University of Glasgow

Real-time and Interactive Applications

Networked Systems (H) – Lecture 7

2019-2020

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow | This work is licensed under the Creative Commons Attribution-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/4.0/ or send a letter to Creative Commons, PO Box 1866,

Lecture Outline

- Real-time media over the Internet
  - What is real-time traffic?
  - Requirements and constraints
  - Quality-of-service and quality-of-experience
- Interactive applications
  - Conferencing application architecture
  - Signalling, session description, and real-time media traffic
- Streaming applications
  - HTTP adaptive streaming for video on demand

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

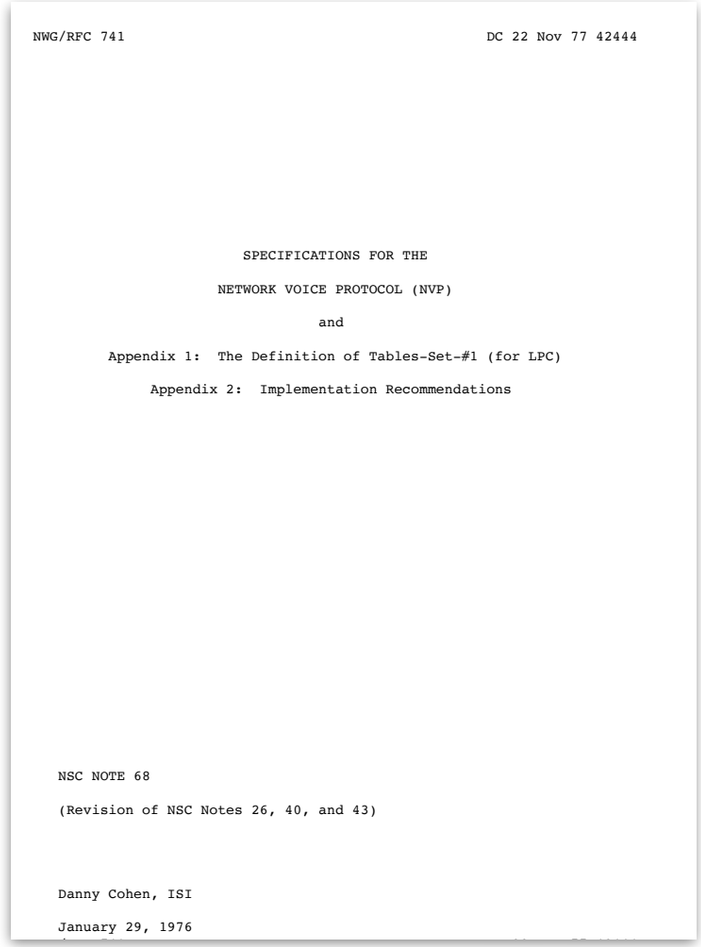
Real-time Media Over The Internet

- What is real-time traffic?
- Requirements and constraints
- Quality of service vs. user experience

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Real-time Media Over The Internet

- Long history of real-time media over the Internet:
  - Telephony and Voice-over-IP (VoIP)
  - Video conferencing
  - Internet radio
  - Streaming video and TV
  - Gaming
  - Sensor data
  - Industrial control
- Initial packet voice experiments → 1970s
  - Danny Cohen – Network Voice Protocol
- Core standards defined → mid-1990s
  - SIP, SDP, RTP, RTSP, H.323
- HTTP adaptive streaming → mid-2010s
  - MPEG DASH



NWG/RFC 741 DC 22 Nov 77 42444

SPECIFICATIONS FOR THE

NETWORK VOICE PROTOCOL (NVP)

and

Appendix 1: The Definition of Tables-Set-#1 (for LPC)

Appendix 2: Implementation Recommendations

NSC NOTE 68

(Revision of NSC Notes 26, 40, and 43)

Danny Cohen, ISI

January 29, 1976

https://datatracker.ietf.org/doc/rfc741/

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

What is Real-time Traffic?

- Defining characteristic: **deadlines**
  - The system fails if data is not delivered by a certain time
    - Railway signalling must deliver data to change a signal before train arrives – **hard real-time**
    - Streaming video should deliver a new frame every 60th of a second, else the movie playback stutters – **soft real-time**
    - What is the probability a deadline is missed? What are the consequences? Engineer accordingly – **no system is 100% reliable**
  - Deadlines can be absolute or relative
    - Data must be delivered before a certain time – **absolute deadline**
    - Data must be delivered periodically, within some time after the previous – **relative deadline**
- Real-time is not necessarily high performance
  - It requires **predictable timing**, not necessarily high bandwidth or low latency

cc BY ND

Requirements for Streaming Applications

- Video on demand has no absolute deadline, but needs regular data to prevent stalls once started
  - Acceptable if a takes a few seconds to start the video
  - Playback should be smooth once it starts
- Live video may have absolute deadlines
  - e.g., live sports playout

NETFLIX

BBC iPlayer

YouTube

- Bit rate depends on desired quality
  - Higher quality is better, up to limits of display
  - Predictable, but lower, quality often preferred to more variable, but on average higher, quality
- For a given bit rate, trade-off between frame rate and frame quality
  - Smoother motion or higher resolution?

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Requirements for Interactivity

- Requirements for interactive applications determined by task and human perception
  - Phone call or video conference
    - One-way mouth-to-ear delay ~150ms maximum for telephony
    - Video conferences want to lip-sync audio and video
      - Audio should be no more than 15ms ahead, or 45ms behind, video

ITU Rec. G.114

ATSC IS-191 “Relative Timing of Sound and Vision for Broadcast Operations”, 2003

  - Lecture style
    - Mostly unidirectional with occasional questions → can tolerate much higher latency
  - Distributed music performance
    - One-way latency ≪50ms desirable
    - Speed of sound: ~15ms to go from one side of a large orchestra to the other


INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU

G.114

(05/2003)

SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

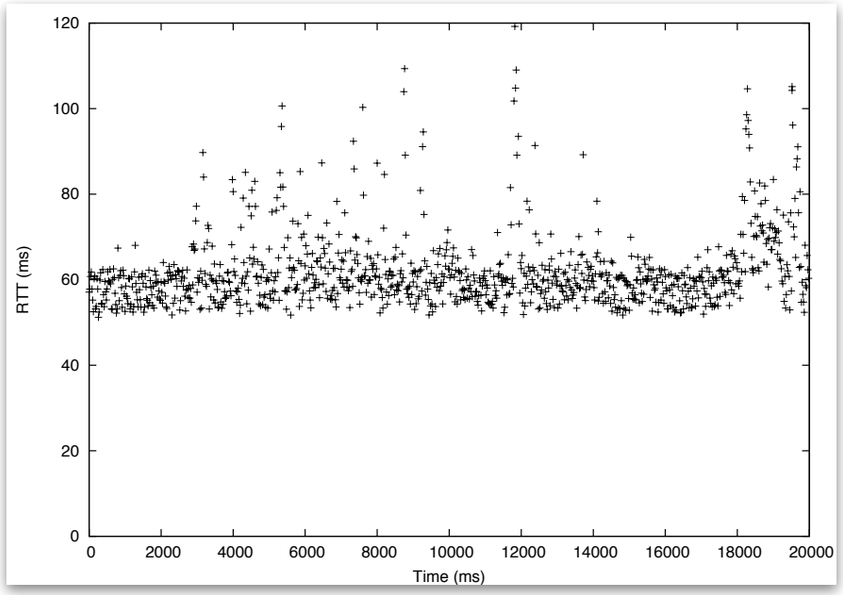
International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection

One-way transmission time

ITU-T Recommendation G.114

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Real-time Traffic Must Be Loss Tolerant



- The Internet is a best effort packet network
  - Timing is not guaranteed – depends on the amount of queueing on the path and on the path taken
  - Packets can be lost – retransmissions take time, and may not arrive before the deadline
  - In some networks these effects can be significant – with careful engineering they can be insignificant
- Real-time applications need to be loss tolerant

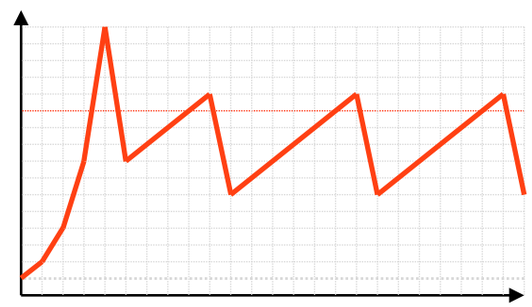


Data transport can be timely or reliable: not possible to be both simultaneously

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Real-time Traffic Has Limited Elasticity

- TCP congestion control adapts transmission speed to match available capacity
  - → Lecture 6
- Many transfers are **elastic** – faster is better, but it doesn't matter what rate the congestion control selects
  - The application can adapt to any chosen rate
- Real-time traffic is **inelastic**
  - Media has a minimum rate, below which it cannot be used
    - e.g., a minimum quality level below which the speech is unintelligible, that needs a certain bit-rate to transmit
    - If the congestion control algorithm cannot sustain this rate – if the network has insufficient capacity – real-time traffic cannot be used
  - Media has a maximum rate → cannot consume more
    - Video limited by capture frame rate and resolution
    - Audio limited by capture sampling rate



Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Quality of Service (QoS) Guarantees

- It may be possible to reserve network capacity for real-time traffic
  - Using RSVP, MPLS, 5G network slicing, …
  - Requires flow setup signalling, authorisation, and accounting
    - Need to tell the network what resources the real-time traffic requires
    - Need to demonstrate that sender is allowed to reserve these resources and can pay for the reserved capacity – reservations prevent other users from accessing the network, so have some cost
  - If the network has capacity for the traffic, reservations don't help
  - If the network does not have capacity, they allow the operator to discriminate in favour of customers who are willing to pay
    - Some customers care enough to pay for reserved capacity
    - Inter-domain reservations and accounting can be difficult to agree and implement
    - Many operators find it cheaper and easier to just buy more capacity, so no need to reserve

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Quality of Experience (QoE)

