



The Missing Link

Peter J. Welcher

Introduction

Hello again! Last month we discussed some background info about Certificate Authorities. You can find the article at: <http://www.netcraftsmen.net/welcher/papers/certauth01.html>

I planned to write about lab work with the Cisco IOS Certificate Authority for this month's article. However, I didn't look ahead, and the Feature Navigator now tells me this feature is supported on 2620 XM but not 2620. It's the usual Cisco IOS image size sort of thing, my convenient lab router has enough RAM (64 M) but not enough flash (16 M, needs 32 M). Time to punt (seeing as it is football season as I write this).

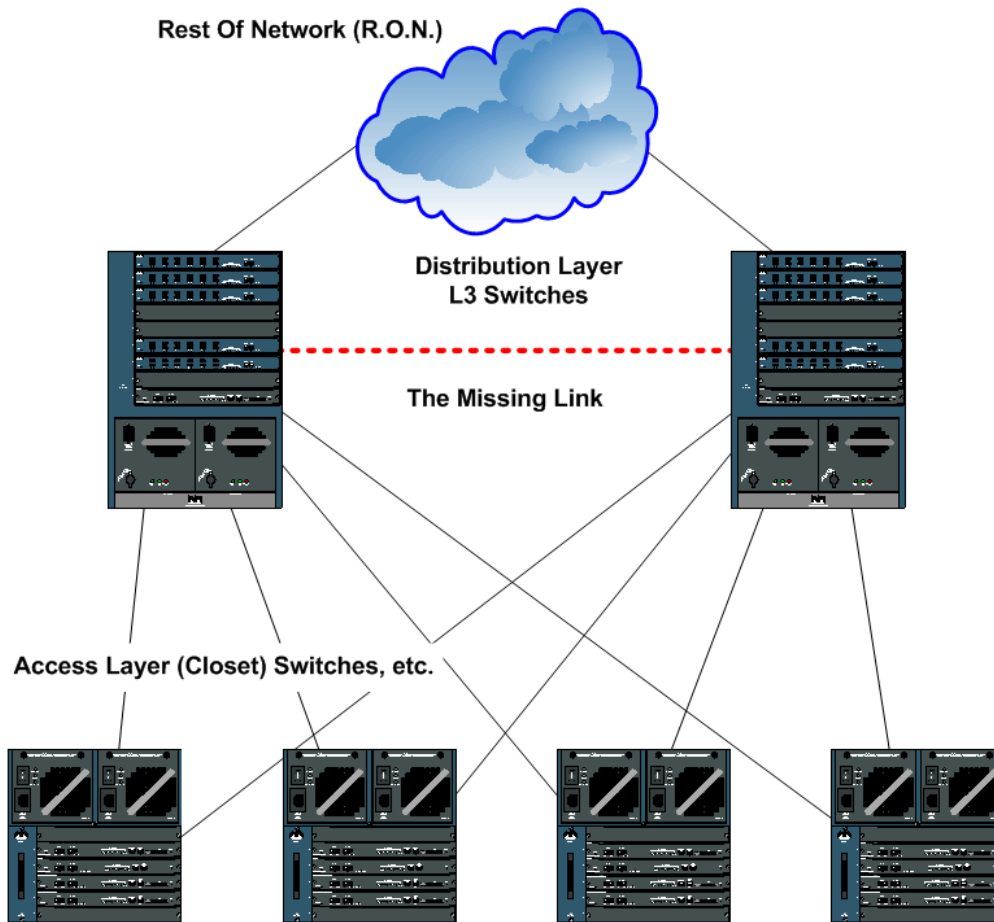
So this month we'll talk about something else.

I'd like to thank Chesapeake Netcraftsmen's David Yarashus for this month's topic. He and I have been working in several large enterprise and government organizations' designs using 3 or 4 layers of campus hierarchy. We've had some good discussions about why or why not should distribution layer switches be linked. Since he and I have both recently had fun evenings troubleshooting situations where the lack of a link was part of the problem, it seems like a timely topic to talk about. The topic may start off sounding a bit trivial, but I hope by the time we're done you'll see that it really does matter! And I do hope that along the way I manage to make it interesting as well.

