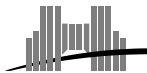


Méthodologies et outils pour l'évaluation des protocoles de transport dans les réseaux très haut débit

Romarc GUILLIER

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

Soutenance de thèse – 29 Octobre 2009

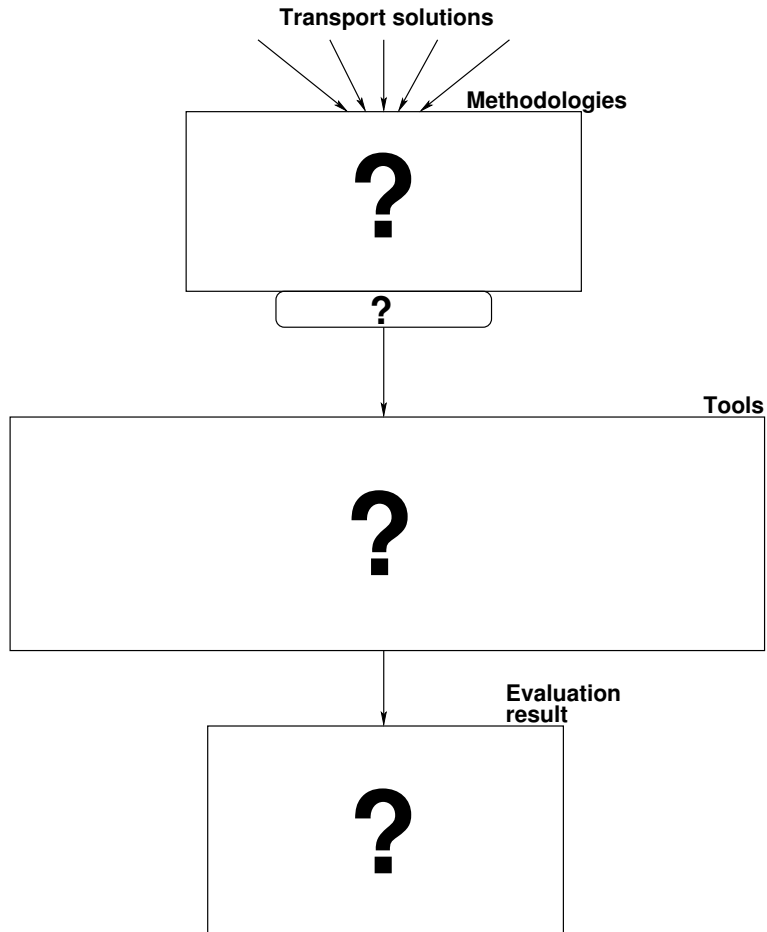


Contexte

Futur de l'Internet

- Omniprésence d'Internet et du protocole TCP.
- Evolution des technologies réseau: sans fil, fibre à la maison
- Evolution des usages: multimédia, cloud
- TCP mal adapté à ces évolutions.
- Emergence d'une grande variété de solutions.
- Quel va être l'impact de ces évolutions sur Internet ?

Problématique



Contexte scientifique

- International Workshop serie on Protocols for Future, Large-Scale & Diverse Network Transports (PFLDNeT) since 2003
- OGF DT-RG (research group): A Survey of Transport Protocols other than “Standard” TCP - informational document : GFD 55 [2005]
- TMRG IRTF workgroup since 2005: Models for the Evaluation of Transport Protocols

Contributions

- ① Synthèse des solutions de transport très haut débit.
- ② Méthodologie d'évaluation.
- ③ Collection d'expériences à large échelle.
- ④ Outil d'automatisation.
- ⑤ Définition d'un benchmark pour transport haut débit.
- ⑥ Outil d'analyse de configuration de transport.

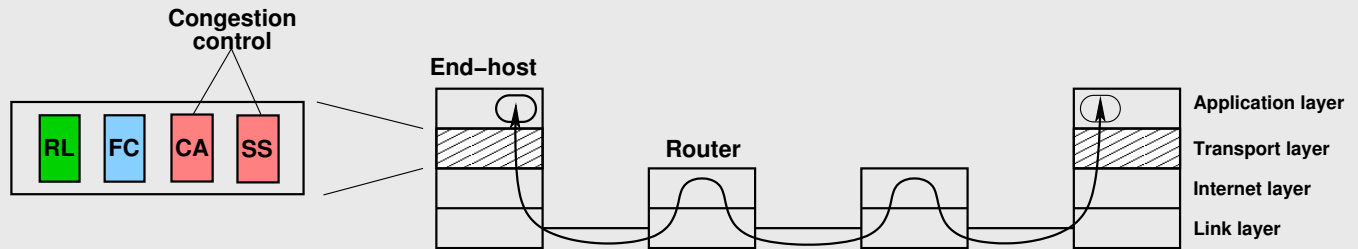
Plan

- 1 TCP dans les réseaux très haut débit
- 2 Méthodologie d'évaluation
- 3 Expériences à grande échelle
- 4 Benchmark orienté utilisateur

Plan

- 1 TCP dans les réseaux très haut débit
- 2 Méthodologie d'évaluation
- 3 Expériences à grande échelle
- 4 Benchmark orienté utilisateur

Rôle fondamental de TCP dans l'Internet



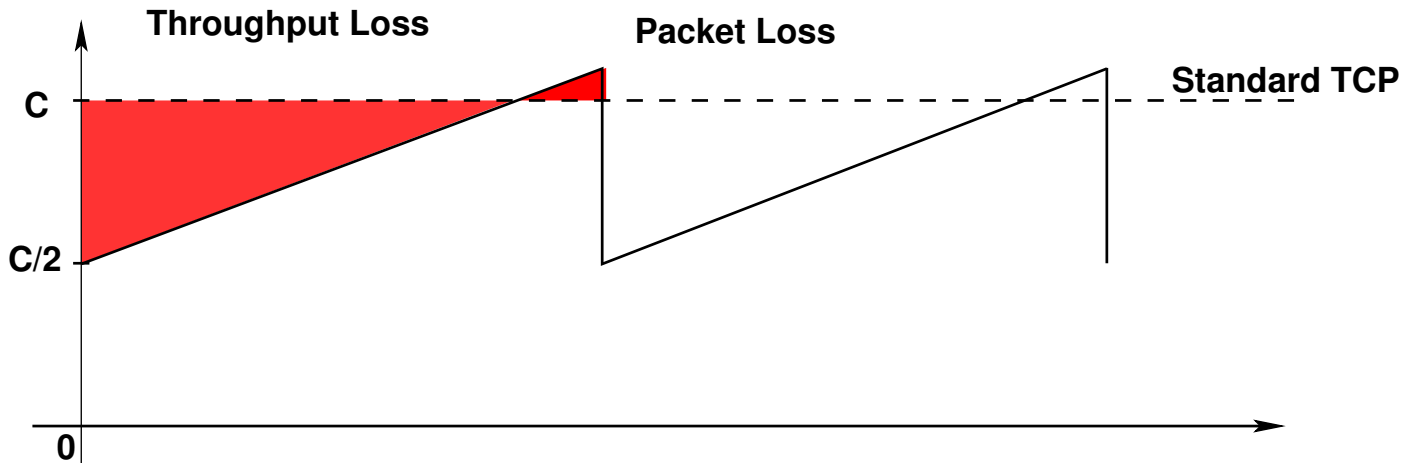
- TCP fournit: transfert **fiable**, contrôle de flux et contrôle de congestion [TCP 81, Cerf 74] de bout-en-bout
- Algorithme complètement **distribué** dans les hôtes d'extrémité (extensibilité)
- **Robuste**
- Partage **équitable** des liens [Bonald 01]
- Contrôle de congestion **stable** [Chiu 89]

Principe du contrôle de congestion dans TCP

Débit régulé par un mécanisme de fenêtre glissante à anticipation.

