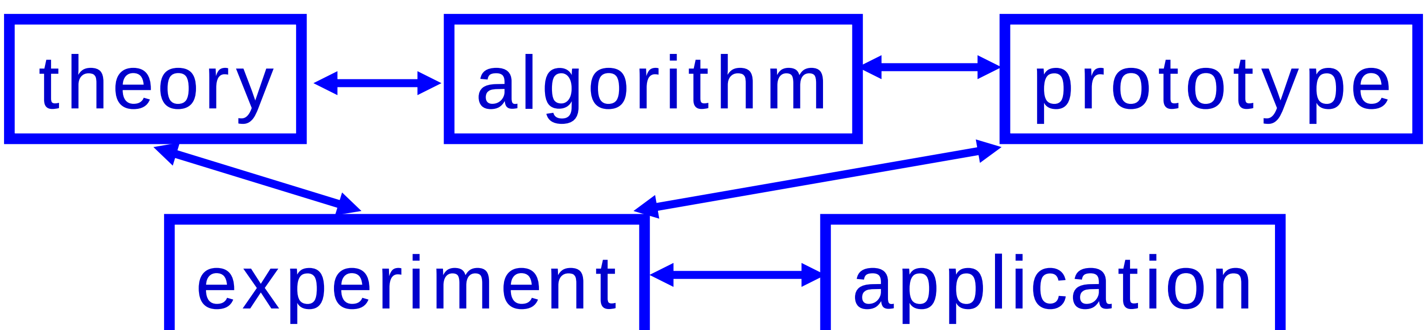
Flow Control Theory for Practitioners

Steven Low
EAS, Caltech



Acknowledgments

- Caltech
  - L. Andrews, J. Doyle, S. Hegde, C. Jin, G. Lee, L. Li, H. Newman, A. Tang, J. Wang, D. Wei, B. Wydrowski
- UCLA
  - F. Paganini
- Princeton
  - M. Chiang, L. Peterson, L. Wang
- KTH
  - K. Jacobsson

Role of (current) theory

- It is not (yet) for
  - Automatic synthesis of new congestion control algorithms
  - Replacing intuitions, experiments, heuristics
- But for providing structure and clarity
  - To refine intuition
  - To guide design
  - To suggest ideas
  - To explore boundaries
  - To assess global structural properties, e.g. scalability
- Risk
  - "All models are wrong"
  - "… some are useful"

Outline

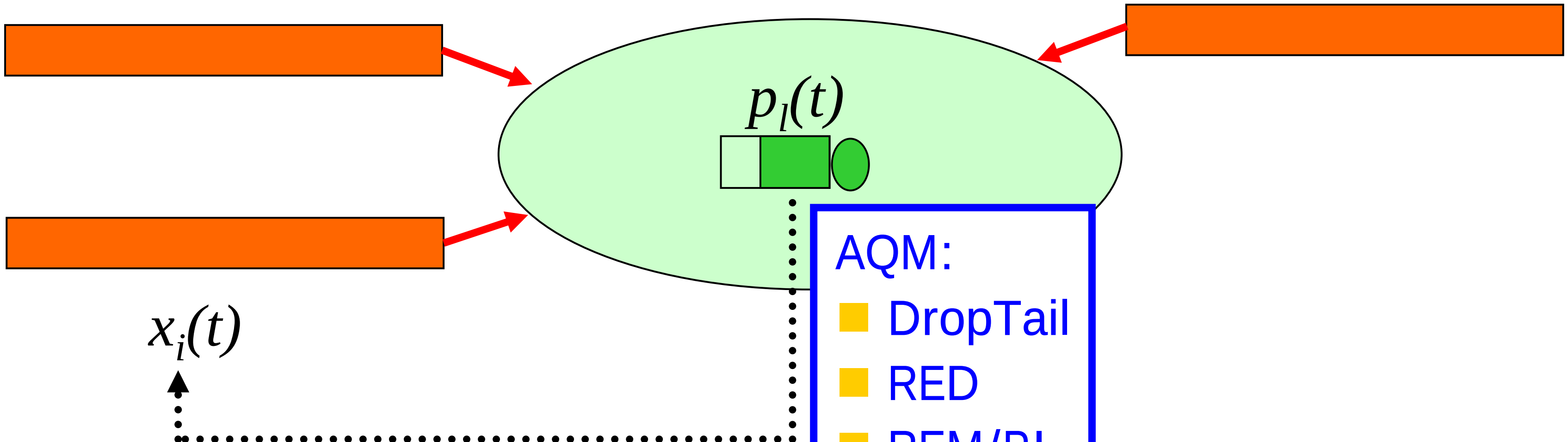
Samples of interactions between theory & experiments

- Duality model of TCP
  - Theory: equilibrium point characterized by an optimization problem
  - Experimental validation: Vegas
- An accurate link model
  - Theory: a new joint link model
  - Application: FAST stability
- Heterogeneous protocols
  - Motivation: FAST+Reno
  - Theory: multiple equilibria, global uniqueness

Congestion control

- Challenge: available info must be end-to-end
- Implicit congestion feedback
  - Loss probability: likelihood of a packet being delivered correctly
  - Round-trip time: time it takes for a packet to reach its destination and for its ack to return to the sender
- Explicit congestion feedback: marks, rates

TCP & AQM



TCP:

- Reno, Vegas
- HSTCP
- FAST
- BIC
- HTCP
- CTCP
- ……

AQM:

- DropTail
- RED
- REM/PI
- AVQ
- ……

Historically

- Packet level implemented first
- Flow level understood as after-thought
- But flow level design determines
  - performance, fairness, stability

Now: can forward engineer

- Sophisticated theory on equilibrium & stability (optimization+control)
- Given (application) utility functions, can design provably scalable TCP algorithms

Packet level

- **Reno**
  AIMD(1, 0.5)

```
ACK:  W  ←  W + 1/W
Loss: W  ←  W – 0.5W
```

- **HSTCP**
  AIMD(a(w), b(w))

```
ACK:  W  ←  W + a(w)/W
Loss: W  ←  W – b(w)W
```

- **STCP**
  MIMD(a, b)

```
ACK:  W  ←  W + 0.01
Loss: W  ←  W – 0.125W
```

- **FAST**

$$\mathrm{RTT}:\ \mathrm{W} \leftarrow \mathrm{W} \cdot \frac{\mathrm{baseRTT}}{\mathrm{RTT}} + \alpha$$

Flow level: Reno, HSTCP, STCP, FAST

- **Common** flow level dynamics!

$$\dot{w}_i(t) \quad = \quad \kappa(t) \quad \cdot \quad \left(1 - \frac{p_i(t)}{U_i'(t)}\right)$$

window adjustment	=	control gain	flow level goal

- **Different** gain κ and utility U_i
  - They determine equilibrium and stability
- **Different** congestion measure p_i
  - Loss probability (Reno, HSTCP, STCP)
  - Queueing delay (Vegas, FAST)

Flow level: Reno, HSTCP, STCP, FAST

- **Similar** flow level equilibrium

Reno $$x_i = \frac{1}{T_i} \cdot \frac{\alpha}{p_i^{0.5}}$$ pkts/sec

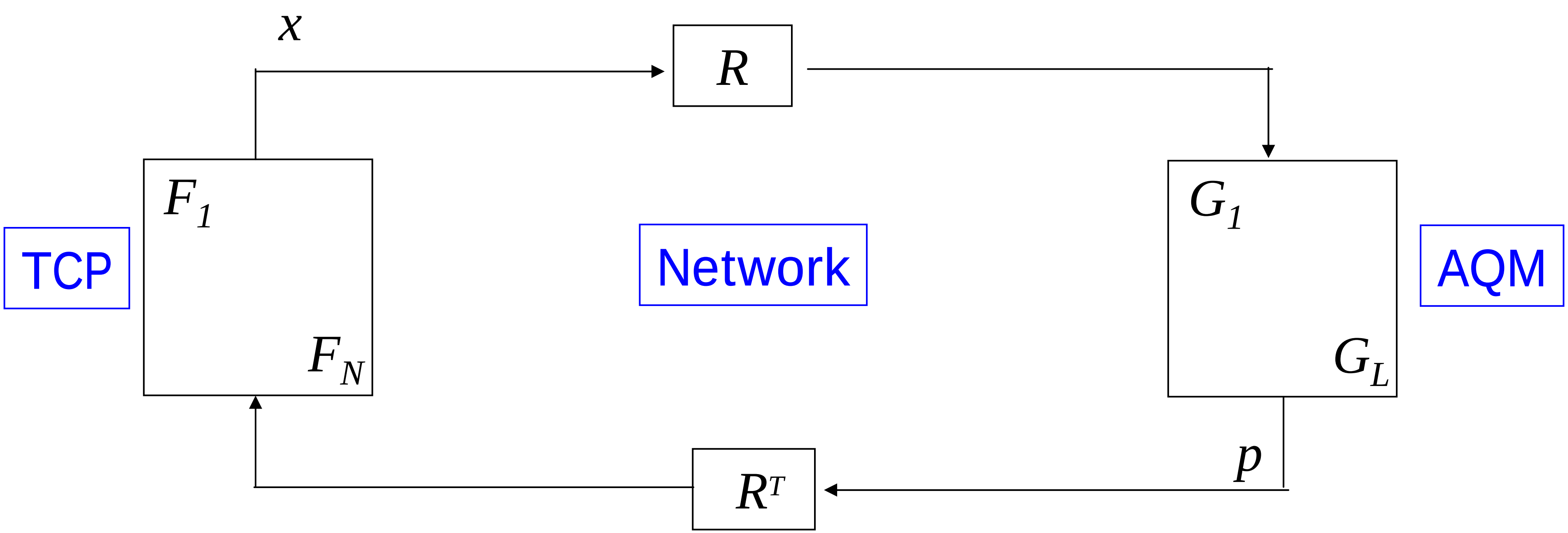
HSTCP $$x_i = \frac{1}{T_i} \cdot \frac{\alpha}{p_i^{0.84}}$$

STCP $$x_i = \frac{1}{T_i} \cdot \frac{\alpha}{p_i}$$

FAST $$x_i = \frac{\alpha}{p_i}$$

α = 1.225 (Reno), 0.120 (HSTCP), 0.075 (STCP)

Network model



$$R_{li} = 1 \quad \text{if source } i \text{ uses link } l$$ ← IP routing

$$x(t+1) = F\,(R^T p(t),\ x(t))$$ ← Reno, Vegas

$$p(t+1) = G\,(p(t),\ Rx(t))$$ ← DT, RED, ...

