AVICI SYSTEMS

Convergence and Restoration Techniques for ISP Interior Routing

Curtis Villamizar <curtis@avici.com>

Abstract:

This talk is about:

- Convergence of IGP and BGP routes in IGP-only usage
- Convergence of MPLS LSP and BGP routing in MPLS usage
- Basic algorithms and time complexity (not too deep)
- Affect of network architecture on time complexity
- Improving algorithms performance (brief)

This talk is NOT about:

- Advocacy of use of IGP-only or use of MPLS
- BGP efficiency of BGP routing techniques
- Comparison of implementations

AVICI SYSTEMS

Extended Abstract - Part 1

Title:

Convergence and Restoration Techniques for ISP Interior Routing

Abstract (very extended abstract):

There are useful and quite general techniques that implementations can take advantage of for both IGP and MPLS convergence. Prior discussions at NANOG focused on incremental SPF. Other techniques can be used as well. For example, a two stage forwarding can allow the IGP route to change and the BGP routes that depend on it to follow with minimal changes to hardware forwarding information.

Tradeoffs exist between using only an IGP vs using MPLS. The IGP SPF takes order($Llog_2N$) which is further reduced by incremental SPF. MPLS present scaling limitation with respect to convergence due to the larger number of CSPF computation that may be required if the set of unique constraints differs. Problems can be avoided through MPLS implementation techniques and network design techniques.

Fast convergence for an IGP is a means to achieve faster restoration when a fault occurs. MPLS has restoration capabilities worth considering.

(for the benefit of those reading slides at home)

AVICI SYSTEMS

Extended Abstract - Part 2

MPLS fast reroute allows fast convergence with a complexity cost in terms of a larger number of LSPs being required. Standby LSP can support sub-second recovery with a bit less complexity. Rerouting LSPs from ingress can be done in seconds for most topologies. This latter case is where MPLS scaling issues come into play. SPF results caching and incremental CSPFs are among the techniques that can alleviate scaling problems but these too have limits.

There are problems to be solved to make incremental CSPF practical such as finding similar CSPF results and quickly determining which links would differ for the purpose of incremental CSPF. Adjusting a CSPF for the current path of an LSP when considering rerouting it and the adjusted CSPF needed for disjoint paths present other complexities for which there are solutions.

Topology and protocol usage also affects scaling on the IGP and of MPLS. Implications of area size and LSP tunneling are discussed.

About the Author:

Curtis Villamizar has been involved in Internet operations, and protocol design and implementation since working for ANS in 1992-1997 in support of the NSF funded T3-NSFNET project and later ANSNET. In 1997-1999 Curtis was part of the UUNET Network Architecture Group. In 1999 Curtis joined Avici where he is presently Principal Design Engineer and responsible for Avici's MPLS/TE implementation.

(for the benefit of those reading slides at home)

AVICI SYSTEMS

Convergence - changing state of the art

Question: Pull or cut the fiber, or reload a router, and how long does it take before all traffic is successfully being delivered?

Answers:

circa 1995: With 40,000 BGP routes, one to three minutes depending on whose router you use or longer if other routers start crashing.

circa 1998: With 150,000 BGP routes, 20-30 seconds to three minutes depending on how you set up your network, whose routers and how old your routers are.

circa 2002: With 250,000 BGP routes, 1-5 seconds to three minutes depending on how you set up your network, whose routers and how old your routers are.

Note: At any given time (including now) you can set up your network sufficiently badly that your network becomes unstable and convergence takes an indefinitely long time.

AVICI SYSTEMS

Hypothetic Network for Discussion

- 250K unique BGP routes
- 1,000 node core
- either:
  - full mesh, or
  - 10 links per router on average
- either:
  - one 1,000 node IGP area, or
  - 10 areas, 100 nodes per area

Note: these are just hypothetic round numbers (for example, mostly powers of 10) but roughly the size of large (tier 1) ISPs.