The rest of this article discusses the question at hand, and then goes through some Real World situations where we've seen this problem arise.

The Basic Setting

To explain what the fuss is about, see the following figure. Recall that L2/L3 Hierarchical Design uses a L3 switches at the distribution layer, which might be for the entire building or for a group of floors, part of the building, etc. This approach has the virtue of modularity, so that you can use a cookie-cutter design.



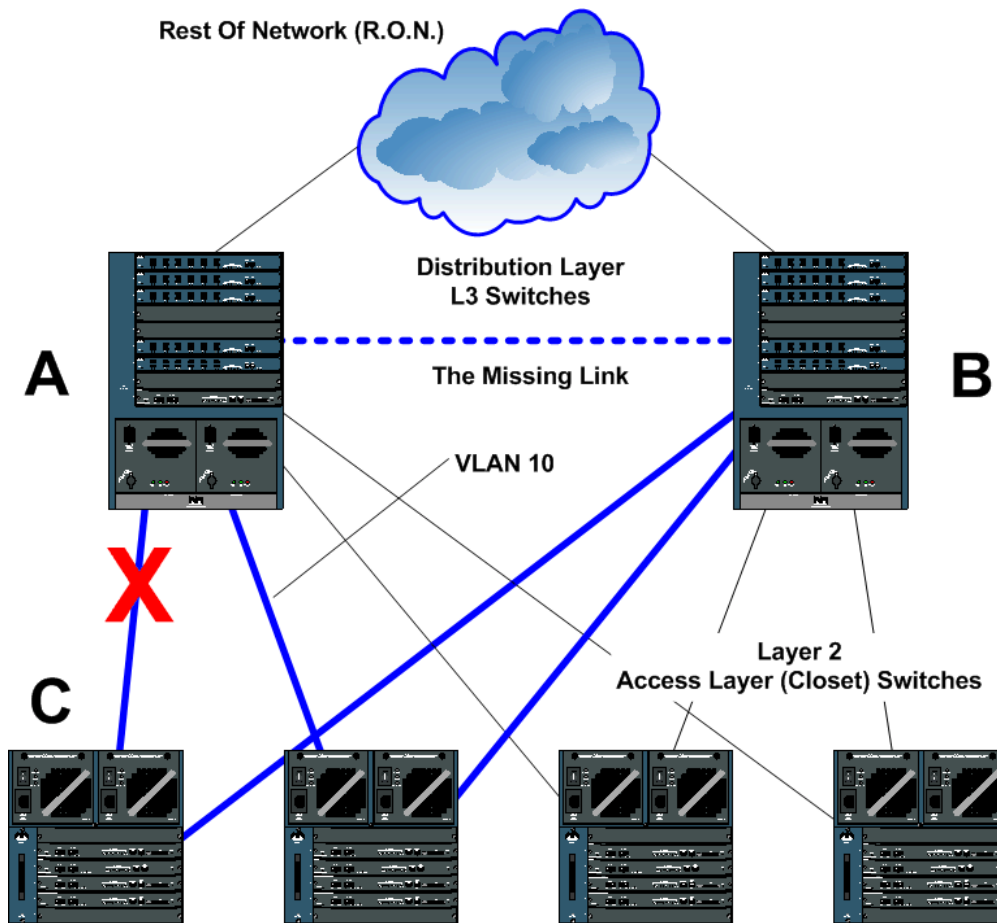
The big virtue to Hierarchical Design is limiting the size of spanning tree domains, so that L2 issues, broadcast storms, etc. don't affect too much of your network. I and my coworkers see security and some other folks now recommending disabling STP (Spanning Tree Protocol) on user and server ports. If you've ever had to troubleshoot a Spanning Tree Loop, you'll know why we strongly advise that you leave STP running on access ports. Think of it as insurance against having a really bad day (or week!). We've seen 3 or 4 enterprise melt-downs now, they aren't pretty, it takes considerable time to track down the cause, and they can be career-impacting events. We've seen too many shops say "we'll be careful and so we don't have to worry about spanning tree loops". Well, other people may connect or rearrange things. They might not be as careful. And it's too easy to have accidents. Conclusion: don't disable STP on switches!

