

Improving the performance of long life flows with ERN protocols in heterogeneous large BDP networks

*Dino LOPEZ
dino.lopez@unice.fr*

Short Biography

- I received my PhD from the ENS (*École Normale Supérieure de Lyon*) of Lyon
 - Title
 - ✓ Propositions for a robust and interoperable eXplicit Control Protocol on Heterogeneous High Speed Networks
 - Advisors
 - ✓ LEFEVRE Laurent
 - ✓ PHAM Congduc
- September 2008 – August 2009: Postdoctorant at the DMI (*Département de Mathématique et Informatique*) Lab. ISAE, Toulouse
 - LOCHIN Emmanuel
- September 2009: Associate Professor at the Signal Team/ I3S Lab/ EPU - University of Nice.

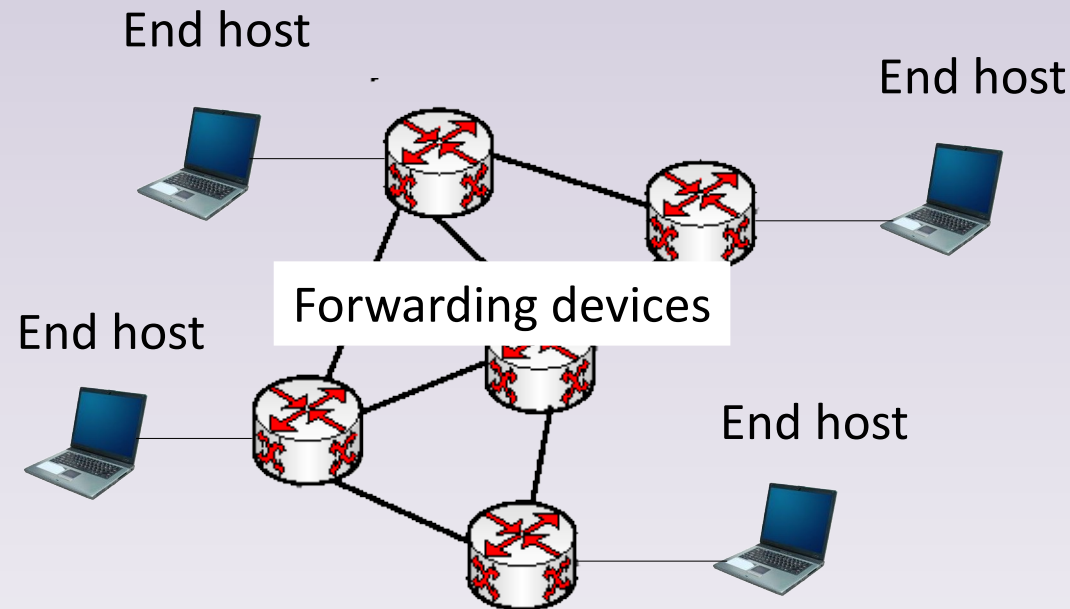
Research Interest

- Transport and congestion control protocols
 - Data flows
 - Multimedia flows
- Interaction between the wireless networks PHY layer and network/transport layer
- Sensor networks
- Internet-based Mobile Ad-hoc Networks (iMANET)
- Delay/Disruption Tolerant Networks (DTN)
- Active Queue Management (AQM)
- Autonomous networks

Introduction

Introduction to Networks

The components

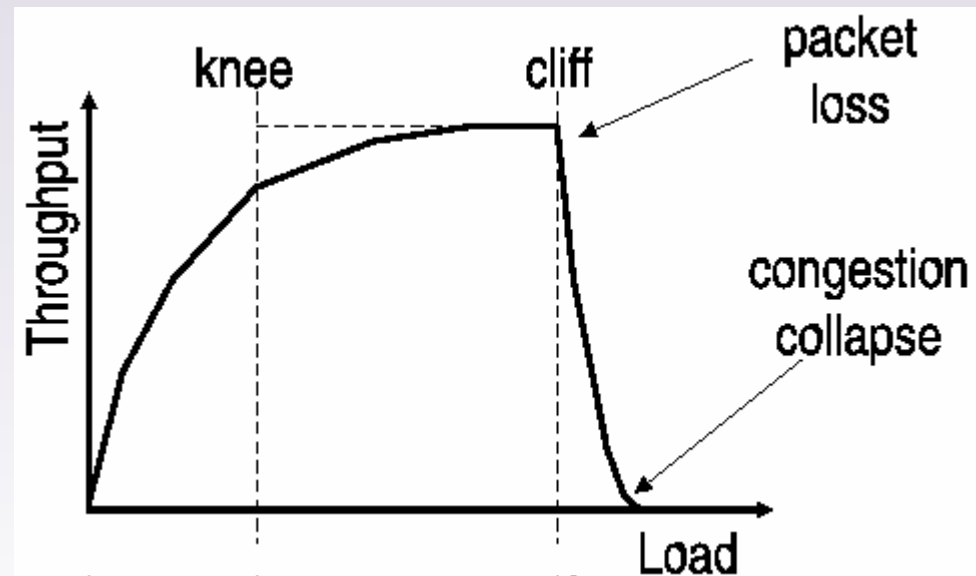


- Networks
 - Allow equipments (end hosts) exchange data packets (video, audio, data).
 - ✓ For reliability, Acknowledgments (ACKs) are sent after n data packets ($n \geq 1$)
 - Link equipments geographically far (forwarding devices).

Introduction to Networks

The Congestion Events

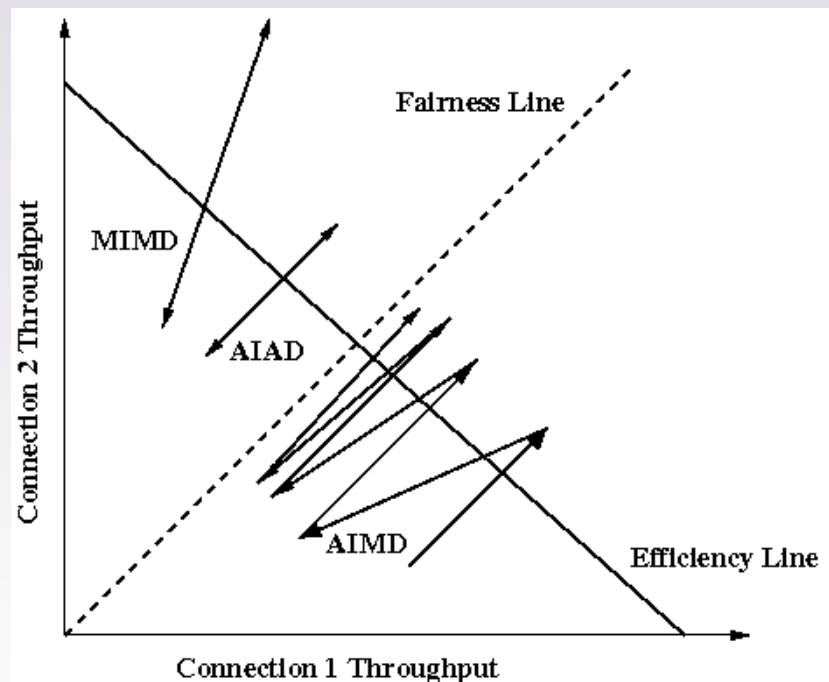
- Big success of networks = Overload of networks (congestion)
- Congestions may prevent the exchange of data
 - Congestion avoidances mechanisms



Introduction to Networks

Fairness

- End users must also fairly share the network resources
- Max-min fairness criteria : flows sharing the same bottleneck are entitled the same amount of bandwidth



Introduction to Networks

Congestion control strategies

- Different strategies to avoid congestion
 - Active Queue Management (AQM) mechanisms
 - Congestion control protocols.
 - ✓ End-to-End (E2E) protocols.
 - ✓ Explicit Rate Notification (ERN) protocols.

