Router participation in Congestion Control

Techniques

- Random Early Detection
- Explicit Congestion Notification

Early congestion notifications

"Early" notifications inform end-systems that the network is "congested" *before* buffers in network elements (routers) overflow.

Sources will take some time to respond to notifications => early notifications may prevent severe congestion (buffer overflow)

Two forms of early notifications:

- Discard: "Active Queue Management (AQM)"
- Explicit Congestion Notifications

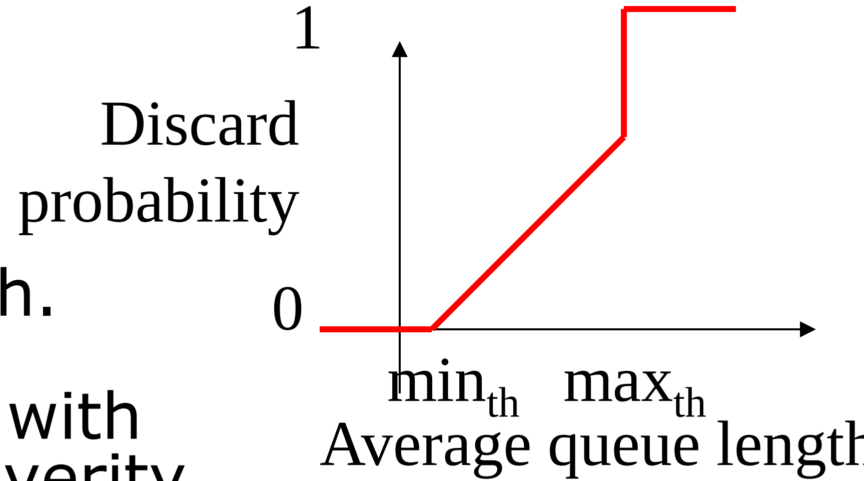
Active Queue Management (AQM)

- Routers drop packets to notify ends of imminent congestion.
  - Notification has same form as buffer overflow => can exploit existing end-system responses.
- The *hope* is that the early discard causes sources to slow down before the buffer overflows (causing high loss rates).
- This *assumes* that sources slow down in response to packet loss.
  i.e. layer interdependence: the network layer depends on certain transport protocols being used.

Details in RFC 2309

Random Early Detection[†] (RED)

Discard arriving packets with a probability that increases[1] linearly with the average[2] queue length.



1. **Increasing** discard rate with queue length indicates severity of congestion and penalises sources that don’t slow down.
2. **Averaging** the queue length rather than using instantaneous measurement allows limited bursts to pass, accommodating TCP’s Slow Start

[†] A form of AQM. Sometimes called “Random Early Discard”/drop

RED variant references

T. Ott, T. Lakshman and L. Wong: “SRED: Stabilized RED,” *Proc. INFOCOM*, pp. 1346–55

J. Aweya, M. Ouellette and D. Montuno: “A Control Theoretic Approach to Active Queue Management,” *Comp. Net.*, vol. 36, issue 2–3, pp. 203–35 [DRED]

B. Wydrowski and M. Zukerman: “GREEN: an active queue management algorithm for a self managed Internet”, Proc. ICC, pp. 2368-72

W. Feng, A. Kapadia and S. Thulasidasan: “GREEN: proactive queue management over a best-effort network”, Proc. Globecom, pp. 1774-78

W. Feng et al., “BLUE: A New Class of Active Queue Management Algorithms,” Technical Report CSE-TR-387-99, Dept. EECS, Univ. MI

G. Chatranon, M. A. Labrador, and S. Banerjee, “BLACK: Detection and Preferential Dropping of High Bandwidth Unresponsvie Flows,” in Proc. IEEE ICC, pp. 664–668.

Outline

Explicit Congestion Notification

Motivation:

× Early packet discard is crude:

  × wastes bandwidth: packets propagate from source to congested router, only to be discarded.

  × increases packet loss

- May be better for routers to explicitly tell sources (through an explicit signal rather than packet loss) that they are getting congested.

