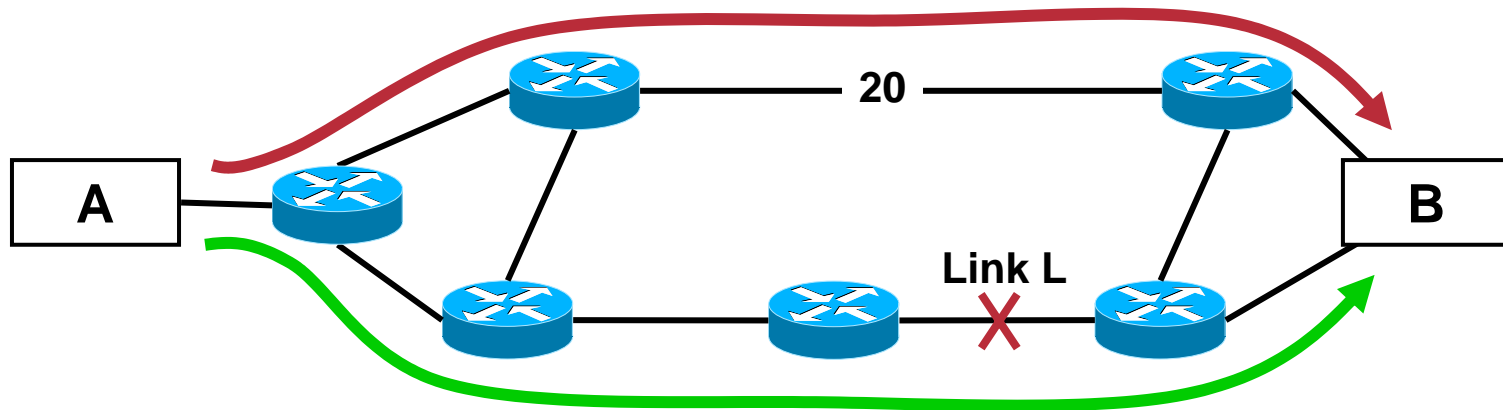# Fast IGP Convergence

John Evans – joevans@cisco.com

Definition of convergence

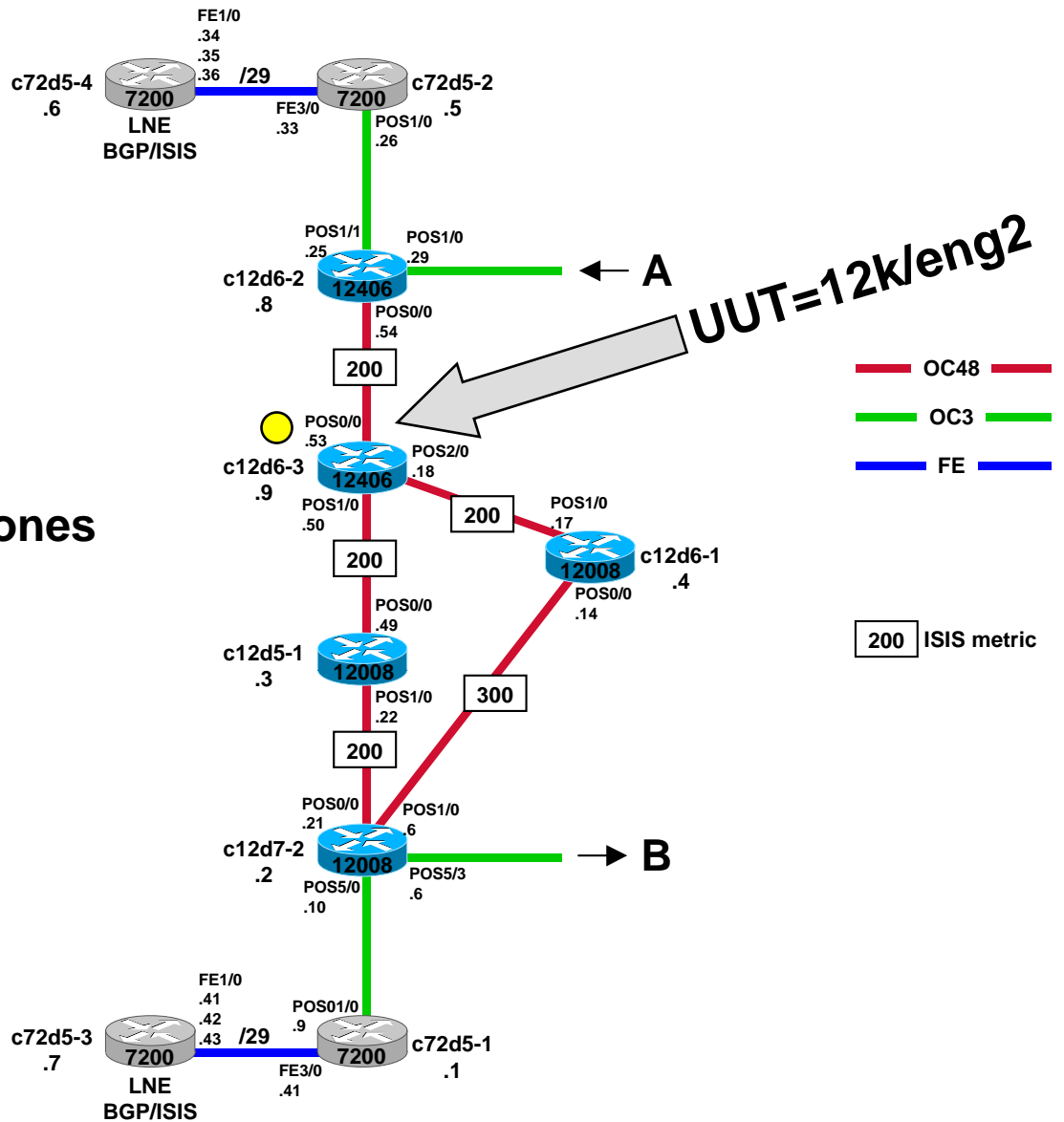**Default metric = 1**

- **Assume a flow from A ➔ B**

- **T1: when link L fails, the best path is impacted and loss of traffic starts**

