Flow label for equal cost multipath routing in tunnels

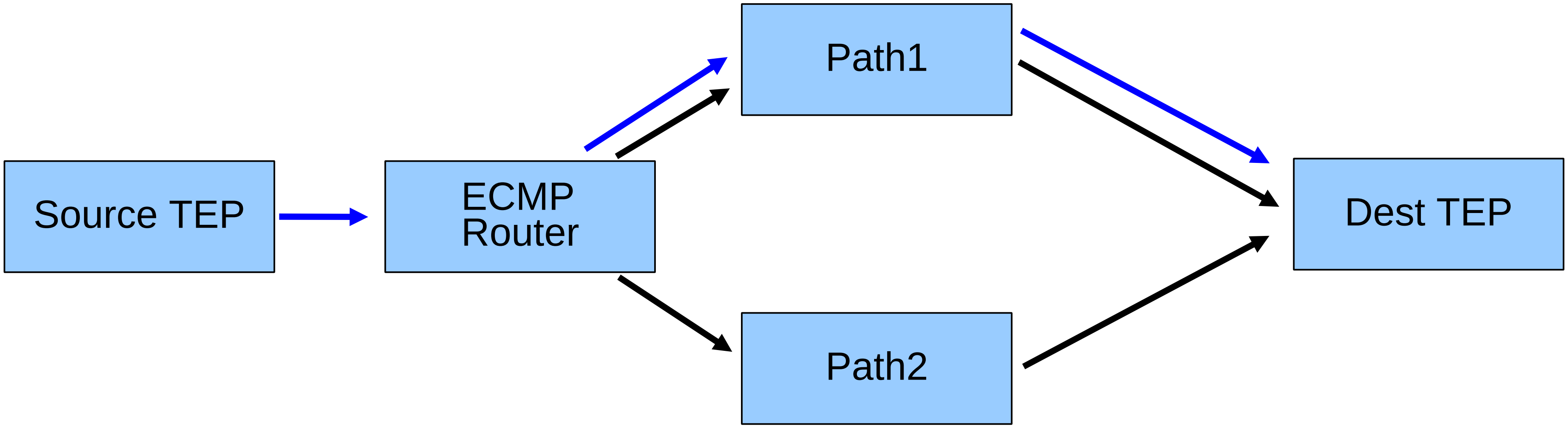
draft-carpenter-flow-ecmp-03

Brian Carpenter
University of Auckland

Shane Amante
Level 3 Communications, LLC

November 2010

The problem with tunnels



Normal traffic split by ECMP.
Tunnel traffic all has same 5-tuple; no split.

Proposed solution

- For foo-in-IPv6 tunnels, the *source TEP* sets a flow label per user flow in the *outer packet*
  - For IP-in-IPv6, the flow label is based on the 5-tuple of the *inner packet*
  - It should be well distributed (pseudo-random)
- Intermediate ECMP or LAG paths use hash based on 6-tuple (the normal 5-tuple plus the flow label)
  - works the same as before for non-tunnel traffic
  - also splits tunnel traffic
  - fully conformant with RFC 3697
- *Caveat:* hashing the flow label would not work in Inter-AS scenarios if it is allowed to have local semantics.

Changes from -02 to -03

- “The flow label in the outer packet SHOULD be set by the sending TEP to a pseudo-random 20-bit value” (was MUST)
  - “Note that this rule is a SHOULD rather than a MUST, to permit individual implementers to take an alternative approach if they wish to do so. Such an alternative MUST conform to [RFC3697].”
- Editorial and clarification fixes

Proposal

- Adopt **draft-carpenter-flow-ecmp** as 6man WG document.

Update to the IPv6 flow label specification

draft-carpenter-6man-flow-update-04

Brian Carpenter
University of Auckland

Sheng Jiang
Huawei

Shane Amante
Level 3 Communications, LLC

November 2010

Why?

- ***RFC 3697 says:***
  - ***Flow label must not be changed en route.***
  - ***Nodes must not assume any mathematical or other properties of Flow Label values***
  - ***Router performance should not depend on the distribution of Flow Label values... Flow Label bits alone make poor material for a hash key.***
- ***These rules have caused difficulty for almost all proposed use cases***.

