Decoupling TCP from IP with Multipath TCP

Olivier Bonaventure
http://inl.info.ucl.ac.be
http://perso.uclouvain.be/olivier.bonaventure

Thanks to Sébastien Barré, Christoph Paasch, Grégory Detal, Mark Handley, Costin Raiciu, Alan Ford, Micchio Honda, Fabien Duchene and many others

April 2016

Agenda

- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases

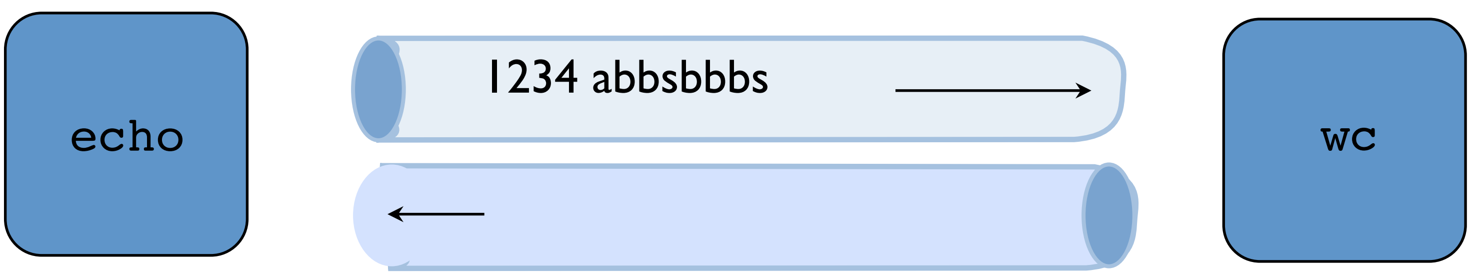
The origins of TCP

Source : http://spectrum.ieee.org/computing/software/the-strange-birth-and-long-life-of-unix

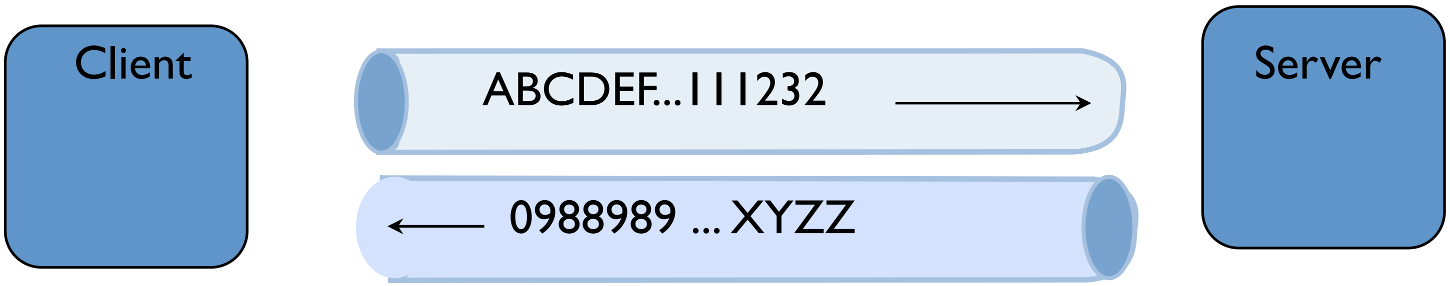
The Unix `pipe` model

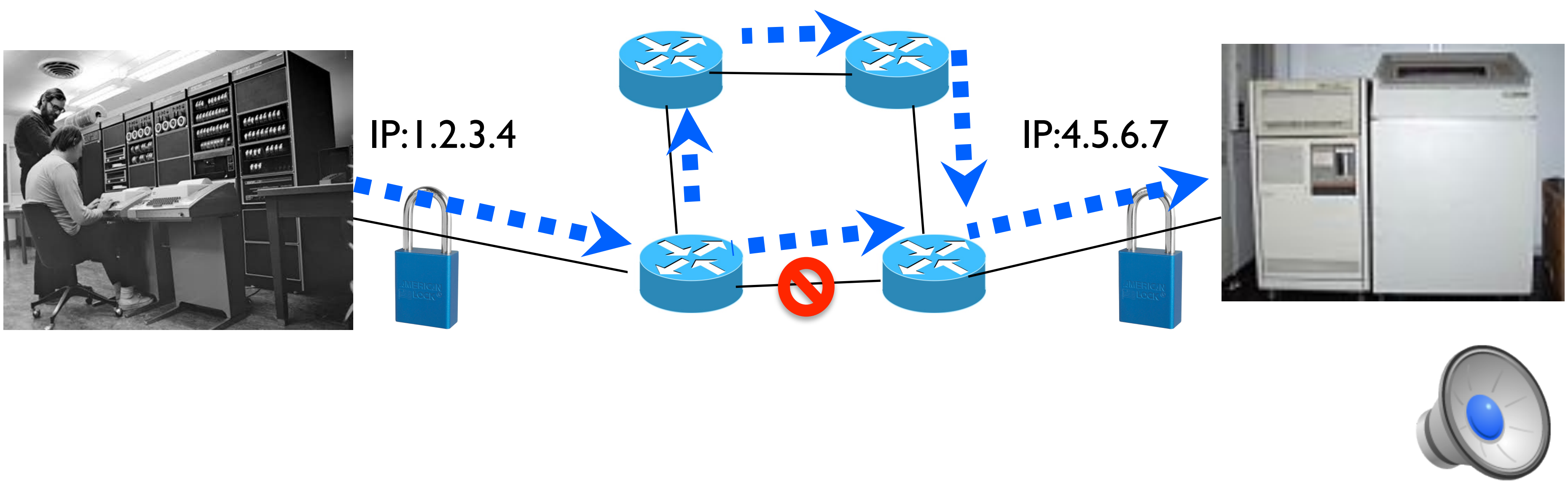
Terminal — bash — 49×7

```
Last login: Tue Nov 13 10:07:47 on ttys006
You have new mail.
mbpobo:~ obo$ echo "1234 abbsbbbs" | wc -c
      14
mbpobo:~ obo$ 
```



The TCP bytestream model





Endhosts have evolved

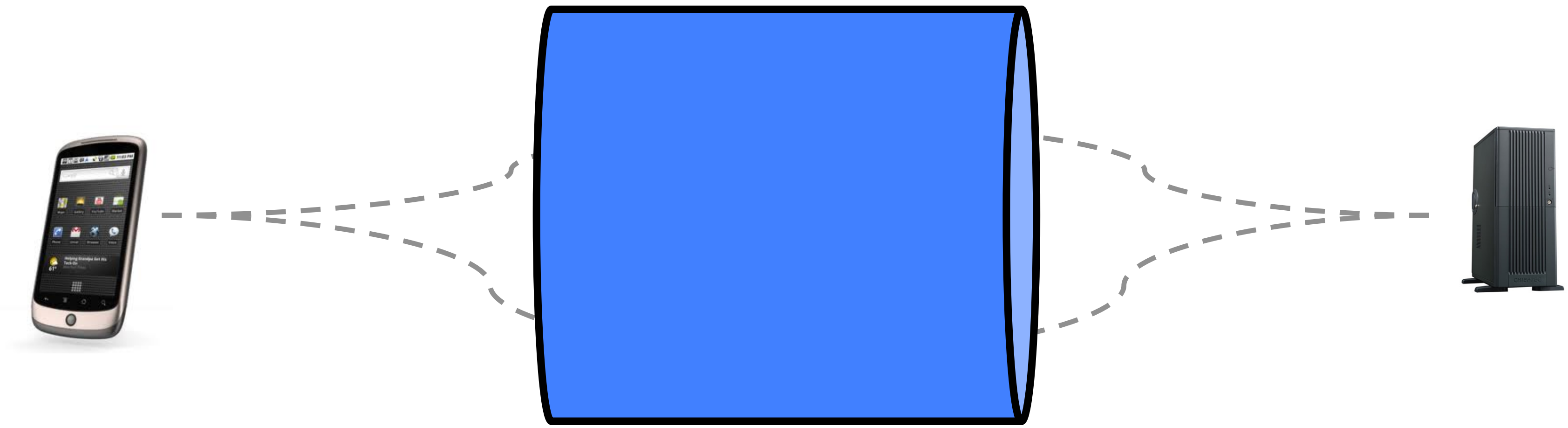




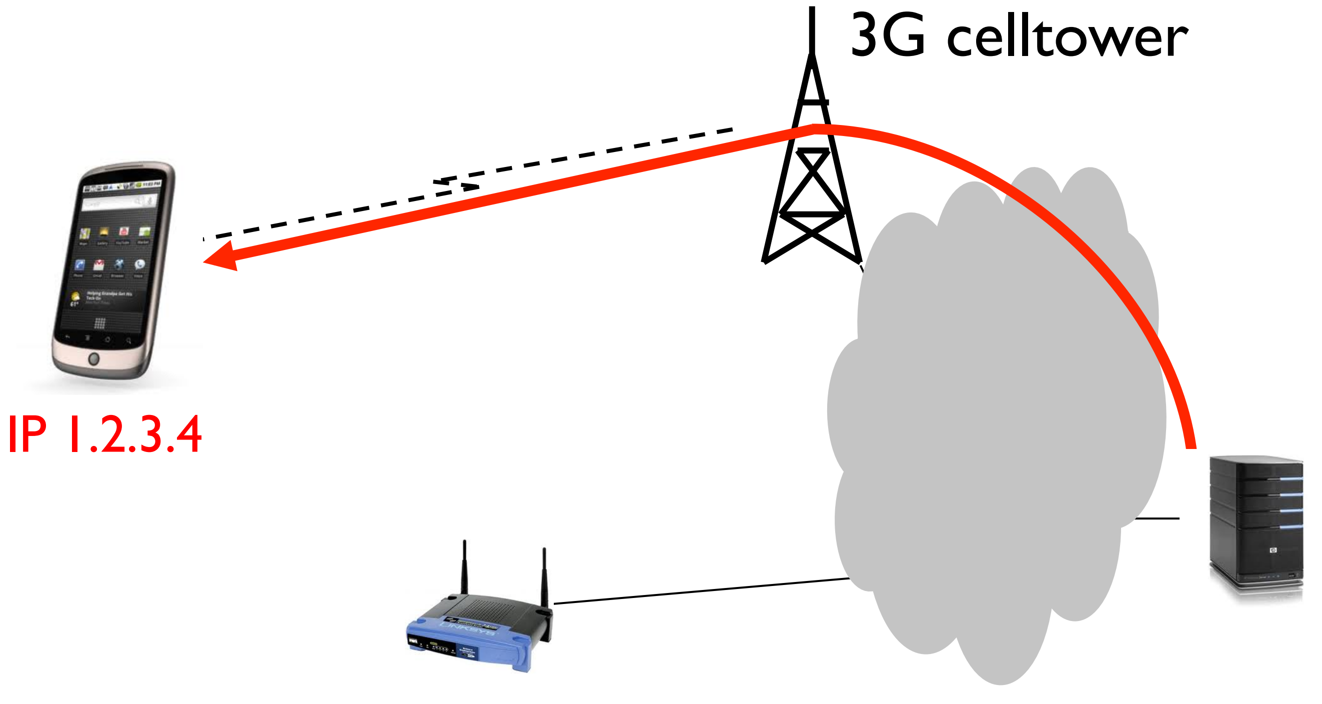


Mobile devices have multiple wireless interfaces

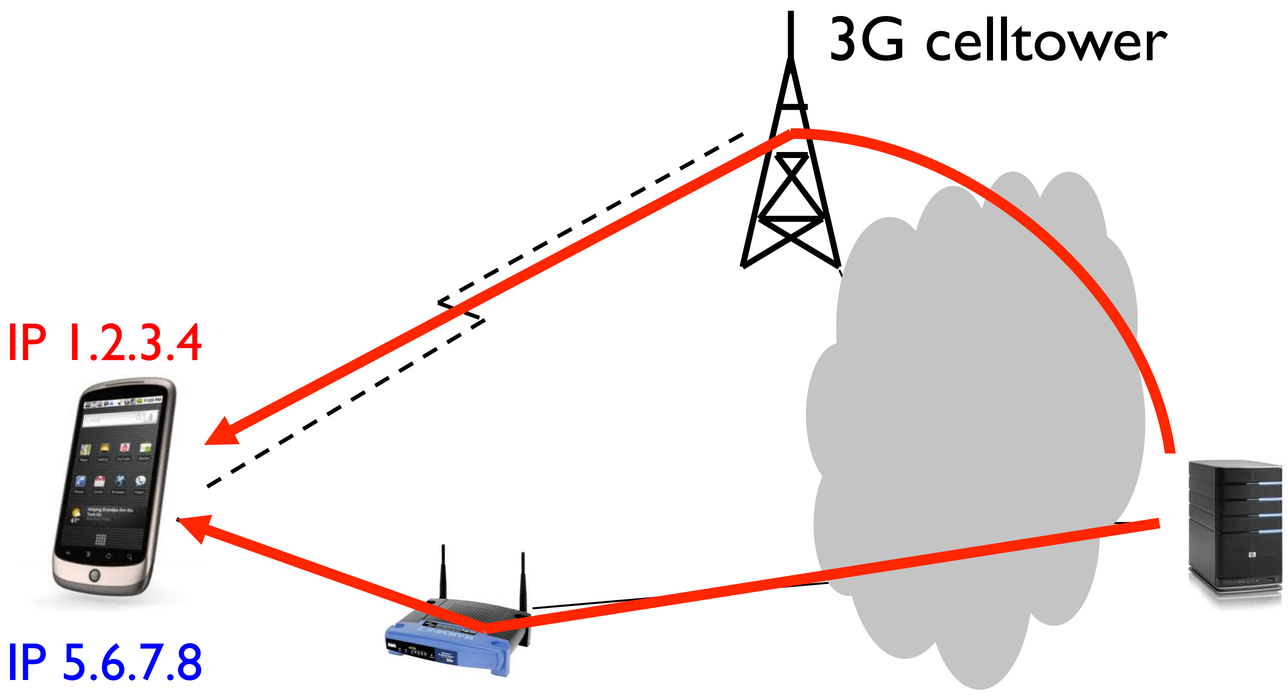
User expectations



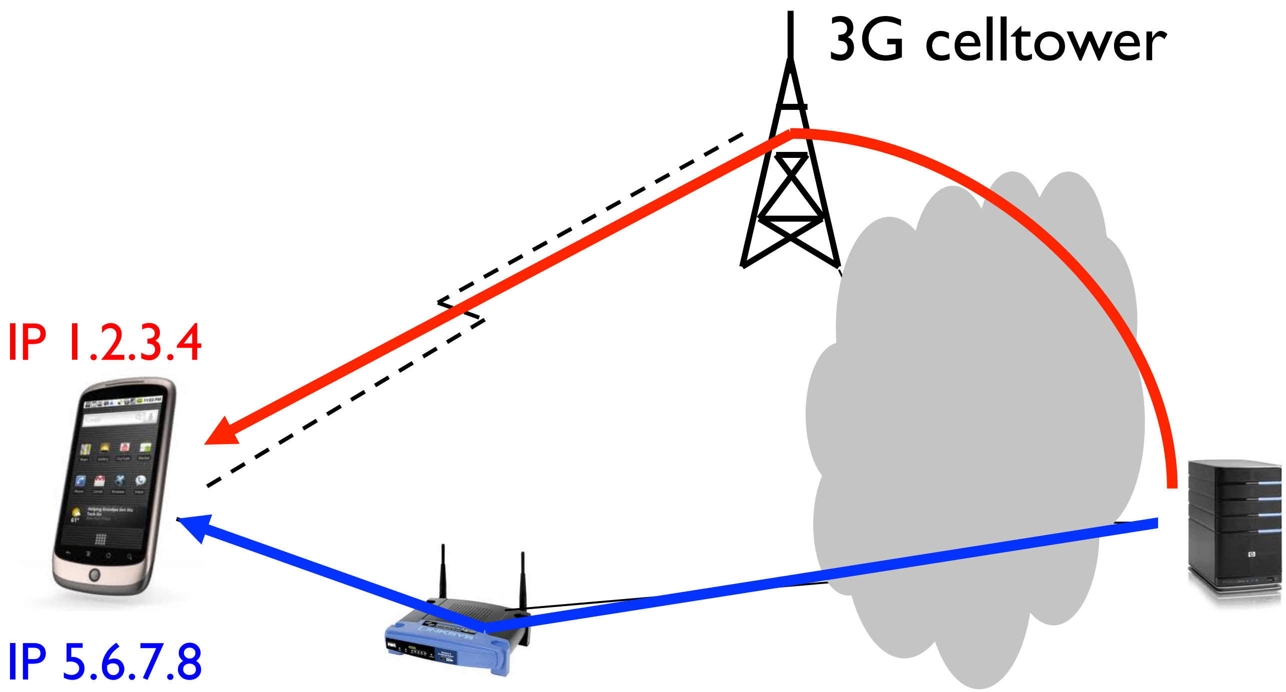
What technology provides



What technology provides



What technology provides



When IP addresses change TCP connections have to be re-established !

Equal Cost Multipath

ECMP implementation

Packet arrival :

$\text{Hash}(IP_{src}, IP_{dst}, Prot, Port_{src}, Port_{dst}) \bmod \#oif$

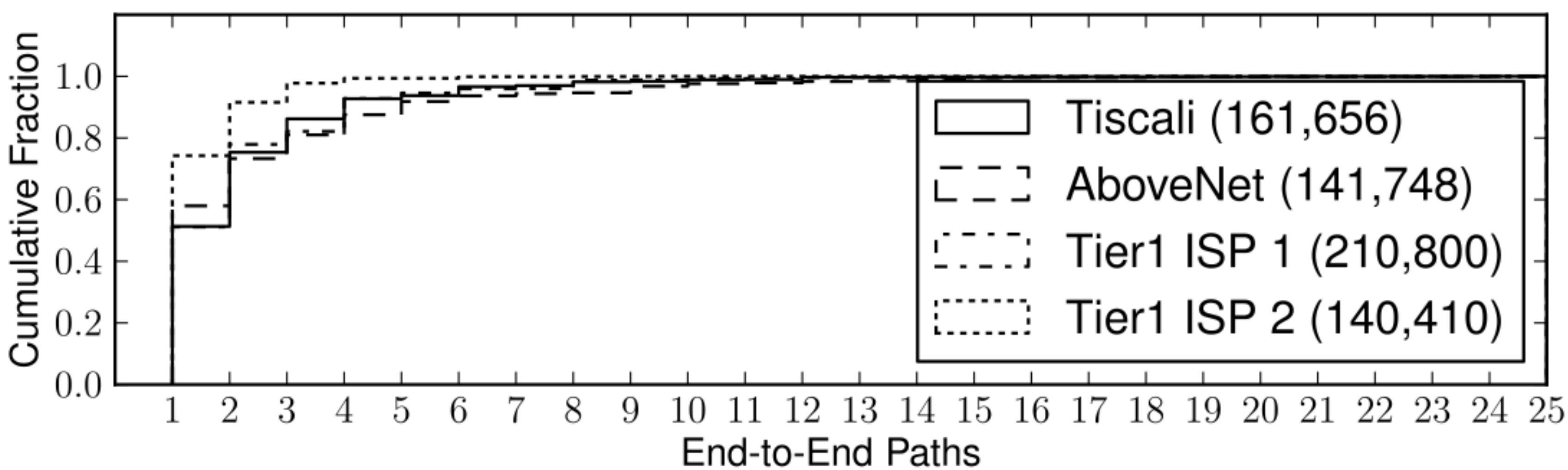
Packets from one TCP connection follow same path

Different connections follow different paths

G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure*, Revisiting Flow-Based Loa[d] Balancing: Stateless Path Selection in Data Center Networks*, Computer Networks, April 2013

How prevalent is ECMP ?

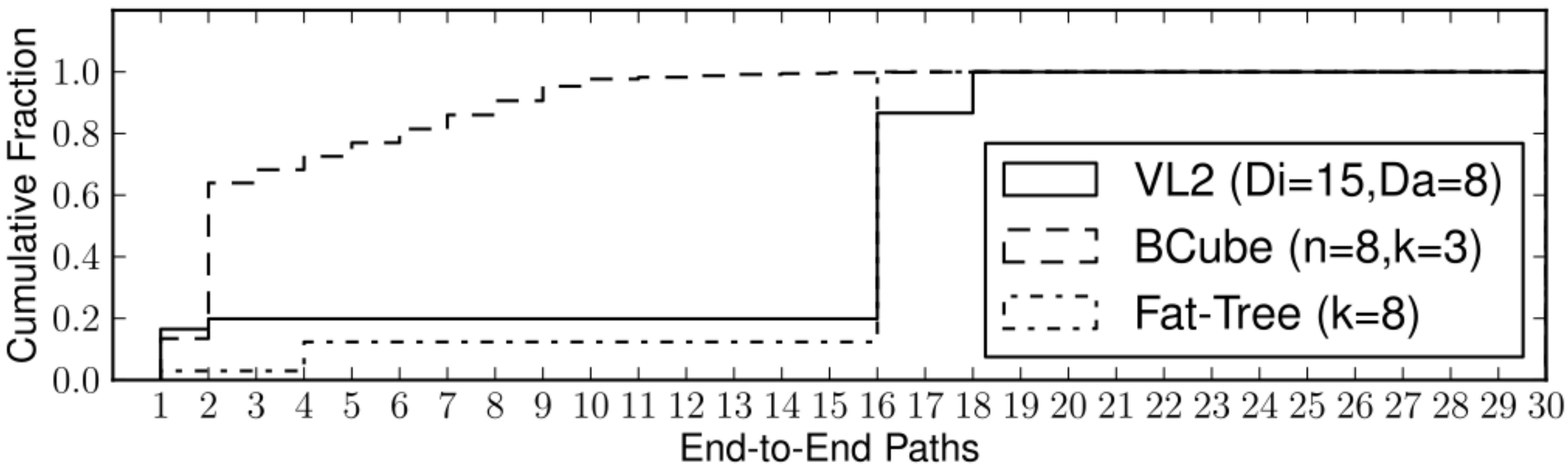
- Analysis of ISP network topologies



G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure, *Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, to appear in Computer Networks

Datacenters

ECMP in datacenters

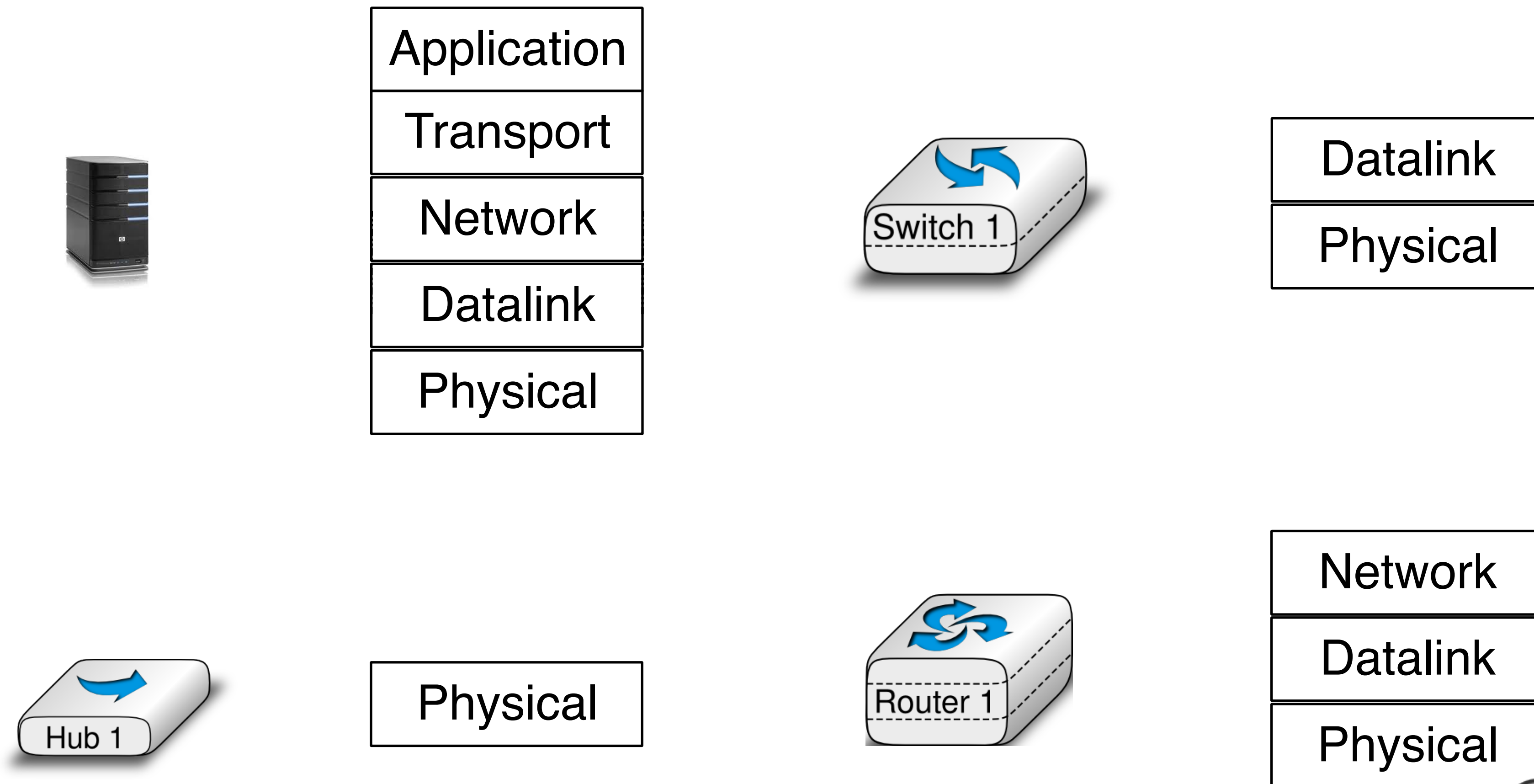


G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure*, Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, Computer Networks, April 2013

Agenda

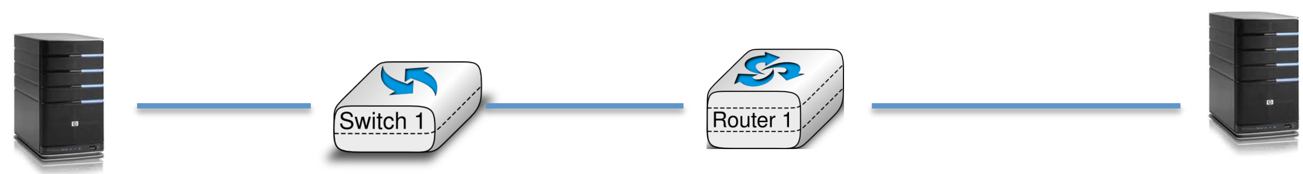
- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases

The Internet architecture that we explain to our students



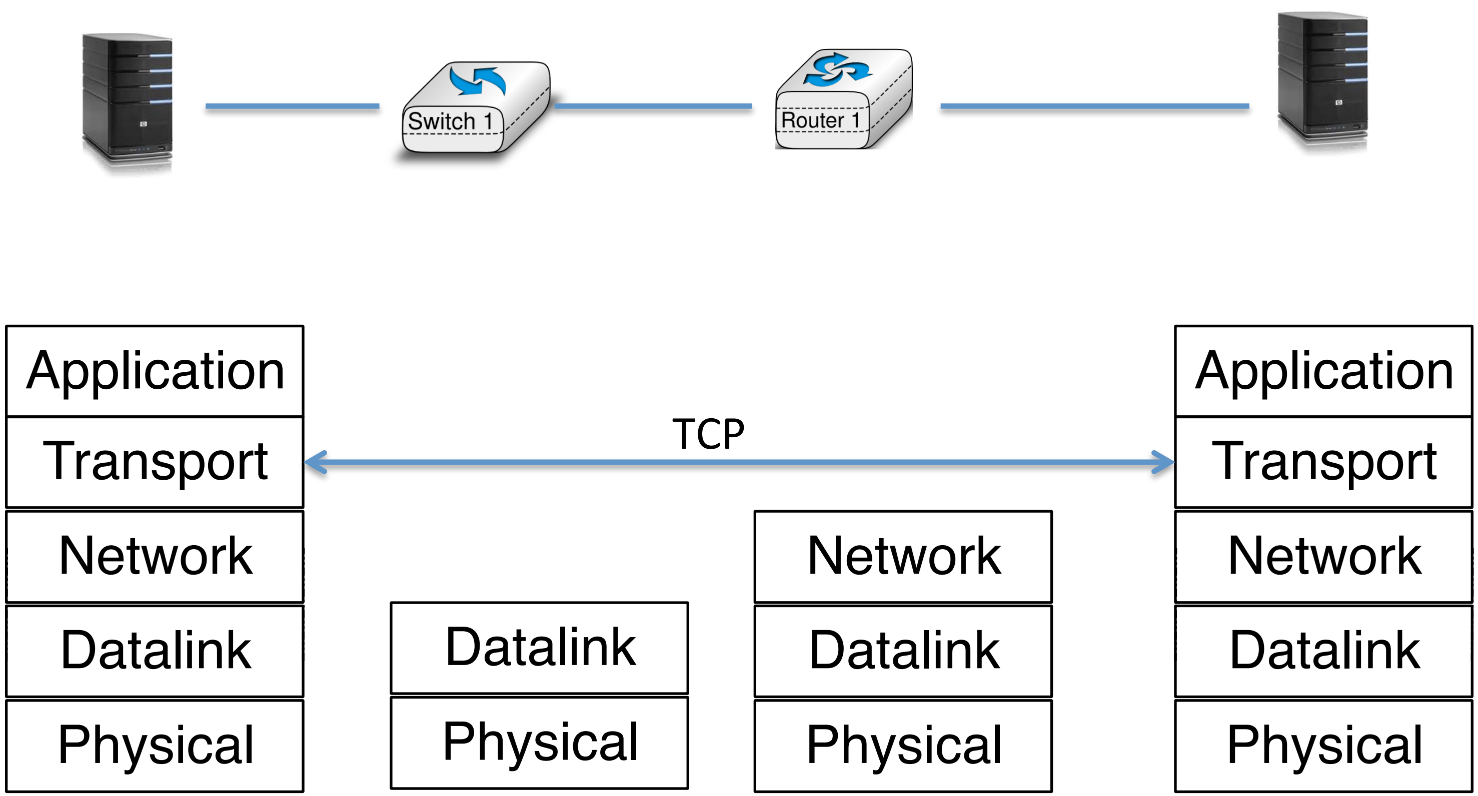
O. Bonaventure, Computer networking : Principles, Protocols and Practice, open ebook, http://inl.info.ucl.ac.be/

A typical "academic" network



Application			Application
Transport			Transport
Network		Network	Network
Datalink	Datalink	Datalink	Datalink
Physical	Physical	Physical	Physical

The end-to-end principle



In reality

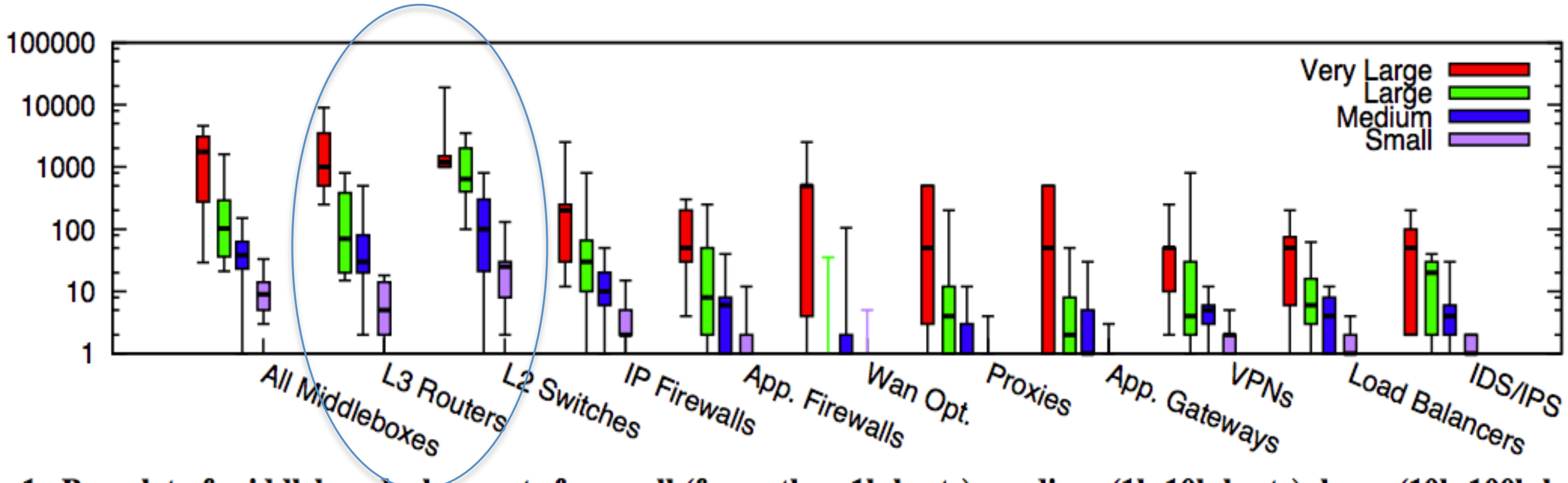
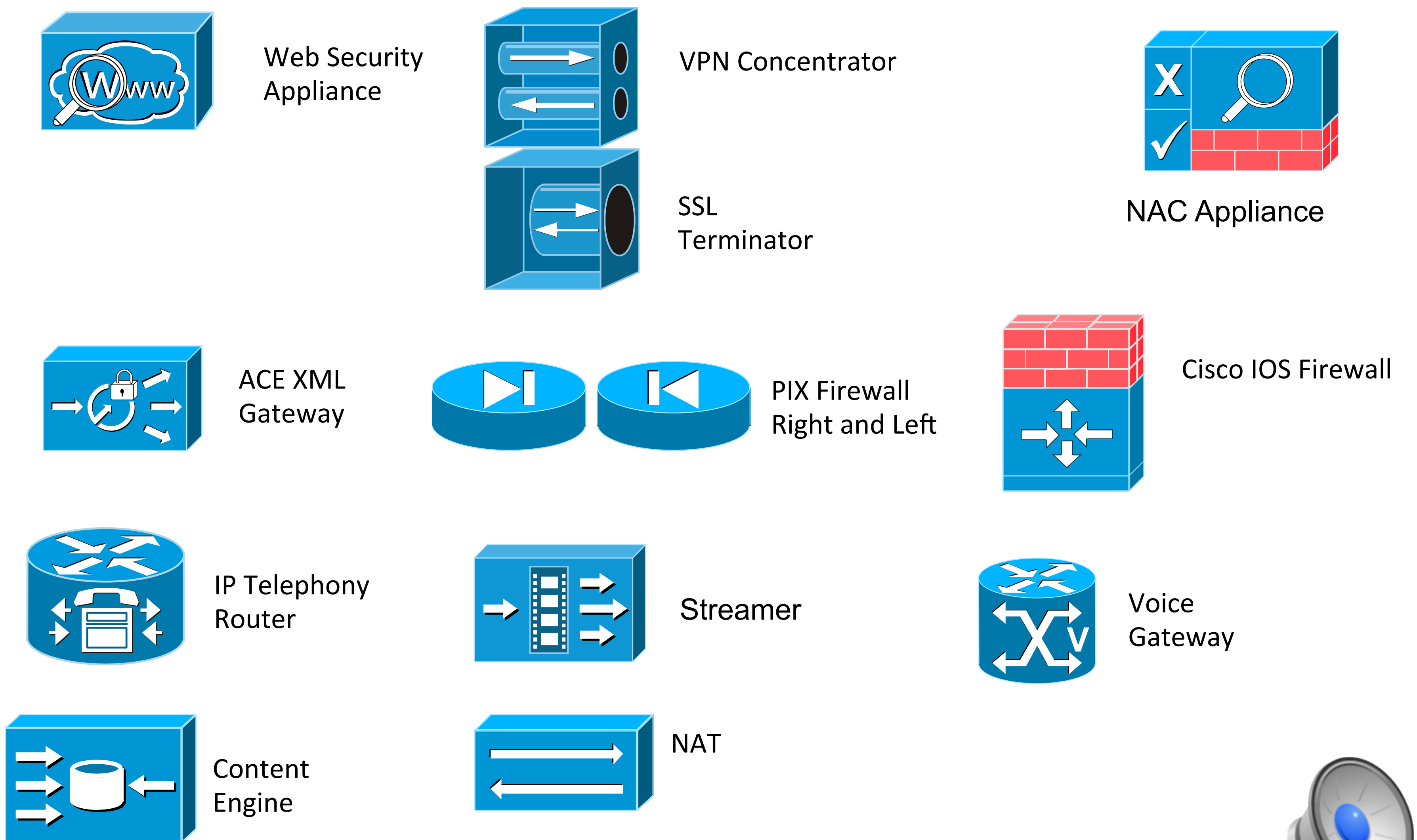


Figure 1: Box plot of middlebox deployments for small (fewer than 1k hosts), medium (1k-10k hosts), large (10k-100k hosts), and very large (more than 100k hosts) enterprise networks. Y-axis is in log scale.

- almost as many middleboxes as routers
- various types of middleboxes are deployed

Sherry, Justine, et al. "*Making middleboxes someone else's problem: Network processing as a cloud service.*" Proceedings of the ACM SIGCOMM 2012 conference. ACM, 2012.

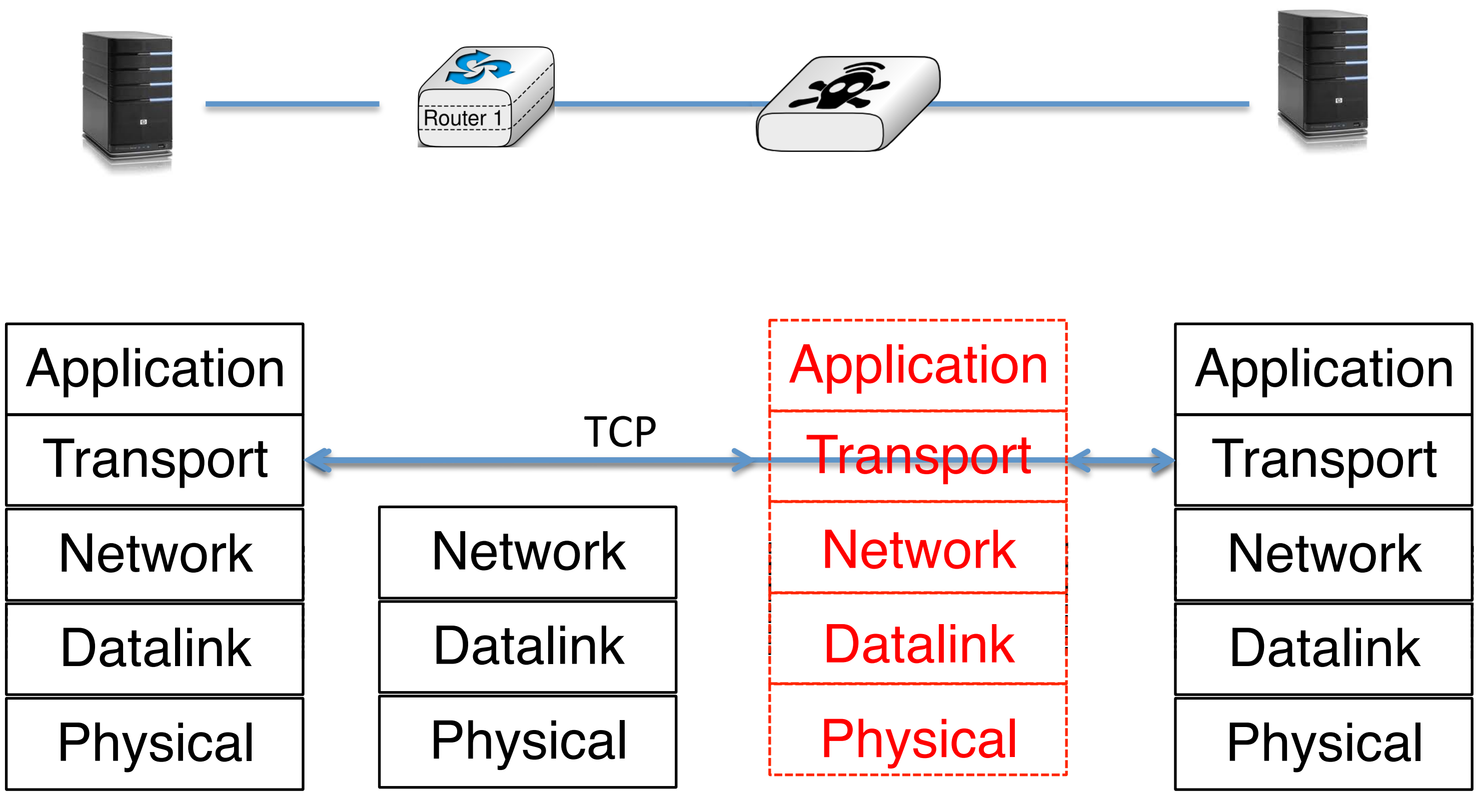
A middlebox zoo



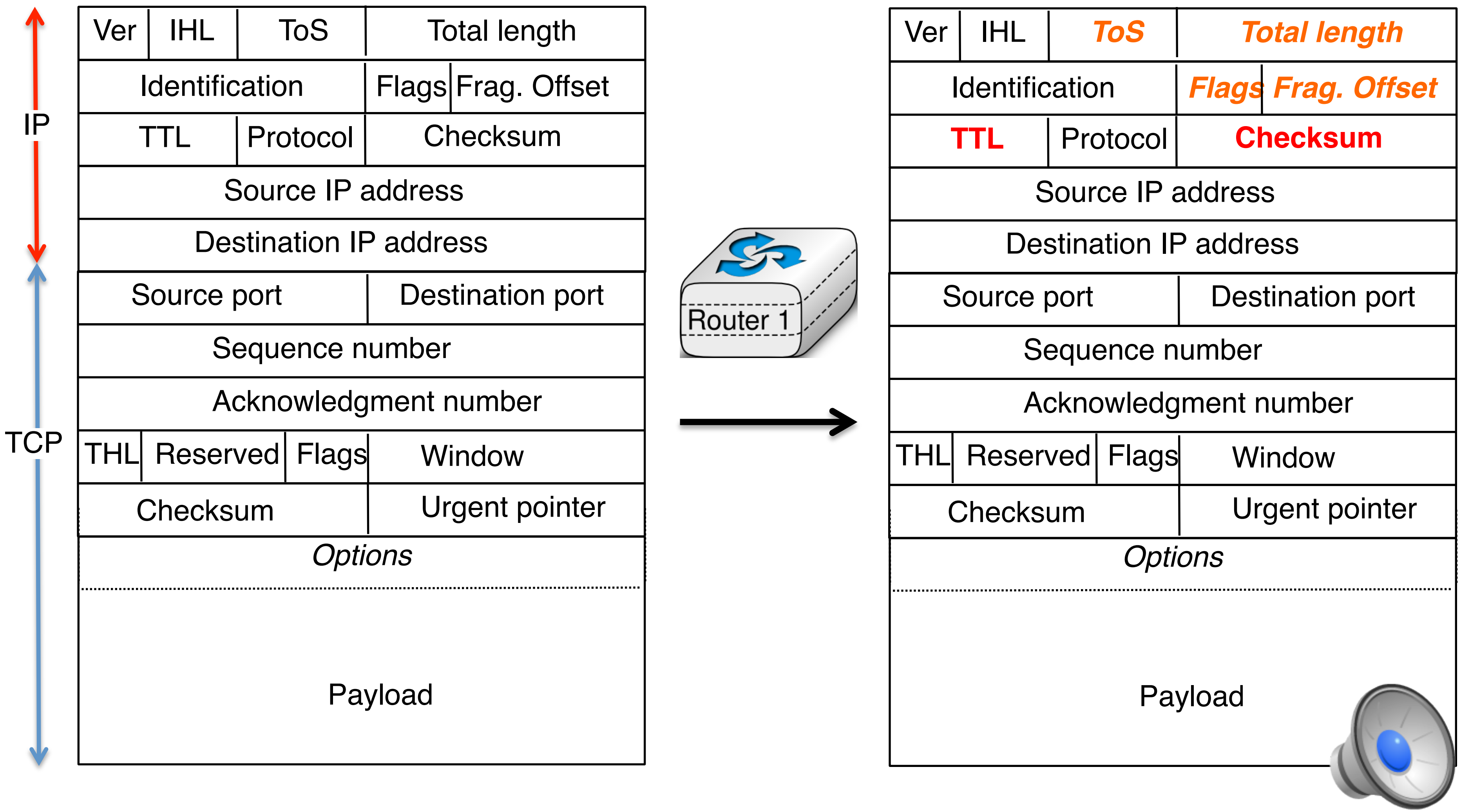
http://www.cisco.com/web/about/ac50/ac47/2.html

How to model those middleboxes ?

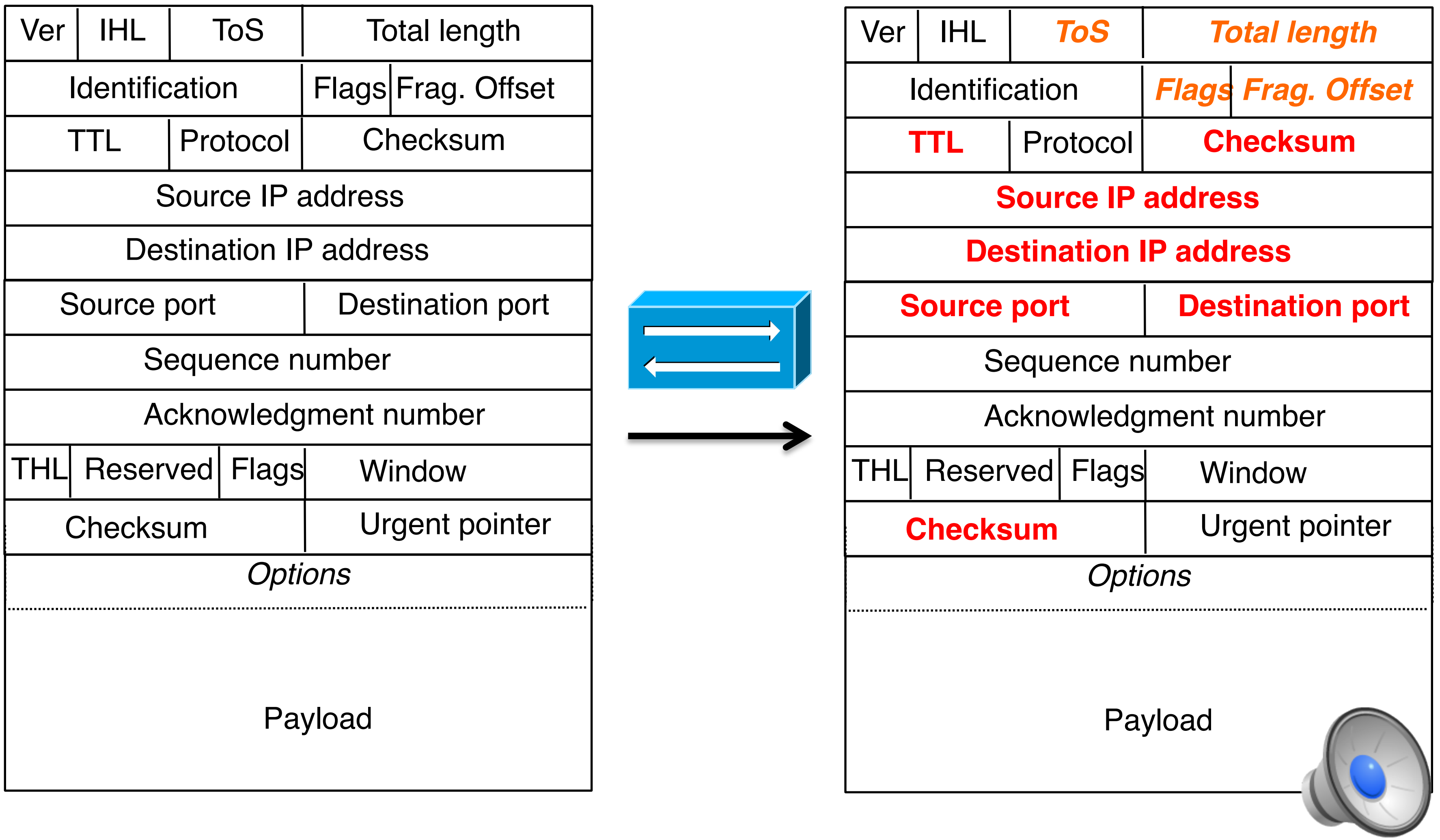
- In the official architecture, they do not exist
- In reality…



TCP segments processed by a router



TCP segments processed by a NAT

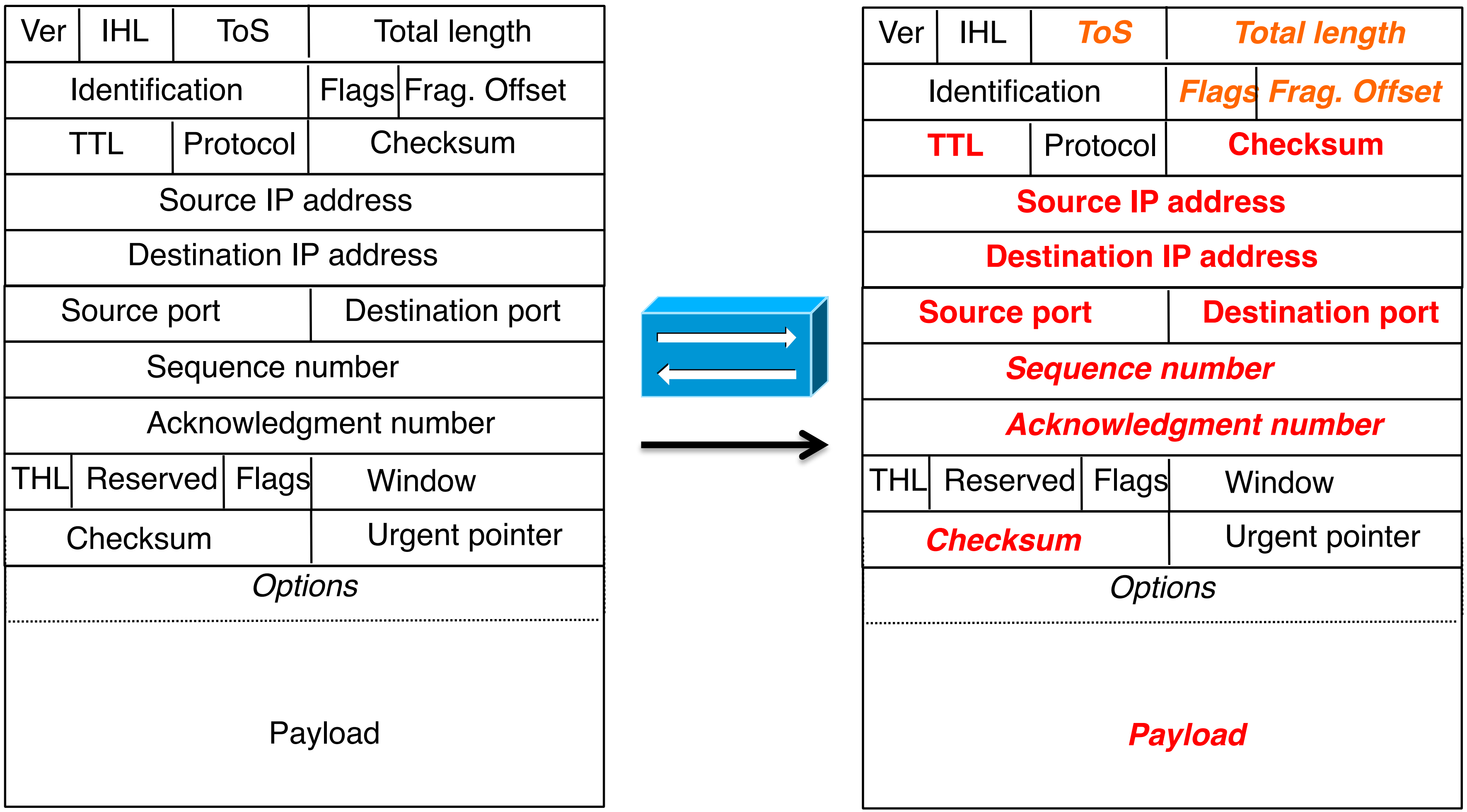


TCP segments processed by a NAT (2)

- active mode ftp behind a NAT

220 ProFTPD 1.3.3d Server (BELNET FTPD Server) [193.190.67.15]
ftp_login: user `<null>' pass `<null>' host `ftp.belnet.be'
Name (ftp.belnet.be:obo): anonymous
---> USER anonymous
331 Anonymous login ok, send your complete email address as your password
Password:
---> PASS XXXX
---> **PORT 192,168,0,7,195,120**
200 PORT command successful
---> LIST
150 Opening ASCII mode data connection for file list
lrw-r--r-- 1 ftp ftp 6 Jun 1 2011 pub -> mirror
226 Transfer complete

TCP segments processed by an ALG running on a NAT



How transparent is the Internet ?

- 25th September 2010 to 30th April 2011
- 142 access networks
- 24 countries
- Sent specific TCP segments from client to a server in Japan

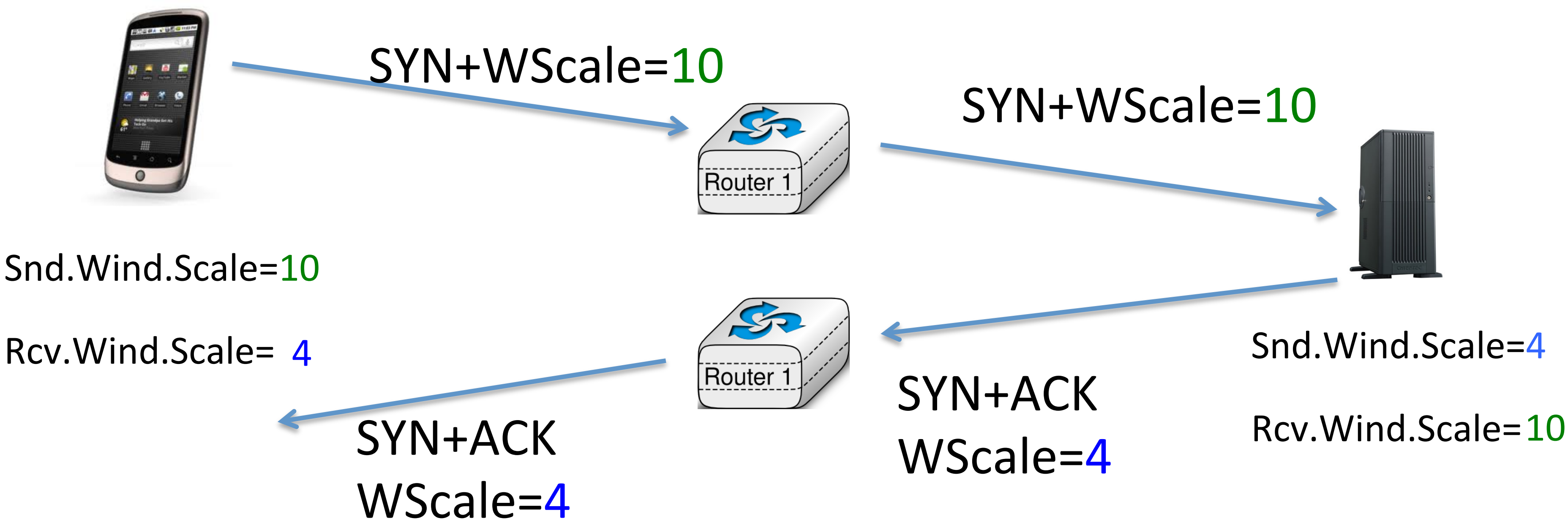
Table 2: Experiment Venues

Country	*Home*	*Hotspot*	*Cellular*	*Univ*	*Ent*	*Hosting*	Total
Australia	0	2	0	0	0	1	3
Austria	0	0	0	0	1	0	1
Belgium	4	0	0	1	0	0	5
Canada	1	0	1	0	1	0	3
Chile	0	0	0	0	1	0	1
China	0	7	0	0	0	0	7
Czech	0	2	0	0	0	0	2
Denmark	0	2	0	0	0	0	2
Finland	1	0	0	3	2	0	6
Germany	3	1	3	4	1	0	12
Greece	2	0	1	0	0	0	3
Indonesia	0	0	0	3	0	0	3
Ireland	0	0	0	0	0	1	1
Italy	1	0	0	0	1	0	2
Japan	19	10	7	3	2	0	41
Romania	1	0	0	0	0	0	1
Russia	0	1	0	0	0	0	1
Spain	0	1	0	1	0	0	2
Sweden	1	0	0	0	0	0	1
Switzerland	2	0	0	0	0	0	2
Thailand	0	0	0	0	2	0	2
U.K.	10	4	4	2	1	1	22
U.S.	3	4	4	0	4	2	17
Vietnam	1	0	0	0	1	0	2
Total	49	34	20	17	17	5	142

Honda, Michio, et al. "*Is it still possible to extend TCP?*" Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

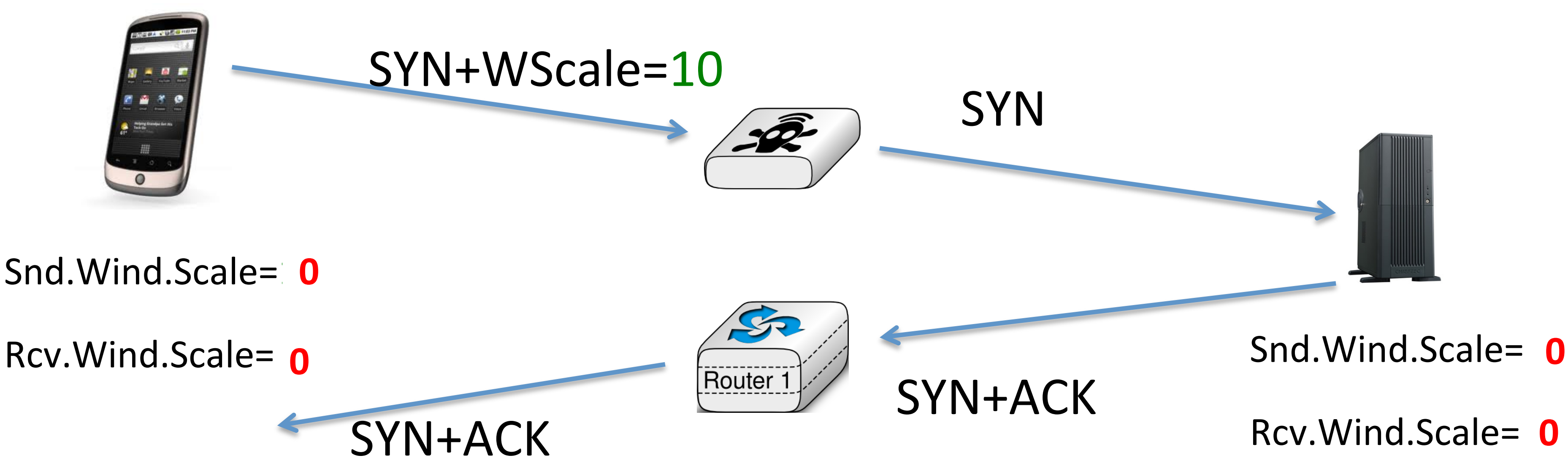
How to extend TCP ? RFC1323

- Large window extension
  - supports >64KBytes windows by shifting window field in TCP segment header



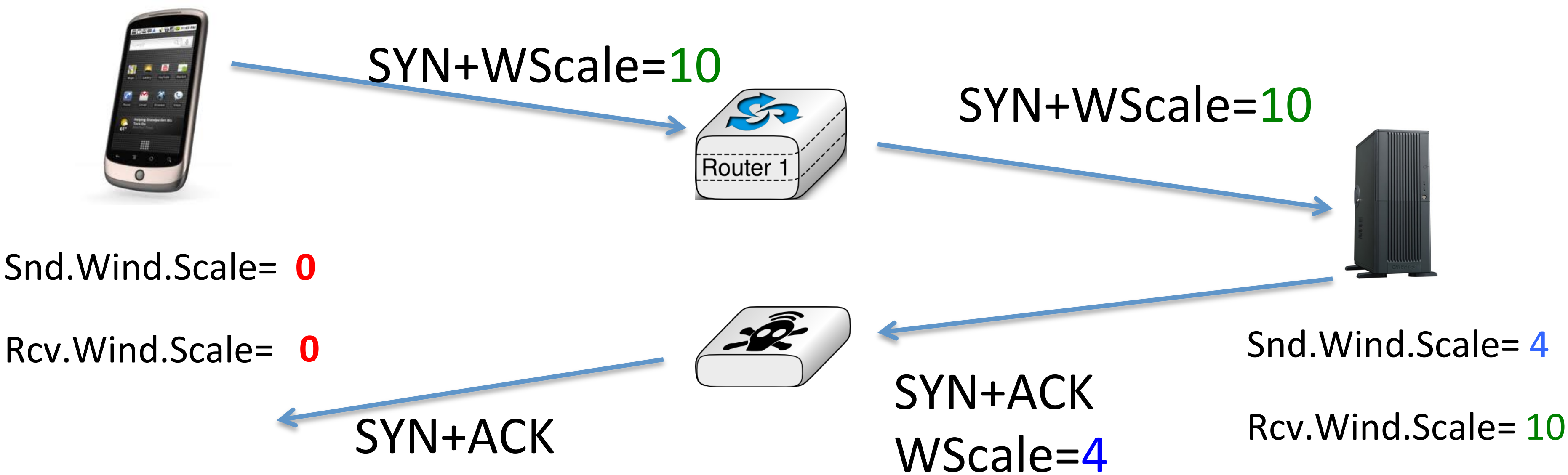
How to extend TCP ? RFC1323

- What happens with middleboxes ?

