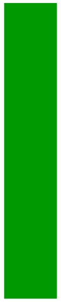
RFC2547 Convergence: Characterization and Optimization

Clarence Filsfils

cf@cisco.com

RFC2547 Convergence - Requirement

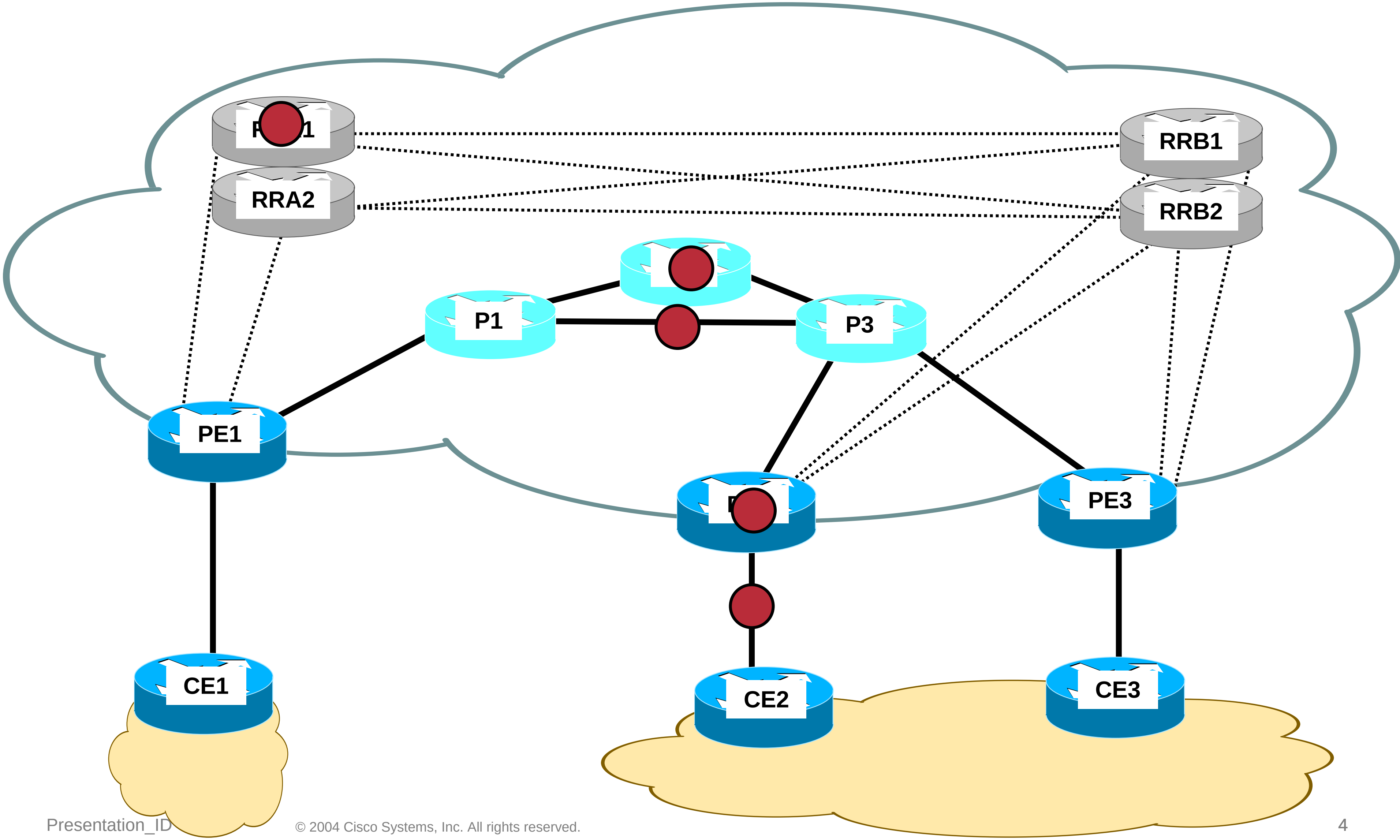
- < 10s
- < 5s
- < 3s
- < 1s
- < 250ms
- < 50ms



RFC2547 – what is possible

- **Once the convergence behavior is optimized, the fundamental parameter is how many prefixes are impacted by the failure**

What failures to consider



The fundamental parameter for Convergence: how many impacted prefixes?

- Core Link/Node Failure
  - # of important impacted prefixes likely < 500
- Edge PE node failure
  - analysis of deployed RFC2547 networks is ongoing
  - for custX: 90% of the PE failures impact less than 250 prefixes across less than 50 vrf … this is rather small and hence more analysis is required to confirm the real numbers
- PE-CE Link Failure
  - custX: 80% of the links advertise less than 250 prefixes and 96% advertise less than 2000 prefixes
  - custY: 90% of the links advertise less than 25 prefixes and 100% advertise less than 250 prefixes
- RR failure
  - multiple 100k's of prefixes are impacted

