Introduction to the Transport and Services Area (TSV)

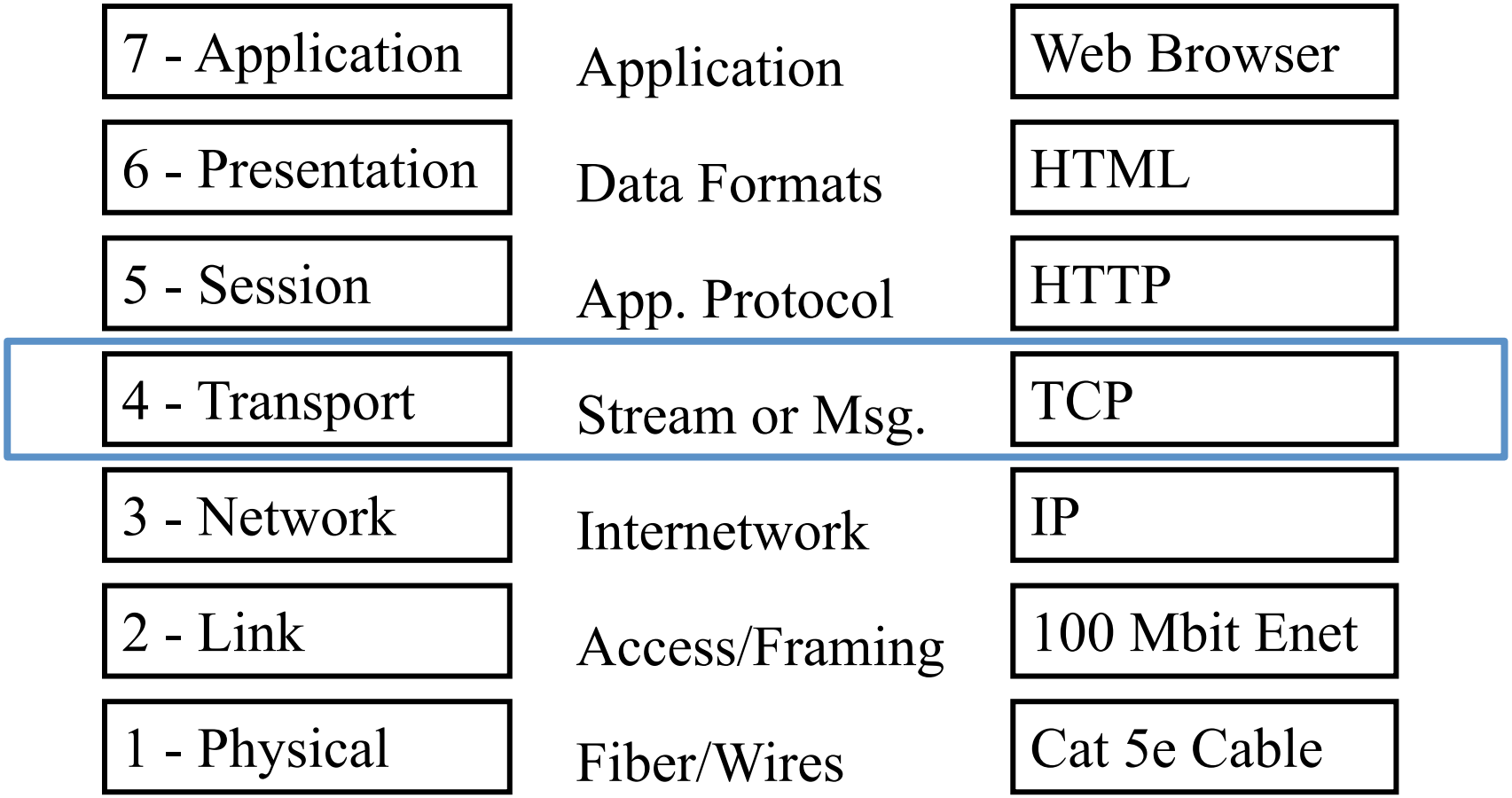
David L. Black, EMC

