Introduction to the Transport and Services Area (TSV)

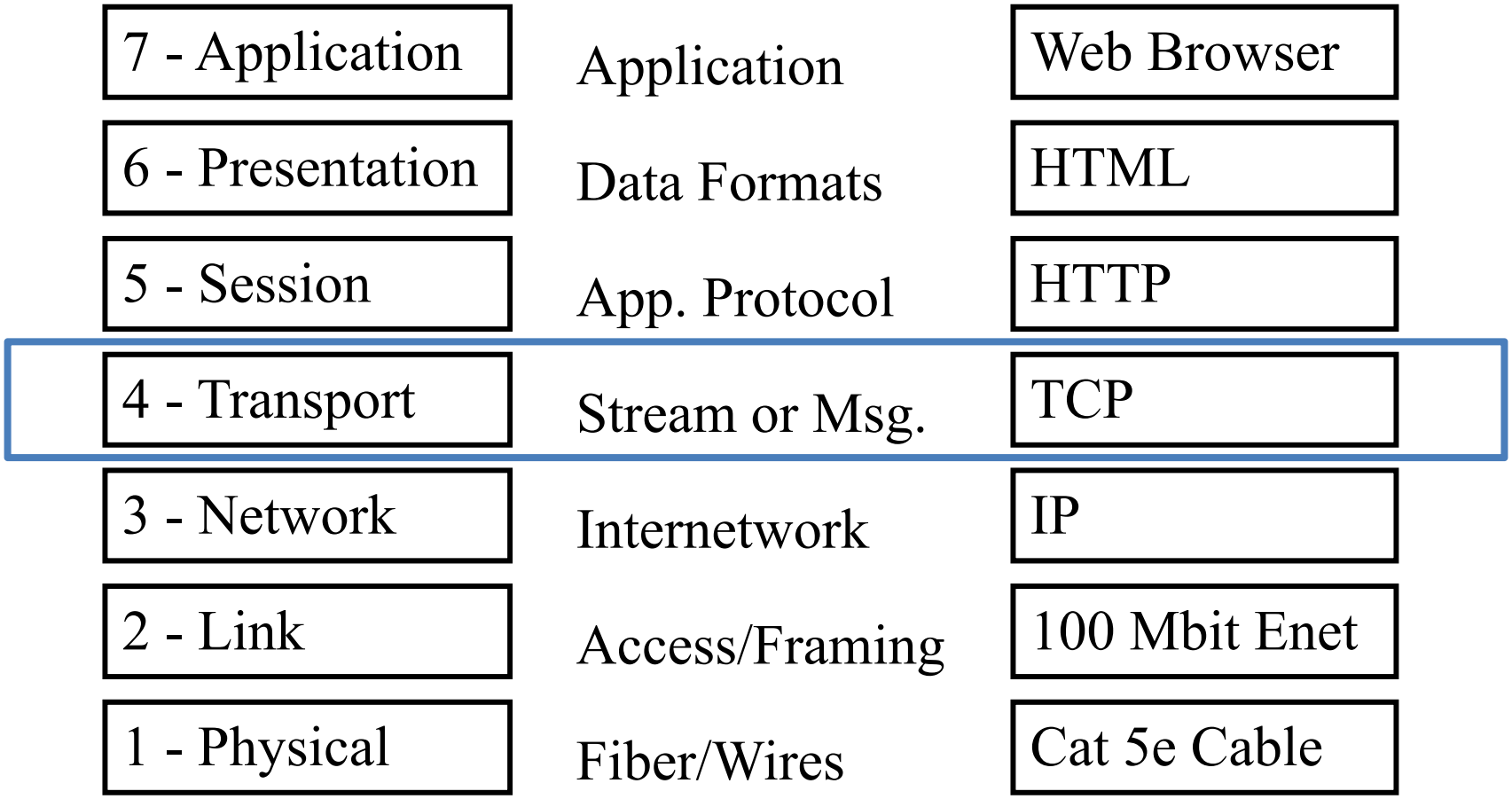
David L. Black, Dell EMC

Mirja Kühlewind, ETH Zurich

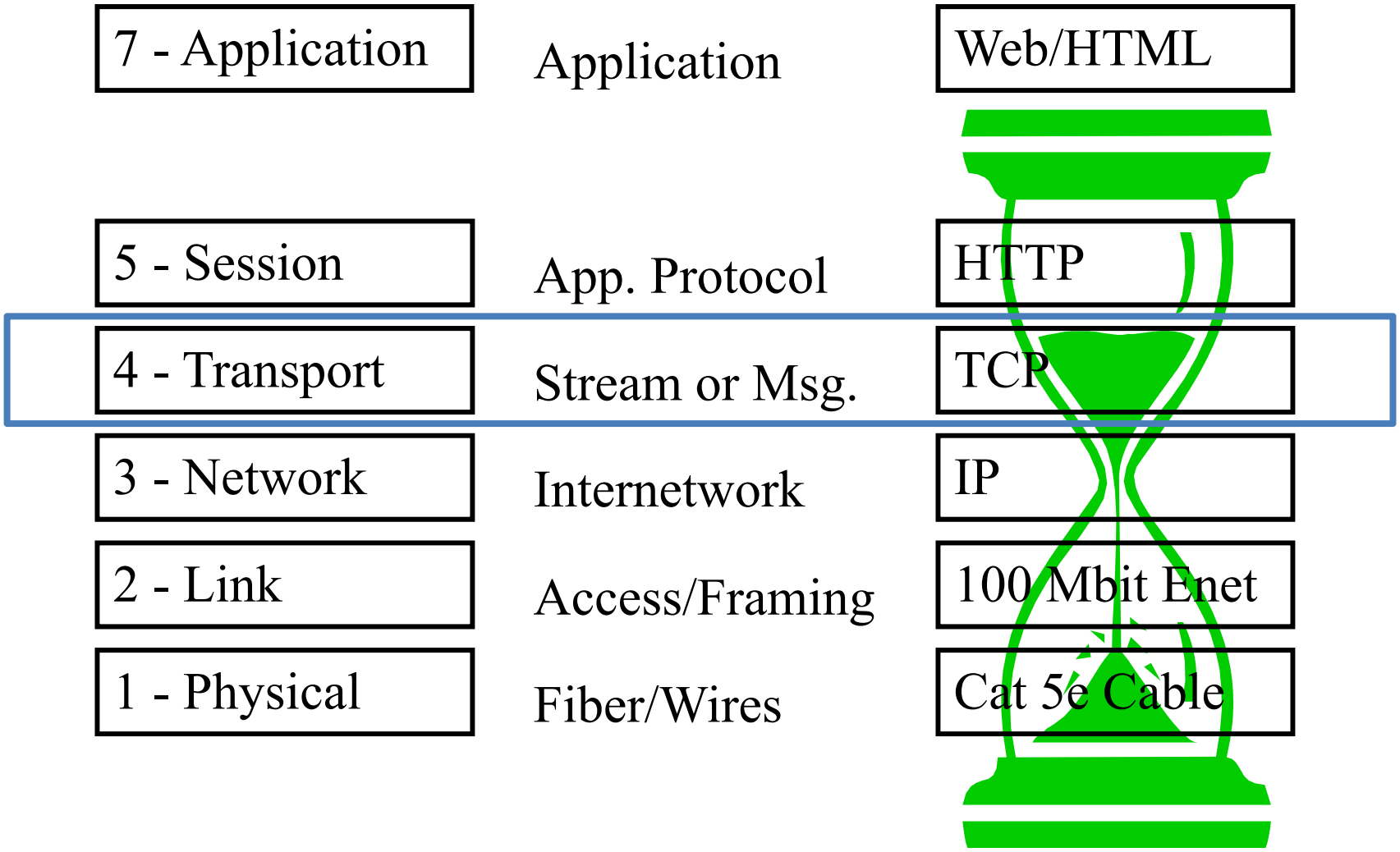
What is TSV (Transport) Area?

