Flow Control in Stochastic Networks

François Baccelli

INRIA & ENS

E-mail: `Francois.Baccelli@ens.fr`

Madison Workshop 2000

Structure of the Course

1. Lecture 1 TCP and its evolution equations;
   joint work with Dohy Hong, Sigcomm '00
2. Lecture 2 The (max,plus) algebra: Spectral properties of matrices and fixed point equations
3. Lecture 3 The (max,plus) algebra: Random media
4. Lecture 4 TCP with cross traffic;
   joint work with Thomas Bonald, Questa (1999)
5. Lecture 5 Beyond (max,plus).

Summary of Lecture 1

1. Part I Flow Control: TCP
   - Reno, Tahoe and co.
   - Stochastic Framework
   - Traffic
2. Part II TCP is max plus linear
   - Max-plus
   - Simplified representation of cross traffic
   - Exact representation of cross traffic under FIFO
3. Part III Main Questions

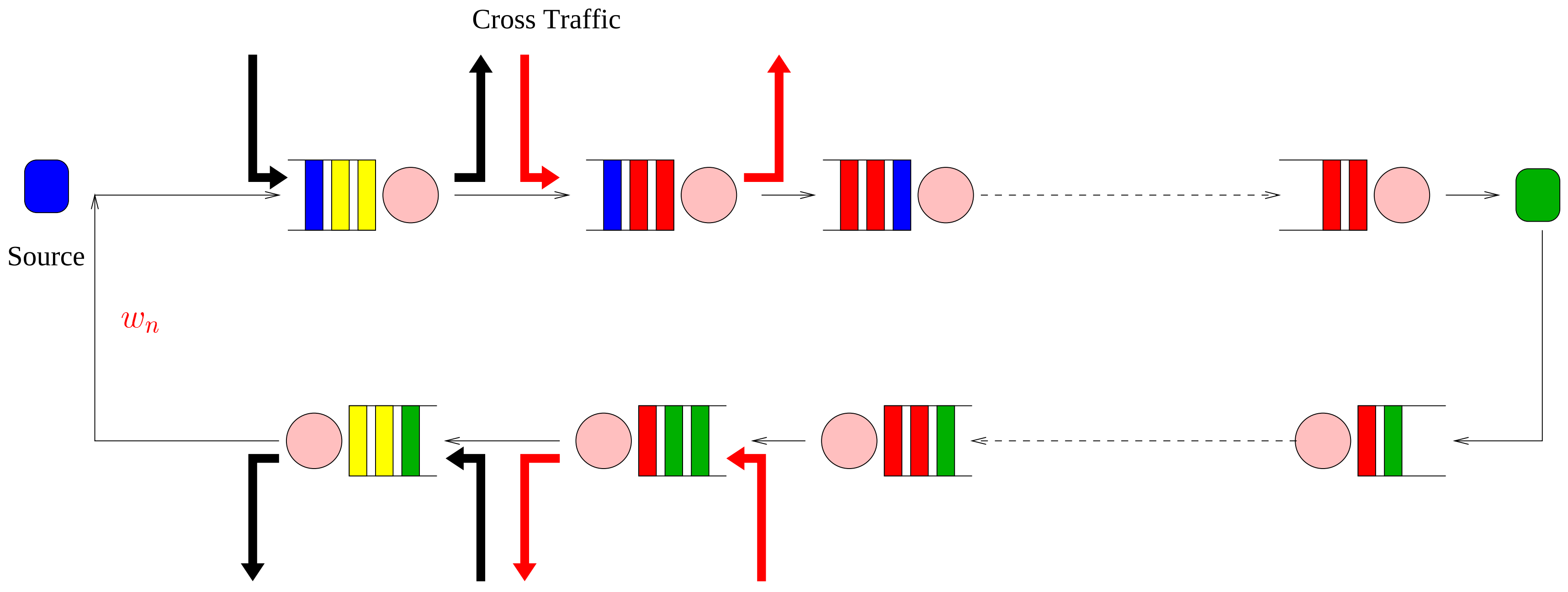
Part I: TCP and its Averaged Throughput

End-to-end control of packet transmission

"TCP over the Internet"