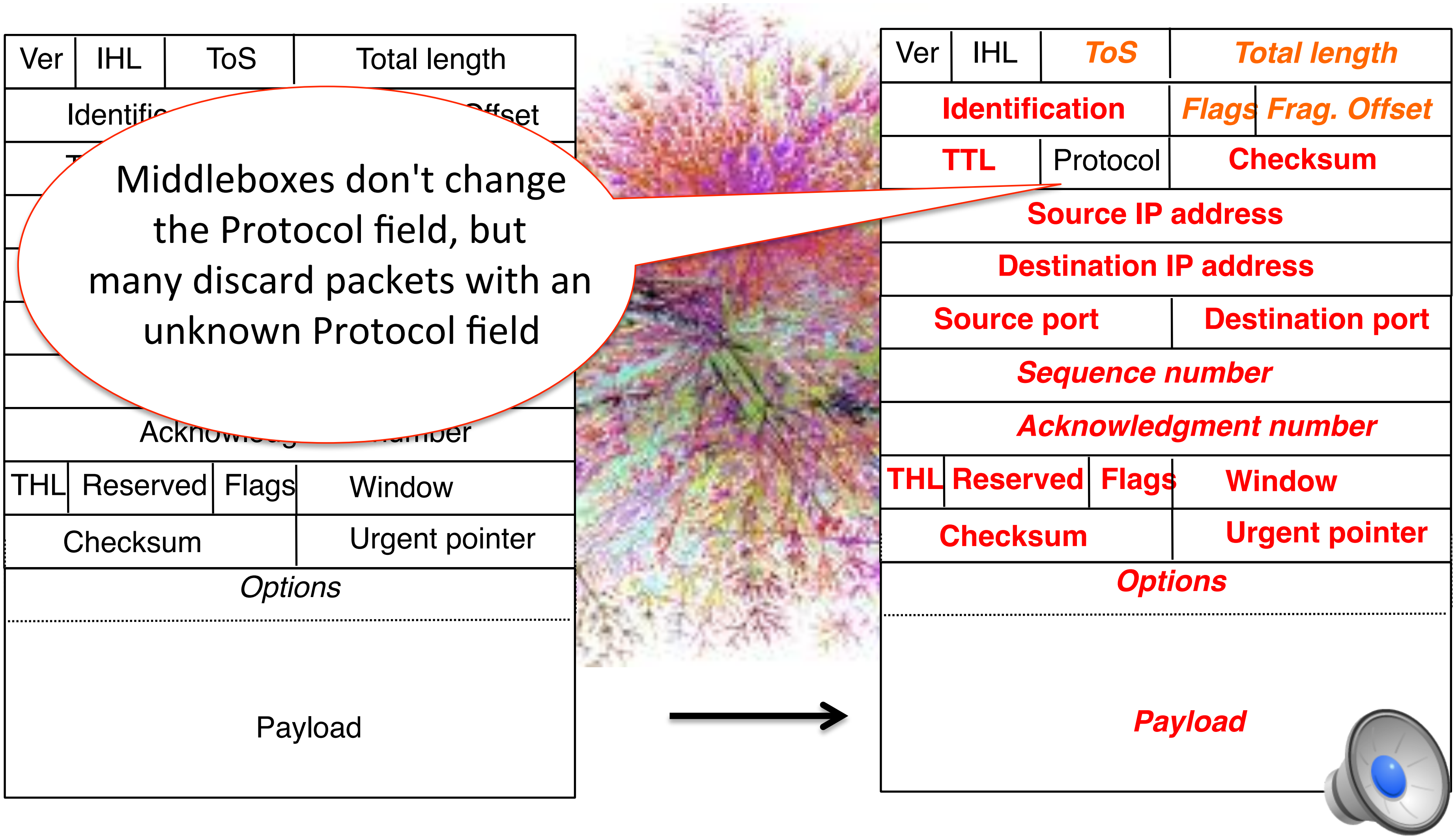


How to extend TCP ?
RFC1323

- What happens with middleboxes ?



End-to-end transparency today



Agenda

- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases

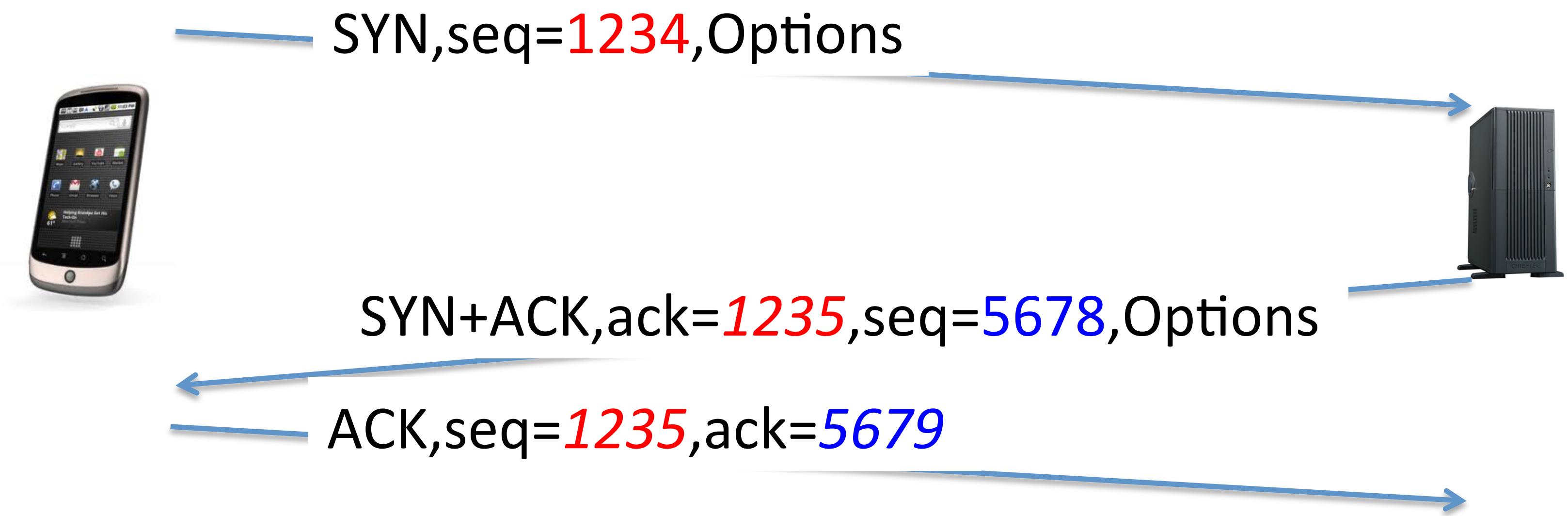
Design objectives

- Multipath TCP is an *evolution* of TCP

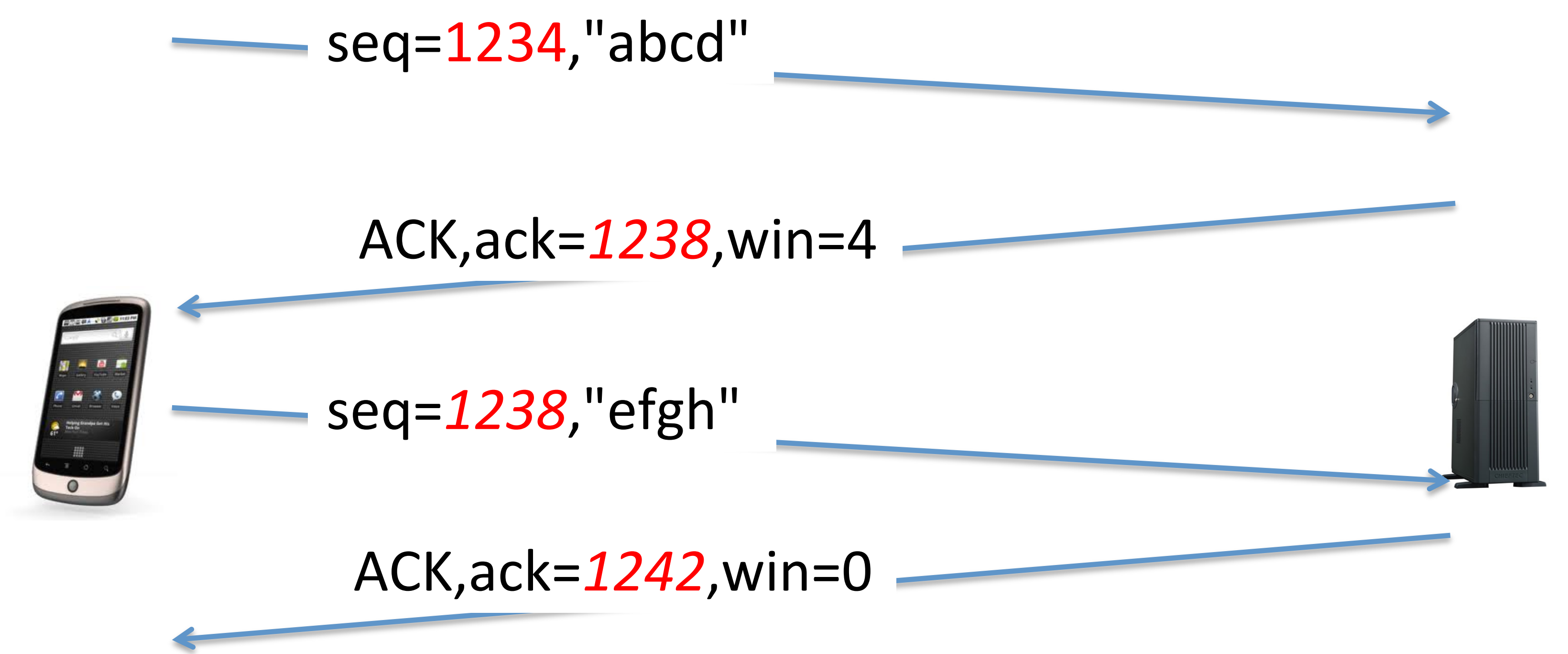
- Design objectives
  - Support unmodified applications
  - Work over today's networks (IPv4 and IPv6)
  - Works in all networks where regular TCP works

TCP Connection establishment

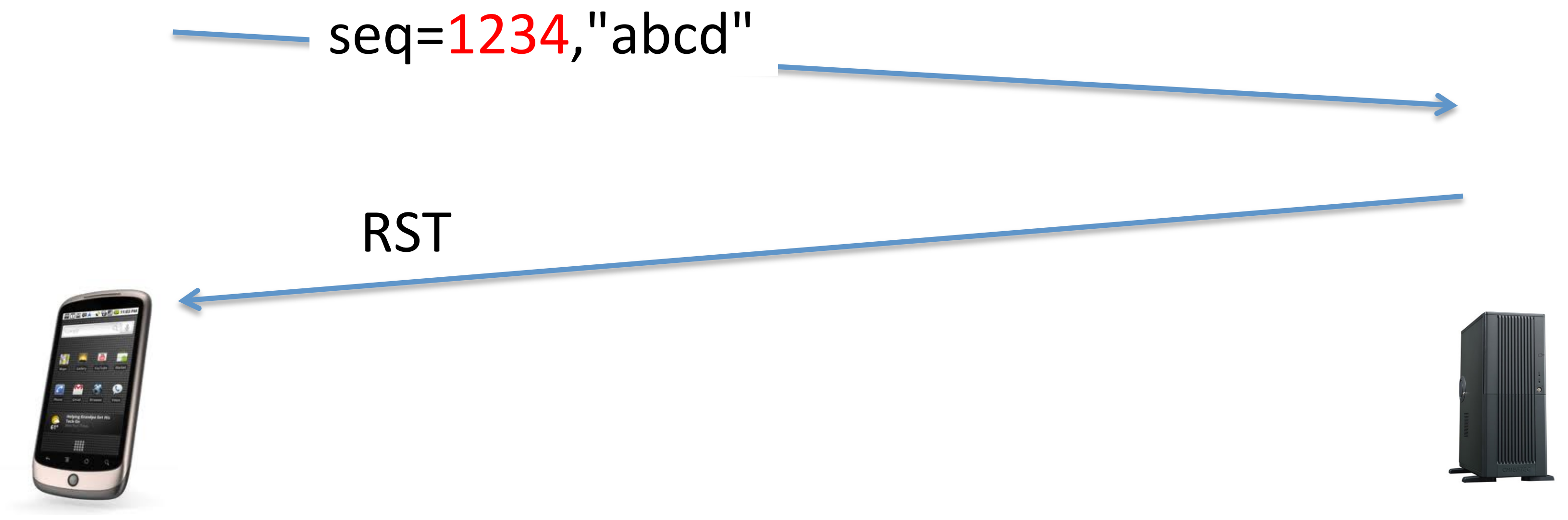
- Three-way handshake



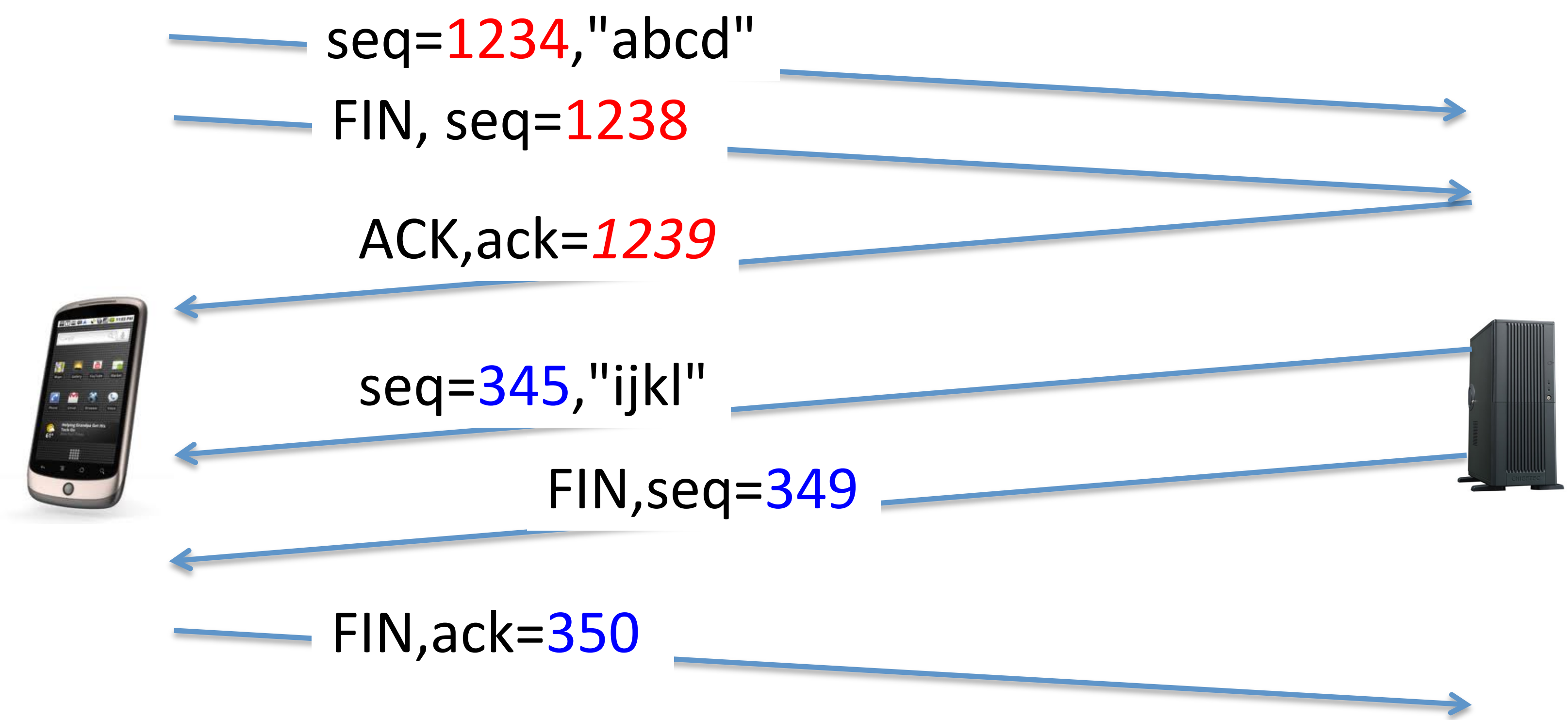
Data transfer



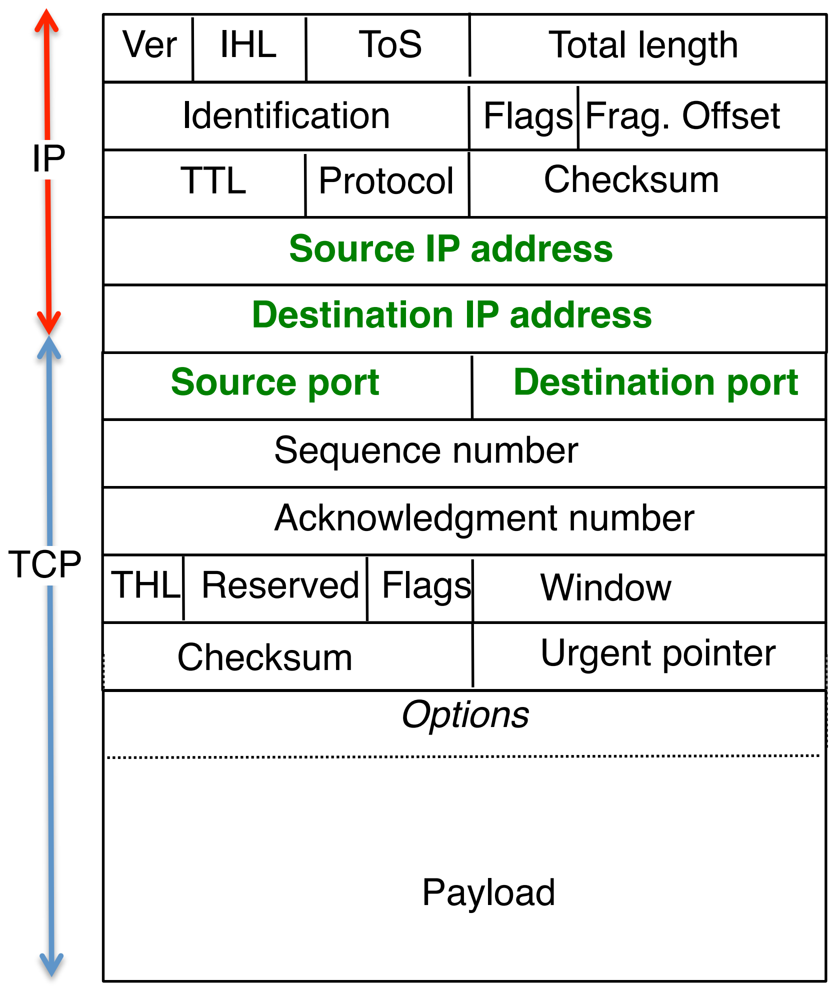
Connection release



Connection release



Identification of a TCP connection

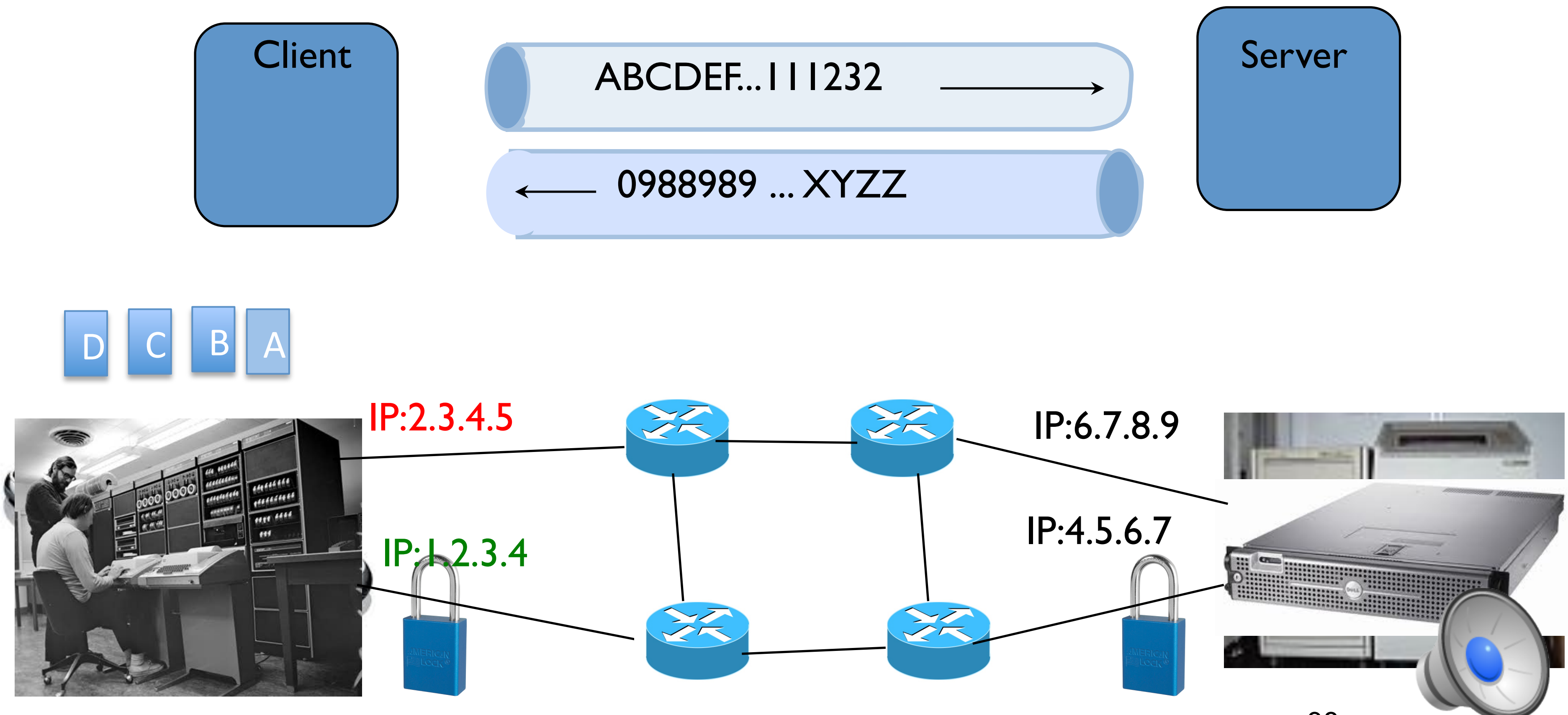


Four tuple

- IP_{source}
- IP_{dest}
- $Port_{source}$
- $Port_{dest}$

All TCP segments contain the four tuple

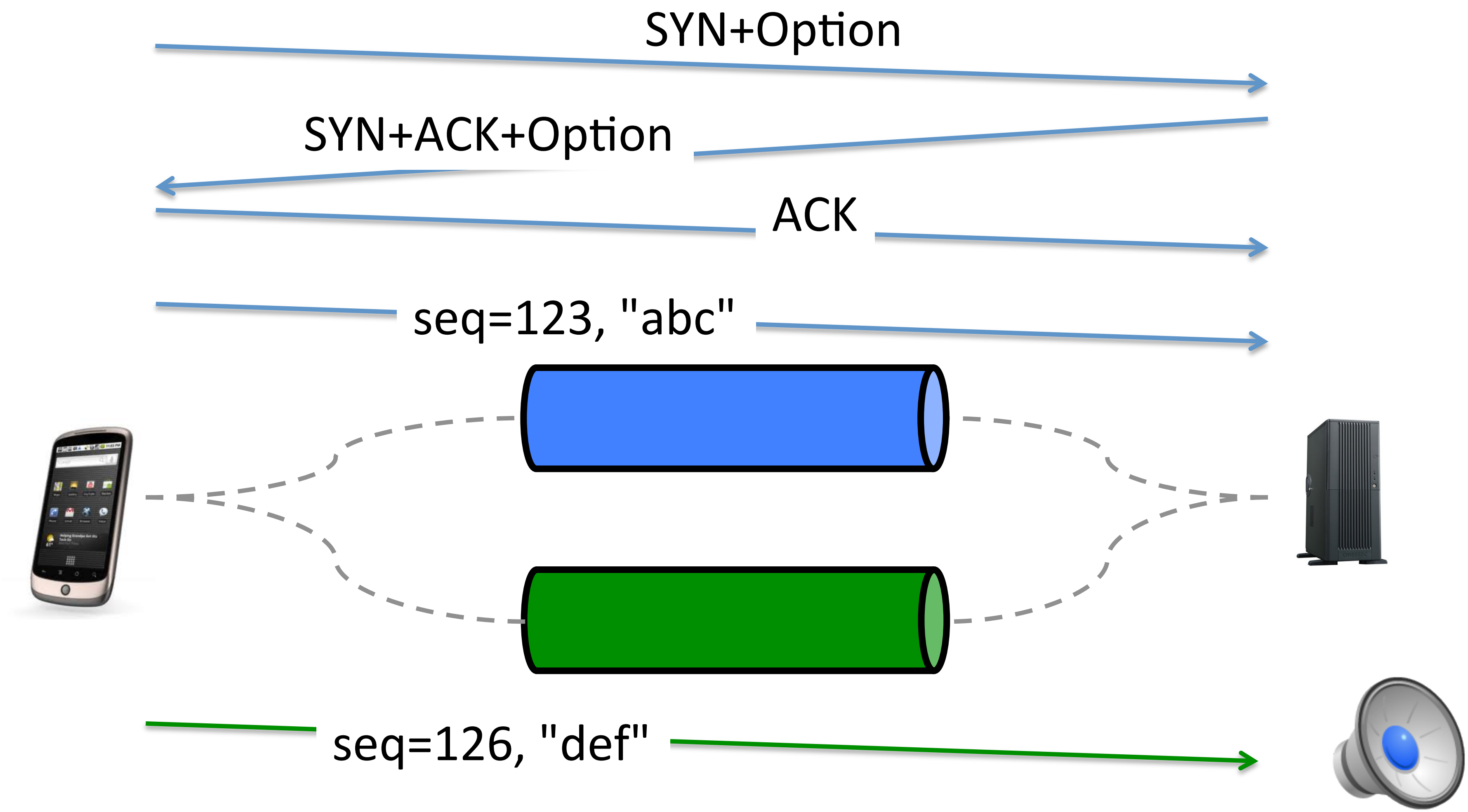
The *new* bytestream model



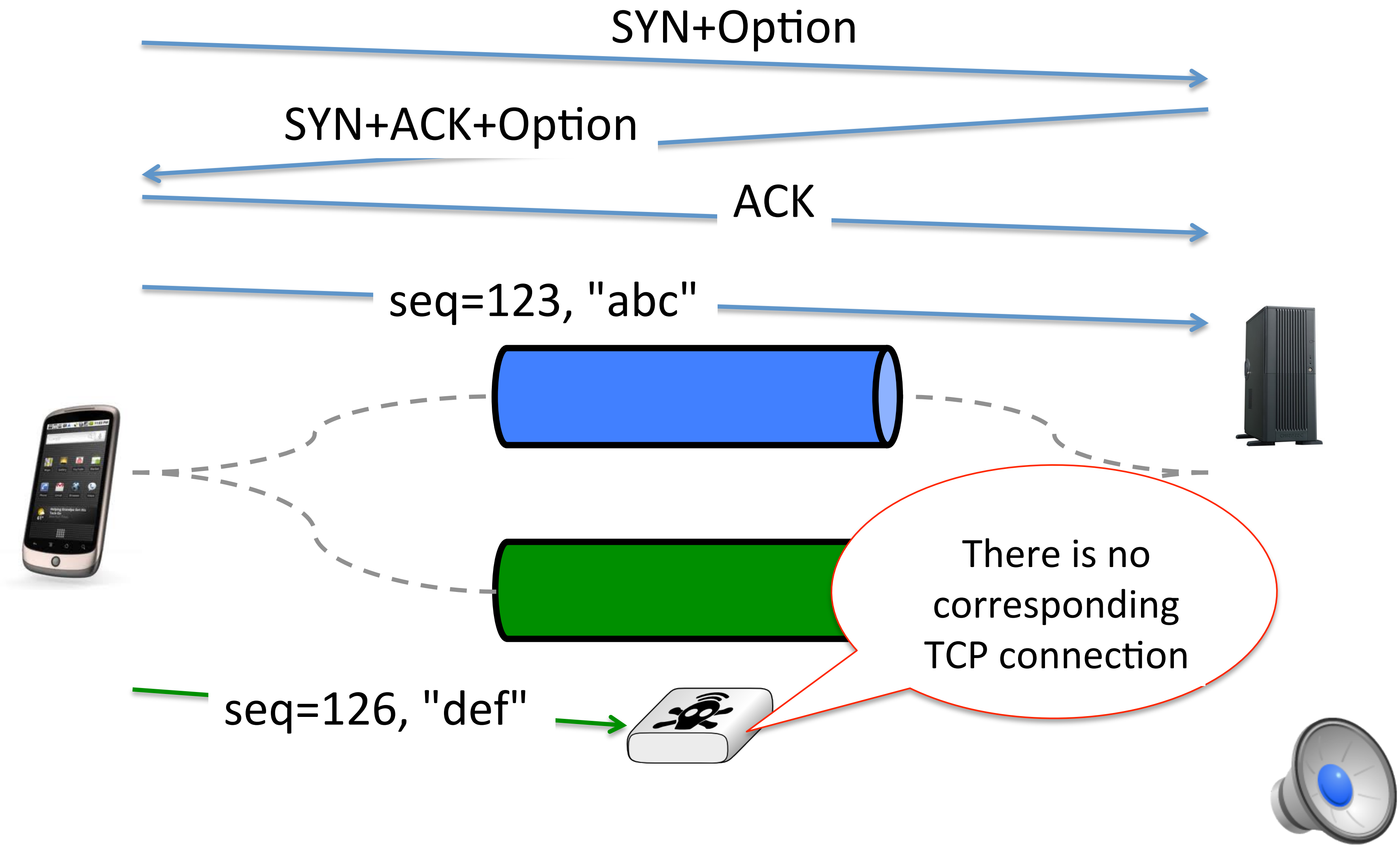
The Multipath TCP protocol

- Control plane
  - How to manage a Multipath TCP connection that uses several paths ?
- Data plane
  - How to transport data ?
- Congestion control
  - How to control congestion over multiple paths ?

A naïve Multipath TCP



A naïve Multipath TCP
In today's Internet ?

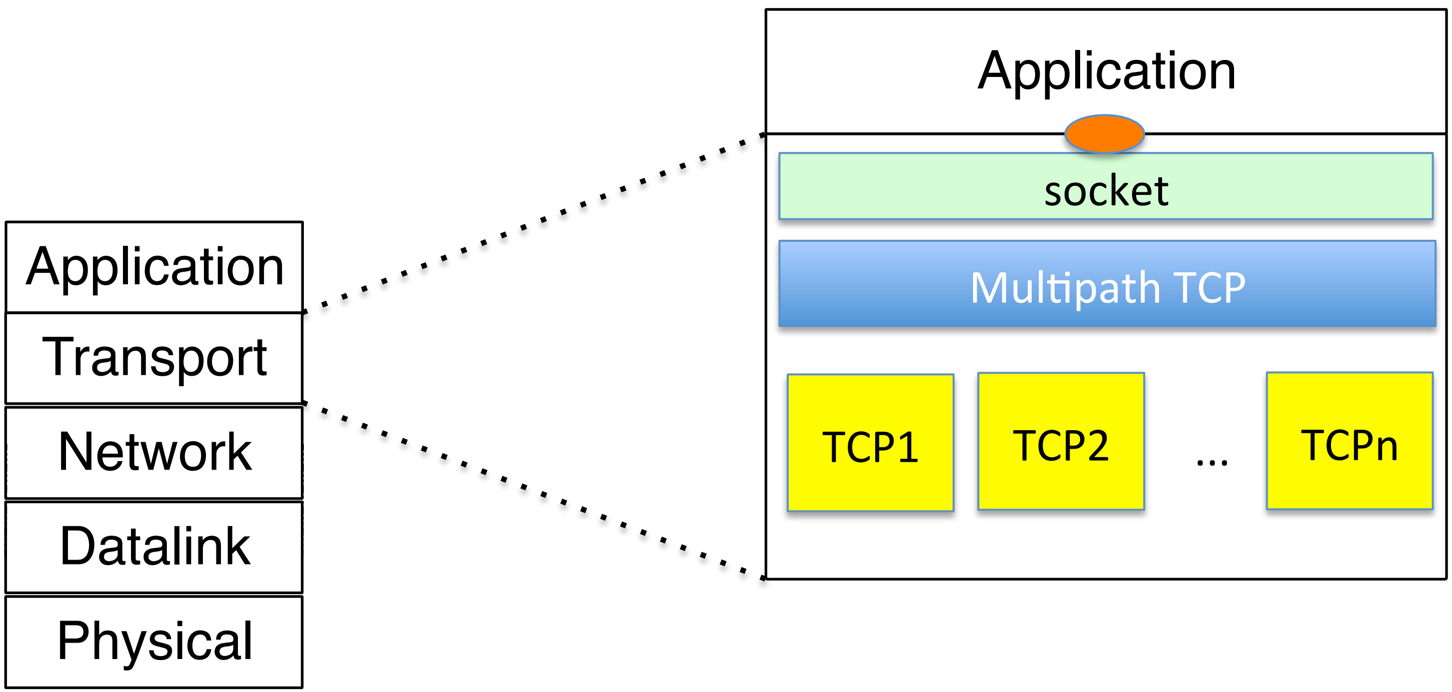


Design decision

– *A Multipath TCP connection is composed of one or more regular TCP subflows that are combined*

- Each host maintains state that glues the TCP subflows that compose a Multipath TCP connection together
- Each TCP subflow is sent over a single path and appears like a **regular TCP** connection along this path

Multipath TCP and the architecture

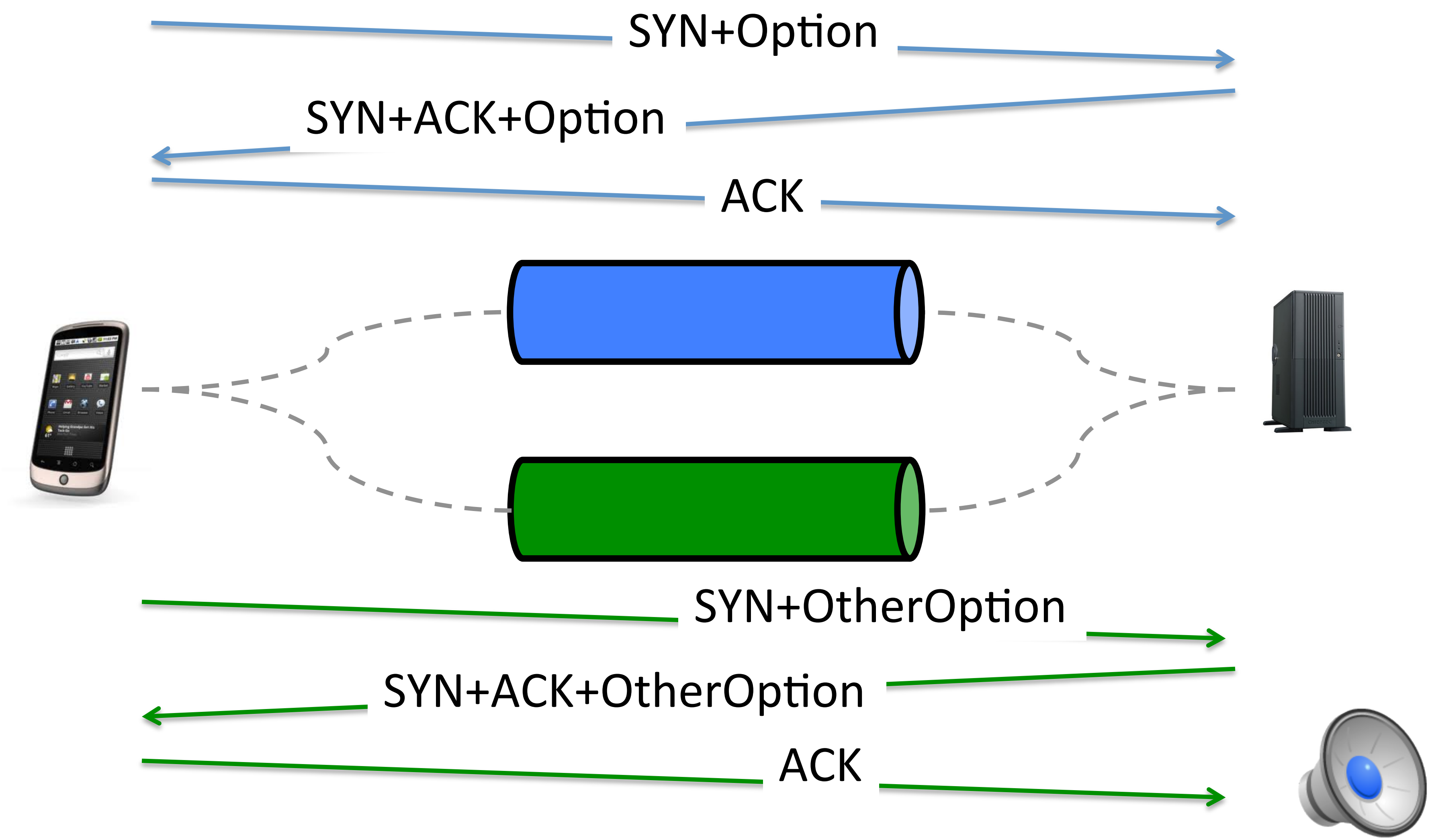


A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, “Architectural guidelines for multipath TCP development", RFC6182 2011.

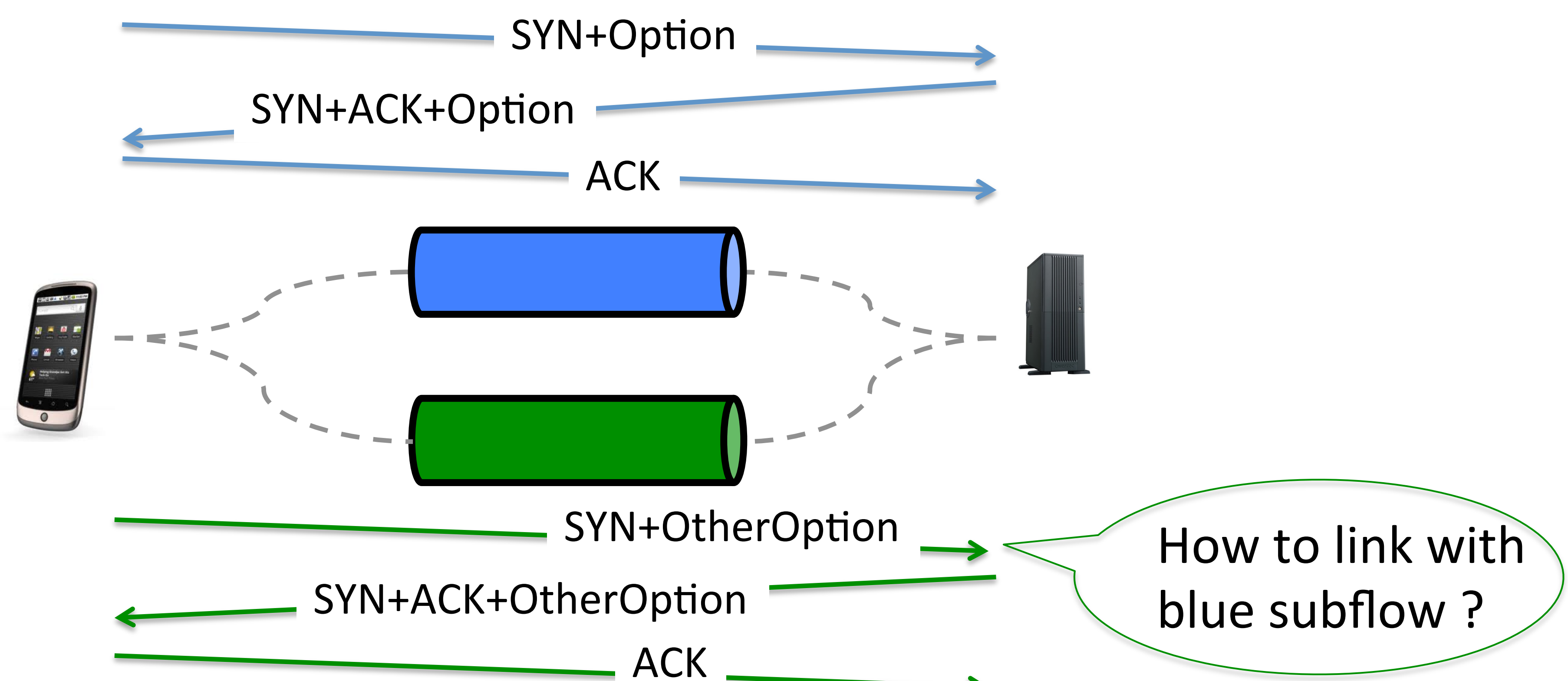
A *regular* TCP connection

- What is a *regular* TCP connection ?
  - It starts with a three-way handshake
    - SYN segments may contain special options
  - All data segments are sent in sequence
    - There is no gap in the sequence numbers
  - It is terminated by using FIN or RST

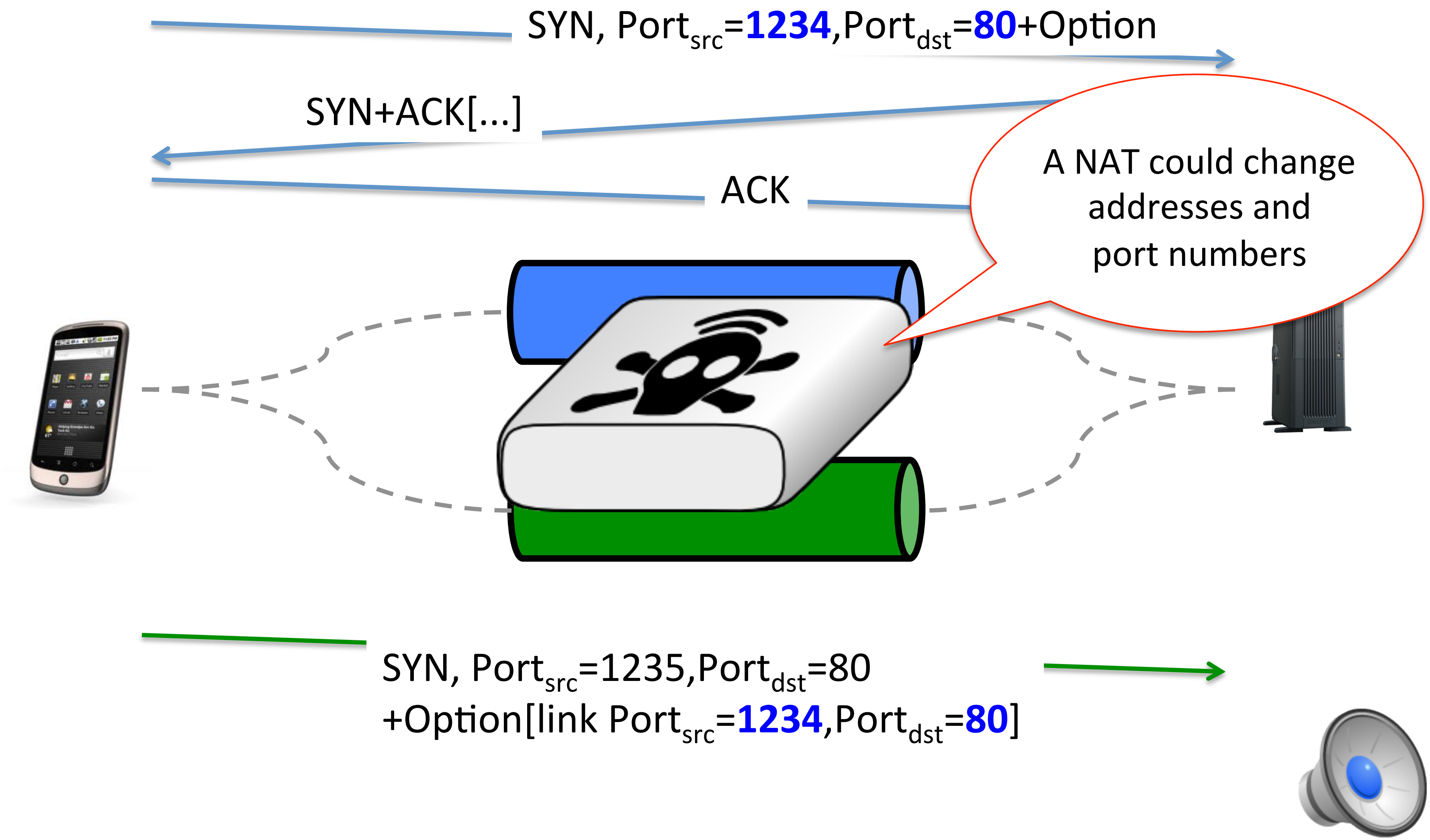
Multipath TCP



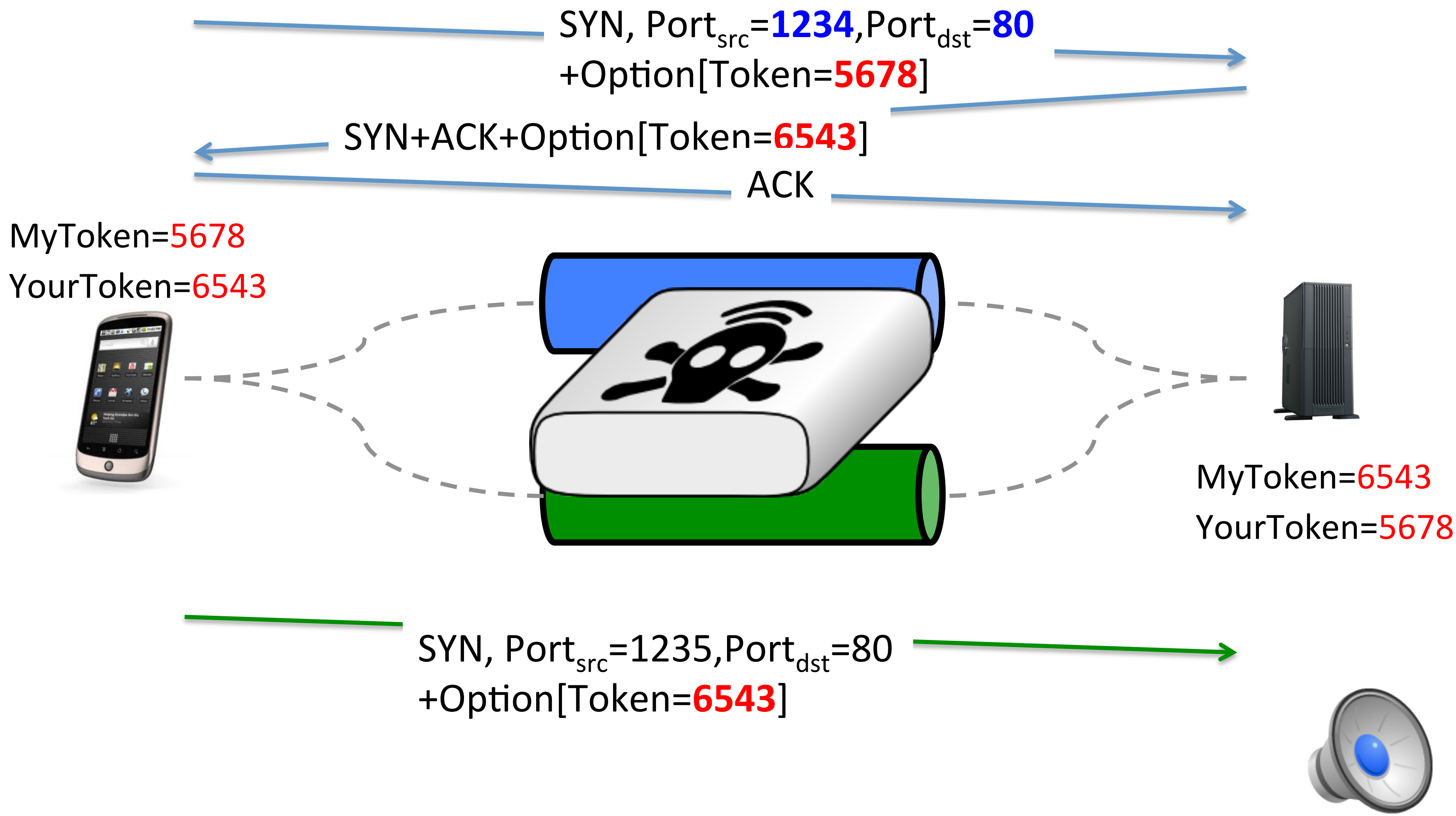
How to combine two TCP subflows ?



How to link TCP subflows ?



How to link TCP subflows ?



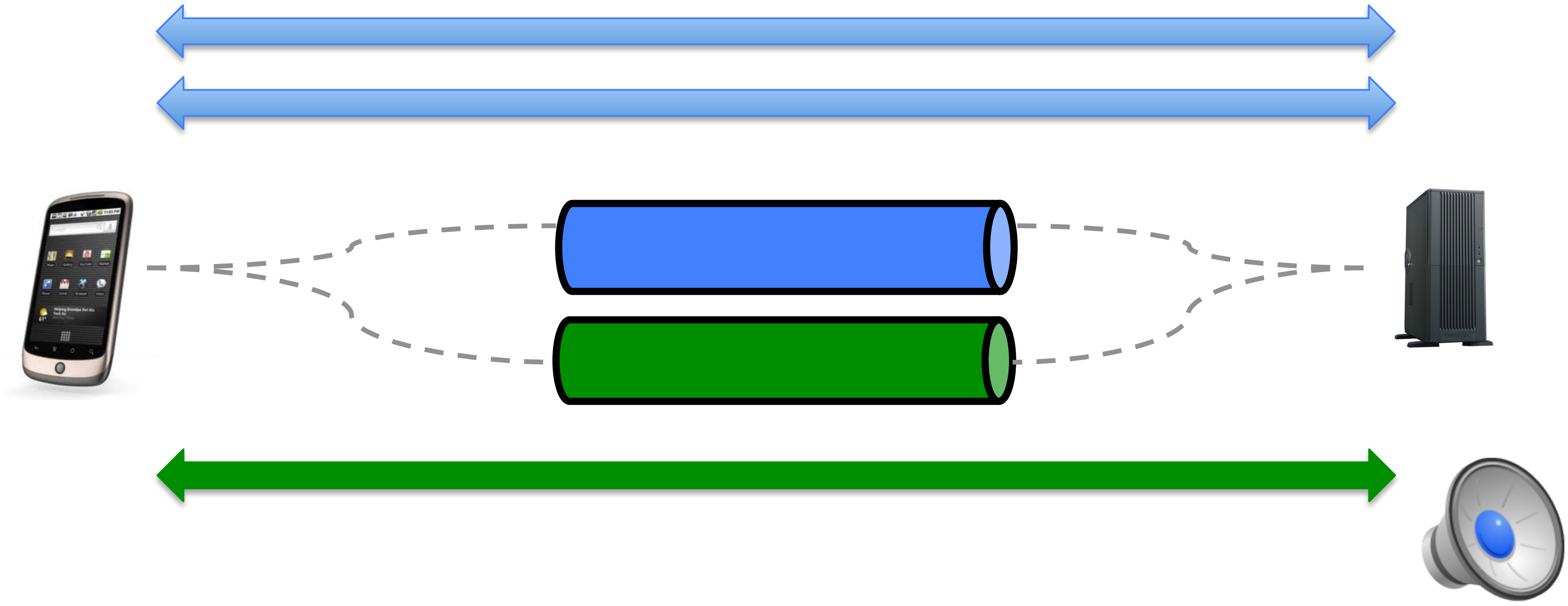
TCP subflows

- Which subflows can be associated to a Multipath TCP connection ?

  – At least one of the elements of the four-tuple needs to differ between two subflows
    - Local IP address
    - Remote IP address
    - Local port
    - Remote port

Subflow agility

- Multipath TCP supports
  - addition of subflows
  - removal of subflows

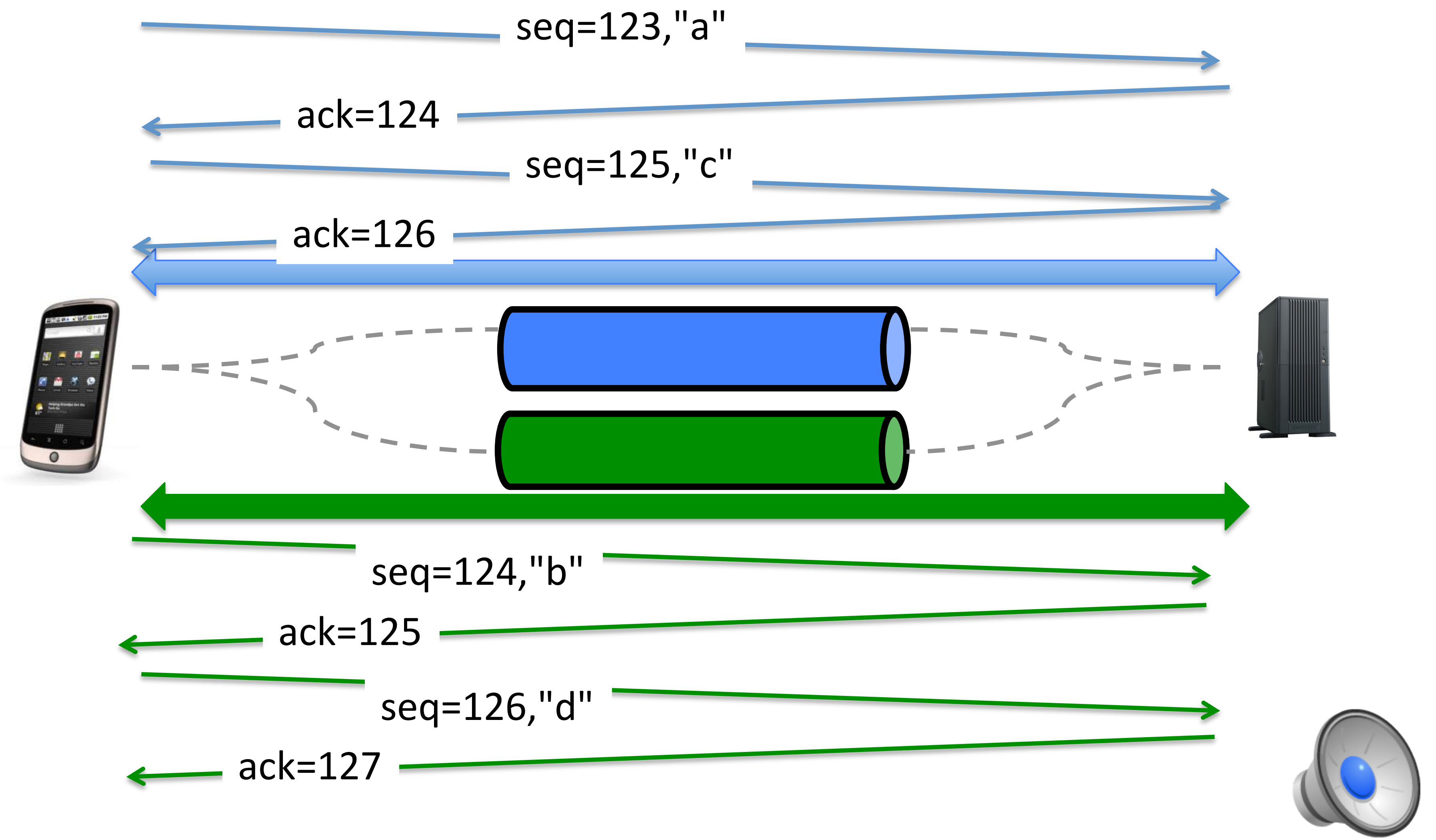


The Multipath TCP protocol

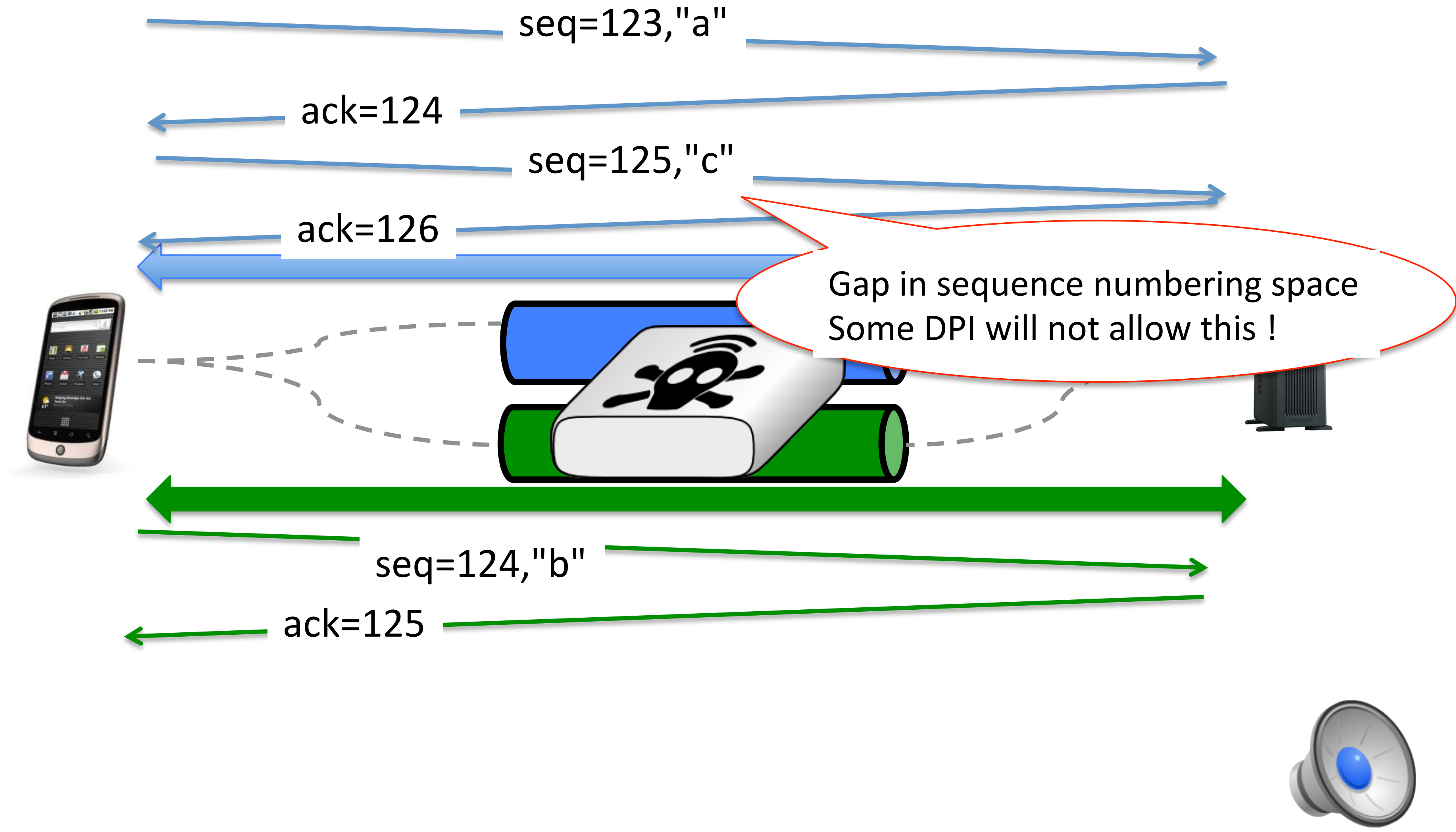
- Control plane
  - How to manage a Multipath TCP connection that uses several paths ?
- **Data plane**
  - How to transport data ?

