Using Edge to Reduce Latency Between End Users and Cloud

Nathan Moore
Director, Performance Engineering
Sep 25, 2019
THE PUNCHLINE

<table>
<thead>
<tr>
<th>Service</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackPath</td>
<td>29ms</td>
</tr>
<tr>
<td>Google Cloud</td>
<td>66ms</td>
</tr>
<tr>
<td>Azure</td>
<td>68.2ms</td>
</tr>
<tr>
<td>AWS</td>
<td>75ms</td>
</tr>
</tbody>
</table>

North America average response times, Cedexis, 2019
LOW LATENCY MEANS HIGH BANDWIDTH

Bandwidth Versus Latency

↑ Bandwidth

← Latency
HOW IT WORKS

StackPath Edge Server, Los Angeles

End User, Los Angeles

Council Bluffs, Iowa
EDGE, A TYPE OF DISTRIBUTED COMPUTING SYSTEM
Protocol Latency Lowering

- TCP Handshake
  - Lower Round Trip Time means less overhead time
  - Lower total time through TCP Slow Start
  - Avoid classical LFN problems
- SSL Handshake
  - Lower Round Trip Time means less overhead time
- Connection Re-use
  - Edge-to-Cloud keep-alives
Connection Termination

Terminating connections at the Edge and re-using Edge-to-Cloud connections alone can result in large latency improvements.

- SSL negotiation can take multiple round trips.
• Keep as much application logic local on the Edge as possible
• Simple example - Log Batching
INTEGRATION WITH OTHER SERVICES

Integrate Application with Additional Services
• Utilize within-Edge Services to reduce remote calls
• Expand capabilities within Edge

Microservices at the Edge