

SPRINTLINK IP NETWORK

Peter Lothberg

<roll@sprint.net>

+1 703 864 7887



<http://radweb>

David Meyer (DMM)



SprintLink and the Philosophy of Building Large Networks

David Meyer

Chief Technologist/Senior Scientist

dmm@sprint.net

October 10, 2002



Agenda

- **Philosophy -- How We Build Networks**
- **SprintLink Architecture/Coverage**
- **What is all of this MPLS talk about?**
- **Putting it all Together**
 - Network Behavior in a Couple Failure Scenarios
- **IPv6**
- **Future**
- **Closing/Q&A**



Build Philosophy

- **Simplicity Principle**

- “Some Internet Architectural Guidelines and Philosophy”, draft-ymbk-arch-guidelines-05.txt

- **Use fiber plant**

- To efficiently provision robust paths
- “1:1 Protection Provisioning”

- **And remember that the job of the core is to move packets, not inspect or rewrite them.**

- Zero Drop, Speed-of-Light-like Latency, Low Jitter
- Side-effect of provisioning approach



Support Philosophy

- **Three S's**
 - Simple
 - NOC Staff can operate it
 - Sane
 - Don't have to be a PhD to understand and troubleshoot the routing
 - Supportable
 - If it takes twelve hours to figure out what's wrong, something isn't right..
- **If upgrading means re-thinking and redesigning the whole support process, something is likely broken**



Aside: System Complexity

- Complexity impedes efficient scaling, and hence is the primary driver behind both OPEX and CAPEX (Simplicity Principle)
- Complexity in systems such as the Internet derives from scale and from two well-known properties from non-linear systems theory:
 - *Amplification*
 - *Coupling*



Amplification Principle

- **In very large system, even small things can (and do) cause huge events**
 - Corollary: In large systems such as the Internet, even small perturbations on the input to a process can destabilize the system's output
 - Example: It has been shown that increased interconnectivity results in more complex and frequently slower BGP routing convergence



■ "The Impact of Internet Policy and Topology on Delayed Routing Convergence",

Labovitz et. Al, Infocom, 2002 <http://radweb>

10/10/2002

- Related: "What is the sound of One Route Flapping", Timothy Griffin, IPAM Workshop on Large Scale Communication Networks, March, 2002

Coupling Principle

- **As systems get larger, they often exhibit increased interdependence between components**
 - Corollary: The more events that simultaneously occur, the larger the likelihood that two or more will interact
 - *Unforeseen Feature Interaction*
 - “Robustness and the Internet: Design and Evolution”, Willinger et al.
- **Example: Slow start synchronization**



Example: The Myth of 5 Nines

- 80% of outages caused by people and process errors [SCOTT]. Implies that at best you have a 20% window in which to work on components
- In order to increase component reliability, we add complexity (optimization), effectively narrowing the 20% window
- i.e., in the quest for increased robustness, you increase the likelihood of people/process failures



Example: The Myth of 5 Nines

- The result is a *Complexity/Robustness Spiral*, in which increases in system complexity create further and more serious sensitivities, which in turn require additional robustness, ...
[WILLINGER2002]
- Keeping in mind that we can always do better...
- What does this say about all of the router HA work?

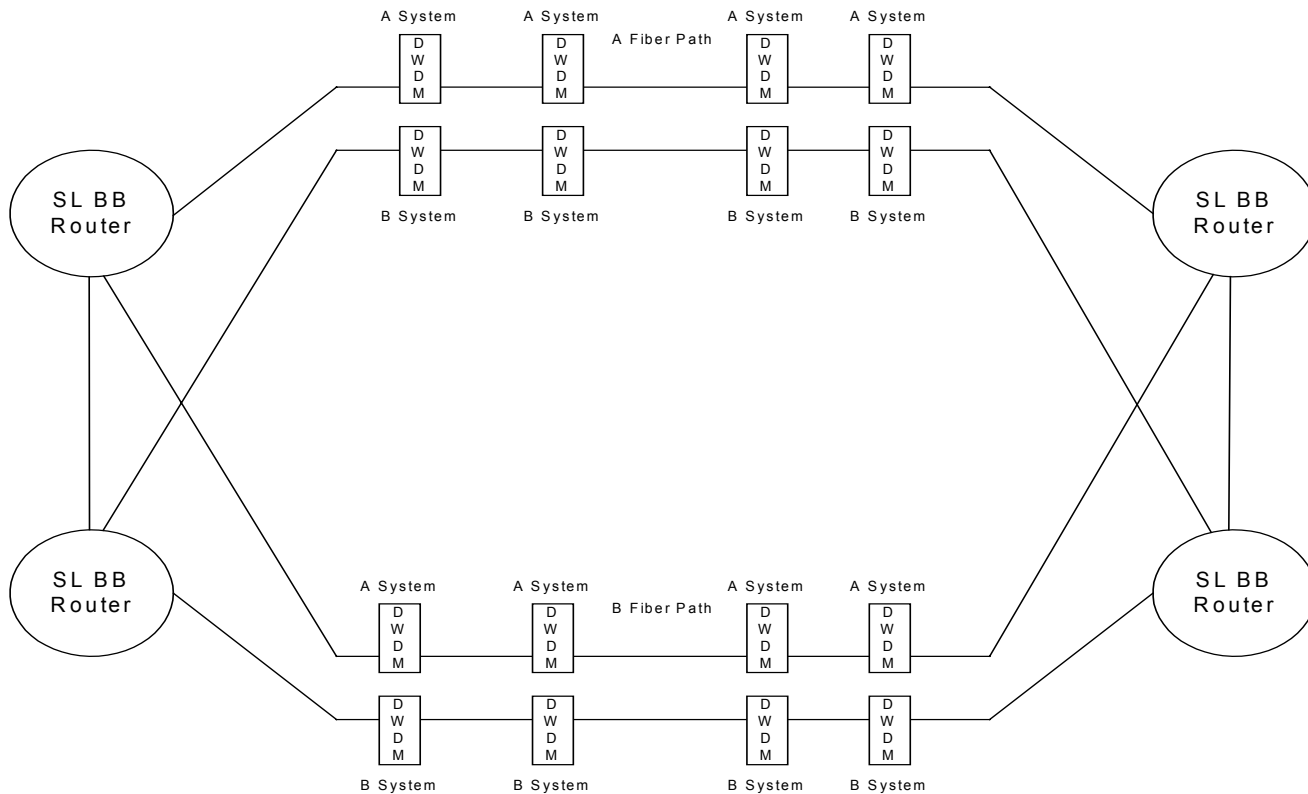


Aside: System Complexity

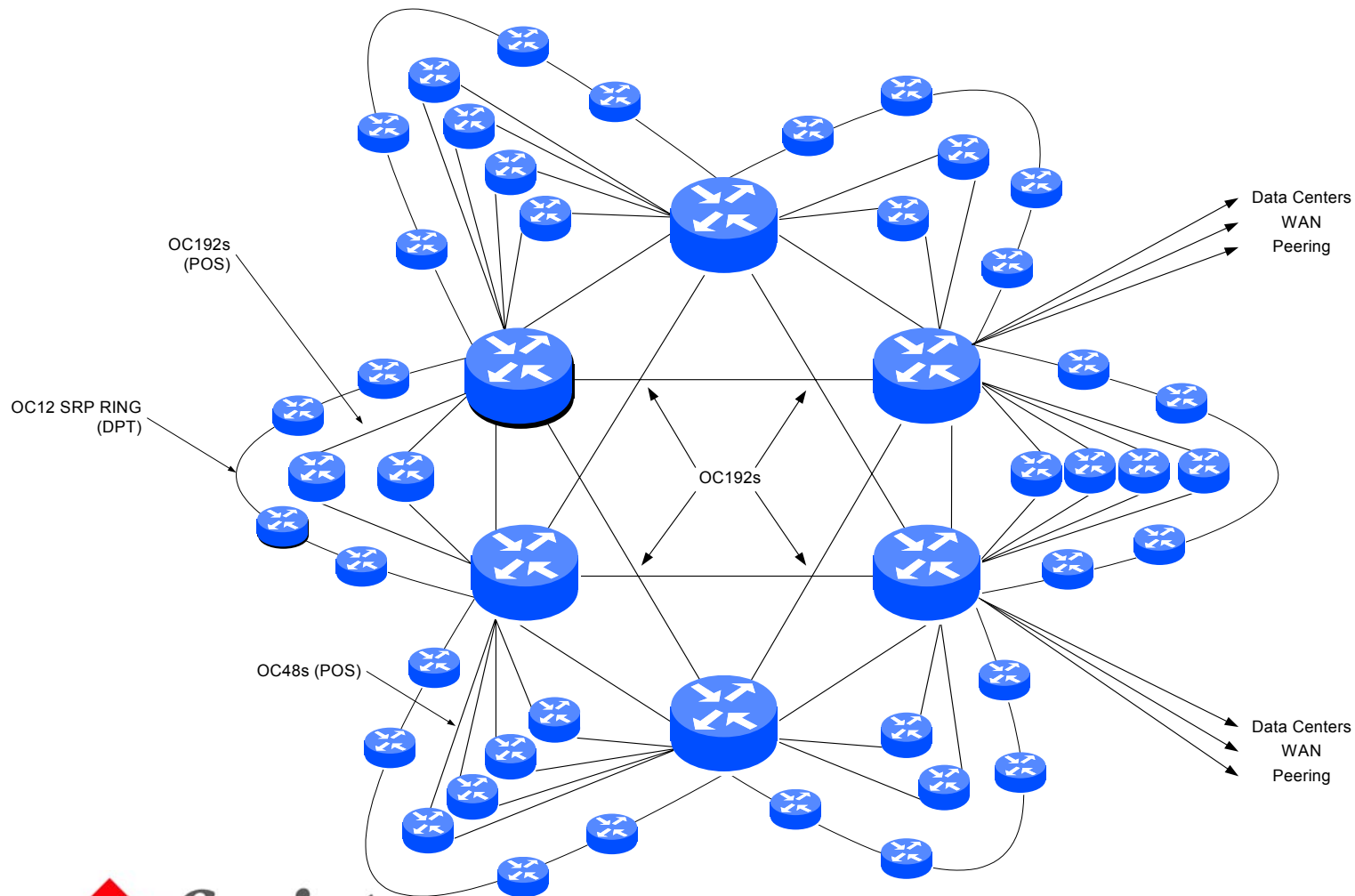
- **Bottom Line: We must manage complexity closely or complexity will quickly overwhelm all other facets of a system**
 - “Some Internet Architectural Guidelines and Philosophy”, Randy Bush and David Meyer, draft-ymbk-arch-guidelines-05.txt, August, 2002
 - Currently in the RFC editor’s queue
 - “Complexity and Robustness”, Carlson, et. al., Proceedings of the National Academy of Science, Vol. 99, Suppl. 1, February, 2002
- **See me if you’d like additional literature for your spare time :-)**



