



Flow control on IEEE 802.3x switch

Mario Baldi

Politecnico di Torino
mario.baldi[at]polito.it
staff.polito.it/mario.baldi

Pietro Nicoletti

Studio Reti
piero[at]studioreti.it
www.studioreti.it

Based on chapter 8 :

M. Baldi, P. Nicoletti, "Switched LAN", McGraw-Hill, 2002, ISBN 88-386-3426-2



Copyright Notice

- This set of transparencies, hereinafter referred to as slides, is protected by copyright laws and provisions of International Treaties. The title and copyright regarding the slides (including, but not limited to, each and every image, photography, animation, video, audio, music and text) are property of the authors specified on page 1.
- The slides may be reproduced and used freely by research institutes, schools and Universities for non-profit, institutional purposes. In such cases, no authorization is requested.
- Any total or partial use or reproduction (including, but not limited to, reproduction on magnetic media, computer networks, and printed reproduction) is forbidden, unless explicitly authorized by the authors by means of written license.
- Information included in these slides is deemed as accurate at the date of publication. Such information is supplied for merely educational purposes and may not be used in designing systems, products, networks, etc. In any case, these slides are subject to changes without any previous notice. The authors do not assume any responsibility for the contents of these slides (including, but not limited to, accuracy, completeness, enforceability, updated-ness of information hereinafter provided).
- In any case, accordance with information hereinafter included must not be declared.
- In any case, this copyright notice must never be removed and must be reported even in partial uses.



Half or full duplex?

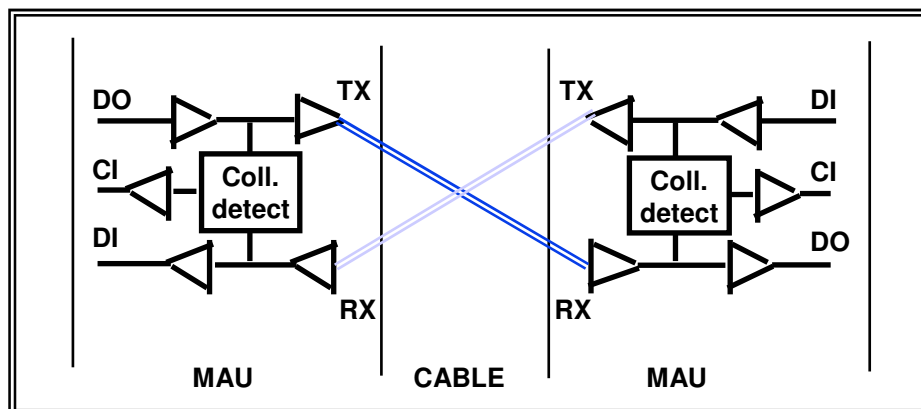
- LANs are intrinsically half-duplex:
 - Only a station at a time can transmit
- Switching strongly reduces shared medium role:
 - often transmissive medium become point to point: only a station is linked to switch
- Point to point transmissive media can be full-duplex:
 - both stations can transmit at same time
 - Transmissions take place on different physical channels



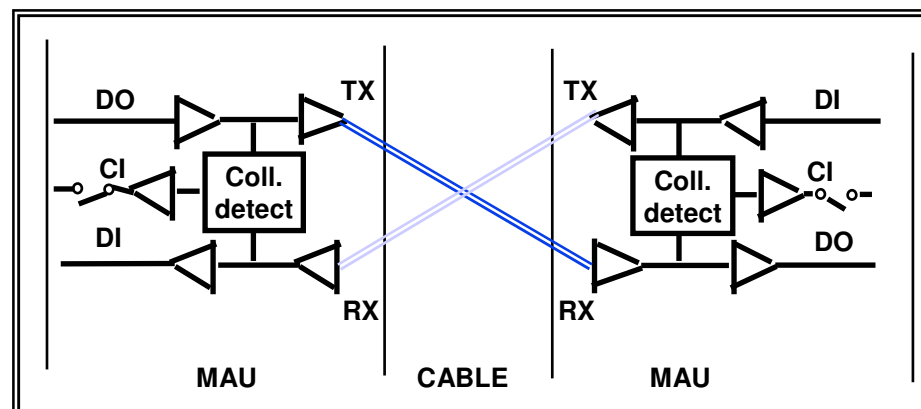
Full duplex functioning modes

- MAC CSMA-CD is no longer used
 - Packets are immediately transmitted by Ethernet stations without sensing the channel
- Always used on switch-switch links, less on switch-station links
- Special transceiver are needed because collision mustn't be detected:
 - ordinary transceiver send a collision signal to the interface when contemporary TX and RX activities are present
- Distance between two Ethernet full-duplex station
 - depends on transmissive channel features only
 - is independent from collision diameter domain

Half and full duplex transceivers



links
between
traditional
transceiver



links
between
Full-Duplex
transceiver

DO = Data Output DI = Data Input CI = Collision Input



Distance limits

- In full-duplex Ethernet :
 - 100 m for telephone twisted pair
 - 2 Km for 62.5/125 μm multimode optic fiber
 - in case of monomode optic fiber and transceiver equipped with category II laser (as in FDDI and Fast-Ethernet), the maximum distance can be 50 Km
 - in Gigabit Ethernet with high power laser:
 - 75 Km for monomode fiber
 - 100 Km for dispersion-shift monomode fiber



Flow control: IEEE 802.3x

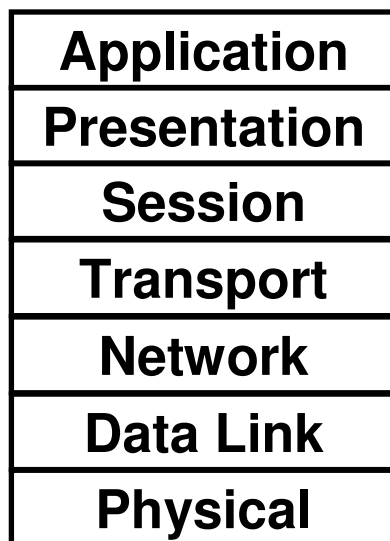
- Define:
 - MAC IEEE 802.3 necessary changes in order to support Full-Duplex mode
 - flow control mechanism for Full-Duplex links
 - available for all Ethernet networks (10/100/1000 Mb/s)
- Mandatory for Gigabit Ethernet, optional for Ethernet and Fast Ethernet



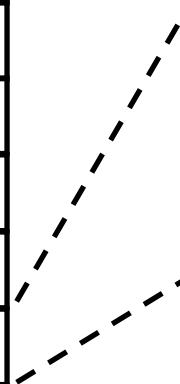
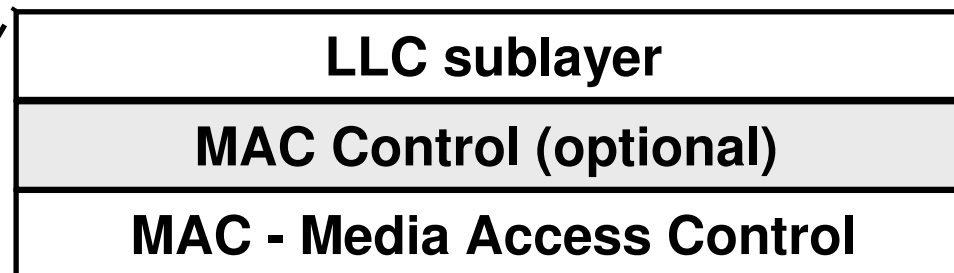
OSI model and IEEE 802.3x

- IEEE 802.3x introduce a sublayer (**MAC Control**) between MAC 802.3 and higher sublayer (Bridge Relay Entity, LLC, ...)

ISO/OSI model



LAN 802 sublayers



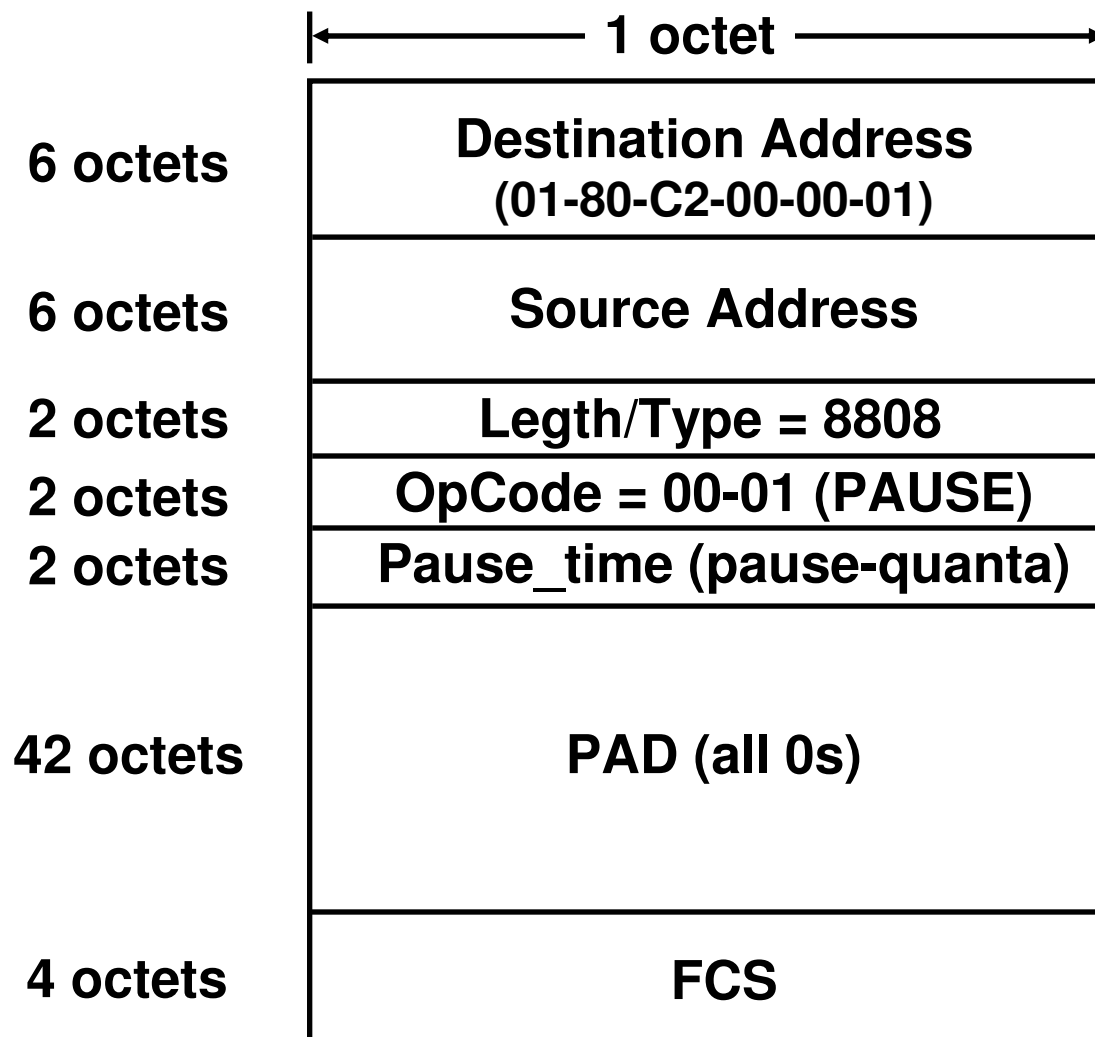


Flow control

- Control flow mechanism is defined in IEEE 802.3x:
 - The device willing to stop transmission send a **PAUSE** packet in multicast to every partner concerned in the trasmission
 - packet contains the amount of time (numero of slot time) that each partner must stop transmission
 - the times can be extended or aborted by sending another PAUSE packet



