Flexible Bandwidth Provision in a Sectored Packet Switch with an Optical Core

Sofia A. Paredes, Trevor J. Hall

Photonic Network Technology Laboratory
Centre for Research in Photonics
School of Information Technology and Engineering
University of Ottawa, Canada

The 10th IEEE International Symposium on Computers and Communications, ISCC 2005
Cartagena, Spain, 27th-30th June, 2005
Outline

• Packet Switch Architectures

• Use of Optical Switching Fabrics

• Architecture of the sectored packet switch with an optical core

• Flexible Bandwidth Provision (FBP)

• Simulation Results

• Conclusions & Remarks
Typical Packet Switch Architecture

• Linecard buffers interconnected by electronic crossbars configured every timeslot (~12 ns memory access time) by a very fast centralized scheduler.
• Crossbar consumes kW of power and the scheduler is the bottleneck to increased line rates.

(Nick McKeown, “Packet Switch Architectures”, lecture at Stanford University)
Use Of Optical Switch Fabrics

• Provides \textit{drastic reduction of power dissipation} density
  – transparent optical core
  – optoelectronic transceivers distributed on linecards

• Further reductions in cost and power dissipation are expected as very-short-reach optical interconnect technologies mature

• Main concern:
  – Large port counts available only at low switch speeds
  – High switching speed only available at low port counts

• Is fast reconfiguration of a transparent optical switch really necessary?
Architecture of the Sectored Packet Switch

- Clos-like architecture

- Groups of traffic ports called sectors, designed as shared memory packet switch modules. Reduction of:
  - memory access contention
  - memory bandwidth requirements
  - complexity requirements

- Interconnection of sectors (or linecards) use reconfigurable photonic technology (ex. liquid crystal, fast MEMs, SLMs)
The concept of Flexible Bandwidth Provision

• The aim is to provide variable bandwidth amongst sectors according to the traffic demand

• The process of *Flexible Bandwidth Provision* (FBP) involves the binding together of several lower rate circuits to form a higher rate logical circuit of variable capacity at a fine granularity

• This is achieved by reconfiguring the switch:
  – state of the central crossbar switches
  – reading operations in the input sectors
Input sectors:
\[ Q_{i,j} \] is the VOSQ for traffic going from input sector \( i \) to output sector \( j \)

Output sectors:
\[ Q_q \] is the dedicated queue for traffic going to output \( q \)

16x16 switch: \( N = n \times l = 16, n=4, l=4, m=8 \)
speed up: \( k = m/n = 2 \)
A traffic matrix is calculated whose elements correspond to the measured/estimated traffic arrivals or requested bandwidth between two sectors.

From the traffic matrix, an inter-sector service rate matrix $S$ is built whose integer entries specify the number of paths to be set up between sector pairs. By varying the number $s_{ij}$, FBP is achieved.

Traffic from input sector $i$ to output sector $j$

$$\Lambda_e = [\lambda_{ij}]$$

Example for 2 sectors:

$$\Lambda_e = \begin{bmatrix} 14 & 35 \\ 43 & 5 \end{bmatrix}$$

The configuration of the central stage is found by solving an edge-colouring problem on a bipartite graph defined by $S$. Each permutation defines the state of a crossbar switch.
A traffic matrix is calculated whose elements correspond to the measured/estimated traffic arrivals or requested bandwidth between two sectors.

From the traffic matrix, an inter-sector service rate matrix $S$ is built whose integer entries specify the number of paths to be set up between sector pairs. By varying the number $s_{ij}$, FBP is achieved.

Example for 2 sectors and 4 crossbars:

$$S = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

Must have a high correlation with the traffic matrix $\Lambda_e$.

The configuration of the central stage is found by solving an edge-colouring problem on a bipartite graph defined by $S$. Each permutation defines the state of a crossbar switch.
A traffic matrix is calculated whose elements correspond to the measured/estimated traffic arrivals or requested bandwidth between two sectors.

\[
S = \begin{bmatrix}
1 & 3 \\
3 & 1 \\
\end{bmatrix}
\]

Decomposition of \( S \):

\[
p_0 = \begin{bmatrix}
1 & 0 \\
0 & 1 \\
\end{bmatrix}, \quad p_1 = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
\end{bmatrix}, \quad p_2 = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
\end{bmatrix}, \quad p_3 = \begin{bmatrix}
0 & 1 \\
1 & 0 \\
\end{bmatrix}
\]

From the traffic matrix, an inter-sector service rate matrix \( S \) is built whose integer entries specify the number of paths to be set up between sector pairs. By varying the number \( s_{ij} \), FBP is achieved.