Ok, if you have a couple of hundred users on two to four switches, this all may seem a bit abstract to you. I do recommend using a two layer hierarchy there, with say one or two core L3 switches and a bunch of access switches, sized to fit number of users, budget, and other needs. The key thing to realize in smaller networks is that networks tend to grow, sometimes quickly, and if you start out modular, you won't have to spend time reworking things. We've seen some ad hoc and daisy-chained switch networks that "just kept growing" over time. They may work, but they can also take substantial time to administer.

As more and more sites use L3 switches in wiring closets and elsewhere, the traditional link between distribution layer switches seems to be vanishing. We also see sites with L2 closets not using the link, which can be a rather bad idea as well, unless you have amazing levels of internal discipline. The next section explains why.

VLAN Black-Holing

For this section, please refer to the following figure.



Suppose the "Missing Link" isn't there. In this story, the access switches are Layer 2 only. Suppose the blue links are say VLAN 10, all in one subnet. Perhaps all the users on the connected access switches are in VLAN 10 as well. The two distribution switches are connected to that subnet. Since they are L3 switches (routers), they advertise the connected subnet to R.O.N. (the Rest of The Network).

The problem I'm about to describe can happen even if VLAN 10 only goes to one closet, but some other port somewhere on a distribution switch is in VLAN 10.

Suppose a failure occurs, as indicated by the red X. Because another port on switch A is in VLAN 10 and is up, the L3 interface VLAN 10 is also considered to be still up. So switch A continues advertising that subnet to RON. Switch B also advertises that subnet, since it too is connected to VLAN 10. As packets are sent from RON back towards users, the packets may go to A or B. But if they go to A, they have no way within VLAN 10 to reach switch C -- VLAN 10 has become discontinuous (not all in one piece). In that case, A has to drop the packets. This is known as "black-holing". They go into A and never come out. Even after the user MAC address ages out, flooding the packets through the parts of VLAN 10 that A is connected to doesn't reach switch C. Eventually, ARP ages out, and the ARP broadcast from A also fails to reach C.

Usually we design for this setting with the Missing Link a trunk between A and B. We then allow VLAN 10 across the trunk. Consider our failure scenario now. When the packet arrives at A, it is connected to the trunk, even after the outage. Flooding frames in the VLAN still reaches C. After STP settles, packets from the user will probably update the MAC tables in A, B, and C, providing an optimal L2 path from A to C and the end user.

Yes, you can plan to absolutely never ever put more than one port on a distribution layer switch into any closet/user VLAN. But will everybody faithfully do that? Will new employees get the word and understand the importance? If somebody fails to follow the plan, you won't notice until there's an outage and you get to figure out why the user lost connectivity, despite the redundant connections. Or maybe there's a department server you hook up to the distribution switch, and put into the user VLAN. In short, planning this way fails to tolerate exceptions. And while we try not to have to create them, some business or networking situations require exceptions, time after time.

By the way, this can also happen with L3 switches if some other port on the distribution switch gets put into the same VLAN as the link to the L3 switch. That's why with the newer Cisco code we prefer to make switch-to-switch links routed

