96-0520: Considerations for Frame-Level Throughput and Latency Measurements of ATM Switches

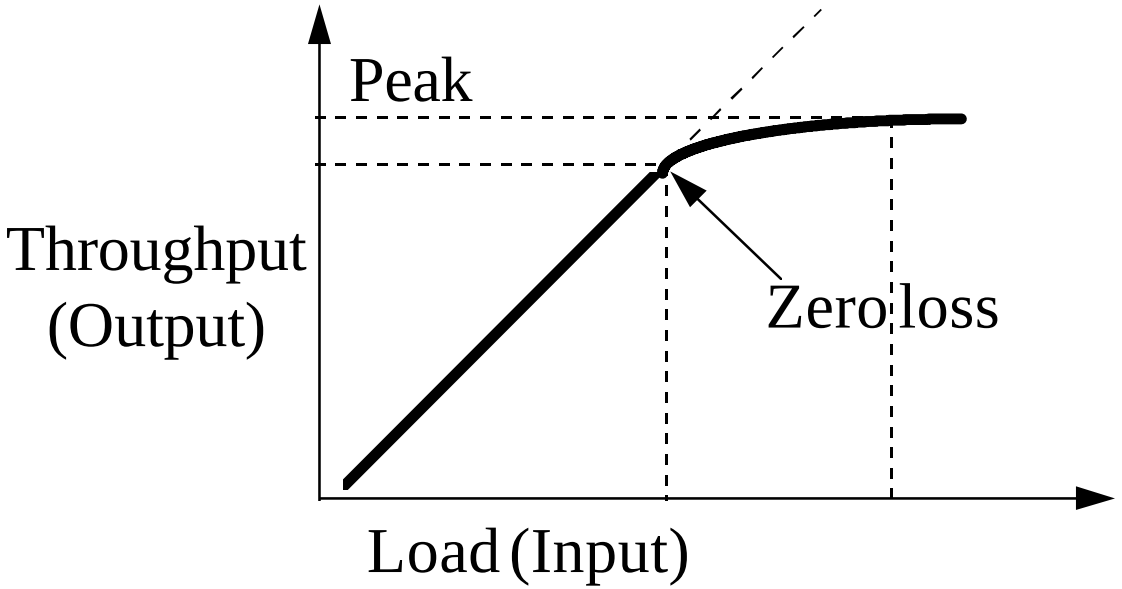
Raj Jain, Gojko Babic, Bhavana Nagendra

Raj Jain is now at
Washington University in Saint Louis
Jain@cse.wustl.edu
http://www.cse.wustl.edu/~jain/