Outline

- 1 End-to-End protocols
 - 1.1) Standard TCP
 - 1.2) High speed variants of TCP
- 2 Active Queue Management & Explicit Congestion Notification
 - 2.1) DropTail
 - 2.2) RED
 - 2.3) ECN
- 3 Explicit Rate Notification protocols
 - 3.1) XCP
 - 3.2) Limits of ERN protocols
 - 3.3) Towards an interoperable ERN protocol
- 4 Conclusions and Perspectives

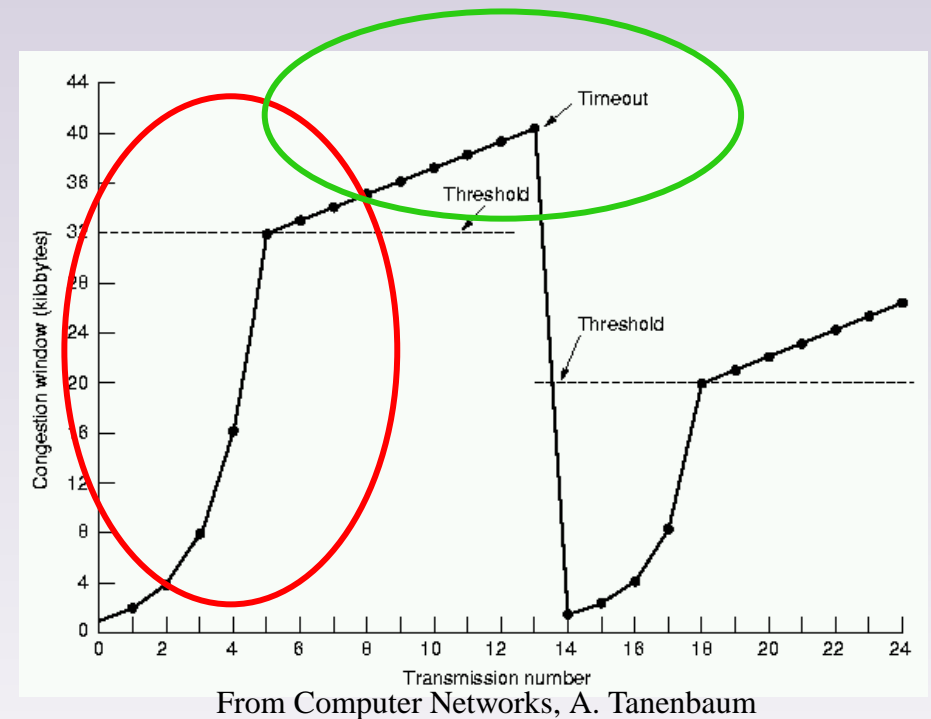
End-to-End protocols

End-to-End protocols

- End-to-End 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

The TCP Congestion Control Protocol

- Defined in [RFC1122]
- Slow-Start
 - $cwnd = cwnd + 1$; (for every ACK)
- Congestion Avoidance
 - $cwnd = cwnd + 1 / cwnd$; (for every ACK)
- In case of losses
 - Fast Retransmit : send the missing packet
 - Fast Recovery :
 - $cwnd = 1 MSS$; (TCP Tahoe : SS)
 - $cwnd = cwnd - (1/2)*cwnd$; (TCP Reno and New Reno : CA)

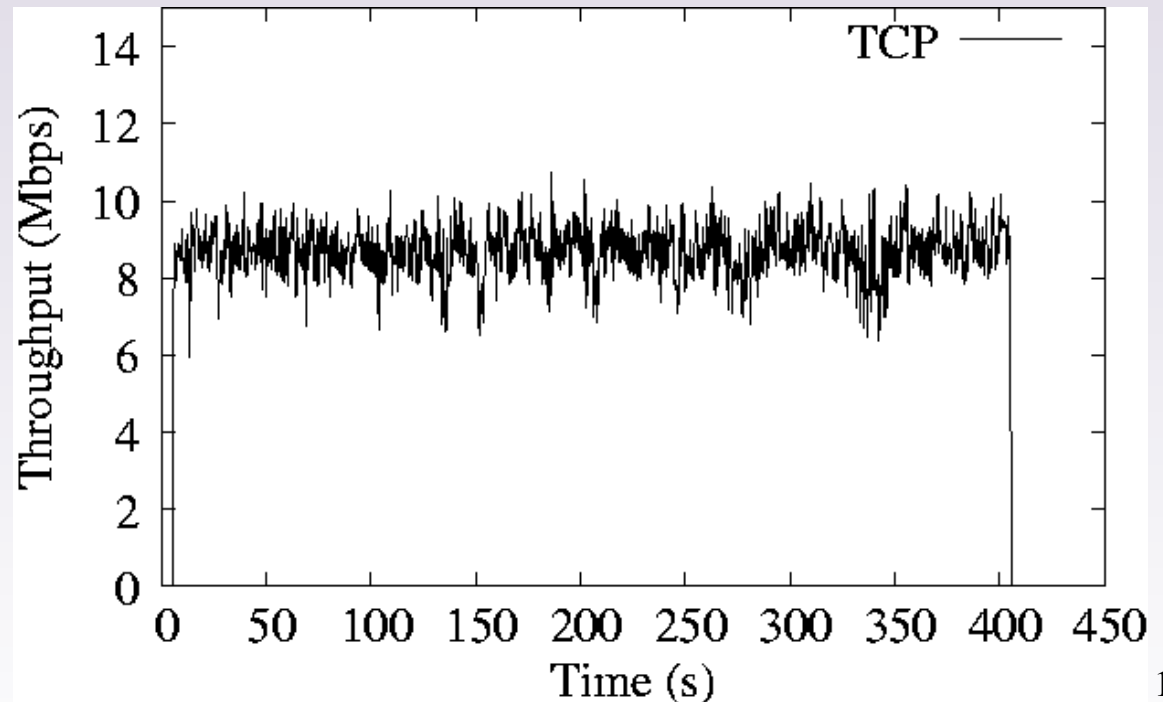


MSS (Maximum Segment Size)

cwnd (Congestion Window) : Amount of data that can be sent before waiting for ACKs

TCP in large bandwidth-delay product (BDP) networks

- TCP stack protocol in a Linux-based system
- Kernel 2.6.17 – Buffer (96KB)
- Bottleneck = 1 Gbps
- Delay emulation using AIST-GtrcNet-1
 - 100ms of base RTT



TCP in large BDP networks

- High capacity is not warranty of high throughput
- Standard TCP is not adapted for large BDP networks
- Long time for transferring some GBs

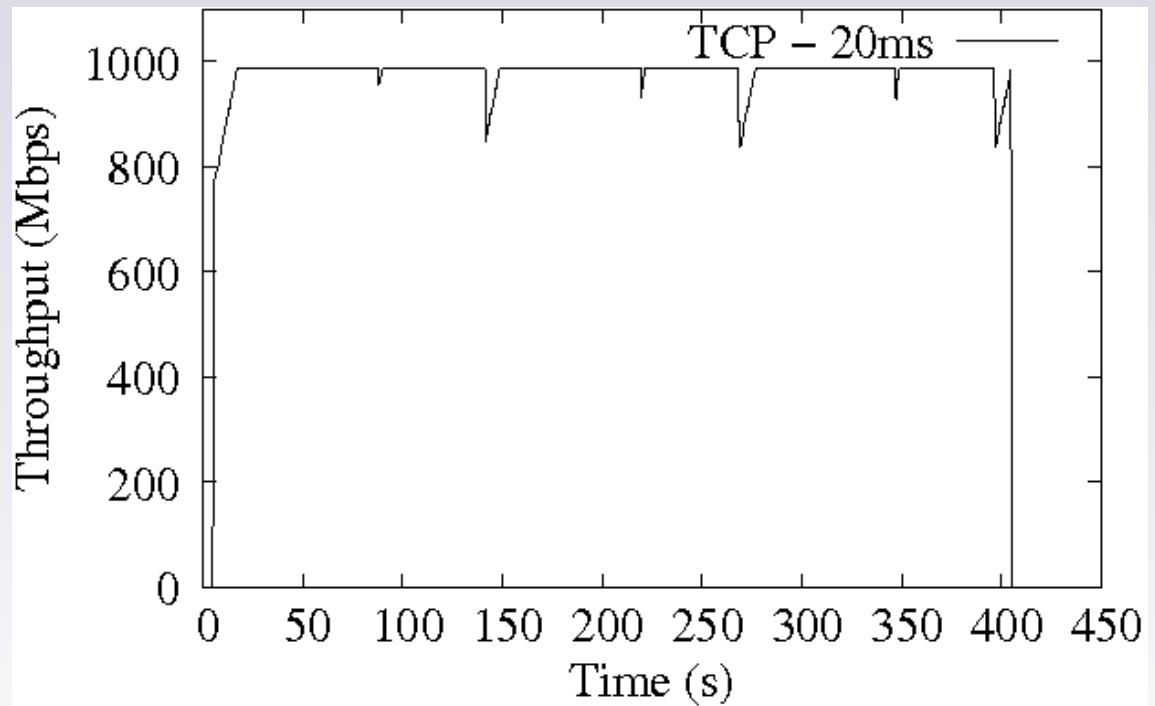
Adapting standard TCP to large BDP networks