- What ultimately matters is subjective quality of experience
  - Does the application meet user needs?
  - Does the application allow users to communicate effectively?
  - Does the application provide compelling entertainment?
- QoE is not a one-dimensional metric
  - What aspect of user experience are you evaluating?
    - e.g., “Does it sound good” is not the same as “Can you understand it?”
  - How does your metric relate to the task being performed?
  - Some aspects of user experience can be estimated from technical metrics
    - e.g., the ITU-T E-model for speech quality evaluation based on latency and packet loss
  - Some aspects are subjective and task dependent – need user trials

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Interactive Applications

- Architecture for video conferencing
- Signalling and media traffic
- SIP, SDP, RTP → WebRTC

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Interactive Applications

- What are interactive applications?
  - Telephony
  - Voice-over-IP (VoIP)
  - Video conferencing



Multi-party video conferencing over IP pilot study, 1998

- Long history of research and standardisation:
  - Network Voice Protocol (NVP)
    - Packet voice experiments over the ARPAnet
    - RFC 741 published in 1976
  - Modern standards development from mid-1990s
    - Mbone conferencing tools
    - SIP, SDP, and RTP protocols
    - Adopted by 3GPP as basis of mobile telephone standards
  - Browser-based conferencing from mid-2010s
    - WebRTC

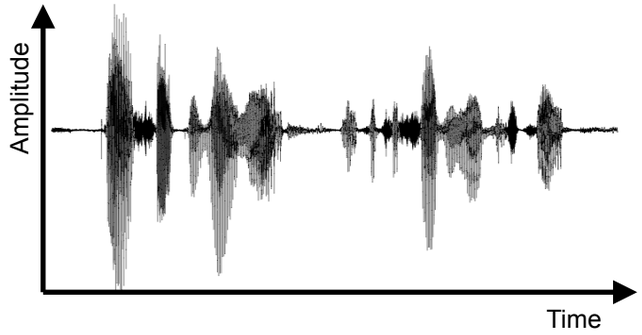


CU-SeeME (source: Wikipedia)

BY ND

Requirements on Timing and Data Rate

- Interactive conferencing has tight latency bounds
  - One-way mouth-to-ear delay ~150ms maximum for telephony
  - Video conferences want to lip-sync audio and video
    - Audio should be no more than 15ms ahead, or 45ms behind, video
  - User experience degrades gracefully
- Data rates depend on media type and encoding options
  - Speech coding typically operates on 20ms packets
    - Captures one frame of speech every 20ms, encodes, transmits
      - Data rate ~tens of kilobits per second
      - Background noise during quiet periods encoded at lower quality than speech, packets sent less often
  - Video frame rate and resolution highly variable
    - High definition video encoding using H.264 is around 2-4Mbps
    - Frame rates between 25 and 60fps commonly used
    - I-frames → tens of packets; P-frames → single or few packets



Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Reliability Requirements

- Speech data is highly loss tolerant
  - Loss concealment algorithms can hide 10-20% random packet loss without noticeable quality impairment
  - Burst losses are less well concealed
- Video packet loss hard to conceal
  - No scene changes to reset decoder state to known good value
    - I-frame requests to reset decoder, but generate bursts of data
  - Retransmissions possible in some cases; forward error correction more typical

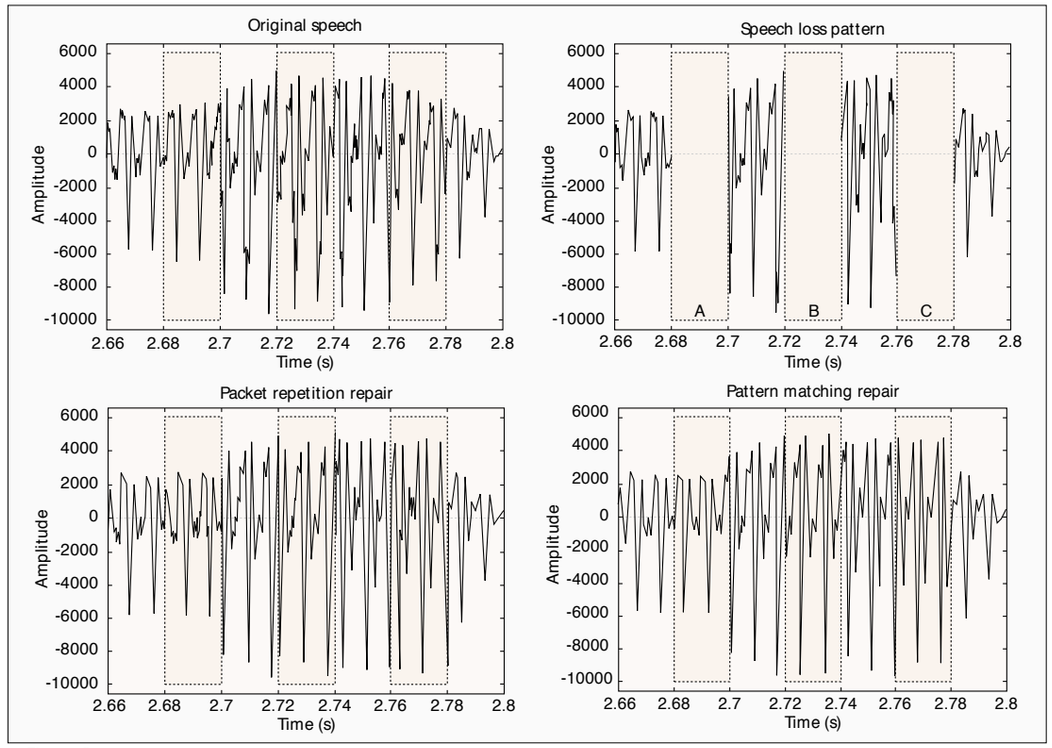
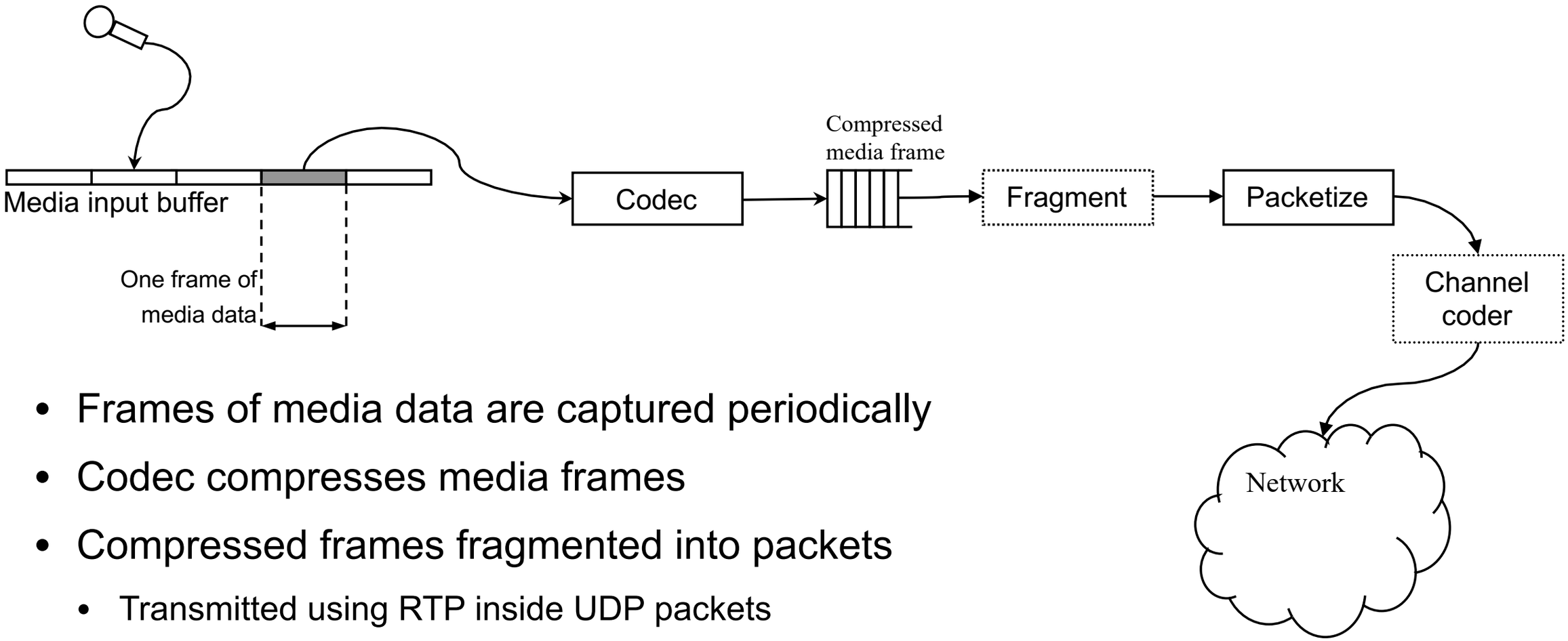


Figure 9. *(a) Sample error concealment techniques: original audio signal; (b) sample error concealment techniques: the loss pattern; (c) sample error concealment techniques; packet repetition; (d) sample error concealment techniques: one sided waveform substitution.*

C. S. Perkins, O. Hodson, and V. Hardman, A Survey of Packet Loss Recovery Techniques for Streaming Media, IEEE Network Magazine, September 1998. DOI:10.1109/65.730750

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Interactive Applications: Media Transmission Path

