RFC2547 Convergence Characterization and Optimization

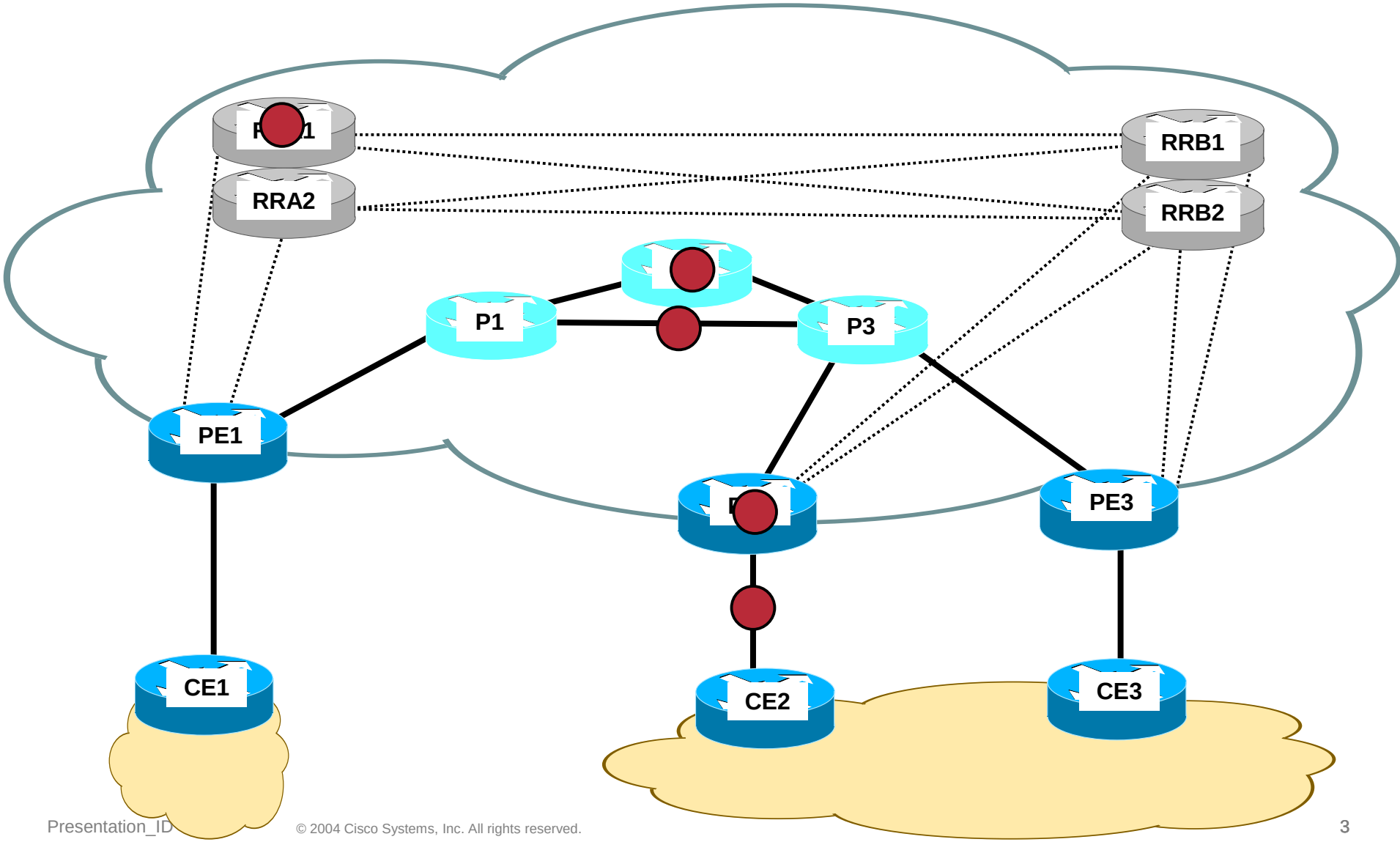
Clarence Filsfils

cf@cisco.com

RFC2547 Convergence - Requirement

- 90%: Typical requirement: <10s
- 9%: More aggressive requirement: <3 to 5s
  - VPN is used to transport Voice
- 1%: Very Aggressive requirement: from <1s to <50ms

What failures to consider



What mechanisms for each failure

- **Core Link/Node failure**: BGP inheritance of IGP Convergence
- **Egress PE node failure**: IGP failure discovery, IGP flooding, event-driven BGP convergence

IGP Convergence is key

- **Egress PE-CE link failure**: local link failure discovery, BGP signalling, BGP convergence
- **RR failure**: clusters are redundant and hence no impact on connectivity. Desire to speed up the BGP reload to minimize the duration when the cluster is non-redundant

Pure BGP signalling and convergence

RFC2547 Convergence does not suffer from the counting-to-infinity problem found in the Internet

- "An Experimental Study of Internet Routing Convergence", Craig Labovitz
  - "…we show that inter-domain routers in the packet switched Internet may take several minutes to reach a consistent view of the network topology after a fault…"
  - "…we show that even under constrained policies, the complexity of BGP convergence is exponential with respect to the number of autonomous systems…"
- Reason: there is only one possible AS path between two customer sites. Big difference between RFC2547 and Internet use of BGP

