Resource pooling issues

Multipath transport meeting

Cambridge

2 July 2010

Frank Kelly

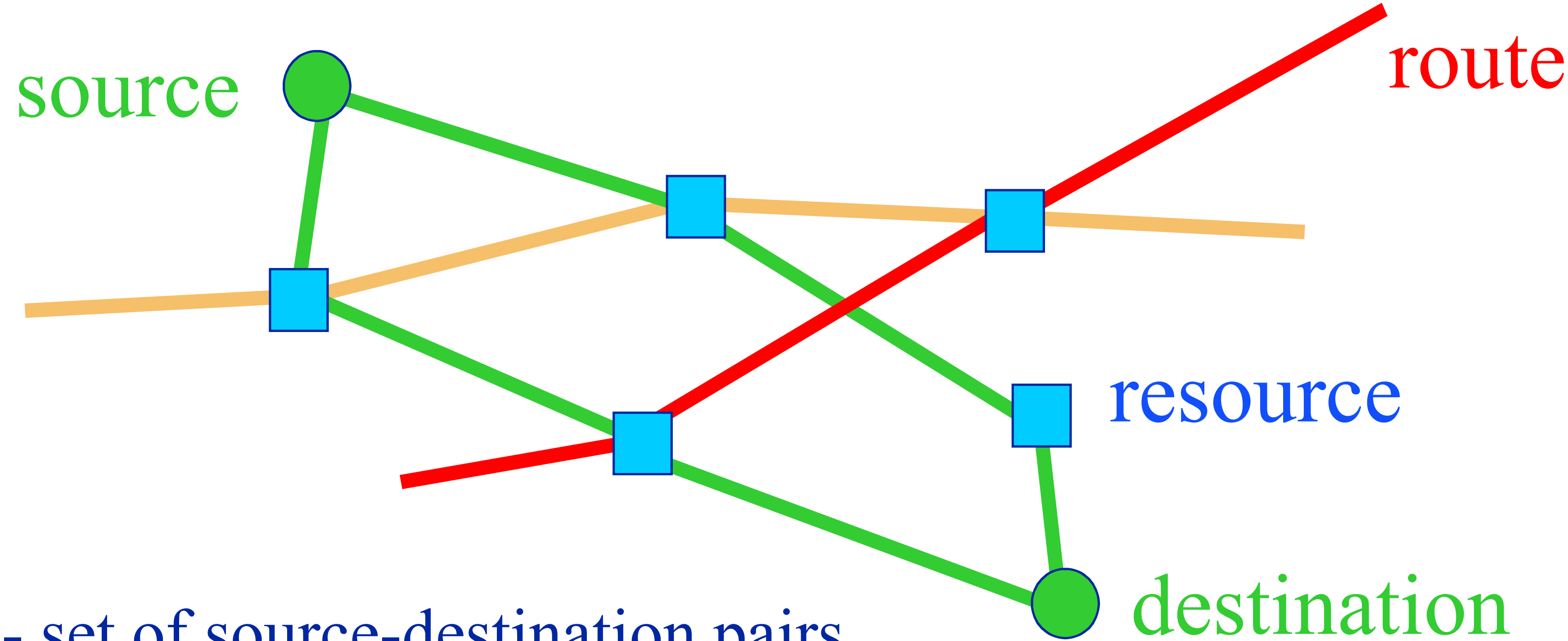
www.statslab.cam.ac.uk

Outline

- Delay stability
- Flow level performance
- Network capacity

Multipath transport

Can we extend flow control to allow load balancing across routes?   Danger: route flap.



S - set of source-destination pairs

$r \in s$ - route r serves s-d pair s

Example of a combined rate control and routing algorithm

On route r

- $x_r(t)$ increased by a / T_r on positive acknowledgement
- $x_r(t)$ decreased by $b_r y_{s(r)}(t) / T_r$ for each congestion indication, where

$$y_s(t) = \sum_{r \in s} x_r(t - T_r)$$

is rate of returning acknowledgements for s-d pair s at time t

- $s = \{r\}$ corresponds to Scalable TCP

Delay stability

Han, Shakkottai, Hollot, Srikant, Towsley 2003, K, Voice 2005

Equilibrium is locally stable (for resource loads) if there exists a global constant β such that

