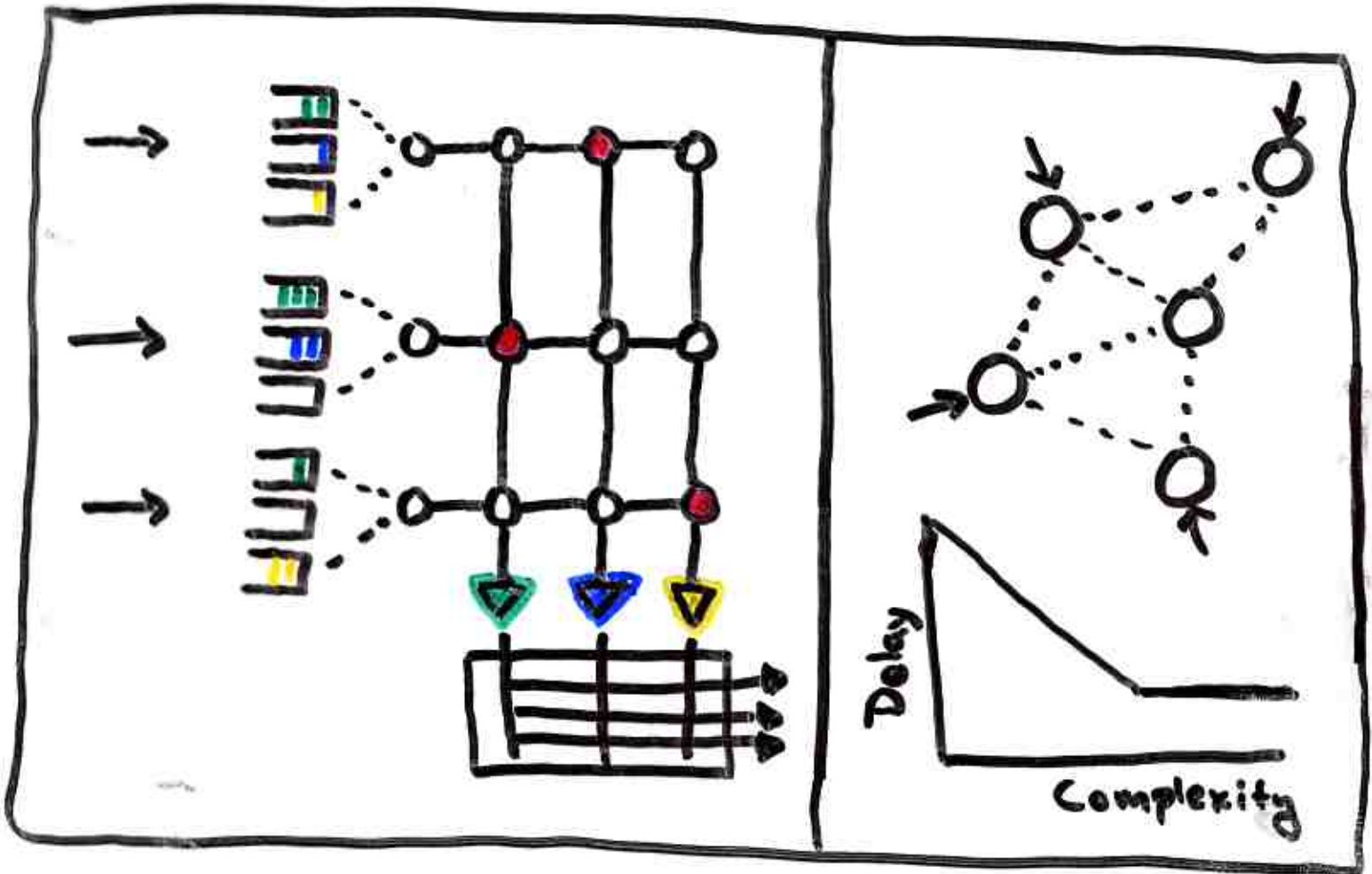
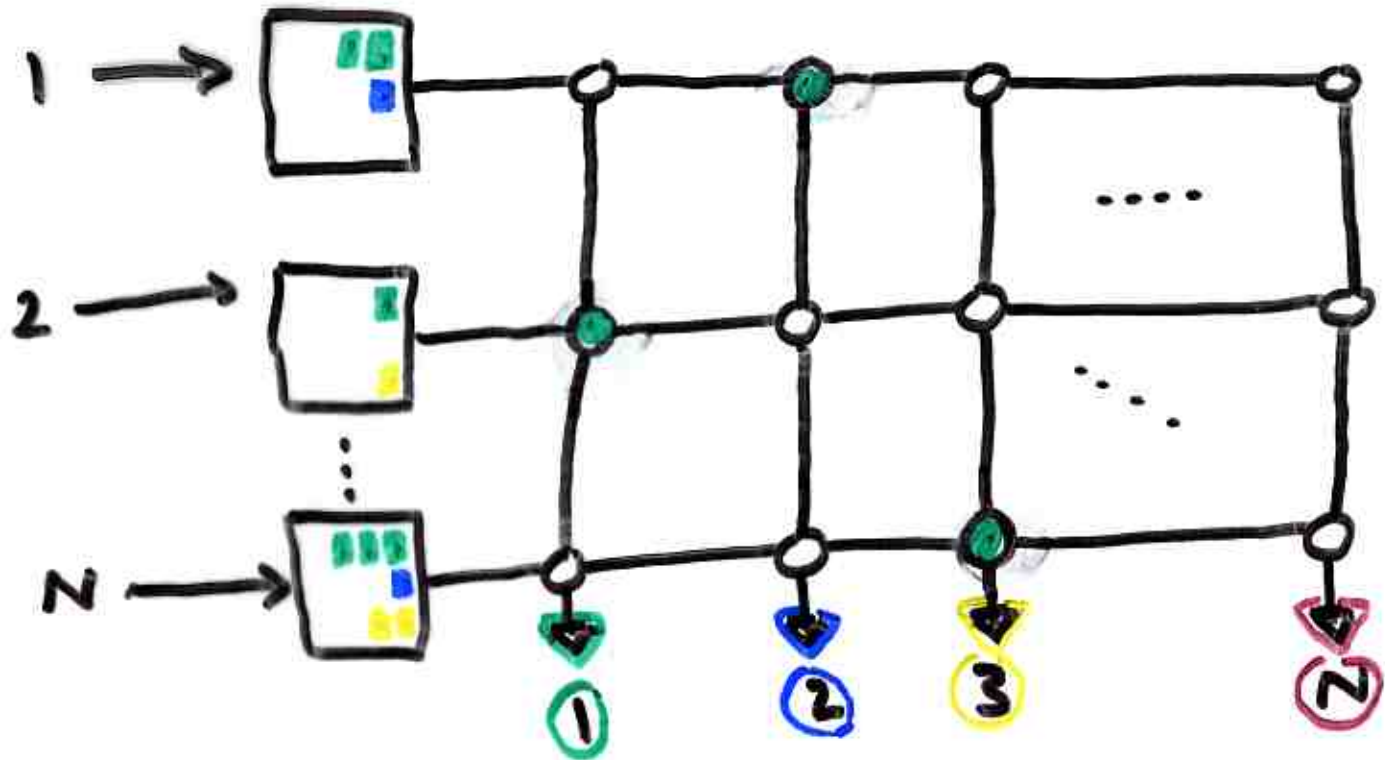


