CS 856 – Latency in Communication Systems

Winter 2010

Latency Challenges

Overview

- Sources of Latency
  - low-level mechanisms
  - services
- Application Requirements
- Latency Measurement
- Summary & Classification

Low-level Mechanisms

- Memory Access (Hierarchy)
- Bus Contention
- Disk Access
- Asynchronous Processing
- CPU Scheduling
- Network Transmission
  - packetization, processing, queueing, propagation

Services

- System Calls
- Transport
- Messaging
- Remote Service Invocation
- Encryption

Application Requirements

- Web Surfing
- Telecommunication
- Interactive Games
- Streaming
- User Interfaces
- Algorithmic Trading
- Monitoring & Control
- High-performance Computing

Low-level Mechanisms

Memory Access

- memory hierarchy
  - registers – O(1) cycles
  - cache – O(10) cycles
  - RAM – O(100) cycles
- relevant for
  - effective CPU throughput
  - e.g. routers:
    per-packet processing budget very small!

Bus Contention

- multiple access & arbitration
- similar to link layer medium access control (later)

Disk Access

- physical characteristics
  - head movement, speedup/slowdown
  - disk rotation
  - I/O transfer via bus

Asynchronous Processing

- cost of context-switching (and/or preemption)
- e.g. network stack
  - socket – queue – network – queue – interface
  - multi-threading
- send side: multiplexing with other traffic
- receive side: traditional interrupt levels

- OS and application programming

CPU Scheduling

- control of asynchronous processing
- priorities & deadlines, real-time
- waiting time on ready queue -> latency

Network Transmission

- transmission
  - serialization of data: (file size / link speed)
  - packetization: discrete units of data
    - cost at each node vs. cut-through processing
- medium access control & back-off
- processing
  - cf. memory latency
  - cf. asynchronous processing

Network Transmission (2)

- queueing
  - well-studied
  - queueing theory
  - packet scheduling: average vs. worst-case
- propagation
  - determined by fundamental physics
  - e.g. global communication: Toronto – Sydney ~10msec propagation (linear distance only)
  - e.g. inter-planetary communication...

Synchronous Networks

- e.g. SONET, CDMA
- no medium access control
- no queueing
- but: fixed upper bound rate
  - cf. transmission delay

Services

System Calls

- blocking system calls
  - typically I/O operations
- cf. asynchronous processing

Transport

- network transmission
- signalling
- reliability
  - timeout
  - retransmission

Messaging

- transport
- signalling
- encoding / decoding
  - might include segmentation & reassembly
- network stack
- process interaction (system calls)

Remote Service Invocation

- messaging and transport
- 2-way request / reply interaction
- e.g. name lookup
- e.g. “web services” (aka RPC)

Transcoding

- e.g. encryption, video transcoding
- computational overhead
- offloading (e.g. special-purpose chip)
- throughput vs. latency gains?

Application Requirements

Web Surfing

- browsing - user expectations
- responsiveness: a few seconds
  - 8- resp. 4-second rules
- commercial web services (B2B)
- retail: shopping, banking, payments
  - avoid duplicate transaction
  - “do not hit reload!” messages...
- online auctions, brokering, etc. -> much faster?

Telecommunication

- telephony
  - call setup and control
    - authentication & authorisation
    - e.g. cell phone roaming
  - data
    - human perception of interactivity
    - good: one-way 150-300 msec, end-to-end!
    - e.g. packetization: 64 kbit in 40x 200-byte packets/sec
      - 25 msec packetization
      - vs. header overhead (cf. ATM)

Telecommunication (2)

- videoconferencing
  - similar delay bound than audio (also: lip-sync)
  - need ~25 frames/sec for continuous presentation

    => upper bound on packetization
- collaborative work
  - e.g. document editing
  - “instantaneous” interaction
  - locking and event ordering important

Interactive Games

- event distribution / roundtrip
- FPS, Car Racing – similar to video/telephony
  - simulate continuous action – trick human brain
- RT Strategy – seconds ok?
- event order relevant
  - ignore/overwrite “old” events

Streaming

- startup delay – several seconds ok
  - cf. VCR, DVD
- live vs. stored content?
  - no difference?
- latency variation => playout buffer

User Interfaces

- asynchronous: multi-threaded or event-based
- need instantaneous response
  - show at least brief immediate effect (button click)
- and/or drive continuous presentation
  - e.g. video presentation + interactivity

Algorithmic Trading

- receive electronic market data
- compute price quotes
- feed back into the market

- everybody else does the same
- speed matters, i.e., fast reaction to market data
- ... at high volumes

Monitoring & Control

- examples
  - monitor factory automation
  - monitor nuclear power plant
  - control anti-lock brake system (ABS)
- reaction speed matters
- tight control loop

High-Performance Computing

- distributed computation
- compete with integrated systems
  - price/performance trade-off
- remote memory access
- communication and synchronization overhead

Wrap-Up

Latency Measurement

- round-trip latency
  - easy to measure with local clock
  - accuracy, statistics?
- one-way latency
  - difficult – requires synchronized clocks
  - derive from round-trip? symmetry?

Summary & Classification

- fundamental physical limitations
  - propagation speed and latency
- contention: bus, MAC, router, etc.
- latency vs. throughput
  - synchronous better for latency
  - asynchronous better for throughput
- average vs. worst-case considerations

Summary & Classification

- performance modelling
  - memory and communication latencies
    - chip, board, chassis, cluster, data center
- latency classes?
  - machine (ns) – computation
  - human subconscious (ms) – interactive
  - human conscious (s) – response time
  - asynchronous (m) – backup
  - human asynchronous (h) – email