OSPF Optimized Multipath (OSPF-OMP)

Curtis Villamizar <curtis@ans.net>

- URLs (including earlier simulations):
  - http://engr.ans.net/mpls-omp
  - http://engr.ans.net/ospf-omp
  - http://engr.ans.net/isis-omp
- Internet Drafts:
  - draft-ietf-ospf-omp-01.txt,ps
  - draft-villamizar-isis-omp-00.txt,ps
  - draft-villamizar-mpls-omp-00.txt,ps
- This is –still– "work in progress".

Draft Status: draft-ietf-ospf-omp-01.txt,ps

- Changes since the -00 version:
  - Document reorganized for better readability
  - Minor changes to parameters from simulation experience
  - Option to relax SPF best path criteria (Dave Ward)
  - Pseudcode included in appendices to aid implementors
- Upcoming changes:
  - Optimization of partial paths
- Multiple vendors considering implementation.
  - Not sure if any have started coding.

OMP Algorithm Highlights

- Flood Loading Information via OSPF or IS-IS
  - Router samples own SNMP counters every 15 seconds
  - Filters and floods depending on load level and change
- Forwarding (OSPF-OMP, ISIS-OMP, MPLS-OMP at ingress)
  - Compute Hash on IP source/destination
  - Select from available paths based on hash value
  - 14-16 bit hash provides fine adjustment granularity
- Load Adjustment (OSPF-OMP, ISIS-OMP, MPLS-OMP)
  - load adjustment through change in hash boundary
  - small initial adjustment
  - exponential increase in adjustment increment
  - increment is halved when adjustment reverses
- Path Setup (MPLS-OMP only)
  - Setup new paths after persistent high utilization
  - Remove extra paths after persistent low utilization
  - TE does not depend on careful configuration of IGP link metrics

Flood Loading Information via OSPF or IS-IS

- Router samples own SNMP counters every 15 seconds
  - Counters are if{In,Out}{Octets,Packet,Discard}
- Filter using a few compare, shift, and add operations
- Compute “equivalent load” as described in OSPF-OMP
- Check for reflooding based on:
  - Time elapsed since last flooding
  - Greater of current load and last flooded value
  - Percent change since last flooded value
- When needed, reflood and record time and value

Forwarding (OSPF-OMP, ISIS-OMP, MPLS-OMP at ingress)

- After SPF create “next hop” data structures
- Compute Hash on IP source/destination per packet
  - 14-16 bit hash provides fine adjustment granularity
  - CRC16 seems to work fine. Others may be used.
- Select from available paths based on hash value
  - Compare hash value to boundaries in “next hop” struct
  - Forwording is like ECMP except load split is unequal
  - Adjustments to boundaries will adjust load split
- Characteristics:
  - No transient routing loops or drops
  - No microflow packet reordering except during adjustment
  - Adjustments are every few minutes (minimal reordering)

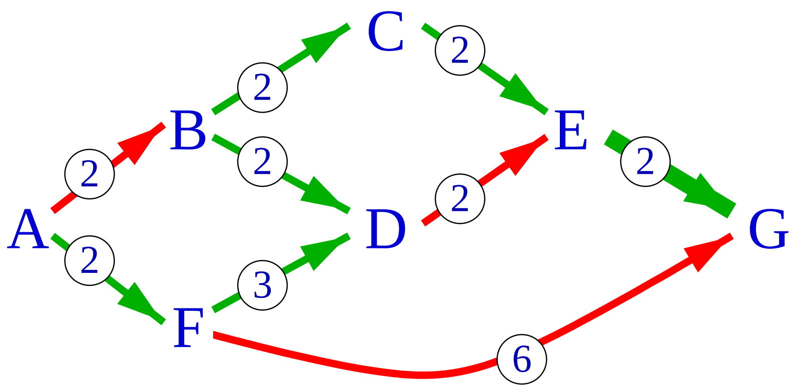
Load Adjustment (OSPF-OMP, ISIS-OMP, MPLS-OMP)