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

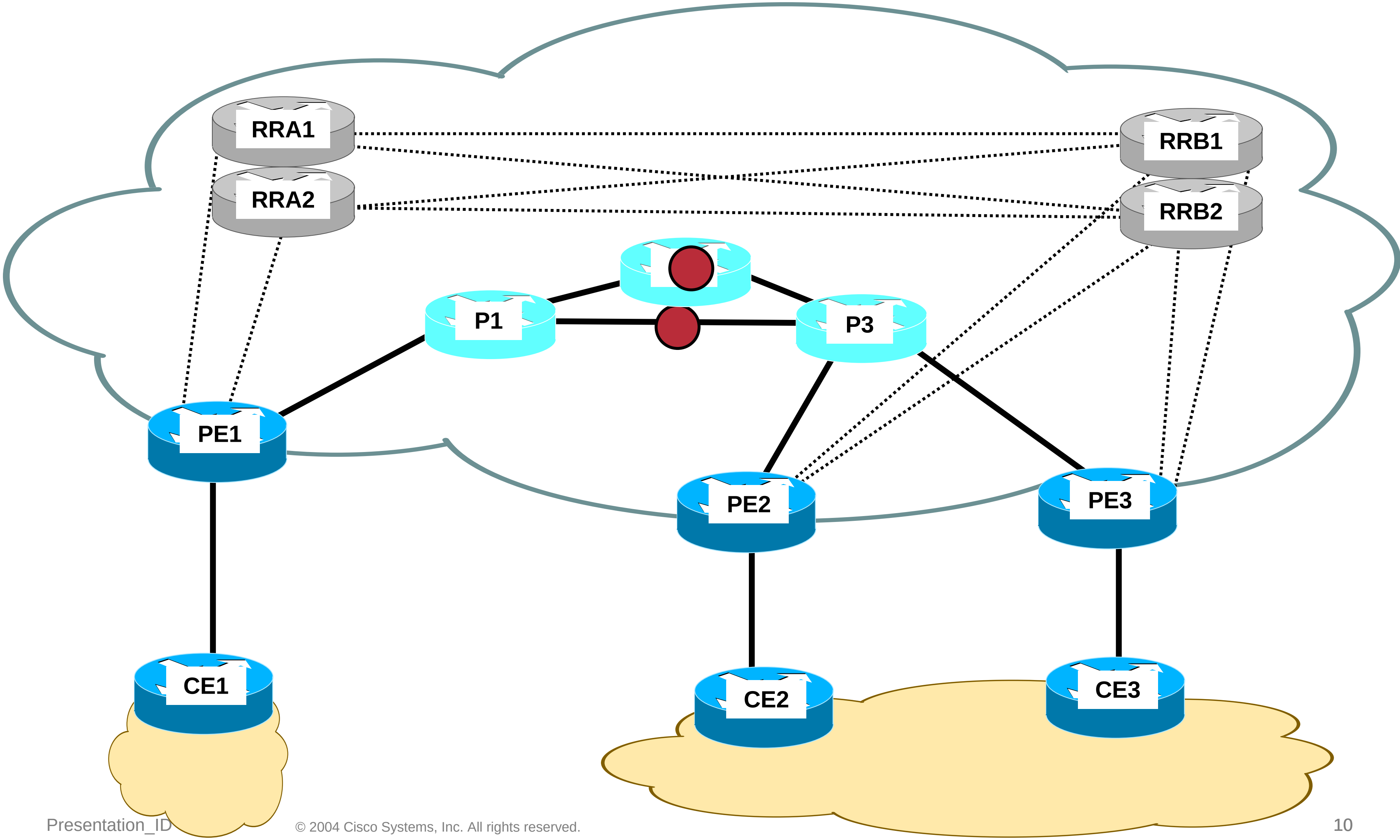
Design Context/Constraint

- Convergence to a redundant site
  - loadsharing or primary/backup policy
- RD allocation technique:
  - RDU: unique RD per VRF
  - RDZ: same RD in all VRF's of the same VPN except for the second VRF connected to a redundant site

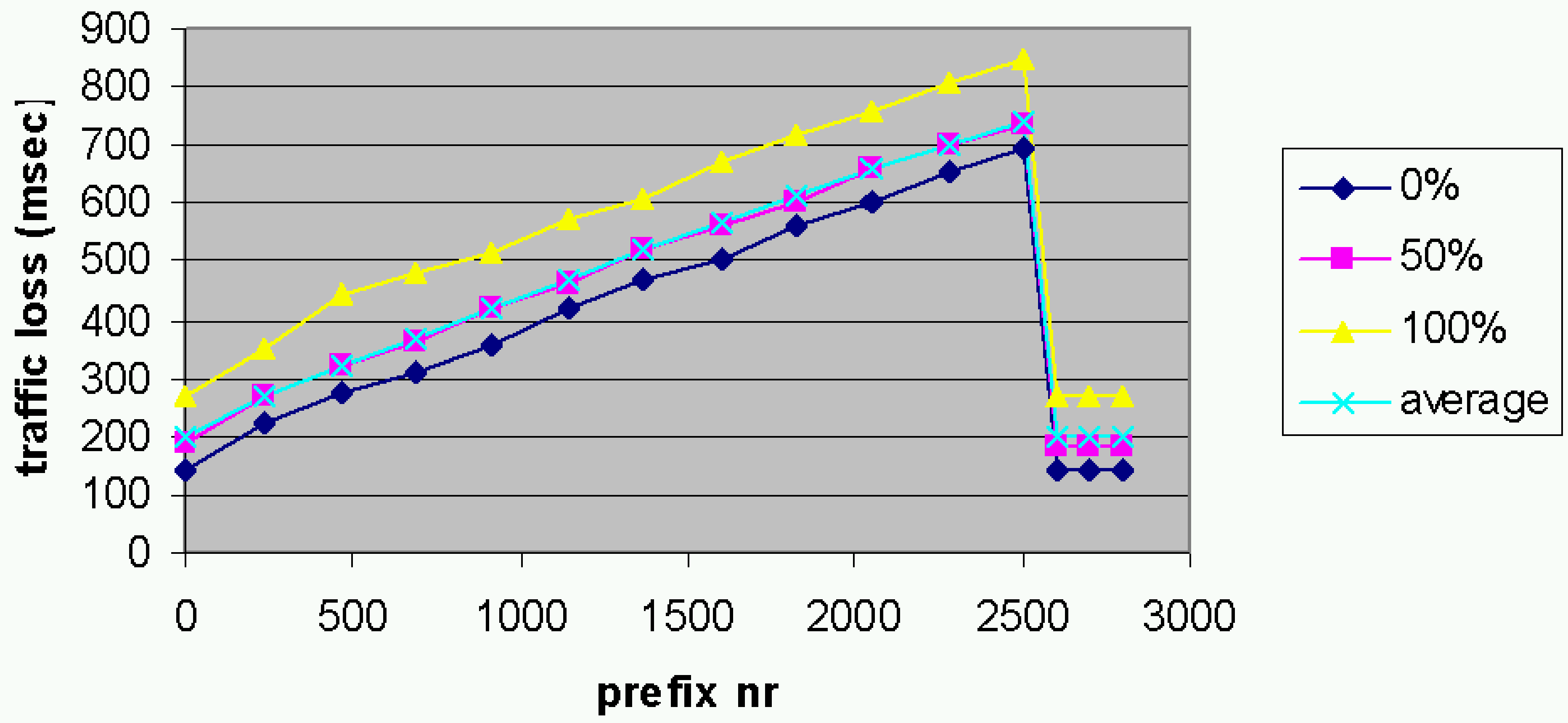
Reality check

- All the results were measured on
  - 12k, PRP1, Eng3, 12.0(31)S
- We limit our talk to technology that exists in 12.0(31)S