$$x p_j'(x) < \beta \, p_j(x), \qquad a\,(1+\beta) < \frac{\pi}{2}$$

condition on sensitivity for each resource j

condition on aggressiveness of sources

Delay stability

Han, Shakkottai, Hollot, Srikant, Towsley 2003, K, Voice 2005

Equilibrium is locally stable (for resource loads) if there exists a global constant β such that

impact of routing

$$x p_j'(x) < \beta \, p_j(x), \qquad a\,(1+\beta) < \frac{\pi}{2}$$

condition on sensitivity for each resource j

condition on aggressiveness of sources

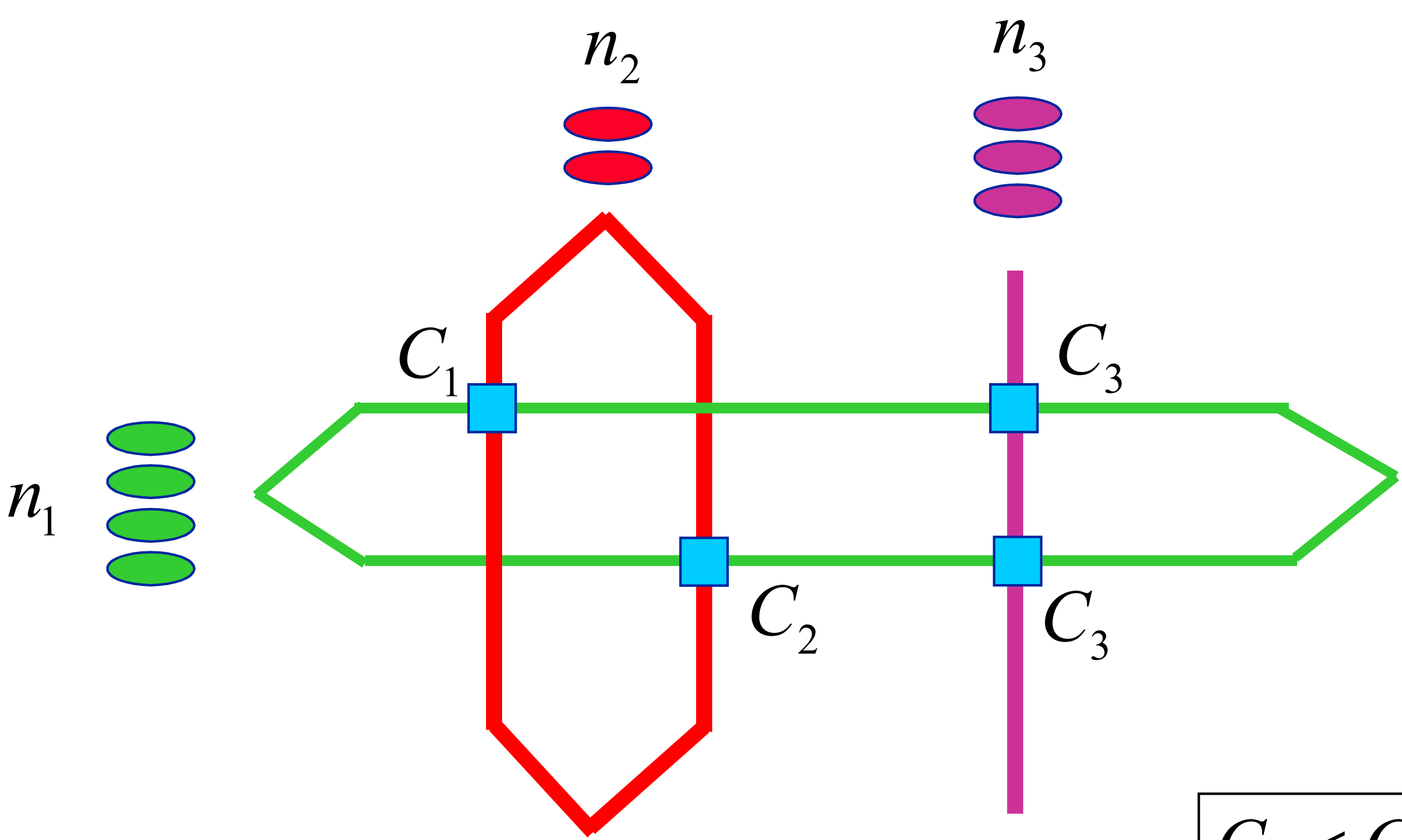
Suggestion on routing?