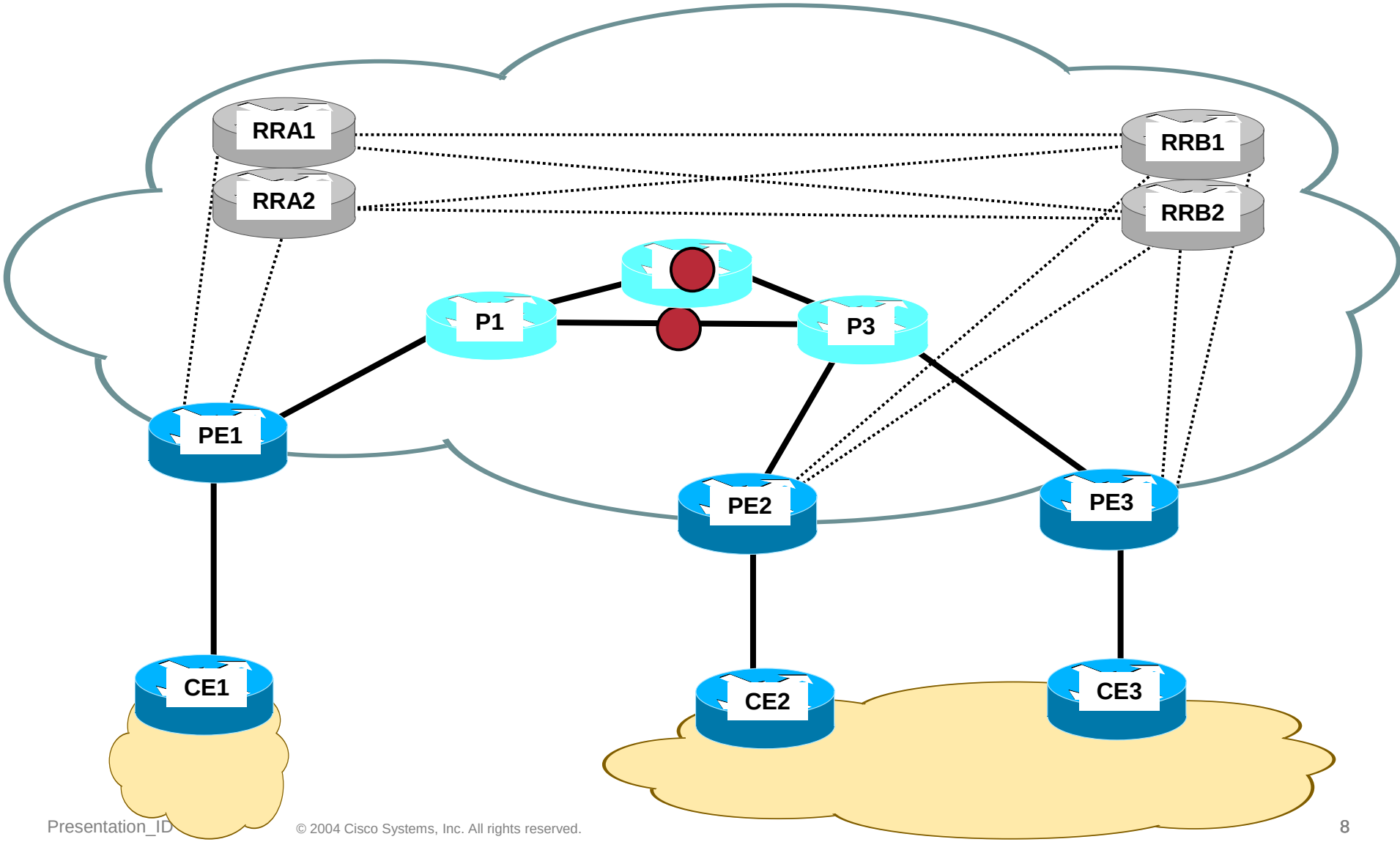
Methodology

- **Same as for the IGP Fast Convergence Project**
  - **Lead customer set requirements, design context and constraints**
  - **Black Box testing to assess behavior as seen by customer. Real traffic is used to measure the Loss of Connectivity (LoC).**
  - **White Box testing to decompose the behavior into its components and hence to allow for implementation optimization. IOS instrumentation is used.**
  - **UUT is in a realistic IGP/BGP setup (700 IGP nodes, 2500 IGP prefixes, 100k VPNv4 routes) and is stressed by 1Mpps and 6 BGP flaps per second**
  - **Black box and white box measurements perfectly match**
  - **20 iterations are used for each tested scenario**
  - **Design Guide**

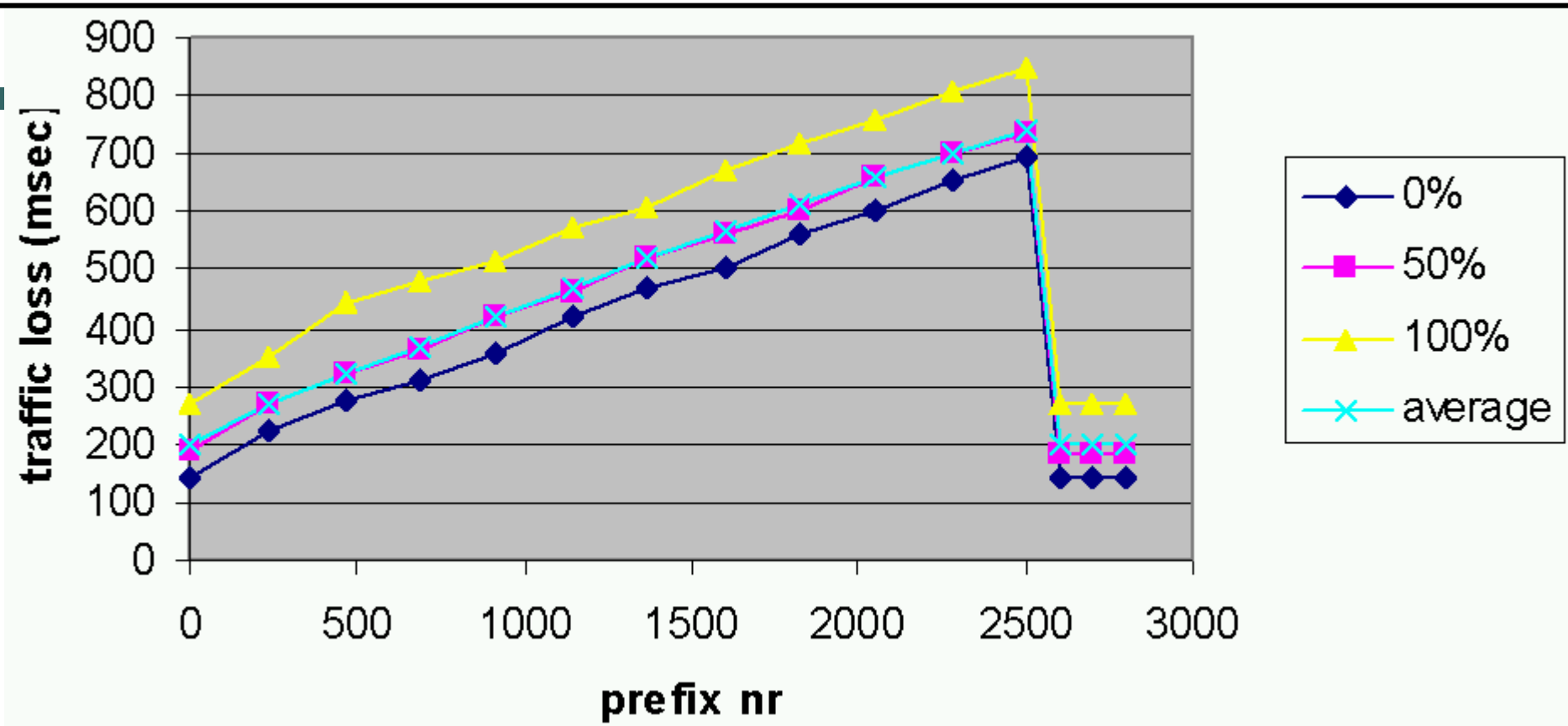
Design Context/Constraint

- Convergence to a redundant site
  - loadsharing or primary/backup config
- A unique RD per PE per VPN
  - remote PE's do learn the two paths, no RR hiding
- 80% of the CE's advertise less than 200 routes
- It is very rare for a CE to advertise more than 1000 routes
- A typical PE selects ? route via a set next-hop
  - we currently test with 2000

