Internet capacity sharing

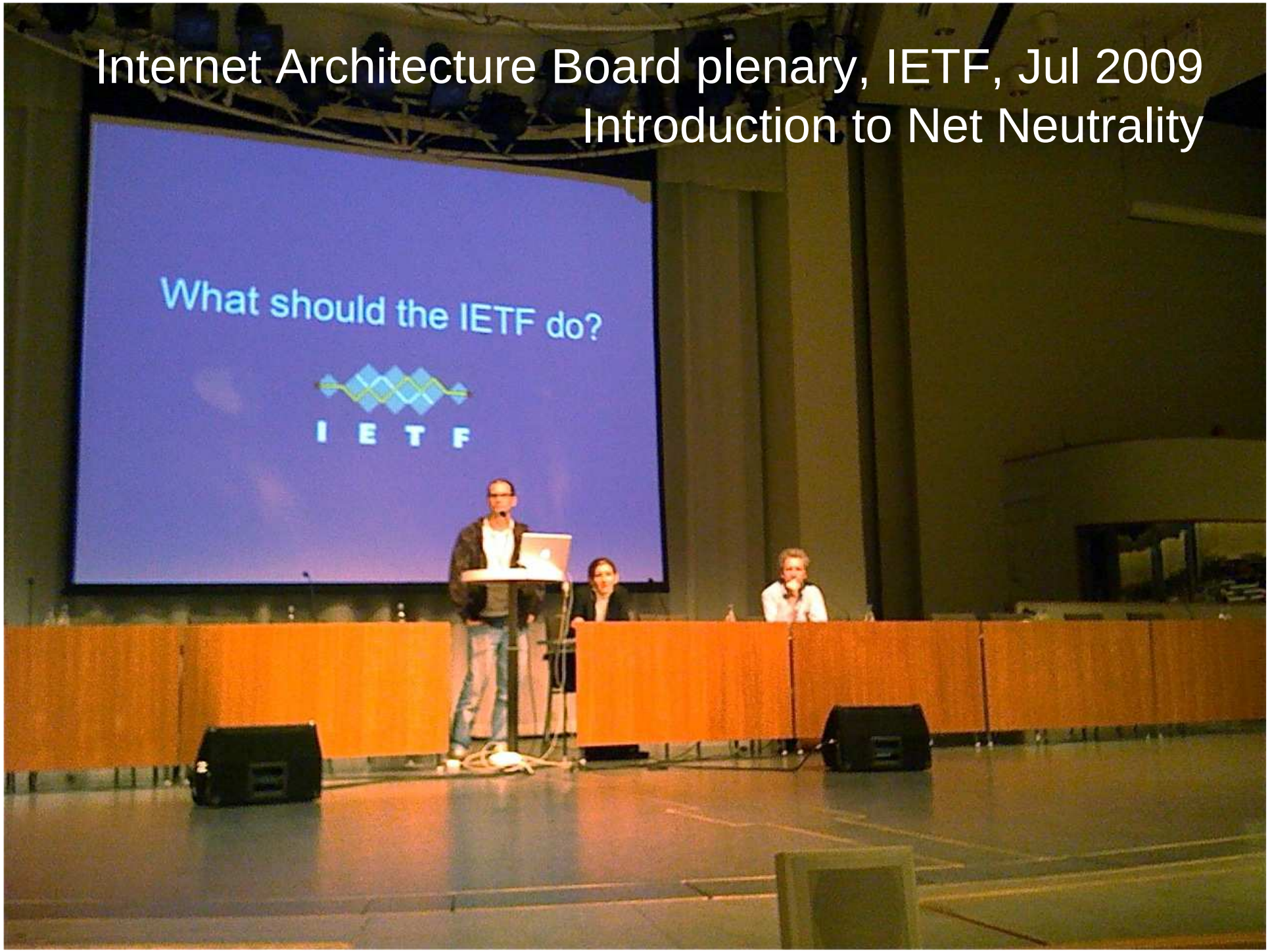
ECN, re-ECN & IETF status report

Bob Briscoe
Chief Researcher, BT

Oct 2009

This work is partly funded by Trilogy, a research project supported by the European Community
www.trilogy-project.org