Han, Shakkottai, Hollot, Srikant, Towsley 2003, K, Voice 2005
Key, Massoulié, Towsley 2006

- Stable, scalable load balancing across paths, based on end-to-end measurements, can be achieved on the same time-scale as rate control
- For load balancing, the key constraint on the responsiveness of each route is the round-trip time of that route.
- While it is natural for structural information to be provided by the network layer, load balancing is more naturally part of the transport layer.

Chiang, Low, Calderbank, Doyle 2007

Example of multipath routing

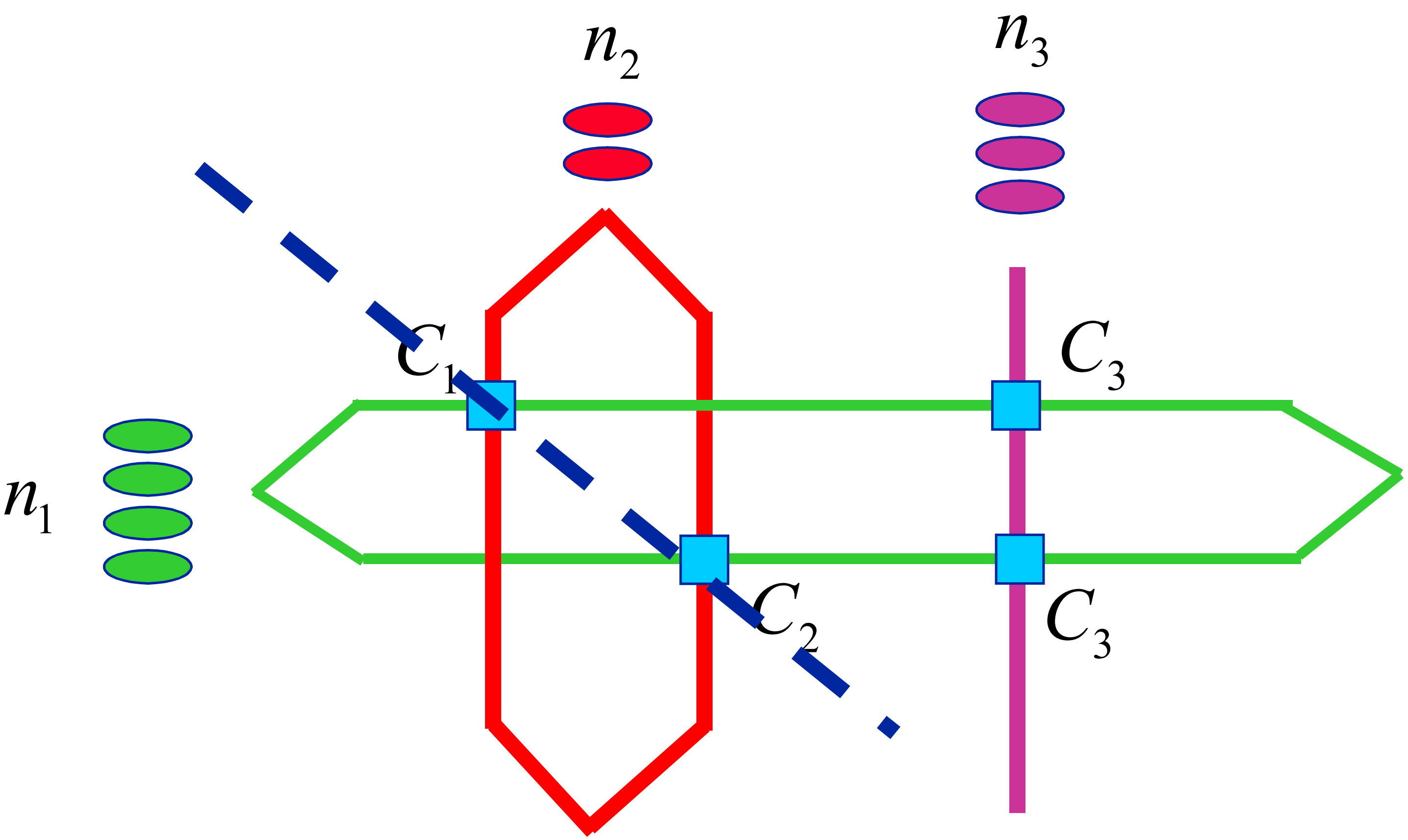


$$C_3 < C_1, C_2$$