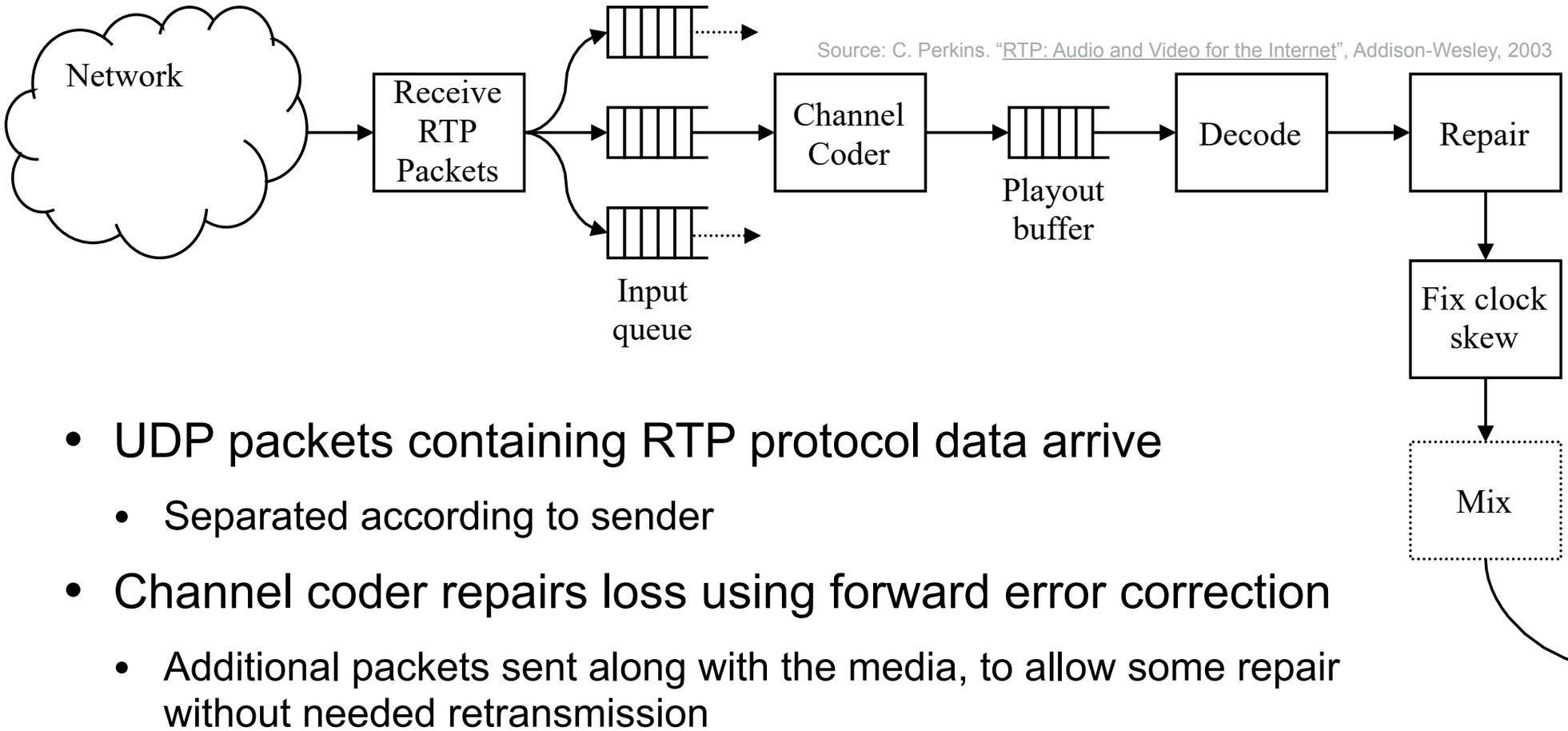


- Frames of media data are captured periodically
- Codec compresses media frames
- Compressed frames fragmented into packets
  - Transmitted using RTP inside UDP packets
  - RTP protocol adds timing and sequencing, source identification, payload identification
- Transmitted over the network

Source: C. Perkins. "RTP: Audio and Video for the Internet", Addison-Wesley, 2003

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Interactive Applications: Media Reception Path

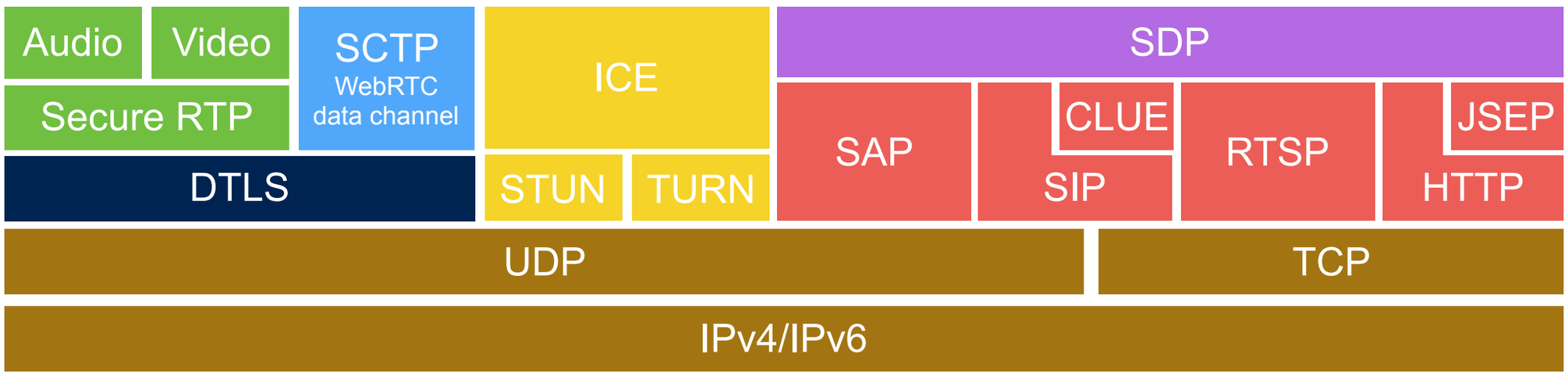


Source: C. Perkins. "RTP: Audio and Video for the Internet", Addison-Wesley, 2003

- UDP packets containing RTP protocol data arrive
  - Separated according to sender
- Channel coder repairs loss using forward error correction
  - Additional packets sent along with the media, to allow some repair without needed retransmission
- Playout buffer used to reconstruct order, smooth timing
- Media is decompressed, packet loss concealed, and clock skew corrected
- Recovered media is rendered to user

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Internet Multimedia Standards

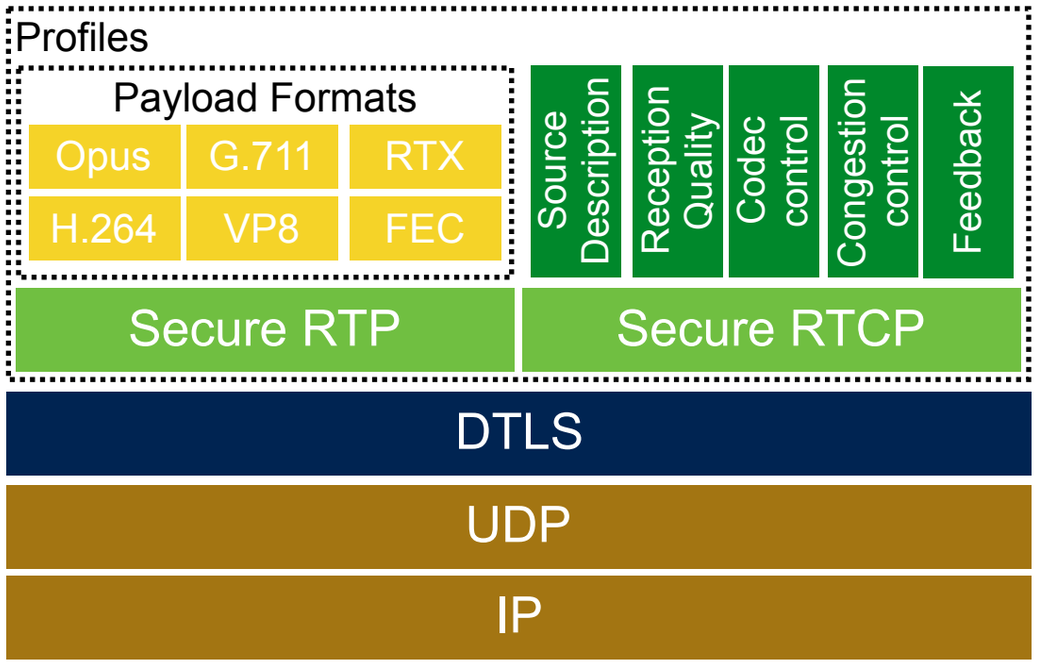


- Media: secure RTP, WebRTC data channel
- Path discovery and NAT traversal: ICE, STUN, TURN → Lecture 2
- Session descriptions: SDP
- Signalling protocols for different purposes
  - Announcing multicast sessions: SAP – *obsolete*
  - Control of streaming media: RTSP
  - Control of interactive conferencing: SIP
  - Control of telepresence: CLUE – *not widely used*
  - Control of web-based interactive media: JSEP (WebRTC)

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

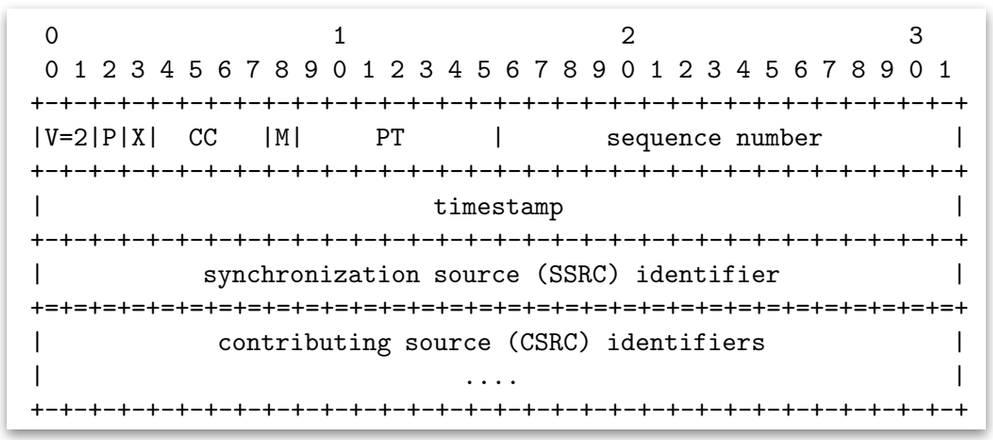
Media Transport: RTP

- Separate data and control channels
  - RTP – media payload formats
  - RTP Control Protocol (RTCP)
    - Source description and caller identity, reception quality feedback, codec control
- Payload formats
  - Codec-specific packet formats → application level framing → robust, but complex
  - Each frame packetised for independent use for low latency
- Datagram TLS – usual TLS handshake but within UDP packets
  - Exchanges keys used to encrypt following RTP data packets
- Extensions
  - Reception quality and user experience monitoring
  - Codec control and other feedback
  - Circuit breakers and congestion control



Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media Transport: RTP

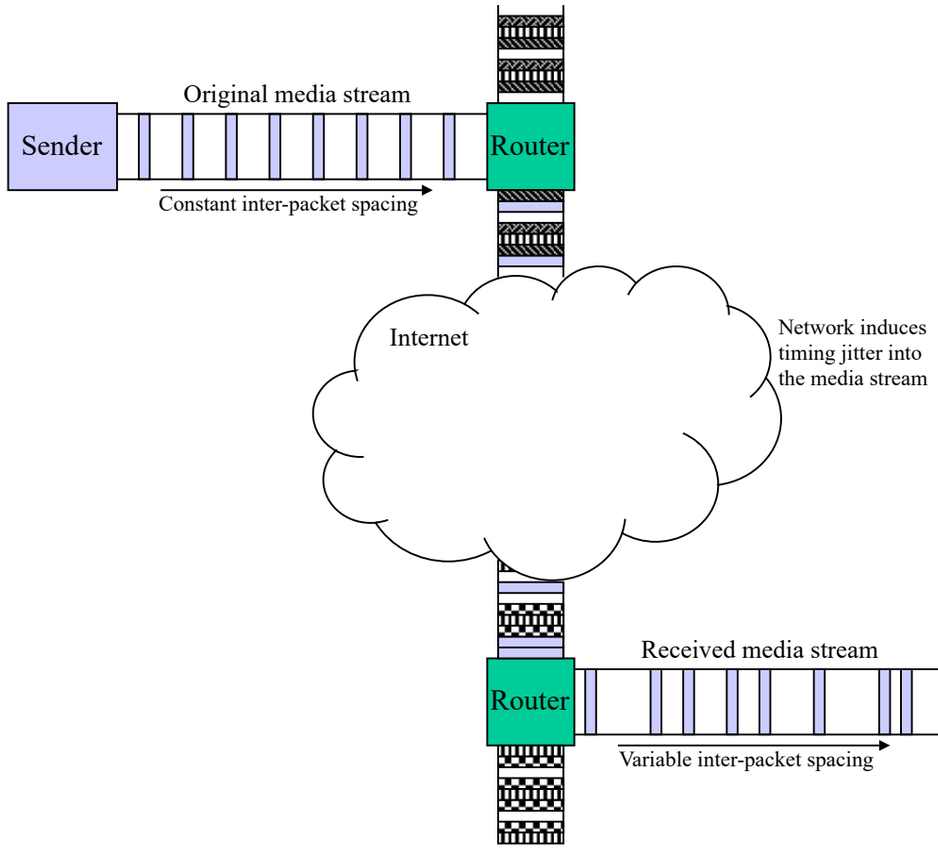


https://datatracker.ietf.org/doc/rfc3550/

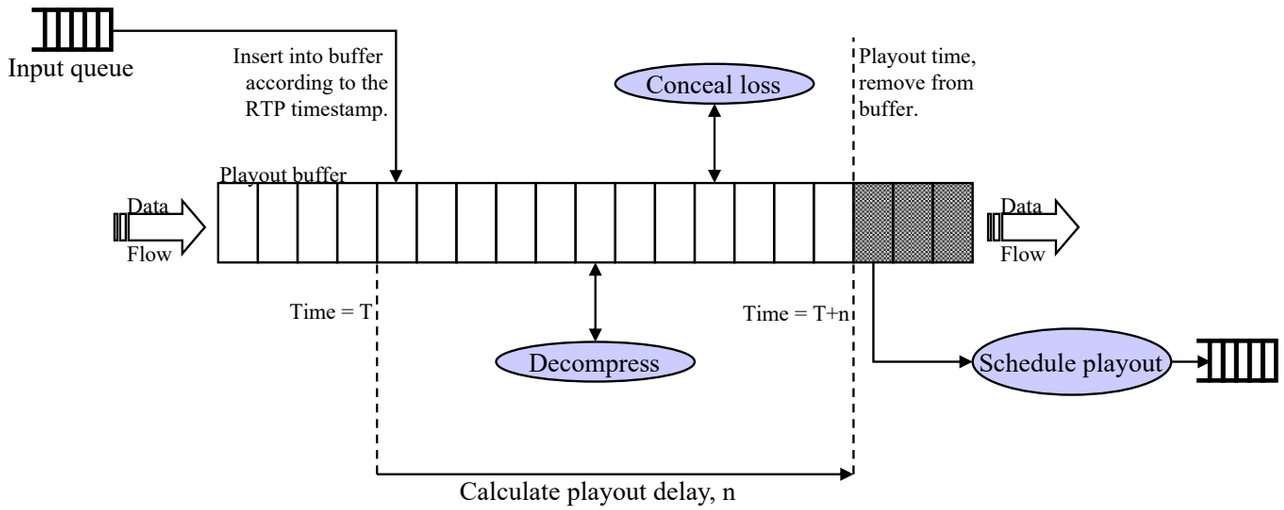
- RTP data packets carried within UDP
- Header information carries:
  - Sequence number and timestamp
    - Allows receiver to reconstruct ordering and timing
  - Source identifiers
    - Who sent this packet – needed for multiparty calls
  - Payload format identifier
    - Does the packet contain audio or video?
    - What compression algorithm is used?

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media Transport: Timing Recovery (1/2)



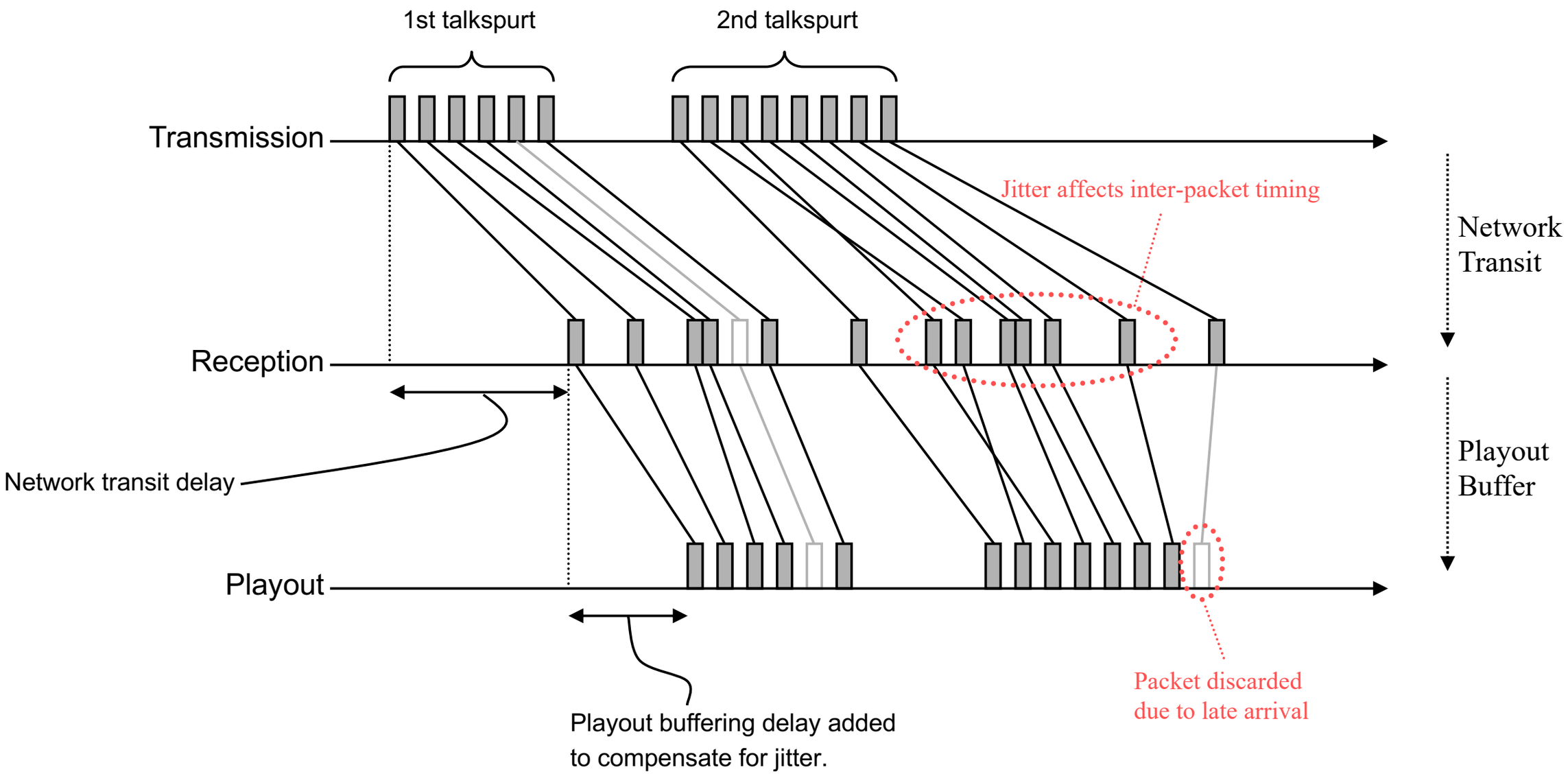
- Variable queueing delays in the network disrupt timing – receiver buffers to smooth out variation
  - If packets played out immediately on arrival, variation in timing leads to gaps
  - Delay playout by more than typical variation in inter-arrival time, to allow smooth play back



Source: C. Perkins. "RTP: Audio and Video for the Internet", Addison-Wesley, 2003

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media Transport: Timing Recovery (2/2)



Source: C. Perkins. "RTP: Audio and Video for the Internet", Addison-Wesley, 2003

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media Transport: Application Level Framing

- Packet loss is possible, so receivers must make best use of packets that do arrive
- RTP **payload formats** define how compressed audio/visual data is formatted into RTP packets
  - Goal: **each packet should be independently usable**
  - If a packet arrives, it should be possible to decode all the data it contains – not always possible, but desirable
- Naïve packetisation can lead to inter-packet dependencies where a packet arrives but can't be decoded because some previous packet, on which it depends, was lost



Architectural Considerations for a New Generation of Protocols

David D. Clark and David L. Tennenhouse
Laboratory for Computer Science, M. I. T.