AVICI SYSTEMS

Two Network Designs

We will consider the following two cases:

1. BGP and IGP-only (no use of MPLS)
2. When MPLS is used for IGP traffic engineering

Note: Lets not get hung up on discussion of which of the above two is better. This is not an advocacy talk.

Terminology: I will use LSA to mean OSPF LSA or ISIS LSP fragment and spell out MPLS LSP to avoid confusion with ISIS LSP.

AVICI SYSTEMS

Scaling of IGP Flooding (Katz Effect)

When a link fails:

1. Two routers send a LSA to "A" adjacent routers.
2. Each of "A" routers sends the LSA to "A" adjacent routers.
3. In total each router receives on the order of A copies of 2 LSA. (order(A)).

When a router fails:

1. Each of "A" adjacent routers originates a new LSA and sends it to the "A" routers it is adjacent to.
2. Each of receiving router sends to the "A" routers it is adjacent to.
3. In total each router receives on the order of A copies of A LSAs. (order(A^2)).

Note: Dave Katz pointed this scaling effect out at a prior NANOG panel on scaling. This effect is believed to be the cause of severe instability of a tier 1 ISP large full mesh cores built over ATM in the mid 1990s. No news here...

AVICI SYSTEMS

Scaling of IGP SPF

- John Moy's OSPF book estimates the SPF to be order($L\ log_2 N$) for L links and N nodes.
- Packet Designs has long advocated use of the faster incremental SPF.
- ISPF only operates on the subtree downstream of the affected link and is much more efficient in all topologies except full mesh.
- Packet Designs has shown that flooding before SPF promotes fast flooding.
- An obvious question is "if flooding is fast and the full SPF is fast (under a second for sloppy implementation in almost any topology), then why does it take 3 minutes for some routers to recover?"

AVICI SYSTEMS

Scaling of BGP route install (IGP-only)

- Packets don't stop dropping when the SPF is done.
- Installing 250K BGP routes dominates recovery time.
- Solution is to make route install faster.
  - just improve transfer and install time
  - and/or two stage route lookup:
    BGP route
    → IGP destination (about 10,000 entries)
    or IGP node (100 with areas or 1,000 without areas)
    → next hop
  - or three stage lookup.
    BGP → IGP dest → IGP node
- Multistage lookup handles route change except where IGP cost causes BGP to change entry point selection.
- Second stage sub-second route install is achievable.

AVICI SYSTEMS

MPLS CSPF Scaling

- SPF requires storage of path information, not just next hop. (Can take longer)
- If constraints differ (ie: different bandwidth or color constraint or priority) CSPF differs.
- Potentially 1 CSPF per MPLS LSP.
- There are ways to get around doing 1 CSPF per MPLS LSP as long as groups of LSPs share common color constraints and priority. (discussed later)

AVICI SYSTEMS

Scaling of route install (MPLS)

- Two stage route lookup is applicable
  BGP route
  - → MPLS tunnel (100 to 1,000 entries)
  - → MPLS LSP instantiation (route)
- MPLS LSP can be configured with fixed LSP cost to avoid any BGP route change at all.
- MPLS LSP can be advertised into the IGP as a fixed cost adjacencies to prevent BGP route change in non-MPLS routers.
- CSPF time and LSP setup time can dominate. Ways to improve this are discussed later.
- FRR or standby LSP eliminate dependency on CSPF time and LSP setup for initial recovery.

AVICI SYSTEMS

MPLS Fast Recovery

- Fast Reroute (aka Local-Protect) can achieve under 50 msec fault recovery using presignaled backup installed at each potential point of local repair (PLR).
- Standby LSPs originate at the ingress, are presignaled and disjoint with the primary and can achieve sub-second repair.
- Precomputed alternate paths save recovery time at the expense of very poor initial recovery layout.
- Computing CPSF and signaling LSPs is typically comparable in recovery speed to IGP-only, but can be slightly better to much worse depending on topology.