- Throughput
- Latency

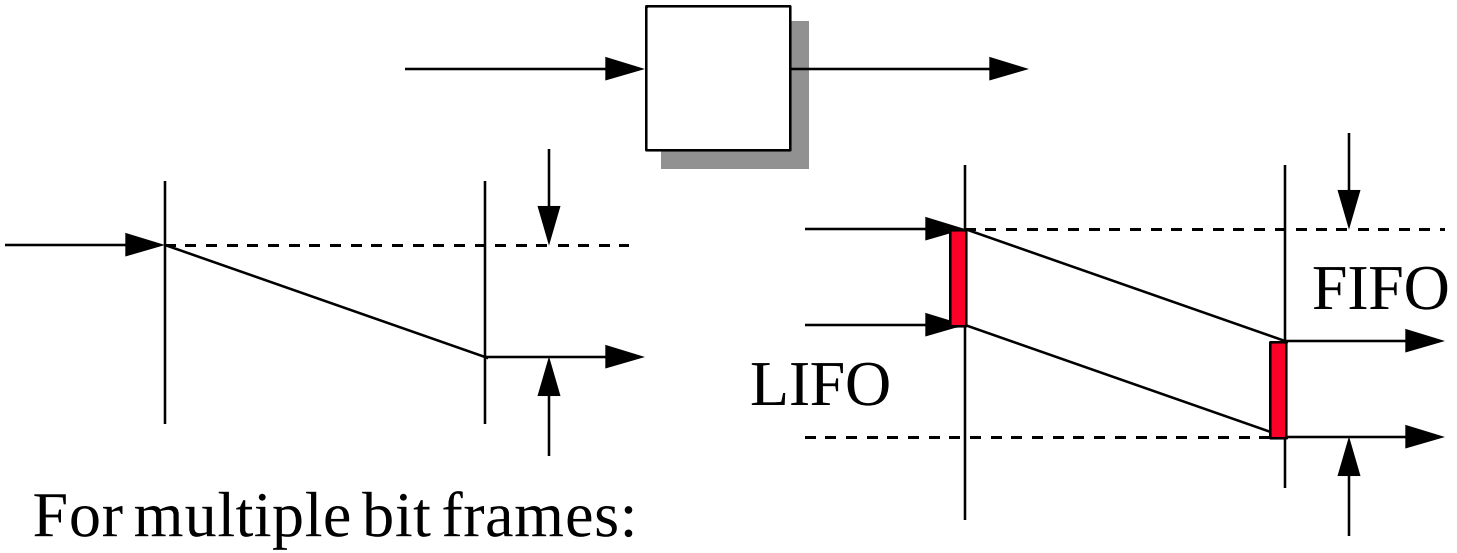
Throughput

- Zero Loss and Peak
- Unit = Frames/sec, cells/sec, or bits/sec
  bits/sec is most intuitive and
  does not require specifying a size.
- Measure With and without background (VBR)



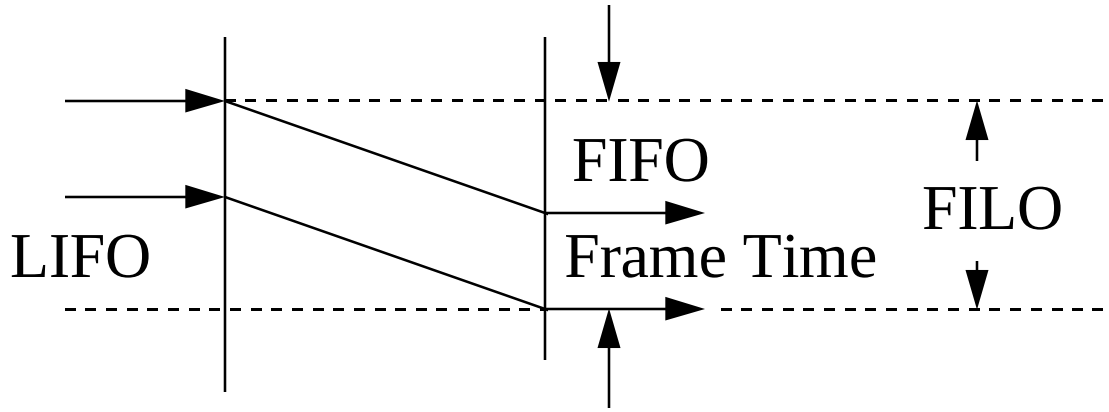
Latency

- For a single bit:
  Total latency = Bit in to bit out = Switch latency



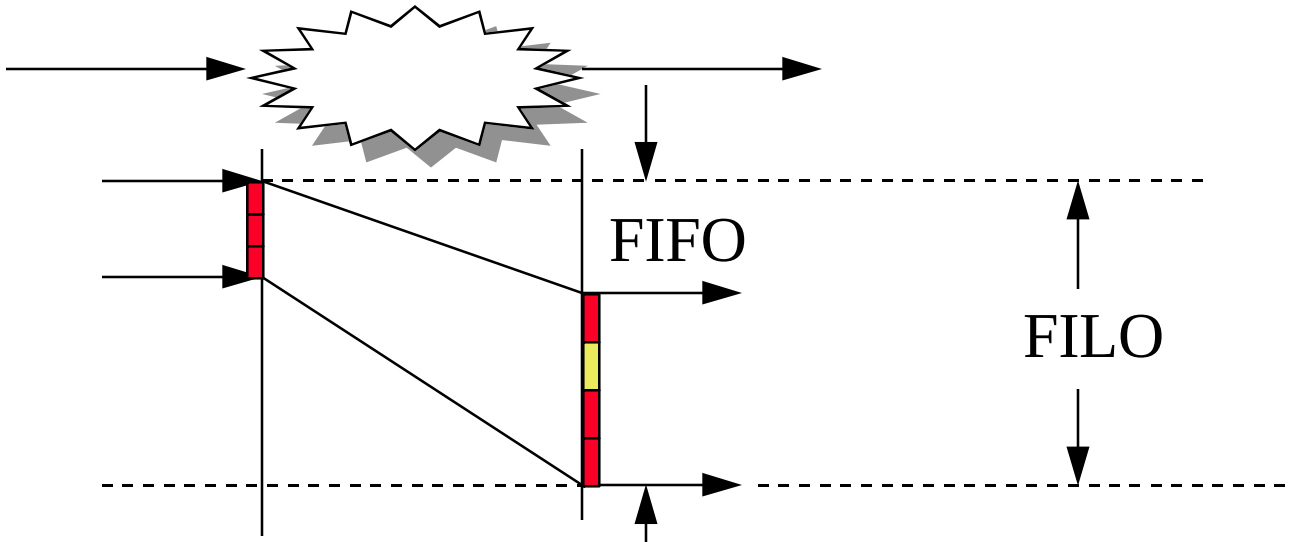
- For multiple bit frames:
  - FIFO = First bit In to First bit Out
  - LILO = Last bit In to Last bit Out
  - FILO = First bit In to Last bit Out
  - LIFO = Last bit In to First bit Out

