Fiber Cuts
- Linecard failure
- Slow restoration time
- Congestion resulting from rerouting
- Inefficient Traffic Engineering mechanisms

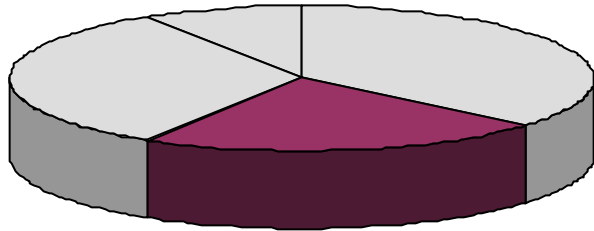
No platform can eliminate link failures, but the best can provide carriers with cost-effective tools to rapidly recover

Link Failure Recovery Mechanisms

Recovery Mechanism	Switchover Time
SONET APS	<50 ms
Link Bundling (e.g. Composite Links)	<45 ms
MPLS Fast ReRoute	<45 ms
MPLS head-end resignalling	< 7 s
IP shortest-path rerouting	< 7 s

Real-time services require SONET-like <50ms switchover times to avoid disruption

Router Failure: (21% of downtime)



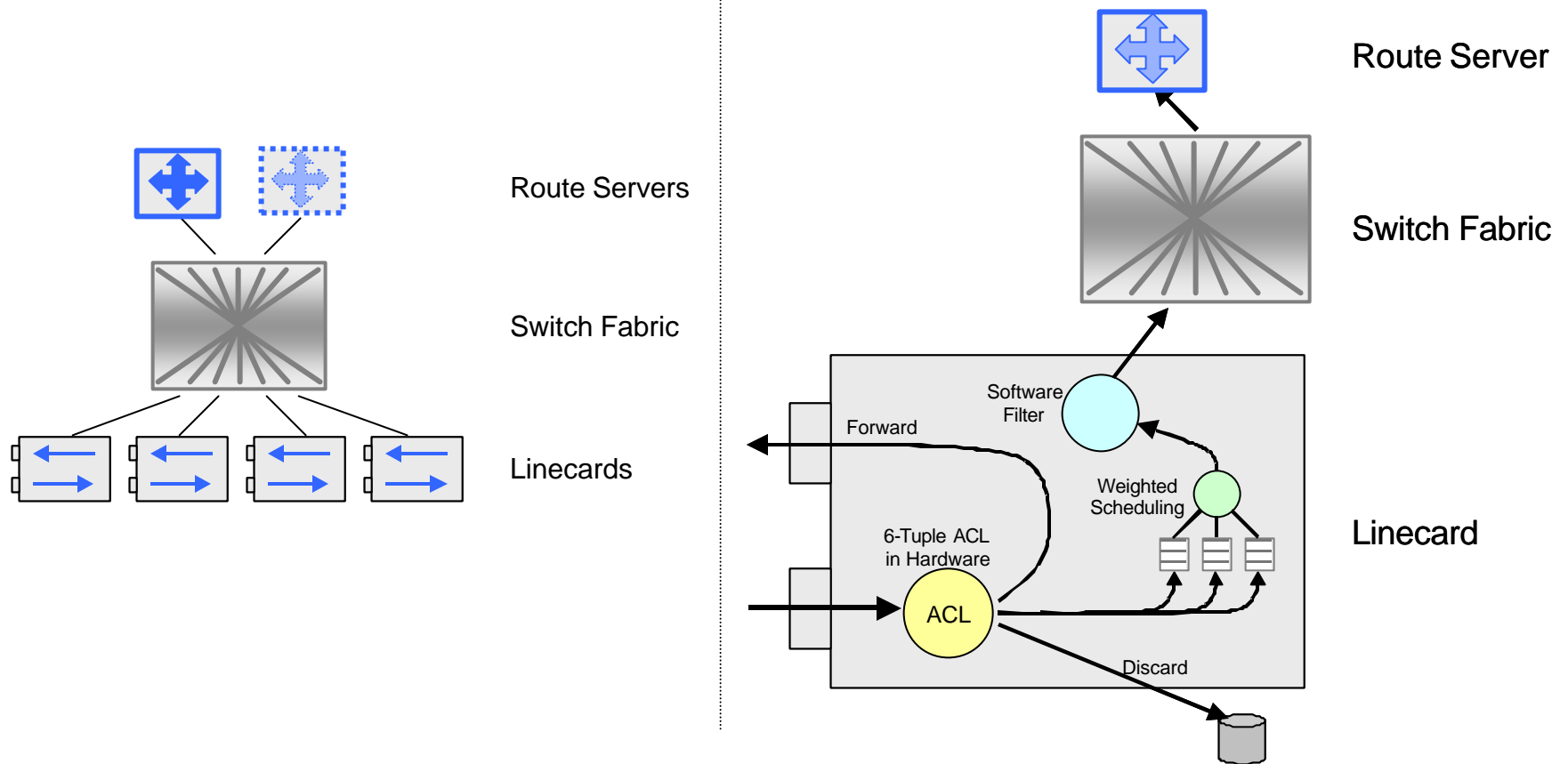
- Hardware Failure
- Denial of Service Attacks
- Software Failure

