Event builder design

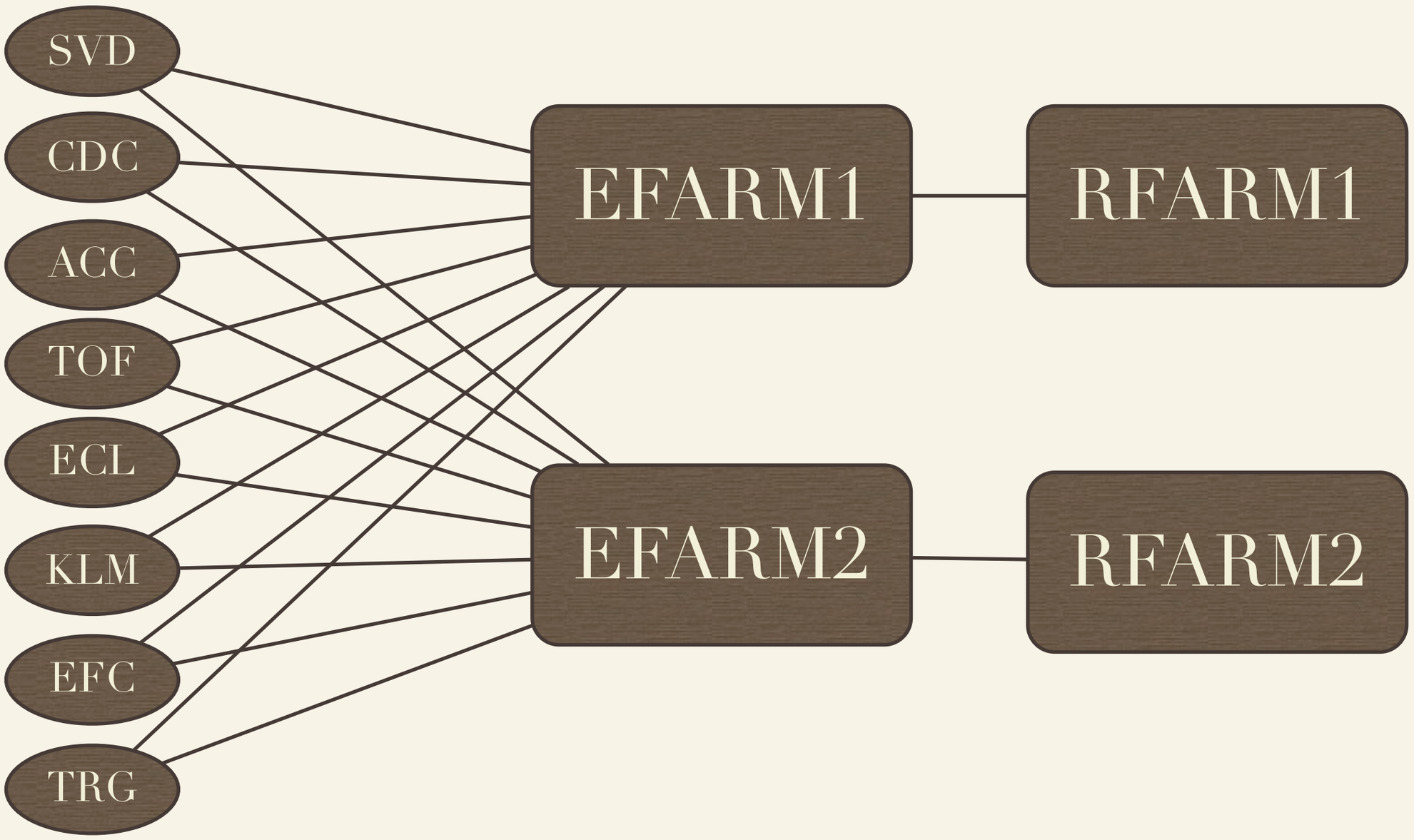
Yamagata
T.Higuchi

Contents

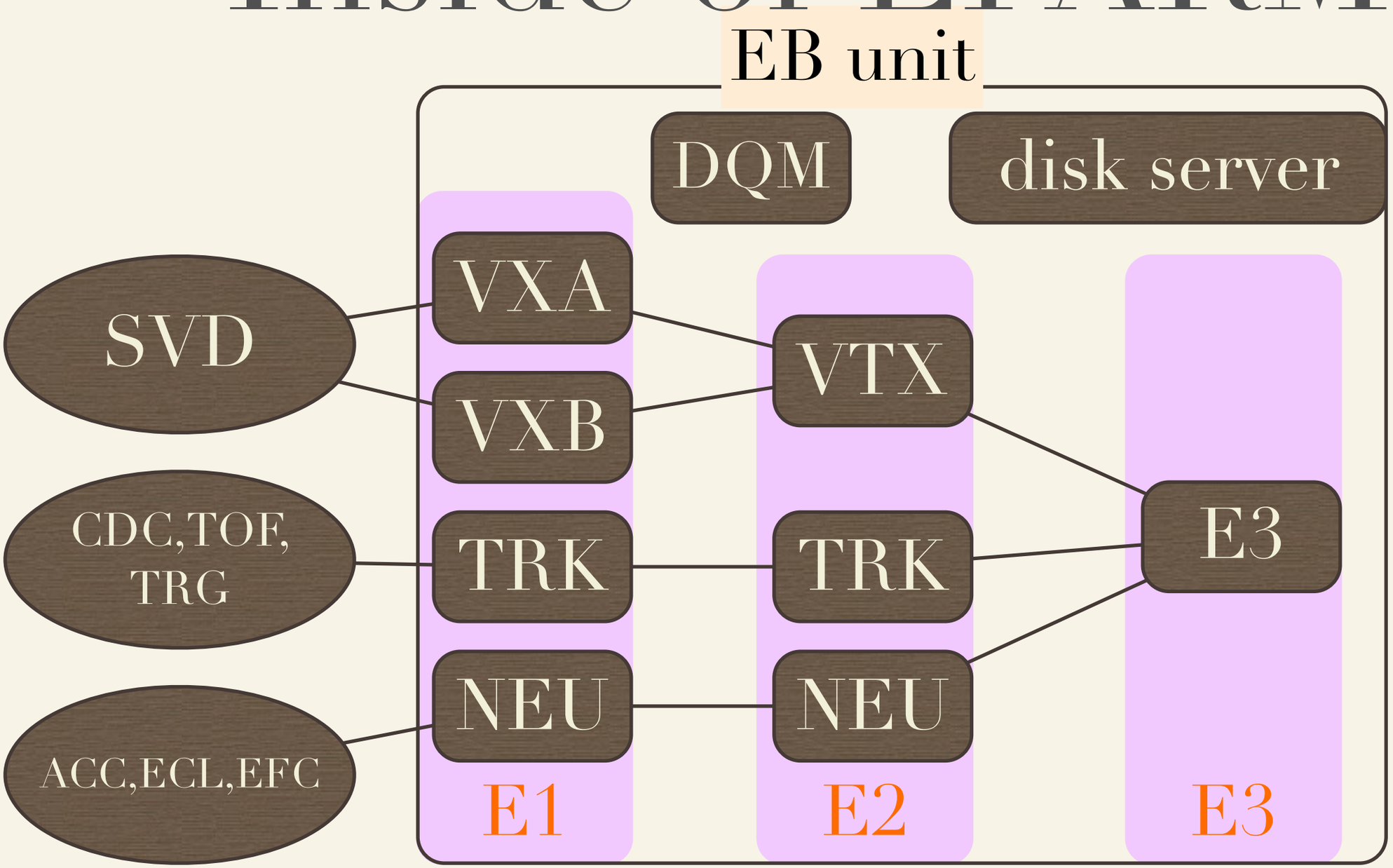
- Current scheme "EB unit"
- Belle II requirement
- Barrel shifter method
- Usage of network switch