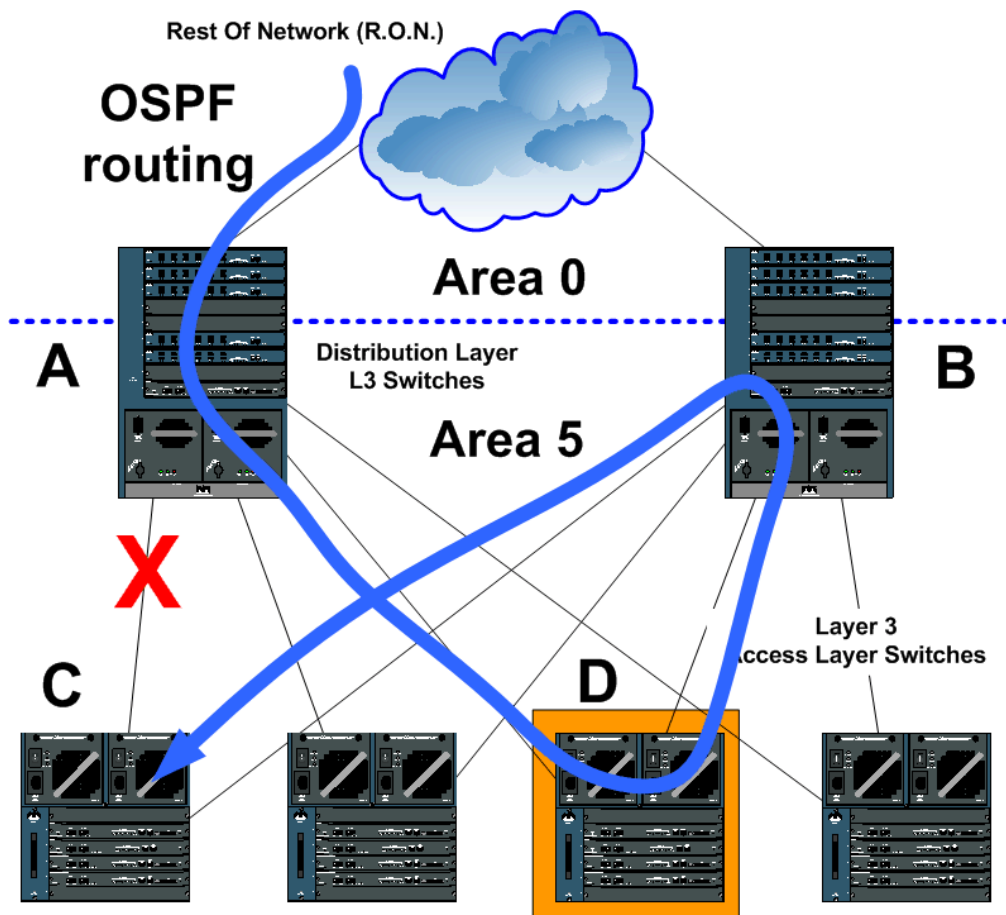
links, i.e. put the address on the Gig interface rather than on a VLAN interface.

Please note that I still do believe in portfast (and specifying non-trunking) for user ports. Portfast short-circuits some of STP startup. Whereas BPDU filtering disables the ability of STP to function, almost the same thing as turning it off.

Summarization Misses That Link!

Change the story now. For now, suppose the routing protocol is OSPF. Suppose some or all of the access layer switches are L3 switches, i.e. routers. Suppose the distribution switches are OSPF Area Border Routers (ABR's). This is commonly done to allow the subnets in the building or L3 module to be summarized outside the building. Suppose you're summarizing like that.

Now suppose a link failure occurs. You may see traffic following a path like that in the following picture.



What happened is that the packet arrived at switch/router A due to the summary route. Router A learned the specific subnet route from D who learned it from B. And B is connected to the relevant subnet, if switch C is L2-only. So traffic takes the indicated round-about path. This can be particularly disconcerting if one intended all the access switches to be non-transit. If the volume of traffic clobbers switch D's links, this can be disconcerting, especially if D happens to connect to the mainframe (as in one outage we recently saw).

Nothing said above is really OSPF-specific, other than the picture. **So the above might in fact also happen with EIGRP, assuming you're doing route summarization at the switches A and B (sort of an EIGRP ABR setting).** The problem is related to the summarization, not the routing protocol.