Routers must become more robust to achieve 99.999% availability

Hardware failure protection

- Dual or better components with failover strategy
 - Power
 - Cooling
 - Processing components (route servers)
 - Linecards with alternate paths (layer 1, 2 or 3)
- In-service field replacement of components

Distributed Architecture – DoS attacks



Distribution of processing functions provides multi-level protection against DOS attacks and single component failure

Router Control Plane Redundancy Approaches

Approach	Description	Service Interrupt Time
No redundancy	restart router on failure	3-10min/hours if hw failure
Hardware-only redundancy	secondary route server in "warm standby"	3-10 min
Fault-Tolerant	backup server maintains lock-step alignment with primary	3-10 min on sw failure
Headless Forwarding	protocol extensions for graceful restart	0
Avici Non-Stop Routing	backup server loosely coupled to primary to minimize fate-sharing	0

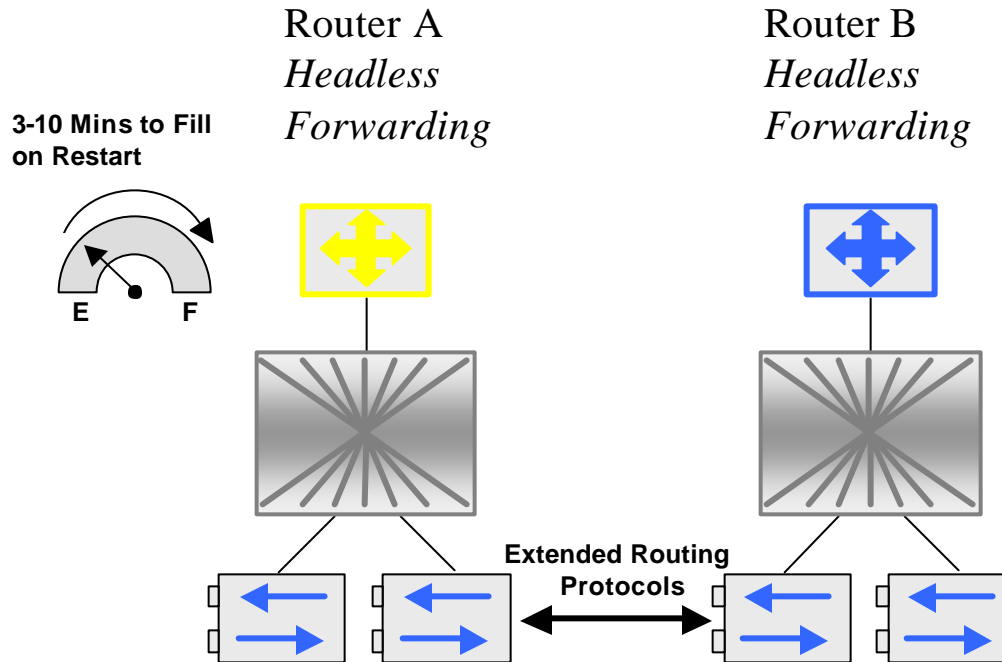
Assumptions:

- Distributed architecture
- Multiple route servers
- Multiple linecards capable of forwarding even when control server is unavailable

