Real-Time Data Communications

Raj Rajkumar
Lecture #17

Adapted from Perkins and Liu

Outline

- Real-time communications
  - Traffic and network models
  - Properties of networks
    - Throughput, delay and jitter
  - Congestion and loss
- Examples
  - Controller Area Networks
  - Ethernet

Real-Time Communications

- In digital data communications, it is highly desirable that network packets reach their destinations dependably
  - One would *like* fast delivery, but *not* at the expense of reliability
    - E.g. web browsing, e-mail, file transfer, twitter, etc.
  - These applications are often referred to as *elastic* applications
    - i.e. time can be "dilated"
- In real-time data communications, network packets must arrive on time
  - Timely delivery may be deemed to be *more* desirable than reliable delivery
  - Different levels of priority may be associated with applications
  - Examples:
    - Anti-lock braking in a car
    - "Fly-by-wire" systems in a modern aircraft
    - Skype internet telephony and IPTV (TV using the Internet Protocol)
      - drop delayed packets

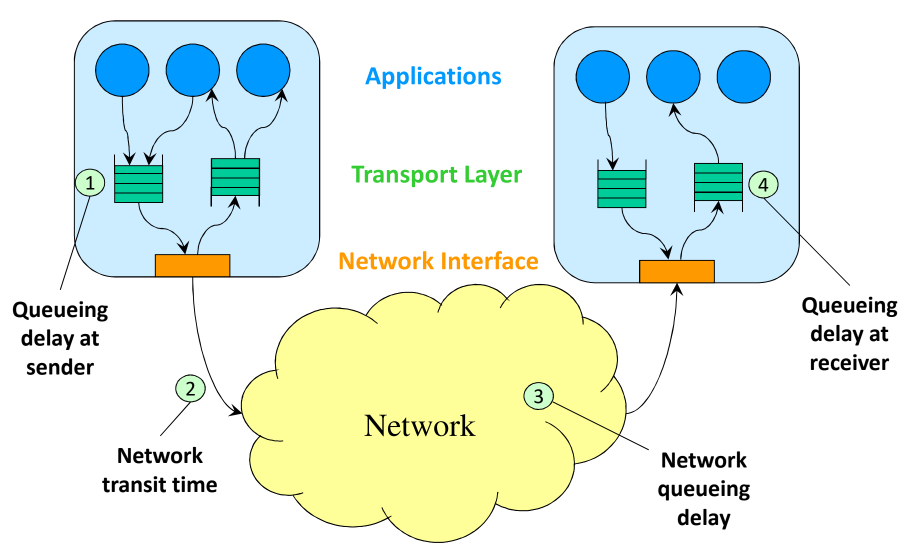
Real-Time Traffic Categories

- Packet-switched traffic falls into two categories:
  - Synchronous periodic messages
    - Produced and consumed in a continual basis, according to some schedule
      - Generally require some performance guarantee
      - Can be generated by periodic tasks
      - Fixed rate ("*isochronous*") flows (e.g. sensor data, speech)
      - Characterize by inter-packet spacing, message length, reception deadline
    - Can be generated by sporadic tasks
      - Variable rate flows (e.g. MPEG-2 video, control traffic)
      - Characterize by average throughput + maximum burst size
  - Aperiodic (asynchronous) messages
    - No deadline, best-effort delivery, but want to keep delays small
    - Characterize by average delivery time

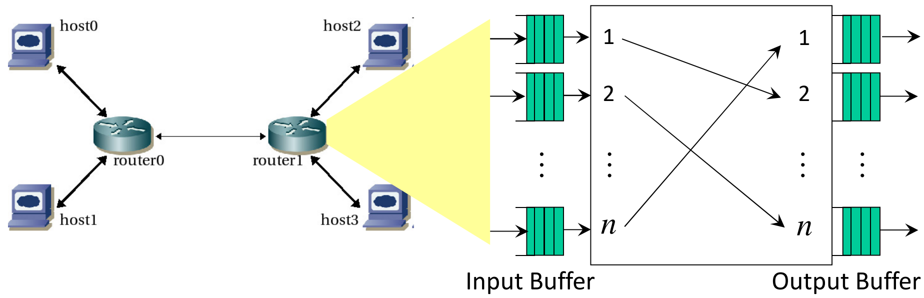
Sources of Message Delays

- Message delays on networks comprise of the following components:
  1. Queueing delay at sender
     - Network not always ready to accept a packet when it becomes available
     - Data may be queued if produced faster than the network can deliver it
  2. Queueing delay in the network
     - Due to cross-traffic or bottleneck links
  3. Network transit time
     - Fixed propagation delay
  4. Queueing delay at receiver
     - Application not always ready to accept packets arriving from network
     - Network may deliver data in bursts

Message Delay Components



Packet-Switched Networks



- Packet-switched networks are comprised of a multitude of network links connected between switches/routers and end-points
- Each network link has a constant *propagation delay*
- A network switch queues data packets for transmission if its output link is busy
  - Such queueing can cause a variable delay
- The *switching policy* on the output links is critical for time-sensitive traffic
  - Packets and/or ports can have priorities
  - "Traffic shaping" and reservations can also be done on network streams