- Congestion control
  - How to control congestion over multiple paths ?

How to transfer data ?

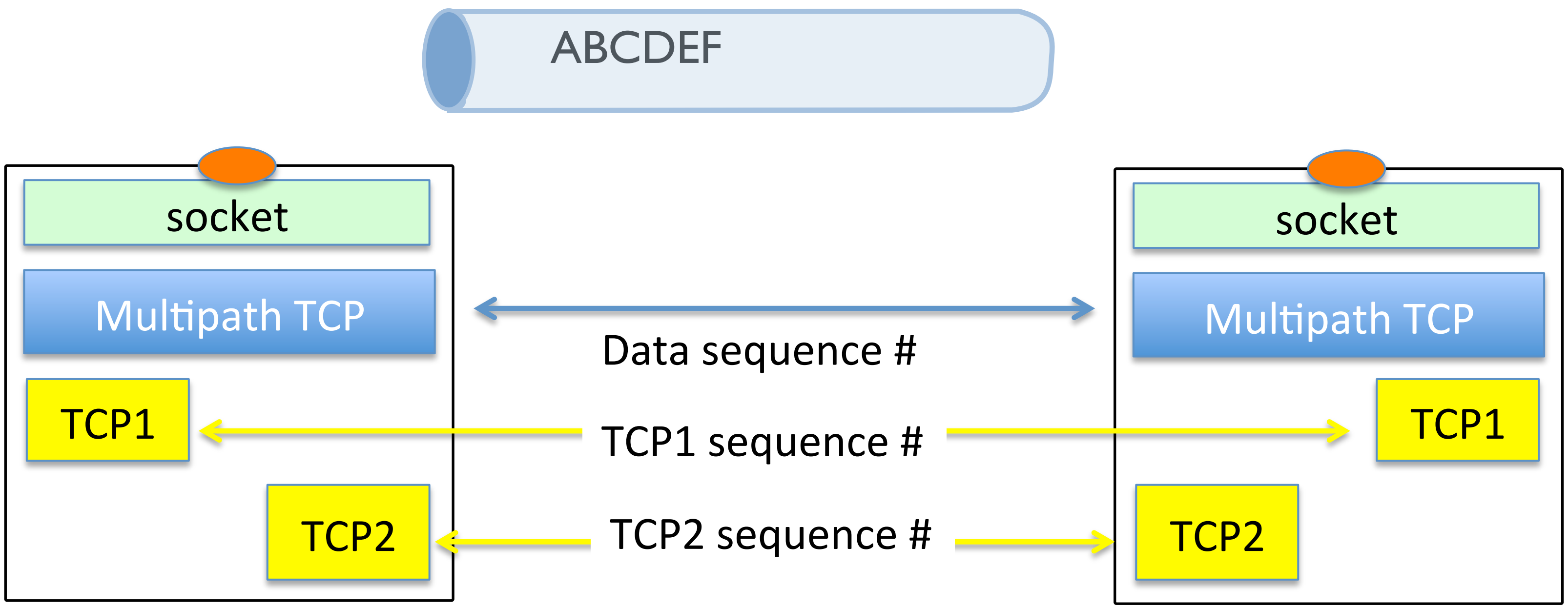


How to transfer data in today's Internet ?

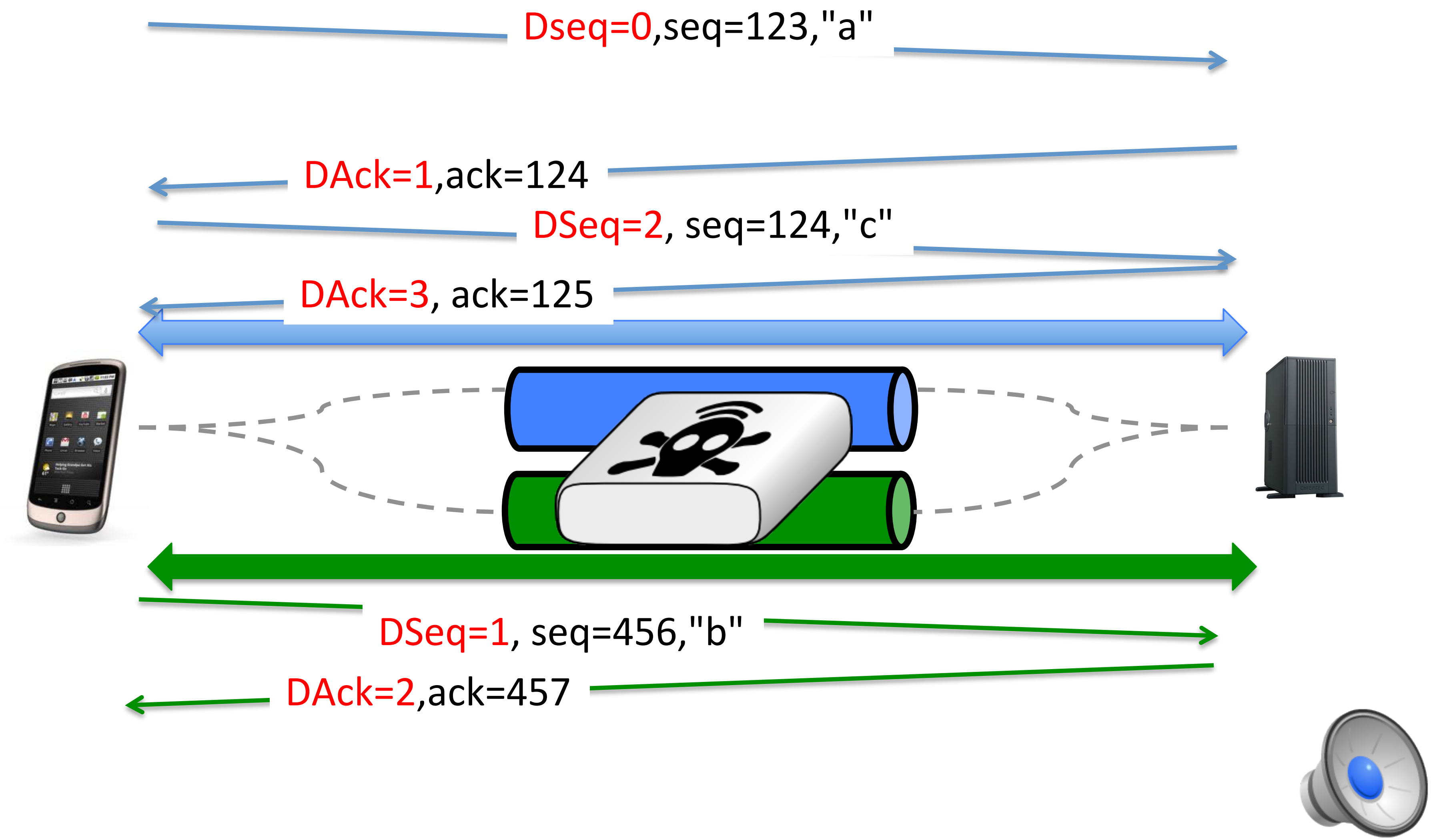


Multipath TCP Data transfer

- Two levels of sequence numbers

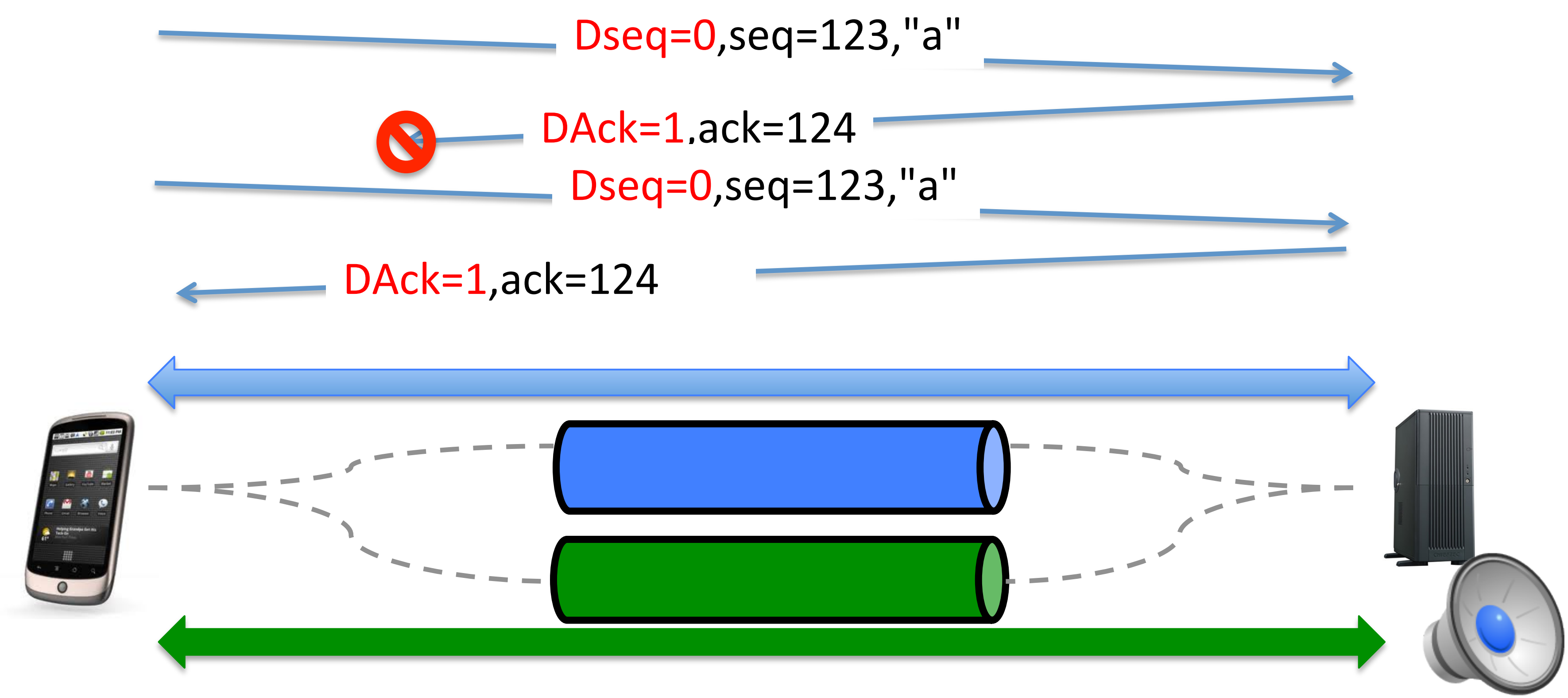


Multipath TCP
Data transfer



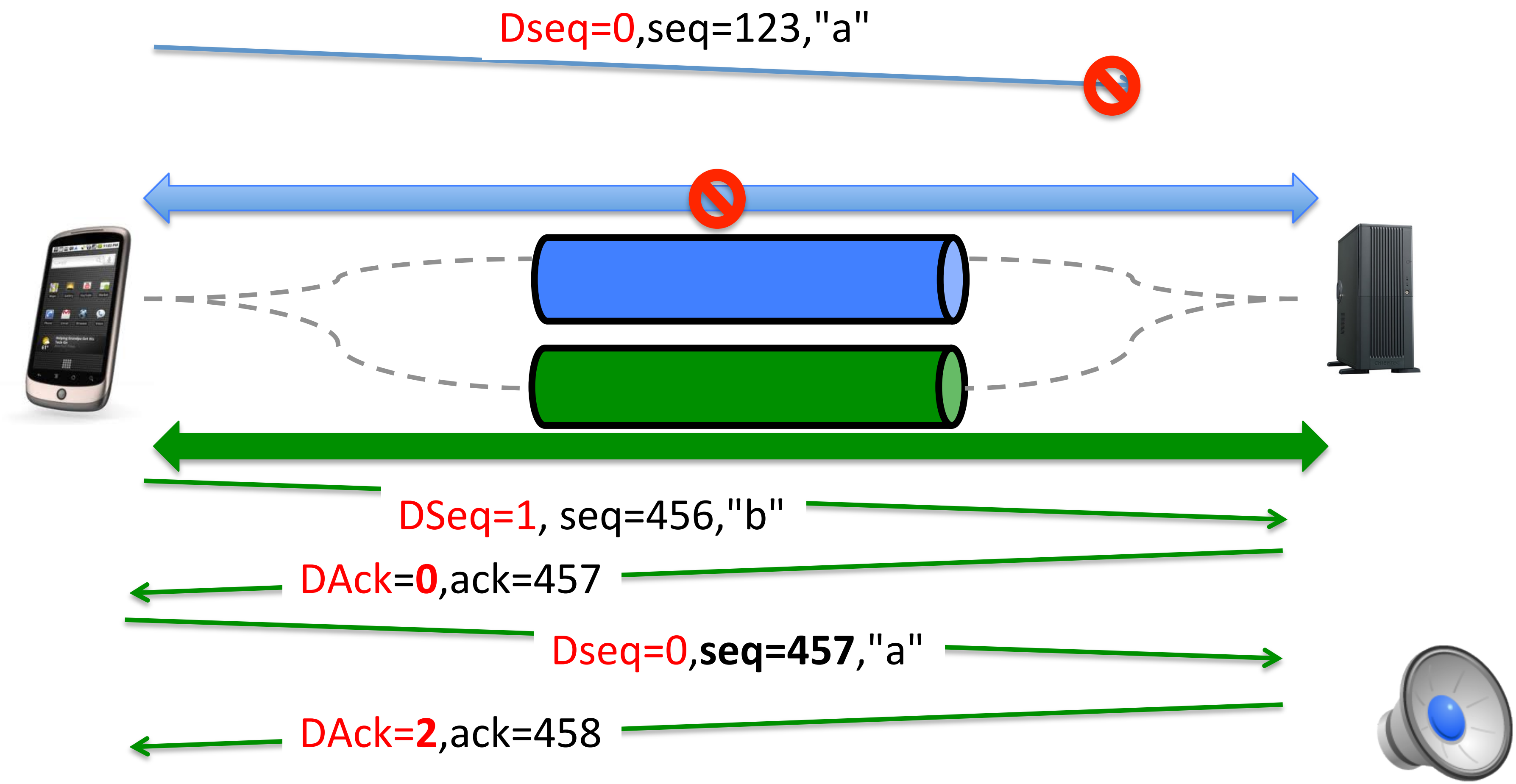
Multipath TCP
How to deal with losses ?

- Data losses over one TCP subflow
  - Fast retransmit and timeout as in regular TCP



Multipath TCP

- What happens when a TCP subflow fails ?



Retransmission heuristics

- Heuristics used by current Linux implementation
  - Fast retransmit is performed on the same subflow as the original transmission
  - Upon timeout expiration, reevaluate whether the segment could be retransmitted over another subflow
  - Upon loss of a subflow, all the unacknowledged data are retransmitted on other subflows

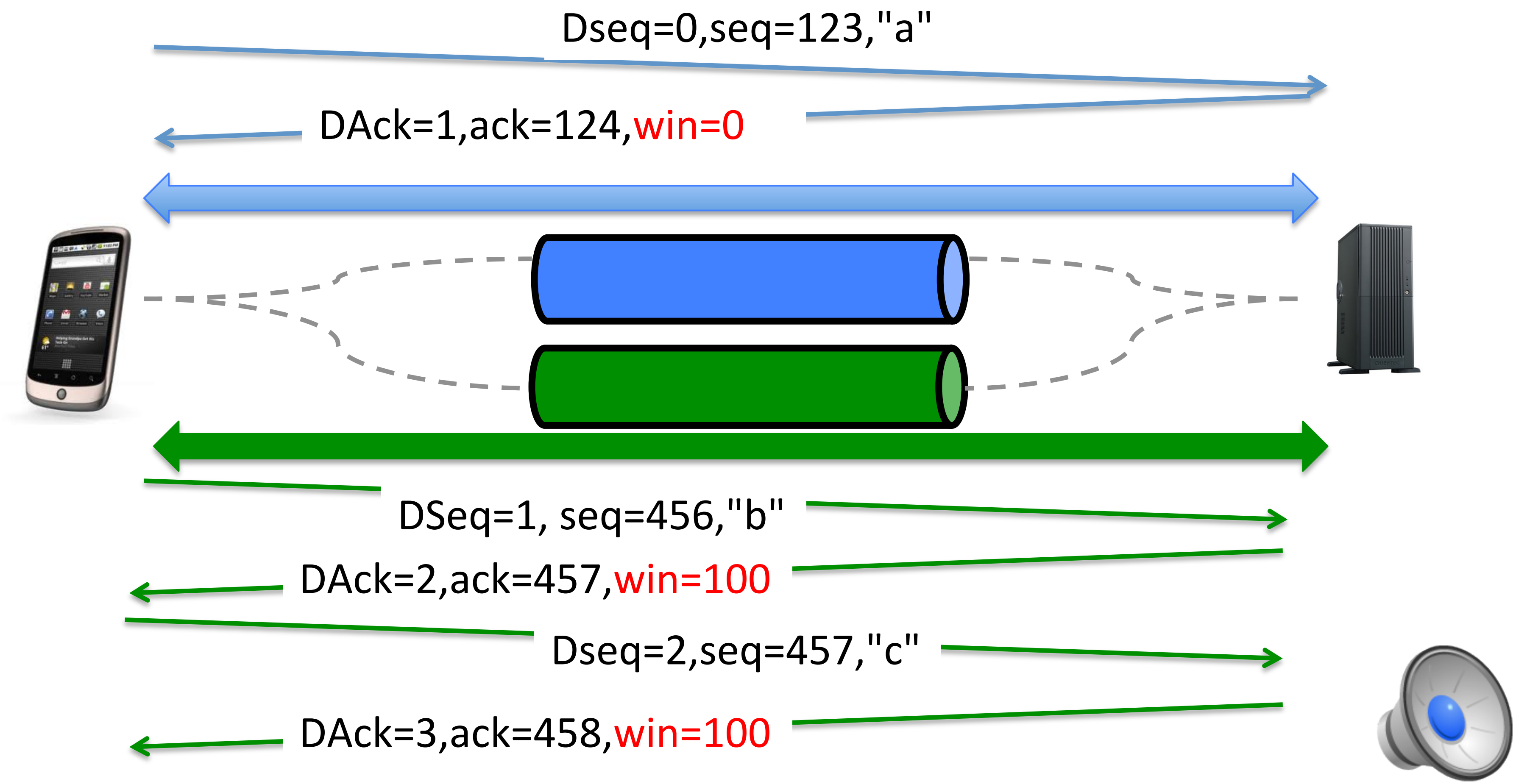
Flow control

- How should the window-based flow control be performed ?

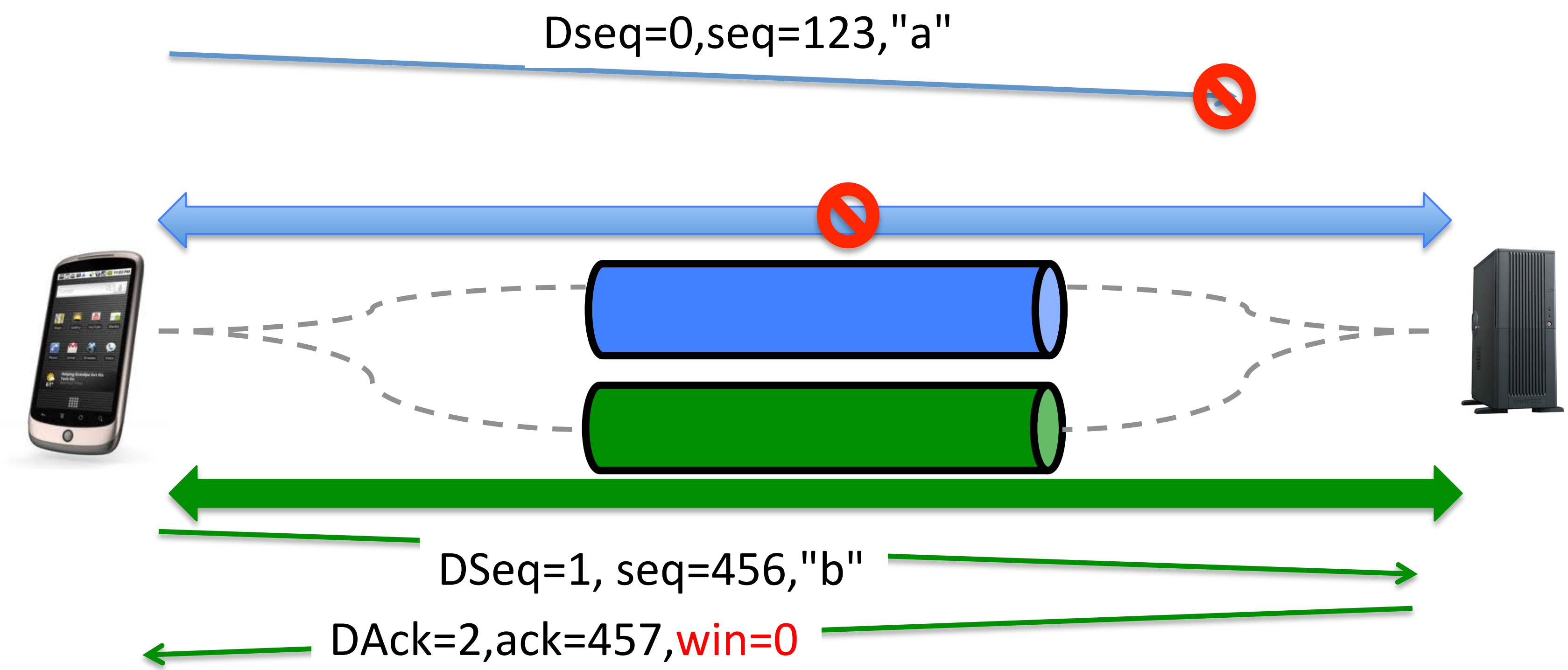
  – Independant windows on each TCP subflow

  – A single window that is shared among all TCP subflows

Independant windows

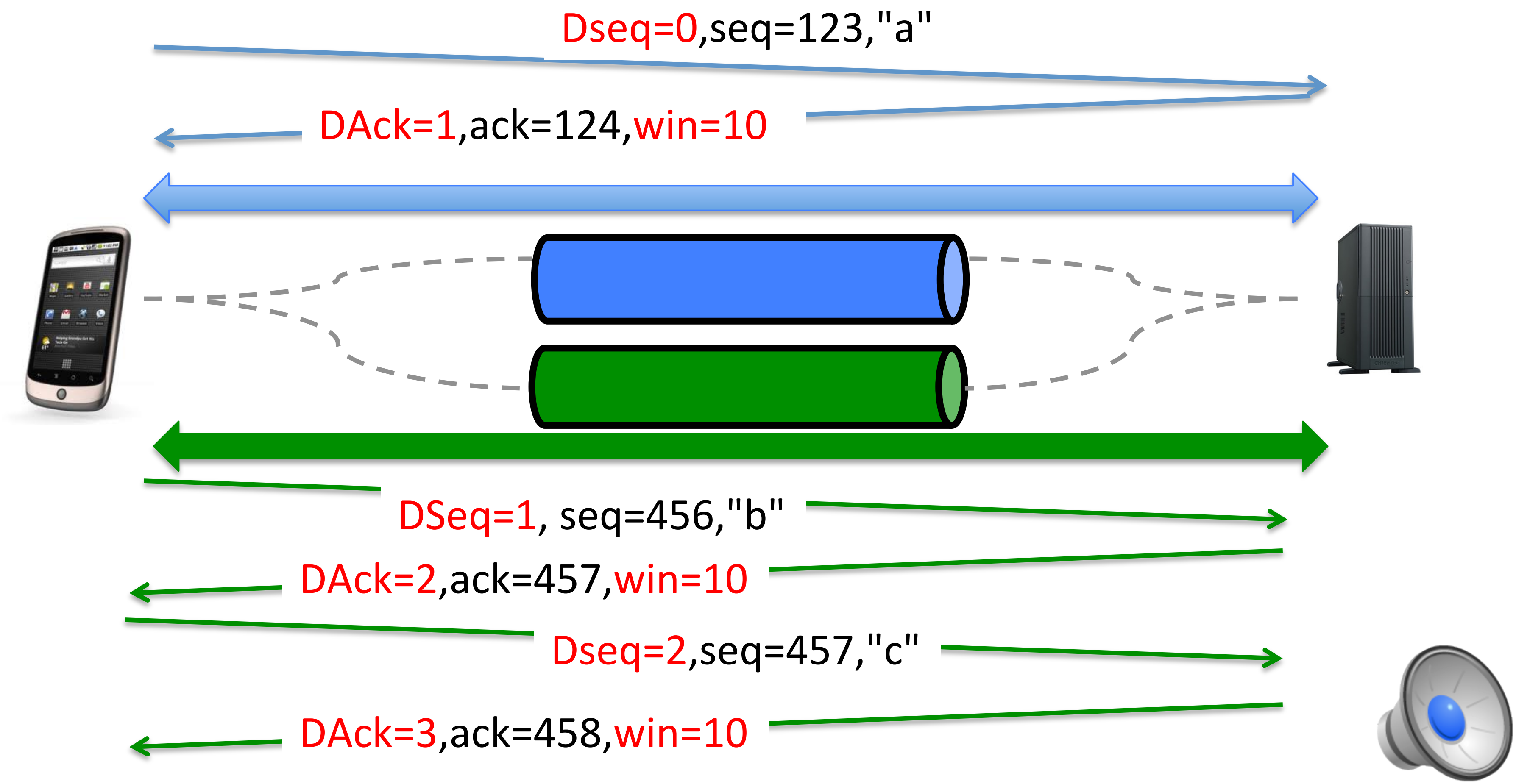


Independant windows possible problem

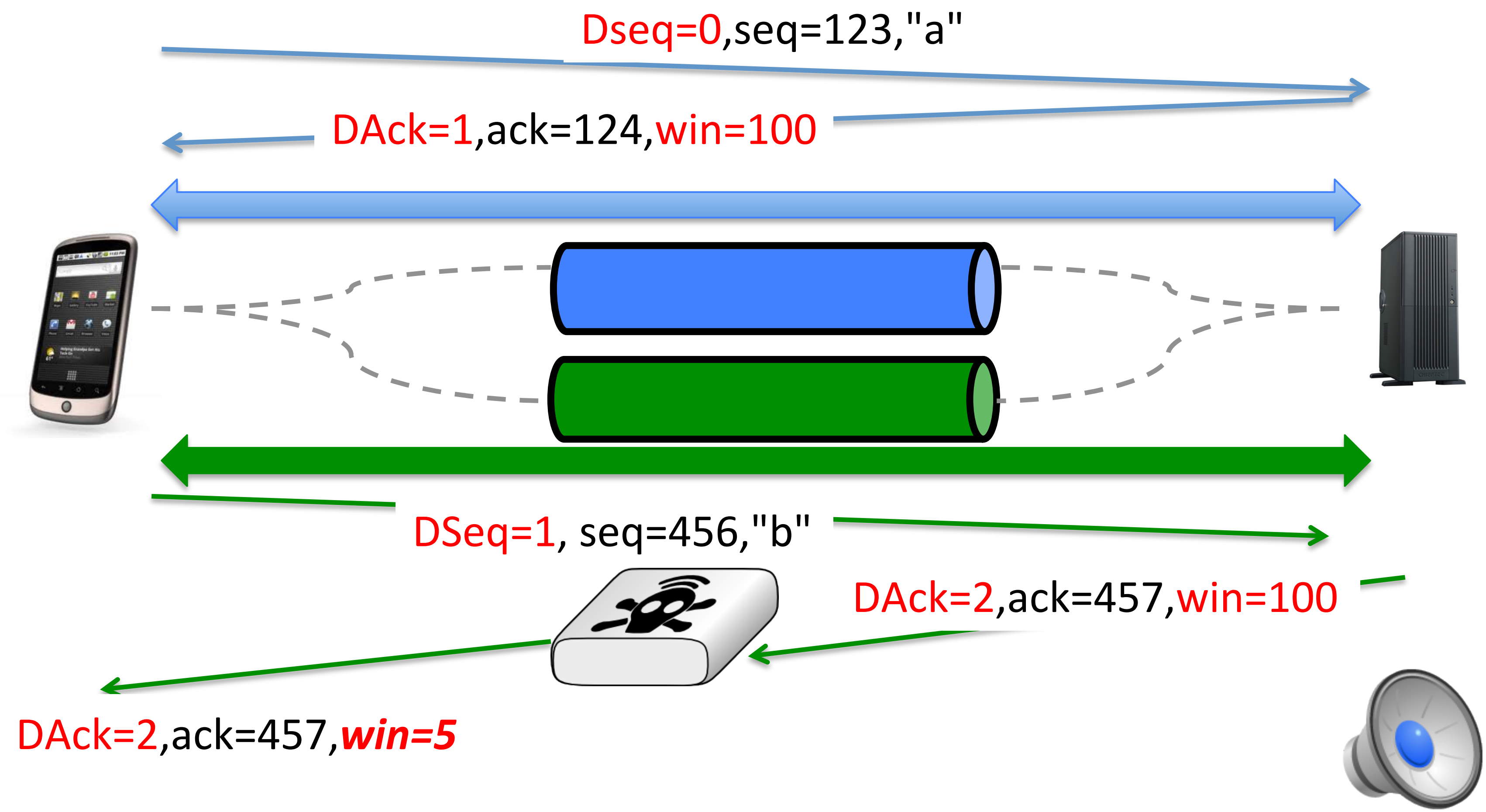


- Impossible to retransmit, window is already full on green subflow

A single window shared by all subflows



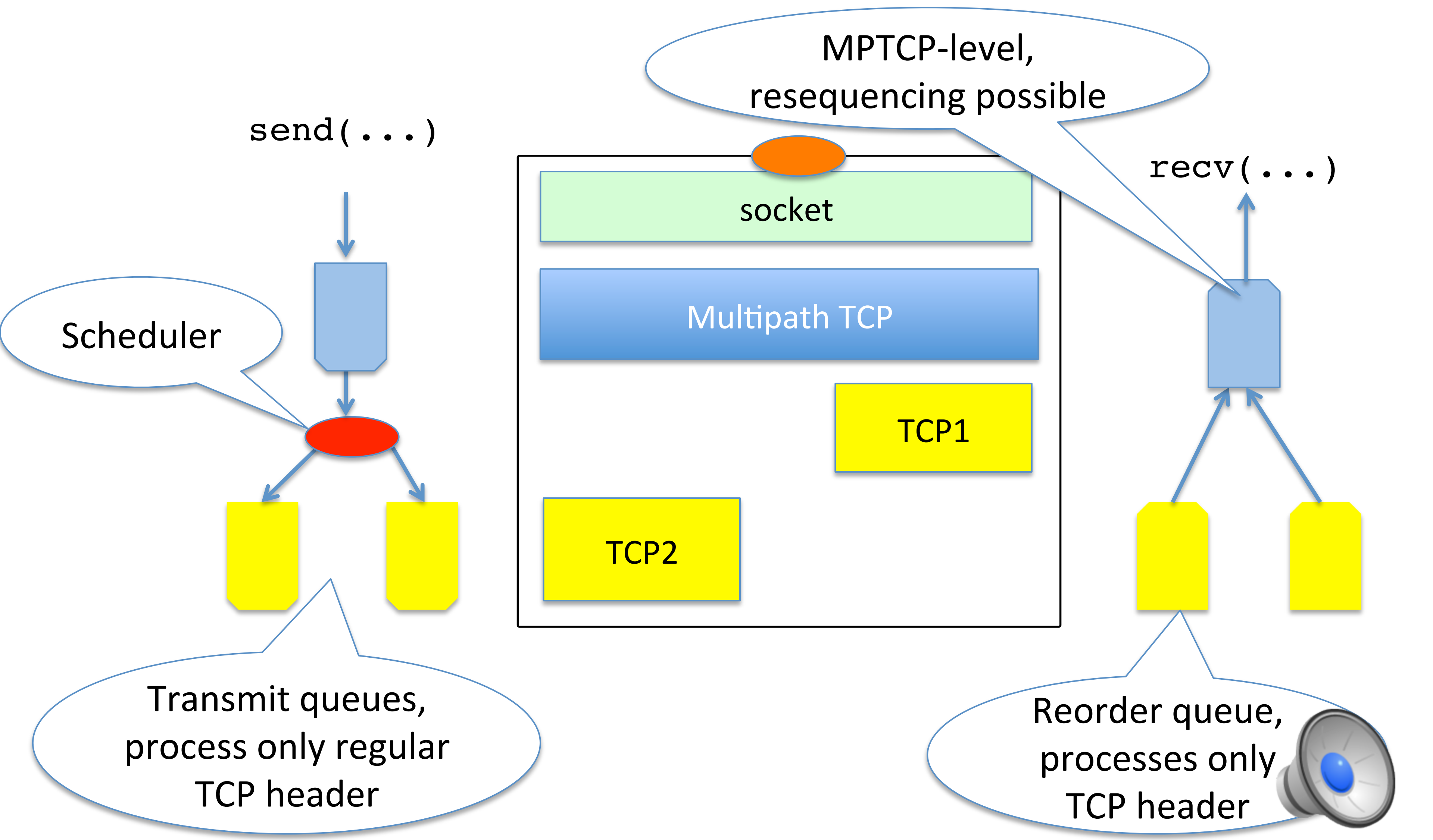
A single window shared by all subflows
Impact of middleboxes



Multipath TCP Windows

- Multipath TCP maintains one window per Multipath TCP connection
  - Window is relative to the last acked data (Data Ack)
  - Window is shared among all subflows
    - It's up to the implementation to decide how the window is shared
  - Window is transmitted inside the `window` field of the regular TCP header
  - If middleboxes change `window` field,
    - use largest `window` received at MPTCP-level
    - use received `window` over each subflow to cope with the flow control imposed by the middlebox

Multipath TCP buffers



Sending Multipath TCP information

- How to exchange the Multipath TCP specific information between two hosts ?

- Option 1
  - Use TLVs to encode data and control information inside payload of subflows

- **Option 2**
  - Use TCP options to encode all Multipath TCP information

Option 1 : Michael Scharf, Thomas-Rolf Banniza , *MCTCP: A Multipath Transport Shim Layer*, GLOBECOM

Multipath TCP with only options

- Advantages
  - Normal way of extending TCP
  - Should be able to go through middleboxes or fallback
- Drawbacks
  - limited size of the TCP options, notably inside SYN
  - What happens when middleboxes drop TCP options in data segments

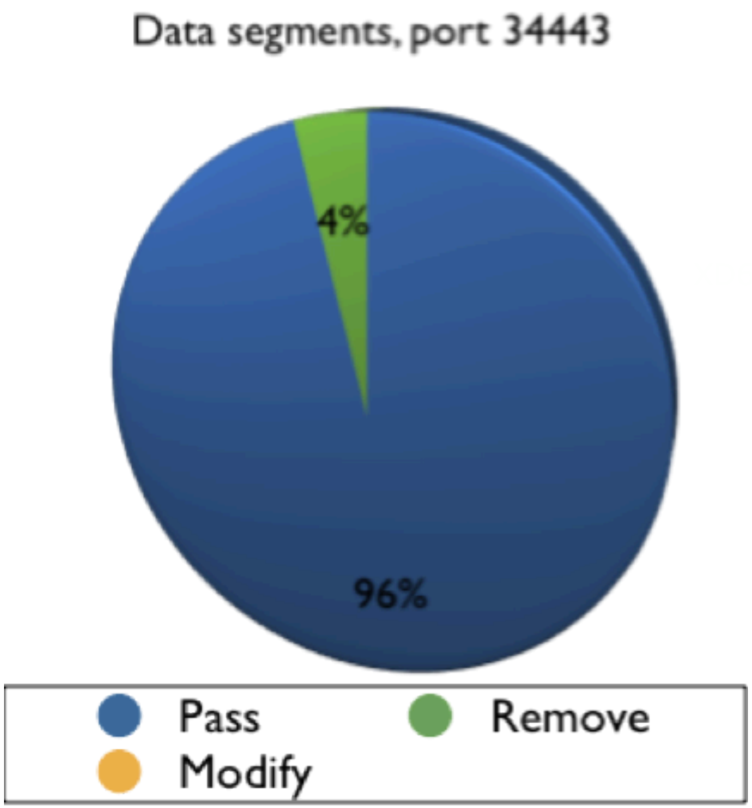
Multipath TCP using TLV

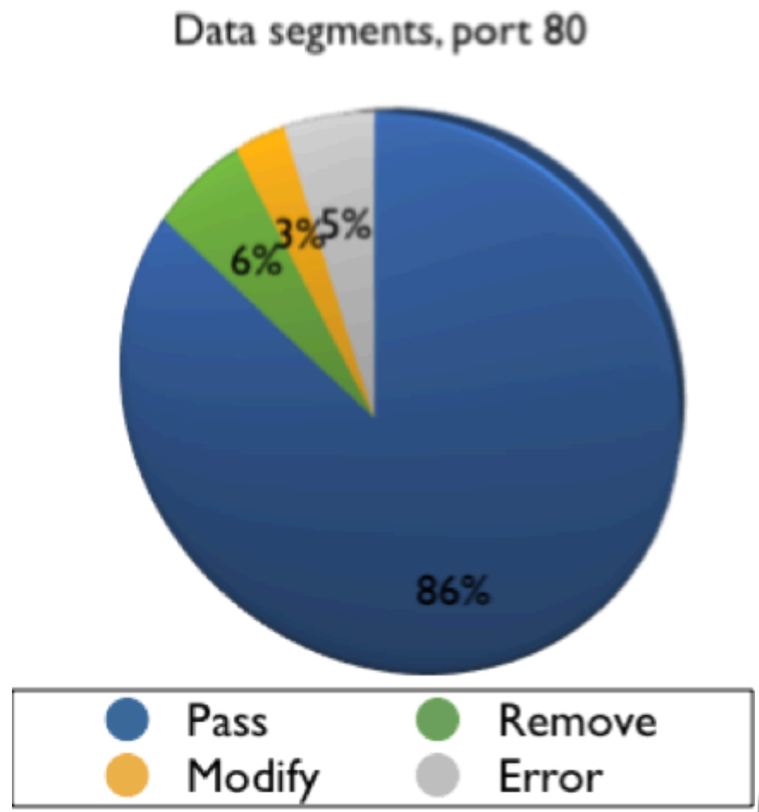
- Advantages
  - Multipath TCP could start as regular TCP and move to Multipath only when needed
  - Could be implemented as a library in userspace
  - TLVs can be easily extended
- Drawbacks
  - TCP segments contain TLVs including the data and not only the data
    - problem for middleboxes, DPI, ..
  - Middleboxes become more difficult

Michael Scharf, Thomas-Rolf Banniza , *MCTCP: A Multipath Transport Shim Layer*, GLOBECOM 2011

Is it safe to use TCP options ?

- Known option (TS) in Data segments

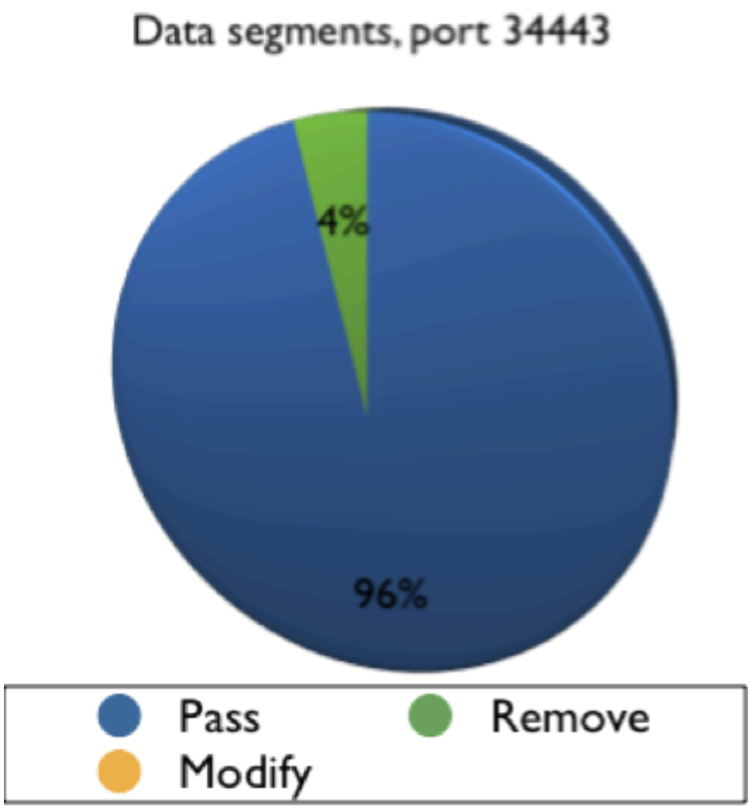


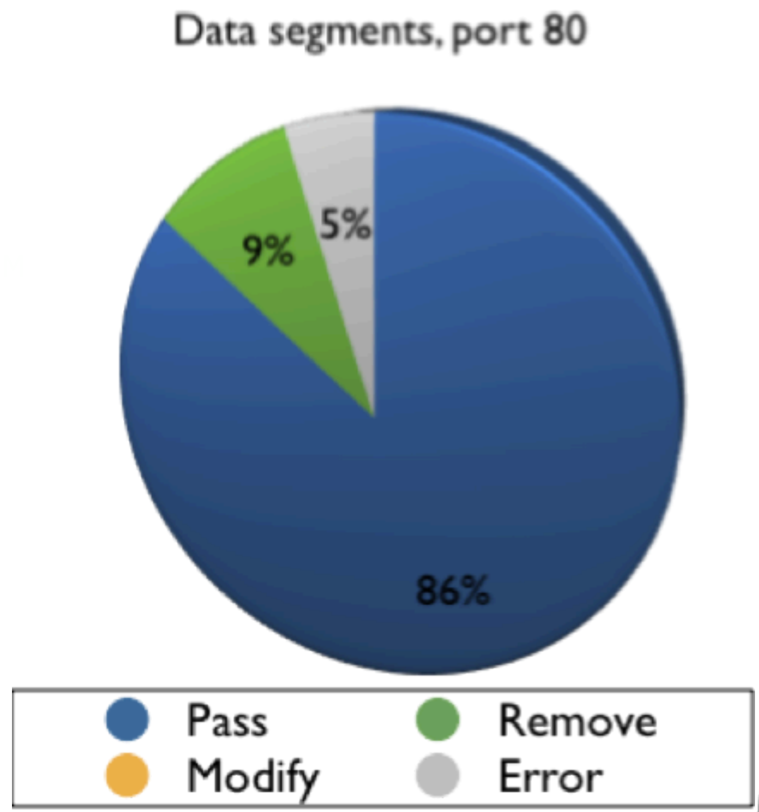


Honda, Michio, et al. "Is it still possible to extend TCP?." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

Is it safe to use TCP options ?

- Unknown option in Data segments





Honda, Michio, et al. "Is it still possible to extend TCP?." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

Multipath TCP options

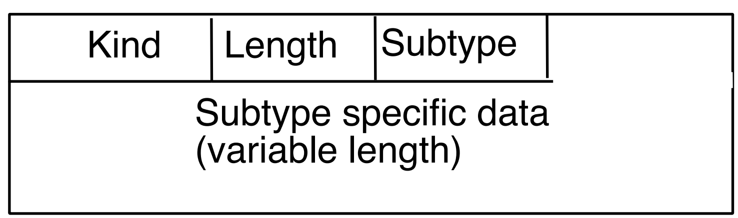
- TCP option format

Kind	Length	Option-specific data

- Initial design
  - One option kind for each purpose (e.g. Data Sequence number)
- Final design
  - A single variable-length Multipath TCP option

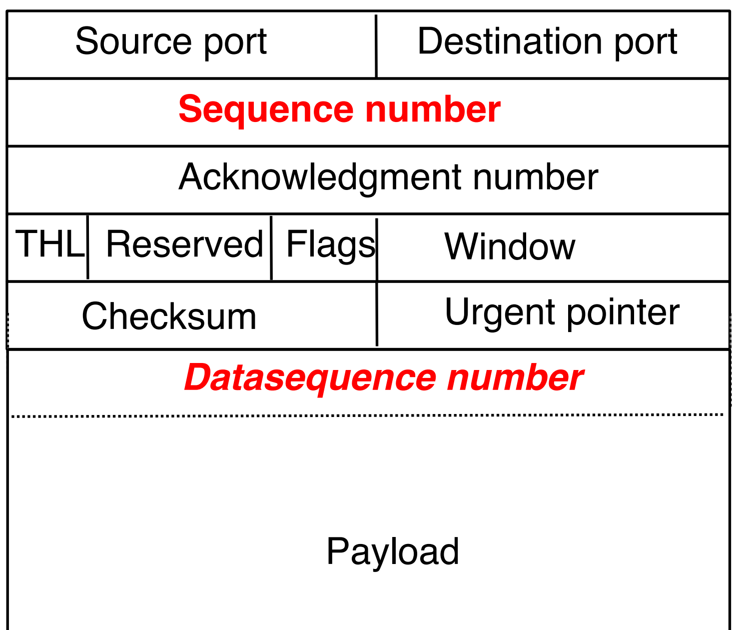
Multipath TCP option

- A single option type
  - to minimise the risk of having one option accepted by middleboxes in SYN segments and rejected in segments carrying data

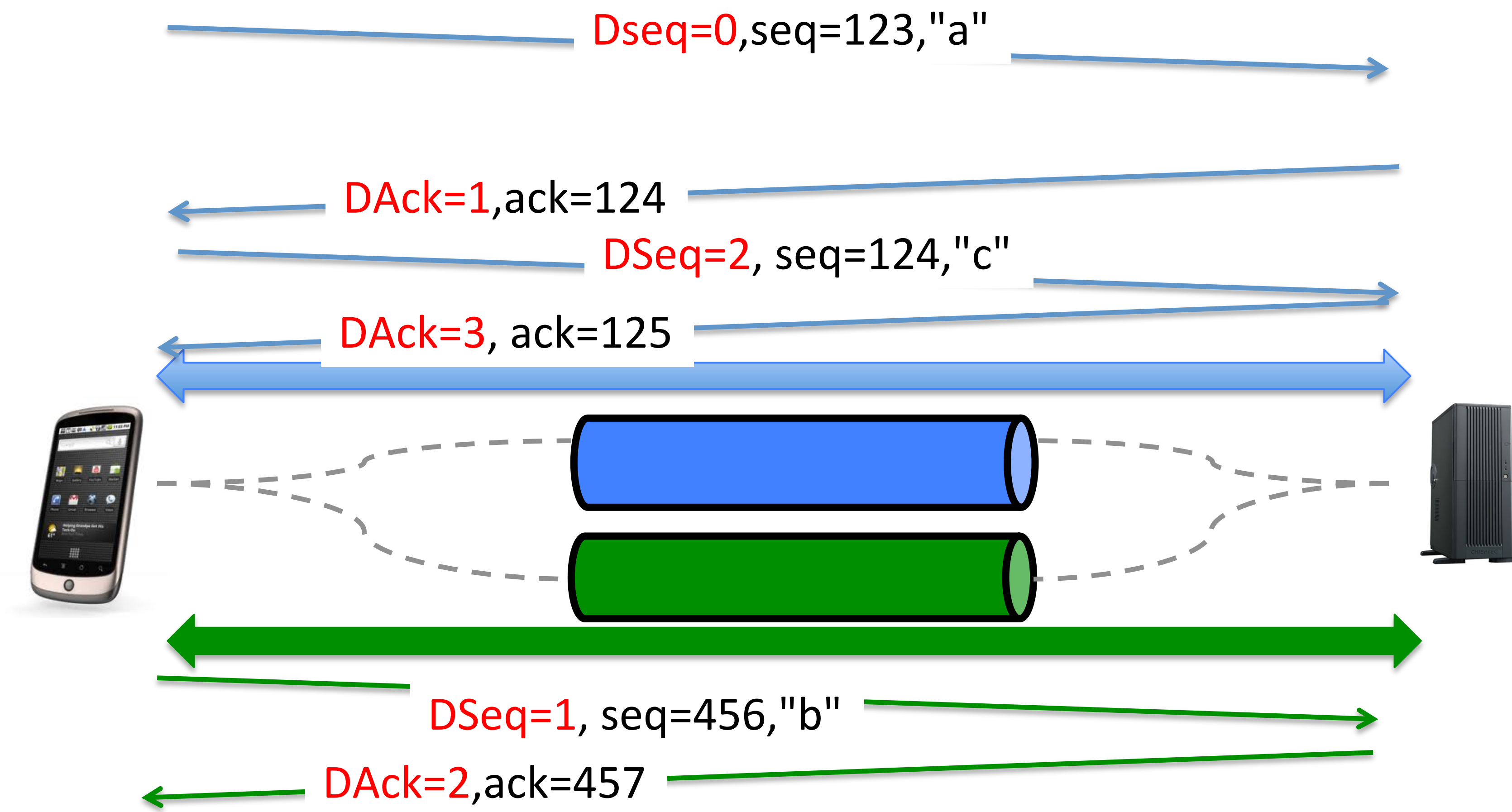


Data sequence numbers and TCP segments

- How to transport Data sequence numbers ?
  - Same solution as for TCP
    - Data sequence number in TCP option is the Data sequence number of the first byte of the segment



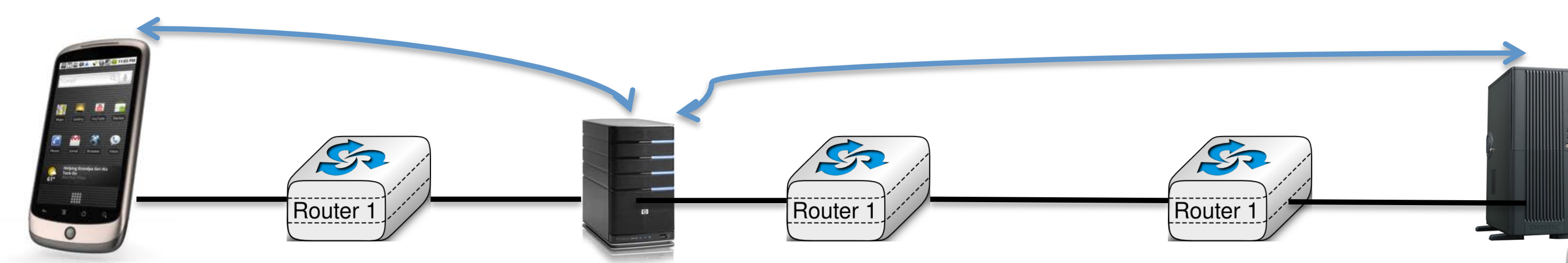
Multipath TCP
Data transfer



Which middleboxes change TCP sequence numbers ?

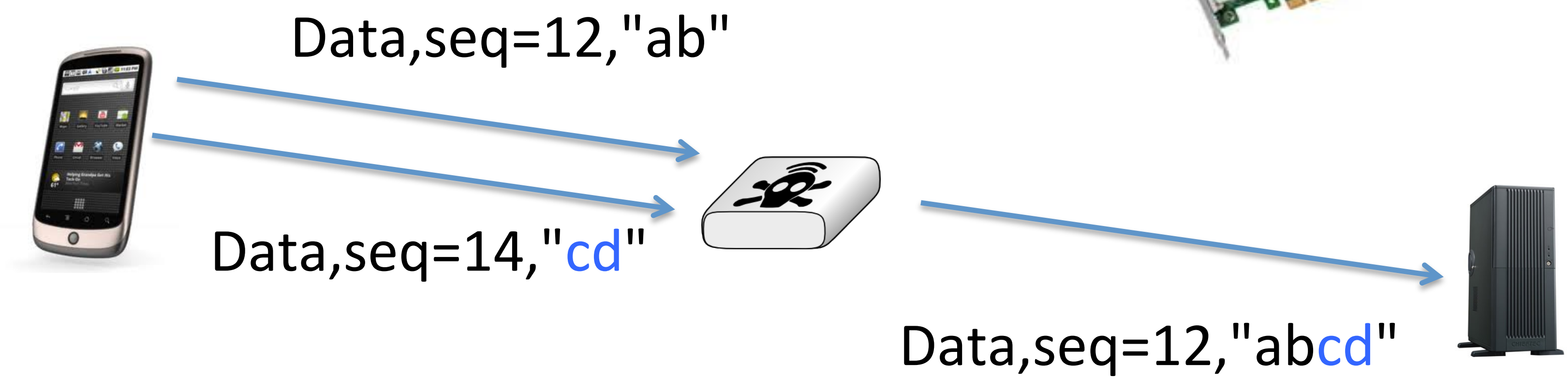
- Some firewalls change TCP sequence numbers in SYN segments to ensure randomness
  - fix for old windows95 bug

- Transparent proxies terminate TCP connections



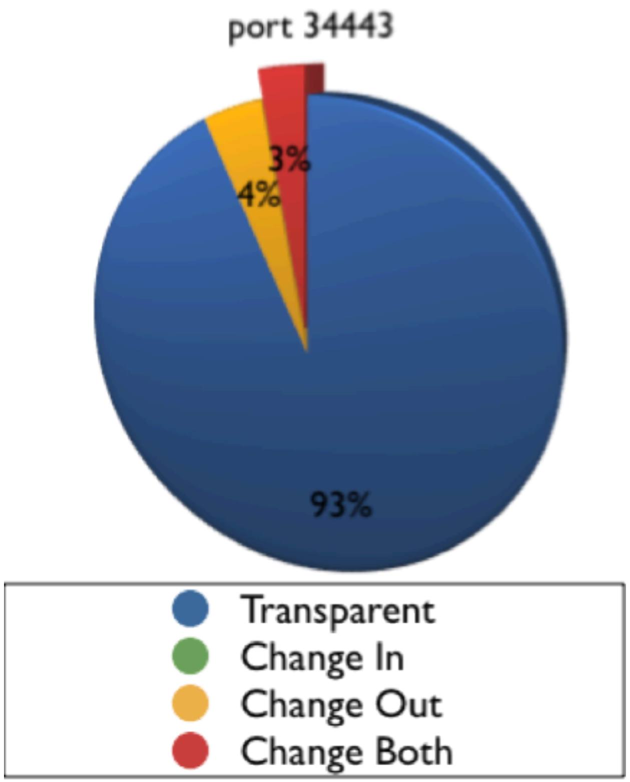
Middlebox interference

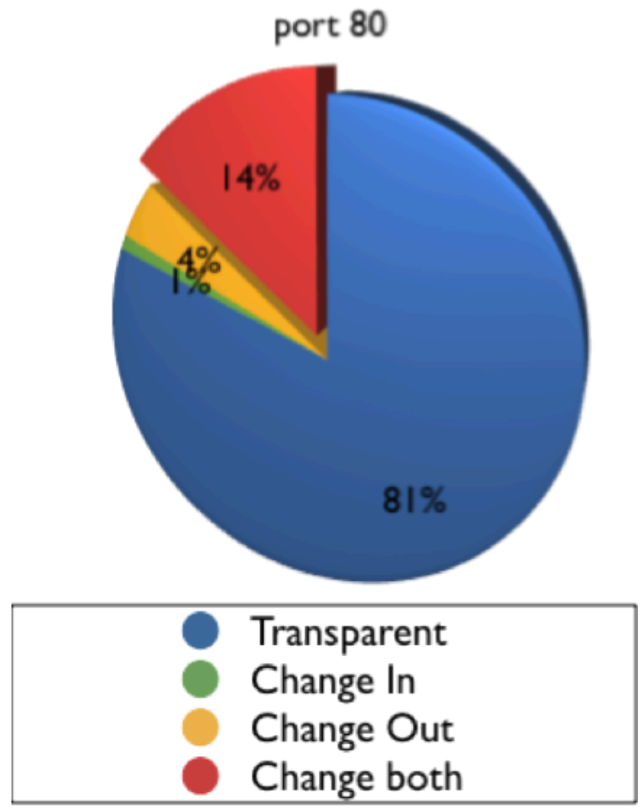
- Data segments



Such a middlebox could also be the network adapter of the server that uses LRO to improve performance.

