Quality of Service

Overview

➤ network impairments and congestion

➤ current status

➤ measurements

(Loosely based on Brian Carpenter's slides)

Fundamental Limits

- Shannon channel capacity with Gaussian noise for error-free transmission: $C = B \log_2(1 + S/N)$, with spectral bandwidth B; e.g., for telephone modem, $B = 3000$ Hz, $S/N = 35$ dB; thus $C = 34.8$ kb/s.

- imperfect detection of symbols with noise $\longrightarrow$ bit errors

- bit error rate (BER) from 10^{-12} for fiber to 10^{-2} for deep space . . .

- packet communications creates bit error multiplier effect: one bit error kills a packet

- bit errors not generally a problem except for wireless

- fundamental trade-off: delay $\leftrightarrow$ loss: in channel with bit errors, can only get perfect channel with infinite delay

- compensate for packet errors by forward error correction (redundancy) or retransmission (TCP)

Congestion

- competition for the same packet transmission

- need perfect coordination or infinite buffers or combinations

- congestion is unavoidable if $\sum \lambda_i > \mu$ where λ is arrival rate over some interval, μ service rate of output link

- *almost* like automobile traffic congestion

- rerouting is **not** the problem!

Internet Performance Problems

- packet delay

- delay *jitter* = variation in packet delay

- packet loss if buffers overflow

- for TCP: throughput variations in time and space: bytes/s $\longrightarrow$ MB/s

- end-system "congestion" $\approx$ network congestion, but single system!

Web Performance Problems

Huitema (Telcordia), Keynote:

- 20% of web page retrievals fail

- *commercial* web page up time of 90-99%

- 15% of HTTP GET > 10 seconds – many per page!

- DNS: 13%, transmit: 42%, connect: 12%, prepare: 33%

Delay
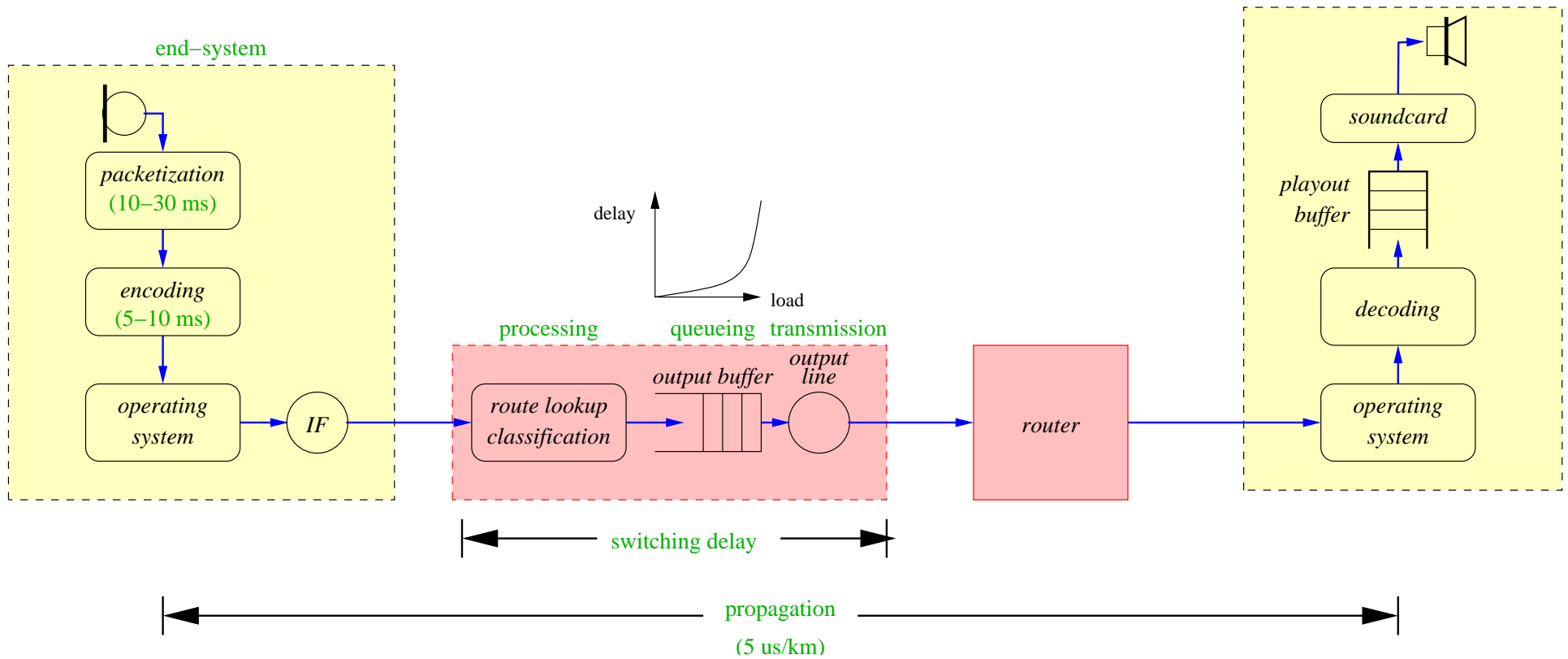
Packet Loss: Impact

- TCP: retransmission ($\longrightarrow$ danger of "congestion collapse"), slow start

- UDP: application-layer retransmit (DNS), audio/video distortion

Queueing Systems

- supermarket checkout

- cafeterias, banks: single line vs. multiple lines

- DMV, CU registration, . . .

- England

- buffers in routers, operating system

- service discipline: FIFO, LIFO, priority queue, SDF, . . .

Introduction to Queueing Models

- science and polling has the Gaussian "bell" curve, networks have *Poisson* model

- probability of event occuring per unit time is constant

- events are independent (i.i.d)

- models: phone calls, arrivals to post offices, ...

- Markov model: discrete time, continuous time – *memoryless*

- birth-death model

- Little's result: $N = \lambda T$, with N = number in system, T = system time (waiting + service)

- also: $N_q = \lambda W$, with W = waiting time

Poisson Distribution

- rate of events λ

- exponential interarrival: $P(T < t) = 1 - e^{-\lambda t}$

- $P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$ = probability of k arrivals in time t

- sum of two Poisson distributions = Poisson

- $M/M/1$ queue = Poisson arrivals, exponential waiting, single server

- more generally: $G/G/1/K$ = general arrivals, general waiting time, K buffers

- renewal paradox

Queueing Behavior

- for $M/M/1$:

$$T = \frac{1}{\mu - \lambda} = \frac{1/\mu}{1 - \rho} \tag{1}$$

- for $M/G/1$ (Pollaczek-Khinchin formula):

$$\bar{q} = \rho + \rho^2 \frac{1 + C^2}{2(1 - \rho)} \tag{2}$$

- reality is not $M/G/n$: "heavy tails"