IPFRR and the Principle of Simplicity

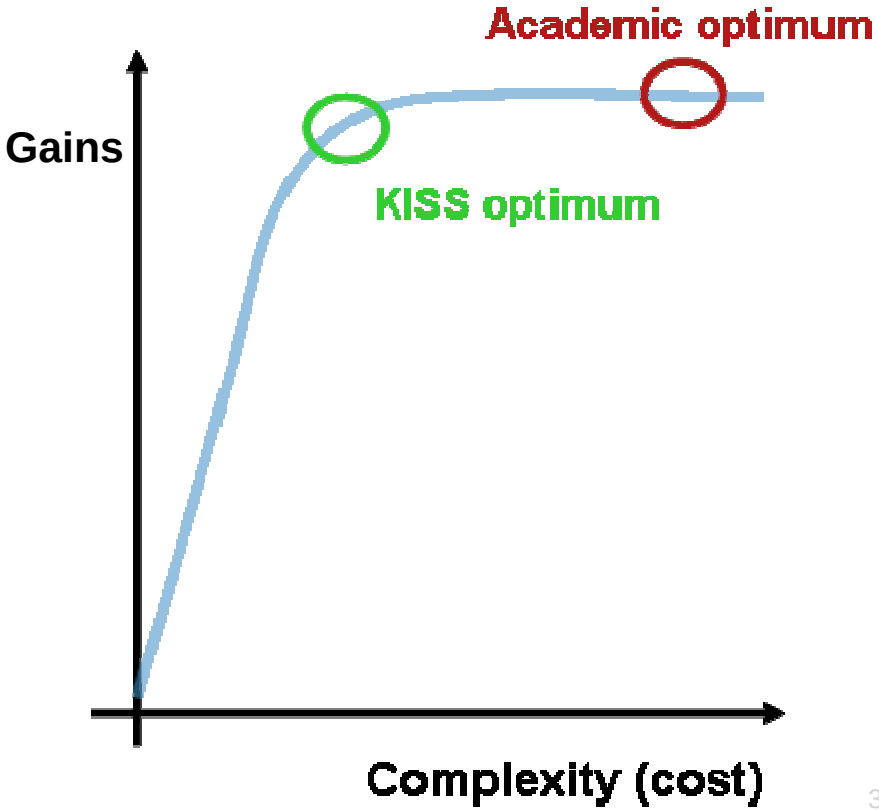
Clarence Filsfils and Pierre Francois

Plan of the talk

- Principle of simplicity
- Terminology
- Requirement for routing convergence
- A fast IGP convergence is always needed.
  - It is the foundation for all routed services.
- Kiss suggests three low-hanging fruits
  - LossLess Local Maintenance
  - IPFRR Per-Link LFA
  - Offline computation for uloop avoidance upon maintenance
- Conclusion

Principle of Simplicity

- “Simplicity is prerequisite for reliability”
  Edsger Dijkstra
- "Simplicity is the ultimate sophistication"
  Leonardo da Vinci
- Kiss: Keep It Simple Straighforward



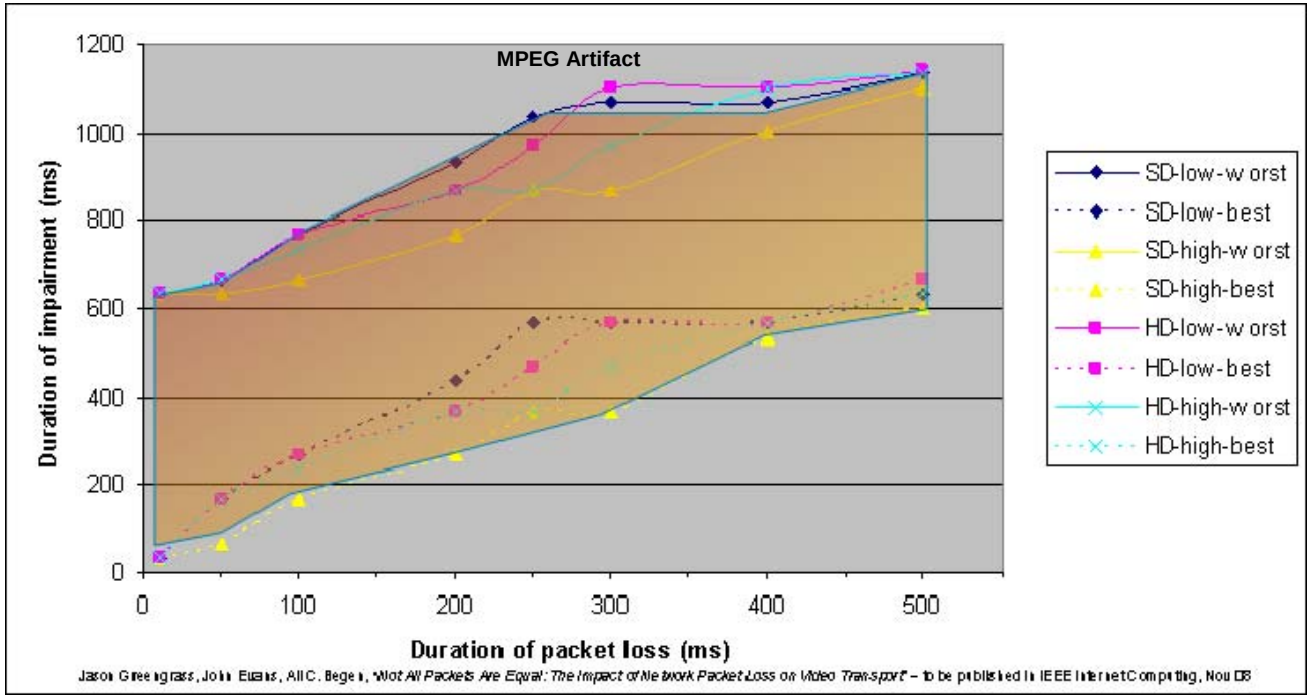
Terminology

- IGP convergence
  - upon managed or unmanaged link or node failures
  - we consider all failures except edge link failures (BGP-related, see PIC presentation at Nanog)
- Local Protection
  - pre-compute and pre-install alternate path
  - no need for remote nodes to know about the failure
- End of Loss of Connectivity
  - time when the data is routed around the failure and connectivity is re-established. The alternate path might not be optimal yet