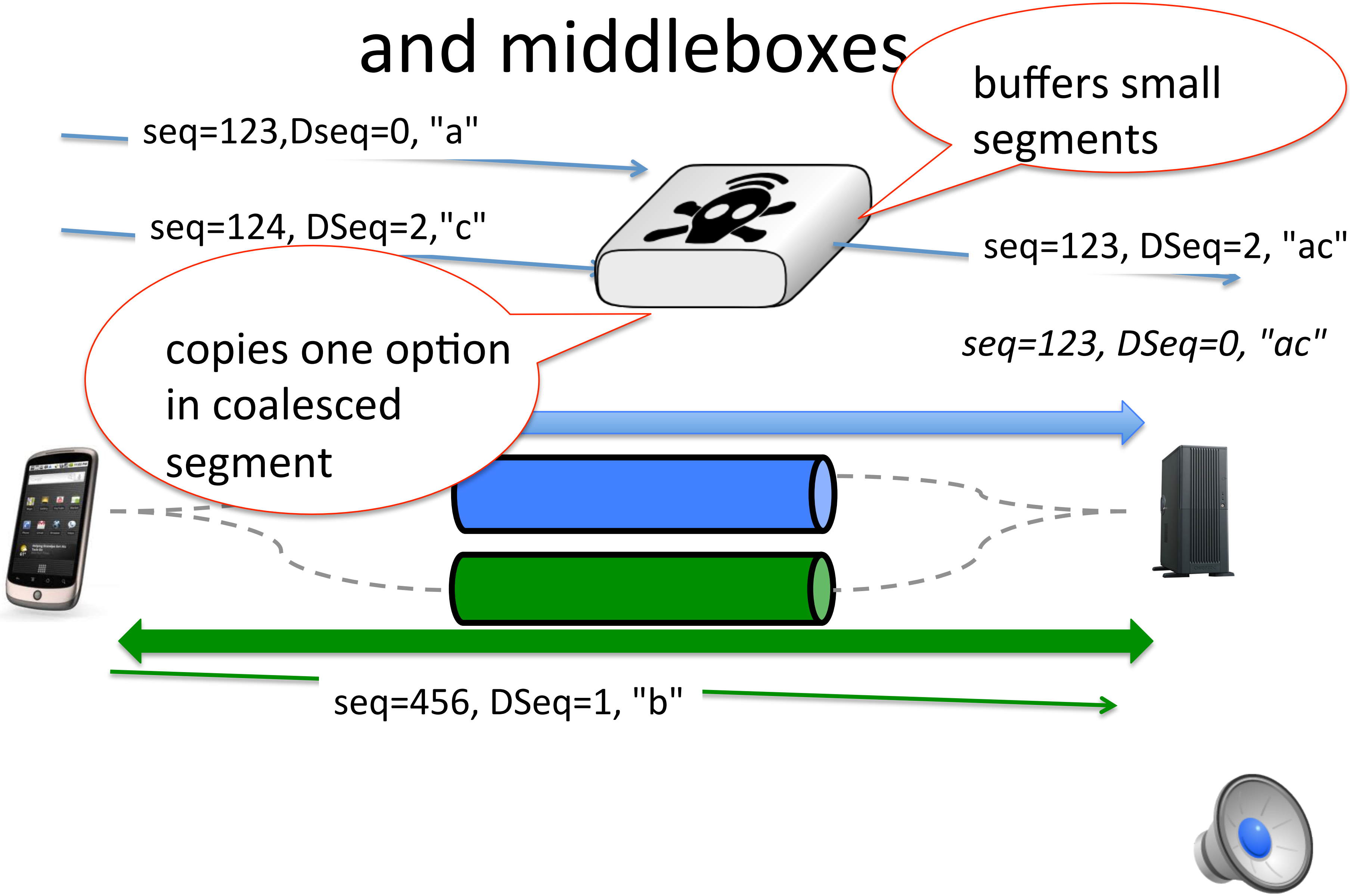
Segment coalescing



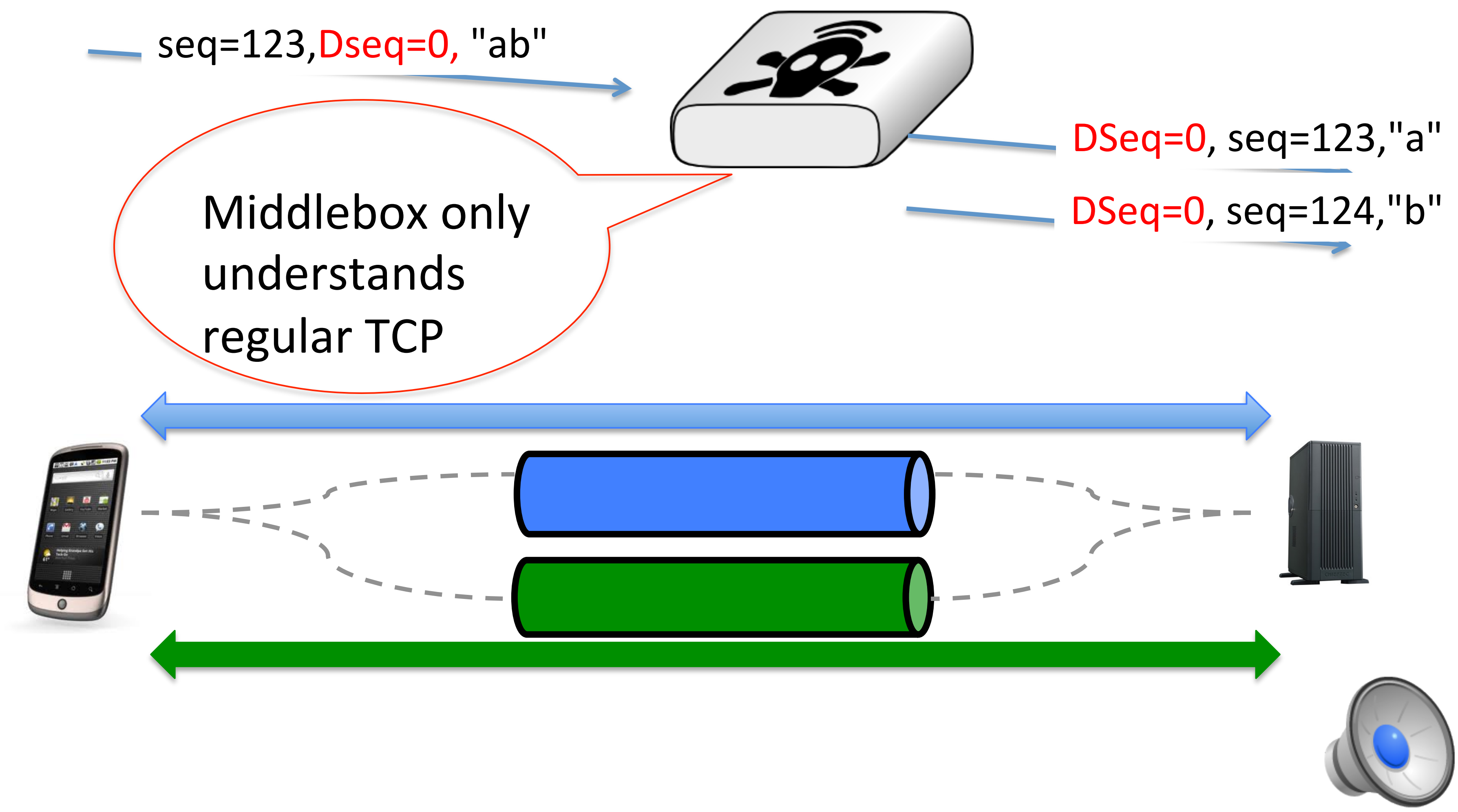


Honda, Michio, et al. "Is it still possible to extend TCP?." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

Data sequence numbers and middleboxes



Data sequence numbers and middleboxes



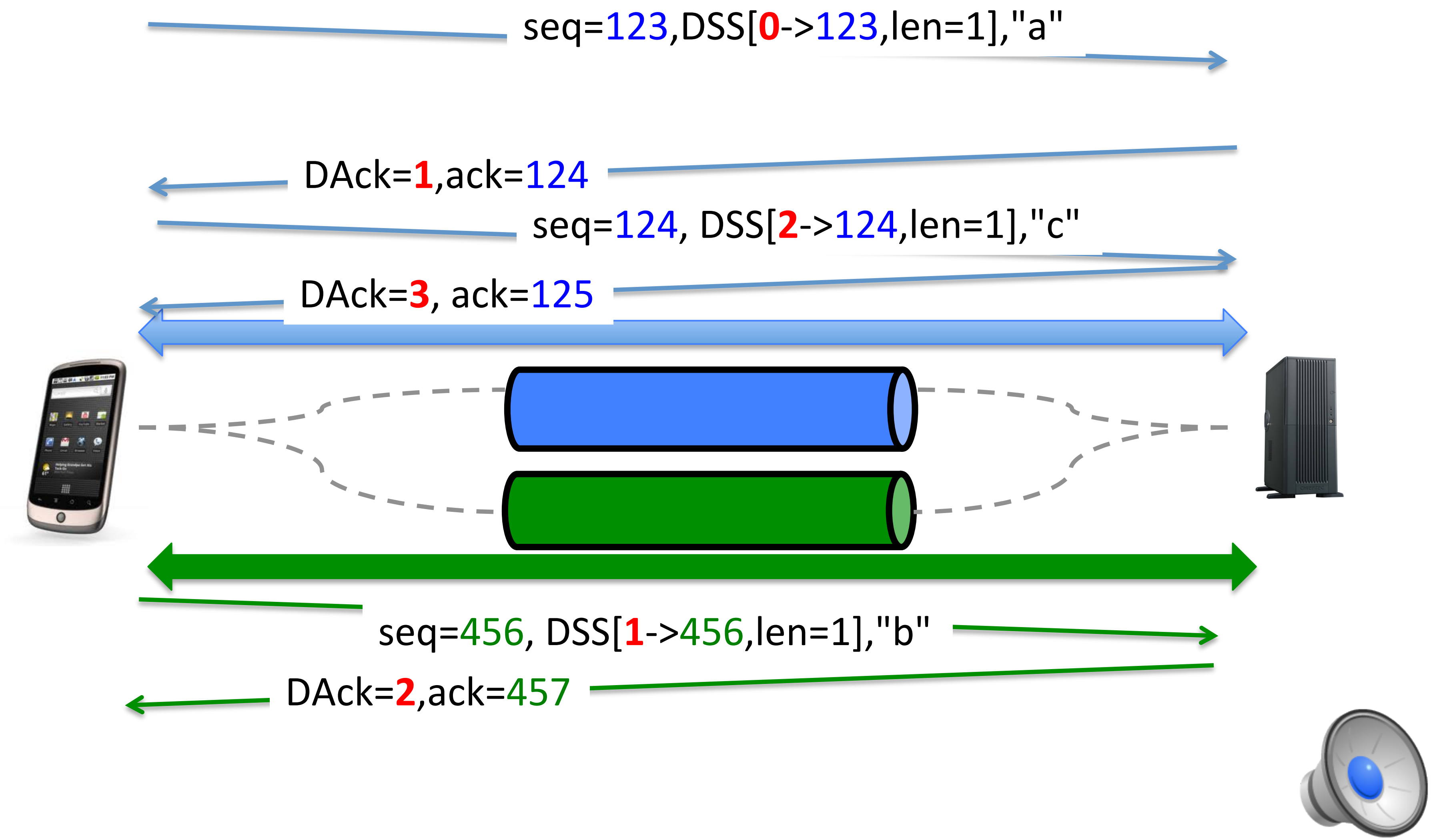
A "middlebox" that both splits and coalesces TCP segments

Data sequence numbers and middleboxes

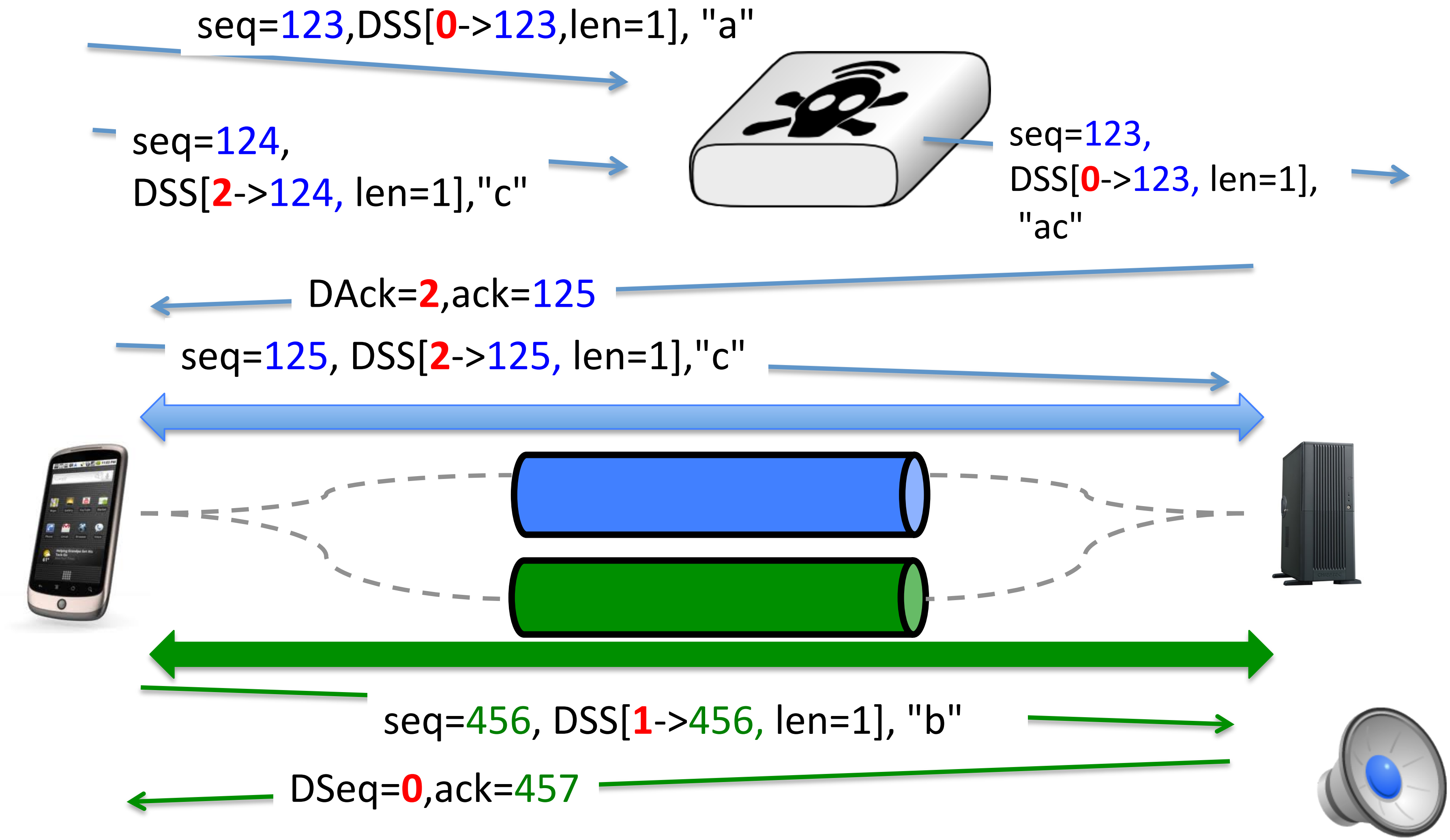
- How to avoid desynchronisation between the bytestream and data sequence numbers ?

- Solution
  - Multipath TCP option carries **mapping** between Data sequence numbers and *(difference between initial and current)* subflow sequence numbers
    - mapping covers a part of the bytestream (length)

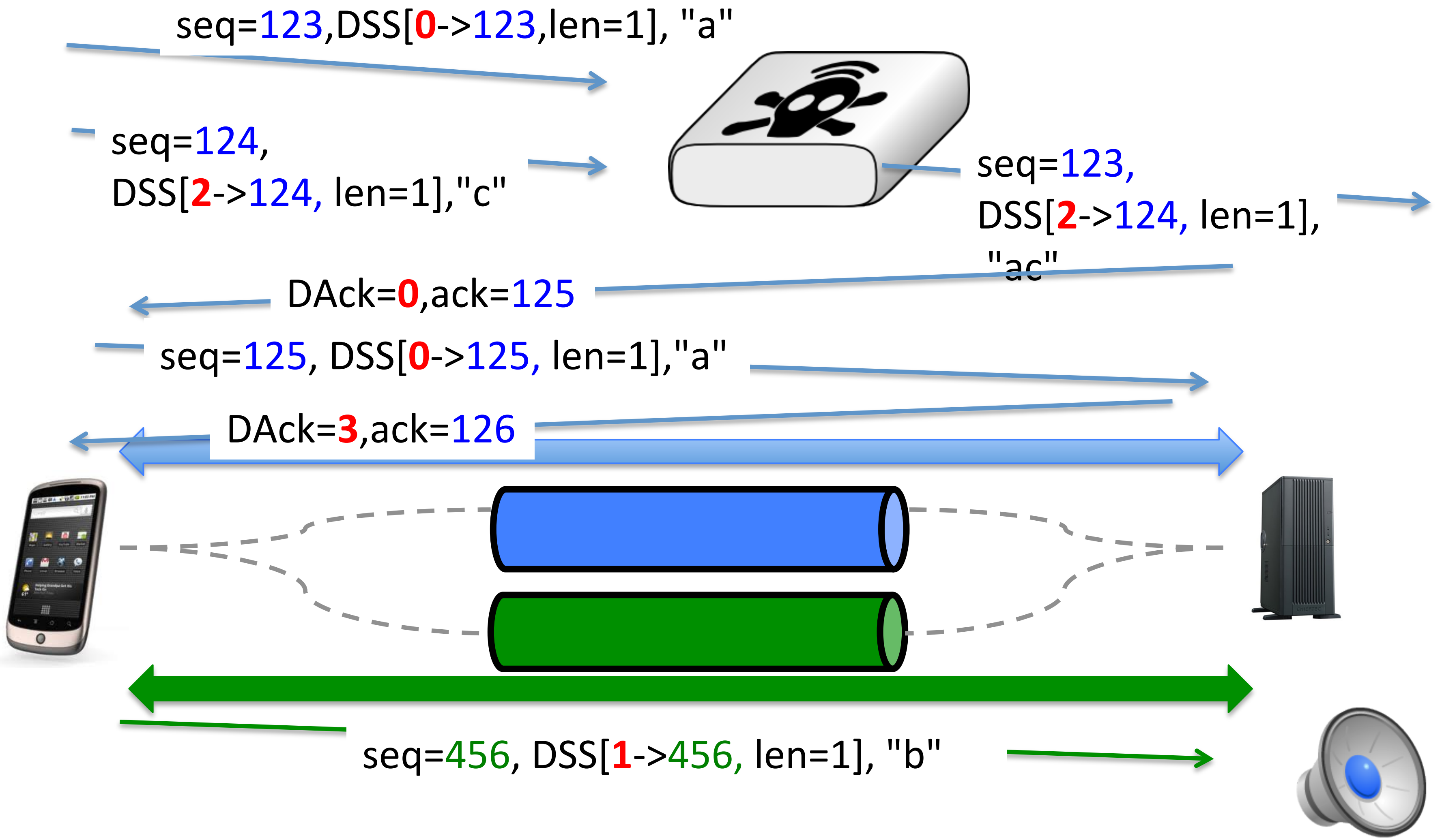
Multipath TCP
Data transfer



Data sequence numbers and middleboxes



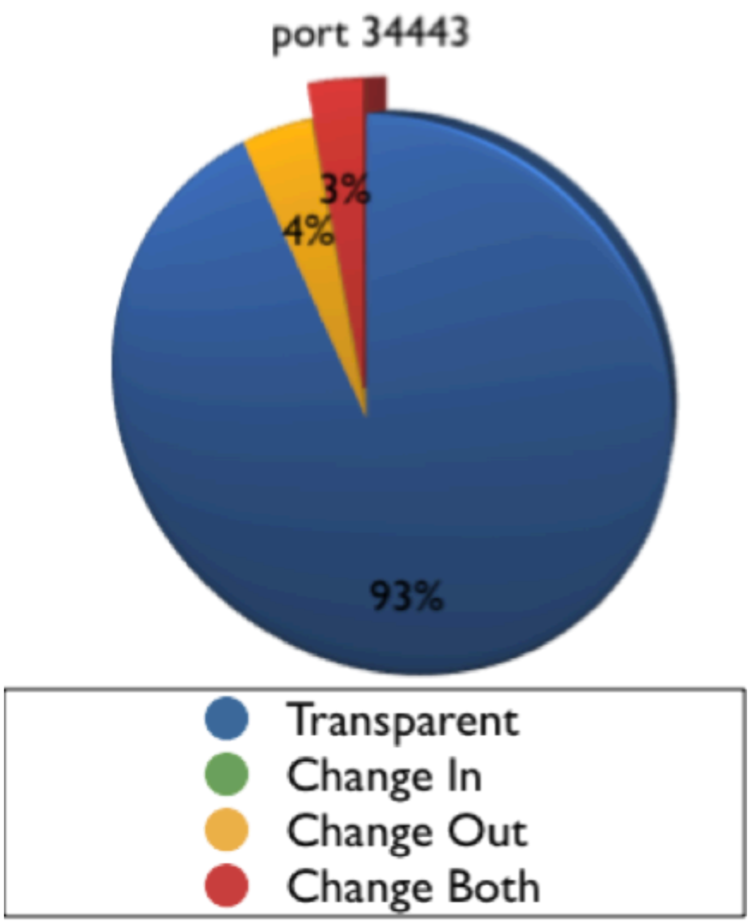
Data sequence numbers and middleboxes

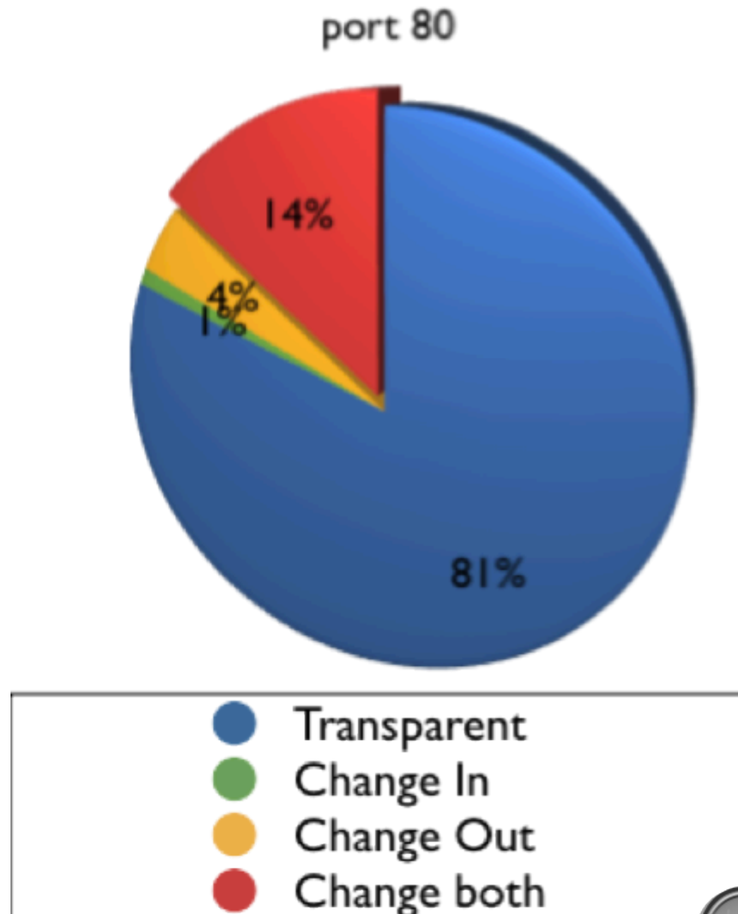


Multipath TCP and middleboxes

- With the DSS mapping, Multipath TCP can cope with middleboxes that
  - combine segments
  - split segments
- Are they the most annoying middleboxes for Multipath TCP ?
  - Unfortunately not

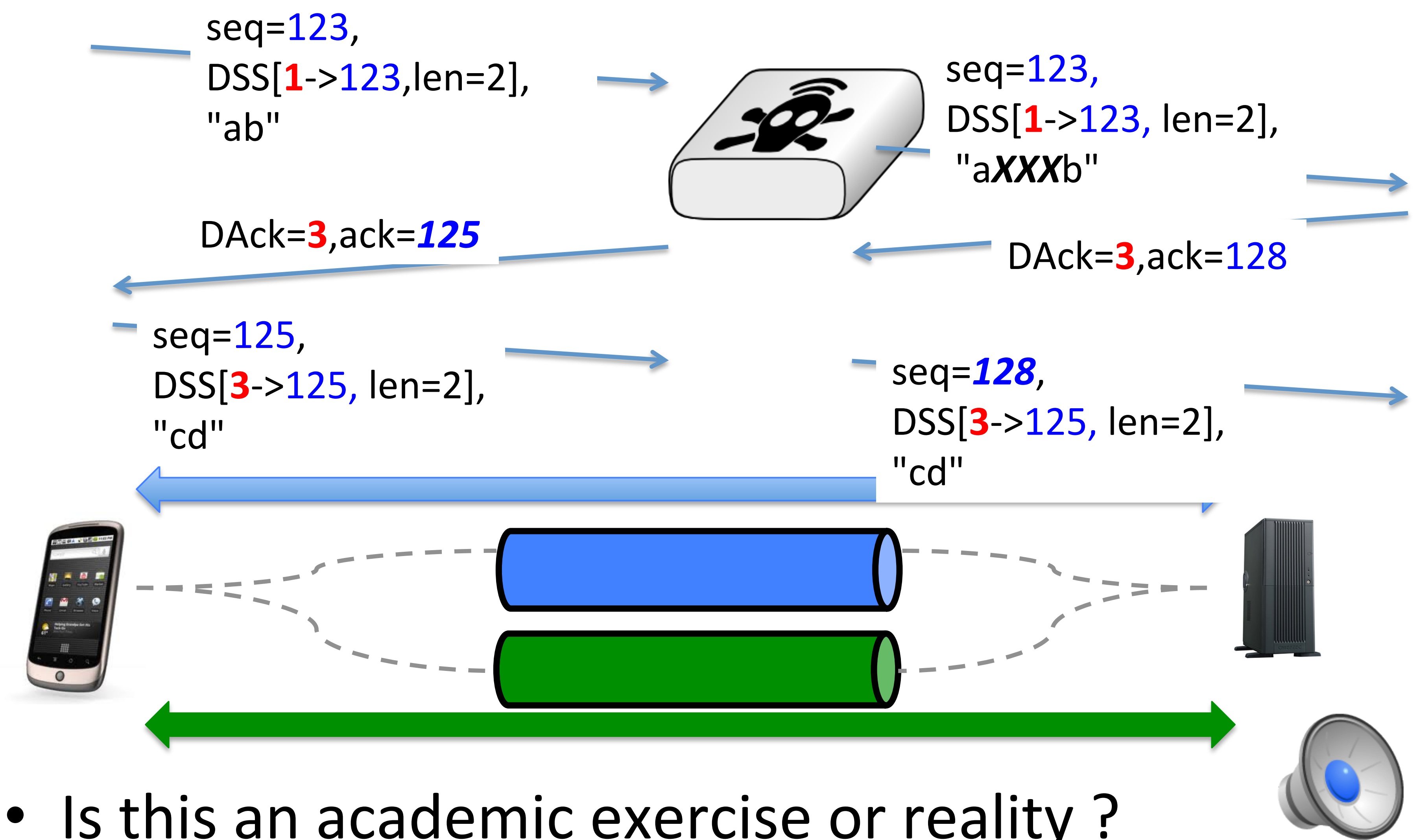
TCP sequence number and middleboxes





Honda, Michio, et al. "Is it still possible to extend TCP?." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

The worst middlebox



- Is this an academic exercise or reality ?

The worst middlebox

- Is unfortunately very old...
  - Any ALG for a NAT

220 ProFTPD 1.3.3d Server (BELNET FTPD Server) [193.190.67.15]
ftp_login: user `<null>' pass `<null>' host `ftp.belnet.be'
Name (ftp.belnet.be:obo): anonymous
---> USER anonymous
331 Anonymous login ok, send your complete email address as your password
Password:
---> PASS XXXX
---> **PORT 192,168,0,7,195,120**
200 PORT command successful
---> LIST
150 Opening ASCII mode data connection for file list
lrw-r--r--   1 ftp   ftp   6 Jun 1 2011 pub -> mirror
226 Transfer complete

Coping with the worst middlebox

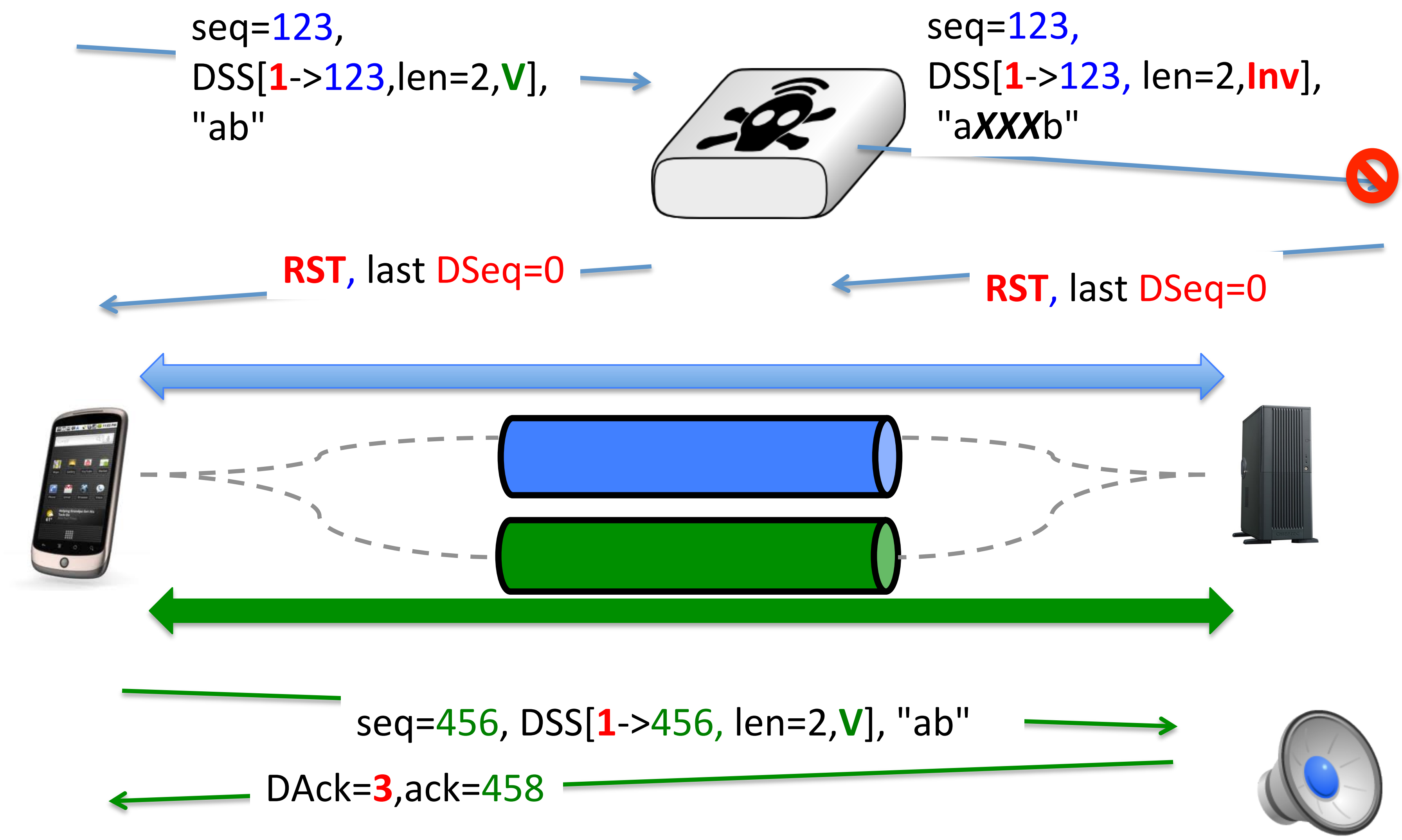
- What should Multipath TCP do in the presence of such a worst middlebox ?
  - Do nothing and ignore the middlebox
    - but then the bytestream and the application would be broken and this problem will be difficult to debug by network administrators
  - Detect the presence of the middlebox
    - and fallback to regular TCP (i.e. use a single path and nothing fancy)

Multipath TCP **MUST** work in all networks where regular TCP works.

Detecting the worst middlebox ?

- How can Multipath TCP detect a middlebox that modifies the bytestream and inserts/ removes bytes ?
  - Various solutions were explored
  - In the end, Multipath TCP chose to include its own checksum to detect insertion/deletion of bytes

The worst middlebox



Multipath TCP
Data sequence numbers

- What should be the length of the data sequence numbers ?

  – 32 bits
    - compact and compatible with TCP
    - wrap around problem at highspeed requires PAWS

  – 64 bits
    - wrap around is not an issue for most transfers today
    - takes more space inside each segment

Multipath TCP
Data sequence numbers

- Data sequence numbers and Data acknowledgements
  - Maintained inside implementation as 64 bits field
  - Implementations can, as an optimisation, only transmit the lower 32 bits of the data sequence and acknowledgements

Data Sequence Signal option

A = Data ACK present
a = Data ACK is 8 octets
M = mapping present
m = DSN is 8

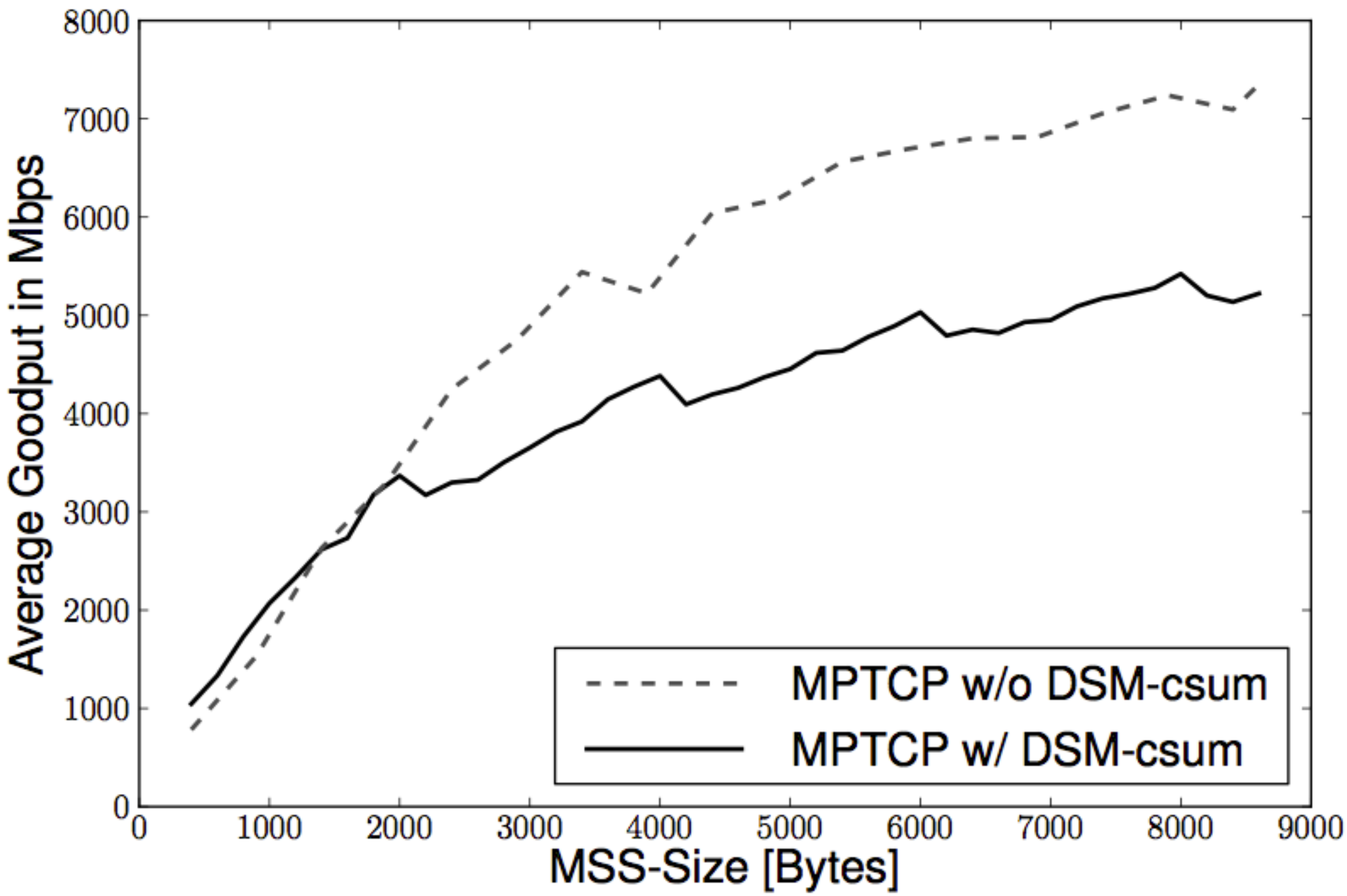
Cumulative Data ack

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+----------------------+
|     Kind      |    Length     |Subtype| (reserved) |F|m|M|a|A|
+---------------+---------------+-------+----------------------+
|          Data ACK (4 or 8 octets, depending on flags)          |
+--------------------------------------------------------------+
|   Data Sequence Number (4 or 8 octets, depending on flags)     |
+--------------------------------------------------------------+
|              Subflow Sequence Number (4 octets)                |
+-------------------------------+------------------------------+
|  Data-level Length (2 octets) |      Checksum (2 octets)       |
+-------------------------------+------------------------------+
```

Length of mapping, can extend beyond this segment

Computed over data covered by entire mapping + pseudo header

Cost of the DSN checksum



C. Raiciu, et al. "How hard can it be? designing and implementing a deployable multipath TCP," NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

The Multipath TCP protocol

- Control plane
  - How to manage a Multipath TCP connection that uses several paths ?
- Data plane
  - How to transport data ?

→ **Congestion control**

  - How to control congestion over multiple paths ?

TCP congestion control

- A linear rate adaption algorithm
  - $rate(t+1) = \alpha_C + \beta_C rate(t)$ when the network is congested
  - $rate(t+1) = \alpha_N + \beta_N rate(t)$ when the network is *not* congested

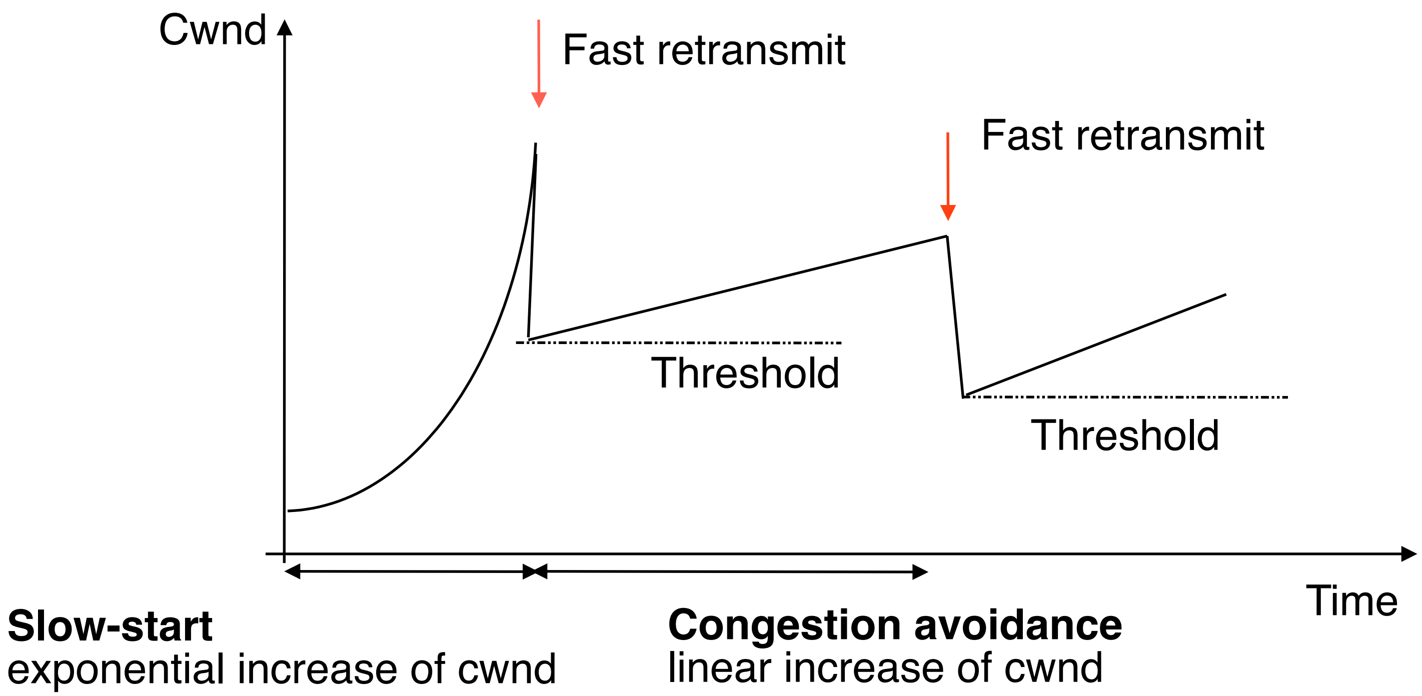
To be fair and efficient, a linear algorithm must use additive increase and multiplicative decrease (AIMD)