- "The transport and services [TSV] area…covers a range of technical topics related to data transport in the Internet."
- Protocol design and maintenance at Layer 4
  - TCP, UDP, SCTP and friends
- Congestion control and (active) queue management
  - Prevent congestive collapse of the Internet:
    - Been there, done that, not going back again ...
  - New concern: Buffer bloat
- Quality of Service and related signaling protocols
  - Examples: Differentiated Services [Diffserv] and RSVP
- Some TSV activities aren't Layer 4 specific (e.g., storage)
  - Located in in TSV for historical reasons

IP Network Layers



IP Network Layers – In Practice



In the beginning…

… there was TCP (well, sort of)

- Transport: One of the oldest IETF Areas
  - Transport protocols (layer 4): key Internet elements
    - TCP, UDP … then later SCTP, DCCP, … and now QUIC
- Transport: Adapt technology to the Internet
  - Making things work over “unreliable” packets
    - At large scale with congestion control
  - Examples: Storage, pseudowires, multimedia

Multimedia and RAI

- Ancient conventional wisdom: Can't obtain reliable service from unreliable packets
  - Disproved: RTP, audio/video codecs (early 1990s)
  - Example: The Rolling Stones on MBONE (1994)
- Broadened to related work
  - IP telephony (motivation for SCTP and SIP)
- Expanded to become separate RAI Area
  - RAI = Real-time Applications and Infrastructure

THE TSV (TRANSPORT) AREA TODAY

Transport Area Scope

- “Core” transport protocols: TCP, SCTP, etc.
- QUIC: New Transport protocol with security
- Congestion Control & Queue Management
- NAT Traversal
- Quality of Service and Signaling
- Storage Networking
- Other topics, e.g., delay tolerant networking, performance metrics for measurement

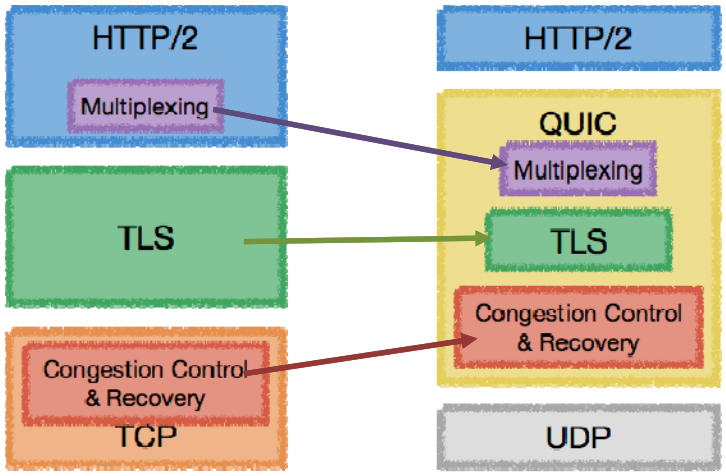
“Core” transport protocols

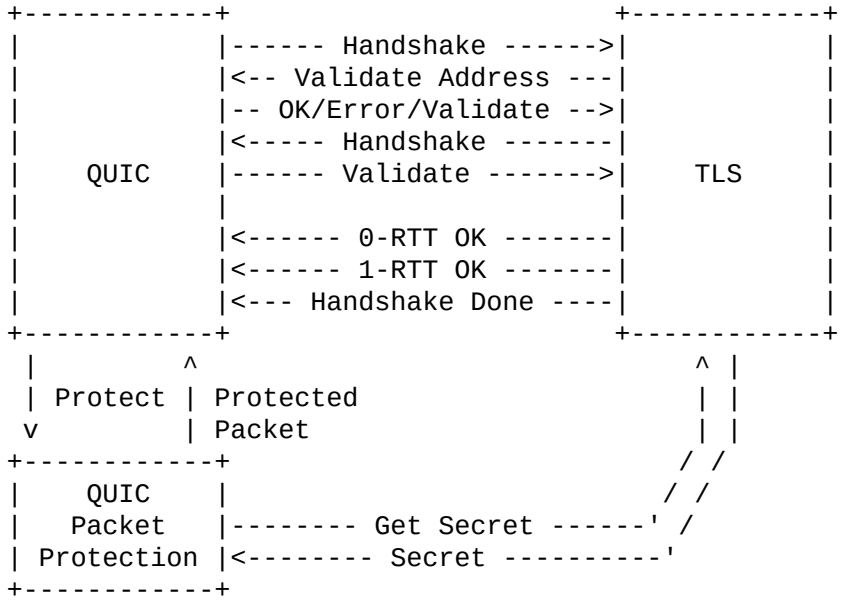
- Transmission Control Protocol (TCP)
  - Connection-oriented, fully reliable stream
- User Datagram Protocol (UDP)
  - Connectionless, ~~unreliable~~ best-effort
  - UDP-Lite adds corruption tolerance
- Datagram Congestion Control Protocol (DCCP)
  - Connectionless, best-effort, congestion-controlled
- Stream Control Transmission Protocol (SCTP)
  - Connection-oriented, multihomed, multistreamed, datagram-preserving, selectably reliable.
- These living protocols require ongoing maintenance
  - TCPM (TCP Maintenance), TSVWG (Transport Area)