Current Scheme

Now



Inside of EFARM



10 PCs per EFARM

Throughput

- Current
  - 30kB x 500Hz〜1kHz => 15MB/s〜30MB/s
  - 1 or 2 units of EFARM operation

Belle II requirement

Throughput

- Current
  - 30kB x 500Hz〜1kHz => 15MB/s〜30MB/s
  - 1 or 2 units of EFARM
- Belle II
  - 100kB x 30kHz => 3000MB/s
  - 50 units of EFARM

of PC

- # of PC / EFARM unit is 50
- To catch up Belle II data rate,
  - 50 unit of EFARM
  - 500 PCs
  - really?

reduce # of PCs

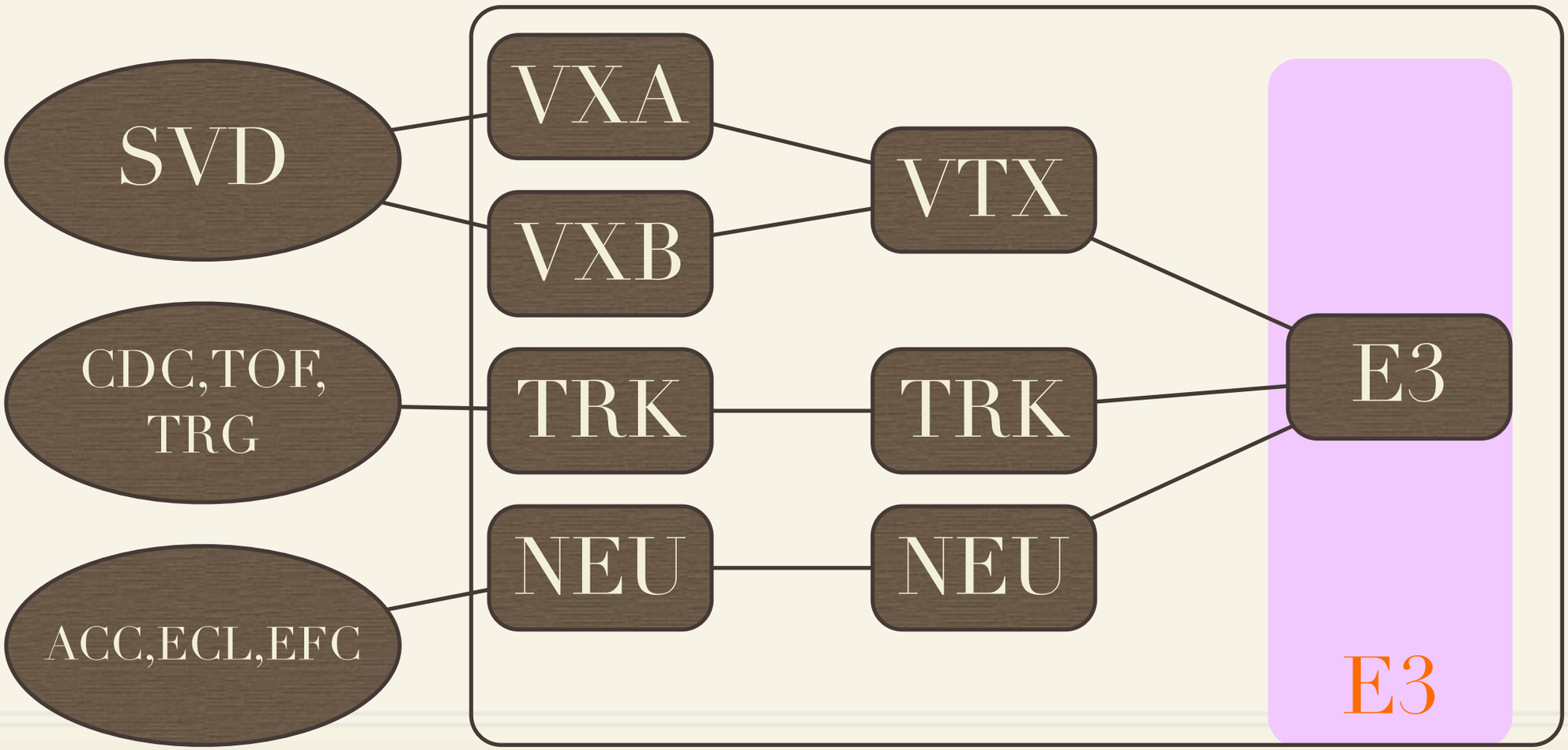
- Now CPU power limits EFARM throughput
- Even if I/O speed is not advanced, processing power will be faster in future.
- We can assume 1Gbps (100MB/s) per EFARM unit.