- Parallel TCP flows
 - But how many?
- Tuning TCP
 - TCP parameters are not adapted for large BDP networks
 - ✓ Increase the buffers capacity at the senders and receivers
 - ✓ Increase the size of TCP packets
 - ◆ Bigger than 1500B
 - ✓ “Remember” the last *ssthresh* value
 - ✓ Initialize *cwnd* with a value bigger than 1MSS
 - ◆ QuickStart

Tuning TCP for a 1Gbps bottleneck and 20ms of base RTT

Test 1

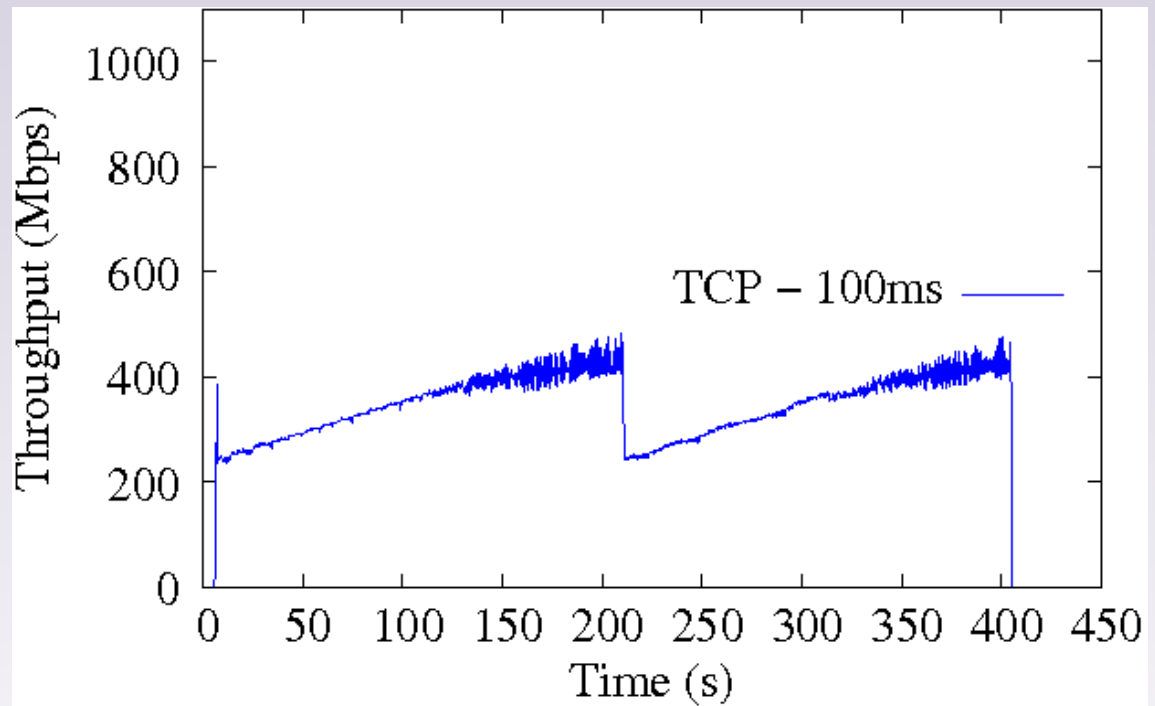
- $ssthresh, buffers \approx bandwidth * delay$
- Bottleneck = 1Gbps
- Base $RTT = 20ms$



Tuning TCP for a 1Gbps bottleneck and 20ms of base RTT

Test 2

- Same sender but different receiver...
 - Base RTT = 100ms



- Tuning TCP protocol stack for one scenario does not solve the problem of large BDP networks

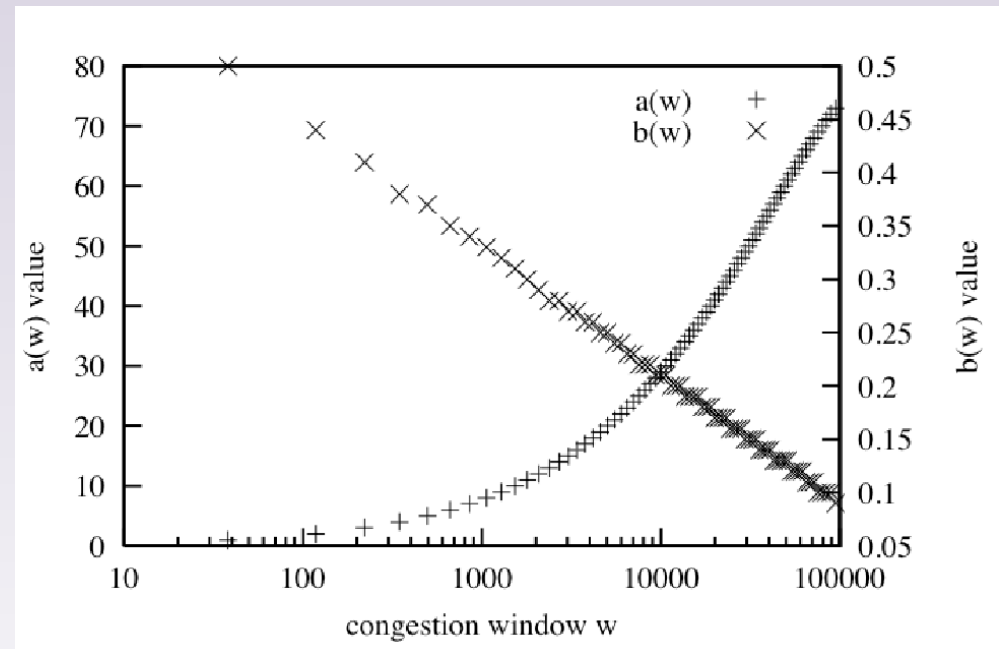
High Speed Variants of TCP

High Speed protocols

- Purely aggressive TCP variants
 - High Speed TCP – Sally Floyd [RFC3649]
 - Scalable TCP – Tom Kelly [Computer Communication Review 2003]
 - Binary Increase (BIC) / CUBIC – Lisong Xu & Injong Rhee [INFOCOM 2004, ACM SIGOPS Operating System Review 2008]
 - ✓ CUBIC is currently used in most Linux-based systems
- TCP-based protocols with available bandwidth estimation
 - TCP Westwood+ – Saverio Mascolo [PFLDNet 2005]
- Delay-based TCP protocols with Early Congestion Detection
 - FAST TCP – Steven Low [IEEE/ACM Transactions on Networking 2007]
- Hybrids TCP
 - Compound TCP – Kun Tan (Microsoft Research Asia) [PFLDNet 2006]
 - ✓ Currently implemented in Windows XP, Windows Vista and Windows Server 2008