Routes, as well as flow rates,
are chosen to optimize

$\sum_s n_s \log(x_s)$ over source-destination pairs s

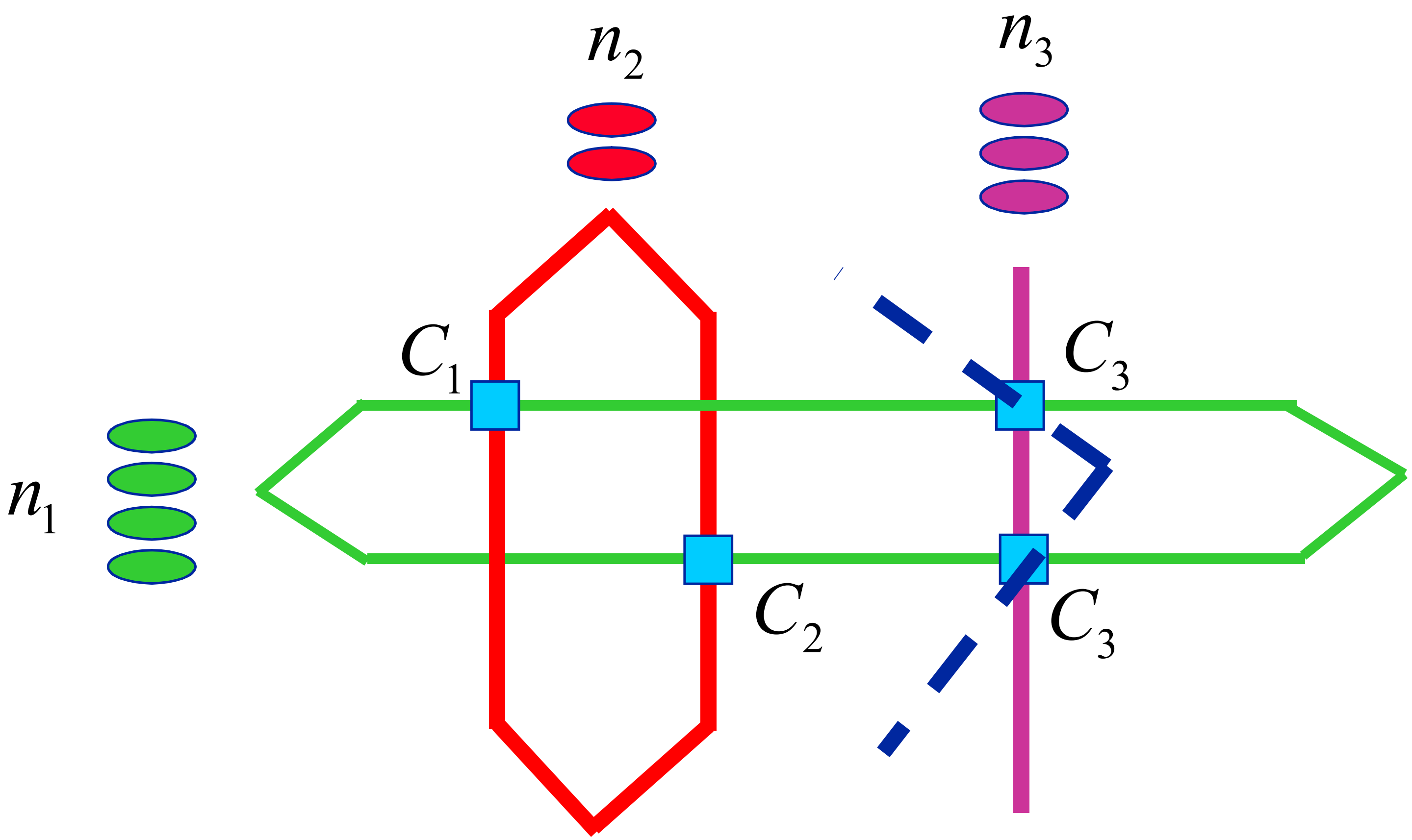
First cut constraint



$$n_1 x_1 + n_2 x_2 \leq C_1 + C_2$$

Cut defines a single *pooled resource*

Second cut constraint



$$\frac{1}{2}n_1x_1 + n_3x_3 \leq C_3$$

Cut defines a *second* pooled resource

Product form

$$\alpha = 1, w_r = 1, \ r \in R$$

In heavy traffic, and subject to some technical conditions, the (scaled) components of the shadow prices p for the pooled resources are independent and exponentially distributed. The corresponding approximation for n is

