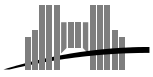
TCP variants and transfer time predictability in very high speed networks

Romaric Guillier, Sébastien Soudan, Pascale Vicat-Blanc Primet

LIP, École Normale Supérieure de Lyon, UMR 5668, France

HSN 2007 – May 11th, 2007





Outline

TCP variants

TCP Congestion window evolution (AIMD)

- ACK : $cwnd \leftarrow cwnd + \frac{\alpha}{cwnd}$
- Drop : $cwnd \leftarrow cwnd - \beta * cwnd$
- Reno[Jacobson88] : $\alpha = 1; \beta = \frac{1}{2}$

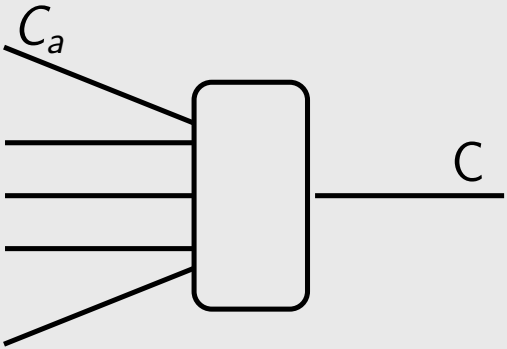
New variants

- HighSpeed TCP[Floyd02] : $\alpha = inc(cwnd); \beta = dec(cwnd)$
- Scalable TCP[Kelly02] : $\alpha = 0.01 * cwnd; \beta = \frac{1}{8}$
- Hamilton TCP[Leith05] : $\alpha = f(last_{loss}); \beta = 1 - \frac{RTT_{min}}{RTT_{max}}$
- Bic[Rhee04] : $\alpha = 1$ if $cwnd < cwnd_{min}$, $bin_{search}(S_{min}, cwnd, S_{max})$ otherwise; $\beta = \frac{1}{8}$
- Cubic[Rhee05] : $\alpha = cub(cwnd, history); \beta = \frac{1}{5}$

Goals: Compatibility with legacy TCP, performance for large BDP networks, etc..

Objectives

Low aggregation level, symetric links
(ex: Grid networks, FTTH)



Aggregation level: $K = \frac{C}{C_a} \simeq 1$

Transfer time predictability:

- Impact of flow inter-arrival?
- Impact of congestion level?
- Impact of reverse traffic level?

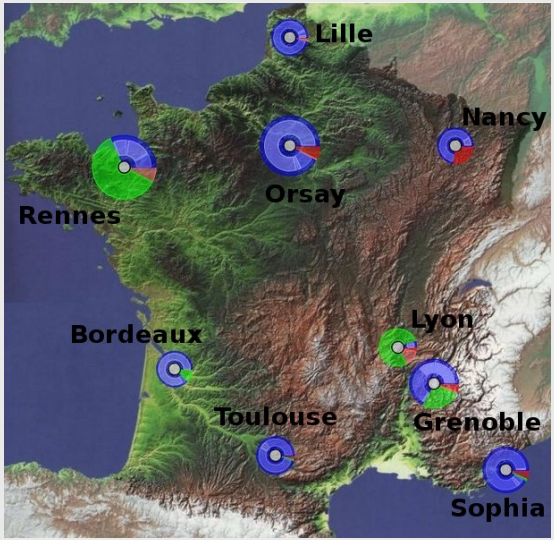
Congestion level: $\frac{\sum C_a}{C}$

Metrics

- Mean completion time : $\overline{T} = \frac{1}{N_{forward}} \sum_{i=1}^{N_{forward}} T_i$
- Max completion time : $T_{max} = max(T_i)$
- Min completion time : $T_{min} = min(T_i)$
- Std deviation of completion time : $\sigma = \sqrt{\frac{1}{N_{forward}} \sum_{n=1}^{N_{forward}} (T_i - \overline{T})^2}$