The configuration of the central stage is found by solving an edge-colouring problem on a bipartite graph defined by \( S \). Each permutation defines the state of a crossbar switch.
Transient response example

- 64x64 switch \( \{n=8, l=8, m=16, \text{ speed-up}(k)=m/n=2\} \), self-similar traffic
- \( \text{tau} = \text{time between reconfigurations (in units of time slots)} \)
- The queues in the input sectors are emptied when the reconfiguration of the switch is changed to adapt to the new traffic pattern
Average delay and delay relative to IOQS

- 64x64 switch \(\{n=8, l=8, m=16, \text{ speed-up}(k)=m/n=2\}\), self-similar traffic

- A closer inspection of the average delay is achieved by making the IOQS delay the horizontal reference \((y=0)\)

- This shows that at 95% load the delay is only 6, 10 and 12 timeslots higher than the IOQS for \(\tau=100, 200\) and \(300\), respectively
Individual delays relative to IOQS

- 64x64 switch \( \{n=8, l=8, m=16, \text{speed-up}(k)=m/n=2\} \), self-similar traffic, 80% load

- The average delay only tells “half of the story”
- A histogram (or probability density function) of the *individual* delays relative to the IOQS reveals that the delay variability is small as well
Is fast reconfiguration of a transparent optical switch necessary?

- Packets are switched in the sectors (linecards) when they are appended to a specific VOQ.
- The core switch provides the connectivity required to transport packets between linecards.
- Therefore the interconnect may be static provided it offers sufficient bandwidth to serve the inter-linecard traffic demand.

- Delay performance is close to that of the “ideal” output queued switch, even for long times between reconfigurations (tau=100, 200, 300 instead of tau=1), using a modest internal spatial speed-up factor of 2

- ns switching speed is not needed: Optical technologies supporting switching speeds in the μs timescale can be used with the FBP architecture (e.g. spatial light modulators)
FBP allows use of not so fast switches

- Time between reconfigurations can be large (in the 100s) even for the most demanding non-stationary/bursty traffic (ideal is \( \tau = 1 \))

- A simple mechanism to trigger reconfiguration could be a defined threshold value of the ratio of input stage queueing to total queueing
  - the aim would be to bring this ratio below 50% and ideally very close to 0%
Conclusions (1)

- This Clos-like switch architecture allows for a partition of labour and capitalizes on the strengths of the two technologies:
  - electronics for switching, buffering and packet processing
  - photonic core for transport

- A per-timeslot scheduler is not warranted:
  - optical switching technologies that are relatively slow (~µs) can be used

- The switch successfully adapts its internal bandwidth to the requirements of incoming traffic using a very simple scheduler

- Delays comparable to the “ideal” output queued switch, achieved at the price of a modest internal spatial speed-up of two
Conclusions (2)

- The central stage mainly transports packets; most of the switching occurs in the sectors:
  - *coarse* switching in the input stage and *fine* switching in the output stage

- Switch configuration overhead may need to be compensated by increasing the internal speedup (spatial or temporal) by a factor equal to:
  \[
  \text{reconfiguration time} / \text{interconfiguration period}
  \]

- This architecture and method allow for interconfiguration periods measured in 100s of time slots
  - the use of switches with reconfiguration times of the same order is possible without the need for excessive speedup.
Remarks

• Switch architecture is scalable to a large number of port counts and high line rates:
  – a 256x256 switch @ 2.5Gbps per port is already feasible using small 16x16 optical switching elements with 10 µs reconfiguration times
  – the total capacity is 640 Gbit/s (80Gbit/s memory, 512-bit packets)

• The application of the FBP paradigm to the next level in the switch hierarchy could similarly adapt the service rate of the output queues and reduce overall delay in a larger scale network

• The switch architecture maps neatly onto the Agile All-Photonic Networks (AAPN) architecture
Thank you for your attention...