Error control: Each packet received by the destination is acknowledged;

Flow control: The number of unacknowledged packets in transit in the network is limited to a maximal value W called the *window*.



packets o $\to$ d **cross trafic** **acks d $\to$ o** **cross trafic**

Traffic: FTP and HTTP Examples

- Traffic of FTP a user in session : saturated source
- Potential Traffic of a HTTP user in session : on off source

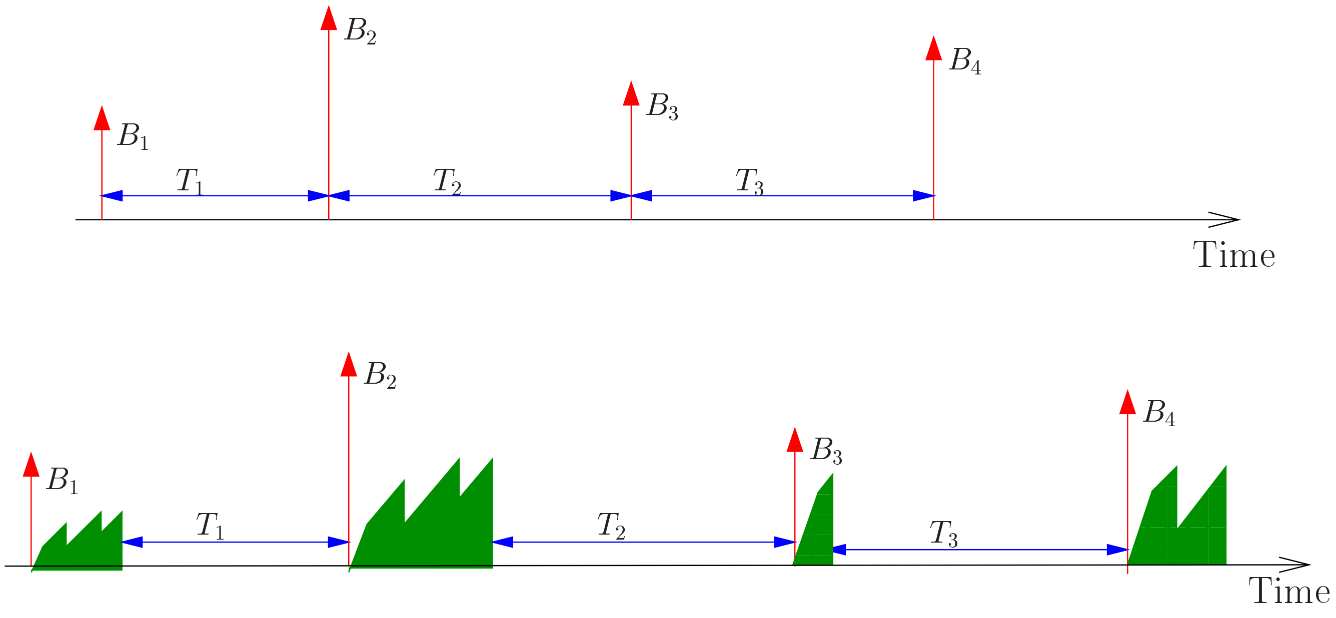


Figure 1: Generated Traffic

The throughput obtained by the controlled connection depends on the throughput obtained by the other sources sharing the access routers and the backbone routers

TCP

- Window flow control if the window size for packet n is w_n, the source can only send packet n after packet $n - w_n$ has been acknowledged.
- Dynamic window size $w_n \in [1, w^*]$ which adapts to network conditions:
- Adaptation scheme in three steps:
  - From feedbacks to window
  - From window to packet transmission $\rightarrow$ max-plus
  - From packet transmission to feedback
- Reference window size $W(n)$
- Effective window size $w_n = \lfloor W(n) \rfloor$

From Feedback to Window

- $ACK(n) \in \{0,1\}$: feedback signal on the state of congestion at the reception of packet n:

$$W(0) = 1, \qquad W(n) = F(W(n-1), ACK(n)),$$

$$\begin{array}{ll} \text{phase} \uparrow & F(W(n-1),1) \geq W(n-1) \\ \text{phase} \downarrow & F(W(n-1),0) < W(n-1). \end{array}$$

- Possible values of 0:

– LO (loss): congestion or random loss of packet n;

– TO (timeout): round trip time (RTT) of packet n larger than $RTO(n)$;

- Other features slow start, congestion avoidance, idleness, updating of RTO RTO $\to RTO(n)$.

