Ethernet Congestion Manager (ECM)

Davide Bergamasco

IEEE 802 Plenary Meeting
Orlando, FL
March 13th, 2007
What is ECM?

- ECM is a *Layer 2 congestion management mechanism*
- Formerly known as BCN

- **Principles**
  - Push congestion from the core towards the edge of the network
  - Use rate-limiters at the edge to “shape” flows causing congestion
  - Control injection rate based on feedback coming from congestion points

- **Inspired by TCP**
  - AIMD rate control
    - TCP window increases linearly in absence of congestion
    - Decreases exponentially (gets halved) at every congestion indication (either implicit or explicit)
  - Self-Clocking Control loop (acknowledgements)
ECM Concepts

Reaction Point: Reaction Component

Congestion Point: Detection Component

Protocol: Signaling Component

Congestion Management Domain
ECM Concepts: Signaling
ECM Concepts: Signaling

<table>
<thead>
<tr>
<th>0</th>
<th>15</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+ DA = SA of sampled frame</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>SA = MAC Address of CP</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>IEEE 802.1Q Tag</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>EtherType = ECM</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>Q</td>
<td>Reserved</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>CPID</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Qoff</td>
<td>Qdelta</td>
<td>+-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Timestamp</td>
<td>Unit</td>
<td>Reserved</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>First N bytes of sampled frame starting from DA</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>+-----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>FCS</td>
<td>+-----------------</td>
<td>-----------------</td>
</tr>
</tbody>
</table>
ECM Concepts: Signaling

```
0  15  31
+---------------------------------------------+
| EtherType = CM-Tag | Version | Reserved |
+---------------------------------------------+
|                             | CPID    |     |
+---------------------------------------------+
|                             | Timestamp | Unit |
+---------------------------------------------+
```
ECM Concepts: Detection

Full Queue

\( Q_{sc} \)

\( Q_{mc} \)

\( Q_{old} \)

\( Q_{len} \)

\( Q_{off} \)

Empty Queue

Sample Frame with Probability P

Frame Sampled?

Yes

No

\( Q_{len} > Q_{sc}? \)

\( Q_{len} > Q_{mc}? \)

Compute Qoff and Qdelta

\( Q_{off} > 0? \)

\( CM-TagPresent\&\&\ CPIDMatch? \)

Send ECM(0,0)

Send ECM-Max

Send ECM(Qoff, Qdelta)

NO

YES

NO

YES

Qoff = Qlen - Qeq [-Qeq, +Qeq]

Qdelta = Qlen - Qold [-2Qeq, +2Qeq]
ECM Concepts: Detection

- Byte-based Sampling
  - P desired frame sampling probability
  - E[L] is the average frame length
  - Sampling interval is \( S = \frac{E[L]}{P} \)
  - E.g., \( P = 0.01, E[L] = 1 \text{ KB} \) ➔ Sample a frame every 100 KB received

- To avoid bias, small random component added to sampling interval
  - \( S = S_f + S_r \) ➔ \( E[S] = S_f + E[S_r] \) ➔ \( E[S_r] = \frac{E[L]}{P} - S_f \)
  - E.g., if \( E[S] = 100 \text{ KB} \) and \( S_f = 90 \text{ KB} \) ➔ \( E[S_r] = 10 \text{ KB} \) ➔ \( S_r \in [0, 20] \text{ KB} \)
ECM Concepts: Detection

- Congestion Point state and configuration variables
  - **State**:
    - Qlen: current queue length
    - Qold: queue length at previous sample
    - Bytecount: # bytes arrived since last sample
  - **Configuration**:
    - Qeq: equilibrium threshold
    - Qmc: medium congestion threshold (BCN-MAX trigger)
    - Qsc: severe congestion threshold (BCN(0,0) trigger)
    - Sf: fixed sampling interval
    - Sr: random sampling interval range
ECM Concepts: Reaction

*Feedback
Fb = -(Qoff + W * Qdelta)

*Additive Increase (Fb > 0)
R = R + Gi * Fb

*Multiplicative Decrease (Fb < 0)
R = R * (1 - Gd * |Fb|)

*Parameters
W = derivative weight
Gi = increase gain
Gd = decrease gain

ECM Frames from congestion point
ECM Concepts: Reaction

- Rate adjustment

```c
handle_bcn_frame( rate_lim, bcn_frame )
{
    Fb = calc_feedback( bcn_frame );
    if ( Fb < 0 )
    {
        rate_lim->R = rate_lim->R * ( 1 - min( - Gd * Fb, alpha ) );
        rate_lim->CPID = bcn_frame->CPID;
    }
    else
    {
        if ( bcn_frame.CPID == rate_lim->CPID )
        {
            rate_lim->R = rate_lim->R + min( Gi * Fb, beta );
        }
    }
}
```
ECM Concepts: Reaction

- Severe congestion reaction: ECM(0,0)
  - Current rate $R$ is set to 0
  - Start random timer $T \in [0, T_{Max}]$:
  - When timer $T$ expires $R \leftarrow R_{Min}$
  - Next ECM(0,0) causes exponential back-off:
    - $T_{Max} \leftarrow T_{Max} \times 2$ and $R_{Min} \leftarrow R_{Min} / 2$
  - Next positive feedback resets $T_{Max}$ and $R_{Min}$
ECM Concepts: Reaction

- Self-increase (aka drift)
  - At regular intervals Td current rate $R \leftarrow R + Rd$
  - Purposes:
    - Speedup recovery from
    - ECM(0,0)
    - Loss of signaling from CP
    - Improve fairness
    - Reclaim a rate limiter at flow termination

- RTT estimator
  - $RTT_{avg} \leftarrow (1 - 2^{-Wrtt}) RTT_{avg} + 2^{-Wrtt} \times RTT$
  - RTT may be used for adjusting ECM parameters as network condition change (still experimental)
ECM Concepts: Reaction

- Reaction Point state and configuration variables
  - **State (per rate-limiter):**
    - **R:** current rate
    - **CPID:** current CPID
    - **RTTavg:** last measured RTT
  - **Opt:**
    - **Rmin:** current minimum back-off rate
    - **Tmax:** current maximum back-off time

- **Configuration:**
  - **W:** weight of derivative component
  - **Gd:** decrease gain
  - **Gi:** increase gain
  - **α:** maximum rate decrease (fraction of current rate)
  - **β:** maximum rate increase (fraction of link capacity)
  - **Wrtt:** weight of the EWMA RTT filter
  - **Td:** self-increase timer
  - **Rd:** self-increase amount
  - **Rmin:** minimum back-off rate
  - **Tmax:** maximum back-of time
Questions?