Latency: Multiple Bit Frames



- FIFO = LILO
- FILO = FIFO + Frame time = FIFO + Frame size m/Speed C_{out}
- LIFO = FIFO - Frame time = FIFO - m/C_{out}
- Total Delay = FILO = Switch latency + Frame time
- This assumes contiguous frames
  $\Rightarrow$ No idle cells intermingled
- Also assumes input and output lines are of same speed.

Latency: General Definition

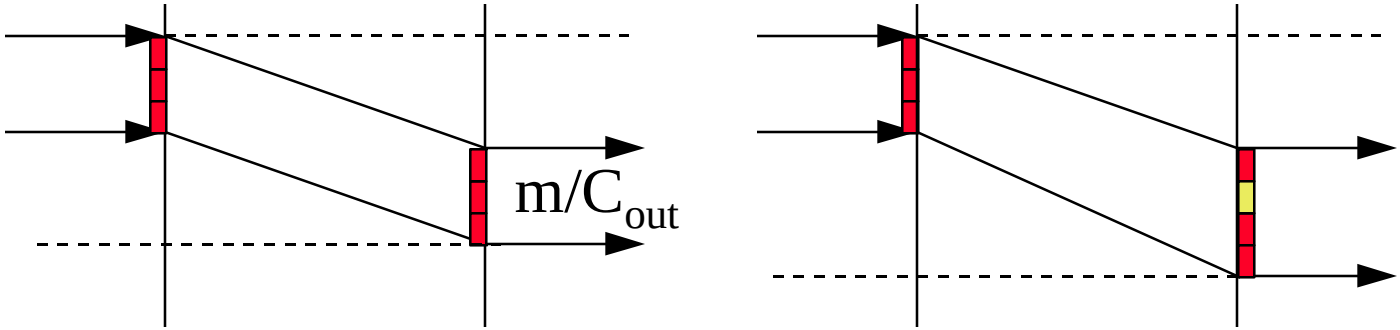


- Frames coming out of a switch (or a network) are not contiguous
- Also, Input/output links may be of different speed.
- Total delay = First bit in to last bit out = FILO
  = Switch latency + Frame time
- Switch Latency = min{LILO, FILO-m/Cout} = MIMO
- MIMO = Message in Message out

Latency Definitions: Comparison

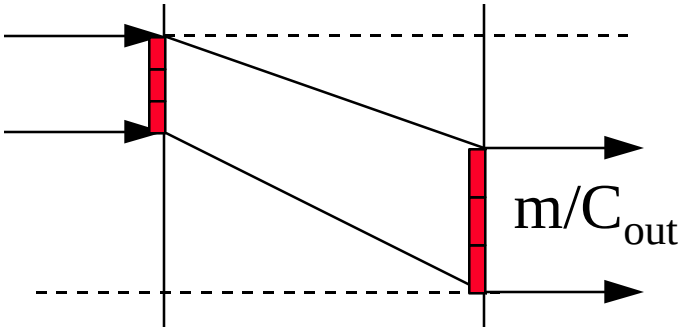
	Case	FIFO	LILO	FILO -m/C_{out}	MIMO
1a	$C_{in} = C_{out}$, Contiguous Frame	√	√	√	√
1b	$C_{in} = C_{out}$, Non-contiguous Frame	X	√	√	√
2a	$C_{in} > C_{out}$, Contiguous Frame	√	X	√	√
2b	$C_{in} > C_{out}$, Non-contiguous Frame	X	X	√	√
3a	$C_{in} < C_{out}$, Contiguous Frame, Zero Switch Latency	X	√	X	√
3b	$C_{in} < C_{out}$, Non-Contiguous Frame, Zero Switch Latency	X	√	X	√
3c	$C_{in} < C_{out}$, Contiguous Frame, Non-zero switch latency	X	√	X	√
3d	$C_{in} < C_{out}$, Non-contiguous Frame, Non-zero switch Latency	X	√	X	√