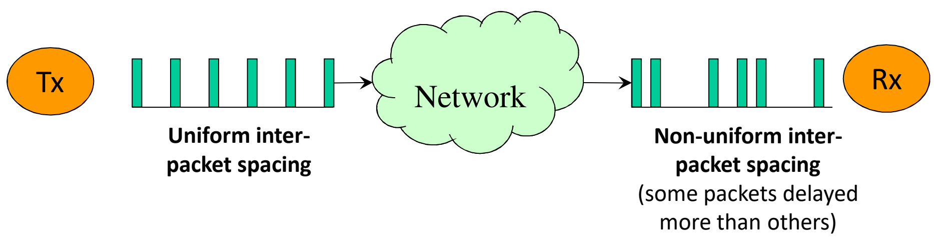
Performance Metrics

- Performance metrics include:
  - **Throughput**: a measure of the number of packets that the network can deliver per unit time
  - **Delay** (latency): time taken to deliver a packet
    - Fixed minimum propagation delay due to speed of light
    - Variation due to queuing on path
  - **Jitter**: Variance of the delay
  - **Buffer requirements**: amount of storage required so as not to drop packets
  - **Packet Miss rate**: ratio of packets that miss their timing constraints
  - **Packet loss rate**: ratio of packets that are not delivered
  - **Packet error rate**: ratio of packets that have an error in them

Throughput, Delay and Jitter

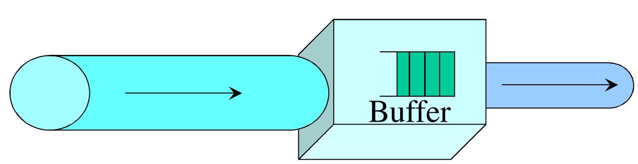
Throughput, delay and jitter vary according to router scheduling techniques.

- Possible to derive bounds for delay/jitter for some policies (e.g. RMA, Round-Robin, Weighted Round-Robin techniques)



Throughput and Delay

- Throughput and delay depend on the capacity of each link, and on the queuing delay at each hop
  - Queuing delay will vary based on the traffic
  - Throughput variations may cause queues to build up at bottleneck links
  - “Cross traffic” will also affect queue occupancy
  - Throughput may be limited by an intermediate link, which cannot be directly observed by sender and receiver
    - How to tell if the throughput is limited by the network, or by other traffic using the network?
    - Cannot know if capacity available, unless requirements signaled in advance

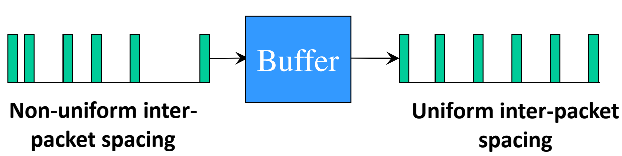


Throughput and Delay

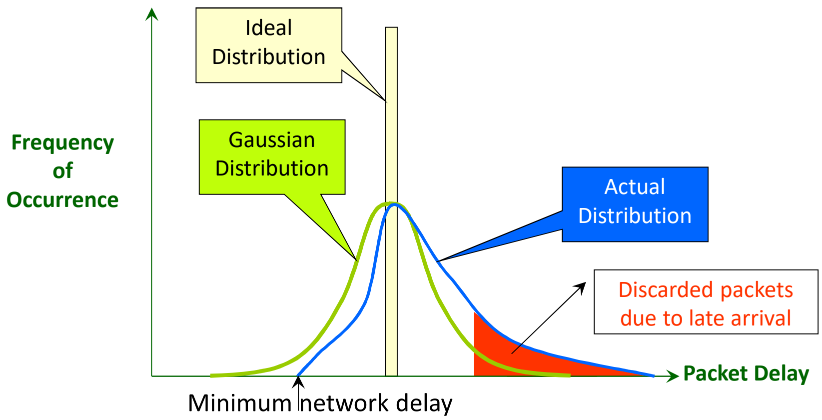
- Delay matters for some applications, but not others
  - Interactive applications need low delay
    - Telephony, video conferencing and games
    - Control applications often need low delay in the sensor ⇒ controller ⇒ actuator path
    - Propagation delay places a lower bound on delay
    - Queuing delay can be substantially more but is (hopefully) controllable
  - Non-interactive applications are less delay-sensitive
    - Watching YouTube/Google videos
    - TV and radio broadcasts
- Throughput is typically important
  - Need to sustain a certain rate in order to support the application
  - May wish to use scheduling algorithms to prioritize which packets are to be sent first (up to a certain limit)
  - **Adaptive** schemes can scale back throughput in case of overload

Jitter and Buffering Requirements

- Delay *jitter* is the variation in delay across a network path
  - For isochronous traffic, often talk about absolute value and standard deviation of packet inter-arrival time
  - Assumes we can characterise the jitter – see examples later
- Jitter imposes requirement for receiver buffering
  - Isochronous applications must be fed correctly-spaced data
    - Need buffer to smooth and reconstruct timing
    - Larger jitter implies more buffering is needed
  - Packet scheduling algorithms can bound jitter



