Scheduling in Input-Queued Cell-Based Switches

Marco Ajmone Marsan
Andrea Bianco
Paolo Giaccone
Emilio Leonardi
Marco Mellia
Fabio Neri

MQOS Workshop – 12-14 gennaio 2000, Courmayeur

Switch Architectures

Input vs. Output Queueing Switches

Output Queueing
⇒ switching fabric and output queues operate at a speed equal to the sum of the rates of all input lines. Speed-up=\( N \).

Input Queueing
⇒ switching fabric and output queues operate at a speed equal to the rate of the fastest input line. Speed-up=\( 1 \).

BUT

Combined Input Output Queueing Switches

Combined Input Output Queueuing
⇒ internal interconnect (switching fabric) and output queues operate at a speed ranging from the line rate to the sum of the rates of all input lines. \( 1 \leq \text{Speed-up} \leq N \).

CIOQ
⇒ performance reduction due to HoL (Head-of-the-Line) blocking in the case of a single FIFO queue per input interface.
⇒ to decide which input ports are enabled to transmit, a scheduling algorithm and internal signalling are needed.
Virtual Output Queueing Architectures

To overcome the HoL blocking ⇒ VOQ (Virtual Output Queueing) or Destination Queuing schemes.

IQ schemes achieve 100 % throughput

It has been shown that [1]:

IQ switches achieve 100 % throughput for unicast traffic, if

- input traffic is admissible: i.e., \( \sum_j \lambda_{ij} < 1 \) \( \forall i \), and \( \sum_i \lambda_{ij} < 1 \) \( \forall j \)
- a VOQ (Virtual Output Queueing) architecture is adopted
- buffer size is infinite
- a MWM (Maximum Weight Matching) scheduling algorithm is adopted, whose weights are either the queue lengths or the head of the queue cell ages

The MWM computational complexity is \( O(N^3 \log N) \).


Matching

IQ Cell Mode Schedulers

Several heuristic scheduling algorithms for IQ cell switches were proposed in the literature.

Performance can be very close to the more hardware-demanding OQ architecture.

We considered some IQ CM scheduling algorithms for IQ switches: ⇒

- we proposed a new cell scheduling algorithm
- we provided a taxonomy
- we (trivially) extended algorithms to a variable packet size environment
- we studied performance by simulation
- we evaluated computational complexity
CIOQ Switches

CIOQ switches require a speed-up ranging from 1 to \( N \).

It is known that CIOQ switches with speed-up \( \leq 2 \) can “emulate” the performance of OQ switches.

We focused on the stability of a general class of cell scheduling algorithms in CIOQ architectures.

Different stability definitions can be used.

CIOQ switches were studied using the Lyapunov function.

IQ Schedulers: Taxonomy

Each algorithm is characterized by:

- metric computation
- heuristic matching algorithm
- contention resolution

Metric computation: Queue Occupancy, Queue Length, Cell Age, MUCS metric (derived from queue occupancy)

Heuristic matching algorithm: Iterative Search, Matrix Greedy, Reservation Vector

Contention resolution: Random, Round Robin (state-dependent), Sequential Search (state-independent)

IQ Cell Schedulers - Complexity

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Additions and subtractions</th>
<th>Products and divisions</th>
<th>Comparisons</th>
<th>Asymptotic complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSLIP</td>
<td>–</td>
<td>–</td>
<td>( 3N^2 \log_2 N )</td>
<td>( O(N^2 \log_2 N) )</td>
</tr>
<tr>
<td>iQF</td>
<td>–</td>
<td>–</td>
<td>( 3N^2 \log_2 N )</td>
<td>( O(N^2 \log_2 N) )</td>
</tr>
<tr>
<td>iOCF</td>
<td>–</td>
<td>–</td>
<td>( 3N^2 \log_2 N )</td>
<td>( O(N^2 \log_2 N) )</td>
</tr>
<tr>
<td>iLPF</td>
<td>( 2N^2 )</td>
<td>–</td>
<td>( N^2 + 2N \log_2 N )</td>
<td>( O(N^2) )</td>
</tr>
<tr>
<td>2DRR</td>
<td>–</td>
<td>–</td>
<td>( N^2 )</td>
<td>( O(N^2) )</td>
</tr>
<tr>
<td>RC</td>
<td>–</td>
<td>–</td>
<td>( \frac{N^2}{2} )</td>
<td>( O(N^2) )</td>
</tr>
<tr>
<td>MUCS</td>
<td>( \approx \frac{N^3}{2} )</td>
<td>–</td>
<td>( \frac{N^3}{2} )</td>
<td>( O(N^3) )</td>
</tr>
<tr>
<td>RPA</td>
<td>( N^2 )</td>
<td>–</td>
<td>( 2N^2 )</td>
<td>( O(N^2) )</td>
</tr>
<tr>
<td>iZIP</td>
<td>–</td>
<td>–</td>
<td>( 2N^2 )</td>
<td>( O(N^2) )</td>
</tr>
</tbody>
</table>