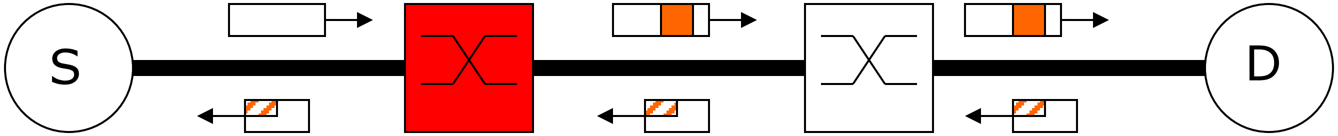
Limitation of ECN:

√ Presence of ECN indicates congestion

× Absence of ECN doesn’t prove non-congestion (congestion might have resulted in ECN signal being lost)

=> sources must still treat loss as a congestion indicator

ECN mechanisms



Router could directly tell the source:

- Send a new packet to the source

  But that would create extra network traffic during times of congestion (Early Internet had "Source quench", but that has been deprecated.)

- Set a field in a packet already going to the source ("Backward ECN")

  Router has to remember which sources are contributing to its congestion (=> scalability problem), while waiting to observe packets being sent to those sources (to piggyback signal on)

  => Used by Frame Relay, but not the Internet

Router could tell destination, and destination tells source (if it receives the packet).

"Forward ECN", used by Frame Relay, and now the Internet.

ECN in the Internet

IP header was:

```
   0     1     2     3     4     5     6     7
+-----+-----+-----+-----+-----+-----+-----+-----+
|   PRECEDENCE    |       TOS       |  0  |  0  |   RFC 791
+-----+-----+-----+-----+-----+-----+-----+-----+

+-----+-----+-----+-----+-----+-----+-----+-----+
|               DSCP                |    CU     |   RFC 2474
+-----+-----+-----+-----+-----+-----+-----+-----+
```

DSCP = Differentiated Services Codepoint   "CU" = "Currently Unused"

Becomes:

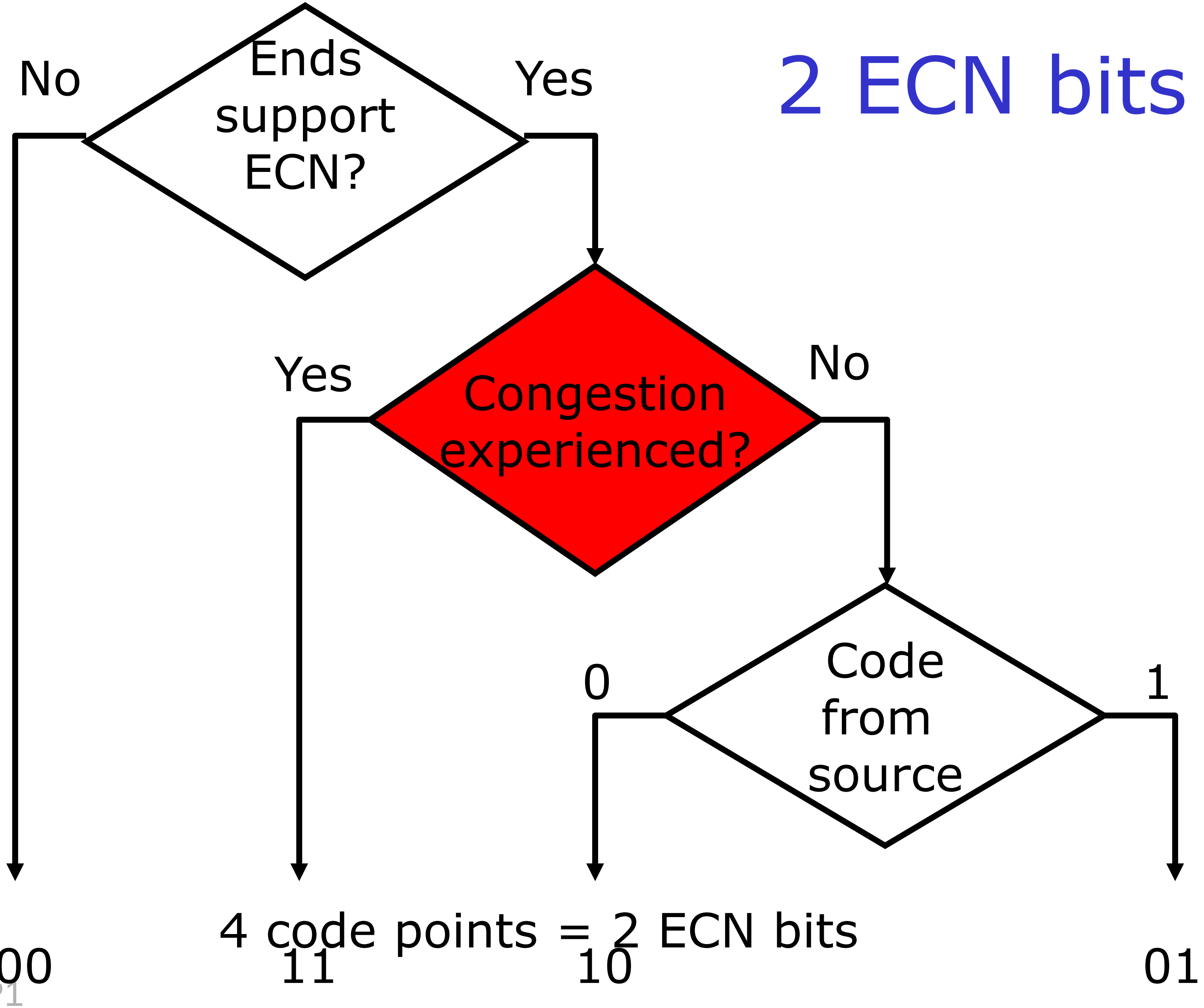
ECT = ECN Capable Transport
CE = Congestion Experienced

```
+-----+-----+
| ECN FIELD |
+-----+-----+
```

