Introduction to the Transport Area (TSV)

石航
清华大学计算机系博士
导师：崔勇
2021年9月29日

清華大學
Tsinghua University

Overview

"The transport and services area … covers a range of technical topics related to data transport in the Internet."

Area Director	职位	单位
Mirja Kühlewind	IAB chair	Ericsson, Sweden
Lars Eggert	IETF chair	NetApp, Finland
Spencer Dawkins	IAB member	Tencent, USA
Martin Duke	QUIC, TAPS	F5 Networks, USA
Zaheduzzaman Sarker	TAPS	Ericsson, Sweden

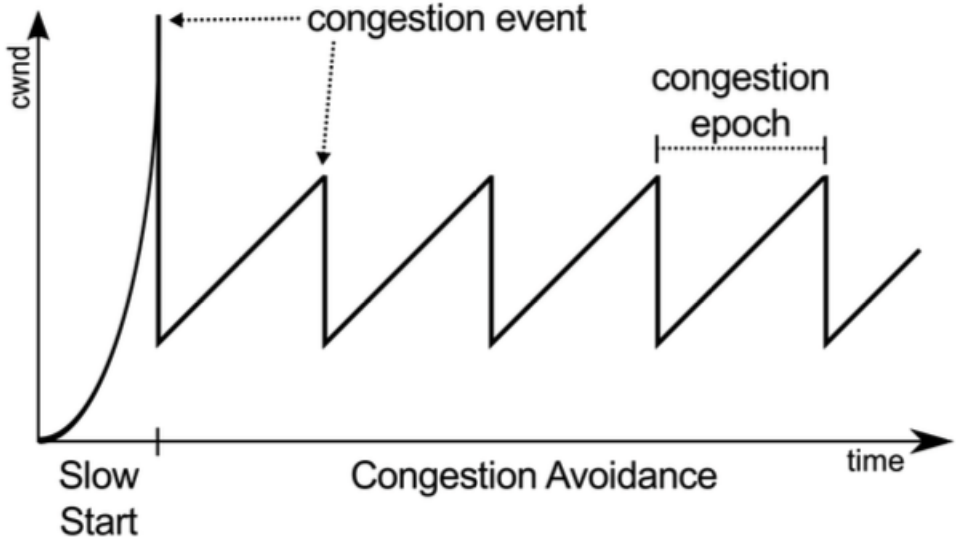
活跃公司

工作组概况

	WG	Name	RFCs	Drafts(WG + personal)	Emails Past Month	Emails Past Year
旧传输协议	tcpm	TCP Maintenance and Minor Extensions	42	7+0	71	856
	taps	Transport Services	5	3+2	17	189
	tram	TURN Revised and Modernized	9	1+0	3	43
新传输协议	quic	QUIC	4	8+23	76	1719
	masque	Multiplexed Application Substrate over QUIC Encryption	0	3+8	19	490
算法优化	rmcat	RTP Media Congestion Avoidance Techniques	9	1+0	1	13
	ippm	IP Performance Measurement	52	16+0	38	692
其他	nfsv4	Network File System Version 4	41	4+5	16	268
	alto	Application-Layer Traffic Optimization	9	4+12	35	438
	dtn	Delay/Disruption Tolerant Networking	0	4+3	30	393
	tsvg	Transport Area Working Group	80	13+13	159	1972

TCP Maintenance and Minor Extensions (tcpm)

- Core TCP protocol
  - 有序可靠的面向连接的传输(5 RFCs)
- 拥塞控制算法
  - Slow start
  - Congestion signal: loss or delay?
  - Loss detection
- 以DCTCP和MPTCP为例讲述论文和标准的关系



DCTCP： 学术论文如何成为标准

Stanford University

Microsoft Azure

Windows Server 2012

FreeBSD

Linux

IETF

2010, 2011 年SIGCOMM　2012 年　2013 年　2014 年　2015 年　2017 年

Data C Windows Se (DCTCP)

Mohammad Alizadeh[†‡], Albe
Parveen Patel[†], Balaji Pra

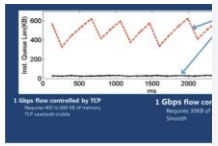
[†]Microso
{albert, dmaltz, padhye,
{aliz

HOME → HARDWARE, MICROSOFT,

Analysis of DCTC

Mohamma
Depa

In Windows Server 2012 Micros



$$F = \frac{\#\ of\ marke}{Total\ \#\ of}$$

IETF87 Berlin

DCTCP imp

FreeBSD

Differential › D604

DCTCP implementation.