- **T2: when the network converges, a next best path is computed and traffic reaches the destination again**

- **Loss of Connectivity = T2 – T1; a.k.a "convergence time"**

- **Analyzed for streams going to IGP and BGP learned prefixes**
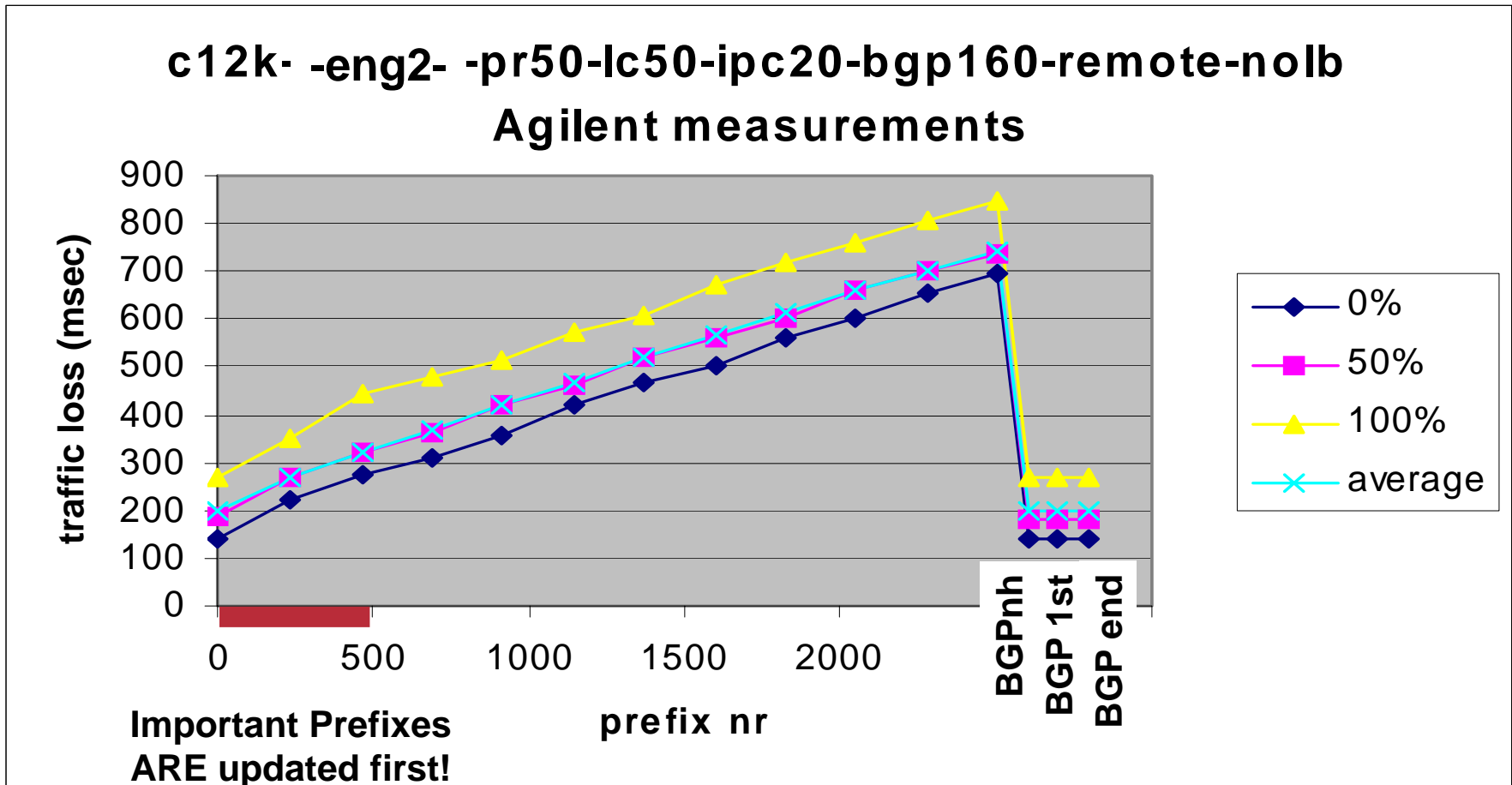
Objective