PAUSE packet





Pause time

- Pause_time field: number of pause-quanta (from 0 to 65535) which indicate the pause time
 - pause-quanta = 512 bit time
 - speed equal or less than 100 Mb/s
 - T-Pause in bit time = pause-quanta * 512
 - Speed greather than 100 Mb/s
 - T-Pause in bit time = pause-quanta * 512 * 2

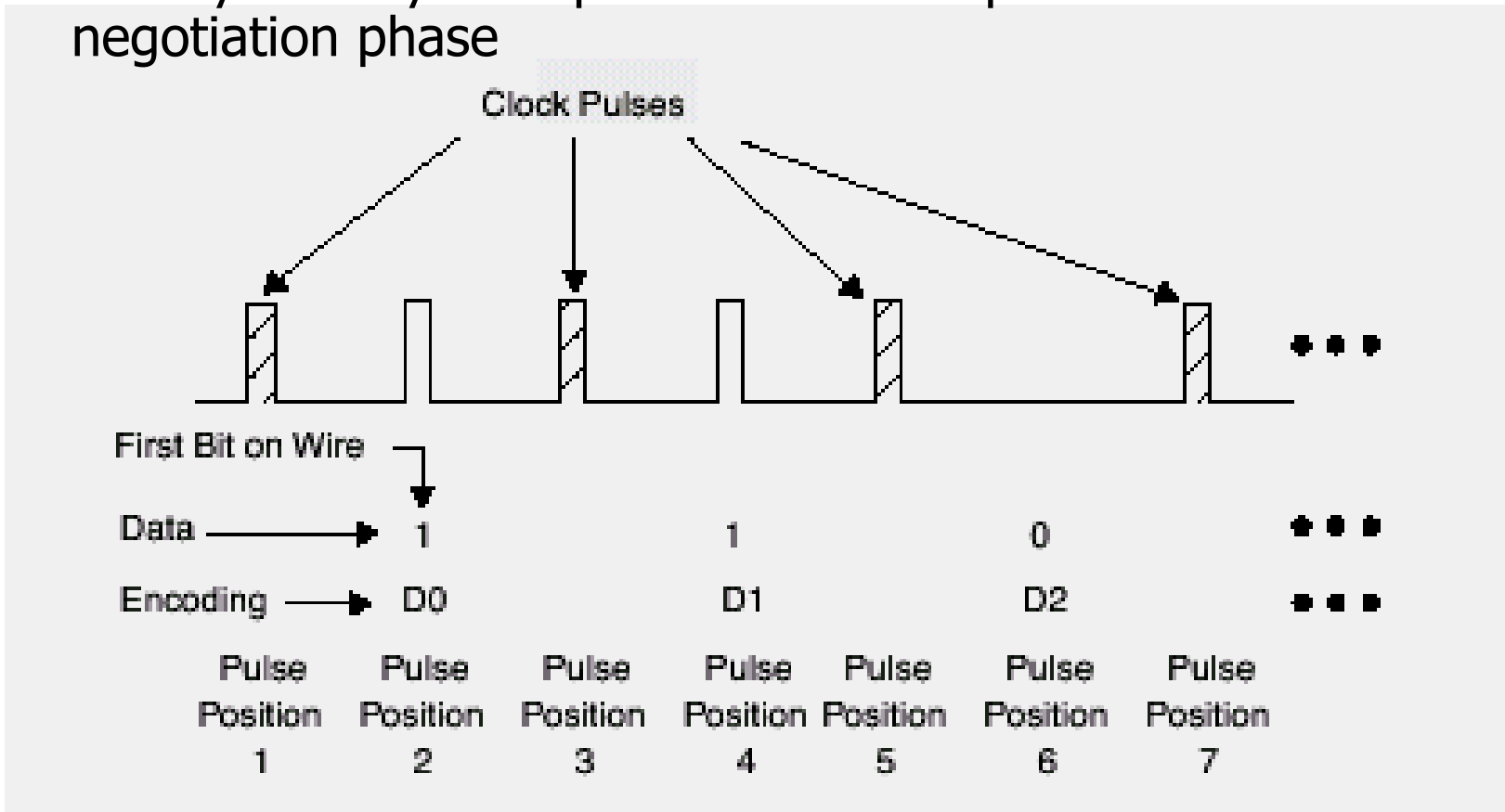


IEEE 802.3x: flow control modes

- Two flow controls mechanism:
 - **asymmetric** mode
 - only one equipment send pause packet, the other just receive the packet and stop transmitting
 - **symmetric** mode
 - both equipment at link's edge can transmit and receive the pause packet

Flow Control negotiation

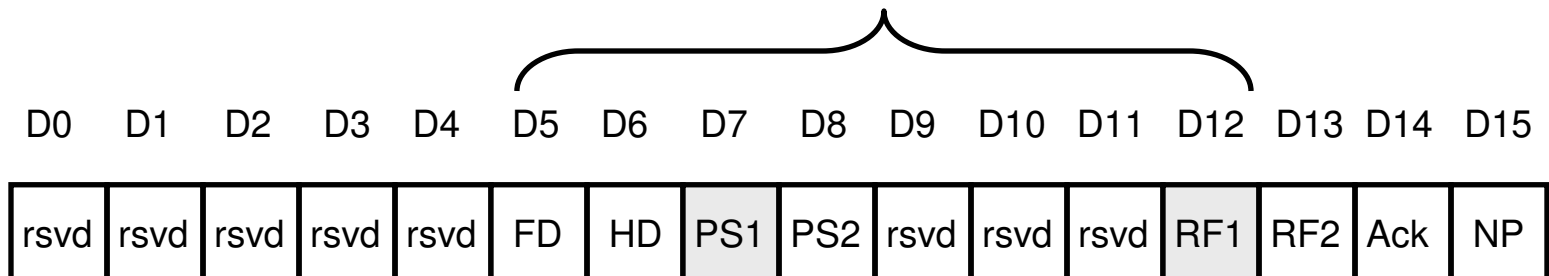
- Flow control negotiation using Burst FLP (Fast Link Pulse) coding
 - initially sent by both partners in link's parameter negotiation phase





Flow control: FLP Reg.4 encoding

Technology Ability Field



0	0	No Pause
0	1	Asymmetric Pause to link partner
1	0	Symmetric Pause
1	1	Symmetric Pause e Asymmetric Pause (tipo "Both") to local device

0	0	No error, Link OK
0	1	Offline
1	0	Link Failure
1	1	Auto-Negotiation Error



Flow control: FLP Reg.4 encoding

Register 4: Auto-Negotiation Advertisement Register

Bit(s)	Name	Description	Default	R/W
15	Next Page	Constant 0 = page transmission with primary capacity	0	RO
14	Reserved	Reserved. Must be set to 0	0	RO
13	Remote Fault	1 = malfunctioning at link's opposite side 0 = nessun malfunzionamento	0	RW
12:5	Technology Ability Field	8 bit field containing info on technologies' specific functionalities identified by selector field	00101111	RW
4:0	Selector Field	5 bit field which identifies the kind of message sent for the negotiation. In the Intel 82559 circuitry this field is read-only and contains 00001b value which stand for IEEE 802.3 standard.	00001	RO

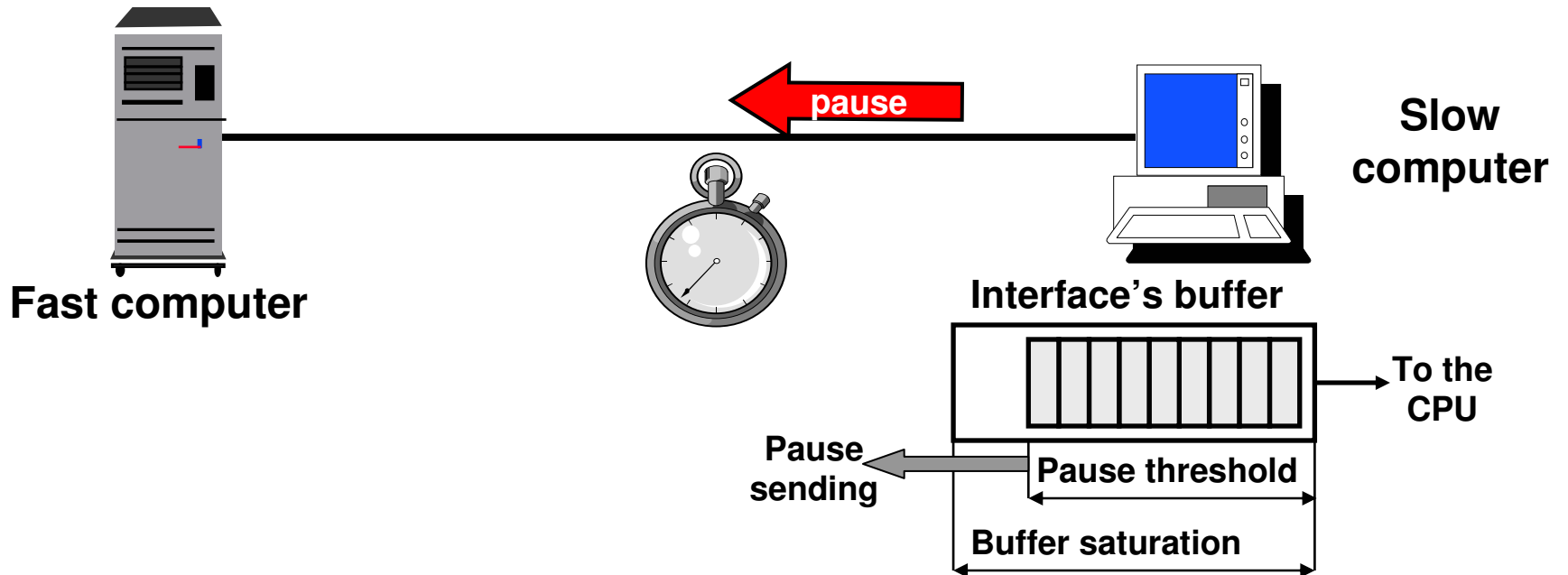
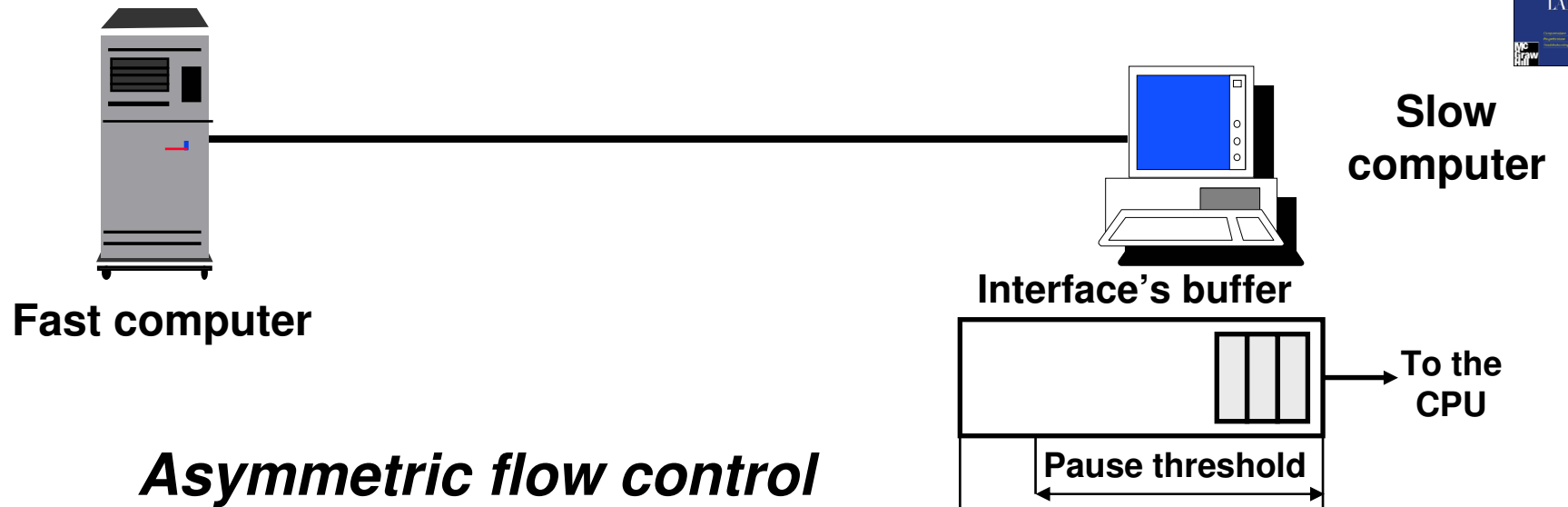