Physical Topology Principle



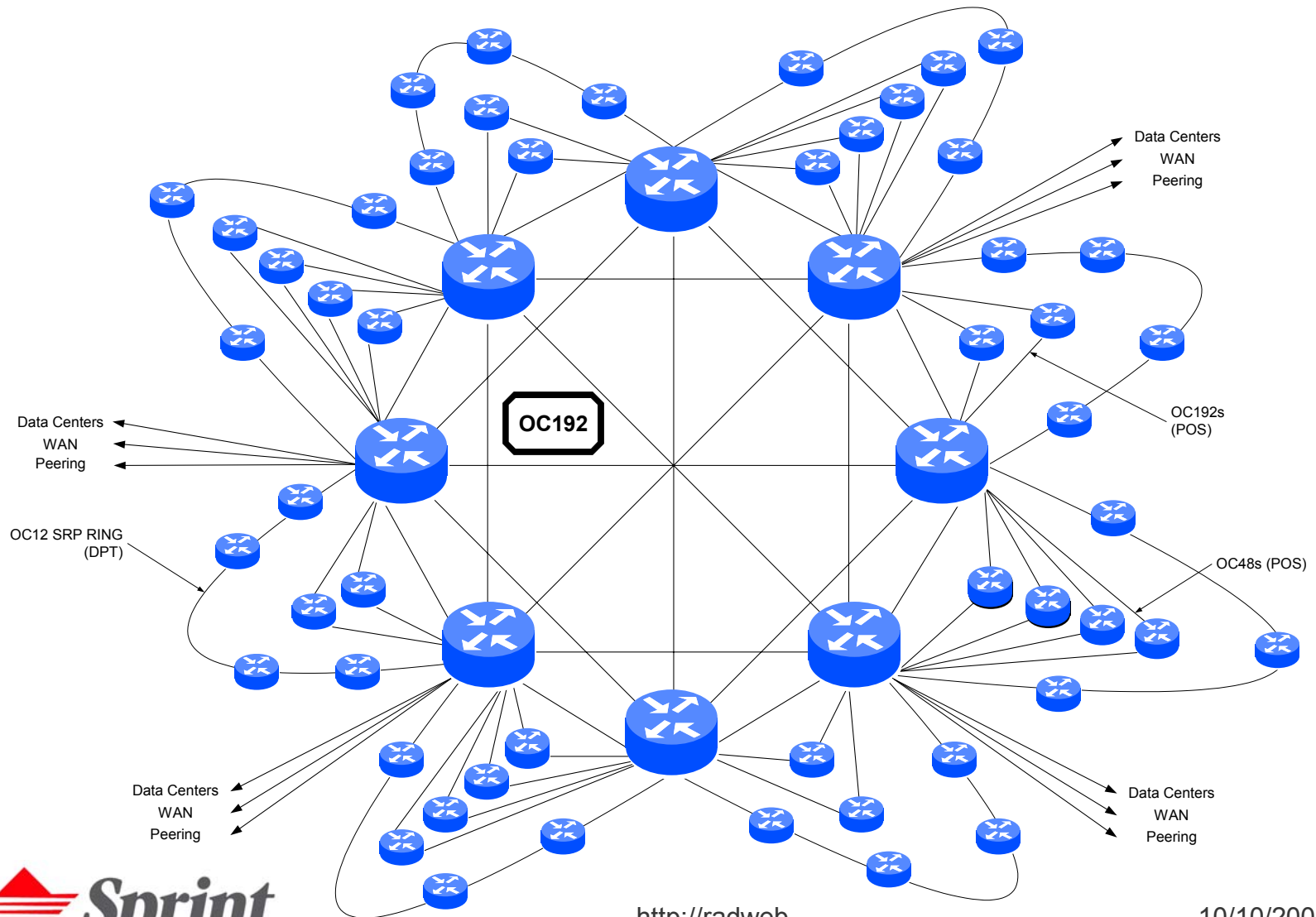
POP Design 2001 – 6 Core Routers



<http://radweb>

10/10/2002

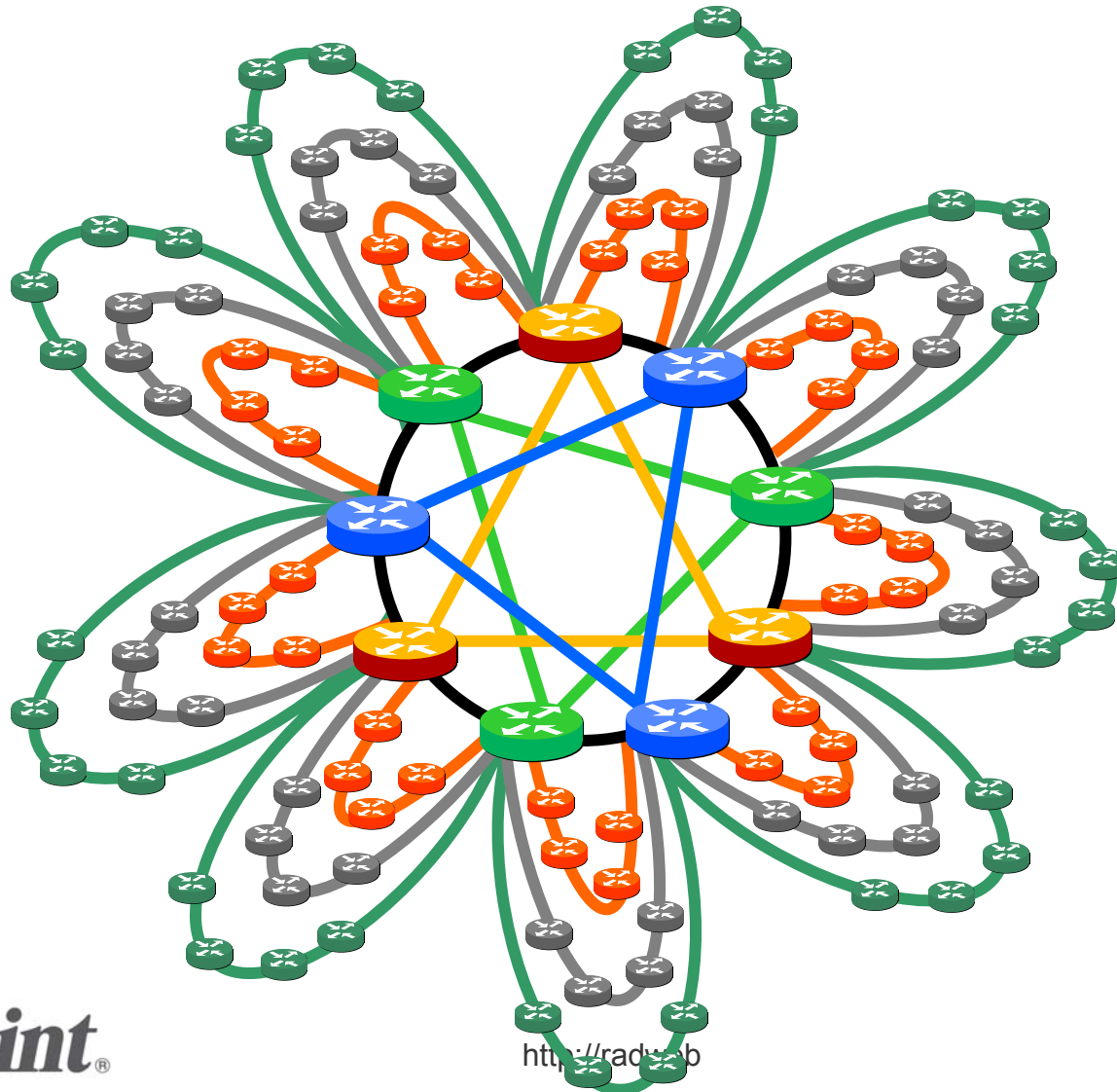
POP Design 2001 – 8 Core Routers



<http://radweb>

10/10/2002

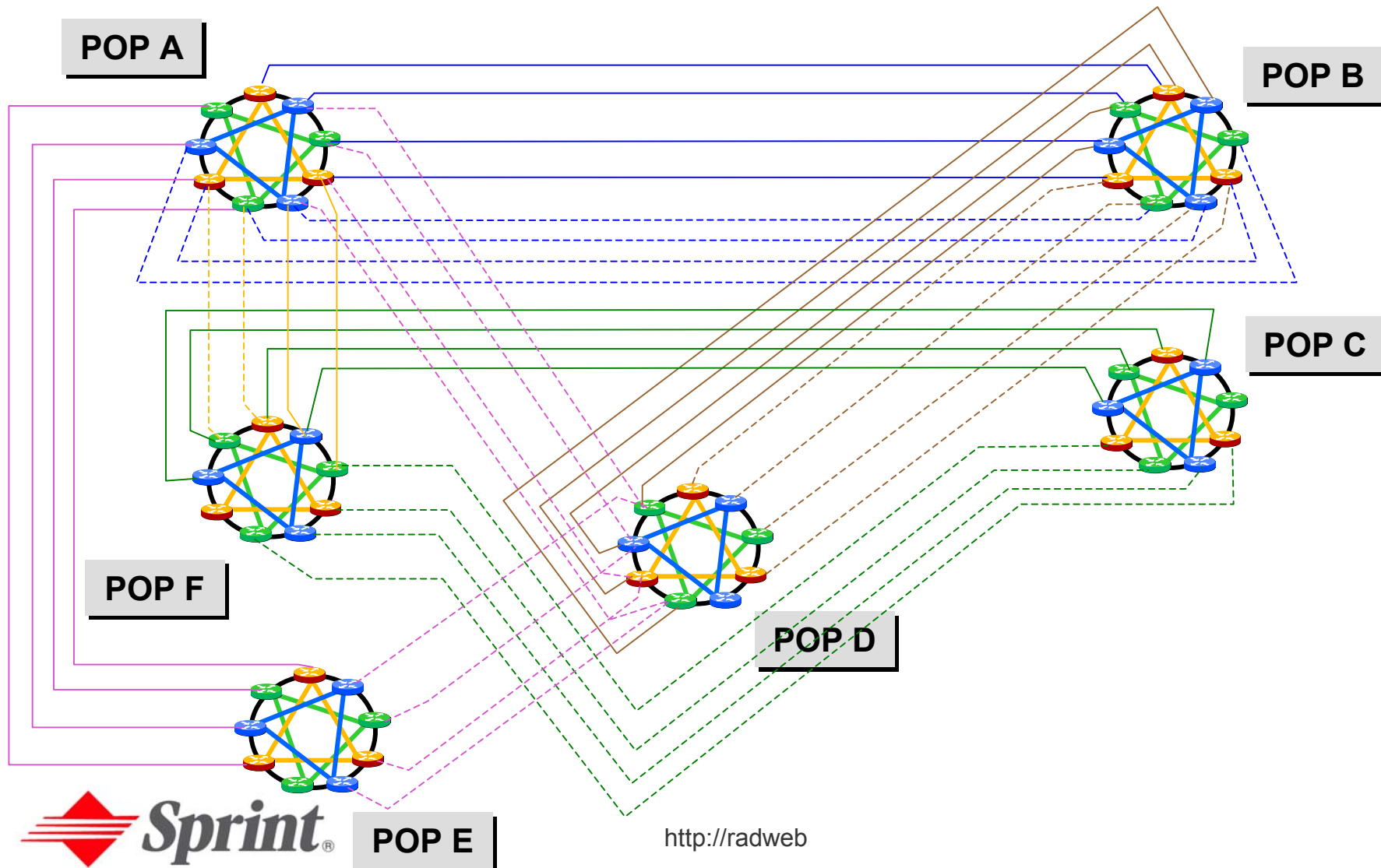
POP Design 2002 – 9 Core Routers



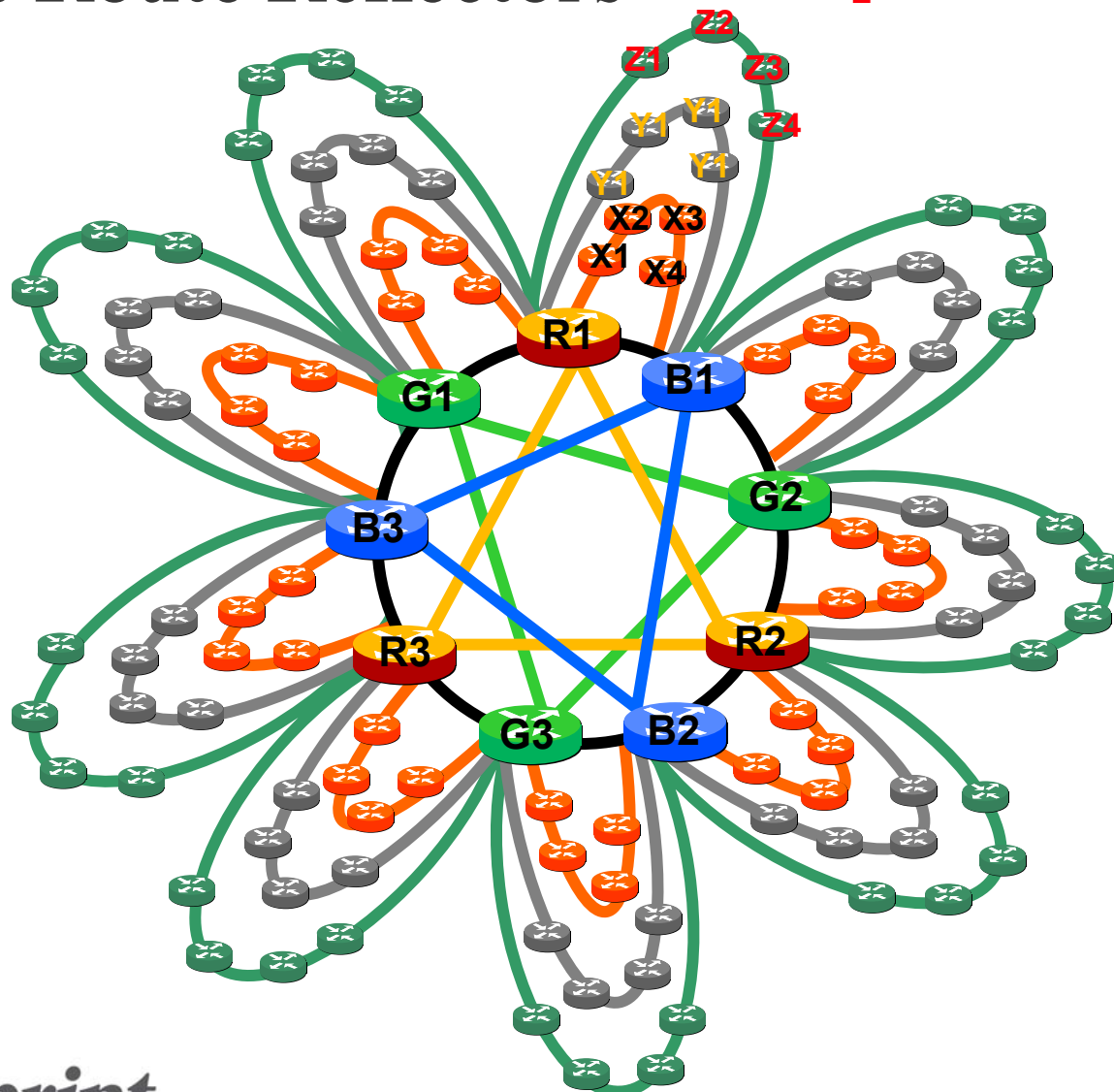
<http://radynb>

10/10/2002

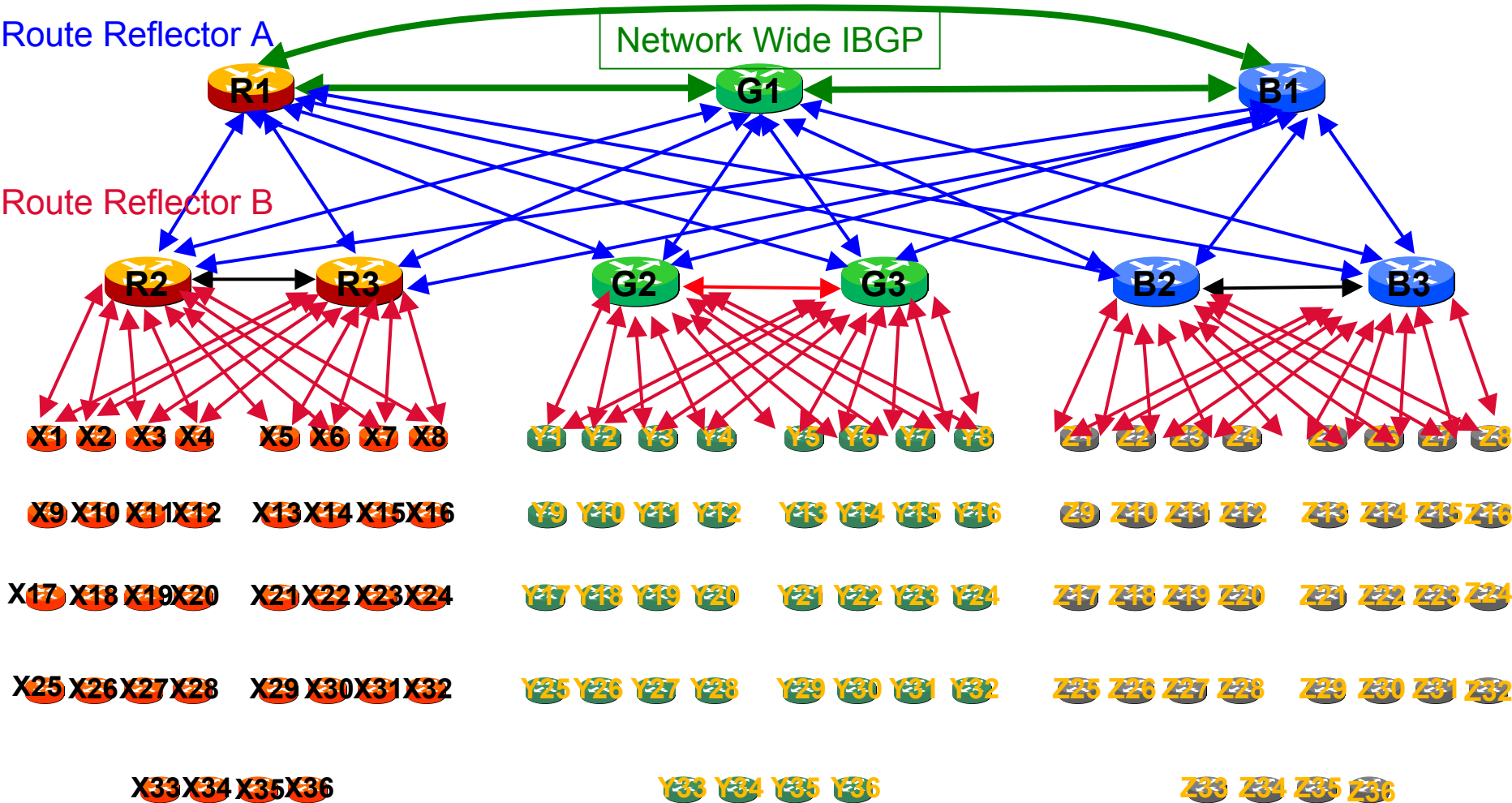
Backbone Principle



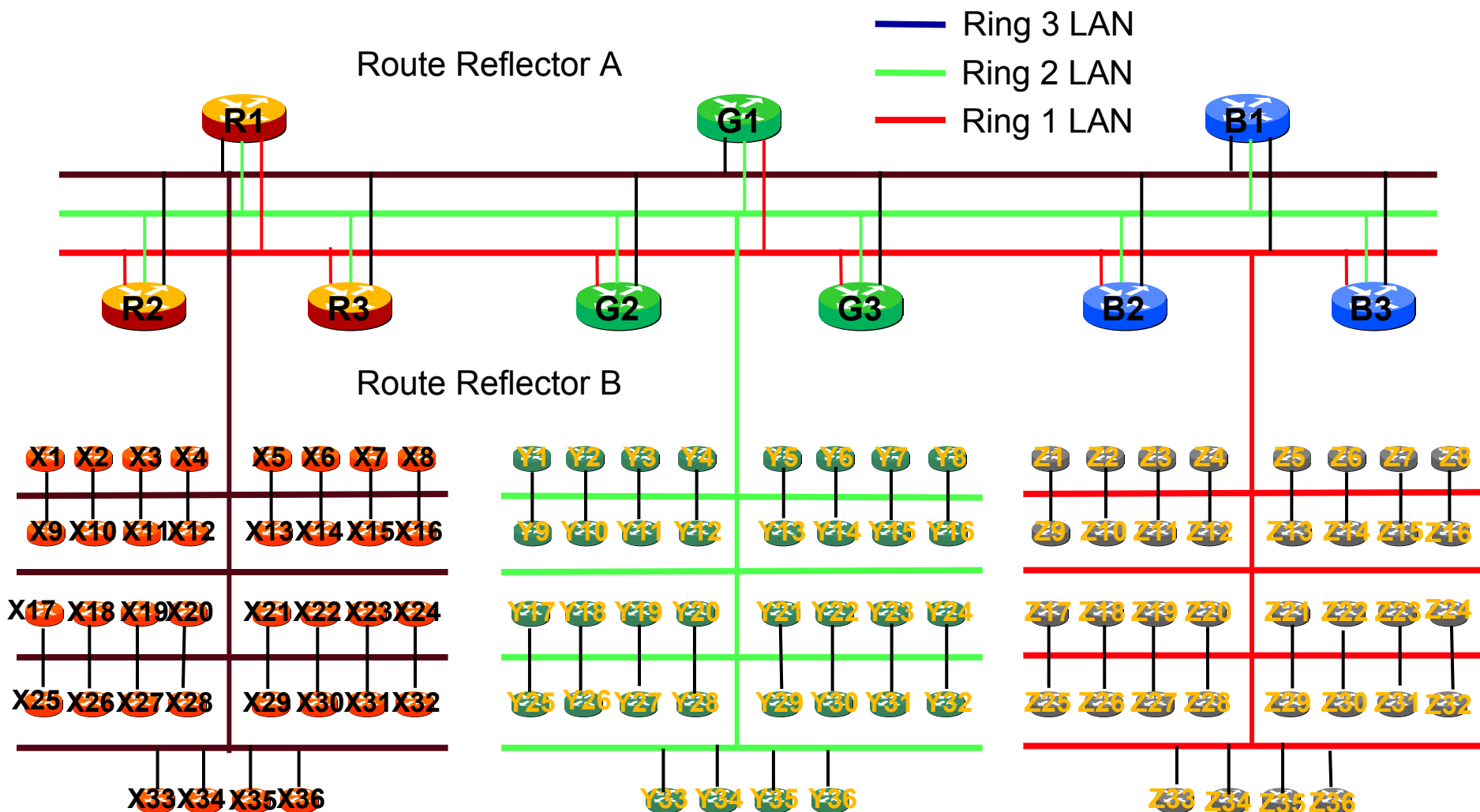
Backbone Route Reflectors



Global I-BGP



POP Logical Topology

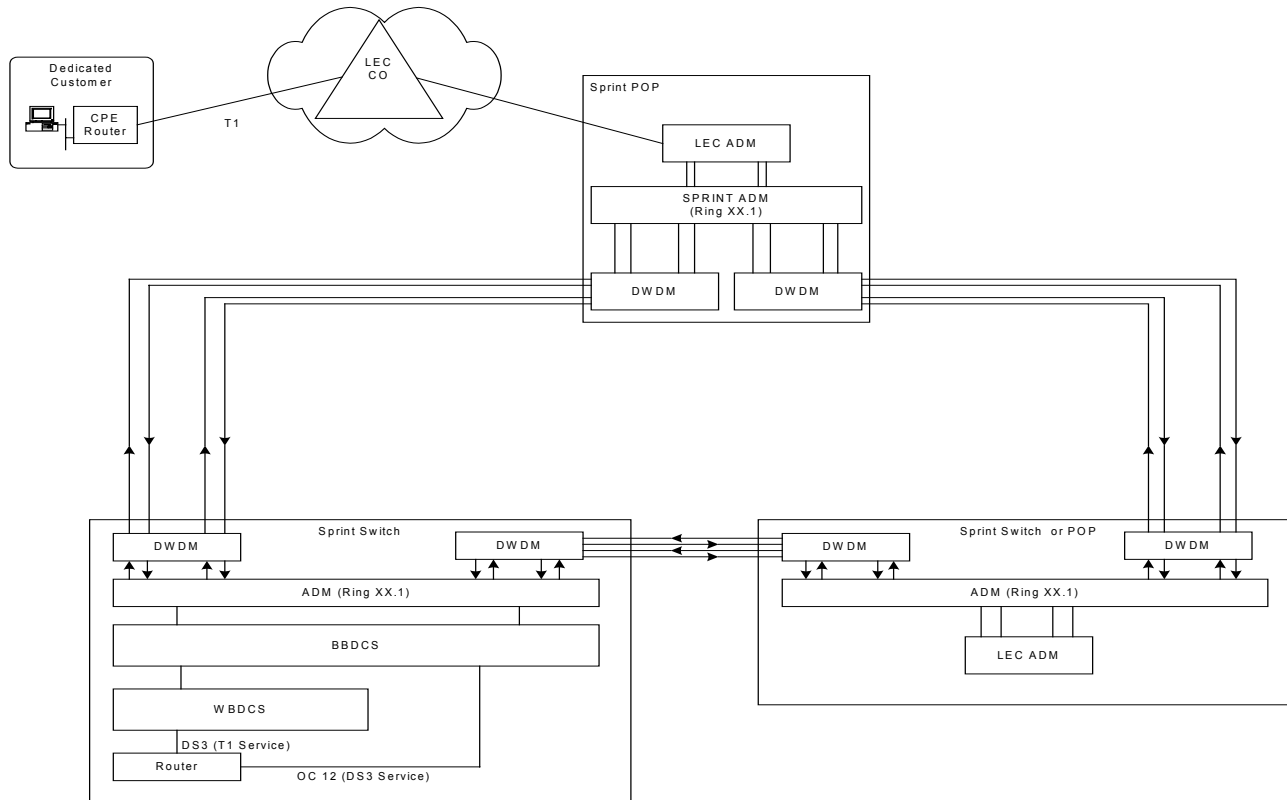


Sprintlink POP, Relay MD (SL-BB20-RLY)

