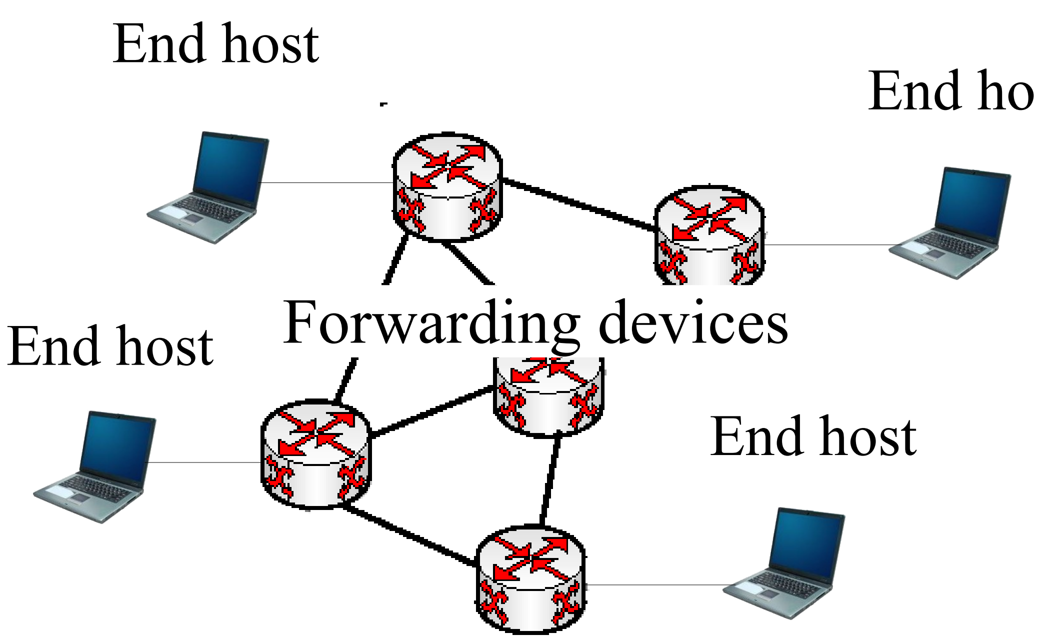
Propositions for a robust and inter-operable eXplicit Control Protocol on Heterogeneous High Speed Networks

PhD dissertation defense by
LÓPEZ PACHECO Dino Martín

This work has been supported by CONACyT

Introduction

Networking today



Web,
Databases,
Remote imaging
Peer-to-peer,
Data Grid...

Networks:

- Allow equipments (end hosts) to exchange data packets (video, audio, data).
- Provide the infrastructure for distributed applications and services.

Network congestion

- Big success of networks = Overload of networks (congestion).
- Congestion may prevent the exchange of data.
- Congestion control protocols:
  - End-to-End (E2E)
  - Routers-assisted

End-to-End protocols

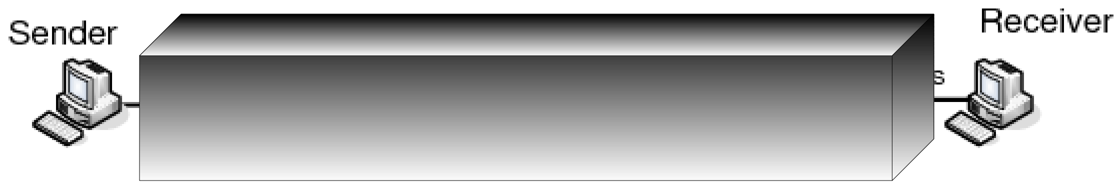
End-to-End (E2E) protocols are the most widely deployed protocols in networks.

- E2E protocols only implements their mechanisms in the end hosts.
- They are independent to the network infrastructure

Several E2E protocols exist today: Transport Control Protocol (TCP) [RFC1122], High Seed TCP [S. Floyd - RFC3649], BIC [L. Xu - INFOCOM2004], Compound TCP [K. Tan - INFOCOM2006], etc.

Limits of E2E protocols

However, networks are like black boxes for E2E protocols.



For this reason, E2E protocols :

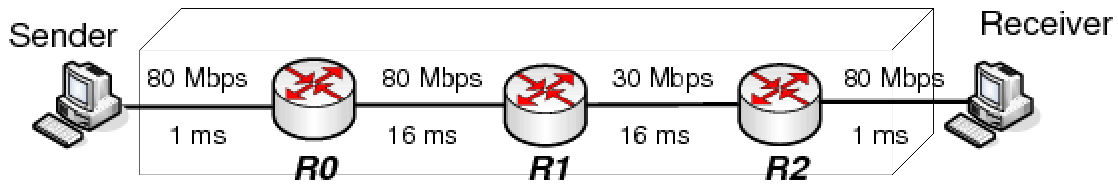
- Are unable to know the real state of the resources.
- Lead to congestion periodically.
- Responsiveness strongly affected by the propagation delay.
- Different RTTs can lead to unfairness.

Efforts to more accurately know the state of the network

Some approaches to control congestion by mean of mechanisms inside the routers were proposed:

- Active Queue Management (AQM) mechanisms: Routers drop randomly packets when congestion is “imminent”. Ex. Random Early Detection (RED) [S. Floyd & V. Jacobson ACM Trans. on Networking 1993]
- Explicit Congestion Notification (ECN [RFC3168]): Routers send a signal to end hosts when congestion is “imminent”.
- Explicit Rate Notification (ERN) protocols: Routers provide explicit sending rate to the senders.

ERN protocols



Since routers provide explicit rate notification :

- ERN protocols are able to fairly share the resources while maximizing their utilization.
- ERN protocols are less affected than E2E protocols by large RTTs.
- Losses of packets rarely happen in fully ERN networks.

Some ERN protocols: XCP [D. Katabi – ACM SIGCOMM 2002], JetMax [D. Leonard – INFOCOM 2006], Quickstart [S. Floyd RFC4782], etc.

Limits of ERN protocols

ERN protocols only work well in fully ERN networks, they are :

- Not inter-operable with current E2E protocols.
- Not inter-operable with current IP routers.
- Very sensitive to feedback loop.

This thesis addresses such problems.

Context

My propositions have been specially designed for :

- Wire-based heterogeneous large *bandwidth-delay product* (BDP) networks.

- Networks where long-life flows are frequent.
For instance: Data Grid networks.

Outline

1. TCP, High Speed TCP & XCP on large BDP networks and Variable Bandwidth Environment (VBE).
2. Propositions to provide XCP-TCP friendliness.
3. A new architecture for a more robust XCP protocol.
4. Propositions to provide interoperability between XCP and non-XCP routers.
5. Discussion & Concluding Remarks.

1. **TCP, High Speed TCP & XCP on large BDP networks and Variable Bandwidth Environment (VBE).**
2. Propositions to provide XCP-TCP friendliness.
3. A new architecture for a more robust XCP protocol.
4. Propositions to provide interoperability between XCP and non-XCP routers.
5. Discussion & Concluding Remarks.

The TCP congestion control

End-to-End (E2E) protocol.

Slow-Start (SS)

- *cwnd = cwnd + 1*.

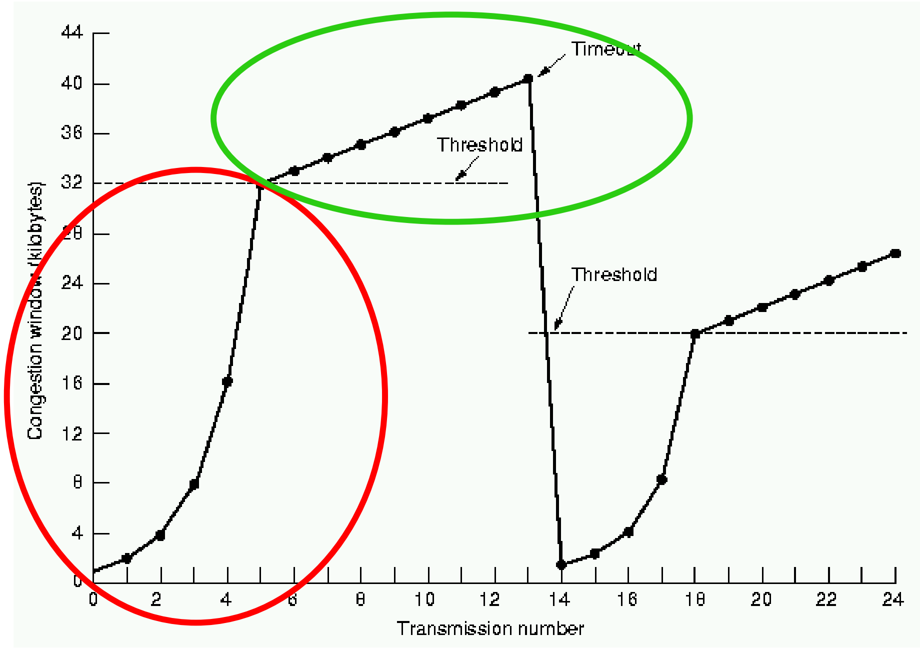
Congestion Avoidance (CA)

- *cwnd = cwnd + 1/cwnd*.