Evolution de la fenêtre de congestion de TCP (AIMD) [Jacobson 88]

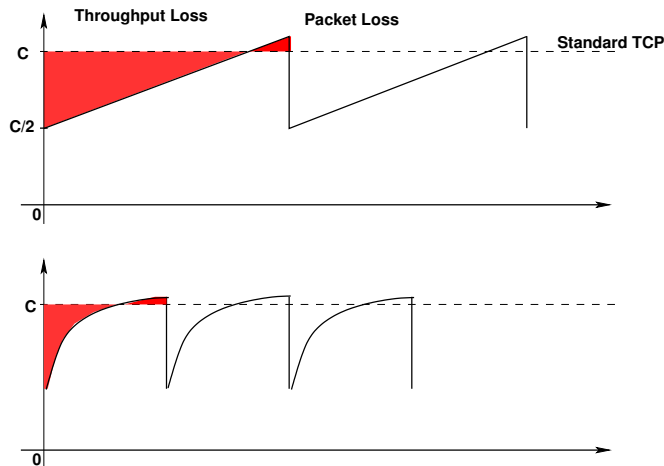
- ACK : $cwnd \leftarrow cwnd + \frac{\alpha}{cwnd}$
- Drop : $cwnd \leftarrow cwnd - \beta * cwnd$
- **Reno**[Jacobson88] : $\alpha = 1; \beta = \frac{1}{2}$



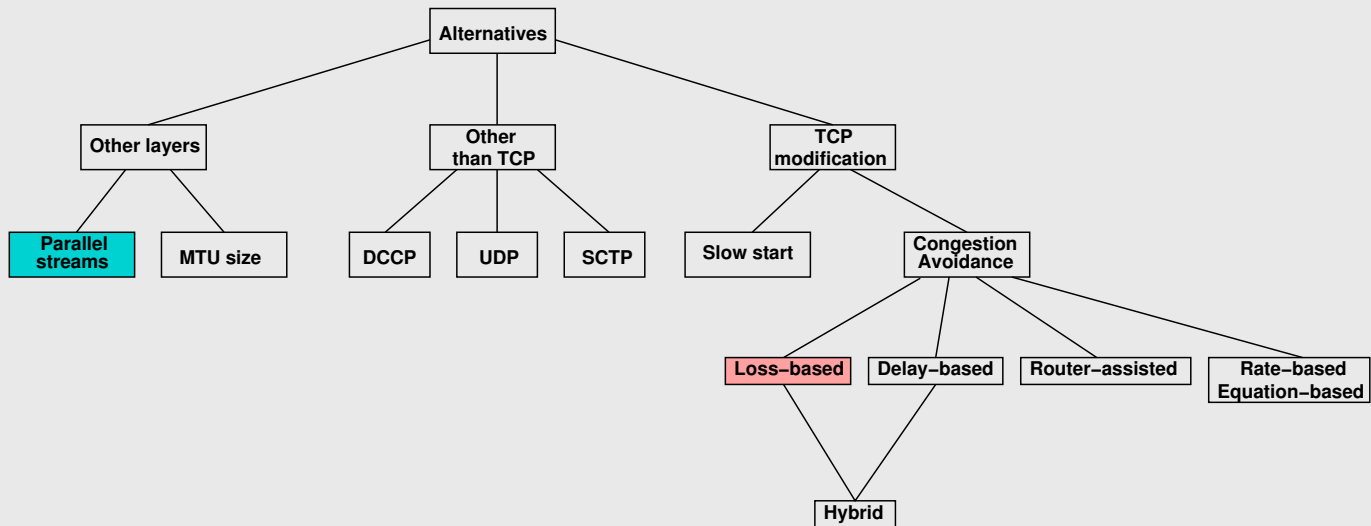
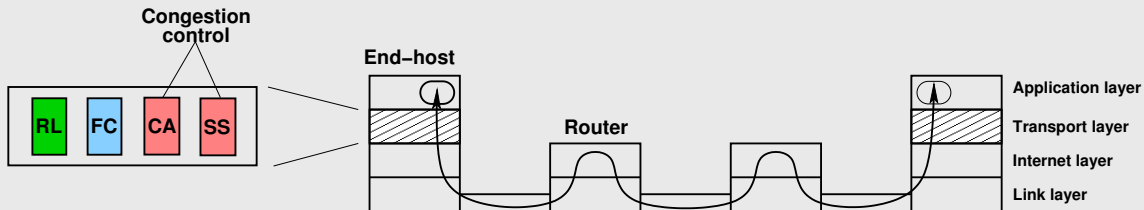
Principe du contrôle de congestion dans TCP

Limites de TCP dans les réseaux à fort produit débit-délai

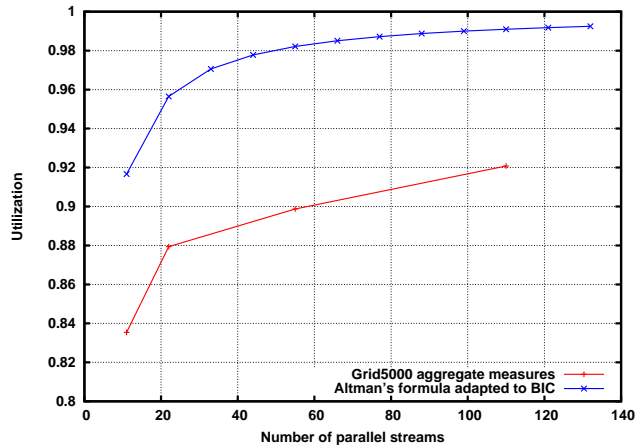
- Modèle simplifié de TCP: $Rate = \frac{MSS}{RTT} \sqrt{\frac{3}{2p}}$ [Padhye 98]
- Avec un RTT de 100 ms et des paquets de 1500 octets, au plus une perte tous les 5e9 paquets pour atteindre un débit stable de 10 Gbps.
- Taille de fenêtre moyenne autour de 83000 paquets à 10 Gbps.
- Une “dent de scie” met environ 1h30.



Solutions alternatives à TCP



Utilisation de flux parallèles



$$\text{Altman's formula [Altman 06]} \quad \bar{x}(N) = C \left(1 - \frac{1}{1 + \frac{1+\beta}{1-\beta} N} \right)$$

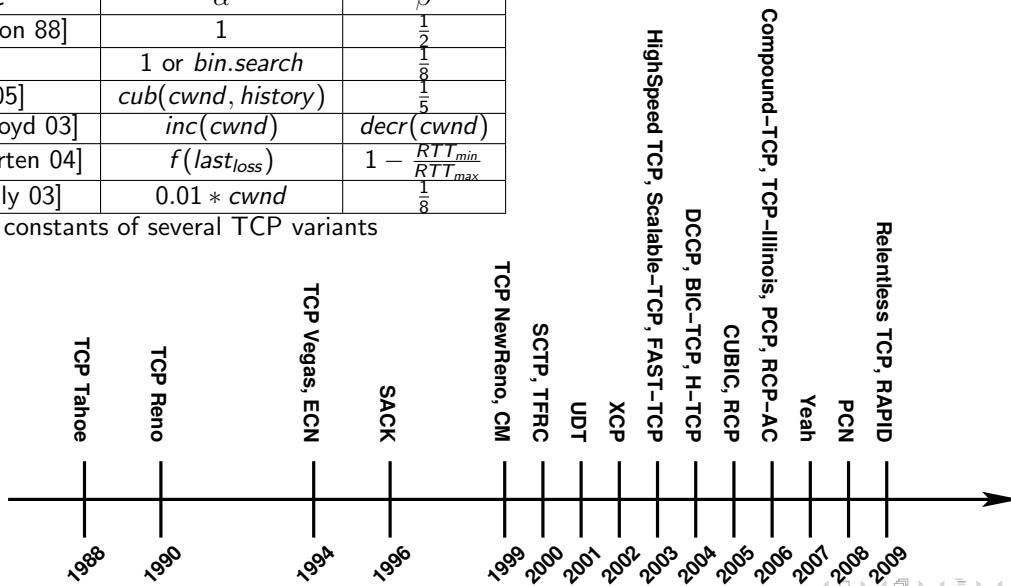
Number of streams by node	1	2	5	10
Total number of streams	11	22	55	110
Mean total goodput (Mbps)	8353.66	8793.92	8987.49	9207.78
Stream mean goodput (Mbps)	761.70	399.83	163.53	83.71
Goodput gain	/	4.9%	7.3%	9.8%
Jain Index per stream	0.9993	0.9979	0.9960	0.9973
Jain Index per transfer	0.9993	0.9994	0.9998	0.9998

Variantes de TCP

- Depuis 2002, plus de 10 variantes de TCP proposées.
- Implémentations disponibles (Linux, Windows)
- Modification de α et β de l'AIMD pour augmenter l'agressivité.
- **Besoin de s'assurer de la conservation des propriétés de TCP**

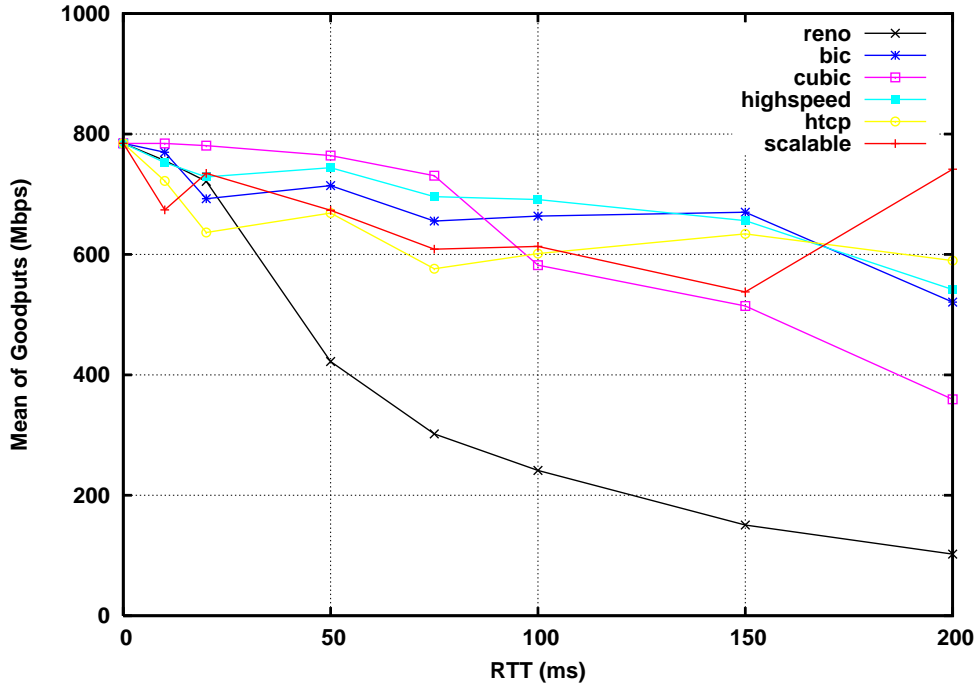
TCP variant	α	β
TCP Reno [Jacobson 88]	1	$\frac{1}{2}$
BIC [Xu 04]	1 or <i>bin.search</i>	$\frac{1}{2}$
CUBIC [Rhee 05]	<i>cub(cwnd, history)</i>	$\frac{1}{5}$
HighSpeed TCP [Floyd 03]	<i>inc(cwnd)</i>	<i>decr(cwnd)</i>
Hamilton TCP [Shorten 04]	<i>f(last_{loss})</i>	$1 - \frac{RTT_{min}}{RTT_{max}}$
Scalable TCP [Kelly 03]	$0.01 * cwnd$	$\frac{1}{2}$