$$n_s \approx \rho_s \sum_j p_j A_{js} \quad s \in S$$

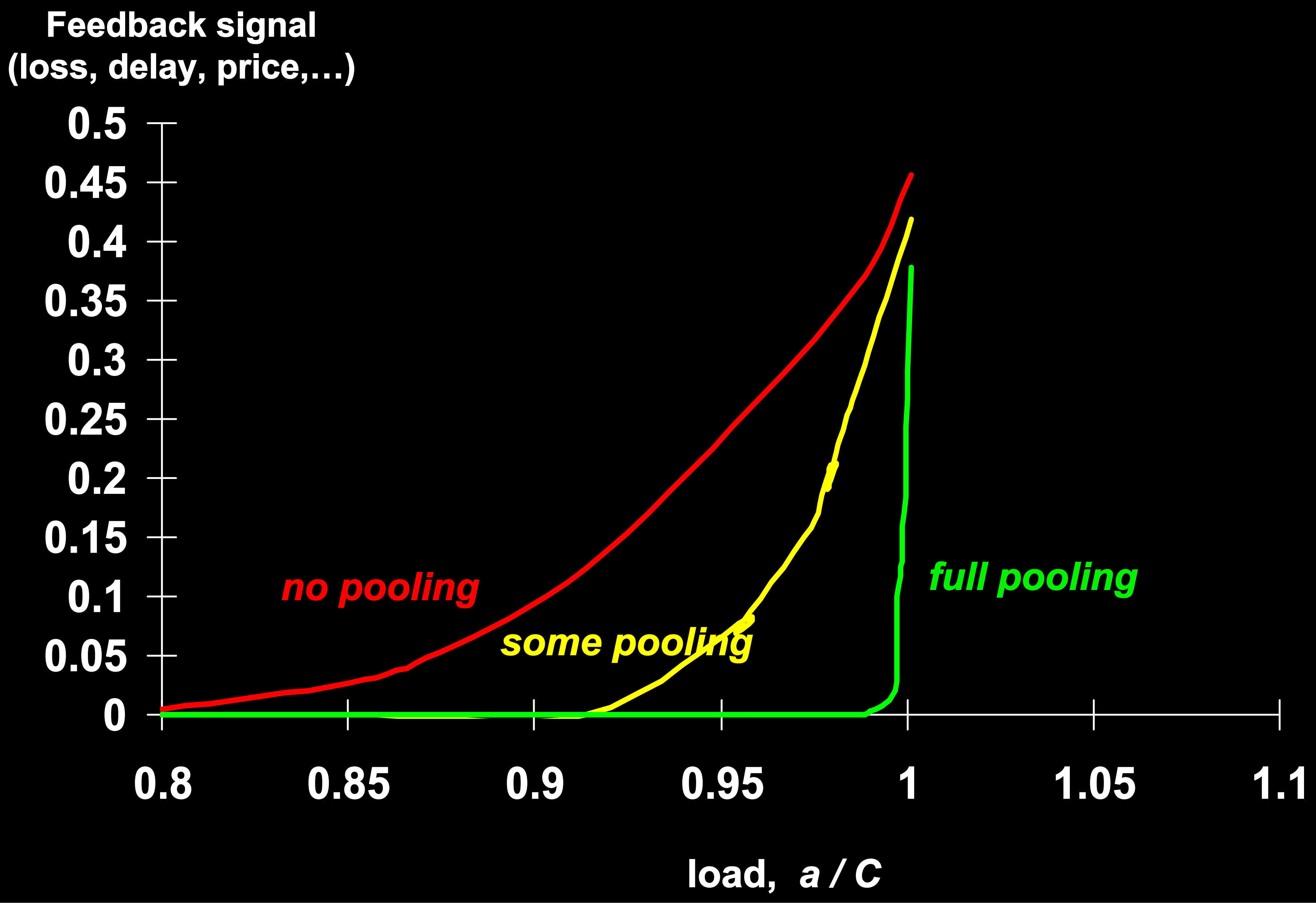
where

$$p_j \sim \text{Exp}(\overline{C}_j - \sum_s \overline{A}_{js} \rho_s) \quad j \in \overline{J}$$

Dual random variables are independent and exponential

Kang, K, Lee and Williams 2009

Sudden impact of capacity



Remarks on resource pooling

- Resource pooling does indeed
  - respond robustly to failures and overloads
  - lessen the impact of forecasting errors
  - make use of spare capacity in the network
  - permit flexible use of network resources
- It can disguise the approach of capacity overload