Delay, Complexity, and Fairness for Packet Switch Networks



Michael J. Neely
University of Southern California
<http://www-rcf.usc.edu/~mjneely>

The Switch and Crossbar (Slotted time system)



$X_{ij}(t)$ = Arrivals to input i destined for output j during slots $\{0, 1, \dots, t\}$

$\lambda_{ij} = \lim_{t \rightarrow \infty} \frac{X_{ij}(t)}{t}$ = Arrival rate for (i, j)

$S_{ij}(t) = \begin{cases} 0 & \text{- connection } (i, j) \text{ disabled} \\ 1 & \text{- connection } (i, j) \text{ enabled.} \end{cases}$

Constraint : $(S_{ij}(t)) \in$ Permutation Matrix

Capacity Region of the Switch:

(λ_{ij}) such that:

$$\sum_i \lambda_{ij} \leq 1 \quad \text{for all inputs } i$$

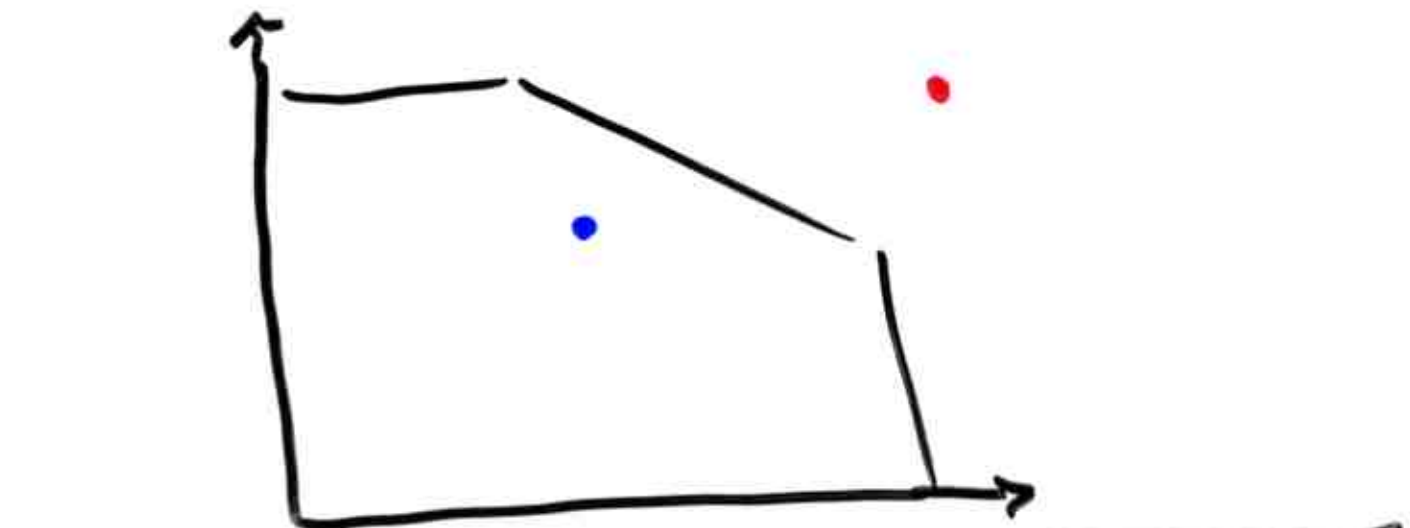
$$\sum_i \lambda_{ij} \leq 1 \quad \text{for all outputs } j$$

Talk:

- ① Fundamental Delay Bounds
[Neely HPSR 2004]
- ② Delay · Complexity Tradeoffs
[Neely CISS 2002]
- ③ Fairness and Optimal
Flow Control for
Switches and networks
[Neely Ph.D. thesis 2003]

inside capacity

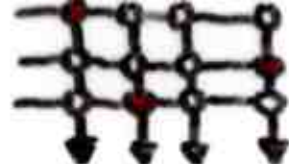
outside capacity



Start with delay...