AIMD constants of several TCP variants

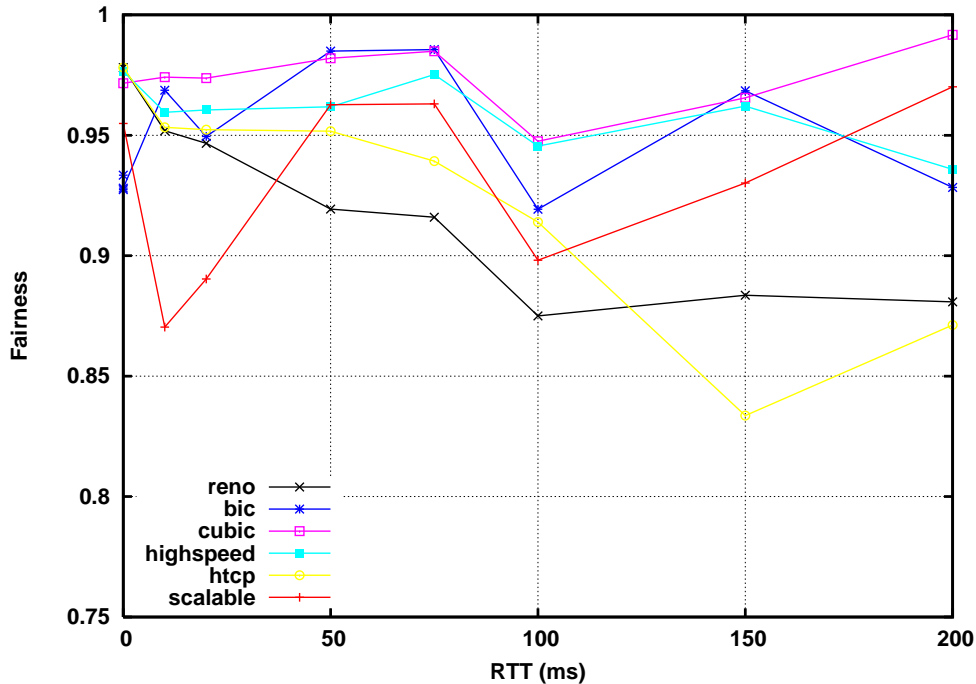


Impact de la latence sur le débit applicatif moyen

12 sources à 1 Gbps partageant un lien à 10 Gbps. Latence émulée.



Impact de la latence sur l'équité (indice de Jain)



Synthèse des expériences préliminaires

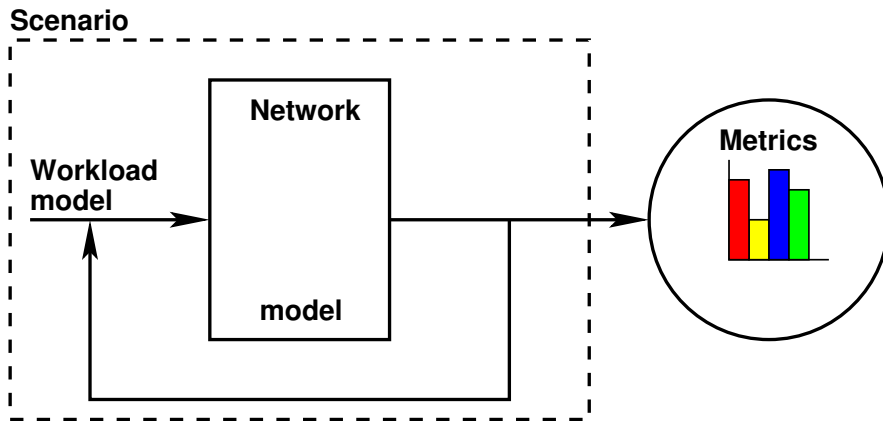
RTT	Flow mean goodput		Mean fairness		Normalised standard deviation	
	11.5 ms	100 ms	11.5 ms	100 ms	11.5 ms	100 ms
Reno	756.0	234.3	0.951	0.918	0.222	0.232
BIC	781.1	653.7	0.969	0.919	0.176	0.306
CUBIC	784.5	534.3	0.974	0.961	0.144	0.140
HS-TCP	753.6	671.9	0.960	0.962	0.069	0.233
H-TCP	722.2	686.1	0.953	0.926	0.230	0.256
Scalable	674.0	540.4	0.870	0.955	0.337	0.317

- Ecarter (ou pas) certaines variantes ?
- Critères de comparaison ?
- Besoin de définir une **méthodologie**

Plan

- 1 TCP dans les réseaux très haut débit
- 2 Méthodologie d'évaluation
 - Métriques
 - Modèles de réseaux
 - Modèles de charge
 - Illustrations
- 3 Expériences à grande échelle
- 4 Benchmark orienté utilisateur

Principes généraux de l'évaluation de solutions de transport



Metriques: valuation des propriétés recherchées

Modèle de réseau: organisation et grandeurs caractéristiques de l'infrastructure physique

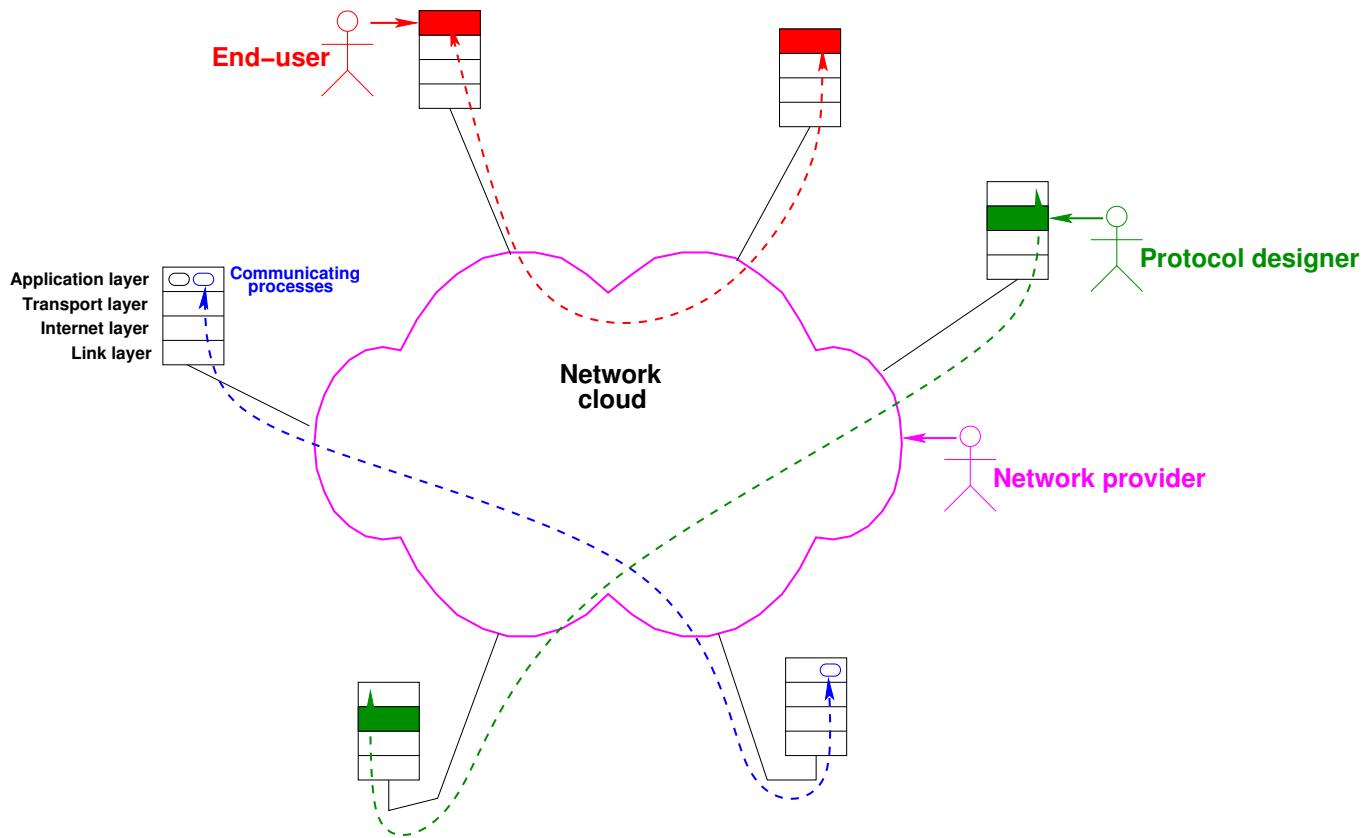
Modèle de charge: caractérisation de l'utilisation de l'infrastructure physique