```
# Additive Increase Multiplicative Decrease
if congestion :
   rate=rate*betaC    # multiplicative decrease, betaC<1
else
   rate=rate+alphaN    # additive increase, v0>0
```

AIMD in TCP

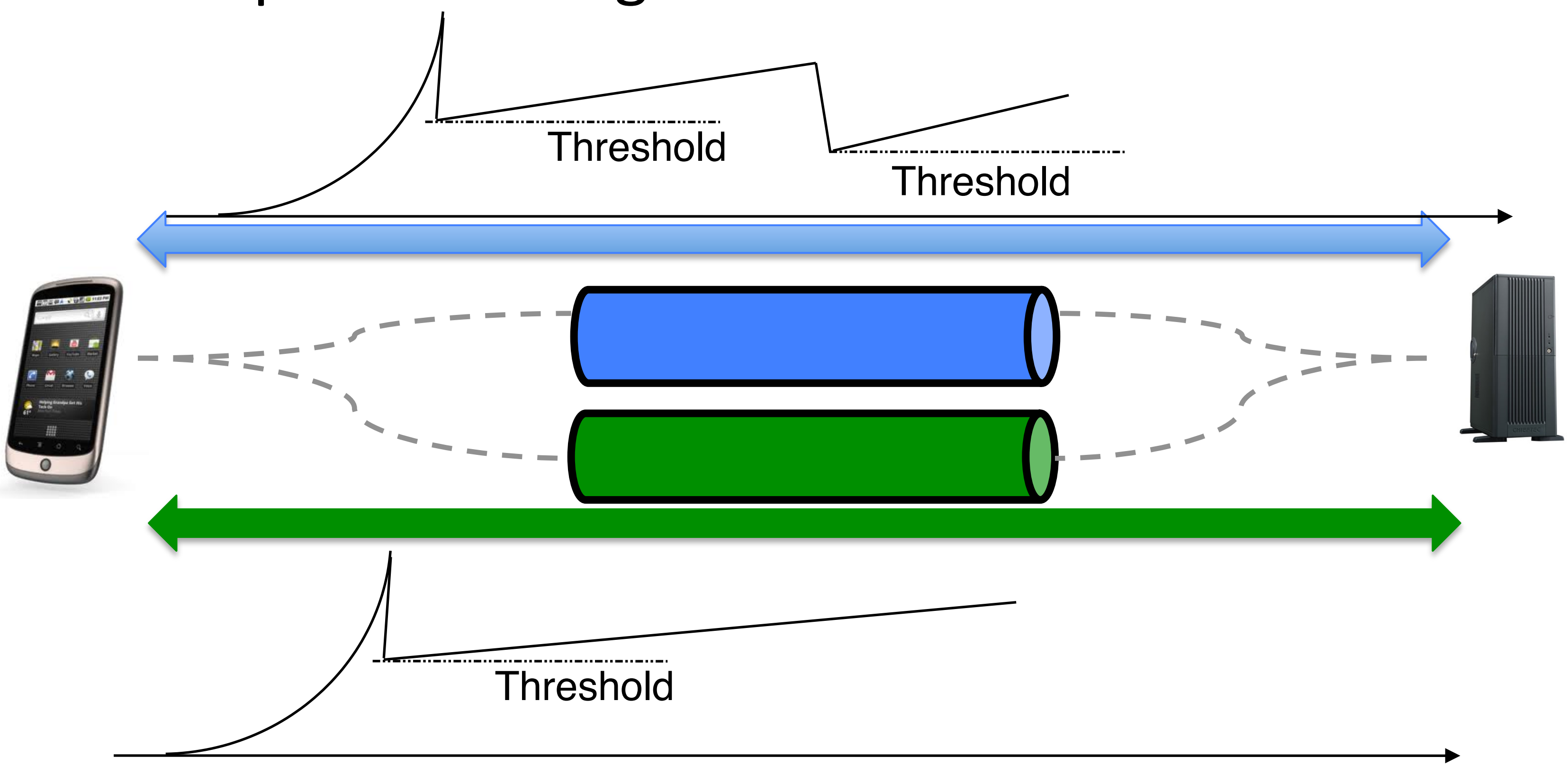
- Congestion control mechanism
  - Each host maintains a *congestion window (cwnd)*
  - No congestion
    - Congestion avoidance (**additive increase**)
      - increase *cwnd* by one segment every round-trip-time
  - Congestion
    - TCP detects congestion by detecting losses
    - Mild congestion (fast retransmit – **multiplicative decrease**)
      - *cwnd*=*cwnd*/2 and restart congestion avoidance
    - Severe congestion (timeout)
      - *cwnd*=1, set slow-start-threshold and restart slow-start

Evolution of the congestion window



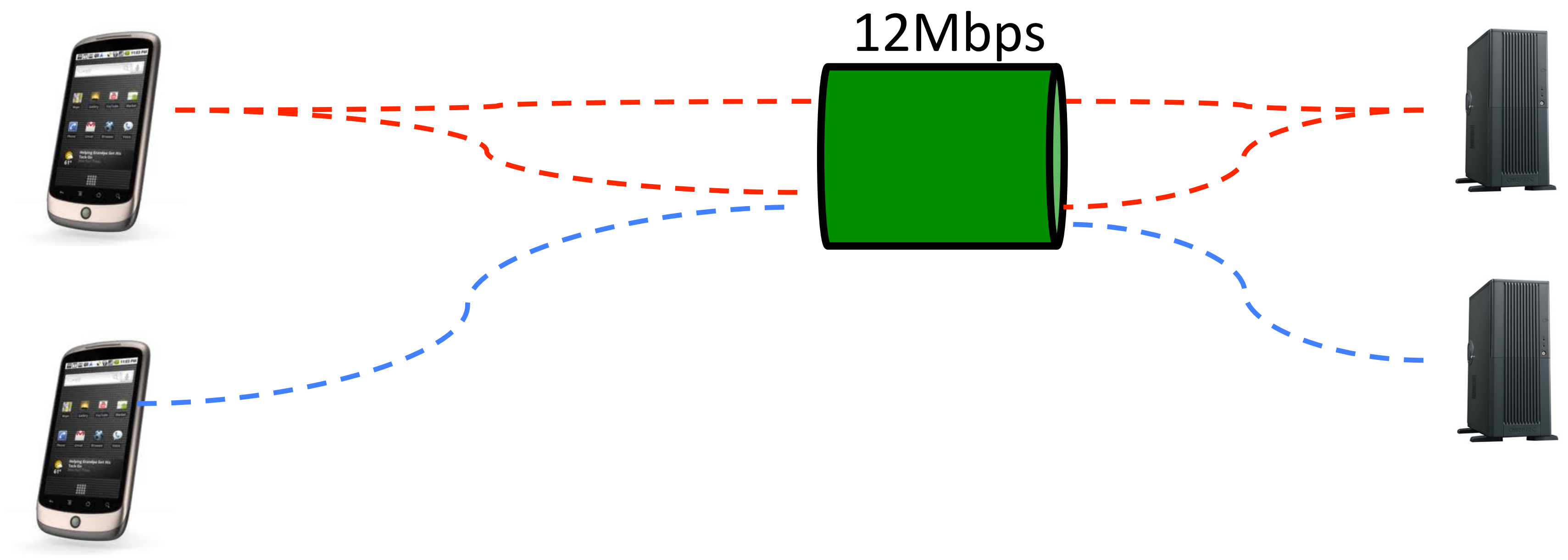
Congestion control for Multipath TCP

- Simple approach
  - independant congestion windows



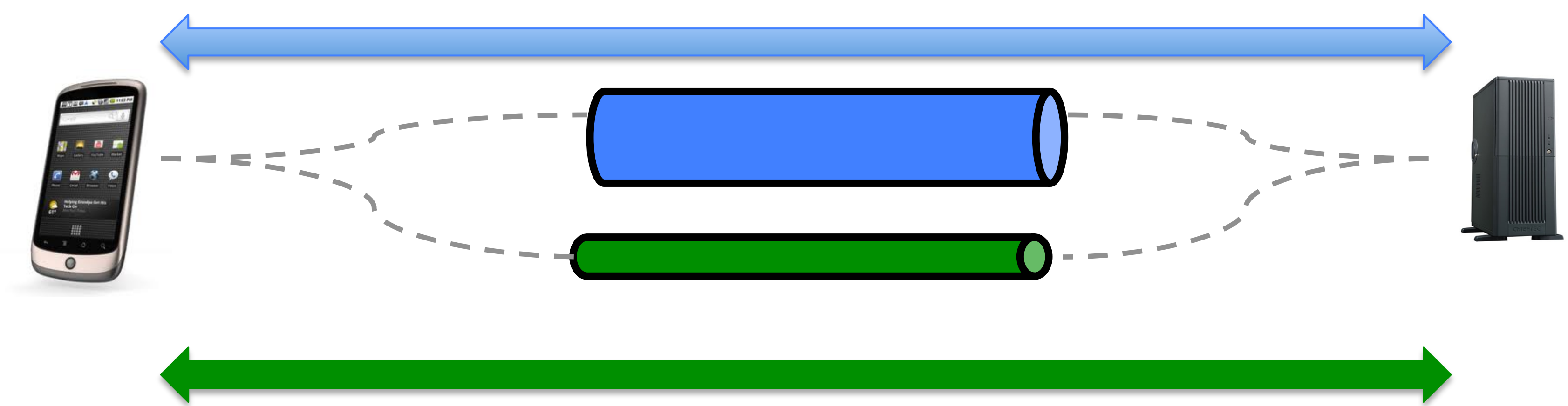
Independant congestion windows

- Problem



Coupled congestion control

- Congestion windows are coupled
  - congestion window growth cannot be faster than TCP with a single flow
  - Coupled congestion control aims at **moving traffic away from congested path**

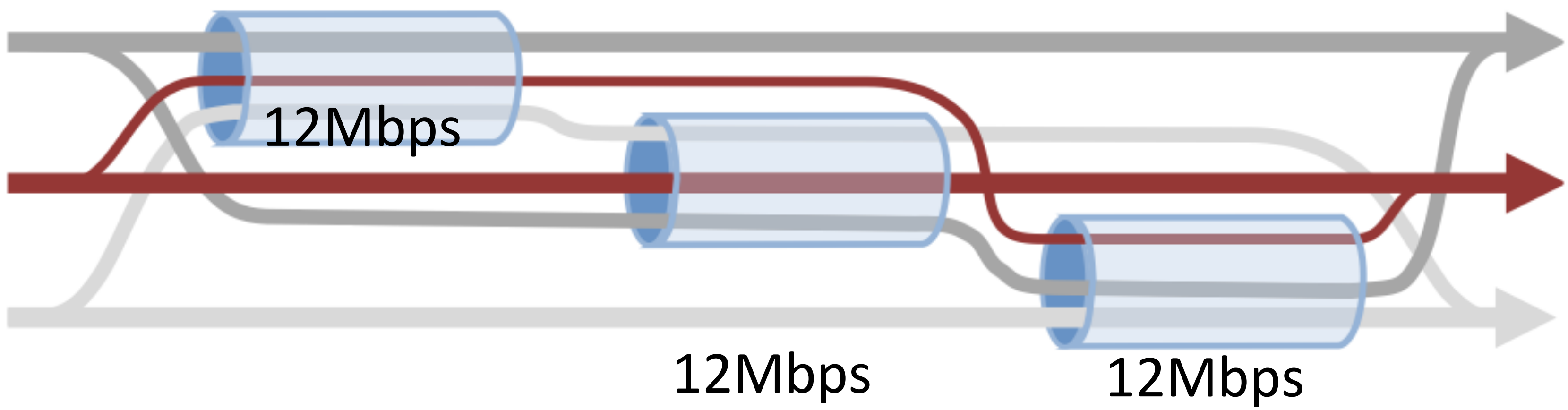


Coupling the congestion windows

- Principle
  - The TCP subflows are not independant and their congestion windows must be coupled
- EWTCP
  - For each ACK on path r, $cwin_r = cwin_r + a/cwin_r$ (in segments)
  - For each loss on path r, $cwin_r = cwin_r/2$
  - Each subflow gets window size proportional to a^2
  - Same throughput as TCP if $a = \frac{1}{\sqrt{n}}$

M. Honda, Y. Nishida, L. Eggert, P. Sarolahti, and H. Tokuda. Multipath Congestion Control for Shared Bottleneck. In *Proc. PFLDNeT workshop*, May 2009.

Can we split traffic equally among all subflows ?



In this scenario, EWTCP would get 3.5 Mbps on the two hops path and 5 Mbps on the one hop path, less than the optimum of 12 Mbps for each Multipath TCP connection

D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, "Design, implementation and evaluation of congestion control for multipath TCP," NSDI'11: Proceedings of the 8th USENIX conference on Networked systems design and implementation, 2011.

Linked increases congestion control

- Algorithm
  - For each loss on path r, $cwin_r = cwin_r / 2$
  - Additive increase

$$cwin_r = cwin_r + \min\left(\frac{\max\left(\frac{cwnd_i}{(rtt_i)^2}\right)}{\left(\sum_i \frac{cwnd_i}{rtt_i}\right)^2}, \frac{1}{cwnd_r}\right)$$

D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, "Design, implementation and evaluation of congestion control for multipath TCP," NSDI'11: Proceedings of the 8th USENIX conference on Networked systems design and implementation, 2011.

Other Multipath-aware congestion control schemes

R. Khalili, N. Gast, M. Popovic, U. Upadhyay, J.-Y. Le Boudec , MPTCP is not Pareto-optimal: Performance issues and a possible solution, Proc. ACM Conext 2012

Y. Cao, X. Mingwei, and X. Fu, "Delay-based Congestion Control for Multipath TCP," ICNP2012, 2012.

T. A. Le, C. S. Hong, and E.-N. Huh, "Coordinated TCP Westwood congestion control for multiple paths over wireless networks," ICOIN '12: Proceedings of the The International Conference on Information Network 2012, 2012, pp. 92–96.

T. A. Le, H. Rim, and C. S. Hong, "A Multipath Cubic TCP Congestion Control with Multipath Fast Recovery over High Bandwidth-Delay Product Networks," *IEICE Transactions*, 2012.

T. Dreibholz, M. Becke, J. Pulinthanath, and E. P. Rathgeb, "Applying TCP-Friendly Congestion Control to Concurrent Multipath Transfer," Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, 2010, pp. 312–319.

The Multipath TCP protocol

- **Control plane**
  - How to manage a Multipath TCP connection that uses several paths ?
- Data plane
  - How to transport data ?
- Congestion control
  - How to control congestion over multiple paths ?

The Multipath TCP control plane

- Connection establishment

- Closing a Multipath TCP connection

- Address dynamics

The Multipath TCP control plane

- Connection establishment
  - Beware of middleboxes that remove TCP options
  - Limited space inside TCP option in SYN
- Closing a Multipath TCP connection
  - Decouple closing the Multipath TCP connection from closing the subflows
- Address dynamics

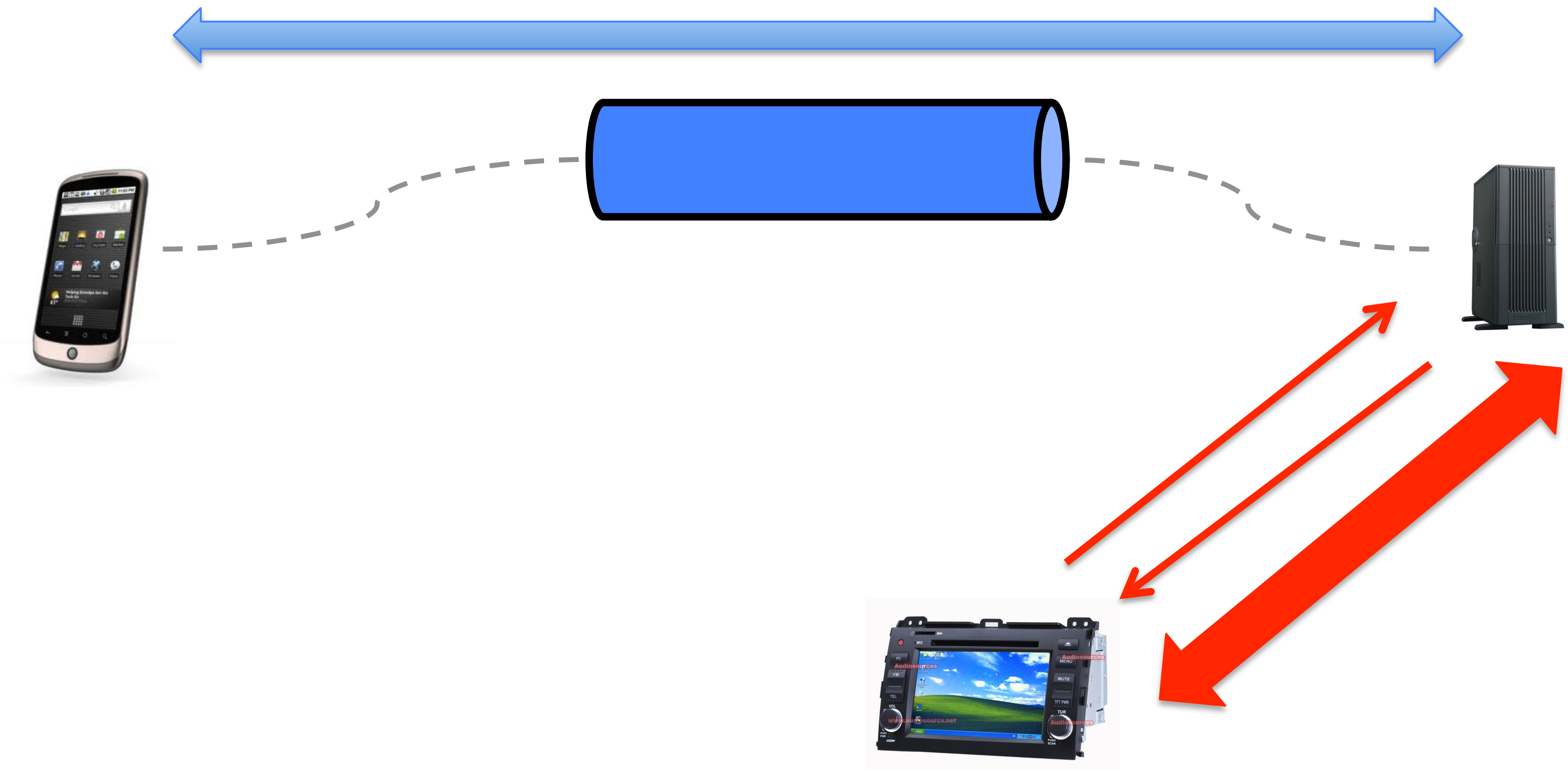
Security threats

- Three main security threats were considered
  - flooding attack
  - man-in-the middle attack
  - hijacking attach

Security goal :
Multipath TCP should not be worse than regular TCP

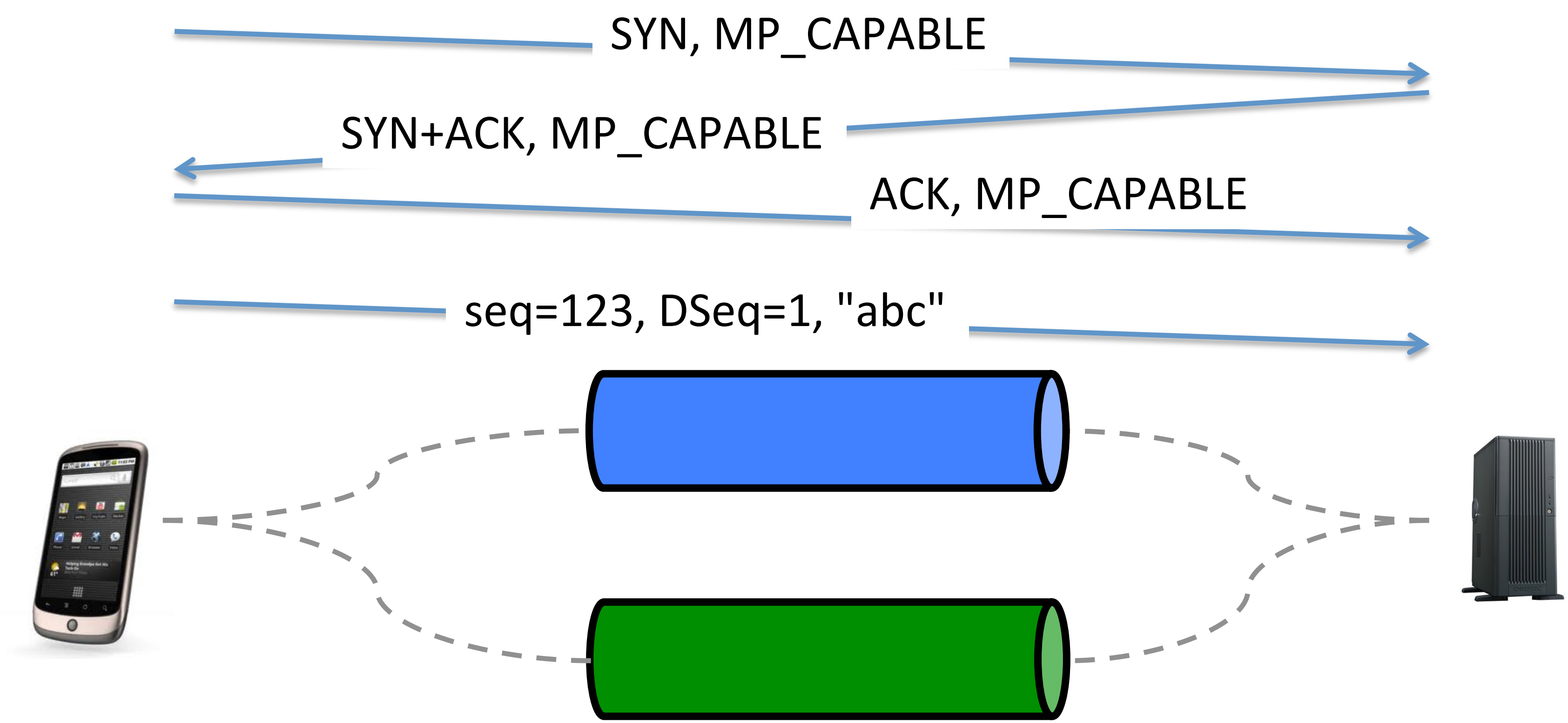
J. Diez, M. Bagnulo, F. Valera, and I. Vidal, "Security for multipath TCP: a constructive approach," *International Journal of Internet Protocol Technology*, vol. 6, 2011.

Hijacking attack



Multipath TCP
Connection establishment

- Principle

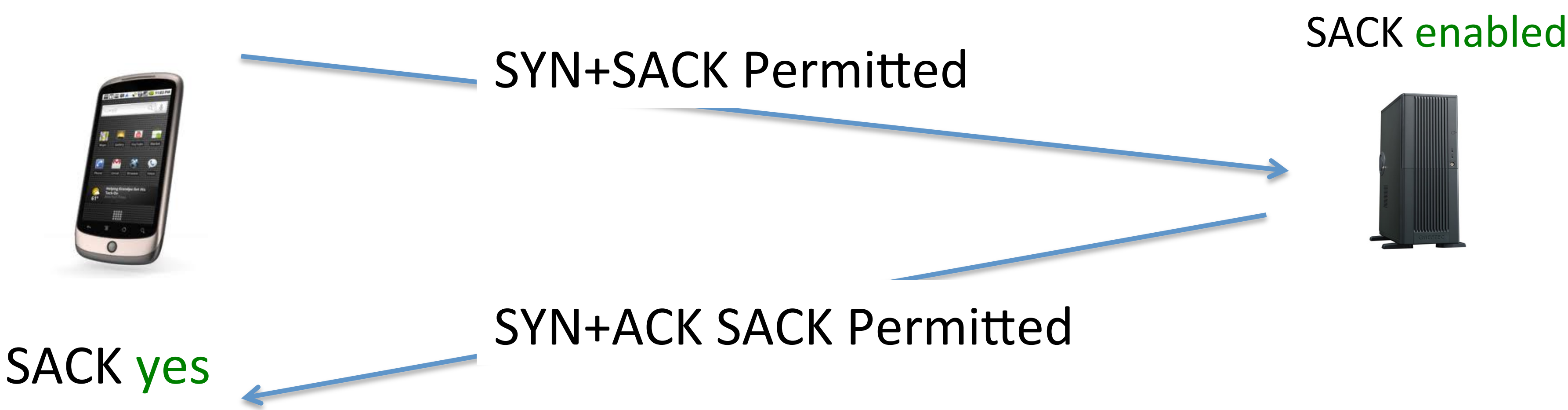


Roles of the initial TCP handshake

- Check willingness to open TCP connection
  - Propose initial sequence number
  - Negotiate Maximum Segment Size
- TCP options
  - negotiate Timestamps, SACK, Window scale
- Multipath TCP
  - check that server supports Multipath TCP
  - propose Token in each direction
  - propose initial Data sequence number in each direction
  - Exchange keys to authenticate subflows

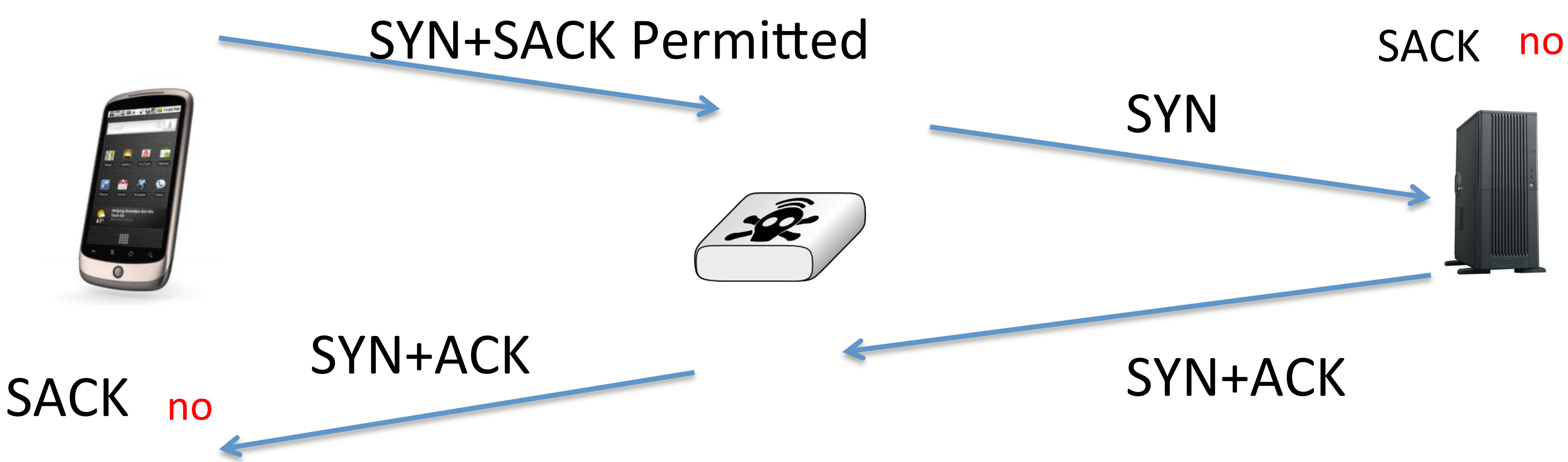
How to extend TCP ?
Theory

- TCP options were invented for this purpose
  - Exemple SACK



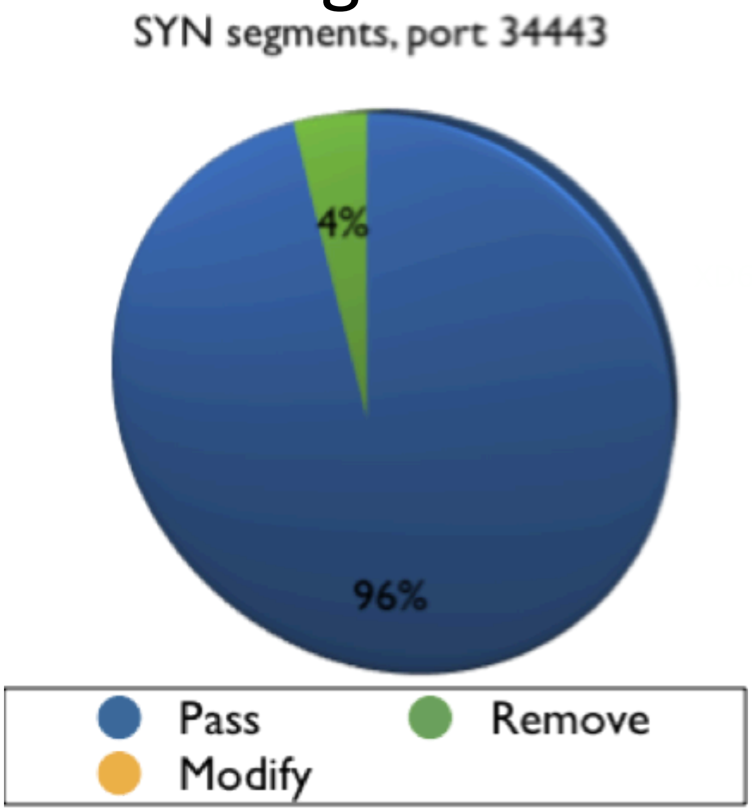
How to extend TCP ? practice

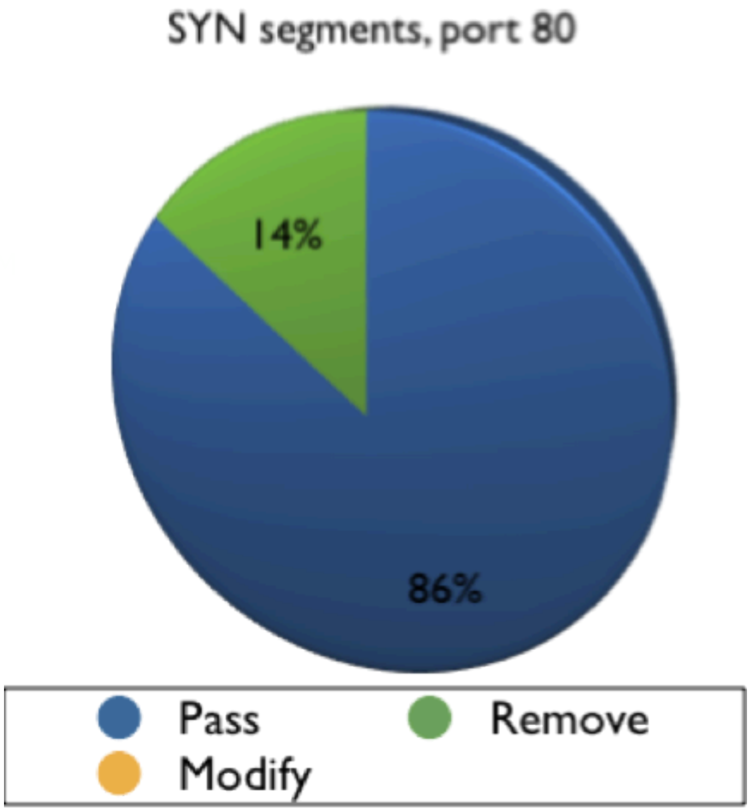
- What happens when there are middleboxes on the path ?