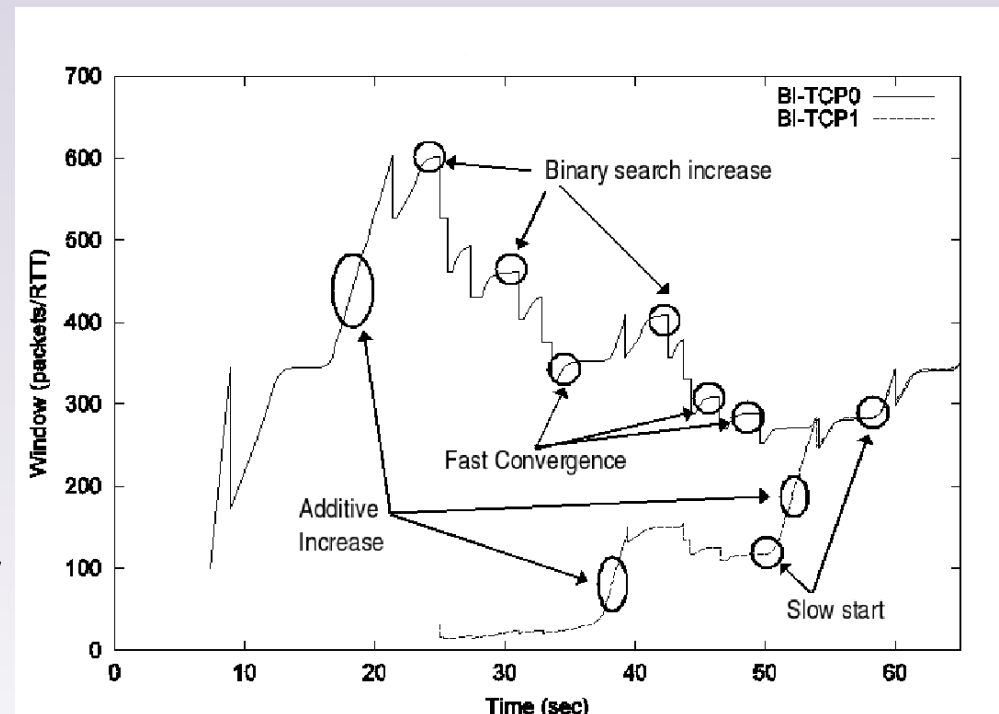
High Speed TCP

- Slow-Start
 - Introduction of “Limited slow-start”
- Congestion Avoidance
 - $cwnd = cwnd + a(cwnd) / cwnd$
- In case of Losses
 - $cwnd = cwnd - b(cwnd) * cwnd$



BIC

- Congestion control is seen like a binary search problem
 - After losses $cwnd$ rate is reduced $\rightarrow min_win$
 - $cwnd$ before losses $\rightarrow prev_max$
 - Next RTT : $cwnd = cwnd + (prev_max - min_win) / 2$



Limits of E2E protocols



- Networks are black boxes for E2E protocols.
- For this reason, E2E protocols :
 - are unable to know the real state of the resources.
 - lead to congestion periodically.
 - are affected by the propagation delay.
 - suffer from unfairness and slow convergence

Congestion control protocols in forwarding devices

- Active Queue Management (AQM) mechanisms: Routers randomly drop packets when congestion is “imminent”. Ex. Random Early Discard (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.

Active Queue Management and Explicit Congestion Notifications

AQM protocols

- Standard TCP : Asymmetric RTTs lead to unfairness
- High speed variants of TCP can affect the performance of standard TCP flows
- Aggressiveness of high speed TCP variants may lead to heavy congestion

Active forwarding devices at the bottleneck could improve the fairness and avoid heavy congestions

Passives and Actives routers

- Passives Routers

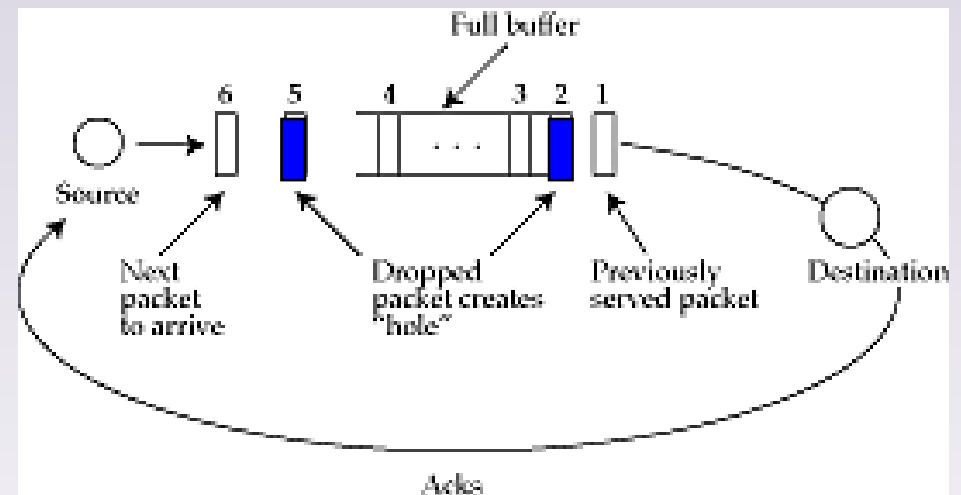
- DropTail

- ✓ Used by default in the routers
 - ✓ When the buffer is full, next incoming packet is discarded

- Actives routers

- Random drop

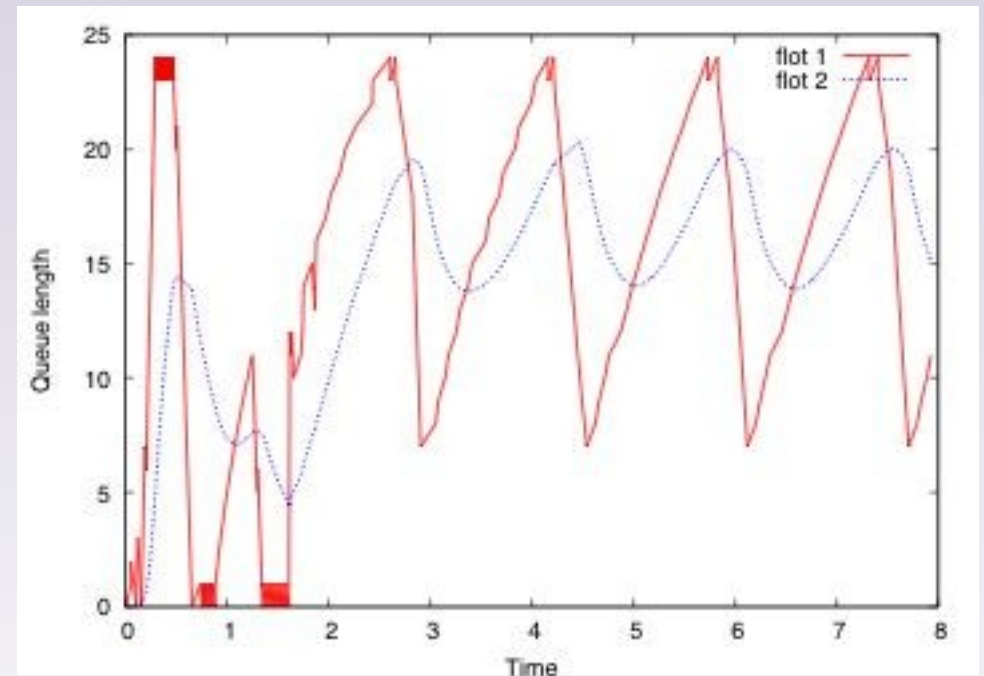
- ✓ Packets are randomly dropped
 - ✓ Punish the more aggressive flows



DropTail

From the point of view of routers

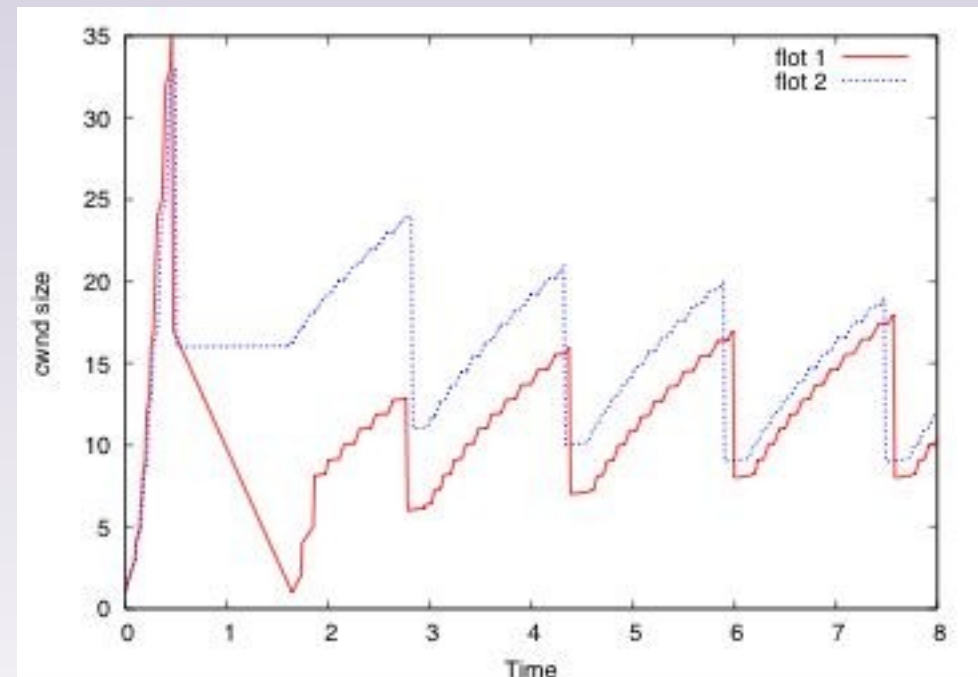
- Important oscillation of the buffer occupancy
- Performance of short-life flows is affected
- Senders receive congestion signaling after congestion occurs