Outline

(λ_{ij})



① Backlog Unaware Algs. : Delay $\geq O(N)$

(Periodic, Randomized scheduling based)
on known rates (λ_{ij})

- Chang et al. (Infocom 2000, 2002)
- Koxsal (MIT Thesis 2002)
- Andrews and Vojnović (Infocom 2003)
- Leonardi et al. (TON 2001)

② Backlog Aware Algs. : $O(\log(N))$ delay is achievable

* MWM : Tassiulas 1992, McKeown 1996

Frame : Weller, Hajek 1998

Andrews, Zhang 2001

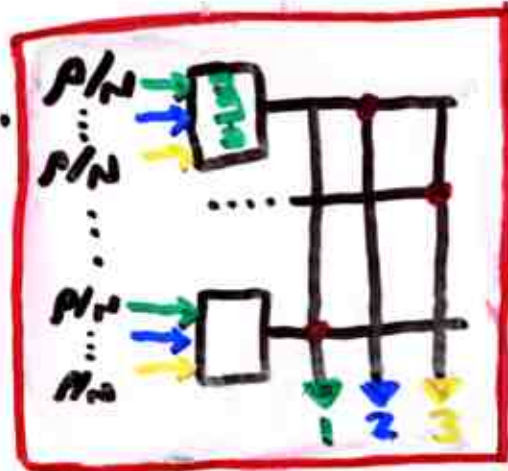
Iyer, McKeown 2002

Previously best known delay for random (Poisson) inputs: $O(N)$ Leonardi et al. 2001

Example Scheduling Algs.

(Backlog Independent)

Uniform Poisson Traffic.
Loading $\rho < 1$.



Randomized Sched:

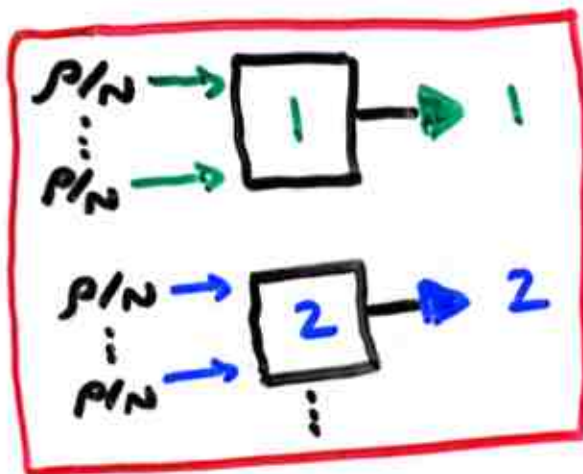
$$\bar{W}_{\text{randomized}} = \frac{N - 1/2}{1 - \rho} + 1$$

Periodic Sched:

$$\bar{W}_{\text{periodic}} = \frac{N}{2(1 - \rho)} + 1$$

Output Queue:

$$\bar{W}_{\text{output queue}} = \frac{1}{2(1 - \rho)} + 1$$



→ Statistical Multiplexing Gains

① Backlog Independent Scheduling

$$X_{1,1}(t) \rightarrow \boxed{\text{■ ■}} \rightarrow S_{1,1}(t)$$

$$X_{1,2}(t) \rightarrow \boxed{\text{■ ■ ■}} \rightarrow S_{1,2}(t)$$

$$\vdots$$

$$X_{1,N}(t) \rightarrow \boxed{\text{■ ■}} \rightarrow S_{1,N}(t)$$

⋮

$$X_{N,1}(t) \rightarrow \boxed{\text{■}} \rightarrow S_{N,1}(t)$$

$$X_{N,2}(t) \rightarrow \boxed{\text{■ ■}} \rightarrow S_{N,2}(t)$$

$$\vdots$$

$$X_{N,N}(t) \rightarrow \boxed{\text{■ ■}} \rightarrow S_{N,N}(t)$$

$(S_{ij}(t)) \in \text{Perm. Matrix}$
(for all t)

Theorem 1: Let $(S_{ij}(t))$ be any stationary sched. alg. that is independent of input streams and backlog. Then:

$$\text{Avg. Delay} \geq O(N)$$

If (λ_{ij}) matrix has $O(N^2)$ entries $\lambda_{ij} \geq O(\frac{\lambda_{av}}{N})$.

Ex: Uniform traffic
 $\lambda_{ij} = \rho/N, \lambda_{av} = \rho$

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & * & \dots & * \\ \lambda_{21} & \lambda_{22} & * & \dots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{N1} & * & \dots & \dots & * \end{pmatrix}$$