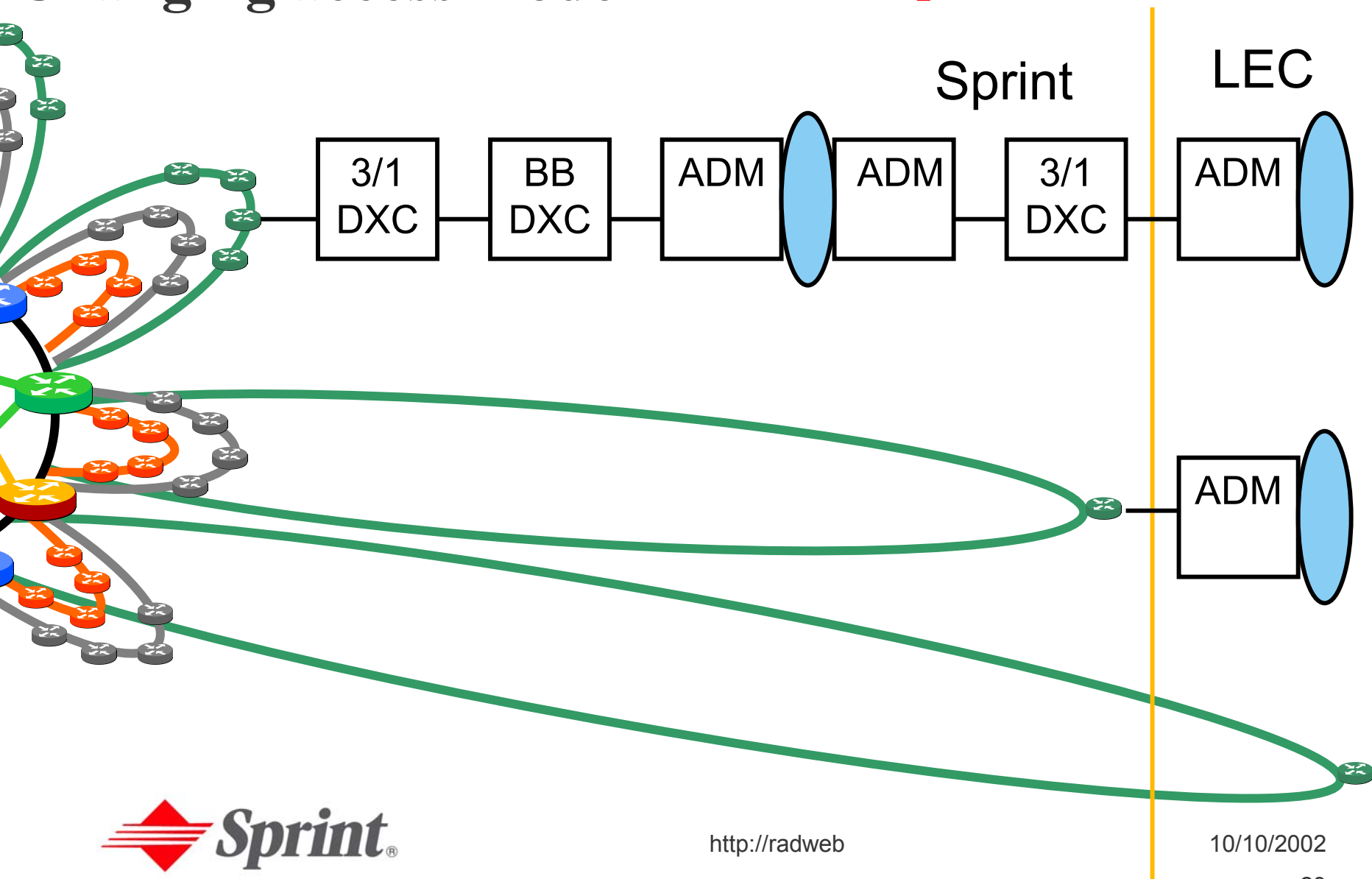


/10/2002

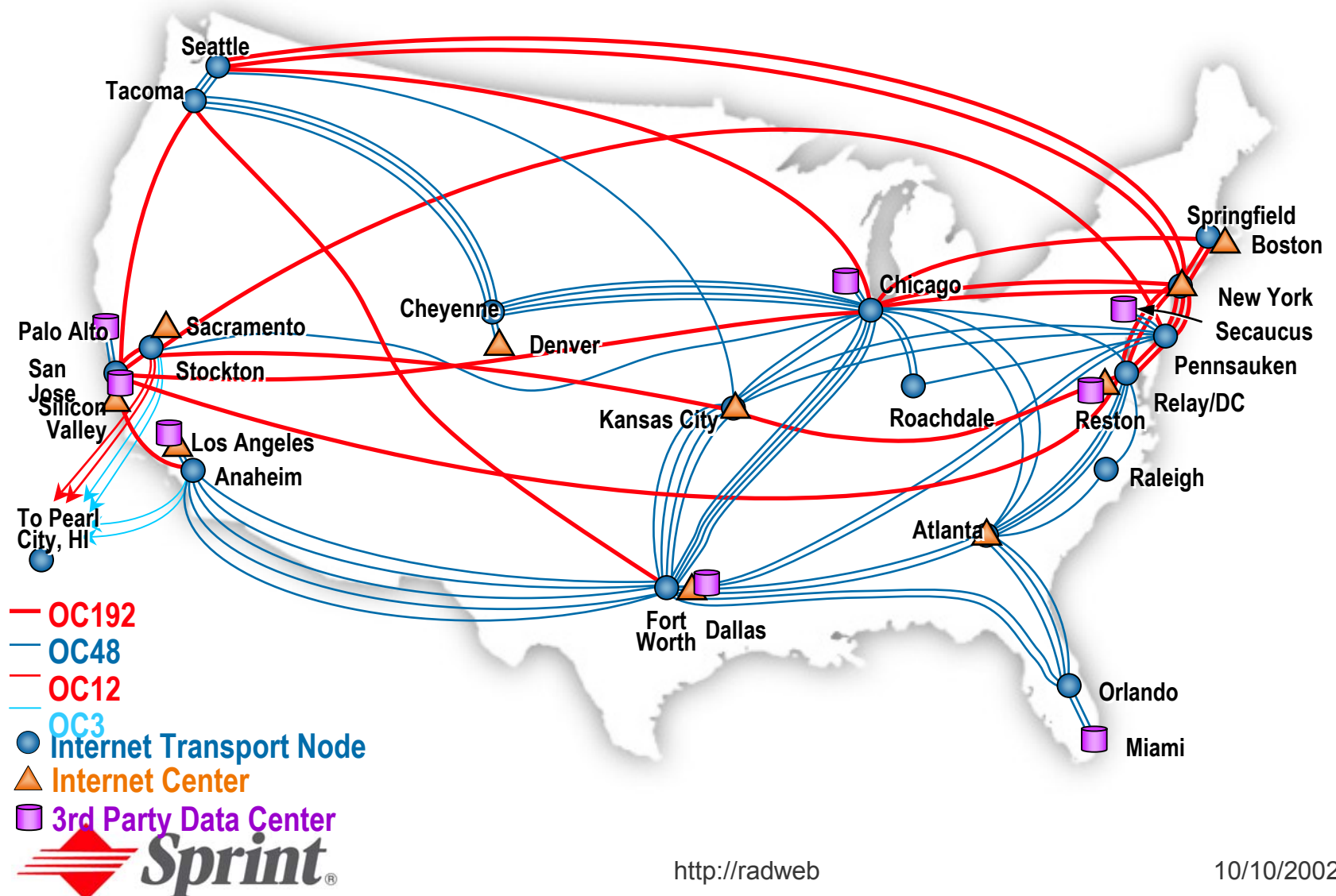
Traditional Access Today



Changing access model



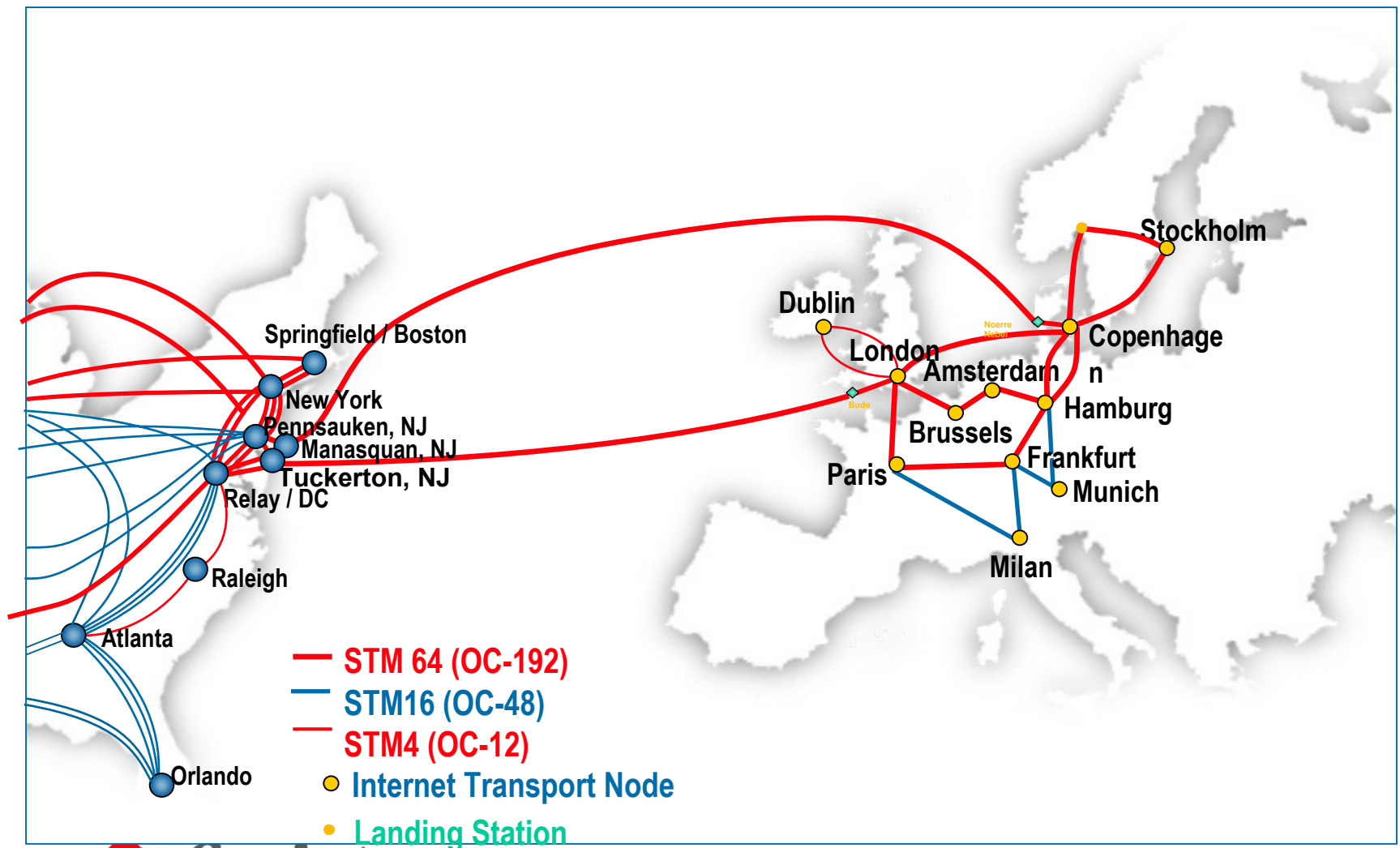
US Network (non technical slide...)