Terminology

- uLoop
  - during IGP convergence, some routers may have inconsistent tables which create loops. These states are transient hence the term MicroLoop (uLoop)
- uLoop Avoidance
  - a technique to avoid uloop

Requirement – Perception vs Reality – Influence on Complexity



- **< 1 or 2 second: human does not bother**
- **< 200msec: human does not notice**
- **< 50msec: human has the perception of being better off**
- **Video Industry requirement: MTBA >= 2 hours**
  - **MTBA due to core failures > 100 hours**

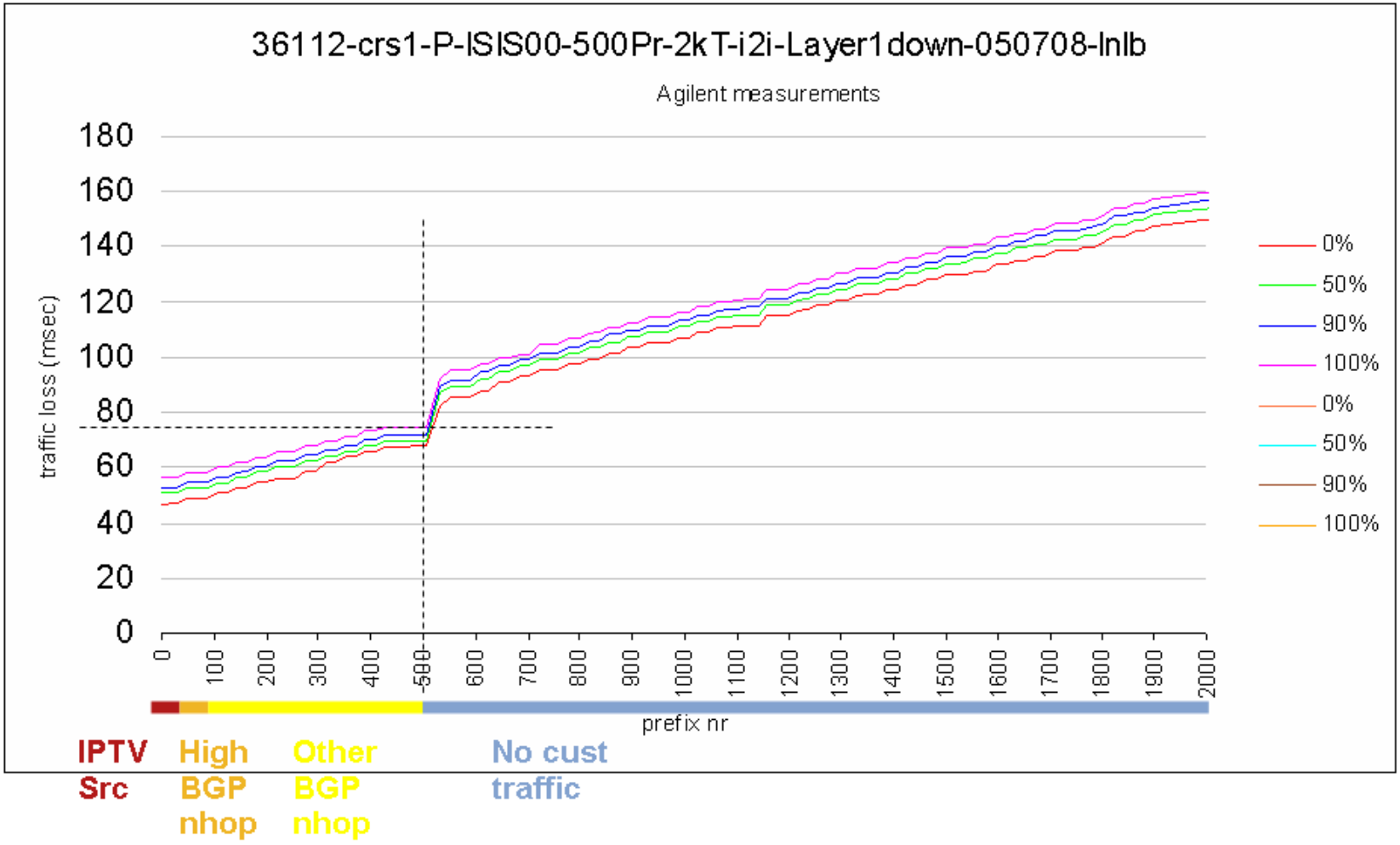
IGP convergence

The foundation

IGP convergence always matters

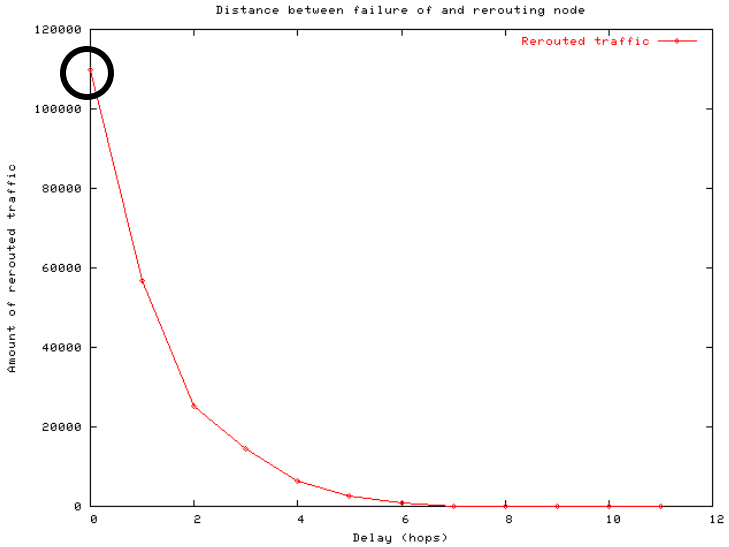
- BGP nhop availability
- PIM source availability
- MPLS TE topology and resource information
  - CAC upon reroute (!)
- Unplanned protection
  - Most MPLS FRR designs only protect link
  - Unknown SRLG
- Catastrophic event