Headless Forwarding

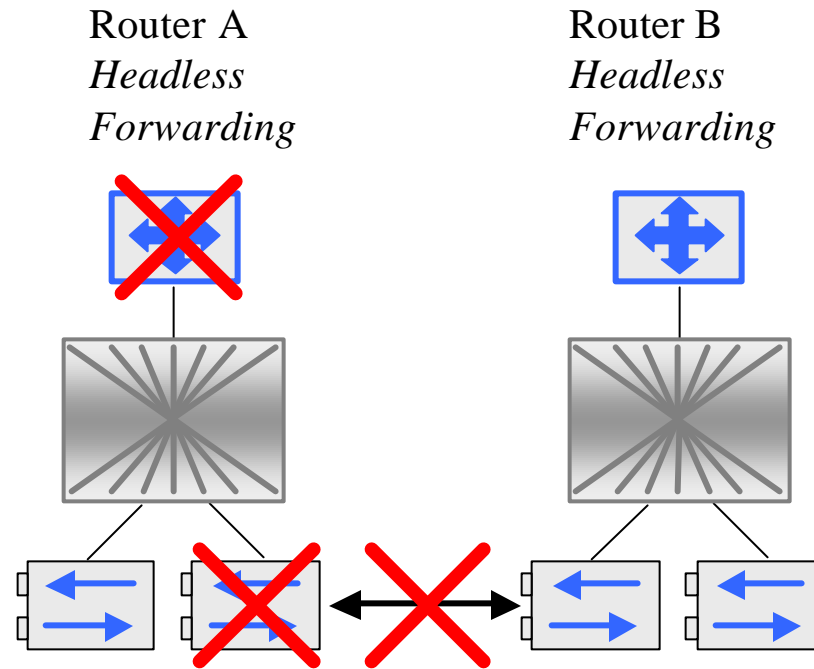
- Relies on Routing Protocol Extensions
 - Several IETF draft proposals
 - BGP draft
 - ISIS draft
 - Two competing OSPF drafts
- Common theme - “I will be back” – Optimistic model
 - Let the peer router know your ability to restart
 - Peer router to maintain and use the forwarding state learned from the failing router while it comes back up
 - Meant to address fail-over with one server
- Drawbacks
 - All routers must implement extensions
 - Vulnerable to extended traffic loss when there are multiple failures

Headless Forwarding Scenarios



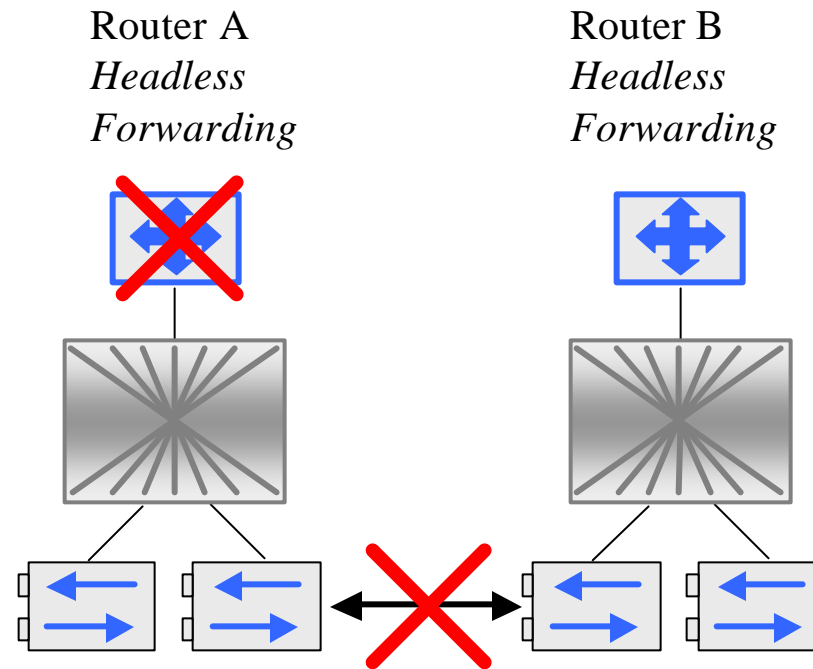
Scenario	Impact
Route server restarts and relearns routing state	Stale forwarding until routes are relearned 3-10mins

Headless Forwarding Scenarios



Scenario	Impact
Linecard fails during route server failure	Traffic loss for minutes