DropTail

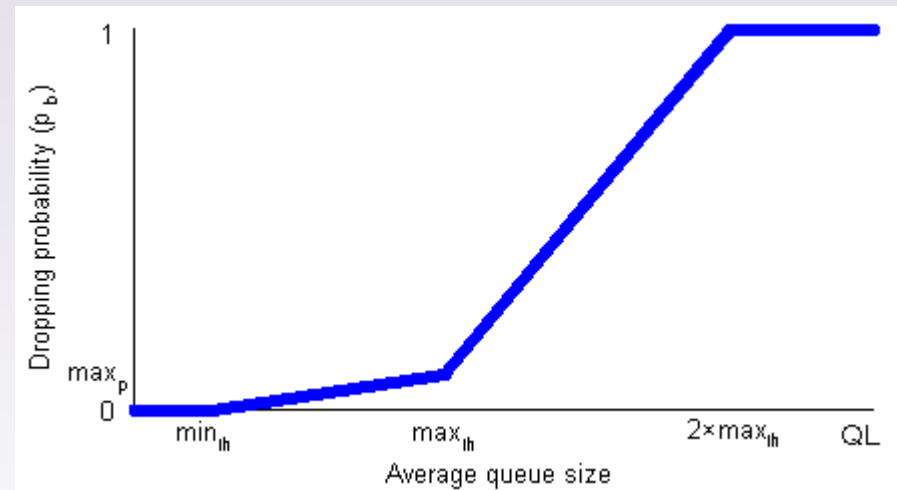
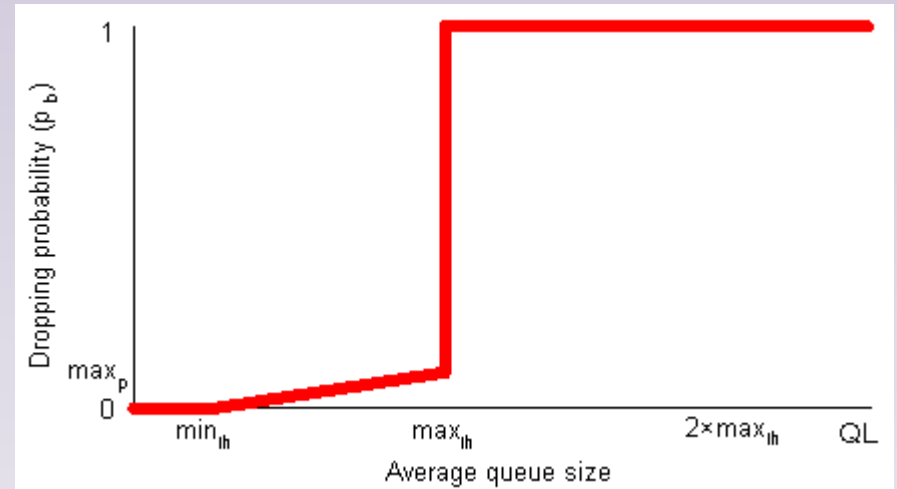
From the point of view of senders

- Possibility of synchronization between senders during congestion events
 - Fast Retransmit / Fast Recovery may lead to an under utilization of network resources
- Burst of losses are frequent
 - Probability of Timeouts increase



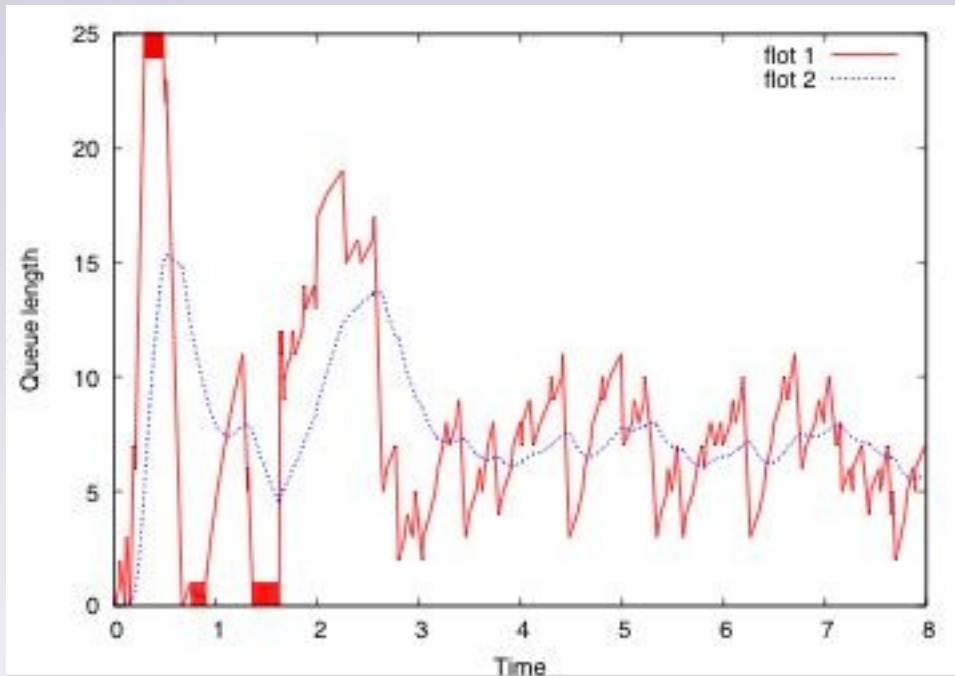
Random Early Discard (RED)

- RED computes the average buffer size (*avq*)
 - When $th_min < avg < th_max$, packets are discarded with a probability p
 - If $th_min > avg$ then do nothing
 - If $th_max < avg$
 - ✓ packets are discarded ($p=1$)
 - ✓ Gentle RED : Packets are discarded with a probability p that linearly increases between max_p and 1