TCP Tahoe

- Tahoe without slow start:

$$F(W(n-1),1) = W(n-1) + \frac{1}{w_{n-1}}, \quad F(W(n-1),0) = 1$$

- Tahoe with slow start: threshold $\Theta(n)$ separating Congestion Avoidance from Slow Start: $\Theta(n) = \phi(\Theta(n-1), W(n-1), 1)$

$$\begin{aligned}
\phi(\Theta(n-1), W(n-1), 1) &= \Theta(n-1), \\
\phi(\Theta(n-1), W(n-1), 0) &= \lfloor \frac{W(n-1)}{2} \rfloor, \\
F(\Theta(n-1), W(n-1), 0) &= 1 \\
F(\Theta(n-1), W(n-1), 1) &= \min(W(n-1)+1, W^*), \text{ if } W(n-1) < \Theta(n-1), \\
&\quad W(n-1), \text{ if } W(n-1) = W^*, \\
&\quad \min(W(n-1) + \frac{1}{w_{n-1}}, W^*), \text{ otherwise.}
\end{aligned}$$

TCP Reno

- Reno without slow start:

$$
\begin{aligned}
F(W(n-1),1) &= W(n-1) + \frac{1}{w_{n-1}}, \\
F(W(n-1),\mathrm{TO}) &= 1, \\
F(W(n-1),\mathrm{LO}) &= \left\lfloor \frac{W(n-1)}{2} \right\rfloor
\end{aligned}
$$

- Reno with slow start; as above with

$$
\begin{aligned}
F(\Theta(n-1), W(n-1), LO) &= \max(1, \lfloor \frac{W(n-1)}{2} \rfloor), \\
F(\Theta(n-1), W(n-1), TO) &= 1.
\end{aligned}
$$

Part II: TCP is Max-Plus Linear

– Deterministic case

– Random case

– FIFO cross traffic

The Max-Plus Algebra

1. Scalar Semi-Field $(\mathbb{R}_{max}, \oplus, \otimes)$
   - $\mathbb{R}_{max} = \mathbb{R} \cup \{\varepsilon = -\infty\}$, $\oplus = \max$, $\otimes = +$
   - $(\mathbb{R}_{max}, \oplus)$ commutative idempotent semi-group $a \oplus b = b \oplus a$, $\quad a \oplus \varepsilon = \varepsilon \oplus a = a$, $\quad a \oplus a = a$;
   - $(\mathbb{R}_{max}, \otimes)$ commutative group neutral element $e = 0$
   - $\otimes$ distributes over $\oplus$ $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$;
2. Matrix Semi-Ring $(\mathbb{R}_{max}^{K,K}, \oplus, \otimes)$
   - Matrix addition Neutral $\mathcal{E}$, $\mathcal{E}_{ij} = \varepsilon$

$$(A \oplus B)_{ij} = A_{ij} \oplus B_{ij}$$

   - Matrix multiplication Neutral E, $E_{ij} = \varepsilon$, $i \neq j$, $E_{ii} = e$