Core Link/Node Failure



IGP Fast Convergence sub-second is conservative

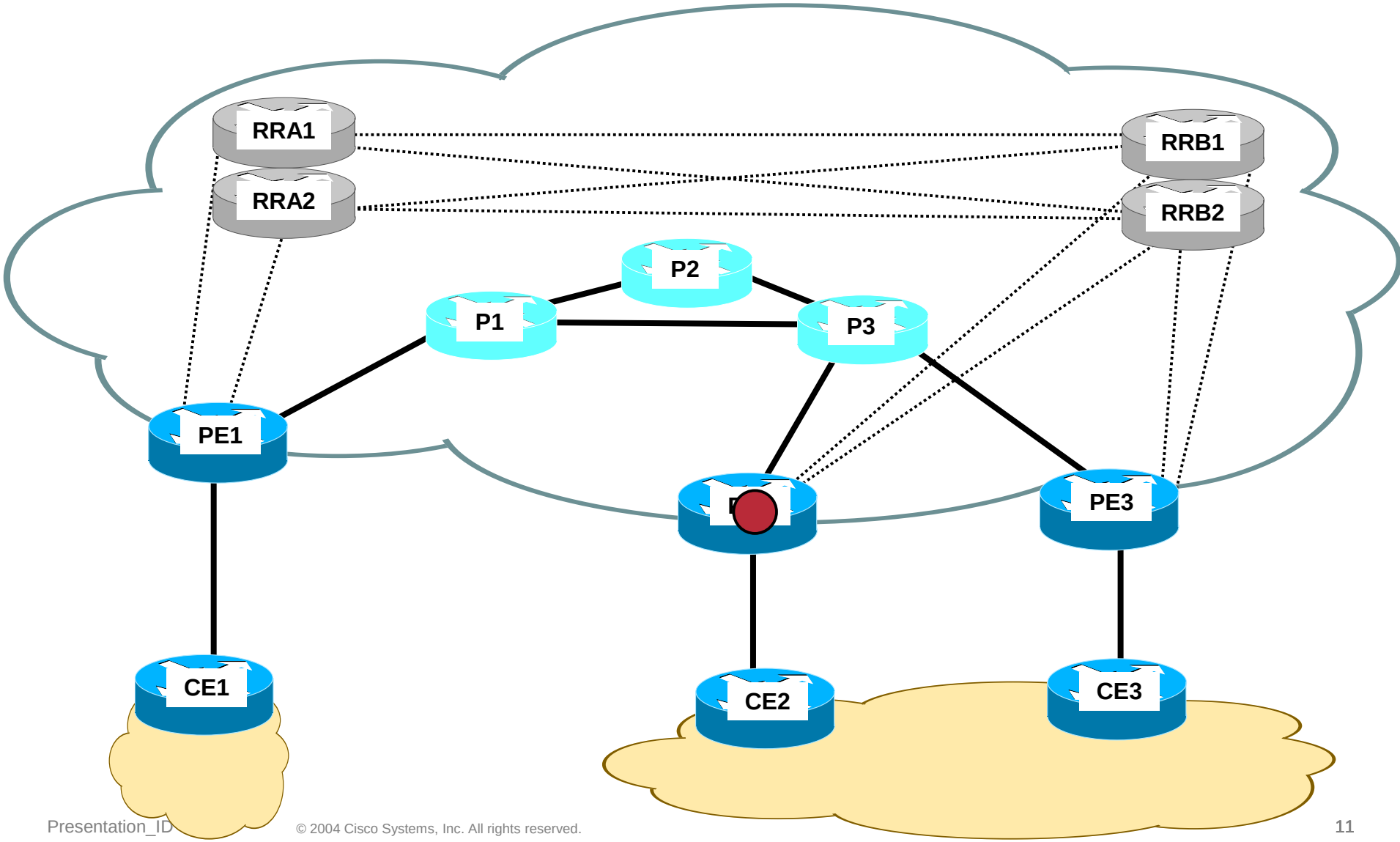


- **For more details, refer to Apricot 2004 presentation**
  - **also at Nanog 29, Ripe 47**
- **Paper under submission**

IGP Fast Convergence - Reminder

- Link Failure Detection: PoS, B2B GE, DPT; if not, BFD
- Fast Origination and Flooding
- SPF optimization: eg. Incremental SPF
- RIB/FIB Prioritization: most important prefix first
- Optimization of Download distribution and HW modify
- BGP fully leverage IGP entries