Closed　Public

Authored by **hiren** on Aug 13 2014, 9:40

Details

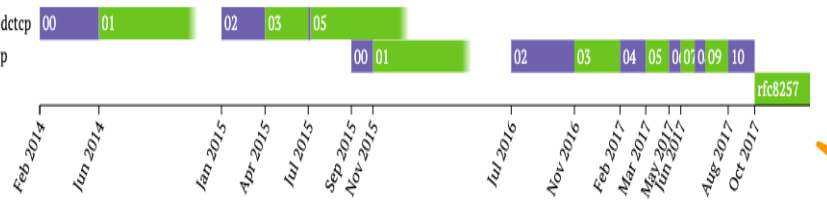
Reviewers　lstewart
gnn
bz

Group Reviewers　network

Commits　rS277054: D

SUMMARY

DCTCP (Datacenter TCP) implementation f
Eggert.



v5.15-rc2 ··· v3.18-rc1

borkmann authored and **davem330** committed on Sep 29, 2014

Showing 6 changed files with 425 additions and 3 deletions.

43 Documentation/networking/dctcp.txt

13 include/uapi/linux/inet_diag.h

26 net/ipv4/Kconfig

1 net/ipv4/Makefile

344 net/ipv4/tcp_dctcp.c

1 net/ipv4/tcp_output.c

aws

Microsoft Azure

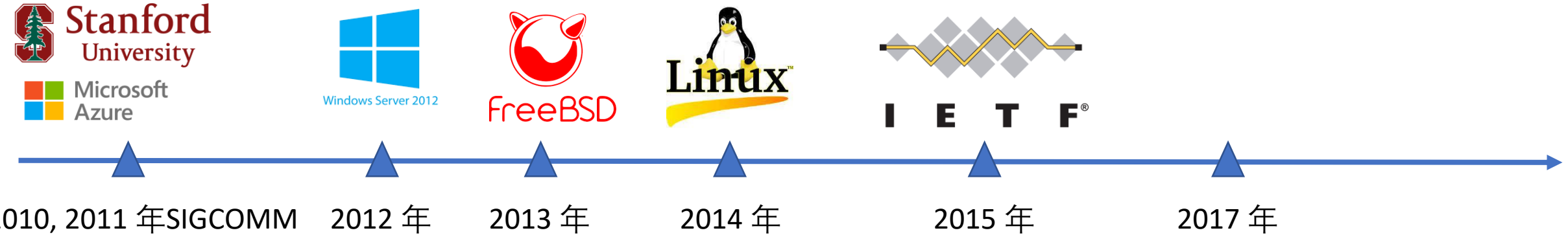
Alibaba Cloud

Google Cloud

➢ **Adaptive window decreases:** $Cwnd \leftarrow (1 - \frac{\alpha}{2})Cwnd$

– Note: decrease factor between 1 and 2.