Abstract

The current generation of protocol architectures, such as TCP/IP or the ISO suite, seem successful at meeting the demands of todays networks. However, a number of new requirements have been proposed for the networks of tomorrow, and some innovation in protocol structuring may be necessary. In this paper, we review some key requirements for tomorrow's networks, and propose some architectural principles to structure a new generation of protocols. In particular, this paper identifies two new design principles, Application Level Framing and Integrated Layer Processing. Additionally, it identifies the presentation layer as a key aspect of overall protocol performance.

1 Introduction

Historically, there has been a sharp distinction between the network architectures used in universal access environments, such as the telephone network, and the architectures developed within the computer communications community. We believe that both communities would benefit from a unified architecture that addresses new communication requirements. We have identified three key requirements that will stress the existing protocol architectures. Future networks will have considerably greater capacity, will be based on a wider selection of technologies, and will support a broader range of services.

The principle design goal for many research projects, and some commercial products, is gigabit operation, both in aggregate over trunks and to individual end points. As networks proceed to higher speeds, there is some concern that existing protocols will represent a bottleneck, and various alternatives have been proposed, such as outboard protocol processors [1, 11] and new protocol designs [2, 14, 12]. Since there is no clear consensus on the real source of protocol overhead, it is hard to validate these various solutions.

Another requirement is that the protocols of tomorrow operate over the range of coming network technology. An obvious example of emerging technology is Broadband ISDN [19, 6] which is based on Asynchronous Transfer Mode (ATM). This network technology provides a switching mode somewhat different from classic packet switching in that the data to be transferred is divided up into small (currently 48 byte) fixed size elements for transmission. A more futuristic network technology is wavelength division over fiber.

Finally, there is the requirement for service integration. In the future, a single end system will be expected to support applications that orchestrate a wide range of media (voice, video, graphics, text) and access patterns (interactive, bulk transfer, real-time rendering). This cross-product of media and access will generate new traffic patterns with performance considerations, such as delay and jitter tolerance, that are not addressed by present architectures.

In summary, future networks must exhibit a significant degree of flexibility and be based on an architecture that admits a wide range of application demands and implementation strategies. In the remainder of this paper, we identify some key principles for the elaboration of such an architecture.

2 Structuring Principles for Protocol Architectures

Most discussions of protocol design refer to the ISO reference model [9], or some other model that uses layering to decompose the protocol into functional modules. There is, however, a peril in using only a layered model to structure protocols into components, which is that layering may not be the most effective modularity for implementation.

It is important to distinguish between the architecture of a protocol suite and the engineering of a specific end system or relay node. The architecture specifies the decomposition into functional modules, the semantics of the individual modules and, the syntax used to effect the protocol. There should be no *a priori* requirement that the decomposition of the engineering design of a given system correspond to the architectural decomposition of the protocol. In the case of layered architectures, practical experience [4, 17] provides strong support for alternative engineering design. Unfortunately, the structure of the protocol architecture may unnec-

*This research was supported by the Defense Advanced Research Projects Agency, monitored by the National Aeronautics and Space Administration under contract No. NAG2-582, and by the National Science Foundation under grant NCR-8814187.

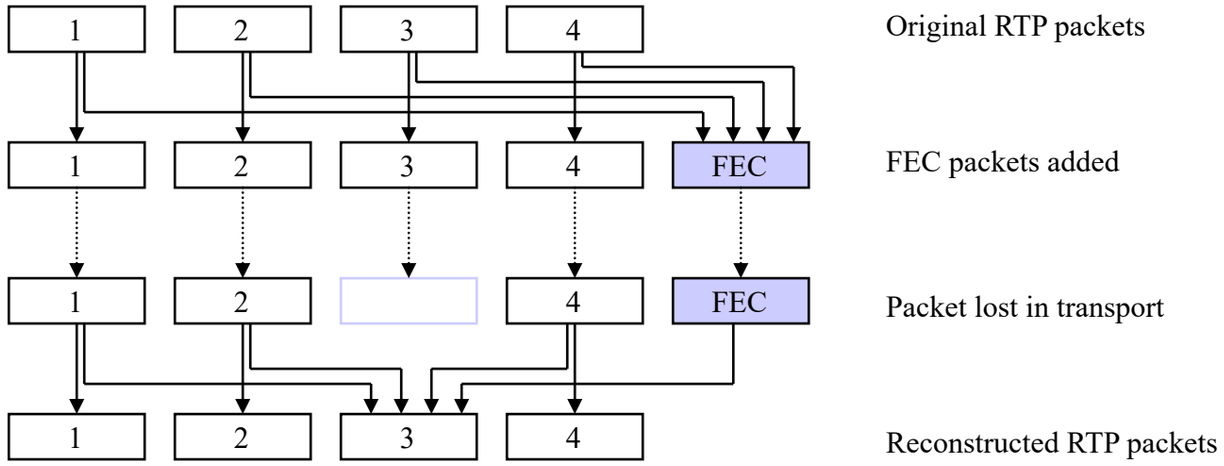
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1990 ACM 089791-405-8/90/0009/0200...$1.50

D. D. Clark and D. L. Tennenhouse. Architectural considerations for a new generation of protocols. In Proceedings of the SIGCOMM Conference, pages 200–208, Philadelphia, PA, USA, September 1990. ACM. https://dx.doi.org/10.1145/99508.99553/

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media Transport: FEC vs. Retransmission

- Retransmission possible, but often takes too long – packet should have been played out before retransmission arrives
- Forward error correction (FEC) often used instead

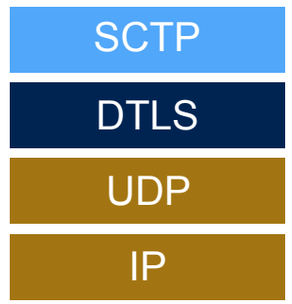


- Additional FEC packets are sent along with the original data
  - Contain error correcting codes
  - e.g., the Exclusive-OR (XOR) of the original packets – many different FEC schemes
- If some original packets are lost but the FEC packets arrive, original data can be reconstructed

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Media: WebRTC Data Channel

- In addition to audio-visual media, WebRTC application provide a peer-to-peer **data channel**
- SCTP in secure UDP tunnel:
  - Predates QUIC, but offers a similar service
  - Will likely migrate to QUIC over time



- Transparent data delivery:
  - Message-oriented abstraction
  - Multiple sub-streams
  - Congestion controlled
- Makes it straight-forward to build P2P applications with WebRTC

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Session Descriptions: SDP

- Media transport flows peer-to-peer for low latency
- A **signalling protocol** is needed to find the peer and establish the paths
- The control protocol needs to describe the communication session expected:
  - Media transport connections required
  - Media formats and compression algorithms
  - IP addresses and ports to use
  - Originator and purpose of session
  - Options and parameters

```
v=0
o=jdoe 2890844526 2890842807 IN IP4 10.47.16.5
s=SDP Seminar
i=A Seminar on the session description protocol
u=http://www.example.com/seminars/sdp.pdf
e=j.doe@example.com (Jane Doe)
c=IN IP4 224.2.17.12/127
t=2873397496 2873404696
a=recvonly
m=audio 49170 RTP/AVP 0
m=video 51372 RTP/AVP 99
a=rtpmap:99 h263-1998/90000
```

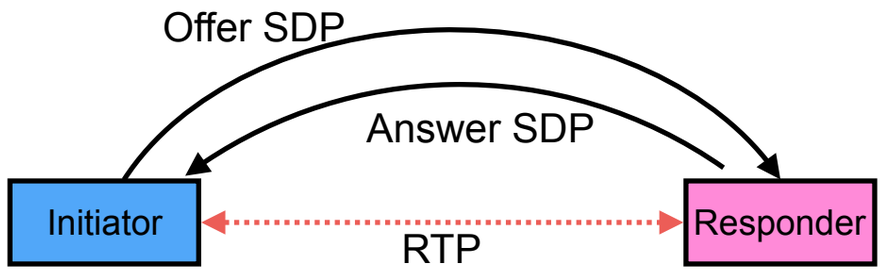
M. Handley, V. Jacobson, and C. S. Perkins, “SDP: Session Description Protocol”, IETF, RFC 4566, July 2006. https://datatracker.ietf.org/doc/rfc4566/

- **Session description protocol** (SDP) provides a standard format for such data

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Session Descriptions: SDP Offer/Answer

- Interactive sessions require negotiation
  - An **offer** to communicate: lists codecs, options and addressing details, identity of caller
  - The **answer** subsets codecs and options to those mutually acceptable, supplies addressing details, and confirms willingness to communicate
  - ICE algorithm (→ Lecture 2) probes NAT bindings, establishes path
  - Audio and video data flows

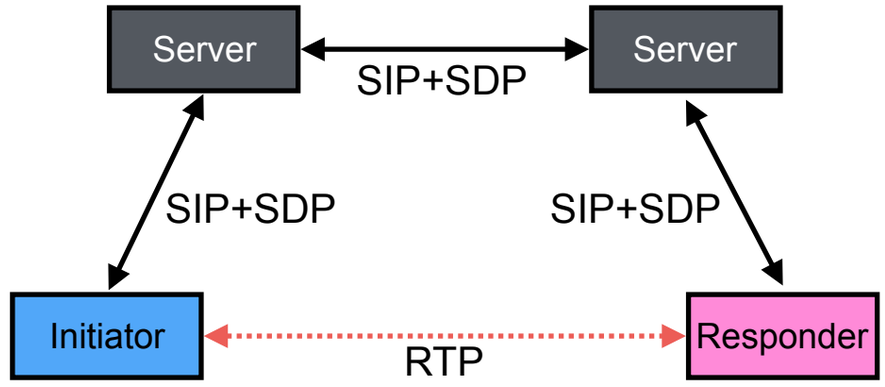


- SDP used as the negotiation format
  - SDP **was not** designed to express options and alternatives
  - Insufficient structure in syntax, semantic overloading
  - Complex → but complexity not initially visible; now too entrenched for alternatives to take off