Egress PE Node failure



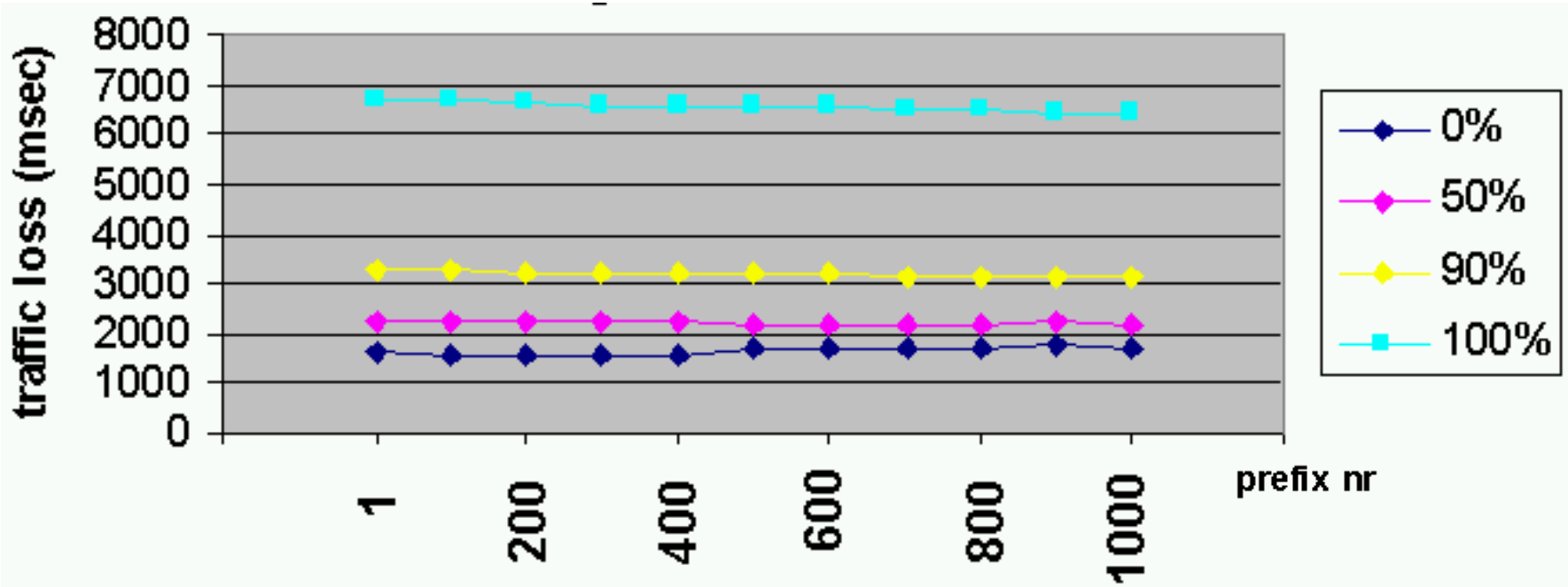
Egress PE node failure

- Adjacent core nodes detect the failure of PE2 (Link or BFD) and flood new LSP's advertising the failure
- PE1's IGP converges and declares PE2 unreachable
- PE1: Unreachable status of a BGP nhop triggers BGP Convergence (ie. use PE3 instead of PE2)
  - "BGP Next-Hop Tracking" Feature

BGP Next-Hop Tracking

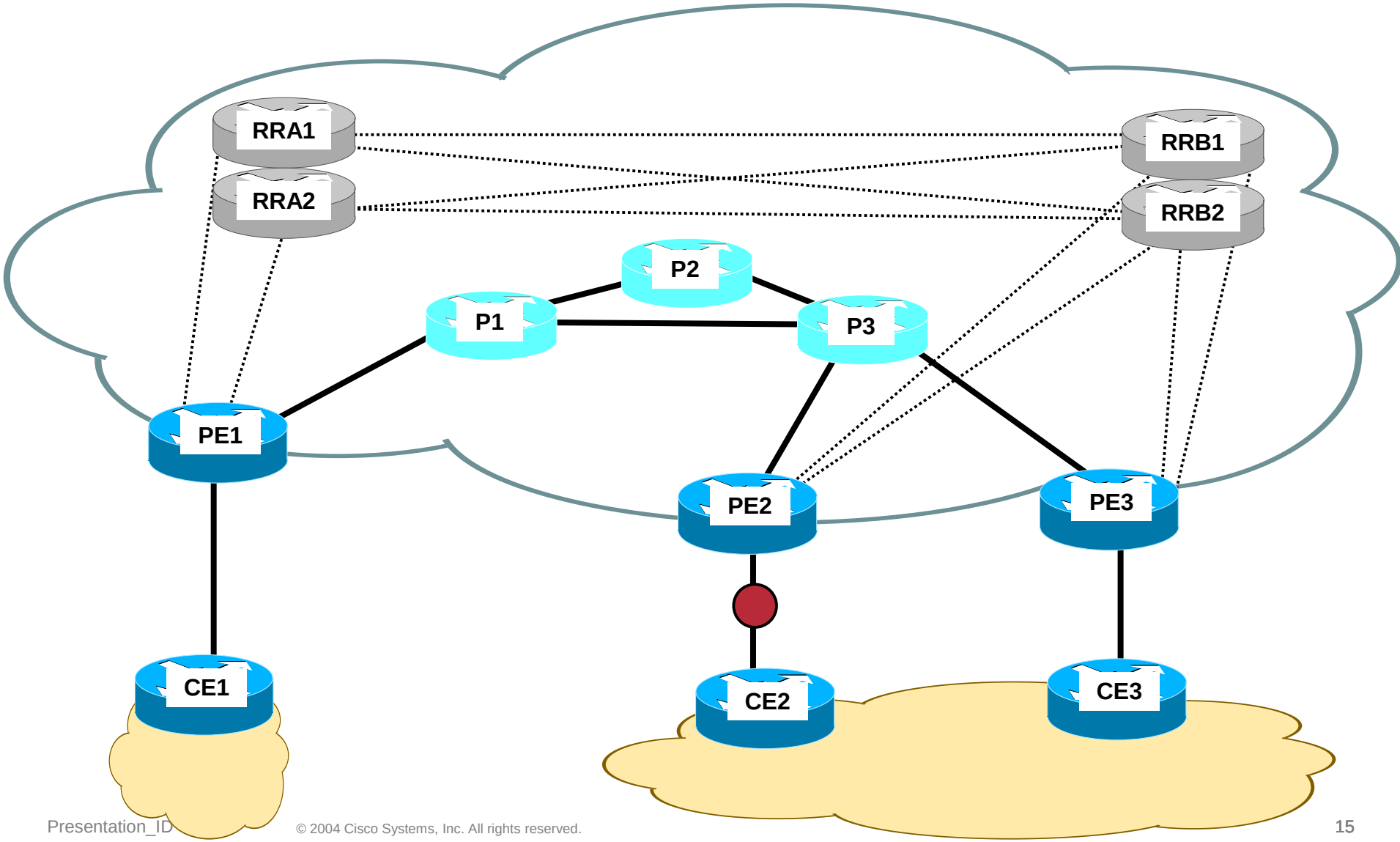
- BGP registers its next-hops with the RIB
- Later, RIB notify BGP when the reachability status of these next-hops change
- Dampening algorithm is used to control how immediate the RIB notification may trigger a BGP reaction

Blackbox Measurement
Egress PE node failure



- PE1 selects 2000 prefixes from PE2
  - 1000 in VPN1, 1 in VPN2, …, 1 in VPN1000
- Traffic is sent to 11 prefixes in VPN1
- Sub-10s for 2000 prefixes is conservative
- Sub-5s is achievable