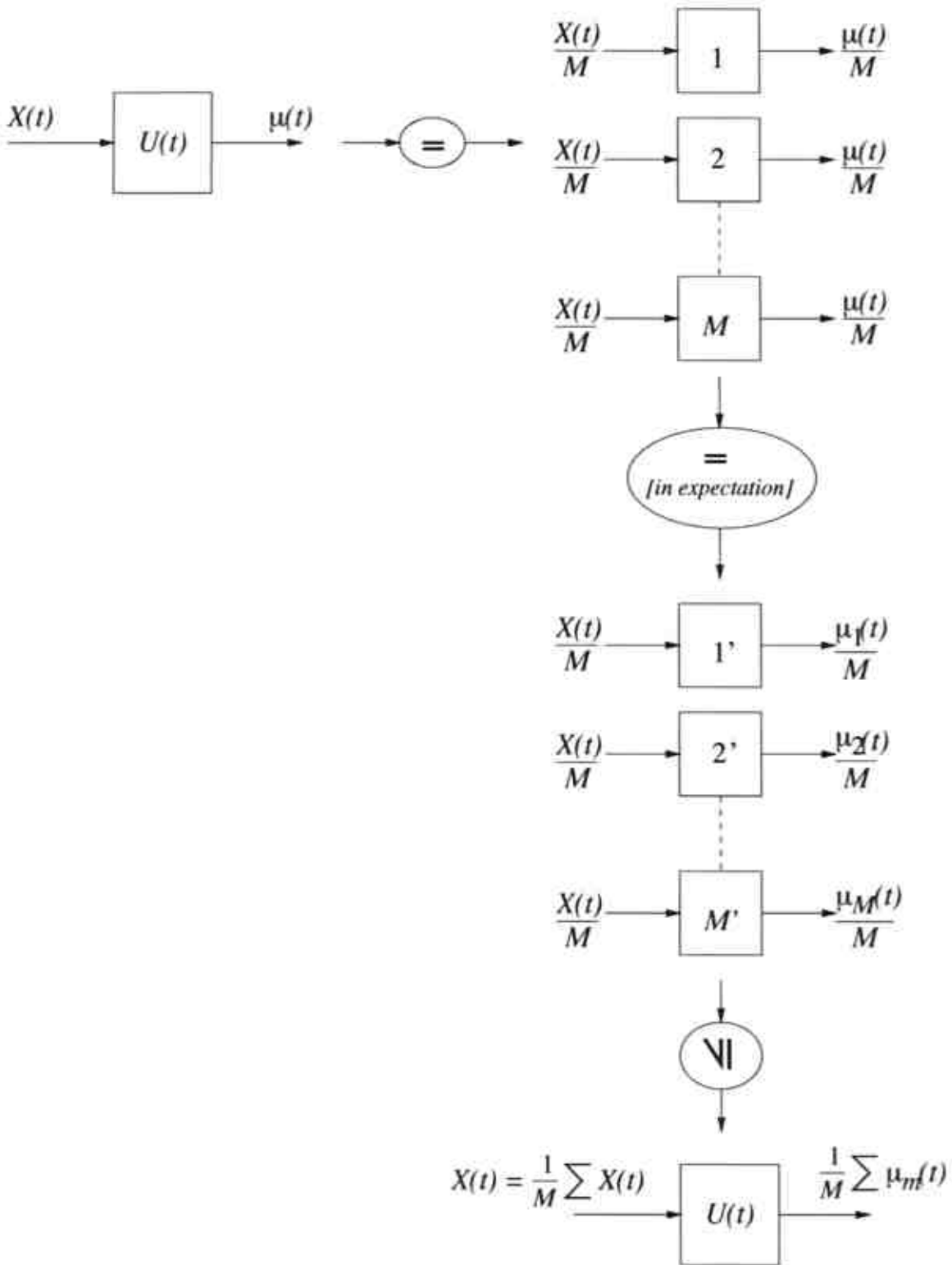
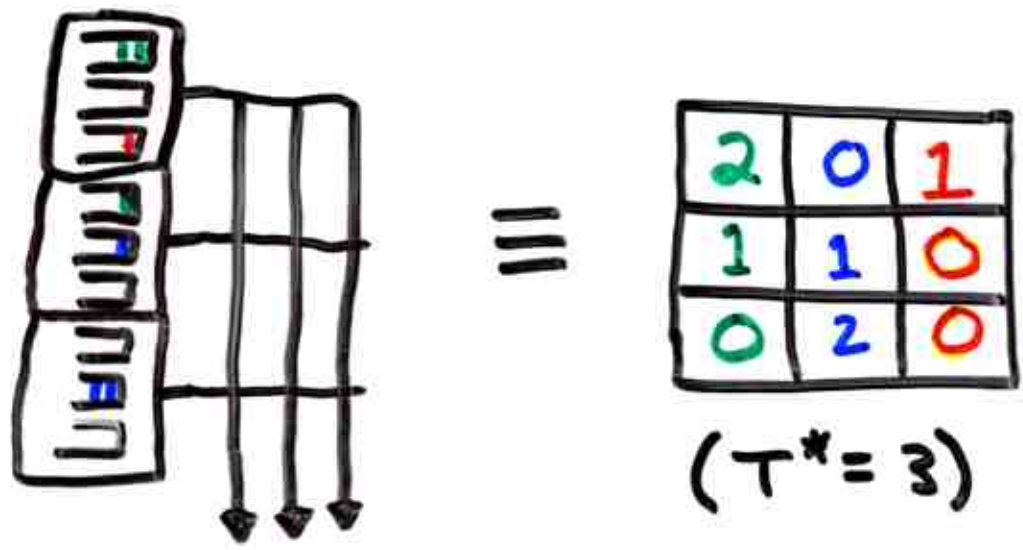


Figure C-1: An illustration proving the Jitter Theorem.

② Backlog Aware Scheduling

Fact 1: Minimum Clearance Time T^*



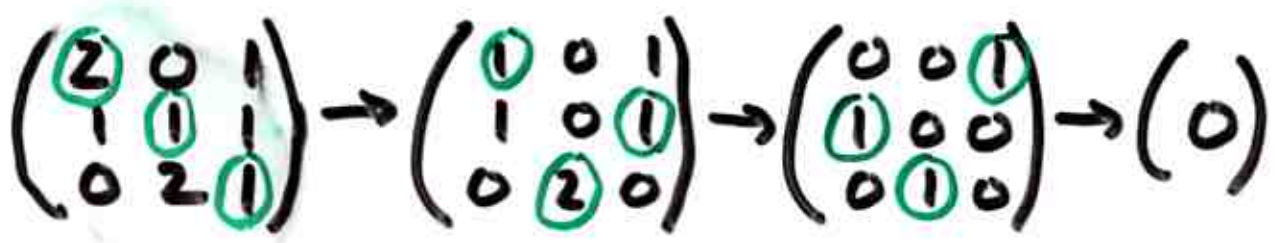
$T^* = \text{Max sum of any row or column. } \square$

(Birkhoff-Von Neumann Thm., Hall's Thm)

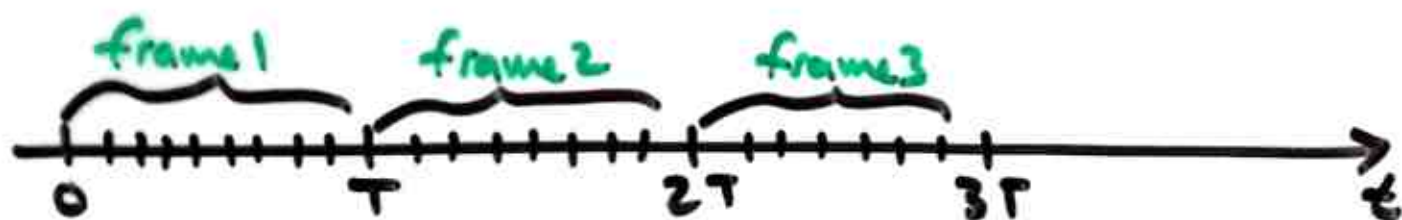
How?