In case of losses

- *cwnd = 1 MSS or*
- *cwnd = cwnd – 1/2*cwnd*

MSS = Maximum Segment Size



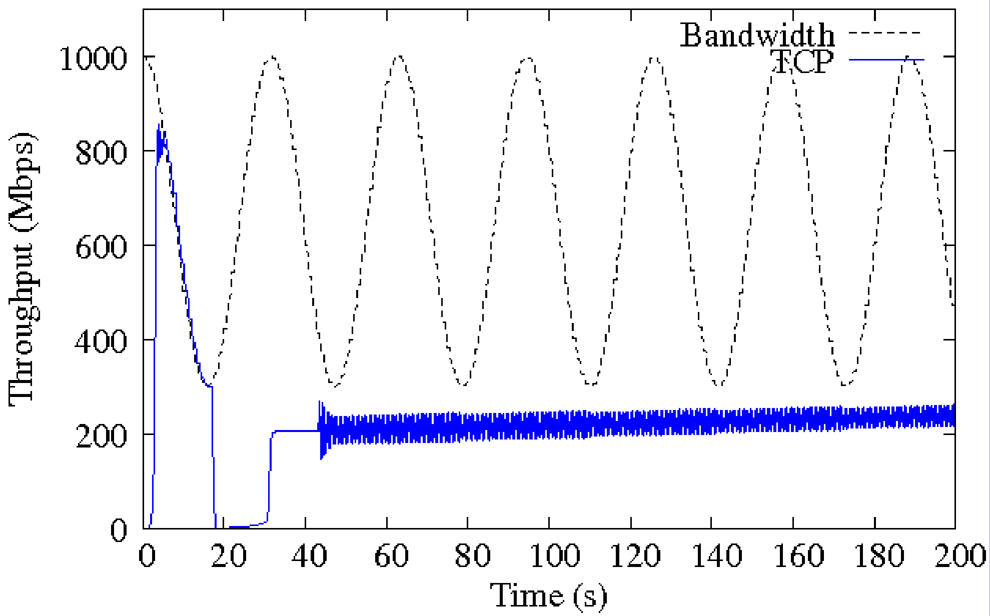
From Computer Networks, A. Tanenbaum

TCP in a large BDP network with VBE

In networks, several factors may lead to Variable Bandwidth Environments (VBE).

We tested TCP (New Reno) in a VBE. Bandwidth variations describing a sinusoidal pattern.

- Minimum bandwidth ≈ 300 Mbps, Maximum bandwidth ≈ 1000Mbps.
- Buffer ≈ 12500 MSS
- *RTT ≈ 200ms*



After losses TCP is unable to quickly recover resources in large BDP networks: Alternatives to TCP have been proposed

High Speed TCP (HSTCP)

TCP-based E2E protocol : One of the first high speed version of TCP.

Slow-Start

- Introduction of "Limited Slow-Start" algorithm.

Congestion Avoidance

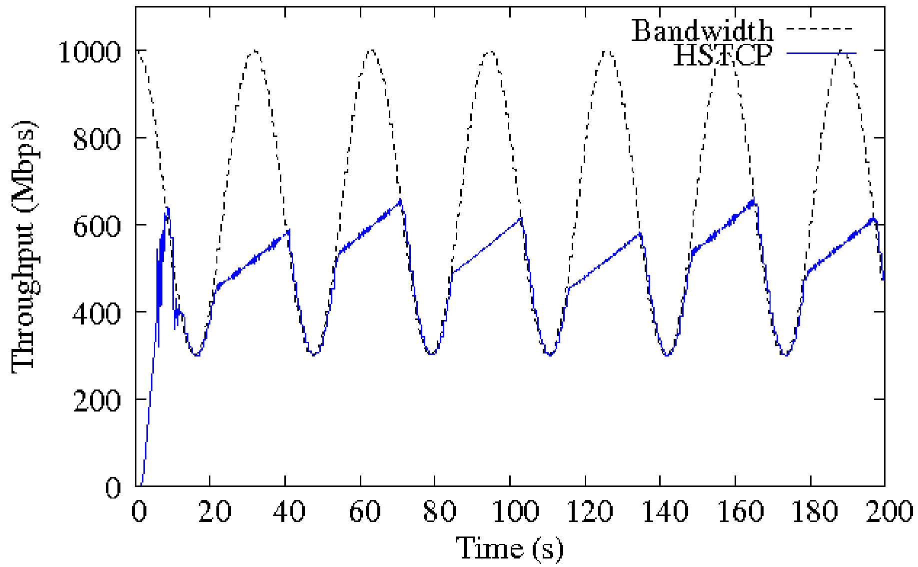
- $cwnd = cwnd + a(cwnd)/cwnd$

In case of Losses

- $cwnd = cwnd - b(cwnd)*cwnd$

HSTCP in a large BDP network with VBE

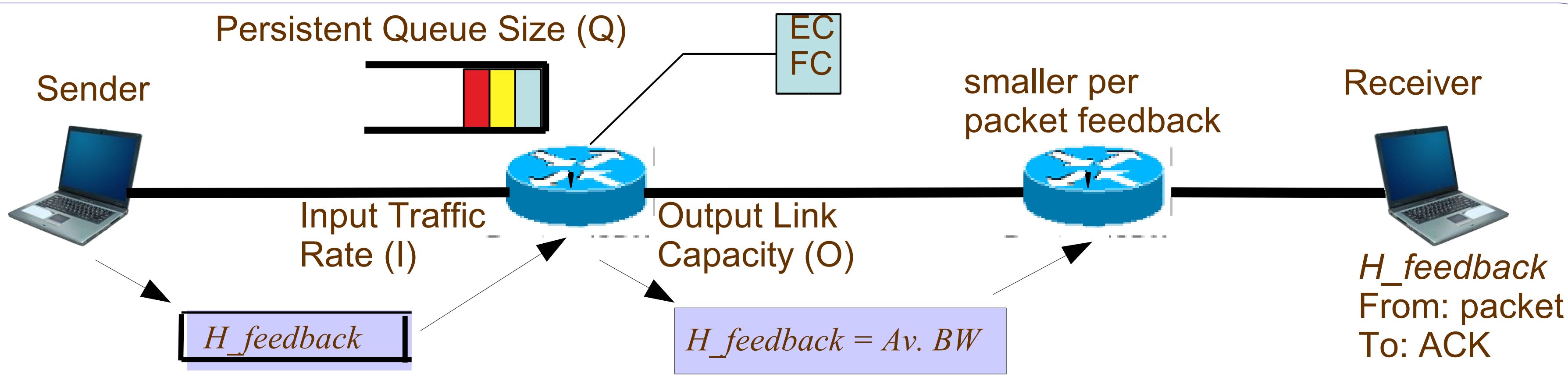
HSTCP under the same conditions as TCP ($RTT \approx 200ms$).



