Transfer initiation: Pull vs push access

Pull: The receiving end initiates the transfer
e.g. Web browsing:
small/minor transfer to server (URL)
large/important transfer to client (page+image)

Push: The source end initiates the transfer
e.g. email

Alternative: initiation by a third party

Has implications for:
• Signalling between switches:
  - Pushing reduces number of passes of signalling messages,
  - But is difficult when multicasting.
• Caching (and hence link dimensioning): Caches are useful only when pulling.

Transmission unit length

No optimal length:
• Transmission overhead: Packet overheads (address, interrupts) dominate for small packets, padding dominates for large (and small payload)
• Multiplexing granularity: Small reduces jitter, reduces packetisation delay, store-and-forward delay

Should the length be variable? …

Transmission unit length (continued)

Benefits of fixed length units (segments, cells, slots):
- Simpler, e.g. buffer management
- Predictable
- Fixed processing time helps concurrency, pipelining

Internal switch fabric can use fixed length units
While external line interfaces use variable-length units
With port processors converting between the two through segmentation & reassembly

To Be Continued in discussion about ATM…
Outline

“Quality of Service”

“Quality of Service” (QoS)

Anything that is desirable contributes to quality.

Service Level Agreements (SLAs) generally encompass measures such as timeliness, reliability, availability (discussed shortly).

Users often want “guarantees” of service ⇒ “QoS” often discussed in contrast to “best effort” alternative.

QoS in the research community vs deployment:
• Work has continued for decades… The difficulty is that all systems on the end-to-end path need to support it – any intermediary can disrupt the QoS
• ISPs don’t like QoS – it costs them money to install the equipment, and they can’t get the return – goes to content providers (ISPs are commodity dealers).

QoS mechanisms only matter when resources are scarce. Can often (not for some wireless) avoid fancy mechanisms by over-dimensioning system.

Timeliness requirements: 1st pass

Superficial dichotomy:
• “real-time”: Interactive systems
  • Continuous-time signals, e.g. voice and video (though only if ends interact “live” e.g. videoconference; tolerate delays for movie playback)
  • Other interactive systems: games, stock ticker, process control
  • “data”: file transfers, email, web browsing
  • The protocols used here have some interactivity – e.g. timeout after a certain period

“Real-time” requires QoS, whereas best effort is sufficient for data

Timeliness requirements in depth

Important aspects of delay:
• Mean - low for interaction
• Range/variation (CDV* - low for streaming media

*Cell Delay Variation - cell = short packet of fixed-length
Reliability

6 aspects of reliability defined:
- **Integrity**: What is received is what was transmitted.
- **Completeness**: Everything transmitted is received.
- **Uniqueness**: Information is only received once.
- **Sequence**: Information is received in the correct order.
- **Relevance**: Information from extraneous sources is not inserted in the midst.
- **Delivery**: Source receives acknowledgement indicating that destination has received information

Ties between switching and reliability

**Integrity**: Affects switch forwarding modes

**Completeness**:
- Packet discard strategies to avoid unnecessary retransmissions.
- Loss priorities for layered video (baseband + enhancement)

**Uniqueness**: Negligible impact on switches.

**Sequence**:
- Switches may misorder when rearranging or recirculating traffic.
- Deflection networks may misorder.
- Preservation of sequence is a benefit of connection-oriented switching.

**Relevance**: Protect addressing information in headers with CRCs

**Delivery**: Middleboxes may break end-to-end semantics

Loss vs delay variation

References in Keshav

Layering: Chapter 5
Addressing: Chapter 10