Outline

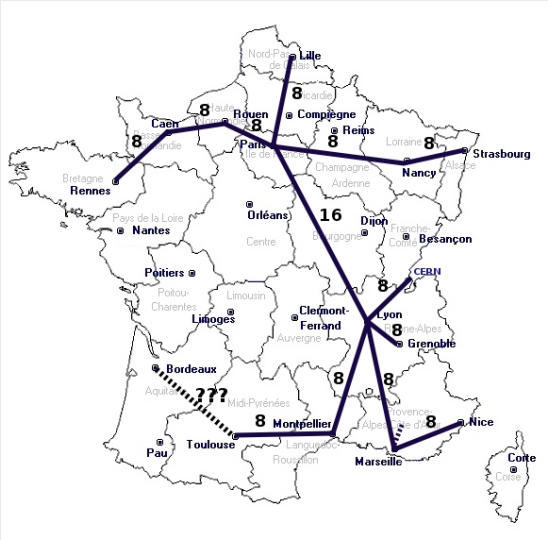
Grid5000: Description

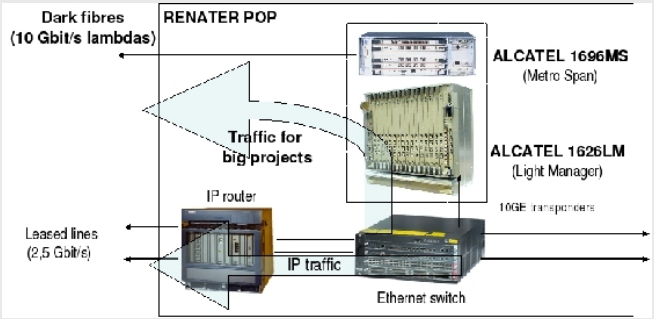


Site	CPU available	CPU scheduled
Bordeaux	198	500
Grenoble	270	500
Lille	146	500
Lyon	252	500
Nancy	94	500
Orsay	684	1000
Rennes	522	522
Sophia	434	500
Toulouse	116	500
Total	2542	5022

- 9 sites in France, 17 laboratories involved
- 5000 CPUs (currently 2500)
- Private 10Gbps Ethernet over DWDM network
- Experimental testbed for Networking to Application layers.

Grid5000: Description



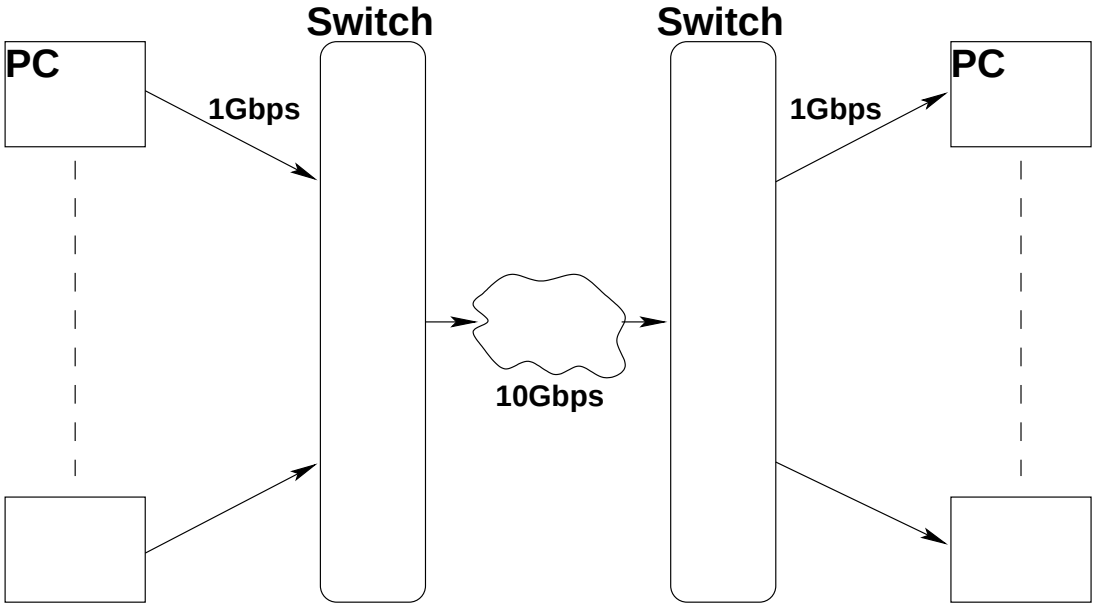


- 9 sites in France, 17 laboratories involved
- 5000 CPUs (currently 2500)
- Private 10Gbps Ethernet over DWDM network
- Experimental testbed for Networking to Application layers.