Brian Trammell, ETH Zurich

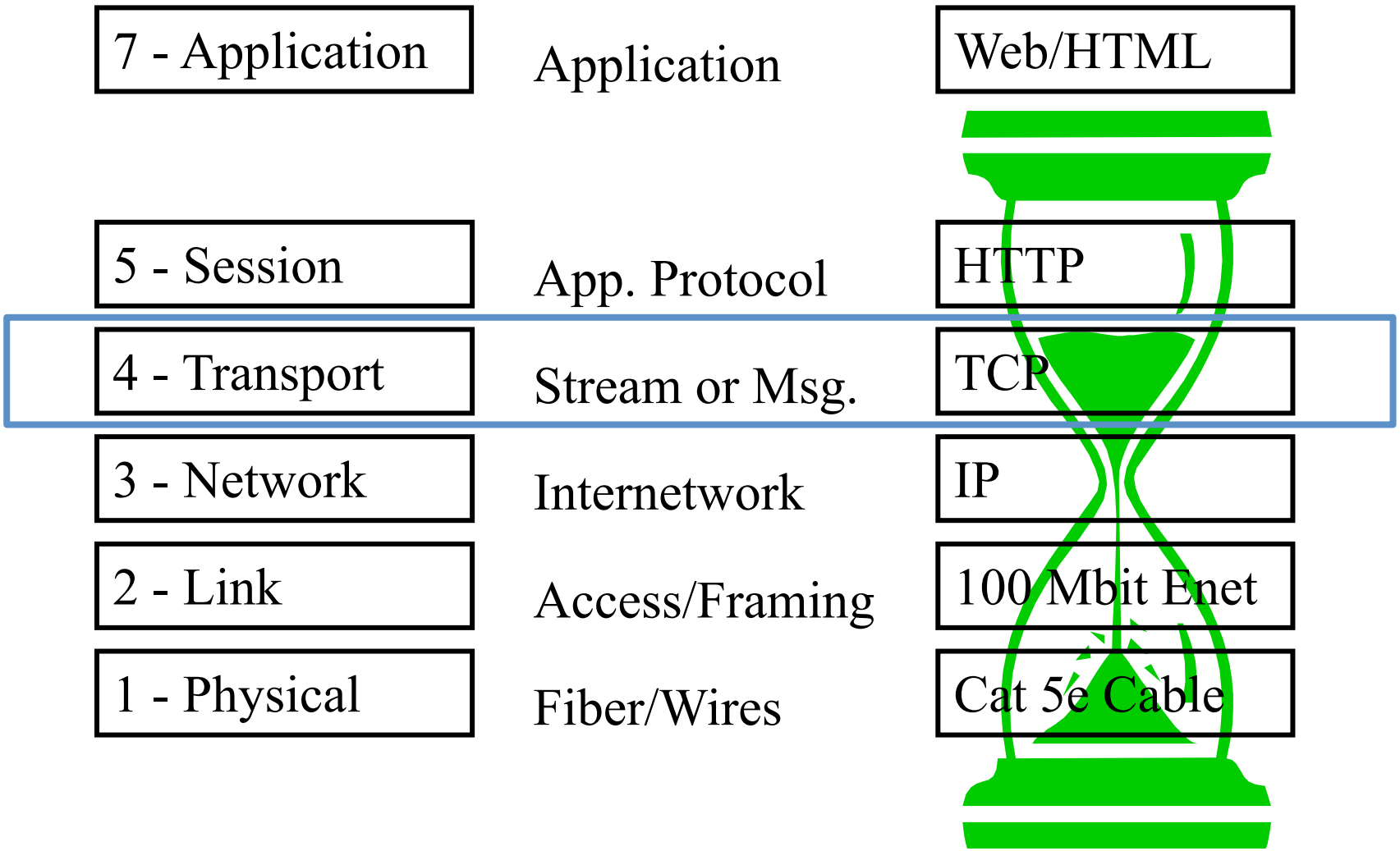
What is TSV (Transport) Area?

