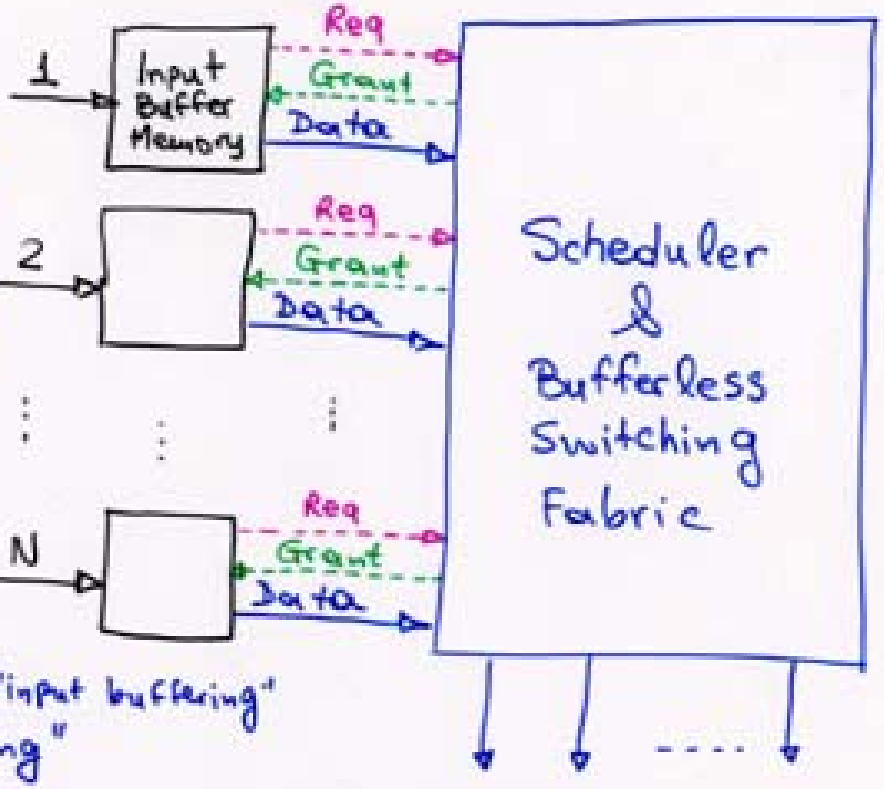
4.2 Input Queueing
(a) the Single-Queue per Input Case

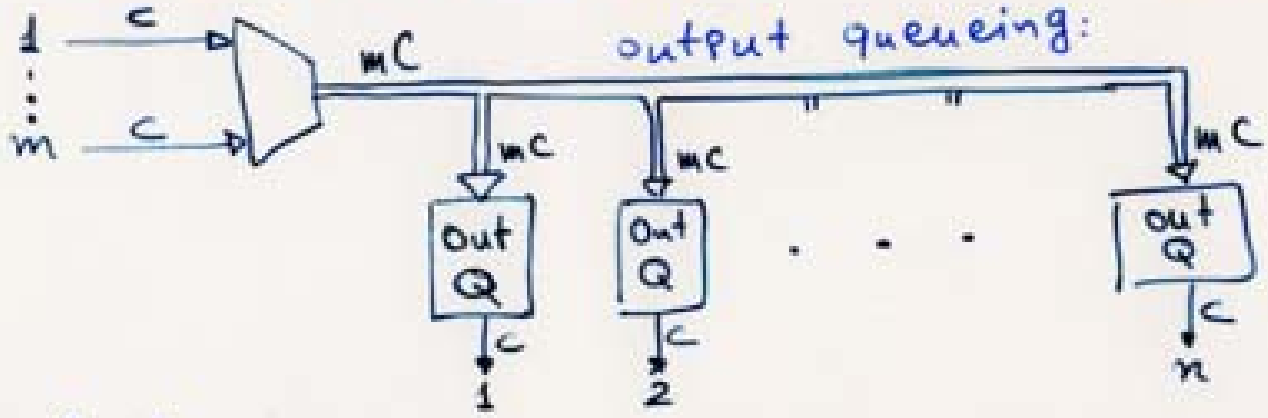
- Crossbar Switch with one Buffer Memory per Input Line
- Throughput per Buffer Memory = 1 (incoming) + 1 (outgoing)
- *Not* the Dual of Output Queueing
  - Output-queueing throughput per buffer memory = *N+1*
  - Output queueing does *not* contain a crossbar:
    a crossbar allows each output to be fed by at most *one* input;
    output queueing allows each output to be fed by *any* (even all) inputs
- Crossbar Scheduling is an issue
  - Single queue per input $\Rightarrow$ independent schedulers per-output (easy)
  - Multiple queues per input (next section) $\Rightarrow$ difficult…