Network model: example

Reno:
Jacobson
1989

```
for every RTT
{   W += 1   }          (AI)
for every loss
{   W := W/2   }        (MD)
```

$$x_i(t+1) = \frac{1}{T_i^2} - \frac{x_i^2}{2} \sum_l R_{li}\, p_l(t)$$

$$p_l(t+1) = G_l\left(\sum_i R_{li}\, x_i(t),\ p_l(t) \right)$$

AI → $\frac{1}{T_i^2}$ term

MD → $\frac{x_i^2}{2} \sum_l R_{li}\, p_l(t)$ term

TailDrop → G_l

Network model: example

FAST:

Jin, Wei, Low 2004
Wei, Jin, Low, Hegde 2007

```
periodically
{
```

$$W := \frac{\text{baseRTT}}{\text{RTT}} W + \alpha$$

```
}
```

$$x_i(t+1) = x_i(t) + \frac{\gamma_i}{T_i}\left(\alpha_i - x_i(t)\sum_l R_{li}\, p_l(t)\right)$$

$$p_l(t+1) = p_l(t) + \frac{1}{c_l}\left(\sum_i R_{li}\, x_i(t) - c_l\right)$$

Reverse engineering

Protocol (Reno, Vegas, RED, REM/PI...)

$$x(t+1) = F\,(p(t),\; x(t))$$
$$p(t+1) = G\,(p(t),\; x(t))$$

Equilibrium

- Performance
  - Throughput, loss, delay
- Fairness
- Utility

Dynamics

- Local stability
- Global stability

Duality model of TCP/AQM

- TCP/AQM $\quad x^* = F(R^T p^*, x^*)$

$$p^* = G(p^*, Rx^*)$$

- Equilibrium (x^*, p^*) primal-dual optimal:

$$\max_{x \geq 0} \sum U_i(x_i) \qquad \text{subject to} \quad Rx \leq c$$

  - F determines utility function U
  - G guarantees complementary slackness
  - p^* are Lagrange multipliers

Kelly, Maloo, Tan 1998
Low, Lapsley 1999

Uniqueness of equilibrium

- x^* is unique when U is strictly concave
- p^* is unique when R has full row rank

Duality model of TCP/AQM

- TCP/AQM

$$x^* = F(R^T p^*, x^*)$$
$$p^* = G(p^*, Rx^*)$$

- Equilibrium (x^*, p^*) primal-dual optimal:

$$\max_{x \geq 0} \sum U_i(x_i) \quad \text{subject to} \quad Rx \leq c$$

  - F determines utility function U
  - G guarantees complementary slackness
  - p^* are Lagrange multipliers

Kelly, Maloo, Tan 1998
Low, Lapsley 1999

The underlying concave program also leads to simple dynamic behavior

Reverse engineering TCP

- Equilibrium *(x*,p*)* primal-dual optimal:

$$\max_{x\geq 0} \sum U_i(x_i) \qquad \text{subject to} \quad Rx \leq c$$

Mo & Walrand 2000:

$$U_i(x_i) = \begin{cases} \log x_i & \text{if} \quad \alpha = 1 \\ (1-\alpha)^{-1} x_i^{1-\alpha} & \text{if} \quad \alpha \neq 1 \end{cases}$$

- $\alpha = 1$: Vegas, FAST, STCP
- $\alpha = 1.2$: HSTCP
- $\alpha = 2$: Reno
- $\alpha = \infty$: XCP (single link only)

Reverse engineering TCP

- Equilibrium (x^*, p^*) primal-dual optimal:

$$\max_{x \geq 0} \sum U_i(x_i) \qquad \text{subject to} \quad Rx \leq c$$

Mo & Walrand 2000:

$$U_i(x_i) = \begin{cases} \log x_i & \text{if} \quad \alpha = 1 \\ (1-\alpha)^{-1} x_i^{1-\alpha} & \text{if} \quad \alpha \neq 1 \end{cases}$$

- $\alpha = 0$: maximum throughput
- $\alpha = 1$: proportional fairness
- $\alpha = 2$: min delay fairness
- $\alpha = \infty$: maxmin fairness

Some implications

- Equilibrium
  - Always exists, unique if R is full rank
  - Bandwidth allocation independent of AQM or arrival
  - Can predict macroscopic behavior of large scale networks

- Counter-intuitive throughput behavior
  - Fair allocation is not always inefficient
  - Increasing link capacities do not always raise aggregate throughput

[Tang, Wang, Low, ToN 2006]

- FAST TCP
  - Design, analysis, experiments

[Wei, Jin, Low, Hegde ToN 2006]

Validation

	Source 1	Source 3	Source 5
RTT (ms)	17.1 (17)	21.9 (22)	41.9 (42)
Rate (pkts/s)	1205 (1200)	1228 (1200)	1161 (1200)
Window (pkts)	20.5 (20.4)	27 (26.4)	49.8 (50.4)
Avg backlog (pkts)	9.8 (10)		

measured → 9.8 ; theory → (10)

- Single link, capacity = 6 pkts/ms
- 5 sources with different propagation delays, α_s = 2 pkts/RTT

[Low, Peterson, Wang, JACM 2002]

Persistent congestion

- Vegas exploits buffer process to compute prices (queueing delays)
- Persistent congestion due to
  - Coupling of buffer & price
  - Error in propagation delay estimation
- Consequences
  - Excessive backlog
  - Unfairness to older sources

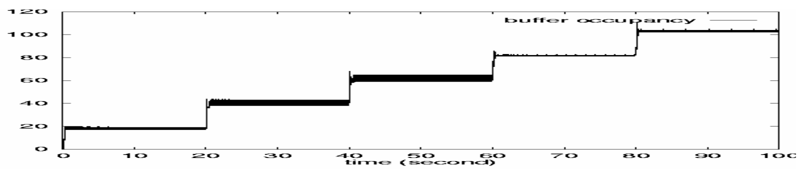
Theorem

A relative error of ε_s in propagation delay estimation distorts the utility function to

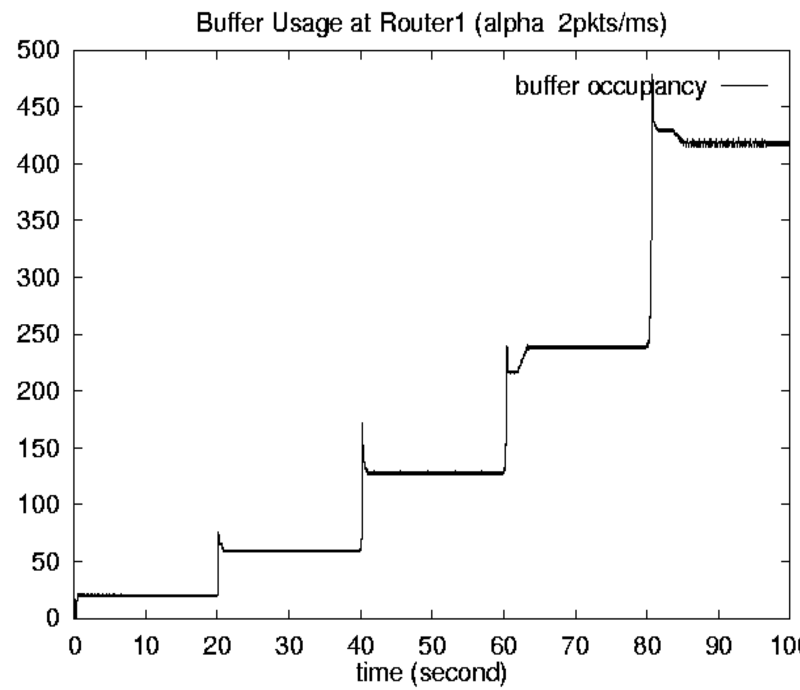
$$\hat{U}_s(x_s) = (1+\varepsilon_s)\alpha_s d_s \log x_s + \varepsilon_s d_s x_s$$

[Low, Peterson, Wang, JACM 2002]

Evidence



Without estimation error



With estimation error

- Single link, capacity = 6 pkt/ms, α_s = 2 pkts/ms, d_s = 10 ms
- With finite buffer: Vegas reverts to Reno

[Low, Peterson, Wang, JACM 2002]

Evidence

Source rates (pkts/ms)

#	src1	src2	src3	src4	src5
1	5.98 (6)				
2	2.05 (2)	3.92 (4)			
3	0.96 (0.94)	1.46 (1.49)	3.54 (3.57)		
4	0.51 (0.50)	0.72 (0.73)	1.34 (1.35)	3.38 (3.39)	
5	0.29 (0.29)	0.40 (0.40)	0.68 (0.67)	1.30 (1.30)	3.28 (3.34)

#	queue (pkts)	baseRTT (ms)
1	19.8 (20)	10.18 (10.18)
2	59.0 (60)	13.36 (13.51)
3	127.3 (127)	20.17 (20.28)
4	237.5 (238)	31.50 (31.50)
5	416.3 (416)	49.86 (49.80)

[Low, Peterson, Wang, JACM 2002]

Outline

- Duality model of TCP
  - Theory: equilibrium point characterized by an optimization problem
  - Experimental validation: Vegas
- An accurate link model
  - Theory: a new joint link model
  - Application: FAST stability [Tang, Jacobsson, Andrew, Low, Infocom 07]
- Heterogeneous protocols
  - Motivatoin: FAST+Reno
  - Theory: multiple equilibria, global uniqueness

FAST TCP

FAST:

Jin, Wei, Low
2004

```
periodically
{
```

$$W := \gamma\left(\frac{\text{baseRTT}}{\text{RTT}}W + \alpha\right) + (1-\gamma)W$$

```
}
```

$$\dot{w}_i = -\gamma \frac{q_i(t)}{\left(d_i + q_i(t)\right)^2} w_i(t) + \gamma \frac{\alpha_i}{d_i + q_i(t)}$$

$$q_i(t) = p(t - \tau_i^b)$$ ← Single Link

Link model 1: integrator model

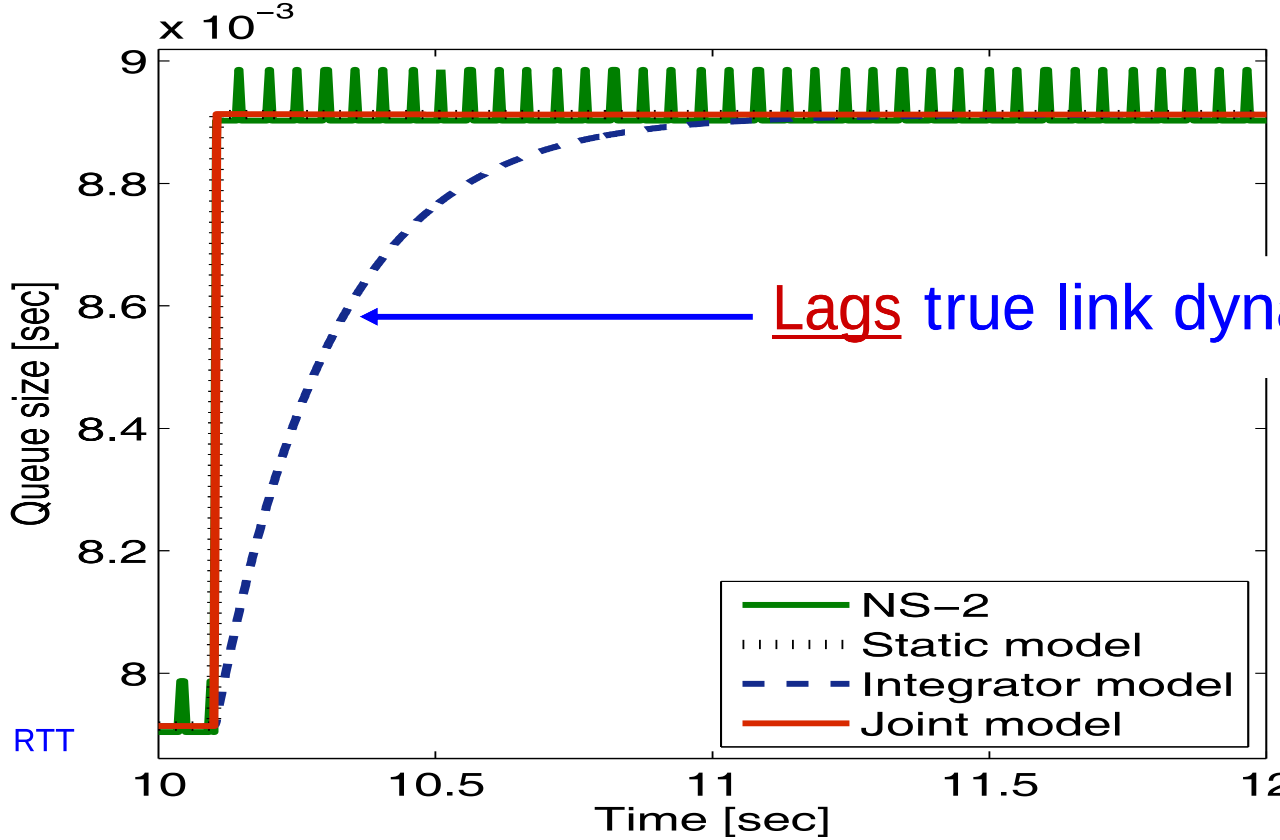
aggregate FAST rate

$$\dot{p} = \frac{1}{c}\left(\sum_i \frac{w_i(t - \tau_i^f)}{d_i + p(t)} + x_0(t) - c \right)$$

cross traffic rate

Link model 1: integrator model

$$\dot{p} = \frac{1}{c}\left(\sum_i \frac{w_i(t-\tau_i^f)}{d_i + p(t)} + x_0(t) - c \right)$$



Lags true link dynamics

Same RTT

Link model 2: static model

D. Wei, 2003:

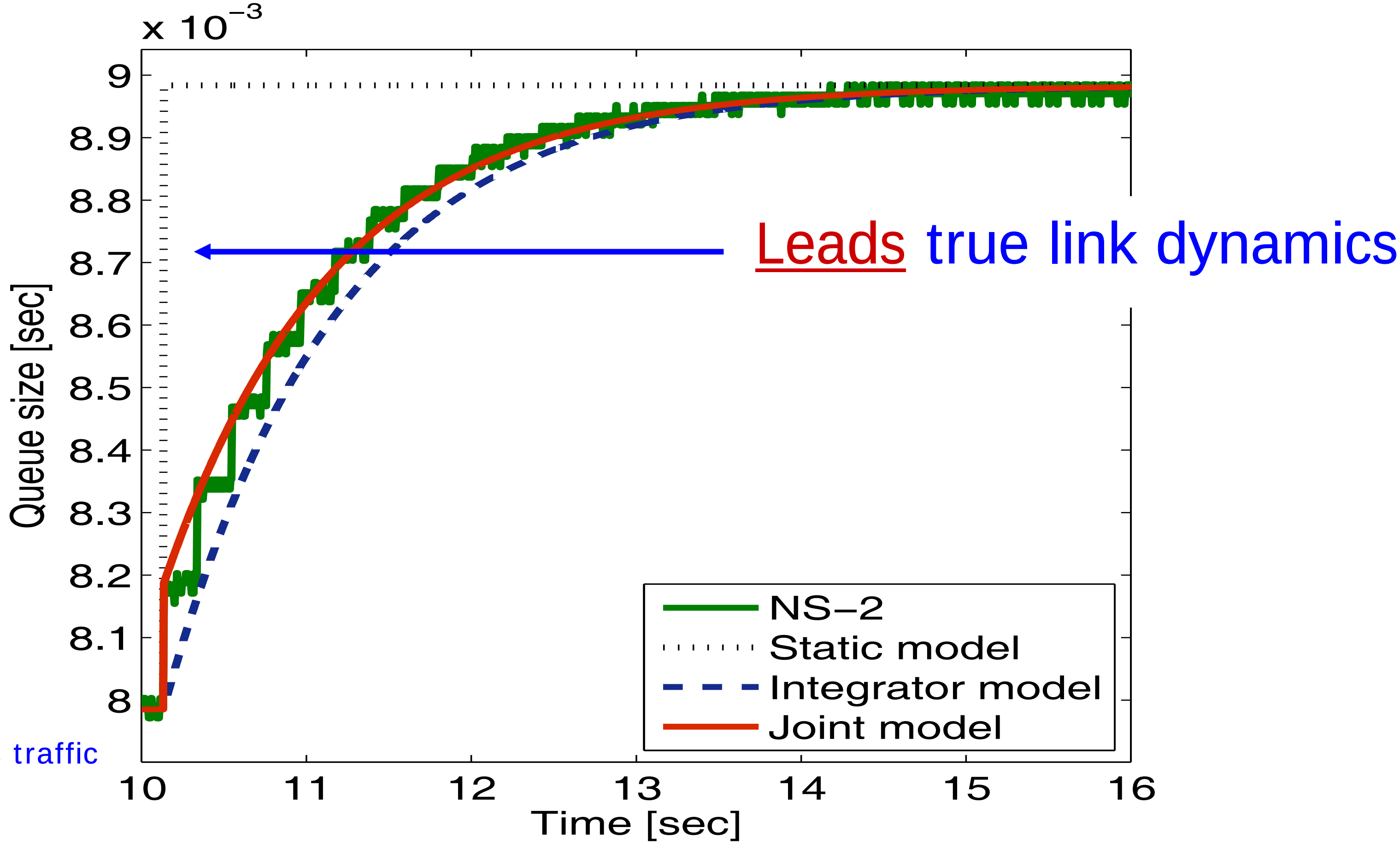
$$\sum_i \frac{w_i(t-\tau_i^f)}{d_i + p(t)} + x_0(t) = c$$

Motivations

- Ack-clocking: input rate = capacity after 1 RTT
- Fast link dynamics

Link model 2: static model

$$\sum_i \frac{w_i(t-\tau_i^f)}{d_i + p(t)} + x_0(t) = c$$



Cross traffic

Link model 3: joint model

K. Jacobsson
etc, 2006:

$$\dot{p} = \frac{1}{c}\left[\left(\sum_i \frac{w_i(t-\tau_i^f)}{d_i + p(t)} + \dot{w}_i(t-\tau_i^f)\right) + x_0(t) - c\right]$$

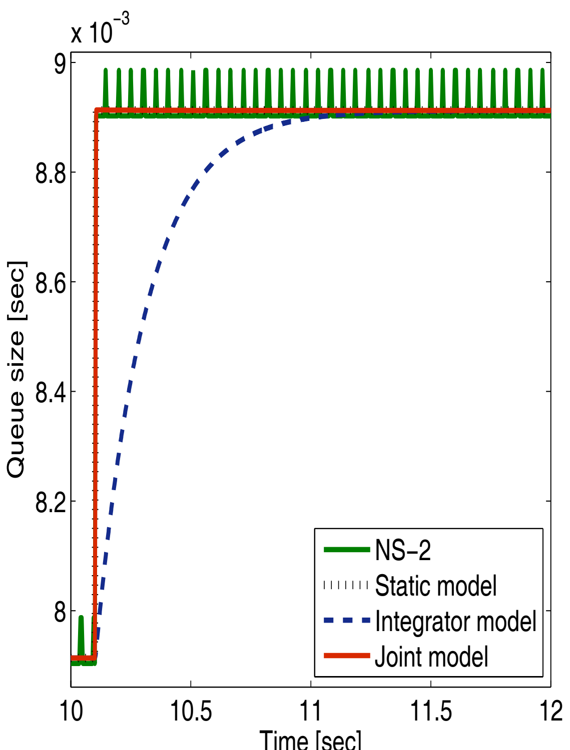
$\dot{w}_i(t-\tau_i^f) = 0$: Reduces to integrator model

<u>and</u> $\dot{p} = 0$: Reduces to static model

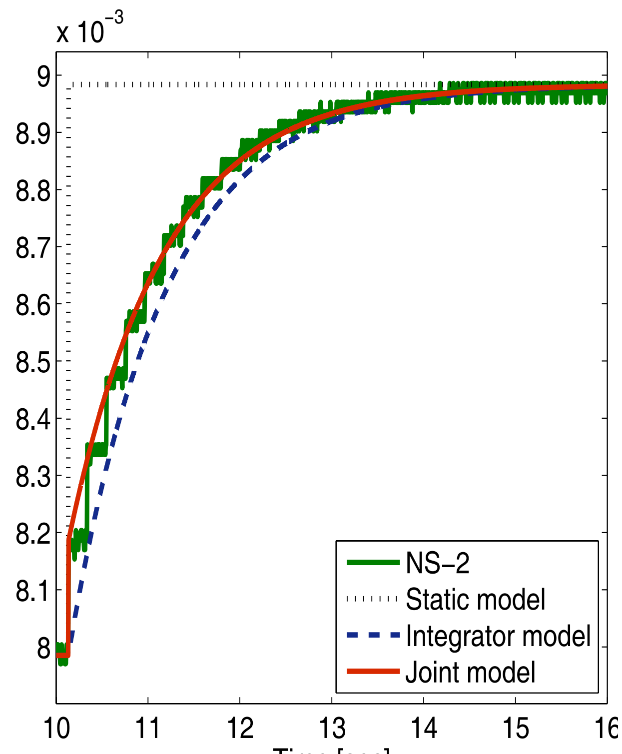
Link model 3: joint model

$$\dot{p} = \frac{1}{c}\left[\left(\sum_i \frac{w_i(t-\tau_i^f)}{d_i + p(t)} + \dot{w}_i(t-\tau_i^f)\right) + x_0(t) - c\right]$$