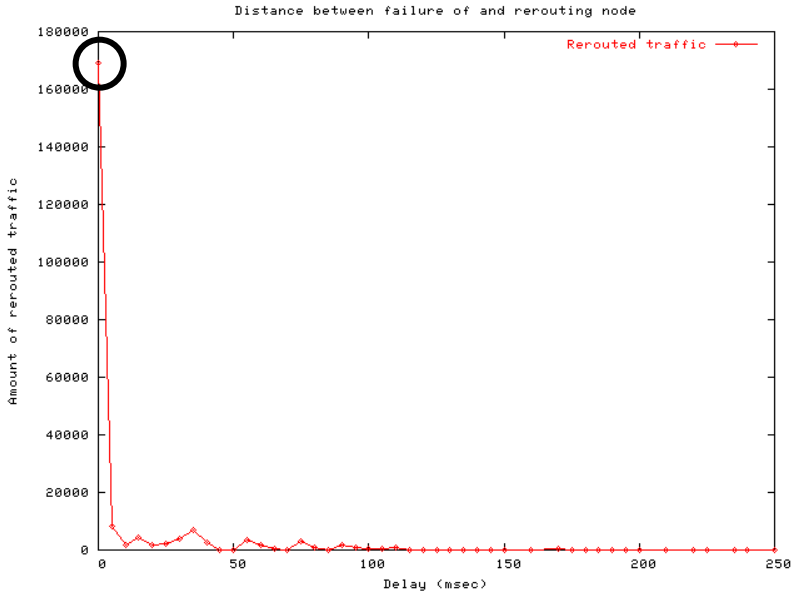
IGP Convergence – Local Failure



No tuning required

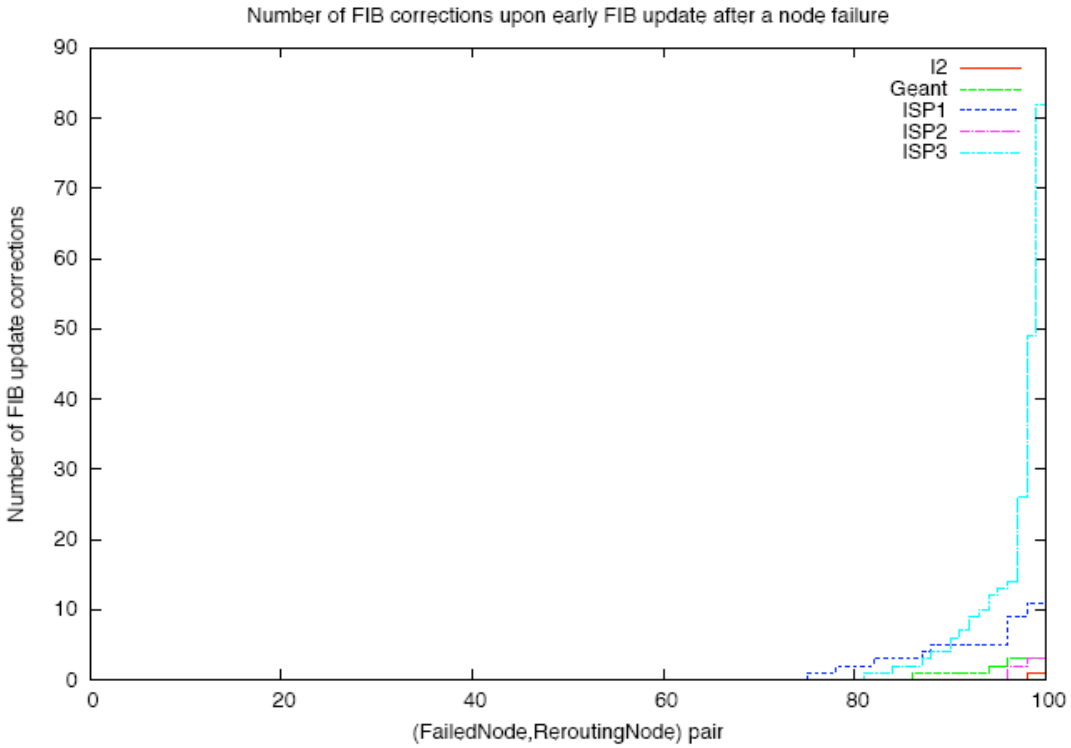
IGP Convergence - Flooding





- **A link failure within a continent very rarely requires a rerouting in a different continent. Propagation is thus bounded by 25msec (5000km of fiber)**
- **It is very rare for a failure to require rerouting further than 5 hops away from the failure. Flooding is thus bounded by 5*5msec.**
- **Intuitively, this rule is expected: designers build networks with resilience in mind**