Input Queueing Family



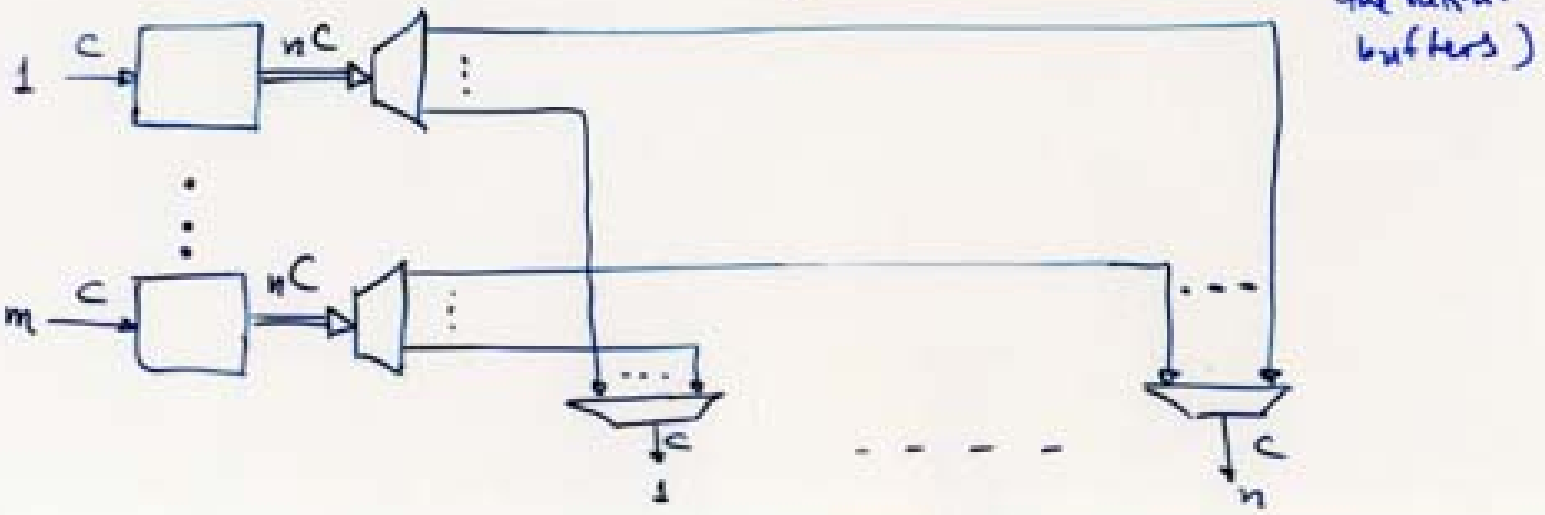
- But. Mem. throughput = const = 2 (each)
  total for all = 2 N = min. possible
- How many Queues per Inp. But. Mem.?
  - single queue: "input queueing"
    ... Head-of-Line Blocking
    ⇒ bad performance (see Ch. 1)
  - many queues (per output or more):
    "advanced input queueing" or "input buffering"
    or "virtual output queueing"
- What Switching Fabric?
  - crossbar .... up to N ~ hundred(s)
  - multistage ... see chapter 5
- What Scheduler? ... What Performance?
  - difficult
  - non scalable
  - big research topic for last >10 years and still... See reading list papers

Note: Input Queueing is NOT the dual of Output Queueing!