Better responsiveness than TCP.
However the RTT value affects the responsiveness of HSTCP

Non E2E alternatives have been proposed

eXplicit Control Protocol (XCP) [Katabi 02]

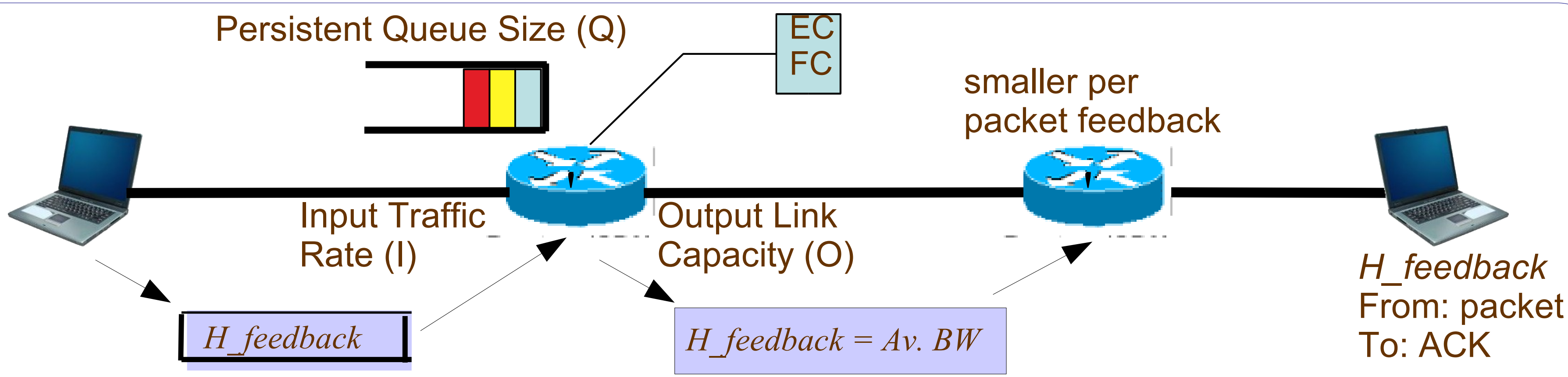


- XCP routers provide Explicit Rate Notification (ERN protocols).
- XCP routers execute two control laws to compute a feedback per packet:
  - Efficiency Controller (EC). Computes the available bandwidth (the general feedback, ϕ).

    $\phi = \alpha .rtt.(O\text{-}I) - \beta .Q \quad rtt = \text{control interval}, \alpha = 0.4, \beta = 0.226$

  - Fairness Controller (FC). Fairly assign resources (bandwidth) between XCP flows.

eXplicit Control Protocol (XCP) [Katabi 02]

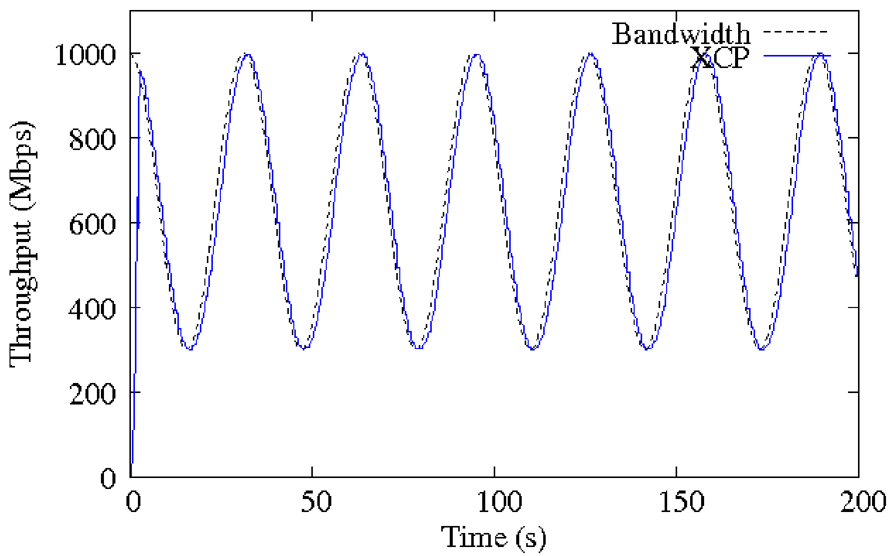


XCP :

- Assigns a portion of bandwidth in every data packet (feedback per packet).
  - Does not keep any state per flow.
- Sends feedback to the sender in every ACK.
  - Does not introduce overhead into the network.
- Data packets do not queue up in routers buffers.

XCP in a large BDP network with VBE

XCP in a fully XCP network under the same conditions as TCP and HSTCP ($RTT \approx 200ms$).



- The responsiveness of XCP is not affected by large RTTs.

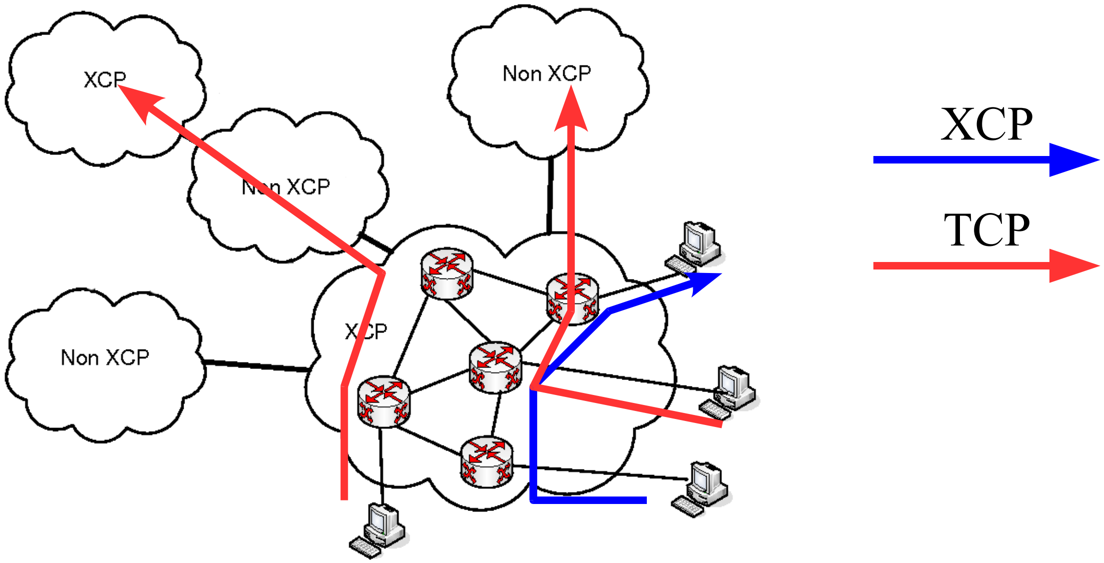
Lessons learned so far

- E2E protocols are sensible to bandwidth variations and RTT values.
- In large BDP networks with VBE, E2E protocols frequently have problems to
  - Correctly grab resources.
  - Correctly yield resources.
  - Fairly share the resources.
- ERN protocols perform well in large BDP networks with VBE.
- Interoperability problems:
  - No friendliness with other E2E protocols (TCP).
  - Non-interoperability with non-ERN equipments.
  - Sensitivity to feedback losses.

1. TCP, High Speed TCP & XCP on large BDP networks and Variable Bandwidth Environment (VBE).
2. **Propositions to provide XCP-TCP friendliness.**
3. A new architecture for a more robust XCP protocol.
4. Propositions to provide interoperability between XCP and non-XCP routers.
5. Discussion & Concluding Remarks.

Deploying XCP in heterogeneous networks

Adding XCP clouds in the network.



In order to exchange data:

- Hosts in the XCP sites will use the XCP protocol.
- Hosts from other sites will use TCP-based protocols.