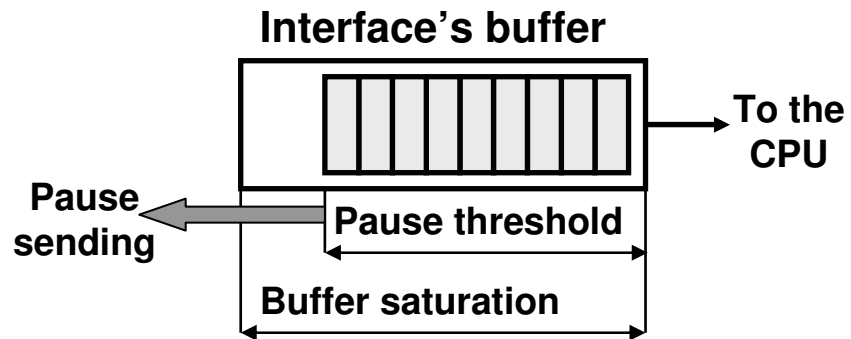
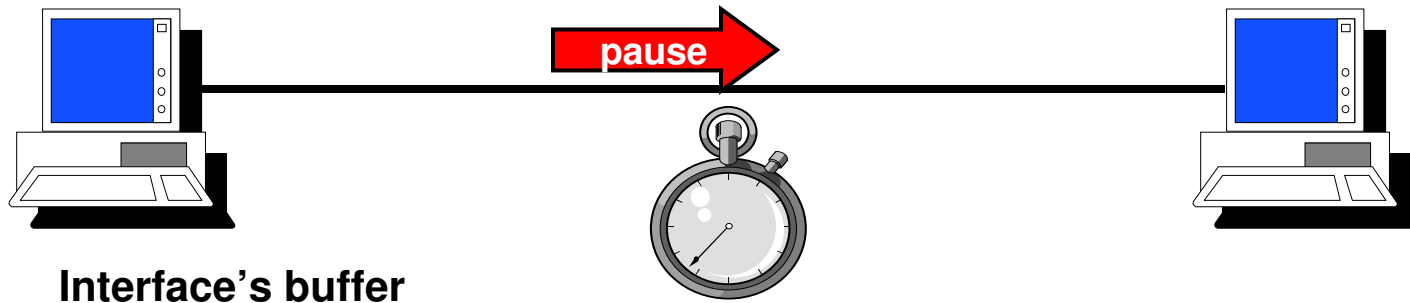
Local Device		Link Partner		Local Device resolution	Link Partner resolution
Bit - PS1	Bit - PS2	Bit - PS1	Bit - PS2		
0	0	Don't care	Don't care	Disable PAUSE TX & RX	Disable PAUSE TX & RX
0	1	0	Don't care	Disable PAUSE TX & RX	Disable PAUSE TX & RX
0	1	1	0	Disable PAUSE TX & RX	Disable PAUSE TX & RX
0	1	1	1	Enable PAUSE TX Disable PAUSE RX	Enable PAUSE RX Disable PAUSE TX
Asymmetric PAUSE		Both Sym & Asym Pause			
1	0	0	Don't care	Disable PAUSE TX & RX	Disable PAUSE TX & RX
1	Don't care	1	Don't care	Enable PAUSE TX & RX	Enable PAUSE TX & RX
Symmetric or Both		Symmetric or Both			
1	1	0	0	Disable PAUSE TX & RX	Disable PAUSE TX & RX
1	1	0	1	Enable PAUSE RX Disable PAUSE TX	Enable PAUSE TX Disable PAUSE RX
Both Sym & Asym Pause		Asymmetric PAUSE			



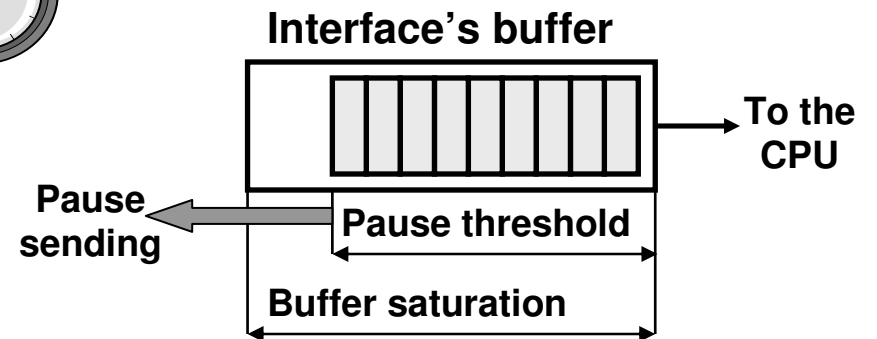
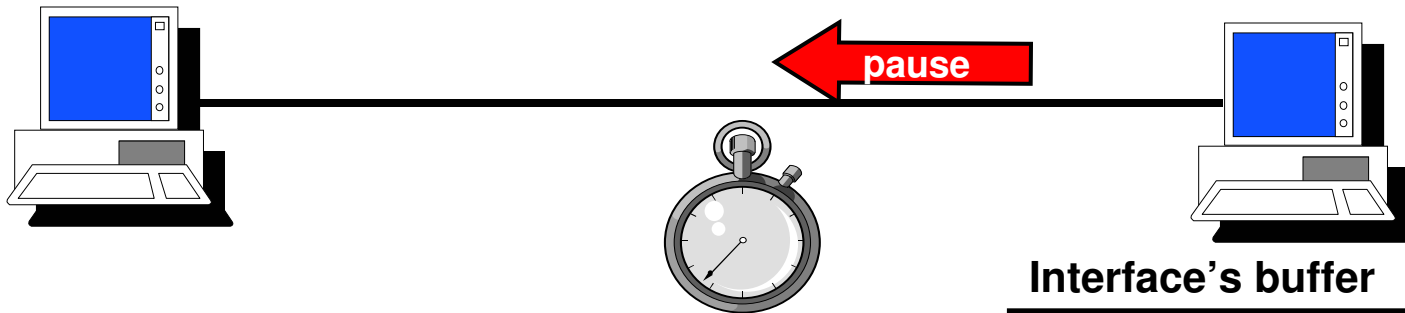
Switch's flow control: output buffer

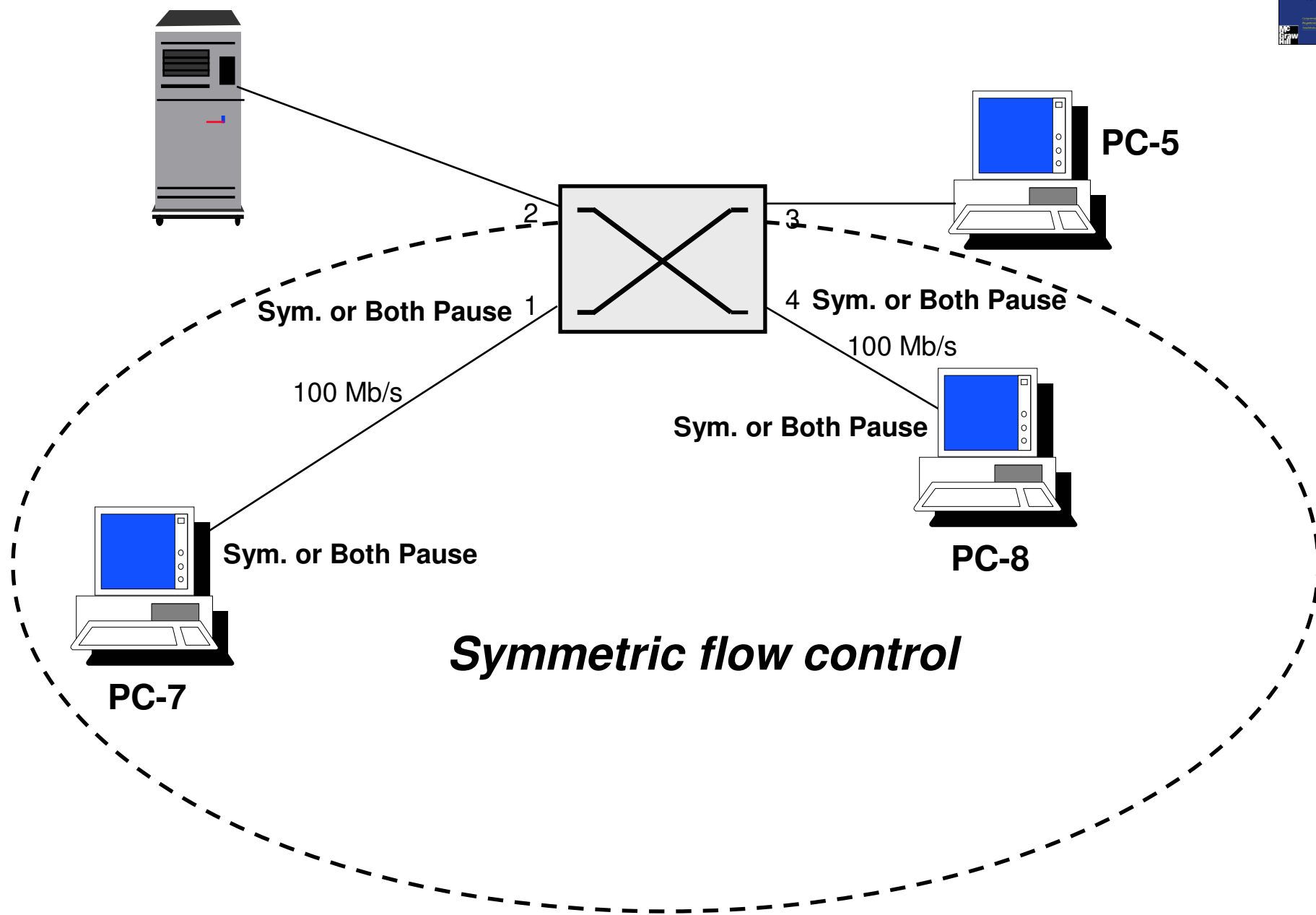
- Switch with output buffer and input port connection
 - output buffer saturation cause the sending of Pause on input port linked with traffic flow
 - solution not implemented by vendors because has more drawback than benefit
 - Pause packet on link between two switches penalize also the traffic flows which don't congest output buffers
- Vendors' solution:
 - switch blocks transmission on port if it receive the Pause packet, but it can't send the pause packet