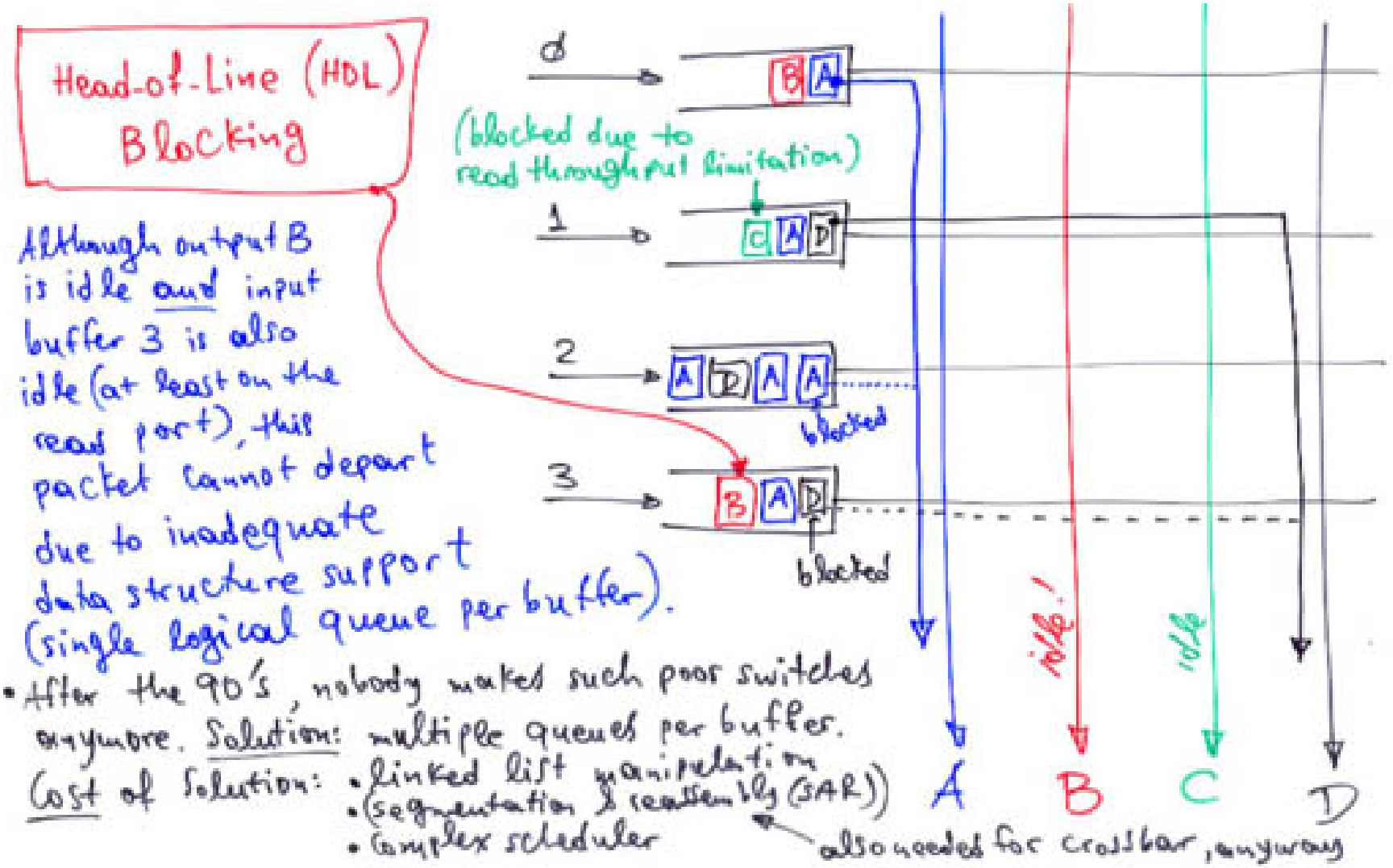


the dual would be:

(like "internal speed-up" of n, but without the output buffers)