Grid5000: Special Features

- A high security for Grid'5000 and the Internet, despite the deep reconfiguration feature
  - ↪ Grid'5000 is confined: communications between sites are isolated from the Internet and Vice versa (level2 MPLS, Dedicated lambda).
- A software infrastructure allowing users to access Grid'5000 from any Grid'5000 site and have simple view of the system
  - ↪ A user has a single account on Grid'5000, Grid'5000 is seen as a cluster of clusters, 9 (1 per site) unsynchronized home directories
- A reservation/scheduling tools allowing users to select nodes and schedule experiments
  - ↪  Reservation engine + batch scheduler (1 per site) + OAR Grid (a co-reservation scheduling system)
- A user toolkit to reconfigure the nodes
  - ↪  Software image deployment and node reconfiguration tool

Topology



Classical dumbbell: $N_{forward}$ and $N_{reverse}$ pairs of 1 Gbps nodes

Network cloud: Grid5000 backbone, 10 or 1 Gbps link

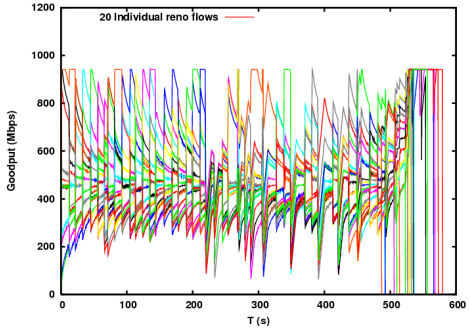
Bottleneck: output port of the L2 switch