- "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, …
- 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

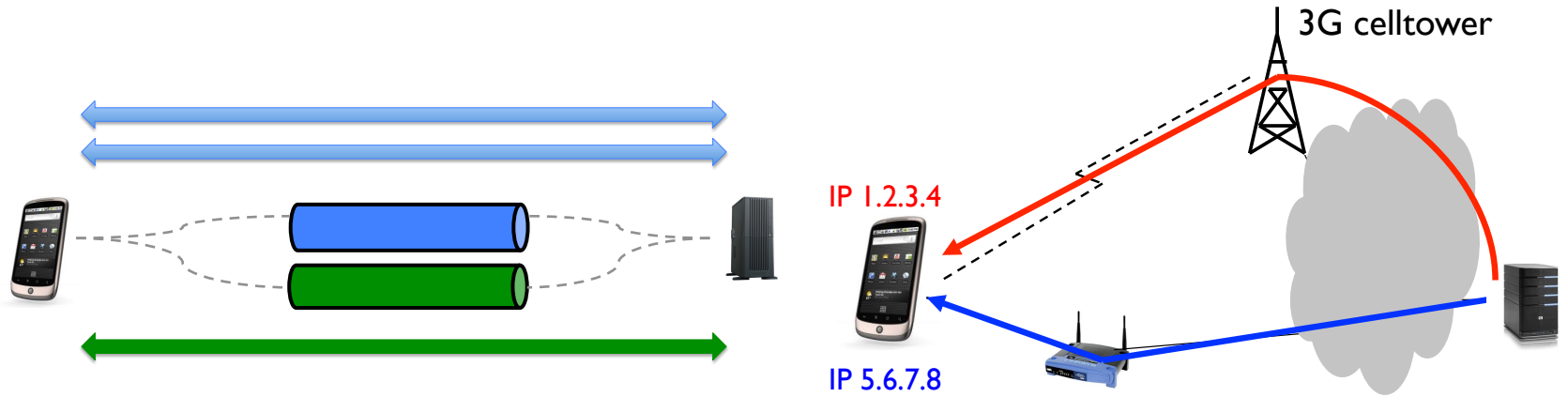
- “Core” transport protocols: TCP, SCTP, etc.
- Congestion Control & Queue Management
- NAT Traversal, UDP Encapsulation
- Quality of Service and Signaling
- Storage Networking
- Other Topics
  - Delay tolerant networking
  - Application Level Transport Optimization
  - TCP Incremental Security

“Core” transport protocols

- Transmission Control Protocol (TCP)
  - Connection-oriented, fully reliable stream
- User Datagram Protocol (UDP)
  - Connectionless, ~~unreliable~~ best-effort
  - “User-space raw sockets with port multiplexing”
  - 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
  - WGs: TCPM (TCP Maintenance), TSVWG (Transport Area)

Multipath TCP

- Bind TCP sessions across multiple interfaces
  - Higher reliability, bandwidth utilization
  - Like SCTP multihoming for TCP
- Multipath TCP (MPTCP) Working Group
  - Experimental protocol in RFC 6824
  - Current work on updates based on deployment experience
- Uses a TCP option for additional signaling in band

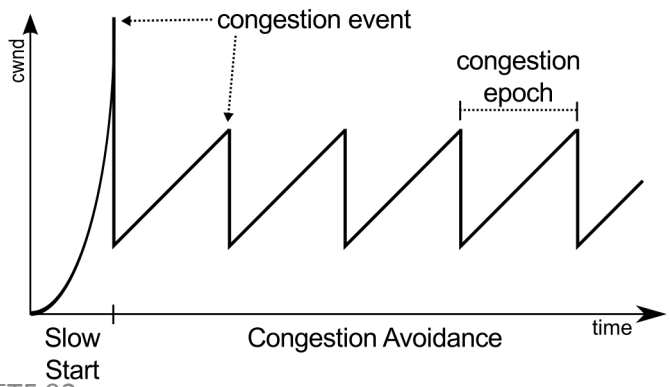


Transport Services and Interfaces

- How to support transport innovation (and deployment of existing diversity) in the present Internet?
- One 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