$$(A \otimes B)_{ij} = \bigoplus_k A_{ik} \otimes B_{kj}$$

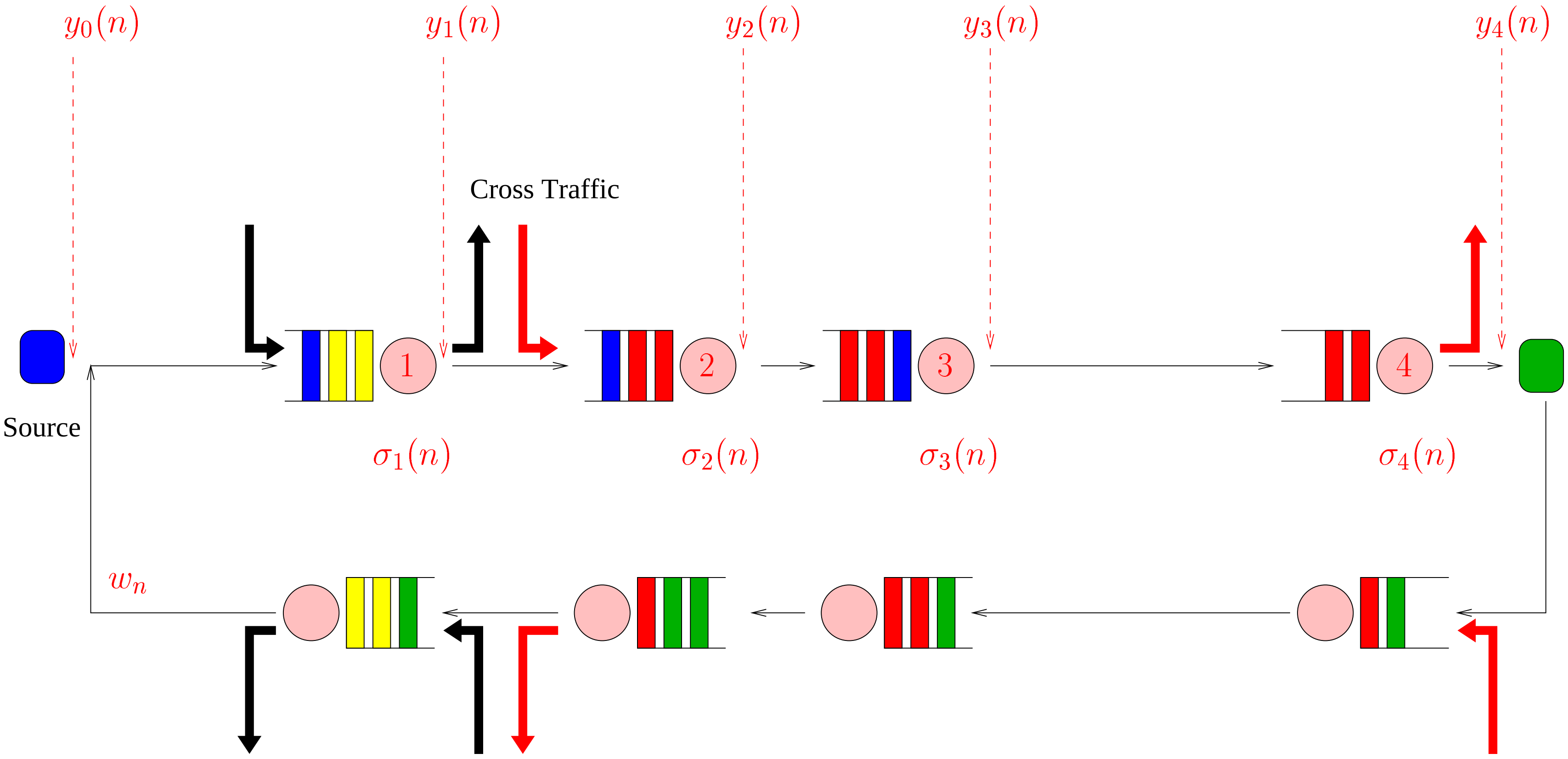


Figure 2: State Variables

From Window to Packet Transmission

State Variables:

- $y_i(n)$ date when packet n leaves router i;
- $\sigma_i(n)$ time that packet n waits to be served by router i once it is head of the line in its flow;
- $d_{i-1,i}$ propagation time from router $i-1$ to router i
- v_n experienced window size window size at the entrance of packet $n+1$;
- $u(n)$ date when the source is ready to send packet n.

$$
\begin{aligned}
y_0(n) &= [y_K(n - v_{n-1}) \otimes d_{K,0}] \oplus u(n), \\
y_i(n) &= [(y_{i-1}(n) \otimes d_{i-1,i}) \oplus y_i(n-1)] \otimes \sigma_i(n), \;\; i = 1, \ldots, K.
\end{aligned}
$$

- Saturated System: $u(n) = \varepsilon$.