IGP Convergence – Node Failure & Flood



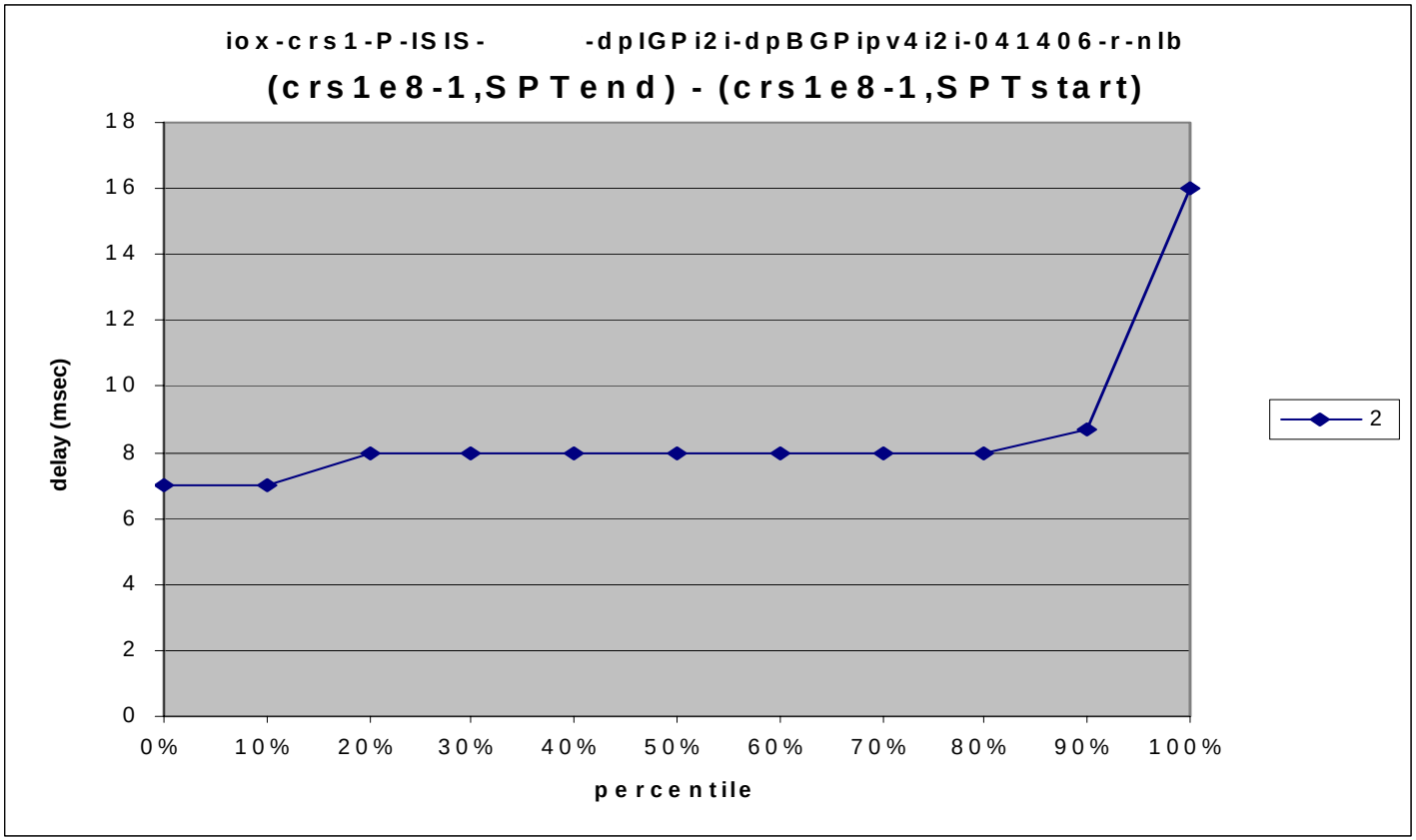
X-Axis : pairs of (node failure, rerouting node for that failure). All rerouting nodes are plotted for each node failures

Y-Axis : The number of prefixes that got updated twice because the first one considering the link failure was wrong

Note: it doesn't account the FIB updates of the second run that were not performed in the first one. This is not “bad”.

- **Aggressive Link-Oriented IGP convergence very rarely incurs overhead in case of node failures**
- **Intuitively, this is expected: very often, once you divert due to a link failure, you completely avoid the previous path and hence avoid the node**

IGP Convergence – SPT computation



- 900-router ISIS network… without leveraging i-SPT
  - with iSPT, most runs under 1msec

Low Hanging Fruit

LossLess Local Maintenance

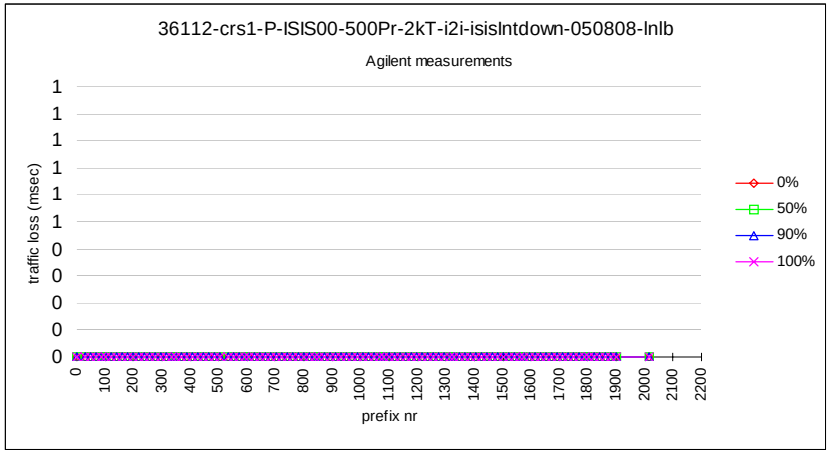
LossLess Local Maintenance

- Before shutting a link, shut the adjacency

HW FIB keeps on forwarding on the link while it is being modified to use the new best path.

Zero local loss instead of 100's of msec of loss.

Better than costing to infinity

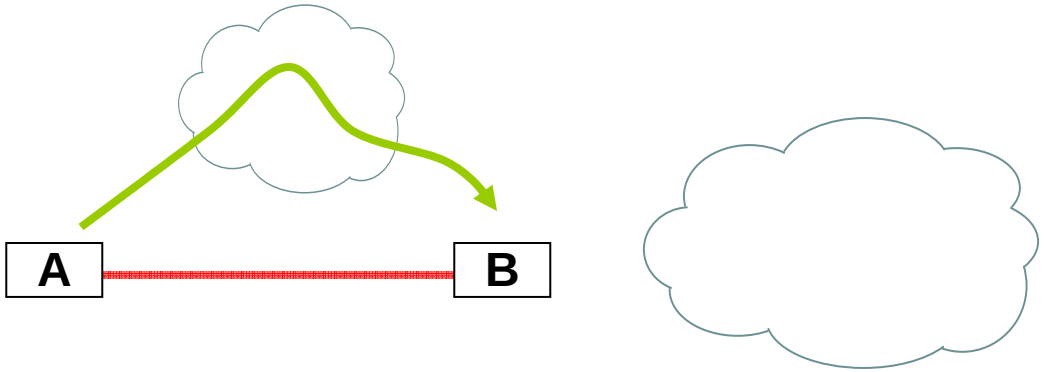



```
RP/0/RP1/CPU0:crs1e8-1(config)#router isis 1
RP/0/RP1/CPU0:crs1e8-1(config-isis)#int GigabitEthernet0/6/0/4
RP/0/RP1/CPU0:crs1e8-1(config-isis-if)#shut
RP/0/RP1/CPU0:crs1e8-1(config-isis-if)#commit
```