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Control of Interactive Conferencing: SIP

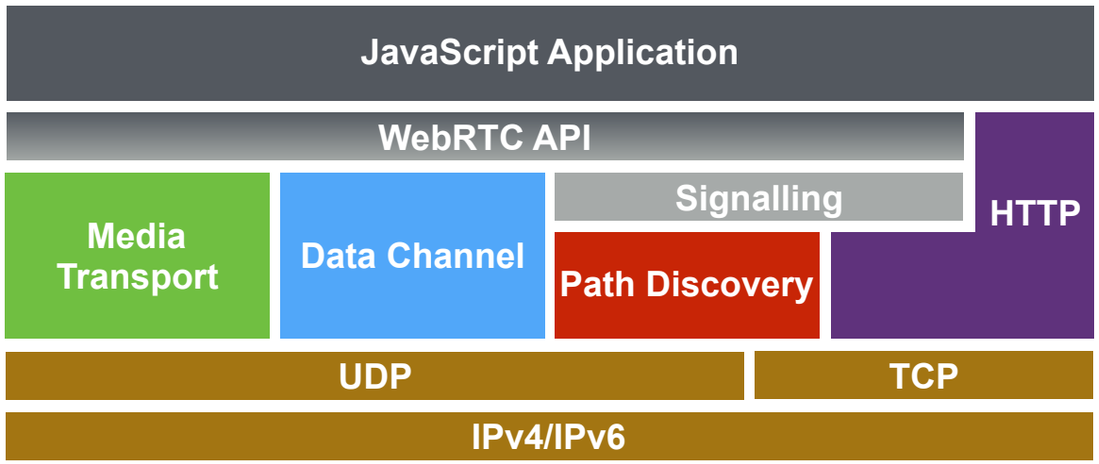
- The signalling protocol sets up the call
  - **Session Initiation Protocol** (SIP) for telephony and video conferencing
  - Conferencing servers with public IP addresses handle calls for a domain
  - SIP messages sent from initiator to responder via servers
    - Determine location of responder
    - Discover NAT bindings
    - Negotiate parameters and options for the call
    - An offer/answer exchange using SDP to describe the session being created
- Alert the responding user – make phone ring! – and agree to setup a call
  - NAT binding discovery and connection probing takes place while alerting
- Media then flows over direct peer-to-peer path

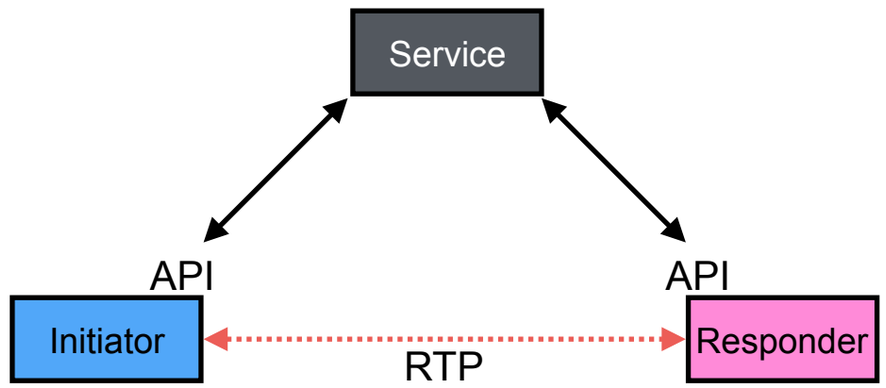


Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Browser-based Conferencing: WebRTC

- **WebRTC** is an alternative signalling protocol
  - Implemented in web browsers
  - Exposes standard JavaScript API





- Signalling messages delivered via HTTP to web server controlling the call
  - Offer/answer exchanges using SDP – conceptually similar to SIP; details different
- Media transport uses RTP – same as SIP
- Adds peer-to-peer data channel
- Designed to integrate video conferencing into web browsers and web applications

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Future Directions for Interactive Applications

- New media types – holographic, tactile, augmented and virtual reality
  - Every stricter requirements on quality and latency
  - All can fit within the basic framework described

- Media over QUIC?
  - Active research and standardisation – expect rapid developments in this space

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Streaming Video

- HTTP Adaptive Streaming

BY ND

Streaming Video Applications

- RTP **should** be usable for streaming video
  - Real-time Streaming Protocol (RTSP) developed by Real Networks for Internet TV in late 1990s
  - Uses unidirectional RTP for video delivery
  - Very low latency, loss tolerant



RTSP 2.0: https://datatracker.ietf.org/doc/rfc7826/

- **Most** video-on-demand, TV, and movies are delivered instead over HTTPS
  - Performance is significantly worse – very high latency – but startup latency is less critical
  - Integrates better with content distribution networks