ECT	CE	
0	0	Not-ECT
0	1	ECT(1)
1	0	ECT(0)
1	1	CE

Bonus marks available to anyone who finds packet captures (e.g. from Wireshark) that show ECN in use.

2 ECN bits



4 code points = 2 ECN bits

ECN in the Internet Protocol

```
+-----+-----+
| ECN FIELD |
+-----+-----+
  ECT   CE
   0     0    Not-ECT
   0     1    ECT(1)
   1     0    ECT(0)
   1     1    CE
```

The 2 ECN bits indicate:

- Traditional Not-ECT
  - Router only changes ECN field if set to ECT(0) or ECT(1)
  - Traditional receivers check that reserved bits are set to 0
- ECN-Capable Transport (ECT)
  - If ECT is set and router is experiencing congestion, router may change ECN field to CE, rather than discarding the packet.
  - Why 2 codepoints? See slide "Detect lying with 2 ECT values" [TH>
- Congestion Experienced (CE)

RFC 3168 (obsoletes RFC 2481)
[Savage99] S. Savage and others: "TCP Congestion Control with a Misbehaving Receiver", *ACM Computer Comm. Review*, October 1999.

Details: Transport reaction to ECN

Nitty gritty details from RFC3168 explain fully, but aren't critical to this course

- Routers set only if ends sure to detect: "an ECT codepoint MUST NOT be set in a packet unless the loss of that packet in the network would be detected by the end nodes and interpreted as an indication of congestion." [3168]
  - e.g. don't set it in pure (not piggybacked) ACKs, since ACKs are cumulative: loss of one may escape notice if next one comes soon
- Source reacts once per RTT to ECN: "end-systems should react to congestion at most once per window of data (i.e., at most once per round-trip time), to avoid reacting multiple times to multiple indications of congestion within a round-trip time." [3168]
- Source reacts to CE as if a single packet were lost: "the congestion control algorithms followed at the end-systems MUST be essentially the same as the congestion control response to a *single* dropped packet." [3168]
- Traditional sources could receive unfair treatment if routers perform early discard to signal congestion rather than marking CE
- End systems need to negotiate whether to use ECN.
  - Source that mistakenly thinks that destination is ECN-capable may penalise dest'n for making false reports about no CE