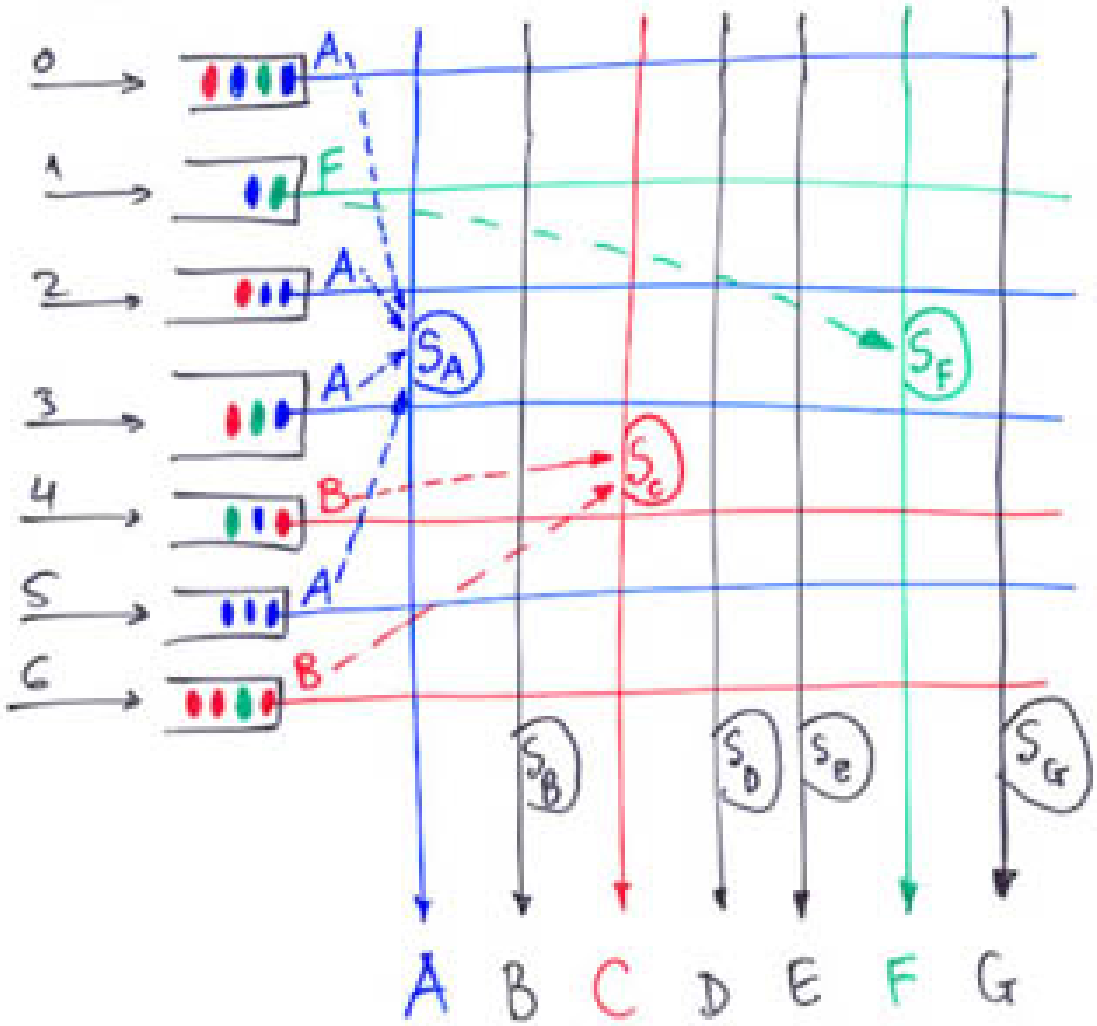
Old-Fashioned Input Queueing: Single Queue per Input



- After the 90's, nobody wanted such poor switches anymore. Solution: multiple queues per buffer.

Cost of Solution:
- linked list manipulation
- (segmentation & reassembly (SAR))
- complex scheduler