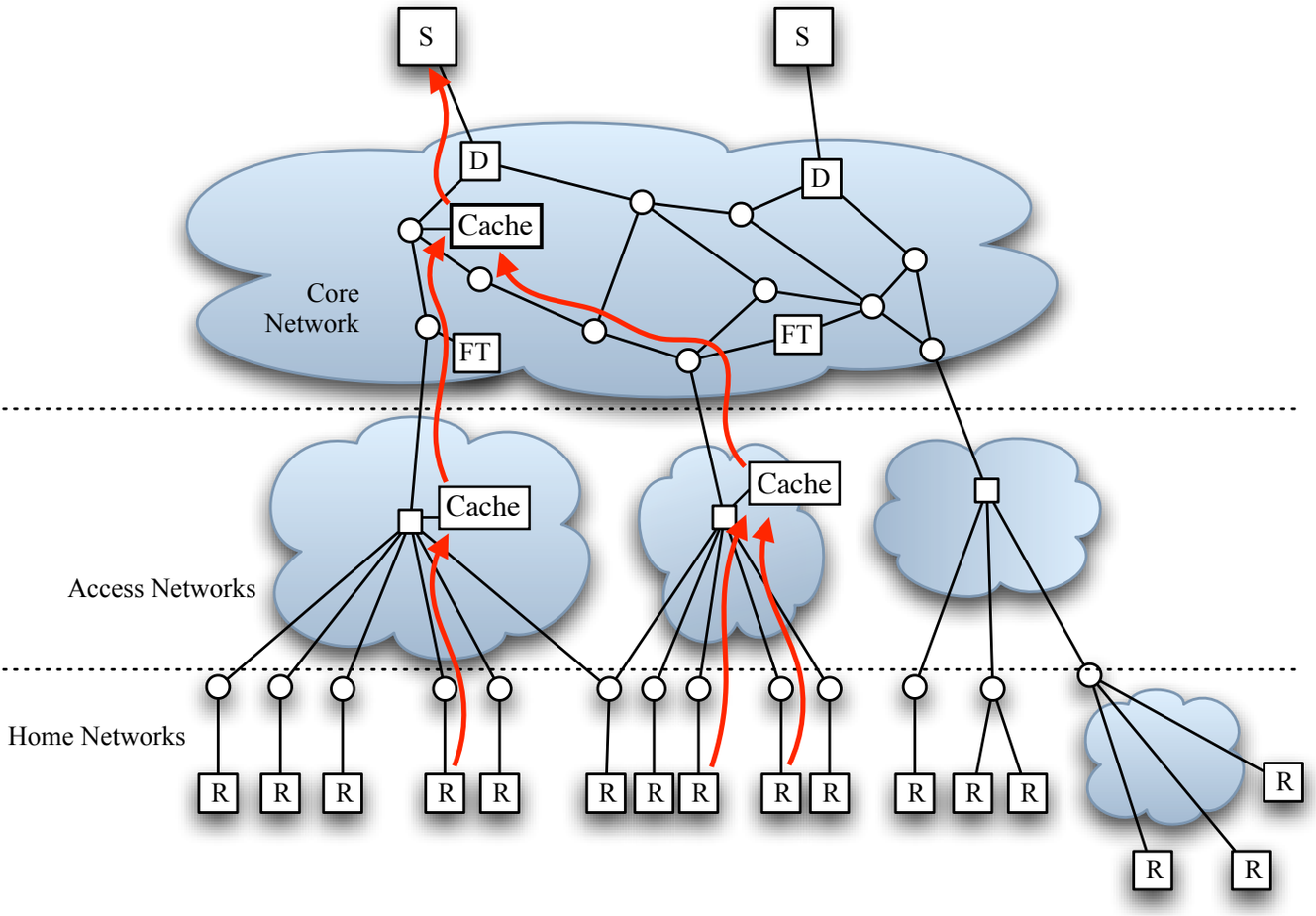






CC BY ND

Content Distribution Networks



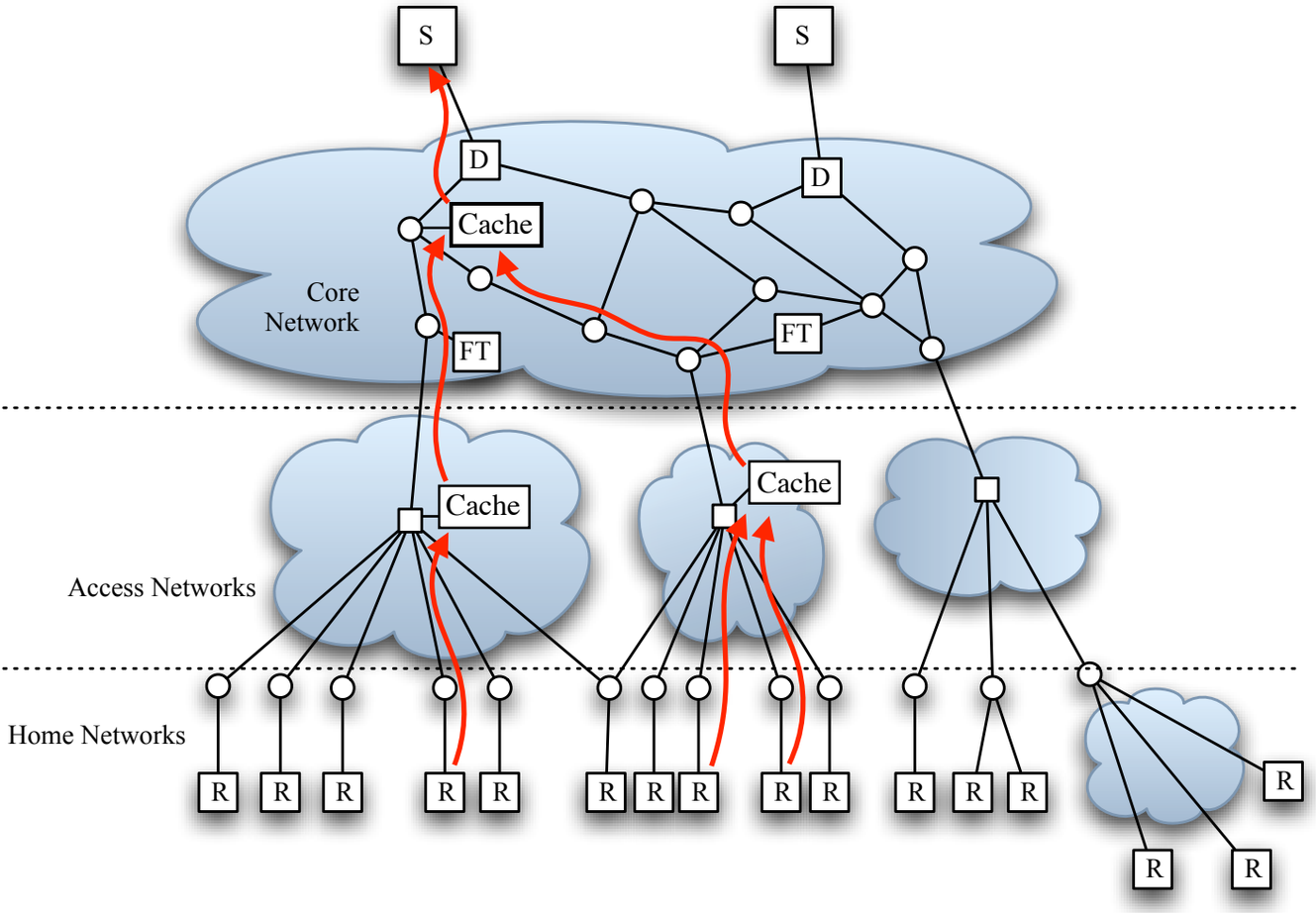
- There is an extensive, well-developed, content distribution infrastructure

Akamai CLOUDFLARE fastly

  - Highly effective for delivering and caching HTTP content
  - Global deployment – agreements with most large ISPs to host caching proxy servers
  - **Only works with HTTP-based content** – does not support RTP-based streaming

cc BY ND

Dynamic Adaptive Streaming over HTTPS (DASH)

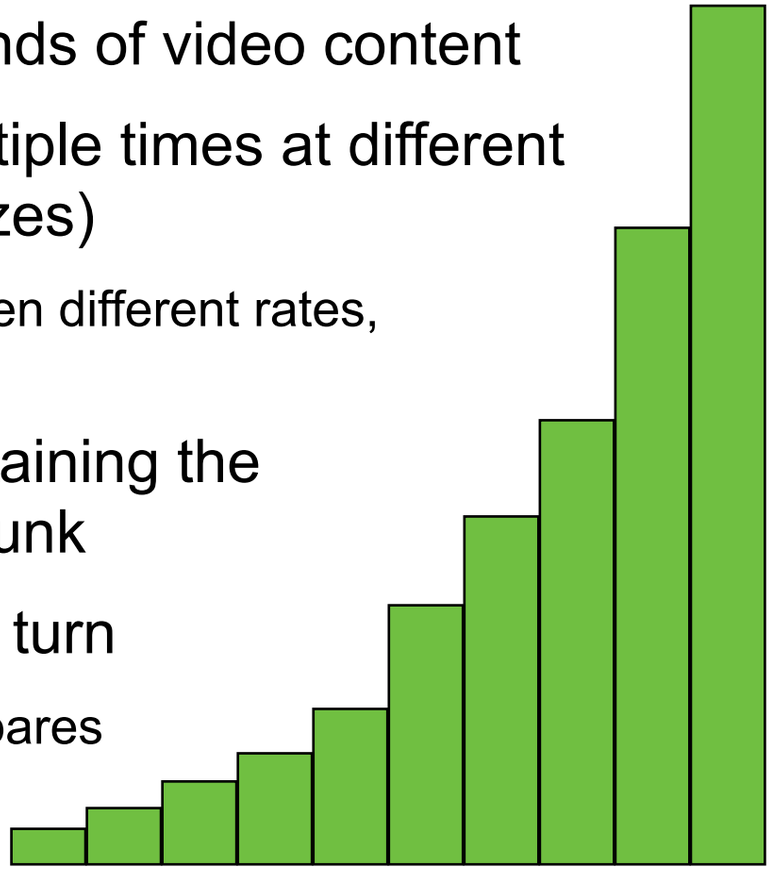


- **Deliver video over HTTPS** to make use of content distribution networks
  - Video content encoded in multiple **chunks**
    - Each chunk encodes ~10 seconds of video
    - Each chunk independently decodable
    - Each chunk made available in many different versions at different encoding rates
- A **manifest** file provides an index for what chunks are available
- Clients fetch the manifest, then download each chunk in turn
  - Standard HTTPS downloads
  - But monitor download rate, and choose what encoding rate to fetch next – adaptive bitrate
  - Playout chunks in turn, as they download

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

DASH: Chunked Media Delivery and Rate Adaptation

- Each chunk encodes ~10 seconds of video content
- Each chunk is compressed multiple times at different encoding rates (i.e., different sizes)
  - e.g., Netflix suggests encoding at ten different rates, as show on the right
- Receiver fetches manifest, containing the list of all versions of all each chunk
- Receiver fetches each chunk in turn
  - Measures download rate and compares to the encoding rate
  - If download rate slower then encoding rate, then switch to lower encoding rate for next chunk
  - If download rate faster than encoding rate, consider fetching higher rate for next chunk
- Chooses encoding rate to fetch based on TCP download rate



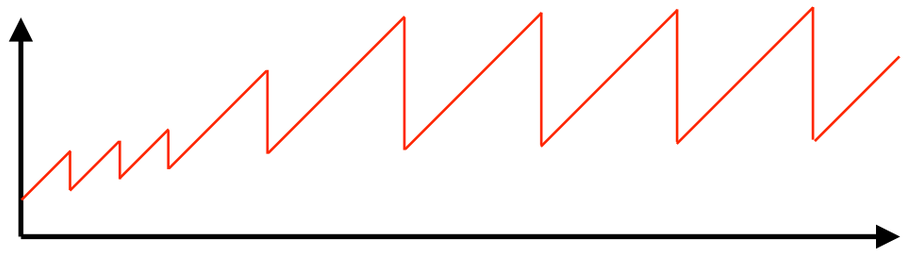
Bitrate (kbps)	Resolution
235	320x240
375	384x288
560	512x384
750	512x384
1050	640x480
1750	720x480
2350	1280x720
3000	1280x720
4300	1920x1080
5800	1920x1080

Source: Netflix

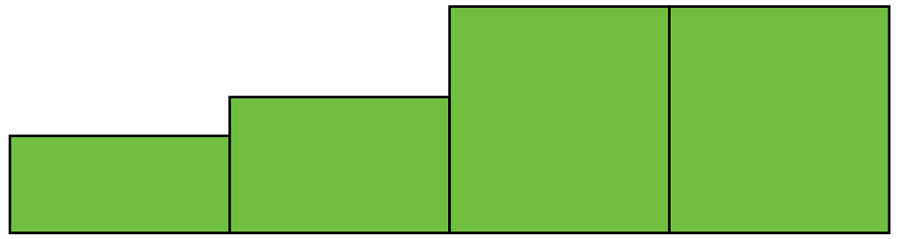
CC BY ND

DASH and TCP Congestion Control

- DASH rate adaptation and TCP congestion control work at different time scales
  - TCP adjusts congestion window once per RTT
    - Continually changing download speed
    - AIMD algorithm – TCP Reno or Cubic



  - DASH receiver measures **average download speed** of TCP connection over ~10 seconds
  - Selects size of next chunk to download based on average TCP download speed