<http://radweb>

10/10/2002

EU Network



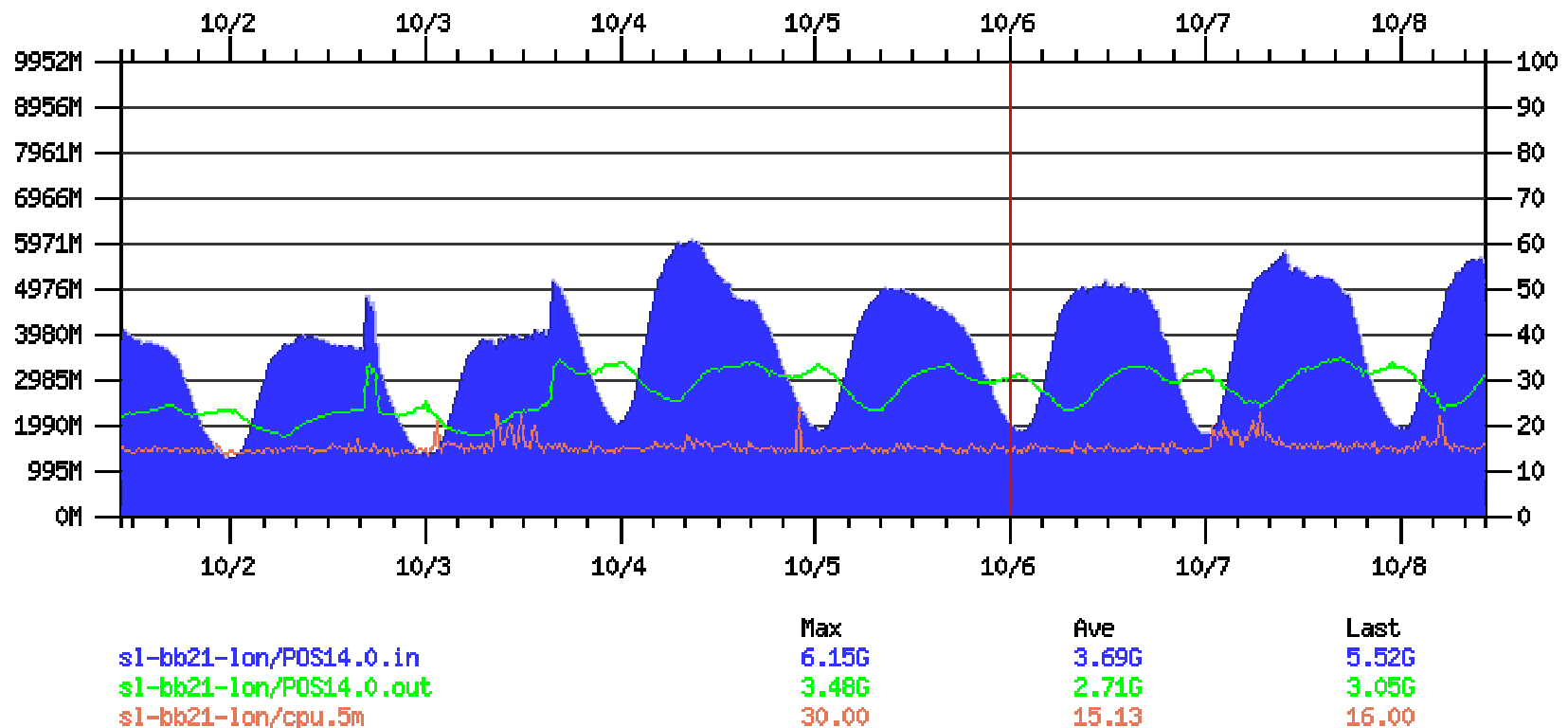
<http://radweb>

10/10/2002

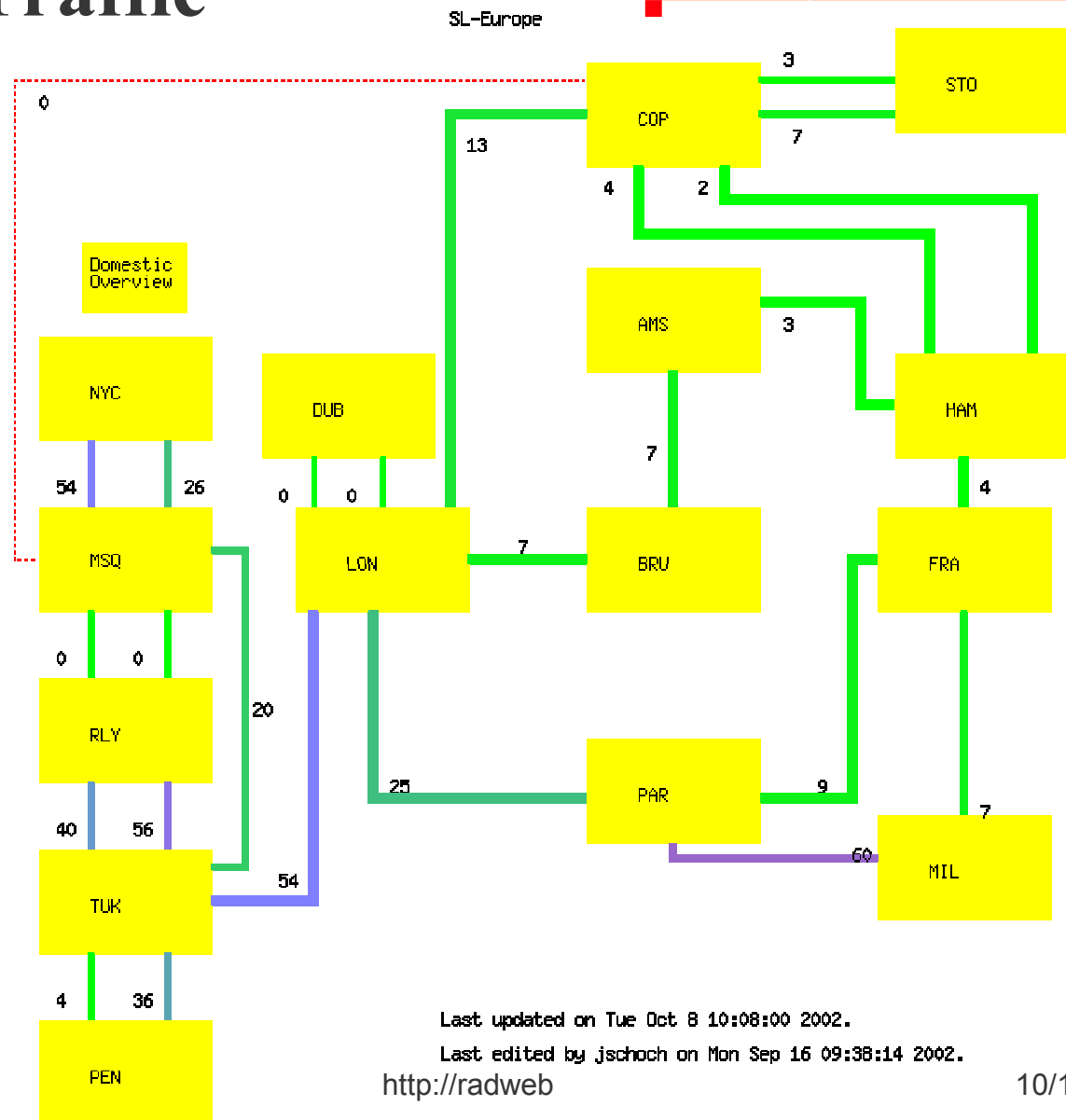
SL-BB20-TUK, SL-BB21-TUK



Tuckerton – London, STM-64



EU Network, Traffic



Last updated on Tue Oct 8 10:08:00 2002.

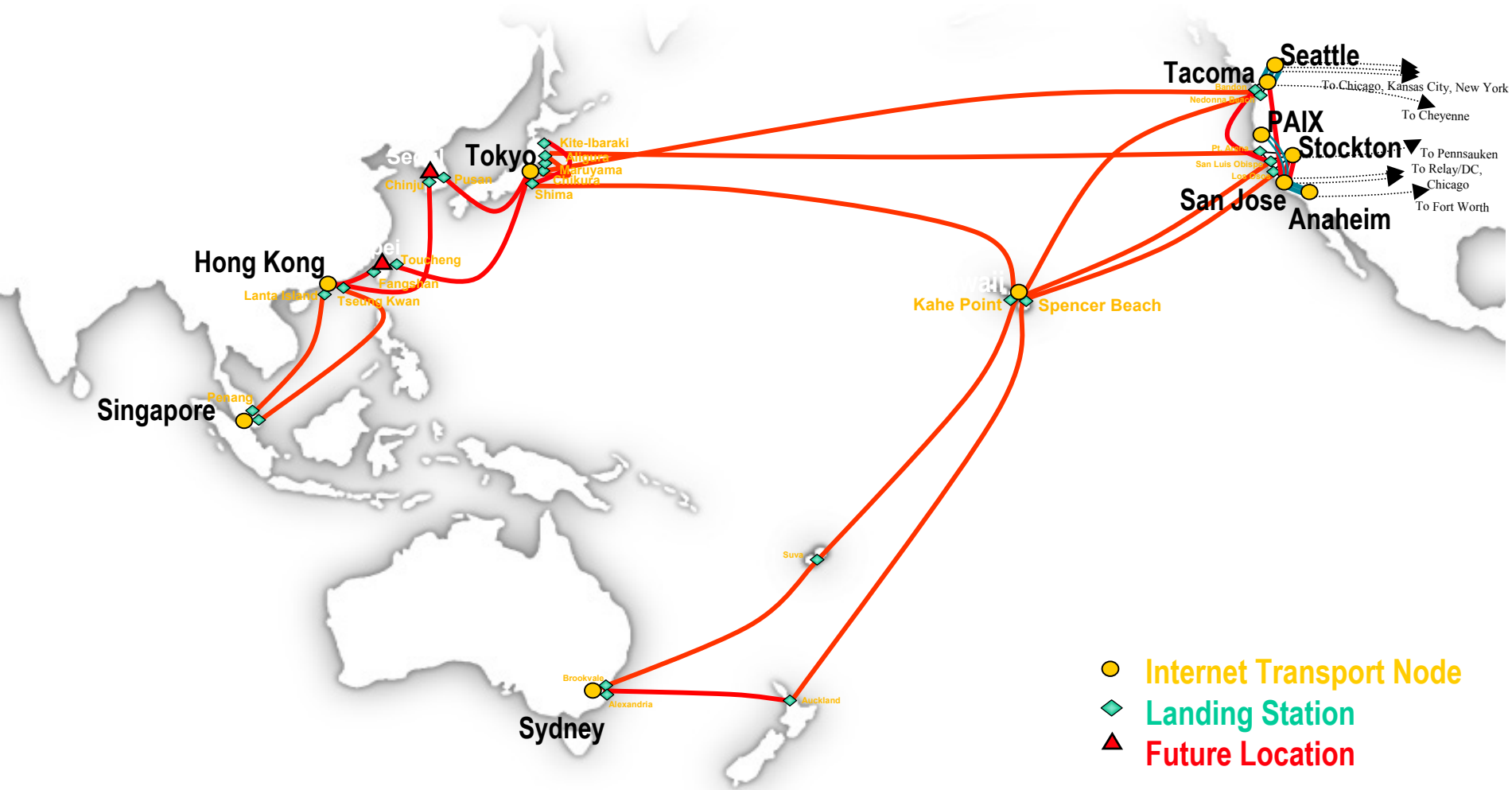
Last edited by jschoch on Mon Sep 16 09:38:14 2002.

<http://radweb>

10/10/2002



2002 Asia Sprint IP Backbone Network



<http://radweb>

10/10/2002

Central and South America Backbone Network



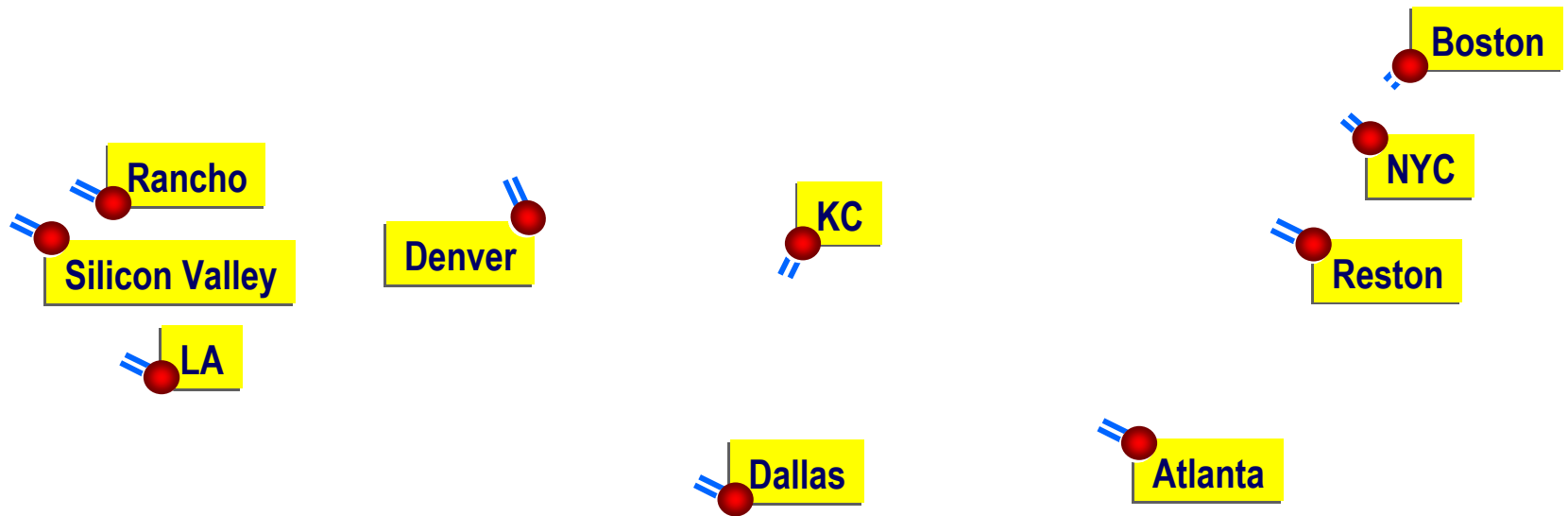
[Back to Navigation Bar](#)



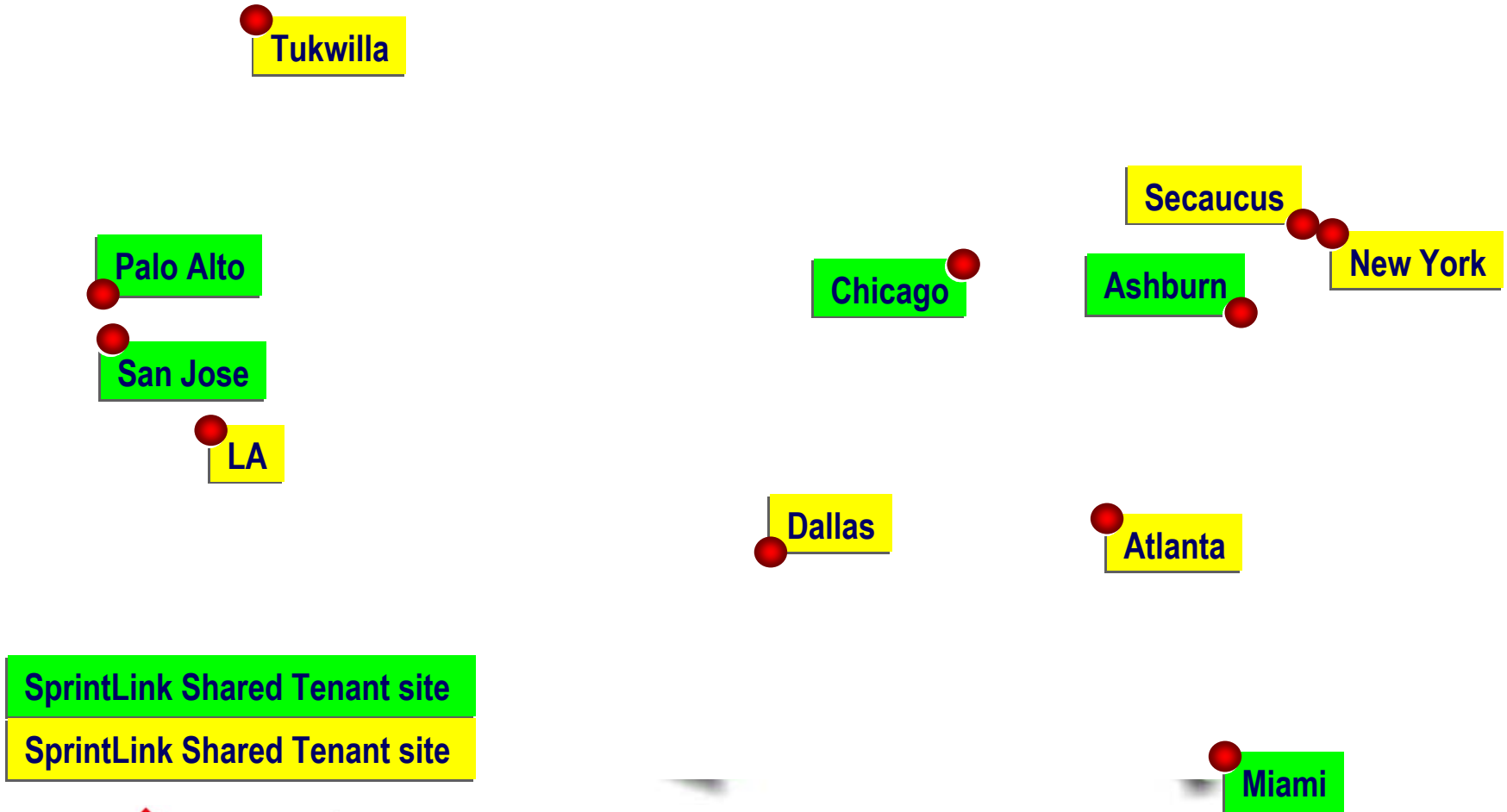
<http://radweb>

10/10/2002

US 10 Internet Centers



10+ Carrier Hotel Sites



SprintLink - Strengths

- **Homogeneous Global Architecture**
- **Single AS Globally (exception: AU)**
- **IP Layer Redundancy Drives Accountability**
 - Accountability equals Customer Service
- **L3/L1 Architecture from Day 1 - No False Starts**
- **Success at Driving New Equipment Development**
- **Leader in Peering Architectures**
- **Robust Architecture Allows for Unsurpassed Stability**
- **Lead in the Introduction of Multicast Technology**
- **Leading SLAs via Zero Loss & Speed of Light Delays**



Agenda -- MPLS

- **Brief MPLS History of the MPLS Universe...**
- **Traffic Engineering**
- **QoS**
- **Convergence/Restoration**
- **Layer 2 Transport/VPN**
- **Layer 3 Transport/VPN**
- **Provisioning**
- **Anything Else?**



Brief History of the MPLS Universe

- This Page Intentionally Left Blank...