Case 1: Input = Output Speed



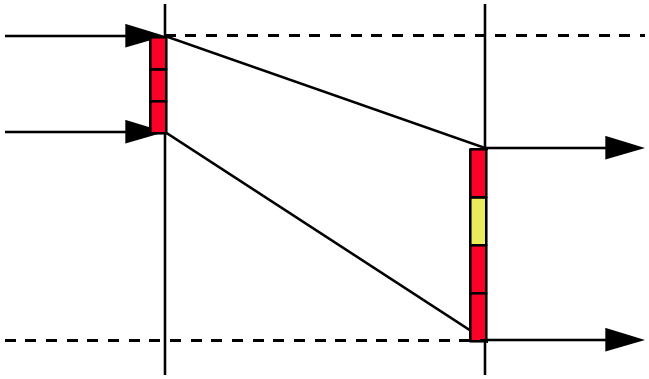
- Total Delay = FILO = Switch Latency + Frame Time
- A. Contiguous Frame on Output:
  - Switch Latency = FILO - m/C_{out} = FIFO = LILO = MIMO
- B. Non-Contiguous Frame on Output:
  - Switch Latency = total delay - frame time
  - FILO - m/C_{out} = LILO = MIMO
  - In this case, FIFO will give wrong answer.

Case 2a: Input > Output Speed Contiguous Frame on Output



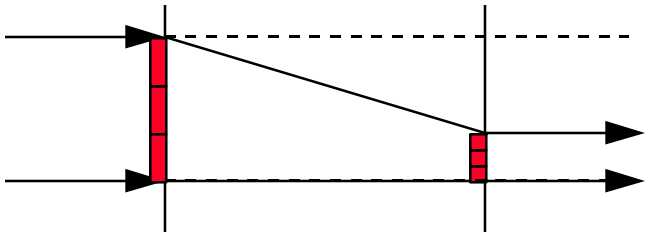
- Total Delay = FILO = Switch Latency + Frame Time
- Switch Latency = FILO - m/C_{out} = FIFO
- In this case, LILO will give wrong answer.
- LILO > FILO - $m/C_{out} \Rightarrow$ MIMO = FILO-m/C_{out}
- MIMO is also correct

Case 2b: Input > Output Speed Non-Contiguous Frame on Output



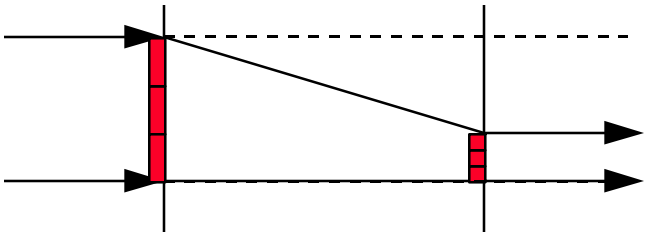
- Total Delay = FILO = Switch Latency + Frame Time
- Switch Latency = total delay - frame time = FILO - m/C_{out}
- FIFO < FILO - m/C_{out} $\Rightarrow$ FIFO is wrong
- LILO > FILO - m/C_{out} $\Rightarrow$ LILO is wrong
- MIMO = min{LILO, FILO-m/C_{out}} = FILO-m/C_{out}
  $\Rightarrow$ MIMO is correct.

Case 3: Input < Output Speed



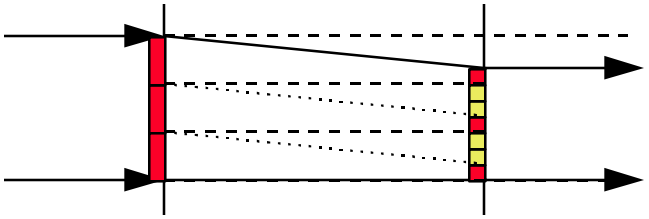
- It it difficult to get contiguous frame on output since output is much faster unless the switch stores the entire frame.
- Some of the flaws of traditional definitions can also be seen by considering a switch with zero latency.
- There are four possible cases:
  - 3a. Contiguous Frame, Zero switch latency
  - 3b. Non-contiguous frame, zero switch latency
  - 3c. Contiguous Frame, non-zero switch latency
  - 3d. Non-Contiguous Frame, non-zero switch latency