B. Max-Size Matches: (Any) (Finish in T^*)



Fair-Frame Algorithm: T slots = 1 frame



1. First Frame \Rightarrow Schedule $(S_{ij}(t))$ Randomly.

2. On Frame $(k+1)$:

• Have $(L_{ij}(kT)) =$ Arrivals from Prev. Frame

• Fair Decomposition:

$$(L_{ij}) = (\tilde{L}_{ij}) + (V_{ij})$$

conforming
packets

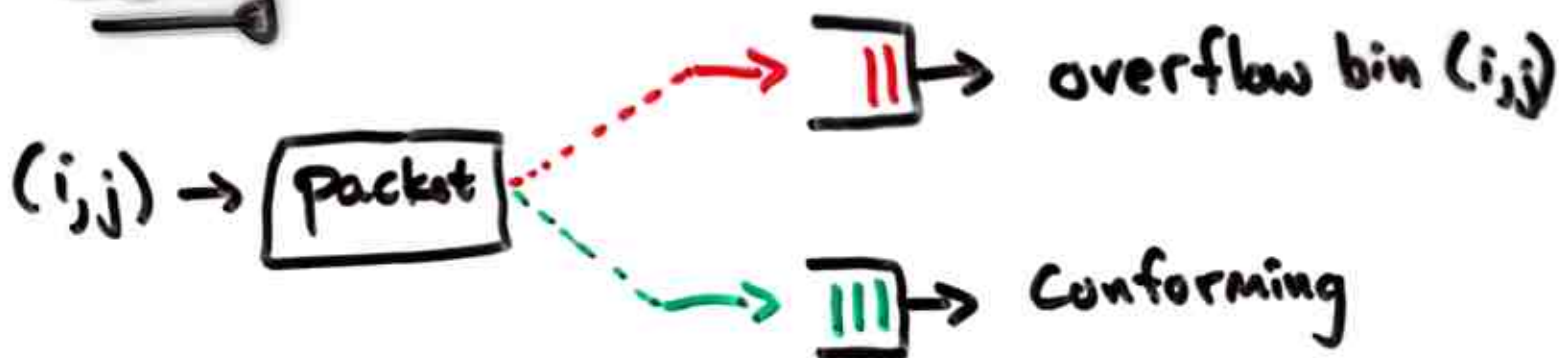
overflow
packets

3. Schedule conforming packets during frame.

4. If finish early, schedule randomly until end of frame, serving overflow of previous frames.

5. Repeat from Step 2.

Delay



$$\text{Avg. Delay} = (\text{Delay conf.}) \cdot P_{\text{conform}} + (\text{Delay overflow}) \cdot P_{\text{overflow}}$$

Theorem: Given Poisson inputs with loading ρ ($\sum_i \lambda_{ij} \leq \rho$ for all j , $\sum_j \lambda_{ij} \leq \rho \forall i$), we can choose a frame size T such that:

- avg. delay $\leq O(\log(N))$
- $T \leq O(\log(N))$
- $P_{\text{conform}} \geq 1 - O(1/N^2)$

Proof Idea:



If non-conforming during frame k , then
at least one of:

$$\sum_i x_{ij}(T) \leq T, \quad j = 1, 2, \dots, N$$

$$\sum_j x_{ij}(T) \leq T, \quad i = 1, 2, \dots, N$$

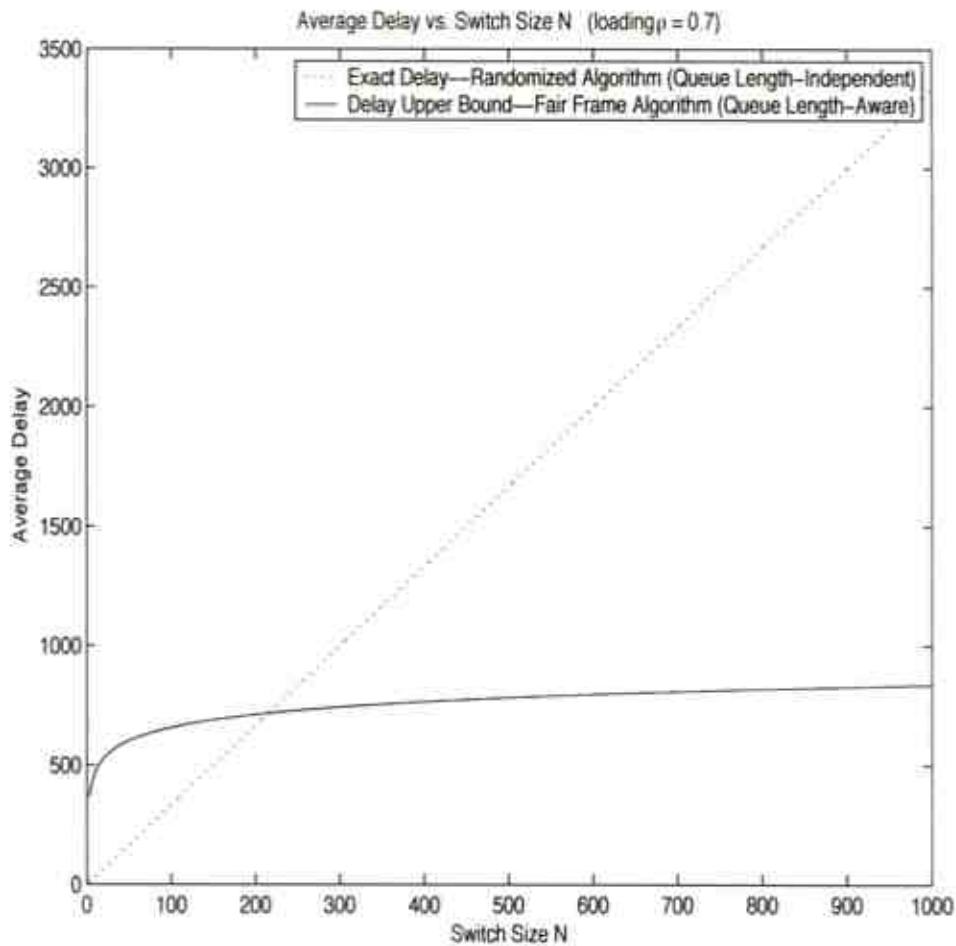
must have been violated.