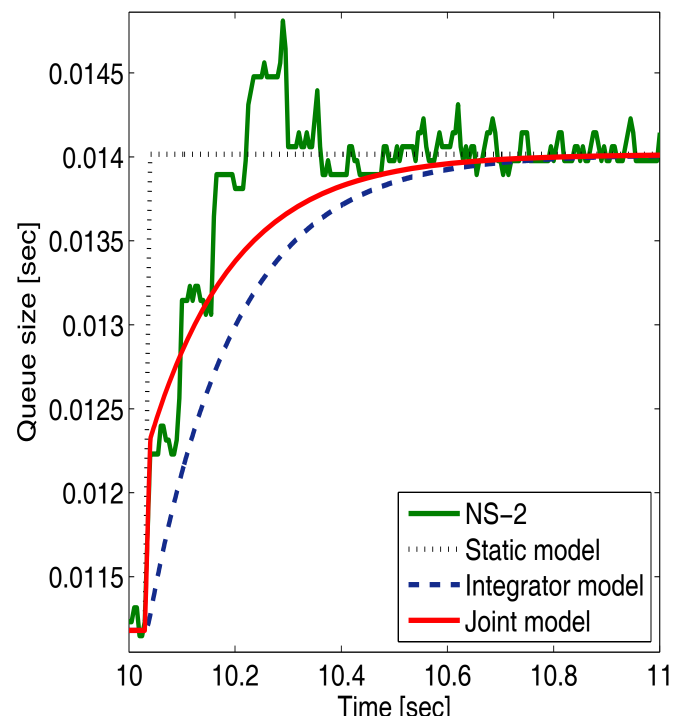
Same RTT, no cross traffic



Same RTT, cross traffic



Different RTTs, no cross traffic



FAST TCP

Source model:

$$\dot{w}_i = -\gamma \frac{q_i(t)}{\left(d_i + q_i(t)\right)^2} w_i(t) + \gamma \frac{\alpha_i}{d_i + q_i(t)}$$

$$q_i(t) = p(t - \tau_i^b)$$ ← Single Link

Link (joint) model:

$$\dot{p} = \frac{1}{c}\left[\left(\sum_i \frac{w_i(t - \tau_i^f)}{d_i + p(t)} + \dot{w}_i(t - \tau_i^f)\right) + x_0(t) - c\right]$$

FAST TCP: linear stability

Theorem

FAST TCP is linearly stable for *arbitrary* delay provided

$$\gamma < 0.94$$

Resolves a major discrepancy between previous predictions and empirical experience

FAST TCP: linearized model

Loop gain:

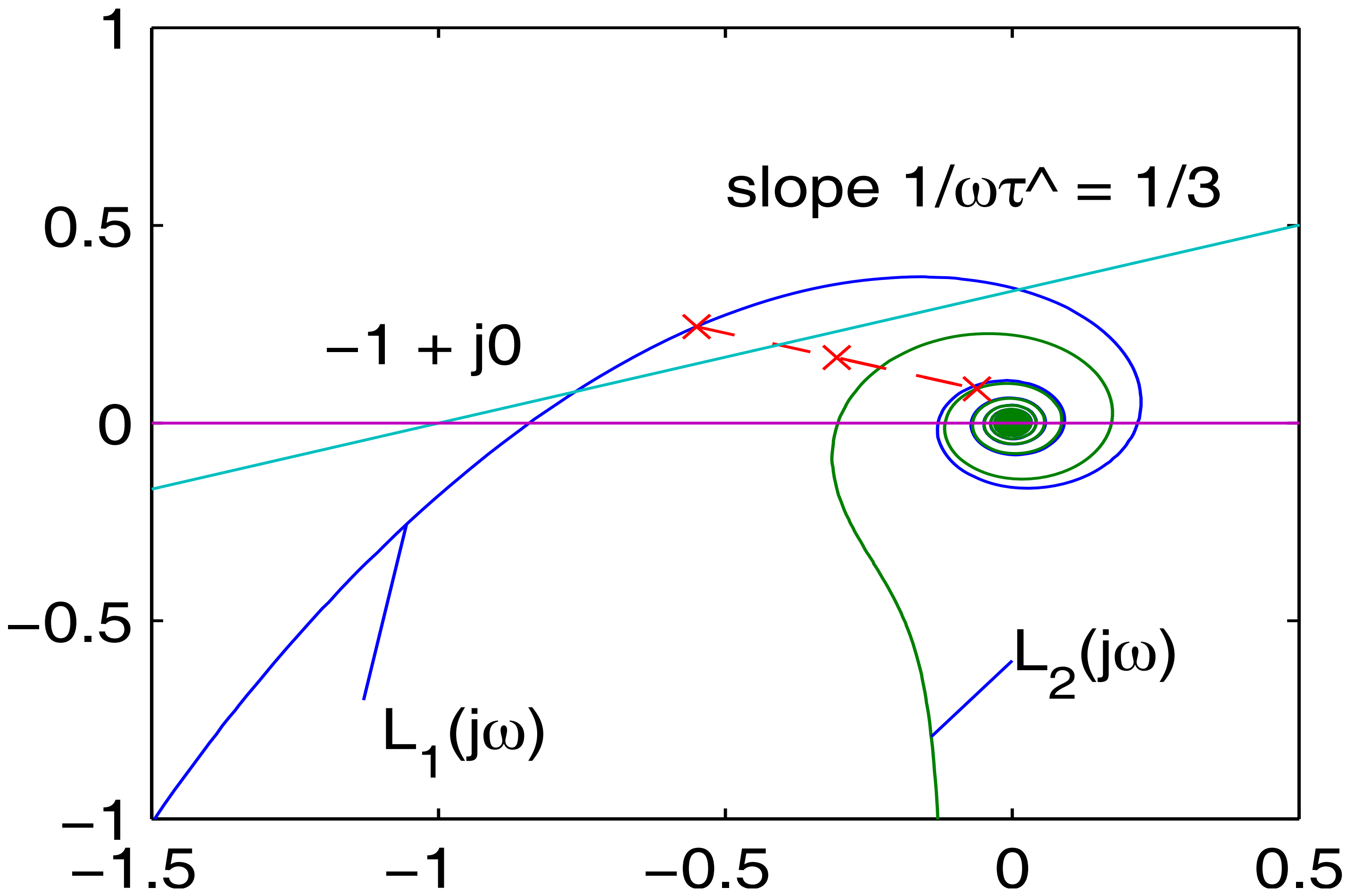
$$L(s) = \sum_i \mu_i L_i(s)$$

$$L_i(s) = \frac{s + \frac{1}{\tau_i}}{s + \frac{1}{\hat{\tau}}} \cdot \frac{\gamma d_i e^{-\tau_i s}}{\tau_i^2 s + \gamma q}$$

$$\mu_i = \frac{\alpha_i}{c \sum_n \alpha_n} \qquad \frac{1}{\hat{\tau}} = \sum_i \mu_i \frac{1}{\tau_i}$$

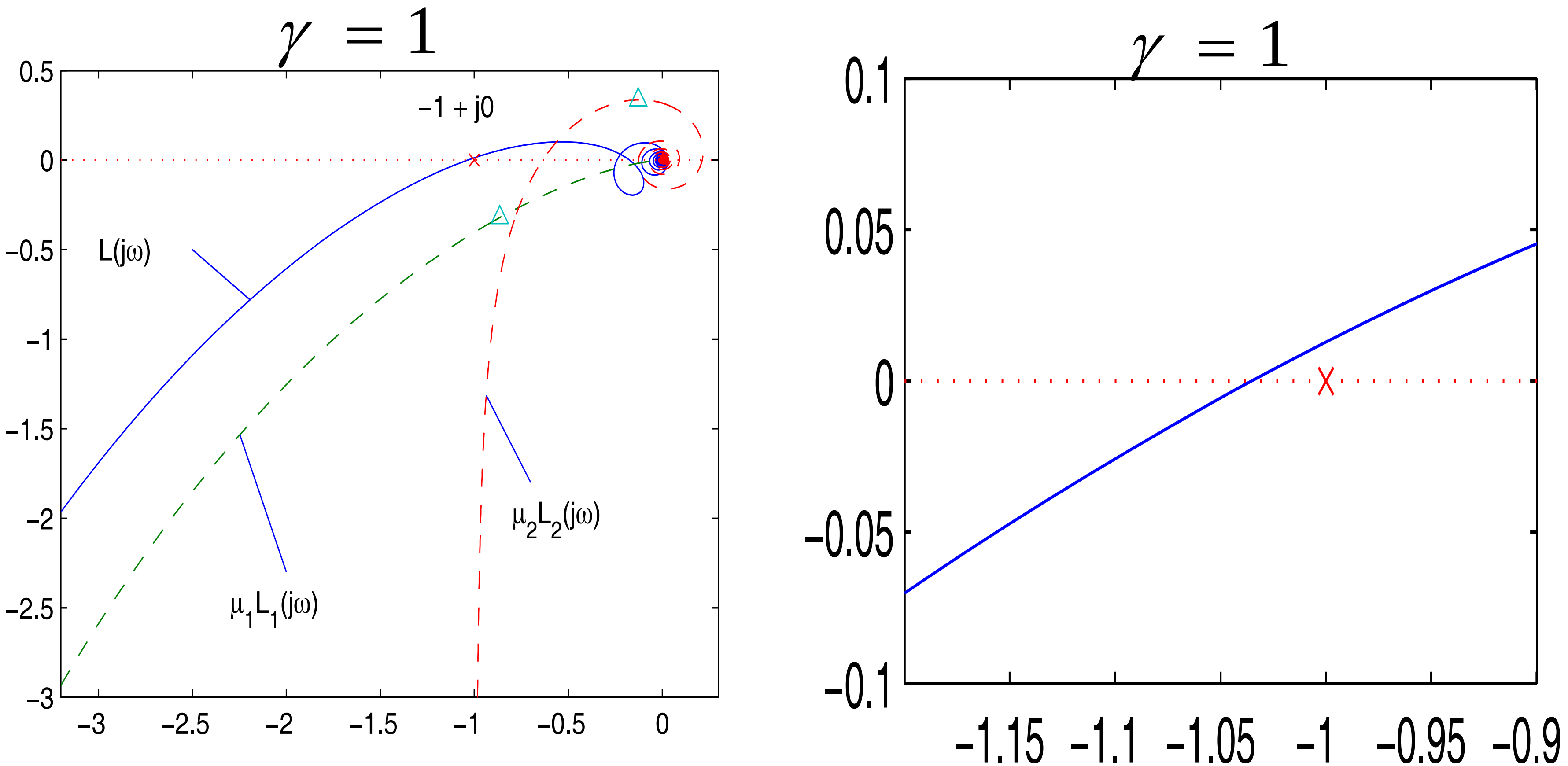
Nyquist stability analysis

$$L(j\omega) = \sum_i \mu_i L_i(j\omega)$$



Stability condition can be "tight"