QUIC

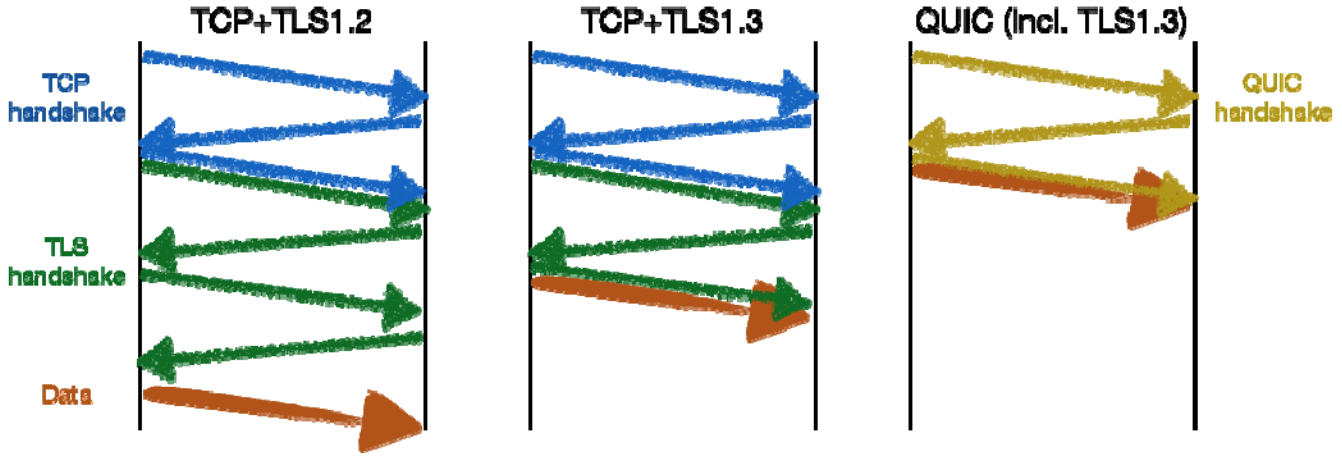
- Low-latency, UDP-encapsulated, encrypted, versioned, multi-streaming, general-purpose connection-oriented transport protocol.
- Developed at Google, standardization within IETF since July 2016 (QUIC Working Group)
- Tight integration with TLS 1.3 for security
- Initial focus: TCP+TLS replacement for HTTP v2.
- Features planned for future versions include partial reliability and multipath.

QUIC in the Protocol Stack





Low-Latency Session Establishment with QUIC



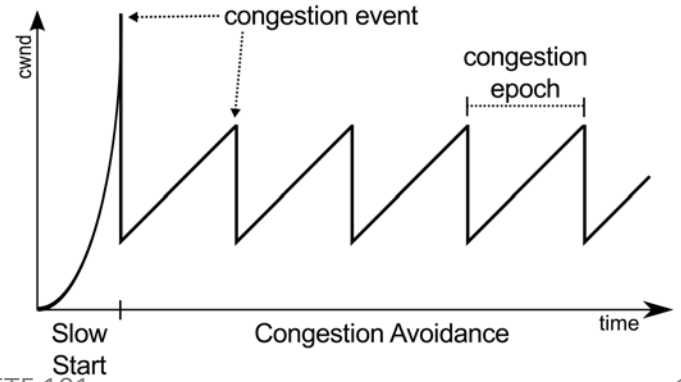
- Transport and cryptographic handshakes occur simultaneously, reducing initial delay to 1 RTT.
- 0 RTT handshake to previously contacted server allows client to send up to one flight of data before handshake completes.

Transport Services and Interfaces

- How to support transport innovation (and deployment of existing diversity) in the present Internet?
- Application-facing approach:
  Transport Services (TAPS) WG
- Common application interface to multiple transport protocols
  - Transport selected based on intersection of requirements defined in terms of services each protocol provides
  - Dynamic measurement of the path to determine which protocols and options will work
  - Based on a survey of available transport services (RFC8095)
- Working on an abstract interface to allow applications to take advantage of transport selection

Congestion Control in the Internet

- Aggressive retransmission by reliable transport protocols can lead to *congestive collapse*
  - traffic becomes dominated by retransmission
  - Settles into stable near-zero goodput state
- Happened repeatedly in 1986-1988
- Result: development and deployment of TCP congestion control
  - Congestion window limits rate, split into slow start and congestion avoidance phases