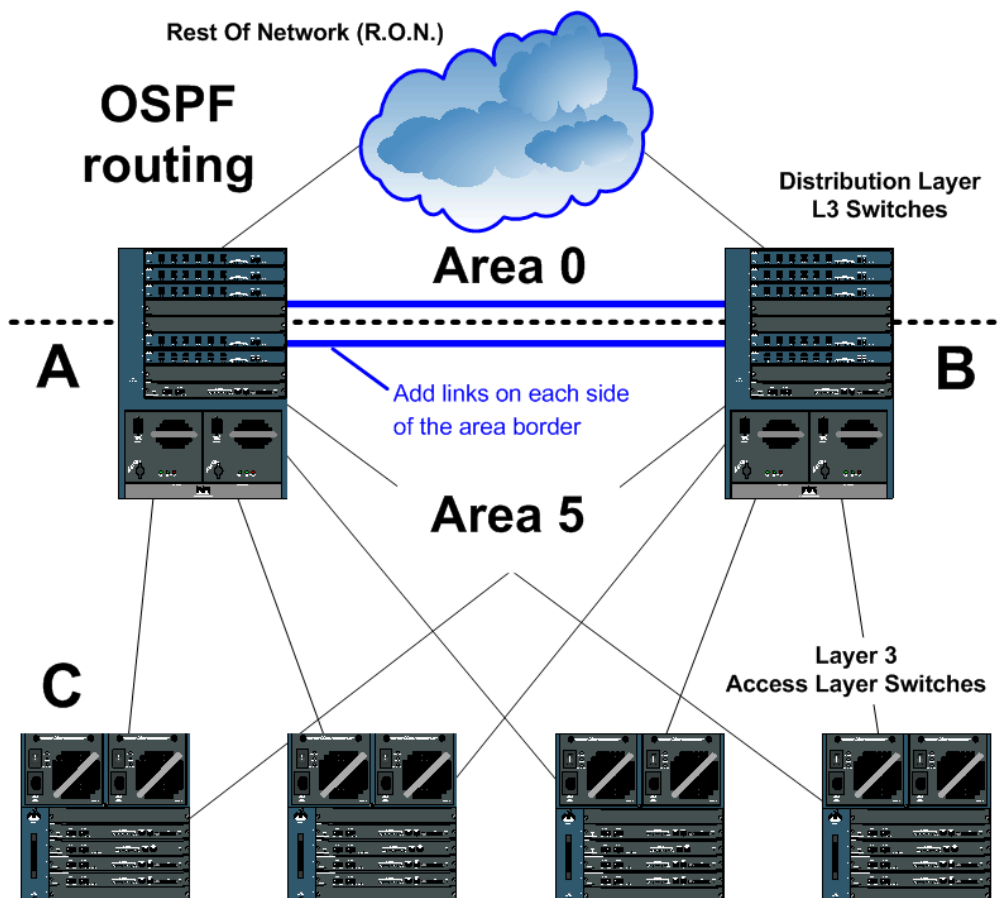
The reason the story started out with OSPF is that OSPF can take some really weird paths when there's an outage. The reason is OSPF's routing rules:

- 1 Traffic within a region cannot exit the region

- Traffic between regions must exit the first region, transit area 0 without detours via another region, then exit and travel the rest of the way within the destination region

The way to prevent OSPF weird paths is to link A and B in the picture, preferably with a link on each side of the area border. That way traffic can get from A to B within area 0 and also within area 5.