Linearly stable if $\gamma < 0.94$



Comparison of 3 link models

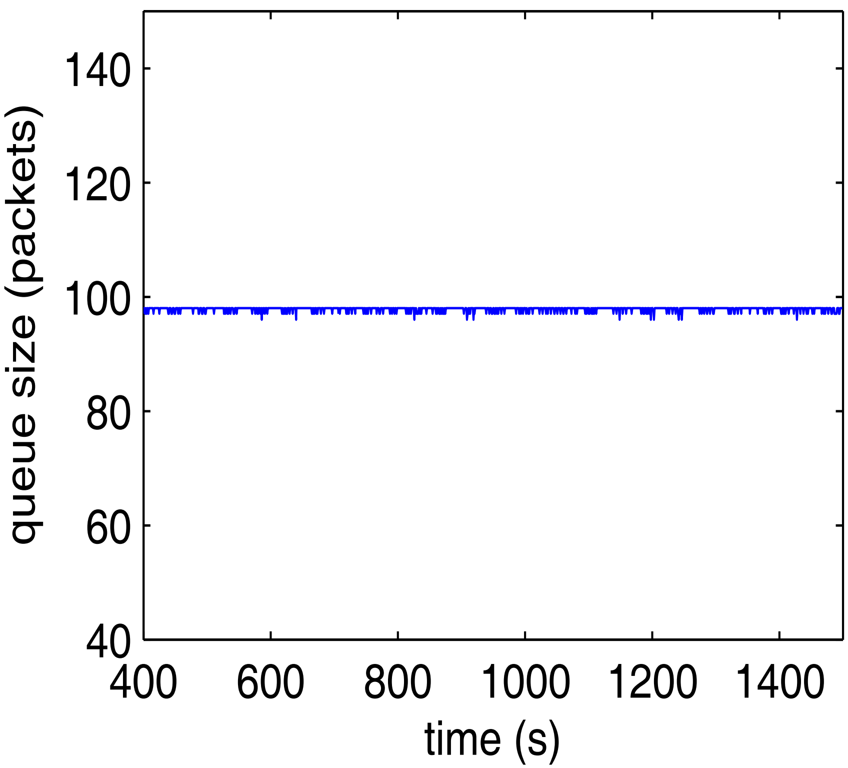
- Single link with capacity 10,000 pkts/s
- Propagation delays: 400ms, 700ms
- α = 50 pkts

Critical step size

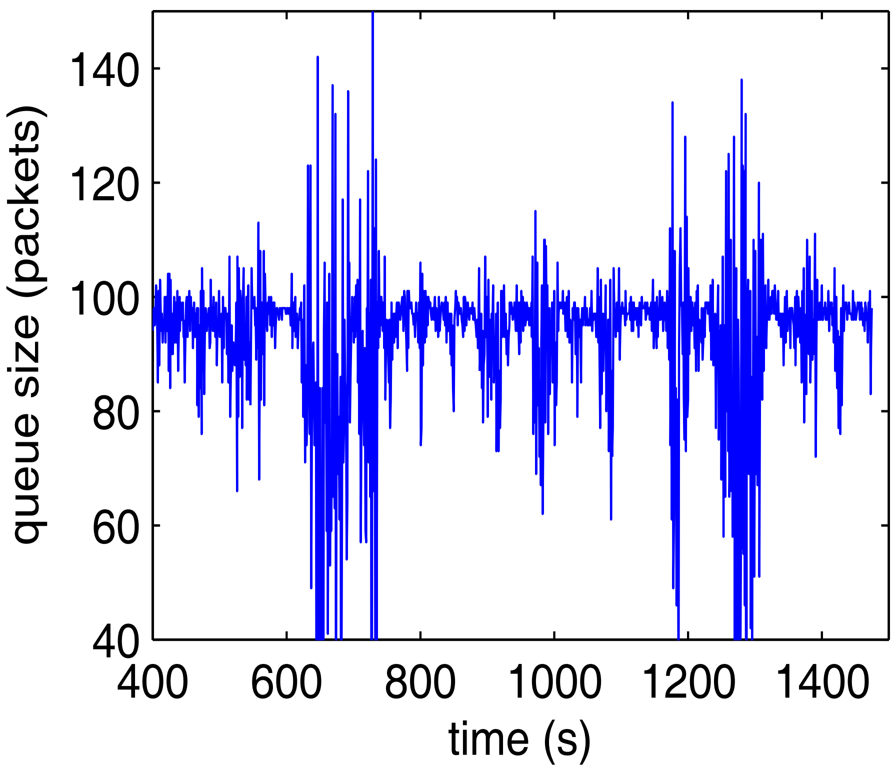
- Integrator model: 1.23
- Static model: 1.80
- Joint model: 1.69