Traffic Engineering

■ MPLS Approach:

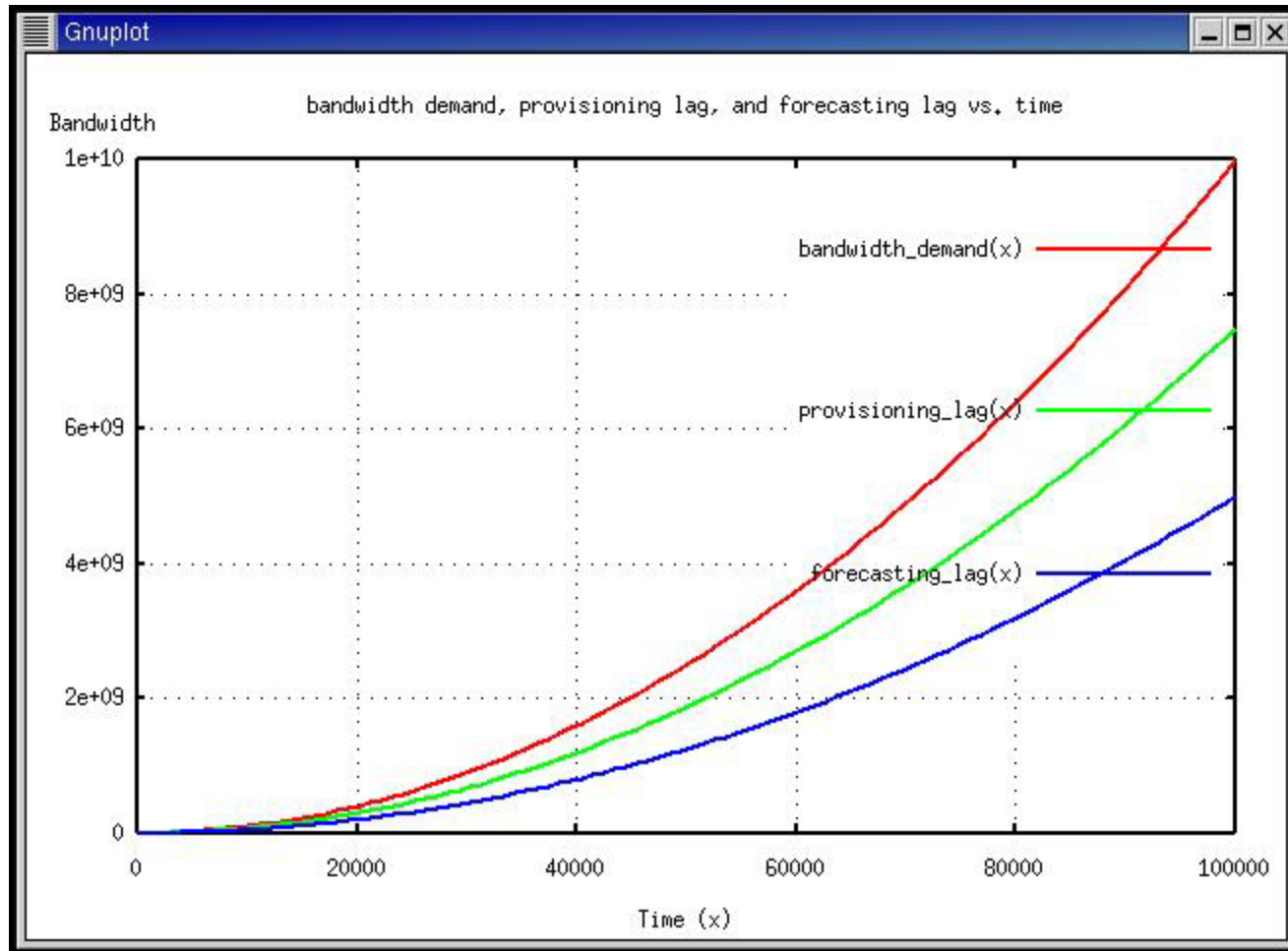
- Off/On-line computation of CoS paths
- RSVP-TE + IS-IS/OSPF-TE
- Tunnel Topology
- Can consider a wide variety of “metrics”

■ Sprintlink Approach

- “1:1 Protection Provisioning”
- Nice side effect: Zero loss, speed-of-light-like latency, small jitter
- Provisioning ahead of demand curve
 - Note demand/provisioning curve deltas



Demand vs. Provisioning Time Lines

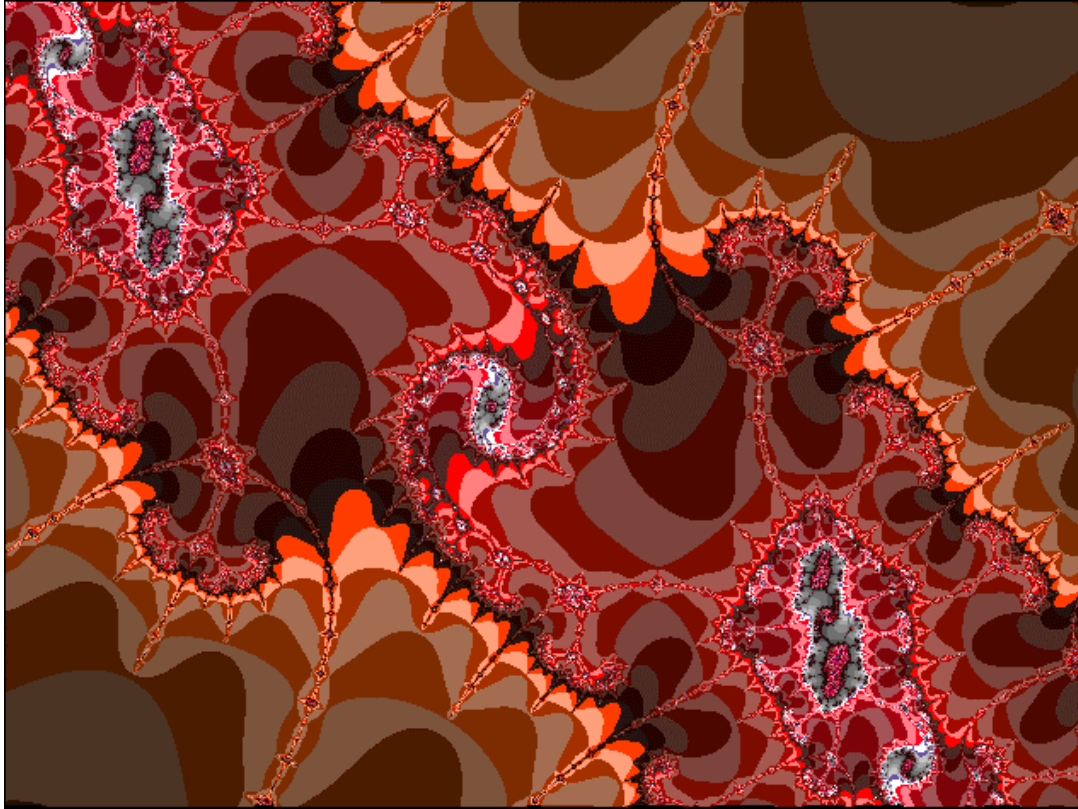


Traffic Engineering

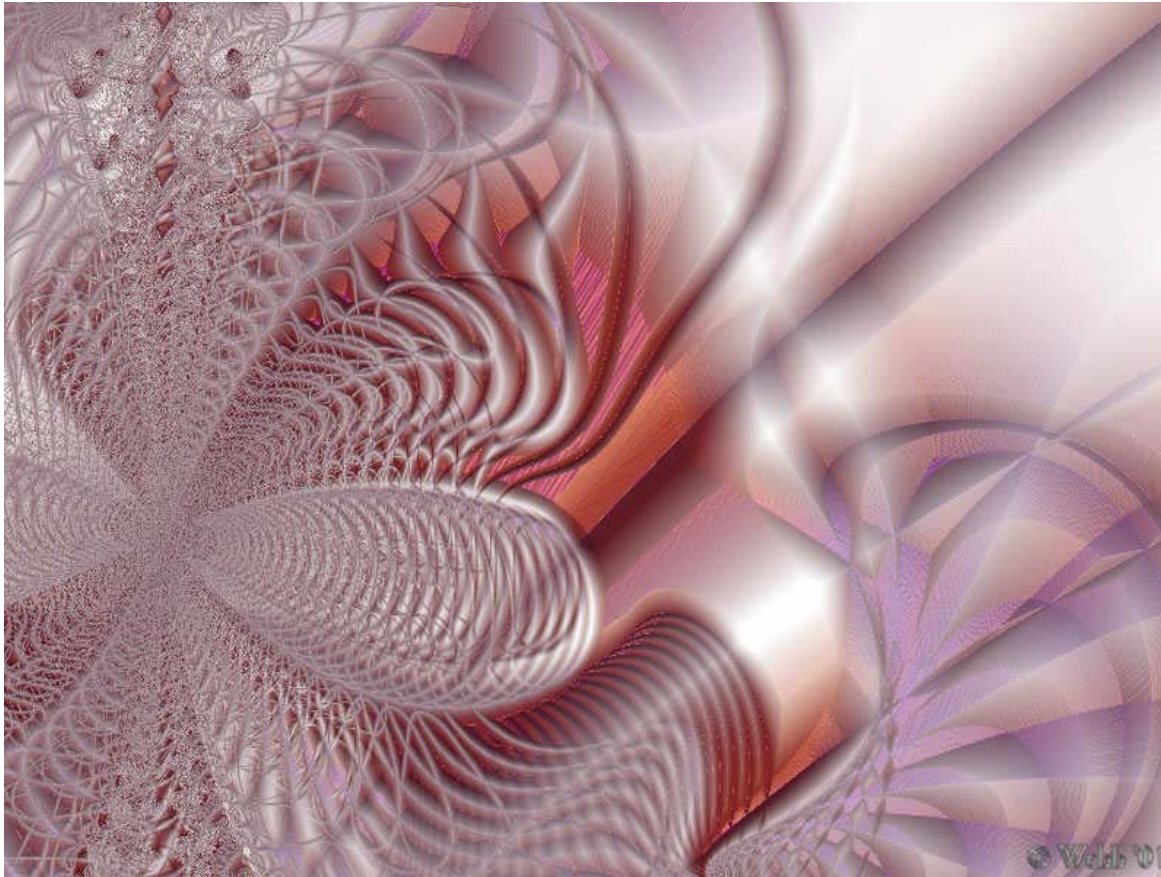
- **Aggregated Traffic in a core network (\geq OC48) is “uncorrelated”, that is, not self-similar**
 - “Impact of Aggregation on Scaling Behavior of Internet Backbone Traffic”, Zhi-Li Zhang, Vinay Riberio, Sue Moon, Christophe Diot, Sprint ATL Technical Report TR02-ATL-020157 (<http://www.sprintlabs.com/ipgroup.htm>)
 - **So you can actually provision to avoid queuing in a core network**
- **With proper network design, you can get within 3% of optimal (utilization)**
 - “Traffic Engineering With Traditional IP Routing Protocols”, Bernard Fortz, Jennifer Rexford, and Mikkel Thorup
 - **So why would you buy the complexity of MPLS-TE?**



Aside: Self-similarity



Aside: Self-similarity



<http://radweb>

10/10/2002

MPLS-TE and Sprintlink

- Engineering Aside -- No Current Need for MPLS-TE
 - All Links Are Same Speed Between All Cities Domestically (two exceptions)
 - 50% of bandwidth is reserved by design on every link for protection (actually 1/n reserved...)
 - If there is no queuing and/or buffering, why do we need a constraint on which packets get forwarded first.
 - More to Follow
 - **We are in the business of delivering ALL packets for ALL of our customers**
 - Too Much State in Your Core Will Eventually Burn You
 - Or Your Edge for That Matter



■ MPLS Approach

- MPLS in and of itself provides no QoS facilities
- Diffserv-aware MPLS-TE, lots of other machinery, state in the core, complexity

■ Sprintlink Approach

- Congestion free core, CoS on edge (“edge QoS”, as access is where congestion occurs)
- As previously mentioned, recent results show that aggregated traffic in the core network “uncorrelated”, which means you can actually provision a core to avoid queuing

■ What does QoS in a core mean anyway?

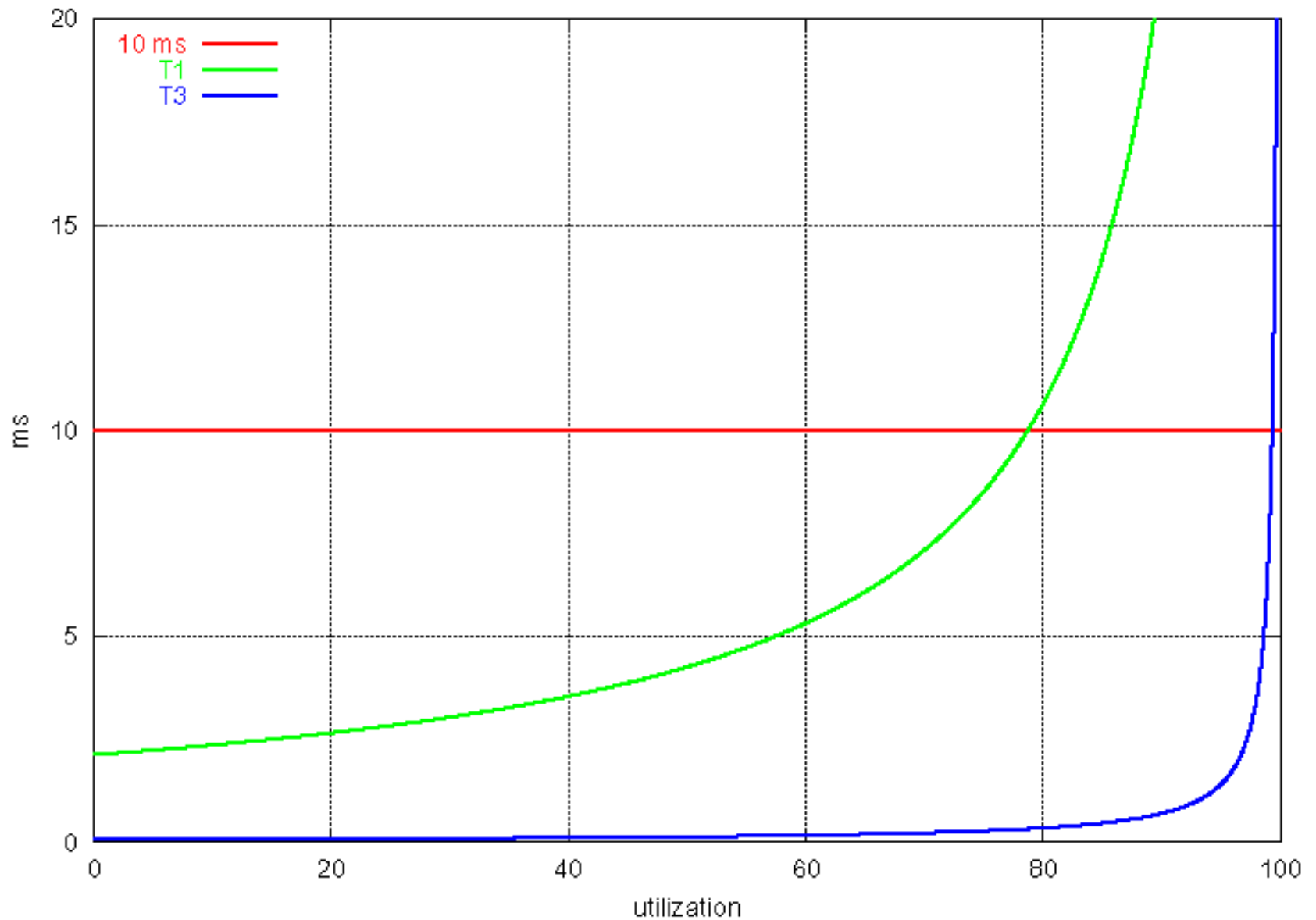


Sprintlink Core SLA

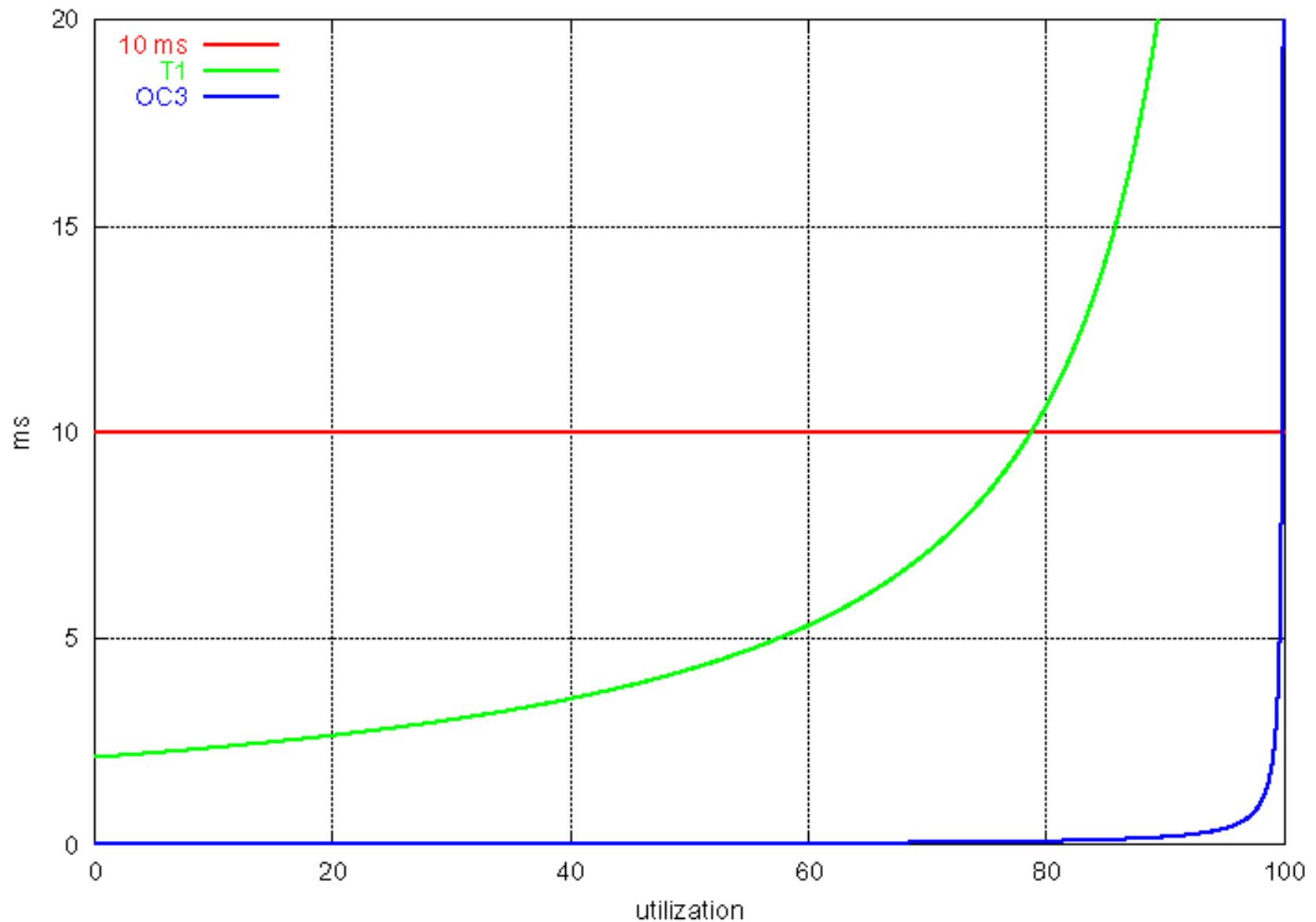
■ Forwarding outages	< 1s
■ Packet loss	0.05%
■ Packet reordering	1%
■ RTT US	100ms
■ RTT World	380ms
■ Jitter	5ms
■ BW/Delay quota	2.4G/350ms
■ MTU	4470B