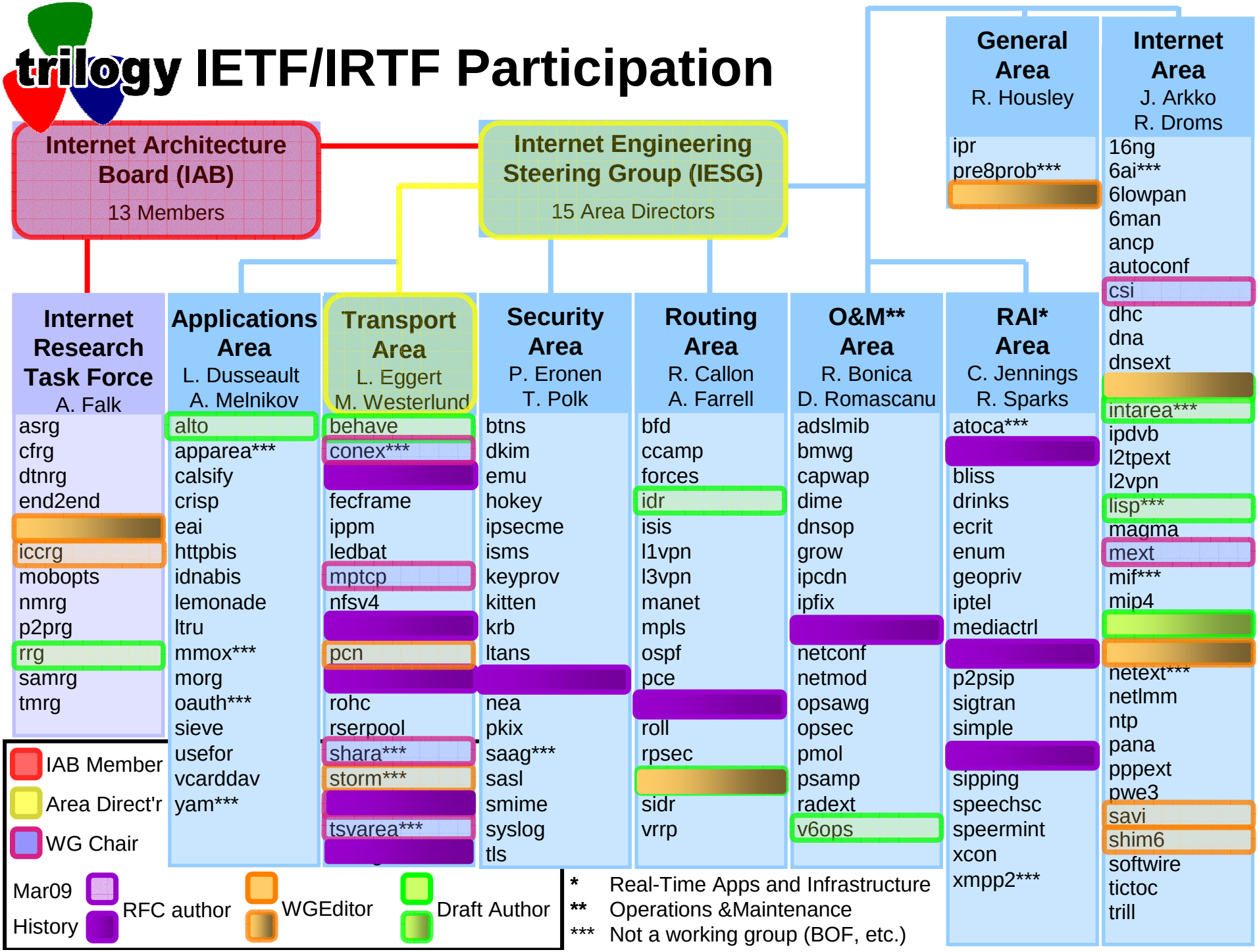
Internet Architecture Board plenary, IETF, Jul 2009
Introduction to Net Neutrality