of EFARM

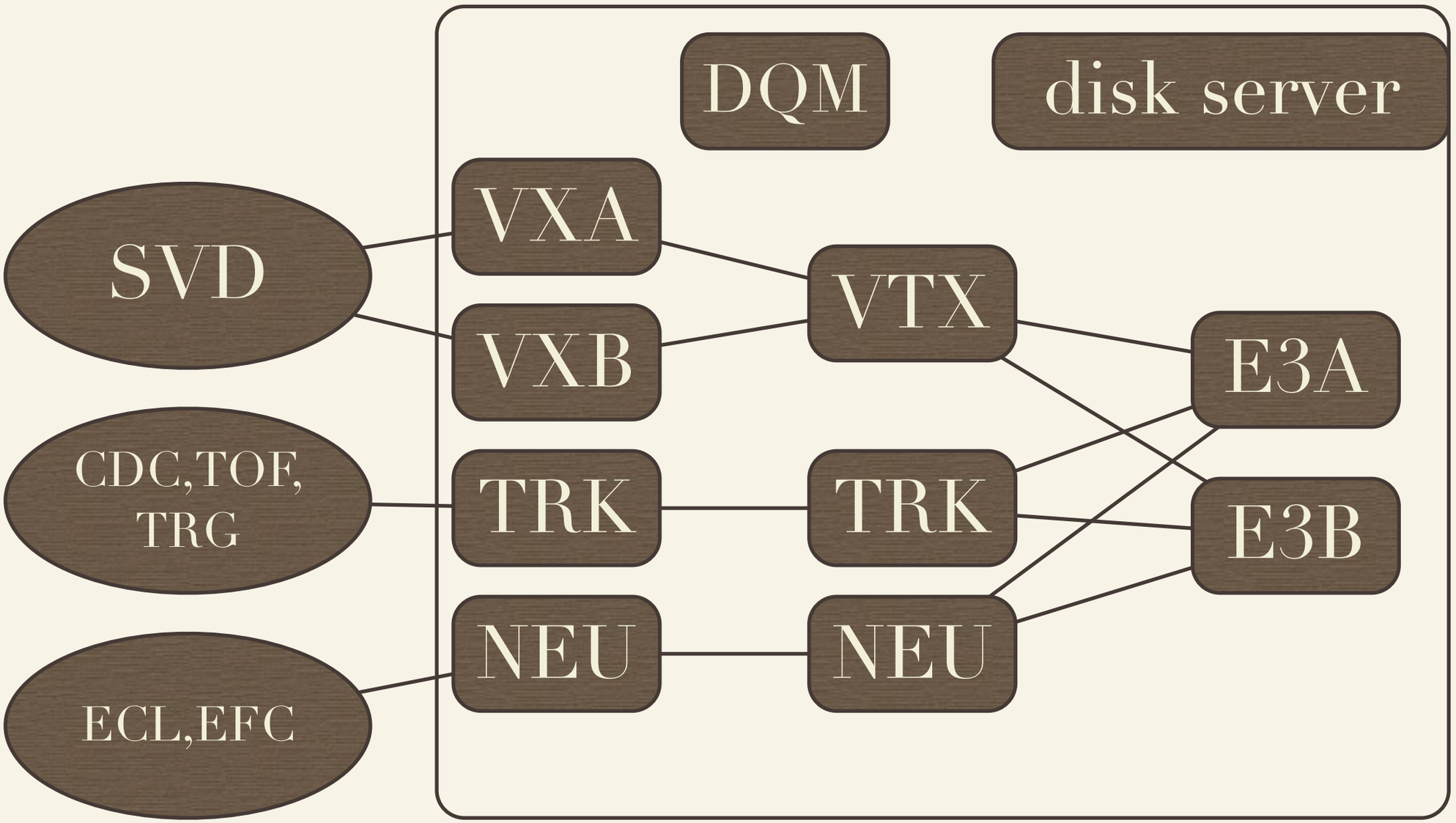
- Total output is 3000MB/s, so 30 units of EFARM will be sufficient
- # of PCs is still 300.
- really?

Bottoleneck

- Current bottleneck is E3
  - E1 and E2 are not saturated.



How about this?

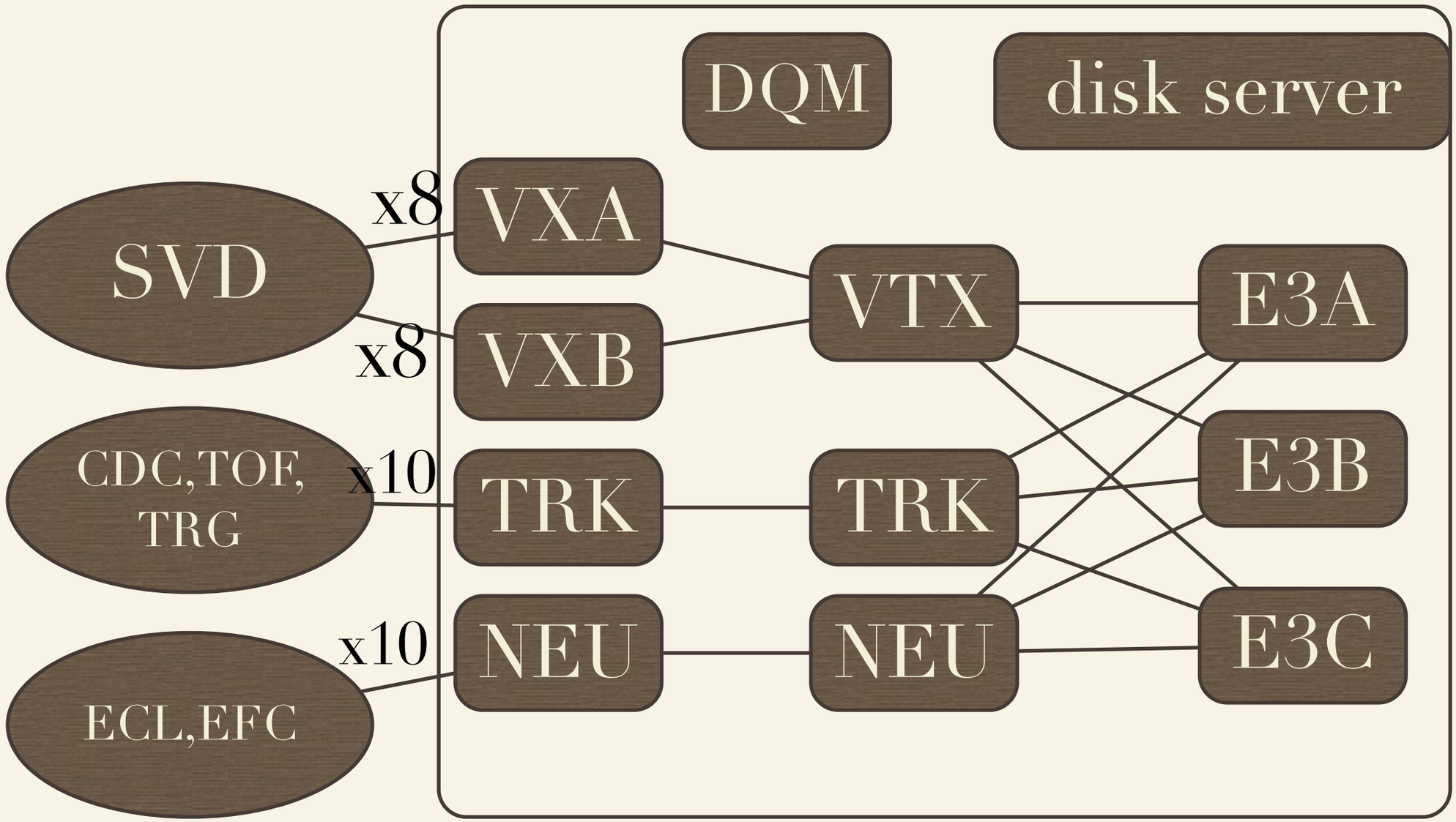


Single unit has two final layer (E3)