Packet Switch

- IP router
- ATM switching engine
- Layer 3 switches
Scheduling in IQ Cell-Based Switches

At each input, an ISM (Input Segmentation Module) segments incoming packets into cells; store and forward operation:

- memory to store a maximum-size packet is required to accept a packet
- segmentation starts only after the complete reception of the packet

At each output, an ORM (Output Reassembly Module) reassembles the packets:

- cells belonging to different packets can be interleaved ⇒ more than one reassembly machine can be active in the same ORM
- at most one cell reaches each ORM in a cell-time ⇒ at most one packet is completed at each ORM in a cell-time

A packet FIFO stores reassembled packets waiting for transmission on the output line.
IQ Packet Mode Schedulers

Adaptation of IQ CM (Cell Mode) schedulers to deal with variable-size packets
⇒ scheduling algorithms constrained to deliver contiguously all the cells deriving from the segmentation of the same packet:
• variable-size packets transformed into ‘trains of cells’
• cells belonging to the same train are transferred to the output on contiguous cell-times, without interleaving with cells of another train.

Advantages of Packet Mode Scheduling

If a packet mode scheduler is used in the switching fabric, the ORM modules are not necessary any longer.

Simulation Results - Source Traffic Model

The arrival of IP datagrams at the input of the packet switch is not explicitly represented.
We instead model the arrival of cell bursts at the input of the internal cell-switch.
Cell bursts are assumed to be originated by the segmentation of a packet.
The cell arrival process at each input is modeled by a two-state model:
• **ON state.** In this state, a packet is being received; packet size in cells is a discrete random variable uniformly distributed between 1 and 192.
• **OFF state.** In this state no cells are generated. The average idle period duration is set so as to achieve the desired input load.
Simulation Results - Traffic Patterns

We consider two types of input/output flows:

**Uniform traffic.**
All input interfaces are equally loaded; at each input, the same amount of traffic is directed to every output.

**Hot-spot traffic.**
All input interfaces are equally loaded, but one of the outputs receives twice as much traffic as any other.

Simulation Results - Performance Indices

**Cell delay.** Time spent by cells in the cell-switch queues.

**Packet delay.** Overall delay of a packet, considering the ISM module, the internal cell-switch queues, the ORM module, and the final packet FIFO.

**Throughput.** Ratio between the total number of cells forwarded to output interfaces, and the total number of cells arrived at input interfaces.

**Maximum number of reassembly machines.** Maximum number of re-assembly machines simultaneously active, i.e. assembling packets, in all the \( N \) ORM modules.
Throughput with CM Scheduling

Uniform Traffic

Packet Delay with PM Scheduling

Uniform Traffic

Packet Delay with Optimized PM Scheduling

Uniform Traffic
Analytical results for CIOQ schemes

If we consider a cell-based switch, with a VOQ architecture at input ports,

- a MWM scheduling algorithm with weights proportional to the queue length is stable with speed-up \( \geq 1 \Rightarrow 100\% \) throughput can be obtained
- any scheduling algorithm belonging to the class of Maximal Size Matching is stable with speed-up \( \geq 2 \)
- many iterative scheduling algorithm with one iteration and a-priori choice for each input are stable with speed-up \( \geq 2 \)
- RPA is stable with speed-up \( \geq 2 \)

Future work

- Weighted Fair Queueing in IQ switches
- Multicast in IQ switches
- Computation of delay bounds in CIOQ architectures using the Lyapunov function