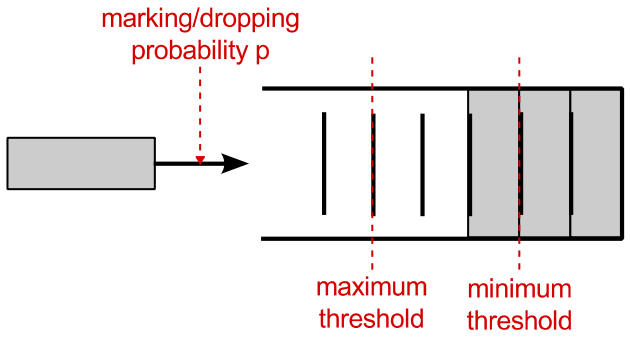


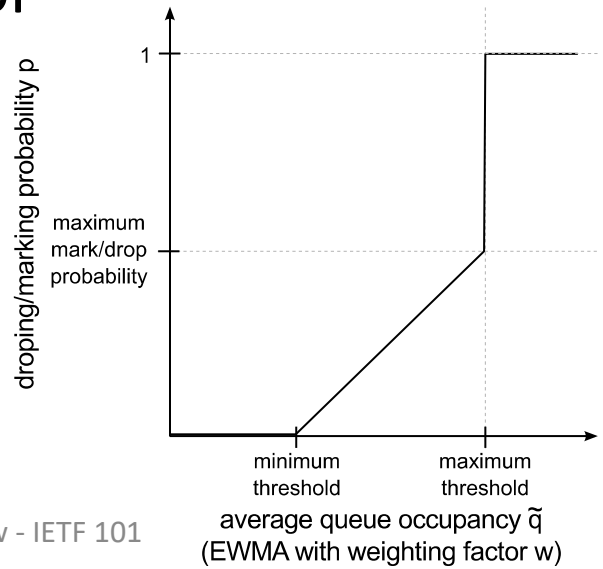
Congestion Control in the Internet

- Common algorithms (NewReno, CUBIC, etc.) use loss as a congestion signal…
  - and therefore underperform on lossy links that aren't congested
- …induce congestion to determine available bandwidth…
  - so interact poorly with buffers sized to prevent loss (buffer bloat)
- …and always (eventually) use as much bandwidth as they can
  - so one must be careful when designing protocols that share bandwidth with TCP.
- Current Activity: Use delay as another congestion signal.
  - BBR (Bottleneck Bandwidth and Round-trip propagation time)
    - Obtain congestion signal from bandwidth and round-trip time
    - Discussed in Internet Congestion Control RG (ICCRG)
  - RMCAT (RTP Media Congestion Avoidance Techniques) Working Group
    - Shared bottleneck detection, coordinate congestion control across flows that share it
    - NADA (Network Assisted Dynamic Adaptation): Add delay change as congestion signal

Active Queue Management (AQM)

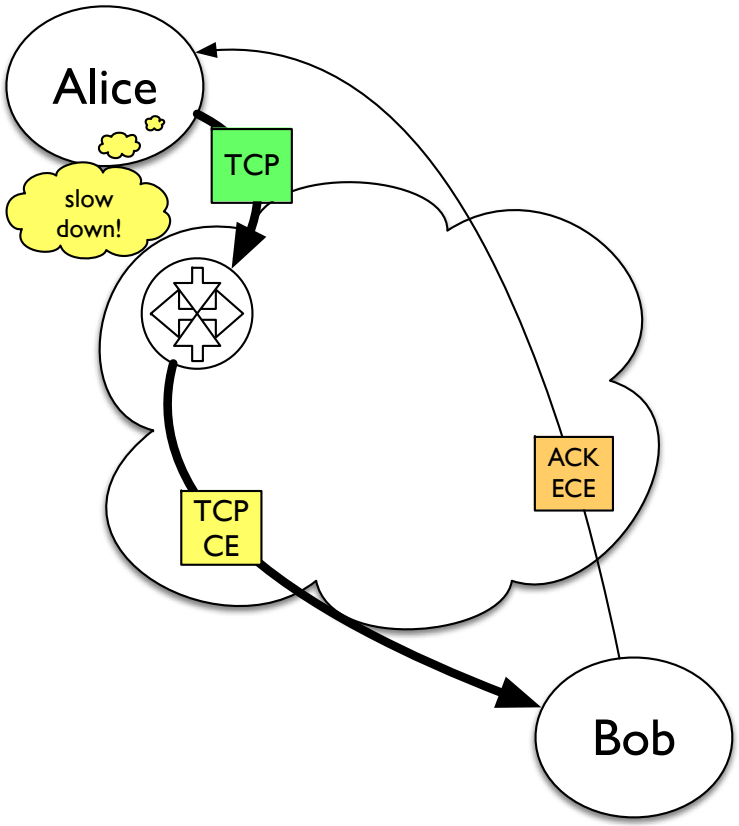
- Loss signals congestion because routers drop packets when their buffers are full.
- Dropping packets *before* the buffers fill can improve overall performance (e.g., RED algorithm)
  - Improving when to drop and when not to: Research topic
- Active Queue Management (AQM) schemes augment end-to-end congestion control





AQM and Explicit Congestion Notification (ECN)

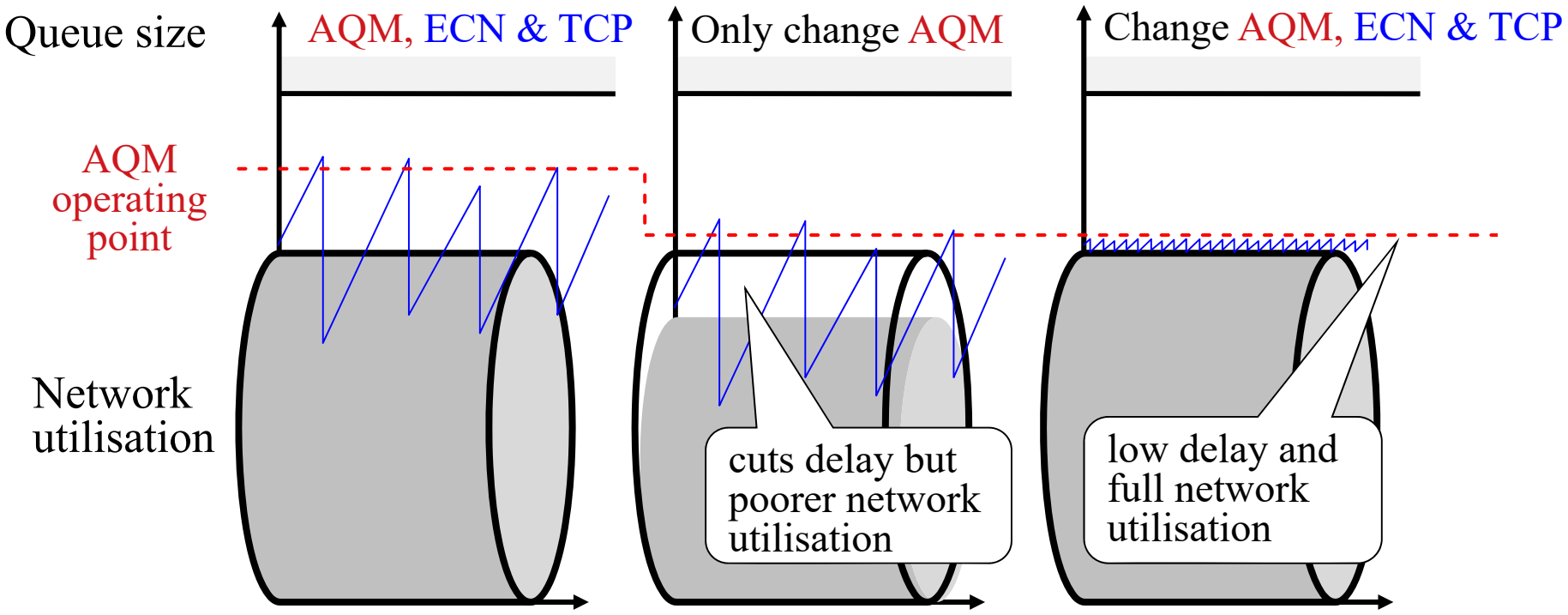
- AQM still drops packets to signal congestion
  - Wouldn't it be nice if we didn't have to do that?
- ECN (RFC 3168) marks IP header and reflects congestion signal in TCP (& other transport protocols), without loss when possible
  - Worst case: can still drop
- Current work to improve behavior when TCP is encapsulated by other protocols.