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

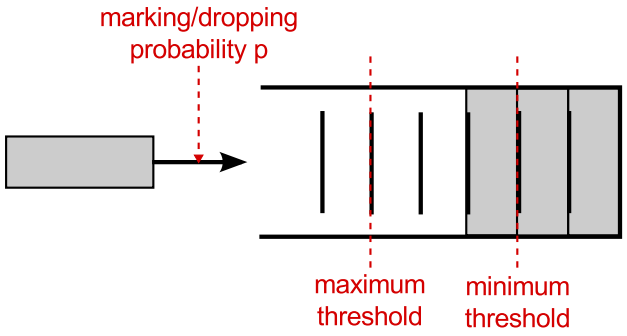


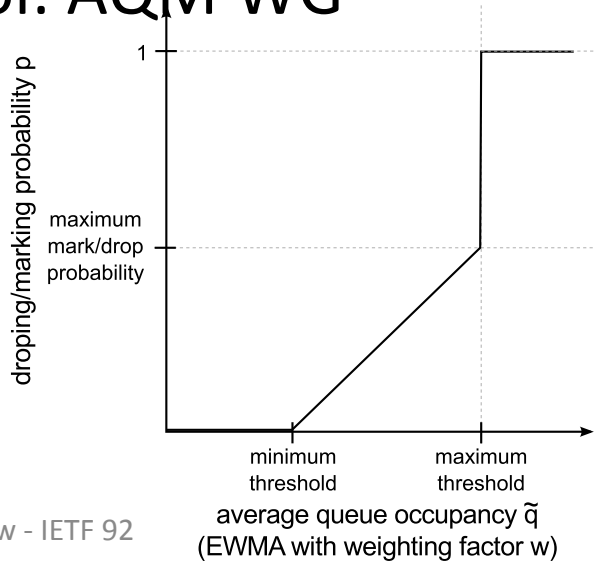
Congestion Control in the Internet

- Widely-deployed algorithms (NewReno, CUBIC, etc.) use loss as a congestion signal…
  - and therefore underperform on lossy links
- …must 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 will share the link with loss-based TCP traffic.
- Current research in Internet Congestion Control RG
- New algorithms for Web RTC (browser-based conferencing)
  - Area of interest: Use delay change as a congestion signal
  - RMCAT WG (RTP Media Congestion Avoidance Techniques)

Active Queue Management (AQM)

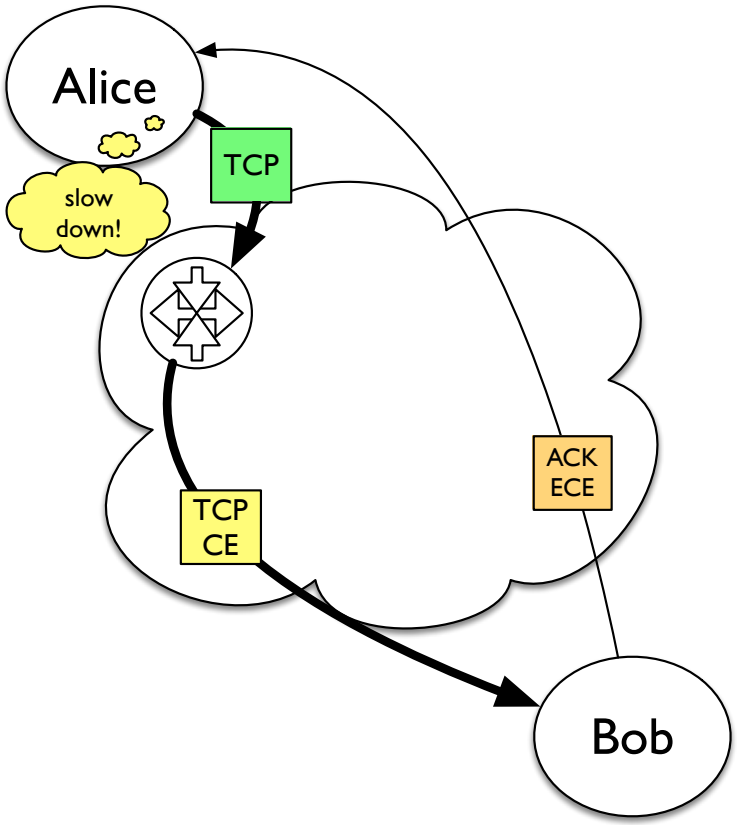
- Loss signals congestion because routers drop packets when their buffers are full.
- Dropping packets *before* the buffers fill can improve overall performance (e.g., RED algorithm)
  - Improving when to drop and when not to: Research topic
- Active Queue Management (AQM) schemes augment end-to-end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





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, without loss when possible
  - Worst case: can still drop
- Current work to increase deployment (currently more or less zero), improve signaling.