Egress PE-CE Link failure



Egress PE-CE Link Failure

- The nhop is PE2 hence IGP + BGP NHT cannot help
- This is a “pure” BGP convergence behavior
  - PE2 locally detects the link failure
  - PE2 updates its BGP, RIB, FIB tables
  - PE2 sends withdraws to its RR cluster
  - B cluster reflects to A cluster
  - A cluster reflects to PE1
  - PE1 modifies BGP, RIB and FIB table

Egress PE-CE Link Failure - Design

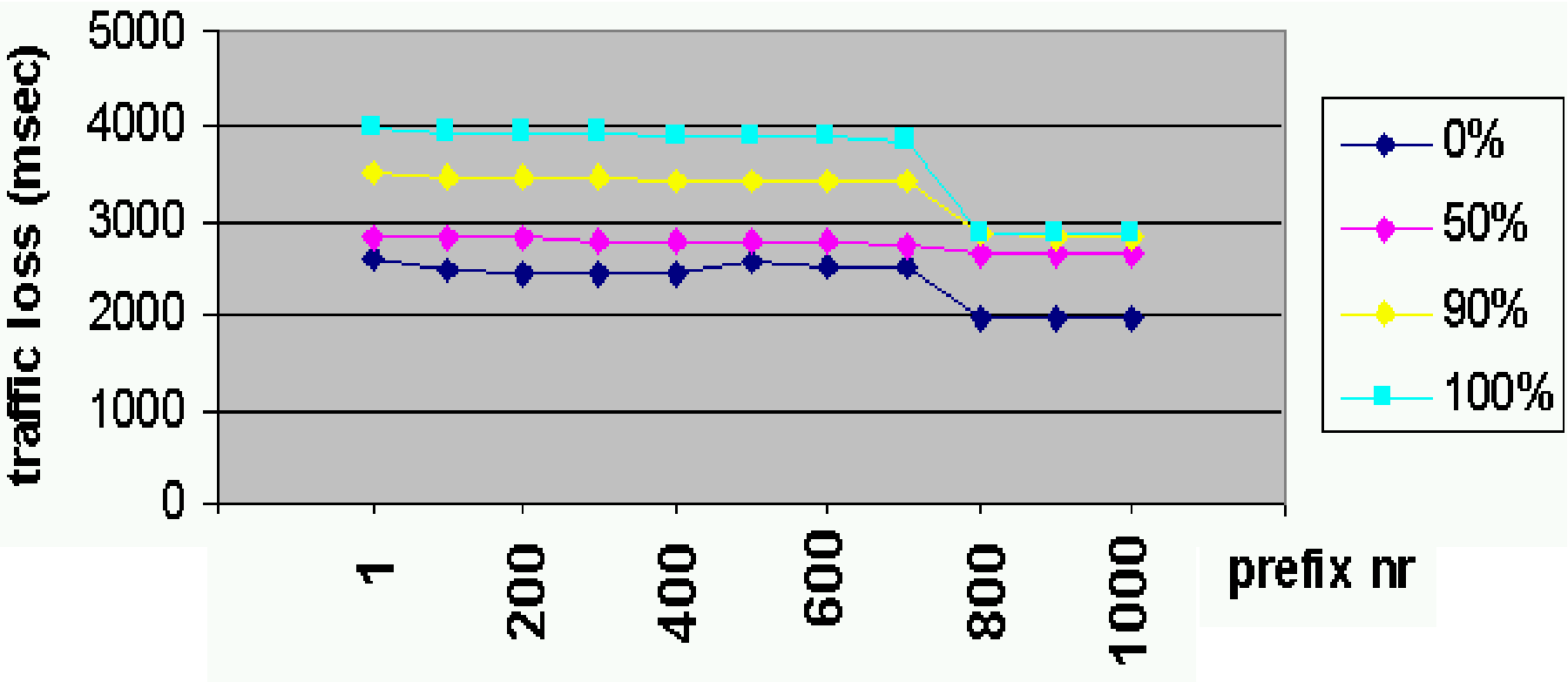
- **Immediate and Stable BGP reaction to Link Failure**
  - **bgp fast-external-fallover:**
  - **interface dampening**
- **Disable Minimum Advertisement Timer for MP-iBGP**
  - **in RFC2547 with unique RD, there is 1! Path per route. Also each VPN has different attributes hence the packing is low. Hence MAT for MP-iBGP brings no real gain.**
  - **default value of 5s would lead to a worst-case impact of 15s with two RR clusters**

```
router bgp
 address-family vpnv4
  neighbor <mp-ibgp neighbor> advertisement-interval 0
```