trilogy
re-architecting the Internet

www.trilogy-project.org

trilogy IETF/IRTF Participation



moving mountains ptI

Internet Engineering Task Force

- Nov 2005
  - proposed replacement resource sharing architecture to IETF
  - general response: "What's the problem? TCP prevalent, so sharing OK"
- Nov 2006
  - Dismantled TCP-Friendliness religion at IETF transport plenary
- Nov 2008
  - thought leaders agree TCP dynamics correct, but sharing goal wrong
  - agreed to draft new Internet capacity sharing architecture
  - IETF delegated process to IRTF design team
  - within Internet Congestion Control Research Group (ICCRG)
    - eventual intent: endorsement by Internet Architecture Board
  - main points in new architecture
    - über-control of resource sharing in network, not end-points
    - dynamic control still primarily in end-points
- Mar 2009
  - straw poll in IETF Transport Area plenary
  - “Is TCP-friendly the way forward?” Y: Zero N: most of the hall

moving mountains ptII

Internet Engineering Task Force

glossary
IETF Internet Engineering Task Force
IESG Internet Engineering Steering Group
IAB Internet Architecture Board
IRTF Internet Research Task Force

- Oct 2009
  - proposed IETF working group: “congestion exposure”
    - candidate protocol: re-ECN (experimental change to IP)
  - IESG / IAB given go-ahead for Hiroshima IETF, Nov’09
    - non-binding vote on working group formation
  - >40 offers of significant help in last few weeks; *individuals* from
    - Microsoft, Nokia, Cisco, Huawei, Alcatel-Lucent, NEC, Ericsson, NSN, Sandvine, Comcast, Verizon, …
    - about 50:50 industry / academia
- Nov 2009
  - will not be asking for endorsement to change to IP
    - defer until support is much wider

moving mountains ptIII
the global ICT industry

- GIIC: ~50 CxOs of the major global ICT corporations
  - Apr 09: then BT CTO, Matt Bross (now Huawei Global CTO)
    - proposed GIIC endorses BT solution
  - commissioners voted for endorsement decision within 30 days of expert review: public policy, commercial & technical
  - 30 Sep 09: favourable expert review in front of and by CxOs
    - all supported, but pointed out known obstacle (ie. ambitious)
  - report due late Oct'09
  - if endorsed, becomes corporate lobbying position, standards position etc

how to share the capacity of the Internet?

- the job of end-to-end L4 protocols (e.g. TCP)?
  - TCP's dynamic response to congestion is fine
  - but the way it shares capacity is very wrong
- ISP's homespun alternatives have silently overridden TCP
  - result: blocks, throttles & deep packet inspection
  - if it's new, it won't get through (if it's big, it won't either)

- IETF transport area consensus reversed since 2006
  - 'TCP-friendly' was useful, but not a way forward
  - rewrite of IETF capacity sharing architecture in process
  - commercial/policy review in process driven by 'captains of industry'

- approach: still pass info up to L4 to do capacity sharing
  - but using weighted variants of existing congestion controls (weighted TCP)
    - similar dynamics, different shares
  - give incentive for apps to set weights taking everyone into account
    - backed by enforcement – simple ingress policing