Transport Area Scope

- "Core" transport protocols: TCP, SCTP, etc.
- Congestion Control & Queue Management
- NAT Traversal, UDP Encapsulation
- Quality of Service and Signaling
- Storage Networking
- Other Topics
  - Delay tolerant networking
  - Application Level Transport Optimization
  - TCP Incremental Security

Network Address Translators (NATs) and Middleboxes (e.g., Firewalls)

- Uhm, well ... back in 2011 ...

What have we done so far?

What have we done so far?

- "NATs are evil. We won't care about them."
- "It will all change with IPv6." ← **Denial**
- "Don't design around middleboxes, that will only encourage them!" ← **Anger**
- "Alright, we'll specify how middleboxes *ought* to behave with different protocols. But they still have to behave." ← **Bargaining**
- "Why build a new transport?? It won't get deployed anyways." ← **Depression***

**Kübler-Ross model: Five stages of grief*

Slide credit: Jana Iyengar

NAT traversal

- 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

UDP Encapsulation

- Encap Motivations: NAT Traversal, multipath header hashes
  - UDP works, is simple: Datagrams with port multiplexing
- Congestion Control: UDP has none
  - Be careful, see RFC 5405 (being revised in TSVWG WG)
- UDP checksum for IPv6
  - IPv6: No header checksum (IPv4 has a header checksum)
  - UDP checksum protects IPv6 header (good)...
    - ... and entire UDP payload (expensive) ...
    - ... and is in header (breaks pipelines, e.g., hardware).
    - So designers want to zero out UDP checksum
  - Ok to zero out if one is **very** careful
    - See RFC 6935, RFC 6936 and draft-ietf-mpls-in-udp (RFC Editor Queue)
- QoS may operate on UDP flows, not encapsulated flows

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
- Recent activity: Differentiated Services for Web RTC

DiffServ and Web RTC (I)

- Web RTC = Web Real Time Conferencing
  - Audio, video, data between browsers
- NAT Traversal: Has to work
  - Every home "router" contains a NAT.
  - Goal: Minimize pinhole punching and maintenance
- Pinhole needed for each local port used
  - So, run different types of traffic on same port
  - UDP encapsulation preferred
- But what about QoS per traffic type?

DiffServ and Web RTC (II)

- Q: When is it ok to vary QoS within a 5-tuple?
  - 5-tuple = 2 IP addresses, protocol (e.g., UDP), 2 ports
  - For Web RTC and other real-time applications
- A: Only when transport protocol is UDP, but ...
  - ... even then, only with care (easy to get wrong)
  - Network may remove QoS differentiation
- See draft-ietf-dart-dscp-rtp (RFC Editor Queue)
  - DART WG (Diffserv Applied to Realtime Transports)
  - Recently completed RAI/TSV cross-area activity

Transport Area Scope

- "Core" transport protocols: TCP, SCTP, etc.
- Congestion Control & Queue Management
- NAT Traversal, UDP Encapsulation
- Quality of Service and Signaling
- Storage Networking
- Other Topics
  - Delay tolerant networking
  - Application Level Transport Optimization
  - TCP Incremental Security