- **Sub-second for**

  the first 500 IGP Prefixes

  all BGP prefixes whose next-hop is within the first 500 IGP prefixes assuming the BGP routes are stable
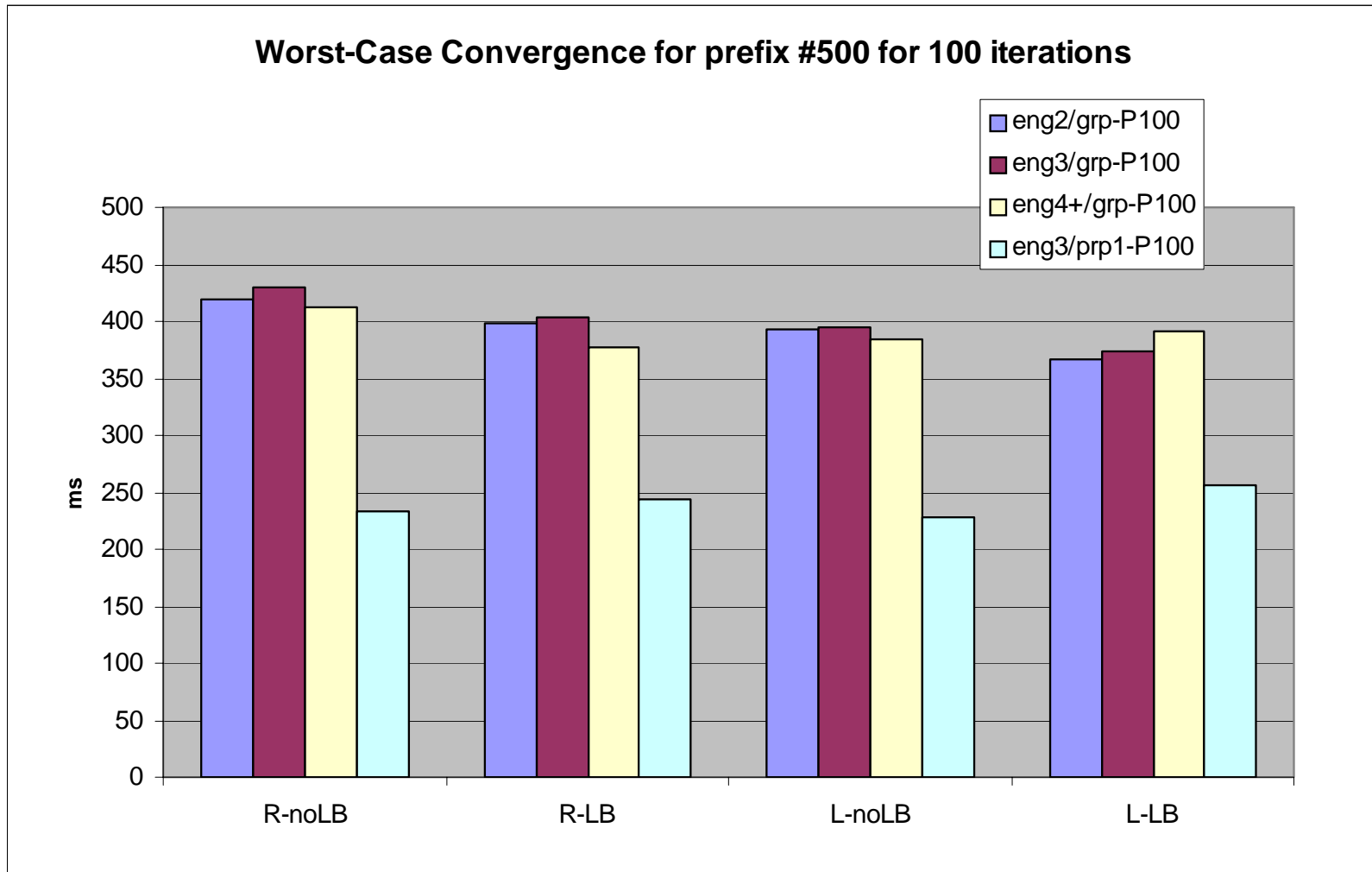
- **IGP: ISIS**