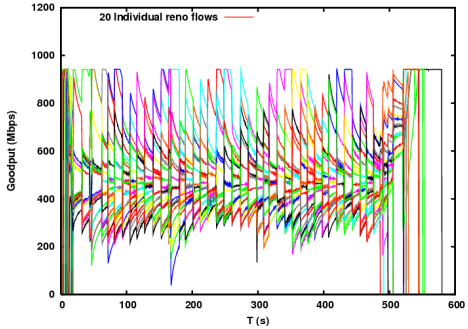
Outline

Inter-arrival < 0 s

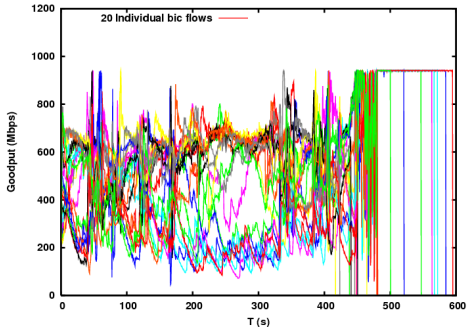
Inter-arrival > 0 s

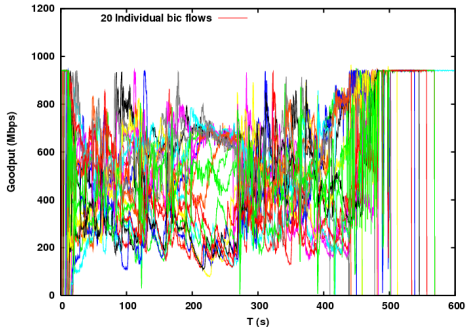
Reno



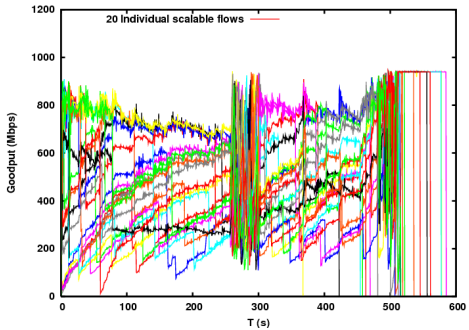


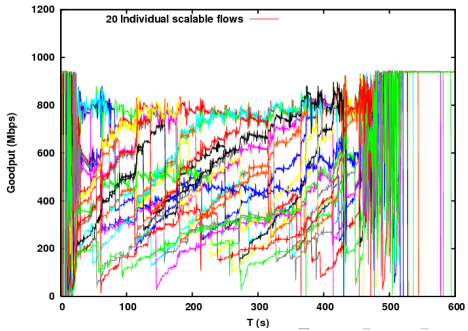
Bic





Scalable





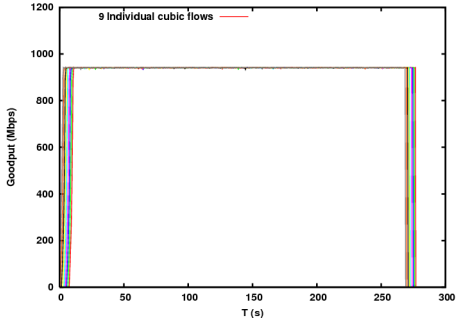
Cong. lvl: $\frac{\sum C_a}{C}$

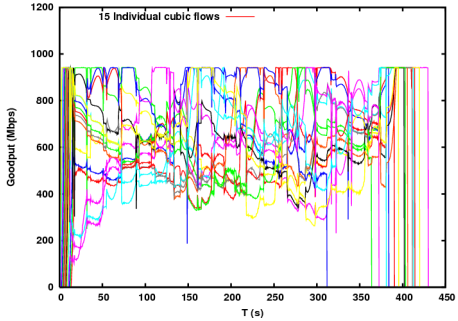
90 %: 280 s/272 s