Métriques

Issu des travaux du groupe TMRG [Andrew 08, Flo 07, Flo 06]

Metric type	Metric name	Formula
Throughput	Aggregate goodput	$G(t) = \sum_{i=1}^N g_i(t)$
	Aggregate througput	$X(t) = \sum_{i=1}^N x_i(t)$
	Efficiency	$E(t) = \frac{G(t)}{X(t)}$
Reactivity	Responsiveness	$R = \min_u \{u x_i(t + u * RTT) \leq \frac{x_i(t)}{2}\}$
	Average responsiveness	$\overline{R}_X = \max_t \{t : \frac{\sum_i x_i(t) - \sum_i \bar{x}_i}{\sum_i \bar{x}_i} > 0.1\}$
	Aggressiveness	$A = \max \{x_i(t + RTT) - x_i(t)\}$
Stability	Goodput variance	$\sigma_i = \sqrt{\frac{1}{T_i} \sum_{t=0}^{T_i} (g_i(t) - \bar{g}_i)^2}$
	Throughput smoothness	$S = \max x_i(t) - x_i(t + \delta_t) $
	Goodput CoV	$CoV_i = \frac{\sigma_i}{\bar{g}_i}$
Fairness	Jain Index	$J = \frac{(\sum_{i=1}^N \bar{g}_i)^2}{N(\sum_{i=1}^N \bar{g}_i^2)}$
Convergence	Convergence speed	$S_{cwnd_i} = \min_u \{u cwnd_i(t + u * RTT) \leq .8 * BDP\}$

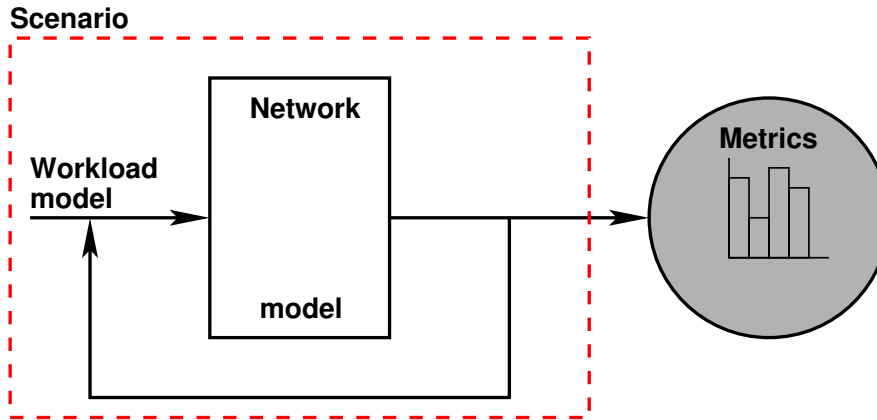
Points de vue



Métriques et points de vue

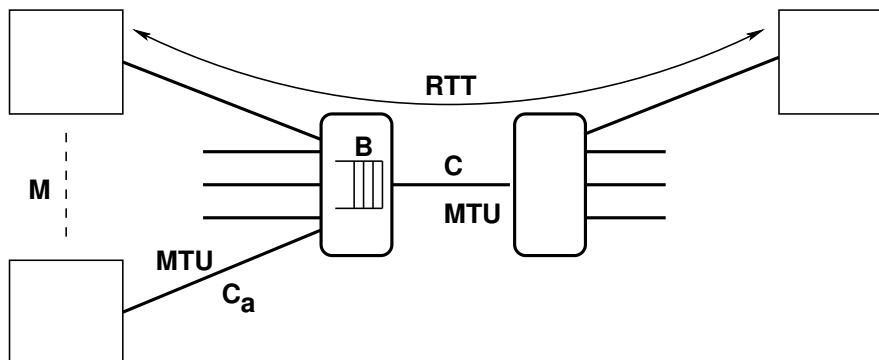
Metrics of	End User	Protocol Designer	Network Provider
Throughput	Goodput G , Completion time T	Cong. window cwnd Throughput X Efficiency E	Throughput X , Link utilization U ,
Delay	RTT	Queueing delay q	Queueing delay q
Loss	Timeouts events t	Packet loss rate p	Packet loss rate p
Reactivity	Aggressiveness A	Responsiveness R Aggressiveness A	Aggressiveness A
Stability	Variance σ	Variance σ Smoothness S	Coeff. of Variation CoV
Fairness	Delta-fair convergence δ_f	Jain Index J	Max-min, Proportional, α -fairness
Convergence	N/A	Convergence Speed s	N/A
Robustness	Retransmission r	Performability p_S [Meyer 80]	Kolmogorov-Smirnov statistic θ_w [England 05]
Deployability	Application modification	Code complexity	Hardware upgrades complexity

Définition des modèles de réseaux et de charge



Capturer l'évolution de la structure et de l'usage du système.

Paramètres réseaux



Parameter	Description	Typical range in the Internet	Actor
RTT	Round Trip Time	0 to 300 ms (land-line) 0 to 800 ms (satellite)	U-D-N
C	Bottleneck capacity	56 kbps to 40 Gbps	D-N
B	Buffer size	20 % BDP to 100 % BDP	D
$K = \frac{C}{C_a}$	Aggregation level	1 to 10000	N
M	Multiplexing factor	1 to 1000	N
MTU	Maximum transmission unit	1500 B to 9000 B	U-D-N

Paramètres de charge

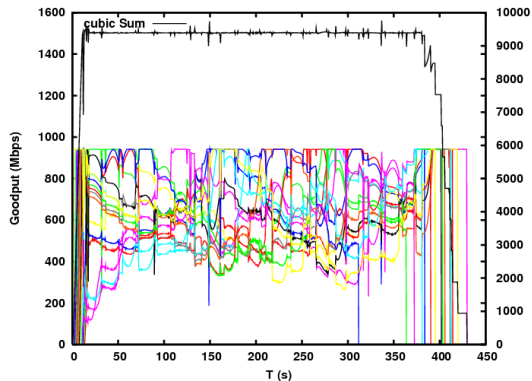
Parameter	Description	Typical Range in the Internet	Actor
C_g	Congestion level	0 to 2.0	D-N
R	Reverse traffic level	0 to 2.0	D-N
B_g	Background traffic	0 to 0.8	D-N
N_s	Parallel streams	0 to 10	U-D

Définition

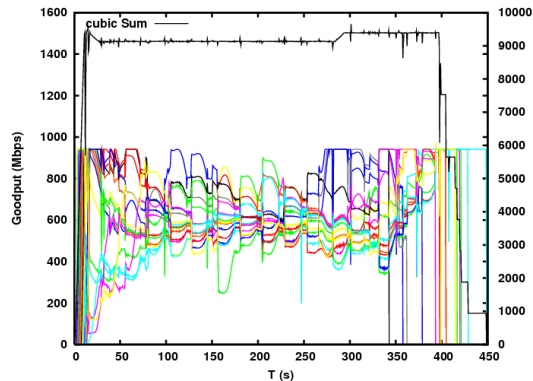
- def1: $C_g = \frac{N_{sources} * C_a}{C}$
- def2: $C_g = \sum_{sources} \frac{C_a}{C} * \frac{\mu_{ON}}{\mu_{OFF} + \mu_{ON}}$

Influence du trafic reverse sur Cubic (150 % cong. lvl)

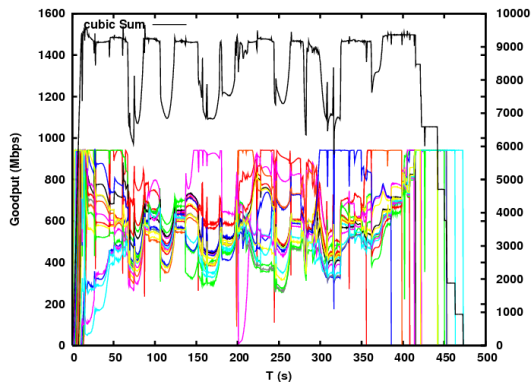
No reverse (395 s)



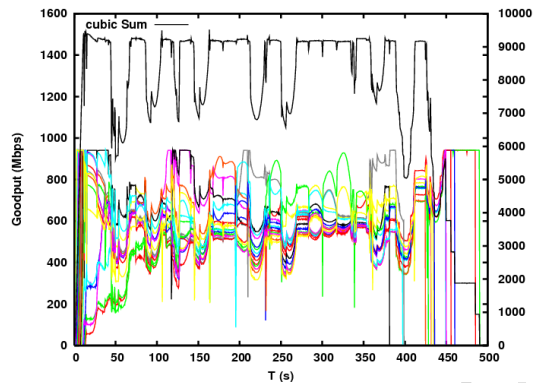
90 % reverse (400 s)