Jitter and Miss-Rate



- Want to characterize jitter distribution
  - Hope for something approximating a Gaussian distribution ⇒ simple statistics to derive the *miss rate*
    - Fraction of packets lost due to jitter
  - Actual distribution is (much?) more complex

Congestion and Loss

- Assumed that no traffic is ever blocked or lost because there is no space in the ready queue when it becomes available for transmission
  - Usually valid for operating systems and LAN communication
  - Not valid for many wide-area communication systems
    - Too expensive to provision buffering in all routers
    - Provision for typical load plus a safety factor, not the worst case
    - Queues may overflow, hence packets are dropped
      - The *loss rate* gives the fraction of packets that are dropped
      - Patterns of loss may also be important: affected by packet scheduling algorithms
- Packets may also be dropped due to corruption or other errors

Congestion and Loss (cont'd)

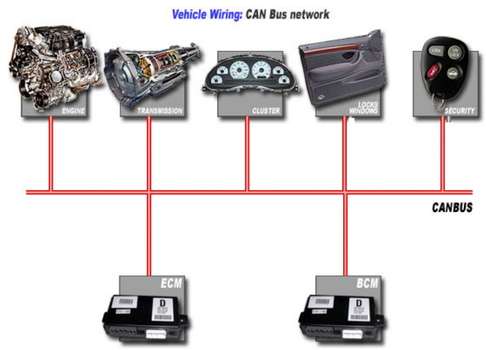
- Both flow characteristics and cross-traffic can cause overloads and congestion
  - Temporary congestion will cause queueing delays
  - Persistent congestion will result in queues that stay full, hence packets may be lost
- How to avoid this?
  - Control the amount of traffic at a bottleneck link
    - Applications need to signal their requirements
    - Network performs admission control
  - Or prioritize traffic to give preference to important flows
    - What scheduling algorithm to use?
      - Fixed-priority schemes are much easier to implement than dynamic priority schemes
      - Weighted round-robin techniques are also available
    - May allow real-time traffic, but discard best-effort data traffic when the network is overloaded

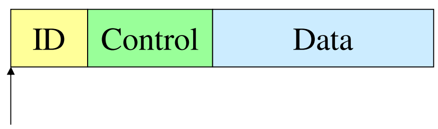
Characterization of Networks

- Given a network technology and implementation, it is desirable that throughput, latency and jitter are within appropriate bounds for the application
- Some network technologies allow this, others do not
  - *Examples*: CAN (**C**ontroller **A**rea **N**etwork), Ethernet

Controller Area Networks (CANbus)

- Shared serial bus, send at 1Mbps, maximum bus length is 50m
- All stations hear transmissions within a fraction of a bit time
- Connections wired together as a logical AND function
- Packets start with an ID, then control and data
- Distributed Priority-based Arbitration:
  - Wait until start of slot, then begin to send with the ID field, but:
  - Stop if you hear a 0 on the bus when you are sending a 1
  - Packet with the smallest ID is transmitted first
    - Priority-based network access





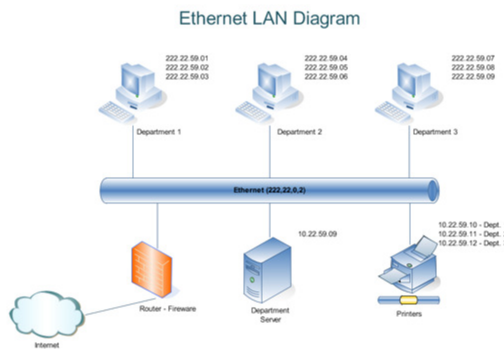
Start of packet

CAN (cont'd)

- Widely used in automotive systems
- Communications can be scheduled using fixed-priority scheduling algorithms
  - Look at the communications patterns, assign deadlines to each message exchange
  - Use deadline-monotonic scheduling to assign priorities
    - 11-bit ID field, implies 2048 priority levels (or message IDs)
  - Treat sporadic messages as periodic messages, according to the worst-case assumptions
    - Wastes capacity, but ensures schedulability
  - CAN cannot preempt or interrupt a message once it has started
  - Low utilization, but can prove that all messages will be delivered before their deadlines and calculate jitter
    - Standard schedulability analysis, as for any set of jobs

Example: Ethernet

- Ethernet uses CSMA/CD with exponential back off
  - Before transmitting, check for a link
  - If not active, try to transmit, liste for collision
  - If a collision occurs, stop sending before retry
  - Random binary exponential bacl
    - After *i* collisions, back-off by 2*i* slots, randomly chosen
- Potentially unbounded delay on network
  - Cannot schedule transmissions t collision
- No prioritization of messages



- Implications:
  - Throughput actually drops at high loads
  - Cannot easily reason about timing properties
  - Difficult to schedule messages to ensure timely delivery

Conclusions

- What is real-time communications?
- Factors that affect real-time communication
  - Throughput, delay and jitter
  - Clock skew
  - Congestion and loss
- Examples of networks and their timing properties
  - Some networks (like CANbus) provide timing guarantees, others (like the ethernet) do not