Independent Per-Output Schedulers



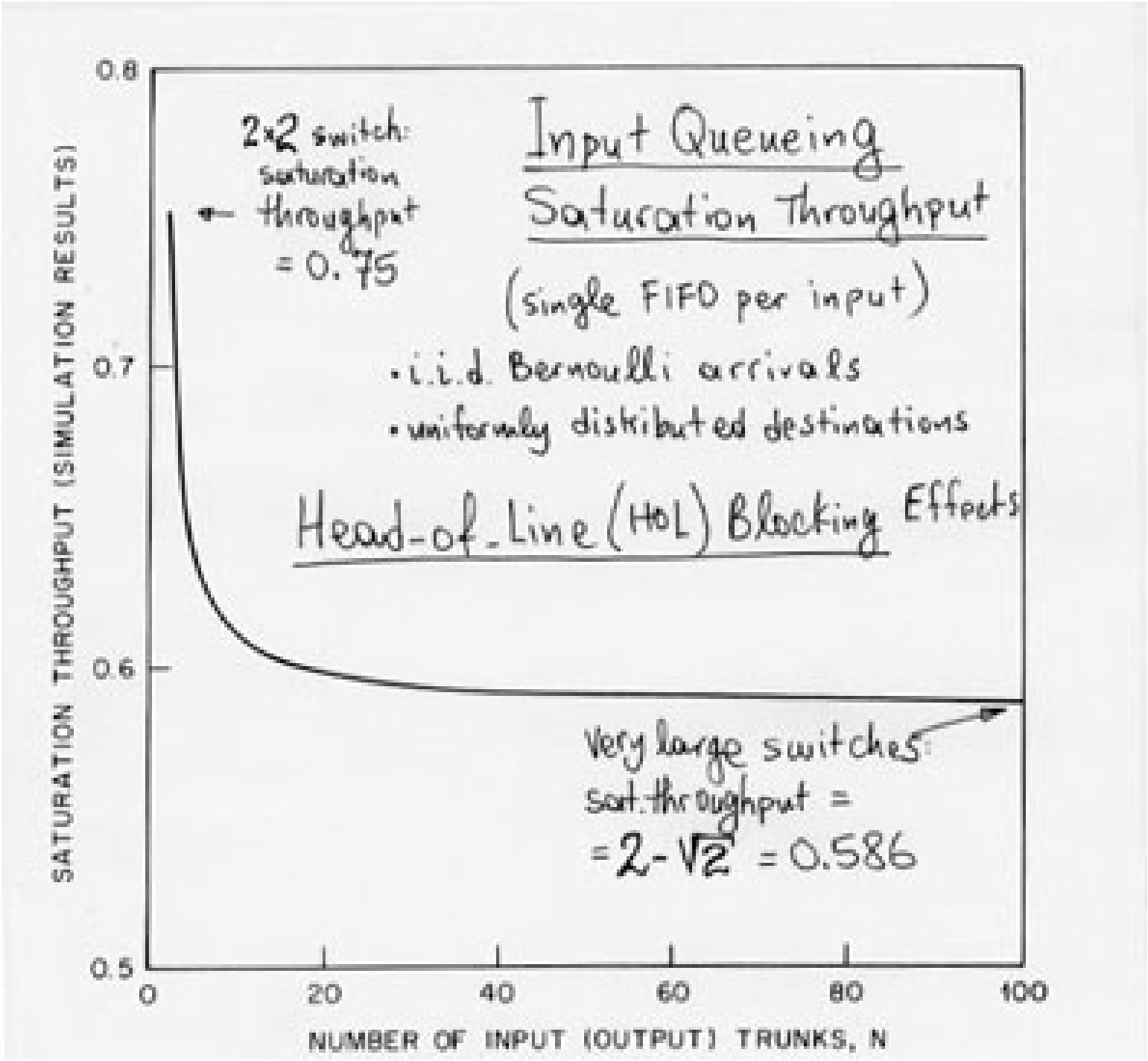
- with old-fashioned input Q.
- (as with crosspoint queueing (similar to output Q or shared buffer)).