Low Hanging Fruit

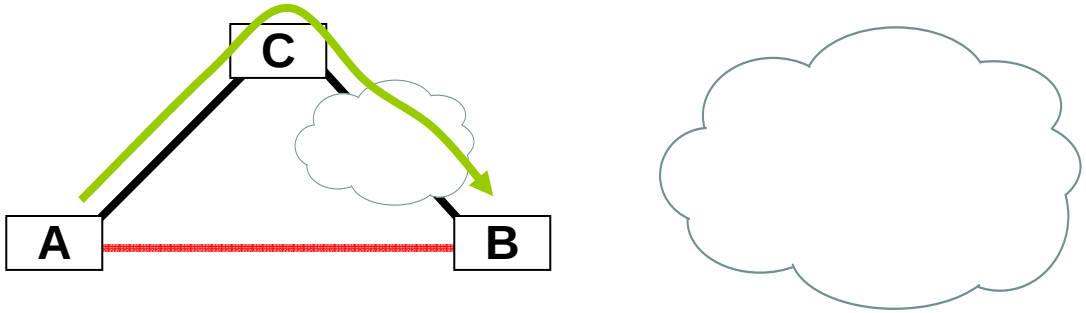
Per-Link LFA

Per-Link Property



- If A finds an alternate path to B
- Then this alternate path is valid for any destinations that A normally routes via B

Per-Link LFA

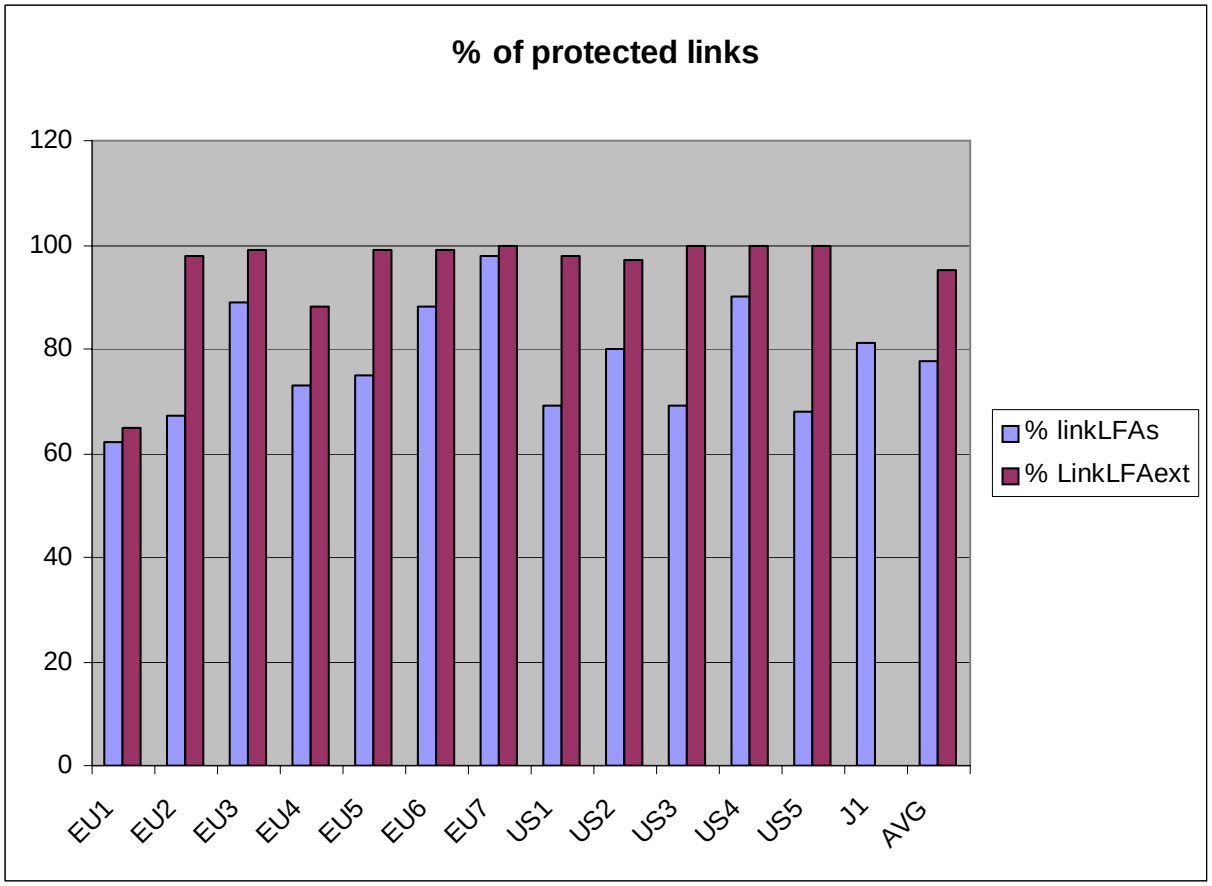


- To protect AB, A may reroute packets via C
  - if, what ever link AB status, the IGP route from C to B avoids AB

Per-Link LFA - Properties

- Automated
- No IETF protocol change
  - all the needed info is already in classical LSDB
- Incremental deployment
- No inter-operability testing
- <25msec, prefix-independent
- Applicable to MPLS LDP networks