Max-Plus Matrix Representation

$$Y(n) = (y_1(n), y_2(n), .., y_K(n)) \in I\!R^K$$
$$Z(n) = (Y(n), Y(n-1), .., Y(n-w^*+1)) \in I\!R^{K.w^*}$$

Theorem If the system is initialy empty, this can be rewritten as the affine dynamical system:

$$Z(0) = (0, .., 0), \quad Z(n) = A_{v_{n-1}}(n) \otimes Z(n-1) \oplus C(n) \otimes U(n), \ \forall n \geq 1$$

- The matrices $A_k(n)$, $k = 1, \ldots, w^*$ and $C(n)$ are functions of the $\sigma_i(n)$ and the $d_{i-1,i}$ only.
- Saturated System: $Z(n) = A_{v_{n-1}}(n) \otimes Z(n-1)$.

Exact Form of the Matrices

Let $M(n)$ et $M'(n) \in I\!R^{K,K}$:

$$(M(n))_{ij} = \sum_{k=j}^{i} \sigma_k(n) + \sum_{k=j}^{i-1} d_{k,k+1}, \quad \text{if } i \geq j, \qquad -\infty, \quad \text{if } i < j,$$

$$(M'(n))_{ij} = \sum_{k=1}^{i} (d_{k-1,k} + \sigma_k(n)) + d_{K,0}, \quad \text{if } j = K, \quad -\infty, \quad \text{if } j < K,$$

$$A_1(n) = \begin{pmatrix} M(n) \oplus M'(n) & \mathcal{E} & \cdots & \cdots & \mathcal{E} \\ I_d & \mathcal{E} & \cdots & \cdots & \mathcal{E} \\ \mathcal{E} & I_d & \mathcal{E} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{E} & \cdots & \mathcal{E} & I_d & \mathcal{E} \end{pmatrix} \in I\!R^{Kw^*,Kw^*}$$

Exact Form of the Matrices (*continued*)

$$A_2(n) = \begin{pmatrix} M(n) & M'(n) & \mathcal{E} & \cdots & \mathcal{E} \\ I_d & \mathcal{E} & \cdots & \cdots & \mathcal{E} \\ \mathcal{E} & I_d & \mathcal{E} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{E} & \cdots & \mathcal{E} & I_d & \mathcal{E} \end{pmatrix} \in I\!R^{Kw^*,Kw^*}$$

...