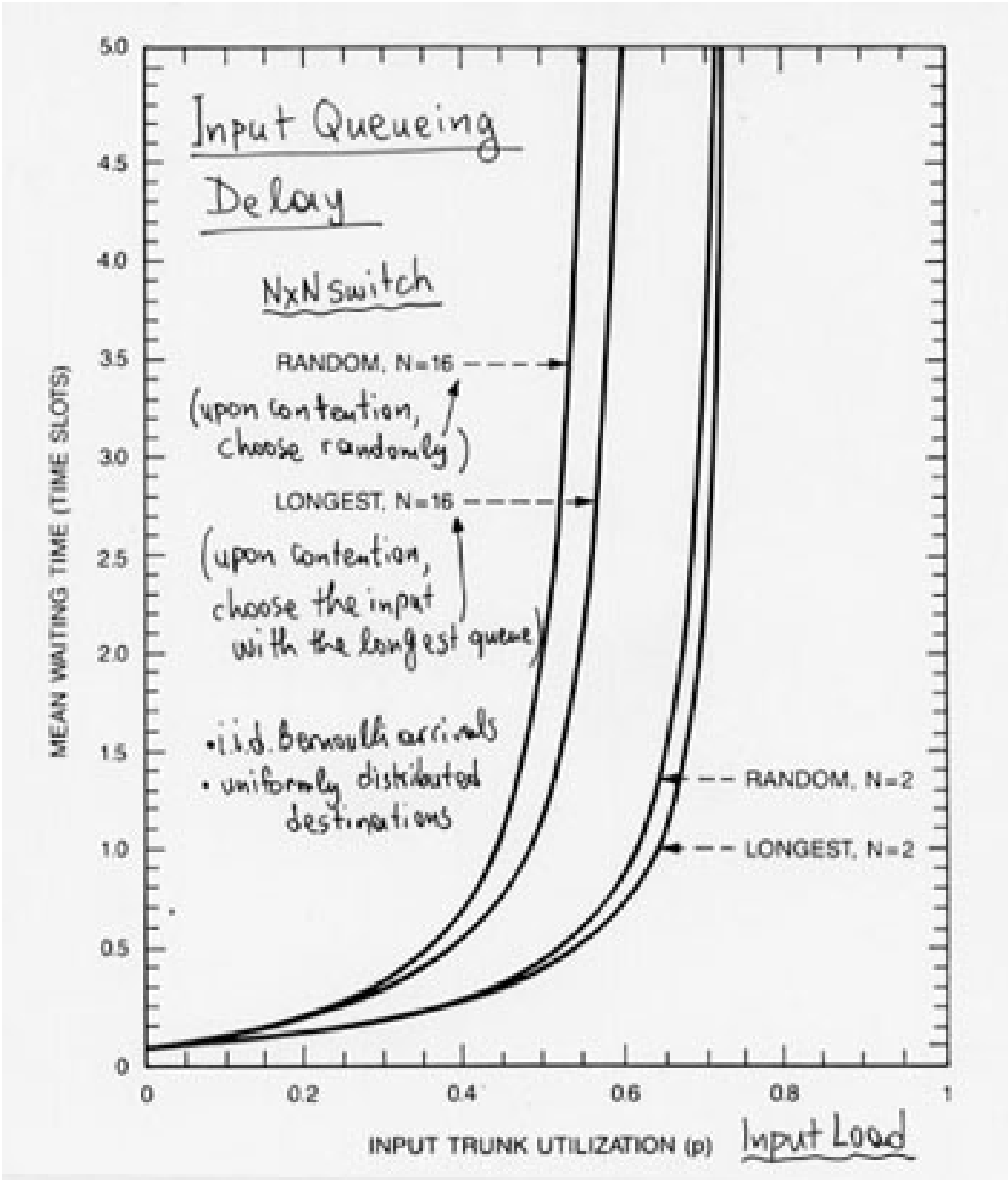
- each output has its own scheduler
- each scheduler sees a distinct set of requests
- each scheduler independently makes a decision
- no better decision is possible by considering the other schedulers' decisions.