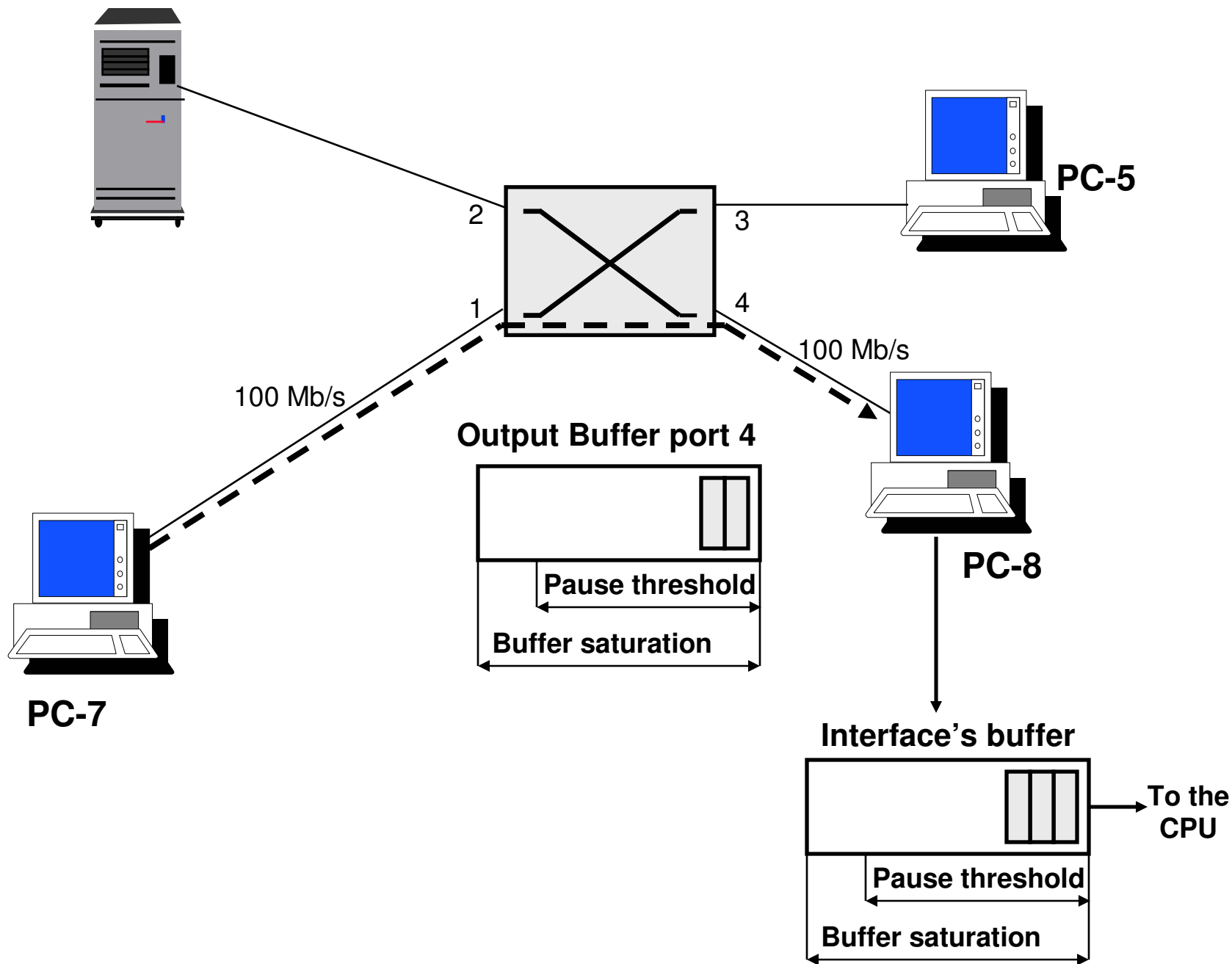


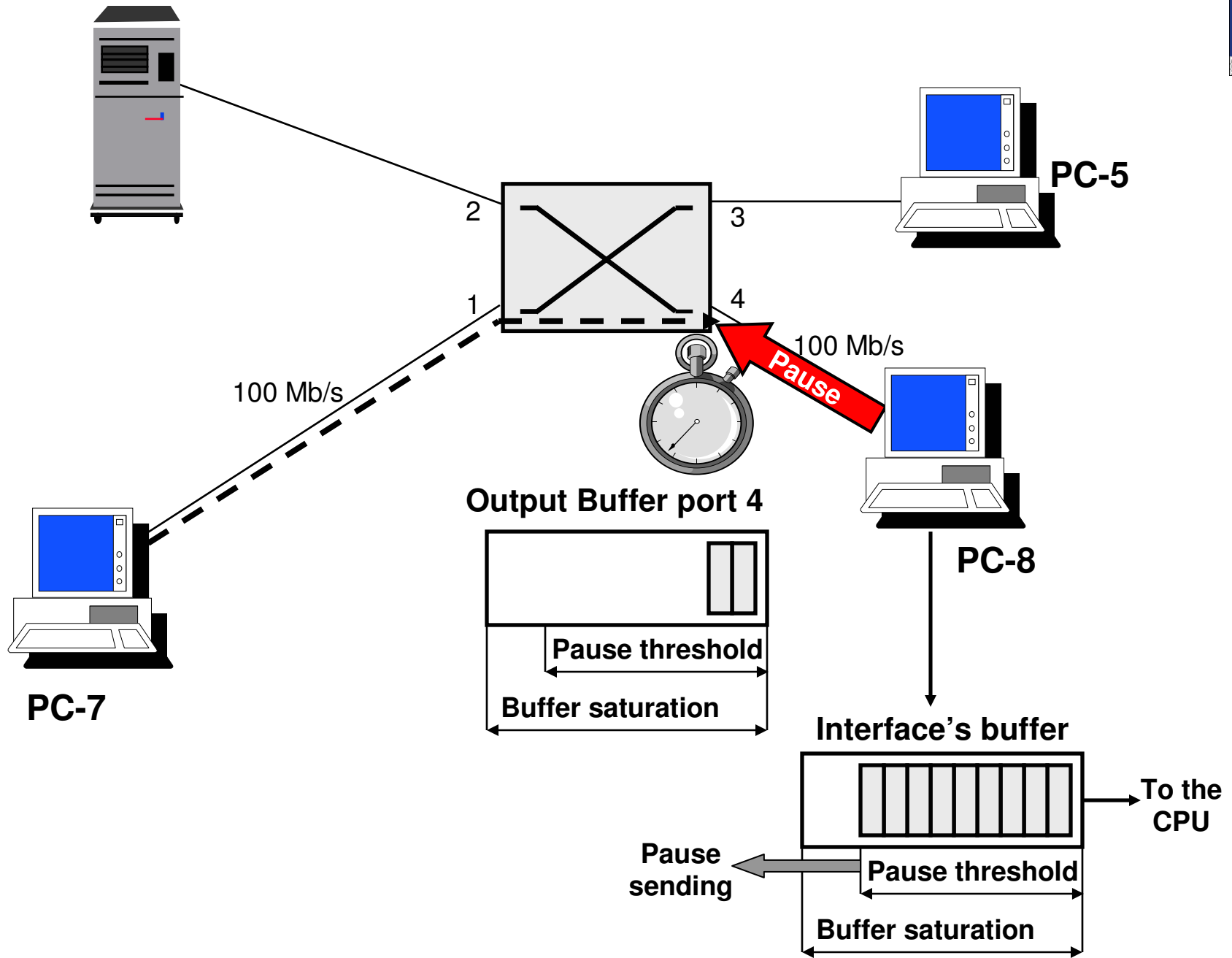
Symmetric flow control

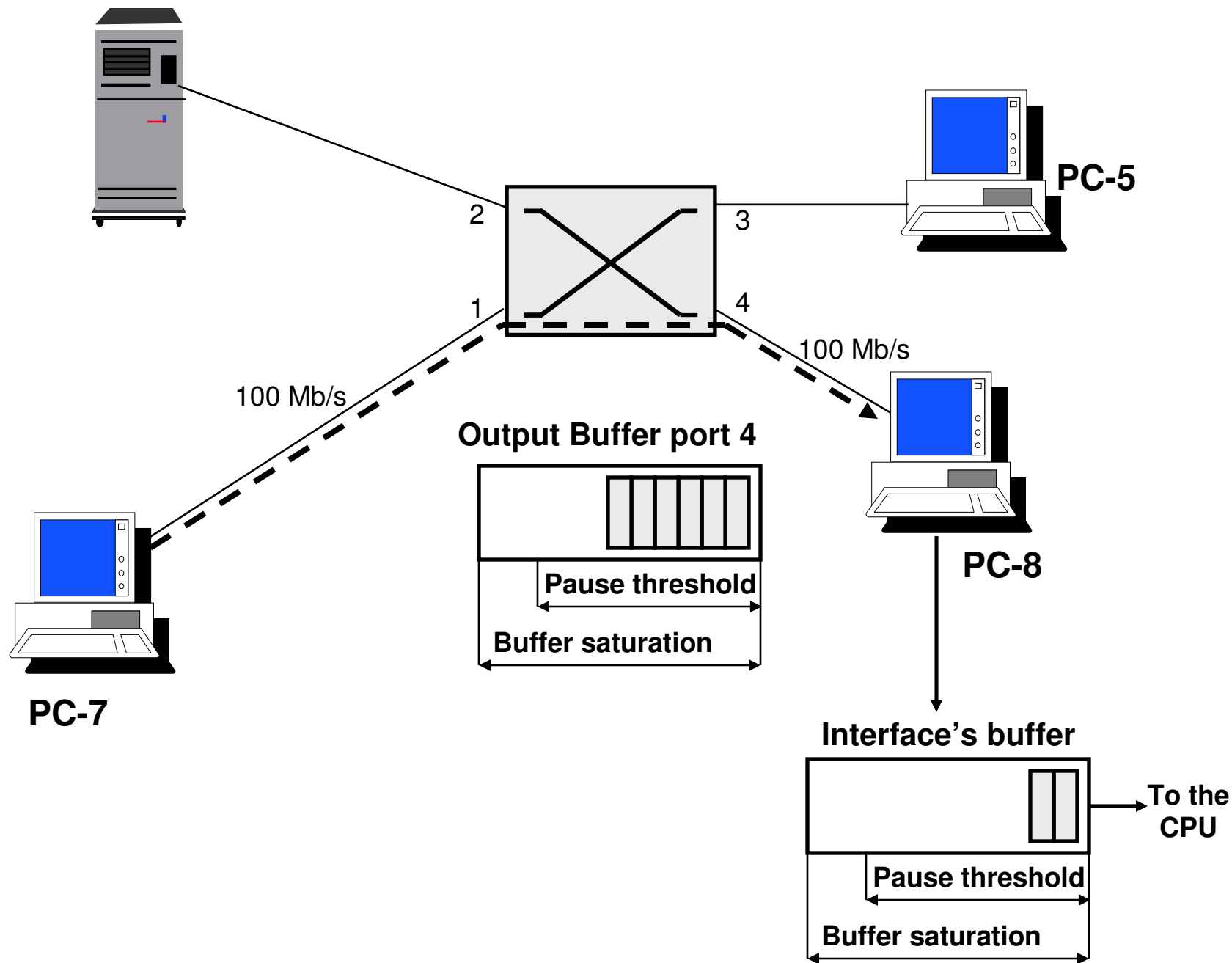