Core Link/Node Failure

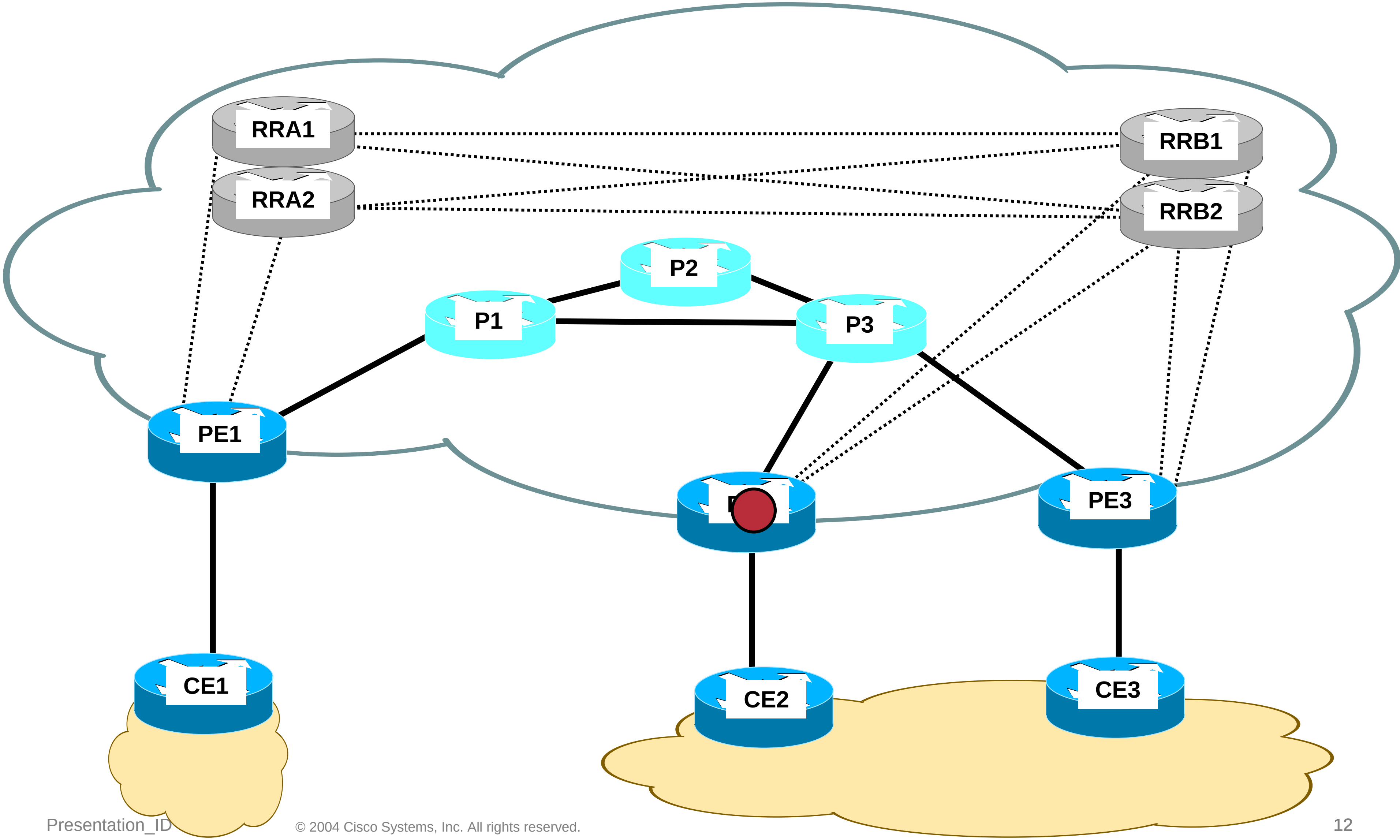


IGP Fast Convergence sub-second is conservative



- **For more details, refer to Apricot 2004 presentation**
  - **also at Nanog 29, Ripe 47, Apricot04, MPLSWorld05**
- **Paper under submission**

