A Latency Taxonomy
and Two Opportunities

Joe Touch
Postel Center Director, USC/ISI
Research Associate Prof., USC CS & EE/Systems
Latency

“Everybody talks about the speed of light, but nobody ever does anything about it.”
– JT, 1988 (0.79 Gsec ago)

• *The* fundamental metric of computing and communication
  – All performance is measured as delay between a question and an answer
  – Everything else is just a means to an end
Proposed Latency Taxonomy

• Measure by impact on the latency budget
  …focus on the reason/mechanism of delay
  – Primary causes – sources consume budget
  – Primary fixes – mitigations reduce source impact

• Consider summary impact
  – Defines limits of improvement

• Ignores:
  – Location
  – Owner
  – “Layer” (when inside a protocol stack)
  – Origin (config., implementation, algorithm, etc.)
Basic Definitions

• Latency budget
  – Hard vs. soft
    • Compares penalty for exceeding
  – Biological vs. computational derivation
    • Determines the expected deadline

• Latency cost
  – Sources – increase cost
  – Mitigations – reduce cost
Sources

• **Generation:**
  – Delay between physical event and availability of data.
  – Physical (audio freq.), source format (video frame), storage (RAM, disk)

• **Transmission:**
  – Inherent in propagating a signal.
  – Signal propagation, initial signal encoding (parallel/serial, striping, bit/symbol)
Sources...

• Processing:
  – Computational translation.
  – Forward, encap/decap, NAT, encrypt, auth., compress, error coding, signal translation

• Multiplexing:
  – Delays needed to support sharing.
  – Shared channel acquisition, output queuing, connection establishment

• Grouping:
  – Reduces frequency of control information and processing.
  – Packetization, message aggregation
Mitigation Approach

• Changes in resources / goals
  – Burn bandwidth, memory, possibly CPU
  – Consider energy impact

• Changes in cost/benefit
  – Costly resources now ‘free’
    • BW, CPU, memory
  – Cost of previously ‘free’ resources
    • CPU, BW (considering energy)
Specific Mitigations

• Relocation
  – Move the endpoints closer (reduces transmission impact)
  – E.g., offload, zero-copy, content distribution centers

• Speedup
  – Increase operations per unit time (reduces processing impact)
  – E.g., faster processor, higher BW path
Specific Mitigations...

- **Dedication**
  - Reserve exclusive resources (reduces *multiplexing* impact)
  - *E.g.*, reserved BW, dedicated circuits, separate network / security processors

- **Partitioning**
  - Split group into individual components (reduces *grouping* impact)
  - Split circuit into packets, split large packets into small cells
General Mitigations

• “Wait loss”
  – Avoid by omission or substitution (reduces impact of all sources)
  – E.g., MPLS, TCP Nagle, AQM, RED

• Anticipation
  – Proactive communication (hides the impact of all sources)
  – E.g., caching, T/TCP, persist-HTTP, ‘Prefetching the means’, TCP control block sharing
Two Opportunities

• Small packets
  – Intermediate between circuit/IP and IP/cell
  – Reduces grouping latency
  – Increases BW, side effects of reordering, *etc.*

• Push anticipation
  – Decouple sender/receiver interaction
  – Latency stays the same, but happens earlier and is thus “hidden”
  – Increases BW, receiver work
Latency Resources

latency.org

(being updated!)