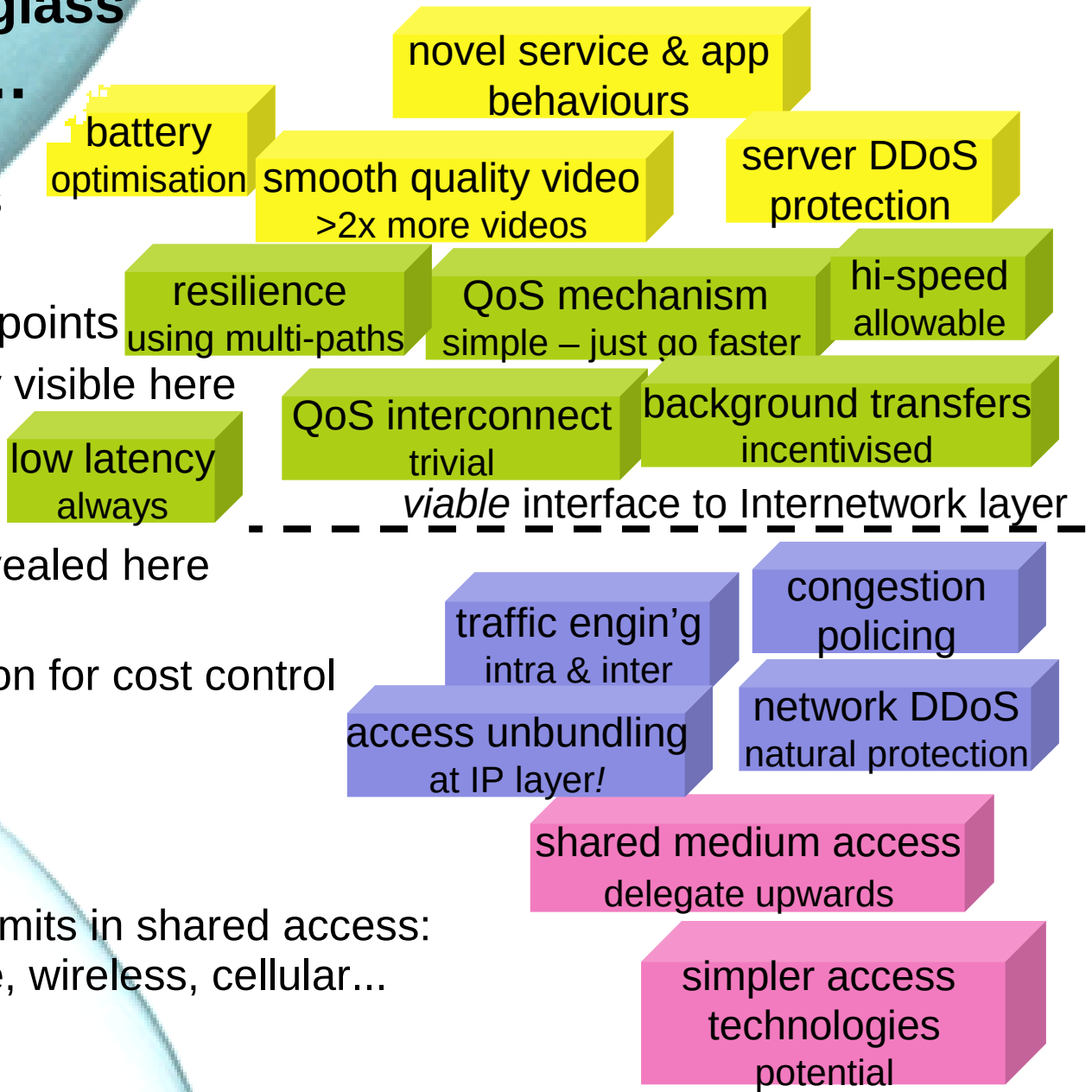
Internet topology visualization produced by Walrus (Courtesy of Young Hyun, CAIDA)

best without effort

- did you notice the interconnected QoS mechanism?
  - *endpoints* ensure tiny queuing delay & loss for all traffic
  - API: if your app wants more bit-rate, it just goes faster
  - effects seen in bulk metric at every border (for SLAs, AUPs)
- simple – and all the right support for operations

the neck of the hourglass
change 1 bit and…

- applications & services
- transport layer on end-points
  - usage costs currently visible here
- internetwork layer
  - once usage costs revealed here
  - ISPs won't need deep packet inspection for cost control
- link layer
  - can remove bit-rate limits in shared access: passive optical, cable, wireless, cellular...



reality check
ECN deployment

quick tutorial on ECN design for partial deployment

- if host ECN enabled, tries to use for all connections
  - if not, ignores ECN part of incoming connection requests
- IP header tells network whether endpoints talk ECN
- congested forwarding element will drop packets
  - if it's ECN-enabled, marks ECN-enabled packets instead
  - dangerous to mark not drop if receiver won't understand

- TCP header negotiates ECN support
  - when ECN client sends TCP SYN (initialisation packet)
    - ECN on in TCP header, off in IP header
    - if server supports ECN, SYN-ACK has ECN on in both
- other TCP-derived e2e transports are similar (DCCP/SCTP)
- UDP-based protocols (e.g. RTP/RTCP used in VoIP)
  - ECN negotiation is undefined (standardisation just starting)

status of ECN in TCP/IP

- although IETF proposed standard since 2001
  - patchy implementation & precious little active deployment
- Windows Vista & Linux
  - on by default for TCP listening sockets (servers)
  - off by default for TCP client sockets
  - cause: if it's new, don't let it through
    - originally random blocking by firewalls & NATs
      - all believed fixed by 2003
    - now it's a few broken models of home gateway
      - TCP/ECN SYN (init packet): 4 drop, 1 crashes
      - note: SYN doesn't turn on ECN in IP, only TCP
- ECN black hole detection (disable ECN if initial pkt dropped)
  - Vista?
  - Linux mainline distribution: philosophically opposed
    - available in distributor patches & default in some distr's