AVICI SYSTEMS

IGP-only Recovery from Fault

1. advertisements are originated
2. advertisements are flooded
3. one SPF is required per router
4. IGP routes installed ← 2 stage forwarding recovery
5. BGP routes are mapped onto IGP routes
6. BGP routes installed ← 1 stage forwarding recovery

AVICI SYSTEMS

MPLS Recovery from Fault

1. fault detected ← FRR recovery
2. advertisements are originated
3. advertisements are flooded
4. ingress infers LSP infeasible ← standby recovery
5. up to one SPF is required per MPLS LSP
6. new LSPs are signaled
7. signaling completed ← recovery w/o backup
8. backup LSPs are established if using FRR or standby
9. LSP paths are optimized (using make-before-break)
10. new backup is established if primary is moved.

AVICI SYSTEMS

Approaches to Improve Scaling

These fall into one of two broad categories:

1. Network Topology.

2. Router Implementation.

AVICI SYSTEMS

IGP Flooding - Impact of Network Topology

advertisements received per router for 1,000 routers

no IGP areas ($N = 1000$)	Full mesh	link down	$(2N)$	2,000
		router down	(N^2)	1,000,000
	Partial mesh 10 links / router	link down	$(2A)$	20
		router down	(A^2)	100
100 routers per IGP area ($N = 100$)	Full mesh	link down	$(2N)$	200
		router down	(N^2)	10,000
	Partial mesh 10 links / router	link down	$(2A)$	20
		router down	(A^2)	100

Q: Why would anyone want to use a full mesh these days?
A: Optical core or Optical/TDM region. (not ATM).

Q: Is there a better way?
A: Some say GMPLS solves everything. (discussion $\gg$ /dev/null)

AVICI SYSTEMS

IGP SPF - Impact of Network Topology

Low level operation for each of 1,000 routers

no IGP areas ($N = 1000$)	Full mesh $L \simeq N^2$	$SPF \simeq N^2 log_2 N$ 10^7
	Partial mesh 10 links / router $L \simeq 10N$	$SPF \simeq 10N\ log_2 N$ 10^5
100 routers per IGP area ($N = 100$)	Full mesh $L \simeq N^2$	$SPF \simeq N^2 log_2 N$ 7×10^4
	Partial mesh 10 links / router $L \simeq 10N$	$SPF \simeq 10N\ log_2 N$ 7×10^3

- Multiplier could be 100 nsec to 10 usec.
- For example, multiplier of 1 usec yields:
  $10^7 \rightarrow 10$ sec, $10^5 \rightarrow 100$ msec, $7 \times 10^3 \rightarrow 7$ msec

AVICI SYSTEMS

MPLS CSPF - Impact of Network Topology

Low level operation for each of 1,000 routers

no IGP areas ($N = 1000$)	Full mesh $L \simeq N^2$	$NCSPF \simeq N^3 log_2 N$ 10^{10}
	Partial mesh 10 links / router $L \simeq 10N$	$NCSPF \simeq 10N^2 log_2 N$ 10^8
100 routers per IGP area ($N = 100$)	Full mesh $L \simeq N^2$	$NCSPF \simeq N^3 log_2 N$ 7×10^6
	Partial mesh 10 links / router $L \simeq 10N$	$NCSPF \simeq 10N^2 log_2 N$ 7×10^5

- For example, multiplier of 1 usec yields: $10^{10} \rightarrow$ 2 hr 46m, $10^8 \rightarrow$ 100 sec, $7 \times 10^5 \rightarrow$ 700 msec
- Looks like a bit of a complication.

AVICI SYSTEMS

MPLS CSPF - a few more complications

- After all LSP are rerouted, they need to be rechecked for optimal routing.
  - Leave feedback and timing for another talk.
  - For each LSP before computing the SPF, bandwidth used by the current path is subtracted.
  - Each SPF is unique after bandwidth is removed.
- With FRR or standby, after rerouting the primary the backup must be rerouted.
  - Computing a disjoint path requires modifying the link set to reflect the primary path and all links that share SRLG.
  - Each of these SPF will be unique.
- Note: Both are candidates for incremental CSPF.

