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)

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

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

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

Constraint: $(S_{ij}(t)) \in \text{Permutation Matrix}$
Capacity Region of the Switch: 

$(2_{ij})$ such that:

\[ \sum_{j} 2_{ij} \leq 1 \quad \text{for all inputs } i \]

\[ \sum_{i} 2_{ij} \leq 1 \quad \text{for all outputs } j \]
Talk:

I. Fundamental Delay Bounds
   [Neely, HPSR 2004]
II. Delay - Complexity Tradeoffs
    [Neely, CISS 2002]
III. Fairness and Optimal Flow Control for
     Switches and Networks
      [Neely, Ph.D. thesis 2003]

Start with delay...
Outline

1. Backlog Unaware Algs.: \( \text{Delay} \geq O(N) \)
   (Periodic, Randomized scheduling based on known rates \( x_{ij} \))
   - Chang et al. (Infocom 2000, 2003)
   - Koksal (MIT Thesis 2002)
   - Andrews and Vojnovi\'c (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:

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

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
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 \((\lambda_{ij})\) matrix has \(O(N^2)\) entries \(\lambda_{ij} = O\left(\frac{2 \lambda N}{n}\right)\).

\[ \begin{pmatrix}
\lambda_1 & \lambda_{12} & \cdots & * \\
\lambda_{12} & \lambda_2 & \cdots & * \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_{n1} & \lambda_{n2} & \cdots & \lambda_n
\end{pmatrix} \]

Ex: Uniform traffic

\[ \lambda_{ij} = \frac{\rho}{n}, \quad \lambda = \rho \]
Figure C-1: An illustration proving the Jitter Theorem.
Backlog Aware Scheduling

Fact 1: Minimum Clearance Time $T^*$

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

($\text{Birkhoff-Von Neumann Thm., Hall's Thm}$)

How?

A. Augment:

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

$\begin{pmatrix}
2 & 0 & 1 \\
1 & 1 & 1 \\
0 & 2 & 0 \\
\end{pmatrix} \rightarrow 
\begin{pmatrix}
1 & 0 & 1 \\
1 & 0 & 1 \\
0 & 2 & 0 \\
\end{pmatrix} \rightarrow 
\begin{pmatrix}
1 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
\end{pmatrix} \rightarrow (0)$
**Fair-Frame Algorithm:** \( T \) slots = 1 frame

1. **First Frame** → Schedule \( (S_{ij}(t)) \) Randomly.

2. **On Frame** \((k+1)\):
   - Have \((L_{ij}(kT))\) = Arrivals from Prev. Frame
   - **Fair Decomposition:**
     \[
     (L_{ij}) = (L'_{ij}) + (V_{ij})
     \]
     - 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

(i,j) → Packet

→ Conforming

→ overflow bin (i,j)

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

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

- \(\text{avg. delay} \leq O(\log(N))\)
- \(T \leq O(\log(N))\)
- \(P_{\text{conform}} \geq 1 - O(1/\sqrt{N})\)
Proof Idea:

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

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

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

must have been violated.

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

(\[ \delta = \rho e^{1-\eta} \])

Union bound: $\Pr[\text{any violated}] \leq 2N\delta^T$
MWM: $O(N)$ analytical bound.
Fair Frame: $O(\log(N))$ analytical bound.
- Suite of MWM-based algorithms over frames:
  
  Complexity = $O(N^\alpha)$ \(\exists\) any $\alpha$ s.t.
  
  Delay = $O(N^{4-\alpha})$ \(0 \leq \alpha \leq 3\)

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

- How does Fair-Frame compare?

  Complexity = $O(N^{1.5} \log(n))$
  
  Delay = $O(\log(N))$
Transition Slide

Lower Complexity thru extra hardware

\[ \begin{array}{c}
    N \times N \\
    \vdots \\
    \rightarrow \\
\end{array} \quad \begin{array}{c}
    N \times N \\
    \rightarrow \\
\end{array} \]

“Uniformization”

“Zero-Complexity” randomized or periodic scheduling leads to 100% throughput.

Chang 2002
Koksal Ph.D. thesis 2002

\[ O(1) \] complexity

\[ O(N) \] delay

What if \((2ij)\) outside capacity?
III. Outside the Capacity Region

Q: Where to put the flow control?

Prefer to have near inputs ble:

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 $g_{ic}(r) = \text{"Satisfaction" user } i \text{ has when send at long-term rate } r.$

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

St.: $r_{ic} \leq \lambda_{ic}$
$(r_{ic}) \in \Lambda$

Capacity region
Algorithm:

Special case: \( g_{ic}(r) = \Theta \cdot r \)  (linear utility)

Flow Control:

\[ X_{1,i} \rightarrow \text{Node 1} \]
\[ X_{2,j} \rightarrow \text{Node 1} \]

* Send packet if:

\[ L_{1i}(t) \leq V \Theta_{1i} \]

(\text{where } V = \text{parameter of control})

* Then choose \( i \) such that \( i = \text{argmax}[V \Theta_{1i} - L_{1i}(t)] \)

In-Network Packet Selection

Differential Backlog Maximizer

\[ \theta = \text{argmax} \left\{ L_{a}(t) - L_{b}(t) \right\} \]
**Thm.** (Assume i.i.d. Bern. inputs $\lambda_{ij}$)

Utility: $E g_{ie}(F_{ic}) \geq E g_{ic}(F_{ic}^{opt}) - \frac{NB}{V}$

Delay: $E T_{ie} \leq \frac{NB}{4\sigma_{sym}} + \frac{\sqrt{g_{\max}} N}{4\sigma_{sym}}$

**Example:**

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

$E \Theta_{ie} F_{ic} \geq E \Theta_{ic} F_{ic}^{opt} - \frac{10 N}{V}$

Avg. Delay $\leq 5N + \frac{NV \Theta_{\max}}{2}$
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 overall reservoir policies.