Comparison of 3 link models

$\gamma = 1.23$



$\gamma = 1.80$

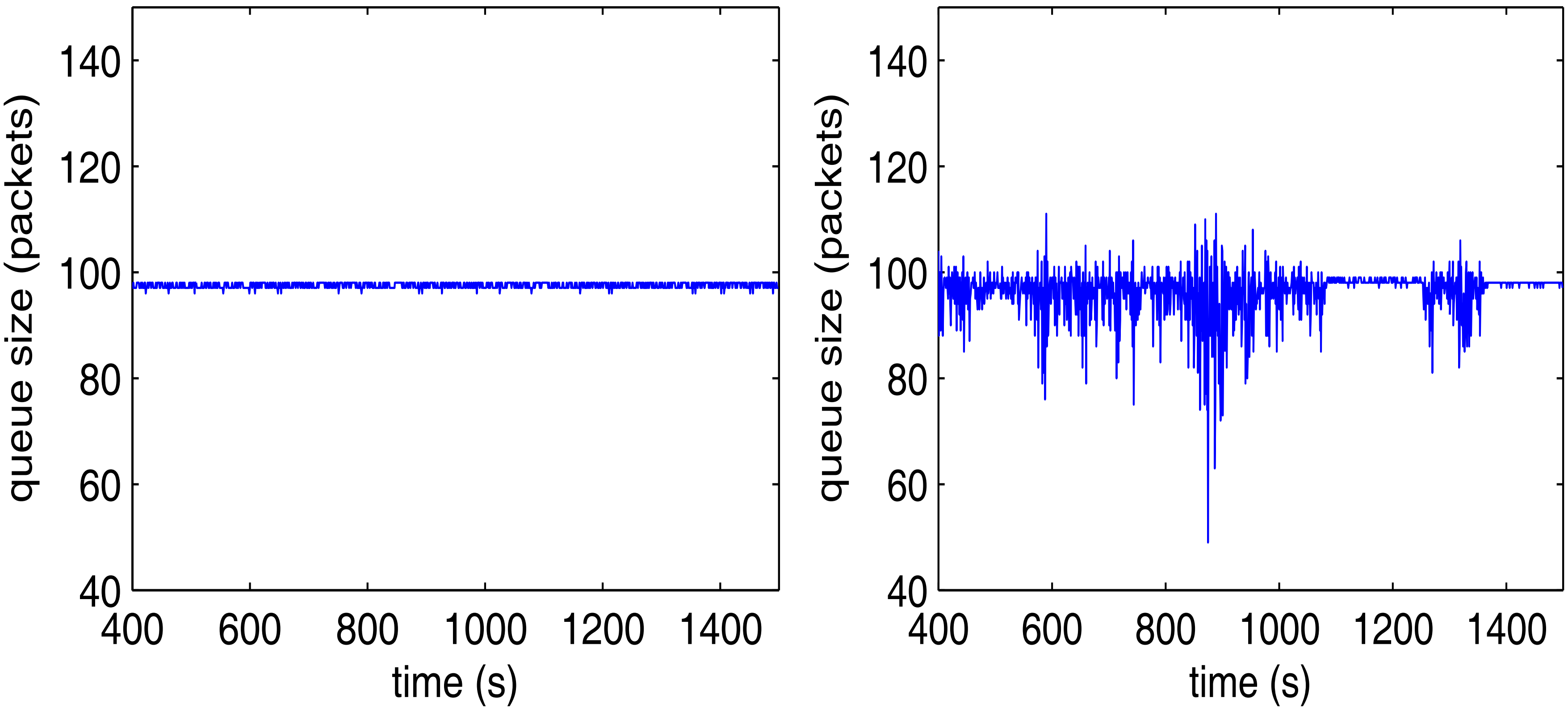


Integrator model too conservative

Static model too aggressive

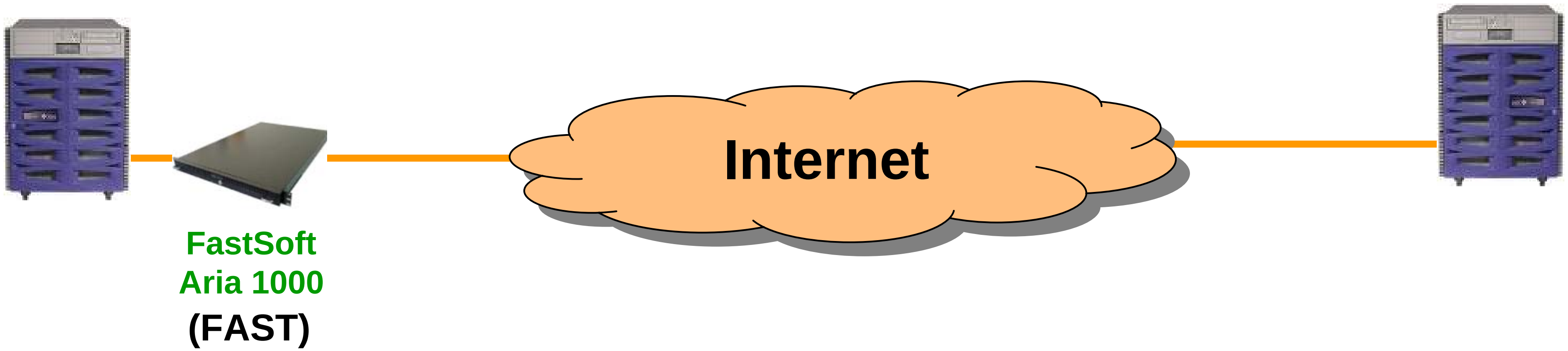
Comparison of 3 link models

$\gamma = 1.65$ $\gamma = 1.75$

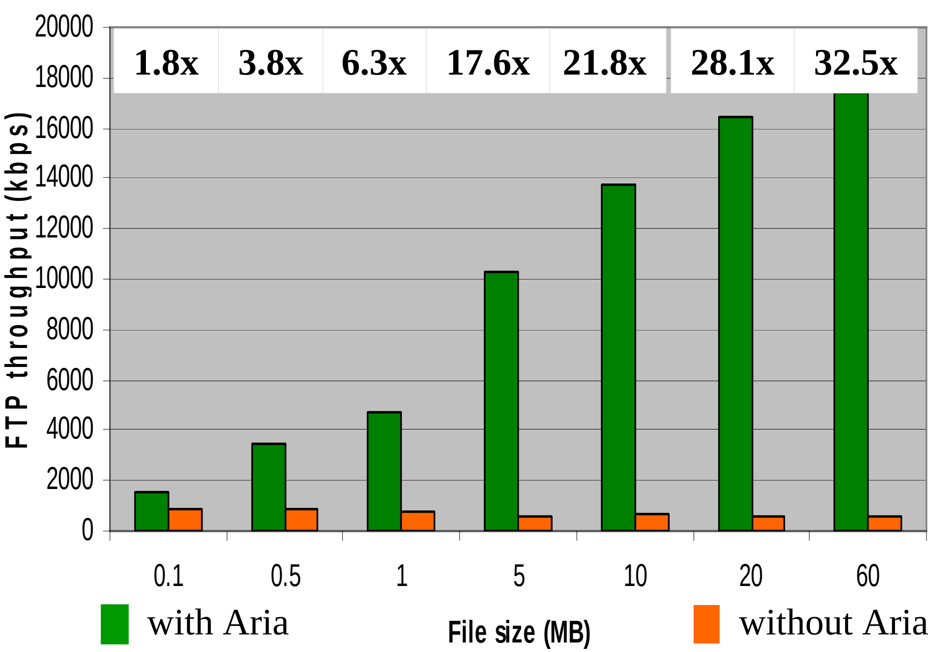


Joint model prediction: $\gamma < 1.69$

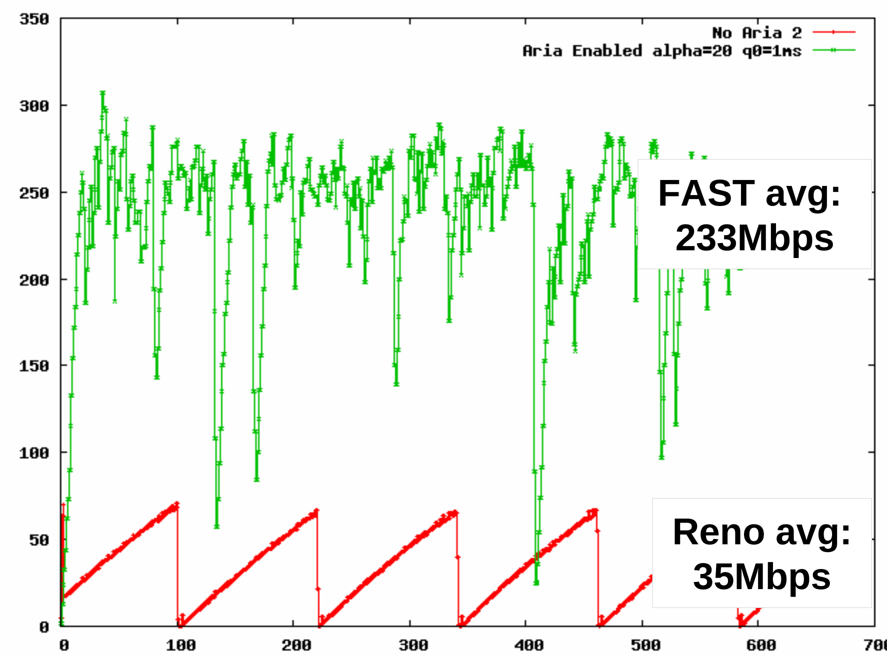
Commercial Deployment: FAST in a box



Throuput: LA → Tokyo



Throuput: San Fran → MIT



Outline

- Duality model of TCP
  - Theory: equilibrium point characterized by an optimization problem
  - Experimental validation: Vegas
- An accurate link model
  - Theory: a new joint link model
  - Application: FAST stability
- Heterogeneous protocols
  - Motivatoin: FAST+Reno
  - Theory: multiple equilibria, global uniqueness

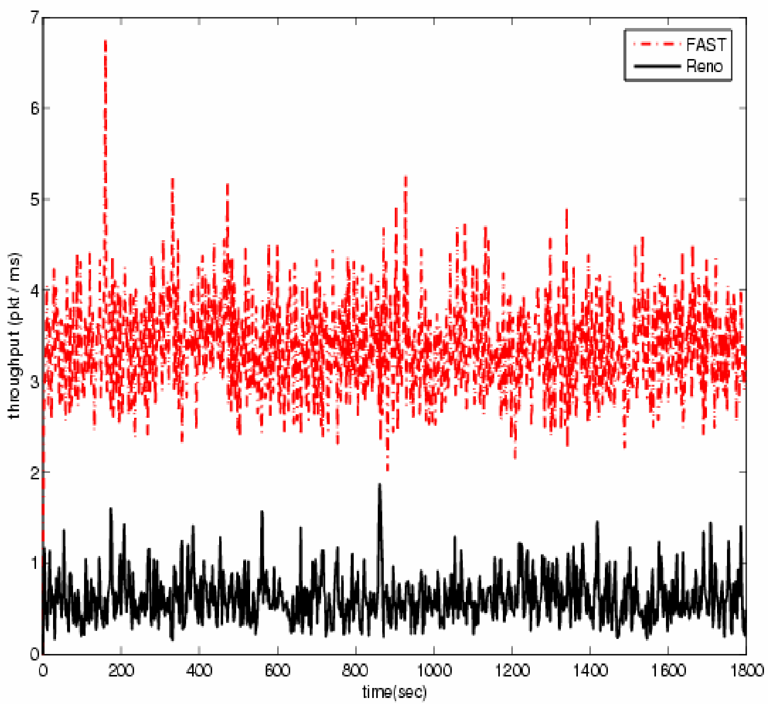
[Tang, Wang, Low, Chiang, ToN 2007]

[Tang, Wang, Hegde, Low, Comp Networks, 2005]

The world is heterogeneous...

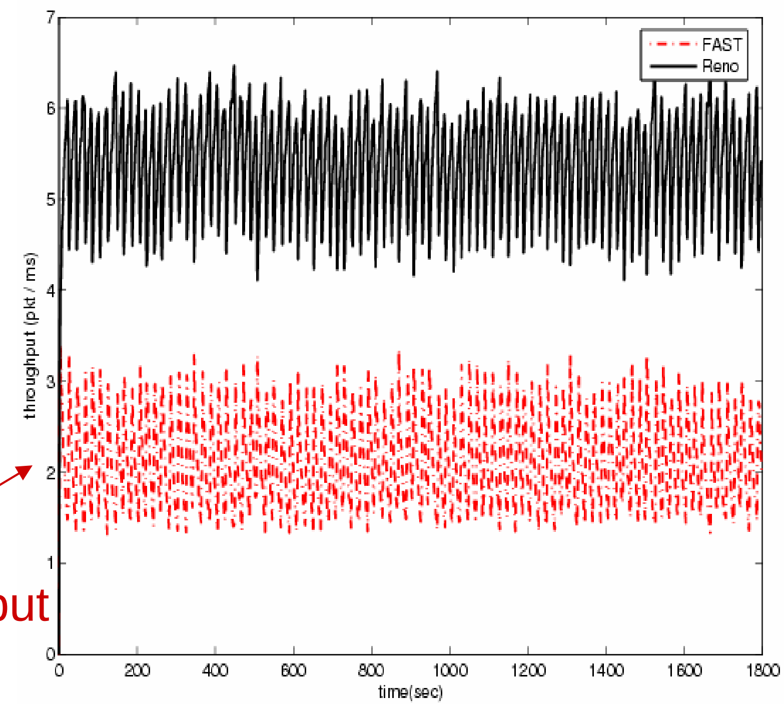
- Linux 2.6.13 allows users to choose congestion control algorithms
- Many protocol proposals
  - Loss-based: Reno and a large number of variants
  - Delay-based: CARD (1989), DUAL (1992), Vegas (1995), FAST (2004), ...
  - ECN: RED (1993), REM (2001), PI (2002), AVQ (2003), ...
  - Explicit feedback: MaxNet (2002), XCP (2002), RCP (2005), ...