AQM, ECN and the Pursuit of Shorter Queues

- Reminder: TCP increases transmission until it sees loss or ECN congestion mark
  - Increases latency by filling queues in the network.
  - AQM helps, but limited by existing TCP reaction to loss or ECN congestion mark
- More radical approach: Change AQM, ECN and TCP together
  - AQM: Start congestion-marking packets when queue is much shorter.
  - ECN: Receiver combines marks from multiple packets into richer congestion feedback.
  - TCP: Sender uses that richer congestion feedback for finer-grain congestion control.
- This works! Much shorter queues, e.g., Data Center TCP (RFC 8257)
  - But … can't share queues with ordinary (TCP) traffic due to AQM marking change.
- So, don't share queues with ordinary traffic: Dual Queue Coupled AQM
  - Split incoming low-latency traffic into separate queue from other traffic
  - Couple AQM behavior across queue pair to induce similar throughput for similar flows
- Result = L4S, Low Latency Low Loss Scalable throughput service
  - Active work in Transport Area Working Group (TSVWG)
  - Work in progress: Evolution of DCTCP to use L4S effectively

Achieving Shorter Queues (DCTCP, L4S)



Transport Area Scope

- “Core” transport protocols: TCP, SCTP, etc.
- QUIC: New Transport protocol with security
- Congestion Control & Queue Management
- NAT Traversal
- Quality of Service and Signaling
- Storage Networking
- Other topics, e.g., delay tolerant networking, performance metrics for measurement

NAT traversal
(NAT = Network Address Translation)

- At first: protocol-specific (e.g., for IKE [ipsec])
- Now: Protocol-independent (STUN/TURN/ICE)
  - Session pinhole punching and maintenance
  - STUN: routable address discovery
    - STUN = Session Traversal Utilities for NATs (RFC 7064)
  - TURN: relay when necessary
    - TURN = Traversal Using Relays around NATs (RFC 7065)
  - ICE: Framework for STUN/TURN usage (e.g., in SIP)
    - ICE = Internet Connectivity Establishment (RFC 5245)
- TURN Revised and Modernized (TRAM) WG
  - Security improvements (e.g., DTLS, authentication
  - TURN: Add server auto-discovery, IPv6 support

Quality of Service (QoS)

- General QoS frameworks: Transport Area
  - Integrated Services (IntServ): Per-flow (poor scaling)
  - Differentiated Services (DiffServ): Traffic class in IP header
    - Limited number of traffic classes
  - Additional framework variants
    - Example: Pre-Congestion Notification (PCN) for real-time non-congestion-responsive flows
- QoS Signaling: RSVP – Resource reSerVation Protocol
- Most QoS work has been completed
  - Current activity: limited development/maintenance
  - DiffServ and RSVP: handled by TSVWG WG

Storage Networking

- Block (SAN) storage: iSCSI and FC/IP
  - In cooperation w/storage standards bodies
    - T10 [SCSI] and T11 [FC=Fibre Channel], respectively
    - [T10 and T11 are historical acronyms]
  - Storage Maintenance (STORM) WG
- File (NAS) storage: NFS (Network File System)
  - NFSv3, then NFSv4
  - Currently NFSv4.2 (close to complete)
  - CIFS and SMB (for Windows): Not IETF protocols
- RDMA protocol suite: iWARP (RDDP WG – concluded)
  - RDMA = Remote Direct Memory Access
  - Often used with storage protocols

Delay-Tolerant Networking (DTN)