  also applicable to OSPF

Lab Setup

- **Pre 12.0(27)S**
- **1000 ISIS nodes**
- **2500 ISIS prefixes**
  - **500 important ones**
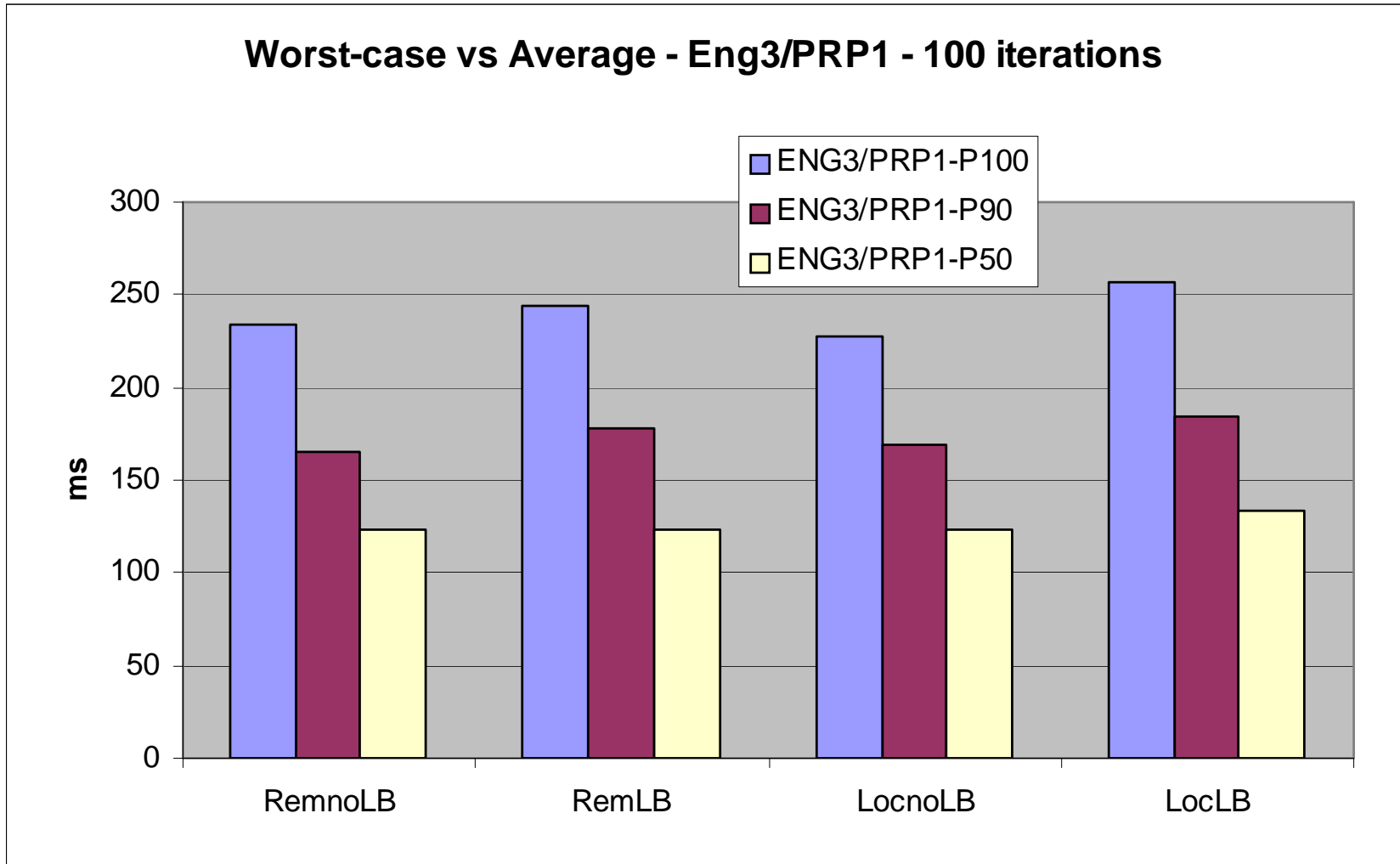 - **2000 non-important ones**
- **160k BGP routes**
- **No flap**
- **POS**

c72d5-4
.6
7200
LNE
BGP/ISIS

FE1/0
.34
.35
.36 /29

FE3/0
.33

c72d5-2
.5
7200

POS1/0
.26

POS1/1
.25

POS1/0
.29 ← A

c12d6-2
.8
12406

POS0/0
.54

200

POS0/0
.53

c12d6-3
.9
12406

POS2/0
.18

200

POS1/0
.17

c12d6-1
.4
12008

POS1/0
.50

200

POS0/0
.14

POS0/0
.49

c12d5-1
.3
12008

POS1/0
.22

200

300

POS0/0
.21

POS1/0
.6

c12d7-2
.2
12008

POS5/0
.10

POS5/3
.6 → B

FE1/0
.41
.42
.43 /29

POS01/0
.9

FE3/0
.41

c72d5-3
.7
7200
LNE
BGP/ISIS

c72d5-1
.1
7200

UUT=12k/eng2

OC48
OC3
FE

200 ISIS metric

Remote noLB – ISIS

c12k- -eng2- -pr50-lc50-ipc20-bgp160-remote-nolb
Agilent measurements

Important Prefixes ARE updated first!

Black-Box measurements: Max(Pref #500)

Worst-Case Convergence for prefix #500 for 100 iterations

Legend:
- eng2/grp-P100
- eng3/grp-P100
- eng4+/grp-P100
- eng3/prp1-P100

6

Max vs average for 500 first prefixes

Worst-case vs Average - Eng3/PRP1 - 100 iterations

Legend:
- ENG3/PRP1-P100
- ENG3/PRP1-P90
- ENG3/PRP1-P50

Y-axis (ms): 0, 50, 100, 150, 200, 250, 300

X-axis categories: RemnoLB, RemLB, LocnoLB, LocLB

Flooding impact

flooding-bgp160-16fps-snmp-pr50-ipc20

A line chart with y-axis labeled "delay (msec)" ranging from 0 to 100 and x-axis labeled "percentile" ranging from 0% to 100%. The delay stays near 0 until about 85% percentile (labeled 2.05), then rises sharply to about 90 msec at 100%.