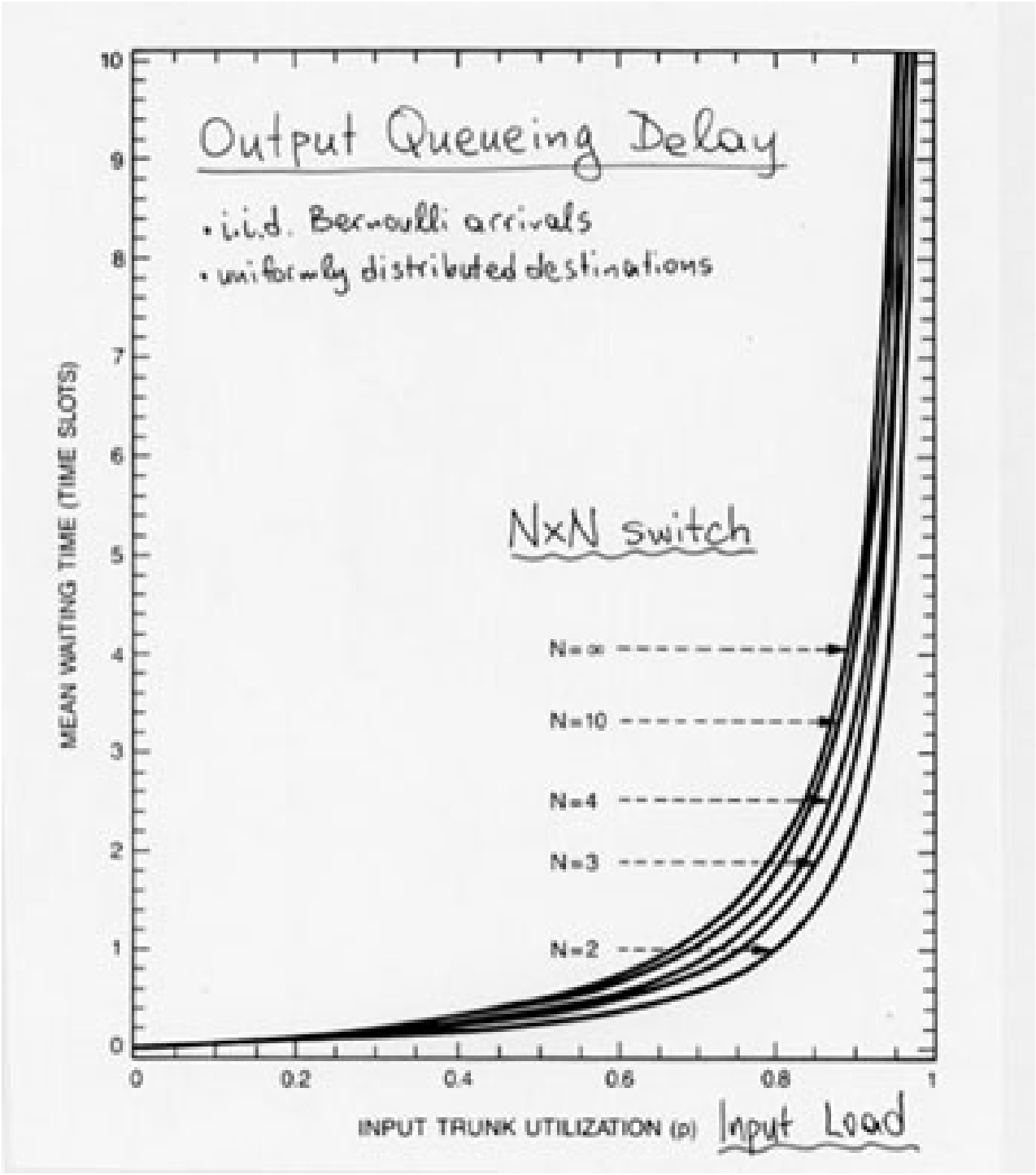
... unlike the next scheme (adv. inp. buffering - VOQ's) ....

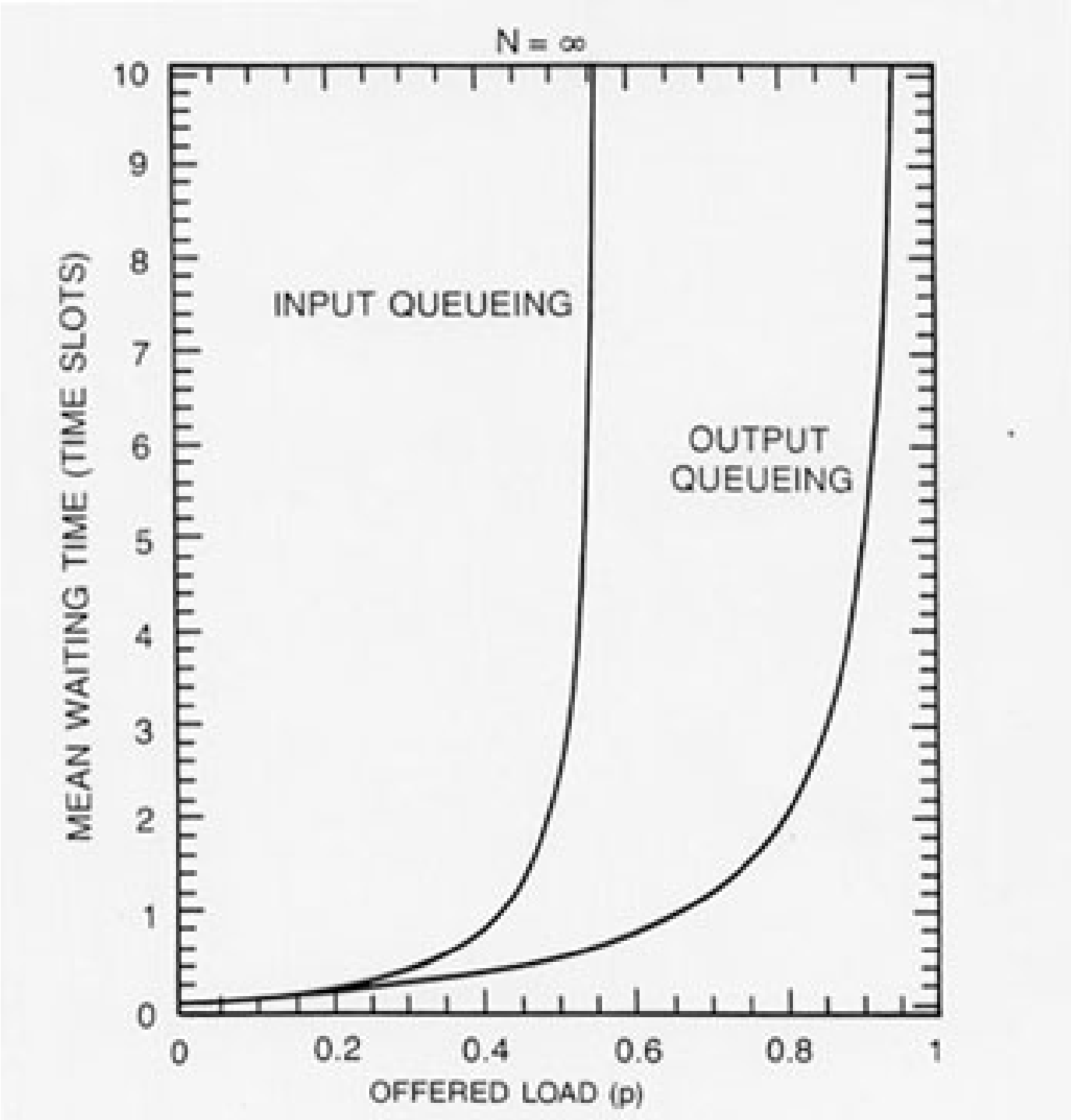
Throughput & Delay under Single-Queue Input Q'ing