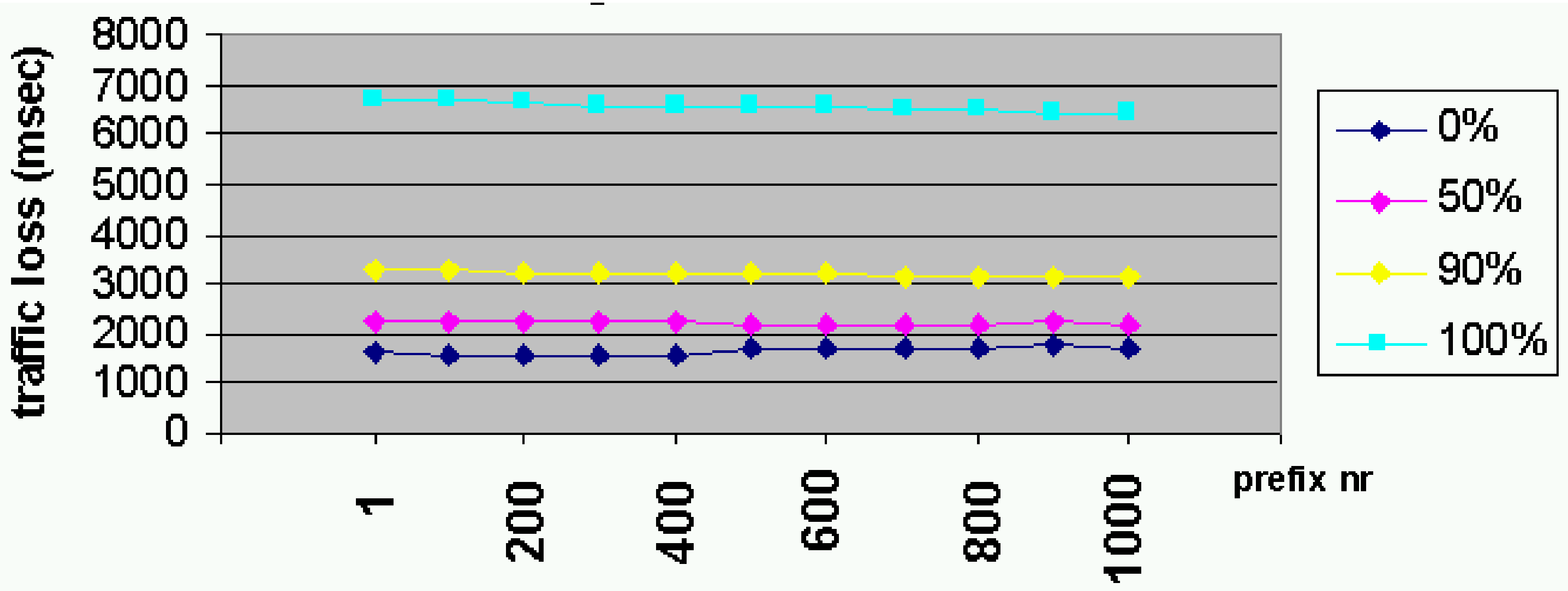
Egress PE Node failure



Egress PE node failure

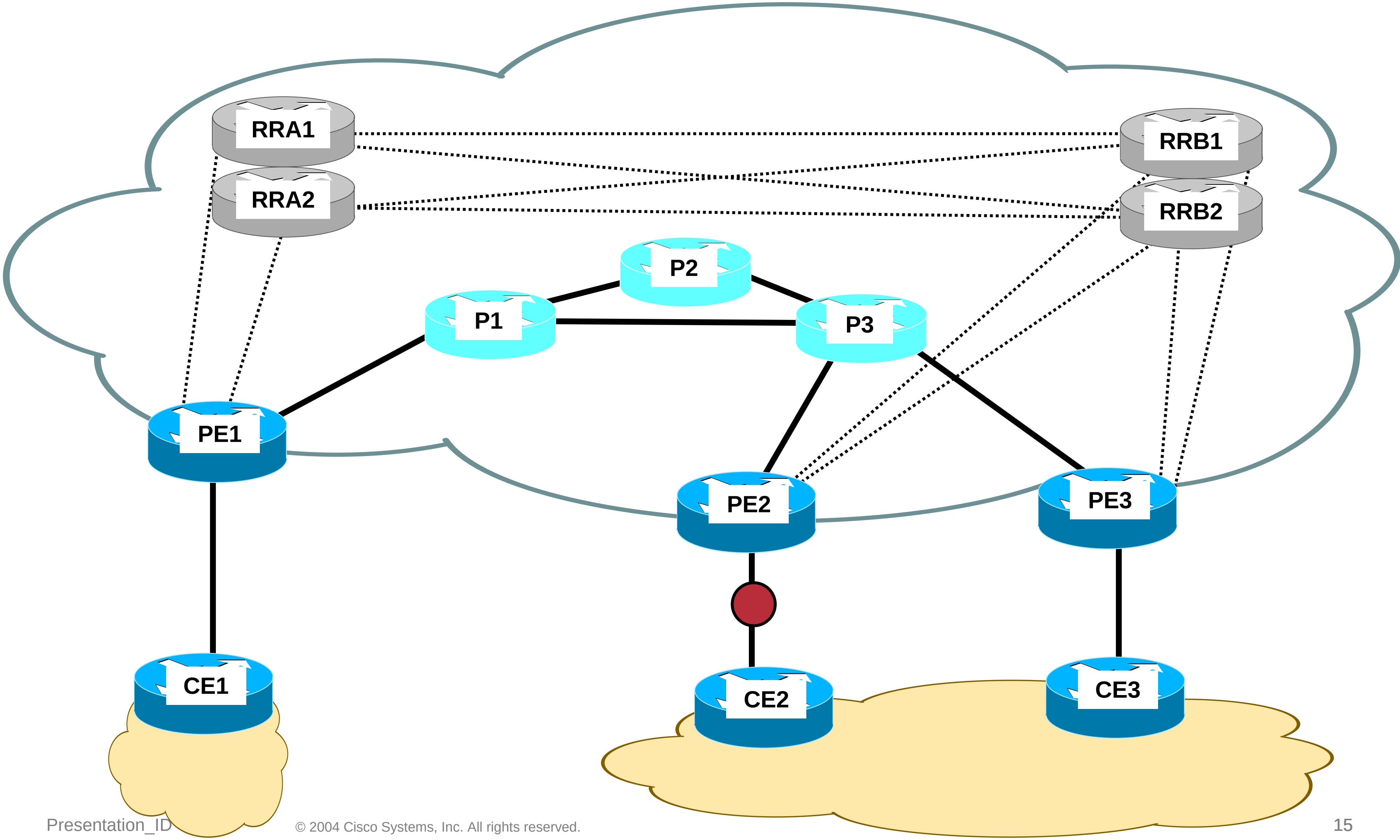
- RDU/RDZ ensures that PE1 knows about the 2 paths prior to the failure
- Adjacent core nodes detect the failure of PE2 and flood new LSP's advertising the failure
- PE1's IGP converges and declares PE2 unreachable
- PE1: Unreachable status of a BGP nhop triggers BGP Convergence which simply consists in invalidating one of the two known paths
- Conclusion
  - no BGP signalling required
  - computation is proportional to number of impacted entries

Blackbox Measurement
Egress PE node failure



- **PE1 selects 1000 prefixes from PE2**
- **Traffic is sent to 11 prefixes**
- **For custX: 90% of the PE failures impact less than 250 prefixes across less than 50 vrf … this is rather small and hence more analysis is required to confirm the real numbers**