- ## **Flooding occurs before SPF**

Propagation distance - analysis

Default metric: 1

Y

10

Reroute
Node
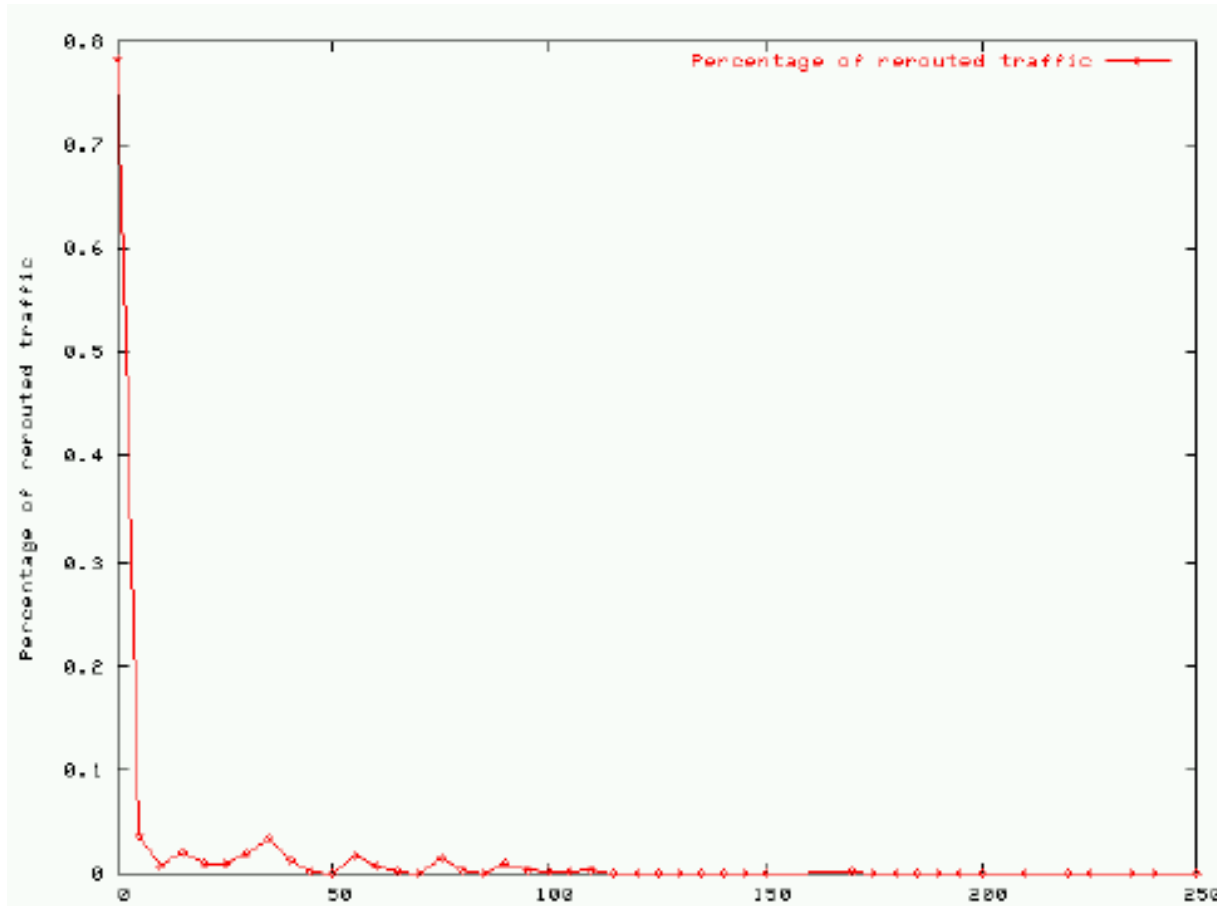
X

F

Z

W

Propagation:
1 hop
X-F ms
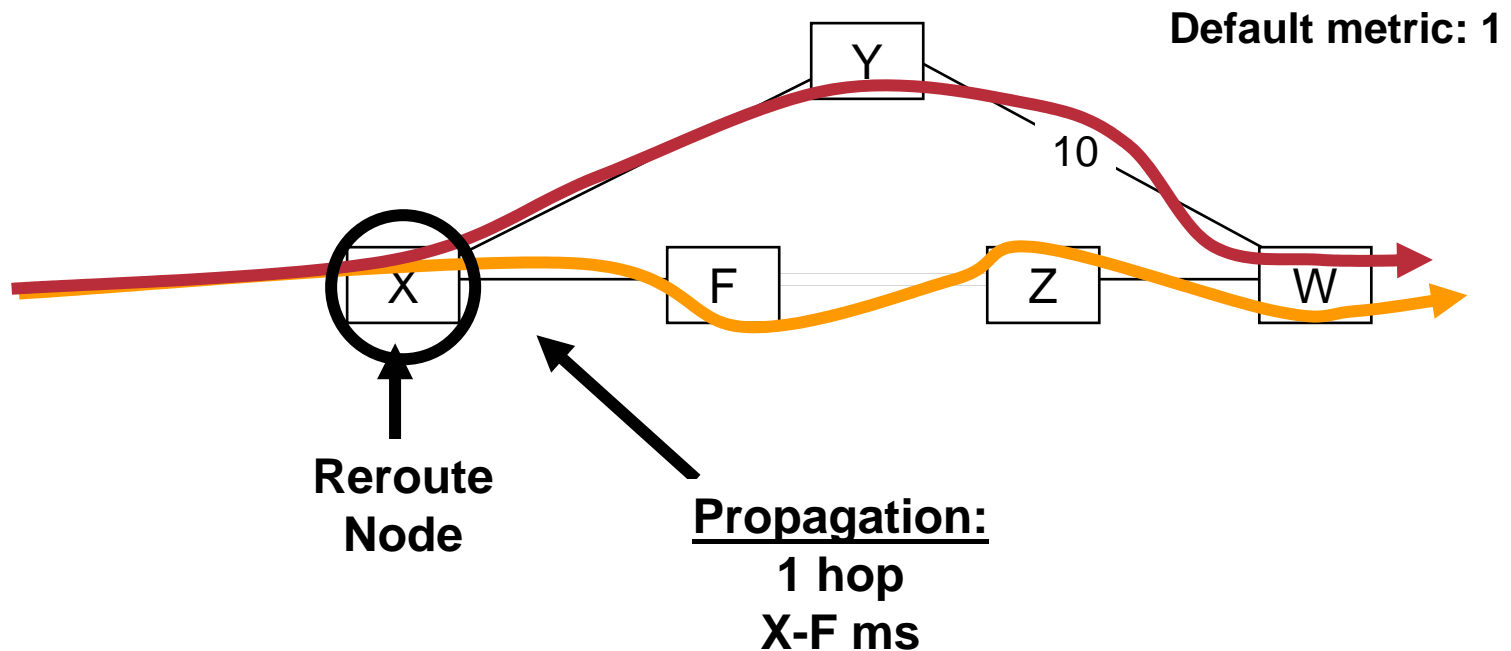
- **R: point where the old and new paths diverge**

  this is a worst-case estimation of P!