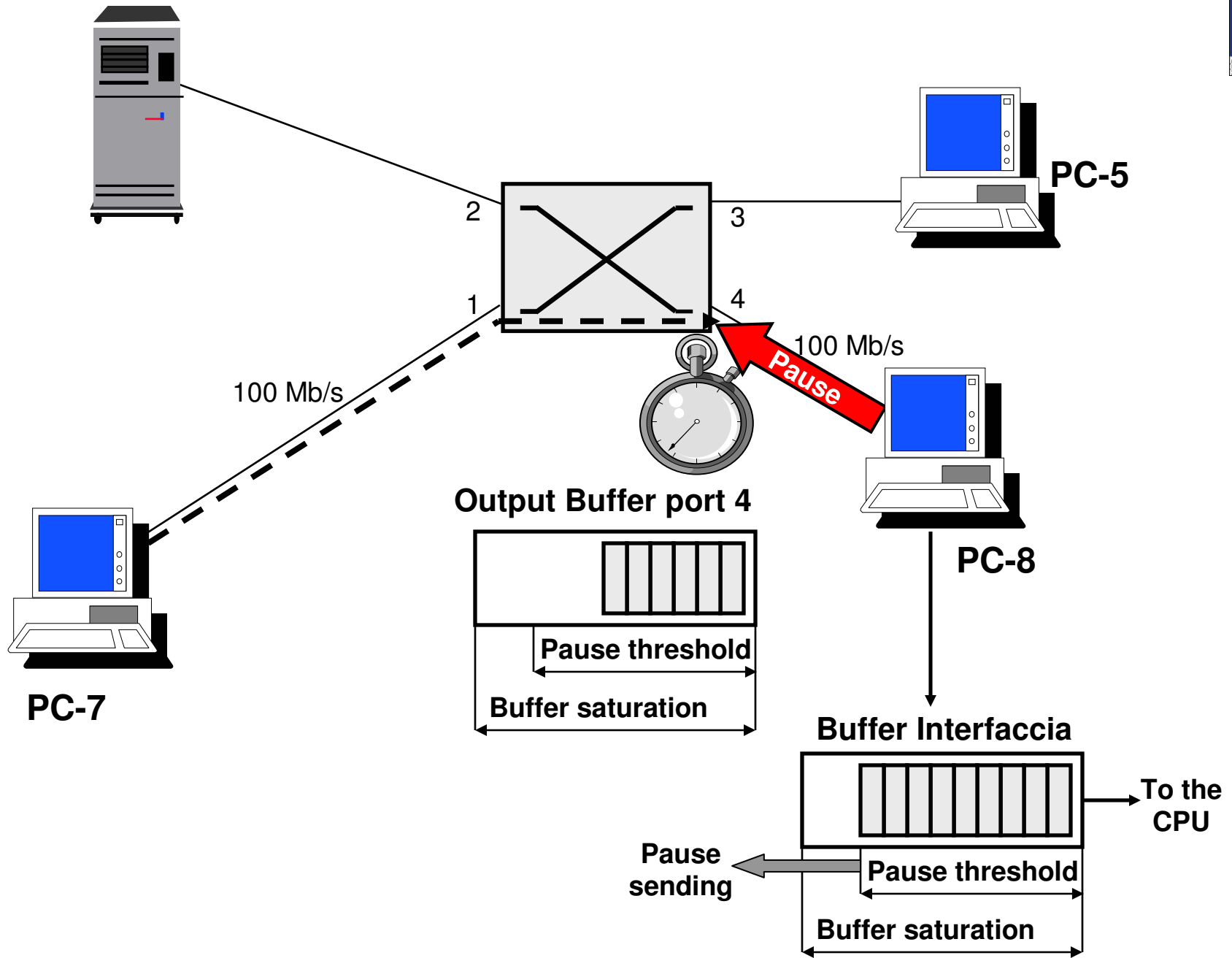


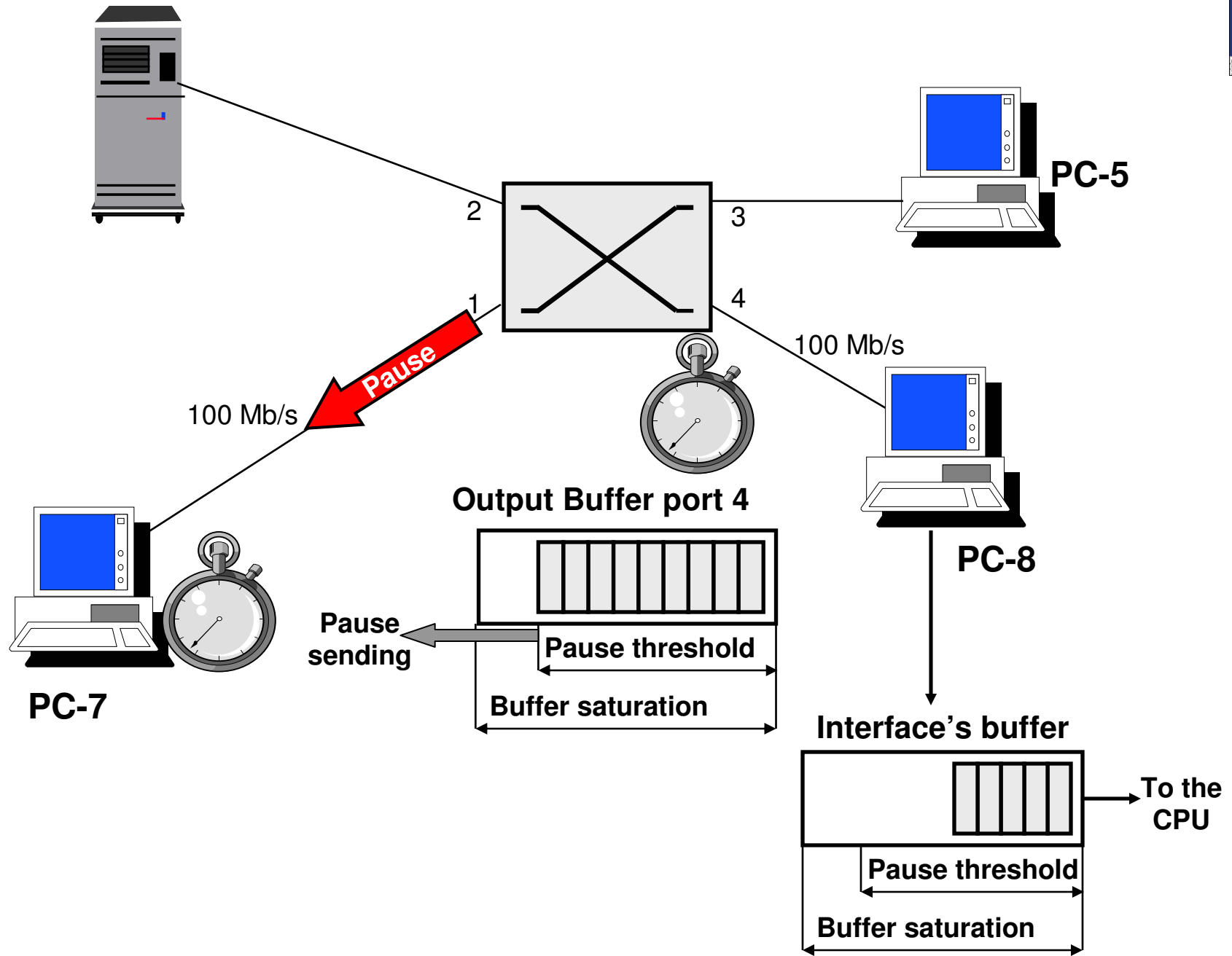
Symmetric flow control

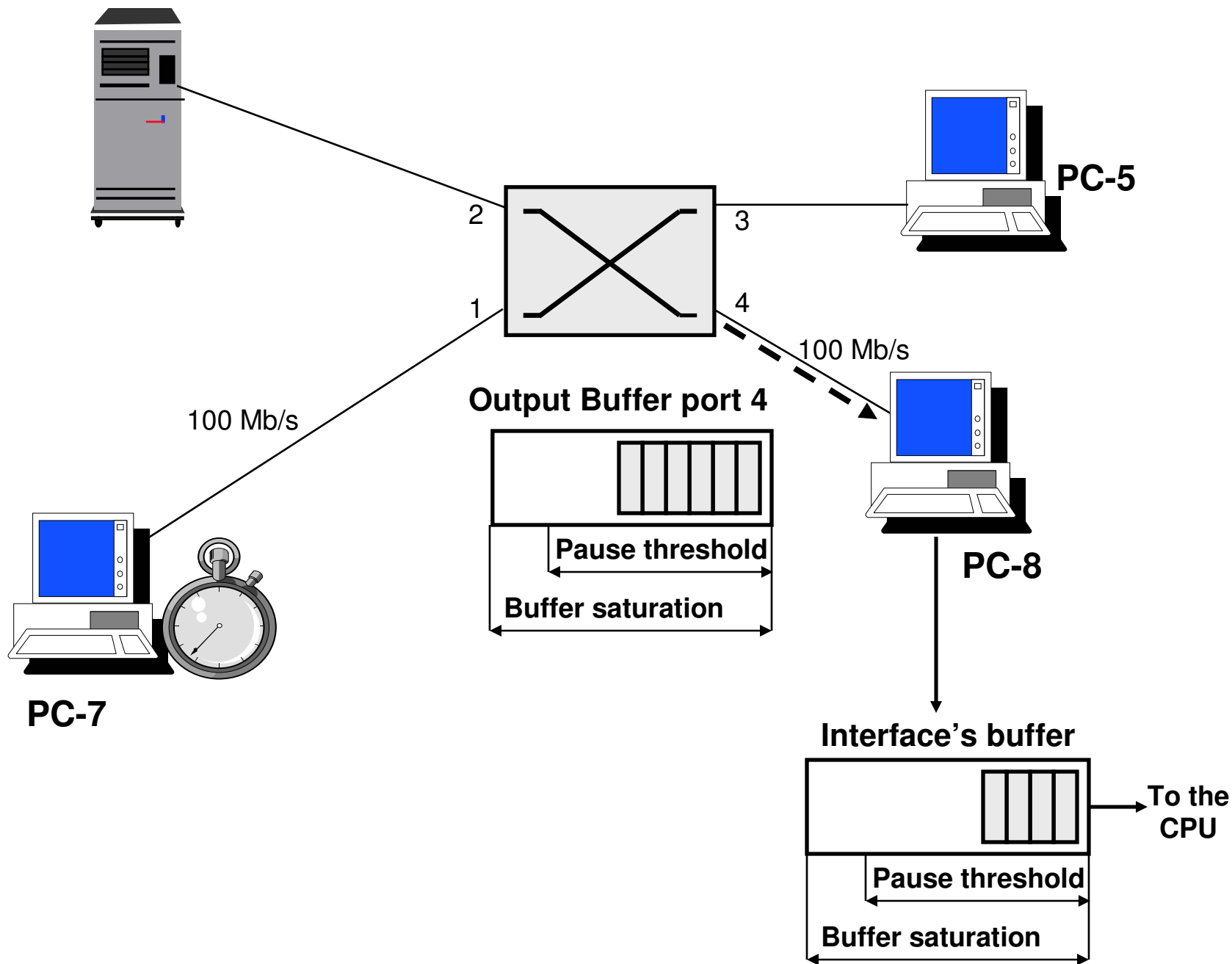