of PCs

- 1 unit of EFARM achieves 200MB/s
- # of EFARM is 15
- # of CPU is still 165

One more E3

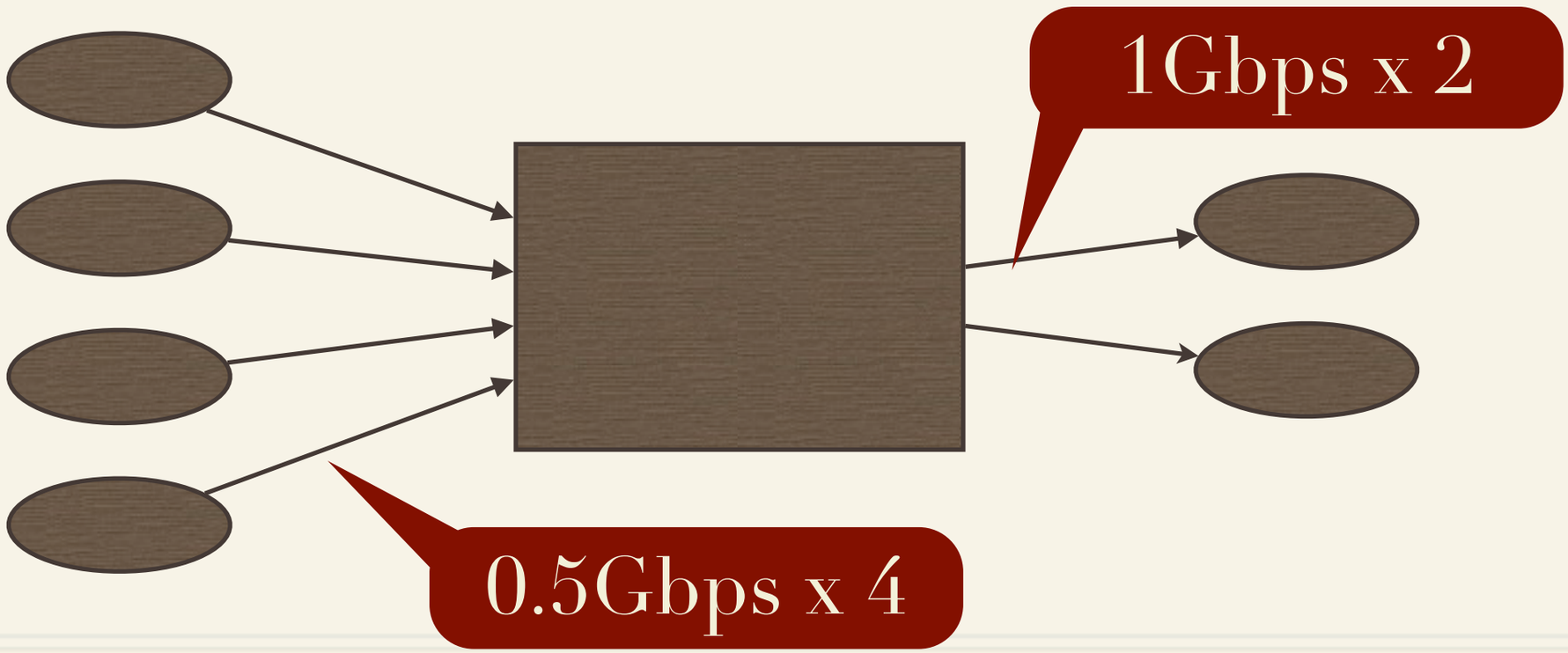

DQM
disk server
SVD
x8
VXA
x8
VXB
VTX
E3A
E3B
E3C
CDC,TOF,
TRG
x10
TRK
TRK
ECL,EFC
x10
NEU
NEU

of PCs

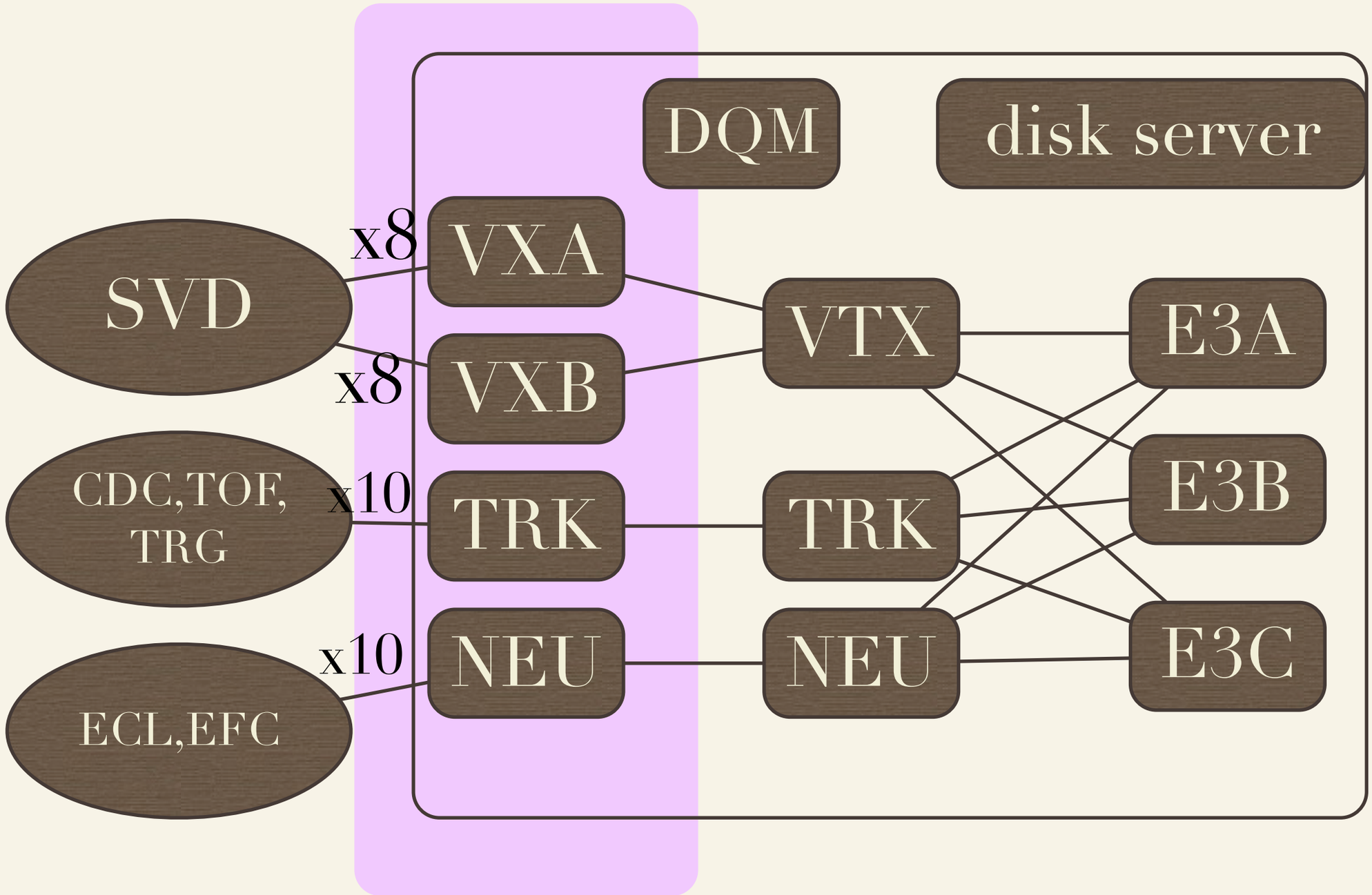
- One EFARM unit achieves 300MB/s
- # of units is 10
- # of PCs is 130
- Still large.
  - This # contains no spare

of in / # of out

- If the CPU processing power doesn't limit the throughput, the bottle neck is where # of input > # of output