DCTCP： 学术论文如何成为标准

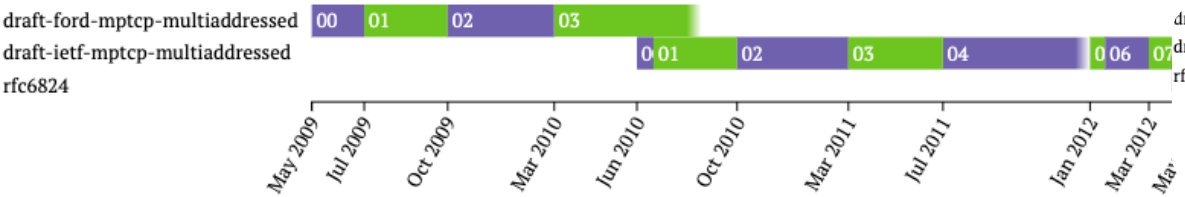


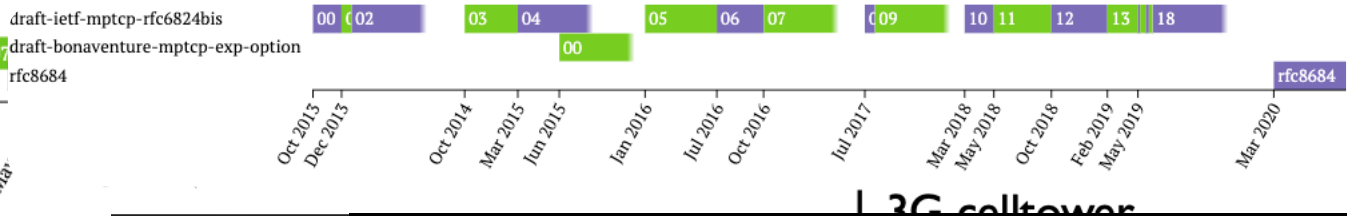
Acadamic

IETF

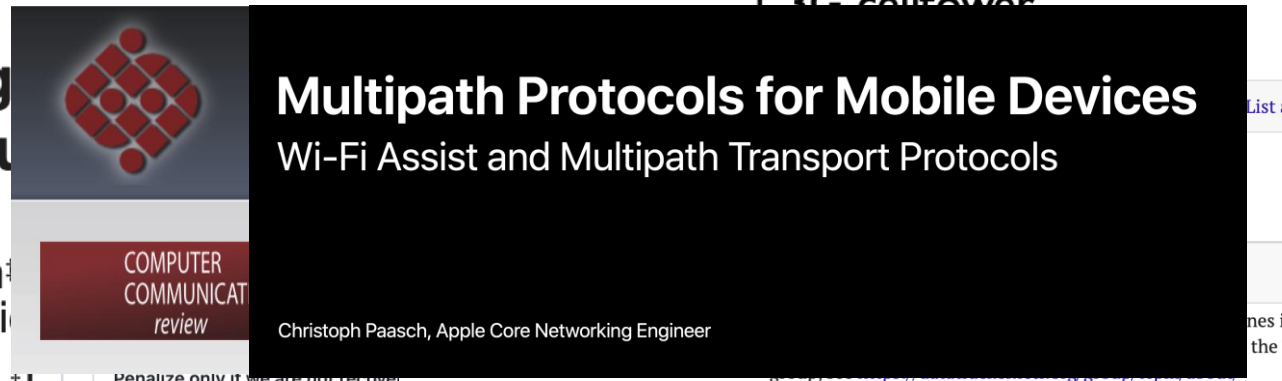
Industry

Multipath TCP：论文和标准并行





- 聚合多条路径(5G+Wi-Fi)的网络资源，提供更好的传输服务