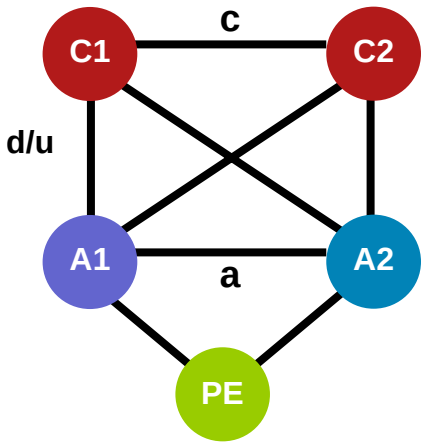
Per-Link LFA – Core Coverage

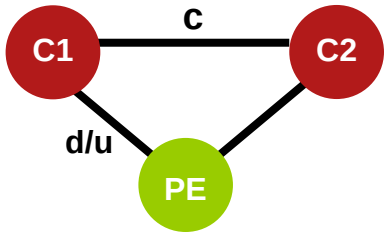


- 75% core events: <25msec (Per-Link LFA)
- 25% core events: <x00msec (IGP Convergence)

Per-Link LFA – PoP coverage

- If c < d + u
- All intra-pop links are protected with Per-Link LFA and no uloop is possible for
  - PE's dual-homed to two core routers
  - PE's dual-homed to two aggregation routers. Aggregation routers are squared&crossed to core routers.

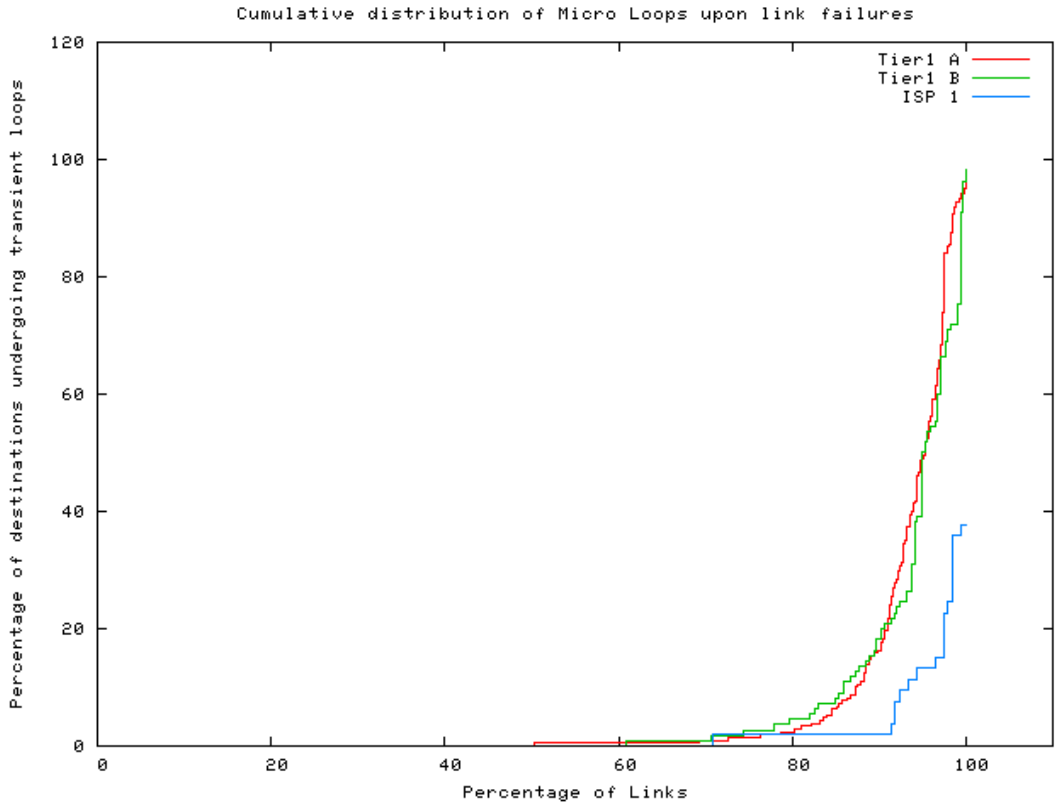




Low/Mid Hanging Fruit

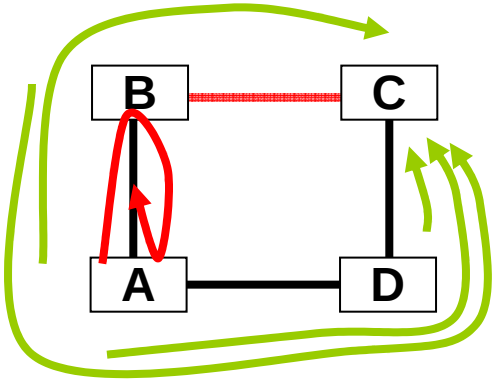
Offline ULoop Avoidance for Maintenance

Is a uLoop possible, topologically?

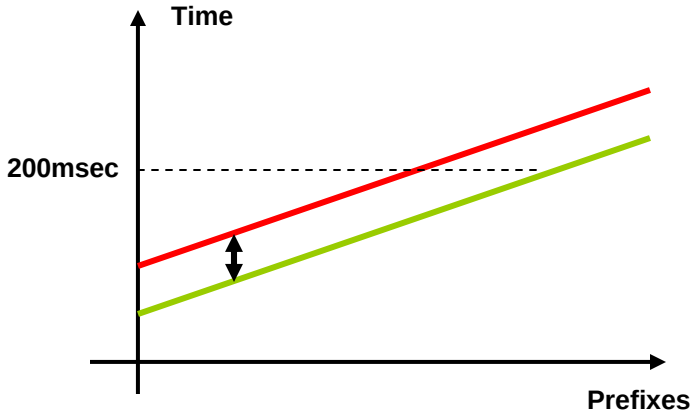


- **Input: Topology**
- **Output: for each link, the proportion of destination which could be subject to a uloop. The occurrence is not certain as it also requires a specific timing pattern of convergence.**

If it occurs, how long does it last?