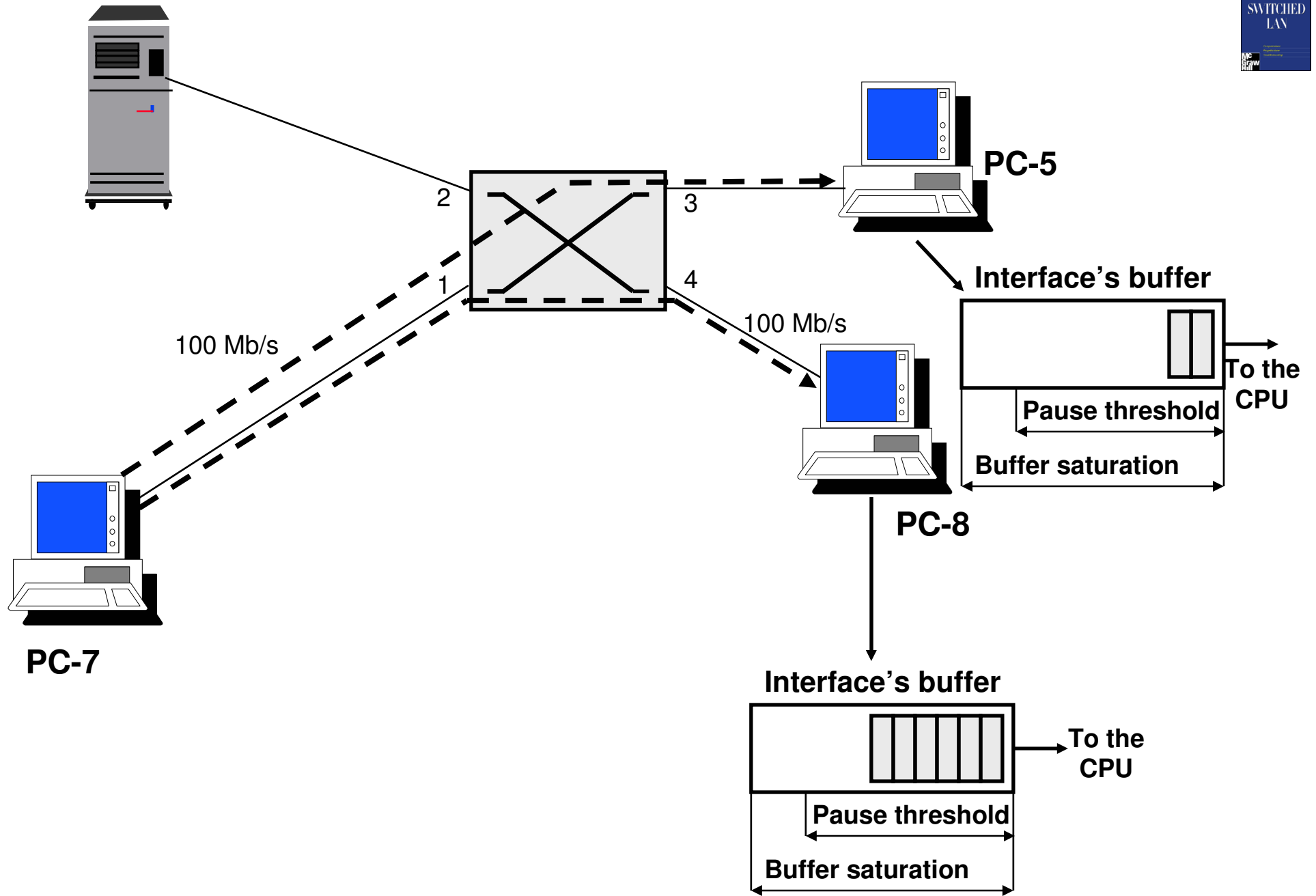


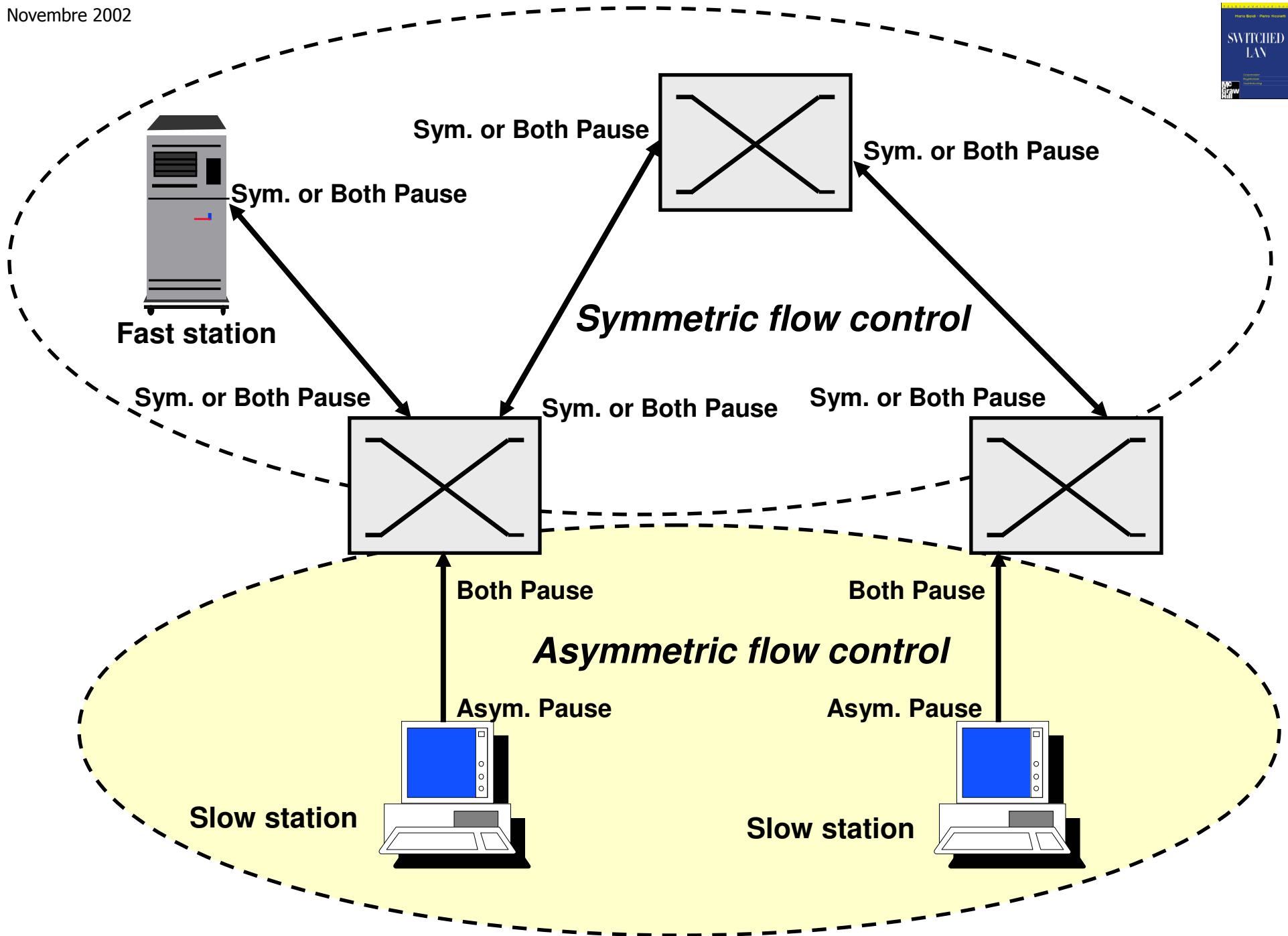


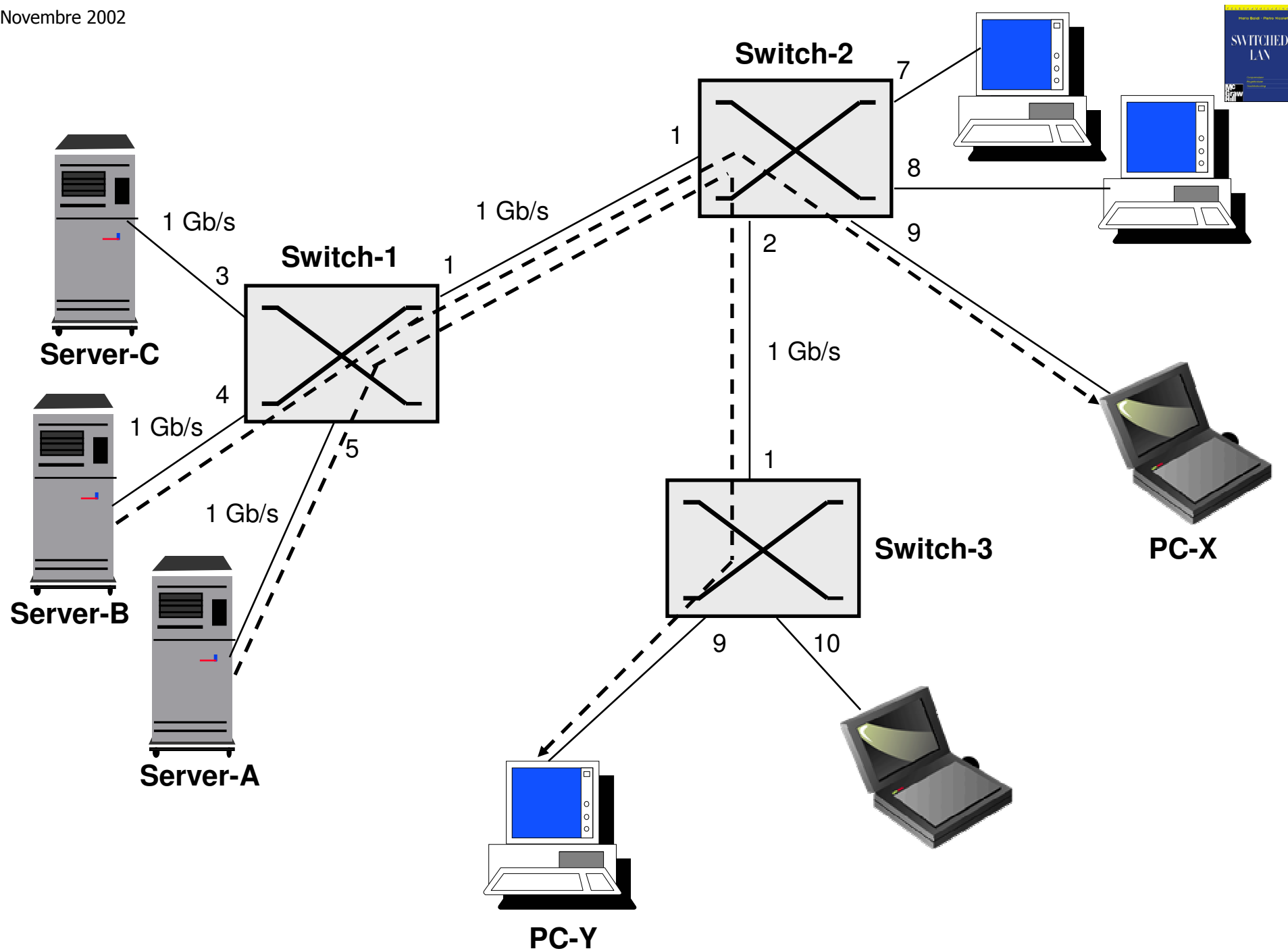