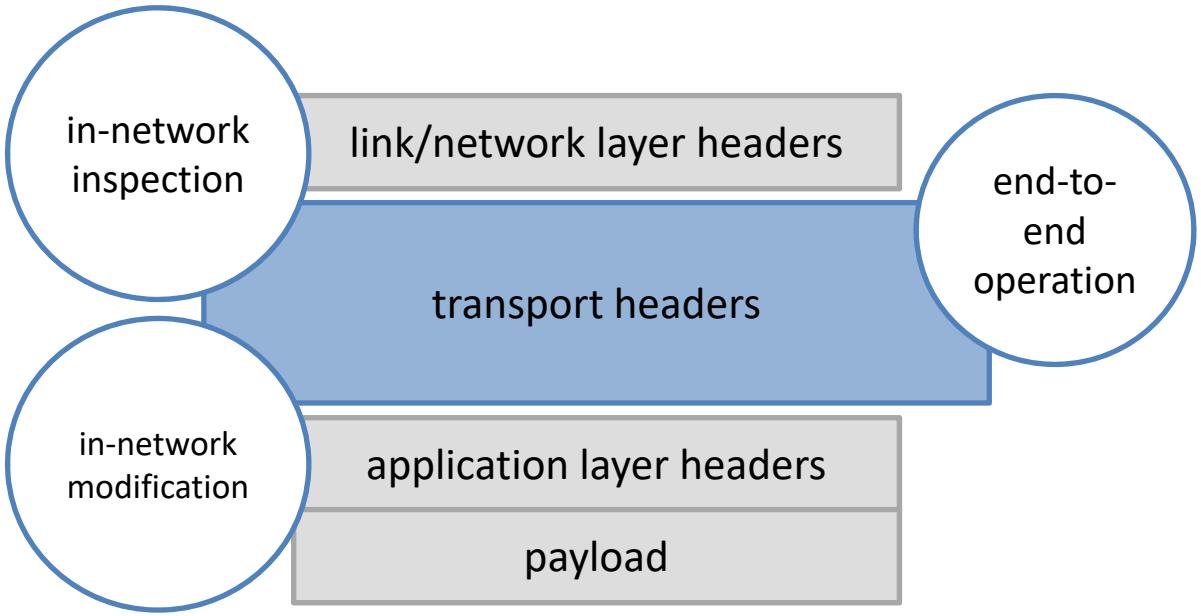
- How to extend the Internet to very-high-delay, low-connectivity environments?
  - disaster recovery, UAV networks, underwater acoustic networks, interplanetary networks, etc.
- Requires new protocols at the transport layer as well as delay-tolerant applications.
- DTNRG defined a set of protocols (Bundle, LTP)
  - LTP = Licklider Transport Protocol (RFC 5326)
- DTN WG updates: implementation experience, support new use cases, on standards track.

Network Performance Measurement

- IPPM WG: Can't manage what you can't measure
  - IPPM = IP Performance Measurement
  - Standard metrics for Internet transport performance
  - Methods to measure metrics and analyze results
- Current (new) focus: hybrid measurement of core and access networks
  - Hybrid measurement = passive observation of traffic generated or modified explicitly to be measurable
  - IOAM (in-band or in-situ Ops/Admin/Mgt) = common data model for adding hybrid-measurable signals to various encapsulation protocols.
- Finishing work on a core registry of standard measurements for common metrics/use cases.

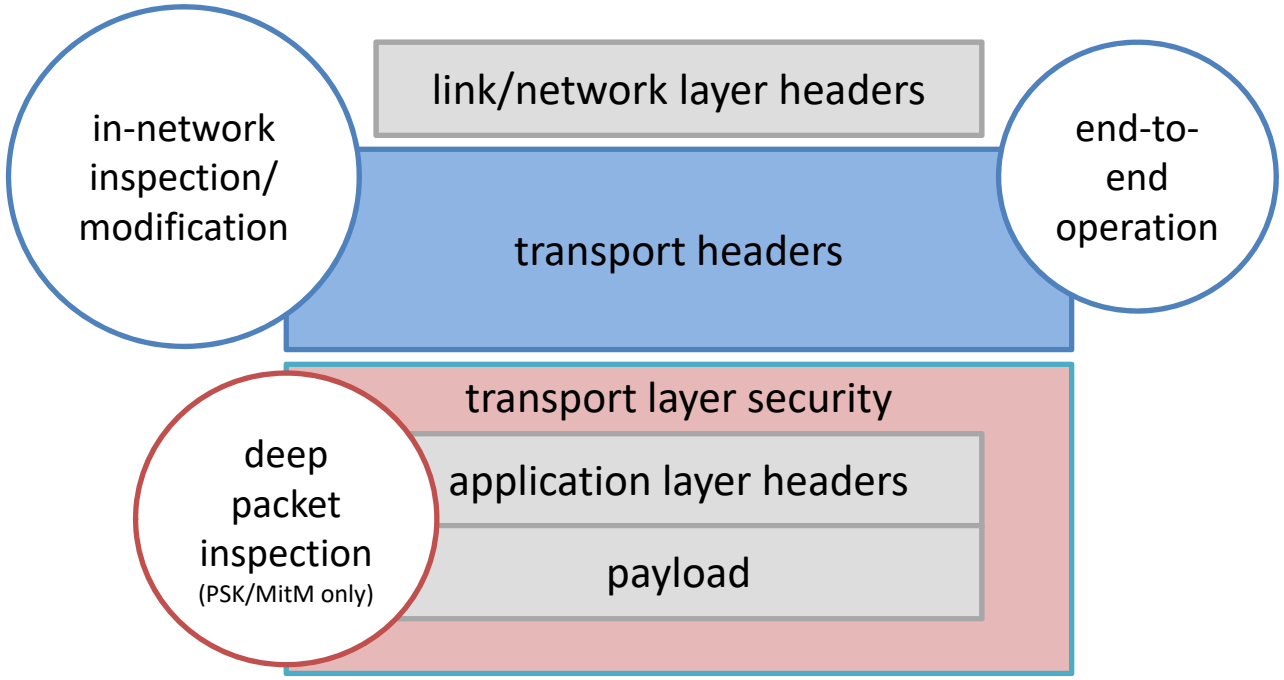
ENCRYPTED TRANSPORT PROTOCOL HEADERS

Protocol Header Separation (Transport Protocol Design, 1980s-style)



- In the beginning, all headers and payload at all layers were visible and modifiable at any point along the path.

Protocol Header Separation
(Transport Protocol Design, now with security!)