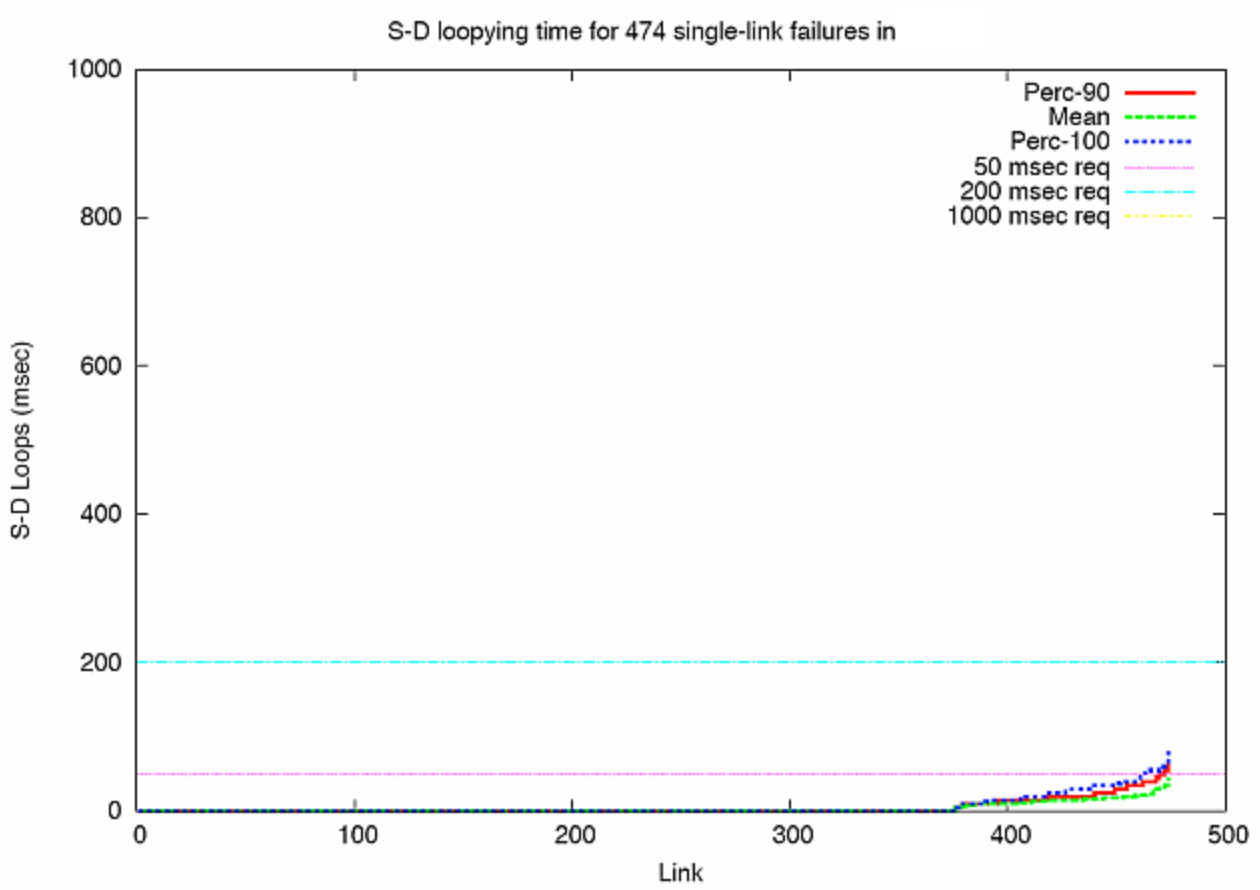


Possible transient state when B already updated the FIB for BC failure while A has not



- **A uloop lasts as long as the difference of convergence between the looped nodes**
- **If IGP converges in ~ x00msec, the uloop lasts much less than x00msec**

Simulation



- **Simulation on Tier1 ISP (900 nodes, single ISIS level) leveraging CRS/IOX3.6 characterization. The result should be a very negative worst-case as 10 /32's where originated by each node, for a total of 9000 important prefixes to converge.**

Kiss Analysis

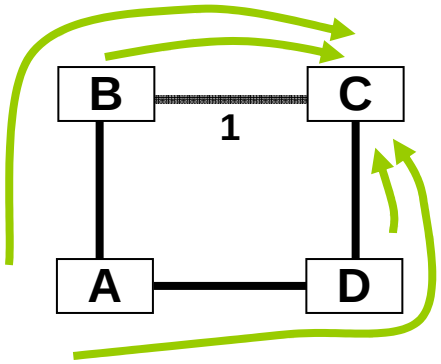
- uLoop, how important?
- if frequent enough and long enough, KISS suggests
  - review IGP convergence status (is it optimized?)
  - low hanging fruit: offline algorithm for maintenance