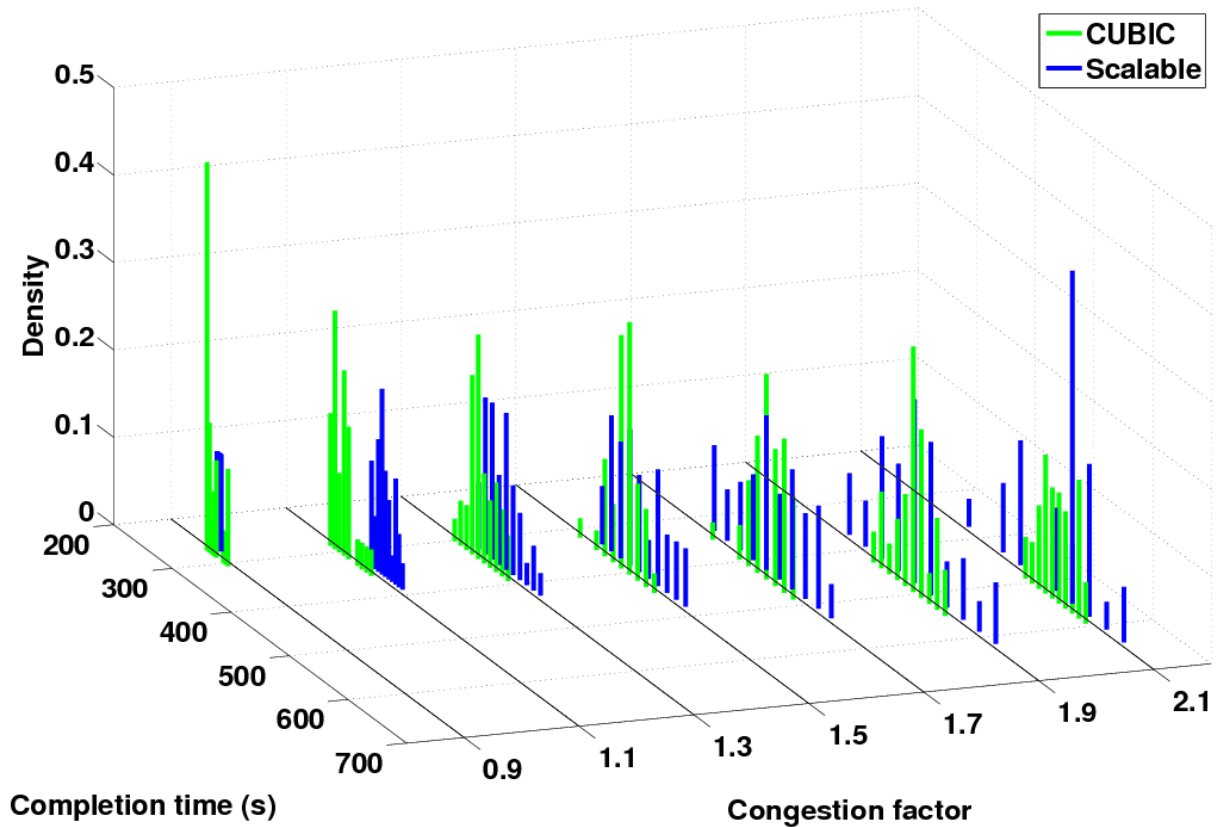
110 % reverse (432 s)



150 % reverse (438 s)



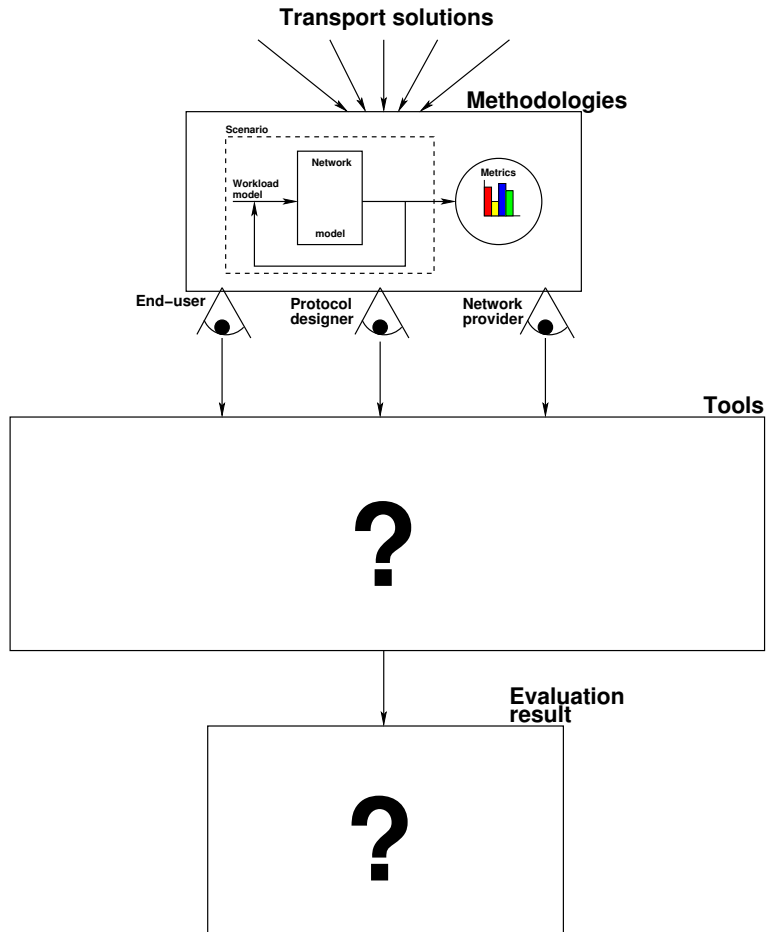
Distribution des temps de complétion



Plan

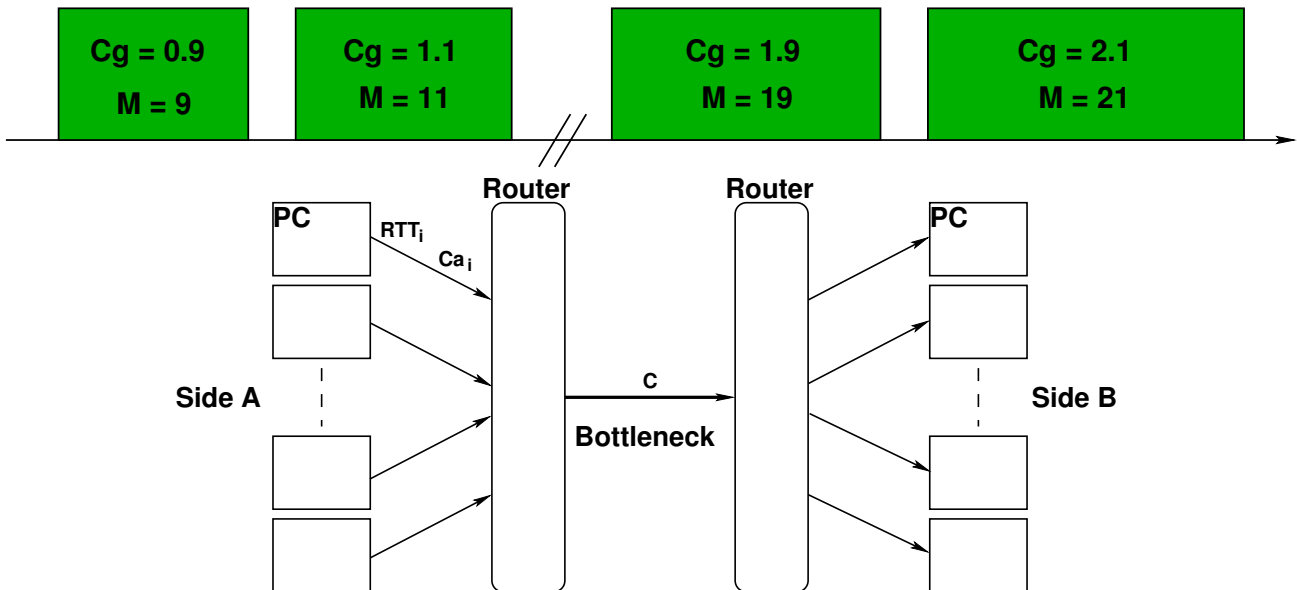
- 1 TCP dans les réseaux très haut débit
- 2 Méthodologie d'évaluation
- 3 Expériences à grande échelle
 - Network eXperiment Engine
 - NS-2 vs. NXE
- 4 Benchmark orienté utilisateur

Retour à la problématique

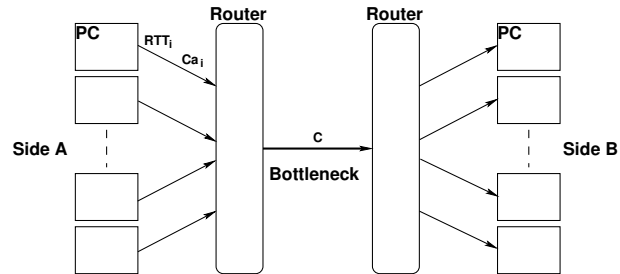


Impact des conditions de trafic

Etude de l'**impact des conditions de trafic** (congestion, trafic reverse) sur le temps de complétion de transferts de fichiers de 3000 MB [Guillier 07c]. Jusqu'à 42 sources indépendantes, emettant simultanément.