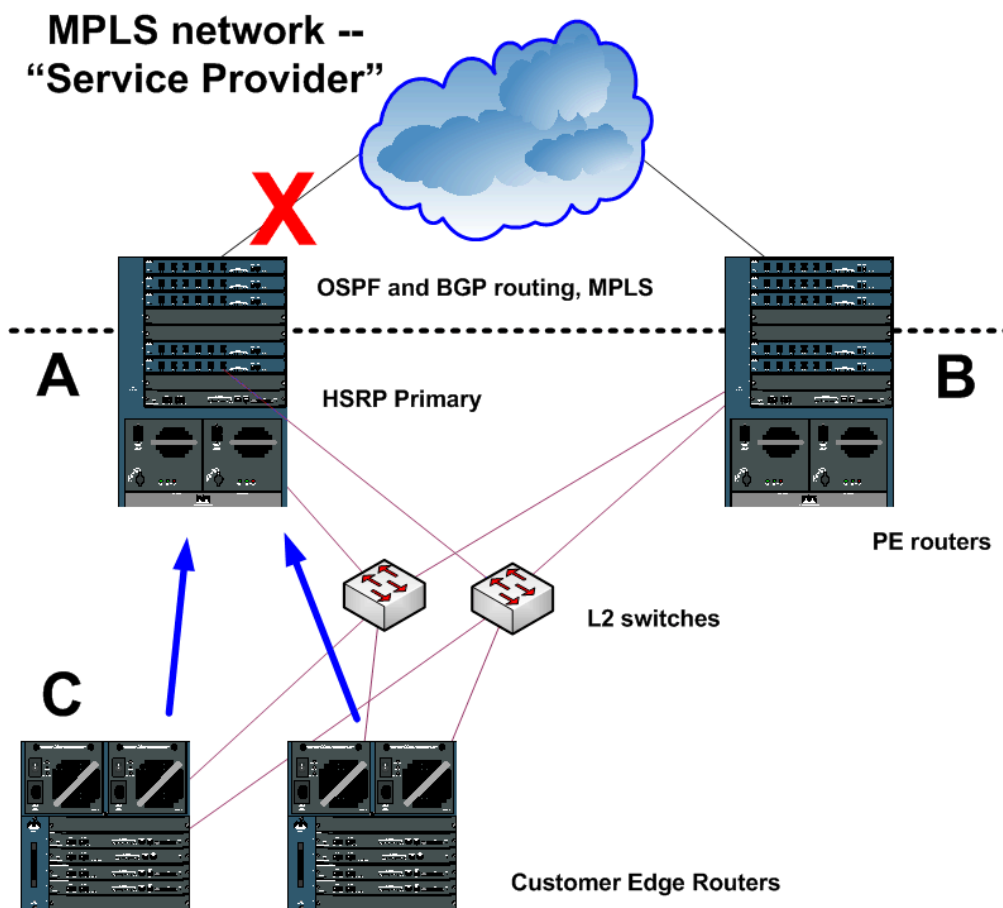
What about with EIGRP? It seems like if you're summarizing, it's a good idea to have a link within the "area 5" part of the above picture, i.e. a link between A and B that does not summarize the building routes. And make any routers in the "area" EIGRP stubs (non-transit routers).



We've seen a lot of sites going to "passive-interface default" for large L3 switches. This keeps from forming a lot of OSPF or EIGRP adjacencies. If you do this, you do have to remember to explicitly "no passive-interface ..." for the link between A and B. This is a nice control for switches with enterprises with most interfaces connected to user or server VLAN's. You un-passive the interfaces that connect to the outside world or the other routers.

HSRP and What It Misses

Now we get to a situation I ran into recently at 4 AM. The following figure illustrates the problem.



The network in question happened to have an MPLS VPN core running over optical gear, but while they added complexity and other factors to troubleshoot, they're irrelevant to the present story. A chain of events led to no OSPF adjacencies over the link with the red X (and one or two other places). My guess at the time was stuck OSPF state due to changes, probably taking out OSPF authentication. (The order in which you delete the commands may matter in some IOS releases).

Router A connected to a GSR and I didn't have permission to bounce the OSPF process on the GSR. The first focus was getting up and running. Router B was happily talking OSPF to the MPLS core network. It had an MPLS VPN configuration error at the time, but that's not really relevant here. For security and simplicity, the Customer Edge routers had static routes and we didn't have access to them at the time because another organization controlled them. But I guessed that the design had the HSRP address as next hop, and later it did turn out the static routes did in fact forward traffic to the HSRP primary, A. Some voice gateways in a separate VLAN had B as HSRP primary, and I was told that they were actually happily forwarding.

The problem was that the customer static routes were forwarding to A, as shown by the blue arrows. But A had no routes to anywhere but connected subnets. So it was dropping packets for other destinations. It had no way to pass them to B, because there was no OSPF on the customer side of routers A and B.