Chernov: $\Pr[\text{one const. violated}] \leq \delta^T$

$$(\delta = \rho e^{-\rho})$$

Union bound:

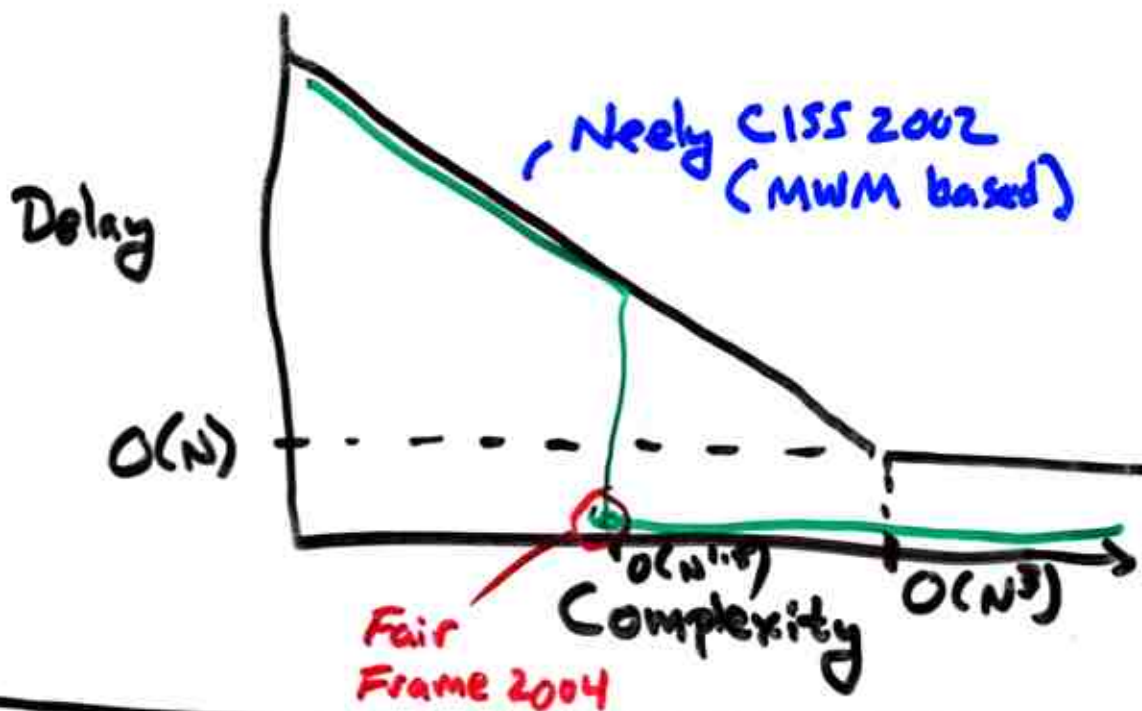
$$\Pr[\text{any violated}] \leq 2N\delta^T$$



MWM: $O(N)$ analytical bound.

Fair Frame: $O(\log(N))$ analytical bound.

II. Complexity · Delay Tradeoffs (1 slide)



- Suite of MWM-based algs. over frames:

$$\left. \begin{array}{l} \text{Complexity} = O(N^\alpha) \\ \text{Delay} \leq O(N^{4-\alpha}) \end{array} \right\} \begin{array}{l} \text{any } \alpha \text{ s.t.} \\ 0 \leq \alpha \leq 3 \end{array}$$

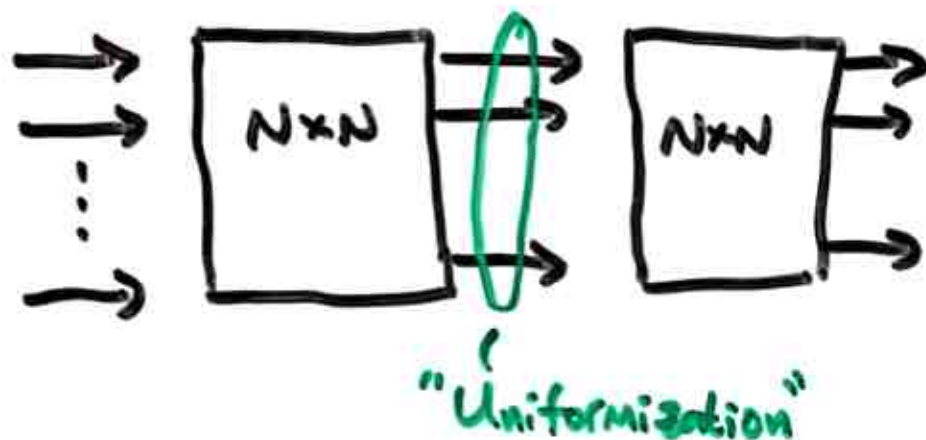
* Similar MWM "approx." alg. independently developed in Shah 2002.

- How does Fair-Frame compare?

$$\begin{array}{l} \text{Complexity} = O(N^{1.5} \log(N)) \\ \text{Delay} \leq O(\log(N)) \end{array}$$

Transition Slide

Lower Complexity thru extra hardware



"Zero-Complexity" randomized or periodic scheduling leads to 100% thruput.