TCP options

- In SYN segments

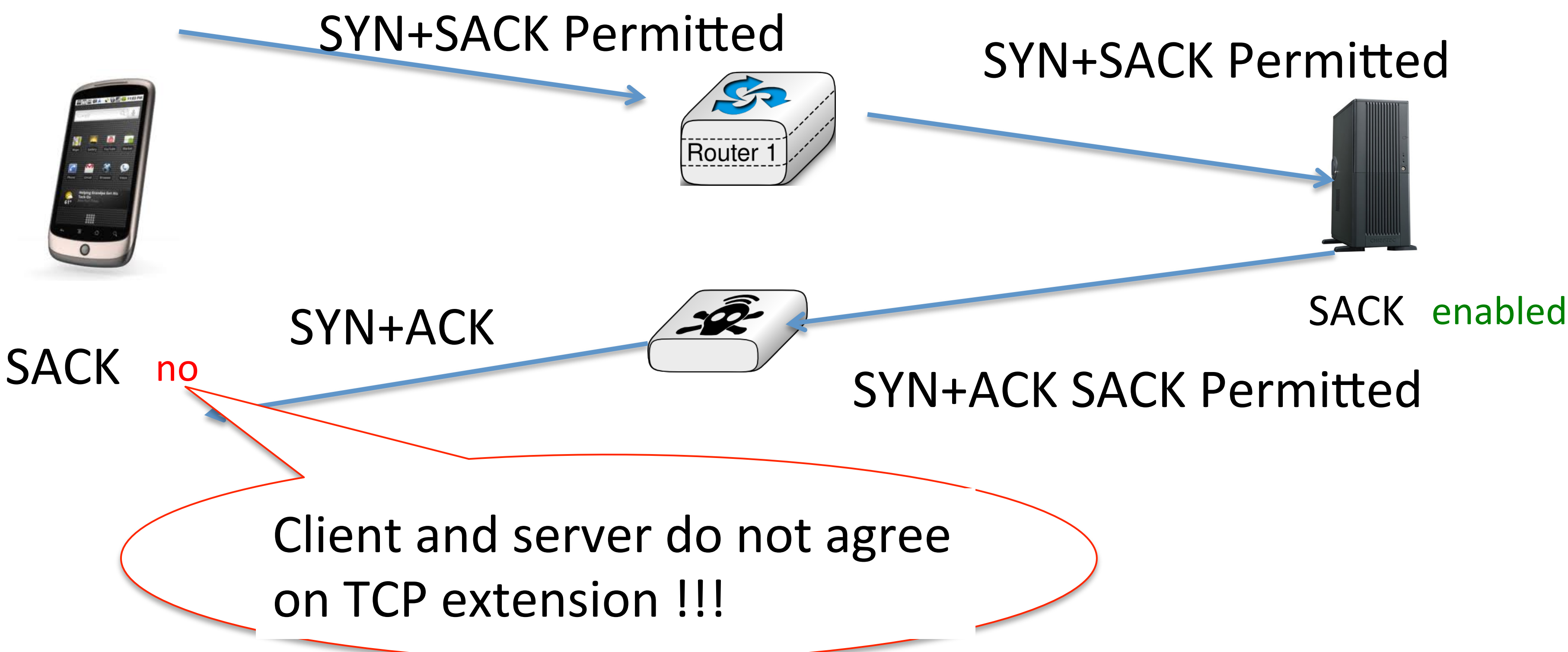




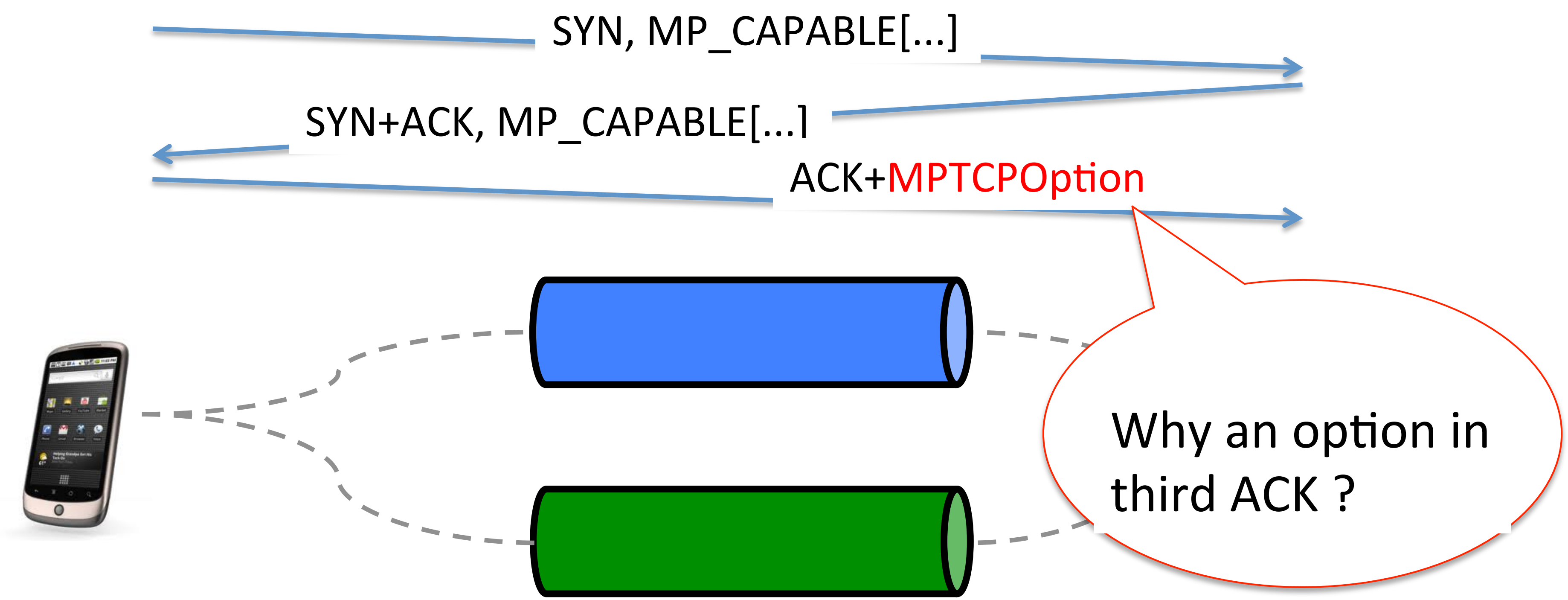
Honda, Michio, et al. "Is it still possible to extend TCP?." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011.

How to extend TCP ?
The worst case

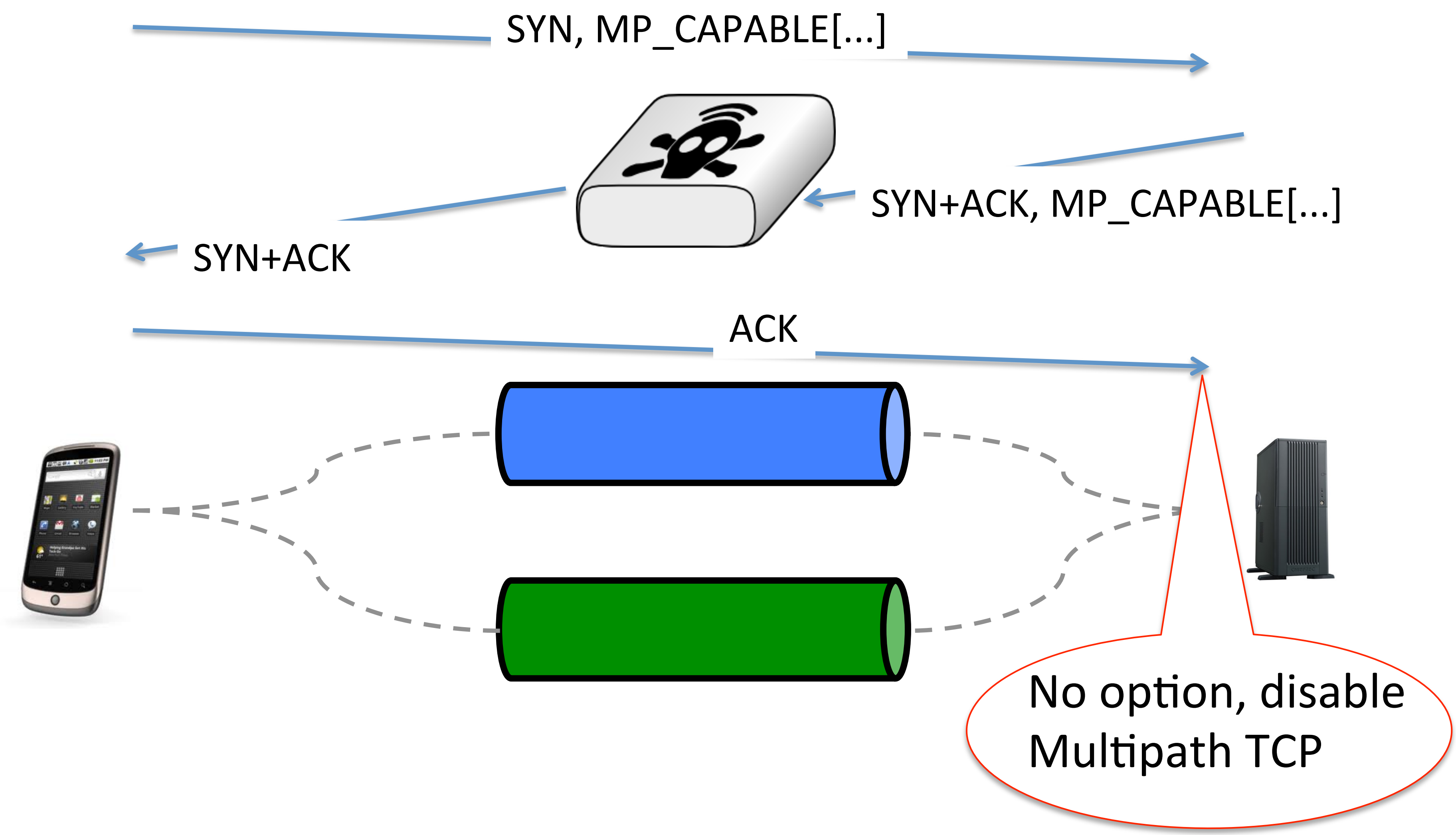
- What happens when there are middleboxes on the path ?



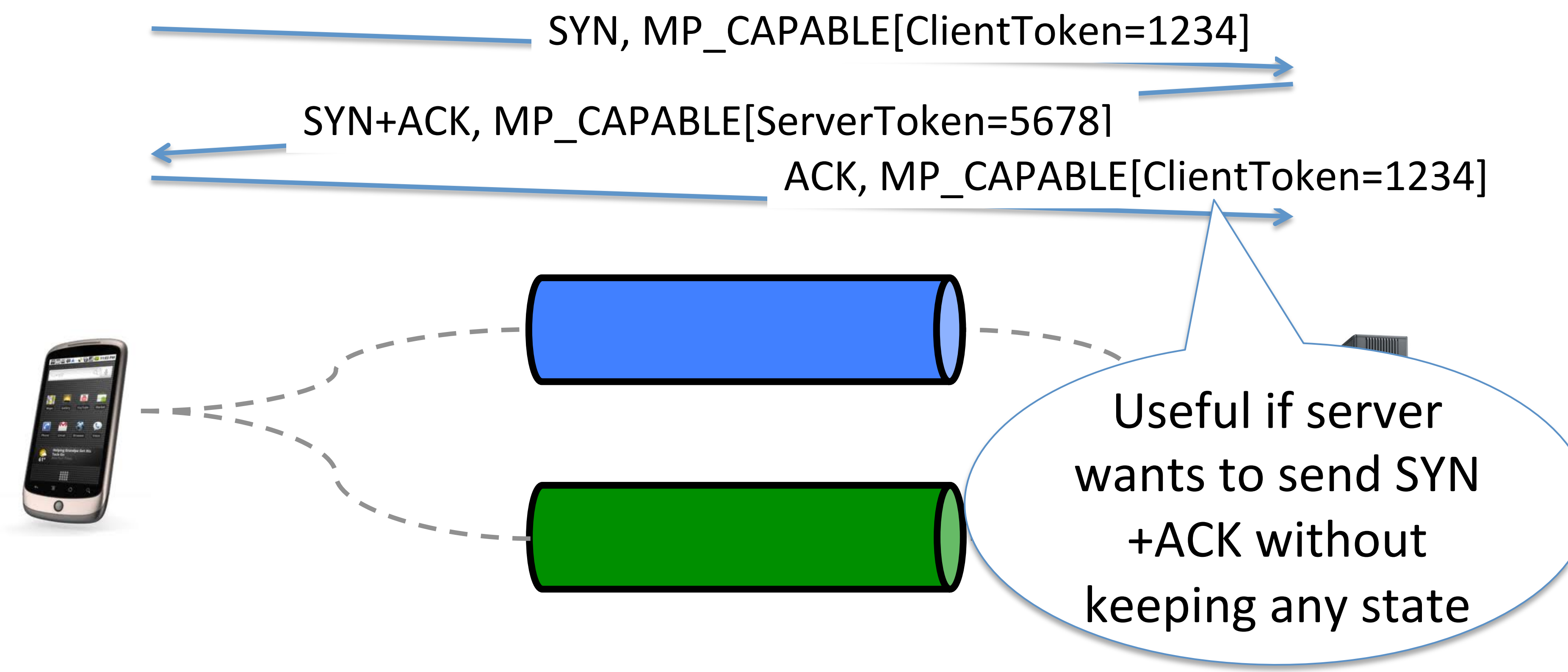
Multipath TCP handshake



Multipath TCP option in third ACK



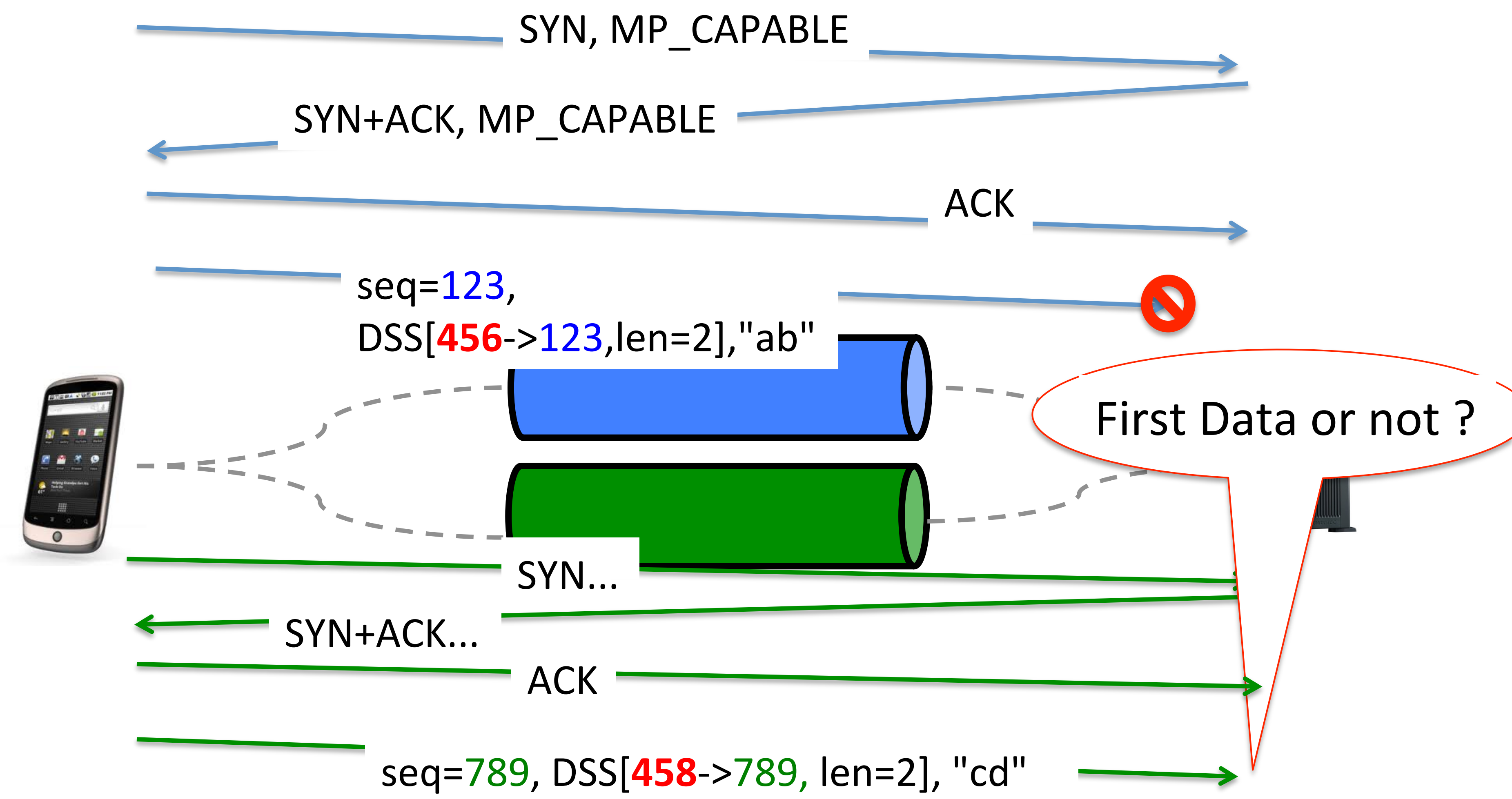
Multipath TCP handshake
Token exchange



Initial Data Sequence number

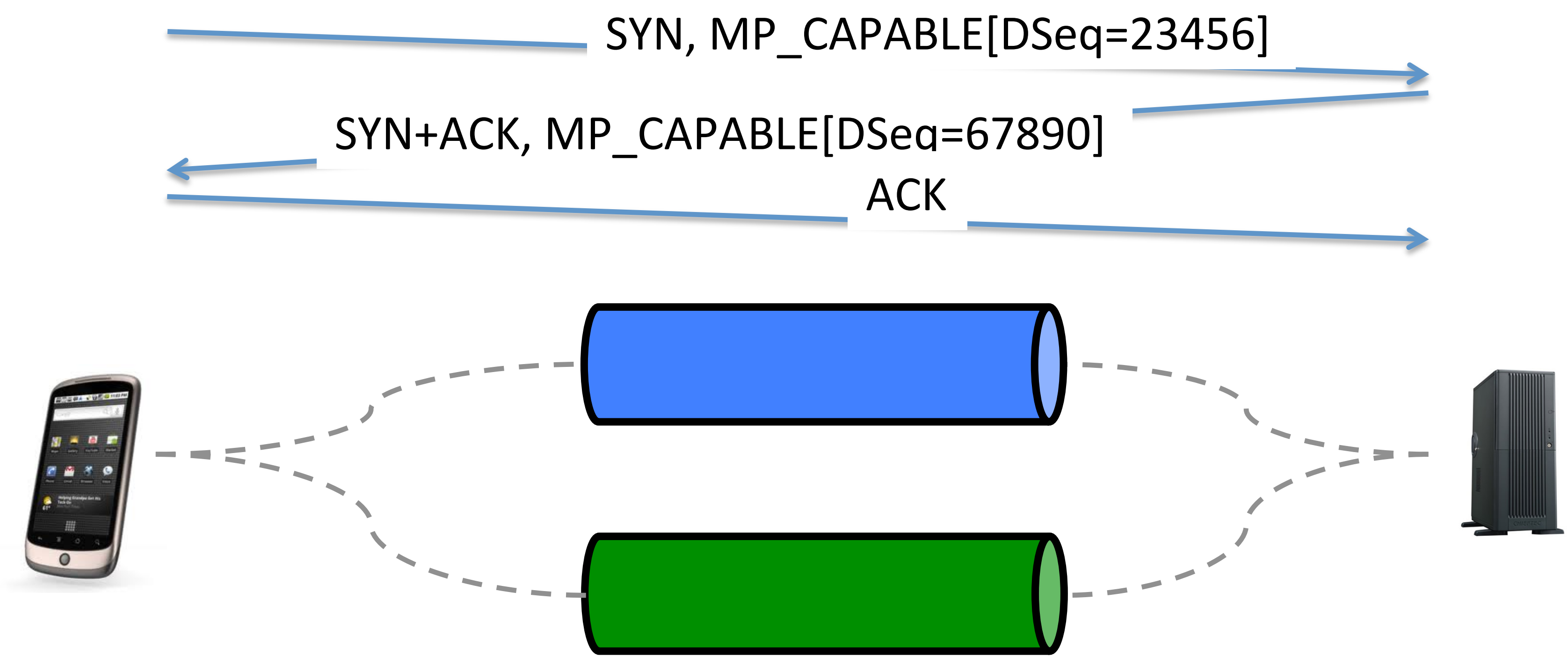
- Why do we need an initial Data Sequence number ?
  - Setting IDSN to a random value improves security
  - Hosts must know IDSN to prevent data injection attacks

Initial Data Sequence number



Initial Data Sequence number

- How to negotiate the IDSN ?



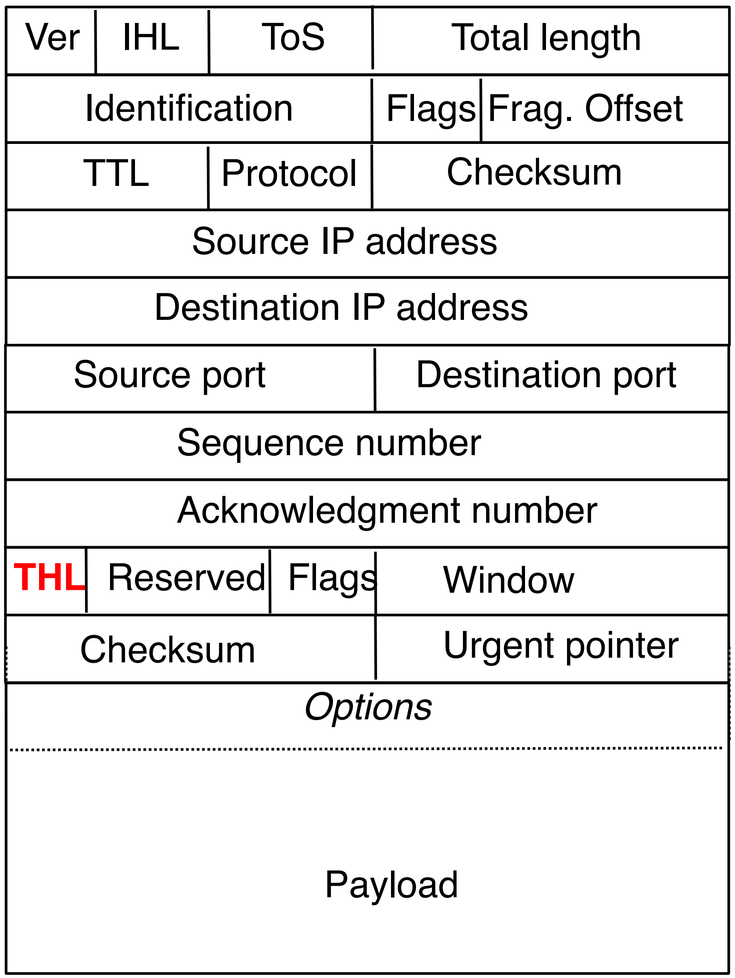
How to secure Multipath TCP

- Main goal
  - Authenticate the establishment of subflows
- Principles
  - Each host announces a key during initial handshake
    - keys are exchanged in clear
  - When establishing a subflow, use HMAC + key to authenticate subflow

Putting everything inside the SYN

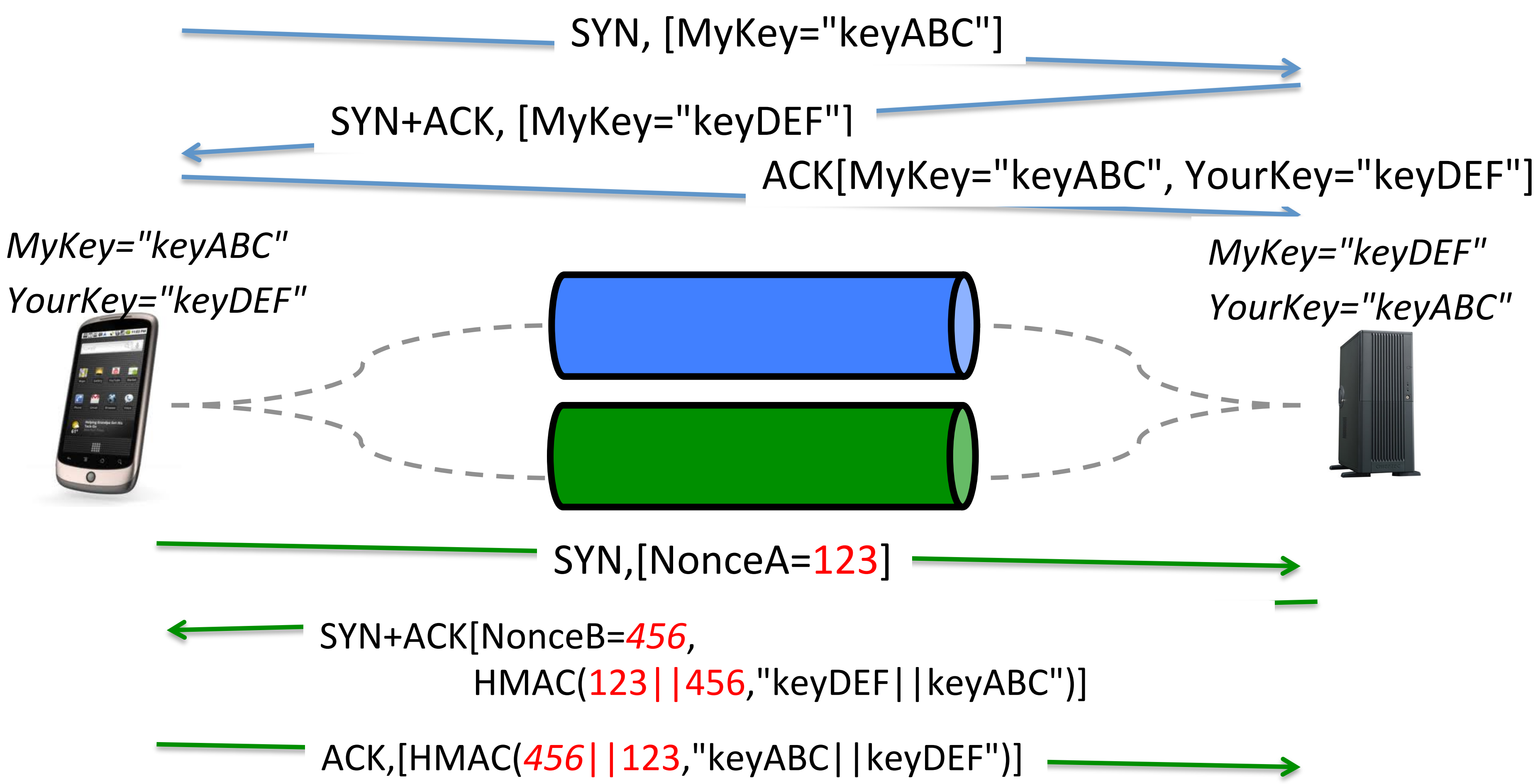
- How can we place inside SYN segment ?
  - Initial Data Sequence Number (64 bits)
  - Token (32 bits)
  - Authentication Key (the longer the better)

Constraint on TCP options



- Total length of TCP header : max 64 bytes
  - max 40 bytes for TCP options
  - *Options* length must be multiple of 4 bytes

Key exchange



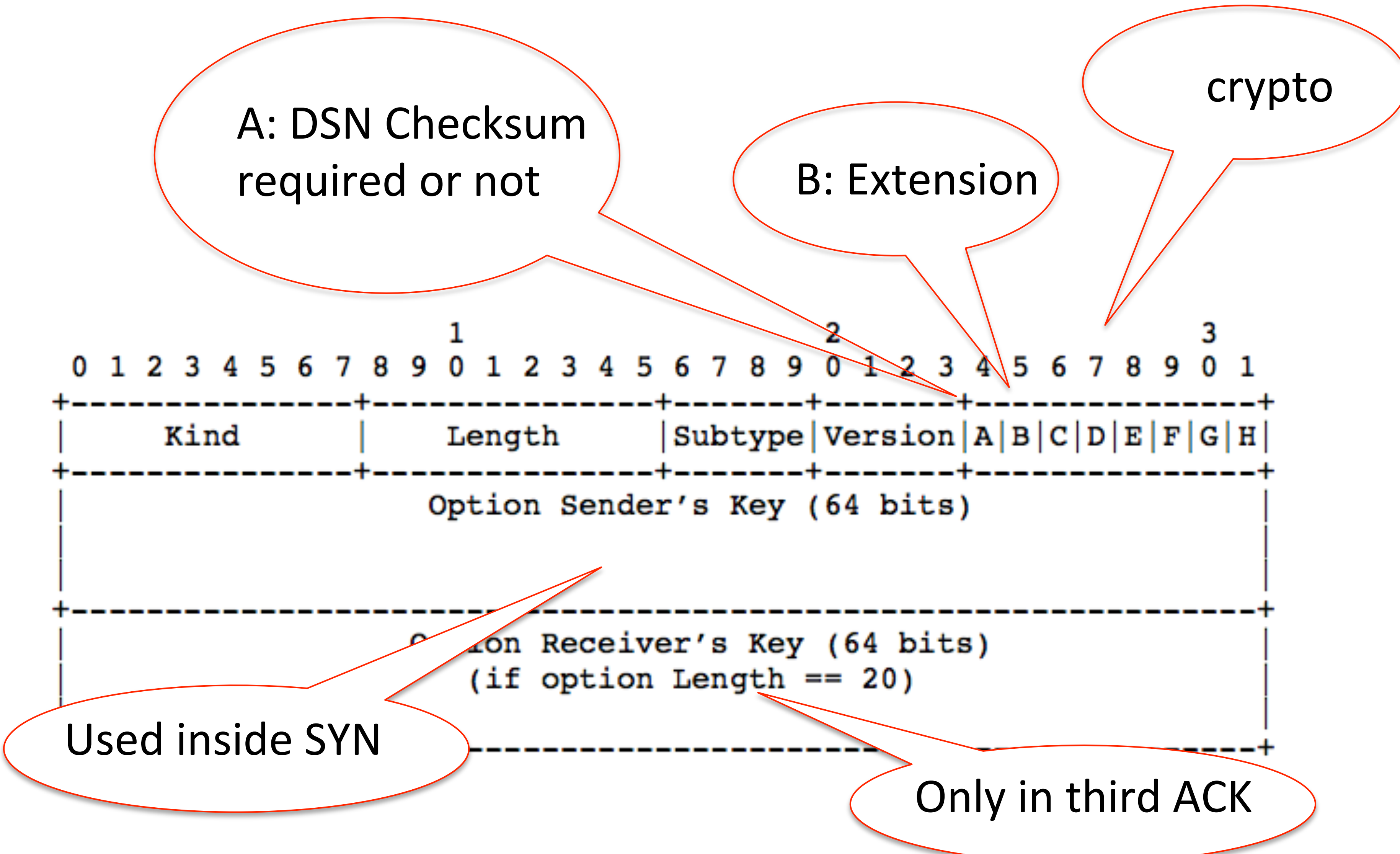
TCP options in the wild

- **MSS** option [4 bytes]
  - Used only inside SYN segments
- **Timestamp** option [10 bytes]
  - Used in potentially all segments
- **Window scale** option [3 bytes]
  - Used only inside SYN segments
- **SACK permitted** option [2 bytes]
  - Used only inside SYN segments
- **Selective Acknowledgements** [N bytes]
  - Used in data segments

Only 20 bytes left inside SYN !

http://www.iana.org/assignments/tcp-parameters/tcp-parameters.xml

The MP_CAPABLE option



Initial Data Sequence Numbers and Tokens

- Computation of initial Data Sequence Number

$IDSN_A = Lower_{64}(SHA\text{-}1(Key_A))$

$IDSN_B = Lower_{64}(SHA\text{-}1(Key_B))$

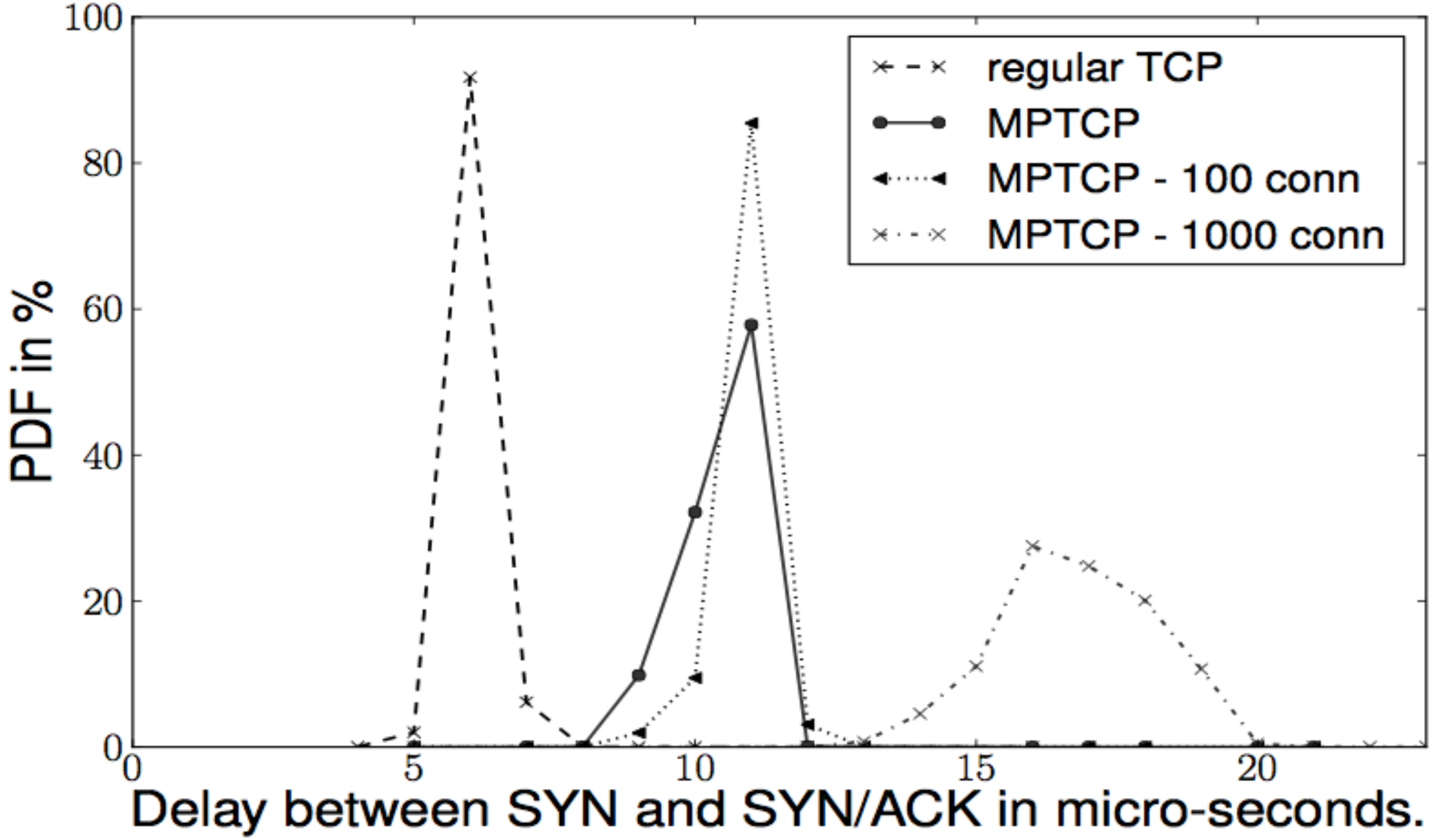
- Computation of token

$Token_A = Upper_{32}(SHA\text{-}1(Key_A))$

$Token_B = Upper_{32}(SHA\text{-}1(Key_B))$

There is a small risk of collision, different keys same token

Cost of the Multipath TCP handshake



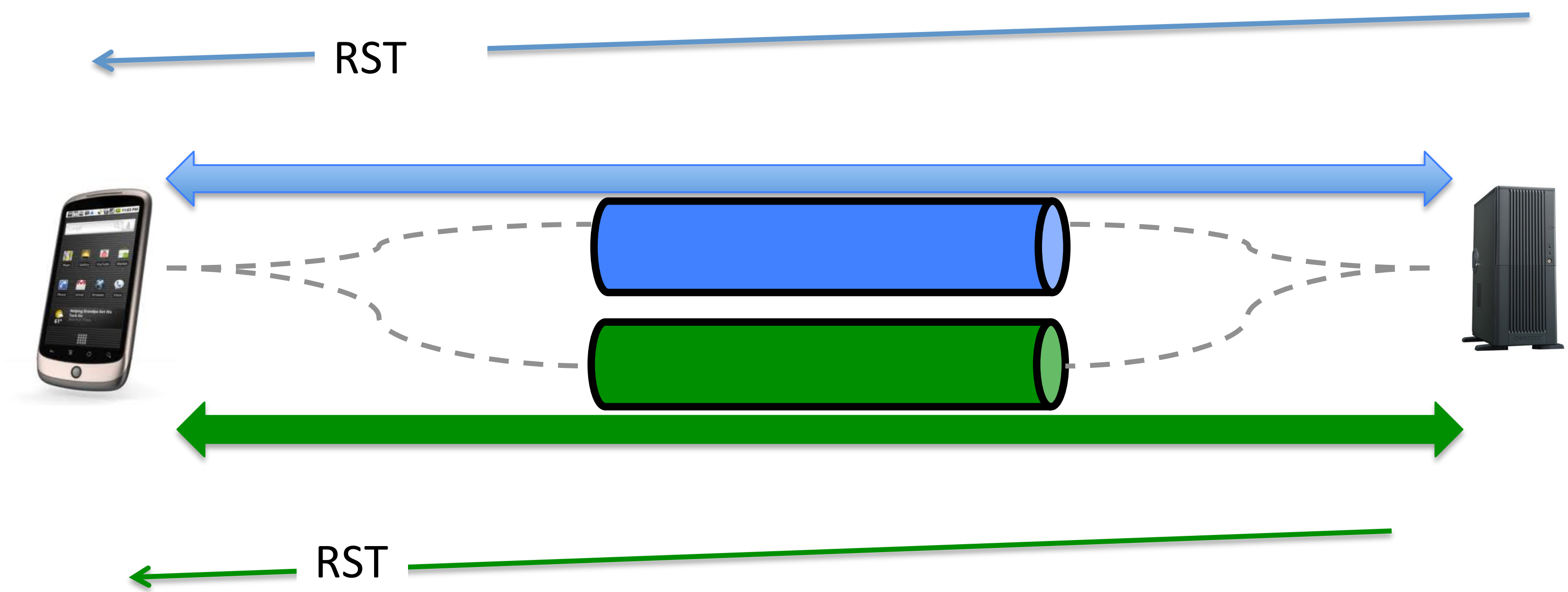
C. Raiciu, et al. "How hard can it be? designing and implementing a deployable multipath TCP," NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

The Multipath TCP control plane

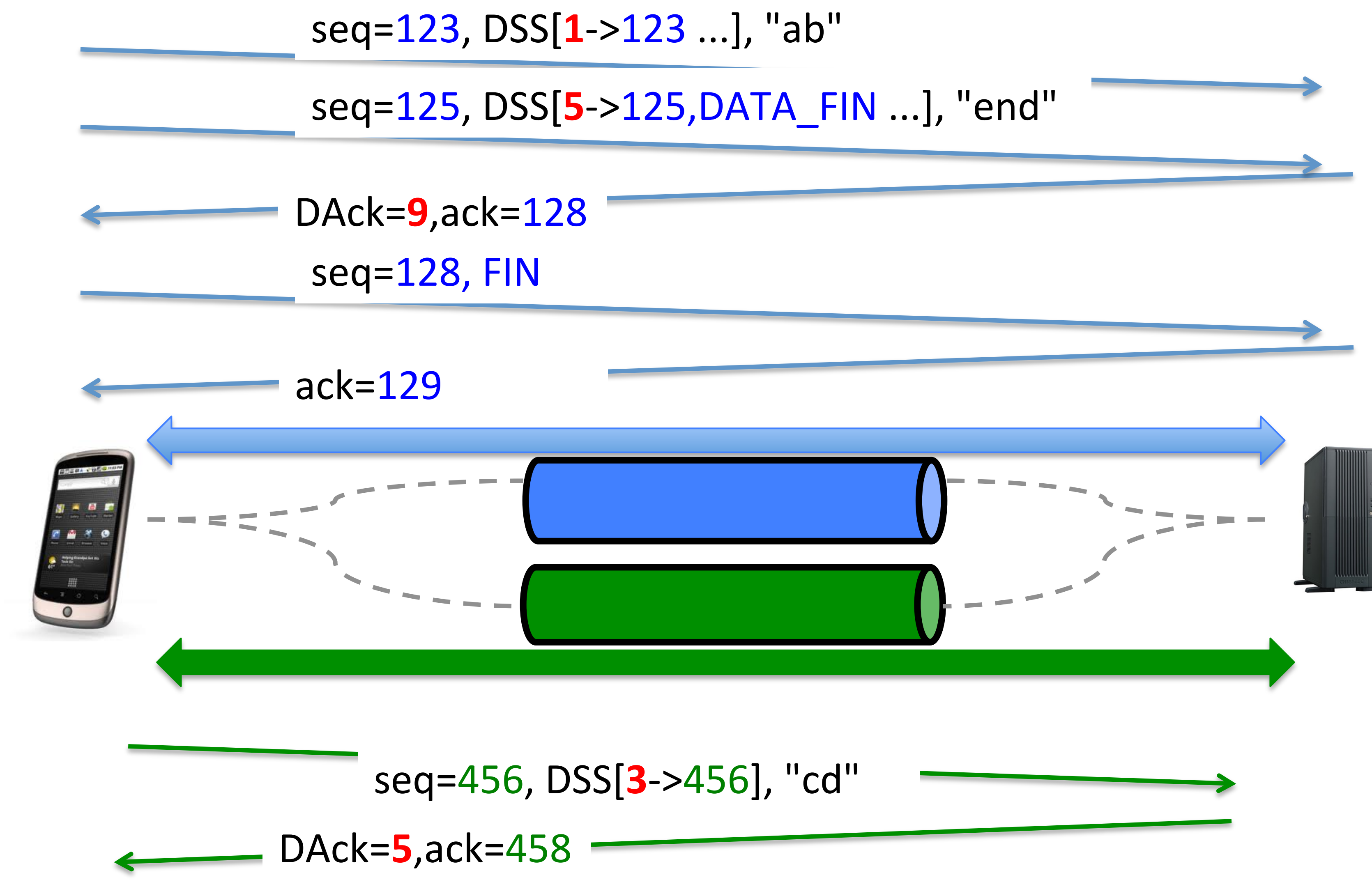
- Connection establishment in details
- Closing a Multipath TCP connection
- Address dynamics

Closing a Multipath TCP connection

- How to close a Multipath TCP connection ?
  - By closing all subflows ?

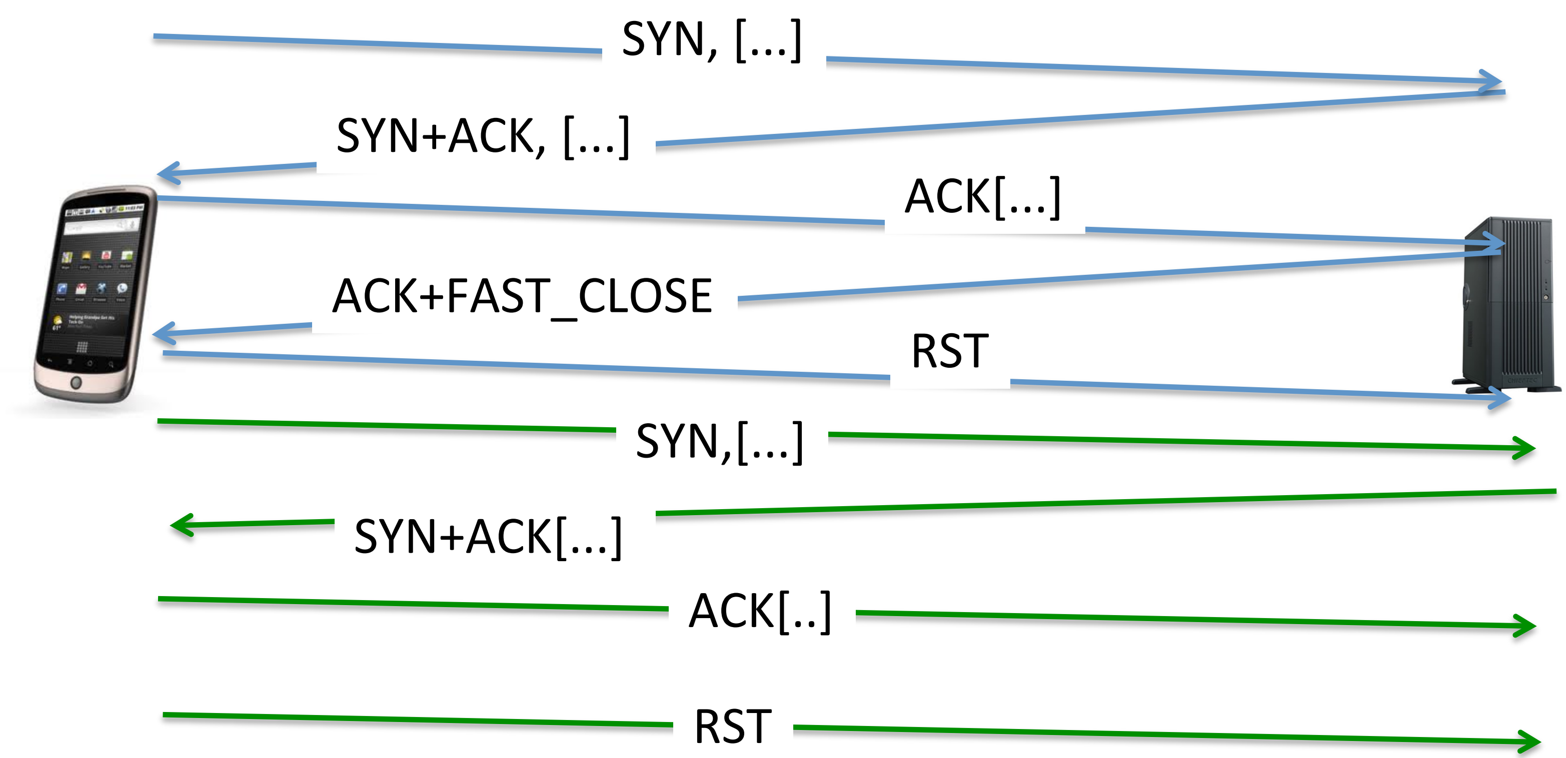


Closing a Multipath TCP connection



Closing a Multipath TCP connection

- FAST Close



The Multipath TCP control plane

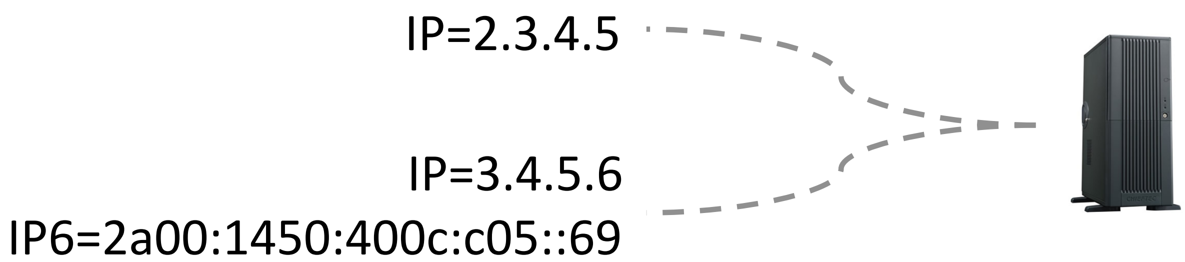
- Connection establishment in details

- Closing a Multipath TCP connection

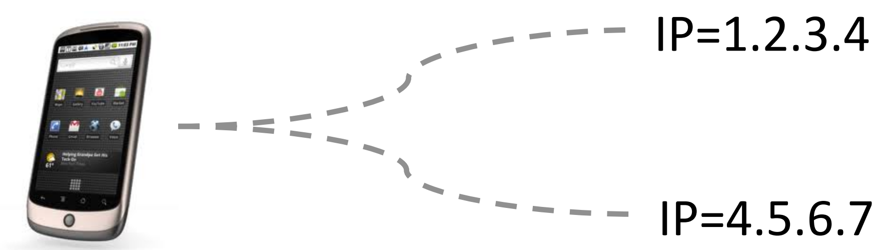
- Address dynamics

Multipath TCP
Address dynamics

- How to learn the addresses of a host ?

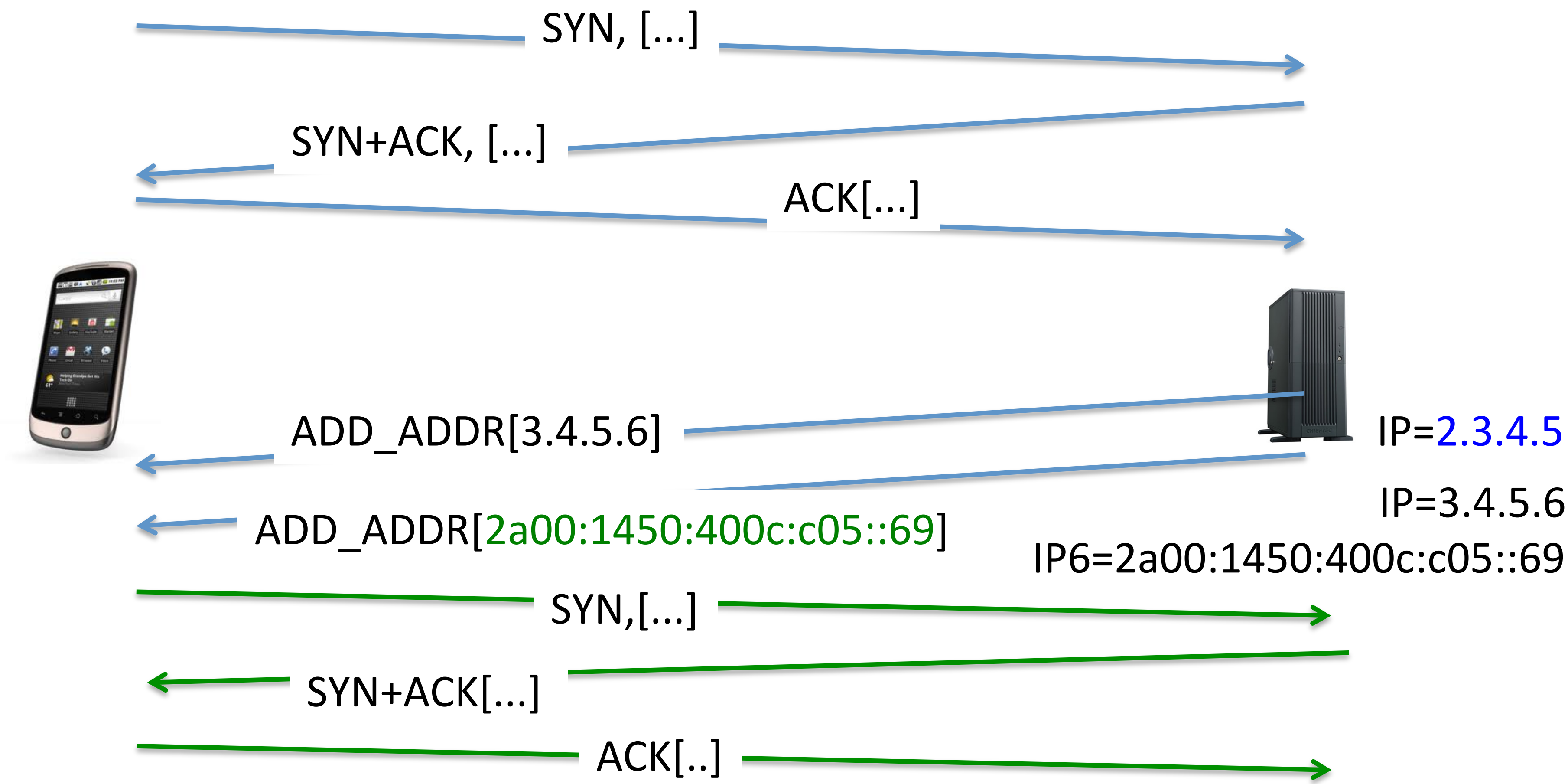


- How to deal with address changes ?



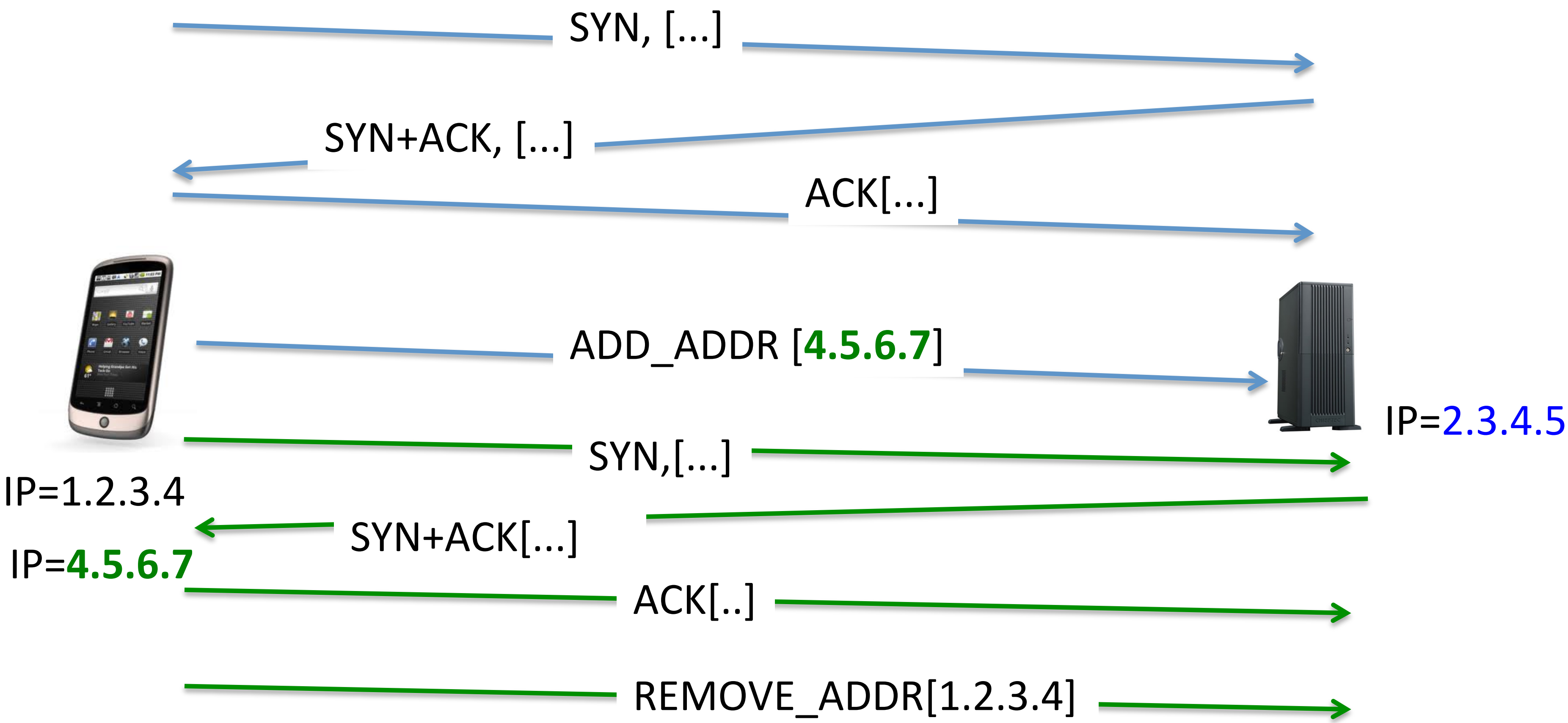
Address dynamics

- Basic solution : multihomed server

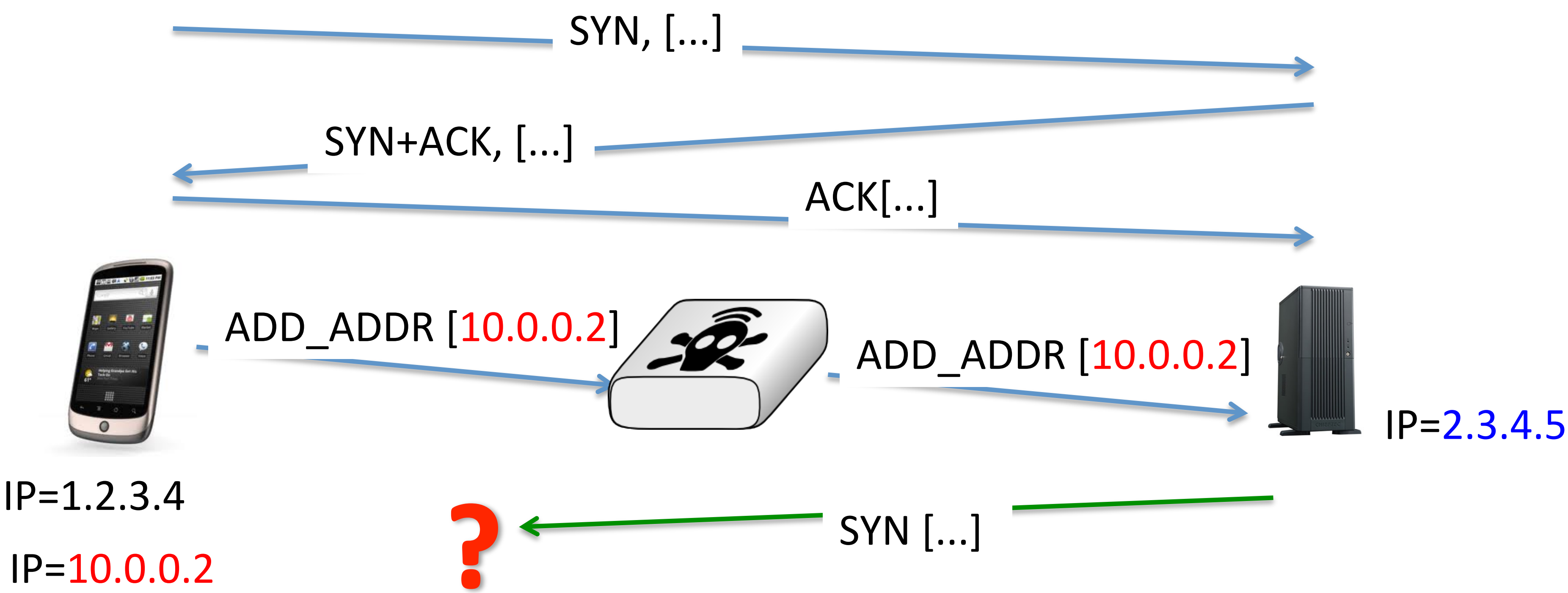


Address dynamics

- Basic solution : mobile client



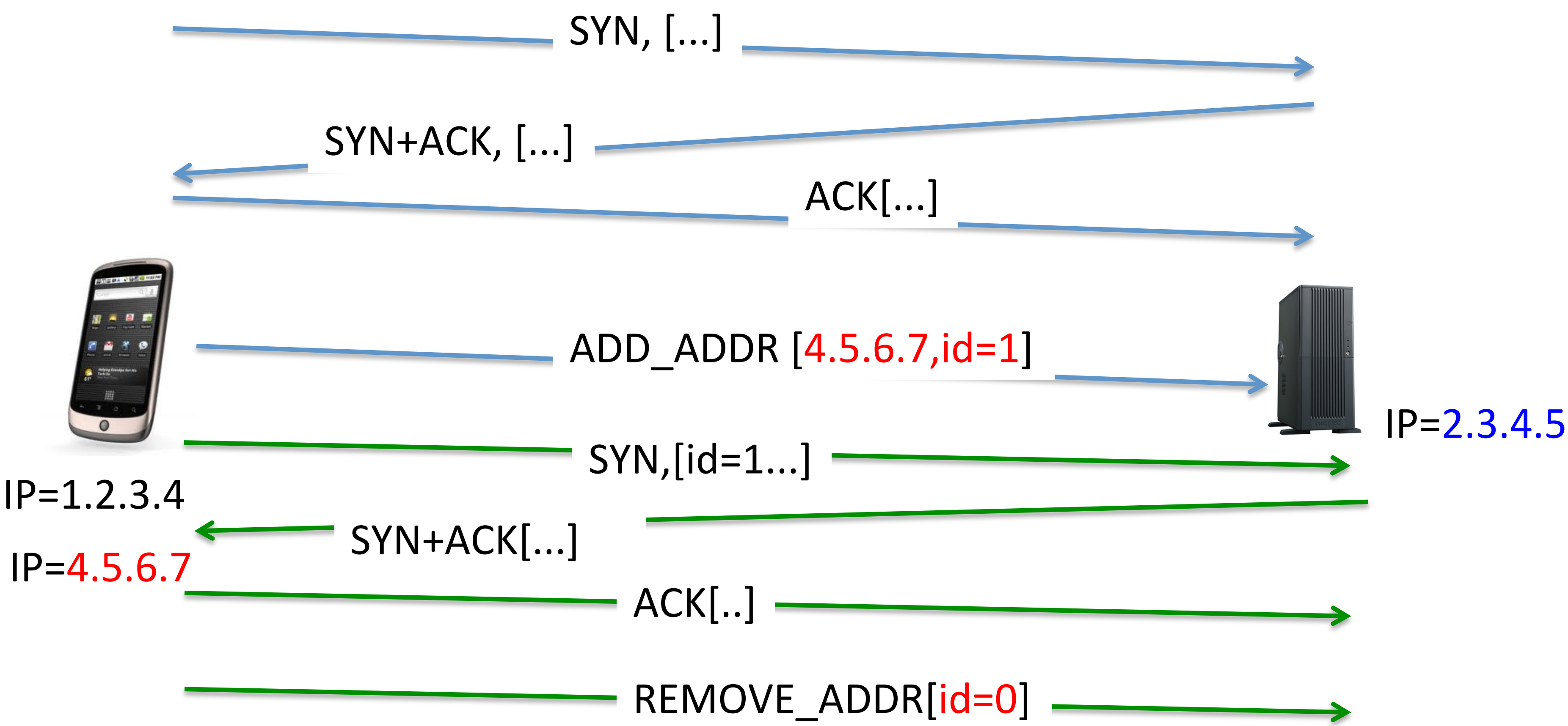
Address dynamics in today's Internet



Address dynamics with NATs

- Solution
  - Each address has one identifier
    - Subflow is established between id=0 addresses
  - Each host maintains a list of <address,id> pairs of the addresses associated to an MPTCP endpoint
  - MPTCP options refer to the address identifier
    - ADD_ADDR contains <address,id>
    - REMOVE_ADDR contains <id>

Address dynamics

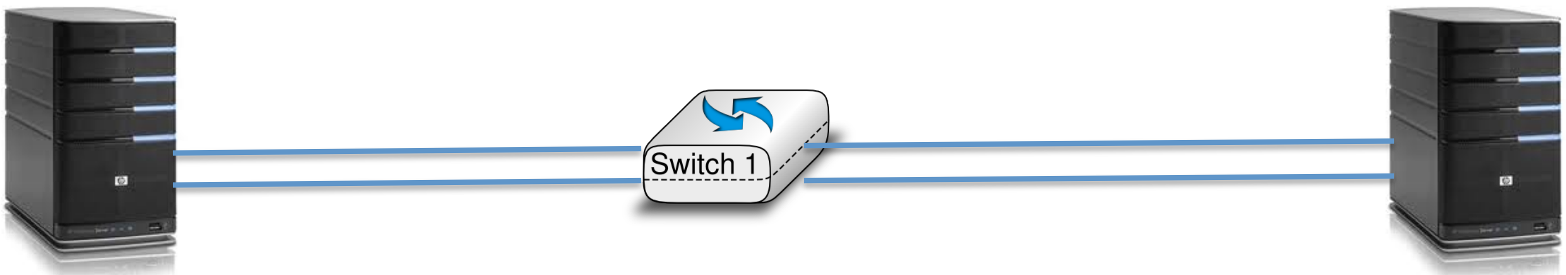


Agenda

- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases
  - Datacenters
  - Smartphones
  - Commercial deployments

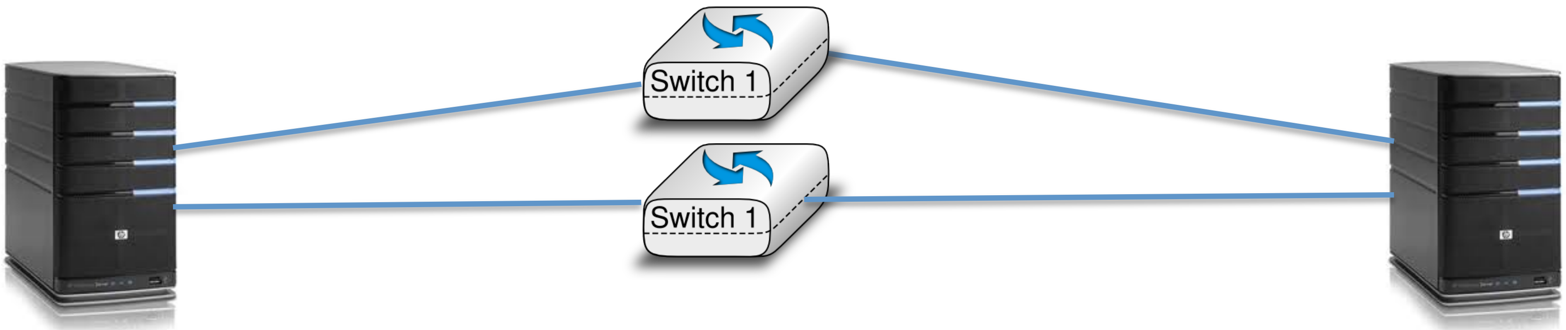
TCP on servers

- How to increase server bandwidth ?



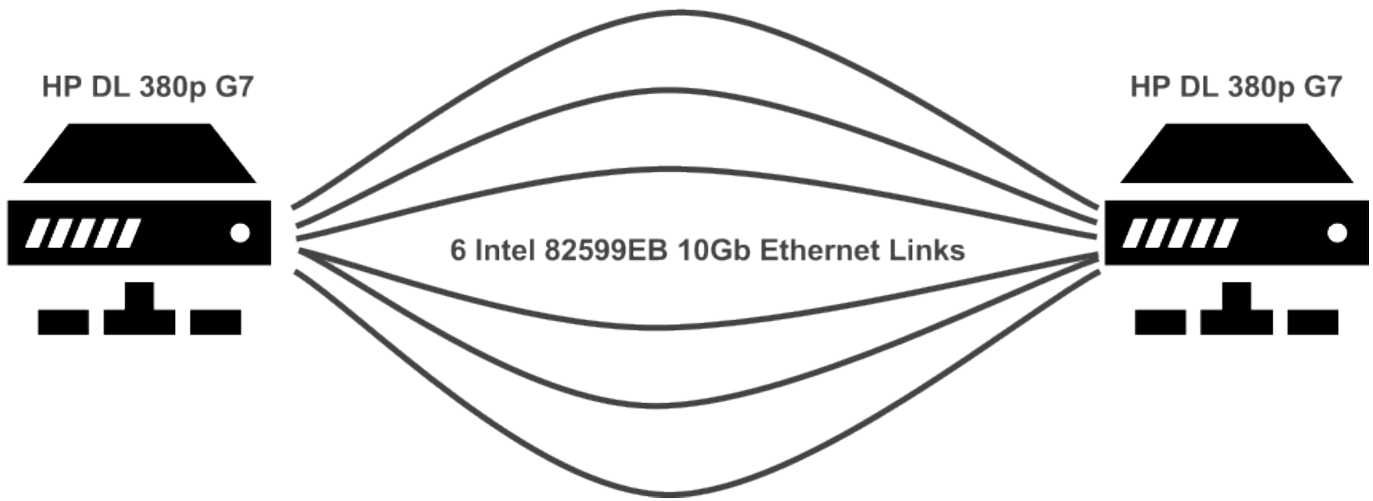
- Load balancing techniques
  - packet per packet
  - per flow load balancing
    - each TCP connection is mapped onto one interface

Increasing server bandwidth with Multipath TCP



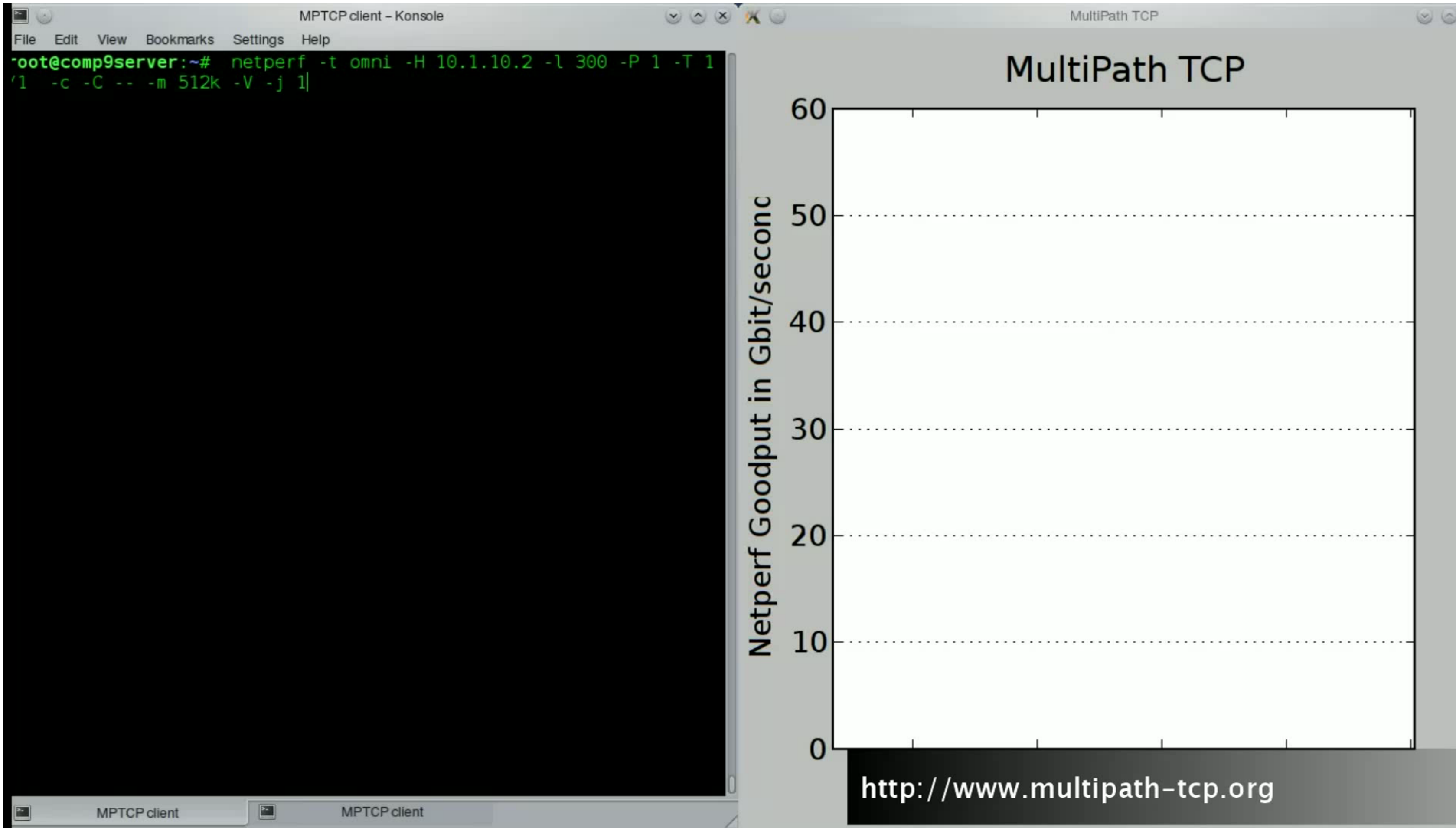
- Load balancing with Multipath TCP
  - Congestion control efficiently uses the two links for each MPTCP connection
  - Automatic failover in case of failures

How fast can Multipath TCP go ?



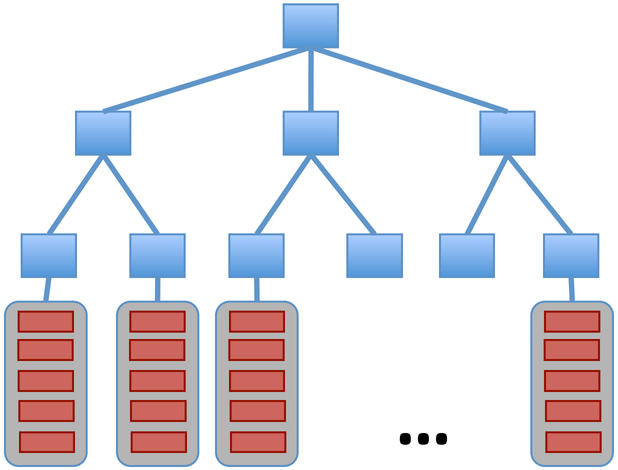
http://linux.slashdot.org/story/13/03/23/0054252/a-50-gbps-connection-with-multipath-tcp

How fast can Multipath TCP go ?



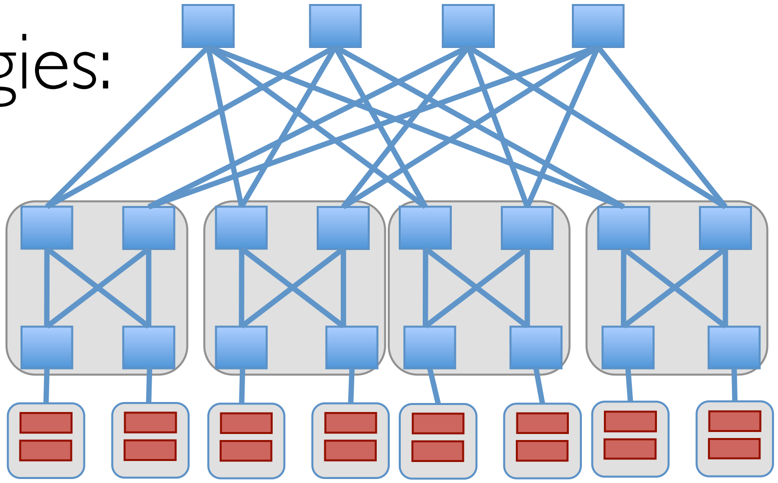
Datacenters evolve

- Traditional Topologies are tree-based
  - Poor performance
  - Not fault tolerant



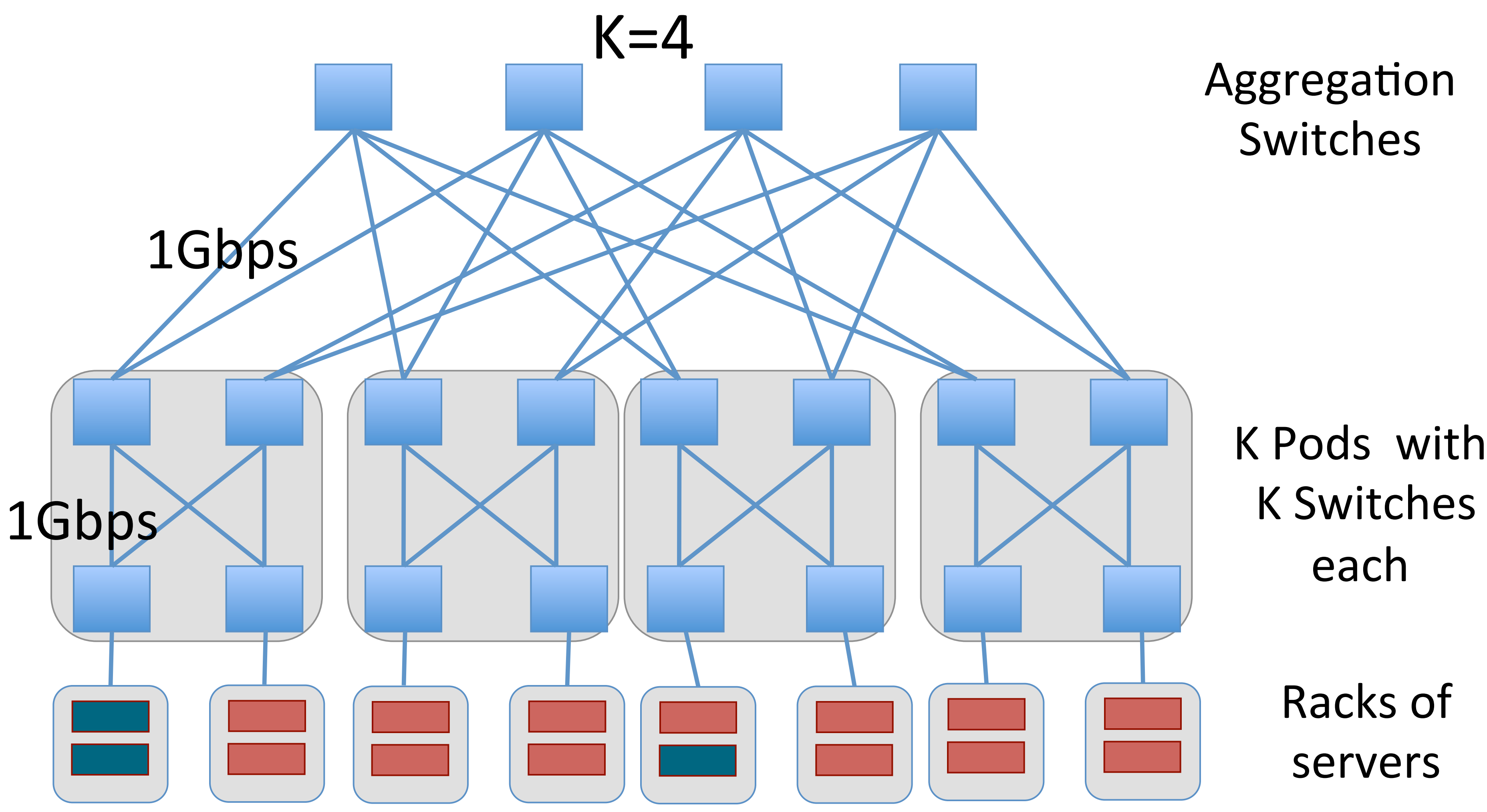
- Shift towards multipath topologies: FatTree, BCube, VL2,

  Cisco, EC2



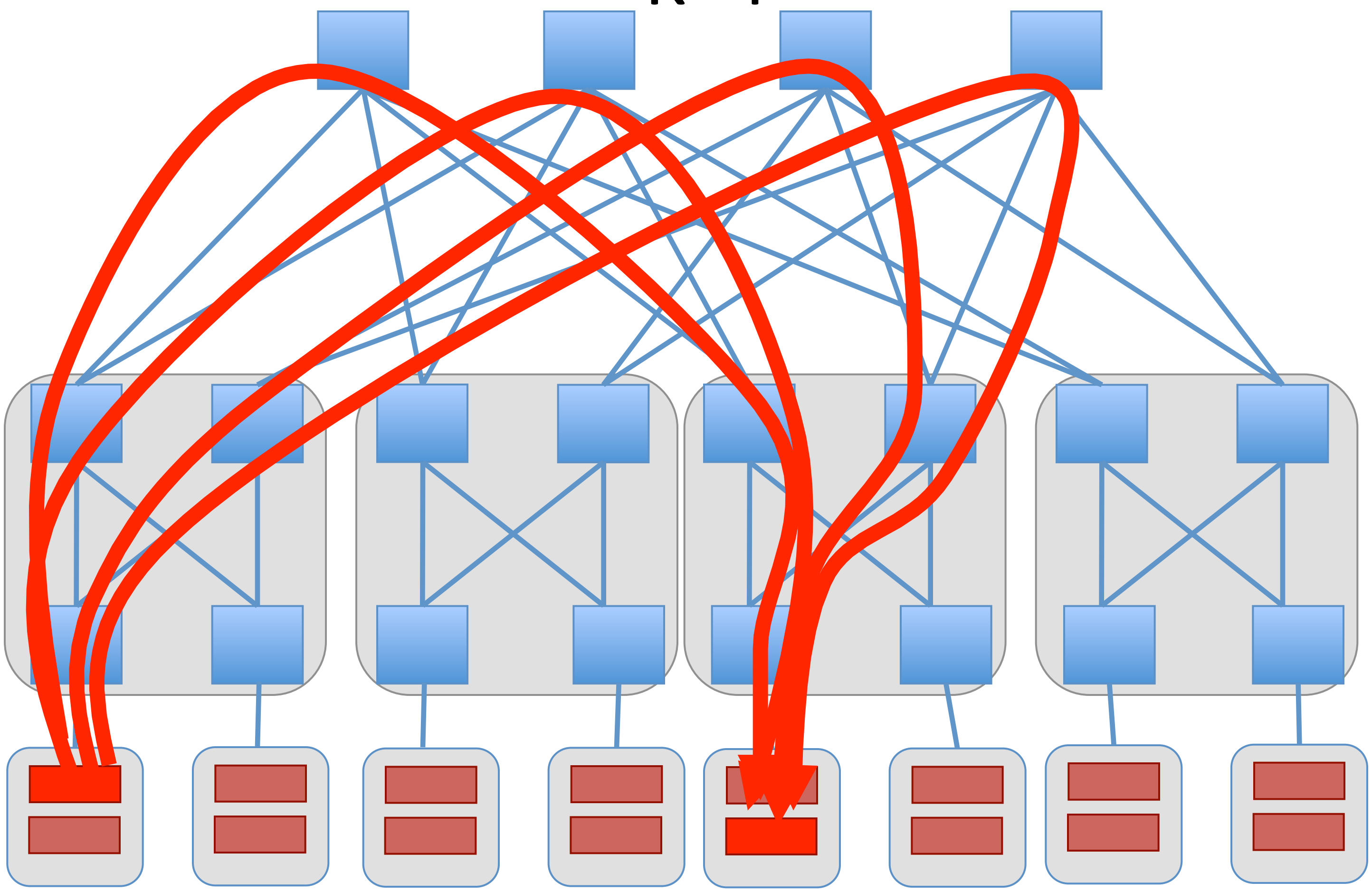
C. Raiciu, et al. “Improving datacenter performance and robustness with multipath TCP,” *ACM SIGCOMM* 2011.

Fat Tree Topology [Fares et al., 2008; Clos, 1953]



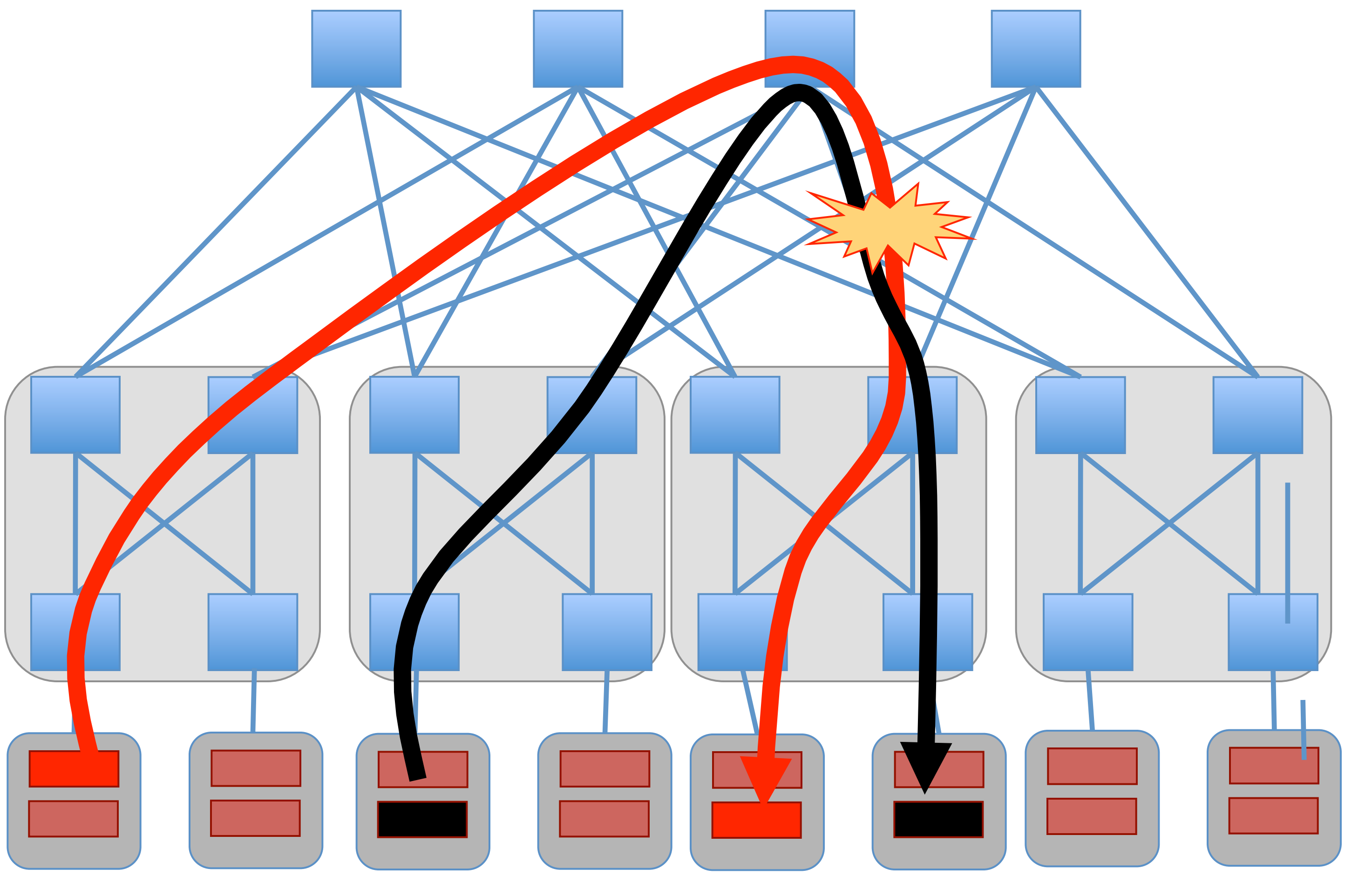