Now bottleneck is here

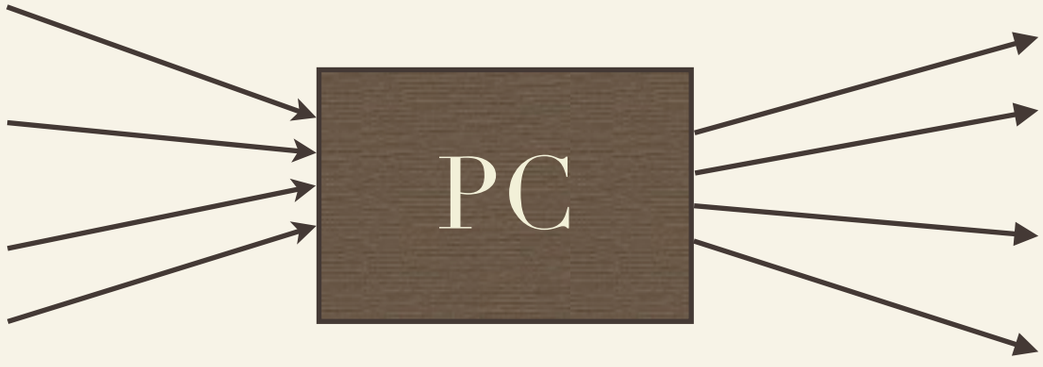


- If # of output is always same of input, I/O bandwidth will be limited CPU power.

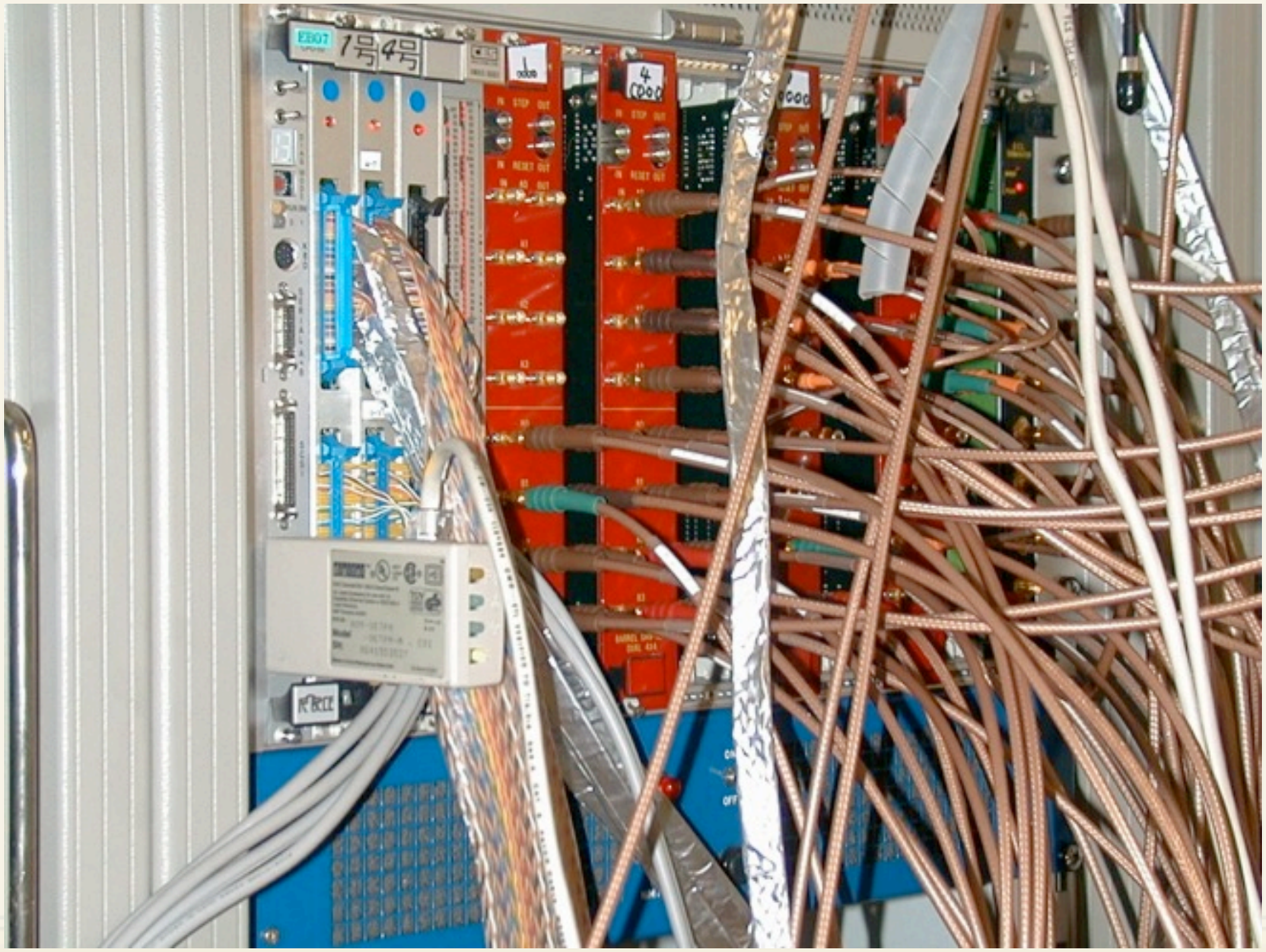
Barrel Shifter method

How to reduce # of PC?

- If the # of PC is most heavy problem, Barrel shifter scheme can be adopted again.
- Use 1 PC as 4x4 Barrel Shifter of GbE