Headless Forwarding Scenarios

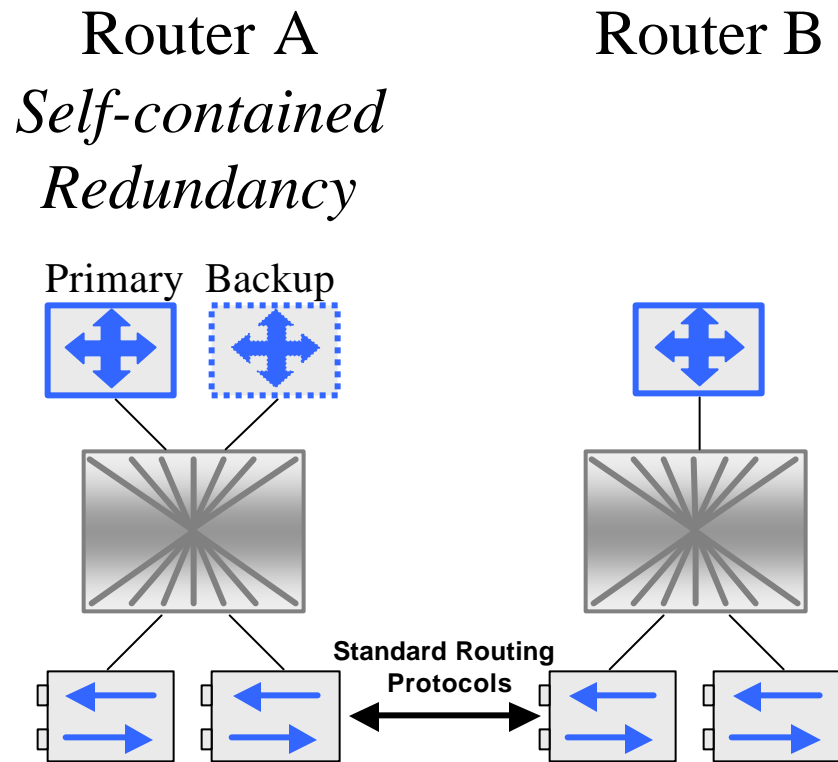


Scenario	Impact
Route server fails to restart	Stale forwarding for minutes

Self-Contained Redundancy

- Self-contained solution
 - No dependence on peer router implementation or protocol extensions
 - Save all pertinent routing state on Backup server at all times
- Advantages
 - Maintain routing protocol connectivity and liveness to peer routers during failover transition
 - Shorter fail-over time
 - Fully resilient to all failure modes – e.g. failover failure