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

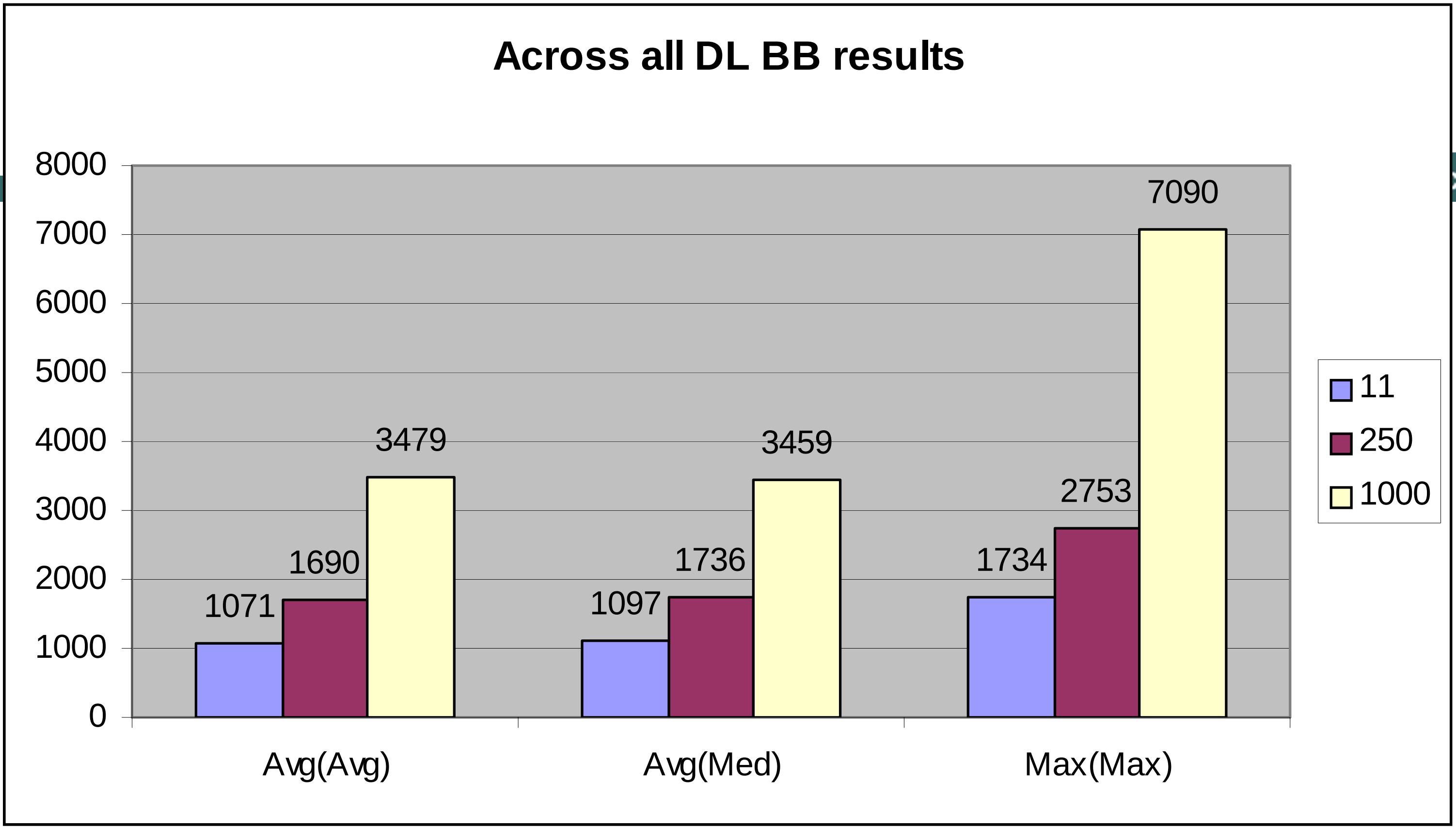
Egress PE-CE Link Failure - Design

- 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

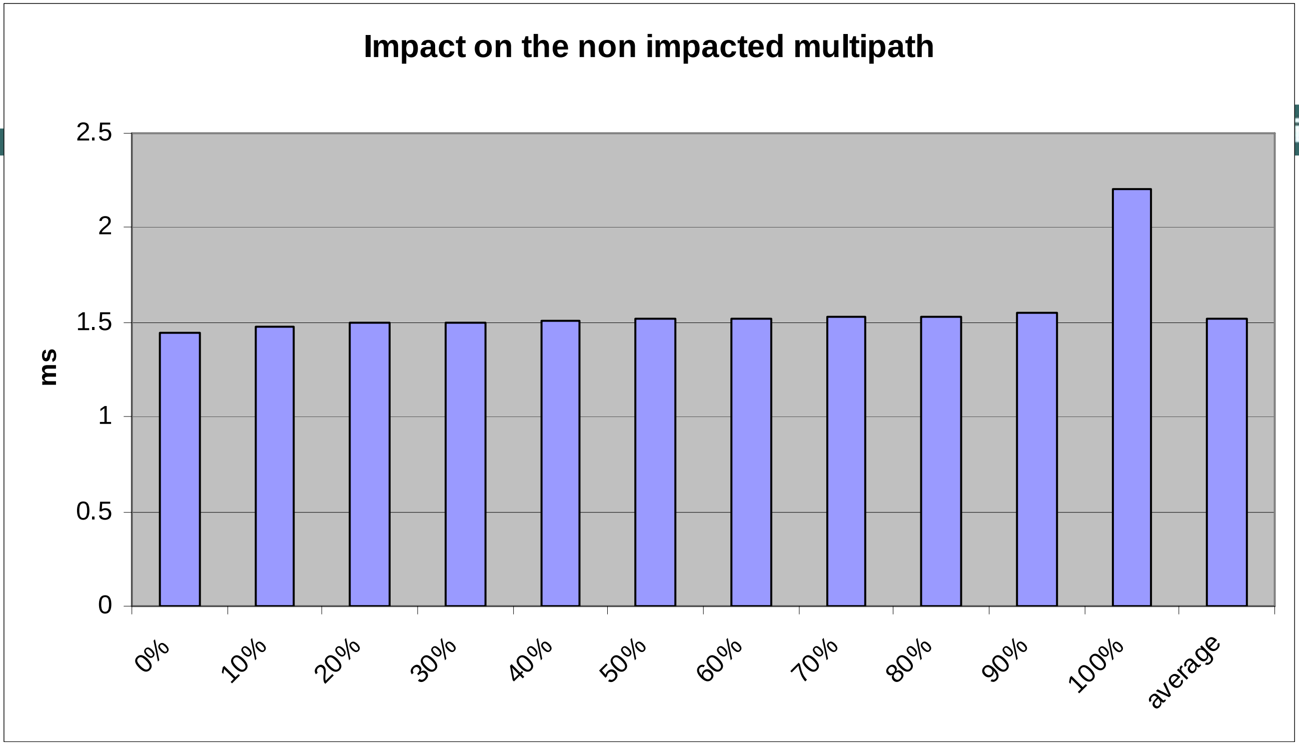
```
router bgp
 address-family vpnv4
  neighbor <mp-ibgp neighbor> advertisement-interval 0
```