Throughputs depend on AQM



buffer size = 80 pkts

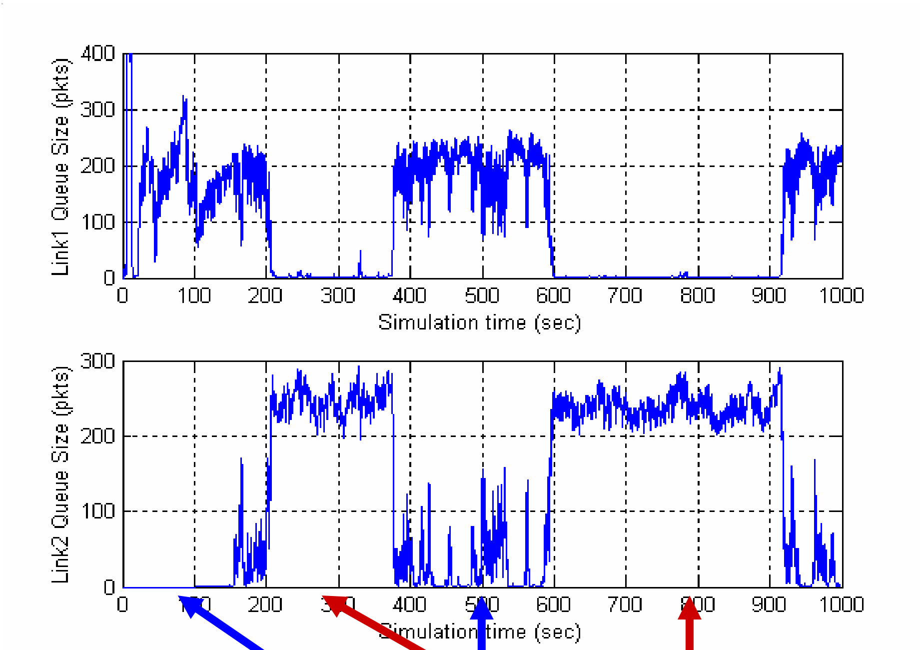
FAST throughput

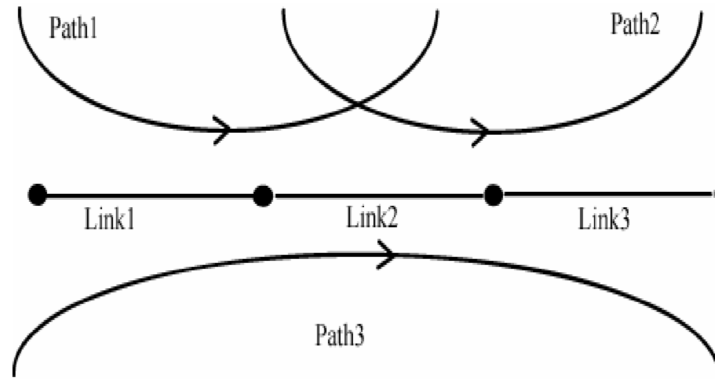


buffer size = 400 pkts

- FAST and Reno share a single bottleneck router
- NS2 simulation
- Router: DropTail with variable buffer size
- With 10% heavy-tailed noise traffic

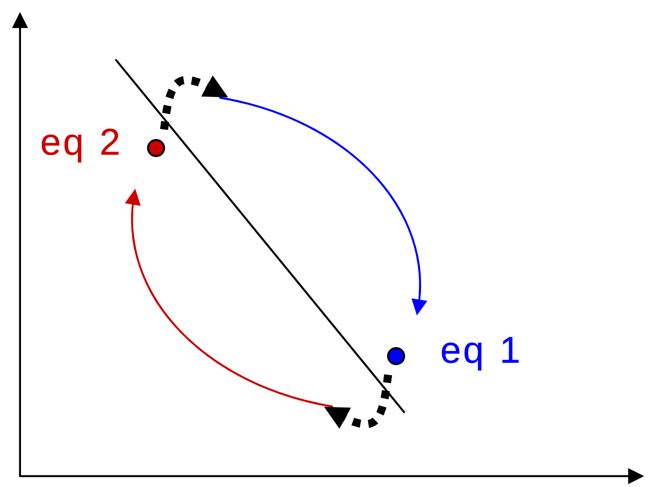
Multiple equilibria: throughput depends on arrival





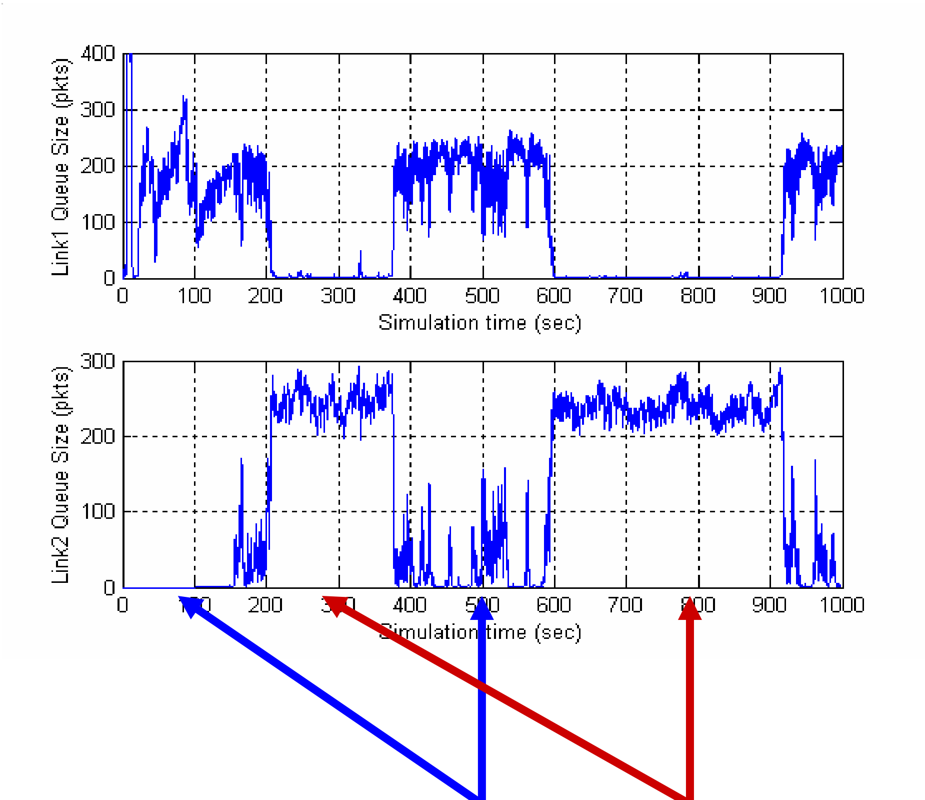
Dummynet experiment

	eq 1	eq 2
Path 1	52M	13M
path 2	61M	13M
path 3	27M	93M

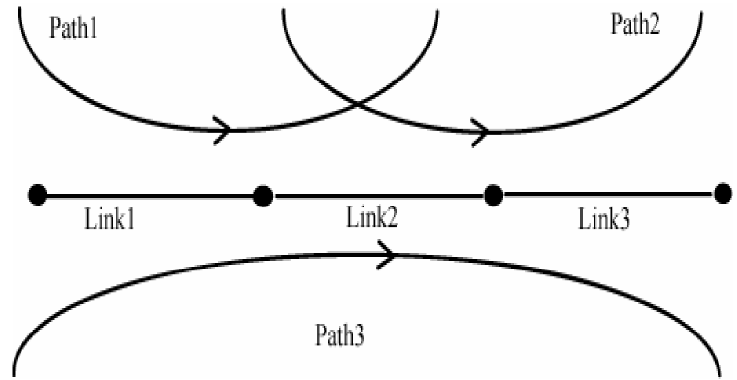


Tang, Wang, Hegde, Low, Telecom Systems, 2005

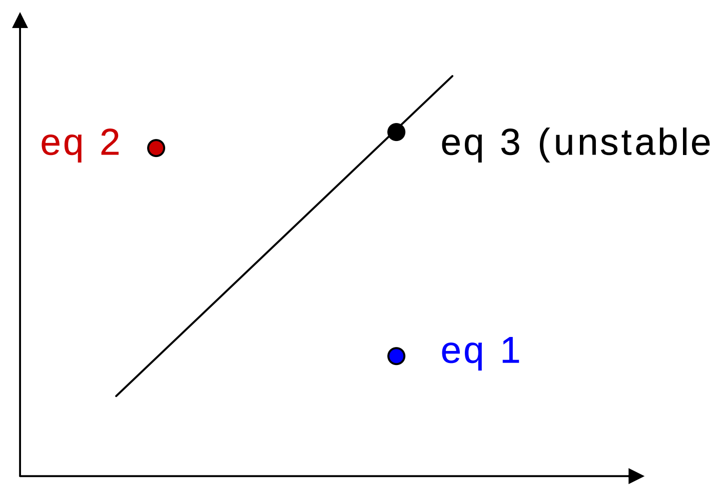
Multiple equilibria: throughput depends on arrival



	eq 1	eq 2
Path 1	52M	13M
path 2	61M	13M
path 3	27M	93M



Dummynet experiment



Tang, Wang, Hegde, Low, Telecom Systems, 2005

Some implications

	homogeneous	**heterogeneous**
equilibrium	unique	non-unique
bandwidth allocation on AQM	independent	dependent
bandwidth allocation on arrival	independent	dependent

- Duality model:

$$\max_{x \geq 0} \sum U_i(x_i) \quad \text{s.t. } Rx \leq c \qquad\qquad x_i^* = F_i\left(\sum_l R_{li} p_l^*, x_i^*\right)$$

- Why can't use F_i's of FAST and Reno in duality model?