Egress PE-CE Link Failure - Design

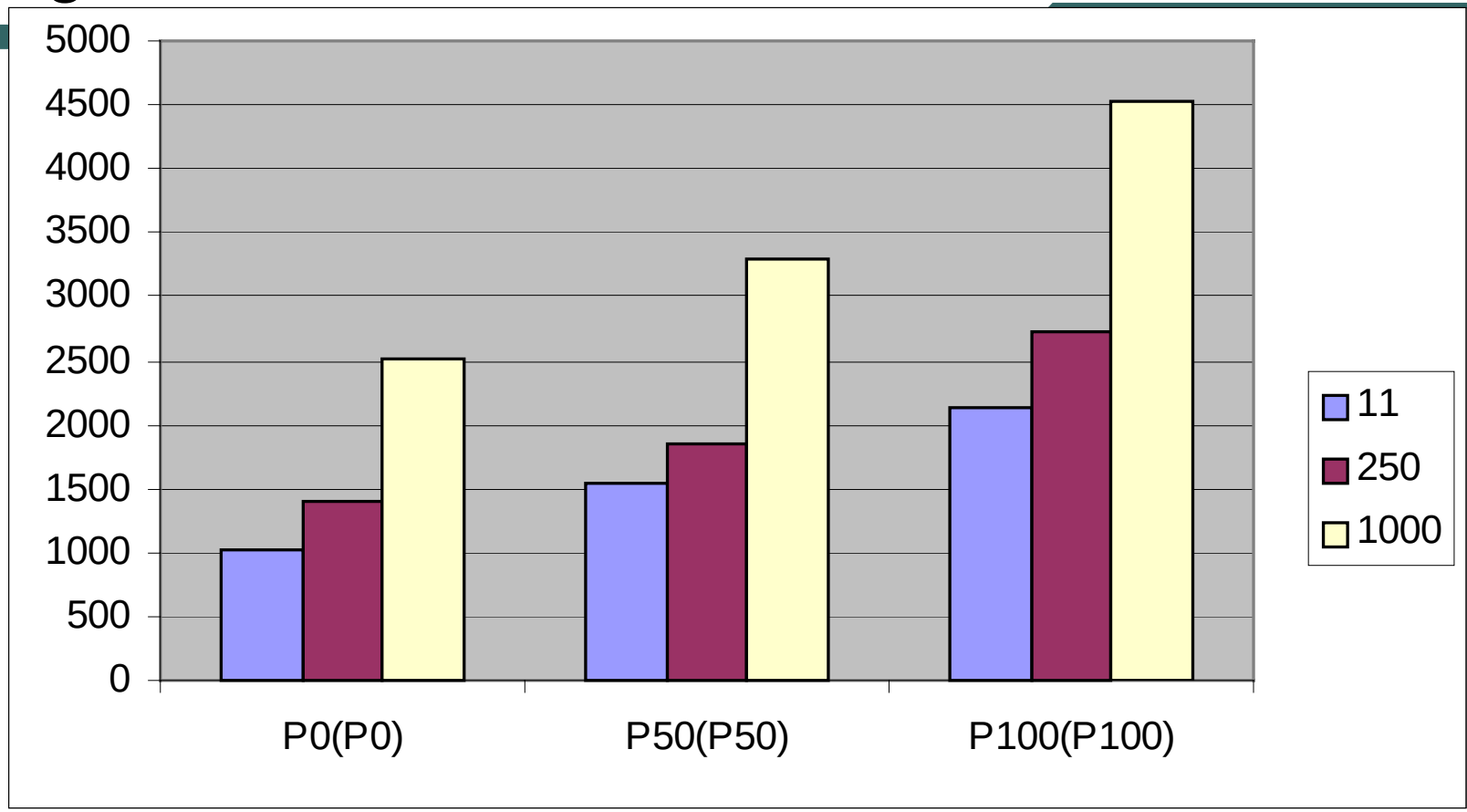
- **Optimize BGP transport goodput**
  - **Large input queue:** `hold-queue <1500-4000> in`
  - **Input Queue Prioriritization (automatic, 22S) (SPD)**
  - **Path MTU discovery:** `ip tcp path-mtu-discovery`
  - **Increase the TCP window size:** `ip tcp window-size`
  - **dynamic update group (automatic, 24S)**
  - **update packing optimization (automatic, 26S)**

Blackbox Measurement
Egress PE-CE Link Failure



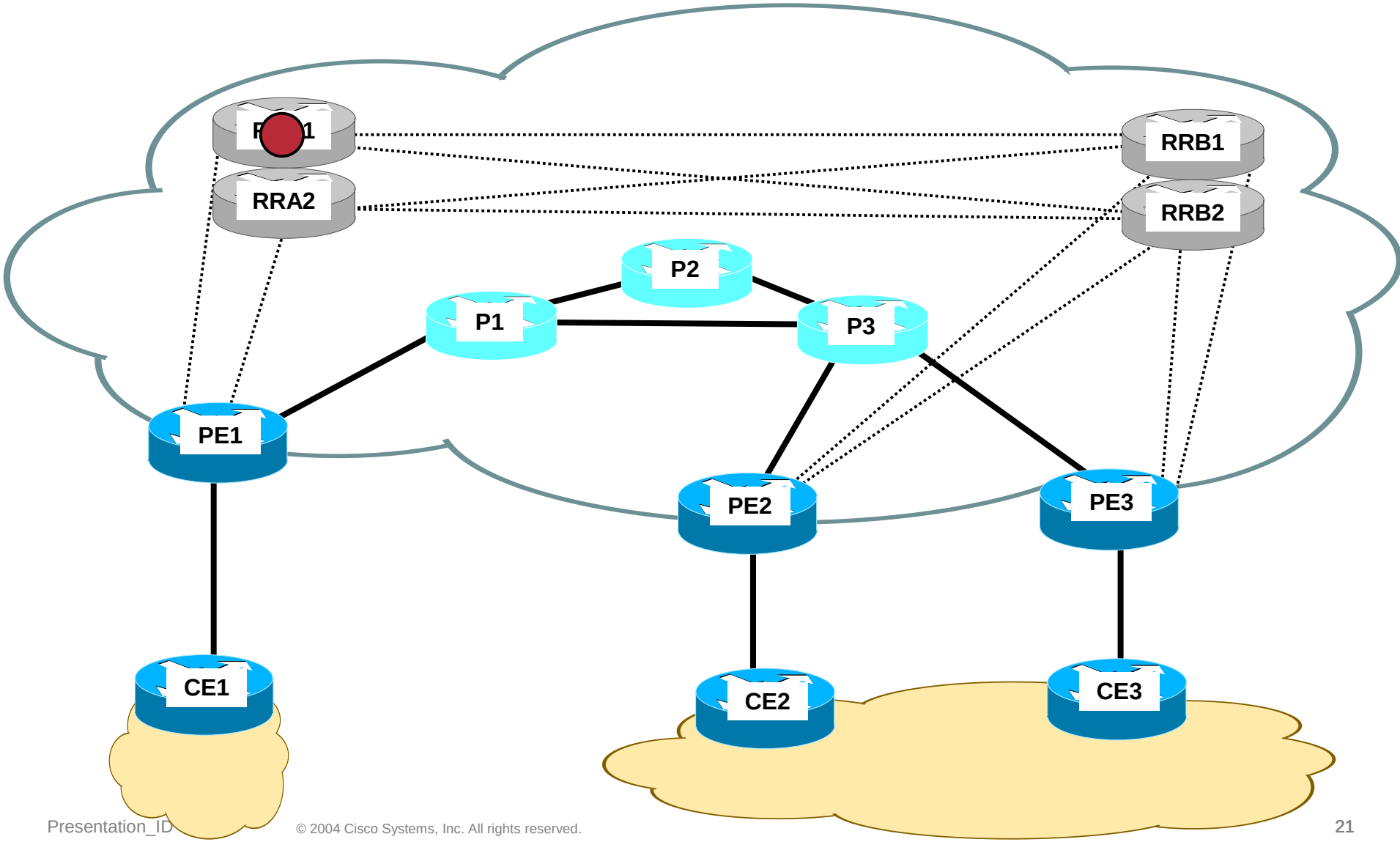
- **P100(1000prefixes): 3953ms**
- **P50(1000prefixes): 2750ms**