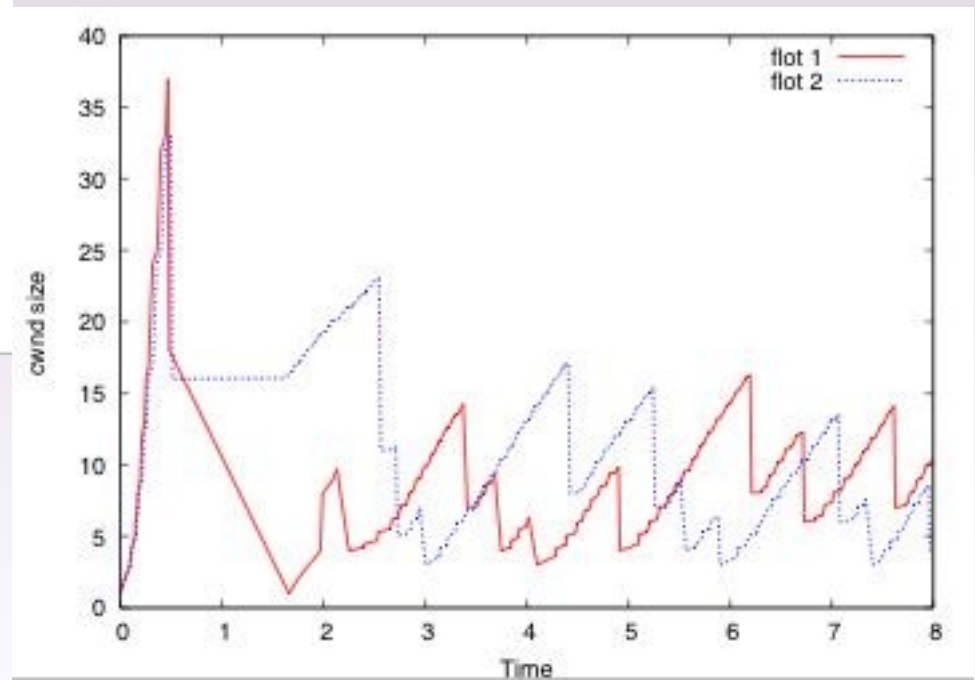


RED

The routers' point of view



The senders' point of view



Advantage and Disadvantage of RED

Advantages

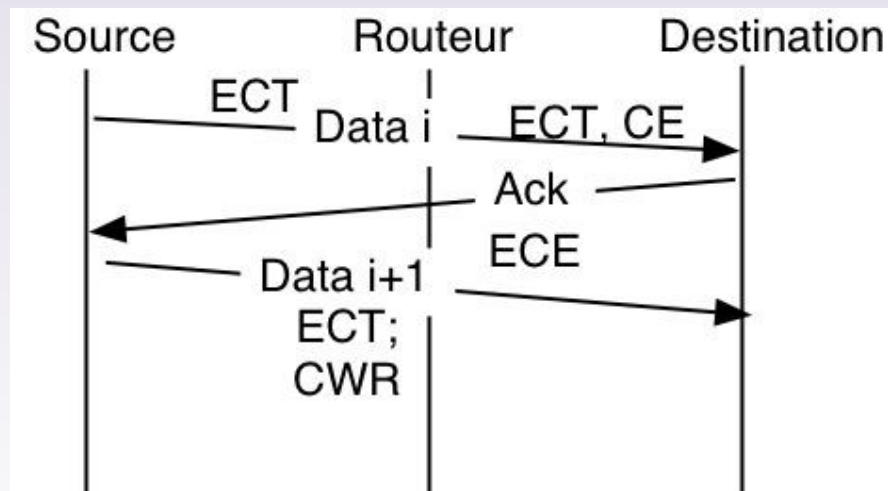
- Decrease the probability of burst of losses
 - Fast Retransmit / Fast Recovery are enough to recover from packet losses
- Decrease of the jitter

Disadvantages

- Difficult to configure
 - Performance of RED depends on the RTT of flows, number of flows, etc.
- High frequency of incoming flows produce congestion before the execution of RED

Explicit Congestion Notification (ECN)

- Instead of dropping packets, routers explicitly indicates when congestion events are imminent
- Some benefits of ECN:
 - Congestion signaling arrives faster to the sender
 - Adaptability of the sending rate without retransmissions

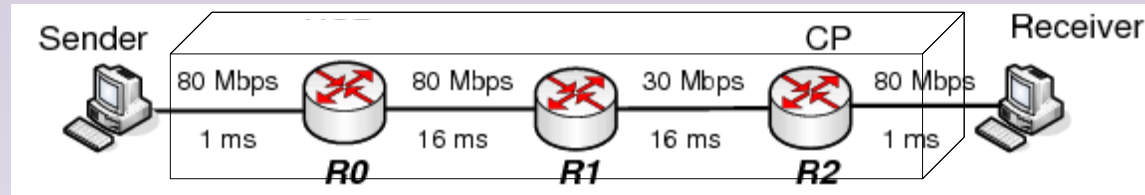


Explicit Congestion Notification (ECN)

- Performance of ECN-capable senders may be lower than the non-ECN-capable senders ones
 - Some nodes may “erase” the congestion event notifications
 - Some senders may ignore the congestion event notifications

Explicit Rate Notification (ERN) protocols

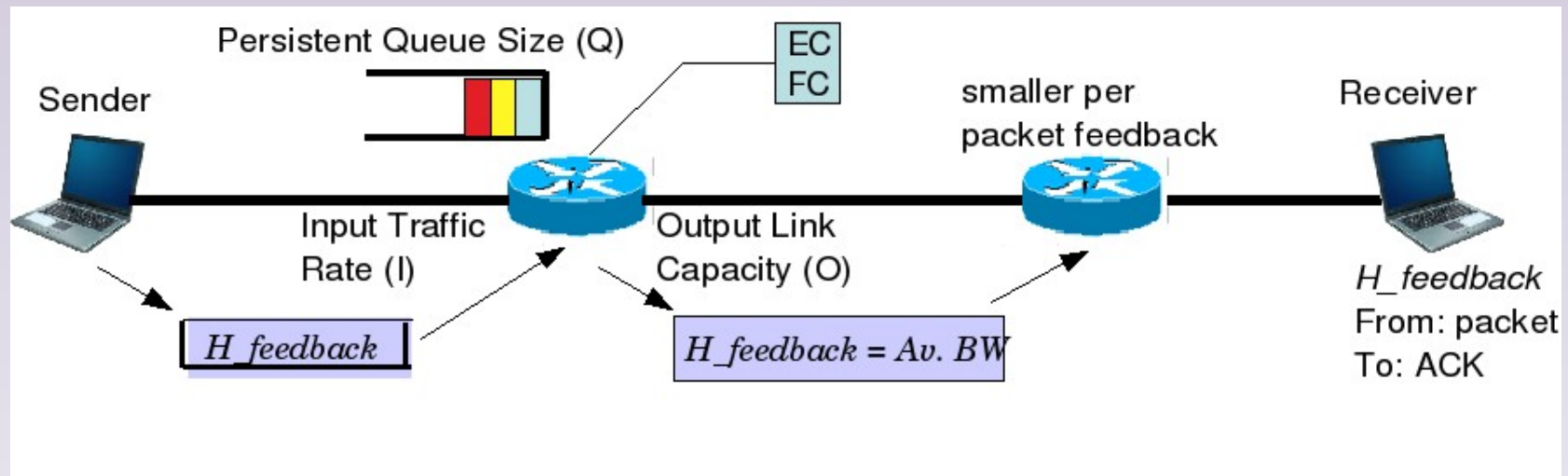
ERN protocols



- Routers (forwarding devices) 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 due to congestion rarely happen in fully ERN wire-based networks.
- Some ERN protocols: XCP [D. Katabi – ACM SIGCOMM 2002], JetMax [D. Leonard – INFOCOM 2006], Quickstart [S. Floyd RFC4782], etc.

eXplicit Control Protocol (XCP)

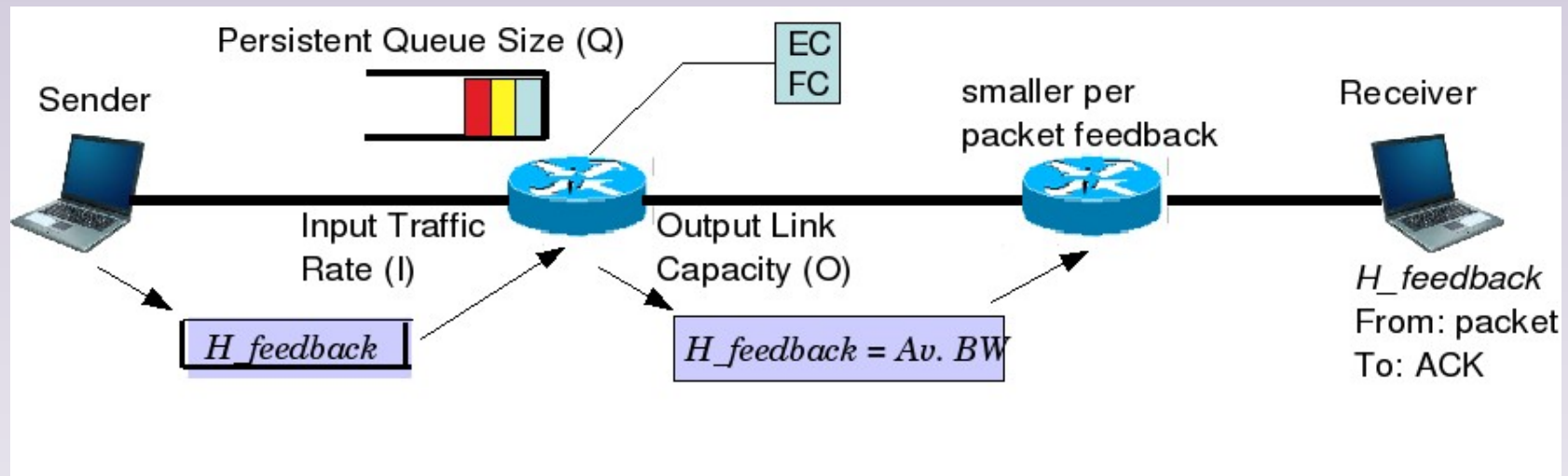
The XCP routers



- XCP routers compute a feedback per packet without keeping per flow states ($H_feedback$)
 - Efficiency Controller : Computes a general feedback (ϕ)
 - ✓ This is the available bandwidth
 - Fairness Controller : Translates ϕ in a feedback per packet
 - ✓ Available bandwidth is fairly shared between XCP flows
 - ✓ Only bandwidth held by others XCP flows may be reassigned

eXplicit Control Protocol (XCP)