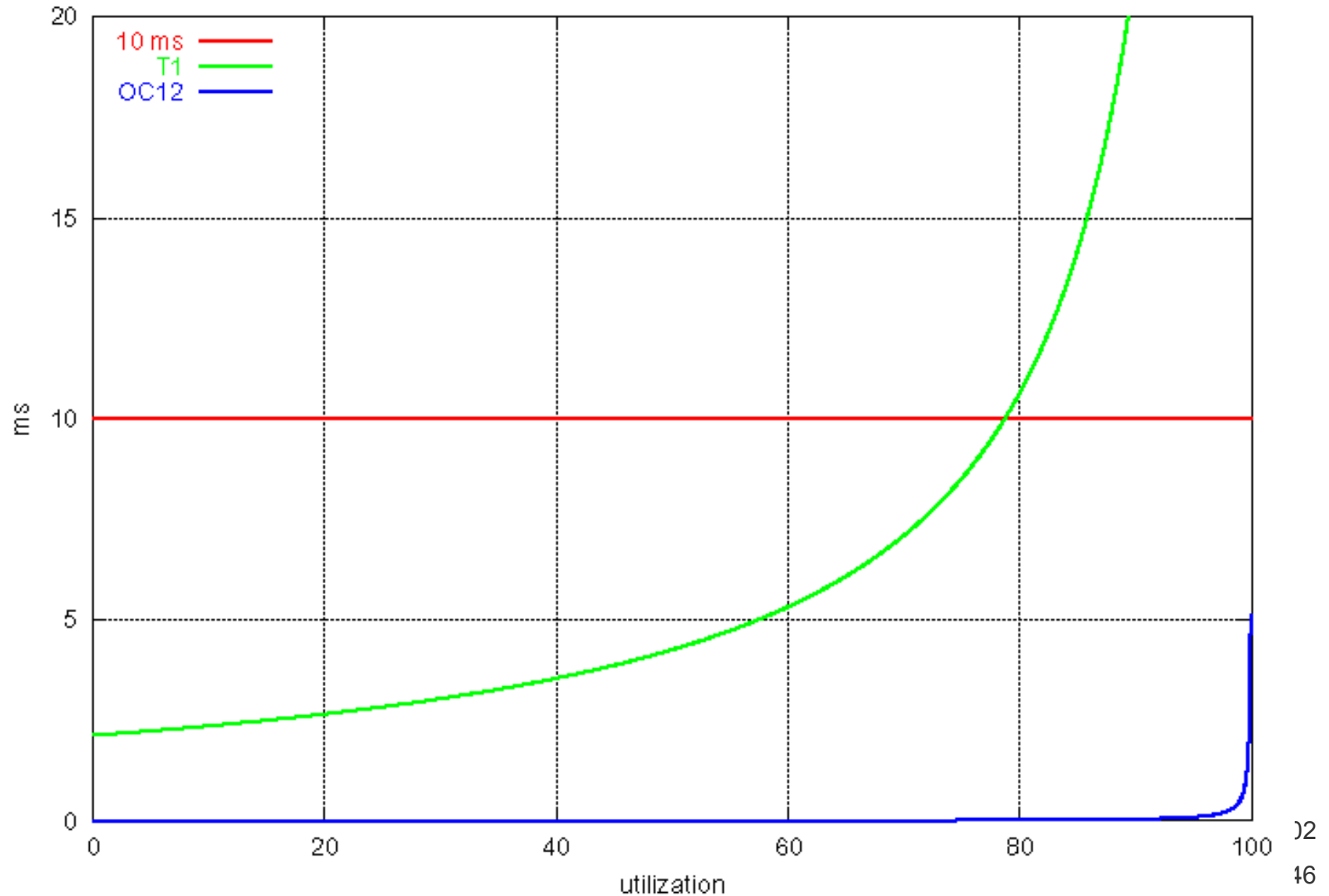
T1 & T3 Queueing Delay



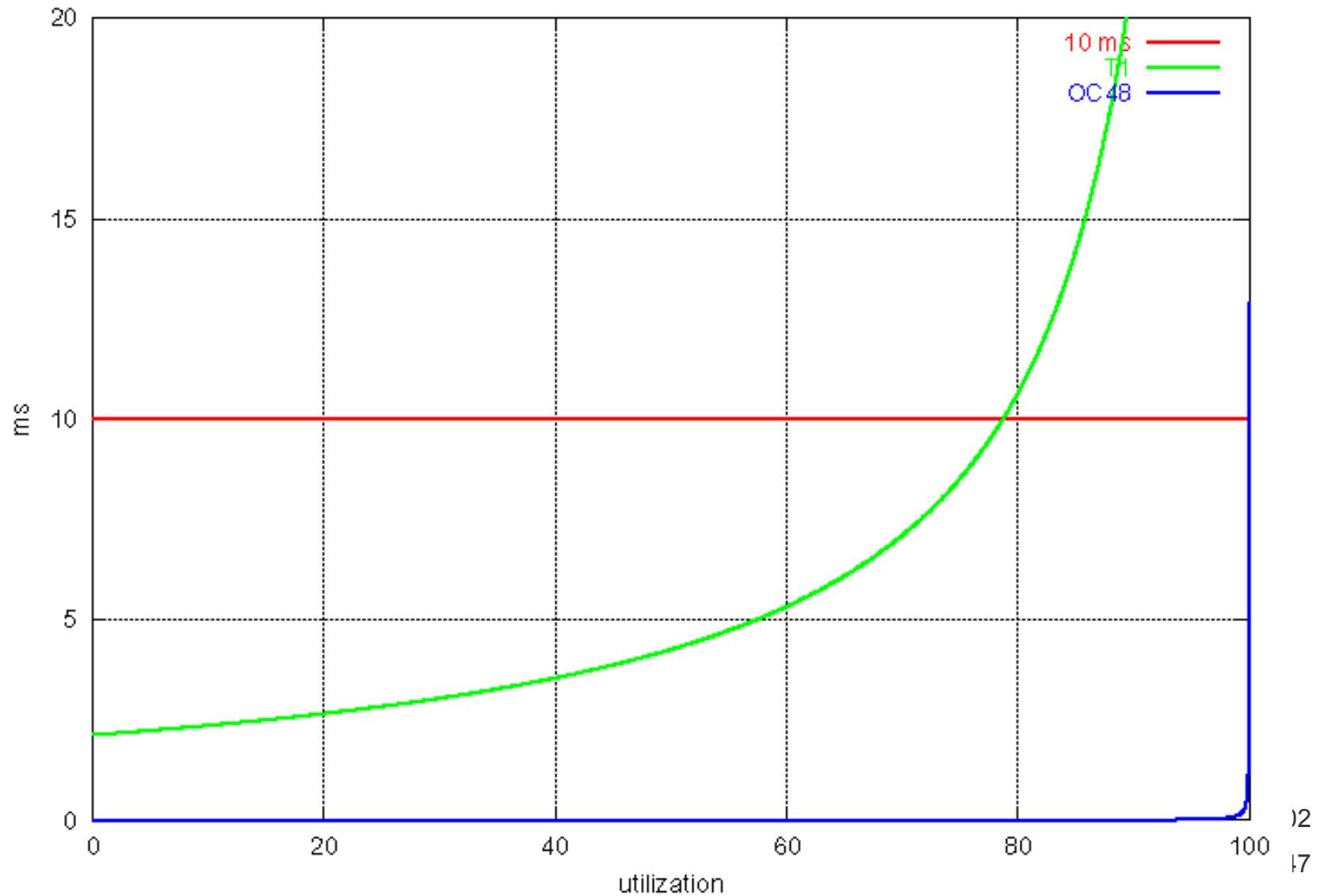
T1 & OC3 Queueing Delay



T1 & OC12 Queueing Delay



T1 & OC48 Queuing Delay



Convergence/Restoration

■ MPLS Approach

- Fast Reroute, with various kinds of protection
- $O(N^2 \cdot C)$ complexity (C classes of service)
- B/W must be available

■ Sprintlink approach

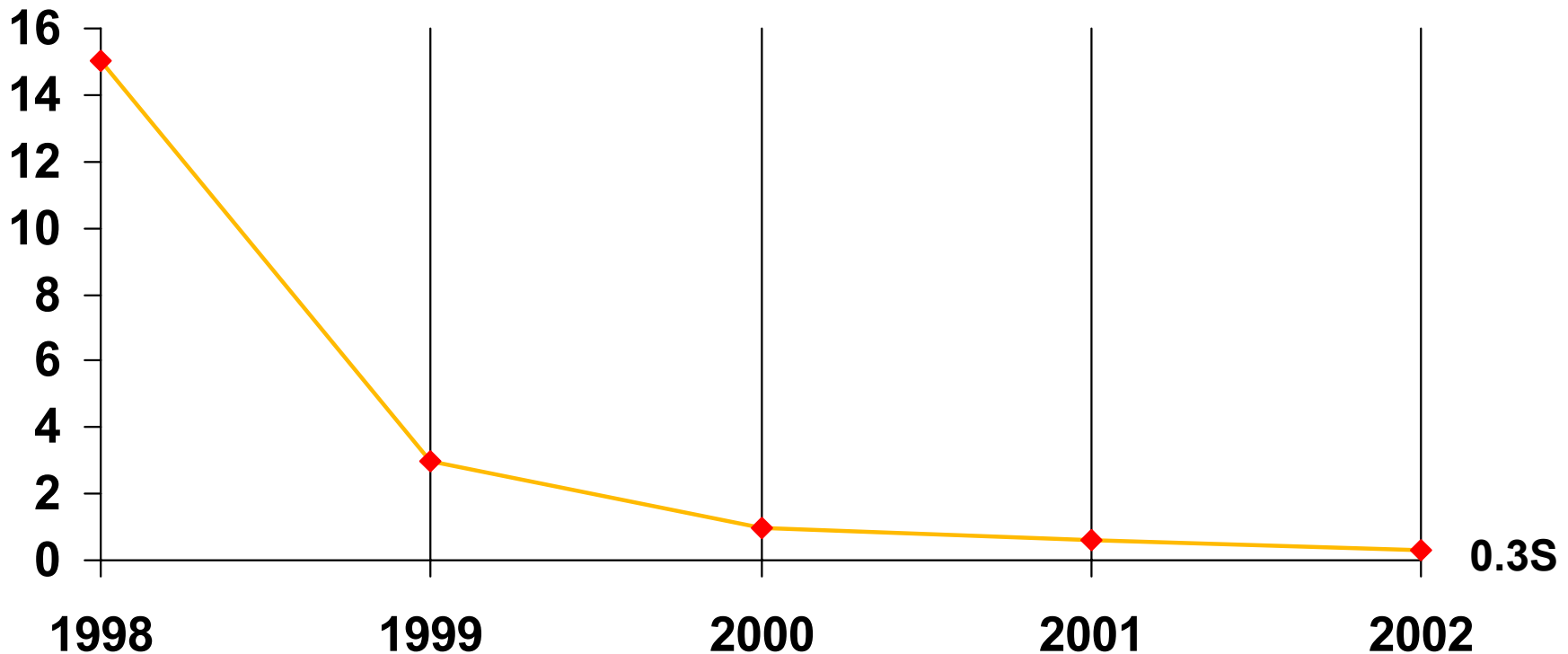
- Simple network design
- Equal cost multi-path/IS-IS improvements for sub-second convergence
- BTW, what is the (service) convergence time requirement?

■ Note: Recent work shows that FIB download dominates service restoration time, so...

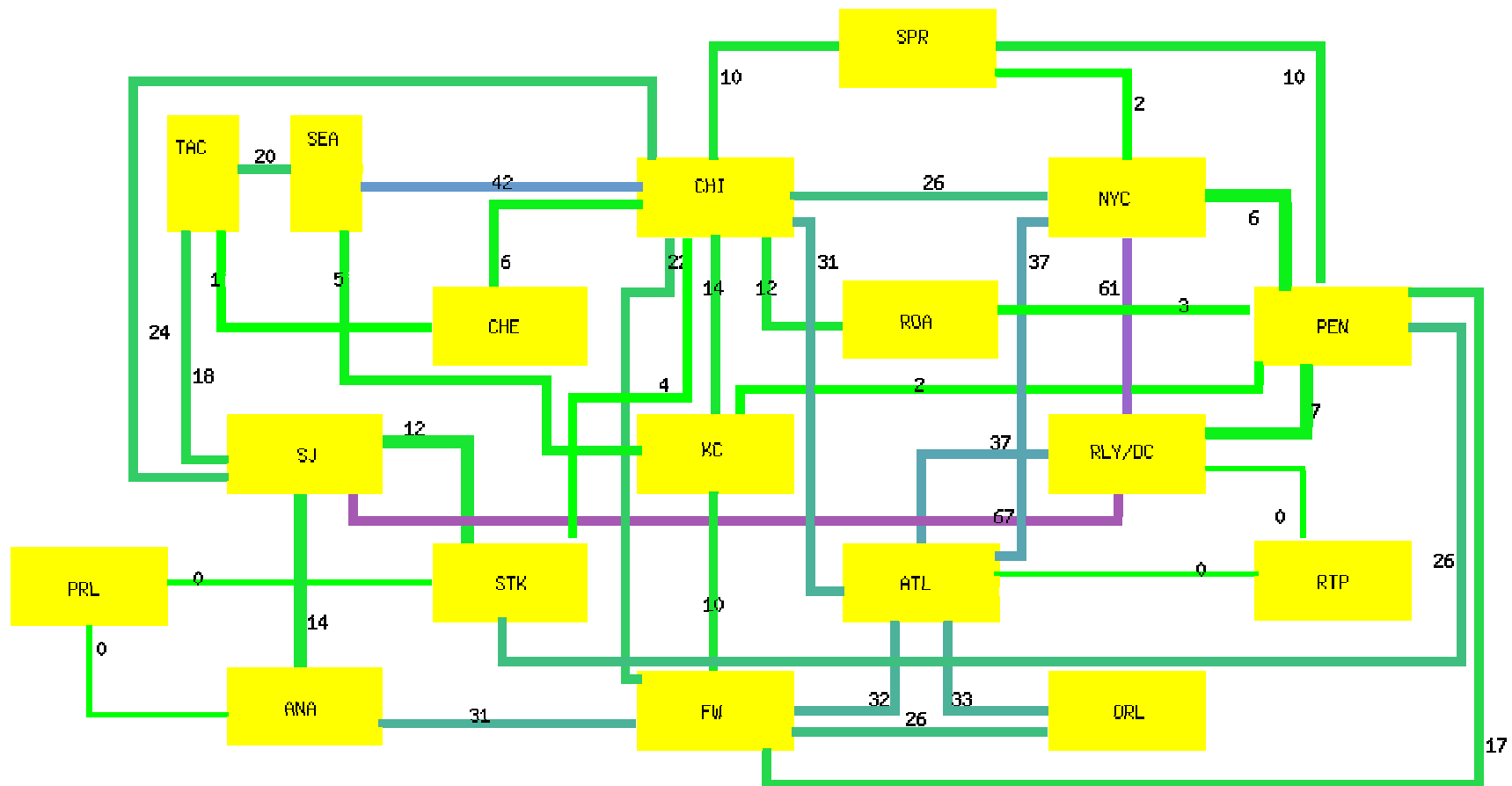


Non Forwarding after topology change

Seconds



US Network (summary)



Last edited on Tue May 28, 2001.
Last updated on Tue Oct 8 10:08:00 2002.



<http://radweb>

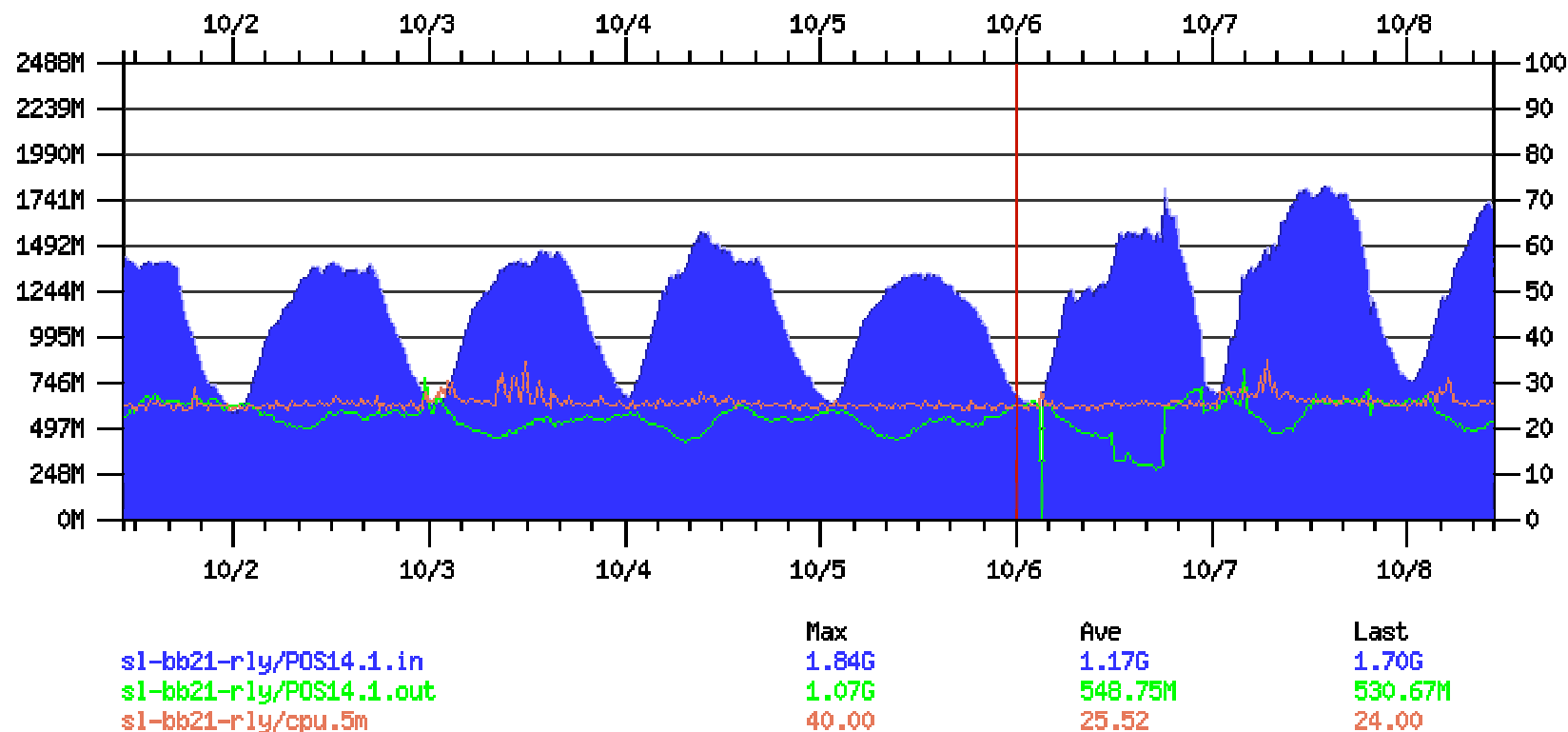
10/10/2002

Sprintlink Engineering Overview

Last updated on Tue Oct 8 10:08:00 2002.