Storage Networking

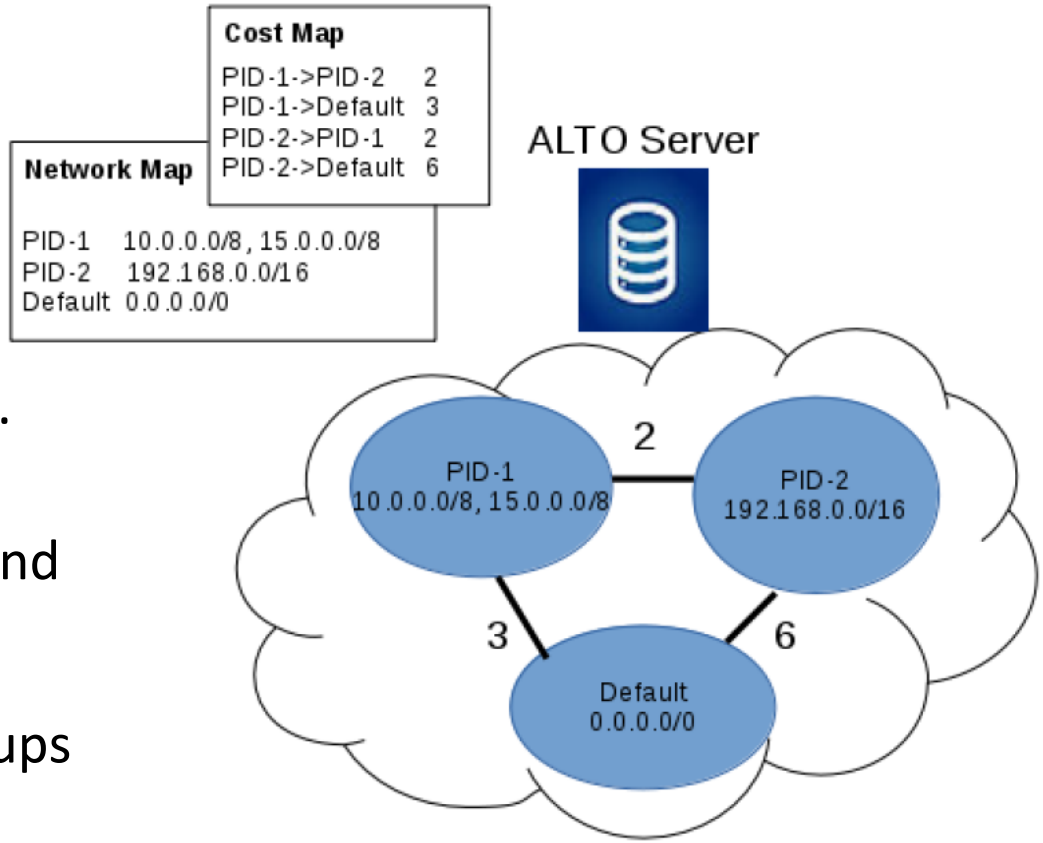
- Block (SAN) storage: iSCSI and FC/IP
  - In cooperation w/storage standards bodies
    - T10 [SCSI] and T11 [FC=Fibre Channel], respectively
    - [T10 and T11 are historical acronyms]
  - Storage Maintenance (STORM) WG
- File (NAS) storage: NFS (Network File System)
  - NFSv3, then NFSv4
  - Currently 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.

App Layer Transport Optimization (ALTO)

- Rendezvous: Select best resource from set of candidates
  - E.g.: Find closest node
  - Occurs in peer-to-peer network, CDNs (content delivery networks), app layer request routing, etc.
- Uses two maps
  - Network map: partition and group endpoints
  - Cost map: Costs between each adjacent pair of groups
- See RFC 7285



Network Performance Measurement

- IPPM WG: Can't manage what you can't measure
  - IPPM = IP Performance Metrics
  - Standard metrics for Internet transport performance
  - Methods to measure metrics and analyze results
- Recent focus: Verify access link performance
  - Following FCC's Measuring Broadband America
  - With LMAP WG
    - LMAP = Large-scale Measurement of Broadband Performance
    - In OPS (Operations and Management) Area
- Current work
  - New methods for bulk transfer capacity measurement
  - Simple metrics registry for comparability.

Opportunistic Transport Security

- Problem: Mass surveillance of insecure TCP connections
  - Adding TLS requires application support
  - Need to improve confidentiality and privacy.
    - But: minimal/no change to apps and no configuration (rules out IPsec)
- TCP Increased Security (TCPINC) WG
  - Add opportunistic security to TCP
  - Opportunistic = zero-config, not-necessarily-authenticated
- Several proposals
  - tcpcrypt, bindings to TLS, TCP-AO extensions
- All leverage TCP options
  - …which may or may not pass through middleboxes
  - …and may require additional options space

TSV: MEETINGS IN DALLAS

Dallas: Transport Area Meetings

- Area-Wide: TSVAREA, TSVWG
- TCP-related: TCPM, MPTCP, TCPINC
- Congestion Control: TCPM, ICCRG, RMCAT, AQM
- New protocols/deployment: TAPS, SPUD BOF
- NAT Traversal: TRAM
- Storage: NFSv4
- Everything Else: ALTO, CDNI, DTN, IPPM, PPSP