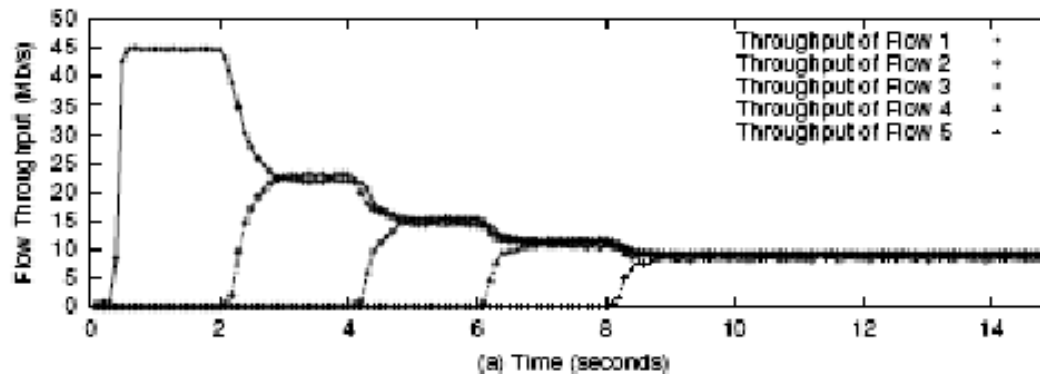
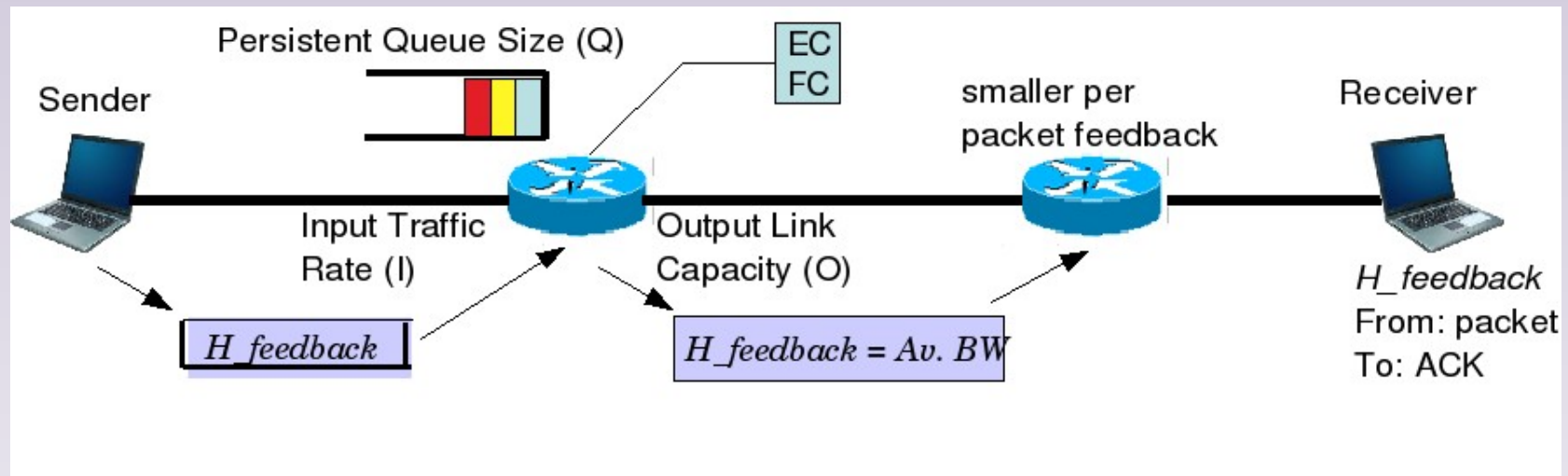
The end hosts



- The receiver copies the $H_feedback$ from the data packet to the ACKs
- The sender only adds the $H_feedback$ value to the $cwnd$ size
 - $cwnd = cwnd + H_feedback$

eXplicit Control Protocol (XCP)

Performance of XCP (ns-2 network simulator)



From Congestion Control for High Bandwidth-Delay Product Networks. Dina Katabi, Mark Handley, and Charlie Rohrs. In the proceedings on ACM Sigcomm 2002.

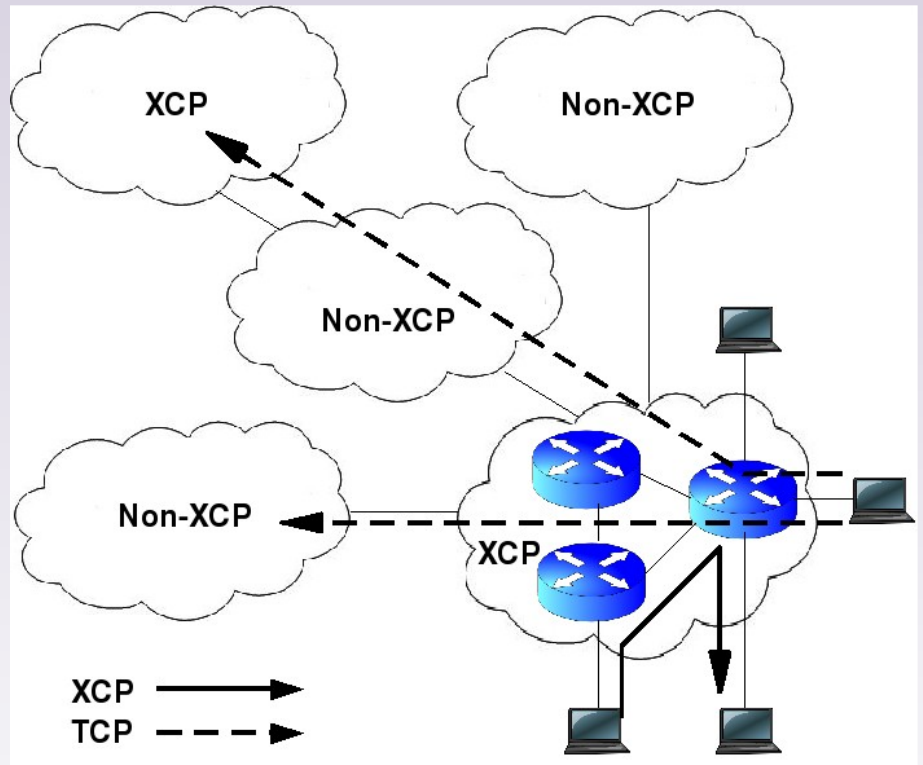
Limits of ERN protocols

- ERN protocols work well in fully ERN networks
 - Not inter-operable with current E2E protocols
 - Not inter-operable with current IP routers
 - Sensitivity to feedback loop