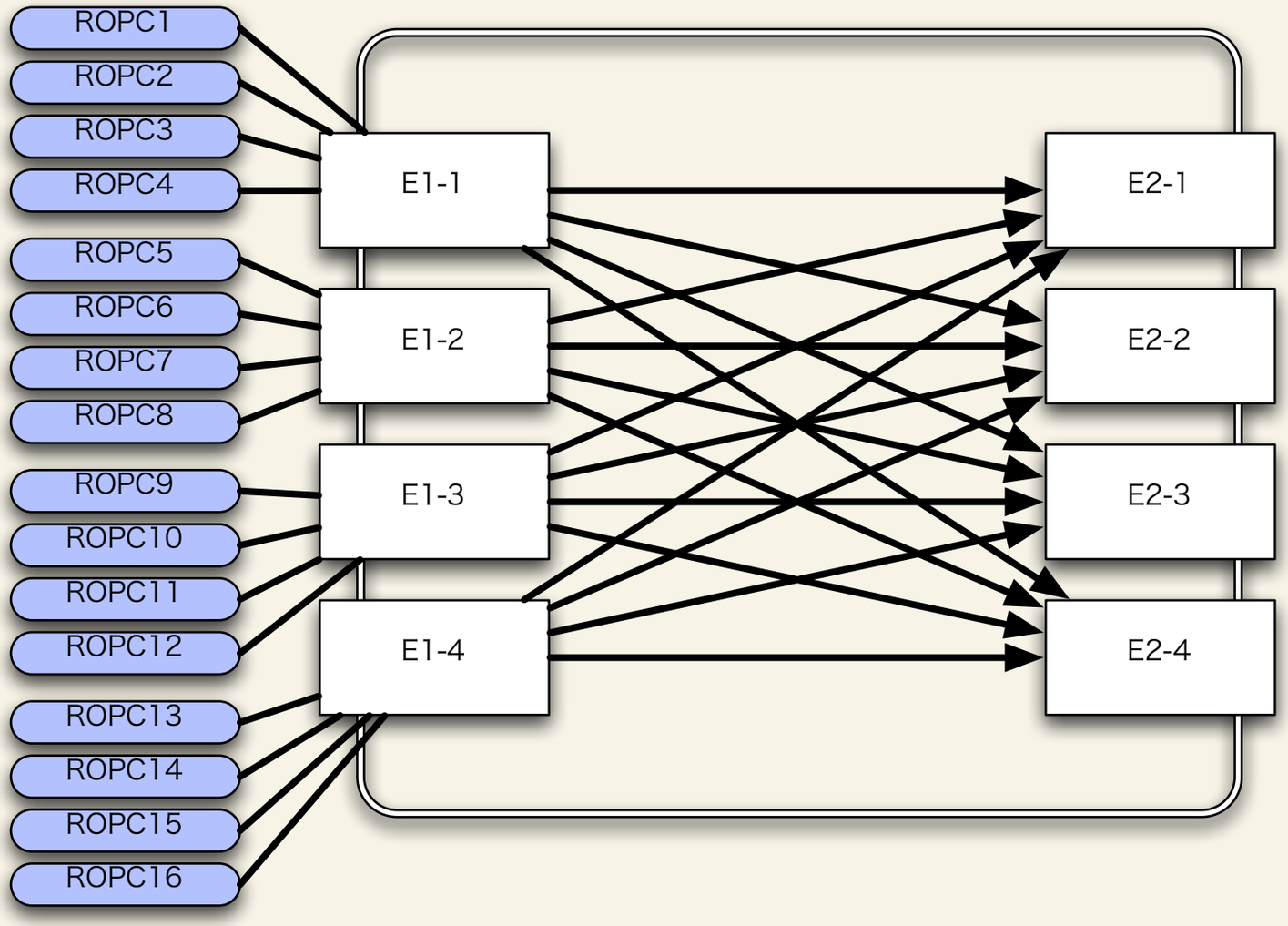


GLINK Barrel Shifter E.B. (~2001)



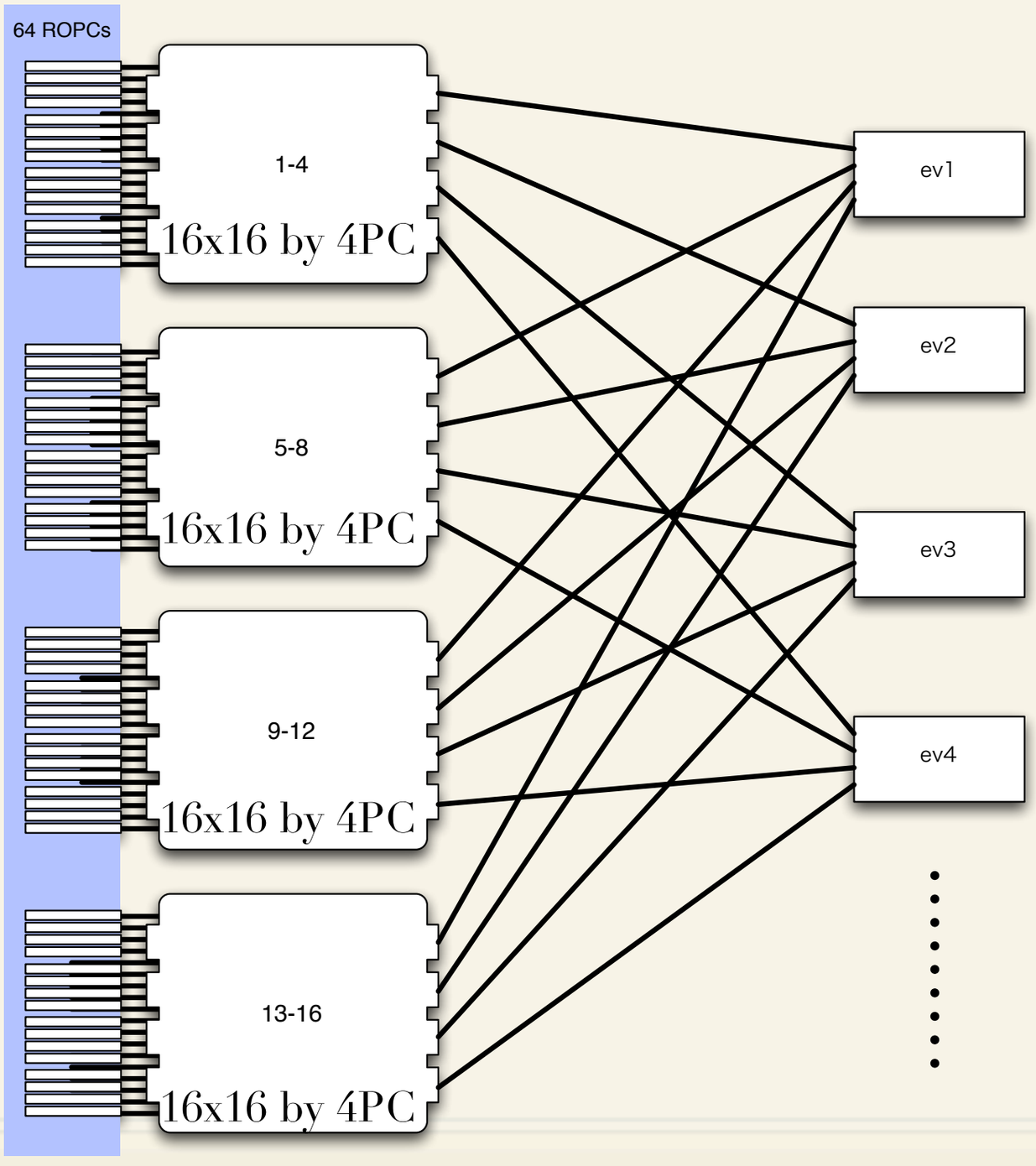
2009年7月7日火曜日

16x16 by 8PCs, 2layer



16 output outs only 160MB/s, so insufficient

64x64 by 48PC by 3layer



Weakness

- There is no redundancy
- In current Belle scheme, EFARM2 can run w/o any problem even if one PC in EFARM1 corrupts.