Espace paramétrique



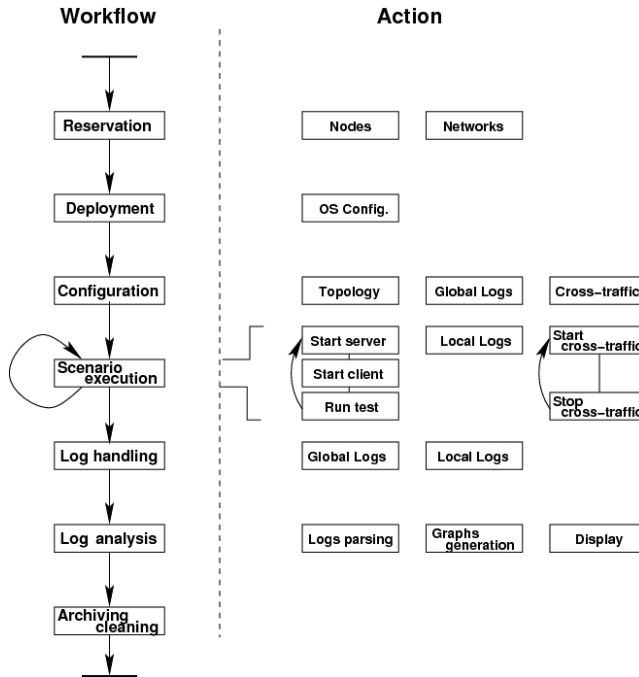
	Parameter	Description	Range
Infrastructure	RTT	Round Trip Time	0 to 200 ms
	C	Bottleneck capacity	1 or 10 Gbps
	$K = \frac{C}{C_a}$	Aggregation lvl	1 or 10
Workload	M	Multiplexing factor	1 to 20
	N_s	Parallel streams	1 to 10
	C_g	Congestion factor	0 to 2.0
	R	Reverse traffic factor	0 to 2.0

- Très vaste espace paramétrique.
- Multiplication des expériences pour validité statistique.
- **Besoin d'un outil pour automatiser ce processus.**

Etapes du déroulement d'une expérience

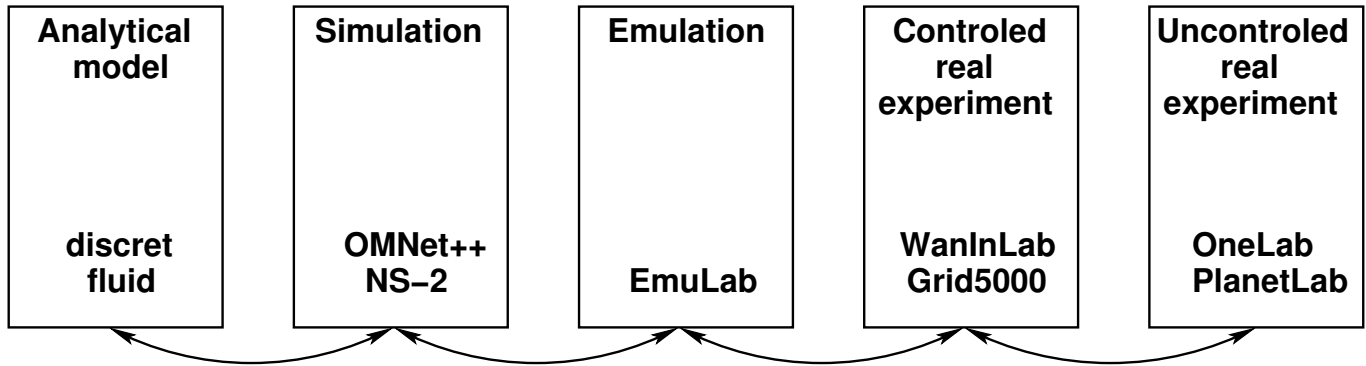
- ① **Réservation:** récupération des ressources.
 - 2 à 100 noeuds
- ② **Deploiement:** phase de configuration des variables internes de l'OS.
 - 5 à 10 variables
- ③ **Configuration:** modification de la topologie, activation de la capture d'informations.
- ④ **Execution du scenario:** plusieurs fois si nécessaire.
 - durée de quelques seconds à plusieurs heures
 - 1 à 10 répétitions
- ⑤ **Gestion des logs:** récupération des logs.
 - volume de quelques KB à plusieurs GB
- ⑥ **Analyse des logs:** calcul des métriques, génération des graphes.
- ⑦ **Archivage**

Outil Network eXperiment Engine



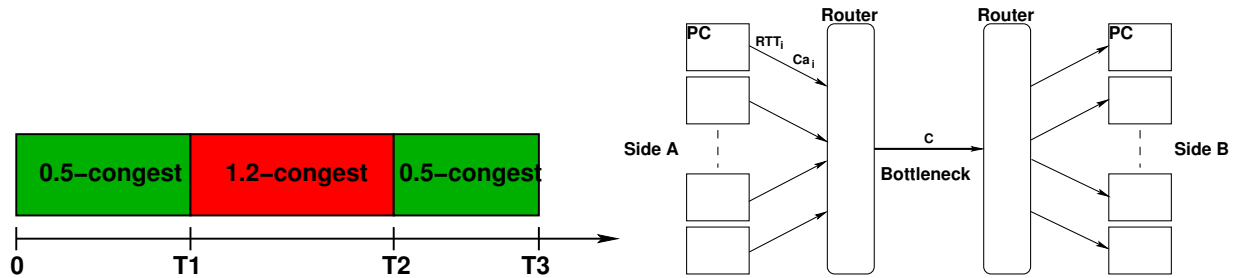
- codé en Python/shell script.
- Utilisateurs dans Grid'5000.
- disponible <http://www.ens-lyon.fr/LIP/RESO/Software/NXE/index.html>

Différentes méthodes d'évaluation



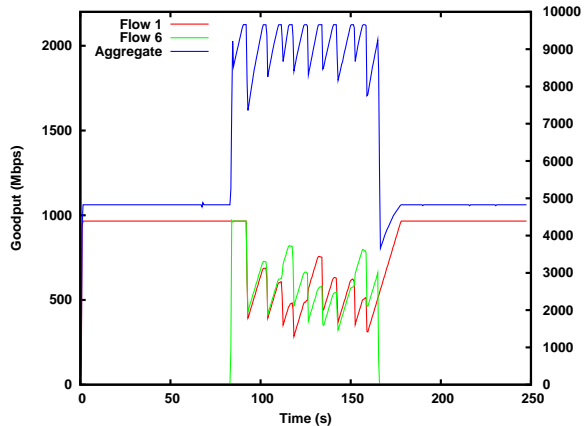
- Avantages/limites de chaque méthode
- **Complémentaires**
- Comparaison possible des différentes approches ?

Comparaison NS-2 and NXE

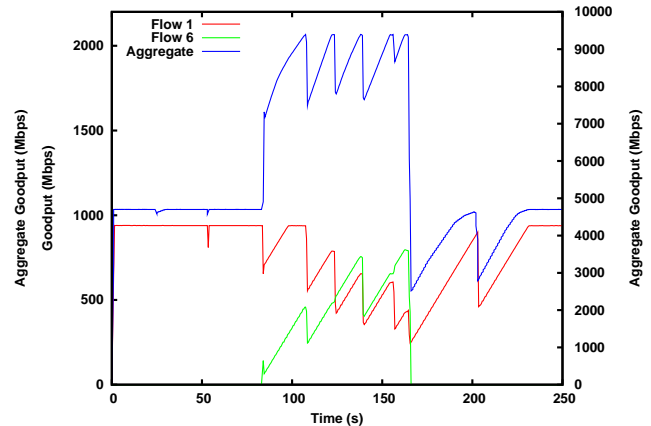


- Jusqu'à 12 sources 1 Gbps partageant un lien à 10 Gbps.
- 20 ms RTT
- Chaque période dure ~ 82 s (total ~ 250 s).
- TCP Reno

Resultats NS-2 et NXE

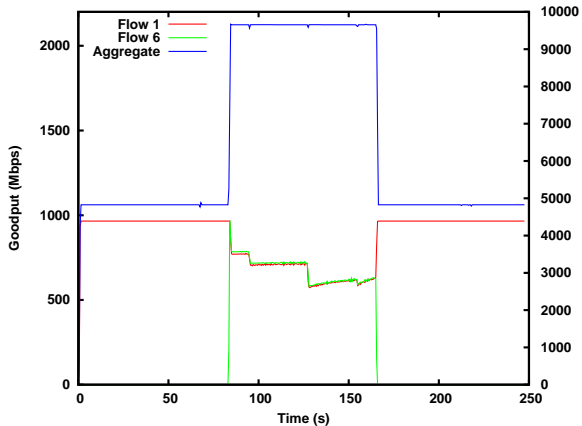


NS-2 (small buffers 170 paquets)