issues arising from multiple versions during incremental deployment

ECN and TCP

Use 2b of TCP’s Reserved field to indicate:

```
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               |               | C | E | U | A | P | R | S | F |
| Header Length |    Reserved   | W | C | R | C | S | S | Y | I |
|               |               | R | E | G | K | H | T | N | N |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```

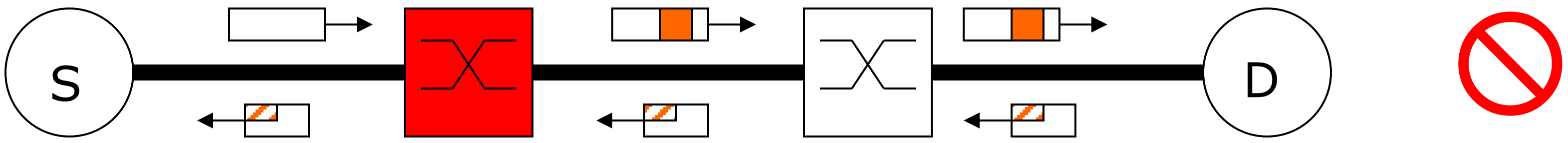
- ECN-Echo (ECE): Echoes to source receipt of CE.

  Destination continues setting ECE in ACKs until it receives CWR from source (in case ACKs are lost).
- Congestion Window Reduced (CWR):

  Source tells destination that window has been reduced. Destination need no longer echo previous CE.

In SYN segments: ECE&CWR => ECN capable

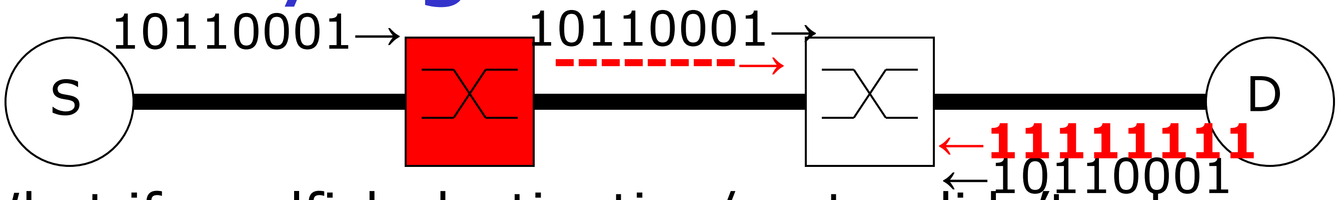
Typical series of events



("IP" refers to 2b IP field, CWR, ECE refer to TCP fields.
Only care about IP field and CWR from source, and ECE from dest)

1. **Source** connects to destination, and **negotiates to use ECN**.
2. **Source sends segments** with IP=ECT, CWR=0
3. **"congested" router changes IP=ECT to CE**. Router:
   - Doesn't process TCP fields
   - Likely marks all segments received while congested with IP=CE. Source will only reduce rate for one segment, but marking all segments expedites signal when segments get lost.
4. **Destination receives segment, sends ACK, ECE=1**
   Keeps setting ECE=1 for all acks until step 6.
5. **Source receives ACK with ECE=1**, reduces congestion window. **Sends segment with CWR=1** (IP=ECT).
6. **Destination receives CWR=1 segment**, and sends ACK with ECE depending on IP field of segment:
   1. IP=CE => router still congested => ECE=1
   2. IP=ECT => router no longer congested => ECE=0

Detect lying with 2 ECT values



- What if a selfish destination/router didn’t echo Congestion Experienced indicators?
  - e.g. if it doesn’t care about network but wants source to keep sending fast to it.
- Two ECT values allow a source to check the destination’s reporting of CE markings.
  - Source varies ECT 0/1 value according to a pattern that it alone knows (e.g. pseudorandom) – “nonce”
  - Destination must return the pattern back to the source, e.g. using a new “Nonce Sum” TCP header field (sum of ECT(x) fields received).

```
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               |           | N | C | E | U | A | P | R | S | F |
| Header Length | Reserved  | S | W | C | R | C | S | S | Y | I |
|               |           |   | R | E | G | K | H | T | N | N |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```

  - Routers overwrite ECT when setting CE, preventing receiver from returning pattern to source => source can detect CE, despite selfish/uncooperative nodes

For details, see RFC 3540

ECN summary

- Reduces waste from loss
- Can’t entirely replace loss as a congestion indicator, since during extreme congestion loss may prevent ECN carriage.
- Router marks packets as having experienced congestion; destination signals to source to reduce rate.
- IP’s coding (using 2b)
  - provides backwards compatibility with non-ECN capable transport.
  - enables a source to detect false reporting.