Problem: No TCP-XCP friendliness mechanism

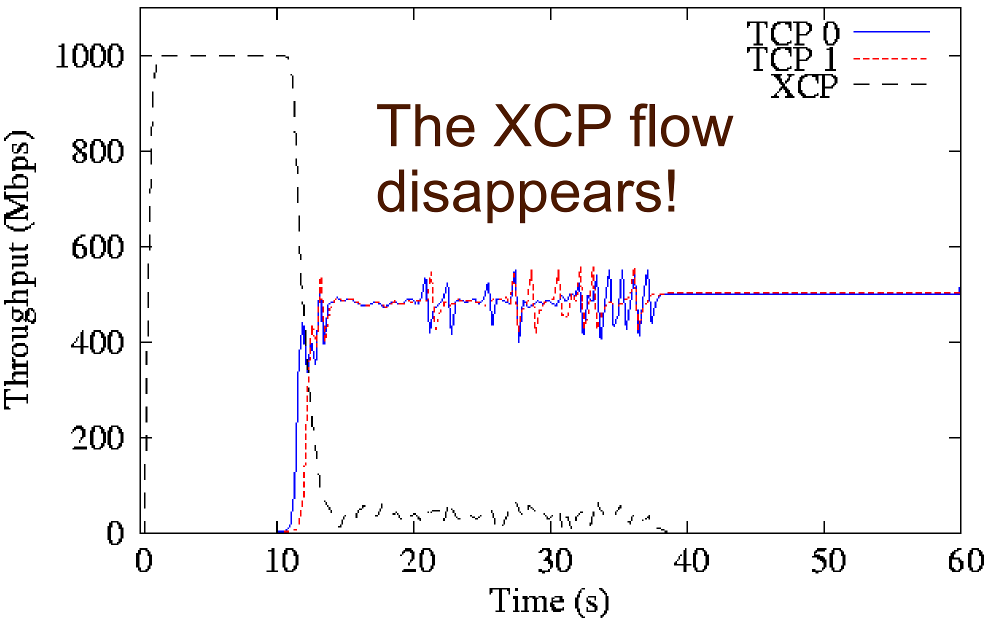
XCP and non-XCP protocols

XCP general feedback equation:

$$\phi = \alpha .rtt.(O\text{-}I) - \beta .Q$$

ϕ decreases as the I increases. However, $I = \sum$ packet size of every incoming packet (XCP, TCP, UDP, etc.)

When I will increase, XCP flows will decrease the rate in profit of non-XCP protocols.



Goals for a new XCP-TCP friendliness solution

I propose a solution which provides XCP-TCP friendliness : ***XCP-f.***

XCP-f is:

- Lightweight in terms of CPU and memory consumption.
- Easy to adapt to others ERN protocols.

[D. Lopez, L. Lefèvre & C. Pham. HSN 2007, IFIP Networking 2008]

Providing XCP-TCP friendliness with XCP-f

- XCP-TCP friendliness is obtained when the bandwidth of XCP, BW_{XCP}

$$BW_{XCP} = \# XCP\ flows \quad * \quad \frac{Link\ Capacity}{\# XCP\ flows + \# TCP\ flows}$$

- To know BW_{XCP}, it is necessary to estimate the # of XCP and non-XCP flows.
- It is difficult and expensive to obtain the accurate number of flows.
- We adapt an SRED-like zombie estimation method [Teunis – INFOCOM 1999], which probabilistically estimates the active number of flows.

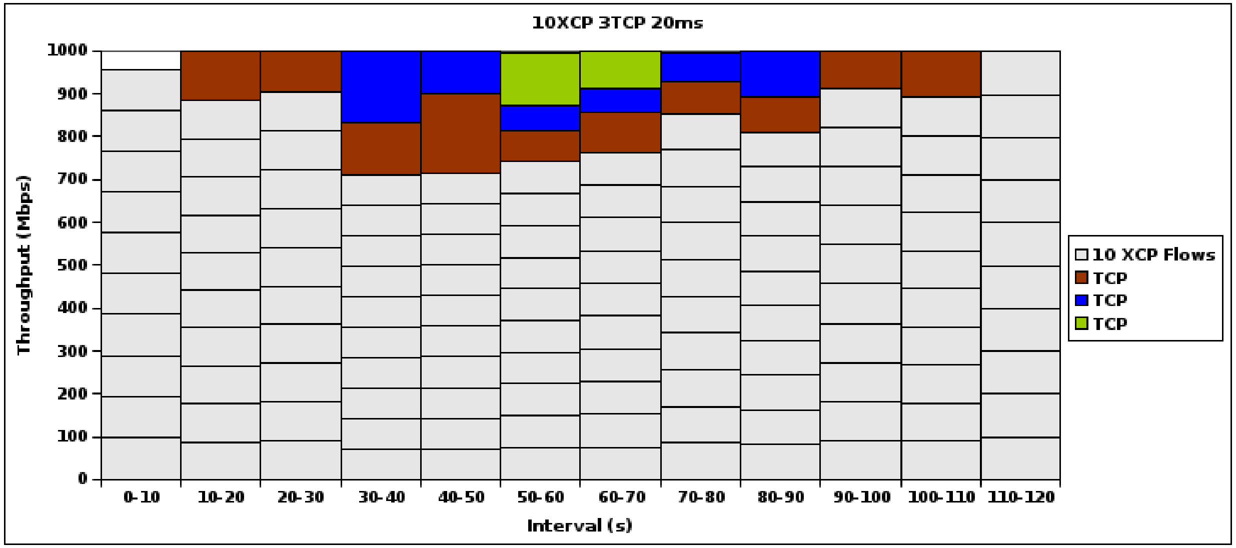
Limiting TCP throughput

- XCP-f compares the bandwidth needed by XCP to get friendliness (BW_{XCP}) with its current throughput, Th_{XCP}.
- When $BW_{XCP} > Th_{XCP}$, drop TCP packets with a probability p.
- Update p as follows at every XCP control interval

If ($BW_{XCP} < Th_{XCP}$) then
$p = p * Ddrop$ // $Ddrop < 1$
else If ($BW_{XCP} > Th_{XCP}$) then
$p = p * Idrop$ // $Idrop > 1$

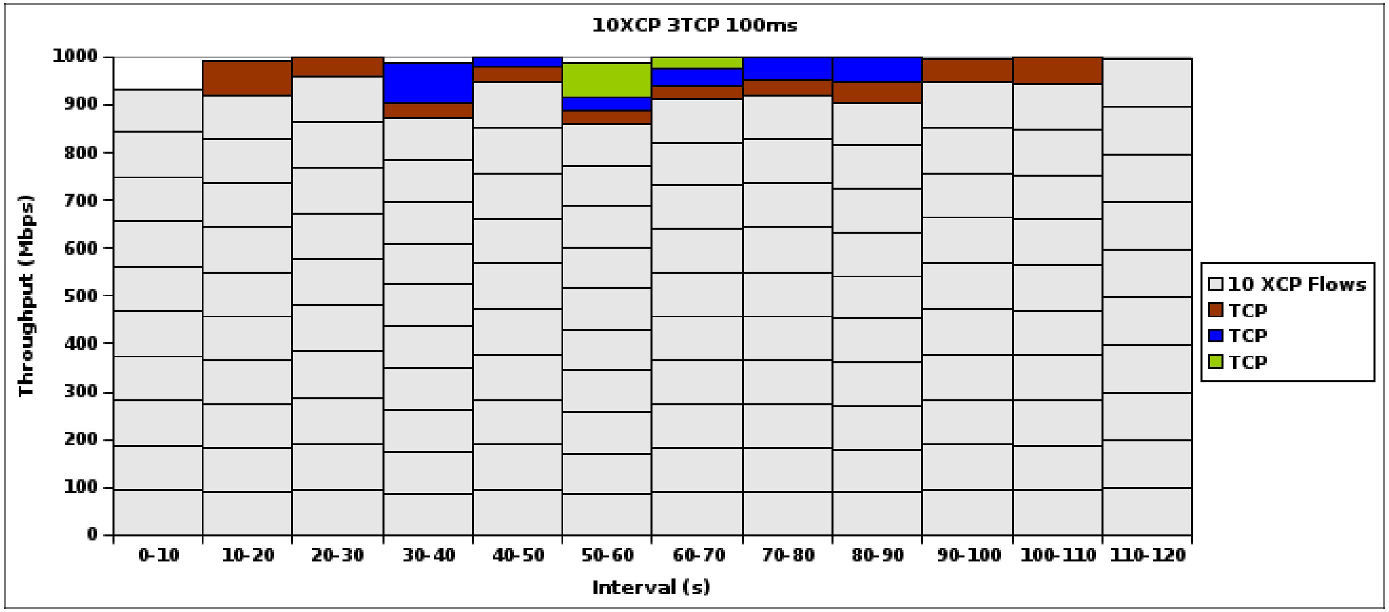
10 XCP-f and 3 TCP flows sharing a bottleneck (RTT ≈ 20ms)

- TCP Flows arriving at seconds 10, 30 and 50 among 10 XCP-f flows.
- Every column contains the average throughput of every active flow during 10s.