Chang 2002

Koksal Ph.D. thesis 2002

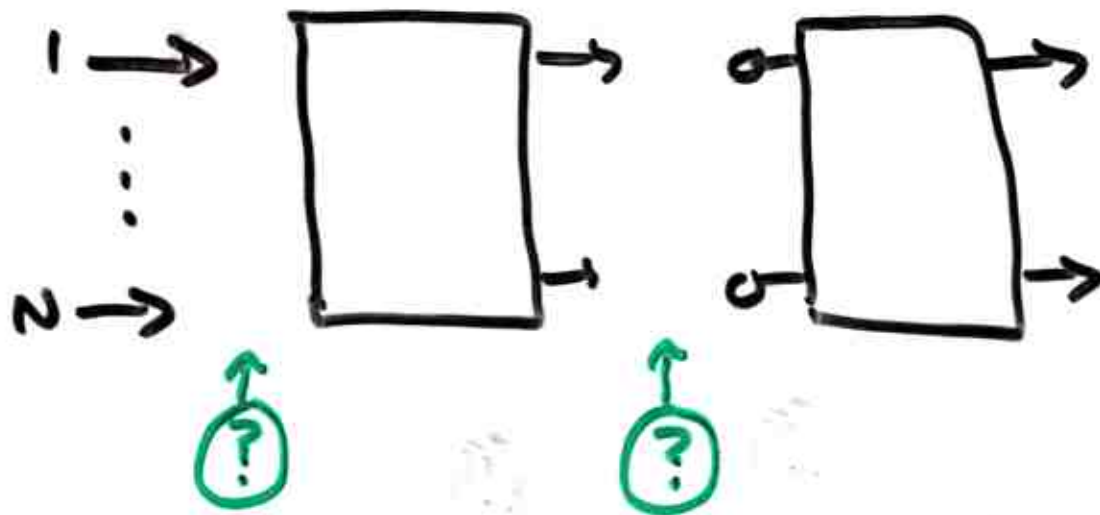
• $O(1)$ complexity

• $O(N)$ delay

What if (λ_{ij}) outside capacity?

III. Outside the Capacity Region

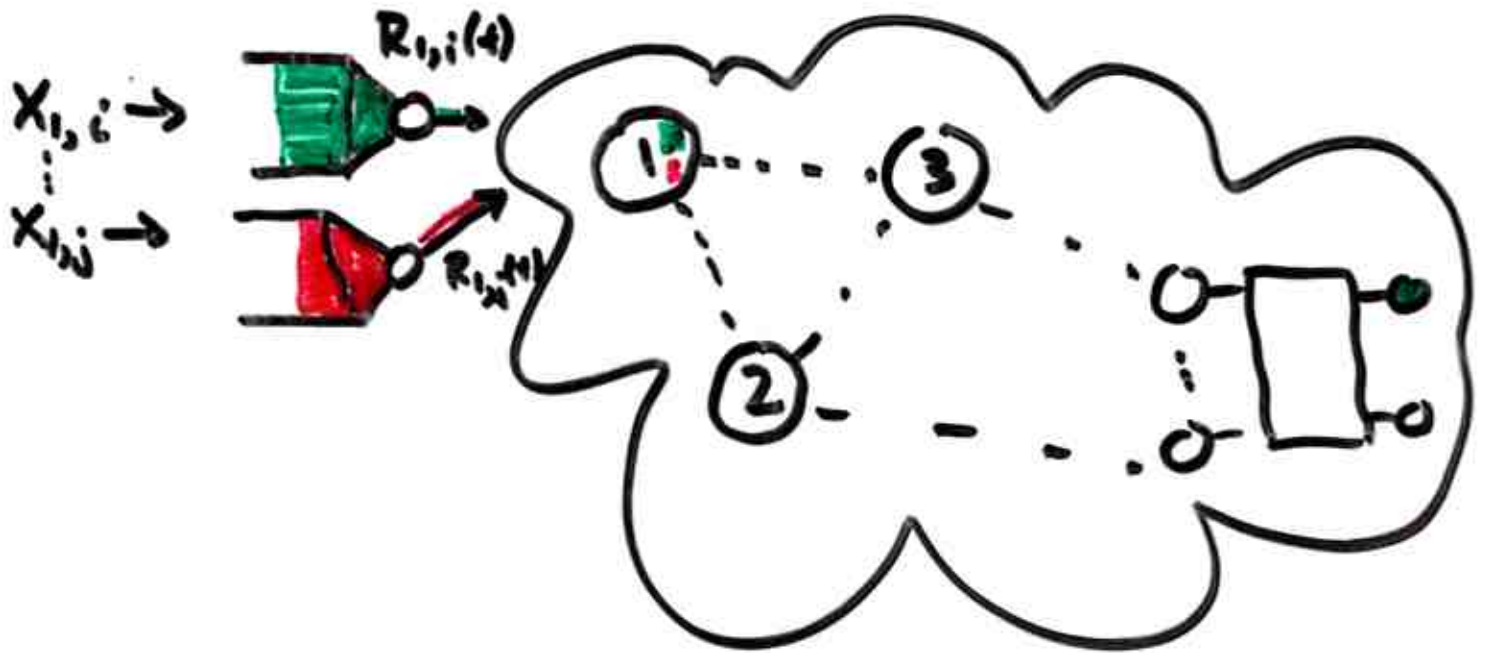
Q: Where to put the flow control?



Prefer to have near inputs b/c:

- ① Fairness across inputs
- ② No wasted transmissions
- ③ Closer to source (for feedback)

Formulation :: General Packet Switch Net
 "Zero-Complexity" randomized link scheduling.

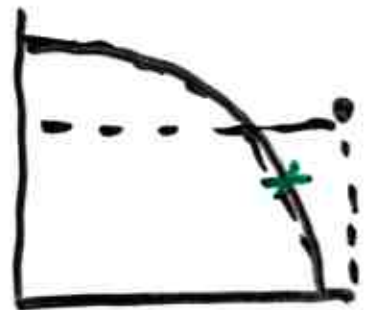


Utility $g_{ic}(r) =$ "satisfaction" user i has when send at long term rate r .