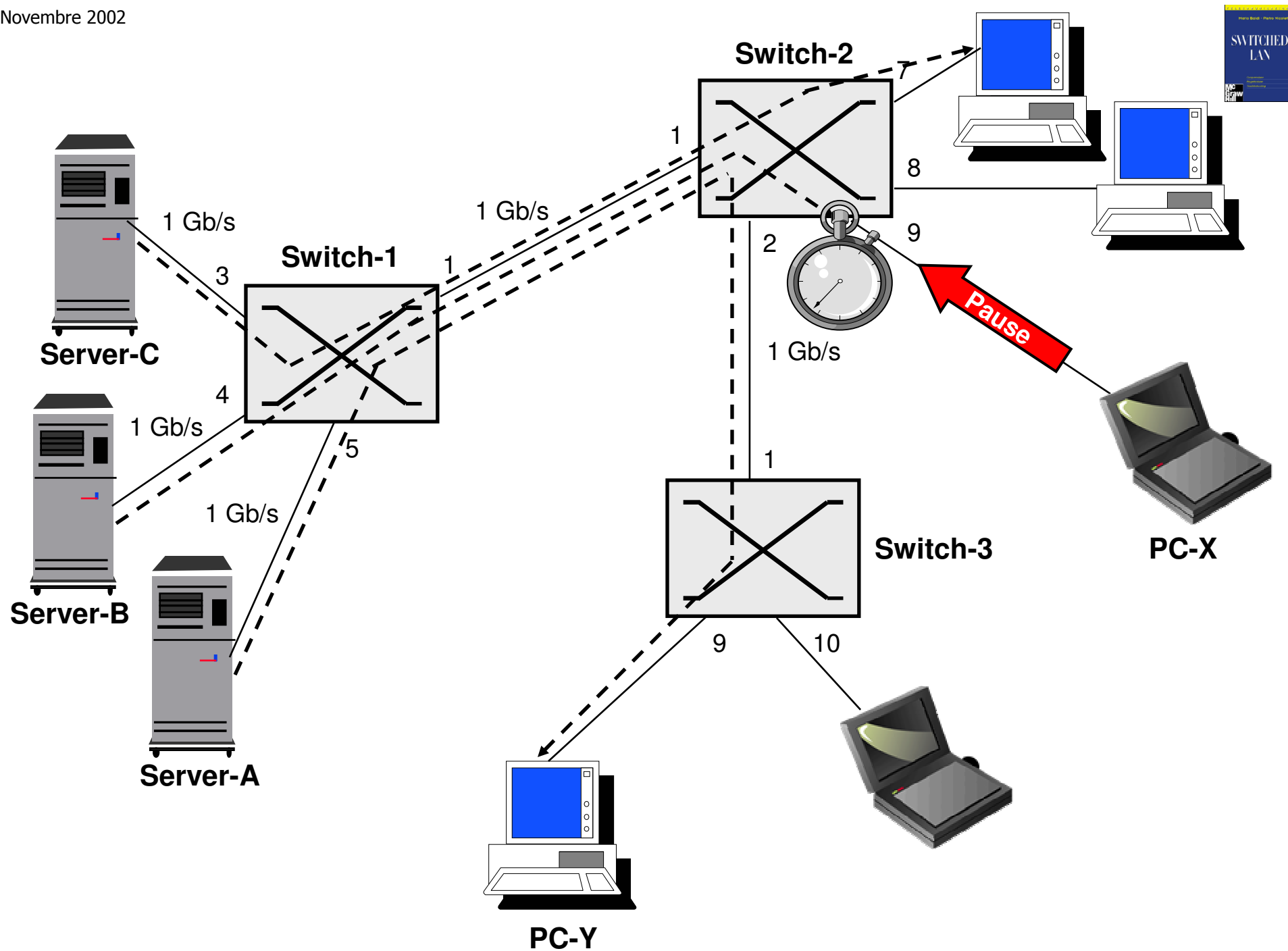


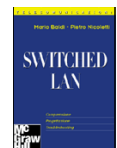
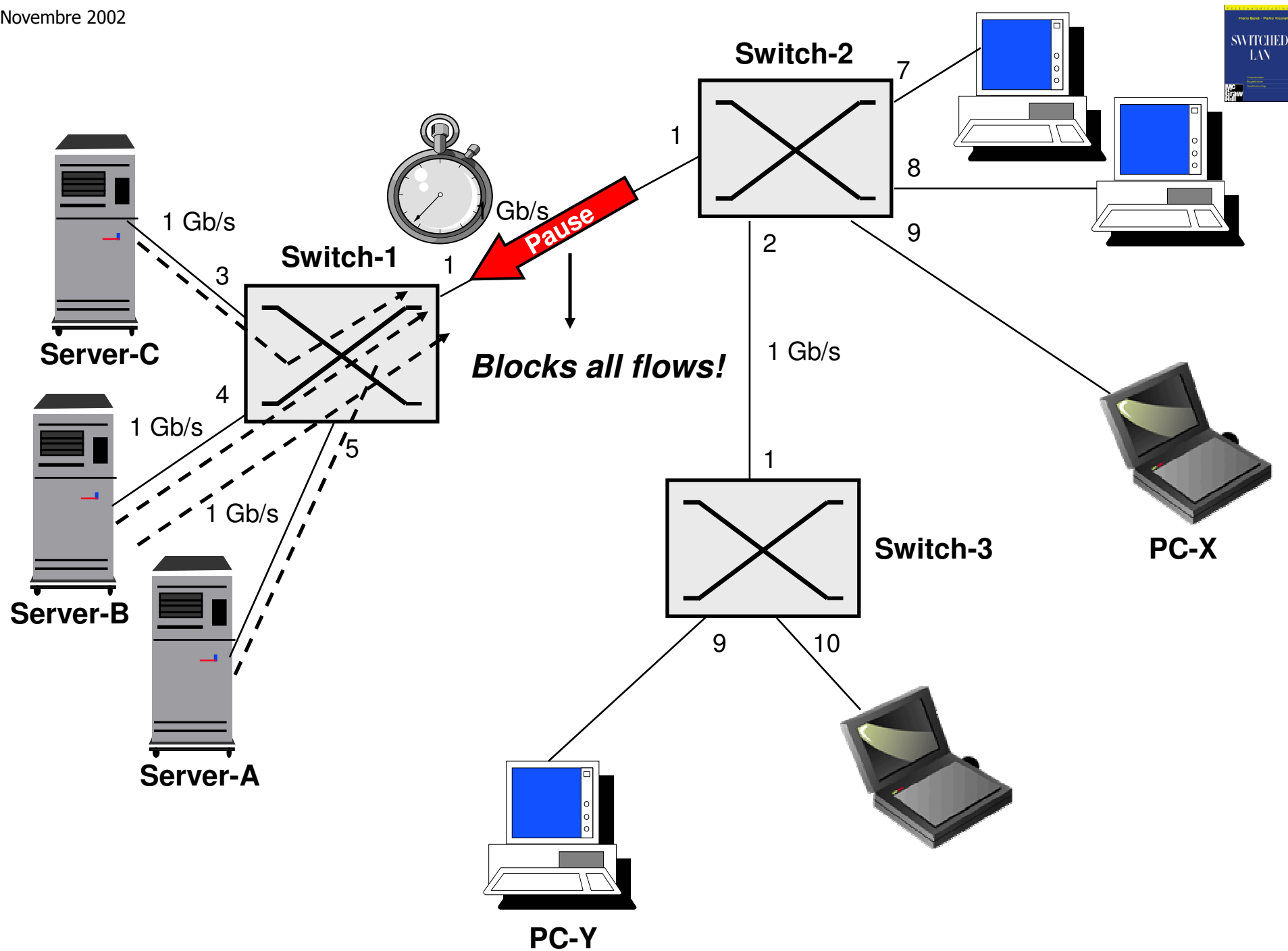