status of ECN support in routers & switches

- standardisation
  - ECN in IP: Mar 2001
  - ECN in MPLS: Jan 2008
  - ECN in IEEE802: work in (early) progress
    - don’t need ECN at L2 if subnet non-meshed or non-blocking
- some large equipment manufacturers
  - Cisco: ECN in many products, but not hi-speed core
  - Huawei: supports ECN in MPLS, but not in IP
  - Juniper: no ECN support AFAIK
  - Ericsson: active on ECN standardisation in 3GPP & IETF

- reason for patchiness: few requests from operators
- reason: incremental performance improvement
  - new product offerings trigger network change
  - ECN gain not sufficient to package as a new product offering
  - competitive performance advantage insufficient
    - except wireless?

tailpiece
ECN in UDP

- early tests seem to reveal a new set of problems
- individual cases of:
  - listening UDP socket not passing ECN from IP to app
  - UCL firewall (?): not just broken but dangerous
    - Jul’09 firewall forwarded ECN in UDP/IP unscathed
    - Aug’09 same firewall *cleared* the ECN field in UDP/IP
    - can suppress congestion indications leading to collapse
    - probably a broken attempt to ‘bleach’ the Diffserv field

more info...

- The whole story in 7 pages
  - Bob Briscoe, “Internet Fairer is Faster", BT White Paper (Jun 2009) ...this formed the basis of:
  - Bob Briscoe, "A Fairer, Faster Internet Protocol", IEEE Spectrum (Dec 2008)
- Slaying myths about fair sharing of capacity
  - [Briscoe07] Bob Briscoe, "Flow Rate Fairness: Dismantling a Religion" ACM Computer Communications Review 37(2) 63-74 (Apr 2007)
- How wrong Internet capacity sharing is and why it's causing an arms race
  - Bob Briscoe et al, "Problem Statement: Transport Protocols Don't Have To Do Fairness", IETF Internet Draft (Jul 2008)
- re-ECN protocol spec
  - Bob Briscoe et al, “Adding Accountability for Causing Congestion to TCP/IP” IETF Internet Draft (Mar 2009)
- Re-architecting the Internet:
  - The Trilogy project <www.trilogy-project.org>

IRTF Internet Capacity Sharing Architecture design team
<http://trac.tools.ietf.org/group/irtf/trac/wiki/CapacitySharingArch>

re-ECN & re-feedback project page:
<http://bobbriscoe.net/projects/refb/>

Congestion Exposure (ConEx) IETF ‘BoF’: <http://trac.tools.ietf.org/area/tsv/trac/wiki/re-ECN>
subscribe: <https://www.ietf.org/mailman/listinfo/re-ecn>, post: re-ecn@ietf.org

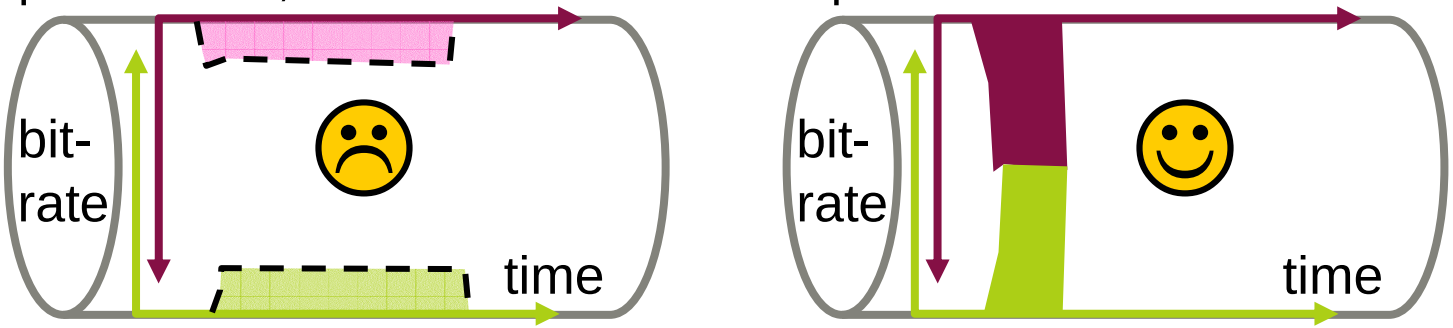
Internet capacity sharing for packets not flows

discuss...