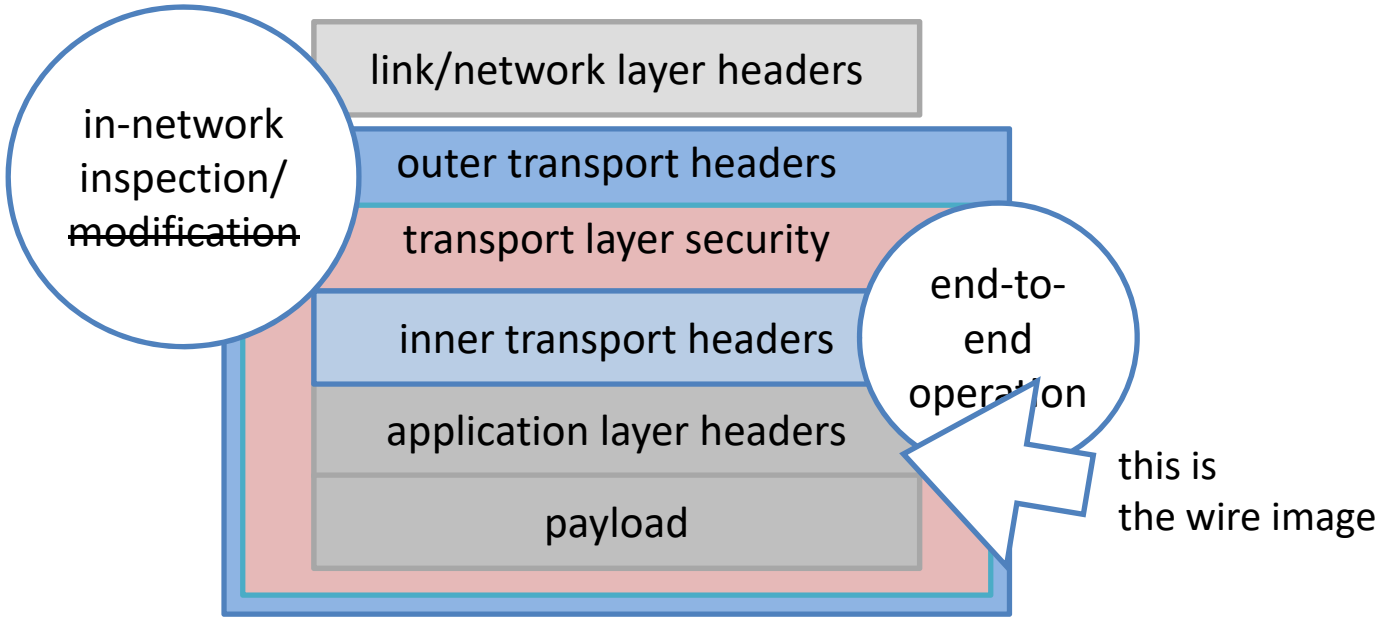


- The widespread deployment of transport-layer security prevents inspection and modification of the application layer, but the transport layer remains unprotected.

Why encrypt protocol headers?

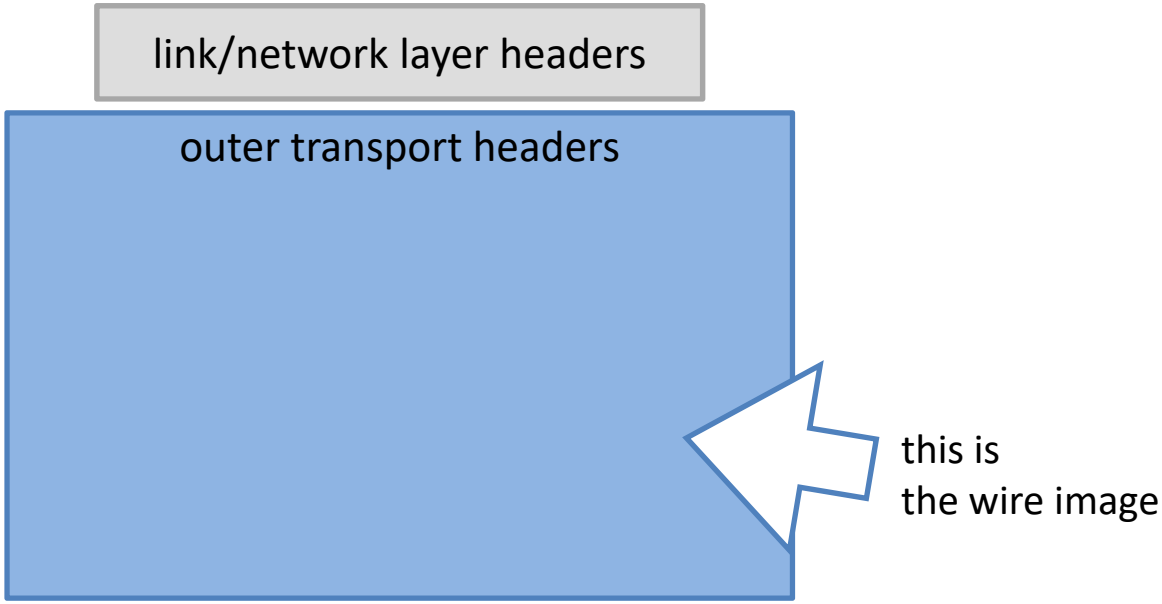
- Most of our inability to deploy TCP enhancements or new transport protocols comes from “off-label” use of the headers by middleboxes.
- Transport-layer metadata can be used to correlate flows and/or fingerprint endpoints.
- General approach (followed by QUIC): encapsulate in UDP for ports/NAT, encrypt above.

Protocol Header Separation (Encrypted Transport Protocol Headers)



- Encryption allows us to separate transport-internal information from on-path signaling
  - E.g. state signaling to firewalls, explicit measurability
- draft-trammell-wire-image and draft-hardie-path-signals discuss the possibilities here.
- Care is needed: the path will use everything it can see.

Protocol Header Separation (Encrypted Transport Protocol Headers)



- Encryption allows us to separate transport-internal information from on-path signaling
  - E.g. state signaling to firewalls, explicit measurability
- draft-trammell-wire-image and draft-hardie-path-signals discuss the possibilities here.
- Care is needed: the path will use everything it can see.

What happens when protocol headers are encrypted?

- Lots of stuff in the network understands TCP
  - Firewalls, (some) DDoS rejection
  - Passive performance measurement
  - On-path debugging via wireshark

- When this stuff can't see the headers, it breaks.