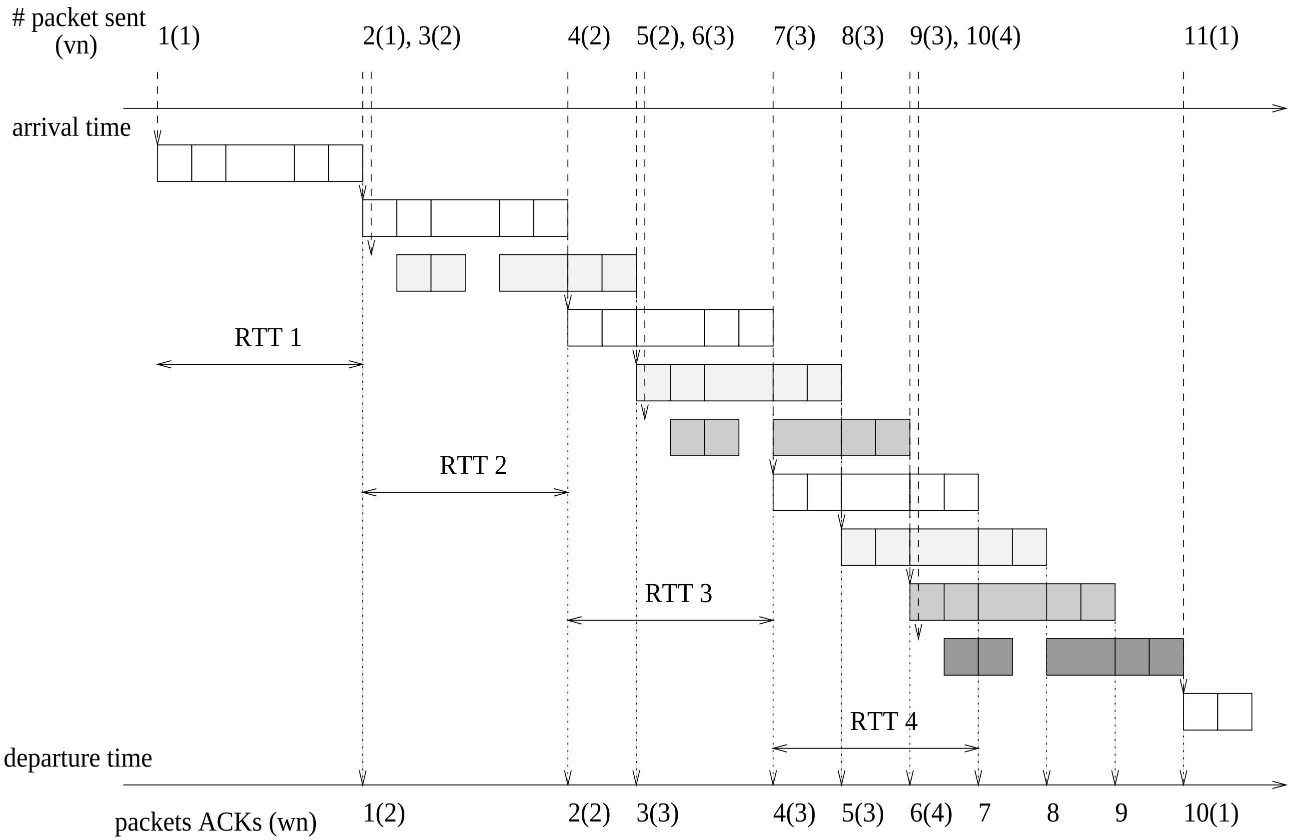
$$A_{w^*}(n) = \begin{pmatrix} M(n) & \mathcal{E} & \cdots & \mathcal{E} & M'(n) \\ I_d & \mathcal{E} & \cdots & \cdots & \mathcal{E} \\ \mathcal{E} & I_d & \mathcal{E} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{E} & \cdots & \mathcal{E} & I_d & \mathcal{E} \end{pmatrix} \in I\!R^{Kw^*,Kw^*}$$

Example of Evolution

- $K = 5$, $w^* = 4$ and $(\sigma_1, \cdots, \sigma_5) = (1, 1, 2, 1, 1)$.
- Periodic Tahoe without slow start:

$$(v_0, v_1, v_2, \ldots) = (1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 1, 1, 2, 2, 2, \ldots).$$

Example of Evolution (*continued*)



Packet # 6 experiences a window size of $v_5 = 3$ 6(3) The admission of packet # 6 takes place when the ack of packet $6 - 3 = 3$ is received.

From Transmission to Feedback

$$ACK(n) = G(v_{n-1}, \overline{Z}(n)).$$

$\overline{Z}$: equivalence class of Z for the equivalence relation $Z \sim Y$ if for all i, $Z_i = Y_i + c$ for some constant c.

Examples

– Rate based loss detection

* $1/\sigma^*(n)$: estimate of the current bottleneck service rate
* $\mathrm{RTT}(n)$: estimate of the current round trip time,

From Little's law

$$(\mathbf{G1}): \quad ACK(n) = 0 \text{ (Tahoe)}, \quad LO \text{ (Reno)}, \quad \text{if } \frac{v_{n-1}}{\mathrm{RTT}(n)} > \frac{1}{\sigma^*(n)}.$$

– Buffer overflow detection

β_i maximal buffer capacity for the reference flow on router i

$$(\mathbf{G2}): \quad ACK(n) = 0 \;\; \text{(Tahoe)}, \quad LO \text{ (Reno)},$$
$$\text{if } \exists 1 \le i \le K, \; y_{i-1}(n) + d_{i-1,i} < y_i(n - \beta_i).$$