Fat Tree Topology [Fares et al., 2008; Clos, 1953]

K=4

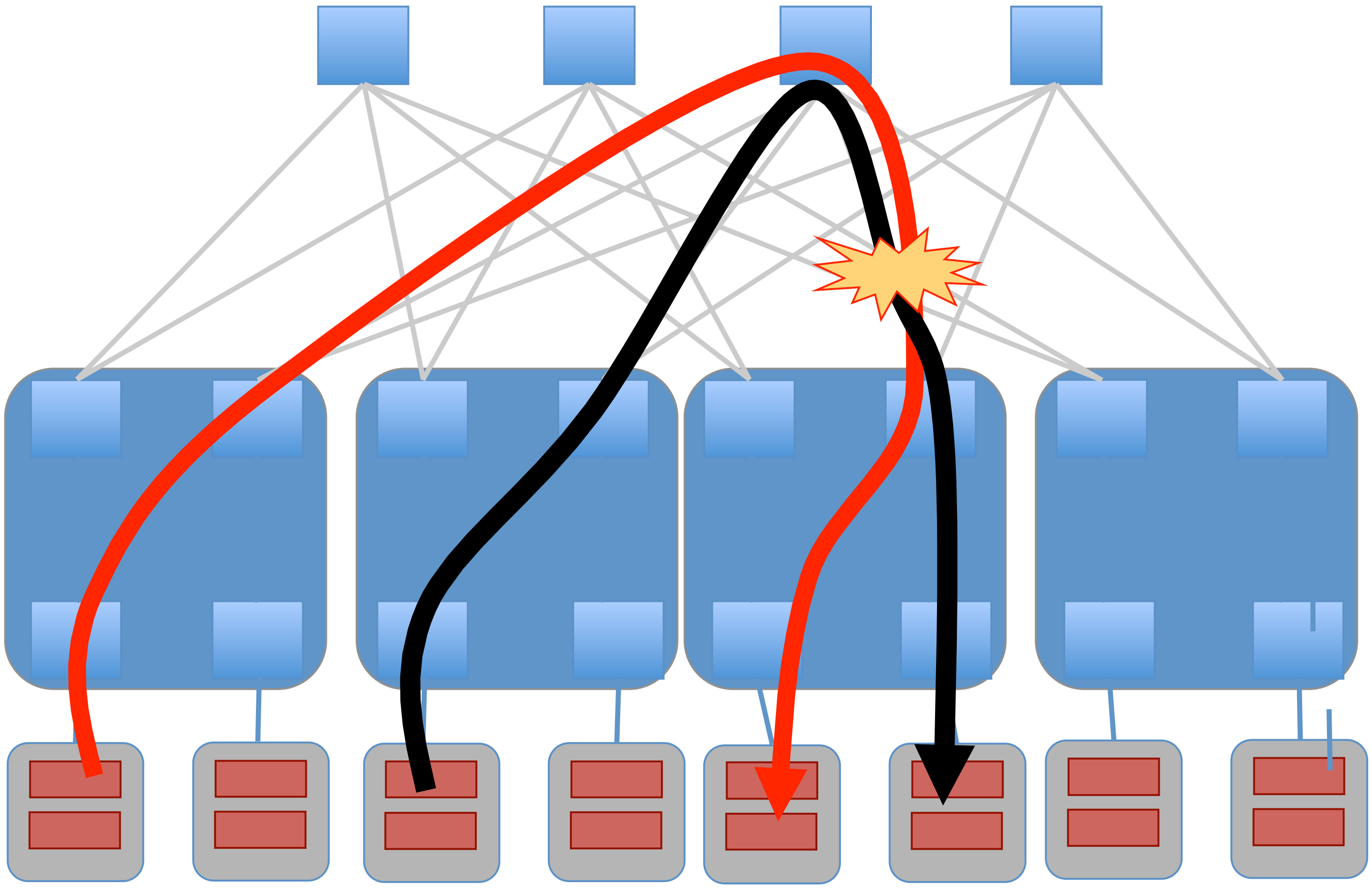


C. Raiciu, et al. "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM* 2011.

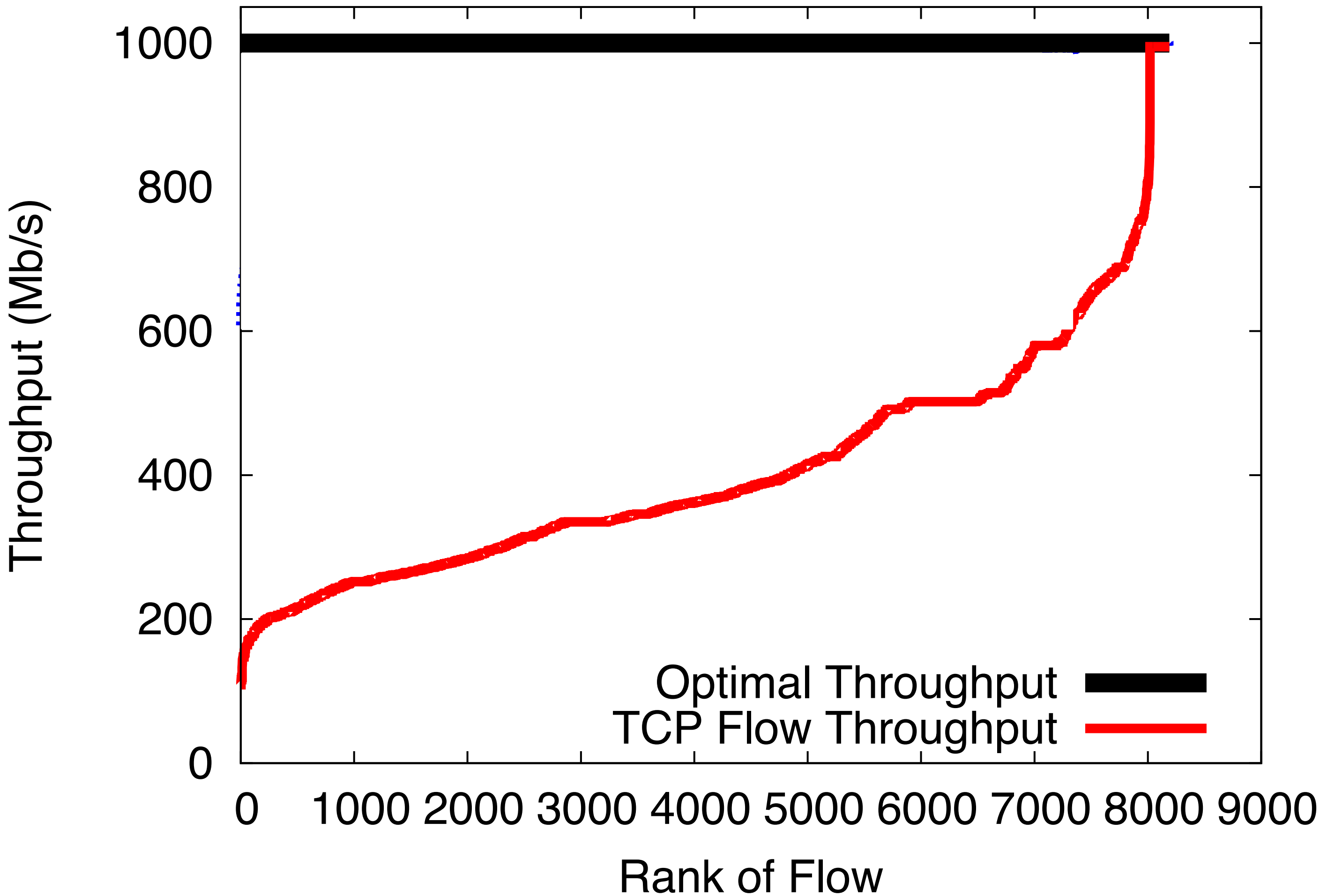
Collisions



TCP in data centers



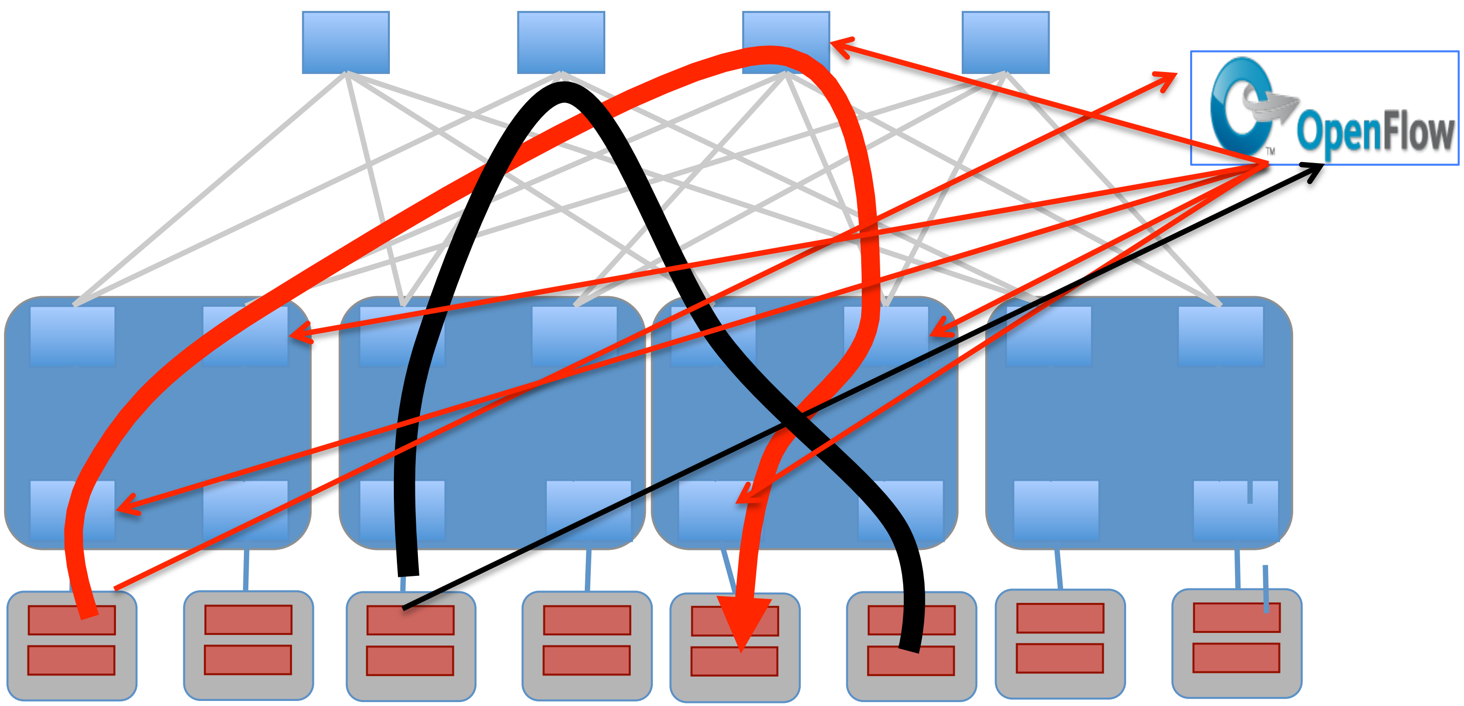
TCP in FAT tree networks
Cost of collissions



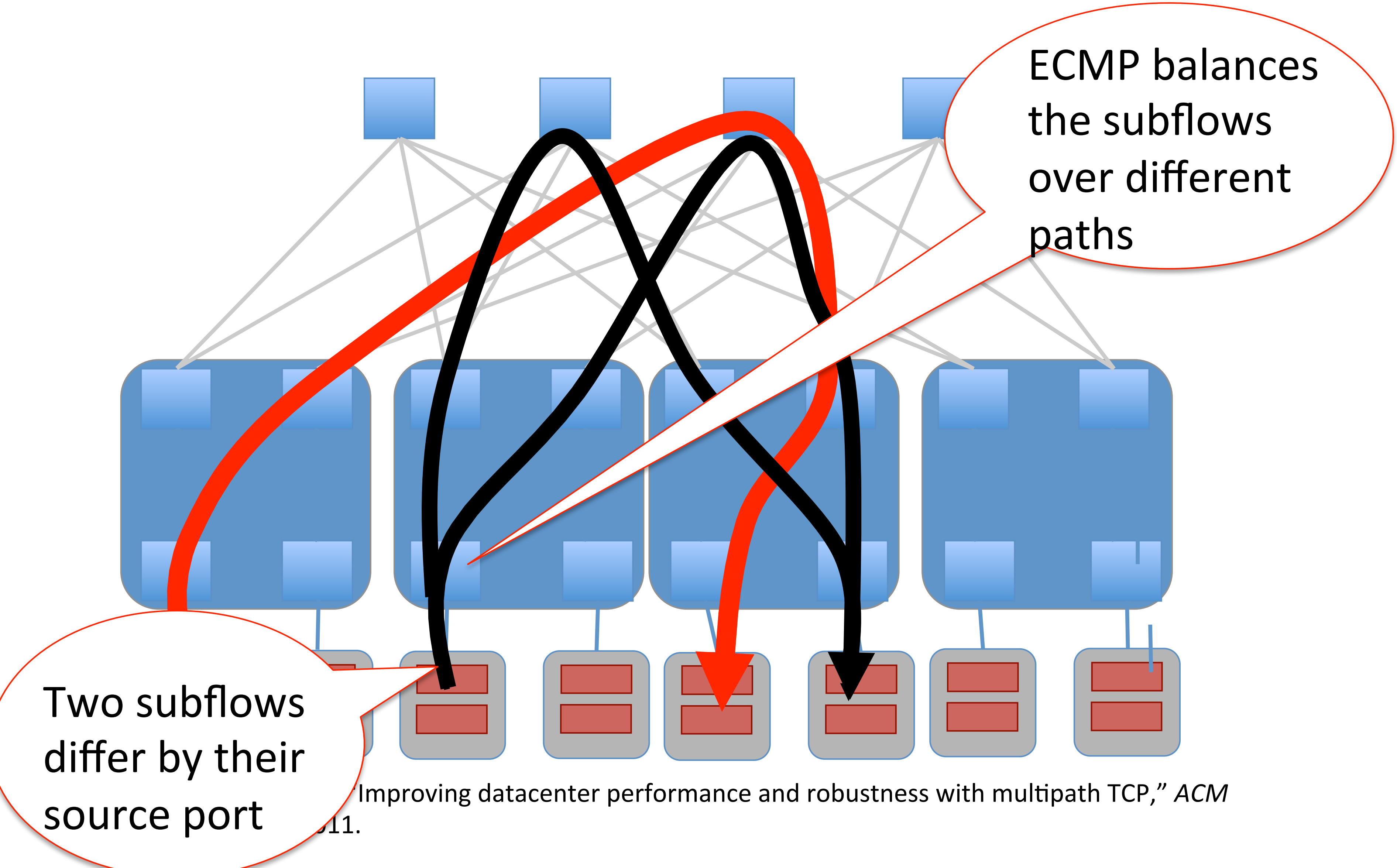
C. Raiciu, et al. "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM* 2011.

How to get rid of these collisions ?

- Consider TCP performance as an optimisation problem

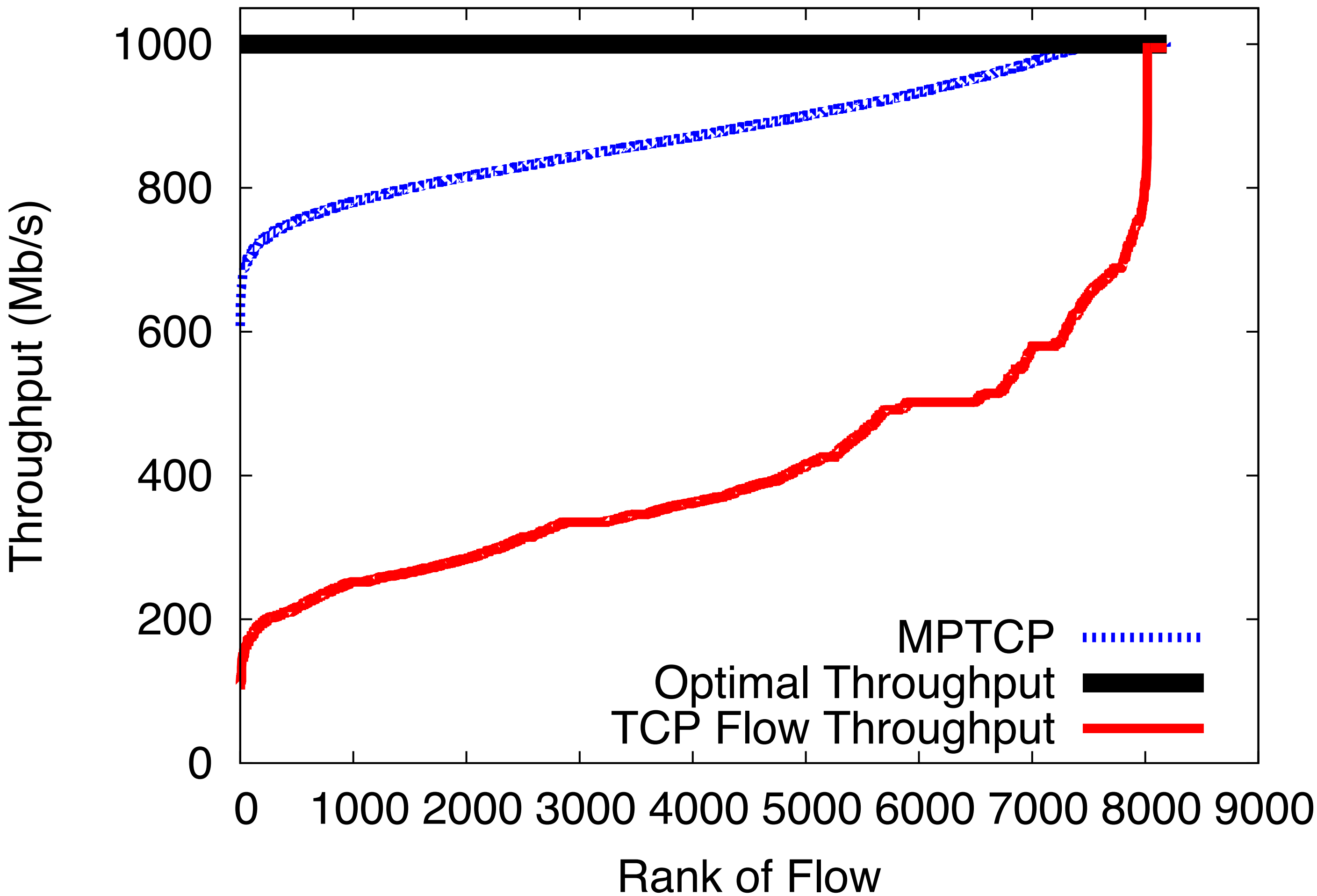


The Multipath TCP way



“Improving datacenter performance and robustness with multipath TCP,” *ACM*
011.

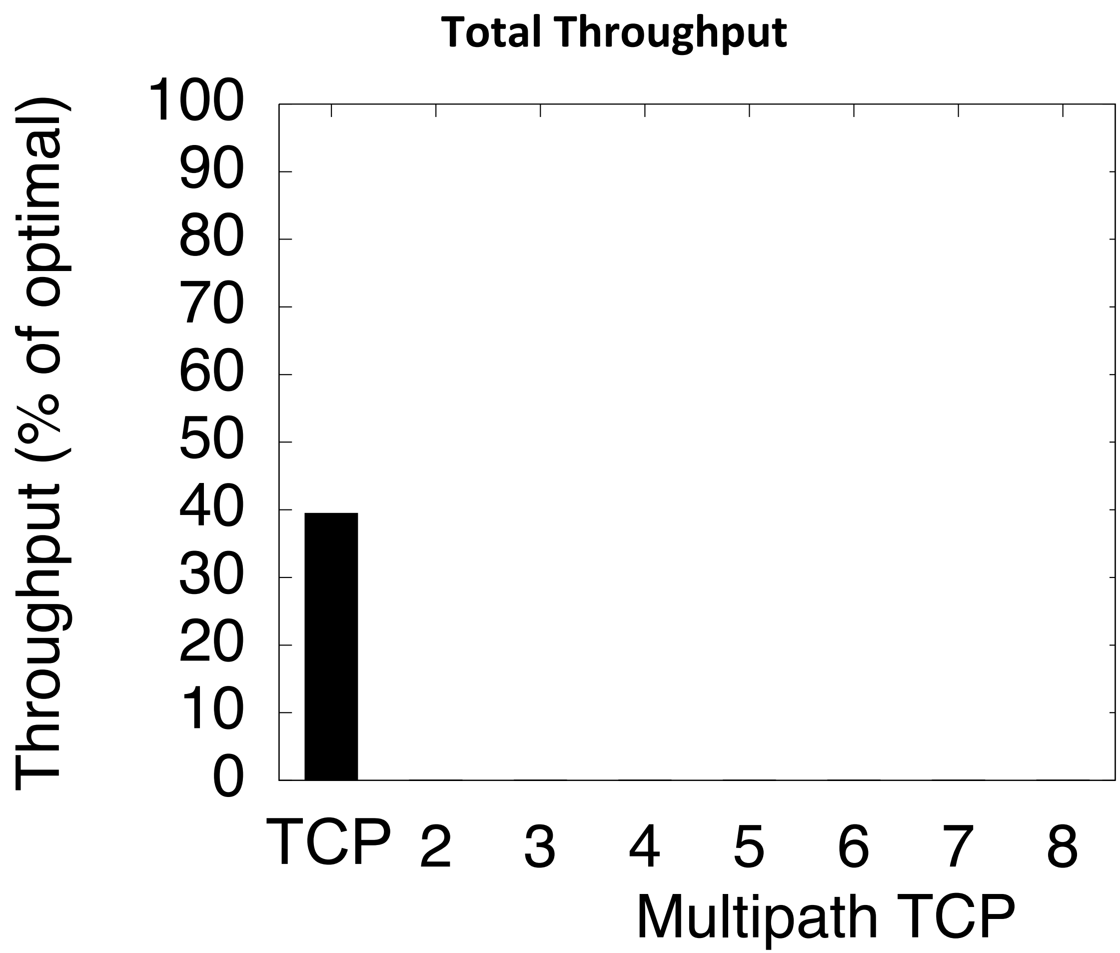
MPTCP better utilizes the FatTree network



C. Raiciu, et al. "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM* 2011.

See also G. Detal, et al. , *Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, Computer Networks, April 2013 for extensions to ECMP for MPTCP

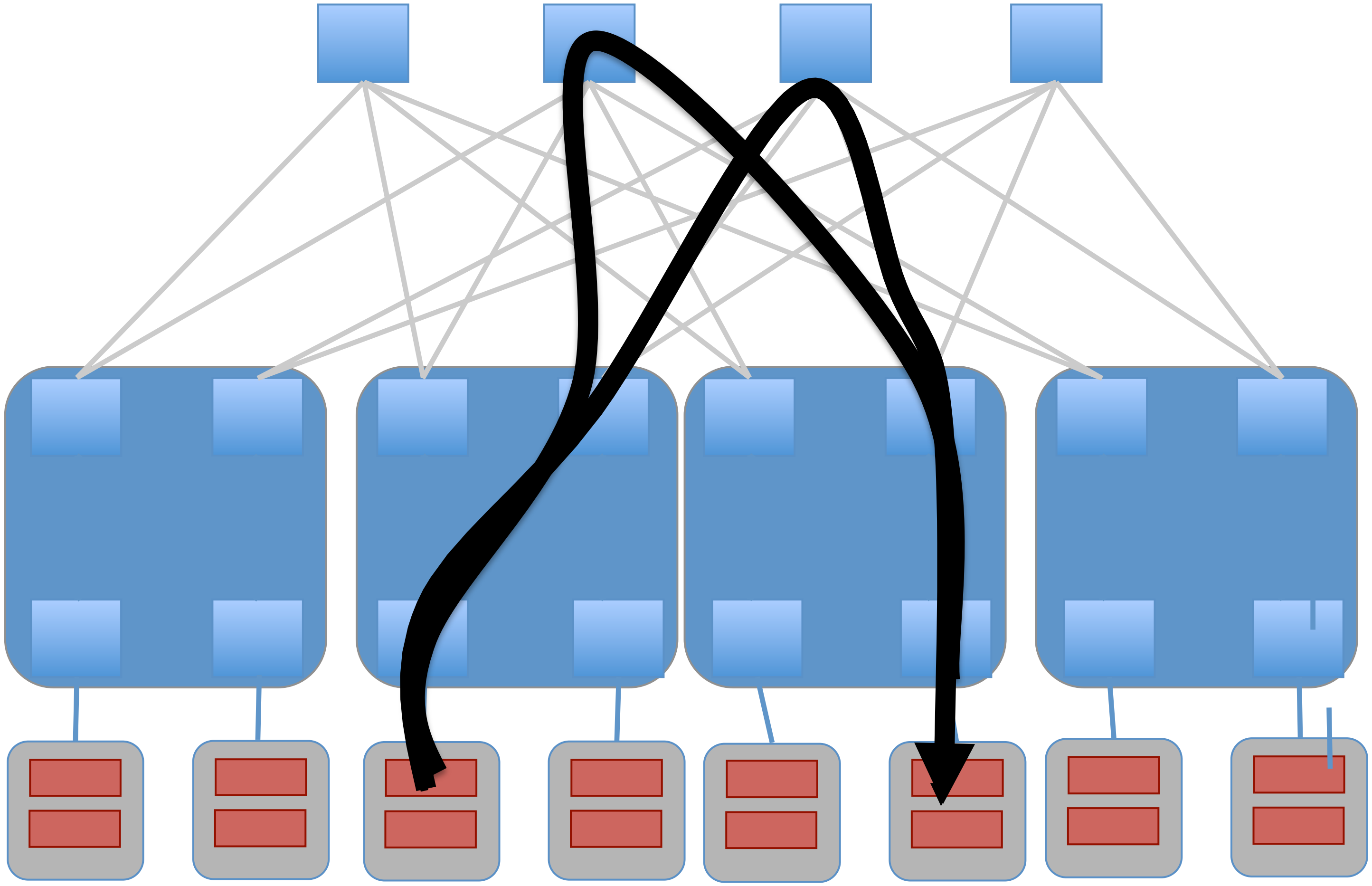
How many subflows does Multipath TCP need ?



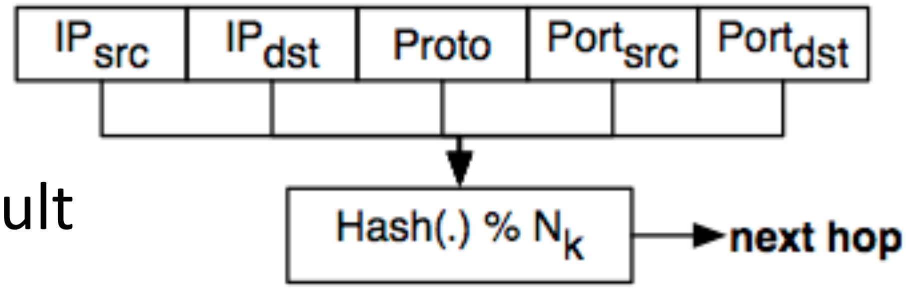
C. Raiciu, et al. "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM* 2011.

Can we improve Multipath TCP ?

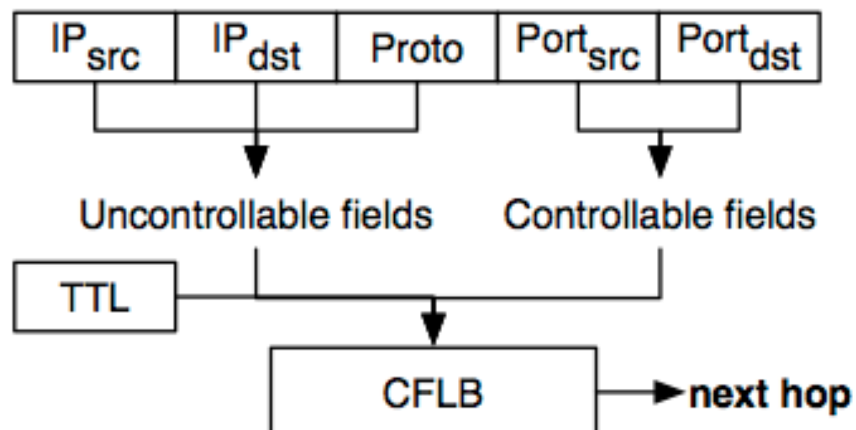
- Two subflows may follow similar paths



Improving ECMP



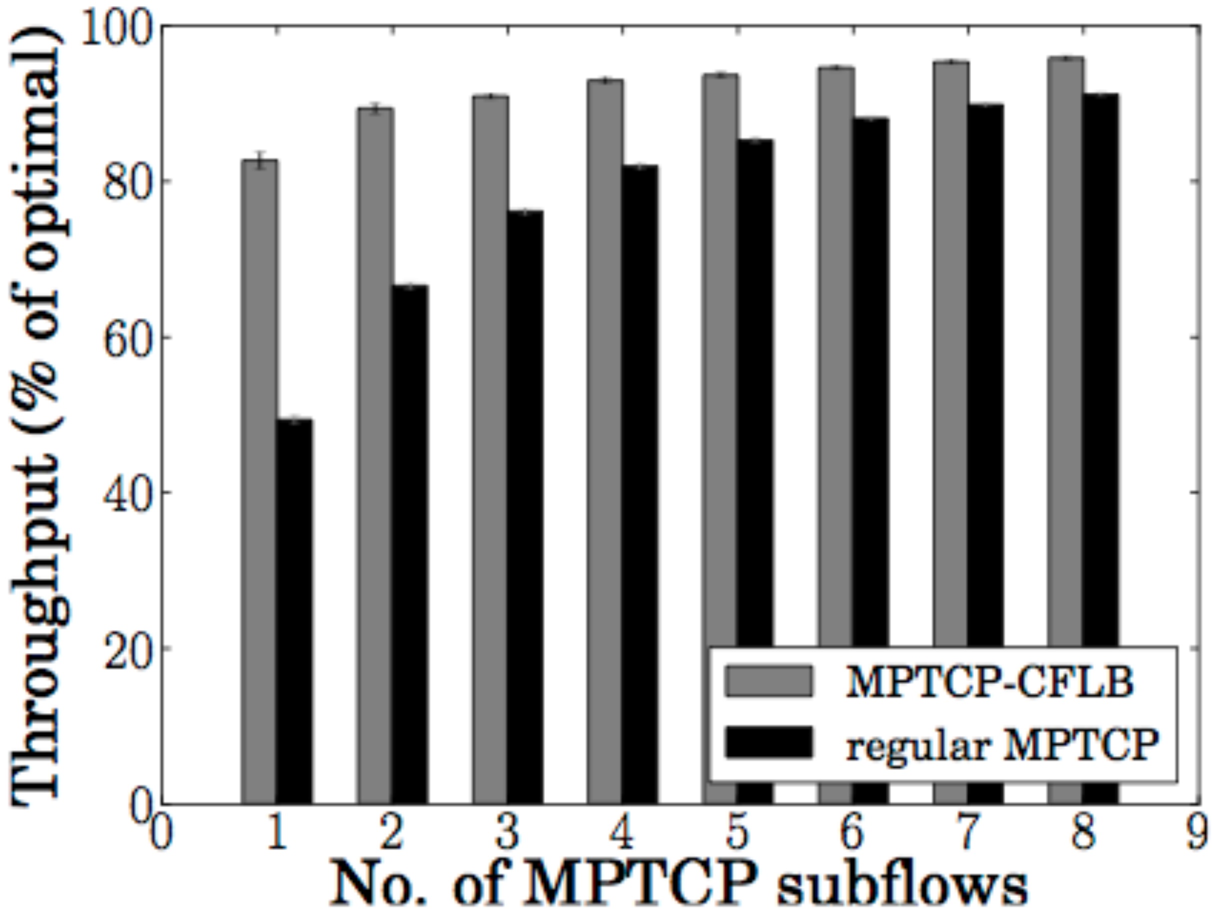
- ECMP's hash
  - good load balancing
  - impossible to predict result



- CFLB
  - replaces hash with block cipher
  - hosts can select paths for Multipath TCP subflows provided they know datacenter topology

G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure, *Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, to appear in Computer Networks

Multipath TCP with CFLB in Fat-Tree



G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure*, Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, to appear in Computer Networks

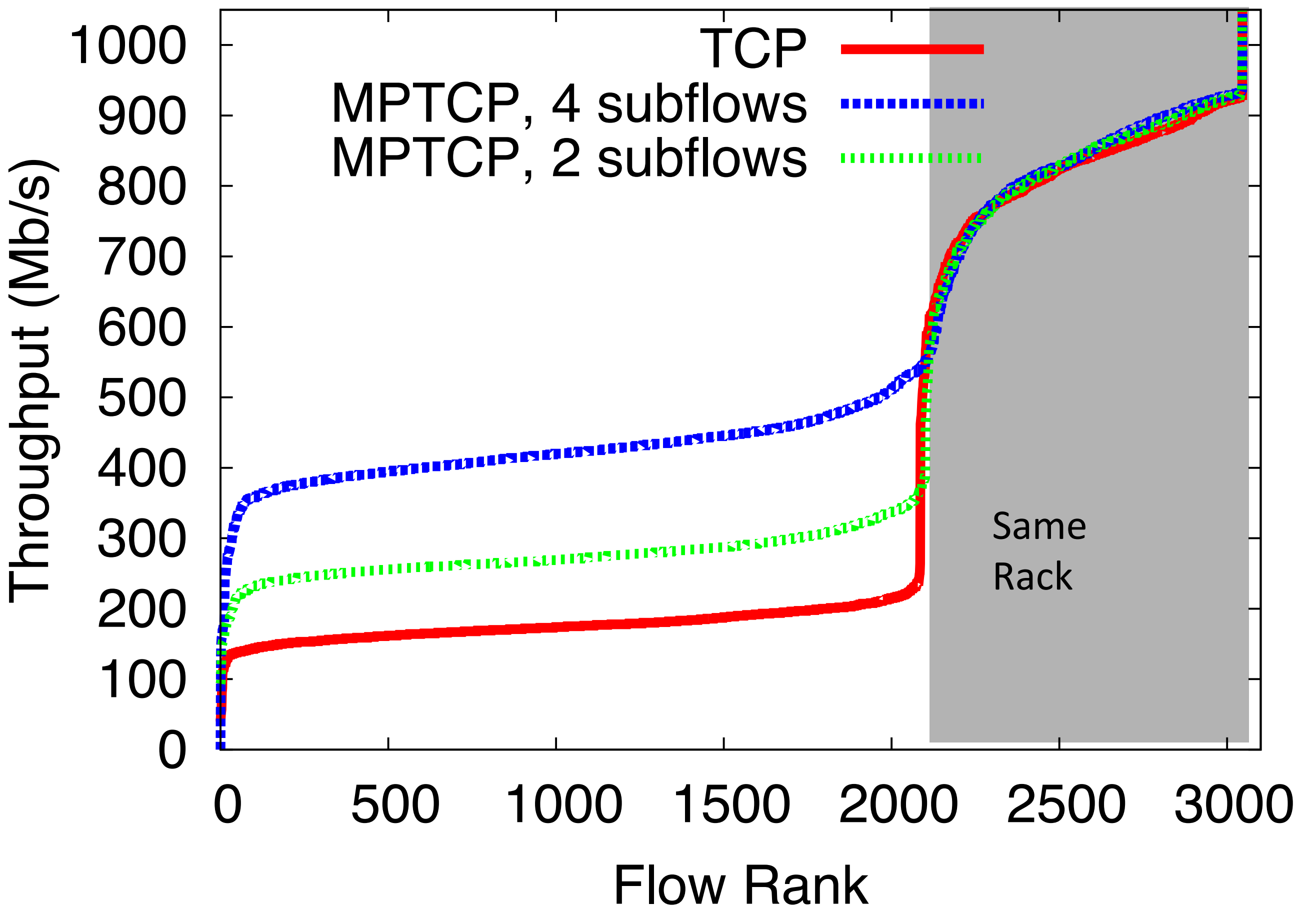
Multipath TCP on EC2

- Amazon EC2: infrastructure as a service
  - We can borrow virtual machines by the hour
  - These run in Amazon data centers worldwide
  - We can boot our own kernel
- A few availability zones have multipath topologies
  - 2-8 paths available between hosts not on the same machine or in the same rack
  - Available via ECMP

Amazon EC2 Experiment

- 40 medium CPU instances running MPTCP
- During 12 hours, we sequentially ran all-to-all `iperf` cycling through:
  - TCP
  - MPTCP (2 and 4 subflows)

MPTCP improves performance on EC2



C. Raiciu, et al. "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM* 2011.

Agenda

- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases
  - Datacenters
  - Smartphones
  - Commercial deployments

Motivation

- One device, many IP-enabled interfaces

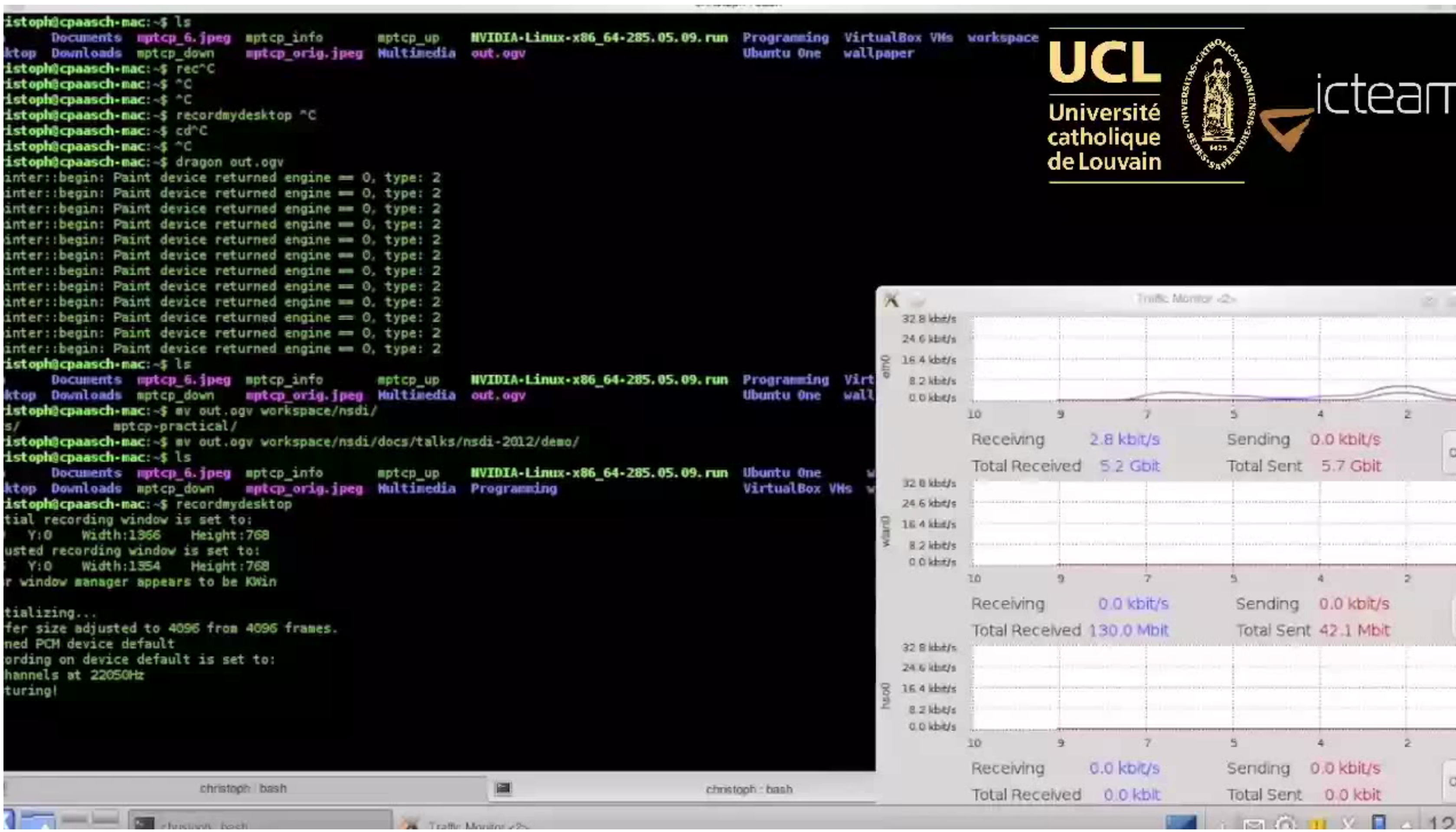
3G

Wi Fi™

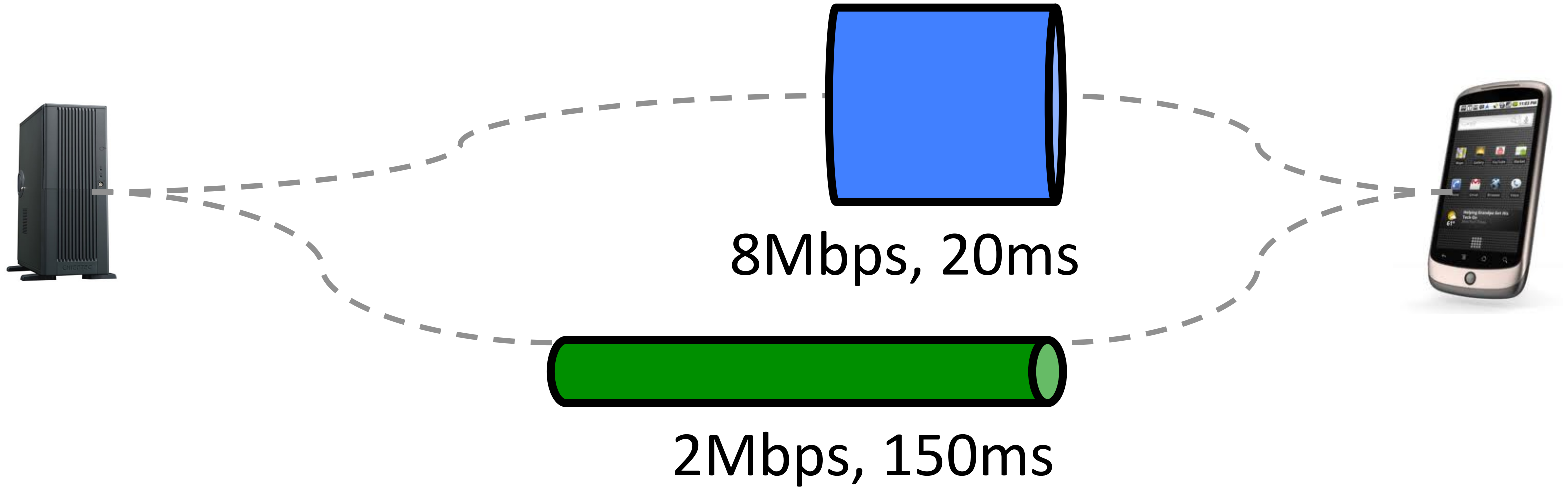
Bluetooth®

Lte™

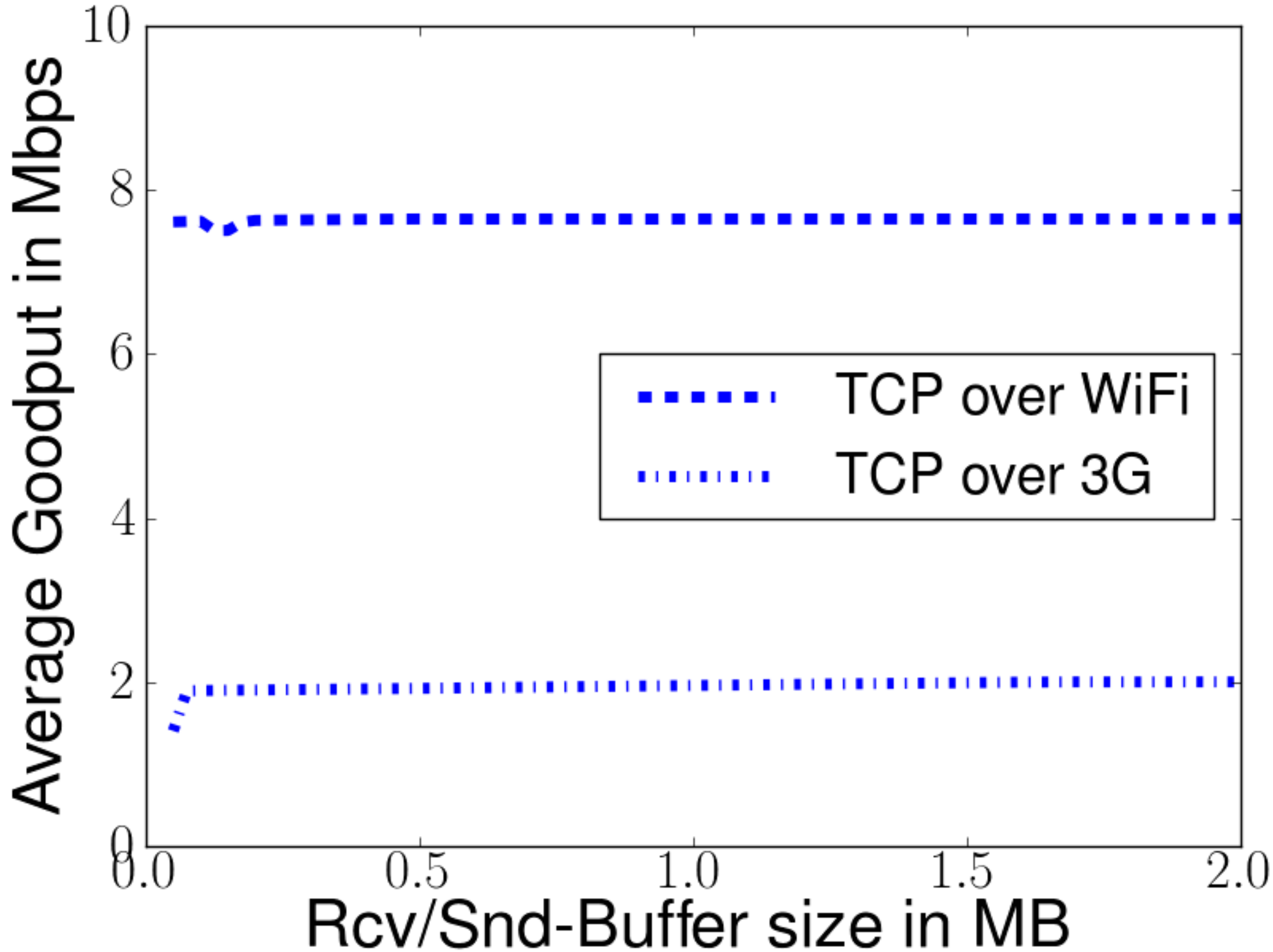
ssh with Multipath TCP



MPTCP over WiFi/3G

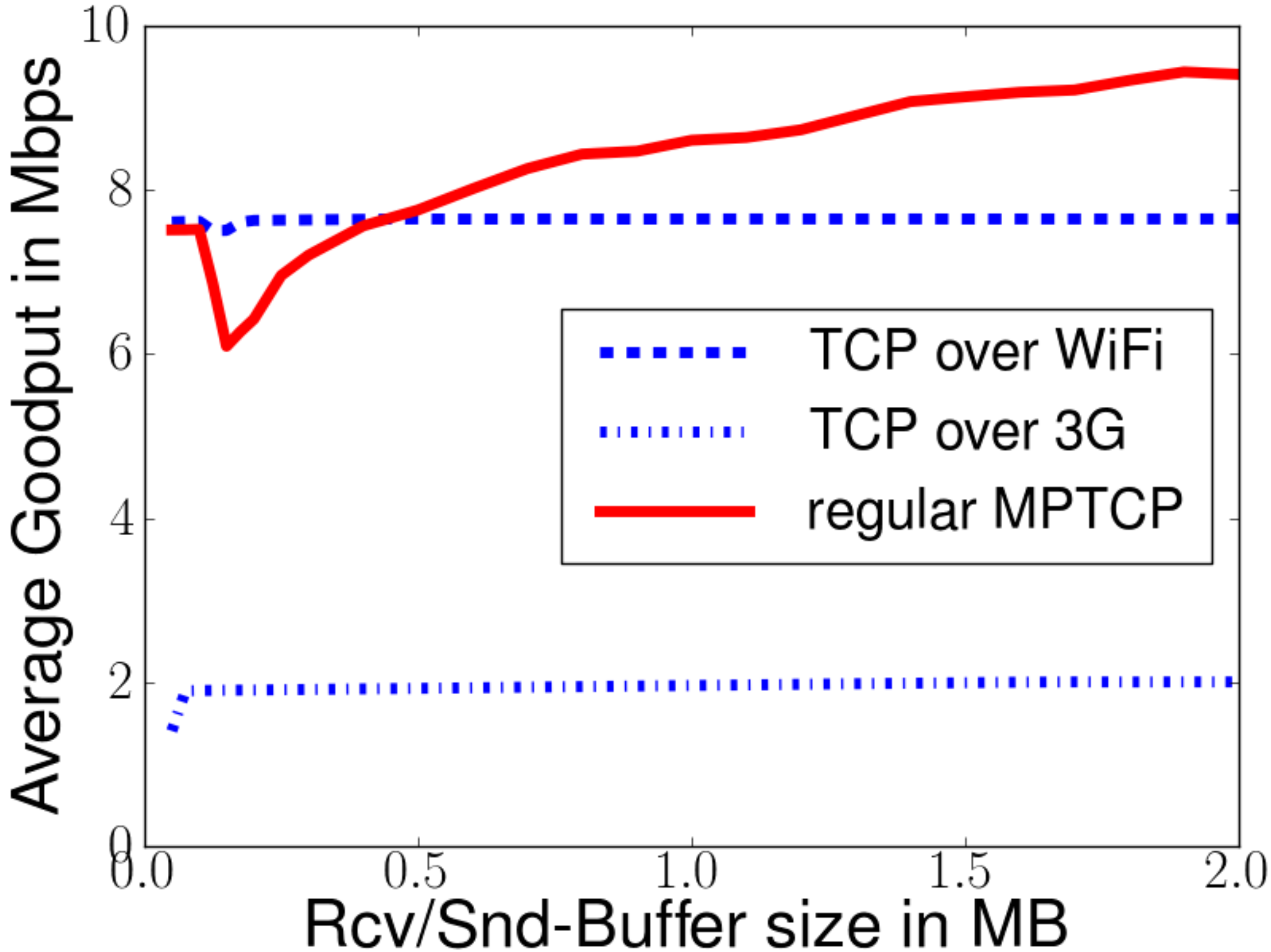


TCP over WiFi/3G



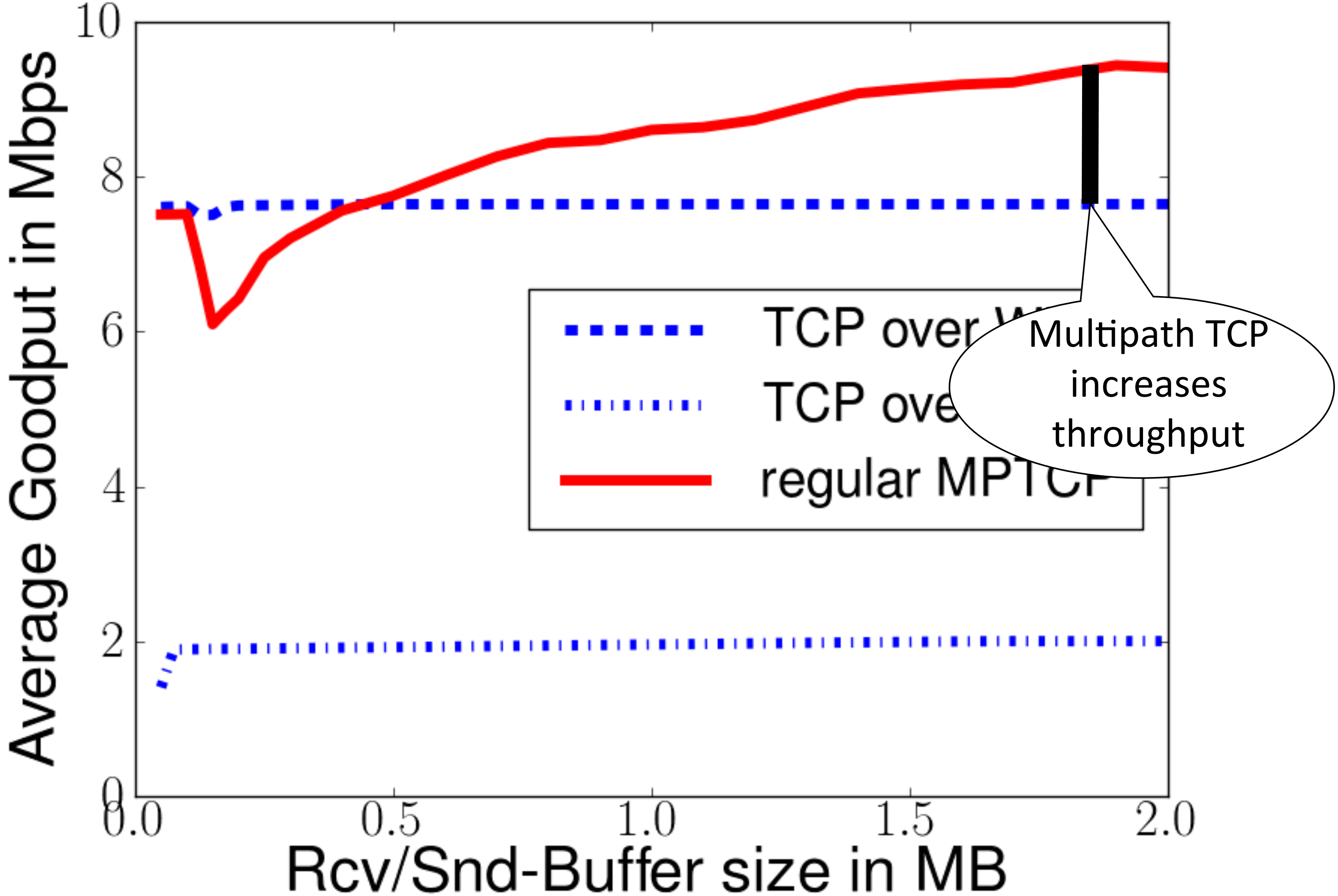
C. Raiciu, et al. "How hard can it be? designing and implementing a deployable multipath TCP," NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

MPTCP over WiFi/3G

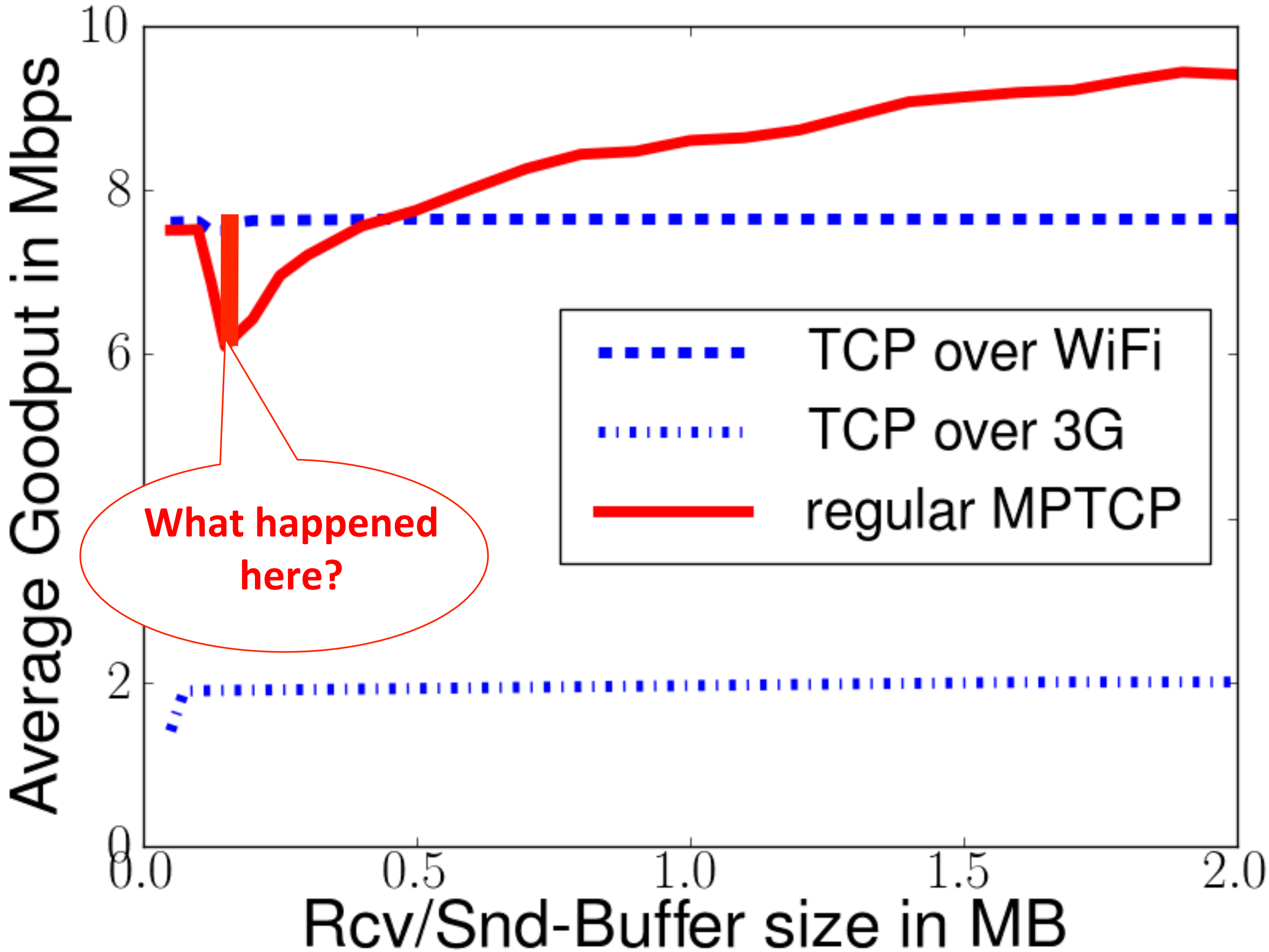


C. Raiciu, et al. "How hard can it be? designing and implementing a deployable multipath TCP," NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

MPTCP over WiFi/3G



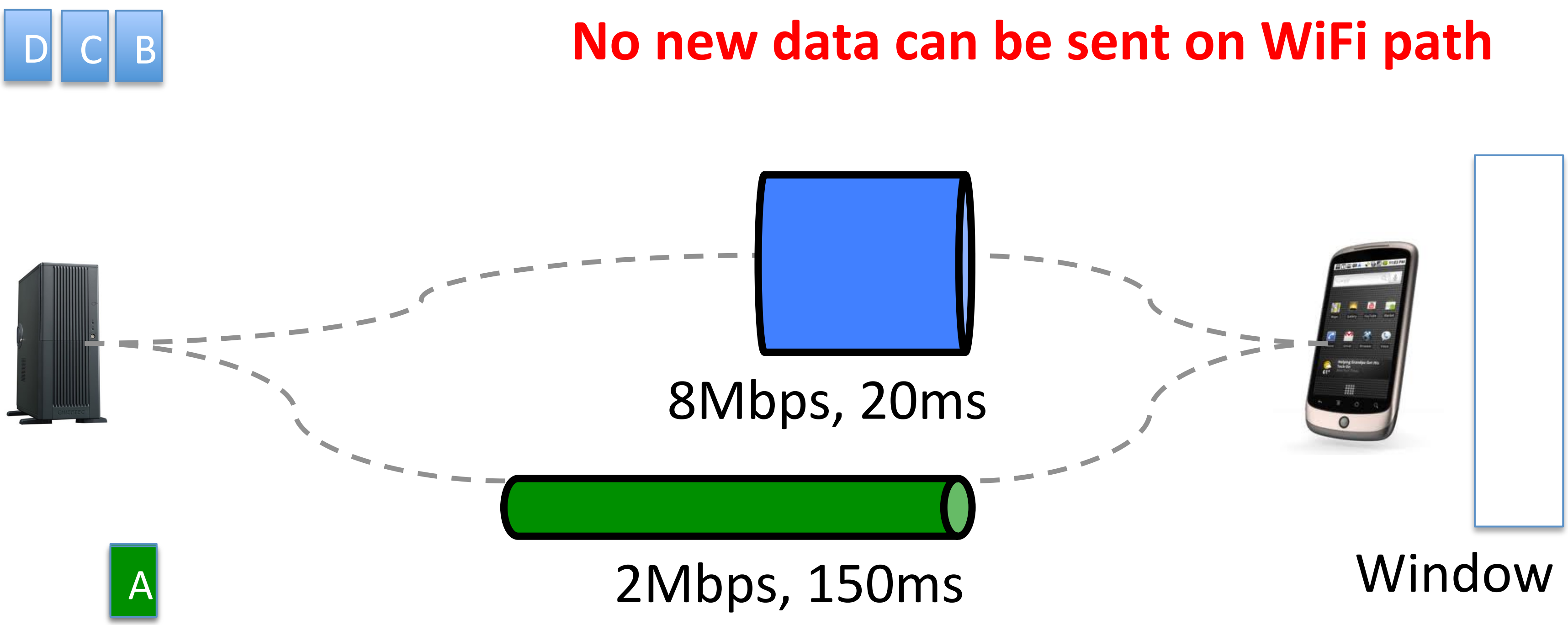
MPTCP over WiFi/3G



Understanding the performance issue

Window full !

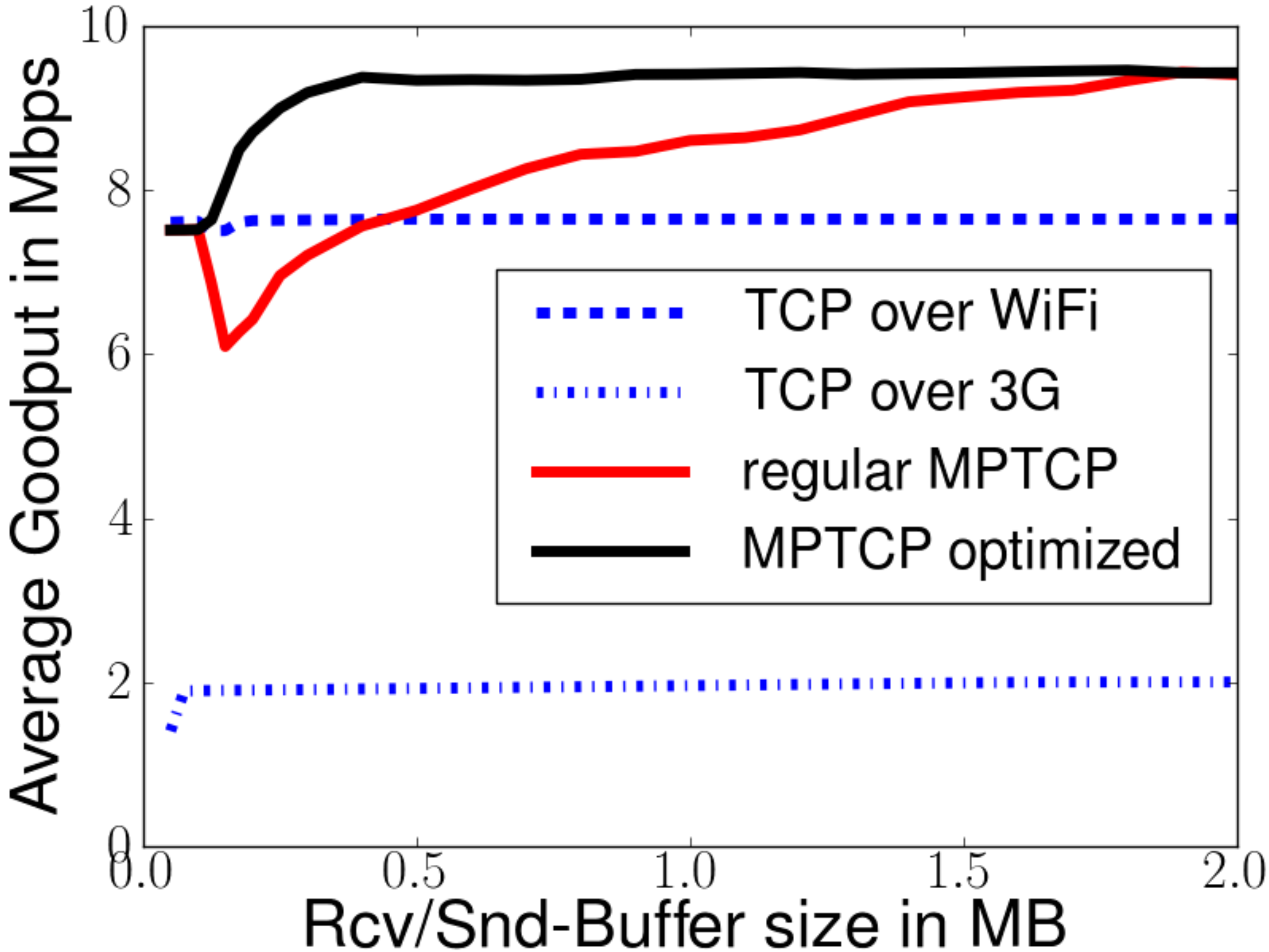
No new data can be sent on WiFi path



Reinject segment on fast path

Halve congestion window on slow subflow

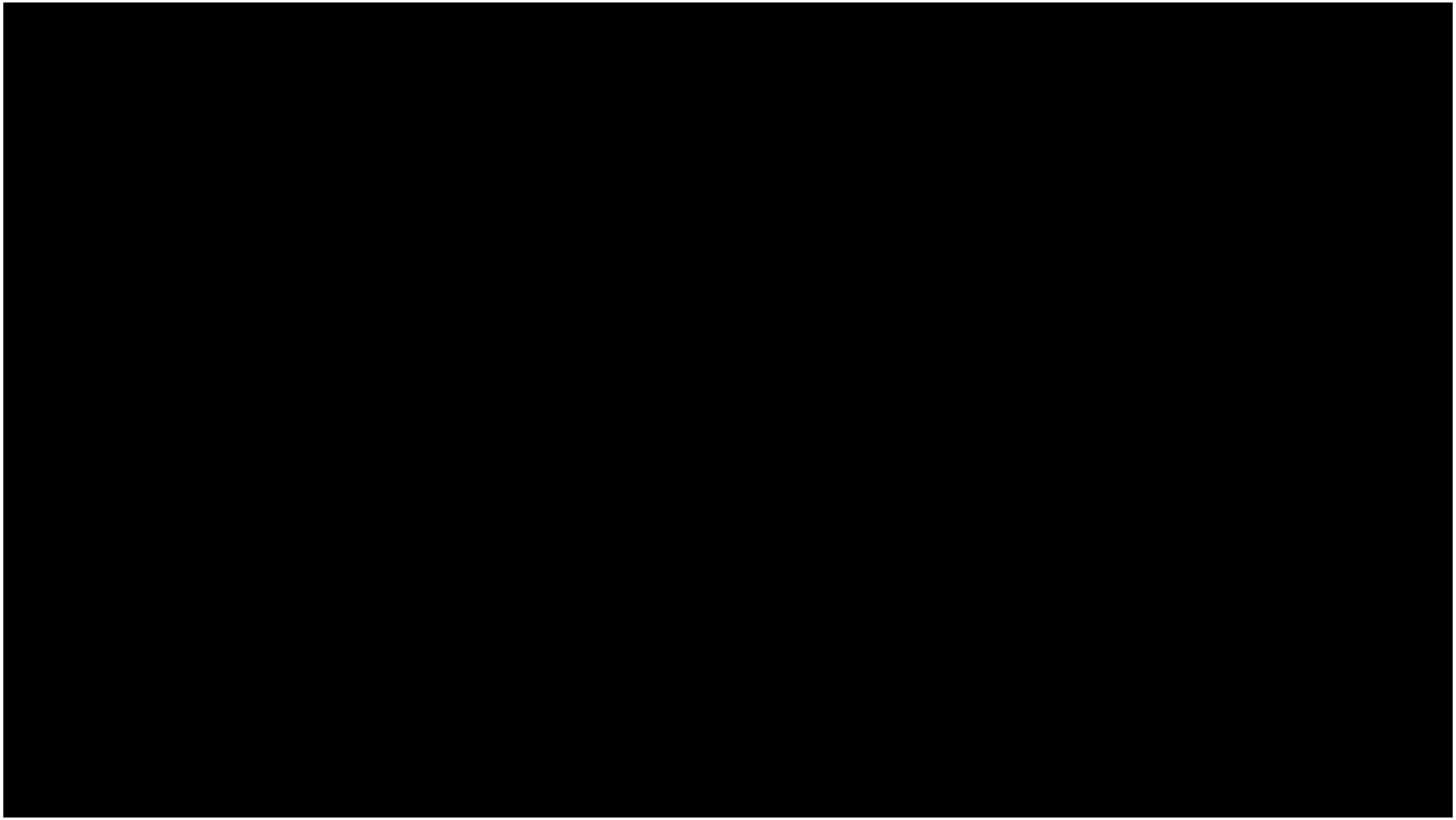
MPTCP over WiFi/3G



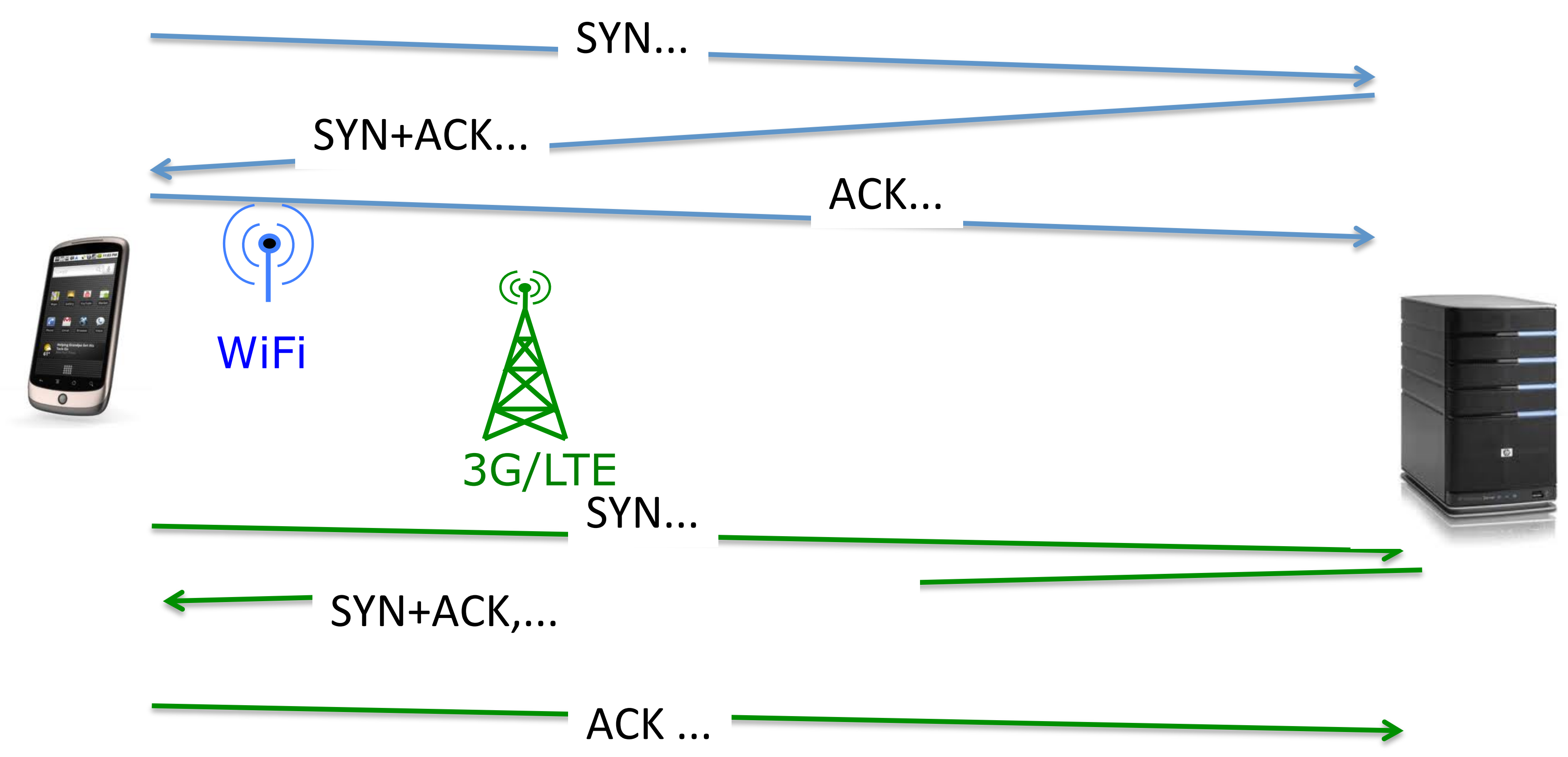
Usage of 3G and WiFI

- How should Multipath TCP use 3G and WiFi ?
  - Full mode
    - Both wireless networks are used at the same time
  - Backup mode
    - Prefer WiFi when available, open subflows on 3G and use them as backup
  - Single path mode
    - Only one path is used at a time, WiFi preferred over 3G

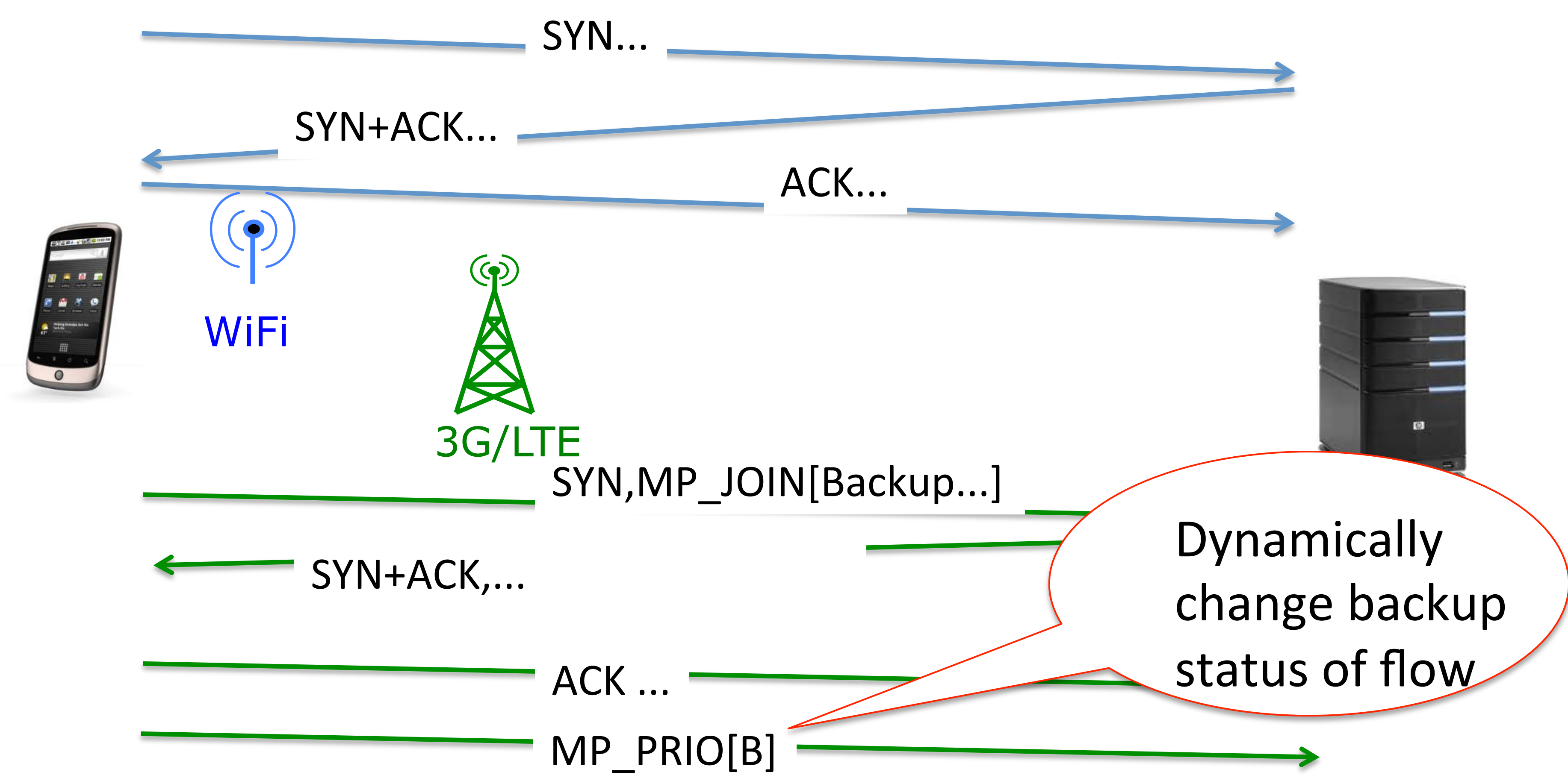
Live streaming



Multipath TCP : Full mode

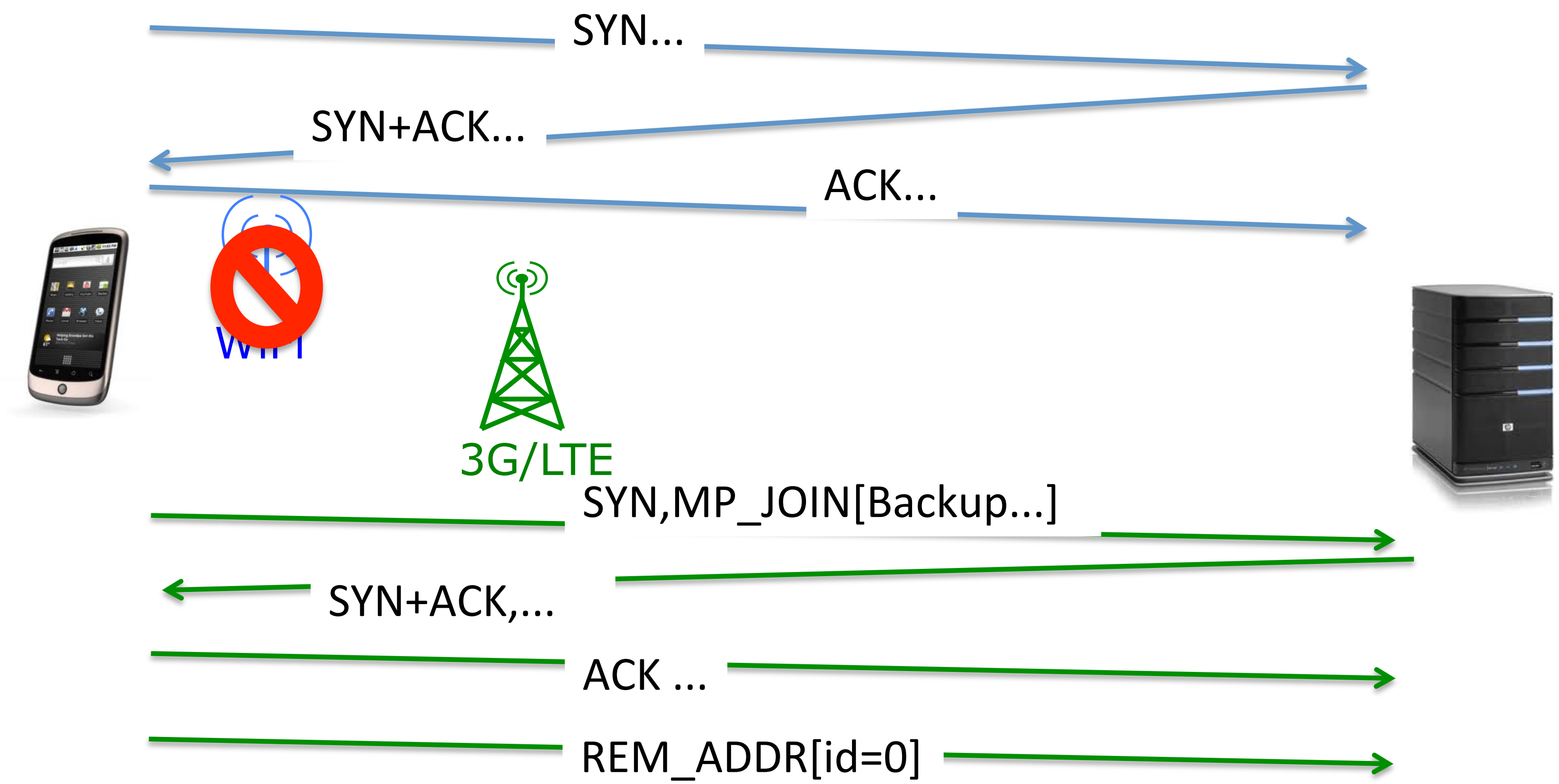


Multipath TCP : Backup mode



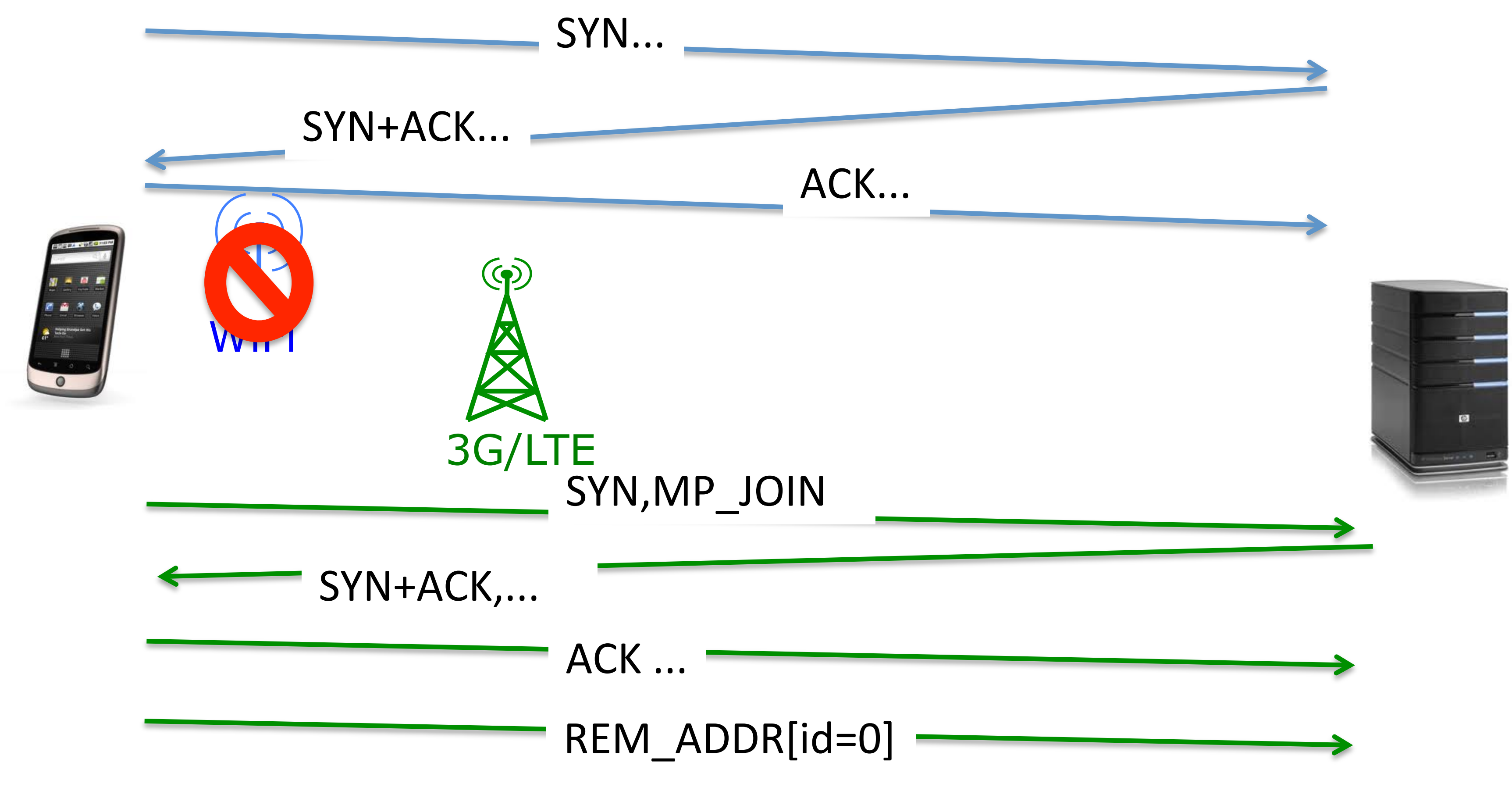
Multipath TCP : Backup mode

- What happens when link fails ?

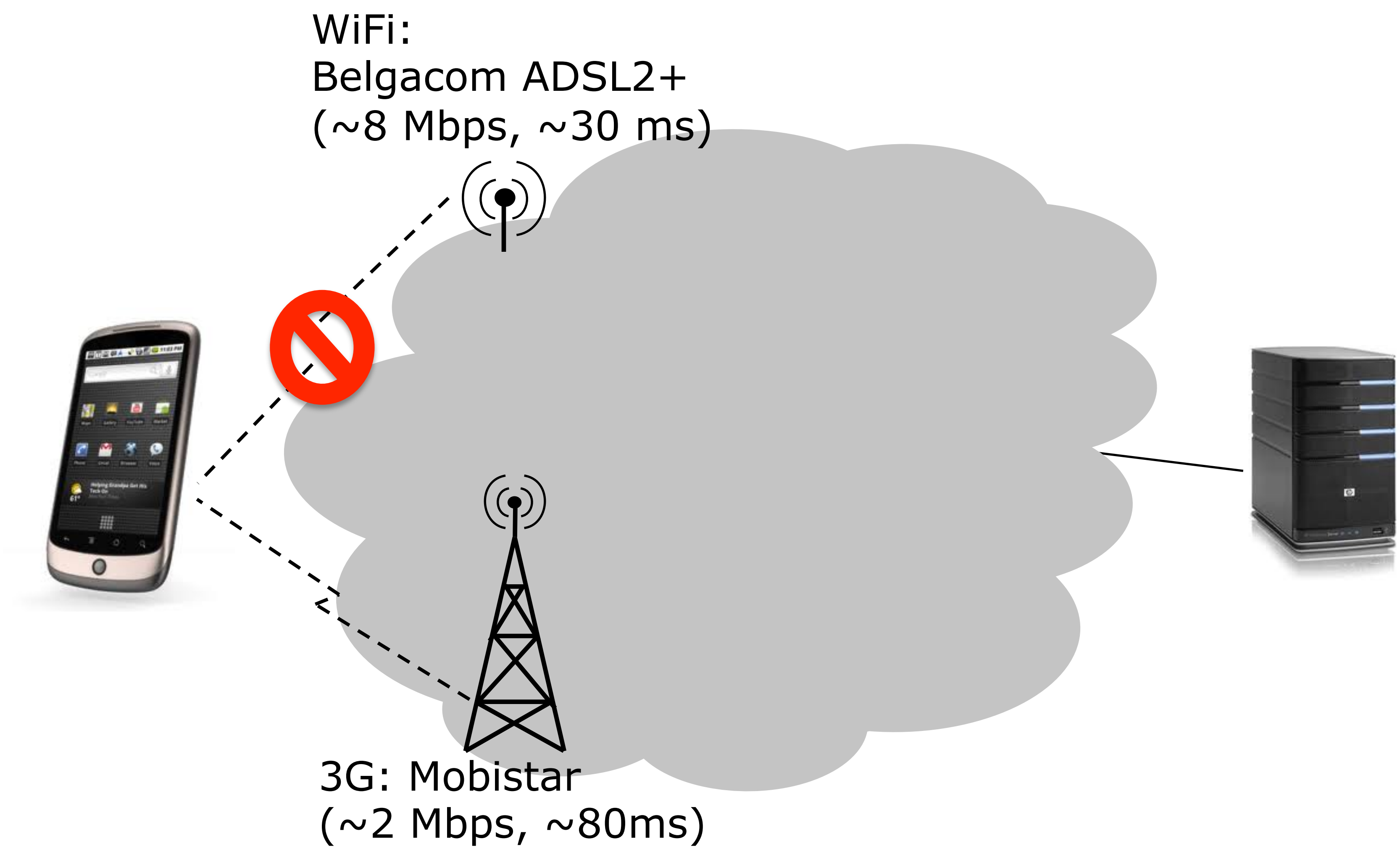


Multipath TCP : single-path mode

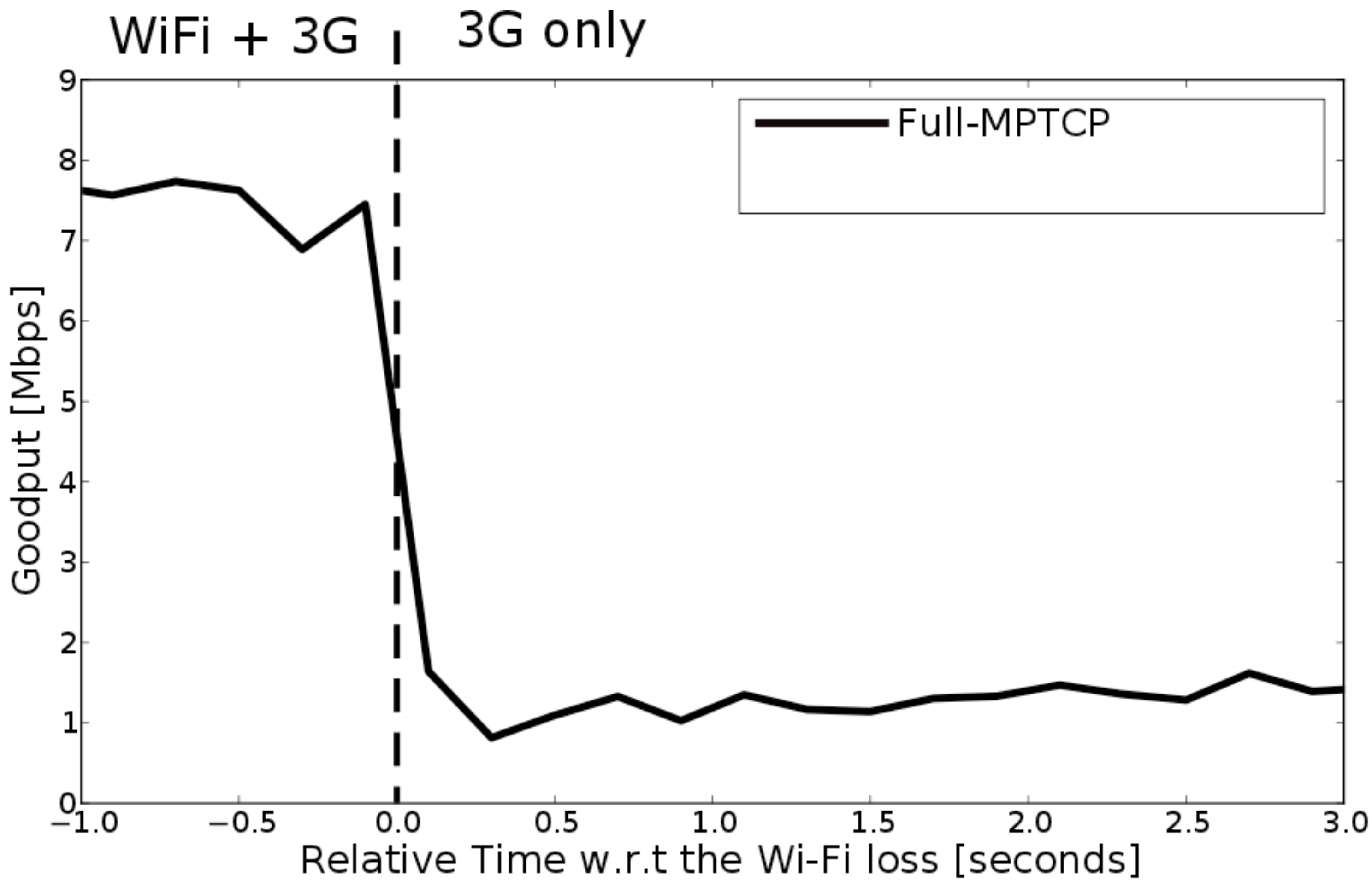
- Multipath TCP supports break before make



Evaluation scenario

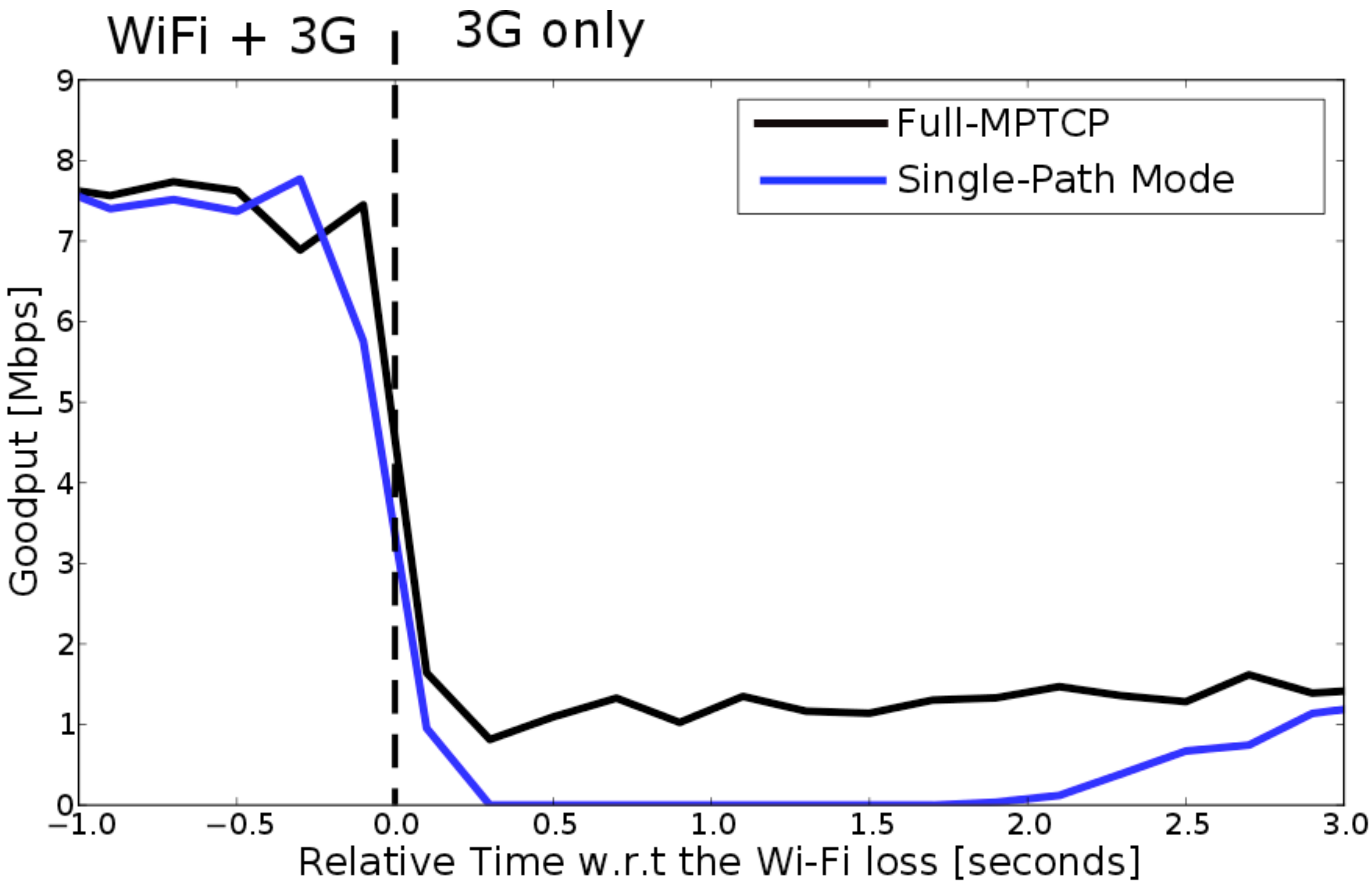


Recovery after failure



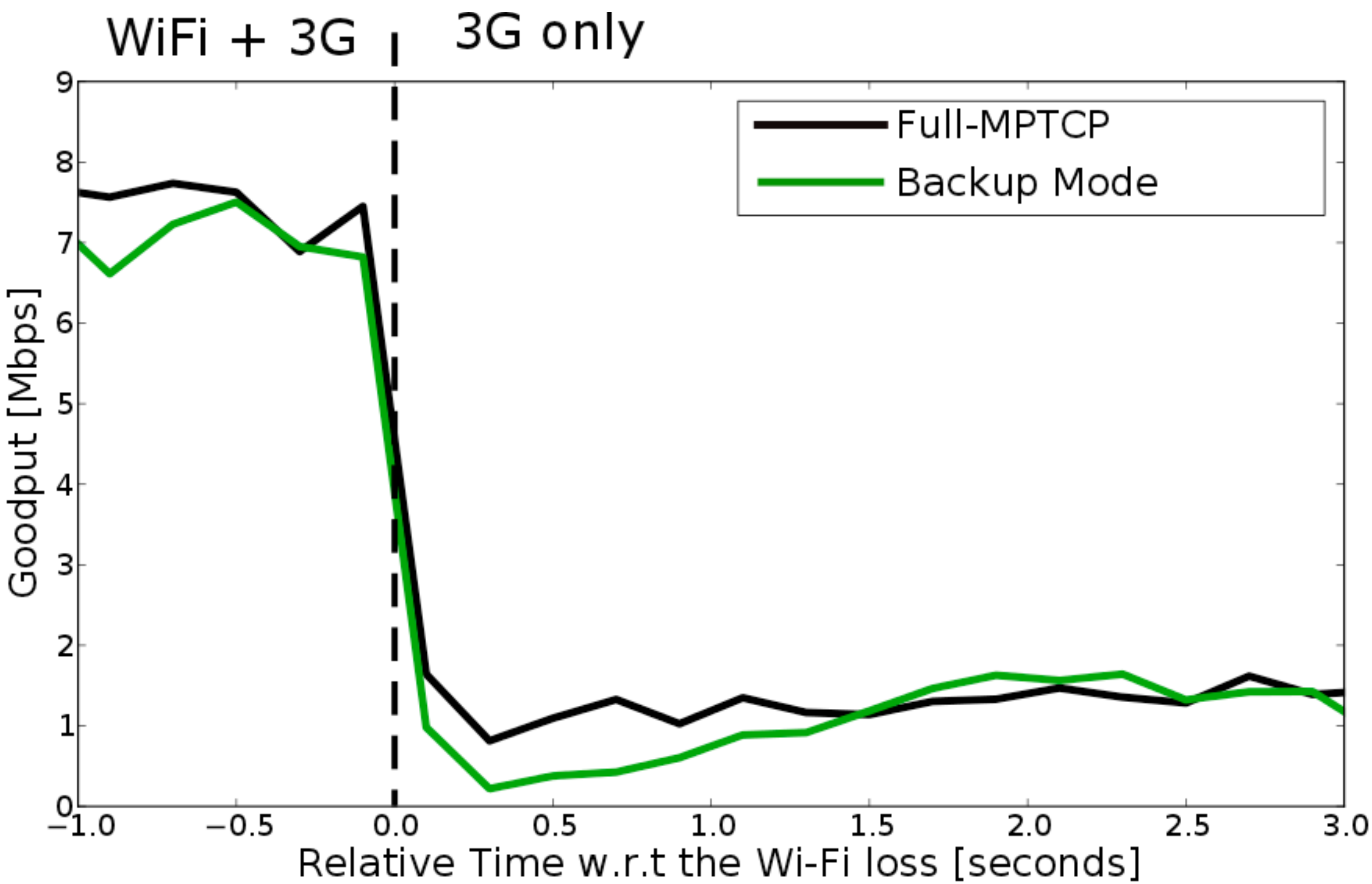
C. Paasch, et al. , “Exploring mobile/WiFi handover with multipath TCP,” presented at the CellNet '12: Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012.

Recovery after failure



C. Paasch, et al. , "Exploring mobile/WiFi handover with multipath TCP," presented at the CellNet '12: Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012.

Recovery after failure

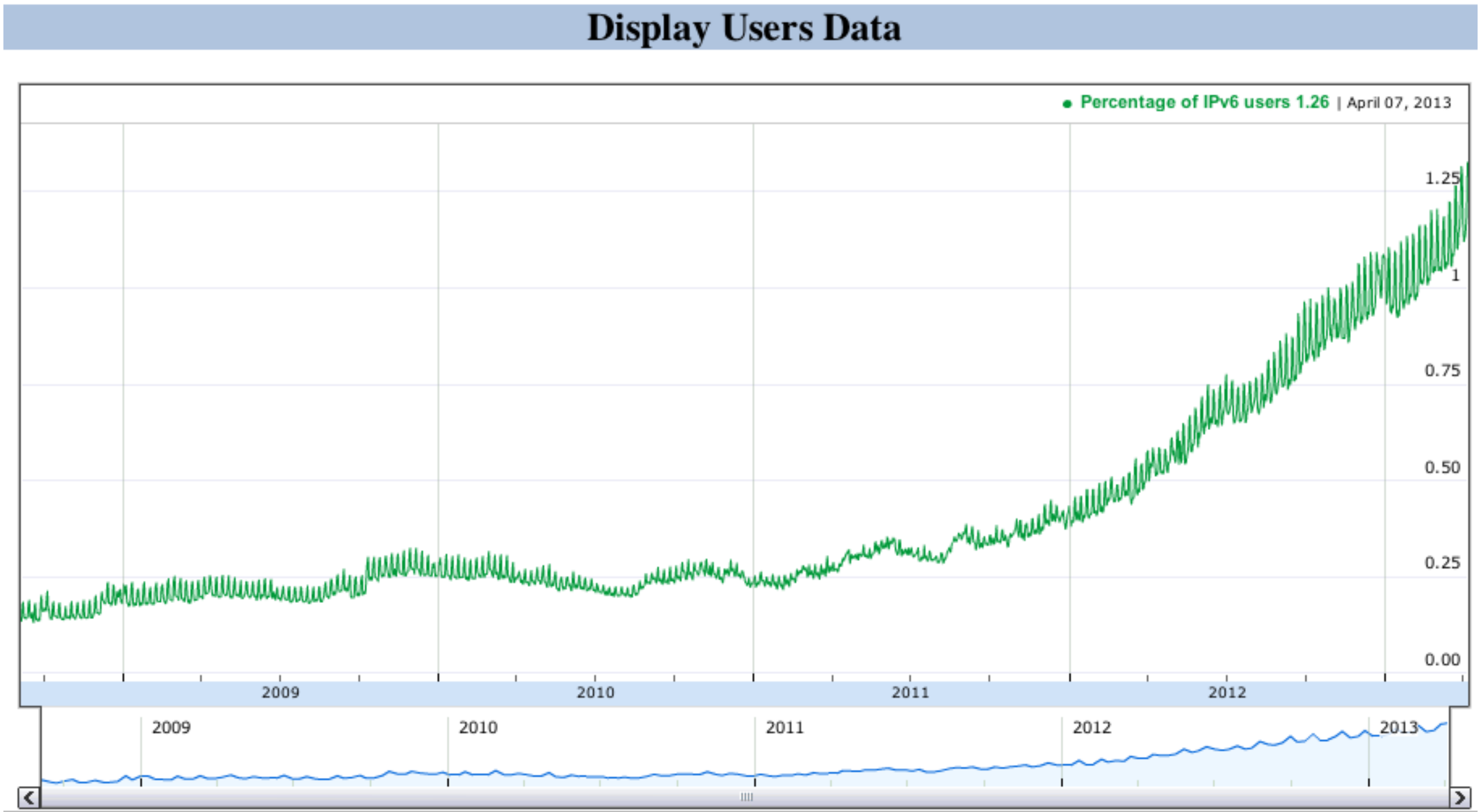


C. Paasch, et al. , "Exploring mobile/WiFi handover with multipath TCP," presented at the CellNet '12: Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012.

Agenda

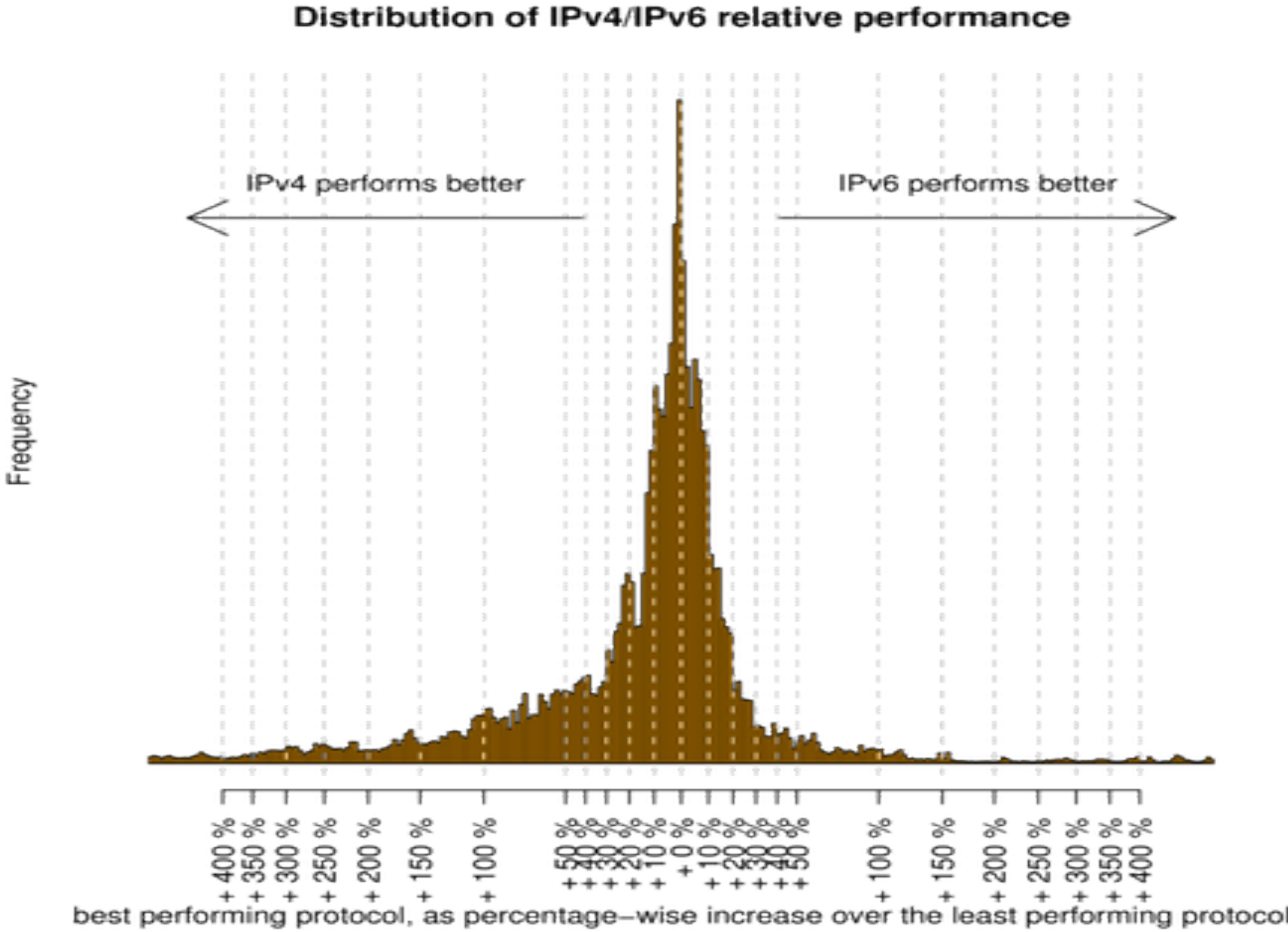
- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases
  - Datacenters
  - Smartphones
  - IPv4/IPv6 coexistence

IPv6 is coming ...



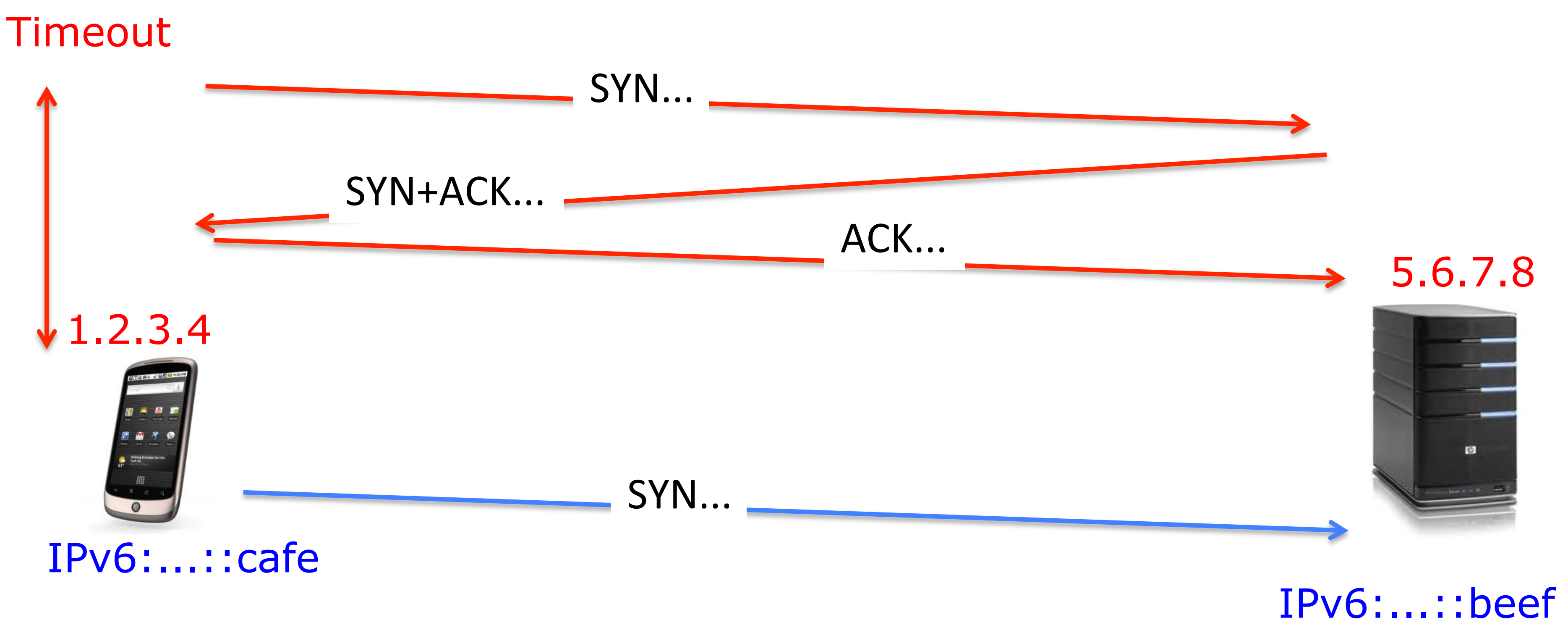
Source http://6lab.cisco.com/stats/cible.php?country=world