- Easy to identify the Slow-Start effect (aggressive behavior of TCP).
- XCP-f successfully limits the TCP throughput.
- After Slow-Start, flows get stability.
- During the seconds 60-70, $BW_{XCP} \approx 787Mbps$ and $Th_{XCP} \approx 750Mbps$

10 XCP-f and 3 TCP flows sharing a bottleneck (RTT ≈ 100ms)



- After dropping TCP packets to limit the TCP throughput, TCP flows suffer to regain bandwidth (due to the RTT).
- During the seconds 60-70, $BW_{XCP} \approx 787Mbps$ *and* $Th_{XCP} \approx 920Mbps$

XCP-TCP cohabitation

- Without XCP-f, XCP only gets the remaining bandwidth (0).
- XCP-f successfully provides XCP-TCP friendliness.
- E2E protocols (TCP) can cohabit with XCP.

In wire-based networks, burst of packets from E2E protocols can produce multiple packet losses in a very short period of time.

Effect of packets losses on E2E & ERN protocols

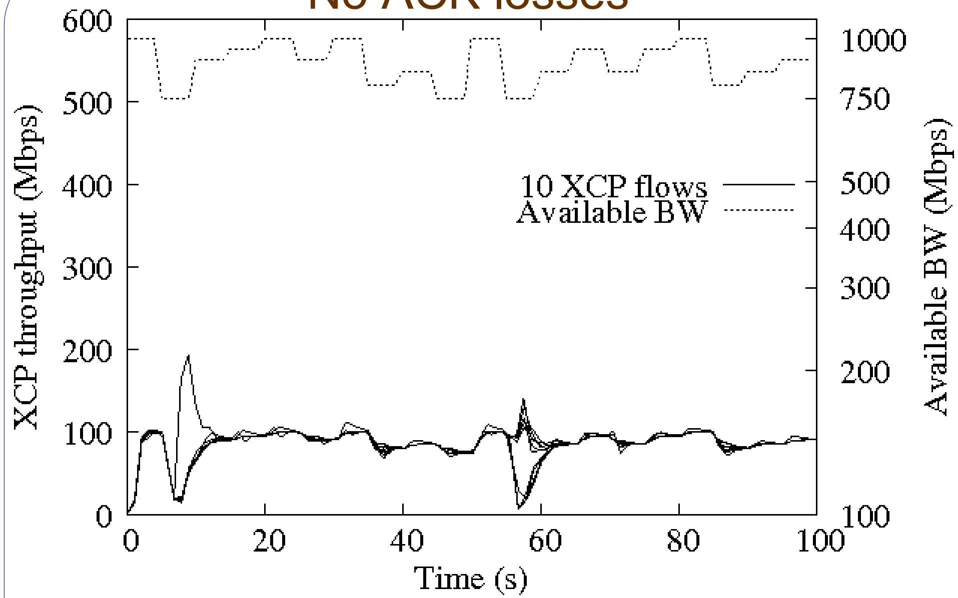
- In E2E protocols, losses of data packets lead to a decrease of the sending rate. In ERN protocols, losses of data packets do not impact the rate of the senders.
- In E2E protocols, losses of ACK only (insignificantly) delay the sliding of the congestion window. In ERN protocols, ACK losses also imply losses about the network state information.

Armor XCP against feedback (ACK) losses.

1. TCP, High Speed TCP & XCP on large BDP networks and Variable Bandwidth Environment (VBE).
2. Propositions to provide XCP-TCP friendliness.
3. **A new architecture for a more robust XCP protocol.**
4. Propositions to provide interoperability between XCP and non-XCP routers.
5. Discussion & Concluding Remarks.

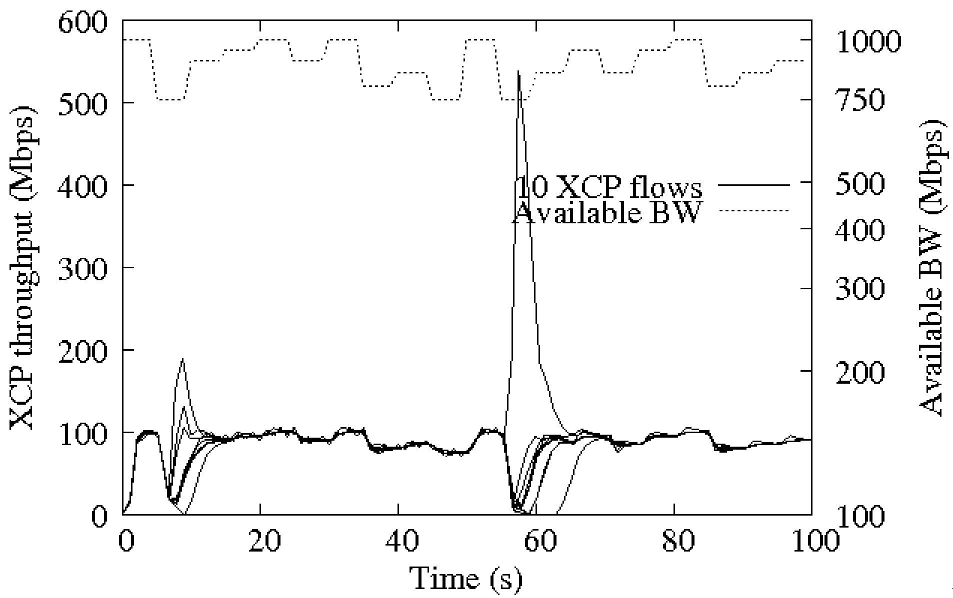
Impact of ACK losses on the XCP behavior

No ACK losses



- 10 XCP flows share the bottleneck
- Variable Bandwidth Environment:
  - 750Mbps < BW < 1Gbps
  - Step-based variation model

10% ACK loss rate



- ACK losses can lead to chaotic behavior of XCP in VBE.

Increasing the robustness with the XCP-r architecture

Since ACK losses lead to a wrong congestion window size in the sender, the ***XCP-r*** architecture:

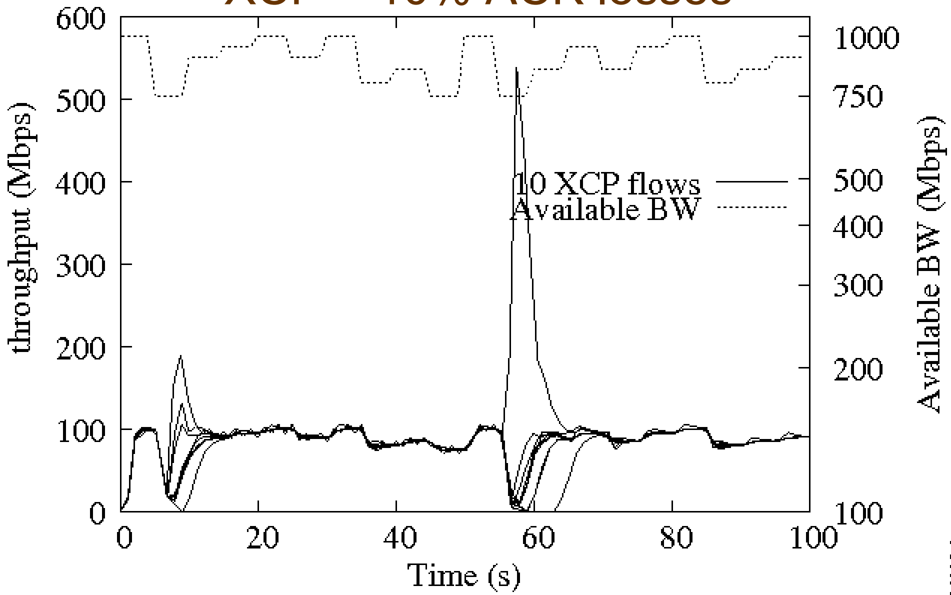
- Transfers a part of the XCP code from the sender to the receiver.
- Computes the congestion window size in the receiver instead of the sender.
- Adds some mechanisms to avoid unsynchronization between the sender and the receiver.