(*Acronym Expansions on Subsequent Slides)

TSV Area Meeting (TSVAREA)

- Venue for discussion of topics of general interest to the entire transport area.
- Dallas topics:
  - IETF Area reorg and likely effects on TSV
  - TSV Area Director duties and workload
  - Technical topics (see posted agenda)
- **Monday 15:20** in **Gold**
- **Friday 11:50** in **Venetian**

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 (primarily for Web RTC)
- UDP guidance update
  - Include encapsulation, multicast considerations
  - Also GRE-in-UDP encapsulation draft
- RSVP (reservation protocol) maintenance
  - Primary usage: MPLS traffic engineering
- QoS topics (circuit breakers, network interconnect, ECN and link layers, Web RTC usage)
- **Tuesday 15:20 & Thursday 13:00,** in **Parisian** (twice)

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: Rechartering to handle alternative congestion control algorithms
- **Tuesday 09:00** in **Oak**

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."
- Current work on:
  - Use cases and operational experience with MPTCP
  - Standards-track revision of the experimental spec
  - Guidelines for MPTCP-enabled middleboxes
  - Guidelines for implementors
- **Tuesday 13:00** in **Far East**

TCP Increased Security (TCPINC)

- "The TCPINC WG will develop the TCP extensions to provide unauthenticated encryption and integrity protection of TCP streams. The WG will define an unauthenticated key exchange mechanism. In addition, the WG will define the TCP extensions to utilize unauthenticated keys, resulting in encryption and integrity protection without authentication."
- Current work on defining requirements and selecting an approach.
- **Thursday 15:20** in **Parisian**

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."
- Dallas agenda: Congestion-related topics
  - Generally ahead of work in IETF Working Groups
- **Monday 09:00** in **Oak**

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 on CC algorithm evaluation results
- **Thursday 09:00** in **Venetian**

Active Queue Management (AQM)

- The Active Queue Management and Packet Scheduling working group (AQM) works on algorithms for managing queues in order to:
  1. minimize the length of standing queues
  2. help senders control their rates without unnecessary loss
  3. protect flows from negative impacts of other flows
  4. avoid synchronization of flows sharing a bottleneck
- Recommendations to be published soon
  - Current work: document and evaluating AQM algorithms
  - Examples: CODEL (and FQ-CODEL), PIE
- **Tuesday 17:30** in **Parisian**

Dallas: Transport Area Meetings

- Area-Wide: TSVAREA, TSVWG
- TCP-related: TCPM, MPTCP, TCPINC
- Congestion Control: TCPM, ICCRG, RMCAT, AQM
- New protocols/deployment: TAPS, SPUD BOF
- NAT Traversal: TRAM
- Storage: NFSv4
- Everything Else: ALTO, CDNI, DTN, IPPM, PPSP

(*Acronym Expansions on Subsequent Slides)

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 decomposing existing transports into the services they provide
- **Monday 13:00** in **Parisian**

Substrate Protocol for User Datagrams (SPUD) BoF

- Complementary approach to de-ossification to that taken by TAPS:
  - Build new transports in user or kernel space encapsulated in UDP (as Web RTC is doing).
  - Selectively expose transport and path information using a common substrate layer over UDP
- Non-WG forming BoF at **09:00 Wednesday** in **International**
  - BoF = Birds of a Feather (discussion session)
- Focus on use cases for middlebox cooperation, determining next steps

TURN Revised and Modernized (TRAM)

- "The goal of the TRAM Working Group is to consolidate the various initiatives to update TURN and STUN to make them more suitable for WebRTC… The work will include the addition of DTLS as an additional transport, authentication mechanisms, and extensions to TURN and STUN."
- New and improved NAT Traversal
  - e.g., support IPv6, DTLS, web origin (for use w/HTTP)
- **Wednesday 15:20** in **Far East**

Network File System version 4 (NFSv4)

- NFS Version 4 is the IETF standard for file sharing.
  - maintain the existing NFSv4, NFSv4.1, Federated Namespace, and related specifications
  - define NFSv4.2 and supporting protocols
  - Collect deployment guidance for NFSv4 FedFS implementations and their interaction with integration with new user authentication models.
- Current work focus: completing NFSv4.2
  - Also looking ahead to NFSv4.3, e.g., additional parallel NFS layout type
- **Thursday 09:00** in **Royal**