Encrypted Protocol Headers: Related Drafts

- Network impacts of pervasive encryption
  - draft-mm-wg-effect-encrypt
- Middleboxes considered useful
  - draft-dolson-transport-middlebox
- Transport protocol considerations & consequences of encrypted headers
  - draft-fairhurst-tsvwg-transport-encrypt

TSV: MEETINGS IN LONDON

London: Transport Area Meetings

- Area-Wide: TSVAREA, TSVWG
- TCP-related: TCPM, MPTCP
- Congestion Control: ICCRG, RMCAT
- New protocols/deployment: QUIC, TAPS
- Everything Else: ALTO, DTN, IPPM

(*Acronym Expansions on Subsequent Slides)

TSV Area Meeting (TSVAREA)

- Venue for discussion of topics of general interest to the entire transport area.
- Primary London topic: TCP encapsulation
  - Enables other protocols to traverse middleboxes, e.g., NATs
  - First example: IKE (IPsec key exchange and setup)
- **Monday 17:40** in **Sandringham**

Transport Area Working Group (TSVWG)

- Catch-all WG for work that needs to be done
  - But that can't sustain its own IETF WG
- SCTP maintenance/extension - NAT traversal, errata
- Diffserv (QoS) – better support for lower-effort (scavenger) traffic
- Forward Error Correction (FEC) update
- UDP - Protocol level path MTU detection (PLPMTUD), UDP options
- ECN - encapsulation behavior, interaction with link-layer congestion detection and congestion isolation
- L4S – Low Latency Low Loss service based on AQM & ECN
- Encryption of transport protocol headers

- **Thursday 15:50 – 19:10,** in **Balmoral** (back-to-back)

TCP Maintenance and Minor Extensions (TCPM)

- “TCP is currently the Internet's predominant transport protocol. TCPM is the working group within the IETF that handles small TCP changes, i.e., minor extensions to TCP algorithms and protocol mechanisms.”
  - Maintenance issues (bugfixes)
  - Moving TCP along the standards track
- Current discussion:
  - ECN alternative backoff (ABE) and accuracy (AccECN)
  - Time-based loss detection (RACK)
- **Monday 09:30** in **Buckingham**

Multipath TCP (MPTCP)

- “The Multipath TCP (MPTCP) working group develops mechanisms that add the capability of simultaneously using multiple paths to a regular TCP session.”
- Primary Focus: Revised standards-track version of Multipath TCP, based on experience
- **Thursday 13:30** in **Viscount**

Internet Congestion Control Research Group (ICCRG) (in IRTF)

- Goal: “move towards consensus on which [new congestion control] technologies are viable long-term solutions for the Internet congestion control architecture, and what an appropriate cost/benefit tradeoff is.”
- Expert congestion control advice to Transport Area
  - Analogous to CFRG (Crypto Forum Research Group) expert crypto advice to Security Area
- London agenda: Congestion-related topics
  - Includes BBR congestion control
  - Generally ahead of work in IETF Working Groups
- **Friday 09:30** in **Buckingham**

RTP Media Congestion Avoidance Techniques (RMCAT)

- “Congestion control algorithms for interactive real time media may be quite different from TCP CC: for example, some applications can be more tolerant to loss than delay and jitter. The set of requirements for such an algorithm includes, but is not limited to:
  - Low delay and low jitter
  - Reasonable bandwidth sharing with RMCAT, other media protocols, TCP
  - Effective use of signals like packet loss and ECN markings to adapt to congestion”
- Current work:
  - Evaluation criteria for RTP congestion avoidance algorithms
  - RTP congestion information feedback via RTCP
- **Wednesday 13:30** in **Palace C**

London: Transport Area Meetings

- Area-Wide: TSVAREA, TSVWG
- TCP-related: TCPM, MPTCP
- Congestion Control: ICCRG, RMCAT
- New protocols/deployment: QUIC, TAPS
- Everything Else: ALTO, DTN, IPPM

(*Acronym Expansions on Subsequent Slides)

QUIC

- "The QUIC working group will provide a standards-track specification for a UDP-based, stream-multiplexing, encrypted transport protocol"
- Current topics:
  - Security topics
  - Invariants (things that will not change in future versions of QUIC, e.g., Connection ID)
  - ECN support
  - SPIN bit (enables observer to measure round-trip time)
- **Monday 13:30** and **Thursday 09:30 in Sandringham**

Transport Services (TAPS)

- “The goal of the TAPS working group is to help application and network stack programmers by describing an (abstract) interface for applications to make use of Transport Services.”
- Current work on:
  - Transport service architecture and interface structure
  - Transport security protocols
- **Wednesday 09:30** in **Park Suite**

Application Layer Traffic Optimization (ALTO)

- "ALTO has developed an HTTP-based protocol to allow hosts to benefit from the network infrastructure by having access to a pair of maps: a topology map and a cost map… ALTO is now being considered as a solution for problems outside the P2P domain, such as in datacenter networks and in content distribution networks (CDN) where exposing abstract topologies helps applications."
- Initial protocol work completed
- London topics: Protocol extensions
  - Generalization of endpoints to domains
  - Using path information with endpoint-based maps
  - Additional extension proposals
- **Monday 15:50** in **Palace C**

Delay Tolerant Networks (DTN)

- “The Delay/Disruption Tolerant Network Working Group (DTN WG) specifies mechanisms for data communications in the presence of long delays and/or intermittent connectivity.”
- Store-and-forward Bundle Protocol deals with long delays and/or intermittent connectivity.
- London Topics:
  - TCP details (when forwarding over TCP)
  - Security
  - Management
- **Friday 09:30** in **Palace C**

IP Performance Metrics (IPPM)

- “The IP Performance Metrics (IPPM) Working Group develops and maintains standard metrics that can be applied to the quality, performance, and reliability of Internet data delivery services and applications running over transport layer protocols (e.g. TCP, UDP) over IP”
- Current work: Core registry of standard measurements for common metrics/use cases.
- **Tuesday 15:50** in **Richmond/Chelsea/Tower**

TSV working groups that are not meeting in London

- NFSV4 (Network File System v4) WG
  - Distributed File System protocol
- TCPINC (TCP INCreased Security) WG
  - Unauthenticated TCP security
  - For authenticated security, use TLS.
- TRAM (TURN Revised and Modernized)
  - NAT Traversal updates & improvements

Acknowledgments

- Olivier Bonaventure
- Scott Bradner
- Bob Briscoe
- Brian Carpenter
- Vijay Gurbani
- Jana Iyengar
- Allison Mankin
- Martin Stiemerling
- Brian Trammell

Brief Survey

- We want your feedback – please complete a very brief (5 questions) survey online at:

https://www.surveymonkey.com/r/101Transport

Thank you!