Towards an interoperable ERN protocol

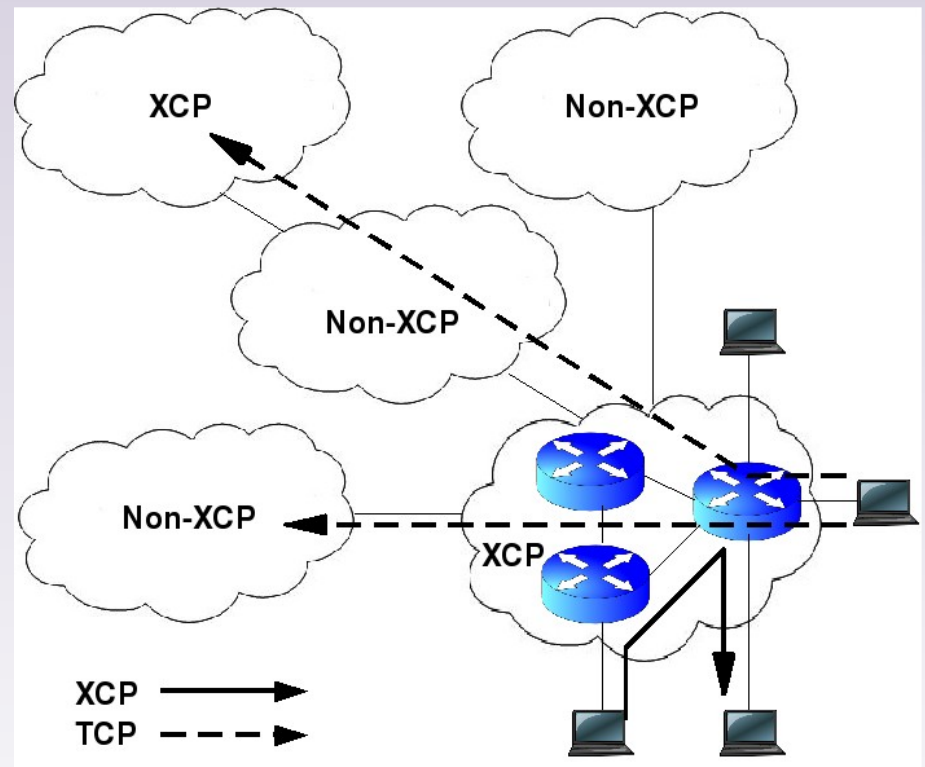
Context

- Our solutions allow an incremental deployment of ERN protocols in :
 - Wire-based heterogeneous large BDP networks
 - Networks where long-life flows are frequent



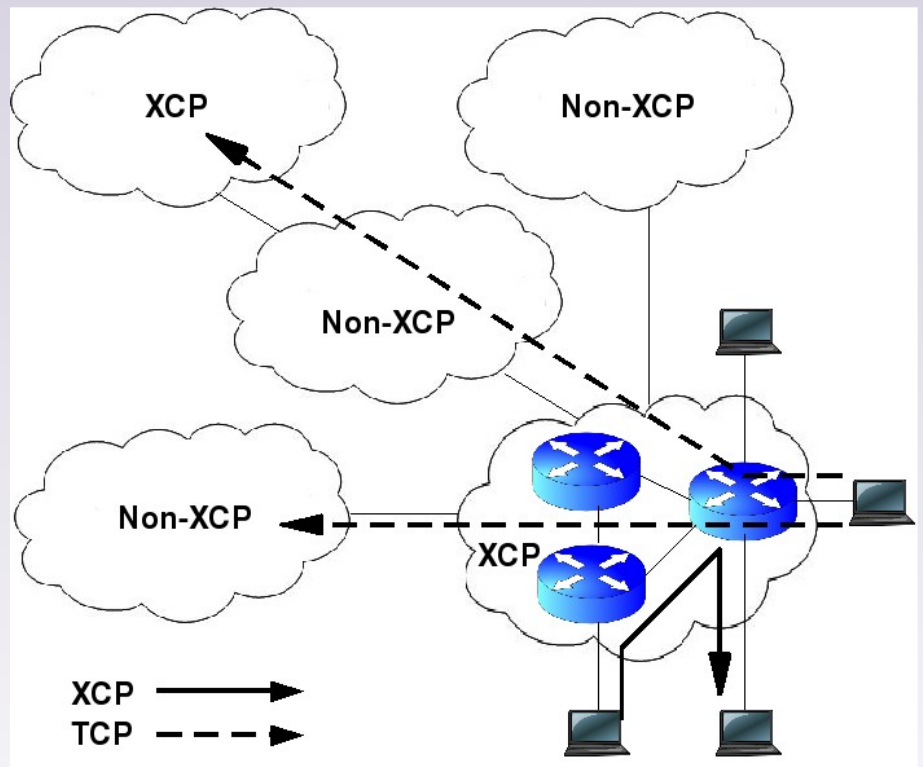
The Proposed Solutions

- XCP-f : Enables XCP-TCP friendliness
 - Estimates the numbers XCP and non-XCP actives flows
 - ✓ The zombie estimator
 - Discards non-XCP packets to limit the throughput of non-XCP flows



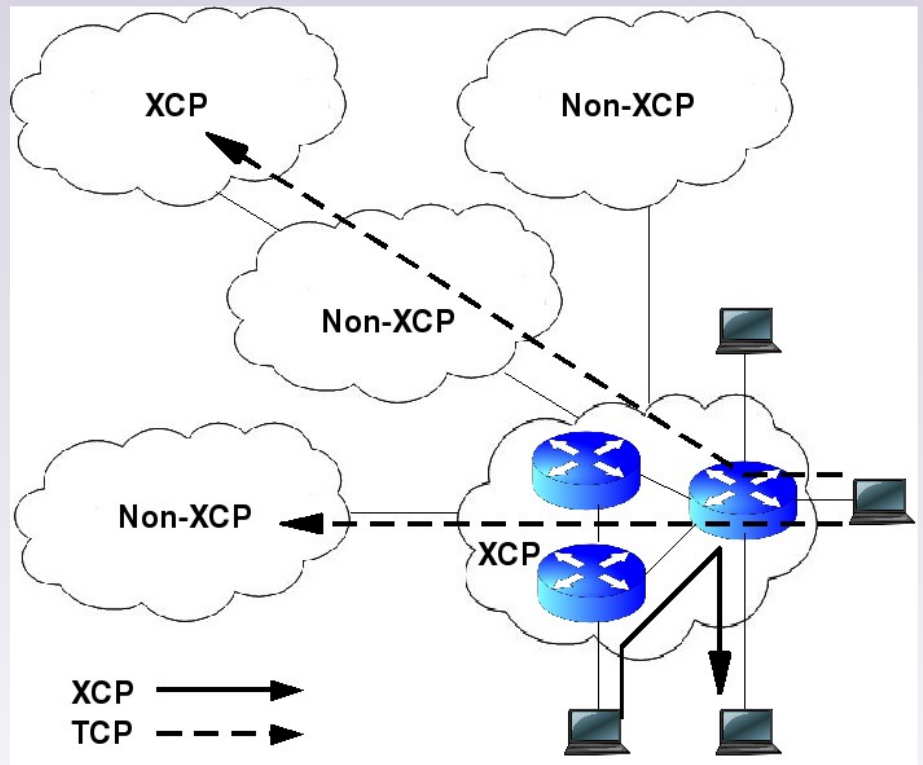
The Proposed Solutions

- XCP-r : Robustness of XCP face to feedback losses
 - Moving the protocol stack from the sender to the receiver
 - ✓ Synchronization protocol



The Proposed Solutions

- XCP-i : The first step towards the interoperability of XCP and non-XCP routers
 - Detection of non-XCP routers
 - Creation of virtual XCP routers



Conclusion and open issues

Conclusions

- In large BDP networks, E2E protocols are unable to correctly grab and fairly share the available resources
 - End hosts and forwarding devices must “cooperate”
 - Cooperation between end hosts and forwarding devices should go far more than simply binary signals
- ERN protocols in large BDP networks :
 - Maximize the link utilization
 - Fairly share resources between flows
 - Are less sensible than E2E protocols to RTT values

Conclusions

- However, ERN protocols are not interoperable with current networks technologies
- In Data Grid Environments, ERN protocols (e.g. XCP) can be deployed :
 - XCP-f provides friendliness between E2E and ERN protocols
 - XCP-r improves the robustness of ERN protocols
 - XCP-i provides interoperability between ERN protocols and non-ERN equipments

New challenges for large ERN adoption

- Security
 - Why must senders trust routers?
 - Why must routers trust senders?
- Propagate the ERN philosophy to layer 2 forwarding devices (e.g. switches)
- Enable ERN protocols in Internet

Thank You