- 可部署：对应用和网络透明



Multipath TCP：论文和标准并行

2009 年
WG成立

2011 年
IMC

2011 年
NSDI

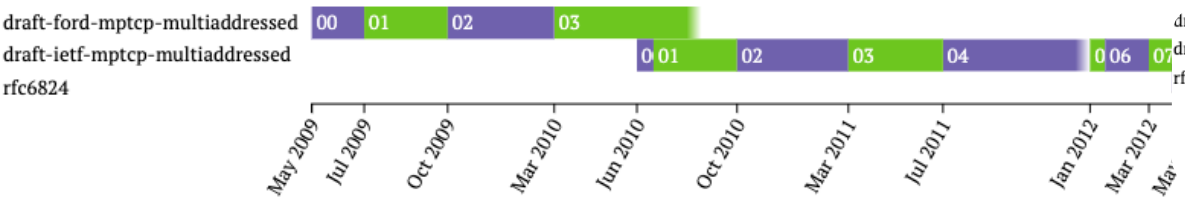
2012 年
NSDI

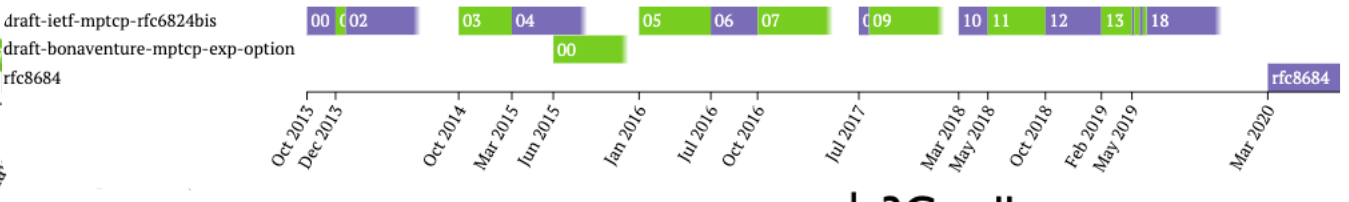
2013 年

2017 年

2019 年

2020 年





Design, imp

How Hard Can It Be? Design