But IPv4 and IPv6 perf. may differ

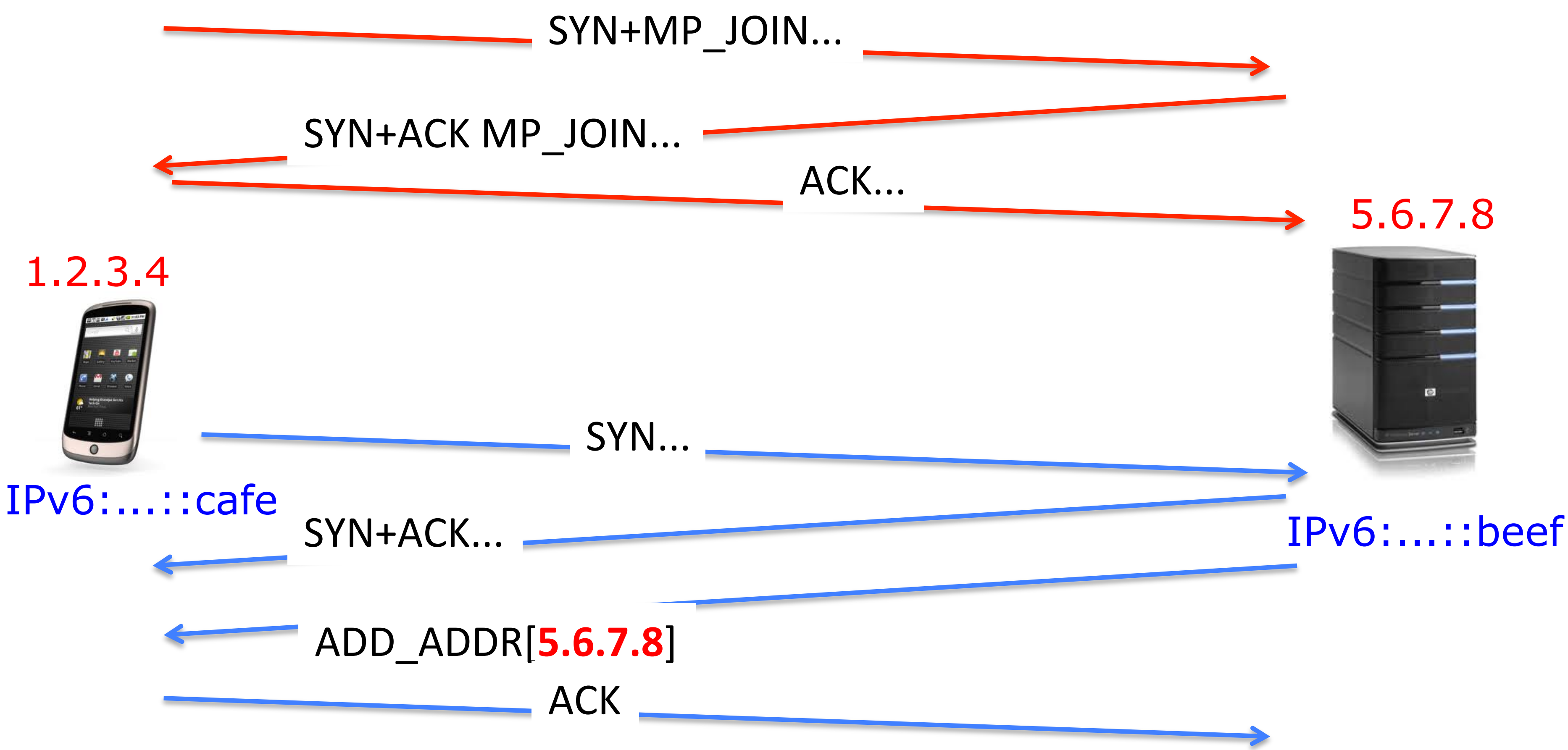


E. Aben, *Measuring World IPv6 Day - Comparing IPv4 and IPv6 Performance*, https://labs.ripe.net/Members/emileaben/measuring-world-ipv6-day-comparing-ipv4-and-ipv6-performance

Happy eyeballs



How to get best of IPv4 and IPv6 ?



Agenda

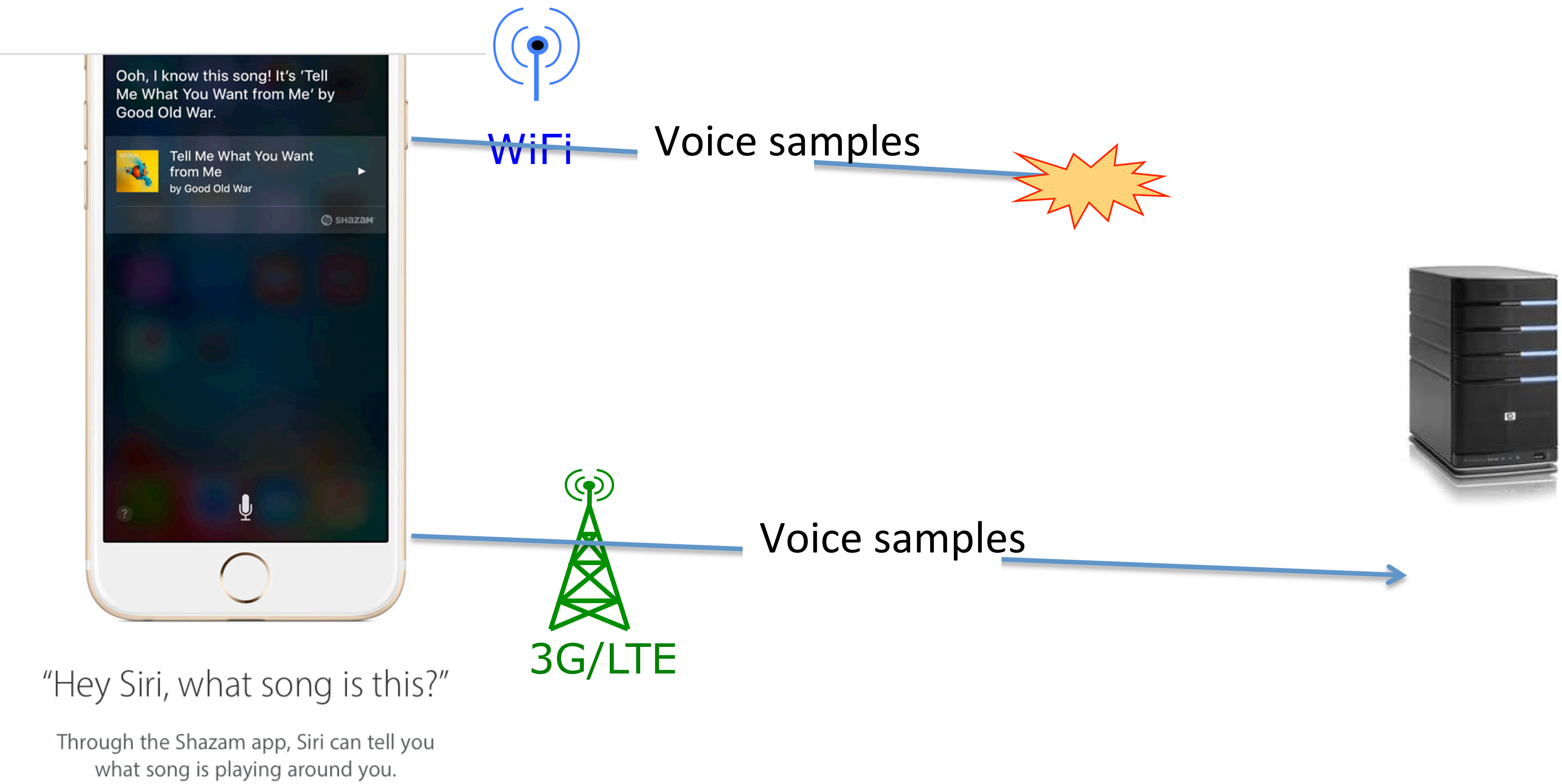
- The motivations for Multipath TCP
- The changing Internet
- The Multipath TCP Protocol
- Multipath TCP use cases
  - Datacenters
  - Smartphones
  - Commercial deployments

Multipath TCP use cases
The beast

VRT lanceert "The beast" voor videojournalisten

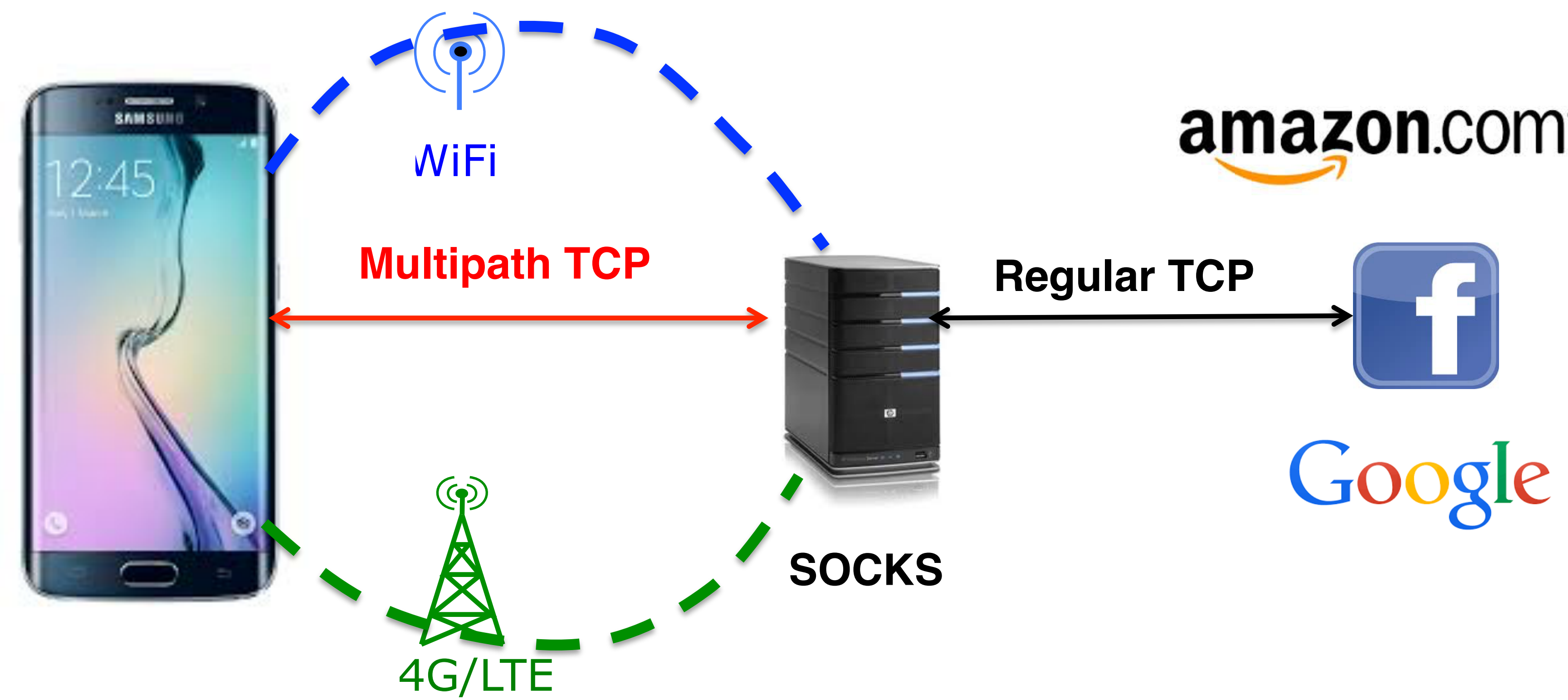
Multipath TCP use cases
Low latency for Siri

- Long-lived TLS connections

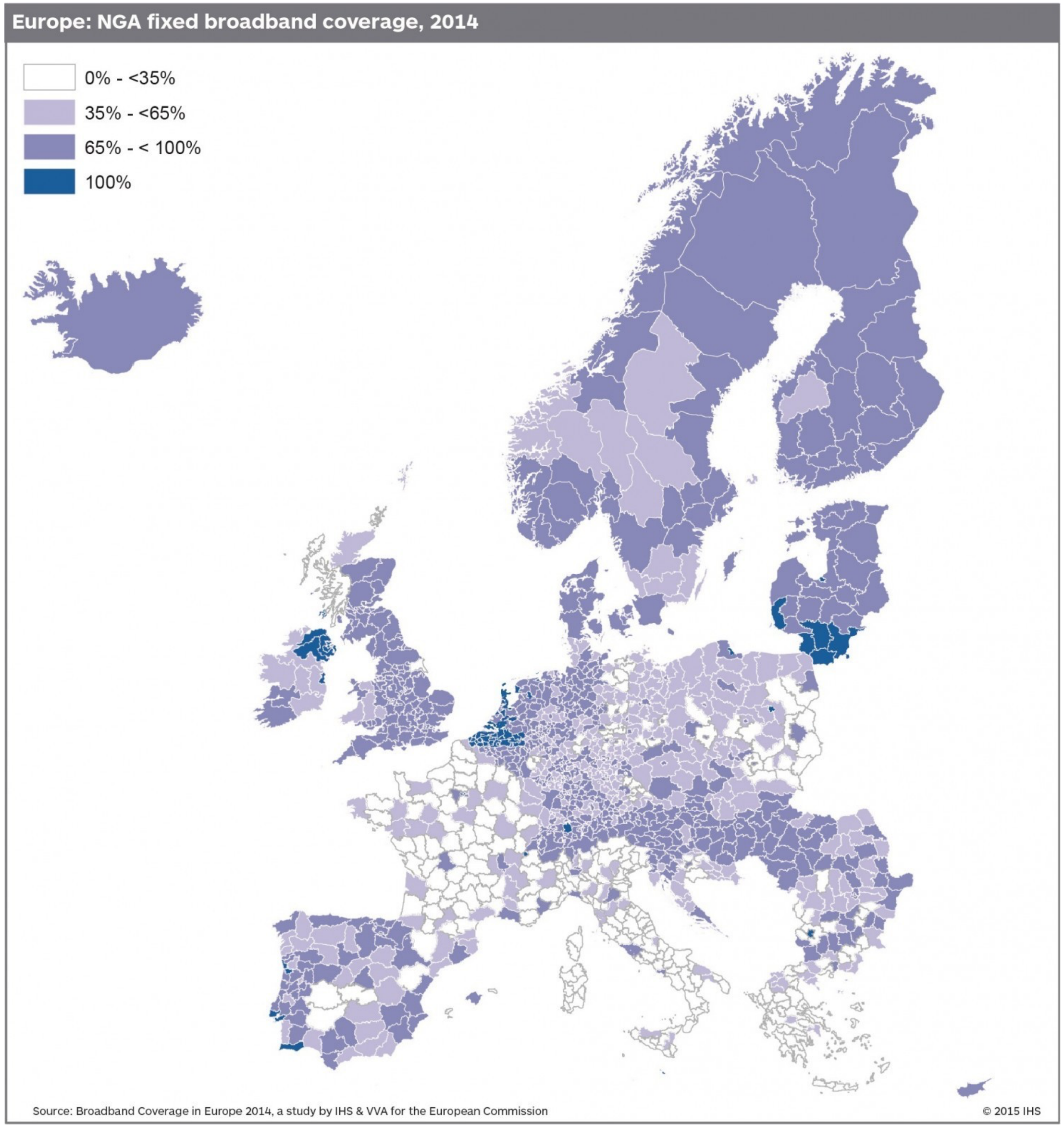


Multipath TCP use cases
High bandwidth on smartphones

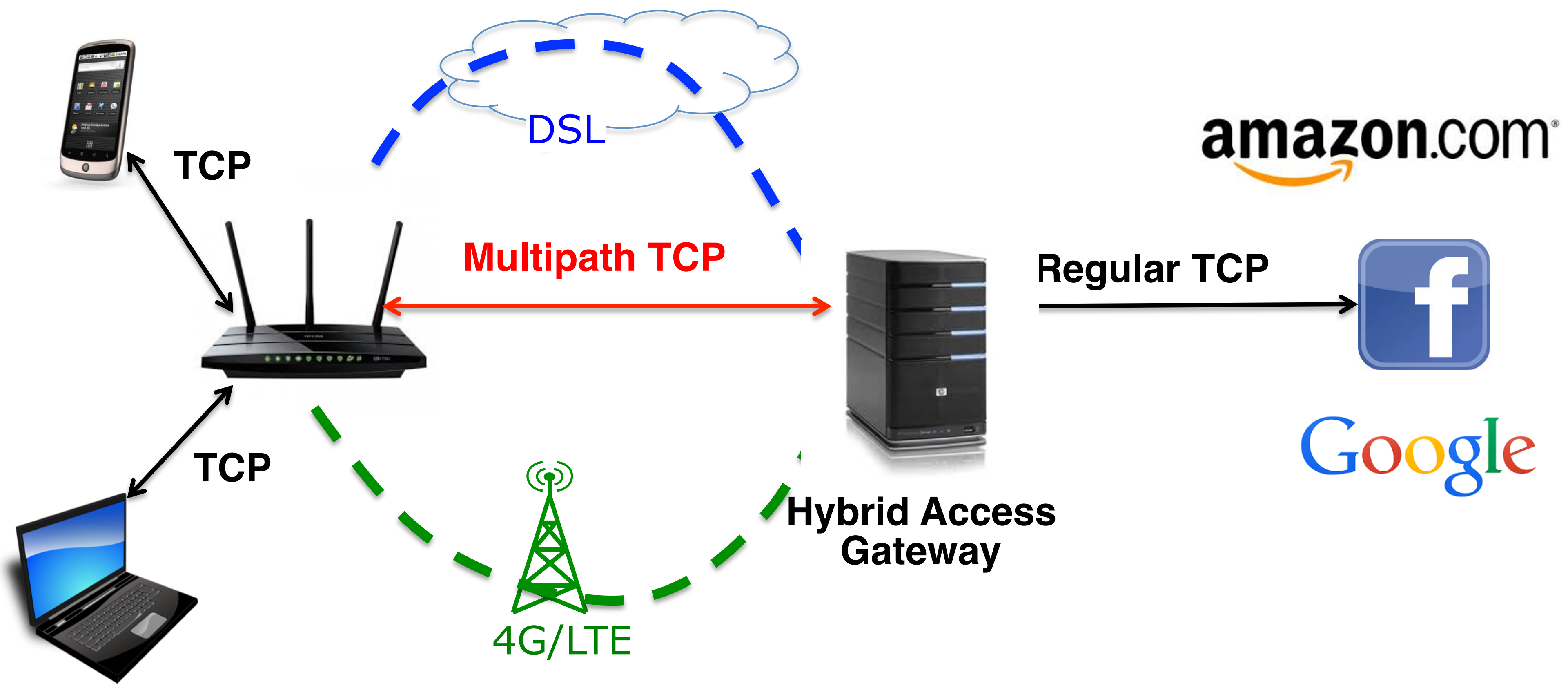
- Koreans want 800+ Mbps on smartphones



Faster broadband networks ?



Multipath TCP use cases
Hybrid Access Networks

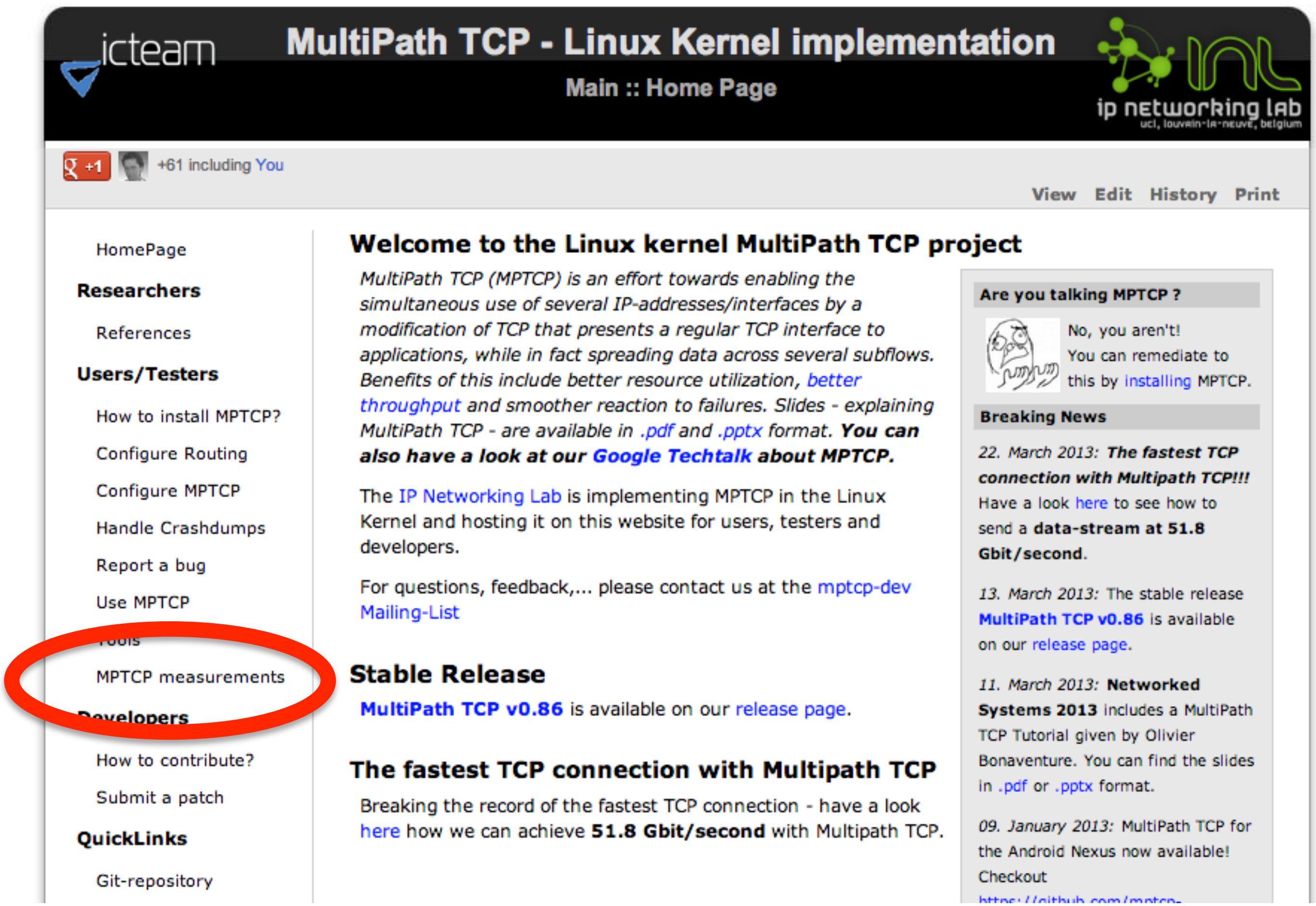


Conclusion

- Multipath TCP is becoming a reality
  - Due to the middleboxes, the protocol is more complex than initially expected
  - RFC has been published
  - there is running code !
  - Multipath TCP works over today's Internet !
- What's next ?
  - More use cases
  - Measurements and improvements to the protocol
    - Time to revisit 20+ years of heuristics added to TCP

Try it by yourself !

http://multipath-tcp.org



References

- The Multipath TCP protocol
  - http://www.multipath-tcp.org
  - http://tools.ietf.org/wg/mptcp/

A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, “Architectural guidelines for multipath TCP development", RFC6182 2011.

A. Ford, C. Raiciu, M. J. Handley, and O. Bonaventure, “TCP Extensions for Multipath Operation with Multiple Addresses,” RFC6824, 2013

C. Raiciu, C. Paasch, S. Barre, A. Ford, M. Honda, F. Duchene, O. Bonaventure, and M. Handley, “How hard can it be? designing and implementing a deployable multipath TCP,” NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

Implementations

- Linux
  - http://www.multipath-tcp.org

    S. Barre, C. Paasch, and O. Bonaventure, "Multipath tcp: From theory to practice," *NETWORKING 2011*, 2011.

    Sébastien Barré. Implementation and assessment of Modern Host-based Multipath Solutions. PhD thesis. UCL, 2011
- FreeBSD
  - http://caia.swin.edu.au/urp/newtcp/mptcp/
- Simulators
  - http://nrg.cs.ucl.ac.uk/mptcp/implementation.html
  - http://code.google.com/p/mptcp-ns3/

Middleboxes

M. Honda, Y. Nishida, C. Raiciu, A. Greenhalgh, M. Handley, and H. Tokuda, "Is it still possible to extend TCP?," IMC '11: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 2011.

V. Sekar, N. Egi, S. Ratnasamy, M. K. Reiter, and G. Shi, "Design and implementation of a consolidated middlebox architecture," *USENIX NSDI*, 2012.

J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratnasamy, and V. Sekar, "Making middleboxes someone else's problem: network processing as a cloud service," SIGCOMM '12: Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, 2012.

Multipath congestion control

– Background

D. Wischik, M. Handley, and M. B. Braun, “The resource pooling principle,” *ACM SIGCOMM Computer …*, vol. 38, no. 5, 2008.

F. Kelly and T. Voice. Stability of end-to-end algorithms for joint routing and rate control. ACM SIGCOMM CCR, 35, 2005.

P. Key, L. Massoulie, and P. D. Towsley, “Path Selection and Multipath Congestion Control,” INFOCOM 2007. 2007, pp. 143–151.

– Coupled congestion control

C. Raiciu, M. J. Handley, and D. Wischik, “Coupled Congestion Control for Multipath Transport Protocols,” *RFC*, vol. 6356, Oct. 2011.

D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, “Design, implementation and evaluation of congestion control for multipath TCP,” NSDI'11: Proceedings of the 8th USENIX conference on Networked systems design and implementation, 2011.

Multipath congestion control

– More

R. Khalili, N. Gast, M. Popovic, U. Upadhyay, J.-Y. Le Boudec , MPTCP is not Pareto-optimal: Performance issues and a possible solution, Proc. ACM Conext 2012

Y. Cao, X. Mingwei, and X. Fu, "Delay-based Congestion Control for Multipath TCP," ICNP2012, 2012.

T. A. Le, C. S. Hong, and E.-N. Huh, "Coordinated TCP Westwood congestion control for multiple paths over wireless networks," ICOIN '12: Proceedings of the The International Conference on Information Network 2012, 2012, pp. 92–96.

T. A. Le, H. Rim, and C. S. Hong, "A Multipath Cubic TCP Congestion Control with Multipath Fast Recovery over High Bandwidth-Delay Product Networks," *IEICE Transactions*, 2012.

T. Dreibholz, M. Becke, J. Pulinthanath, and E. P. Rathgeb, "Applying TCP-Friendly Congestion Control to Concurrent Multipath Transfer," Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, 2010, pp. 312–319.

Use cases

– Datacenter

C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. J. Handley, “Improving datacenter performance and robustness with multipath TCP,” *ACM SIGCOMM* 2011.

G. Detal, Ch. Paasch, S. van der Linden, P. Mérindol, G. Avoine, O. Bonaventure*,* *Revisiting Flow-Based Load Balancing: Stateless Path Selection in Data Center Networks*, Computer Networks, April 2013

– Mobile

C. Pluntke, L. Eggert, and N. Kiukkonen, “Saving mobile device energy with multipath TCP,” *MobiArch '11: Proceedings of the sixth international workshop on MobiArch*, 2011.

C. Paasch, G. Detal, F. Duchene, C. Raiciu, and O. Bonaventure, “Exploring mobile/WiFi handover with multipath TCP,” CellNet '12: Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012.