    No router HW/SW requirement

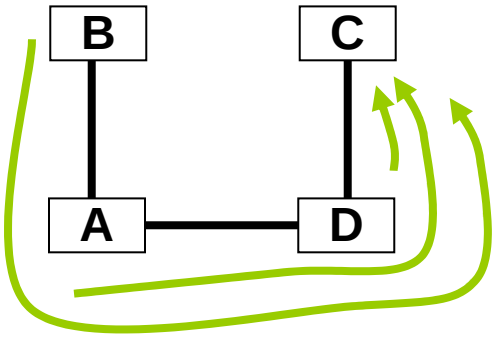
    No IETF dependency

    No feature deployment across network

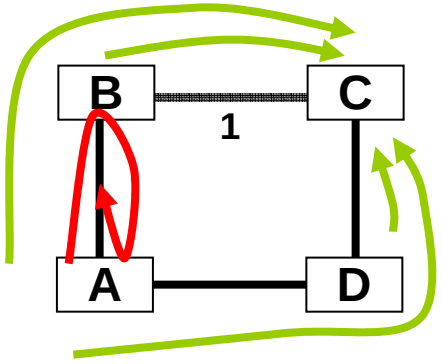
uLoop upon Maintenance



Initial State



New state for B



Possible transient state for A

Default Link Metric = 1

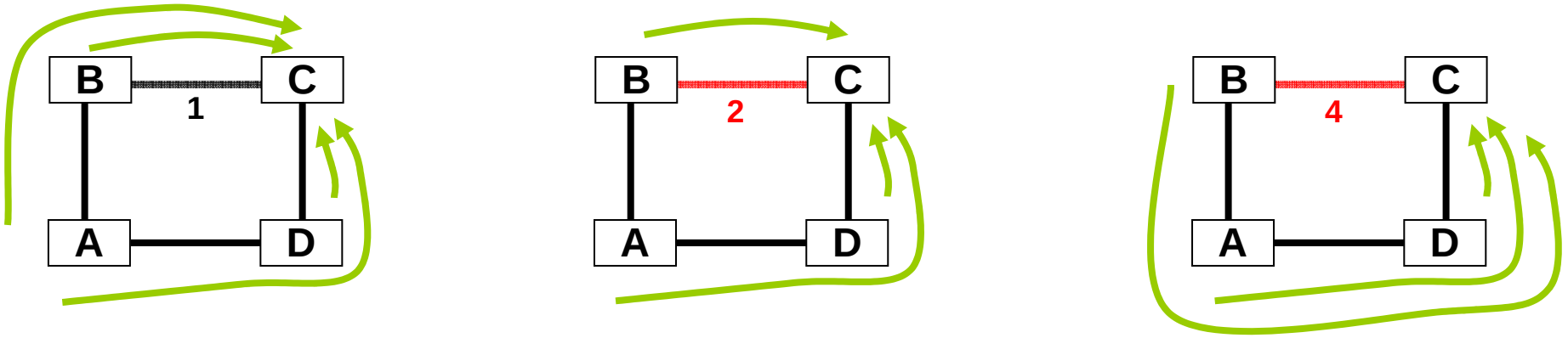
- If one disables the BC adjacency (to eventually shut the link)
  - B switches over without loss
  - But A may still route to C via B and hence a transient loop occurs
- MicroLoop (uLoop)
  - transient loop resulting from the distributed nature of the IGP conv.

Offline ULoop Avoidance for Maintenance

- Computation
  - offline
- Input
  - topology
  - new target metric T for Link L (current metric C)
- Output
  - a set of k interim metrics (M1, M2, Mk) such that any transition from Mi to Mi+1 is loop-free

[Francois07]: “Disruption-free topology reconfiguration in OSPF Networks. IEEE INFOCOM 2007, Pierre François, Mike Shand and Olivier Bonaventure.

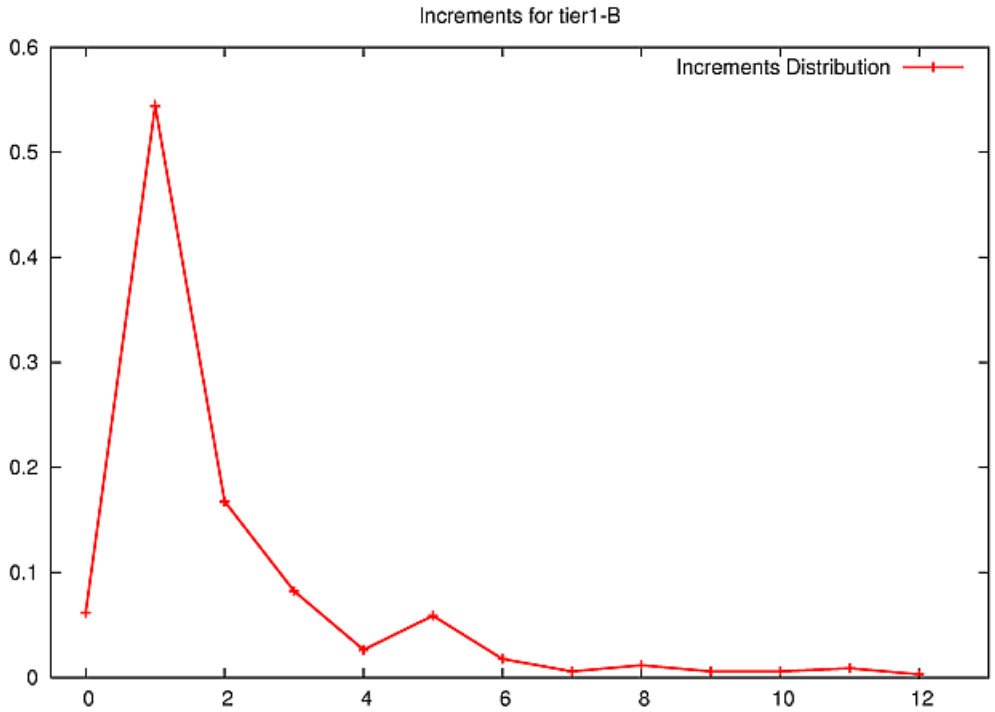
IPFRR – uLoop Avoidance Incremental Metric



- To bring BC out of service without uloop: change the metric to 2 then 4 and finally shut the link.

Default Link Metric = 1

IPFRR – uLoop Avoidance
Incremental Metric



- Sequence of metrics are usually short
- KISS hint: some steps might be avoided if they avoid uloop for a very small % of the traffic matrix

Conclusion

Conclusion

- Requirements influence Complexity and Cost
- Simplicity is prerequisite for reliability, Dijkstra
- IGP Convergence is the always-needed foundation
  - ~ x00 msec is feasible
- Three low-hanging fruits to consider: in order
  - LossLess Local Maintenance
  - Per-Link LFA Protection
  - Offline uloop avoidance for Maintenance
- No religion
  - Per-Link LFA in the PoP, MPLS TE Full-Mesh in the core
  - Per-Link LFA applies to IGP/LDP
  - Do not re-invent the wheel
- Use tools

A combination to consider

- MPLS FRR Full-Mesh in the core
  - Full Mesh of TE tunnels from PoP to PoP
  - 50msec protection without uloop for all core events
  - smaller full-mesh, less complex
- Per-Link LFA for IGP/LDP in the PoP
  - automated per-link LFA for all intra-PoP links
  - 50msec protection without uloop for all pop events
  - no mesh in the PoP, less complex
- Fast IGP Convergence in x00msec
  - BGP nhop and PIM source availability, MPLS TE CaC upon reroute, Unplanned protection, unknown SRLG, Catastrophic event

Tools

- We have worked with cariden to add basic IGP convergence and IPFRR functionality
- We have developped more-detailed tools (algorithmic topology analysis, simulation) to help with network design