US Cross Country Link



L2 Transport/VPN

- **MPLS Approach**

- PWE3 consolidated approach (e.g. martini encap)
- CoS/QoS Capabilities

- **Sprintlink Approach**

- L2TPv3 (UTI) + Edge QoS
- Already doing (I)VPL, Ethernet, and Frame Relay

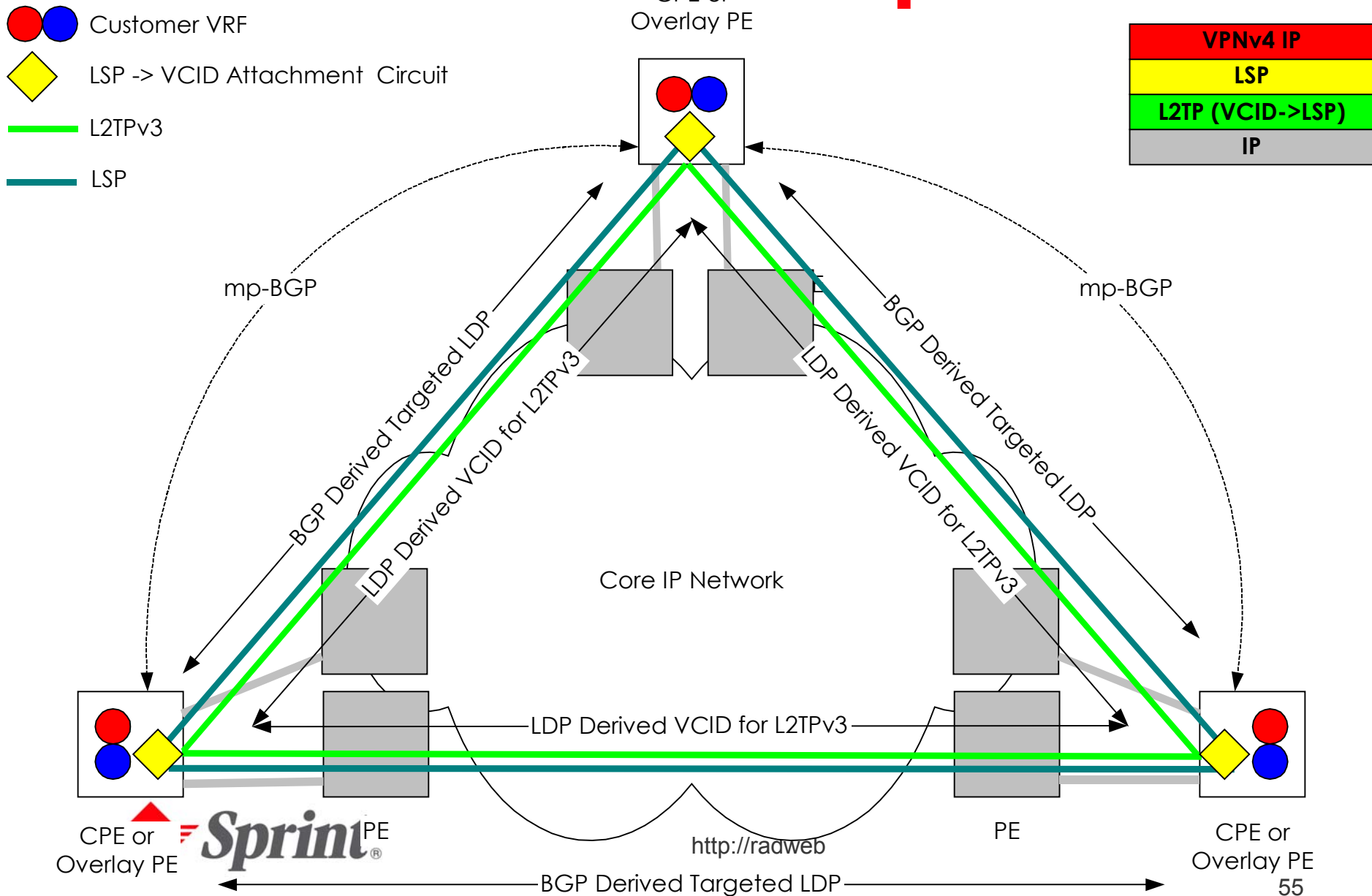


L3 Transport/VPN

- **MPLS Approach**
 - RFC 2547 (MPLS/BGP VPN)
- **Sprintlink Approach**
 - CPE Based and VR based (network based)
- **Interestingly, although many customers seem to be asking for 2547 VPN, there is no artifact that will allow users to distinguish between a VR VPN and a 2547 VPN**
 - See also “Integrity for Virtual Private Routed Networks”, Randy Bush and Tim Griffin, INFOCOMM 2003
 - Result: 2547 cannot provide isolation (“security”) in the multi-provider (inter-domain) case



MPLS/LSP through L2TPv3 Attachment Circuits



Comment on VPN “Security”

- **Many providers are claiming**
 - **Isolation == Security**
 - This is the “Private network argument”
 - In particular, from DoS like attacks
- **Reality Check --> Isolation != Security**
 - This is the **Security by Obscurity** argument!
 - On a public infrastructure...
 - you would have to trace the tunnel(s)
 - end points are RFC 1918, so not globally visible
 - and not even addressed in L2 VPN
 - On “Isolated” infrastructure...



Isolated Infrastructure...

- Well, as soon as > 1 customer, we're no longer "isolated"
- What happens when someone puts up a public internet g/w?
 - Appears to be some kind of false security
- Isolation \neq Security (of any real kind)



Provisioning/Optical Control Planes

- **MPLS Approach**

- GMPLS or some variant (ASON)

- **Sprint Approach**

- Support the deployment of an optical layer control plane
- Integration into backoffice/OSS systems still under study
- Reliability/Robustness must be proven before deployment

- **There is, however, reason to be skeptical of optical control planes like GMPLS...**



What is there to be skeptical about?

- **Well, a fundamental part of the IP architecture is “broken” (decoupled) by GMPLS**
 - Basically, the “decoupling” means that one can no longer assume that a control plane adjacency implies a data plane adjacency, so you need a convergence layer (RSVP-TE+LMP)
 - What are the implications of this?
- **Aside: We know that IP doesn’t run well over a control plane that operates on similar timescales (cf. IP over ATM with PNNI)**



MPLS – Bottom Line

- **If you have 5 OC48s Worth of Traffic...**

- You need 5 OC48s...
 - **none of these TE or {C,Q}oS techniques manufactures bandwidth**
- If the path that carries those 5 OC48s (or a subset of breaks)...
- Then you better have 5 more (or that subset) between the source and destination...
- Its that simple for a true tier 1 operator.

- **If the above is not the case...**

- Then be prepared to honor your SLAs and pay out (waive the fees)

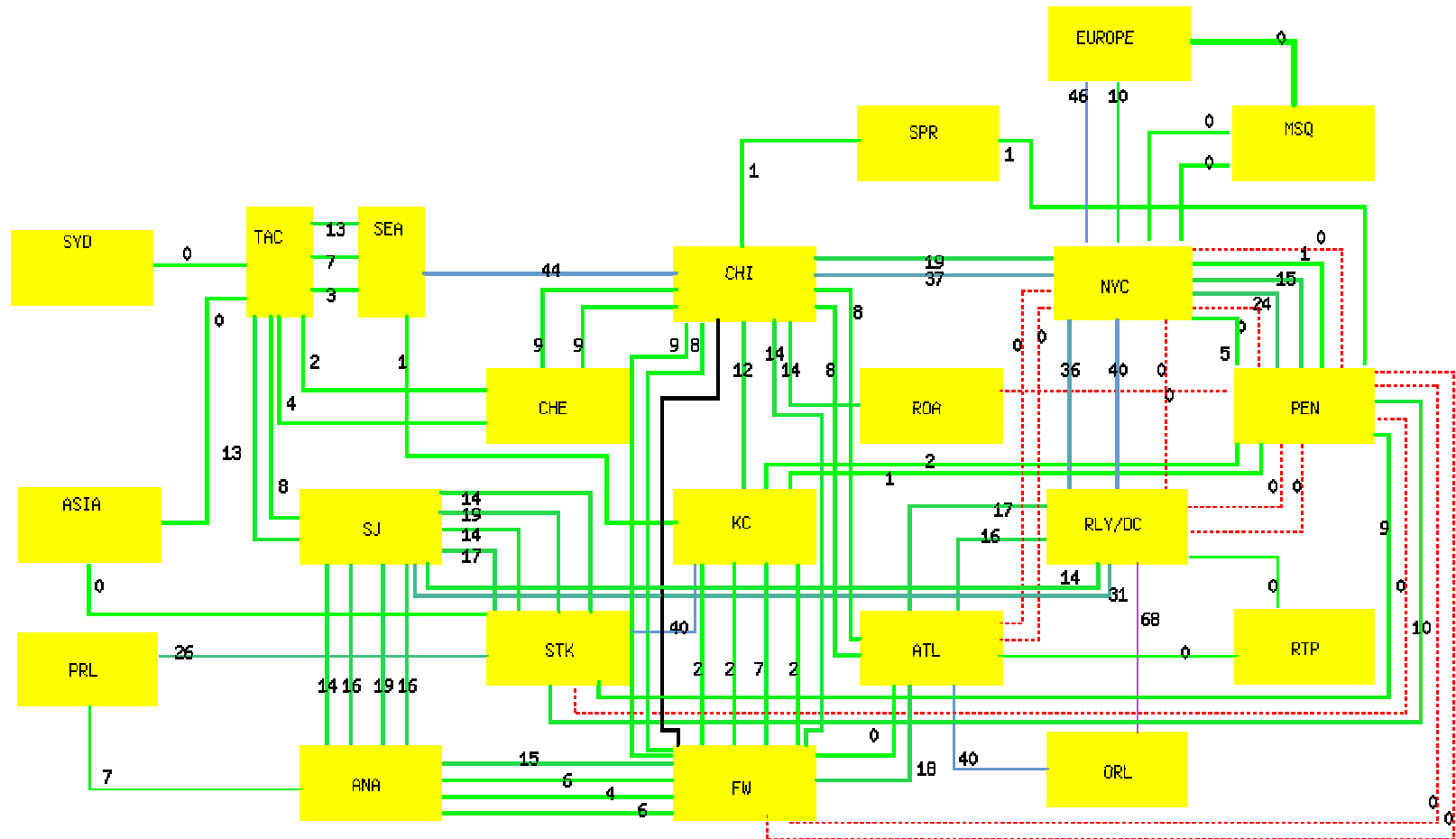


A Brief Look...

- **At a couple of high profile failure scenarios**
- **Baltimore Tunnel Fire**
- **Other Fiber cuts**



Baltimore Train Tunnel Fire



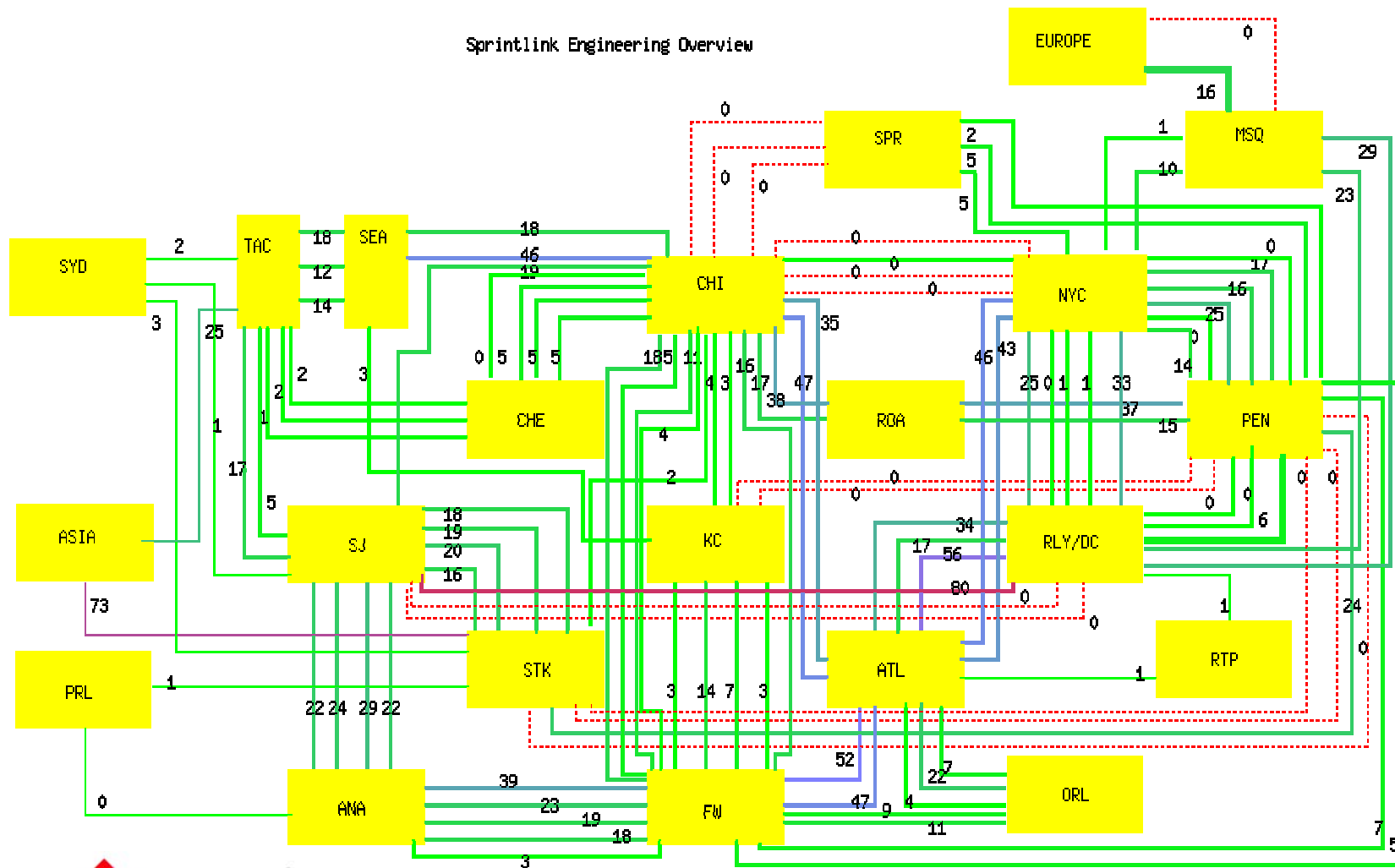
Last edited on Mon Jul 9 11:51:14 2001.
Last updated on Fri Jul 20 14:18:00 2001.



<http://radweb>

10/10/2002

Train Derailment, Major Fiber Cut In Ohio April 25




Last edited on Mon Apr 15 16:37:09 2002.
Last updated on Thu Apr 25 12:18:00 2002.

<http://radweb>











10/10/2002



“WorldCom officials blame the problem on a train derailment that occurred in Ohio, 50 miles south of Toledo, resulting in fiber cuts. Meanwhile, independent engineers pointed to [Cisco Systems Inc.](#) (Nasdaq: [CSCO](#) - [message board](#)) routers, which Cisco officials later confirmed. But the bottom line may be: If there's a fiber cut or router problem, isn't the network supposed to stay up anyway?”