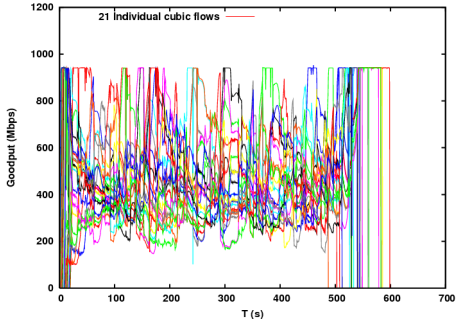
150 %: 395 s/398 s

210 %: 545 s/535 s

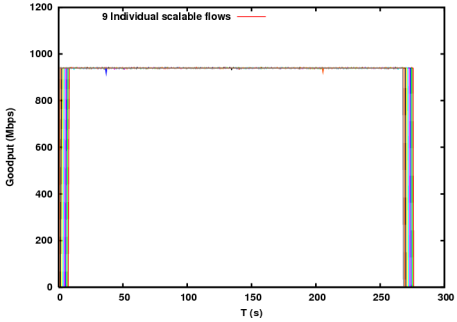
Cubic

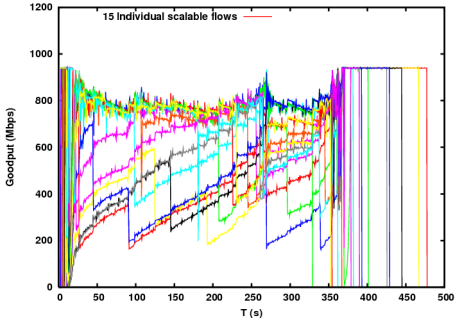


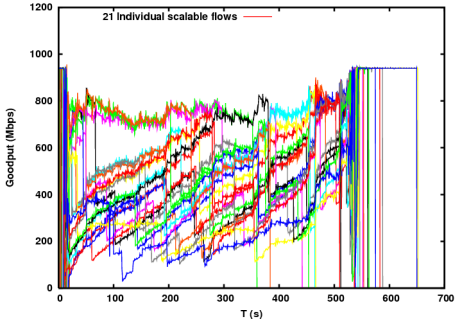




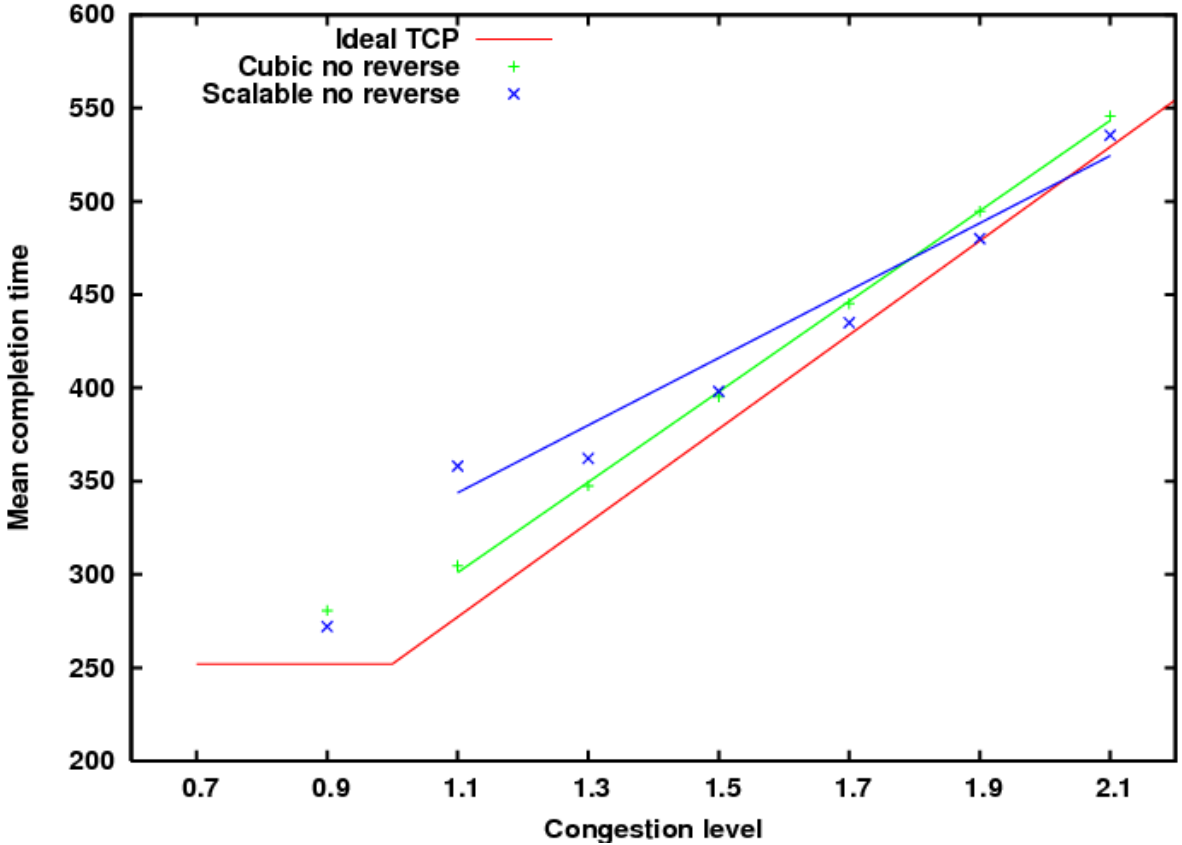
Scalable







Mean completion time of Cubic and Scalable

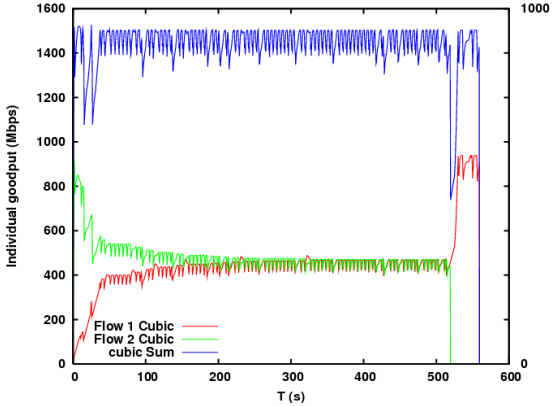


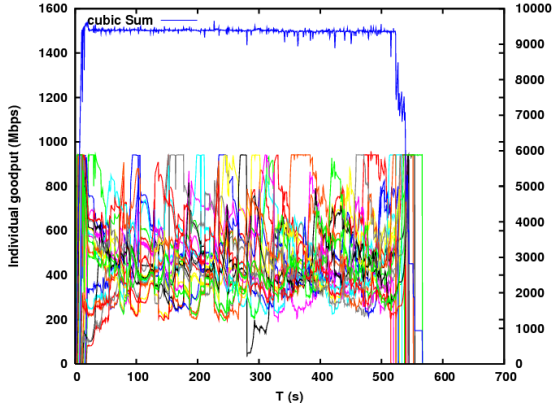
The multiplexing factor (200 % congestion level)