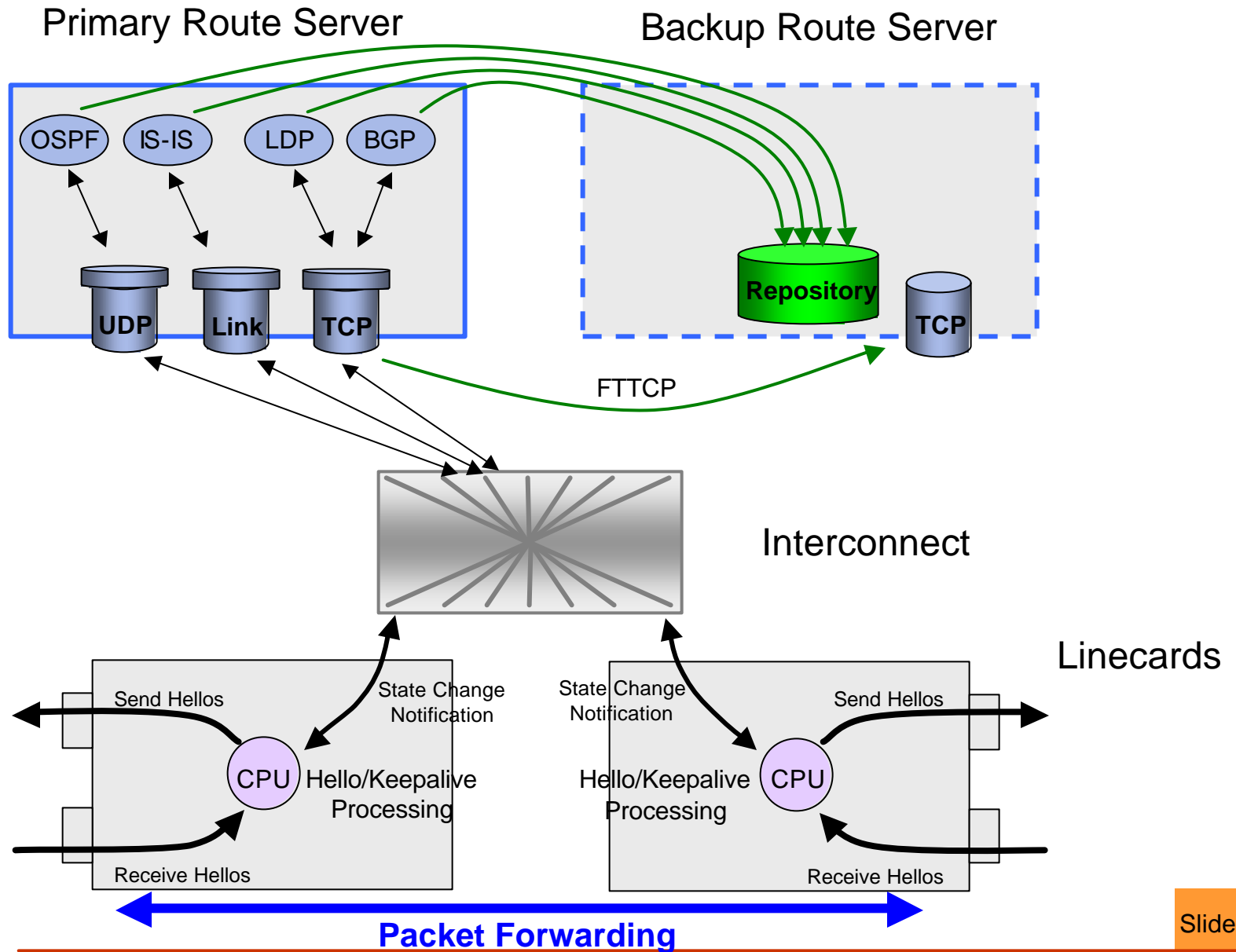
Self Contained Model



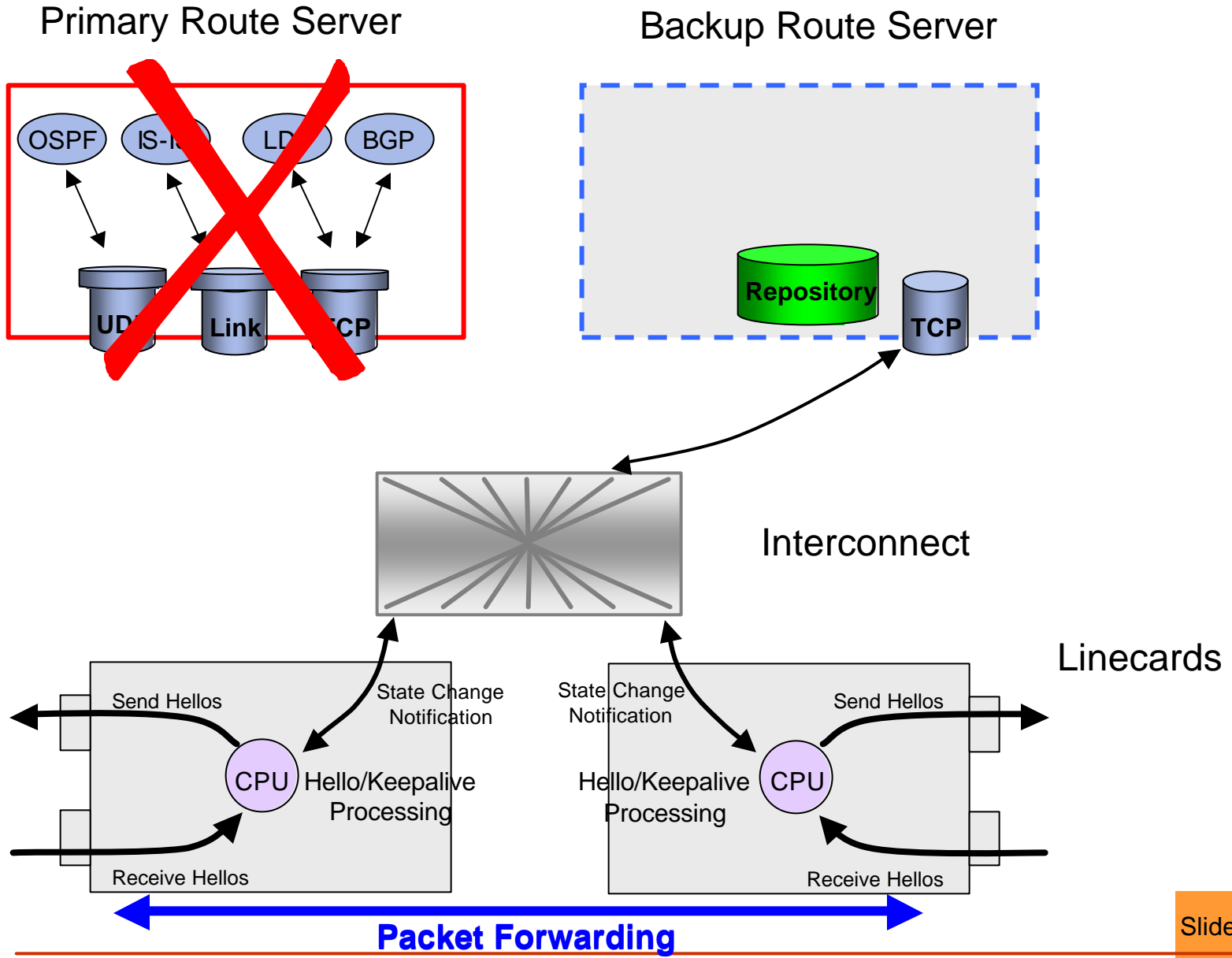
Self-Contained – State Replication

- Routing state (operational state)
 - Magnitude ~300MBytes BGP, ~1MByte IGP
 - TCP transparency (end state atomic replication)
- Self monitoring – automatic internal consistency checking
- No single point of failure for critical state
 - E.g. router config file replicated