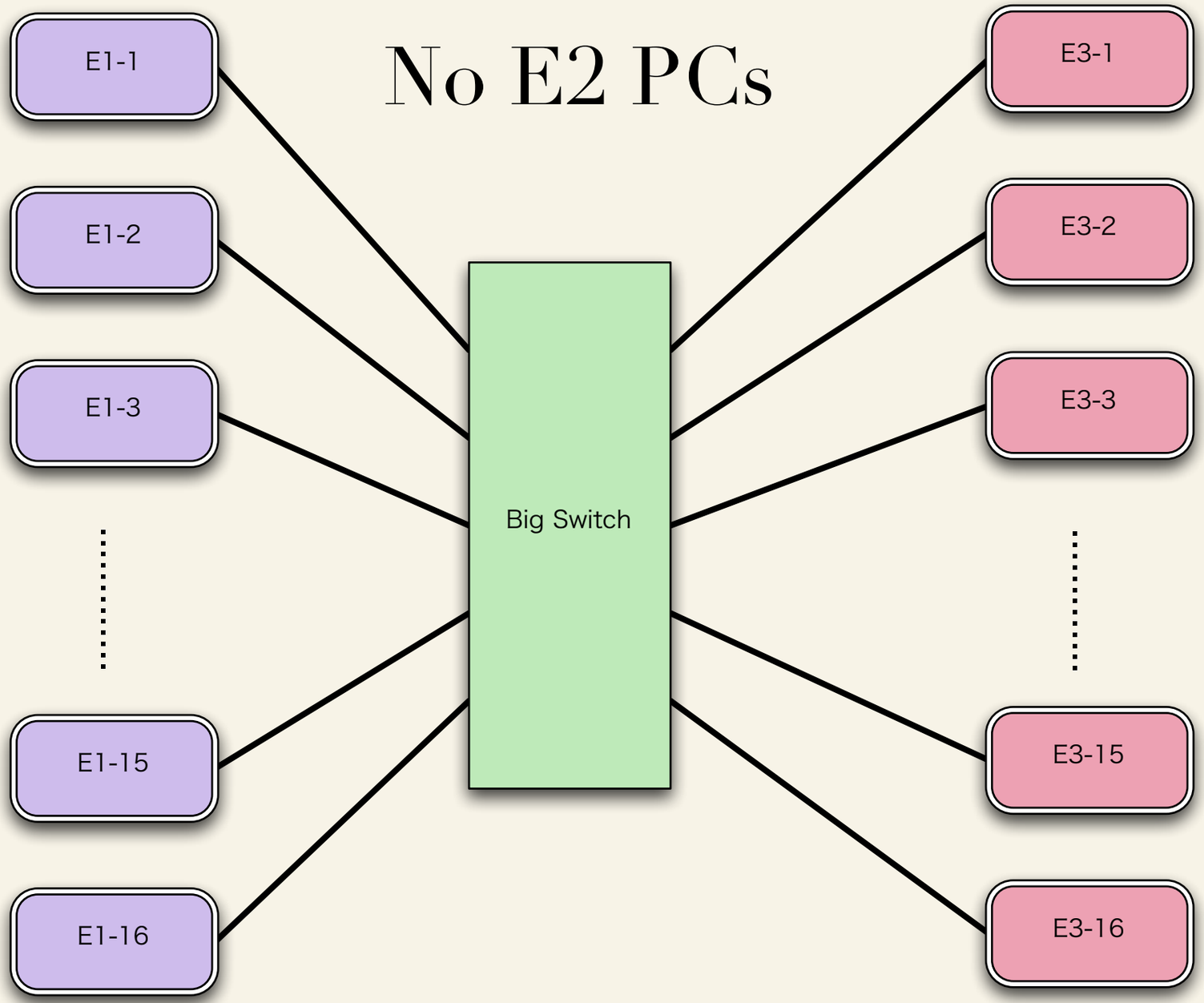
Application of network switch

noDAQ guys say,

- Why don't you use network switch?
  Recent switch has sufficient speed.
- If you use, E.B. topology will be simple and you will get both of redundancy and flexibility.