congestion is not evil

congestion signals are healthy

- no congestion across whole path ⇒ feeble transport protocol
  - to complete ASAP, transfers should sense path bottleneck & fill it



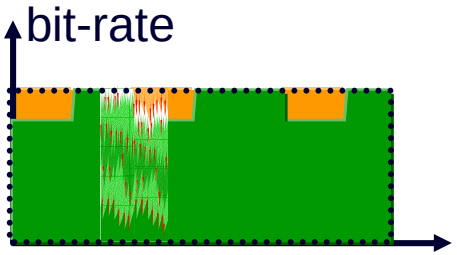
the trick

congestion signal *without* impairment

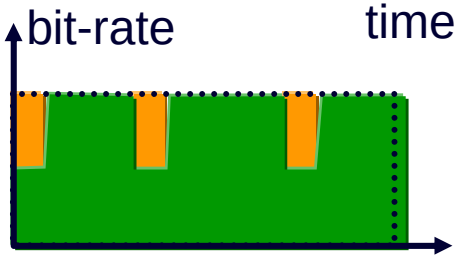
- explicit congestion notification (ECN)
  - update to IP in 2001: mark more packets as queue builds
- then tiny queuing delay and tiny tiny loss for all traffic
- no need to overprovision (whether core, access or borders) to prevent impairment

no traditional sharing approaches harness end-system flexibility… over time

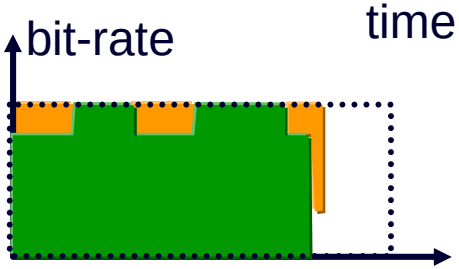
1. TCP



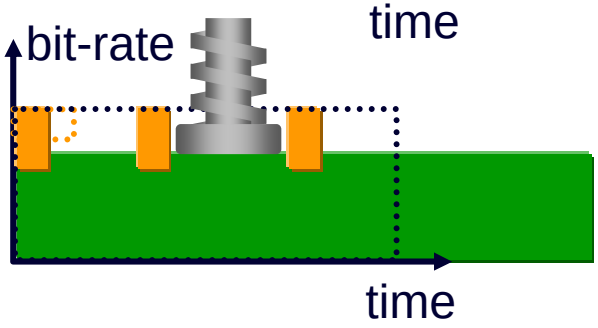
2. WFQ



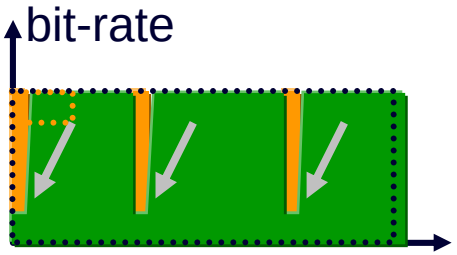
3. volume cap

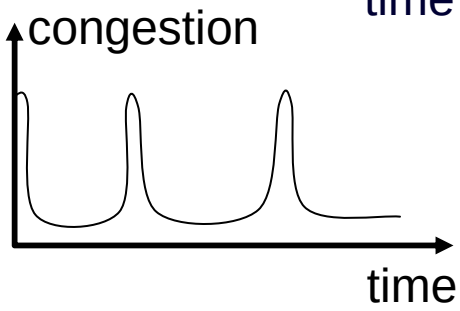


4. DPI



weighted sharing





- light usage can go much faster
- hardly affects completion time of heavy usage

NOTE: weighted sharing doesn't imply differentiated network service

- just weighted aggressiveness of end-system's rate response to congestion cf. LEDBAT

measuring marginal cost

- user's contribution to congestion
  = bytes marked
- can transfer v high volume
  - but keep congestion-volume v low
  - similar trick for video streaming