P: Propagation in number of hops
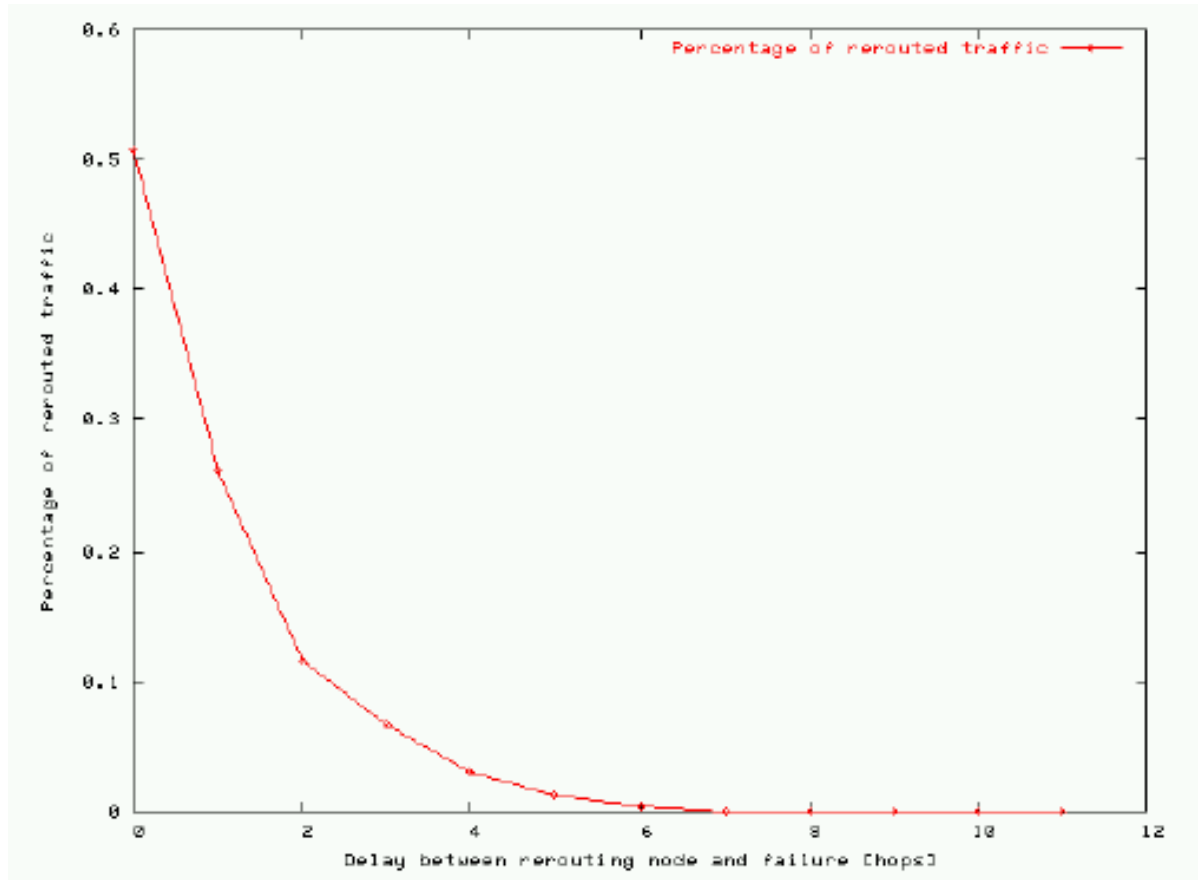
- **Worldwide ISP with traffic matrix – summary for the failures of the 340 most loaded links. Pessimistic definition of R**

P: Propagation in ms (light speed)

- **Worldwide ISP with traffic matrix – summary for the failures of the 340 most loaded links. Pessimistic definition of R**

Conclusion

- **Sub-Second objective is realistic**

    Conservative

- **Technology has significantly improved**

CISCO SYSTEMS

Why is it possible?