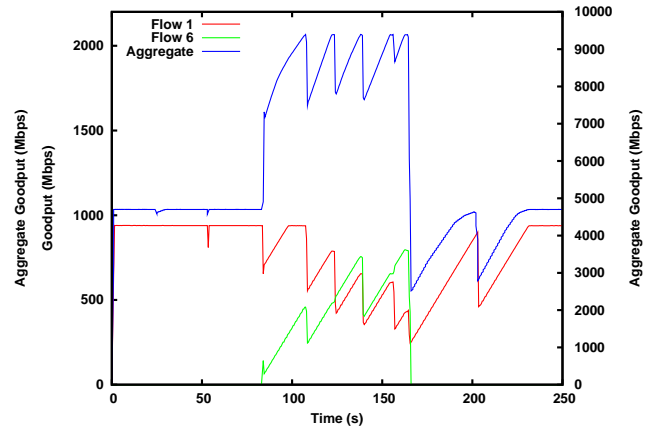


NXE

Resultats NS-2 et NXE



NS-2 (large buffers 17000 paquets)



NXE

Comparaison NS-2 et NXE: temps d'exécution

Tool	Experiment setup (s)	Experiment execution (s)	Experiment post-processing (s)	Total (s)
NS-2	1	2360	5	2366
NXE	460 (12)	260	10	730 (282)

- NS-2: temps d'exécution exponentiel [Wei 06]
- coût machine pour 10 répétitions:
 - NS-2: 1 noeud pendant 6h40mn → 6h40mn
 - NXE: 12 noeuds pendant 54mn → 10h48mn
- NXE: conditions de la plateforme expérimentale
 - expérience avec des vraies applications.
 - interactions entre le logiciel et le matériel.
 - influence de paramètres non capturés par la simulation.

Plan

- 1 TCP dans les réseaux très haut débit
- 2 Méthodologie d'évaluation
- 3 Expériences à grande échelle
- 4 Benchmark orienté utilisateur**

Lignes directrices

But

Comparaison de la performance relative de plusieurs solutions alternatives.
Exemple: NAS pour les implémentations de MPI.
TMRG test-suite orienté protocol-designer de TCP.

Propriétés voulues

Facile à utiliser

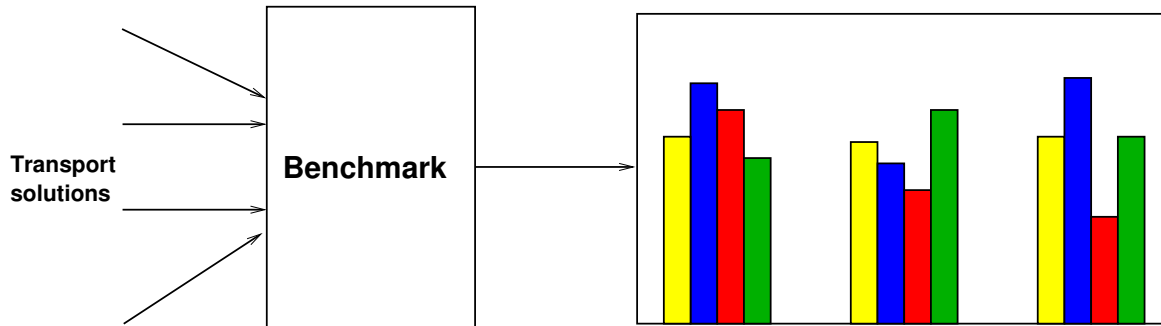
Representatif

Portable

Reproductabilité

Bien défini

Cahier des charges



Mean completion time: $\bar{T} = \frac{1}{N} \sum_{i=1}^{N_f} T_i$

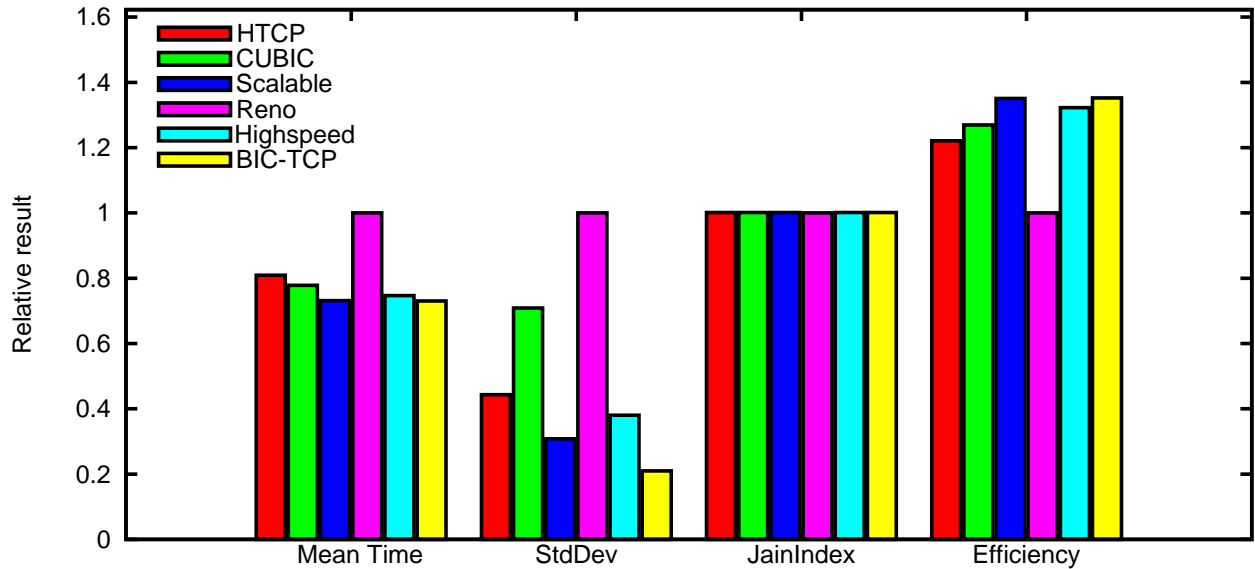
Standard deviation of completion time: $\sigma_{T_i} = \sqrt{\frac{1}{N} \sum_{n=1}^{N_f} (T_i - \bar{T})^2}$

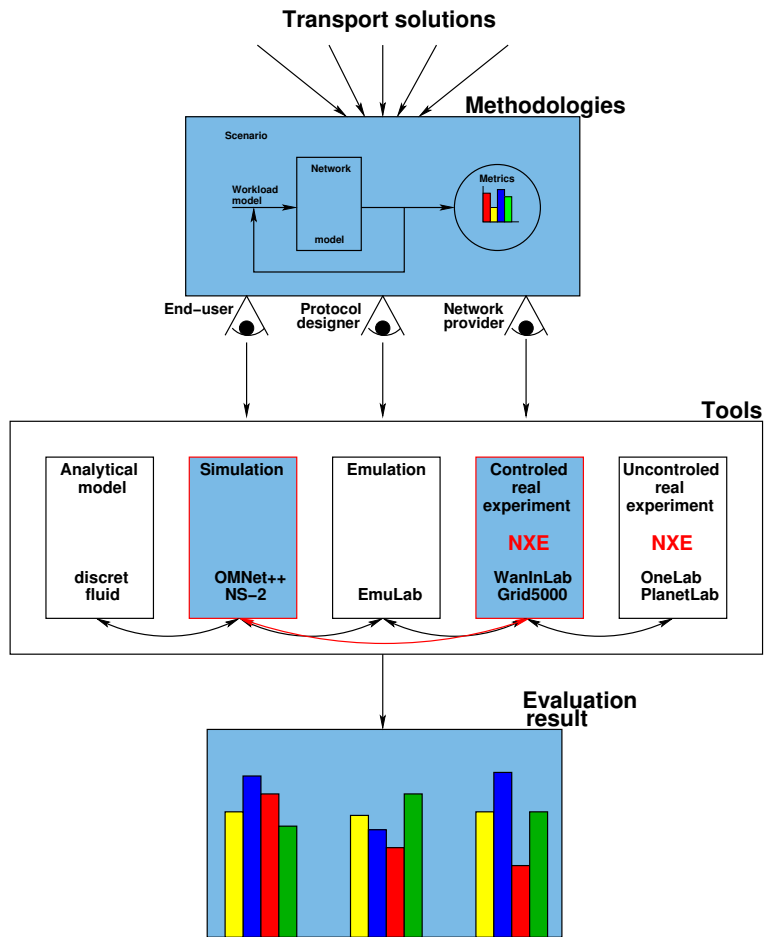
Fairness: $J = \frac{(\sum_{i=1}^N \bar{T}_i)^2}{N(\sum_{i=1}^N \bar{T}_i^2)}$

Efficiency: $E(t) = \frac{G(t)}{C}$

Exemple [Guillier 07b]

HSTTS Result graph for BU with 19.8ms RTT, Cg=1.0 and R=1.0 with reference Reno












Conclusions

- Impact non négligeable des problèmes de configuration (wizard gap)
 - Solutions qui marchent bien: flux parallèles, multipath mais agressives
 - Nouvelles variantes: performances dépendent de certaines caractéristiques non maitrisables du réseau.
 - Partage équitable de la bande passante bloquant ?
-
- Fournir les outils pour détecter (expérimentalement) quelle solution convient à une situation donnée.