Egress PE-CE Link Failure - Design

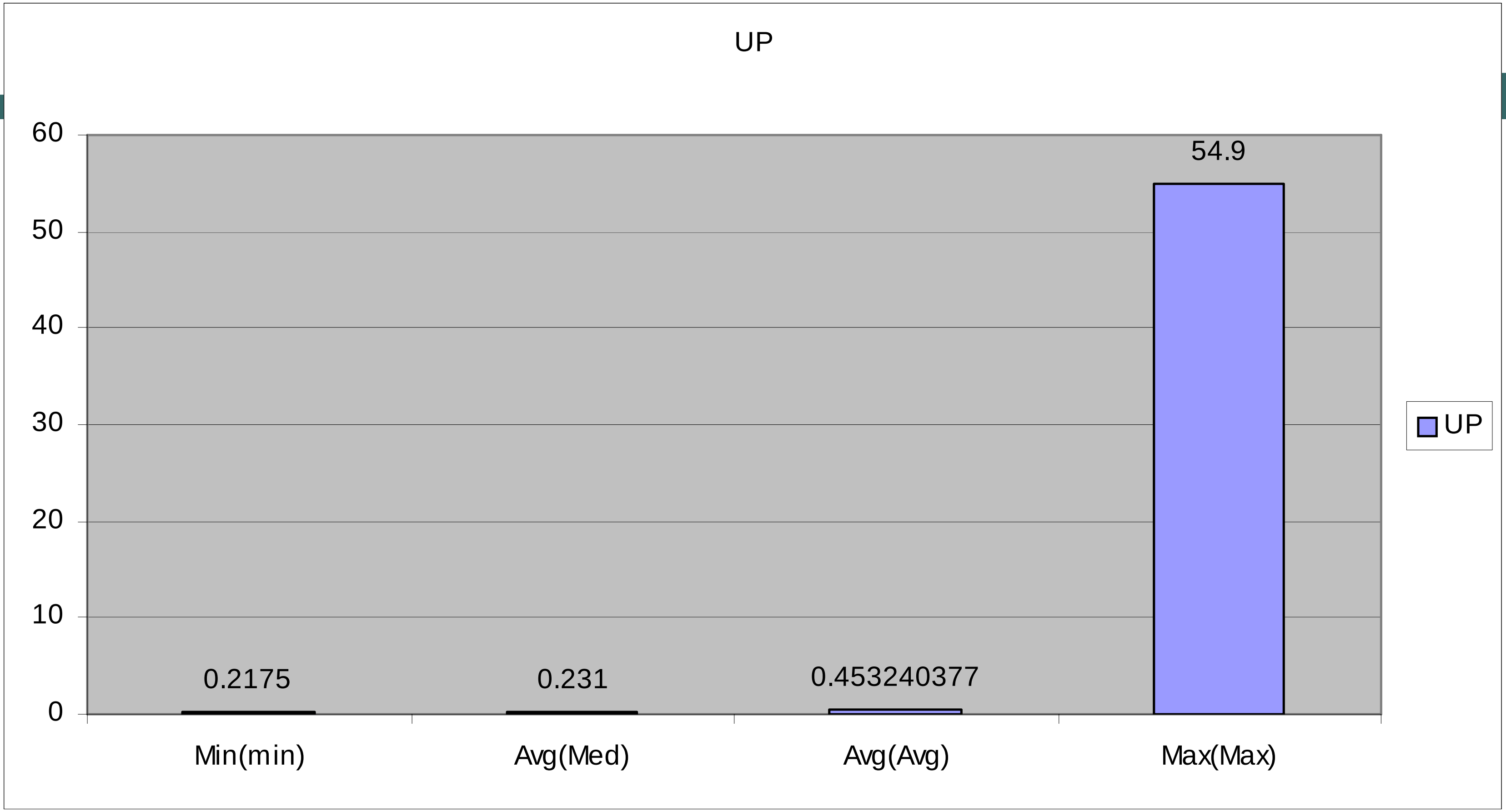
- Optimize BGP transport goodput
  - Large input queue: `hold-queue <1500-4000> in`
  - Input Queue Prioriritization (automatic, 22S) (SPD)
  - Path MTU discovery: `ip tcp path-mtu-discovery`
  - Increase the TCP window size: `ip tcp window-size`
  - dynamic update group (automatic, 24S)
  - update packing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



- **custX: 80% of the links advertise less than 250 prefixes and 96% advertise less than 2000 prefixes**
- **custY: 90% of the links advertise less than 25 prefixes and 100% advertise less than 250 prefixes**
- **VoIP VPN design: a few MGW's per site ➔ << 10 prefixes per site**