13

# Components contributing to loss of connectivity

- **D: Failure is detected**

- **O: New LSP is originated**

- **QSP: cumulative queueing, serialization, propagation**

- **h*F: LSP is flooded up to rerouting node**

- **SPT: SPT is updated**

- **RIB: RIB/FIB is updated**

- **DD: LCs are updated**

- **BGP recursion is fixed**

$$LoC = D + O + QSP + (h * F) + SPF(n) + Rib(p) + DD + CRR$$

D: POS – excellent for Convergence

- **Very fast Link Failure detection**

  no need for fast IGP hello's

- **Various timers to order protection techniques**

  SONET/SDH optical protection

- **Native anti-flap property**

  Down info is signalled very fast

  Up info is confirmed for 10s before relaying to interface

D

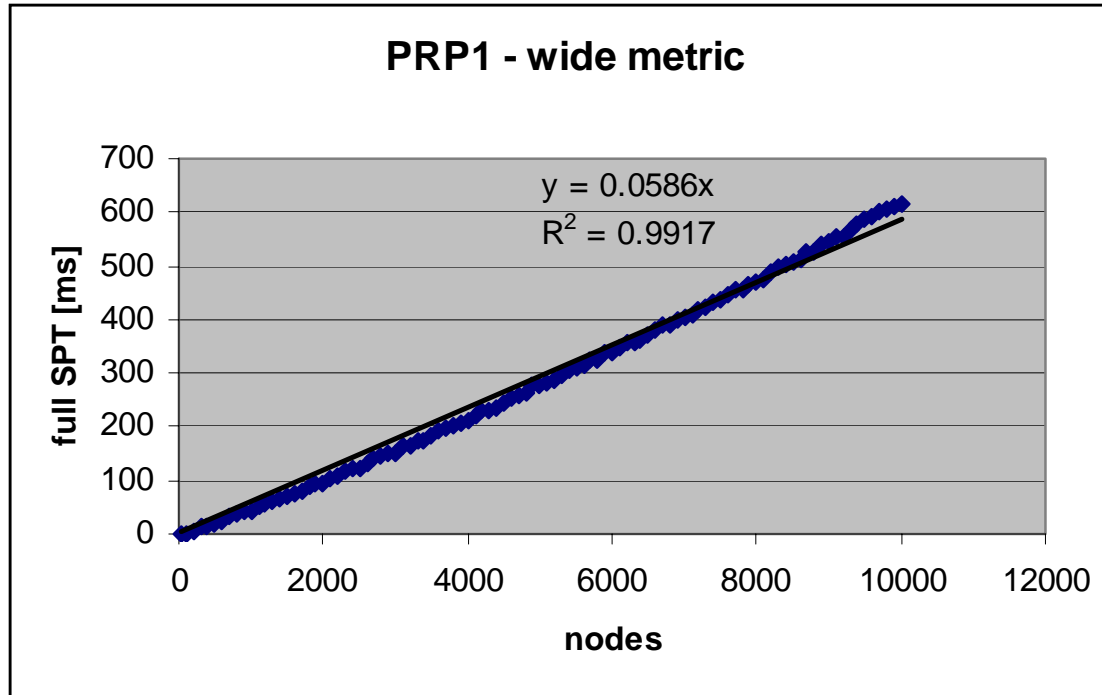**D - ENG2/POS: SNMP load, 8 BGP flaps/s**
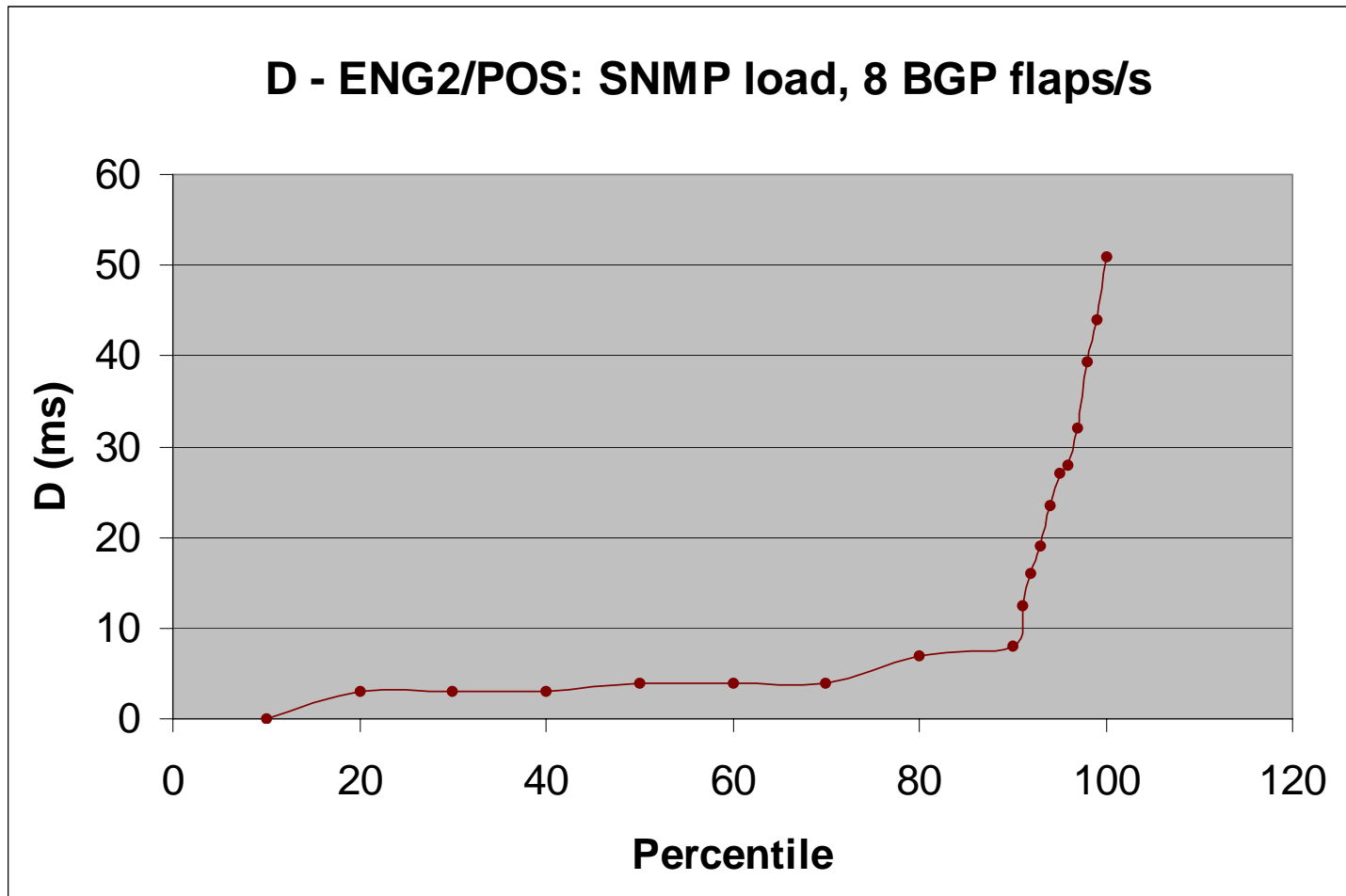
Probability of the worst-case