2x Cubic 1 Gbps flows, 1Gbps bottleneck

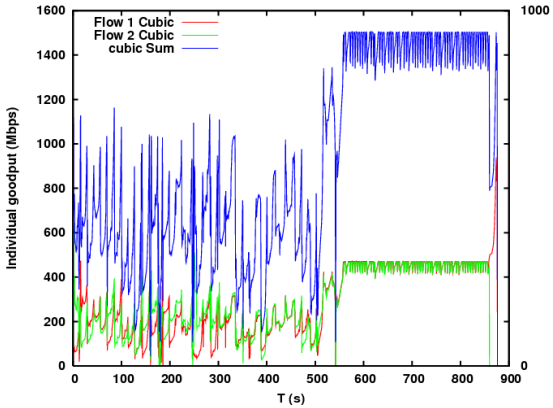
20x Cubic 1 Gbps flows, 10Gbps bottleneck

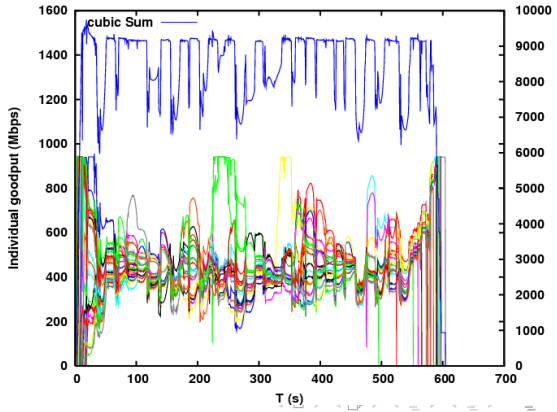
No reverse: 562 s/567 s





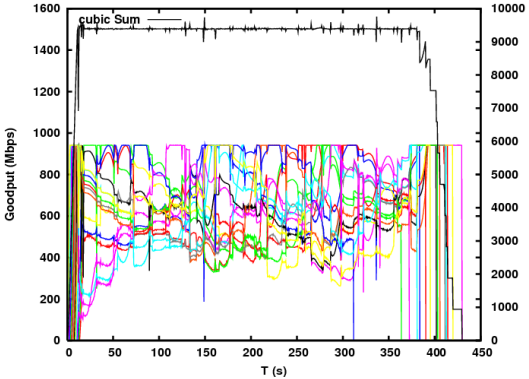
200 % reverse: 875 s/605 s



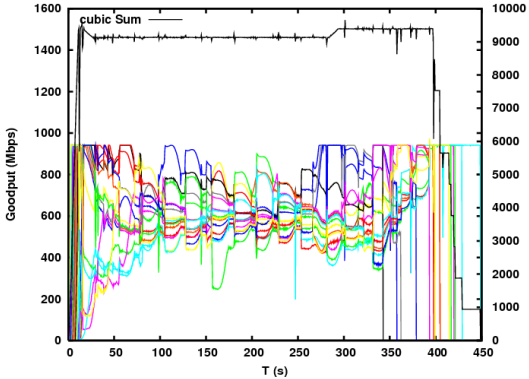


Influence of reverse traffic on Cubic (150 % cong. lvl)

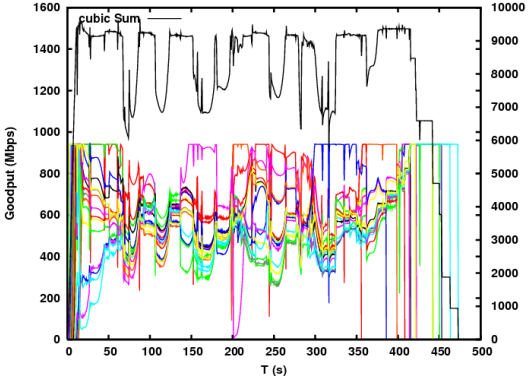
No reverse (395 s)



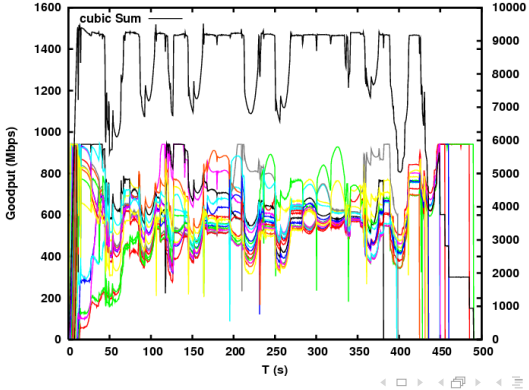
90 % reverse (400 s)