Examples *(continued)*

– Timeout detection

$$\begin{aligned}(\mathbf{G3}):\quad & ACK(n) = 0 \;\; \text{(Tahoe)}, \quad TO \;\; \text{(Reno)},\\ & \text{if} \;\; y_K(n) - y_0(n) + d_{K,0} > RTO(n).\end{aligned}$$

Since $y_0(n) = y_K(n - v_{n-1}) + d_{K,0}$, this condition can be retrieved from $(v_{n-1}, \overline{Z}(n))$;

– Large service times or RTT

$$(\mathbf{G4})\text{-Tahoe}: ACK(n) = 0, \;\text{if}\; \sigma(n) \in \mathcal{B}, \;\; 1 \;\text{otherwise},$$

where $\sigma(n) = (\sigma_1(n), \ldots, \sigma_K(n))$ and where $\mathcal{B}$ is a subset of $I\!R^K$: (e.g. $\Sigma_i \sigma_i > X$ for timeout or $\sigma_i > Y_i$.

$$\begin{aligned}(\mathbf{G4})\text{-Reno}: ACK(n) \;=\; & LO, \;\; \text{if} \;\; \sigma(n) \in \mathcal{C},\\ & TO, \;\; \text{if} \;\; \sigma(n) \in \mathcal{D},\\ & 1, \;\; \text{otherwise},\end{aligned}$$

where $\mathcal{C}$ and $\mathcal{D}$ are subsets of $I\!R^K$ in the same vein as above.

Classes

– Case where one can study the window first and then the network:

  * G1: $ACK(n) = g(v_{n-1})$; periodic window size
  * G4: $ACK(n) = g(\sigma_n)$: Markov window size based on stopping times of the σ's

– Case where one cannot separate the two levels:

  * G2: $ACK(n) = g(\overline{Z}_n)$: stopping rule based on $\overline{Z}_n \in \mathcal{S}$
  * G3: $ACK(n) = g(v_{n-1}, \overline{Z}_n)$: stopping rule based on $(v_{n-1}, \overline{Z}_n) \in \mathcal{S}$

Stochastic Feedback

- The feedback signals $\{ACK(n)\}$ are also function of random perturbations $\{\xi(n)\}$ i.i.d. and $\{0,1\}$-valued:

$$ACK(n) = \Gamma(v_{n-1}, \overline{Z}(n), \xi(n)),$$

with $\Gamma(v_{n-1}, \overline{Z}(n), 1) = G(v_{n-1}, \overline{Z}(n))$ for both Tahoe and Reno and, for Tahoe $\Gamma(v_{n-1}, \overline{Z}(n), 0) = 0$, whereas for Reno

$$\begin{aligned} \Gamma(v_{n-1}, \overline{Z}(n), 0) \; = \; & LO, \text{ if } \; G(v_{n-1}, \overline{Z}(n)) \text{ is 1 or } LO; \\ & TO, \;\; \text{if } \; G(v_{n-1}, \overline{Z}(n)) = TO. \end{aligned}$$

p probability that $\xi(1) = 0$ (random loss probability).

Global Dynamics – Saturated Case

- Simplified Dynamics approximation $w_n = v_n$ for all n.

Theorem 1 *Under the foregoing assumptions, the sequence of vectors* $\{Z(n), W(n)\}$ *satisfies the recurrence relation*

$$\begin{aligned} Z(n) &= A_{\lfloor W(n-1) \rfloor}(n) \otimes Z(n-1), \\ W(n) &= F\left(W(n-1), \Gamma(W(n-1), \overline{Z}(n), \xi(n))\right), \end{aligned}$$

- Basis for an algebraic simulation scheme

Global Dynamics – Saturated Case (*continued*)

- Exact Dynamics History of the reference window size

$$\mathcal{W}(n) = (W(n), \ldots, W(n - w^* + 1))$$

- The experienced window size v_n is then obtained by picking the integer part of the appropriate coordinate of $\mathcal{W}(n)$ vector.

- Procedure for the increasing phase: $\rightarrow v_n = a(\mathcal{W}(n))$.