Blackbox Measurement
Egress PE-CE Link Failure



- **Data VPN: 80% of the CE's advertise less than 200 routes. It is very rare for a CE to advertise more than 1000 routes**
- **Voice VPN CE's would typically advertise < 10 prefixes**

RR failure within a redundant cluster



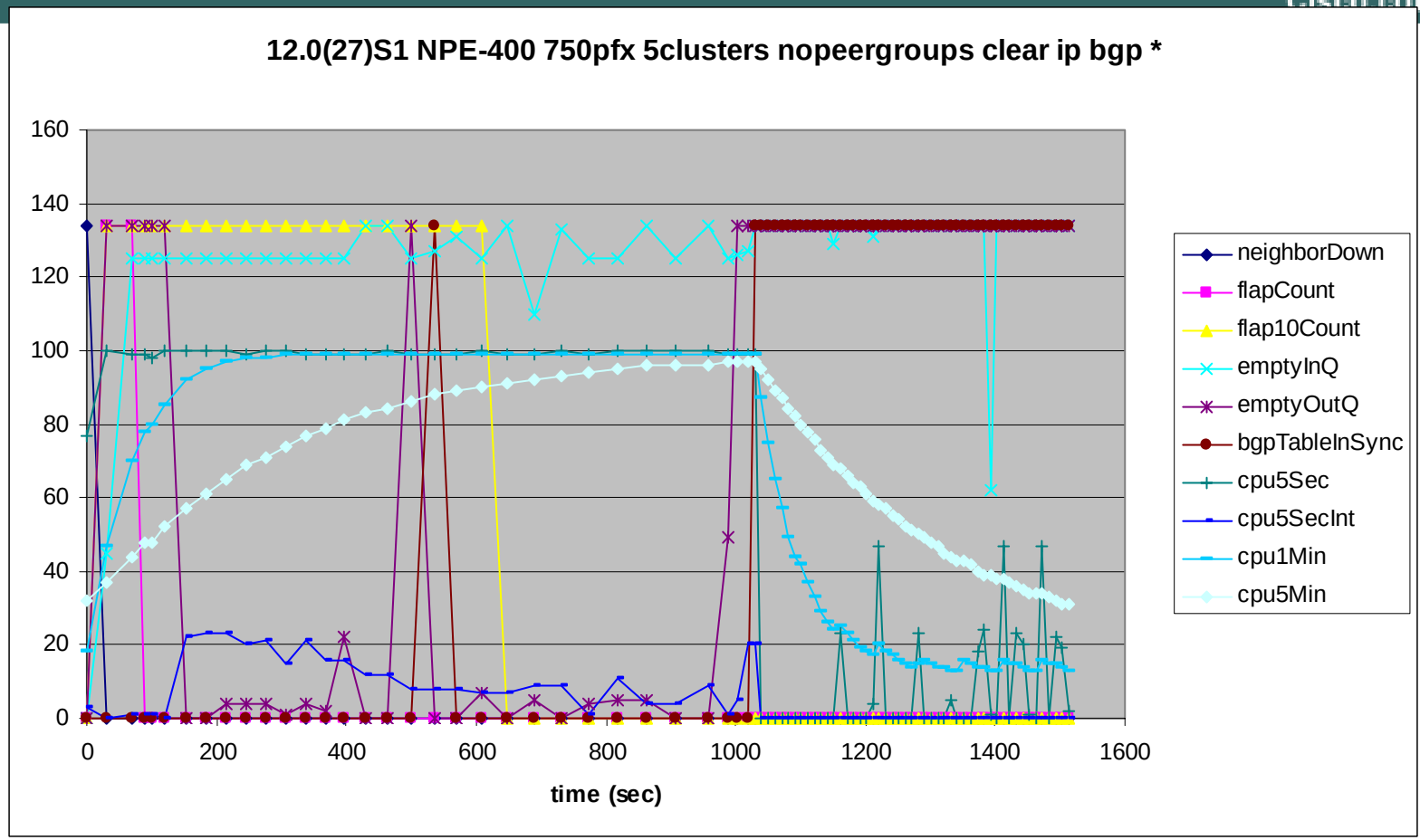
RR failure within a redundant cluster

- PE1 will discover the adj down after ~120/180s
- PE1 will then switch onto the same exact path but received from the other RR of the same cluster
- No Dataplane impact
- When RR comes back up, sessions must be reestablished with all peers and clients and BGP convergence must occur
  - we would like to optimize this ‘bring up’ time to minimize the non-redundancy period