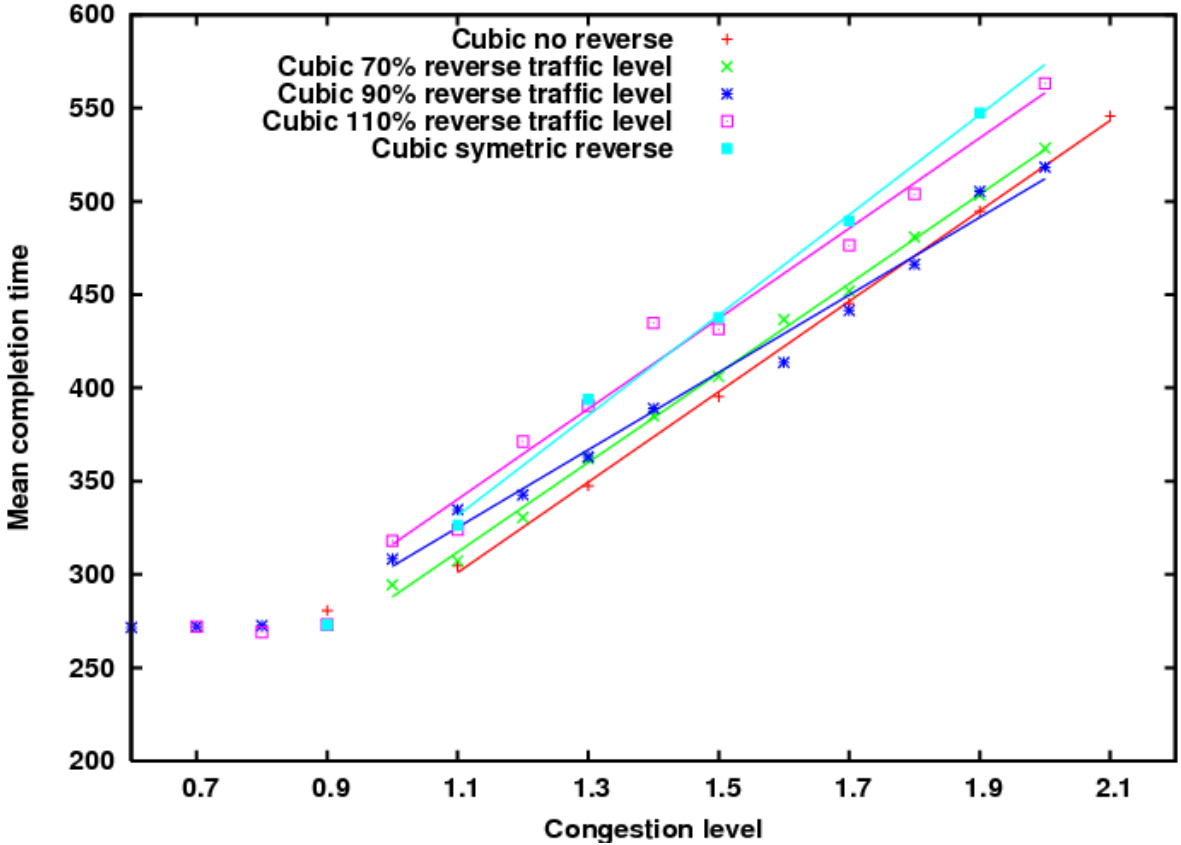


110 % reverse (432 s)



150 % reverse (438 s)




Cubic no reverse
Cubic 70% reverse traffic level
Cubic 90% reverse traffic level
Cubic 110% reverse traffic level
Cubic symetric reverse
Mean completion time
Congestion level
600
550
500
450
400
350
300
250
200
0.7
0.9
1.1
1.3
1.5
1.7
1.9
2.1

Conclusion

- Transfers should not be started simultaneously.
- Congestion level has a linear impact on the mean completion time.
- Multiplexing helps reducing completion time for a given congestion level (30 % in our example).
- Reverse traffic has a linear impact on the mean completion time when it is congesting the reverse path. The impact is about 1 % when there is no congestion.
- Most TCP variants behave similarly for the range of RTT studied, except Scalable which is unstable and displays a huge variability.
- Further aspects must be studied: heterogeneous workload, reverse traffic modelling, predictability service

Questions?

Thanks for your attention...