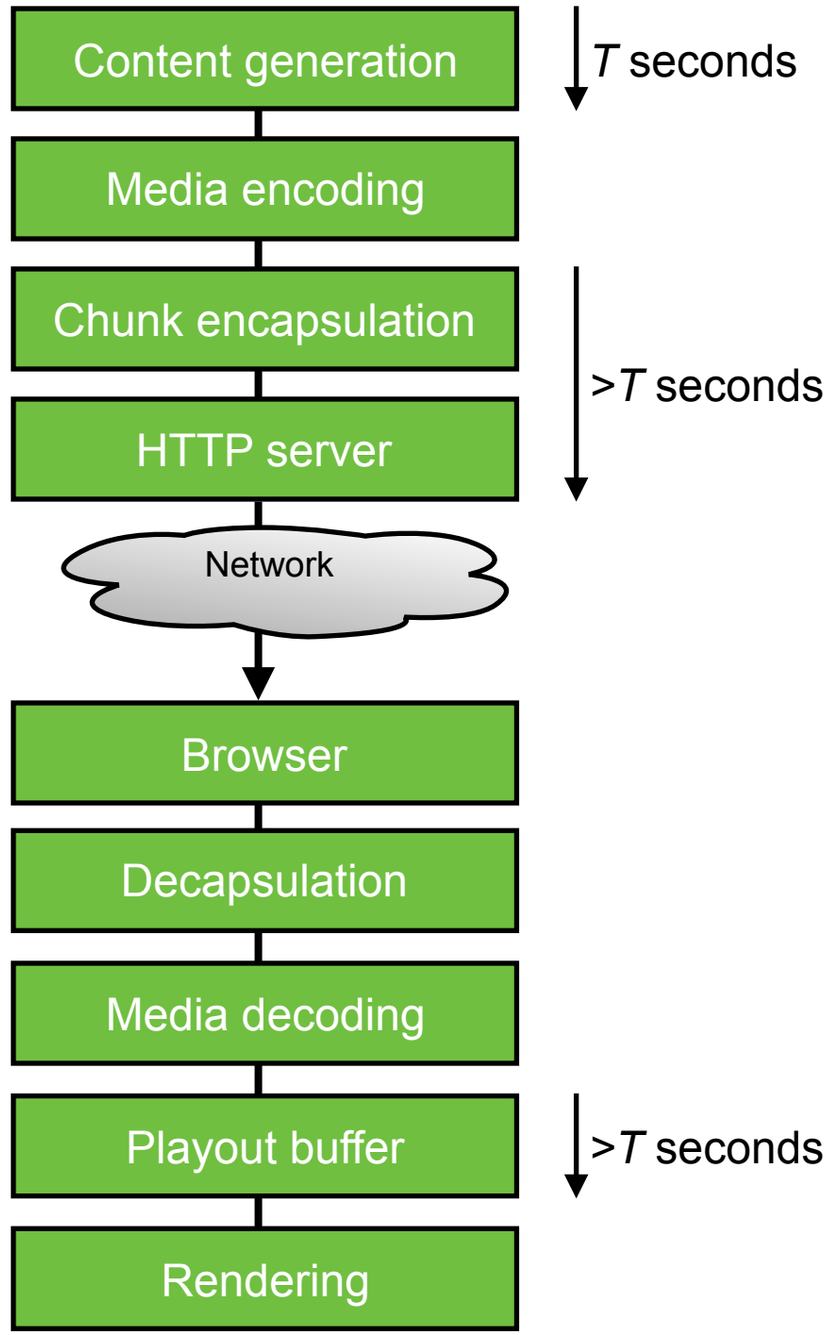
- Videos played using DASH often have relatively poor quality for first few seconds
  - Receiver picked a conservative download rate – relatively poor quality, small size – to start
  - Uses that to measure connection download speed, then switches to more appropriate rate

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

DASH Latency

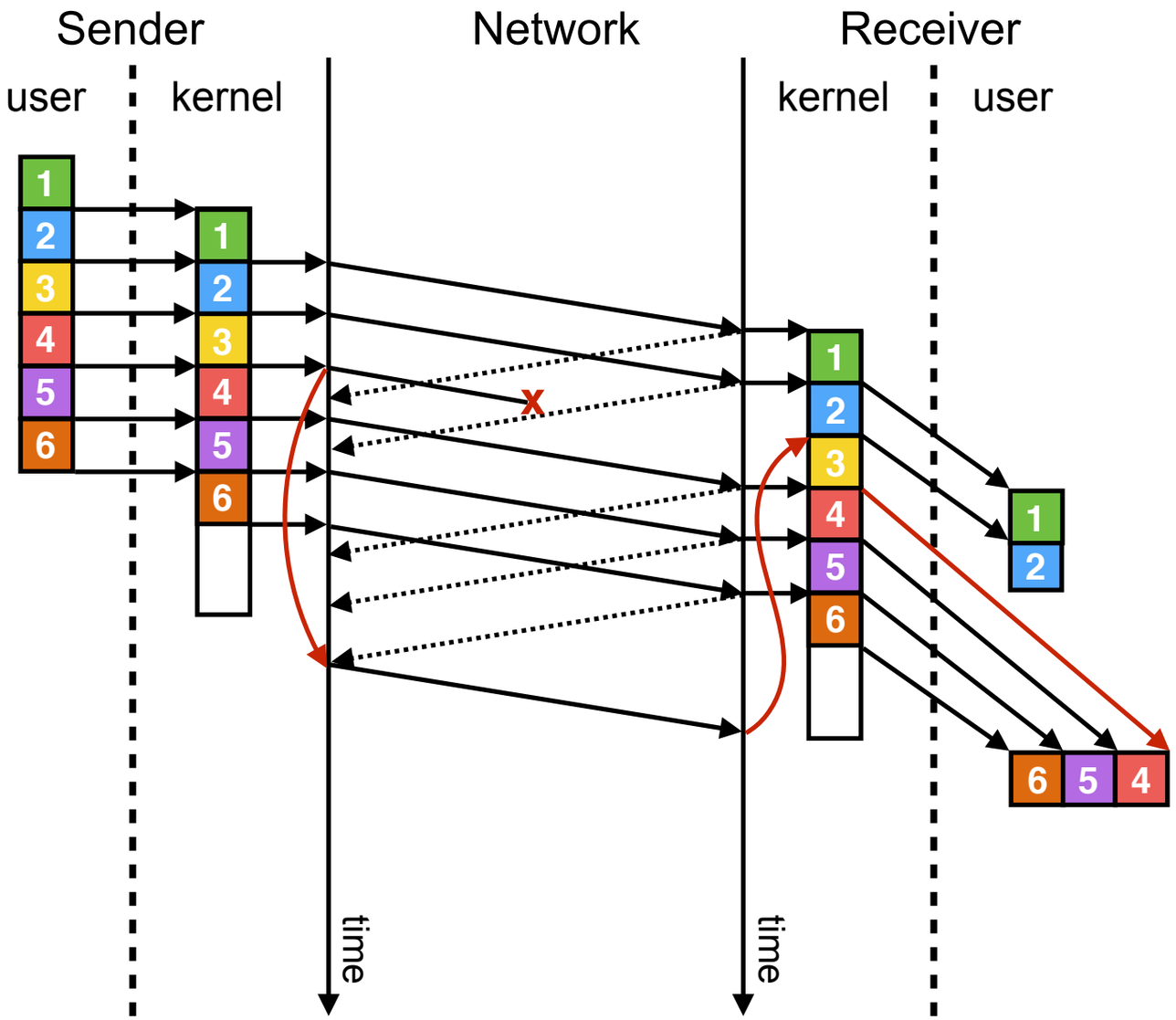
- Multi-second playout latency common, due to:
  - Chunk duration
  - Playout buffering
  - Encoding delays for live streaming

- **Chunk duration is a key** – must download complete chunk before starting playout
  - Download and decompress chunk *n*
  - Start playing chunk *n* and while playing download chunk *n+1*
  - No packets lost → latency equals chunk duration



Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Sources of Latency: TCP



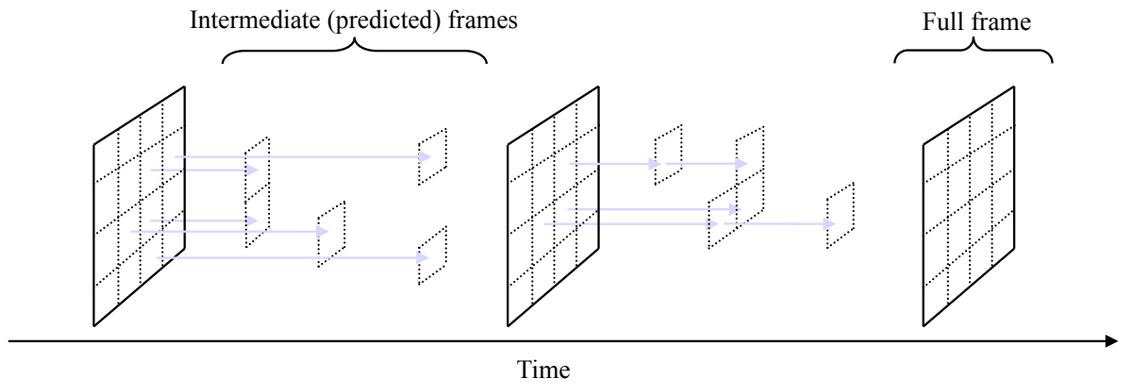
$$T_{rexmit} = 4 \times T_{framing} + T_{rtt}$$

- If a packet is lost, TCP retransmits after triple duplicate ACK
  - This takes time, and delays download of the chunk:
    - 4× packet transmission latency (the lost packet, plus three following that generate duplicate ACKs)
    - 1× network RTT to retransmit packet
  - Each packet loss and retransmission will increase chunk download time – and causes TCP to reduce its window
- Both add latency if using TCP

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Sources of Latency: Chunks and Video Compression

- Each chunk of video is independently decodable
  - Can’t compress based on previous chunk, because previous chunk depends on the network capacity → unpredictable



- Each chunk therefore starts with an I-frame, followed by P-frames
  - I-frames are large: often 20× size of P-frames
  - If chunks are smaller duration → contain fewer P-frames compared to number of I-frames; video compression ratio goes down
- Trade-off between chunk size and compression overhead
  - Large chunks require less data, but add latency
  - Small chunks reduce latency, but need more data
  - Overheads become excessive for chunk duration ≲2 seconds

Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Future Directions for Streaming Video

- Streaming video over WebRTC?
  - Web browsers now all support WebRTC
  - Intended for interactive conferencing, but entirely usable for RTP-based video streaming to browser
  - Much **lower latency** than DASH
  - Will this encourage CDNs to support RTP?



- Streaming video over QUIC?
  - HTTP/3 is a drop-in replacement for existing HTTP stack, and delivers requests over QUIC
  - YouTube already delivers video over QUIC this way – expect to see much more deployment and special purpose QUIC extensions for video



Colin Perkins | https://csperkins.org/ | Copyright © 2020 University of Glasgow

Real-time and Interactive Applications

- Real-time traffic and its constraints
- Interactive applications → WebRTC
- Streaming applications → DASH
- Custom infrastructure for low latency; CDN and HTTP if latency less critical

BY ND