Layered Encapsulation of Congestion Notification

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

Bob Briscoe, BT
IETF-73 pcn 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')
## Remider (exec summary)

### Encapsulation at tunnel ingress

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Incoming header (also = outgoing inner)</th>
<th>RFC3168 ECN limited functionality</th>
<th>RFC3168 ECN full functionality</th>
<th>RFC4301 IPsec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not-ECT</td>
<td>Not-ECT</td>
<td>Not-ECT</td>
<td>Not-ECT</td>
<td></td>
</tr>
<tr>
<td>ECT(0)</td>
<td>Not-ECT</td>
<td>ECT(0)</td>
<td>ECT(0)</td>
<td></td>
</tr>
<tr>
<td>ECT(1)</td>
<td>Not-ECT</td>
<td>ECT(1)</td>
<td>ECT(1)</td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>Not-ECT</td>
<td>ECT(0)</td>
<td>CE</td>
<td></td>
</tr>
</tbody>
</table>

### Decapsulation at tunnel egress

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

**Not-ECT**
- RFC3168 ECN limited functionality
- RFC3168 ECN full functionality
- RFC4301 IPsec

**ECT(0)**
- RFC3168 ECN limited functionality
- RFC3168 ECN full functionality
- RFC4301 IPsec

**ECT(1)**
- RFC3168 ECN limited functionality
- RFC3168 ECN full functionality
- RFC4301 IPsec

**CE**
- RFC3168 ECN limited functionality
- RFC3168 ECN full functionality
- RFC4301 IPsec
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...?
• 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

(outgoing inner)

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

(outgoing outer)

Outgoing header (RFC3168 & RFC4301)

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

encapsulation at tunnel ingress

decapsulation at tunnel egress

- 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

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

Outgoing header (proposed update)
(bold = proposed change for all IP in IP)

(!!!) = illegal combination, egress MAY raise an alarm
new comprehensive decap rules
pros & cons of ways to introduce them

<table>
<thead>
<tr>
<th>Default for all PHBs</th>
<th>Adv: no config as old behaviour was unusable</th>
<th>Recommended. Can fall back on expt track if stall</th>
<th>More likely to get through</th>
</tr>
</thead>
<tbody>
<tr>
<td>For PHBs that need it</td>
<td>Disadv: no motivation for unused fork</td>
<td>reject</td>
<td>reject</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adv: no motivation for unused fork</th>
<th>Disadv: eventually extra mode of tunnel to be compatible with</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>within tsvwg-ecn-tunnel stds track</th>
<th>new pcn-tunnel-... expt track</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disadv: may never need change</td>
<td></td>
</tr>
</tbody>
</table>
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 rvw)

• 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
Layered Encapsulation of Congestion Notification

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

Q&A
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

<table>
<thead>
<tr>
<th>30% marked</th>
<th>42% marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encapsulation at tunnel ingress</td>
<td>Decapsulation at tunnel egress</td>
</tr>
</tbody>
</table>

ECN marking across tunnel

Problem:
- Tunnel marks some packets that were already marked

\( p_t \)

30

0% 30% 100%

Inner header ECN marking
(Already marked before ingress)
# Backward & Forward Compatibility

<table>
<thead>
<tr>
<th>ingress</th>
<th>mode</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>IPsec-like</td>
<td>I-D.ecn-tunnel</td>
<td>normal</td>
<td>'copy'</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>compat</td>
<td>'zero'</td>
<td>inner</td>
<td>inner</td>
<td>n/a</td>
</tr>
<tr>
<td>'3g IPsec'</td>
<td>RFC4301</td>
<td>4301</td>
<td>'copy'</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>ECN</td>
<td>RFC3168</td>
<td>full</td>
<td>'reset CE'</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>limited</td>
<td>'zero'</td>
<td>inner</td>
<td>inner</td>
<td>n/a</td>
</tr>
<tr>
<td>ECN expt</td>
<td>RFC2481</td>
<td>2481</td>
<td>'copy'?</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>limited?</td>
<td>'zero'</td>
<td>inner</td>
<td>inner</td>
<td>n/a</td>
</tr>
<tr>
<td>'2g IPsec' IP in IP</td>
<td>RFC2401 RFC2003</td>
<td>-</td>
<td>'copy'</td>
<td>C</td>
<td>B</td>
<td>n/a</td>
</tr>
</tbody>
</table>

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

Broken: Loses CE