AVICI SYSTEMS

IGP Scaling - Router Implementation

- Mesh groups help (but not solve) the full mesh case.
- Give priority to IGP reflooding over IGP SPF (observation made by Packet Design).
- Implement incremental SPF. (observation made by many, pointed out by Packet Design).
- Flooding can be faster if handled by the line cards.
- Decoupling IGP and BGP processing can help.
- Improve route install time.
- Use two or three stage lookup if possible.

AVICI SYSTEMS

MPLS Scaling - Router Implementation

1. Cache the results of applying constraints to the links (note: 10,000 links if areas are not used). These are called "link sets".
   - Adjust link sets as links change rather than recompute from scratch.
   - Obtain link set deltas if LSP bandwidth differs but all other constraints are the same.
   - Obtain link set deltas after adjusting for current path or for disjoint path computation.
2. Cache SPF results. Allow cache search to obtain the most similar result differing only in bandwidth.
3. Use incremental CSPF to convert from link set with known CSPF result and delta to new CSPF result.

AVICI SYSTEMS

Brief Description of CSPF Caching

1. Find the link set cache entry for the set of constraints with the closest bandwidth available.
2. Apply constraints to the links that have changed since the link set was created. This yields a delta due to change in availability or metric and a new link set.
3. Look up CSPF result for prior link set.
4. Apply deltas and do an incremental SPF.
5. Cache the new link set and CSPF result.
6. If needed adjust for existing path and disjointness and apply very small additional deltas and do an incremental SPF.

AVICI SYSTEMS

Observations on CSPF Caching

1. If constraints are identical and feedback is slow exactly one SPF is needed for initial recovery.
2. If constraints differ only in bandwidth and network is underutilized deltas are the null set.
3. If feedback arrives and links utilization is still low, link set are updated but deltas are the null set.
4. In trying simulate customer topologies and network scenarios, cache hits rates are extremely high. Very few SPFs are run. (Simulation test cases are somewhat limited, though 1000+ node simulations have been done).
5. Performance becomes limited by efficiency of link set processing (which is order(L) but with smart coding can be reduced to order(ΔL).

AVICI SYSTEMS

MPLS Scaling - Alternate Approaches

1. Sort LSP that need SPF according to similar constraints.
2. Apply constraints and compare to previous link set.
3. Use incremental CSPF to convert from link set with previous CSPF result and delta to new CSPF result.

- By only saving one CSPF result, avoid any dynamic allocation needed to save CSPF results in cache.
- Some dynamic allocation may still be needed for adequate support of multipath.

AVICI SYSTEMS

MPLS Scaling - Hierarchical LSPs

- Why bother?
  (Good question.)
- Areas provide good scaling. Hierarchical LSPs within an area (a core subset of the area) can provide further scaling.
  (Anybody that big yet?)
- LSP hierarchy can achieve E2E TE LSPs.
  (Does anyone want that?)
- Some way is needed to support Optical and Optical/TDM technologies.
  (Assuming someone wants that.)
- Technology ahead of its time?
  (Good stuff though - IMHO).

AVICI SYSTEMS

Conclusions

wouldnt be a nanog if it didnt have at least one research network explaining what they are currently smoking/injecting. – *Ed Kern*

- This was arguably a little technical for NANOG.
- Some background info on algorithms never hurt.
- It helps to know the impact of topology decisions. Are you unnecessarily beating on your routers?
- Algorithm discussion was intended to be brief.
- Test and simulation results were not presented to maintain NANOG no product plug traditions.
- For those who are interested, there may be a detailed follow up at the MPLS Conference in October.

(OK Ed - I know its not about a research network.)