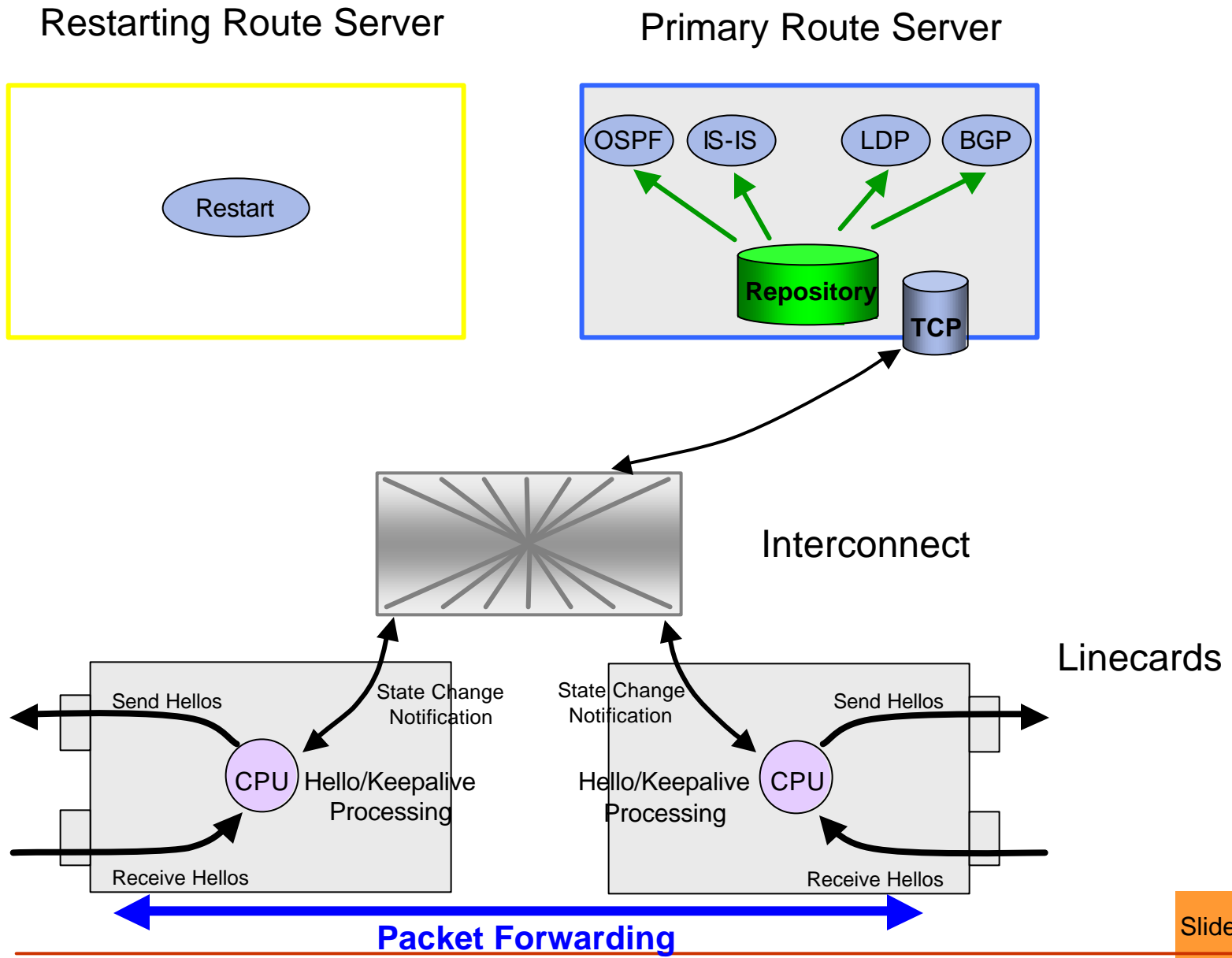
Steady State Protection



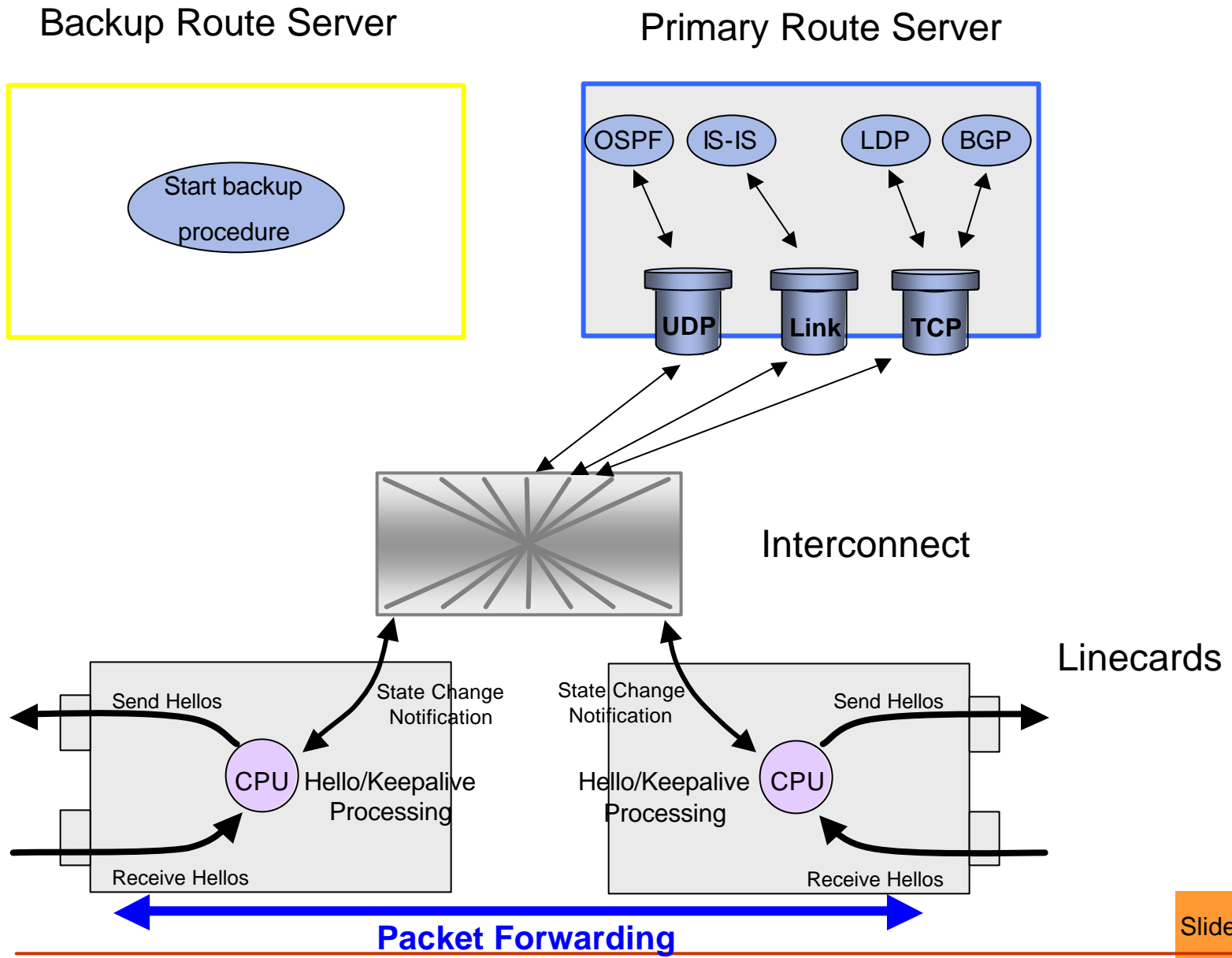
Failure Event



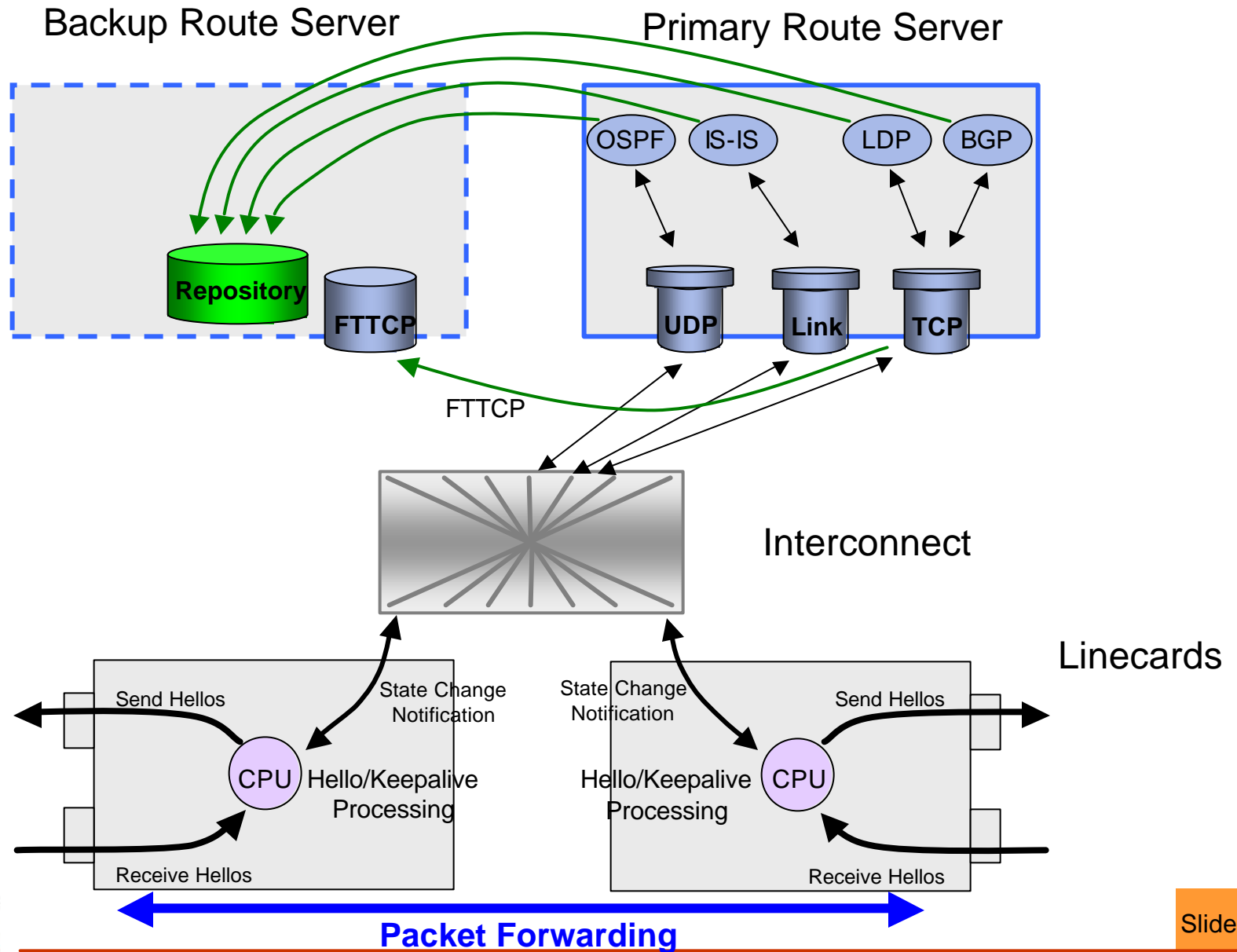
Transition State



Transition Complete



Steady State Protection



Self-contained - Management

- Support different software images on servers
- Provide manual switch to perform hitless in-service software image upgrade/downgrade
- Configuration management
 - Support different config on backup compared to primary (eliminate new config failure propagation)
 - User-controlled config replication commit mechanism
- Management – manual consistency checking, full control of multi-server configuration

Summary

- Fully Redundant Hardware Available Now
- IP Service growth and competitive SLAs require reliability improvement
- Achieving IP router reliability of 99.999% requires solutions to many problems
 - avoidance of PoP churn by providing scalability
 - avoidance of downtime due to software upgrades
 - avoidance of service loss due to failure of route server
- These solutions are available now from Avici Systems

A photograph of a long, straight asphalt road with two yellow lines down the center, stretching towards a clear blue sky and distant mountains.

Thank You

A large graphic consisting of a blue vertical bar on the left and a large orange rectangle on the right.

Reliable Routing for the Internet