Max: $\sum_{i \in c} g_{ic}(r_{ic})$

St: $r_{ic} \leq \lambda_{ic}$

$(r_{ic}) \in \Delta$

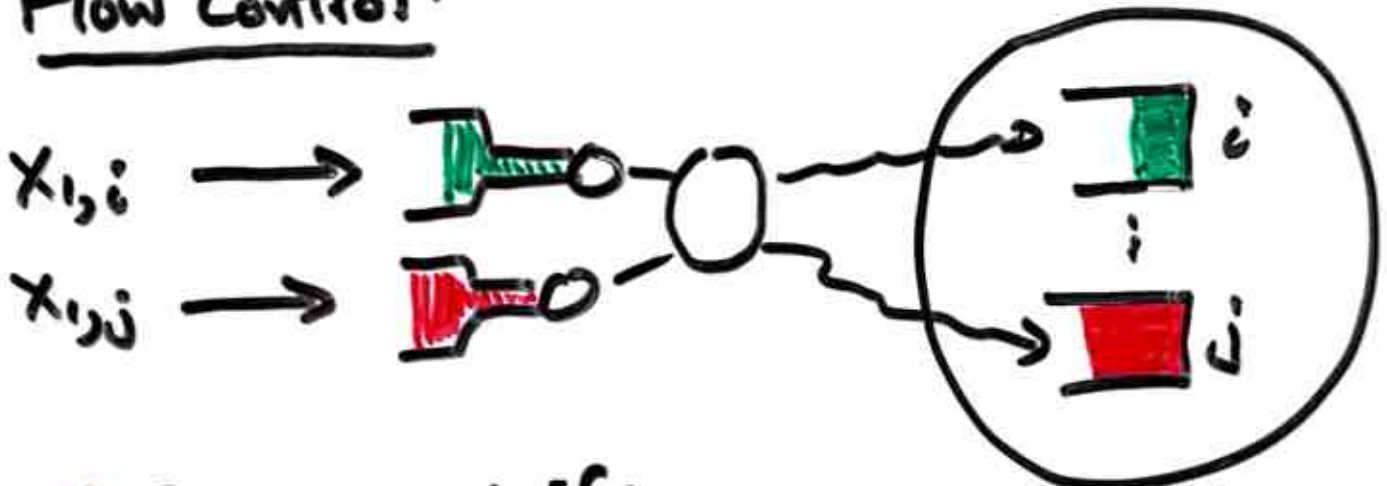


Capacity region

Algorithm:

Special case: $g_i(c) = \theta_i c^r$
(linear utility)

Flow Control:



* Send packet if:

$$L_{1i}(t) \leq v \theta_{1i}$$

(where v = parameter of control)

* Then choose i such that $i = \operatorname{argmax} [v \theta_{1i} - L_{1i}(t)]$

In-Network Packet Selection

Differential Backlog Maximizer

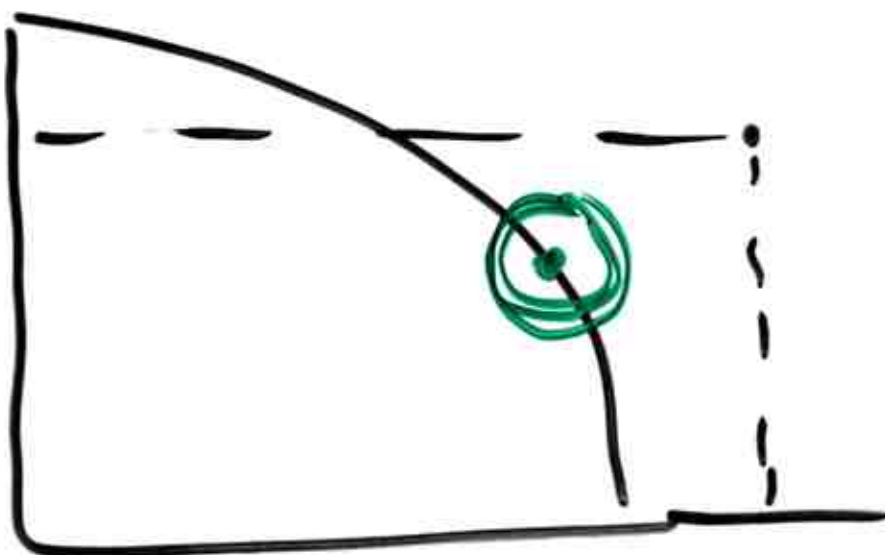


$$c = \operatorname{argmax}_c \{ L_a^{(c)}(t) - L_b^{(c)}(t) \}$$

Thm: (Assume i.i.d. Bern. inputs λ_{ij})

Utility: $\sum g_{ic}(\bar{r}_{ic}) \geq \sum g_{ic}(r_{ic}^{opt}) - \frac{NB}{\nu}$

Delay: $\sum \bar{L}_{ic} \leq \frac{NB}{4\lambda_{sym}} + \frac{\nu g_{max} N}{4\lambda_{sym}}$



Example:

For 2-tandem $N \times N$ switch:

$$\sum \theta_{ic} \bar{r}_{ic} \geq \sum \theta_{ic} r_{ic}^{opt} - \frac{10N}{\nu}$$

$$\text{Avg. Delay} \leq 5N + \frac{N\nu\theta_{max}}{2}$$

Conclusions:

I. Logarithmic Delay is Achievable

$$O(1) \leq \text{optimal delay} \leq O(\log(N))$$

II. Arb. Low Complexity is Achievable

(Delay · Complexity tradeoff



Paradigm shift:

Stability is an incomplete metric for performance.

III. Network Control Outside of Capacity Region

- Introduce reservoir

- Dynamic optimization over all reservoir policies.