- **D: there are two detection points**

    Worst-case must occur at the same time at both points

- **F: there are many flooding paths**

    Worst-case must occur at each hop for the same LSP along all possible paths

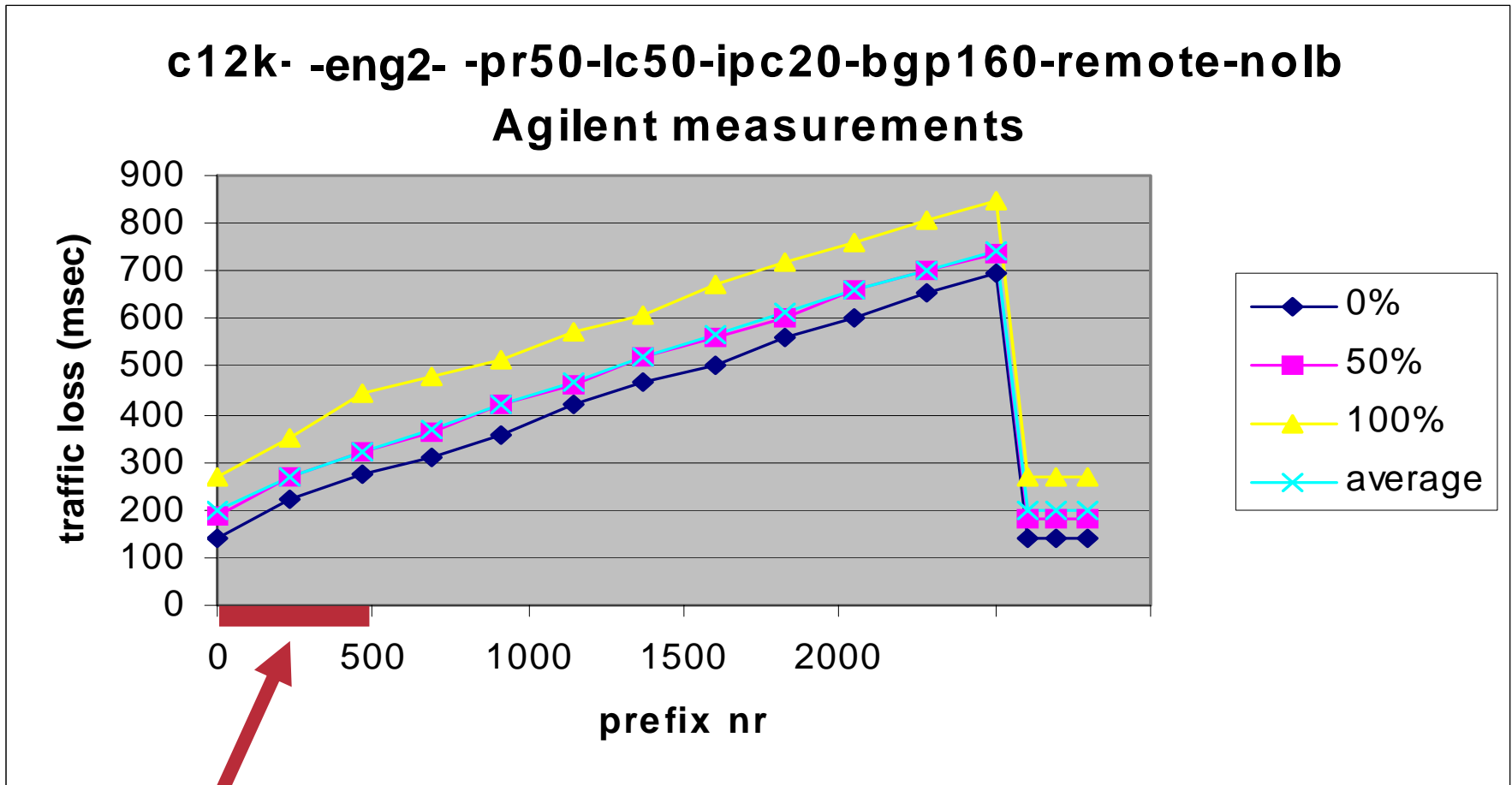- **Probability of worst case is unlikely**

SPT computation

PRP1 - wide metric

$y = 0.0586x$
$R^2 = 0.9917$

full SPT [ms] vs nodes

- **Full SPT (wide metric): 600 nodes => 35 ms**

- **Incremental-SPF benefits come on top of this**

  roughly: only the nodes impacted by the failure do matter as opposed to all the nodes of the topology for a 'normal' SPF

RIB update – prioritized update

c12k- -eng2- -pr50-lc50-ipc20-bgp160-remote-nolb
Agilent measurements

Conclusion

- **Sub-Second objective is realistic**

  Conservative

- **Technology has significantly improved**