RR failure within a redundant cluster Design

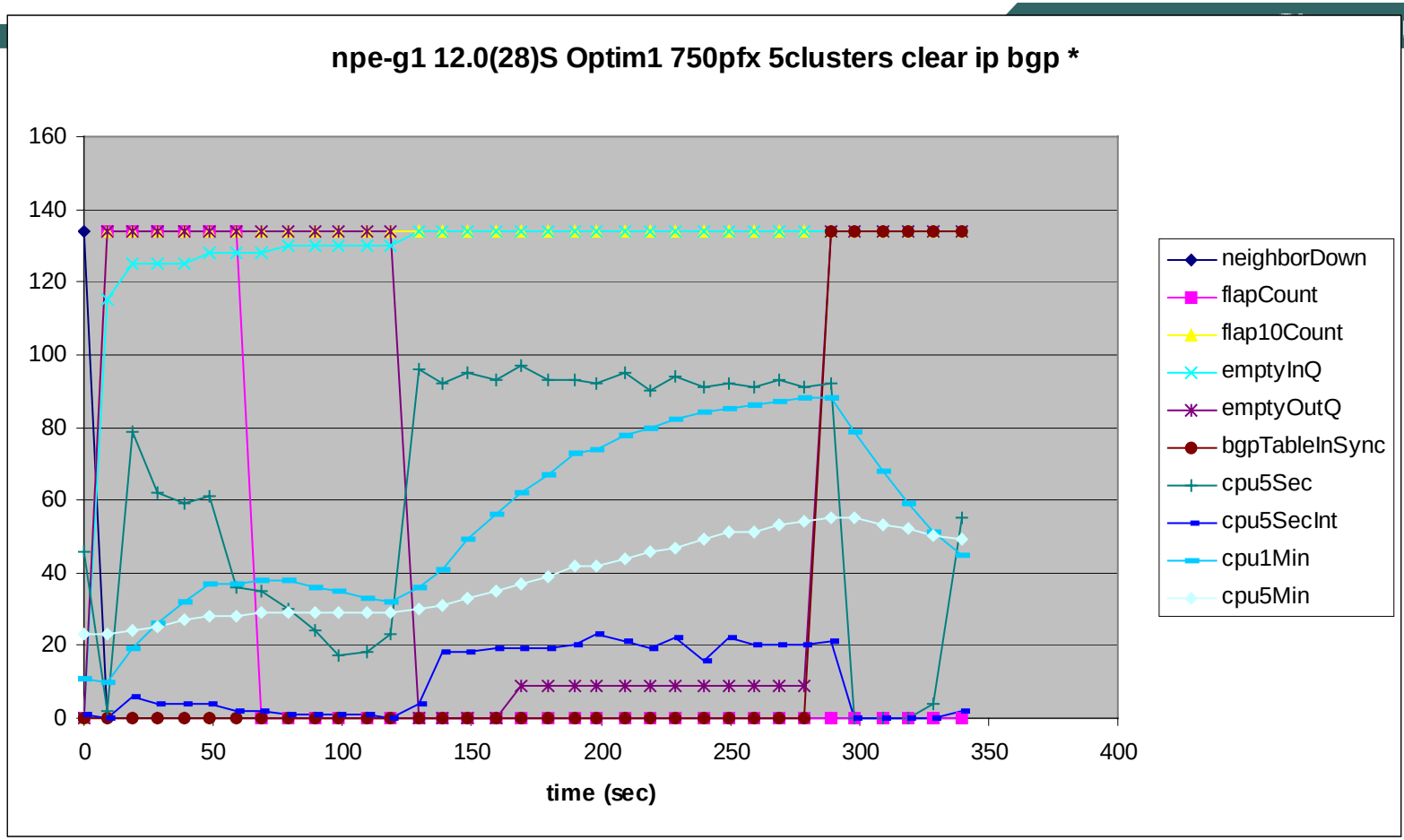
- **No dataplane impact**
  - **ensure that both paths are imported in the local VRF's**
- **Optimization of the RR 'bring up'**
  - **implementation optimization for BGP goodput (ie 26S)**
  - **key optimization of VPNv4 BGP table in 28S1**
  - **more CPU power means faster bring up (very cpu intensive)**

RR failure within a redundant cluster Measurement



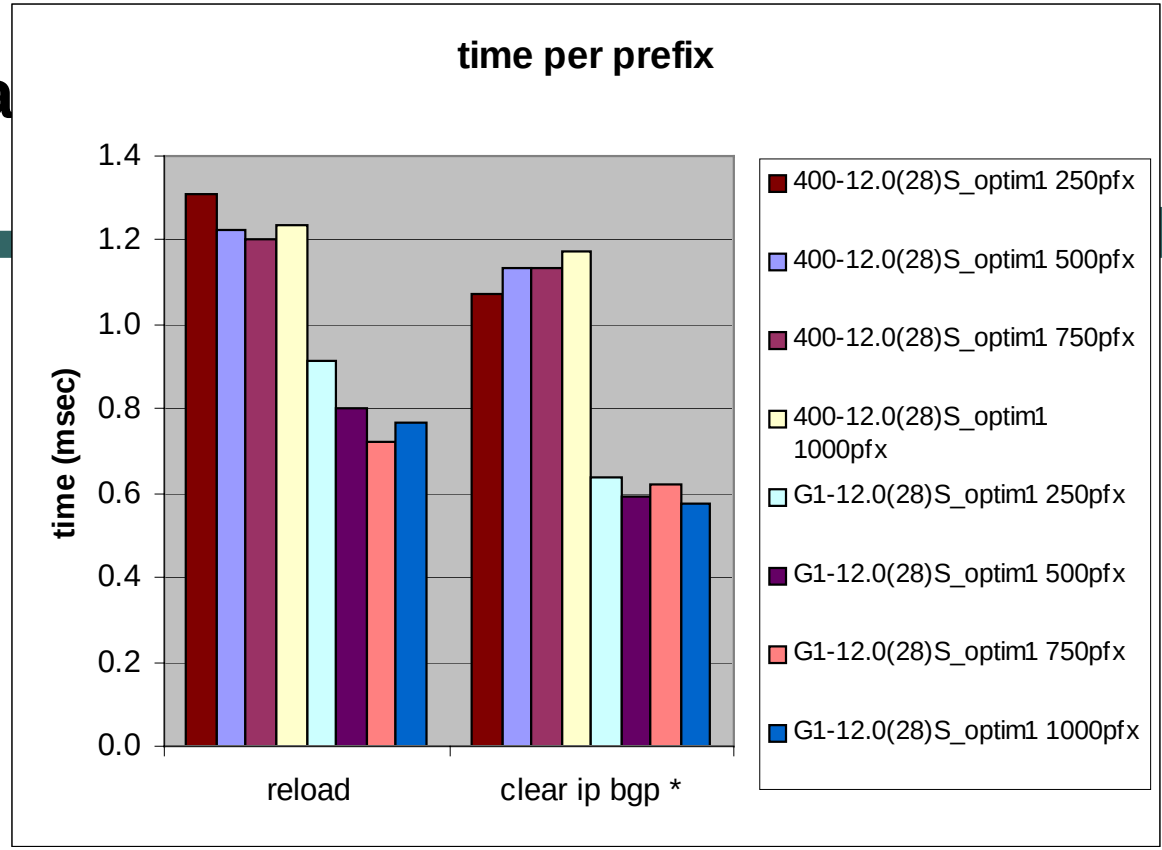
- **RR_Convergence(468750, npe400, 27S1) ~ 18 min**

RR failure within a redundant cluster Measurement



- **RR_Convergence(468750, npe400, 27S1) ~ 18'**
- **RR_Convergence(468750, npeG1, 28S1) ~ 4'40''**

RR failure within a
Measurement



- **NGE1 twice as performance as NPE400**
  - **~ factor 2 speed up in bring up time per prefix**
- **Key optimization in 28S1 (lab tests show 2 to 3 factor gain)**
  - **468750 * 0.6ms ~ 4min40sec**

Conclusion

- Sub-1s for core node/link failure
- No impact from RR failure
  - RR bring up for 500k vpnv4 routes in 4min40''
- PE node failure, PE-CE link failure
  - Prefix dependent
  - Sub-10s is conservative for most VPN's
  - Sub-5s is achievable with careful design
- Additional ideas to further optimize…