Contributions

- ① Synthèse des solutions de transport pour le très haut débit.
 - 1 article journal (en cours)
 - 1 chapitre de livre (en cours)
- ② Méthodologies d'évaluation.
 - 1 article journal (en cours)
 - 1 publication internationale [Guillier 08b]
- ③ Expériences à large échelle.
 - 4 publications internationales [Guillier 07a, Soudan 07a, Soudan 07b, Guillier 07d]
 - 2 rapports de recherche [Guillier 06, Guillier 09a]
 - 1 démonstration [Guillier 08a]
- ④ Outil d'automatisation.
 - 1 publication internationale [Loiseau 09a]
 - 1 logiciel libre [Guillier 09c]
 - 1 démonstration [Loiseau 09b]
- ⑤ Définition de benchmarks.
 - 2 publications internationales [Andrew 08, Guillier 09b]
 - 1 démonstration [Guillier 07e]
- ⑥ Outil d'analyse de configuration.
 - 1 brevet (valorisation start-up LinkTiss) [Guillier 09d]









References I

-  Eitan Altman, Dhiman Barman, Bruno Tuffin & Milan Vojnovic.
Parallel TCP Sockets: Simple Model, Throughput and Validation.
In Proceedings of the IEEE INFOCOM, 2006.
-  Lachlan Andrew, Cesar Marcondes, Sally Floyd, Lawrence Dunn, Romaric Guillier, Wang Gang, Lars Eggert, Sangtae Ha & Injong Rhee.
Towards a Common TCP Evaluation Suite.
In PFLDNet, march 2008.
-  Raphaël Bolze, Franck Cappello, Eddy Caron, Michel Daydé, Frederic Desprez, Emmanuel Jeannot, Yvon Jégou, Stéphane Lanteri, Julien Leduc, Noredine Melab, Guillaume Mornet, Raymond Namyst, Pascale Vicat-Blanc Primet, Benjamin Quetier, Olivier Richard, El-Ghazali Talbi & Touché Irena.
Grid'5000: a large scale and highly reconfigurable experimental Grid testbed.
International Journal of High Performance Computing Applications, vol. 20, no. 4, pages 481–494, November 2006.
-  T. Bonald, S. Ben Fredj, A. Proutiere, G. Regnie & J.W. Roberts.
Statistical Bandwidth Sharing: A Study of Congestion at Flow Level.
In SIGCOMM. ACM, Aug. 2001.
-  V. Cerf & R. Kahn.
A Protocol for Packet Network Intercommunication.
In IEEE Transactions on Communications, volume 22, pages 637–648, may 1974.
-  D. Chiu & R. Jain.
"Analysis of the Increase/Decrease Algorithms for Congestion Avoidance in Computer Networks.
Journal of Computer Networks and ISDN, vol. 17, no. 1, pages 1–14, June 1989.
-  Darin England, Jon Weissman & Jayashree Sadagopan.
A new Metric for Robustness with Application to Job Scheduling.
In High Performance Distributed Computing, Jul. 2005.









References II

-  *Tools for the Evaluation of Simulation and Testbed Scenarios.*
In Sally Floyd & E Kohler, editeurs, <http://www.icir.org/tmrg/draft-irtf-tmrg-tools-03.txt>, December 2006.
-  *Metrics for the evaluation of Congestion Control Mechanisms.*
In Sally Floyd, editeur, <http://www.icir.org/tmrg/draft-irtf-tmrg-metrics-11.txt>, October 2007.
-  Sally Floyd.
RFC 3649: HighSpeed TCP for Large Congestion Windows.
RFC 3649, December 2003.
experimental.
-  Romaric Guillier, Ludovic Hablot, Pascale Vicat-Blanc Primet & Sebastien Soudan.
Evaluation des liens 10 GbE de Grid'5000.
Research Report 6047, INRIA, 12 2006.
-  Romaric Guillier, Ludovic Hablot, Yuetsu Kodama, Tomohiro Kudoh, Fumihiko Okazaki, Ryousei Takano, Pascale Vicat-Blanc Primet & Sebastien Soudan.
A study of large flow interactions in high-speed shared networks with Grid5000 and GtrcNET-10 instruments.
In PFLDnet 2007, Feb. 2007.
-  Romaric Guillier, Ludovic Hablot & Pascale Vicat-Blanc Primet.
Towards a User-Oriented Benchmark for Transport Protocols Comparison in very High Speed Networks.
Research Report 6244, INRIA, 07 2007.
Also available as LIP Research Report RR2007-35.
-  Romaric Guillier, Sebastien Soudan & Pascale Vicat-Blanc Primet.
TCP variants and transfer time predictability in very high speed networks.
In Infocom 2007 High Speed Networks Workshop, May 2007.
-  Romaric Guillier, Sebastien Soudan & Pascale Vicat-Blanc Primet.
TCP variants and transfer time predictability in very high speed networks.
In Infocom 2007 High Speed Networks Workshop, May 2007.






References III

-  Romaric Guillier & Pascale Vicat-Blanc Primet.
High Speed Transport Protocol Test Suite.
poster, SuperComputing 2007, November 2007.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
Congestion Collapse in Grid5000.
demo, "The Future of TCP: Train-wreck or Evolution?", Stanford Congestion Collapse workshop, April 2008.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
Methodologies and Tools for Exploring Transport Protocols in the Context of High-Speed Networks.
In IEEE TCSC Doctoral Symposium, May 2008.
-  Romaric Guillier, Sebastien Soudan & Pascale Vicat-Blanc Primet.
UDT and TCP without Congestion Control for Profile Pursuit.
Research Report 6874, INRIA, 03 2009.
Also available as LIP Research Report RR2009-10.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
A User-Oriented Test Suite for Transport Protocols Comparison in DataGrid Context.
In ICOIN 2009, January 2009.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
NXE, 2009.
Software, APPcode: IDDN.FR.001.030005.000.S.P.2009.000.10800.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
PATHNIF.
Patent under study, 2009.
-  Romaric Guillier & Pascale Vicat-Blanc Primet.
PathNIF, 2009.
Software, APPcode: IDDN.FR.001.260002.000.S.P.2009.000.10800.

References IV

-  Van Jacobson.
Congestion Avoidance and Control.
In SIGCOMM'88, 1988.
-  Tom Kelly.
Scalable TCP: Improving Performance in Highspeed Wide Area Networks.
In Computer Communication Review, volume 32, April 2003.
-  Patrick Loiseau, Paulo Gonçalves, Romaric Guillier, Matthieu Imbert, Yuetsu Kodama & Pascale Vicat-Blanc Primet.
Metroflux: A high performance system for analyzing flow at very fine-grain.
In TridentCom, Apr. 2009.
-  Patrick Loiseau, Romaric Guillier, Oana Goga, Matthieu Imbert, Paulo Goncalves & Pascale Vicat-Blanc Primet.
Automated Traffic Measurements and Analysis in Grid'5000.
demo, SIGMETRICS/Performance 2009, Recipient of Best Student Demo Award, June 2009.
-  John Meyer.
On Evaluating the Performability of Degradable Computing Systems.
In IEEE Transactions on Computers, volume 29, Aug. 1980.
-  J. Padhye, V. Firoiu, D. Towsley & J. Kurose.
Modeling TCP Throughput: A Simple Model and its Empirical Validation.
In ACM SIGCOMM '98, 1998.
-  Injong Rhee & Lisong Xu.
CUBIC: A New TCP-Friendly High-Speed TCP Variants.
In PFLDnet, 2005.
-  R.N. Shorten & Doug Leith.
H-TCP: TCP for high-speed and long-distance networks.
In PFLDnet'04, Argonne, Illinois USA, February 2004.

References V

-  Sebastien Soudan, Romaric Guillier, Ludovic Hablot, Yuetsu Kodama, Tomohiro Kudoh, Fumihiro Okazaki, Ryousei Takano & Pascale Vicat-Blanc Primet.
Investigation of Ethernet switches behavior in presence of contending flows at very high-speed.
In PFLDnet 2007, Feb. 2007.
-  Sebastien Soudan, Romaric Guillier & Pascale Vicat-Blanc Primet.
End-host based mechanisms for implementing Flow Scheduling in GridNetworks.
In GridNets 2007, Oct. 2007.
-  *Transmission Control Protocol.*
RFC 793, september 1981.
-  David X. Wei & Pei Cao.
NS-2 TCP-Linux: an NS-2 TCP implementation with congestion control algorithms from Linux.
In WNS2 '06: Proceeding from the 2006 workshop on ns-2: the IP network simulator, page 9, New York, NY, USA, 2006. ACM Press.
-  Lisong Xu, Khaled Harfoush & Injong Rhee.
Binary Increase congestion Control for Fast Long-Distance Networks.
In INFOCOM, 2004.