- *References:*
  - M. Hluchyj, M. Karol: "Queueing in High-Performance Packet Switching", IEEE Journal on Sel. Areas in Commun. (JSAC), Dec. 1988, pp. 1587-1597.
  - M. Karol, M. Hluchyj, S. Morgan: "Input versus Output Queueing on a Space-Division Packet Switch", IEEE Tr. on Communic., Dec. 1987, pp. 1347-1356.
  - J. Hui, E. Arthurs: "A Broadband Packet Switch for Integrated Transport", IEEE Journal on Sel. Areas in Commun. (JSAC), Oct. 1987, pp. 1264-1273.
- *Attention:* results for i.i.d. Bernoulli (non-bursty) arrivals, with uniformly-distributed destinations (no overloaded hot-spots), are only useful for gaining a rough, first insight into the behavior of systems, but are often not representative of the real behavior of systems under real traffic!...


0.8
0.7
0.6
0.5
SATURATION THROUGHPUT (SIMULATION RESULTS)
2x2 switch: saturation throughput = 0.75
Input Queueing Saturation Throughput
(single FIFO per input)
• i.i.d. Bernoulli arrivals
• uniformly distributed destinations
Head-of-Line (HoL) Blocking Effects
Very large switches: Sat. throughput = $2-\sqrt{2}$ = 0.586
0
20
40
60
80
100
NUMBER OF INPUT (OUTPUT) TRUNKS, N


Input Queueing
Delay
NxN switch
RANDOM, N=16
(upon contention, choose randomly)
LONGEST, N=16
(upon contention, choose the input with the longest queue)
• i.i.d. Bernoulli arrivals
• uniformly distributed destinations
RANDOM, N=2
LONGEST, N=2
MEAN WAITING TIME (TIME SLOTS)
INPUT TRUNK UTILIZATION (p)
Input Load
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0.2
0.4
0.6
0.8
1


Output Queueing Delay
• i.i.d. Bernoulli arrivals
• uniformly distributed destinations
NxN switch
N=∞
N=10
N=4
N=3
N=2
MEAN WAITING TIME (TIME SLOTS)
INPUT TRUNK UTILIZATION (p)
Input Load
0
1
2
3
4
5
6
7
8
9
10
0.2
0.4
0.6
0.8
1


N = ∞
MEAN WAITING TIME (TIME SLOTS)
10
9
8
7
6
5
4
3
2
1
0
INPUT QUEUEING
OUTPUT
QUEUEING
0
0.2
0.4
0.6
0.8
1
OFFERED LOAD (p)