Deployable Mu

Michi

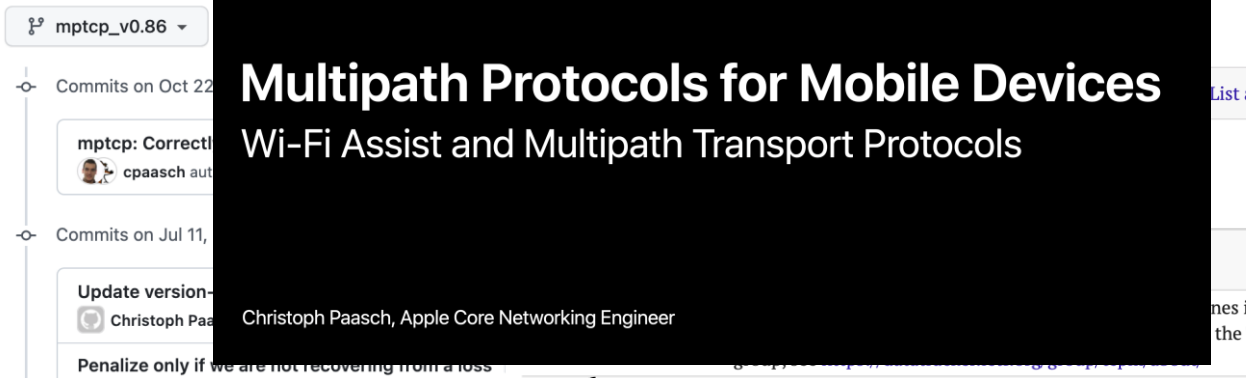
Dam

Costin Raiciu†, Christoph Paasch‡,
Michio Honda◇, Fabien Duchene‡, Olivie

Keio University,

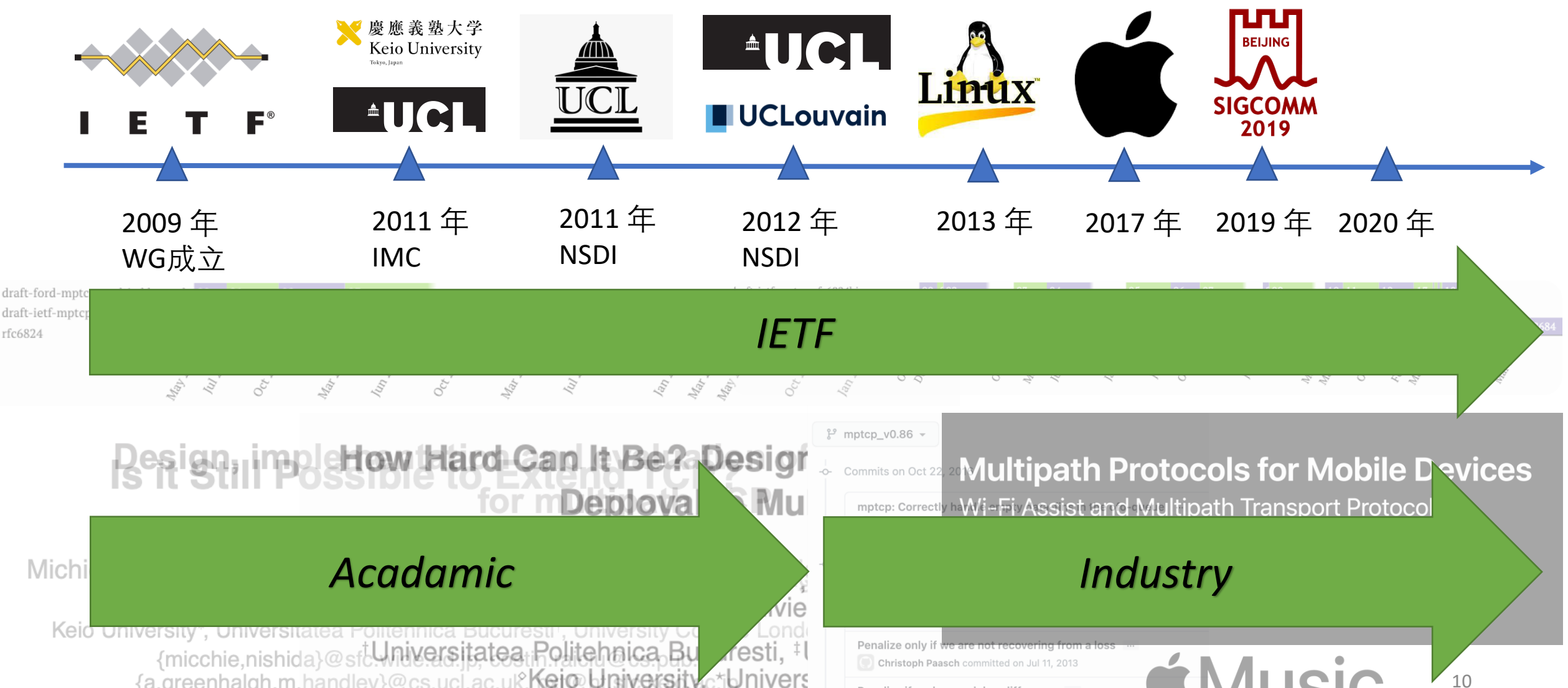
{micchie,nishic
{a.greenhalgh,m.