Lightreading – 4/26/02

More Stats – 3rd Party

Name	Type	Indicator	Metrics					Start Date/End Date
sprint nyc->san jose top	vector	Jitter		<input checked="" type="checkbox"/>	0.005	msec	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint nyc->san jose top	vector	Delay		<input checked="" type="checkbox"/>	28.824	msec	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint nyc->san jose top	vector	Packet Loss		<input checked="" type="checkbox"/>	0	%	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint nyc->san jose top	vector	Outages		<input checked="" type="checkbox"/>	0	%	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint san jose->nyc top	vector	Jitter		<input checked="" type="checkbox"/>	0.006	msec	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint san jose->nyc top	vector	Delay		<input checked="" type="checkbox"/>	28.808	msec	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint san jose->nyc top	vector	Packet Loss		<input checked="" type="checkbox"/>	0	%	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT
sprint san jose->nyc top	vector	Outages		<input checked="" type="checkbox"/>	0	%	Avg	24/07/2002 08:35:00 EDT 25/07/2002 08:35:00 EDT



Closing

- **Robust, yet simple, and built (day 1) on native Packet-Over-SONET/SDH framing infrastructure**
 - Ask me about *HOT* (*Highly Optimized Tolerance*) models of complex systems if we wind up with time
 - Basic result: Complex systems such as the Internet are characterized by *Robust yet Fragile* behavior
- **Load-sharing is done by a per-destination caching scheme**
 - I.E. traffic flows take only ONE best path across the SprintLink Network
 - Minimized packet re-ordering, reduced fiber-path induced jitter.
- **IP traffic growth is still doubling ~yearly**
 - Easier to provision the network to ensure no congestion in the core, more cost-effective than fancy queuing in the core.
 - Simple means reliable, fixable, and more stable.



Closing 2

- **Queuing only needed at the edge, where packet/frame sizes are ‘large’ in proportion to the ingress bandwidth.**
 - Stays with Simplicity Principle
 - Frees up Core routing system’s resources
- **Aside: Recent work in the complex systems field is leading to a deep understanding of the Complexity/Robustness tradeoffs in large (non-linear) systems. Let me know if you’d like more literature on this one...**



SprintLink and IPv6, Sure!

So once UMTS/3G
is deployed, there is
a worldwide IPv6
network you may
not be able to talk
with.....

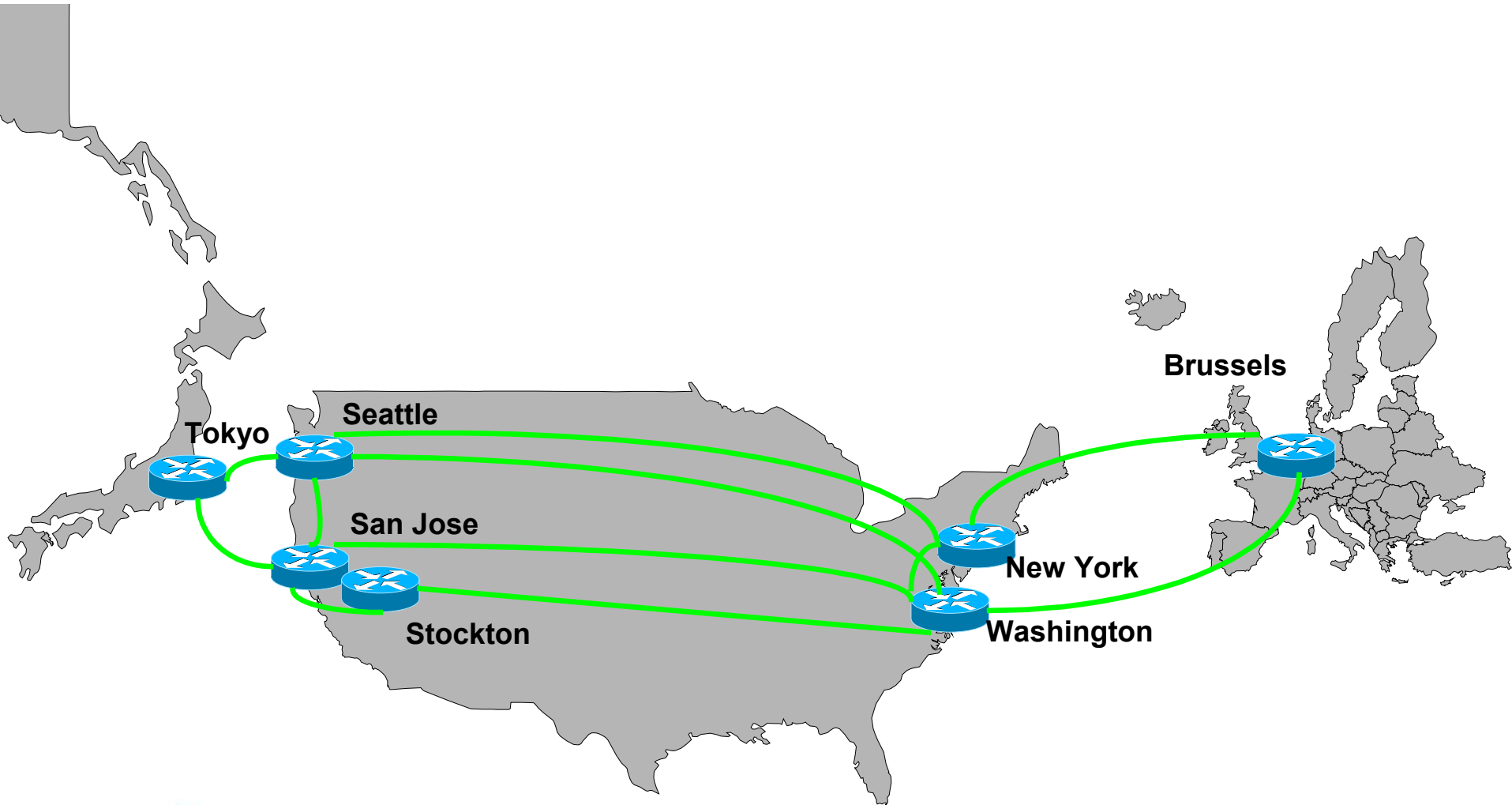
SprintLink IPv6 history

- 1997: Obtained 6bone address space (3ffe:2900::/24)
 - Original router under Robs desk ☺
- 1998: Totaling 15 customers using tunnels to 6bone
- 1999: Totaling 40 customers using tunnels to 6bone
 - Move router out to the network...
- 2000: Obtained ARIN space (2001:440::/35)
 - Totaling 110 customer using tunnels to 6bone.
- 2001-2002: Added 4 more IPv6 capable PoP's
 - Brussels, Washington DC, San Jose, New York
 - Member of the NY6IX exchange
 - Adding 1-2 connections/week!



10/10/2002

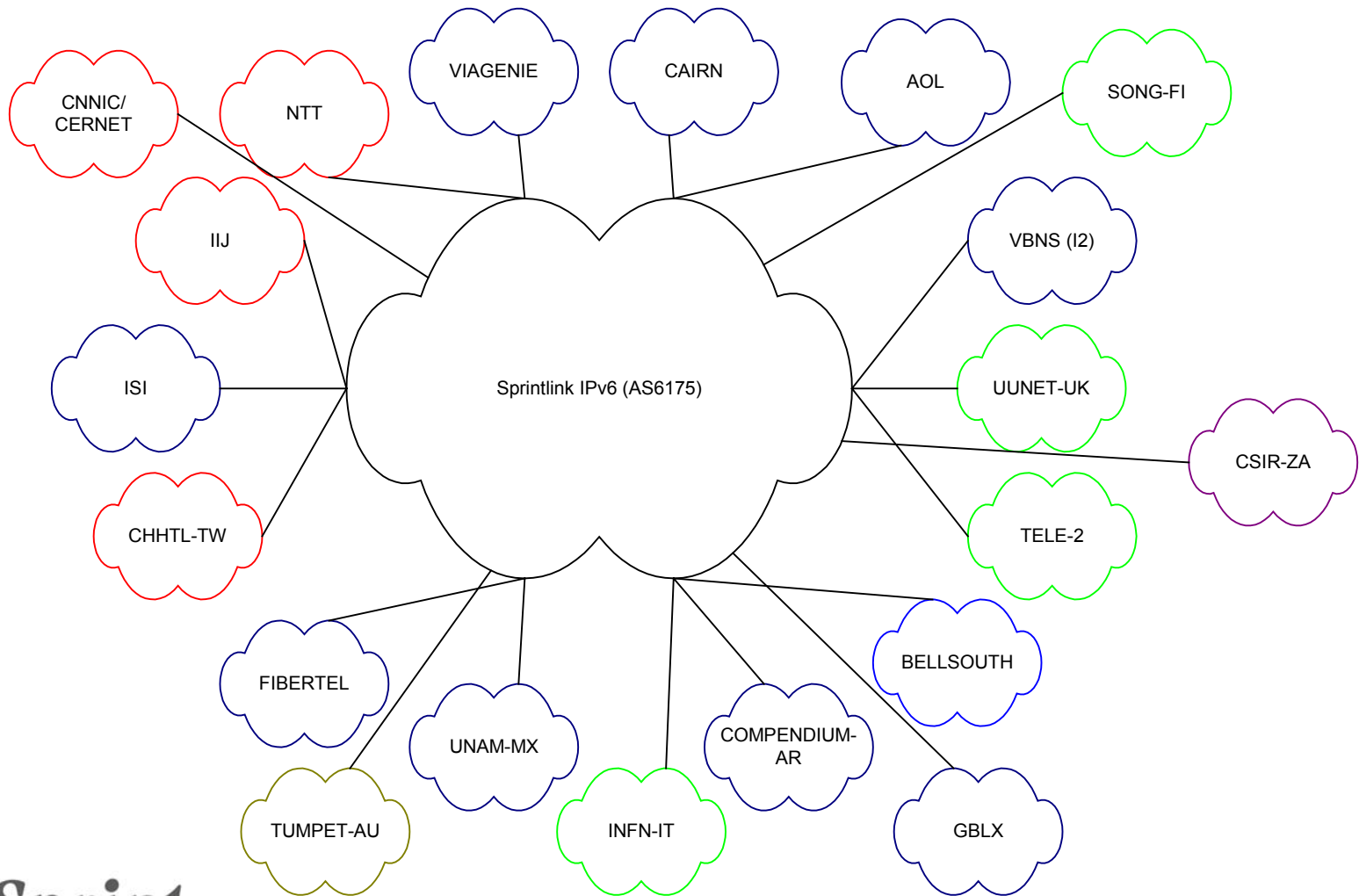
IPv6 Backbone Today



<http://radweb>

10/10/2002

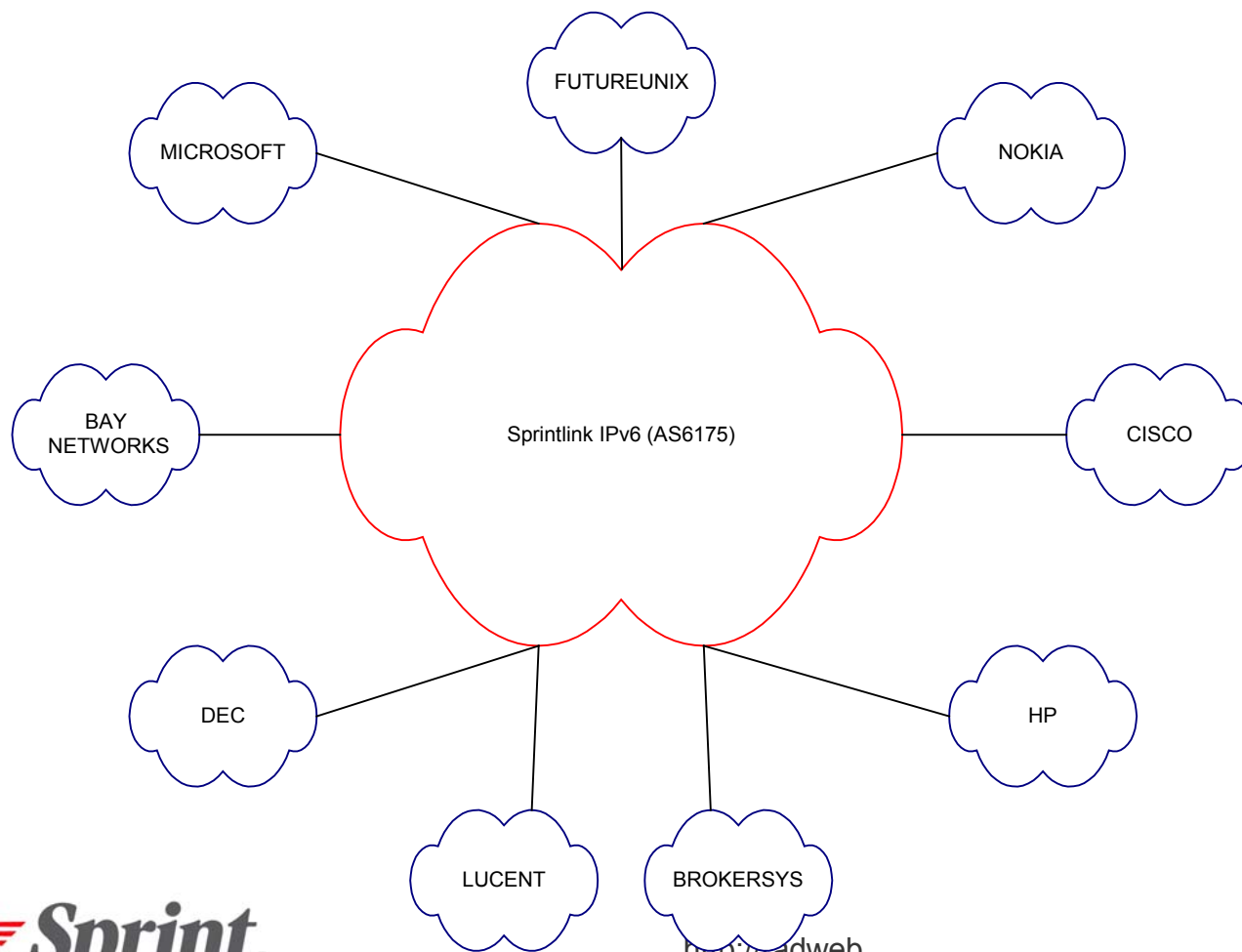
External IPv6 Connectivity



<http://radweb>

10/10/2002

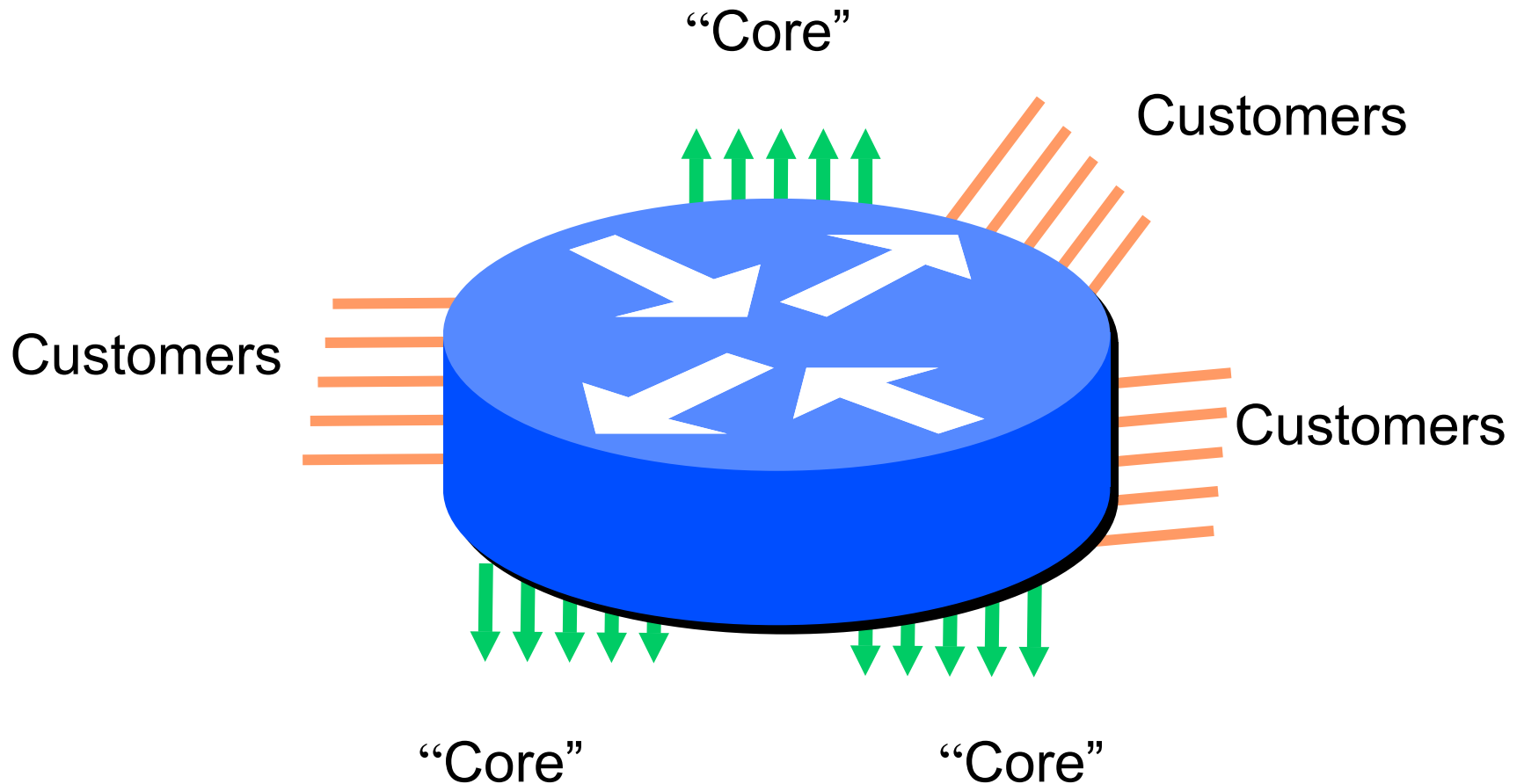
SprintLink IPv6 connectivity to HW & SW vendors



<http://adweb>

10/10/2002

Looking in to 2003





Questions?

Thank You

**Interested in switching
over to a real IP network?**

**Call or send E-Mail for a
Quote! roll@sprint.net**

+1 703 864 7887