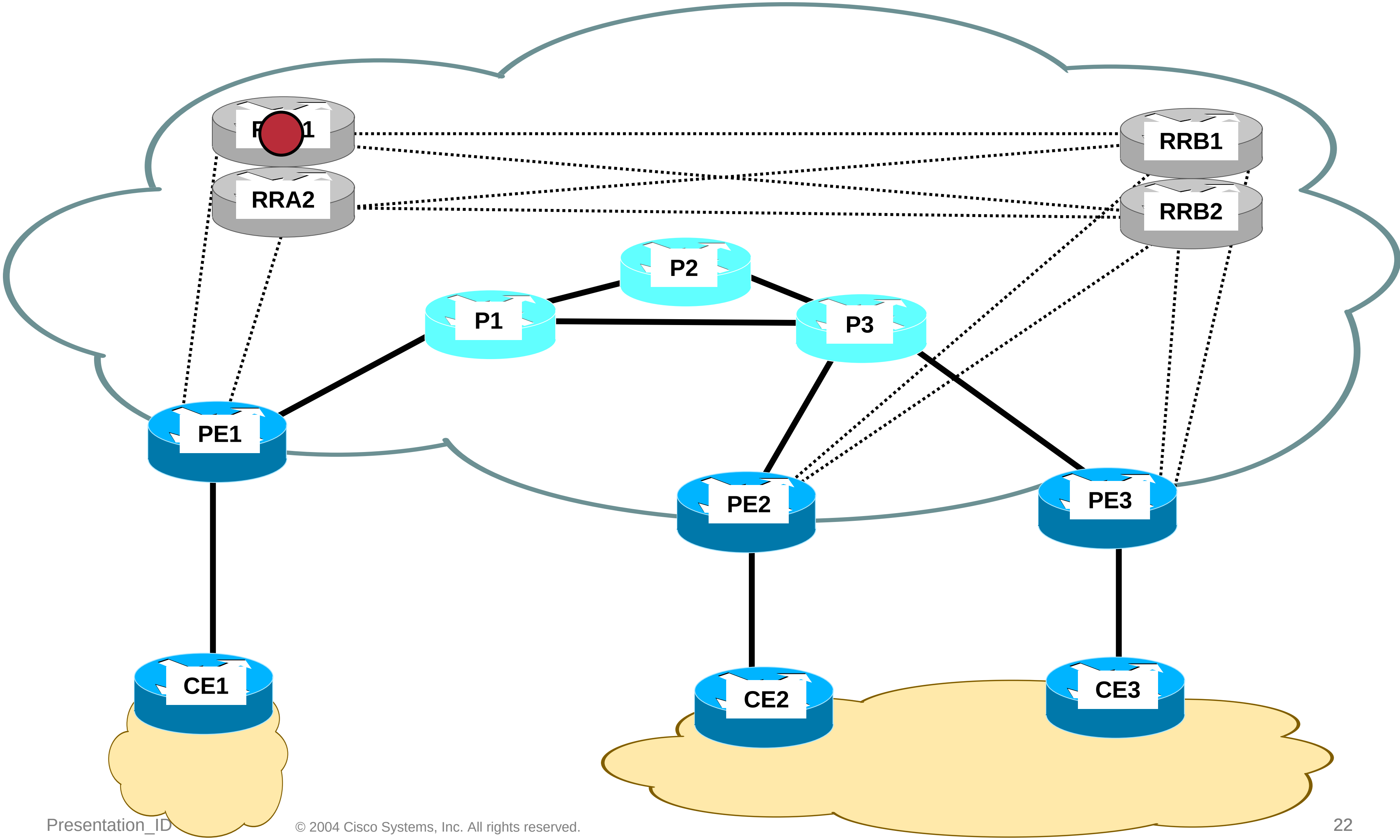


- **Negligible impact (~ 1ms)**



- **No Loss on Link Up (negligible)**

RR failure within a redundant cluster



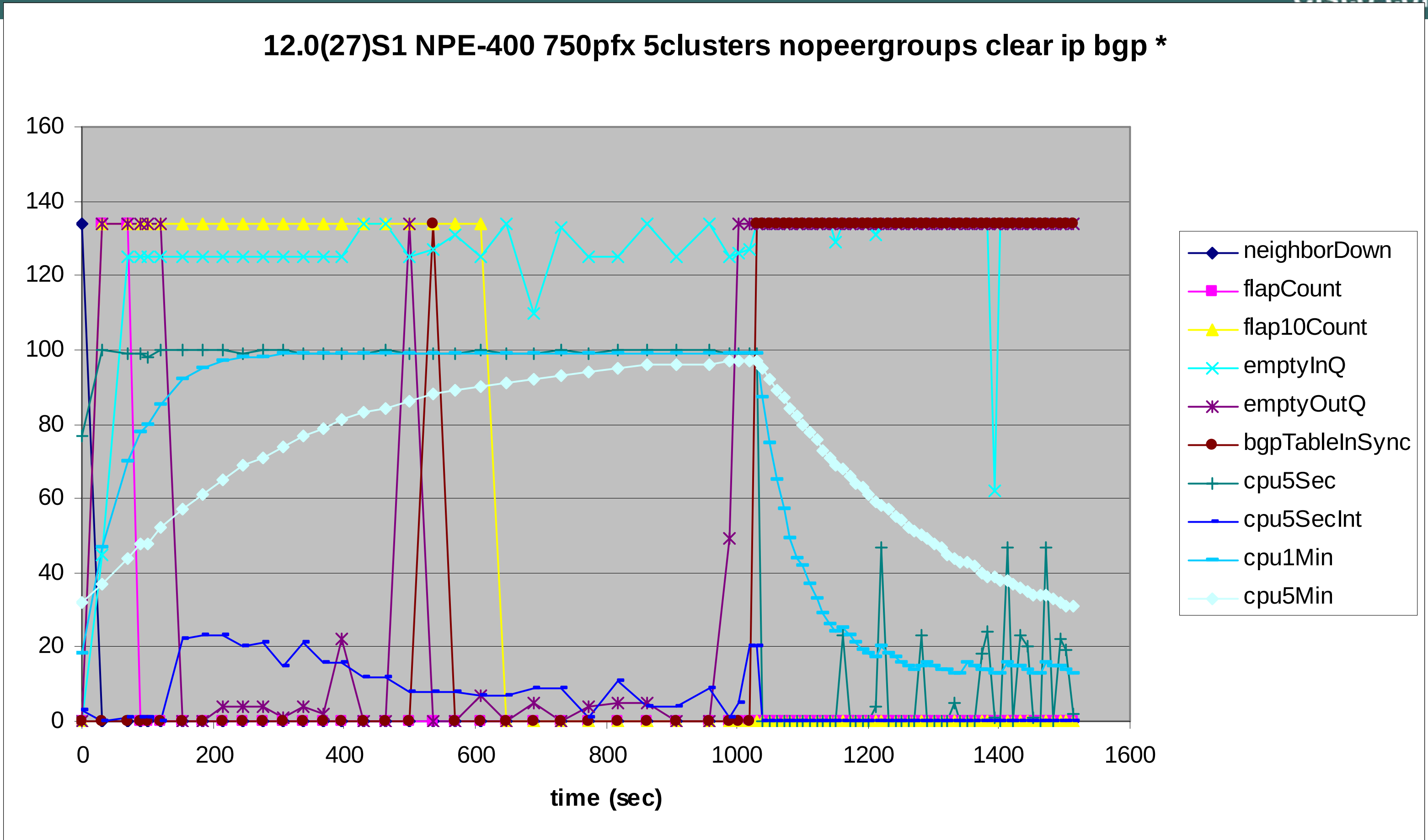
RR failure within a redundant cluster

- **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 provided we import the necessary paths**
- **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

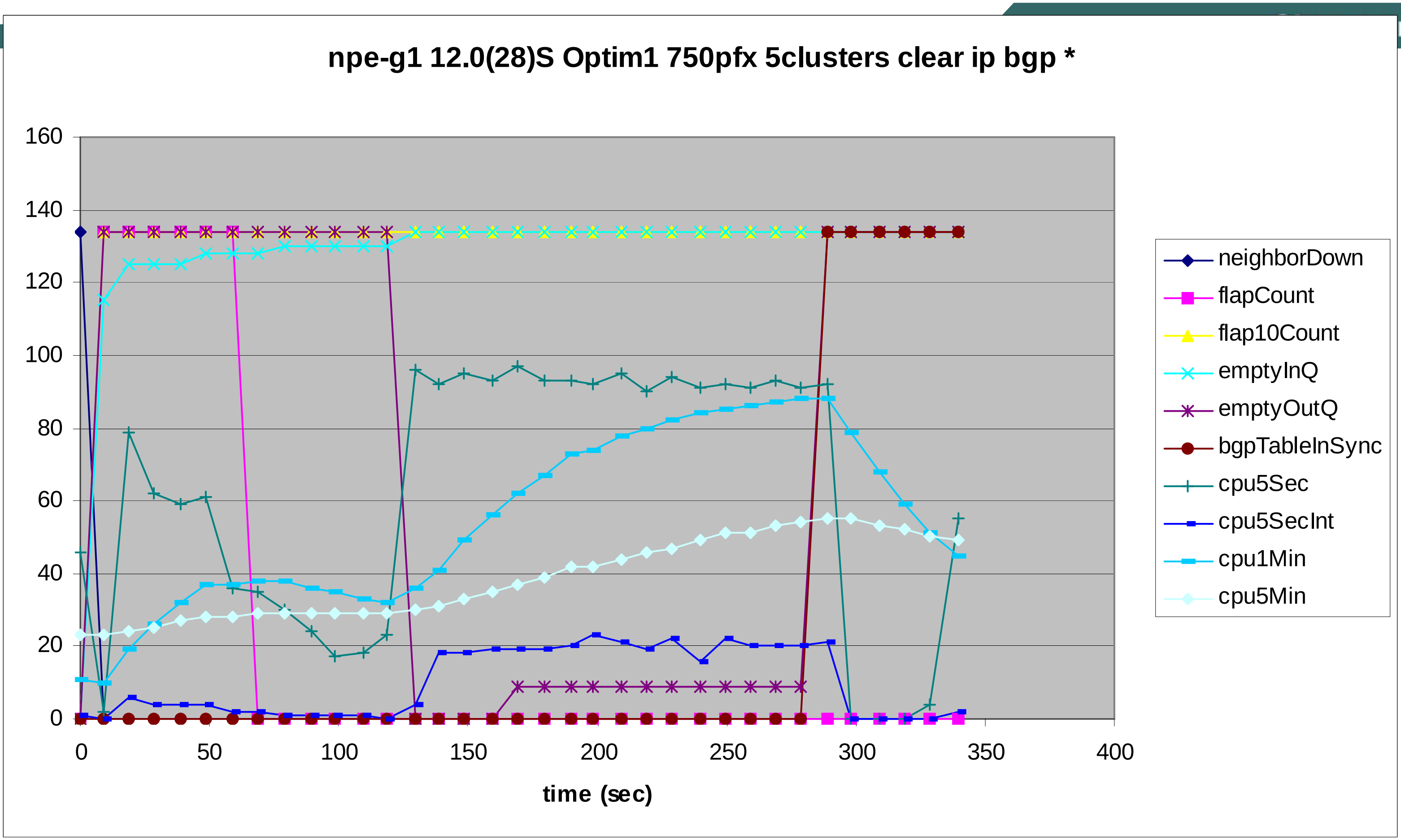
- No dataplane impact
  - ensure that both paths are imported in the local VRF’s
- Optimization of the RR ‘bring up’
  - implementation optimization for BGP goodput (ie 26S)
  - key optimization of VPNv4 BGP table in 28S1
  - more CPU power means faster bring up (very cpu intensive)

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’’

Conclusion

- Based on the number of impacted prefixes discussed previously and the test results:
  - Core node/link failure: <1s is achievable
  - PE-CE Link: <2 to 3s is achievable
  - PE node failure: < 10s is achievable
  - RR bring up (500k pref): < 5min, no dataplane impact
- We have additional ideas to further optimize…
- Please give your requirement/feedback