Layered Encapsulation of Congestion Notification

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

Bob Briscoe, BT
IETF-72 tsvwg Jul 2008
updated draft

• Layered Encapsulation of Congestion Notification
  • updated draft: draft-briscoe-tsvwg-ecn-tunnel-01.txt
  • intended status: standards track
  • immediate intent: move to WG item discuss widening scope

• exec summary
  – bring ECN IP in IP tunnel ingress [RFC3168] into line with IPsec [RFC4301]
    • all tunnels can behave the same, revealing full congestion info
    • only wire protocol processing, not marking or response algorithms
  – thorough analysis of implications: security, control, & management
    • guidance on specifying ECN behaviour for new links, alternate PHBs
  – ideally fix egress too (currently only 'for discussion')
one main update to RFC3168 ECN

encapsulation at tunnel ingress
decapsulation at tunnel egress

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>ECT(0)</td>
<td>Not-ECT</td>
<td>ECT(0)</td>
<td>ECT(0)</td>
</tr>
<tr>
<td>ECT(1)</td>
<td>Not-ECT</td>
<td>ECT(1)</td>
<td>ECT(1)</td>
</tr>
<tr>
<td>CE</td>
<td>Not-ECT</td>
<td>ECT(0)</td>
<td>CE</td>
</tr>
</tbody>
</table>

propose unchanged compatibility state for legacy

'reset' CE no longer used

'copy' CE becomes normal state for all IP in IP

---

DS

ECT

ECN

Not-ECT

ECT(0)

ECT(1)

CE

I

E

I
why update ECN RFC3168 now?

• sequence of standards actions led to perverse position
  – despite everyone’s best intentions
  – 2001: RFC3168 tunnel ingress specified cautiously due to security concerns
  – 2005: RFC4301 IPsec decided caution wasn't necessary
    • IETF Security Area decided 2-bit ECN covert channel can be managed

• vestige of security no longer used by IPsec
  now limits usefulness of non-IPsec tunnels
  • already PCN "excess rate marking" says "doesn't work with 3168 tunnels"
  • anyway, copying of whole ECN field is simpler
activity from initial -00 to -01 draft

• general agreement on list with 'copy on encap'
• concern on list (a year ago) over a couple of details
  – exception for in-path load regulators (resolved by removing it)
    • conceptual model from RFC2983 avoids need for exception
    • Appx D: Non-dependence of tunnelling on in-path load regulation
      – reconstructing precise cross-tunnel congestion metric (resolved)
        • Appx B: suggested precise cross-tunnel measurement technique
          • since replaced with really simple technique [for -02 after IETF-72]
• that was 1 year ago
  • agreed to go dormant until PCN wire protocol clearer...
current egress behaviour OK(ish)

- works for current ECN
- propose only one state at egress
  - same behaviour for both ingress states
- but any changes to ECT lost
  - effectively wastes ½ bit in IP header
- PCN tried to use ECT codepoints
  - having to waste DSCPs instead

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

Outgoing header (RFC3168 full & RFC4301)
(bold red = proposed for all IP in IP)

(!!!) = illegal transition, E MAY raise an alarm
ideally fix egress too (only 'for discussion')

- change egress at same time?
- backwards compatible
  - just previous tunnels wouldn't propagate changes to ECT
- this is not currently part of proposal
  - but documented in an appendix

Encapsulation at tunnel ingress

decapsulation at tunnel egress

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

Outgoing header (RFC3168 full & RFC4301)
(bold red = proposed for all IP in IP)

(!!!) = illegal transition, E MAY raise an alarm
next steps

• would like to request as WG item
  • PCN w-g needs to know if proposal is likely to happen
  • also implications for PWE3 (if using ECN)
  • will need IPsec to be happy that they aren't affected

• also to discuss (here or on list):
  • should we change the egress at the same time?
Layered Encapsulation of Congestion Notification

draft-briscoe-tsvwg-ecn-tunnel-01.txt
## backward & forward compatibility

<table>
<thead>
<tr>
<th>ingress</th>
<th>egress</th>
<th>pro-</th>
<th>RFC 4301</th>
<th>RFC 3168</th>
<th>RFC 2481</th>
<th>RFC2401 RF2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>action</td>
<td>calc B</td>
<td>calc B</td>
<td>calc B</td>
<td>inner</td>
<td>calc A</td>
<td>inner</td>
</tr>
<tr>
<td>I-D.ecn-</td>
<td>normal</td>
<td>'copy'</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>n/a</td>
</tr>
<tr>
<td>tunnel</td>
<td>compat</td>
<td>'zero'</td>
<td>inner</td>
<td>n/a</td>
<td>n/a</td>
<td>inner</td>
</tr>
<tr>
<td>'3g IPsec'</td>
<td>RFC4301</td>
<td>'copy'</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>n/a</td>
</tr>
<tr>
<td>ECN</td>
<td>RFC3168</td>
<td>full</td>
<td>'reset CE'</td>
<td>B</td>
<td>n/a</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>limited</td>
<td>'zero'</td>
<td>inner</td>
<td>n/a</td>
<td>n/a</td>
<td>inner</td>
</tr>
<tr>
<td>ECN expt</td>
<td>RFC2481</td>
<td>'copy'?</td>
<td>B</td>
<td>n/a</td>
<td>B</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>limited?</td>
<td>'zero'</td>
<td>inner</td>
<td>n/a</td>
<td>n/a</td>
<td>inner</td>
</tr>
<tr>
<td>'2g IPsec'</td>
<td>RFC2401 RF2003</td>
<td>'copy'</td>
<td>B</td>
<td>n/a</td>
<td>n/a</td>
<td>inner</td>
</tr>
</tbody>
</table>

**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
tunnel contribution to congestion, $p_t$

Q. how to measure $p_t$ at egress?

A. $p_t = \frac{12}{70} \approx 17\%$

• just monitor the 70 packets without the inner header marked

The large square represents 100 packets

ECN marking across tunnel

problem: marks some packets marked already

(already ECN marked before ingress)
conflicting design constraints

security vs. management & control

- information security constraint (lesser known IPsec reqm’t)

  - I can prevent covert channel A→M with encryption
  - E can prevent covert channel M→B with integrity checking

- tunnel ingress control / management constraints

  - marking algorithm at M may depend on prior markings (since A)
    - e.g. a number of PCN marking proposals work this way
  - M may need to monitor congestion since A
    - e.g. if M is monitoring an SLA at a border

- IPsec crypto cannot cover mutable fields (ECN, DS & TTL)
  - if ‘I’ copies ECN CE, it opens up 2-bit covert channel A→M or R→M
conflicting design constraints
security vs. congestion control

- information security constraint (lesser known IPsec reqm’t)
  - I can prevent covert channel A→M with encryption
  - E an prevent covert channel M→B with integrity checking

- tunnel egress control constraint
  - explicit congestion notification control channel M→B→A

- IPsec crypto cannot cover mutable fields (ECN, DS & TTL)
  - if E copies ECN CE, it opens up 2-bit covert channel M→B