History

- Versions -00 to -03
  - Allow local semantics for flow-label
  - Required reset of flow-label on exit from a domain
  - Downstream AS could easily misinterpret label
  - Vigorous discussions at two IETFs and on 6man list
  - Judged operationally challenging, no consensus
- Now a -04 version
  - Goodbye local semantics
  - Recognise consequences of flow label being unprotected (forgeable)
  - Recognise preferred usage for load balancing
  - Specific but modest changes to RFC 3697

Several challenges with IPv6 flow-label

- (-) Largely unused by both hosts and routers
- (-) No integrity 'guarantee' of flow-label
  - Not protected by header checksum
  - (Outer header) flow-label not protected by IPSec
- (+) Fixed location in header make it straightforward for [very] high-speed routers to use as input-key for LAG and/or ECMP versus:
  - (-) Variable offset of "Next Header" containing Transport protocol info {proto, src_port, dst_port}
  - (-) Brittle nature of existing "Next Header" that do not have TLV structure. Thus, unknown next-headers *cannot* easily be skipped over to find input-keys for ECMP or LAG[1].

[1] draft-krishnan-ipv6-exthdr could fix this, assuming it is moving forward (?)

Tentative conclusion

1. Local flow label semantics considered harmful
   - Operationally challenging to restore or reset flow label at FL domain *exit* routers
   - Nowhere to store an existing flow label value inside a packet at FL domain ingress
   - No guarantee FL *exit* router will (be properly configured to) restore/reset flow label
2. No integrity protection of IPv6 flow label
   - Therefore, flow label viewed as suspect at a security boundary
3. Conclusion: the flow label is a best effort field with best effort end-to-end semantics

Recommendations in 04 (1)

- Redefine a flow as "a sequence of packets sent from a particular source to a particular unicast, anycast, or multicast destination that a node desires to label as a flow."
  - Change from RFC 3697 is node instead of source, so that an ingress router may set the flow label.
- RECOMMENDED that source hosts set the flow label field for all packets of a flow to the same pseudo-random value.
  - Change from RFC 3697 is to specify a pseudo-random value as the preferred method.
  - The draft-gont-6man-flowlabel-security algorithm MAY be used

Recommendations in 04 (2)

- A node forwarding a flow, whose flow label in arriving packets is zero, MAY set the flow label value. It is RECOMMENDED to set the flow label to a pseudo-random value.
  - New compared to RFC 3697.
- In general, a forwarding node MUST NOT change the flow label in an arriving packet if it is non-zero. But:
  - A domain border device MAY be configured to set the flow label value in incoming packets to zero. *[Should we say this? It’s contentious, but firewalls might do it anyway. Nullifies inter-AS usage.]*
  - A network domain MUST NOT forward packets outside the domain whose flow labels are other than zero or pseudo-random. *[Backstop rule for sites that break other rules.]*
    - New compared to RFC 3697.

Recommendations in 04 (3)

- IPv6 nodes MUST NOT assume that the Flow Label in an incoming packet is identical to the value set by the source node.
  - Even though the flow label is in general immutable, this is not guaranteed in real life, hence this rule.
  - Replaces a wishy-washy rule in RFC 3697.
- Nodes such as load balancers MUST NOT depend only on Flow Label values being randomly distributed.
  - In usage like a hash for load balancing, the Flow Label bits MUST be combined with other bits in the packet to produce a good distribution of hash values.
  - Replaces another wishy-washy rule in RFC 3697.

Discussion

- These proposals modify strict immutability, but in a restricted way:
  - A network domain can include routers that set flow labels on behalf of hosts that don't.
  - A domain can be protected at its border (if desired) by clearing untrustworthy flow labels.
  - Flow labels exported to the Internet must always be either zero or pseudo-random.
- Hosts and routers that ignore the flow label will be unaffected.
- The flow label is no longer asserted to be strictly e2e immutable (as a matter of realism).
- The expected default usage of the flow label is some form of load balancing, e.g. ECMP/LAG

Proposal

- Adopt **draft-carpenter-6man-flow-update** as 6man WG document (Informational)
- Then start work on RFC3697bis (Standards track)

Thank You!