The solution at the time was to change the HSRP priorities in A and B to make B primary and allow B to preempt A. That swung the traffic from the customer edge routers over to B and got the customer up and running. That's how I determined that the guess about the HSRP address being the static route next hop was correct.

The interesting thing to me here is the design point it makes. There are two design methods that come to mind for avoiding this black-holing problem:

- 1 Connect A and B with a link inside area 0 (the provider side area for OSPF)
- 1 Have HSRP track state

The first of these is basically what we just went over in the Summarization section above. If A and B were connected and

running OSPF, then A would be able to route outwards via B. One reason this hadn't been done was apparently that A and B were 7301 routers, with 3 Gig links. These had been used up, one to the MPLS side and two to the local switches and VLAN's. So creating a link inside OSPF area 0 would have required adding another VLAN and running backbone traffic through the customer-side trunks and switches. This is enough of an exception that one can imagine it leading to difficulties, possible security concerns, etc.

The other approach would be to use HSRP tracking. This feature in the past allowed you to configure HSRP to track the state of a link. If the link went down, the HSRP priority would be lowered, allowing the other router to become the primary for HSRP. In this particular incident, the routers considered the A to Core link to be up. In fact, we used it to telnet to A from the NOC (telnet to the core GSR, then to the directly connected interface). However, HSRP can now track state of objects. So loss of routes or connectivity could be used to trigger HSRP failover. In the case above, the customer routers needed to establish a GRE tunnel to run IBGP over (I'm not going to go into why). So loss of a route to the other tunnel endpoint might have been an interesting object to track. Tracking routing state would not have sufficed, routers A thought OSPF was active but saw no neighbors.

We don't have space to go into details here. The object tracking feature was added beginning in Cisco IOS 12.2(15) T, with enhancements in 12.3 as well. Things you can track:

- 1 IP routing state of an interface
- 1 Line protocol state of interface
- 1 IP route reachability
- 1 Threshold for IP route metrics
- 1 Lists of the above and boolean expressions
- 1 SAA round-trip information

There's a new acronym to go with this: FHRP = First Hop Routing Protocol (HSRP, GLBP, VRRP). They can all track objects in the new code. Static and Policy-Based Routing can also do this!

Summary

I haven't seen most of this anywhere in print / web form, with the exception of the HSRP design issue noted above. Here are the links concerning FHRP's and tracking objects:

Tracking objects: <http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122newft/122t/122t15/ftshrptk.htm>

Standby track command:

http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122cgcr/fipras_r/1rfip2.htm#wp1021373

FHRP's and SAA:

http://www.cisco.com/en/US/products/sw/iosswrel/ps5207/products_feature_guide09186a00801d2d74.html

Objects and routing:

http://www.cisco.com/en/US/products/sw/iosswrel/ps5207/products_feature_guide09186a00801d1e95.html

http://www.cisco.com/en/US/about/ac123/ac114/ac173/Q2-04/department_techtips.html

http://www.cisco.com/en/US/about/ac123/ac114/ac173/Q2-04/department_techtips.html

This topic definitely feels like an article I'm going to want to write, after some lab time!

Your comments, questions, and suggestions for future articles are of course welcome! See below to decipher my email address.

Dr. Peter J. Welcher (CCIE #1773, CCSI #94014, CCIP) is a Senior Consultant with Chesapeake NetCraftsmen.

NetCraftsmen is a high-end consulting firm and Cisco Premier Partner dedicated to quality consulting and knowledge transfer. NetCraftsmen has ten CCIE's, with expertise including large network high-availability routing/switching and design, VoIP, QoS, MPLS, IPSec VPN, wireless LAN and bridging, network management, security, IP multicast, and other areas. See <http://www.netcraftsmen.net> for more information about NetCraftsmen. Pete's links start at <http://www.netcraftsmen.net/welcher> . New articles will be posted under the Articles link. Questions, suggestions for articles, etc. can be sent to [pjw <at> netcraftsmen <dot> net](mailto:pjw@netcraftsmen.net) (formatted this way to fool email harvesting software).

11/4/2004

Copyright (C) 2004 Peter J. Welcher