Case 3a: $C_{in}<C_{out}$, Contiguous Frame, Zero Switch Latency



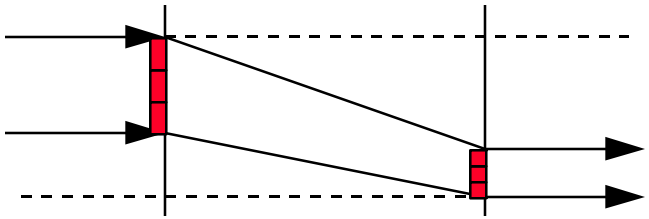
- FIFO is non-zero $\Rightarrow$ Not a correct measure of switch latency
- FILO - m/C_{out} = FIFO = non-zero. Also incorrect.
- LILO is zero. So it is correct.
- MIMO = Min{LILO, FILO - m/C_{out}} is zero.
  It is also correct.

Case 3b: $C_{in}<C_{out}$, Contiguous Frame, Zero Switch Latency



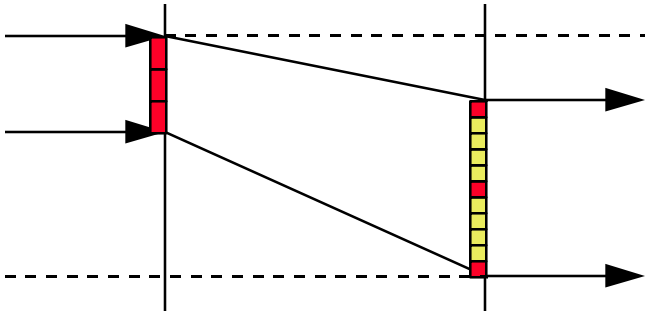
- FIFO is non-zero. So it is incorrect.
- FILO - m/C_{out} is non-zero. So it is incorrect.
- LILO is zero. So it is correct
- MIMO = Min{LILO, FILO - m/C_{out}} is zero. So it is also correct.

Case 3c: $C_{in}<C_{out}$, Contiguous Frame, Non-Zero Switch Latency



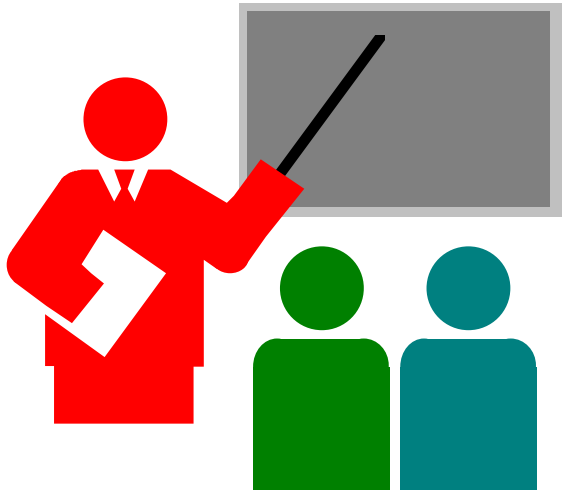
- FIFO can be made arbitrarily large by increasing the output link speed (and not changing the switch at all).
  FIFO is not an incorrect measure of switch latency.
- FILO - m/C_{out} = FIFO is similarly incorrect.
- LILO is the only metric that can be argued to be the correct measure of switch latency.
- LILO < FILO - m/C_{out}
  MIMO = Min{LILO, FILO - m/C_{out}} = LILO
  MIMO is also a correct measure.

Case 3d: $C_{in}<C_{out}$, Non-Contiguous Frame, Non-Zero Switch Latency



- FIFO can be made arbitrarily large by increasing the output link speed (and not changing the switch latency at all).
- FIFO can also be made small by sending the first cell fast but introducing idle cells later ⇒ FIFO is not correct.
- FILO - m/C_{out} > FIFO is similarly incorrect.
- LILO is the only metric that can be argued to be correct.
- LILO < FILO - m/C_{out}
  MIMO = Min{LILO, FILO - m/C_{out}} = LILO

Summary



- Throughput: Zero-loss throughput and peak throughput
- Latency = Min{LILO, FILO-m/C_{out}} = MIMO