[D. Lopez & C. Pham. MICC-ICON 2005, ICN 2006]

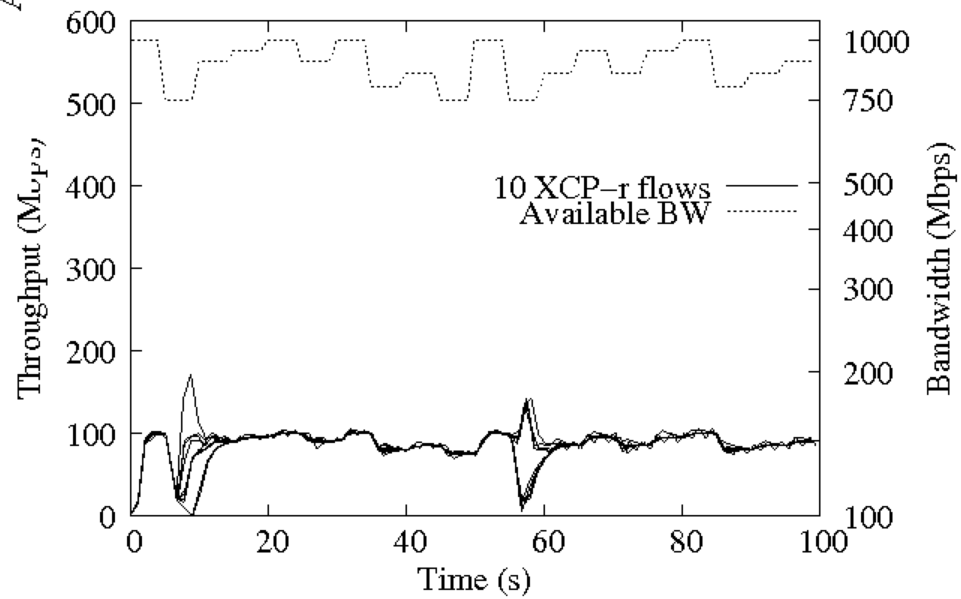
XCP-r is easy to adapt to other ERN protocols.

Benefits of XCP-r

XCP – 10% ACK losses



XCP-r – 10% ACK losses



- The XCP-r architecture provides robustness to XCP in presence of ACK losses in a VBE.
- Less chaotic behavior of flows.

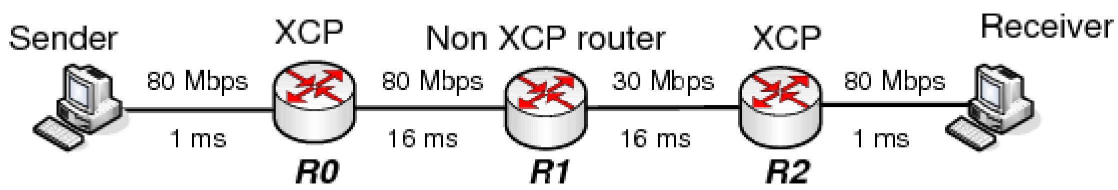
Interoperability issues

We have a robust XCP protocol able to cohabit with TCP. However, Full ERN networks only exist in labs but not in real networks.

We need solutions to provide the interoperability between XCP and non-XCP routers

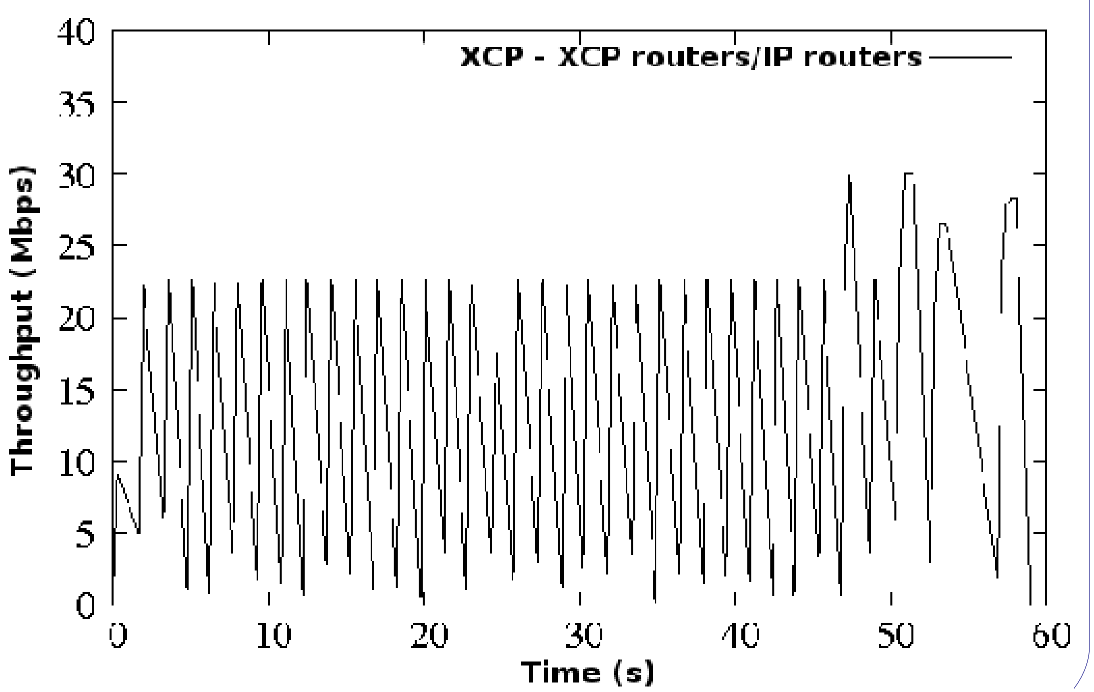
1. TCP, High Speed TCP & XCP on large BDP networks and Variable Bandwidth Environment (VBE).
2. Propositions to provide XCP-TCP friendliness.
3. A new architecture for a more robust XCP protocol.
4. **Propositions to provide interoperability between XCP and non-XCP routers.**
5. Discussion & Concluding Remarks.

XCP in the presence of legacy IP routers



- Unknown bottleneck capacity due to the presence of a non-XCP router.

- Very unstable behavior



Interoperability between XCP and non-XCP routers with XCP-i

XCP-i is the first step towards the interoperability between XCP and non-XCP equipments.

XCP-i:

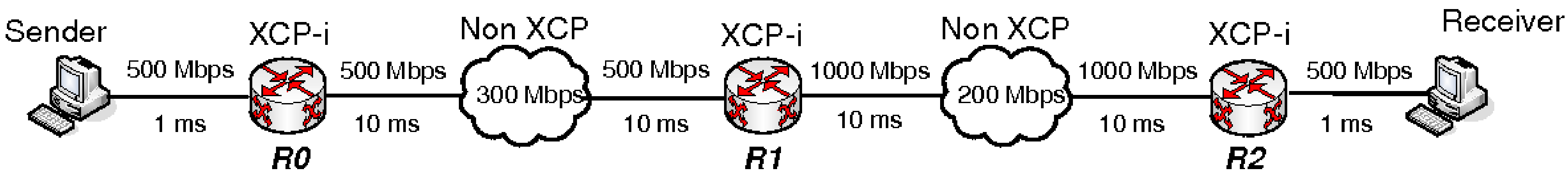
- Keeps the XCP algorithm as in the original model.
- Reduces as much as possible the use of memory resources.
- Avoids keeping per flow states.
- Is easy to adapt to other ERN protocols.

[D. Lopez, L. Lefèvre & C. Pham. Globecom 2006, CFIP 2006]

Some definitions:

1. XCP-i: XCP router supporting our algorithm.
2. Non XCP cloud: Set of n contiguous non-XCP routers, where $n > 0$

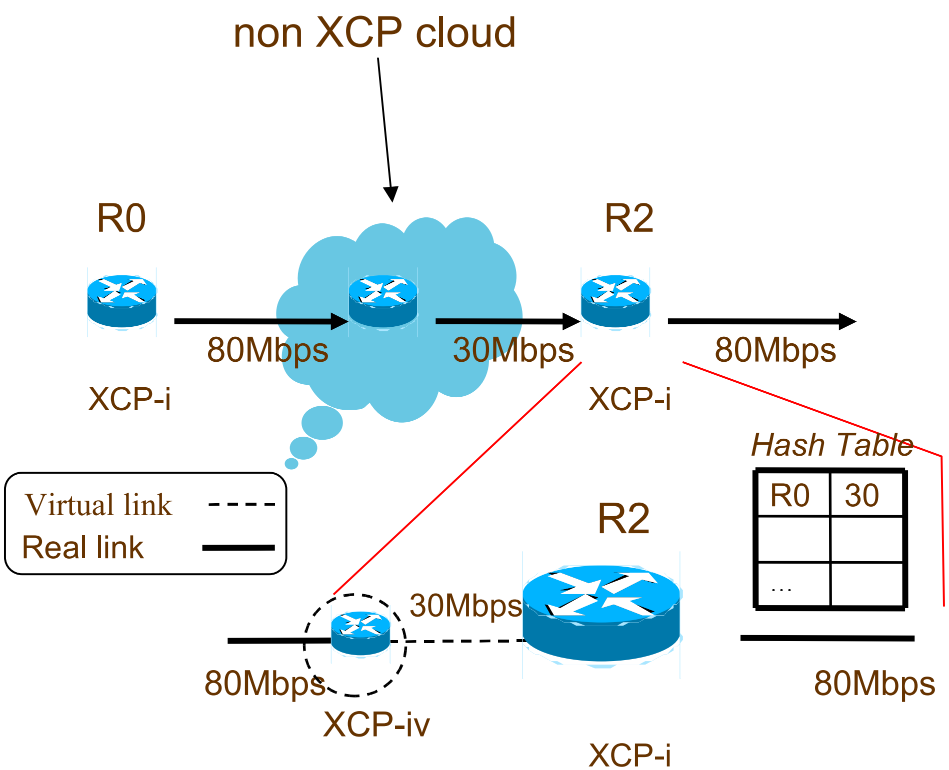
Core mechanisms of XCP-i



XCP-i :

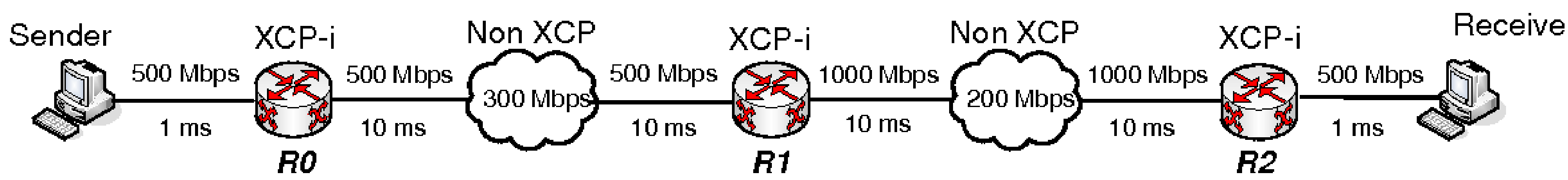
- Detects the non-XCP clouds.
  - The dual-TTL strategy
- Estimates the resources <u>only in the non-XCP clouds</u>.
  - Identify the edge XCP-i routers of the non-XCP clouds.
  - Estimate the available bandwidth into the non-XCP cloud.
- Provides a feedback which reflects the state of the non-XCP clouds.
  - The virtual XCP-i router.

Creating a virtual XCP-i router

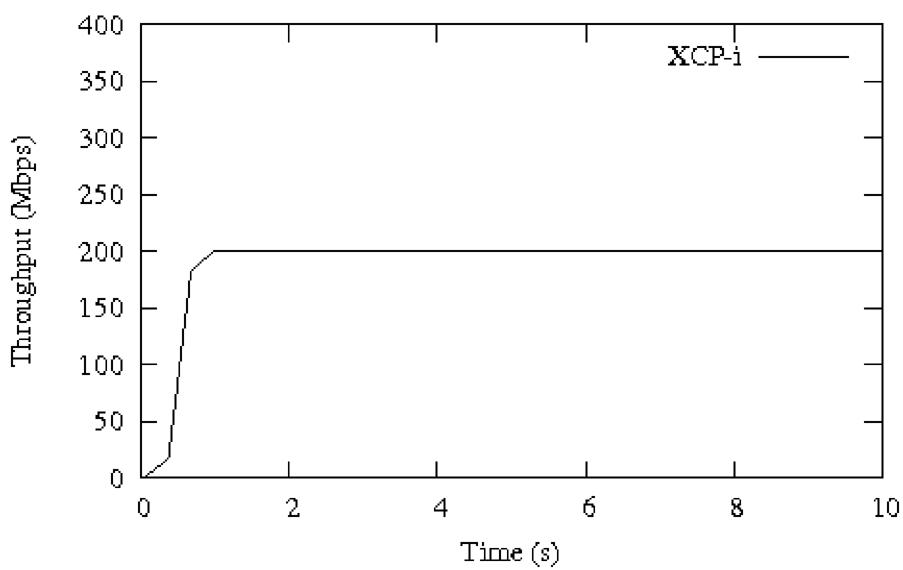


- Router discovering the non-XCP cloud must create a virtual XCP-i router.
- Virtual XCP-i routers compute a feedback reflecting the state in the non-XCP clouds.
- Advantage : Virtual routers can simply reuse the code of the XCP routers.

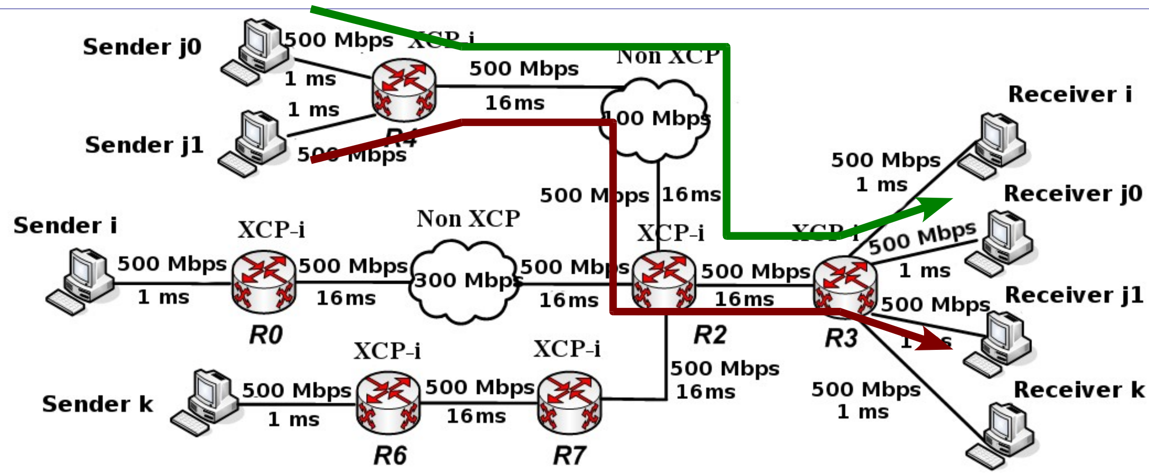
Validating XCP-i (1)



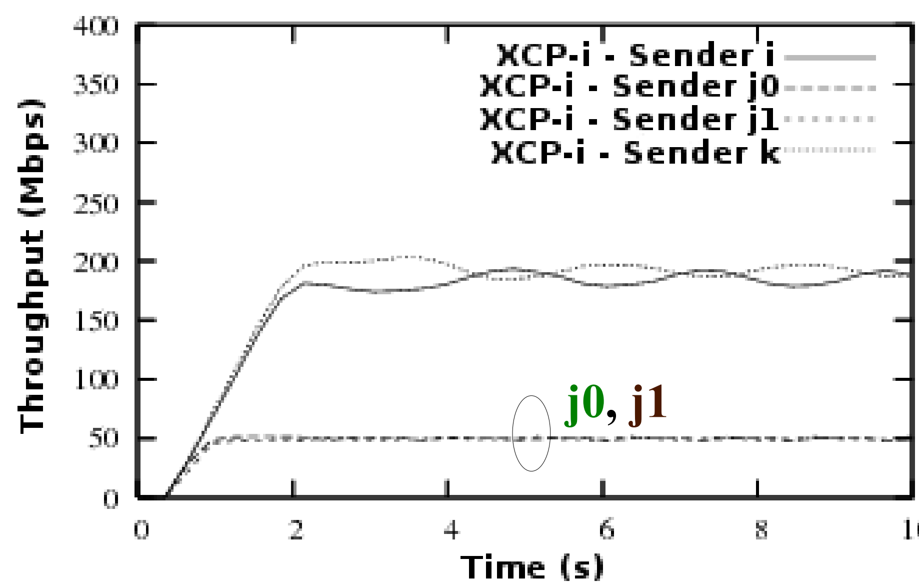
- XCP-i correctly detects the non-XCP clouds and provides accurate feedback.



