Update on
draft-ietf-soc-overload-rate-control

SOC Rate team:
Eric Noel, AT&T Labs, Inc
Janet Gunn, CSC
Philip Williams, BT
Introduction

• Last SOC meeting IETF83
  – Phil gave an overview of draft-ietf-soc-overload-rate-control

• Since then we received a few comments
  – Editorial changes
  – Clarification of Leaky Bucket parameters selection
  – Addition of pseudo code

• All comments were addressed in draft-ietf-soc-overload-rate-control-02
Refresh

• Rate-based overload control approach
  – Mitigates congestion in SIP networks
  – Addresses loss-based control limitations
    • capacity guarantees...
  – Conforms to draft-ietf-soc-overload-control signalling scheme

• draft-ietf-soc-overload-rate-control-02 available
Refresh: Commonality & Differences

loss-based, rate-based

Client
Throttling:
max Request rate
[loss-based: rejects %]

Same parameters
Different values/interpretation

Server (overload)
Periodically calculates:
max Request rate
[loss-based: % rejected]

based on internal measurements e.g.
message rate, CPU utilisation, queueing delay
Refresh: rate control Server operation

• Server MUST periodically evaluate its overload state, estimate a target SIP request rate for each client, and send the new target rate to the specific client.
  – Algorithm for estimating rate is out of scope
Refresh: rate control Client operation


• Two parameters: T and TAU
  – Target inter-arrival time $T = 1 / \text{[Server target rate]}$
  – Tolerance parameter TAU
  – Priority scheme relies on two tolerance parameters
    • TAU1 & TAU2

• Target rate is sent from the Server to the Clients using the proposed mechanism

• The values TAU, TAU1 and TAU2 are static and set in advance (or if they need to be changed, that is done by some other mechanism)
Sensitivity on parameters: no priority

- Tolerance parameter $\text{TAU}$ can assume any positive value
- A new SIP request is forwarded to the server iff the provisional content of the bucket is less than or equal to the limit value $\text{TAU}$
  - The larger $\text{TAU}$ the more tolerance to deviations from the inter-departure interval $\text{T}$ and the larger the tolerance to burst size

\[ \begin{align*}
\text{max fill } & T+\text{TAU} \\
\text{current fill (at last conforming arrival)} & = \text{provisional fill after arrival} \\
\text{fill after admission (iff current fill - } \Delta t < \text{TAU)} & = \text{current fill (at last conforming arrival) - } \Delta t \\
\end{align*} \]

\[ T \]

\[ \Delta t \]
Sensitivity on parameters: priority

- Tolerance parameters TAU1 & TAU2 can assume any positive value
  - TAU1 (non-priority traffic) \( \leq \) TAU2 (priority traffic)
  - TAU1 = TAU2 is equivalent to no priority
  - The larger TAU2 – TAU1, the closer to strict priority
- TAU1 influences combined throttle rate the same as TAU does when no priority are set
Conclusions

• Open discussion on draft-ietf-soc-overload-rate-control