†Universitatea Politehnica Bucuresti, ‡
◇Keio University, ⋆Univers



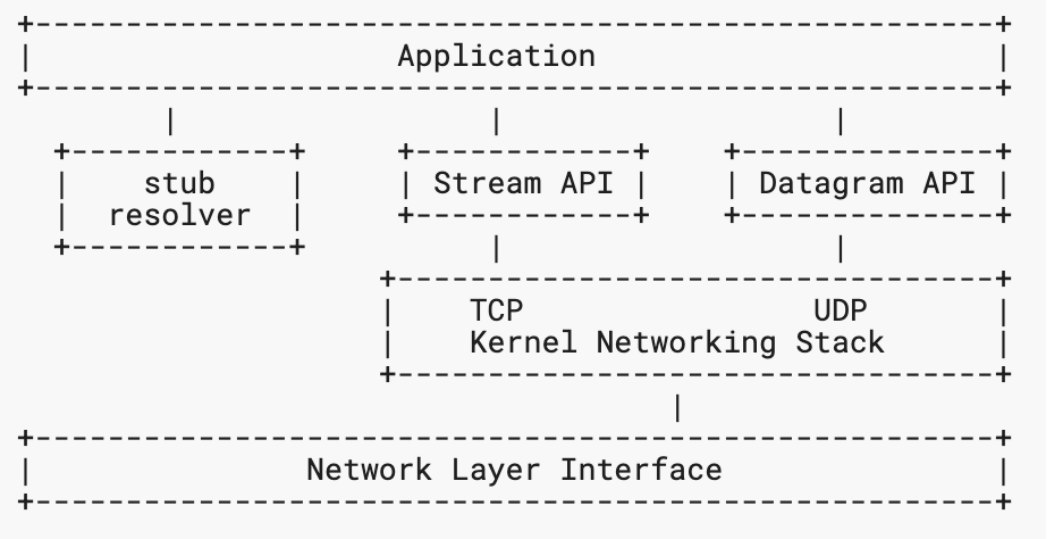
Music

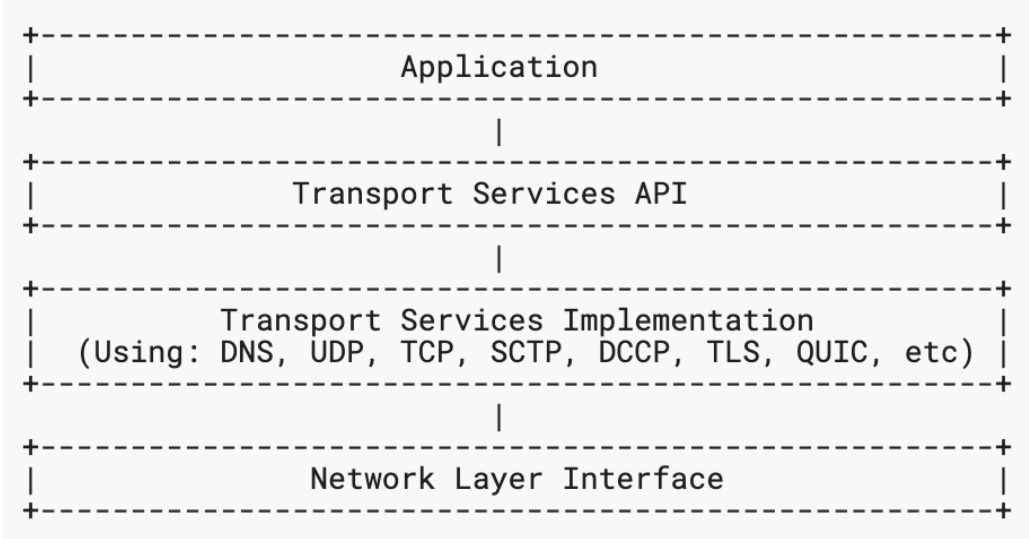
Multipath TCP：论文和标准并行



Transport Services (taps)

- 新传输协议=新API=部署困难，如何加快新传输协议的部署？
- API 面向传输服务而不是单个传输协议
  - 应用表达传输需求：可靠性，流式传输，安全性。Service层选择具体协议栈来满足需求。
  - 统一连接建立/断开，数据收发，事件处理API





TURN Revised and Modernized (tram)

- 如何处理NAT和Firewall等Middlebox
  - NAT 是邪恶的，有了IPv6就不需要了（否认）
  - 不能惯着Middlebox，忽略他们（愤怒）
  - Middlebox是去不掉了，让我们讨论如何规范他们的行为吧（讨价还价）
  - Middlebox也不听我们的，我们的新传输协议都无法部署了（绝望）
- NAT穿透问题
  - STUN：公网地址发现
  - TURN：NAT中继穿透
  - ICE：STUN+TURN使用框架，用于SIP
  - Near concluded

QUIC(quic)

- 截止2017年，互联网上7%的数据使用QUIC进行传送
- 2018年，IETF宣布，HTTP/3将弃用TCP协议，改为使用QUIC协议实现
- 2020年，华为在最新的HMS core网络加速套件中推出hQUIC
- 2020年， Facebook 宣布其超过75% 的网络流量使用 QUIC；……

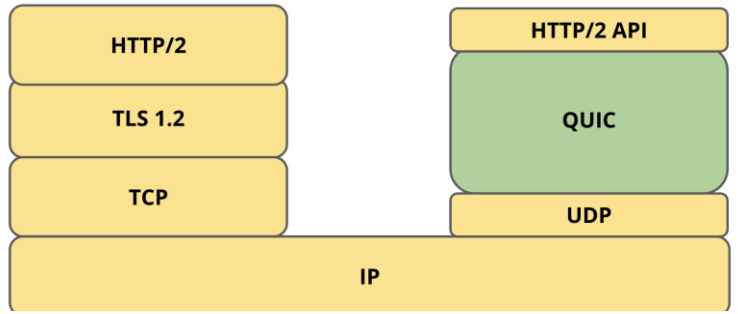
- iQUIC vs gQUIC
- 前任chair：Lars Eggert
- 前active contributor Martin Duke 现任Transport AD

The QUIC Transport Protocol: Design and Internet-Scale Deployment

Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, Jeff Bailey, Jeremy Dorfman, Jim Roskind, Joanna Kulik, Patrik Westin, Raman Tenneti, Robbie Shade, Ryan Hamilton, Victor Vasiliev, Wan-Teh Chang, Zhongyi Shi *
Google
quic-sigcomm@google.com