They use different prices!

$$F_i = x_i + \frac{\gamma_i}{T_i}\left(\alpha_i - x_i \sum_l R_{li} p_l\right)$$ ← delay for FAST

$$F_i = \frac{1}{T_i^2} - \frac{x_i^2}{2} \sum_l R_{li} p_l$$ ← loss for Reno

- Duality model:

$$\max_{x \ge 0} \sum U_i(x_i) \quad \text{s.t. } Rx \le c \qquad\qquad x_i^* = F_i\left(\sum_l R_{li} p_l^*, x_i^*\right)$$

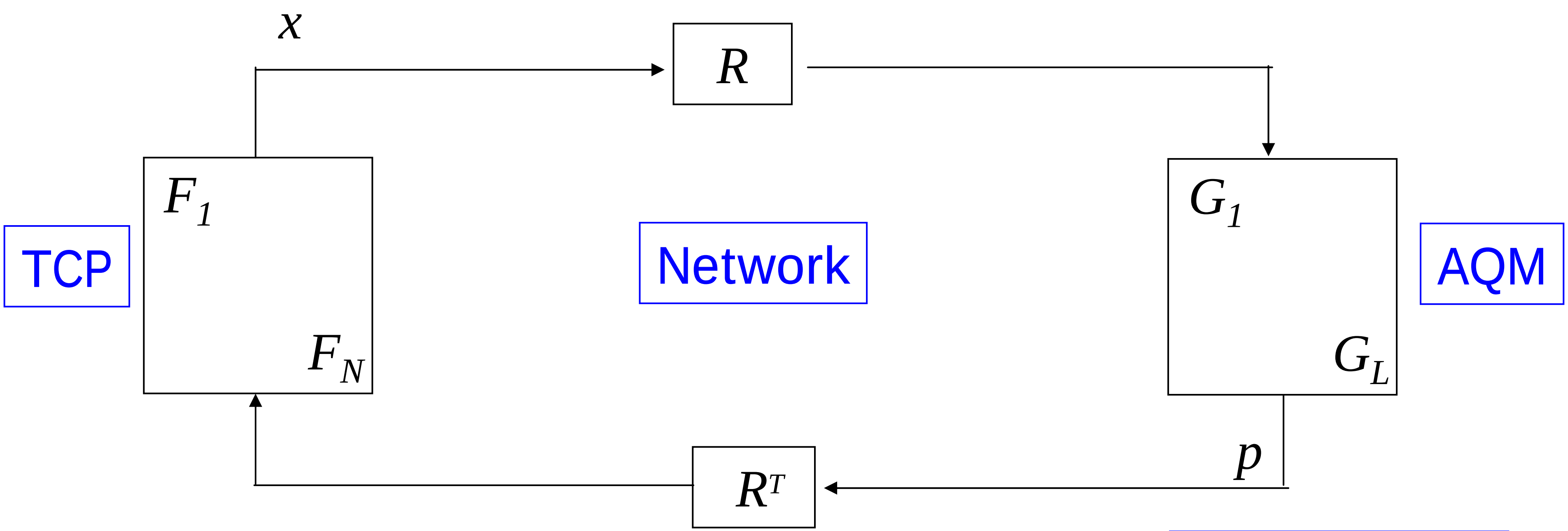
- Why can't use F_i's of FAST and Reno in duality model?

They use different prices!

$$F_i = x_i + \frac{\gamma_i}{T_i}\left(\alpha_i - x_i \sum_l R_{li} p_l\right) \qquad \leftarrow \qquad \dot{p}_l = \frac{1}{c_l}\left(\sum_i R_{li} x_i(t) - c_l\right)$$

$$F_i = \frac{1}{T_i^2} - \frac{x_i^2}{2} \sum_l R_{li} p_l \qquad \leftarrow \qquad \dot{p}_l = g_l\left(p_l(t), \sum_i R_{li} x_i(t)\right)$$

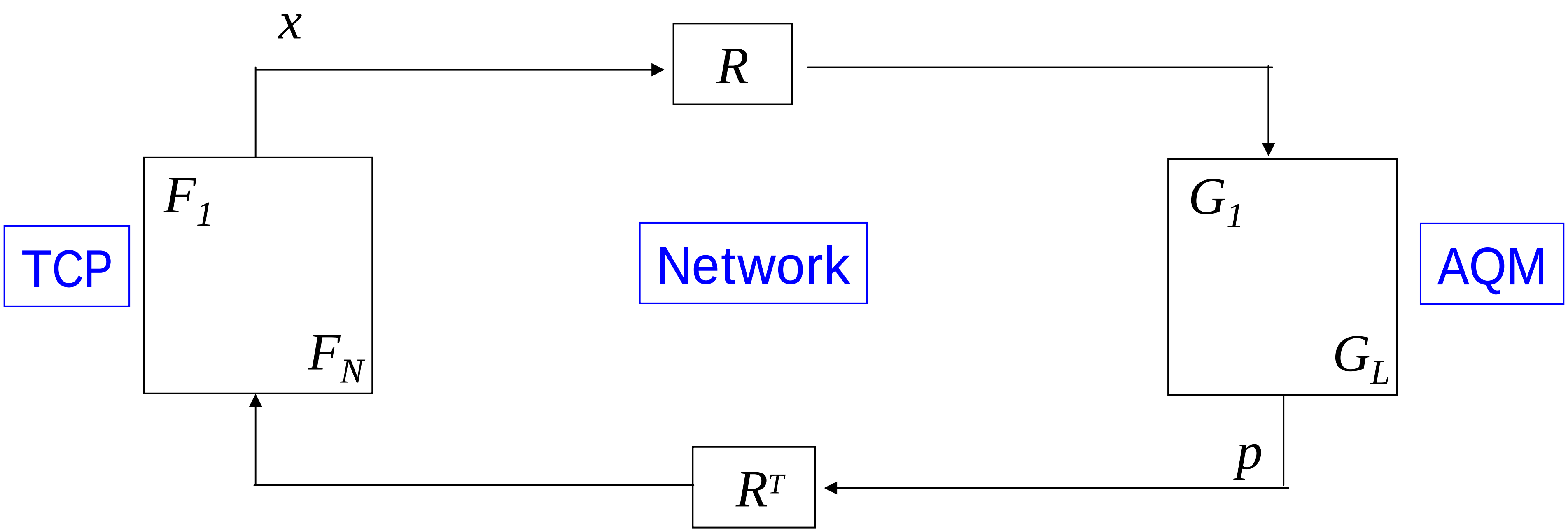
Homogeneous protocol



same price
for all sources

$$x_i(t+1) \;=\; F_i\left(\sum_l R_{li}\, p_l(t),\; x_i(t)\right)$$

Heterogeneous protocol



$$x_i(t+1) = F_i\left(\sum_l R_{li}\, p_l(t),\; x_i(t)\right)$$

$$x_i^j(t+1) = F_i^j\left(\sum_l R_{li}\, m_l^j\big(p_l(t)\big),\; x_i^j(t)\right)$$

heterogeneous prices for type *j* sources

Heterogeneous protocols

- Equilibrium: p that satisfies

$$x_i^j(p) = f_i^j\left(\sum_l R_{li} m_l^j(p_l)\right)$$

$$y_l(p) := \sum_{i,j} R_{li}^j x_i^j(p) \begin{cases} \leq c_l \\ = c_l & \text{if } p_l > 0 \end{cases}$$

Duality model no longer applies !

- p_l can no longer serve as Lagrange multiplier

Heterogeneous protocols

- Equilibrium: p that satisfies

$$x_i^j(p) = f_i^j\left(\sum_l R_{li} m_l^j(p_l)\right)$$

$$y_l(p) := \sum_{i,j} R_{li}^j x_i^j(p) \begin{cases} \le c_l \\ = c_l \quad \text{if } p_l > 0 \end{cases}$$

Need to re-examine all issues

- Equilibrium: exists? unique? efficient? fair?
- Dynamics: stable? limit cycle? chaotic?
- Practical networks: typical behavior? design guidelines?

Heterogeneous protocols

- Equilibrium: p that satisfies

$$x_i^j(p) \;=\; f_i^{\,j}\left(\sum_l R_{li} m_l^j(p_l)\right)$$

$$y_l(p) := \sum_{i,j} R_{li}^j x_i^j(p) \begin{cases} \le c_l \\ = c_l \quad \text{if} \;\; p_l > 0 \end{cases}$$

- Dynamic: dual algorithm

$$x_i^j(p(t)) \;=\; f_i^{\,j}\left(\sum_l R_{li} m_l^j(p_l(t))\right)$$

$$\dot{p}_l \;=\; \gamma_l\left(y_l(p(t)) - c_l\right)$$

Existence

<u>Theorem</u>

Equilibrium p exists, despite lack of underlying utility maximization

- Generally non-unique
  - There are networks with unique bottleneck set but infinitely many equilibria
  - There are networks with multiple bottleneck set each with a unique (but distinct) equilibrium

Regular networks

Definition

A *regular network* is a tuple (R, c, m, U) for which all equilibria p are locally unique, i.e.,

$$\det \mathbf{J}(p) := \det \frac{\partial y}{\partial p}(p) \neq 0$$

Theorem

- Almost all networks are regular
- A regular network has finitely many and odd number of equilibria (e.g. 1)

Global uniqueness

$$\dot{m}_l^j \in [a_l, 2^{1/L} a_l] \quad \text{for any } a_l > 0$$

$$\dot{m}_l^j \in [a^j, 2^{1/L} a^j] \quad \text{for any } a^j > 0$$

Theorem

- If *price heterogeneity* is small, then equilibrium is globally unique

Corollary

- If price mapping functions m_l^j are linear and link-independent, then equilibrium is globally unique

e.g. a network of RED routers with slope inversely proportional to link capacity almost always has globally unique equilibrium

Global uniqueness

$$\dot{m}_l^j \in [a_l, 2^{1/L} a_l] \quad \text{for any } a_l > 0$$
$$\dot{m}_l^j \in [a^j, 2^{1/L} a^j] \quad \text{for any } a^j > 0$$

Theorem

- If *price heterogeneity* is small, then equilibrium is globally unique

Remarks:

- Condition independent of U, R, c
- Depends on m and size L of network
- "Tight" from Index Theorem

Local stability: `uniqueness' → stability

$$\dot{m}_l^j \in [a_l, 2^{1/L} a_l] \quad \text{for any } a_l > 0$$

$$\dot{m}_l^j \in [a^j, 2^{1/L} a^j] \quad \text{for any } a^j > 0$$

Theorem

- If *price heterogeneity* is small, then the unique equilibrium p is locally stable

Linearized dual algorithm: $\delta\dot{p} = \gamma\, \mathbf{J}(p^*)\ \delta p(t)$

Equilibrium p is *locally stable* if

$$\text{Re}\, \lambda\big(\mathbf{J}(p)\big) < 0$$

Local stability: `converse'

Theorem

- If all equilibria p are locally stable, then it is globally unique

Proof idea:

- For all equilibrium p: $I(p) = (-1)^L$
- Index theorem:

$$\sum_{\text{eq } p} I(p) = (-1)^L$$

Future directions

- Dynamics of TCP
  - Global stability of networks in the presence of delay
  - Rate of convergence
  - Characterize/bound instability
- Heterogeneous congestion control protocols
  - Local and global stability in the presence of delay
  - Stability with slow-timescale control
  - Dynamic behavior in the presence of multiple equilibria
- Non-convex utility functions
  - Estimating duality gap and asymptotic behavior
  - Instability of dual algorithm as network size tends to infinity

Future directions

- TCP/IP interactions
  - Connection between duality gap and NP hardness
  - Connection between duality gap and multi-path gain
- Routing/economics interactions
  - Inter-domain routing: interplay between routing protocols and economics
  - Optimizations and games over routes, traffic demands, and pricing