- Load adjustment through change in hash boundary
- Initial adjustment are very small (default is 1%)
- Additional adjustments are made:
  - When loading on the most heavily loaded link is reflooded
  - After timers expire and no change is reported
- The adjustment increment increases exponentially
  - When significant adjustment occurs, flooding is forced
  - Flooding will either accellerate or reverse adjustment
- Some overshoot can occur when traffic rapidly ramps up
- When adjustment reverses, adjustment increment is halved
  - Halving the rate on reversal insures stability
  - Stability has not been mathematically proven, but simulation results strongly indicate stability
  - When load stabilizes, flooding rate also drops

OSPF-OMP when used with MPLS-OMP

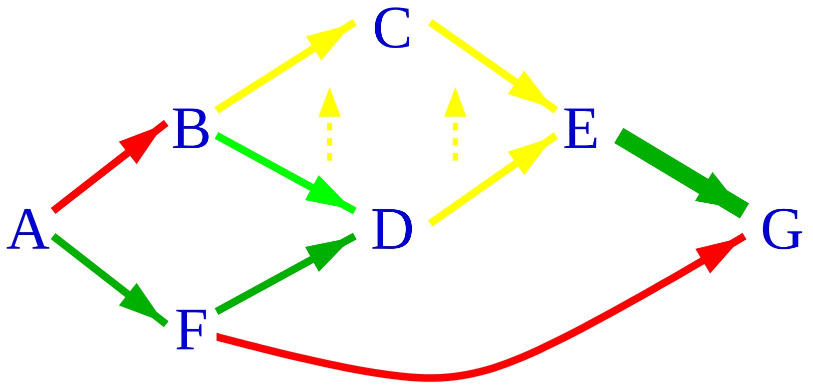
- Besides doing what OSPF normally does, an interior router in an MPLS-OMP domain does the following:
  - Sample its own SNMP counters every 15 seconds.
  - Apply simple filter to SNMP sampled data.
  - Determine when to flood flitered result according to guidelines in OSPF-OMP
- Ingress routers must also do the following:
  - Setup MPLS LSP path sets according to MPLS-OMP
  - Adjust loading on path sets according to OSPF-OMP
  - Hash on src/dst and forward according to OSPF-OMP

A Simple Example



- Major ingress and egress are A, F and E, G
- Major flows are A-E = F-G = F-E = 0.5, A-G = 1
- Link E-G is double capacity of others
- Link costs are as shown in the circles
- Utilizations: Red = 1, Green = 0.5