QUIC 的野心

- QUIC 谐音quick，高性能且快速迭代升级的传输协议
  - HTTP连接复用TCP会出现头部阻塞 -> 无阻塞的多流复用
  - TLS 握手+TCP 握手时间长 -> 0-RTT 握手
  - 整合TCP几十年的经验：快速恢复，尾包探测...
  - TCP 的改动部署困难
    - Middlebox 干涉 -> 强加密
    - 用户侧kernel难以升级 -> 基于UDP
- 20+开源实现
  - Apple/Google/FB/Apache/Nginx
- Datagram Over QUIC(Apple, Google)



QUIC
Internet-Draft
Intended status: Standards Track
Expires: 12 March 2022

T. Pauly
E. Kinnear
Apple Inc.
D. Schinazi
Google LLC
8 September 2021

An Unreliable Datagram Extension to QUIC
draft-ietf-quic-datagram-04

Abstract

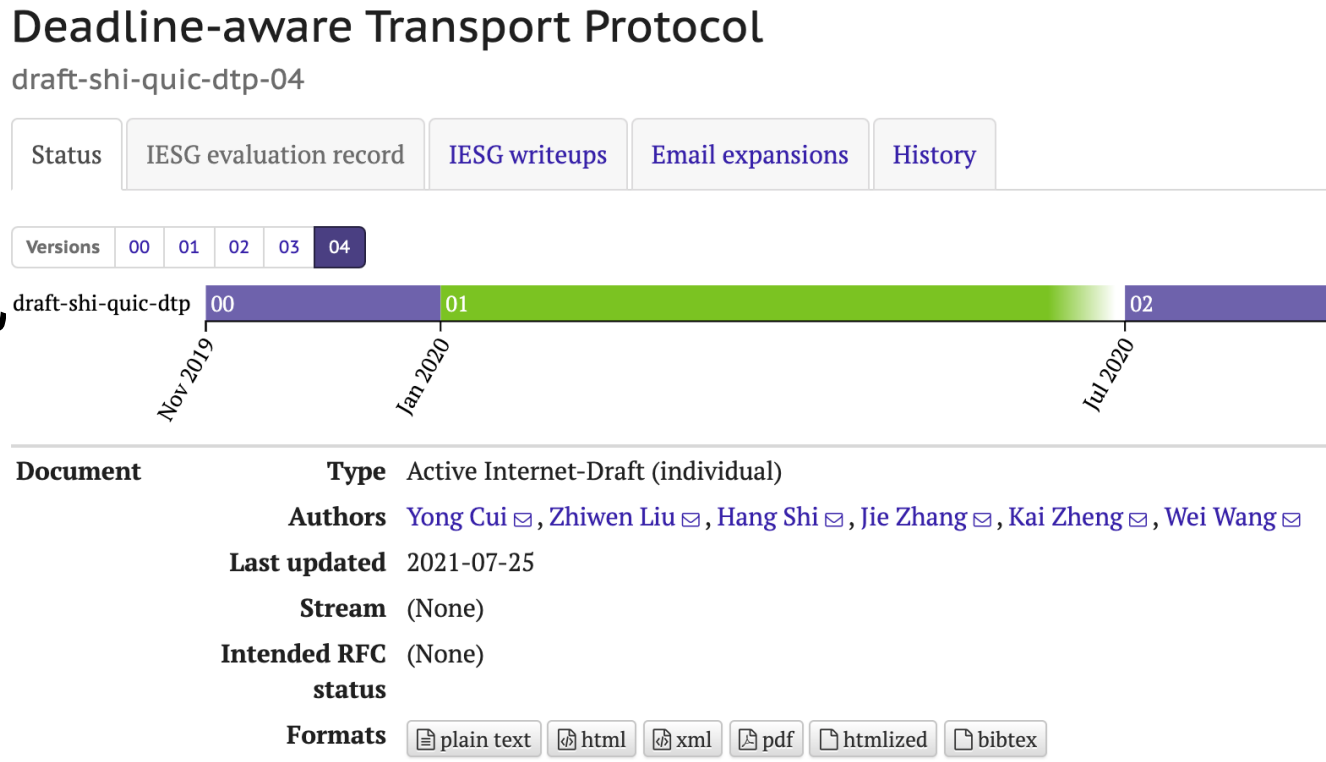
This document defines an extension to the QUIC transport protocol to add support for sending and receiving unreliable datagrams over a QUIC connection.