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
- Dallas topics: Topology, deployments, enhancements
- **Thursday 15:20** in **Continental**

Content Delivery Network Interconnection (CDNI)

- "The goal of the CDNI Working Group is to allow the interconnection of separately administered CDNs in support of the end-to-end delivery of content from CSPs through multiple CDNs and ultimately to end users (via their respective User Agents)"
- Framework: complete
- Current focus: completing interface definitions
- **Wednesday 13:00** in **Far East**

Delay Tolerant Networks (DTN)

- "The Delay/Disruption Tolerant Network Working Group (DTNWG) specifies mechanisms for data communications in the presence of long delays and/or intermittent connectivity."
- Current work on use case evaluation for updates to Bundle, LTP, etc.
- **Thursday 17:40** in **Oak**

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 on active measurement protocol maintenance, advancing basic metrics along the standards track, defining a metric registry and bulk transfer metrics for LMAP
- **Friday 09:00** in **Continental**

Peer to Peer Streaming Protocol (PPSP)

- "The Peer-to-Peer Streaming Protocol (PPSP) working group develops two signaling and control protocols for a peer-to-peer (P2P) streaming system for transmitting live and time-shifted media content with near real-time delivery requirements."
- Current work: Completing base tracker protocol
- **Tuesday 13:00** in **Royal**

TSV working groups that are not meeting in Dallas

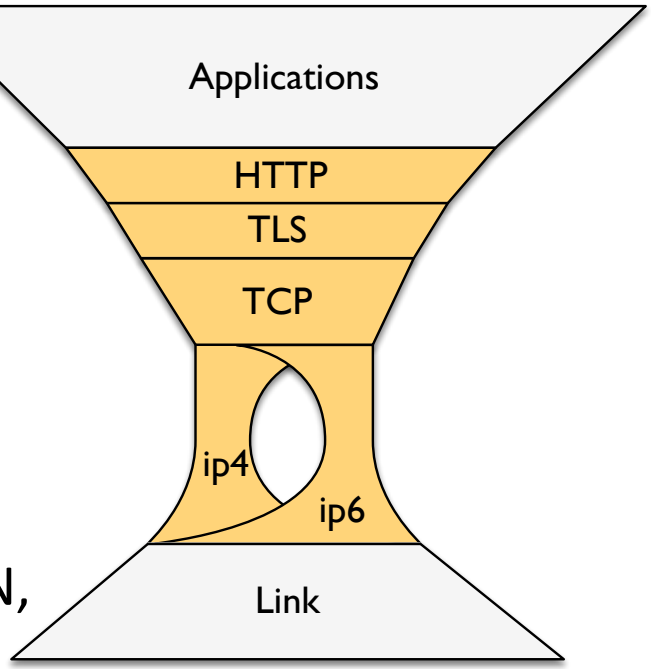
- CONEX (Congestion Exposure) WG
- STORM (STORage Maintenance) WG

- Both WGs have effectively completed their work and are likely to be closed soon.

TRANSPORT FUTURES: IAB SEMI WORKSHOP

Problem: Transport Layer Ossification

- Different transport protocols for a variety of use cases
  - But Internet runs over TCP
  - (and increasingly over HTTPS)
- Narrow interfaces: BSD sockets make the network look like a file descriptor, which is only half right
- Opaque paths: middleboxes assume lowest common denominator traffic
  - interferes with MPTCP, TCPINC, ECN, etc.
- What can be done?



IAB SEMI Workshop Report: Technical Plenary

- What can be done? IAB SEMI Workshop focus
  - Stack Evolution in a Middlebox Internet [SEMI]
  - Update at Monday evening plenary
- Productive workshop (Zurich, February)
  - One output: SPUD (Substrate Protocol for UDP Datagrams) BoF and related activity
  - Technical Plenary: Report on everything else
- **Monday 17:10** in **International**

Acknowledgments

- Olivier Bonaventure
- Scott Bradner
- Brian Carpenter
- Vijay Gurbani
- Jana Iyengar
- Mirja Kühlewind
- Allison Mankin
- Martin Stiemerling