First Opportunity for Load Adjustment

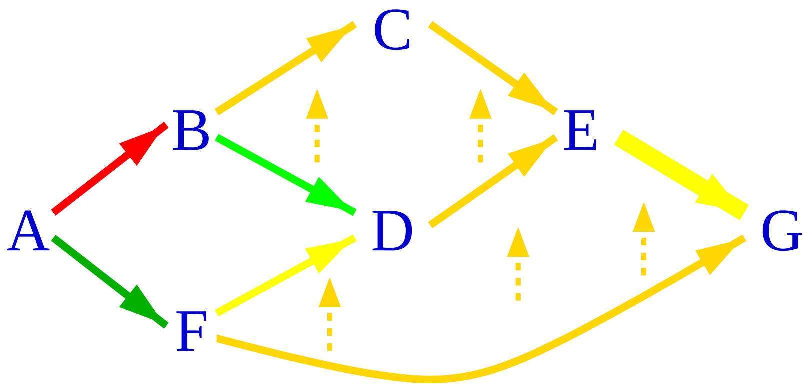


- Node B can move load from B-D-E to B-C-E

- Utilizations of B-C, C-E, and D-E approach 0.75

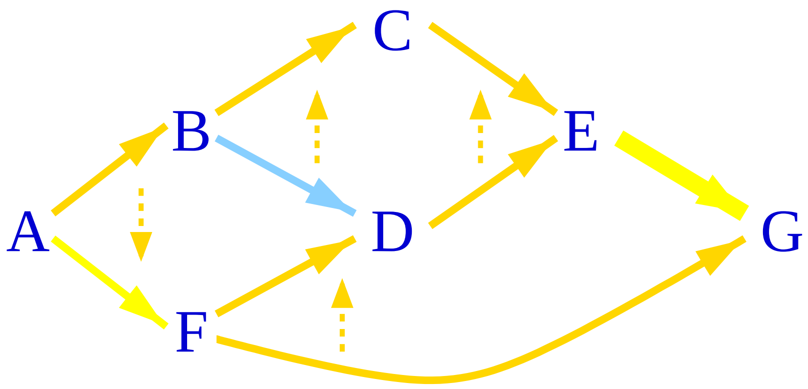
- Utilizations of B-D drops to 0.25

Second Opportunity for Load Adjustment



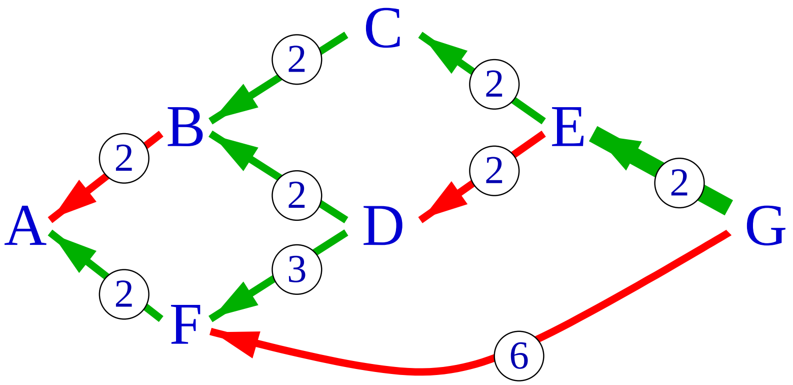
- Node F can move load from F-G to F-D-E
- Node B will continue to move load from B-D-E to B-C-E
- Utilizations of B-C, C-E, and D-E, F-G approach 0.83
- Utilizations of F-G and E-G approach 0.67
- Node F will actually not wait until D-E loading has reached 0.75, it will start moving load when D-E loading is noticed to be lower than F-G

Second Opportunity for Load Adjustment



- Node A can move load from A-B-{CD}-E-G to A-F-G
- Node F will continue to move load fro F-G to F-D-E
- Node B will continue to move load from B-D-E to B-C-E
- Utilization of A-B will approach 0.83
- Utilizations of F-G and E-G also approach 0.83
- Node A will start moving load when F-G loading is noticed to be lower than A-B
- A-F goes to 0.67, B-D approaches zero

The Need for Partial Path Optimization



- Consider traffic in the reverse direction
- Worst loading on the E-C-B-A path load of 1.0 on D-A
- Worst loading on the E-D-B-A path is on D-A and E-D
- Moving load from E-D-B-A to E-C-B-A does not reduce load on the link D-A so it does not reduce the load on E-D-B-A.
- Node E will not move load from E-D-B-A to E-C-B-A

Validating the Algorithms

- Simulations are at http://engr.ans.net/ospf-omp
  - tutorial directory has simple examples
  - simulations directory has larger topologies
  - simulations directory has adverse conditions cases
    - link failure
    - fast rise in offerred load
    - high noise in offerred load
    - large drift over time
- Simulations coverage:
  - OSPF-OMP is completely covered.
  - ISIS-OMP is not implemented at all.
  - MPLS-OMP LSP deletion is not implemented.
  - MPLS-OMP link failure is not implemented.
  - MPLS simulations are not yet on the web page.
  - If UUNET simulations cannot be made available, simulations using a complex hypothetical topology are needed.

Summary

- Algorithms are being validated through simulations.
  - Most are available at http://engr.ans.net/ospf-omp
- Status: draft-ietf-ospf-omp-01.txt,ps
  - Little change to the algorithms since -00 version.
  - Substantial improvement to the document since -00.
  - Optimization of partial paths needs to be added.
  - Another iteration of the draft is needed.
  - Comments would be nice. Implementations nicer.