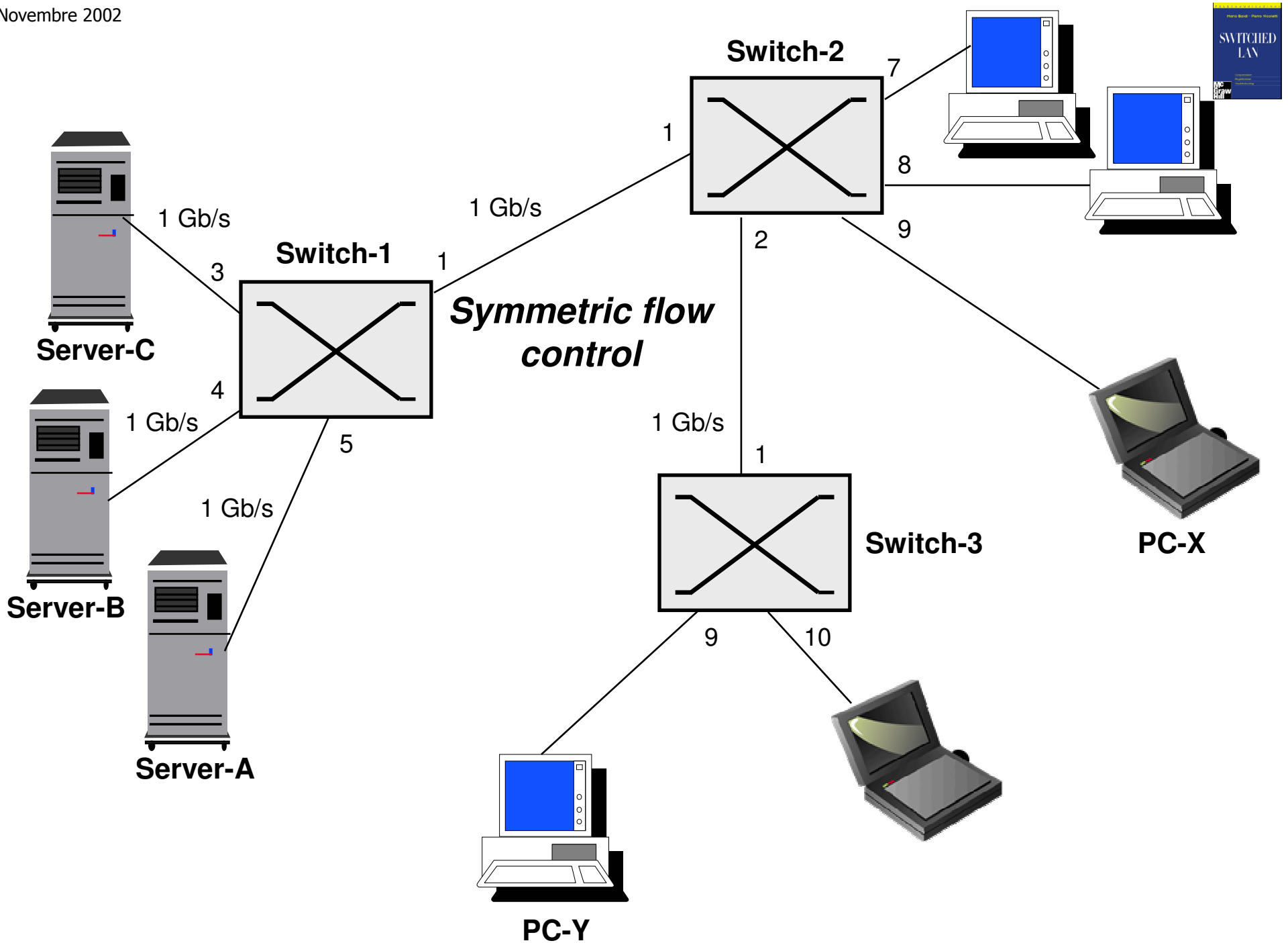


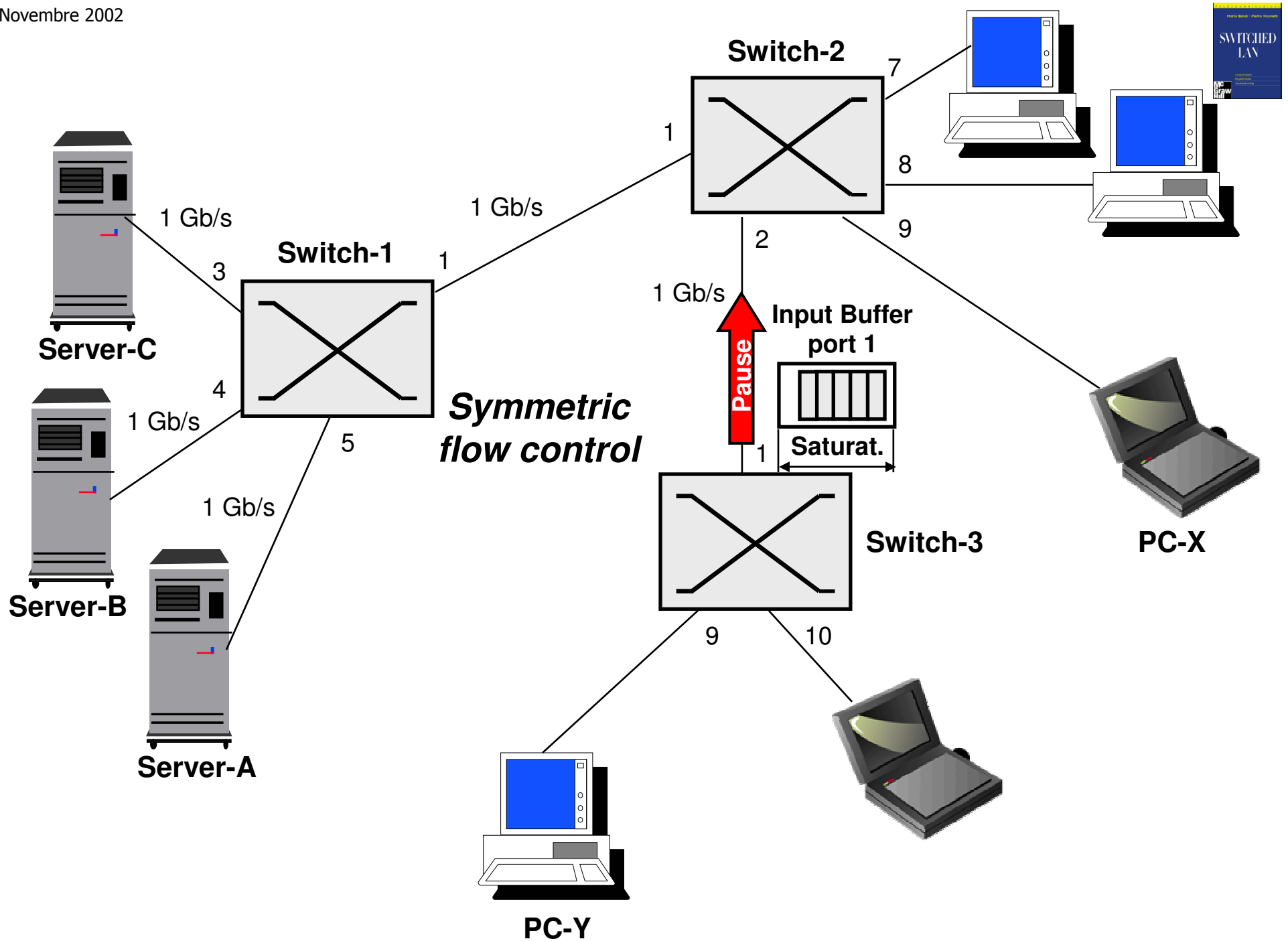


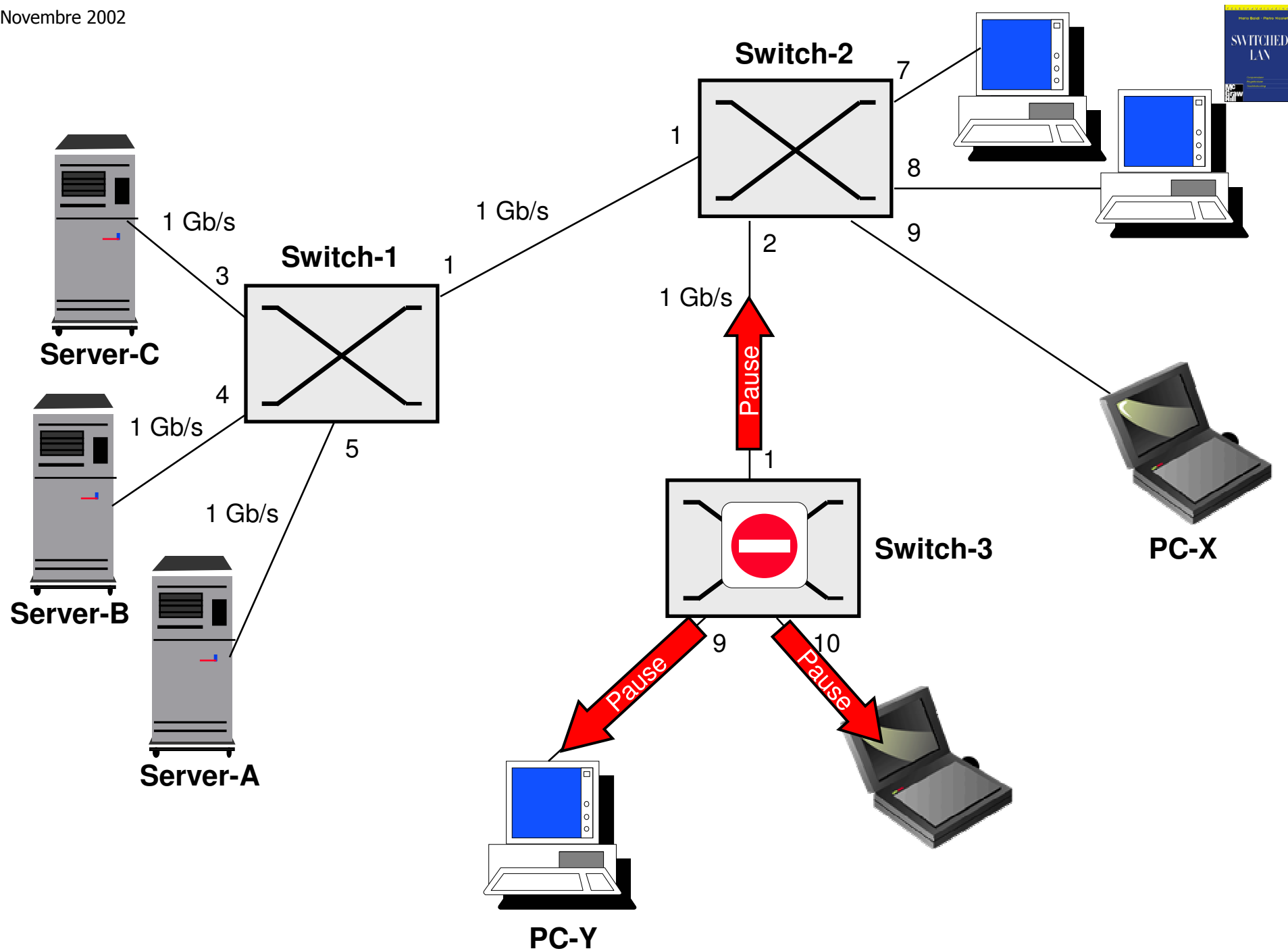


Flow control on switch with input buffer

- If an input buffer met is saturation because of input port congestion, it send a Pause packet in the concerned port
 - If switching matrix is blocking
 - If exist contention on output ports
 - No output buffer
 - Output buffer are full
- If switching matrix is congested, the event unleash the sending of Pause packet on all ports
 - Pause packet sending occurs only on switch with blocking matrix









Realistical approach on switch with output buffer

- Asymmetric flow control is enabled only on switch-station connections
 - make up temporary congestions on input buffers of station's network interfaces
 - not an ideal solution, but can be a good deal in certain traffic condition

