Layered Encapsulation of Congestion Notification

draft-briscoe-tsvwg-ecn-tunnel-01.txt

Bob Briscoe, BT
IETF-73 tsvwg Nov 2008
status

• Layered Encapsulation of Congestion Notification
  • new WG draft: draft-ietf-tsvwg-ecn-tunnel-01.txt as of late Oct'07
  • previously: draft-briscoe-tsvwg-ecn-tunnel-01.txt
  • intended status: standards track
  • RFC pub target: ? TBA
  • immediate intent: discuss including fix to decap as well as encap
    get people to sign up to review
  • w-gs & r-gs affected: TSVWG, PCN, ICCRG, IPsec, Internet Area?
reminder (exec summary)

- **scope**
  - solely wire protocol processing of tunnelled ECN, not marking or response algorithms

- **sequence of standards actions led to perverse position**
  - non-IPsec ECN tunnels [RFC3168] have vestige of stronger security than even IPsec [RFC4301] decided was necessary!
  - limits usefulness of 3168 tunnels
    - e.g. PCN "excess rate marking" works with 4301 but not 3168 tunnels

- **bring ECN IP in IP tunnel ingress** [RFC3168] into line with IPsec [RFC4301]
  - all tunnels can behave the same, revealing full congestion info
  - anyway, copying of whole ECN field is simpler

- **thorough analysis of implications:**
  - security, control, & management
  - guidance on specifying ECN behaviour for new links, for alternate PHBs

- ideally fix egress too (currently only 'for discussion')
**reminder (exec summary)**

**encapsulation at tunnel ingress**

<table>
<thead>
<tr>
<th>Incoming header (also = outgoing inner)</th>
<th>outgoing outer</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC3168 ECN limited functionality</td>
<td>RFC3168 ECN full functionality</td>
</tr>
<tr>
<td>Not-ECT</td>
<td>Not-ECT</td>
</tr>
<tr>
<td>ECT(0)</td>
<td>Not-ECT</td>
</tr>
<tr>
<td>ECT(1)</td>
<td>Not-ECT</td>
</tr>
<tr>
<td>CE</td>
<td>Not-ECT</td>
</tr>
</tbody>
</table>

**decapsulation at tunnel egress**

- 'reset' CE no longer used
- 'copy' CE becomes normal state for all IP in IP

- proposal

- unchanged compatibility state for legacy
text updates since IETF-72

[draft-briscoe-tsvwg-ecn-tunnel-01.txt]  
→ [draft-ietf-tsvwg-ecn-tunnel-00.txt]  
→ [draft-ietf-tsvwg-ecn-tunnel-01.txt]

- much simpler method to monitor tunnel's contribution to congestion
  - see spare slide or Appendix B

- all significant edits concern decap – encap has stayed stable

- documented full set of illegal combinations of inner & outer at egress
  - on which egress should (optionally) raise a management alarm

- generalise egress behaviour while we're at it?
  - currently just in appendix 'for discussion' – says 'not normative'
  - problem: current egress behaviour discards changes to ECT(0) or ECT(1)
    - space for 2 congestion levels (e.g. PCN) but can't use it
    - effectively wastes half a bit of the IP header
  - now written up pros & cons of change (Appx C)
    - convinced myself this change should be in normative part of draft
    - what do you think...?
current egress behaviour

- OK for current ECN
- but any changes to ECT lost
  - effectively wastes ½ bit in IP header
  - again for safety against marginal threat
    that IPsec decided was manageable
- PCN tried to use ECT(0/1)
  - but having to waste DSCPs instead
  - or a limited scheme where it’s arranged for the
    egress to already know which of ECT(0/1) the
    ingress originally sent

<table>
<thead>
<tr>
<th>incoming inner</th>
<th>incoming outer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Not-ECT</td>
<td>Not-ECT</td>
<td>drop (!!!)</td>
</tr>
<tr>
<td>ECT(0)</td>
<td>ECT(0)</td>
<td>ECT(0) (!!!)</td>
</tr>
<tr>
<td>ECT(1)</td>
<td>ECT(1) (!!!)</td>
<td>ECT(1)</td>
</tr>
<tr>
<td>CE</td>
<td>CE</td>
<td>CE (!!!)</td>
</tr>
</tbody>
</table>

Outgoing header (RFC3168 & RFC4301)

(!!!) = illegal combination, egress MAY raise an alarm
'comprehensive' egress rules (only 'for discussion')

- recall: proposed change to ingress
  - brings RFC3168 into line with RFC4301
- if we also changed the egress
  - it would be a new update to both RFCs
- but no effect on any existing tunnels
  - adds a new capability using a previously illegal combination of inner & outer
  - only tunnels that need the new capability would need to comply
  - and update, not a fork
- note well: change to egress is currently not in the normative part of this proposal
  - but documented in appendix C 'for discussion'
  - however I'll make it normative if no-one objects

(!!!) = illegal combination, egress MAY raise an alarm
next steps

- should we change the egress at the same time?
  - tunnel stuff makes people's heads hurt
  - needs careful list discussion
  - remember, these are nuances to the behaviour of the neck of the hour-glass
  - will need to assure IPsec folks that they don't have to change (again)
  - I'll only make comprehensive egress rules normative if consensus to do so
  - I'll also add reasoning for original egress behaviour (requested in Anil Agarwal's review)

- plan to split out guidelines for new ECN encapsulations
  - for those adding congestion notification to alternate PHBs or to layer 2 technologies (incl. non-IETF, e.g. IEEE 802.1)
  - better in a separate (informational) I-D – just stds track IPinIP stuff in this one
  - and improve structure of this draft at same time (Michael Menth's comments)

- need people to sign up to review this draft
  - will need reviews once all the above settled
contribution to congestion across tunnel

**complaint:**

- if CE copied at ingress, operators can't distinguish congestion added since tunnel ingress
- it's not 12%
- new method in Appendix B
  - it's $\frac{12}{100-30} \approx 17\%$
  - just monitor the 70 packets without the inner header marked

The large square represents 100 packets

- ECN marking across tunnel
- problem: tunnel marks some packets that were already marked
- inner header ECN marking (already marked before ingress)

$$p_t$$
### Backward & Forward Compatibility

<table>
<thead>
<tr>
<th>IPsec-like</th>
<th>Egress</th>
<th>RFC 4301</th>
<th>RFC 3168</th>
<th>RFC 2481</th>
<th>RFC 2401/2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-D.ecn-tunnel</td>
<td>Normal</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>'Copy'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compat</td>
<td>inner</td>
<td>inner</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>RFC4301</td>
<td>'Copy'</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>RFC3168</td>
<td>Full</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>RFC2481</td>
<td>'Copy'?</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>A</td>
</tr>
<tr>
<td>RFC2401/RFC2003</td>
<td>Limited</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>inner</td>
</tr>
<tr>
<td>'2g IPsec'</td>
<td>Limited?</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
<td>inner</td>
</tr>
</tbody>
</table>

#### Calculation Descriptions:

- **C**: calculation C (more severe multi-level markings prevail)
- **B**: calculation B (preserves CE from outer)
- **A**: calculation A (for when ECN field was 2 separate bits)
- **inner**: forwards inner header, discarding outer
- **n/a**: not allowed by configuration

**Breaking Condition:** loses CE