  $v = \lfloor W(n) \rfloor;$

  `for` $(k = 1; k < v; k$`++`$)$

      `if` $(\lfloor W(n-k+1) \rfloor == \lfloor W(n-k) \rfloor + 1)$

          v`-`;

Global Dynamics – Saturated Case (*continued*)

- New form of the main theorem:

$$
\begin{aligned}
Z(n) &= A_{a(\mathcal{W}(n-1))}(n) \otimes Z(n-1), \\
\mathcal{W}(n) &= \mathcal{F}\left(\mathcal{W}(n-1), \mathcal{G}(\mathcal{W}(n-1), \overline{Z}(n), \xi(n))\right),
\end{aligned}
$$

$\mathcal{F}$ and $\mathcal{G}$: extensions of the F and G mappings to the histories of the variables.

Part III: Classes of Models & Main Problems

Classes of Models

- Basic dichotomies:
  - fixed versus adaptive window size
  - exact versus simplified description of cross traffic
  - deterministic versus stochastic service times

Exact Cross Traffic Descriptions: FIFO Case, Fixed Window Size

W: window size (fixed), K: number of routers

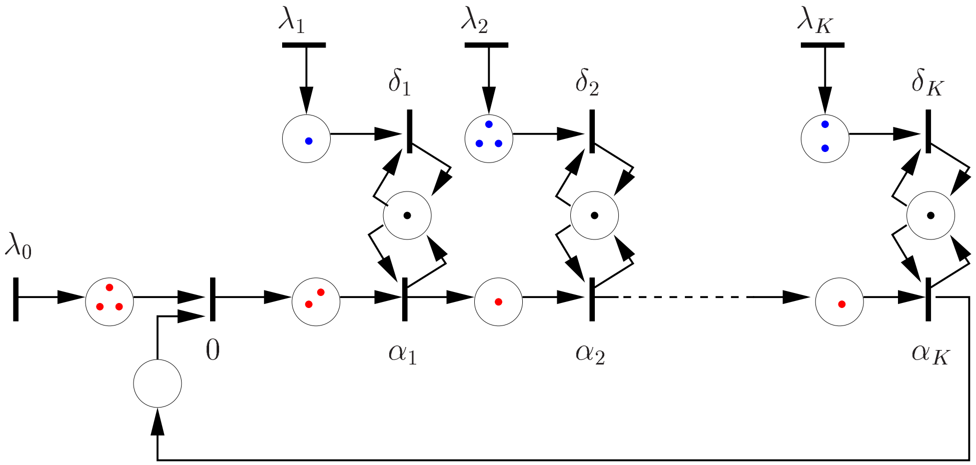


Figure 3: Controlled flow Cross flows

Examples of Models to be Studied

- fixed window size and deterministic service times: spectral theory of matrices (M)
- fixed window size and random service times: random media, expansions (HM)
- fixed window size and true cross traffic: ergodic theorems, bounding techniques (HM)
- adaptive window size: tous les coups sont permis from (M) to (HM*)

Main Problems

– What is the ergodic averaged throughput of a long FTP session?

– What is the stationary distribution of the instantaneous throughput?

– What is the stationary distribution of the reponse time to a click?

– What is the loss or the timeout probability?

Classical Estimate for the Throughput of a Controlled Connection

– Loss Probability p_{loss}

– Round Trip Time RTT

The throughput for a single TCP connection is of the form:

$$D_{th} = \frac{c_o}{RTT\sqrt{p_{loss}}},$$

where c_o is a real constant.

Mathis, M. Semske, J. Mahdavi, J and Ott, T. (1997) The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. *Computer Communication Review.*

Classical Estimate for the Throughput of a Controlled Connection (*continued*)

Limitations: 1 router!

- p_{loss} cannot be predicted in general;
- The formula is inadequate for short bursts as in HTTP traffic.

Padhye, J. and Firoiu, V. and Towsley, D. and Kurose, J., (1998) Modeling TCP throughput: a simple model and its empirical validation., *SIGCOMM 98*.

Lecture I: References

- R. Stevens, TCP Illustrated Vol. 1 and 2, 1994-195
- FB & D. Hong, Sigcomm 2000