ATS Capacity Tuning & Concurrency Limiter

Sudheer Vinukonda
June 2020
ATS Capacity Tuning - Today

❖ Overall connection Limit
  ➢ Includes inbound, outbound
  ➢ Needs to be set very high
❖ Max Outbound Connections
❖ Transaction, Connection timeouts
  ➢ Inactivity/Active Timeouts
  ➢ Malicious, rogue or broken client, network behavior
❖ Connection Limits per Origin
  ➢ Service latency issues or failures

- proxy.config.net.connections_throttle: 300000
- proxy.config.http.server_max_connections: 60000
- proxy.config.http.transaction_no_activity_timeout_in: 75
- proxy.config.http.transaction_no_activity_timeout_out: 75
- proxy.config.http.transaction_active_timeout_in: 180
- proxy.config.http.transaction_active_timeout_out: 0
# ATS Capacity Tuning - Today

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Connection Limit</td>
<td>● Works for HTTP/1.1 Traffic</td>
<td>● Concurrency is vastly different for HTTP/2 due to request multiplexing</td>
</tr>
<tr>
<td>Connection/Transaction Timeouts</td>
<td>● Protects against Runaway Transactions, Connection leaks etc</td>
<td>● Needs to be tuned relatively low to optimize for capacity/throughput</td>
</tr>
<tr>
<td></td>
<td>● Efficient resource utilization</td>
<td>● Upload, long poll etc use cases require higher timeouts</td>
</tr>
<tr>
<td>Origin Connection Limits</td>
<td>● Isolation/resiliency against failures or latency spikes in individual Services</td>
<td>● “Too late” because majority of the “work” is already completed</td>
</tr>
</tbody>
</table>
ATS Capacity Tuning - Concurrency Limiter

- Available resources/capacity
  - CPU, Memory, Network bandwidth
- Using active connection limits
  - Works well for HTTP/1.1
  - HTTP/2 poses a challenge with Multiplexing
- Using Max Active requests (Concurrency Limiter)
  - Streams in HTTP/2
- Reduces/eliminates dependency on timeout configs
  - Active/Inactive/keep-alive timeouts
  - Hard to get right
# ATS Capacity Tuning - Changes

<table>
<thead>
<tr>
<th>Current Config</th>
<th>Proposed Config</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>proxy.config.net.max_active_connections_in</td>
<td>proxy.config.net.max_requests_in</td>
<td>max number of in-flight inbound requests (concurrency)</td>
</tr>
<tr>
<td>proxy.config.http.server_max_connections</td>
<td>proxy.config.net.max_connections_out</td>
<td>max number of outbound connections (active + keepalive)</td>
</tr>
<tr>
<td>proxy.config.net.max_connections_in</td>
<td>proxy.config.net.max_connections_in</td>
<td>max number of in-flight requests (concurrency)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ idle (socket) connections</td>
</tr>
<tr>
<td>proxy.config.net.connections_throttle</td>
<td>proxy.config.net.connections_throttle</td>
<td>max number of open connections (FD, rlimits etc)</td>
</tr>
</tbody>
</table>
ATS Capacity Tuning - Changes

❖ Should connection be closed when throttling a request?
  ➢ Status code 503?
  ➢ HTTP/1.1
    ■ Connection: close
  ➢ HTTP/2
    ■ GoAway
❖ Internal requests (PluginVC/FetchSM)?
❖ Anything else…?