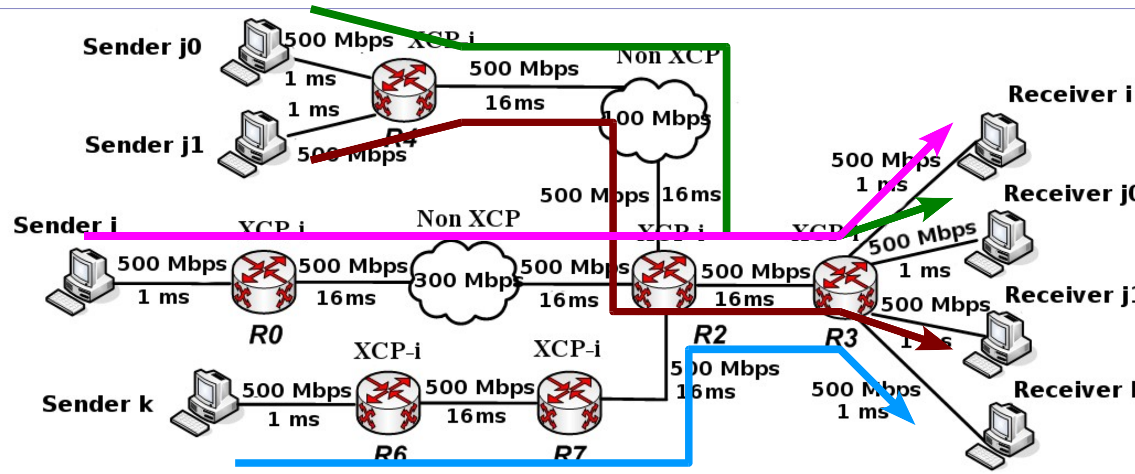
Validating XCP-i (2)



- Flow j0 and j1 ≈ 50Mbps.
- Good fairness and stability properties without a full XCP network.

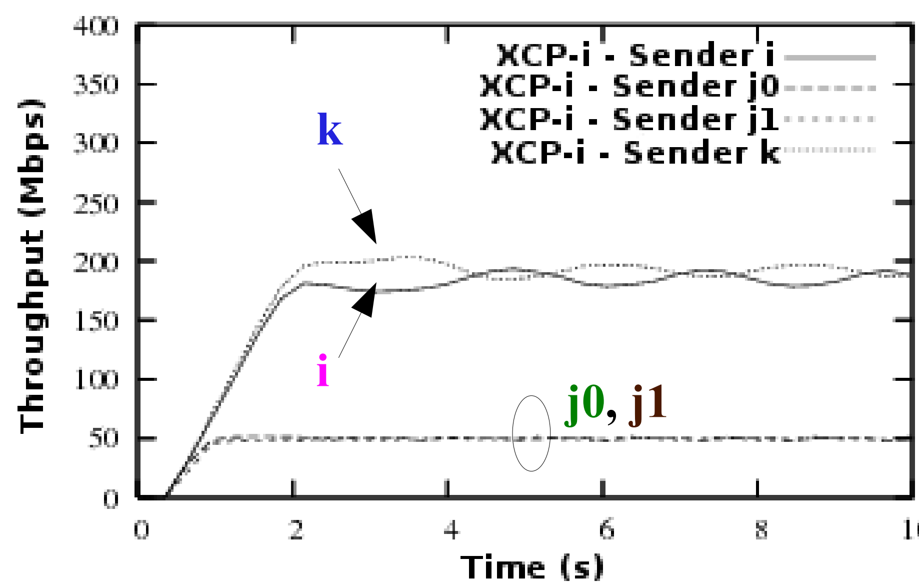


Validating XCP-i (2)

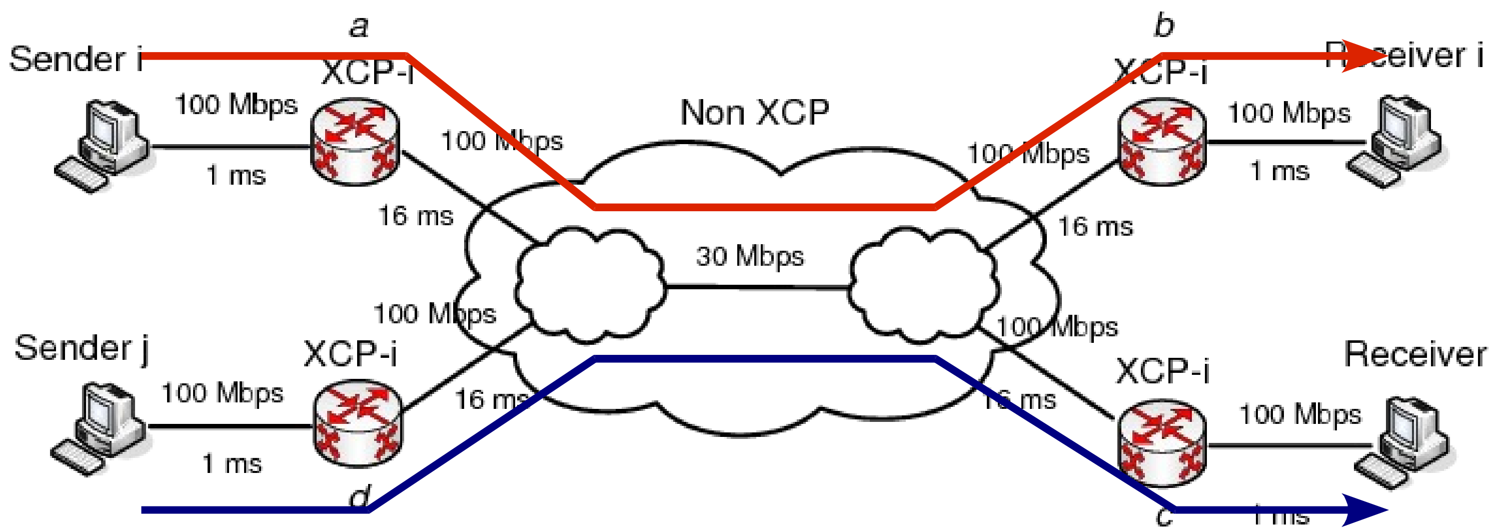


- Flow j0 and j1 ≈ 50Mbps.
- Flow i and k ≈ 200Mbps.

- Good fairness and stability properties without a full XCP network.



Future works for XCP-i



- In some complex topologies, it is difficult to detect when several XCP flows share the same bottleneck.
  - 1 XCP flow can take most of the resources preventing the other one.
- Preliminary solutions:
  - Develop tools to detect the bottleneck.
  - Use broadcast to communicate the bottleneck between the edge XCP-i virtual routers.

1. TCP, High Speed TCP & XCP on large BDP networks.
2. Interoperability of XCP with current technologies
    2.1. Propositions to provide XCP-TCP friendliness.
    2.2. A new architecture for a more robust XCP protocol.
    2.3. Propositions to provide interoperability between XCP and non-XCP routers.
3. **Discussion & Concluding Remarks.**

Conclusions

ERN protocols in large BDP networks with VBE:

- Maximize the link utilization.
- Fairly share resources between flows.
- Are less sensitive than E2E protocols to RTT values.

However, ERN protocols are not interoperable with current technologies. Therefore, I proposed:

- XCP-f which provides friendliness between E2E and ERN protocols.
- XCP-r which improves the robustness of ERN protocols.
- XCP-i which provides interoperability between ERN protocols and non-ERN equipments.

Perspectives

Implement our solutions in real equipments

Concerning XCP-f :

- To update the probability of dropping non-XCP packets in an elastic way
  - The constants to increases/decreases the probability for dropping non-XCP packets could strongly penalize TCP flows with large RTTs.
  - High speed version of TCP could not be correctly limited (too aggressive).
- Test XCP-f in more complex scenarios.

Concerning XCP-i :

- Non-XCP clouds with complex topologies.
- Detection of non-ERN layer 2 devices.

New challenges for large ERN adoption

- Security
  - How can we trust the information from routers?
- Propagate the ERN philosophy on others equipments (e.g. switches).
- Convince people about the benefits of ERN protocol.
  - Equipment manufacturers.
  - Network administrators.
  - Network operators.
  - Network services providers.

Questions?