Discussion of this work is encouraged to happen on the QUIC IETF mailing list quic@ietf.org (mailto:quic@ietf.org) or on the GitHub repository which contains the draft: https://github.com/quicwg/datagram (https://github.com/quicwg/datagram).

Status of This Memo

QUIC拓展: Deadline-aware Transport Protocol（DTP）

- 满足应用在截止时间之前完成块传输的需求
  - 块：视频帧，游戏指令，网页元素
  - 完成时间 < deadline
- 避免丢包重传时延：增加冗余
- 避免排队时延：
  - 拥塞控制，避免自己堵死自己
  - 丢弃低优先级/过期数据
- 实现为QUIC的拓展，性能10x
- 参加开源软件供应链点亮计划
- 计划在QUIC工作组内推广



Multiplexed Application Substrate over QUIC Encryption (masque)

- HTTP 代理只能基于TCP，SOCKS代理无加密
- HTTP/3 over QUIC适合做代理：
  - 加密，多流复用，连接迁移等特性
- 年轻的工作组：
  - 2020年3月成立
  - 3个 WG drafts，暂无RFC
- 有前景：
  - QUIC chair, Google, Cloudflare支持
  - Mozilla 参与

Document
Active Internet-Drafts (3 hits)
draft-ietf-masque-connect-udp-04 **The CONNECT-UDP HTTP Method**
draft-ietf-masque-h3-datagram-03 **Using Datagrams with HTTP**
draft-ietf-masque-ip-proxy-reqs-03 **Requirements for a MASQUE Protocol to Proxy IP Traffic**

Document
Related Internet-Drafts (8 hits)
draft-cms-masque-connect-ip-02 **The CONNECT-IP HTTP Method**
draft-cms-masque-connect-ip-ext-routes-00 **A Routing Extension to CONNECT-IP**
draft-kuehlewind-masque-connect-ip-01 **The CONNECT-IP HTTP method for proxying IP traffic**
draft-pardue-masque-dgram-priority-01 **HTTP Datagram Prioritization**
draft-pauly-masque-quic-proxy-01 **QUIC-Aware Proxying Using CONNECT-UDP**
draft-schwartz-masque-h3-datagram-ping-00 **HTTP Datagram PING**
draft-tbd-masque-connect-ip-ext-flow-00 **A Flow Forwarding Mode Extension to CONNECT-IP**
draft-westerlund-masque-transport-issues-02 **Transport Considerations for IP and UDP Proxying in MASQUE**

实时拥塞控制算法及QoS

- RTP Media Congestion Avoidance Techniques (rmcat)
  - 实时音视频的需求不同：低时延，容忍丢包，小抖动。
  - Video traffic pattern 的特性是否可以利用?
  - 非丢包的拥塞控制算法之间的公平性问题?
  - RTP 和 RTCP配合实现拥塞控制
- IP Performance Measurement (ippm)
  - 路径的QoS如何表征：Loss pattern，Packet reordering
  - 如何测量：one/two-way active measurement protocol

其他

- Network File System Version 4 (nfsv4)
  - Network file system，RDMA，RPC标准
- Application-Layer Traffic Optimization (alto)
  - 根据 performance，cost 选择最佳的content 节点
- Delay/Disruption Tolerant Networking (dtn)
  - 火星星际通信超长时延需要存储转发节点
  - Bundle protocol：存储转发的基本单位，可以基于TCP/UDP/others
- Transport Area Working Group (tsvwg)
  - 所有其他的drafts
  - SCTP，DiffServ，RSVP

总结

- Thanks to QUIC 传输层大有可为
  - 把(MP)TCP上的工作再做一遍
  - 新的传输服务拓展 QUIC，比如DTP:加入我们!
- 与中间网络设备配合?

谢谢大家