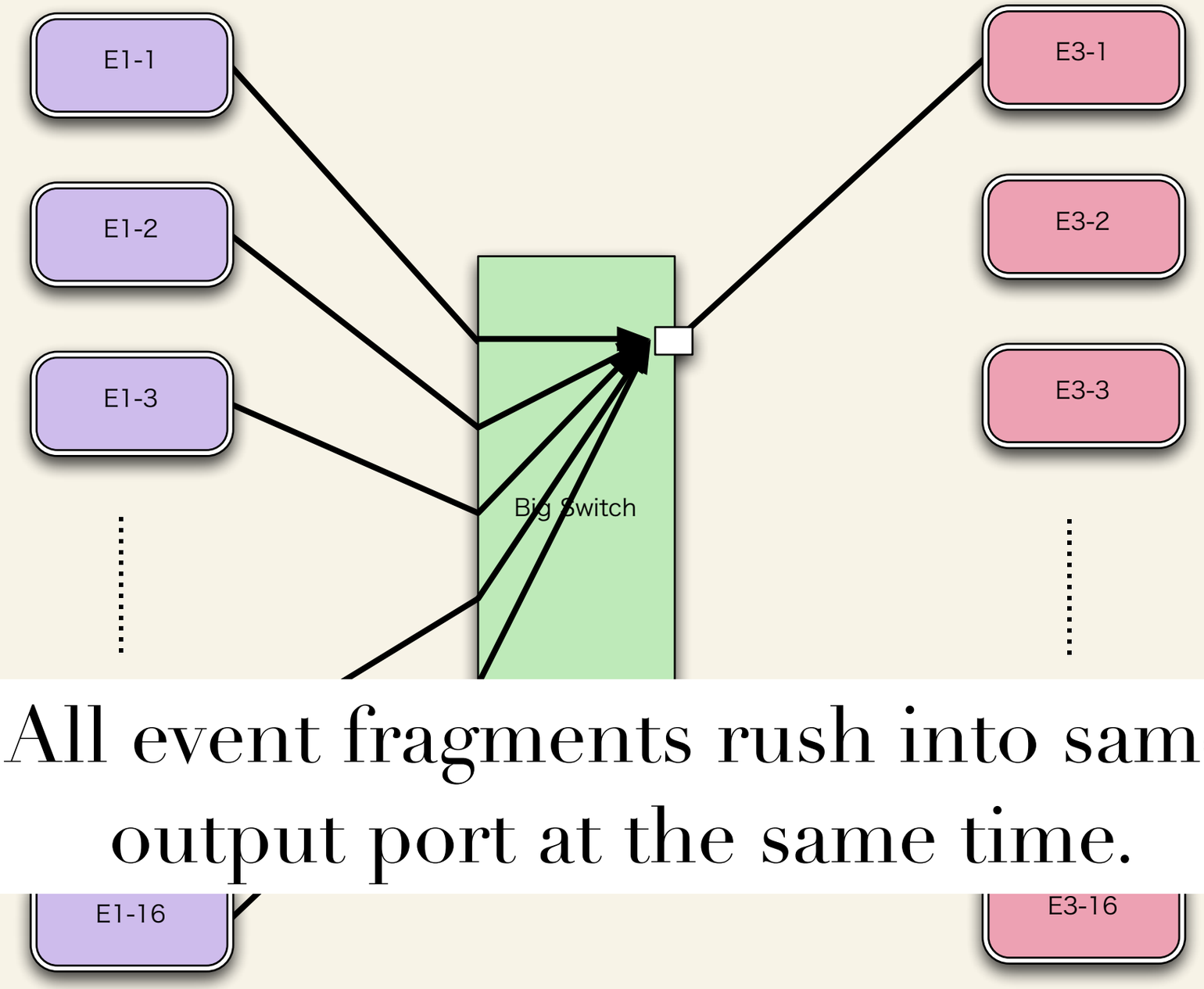
Simple topology

No E2 PCs



# of PC is only 32, also hot-standby PC can be adopted.

Problem: Traffic jam

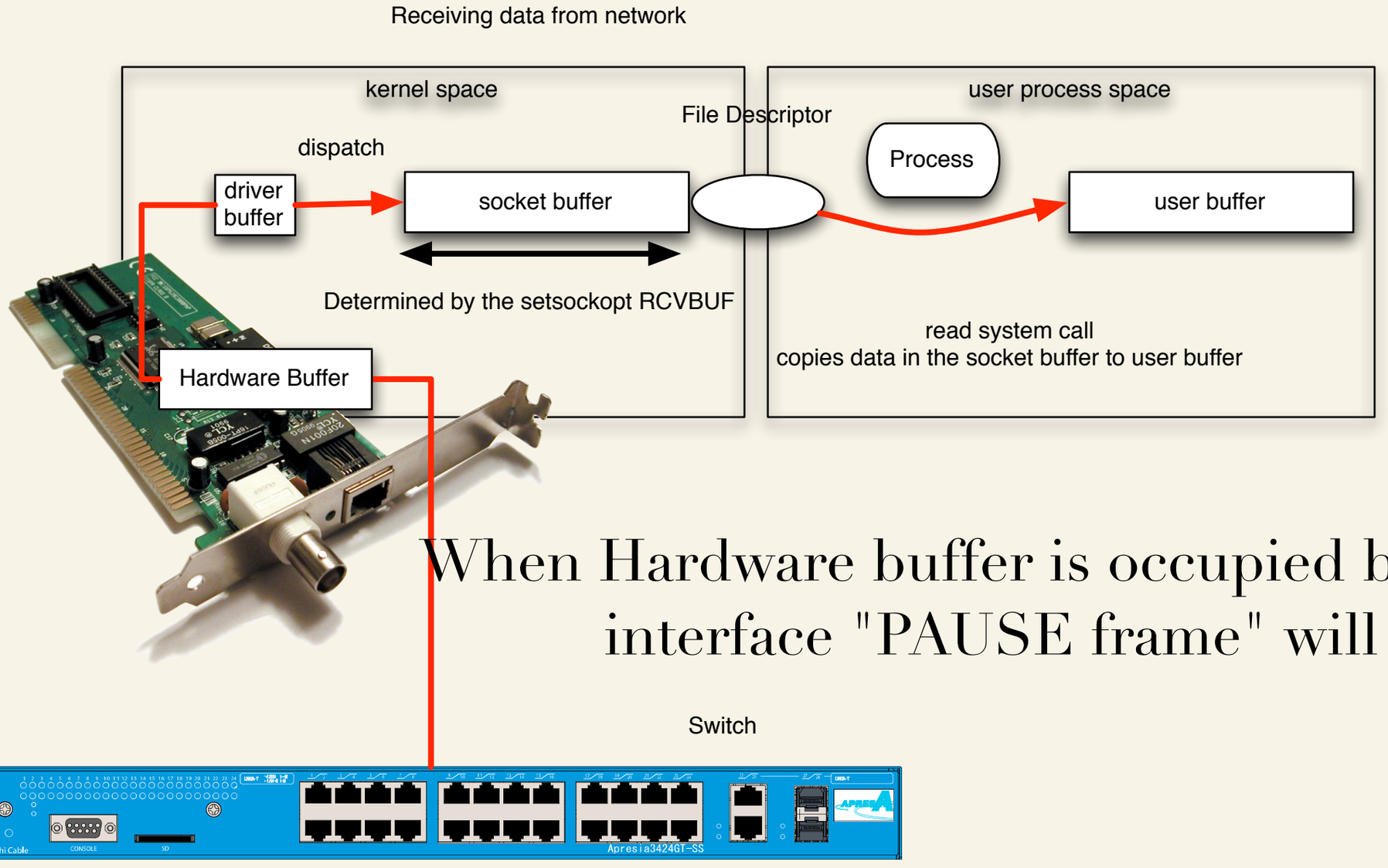


All event fragments rush into same output port at the same time.

Back Pressure

- To avoid collision, Ethernet has 802.3x flow control specification.

Back Pressure



When Hardware buffer is occupied by unread data, interface "PAUSE frame" will be sent

Buffer overflow

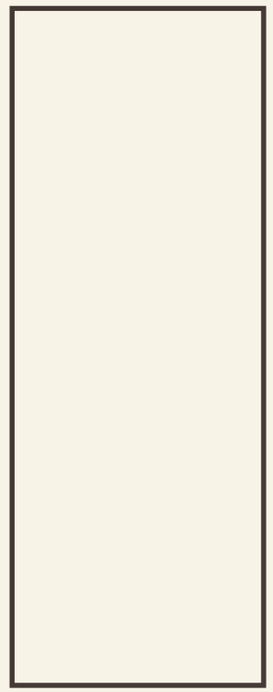
- Most shortest buffer is hardware.
- So kernel must read data from hardware to kernel buffer very very frequently.
- Also user process must read very quickly the kernel buffer.
- Once kernel or process neglect the timing, the hardware buffer will be overflowed, PAUSE will be sent.

is it supported?

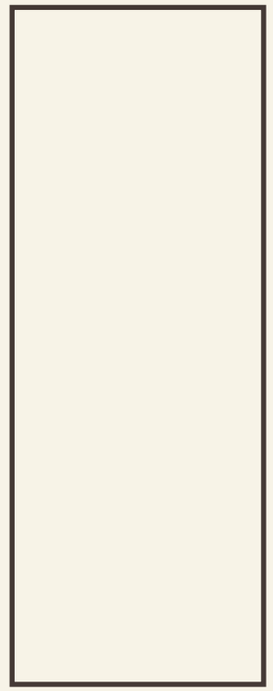
- Receiving and waiting are supported
- Linux sends PAUSE
  - can be turned on/off by ethtool
- Switch never sends PAUSE
  - Principally it can send, but no such production

Buffer

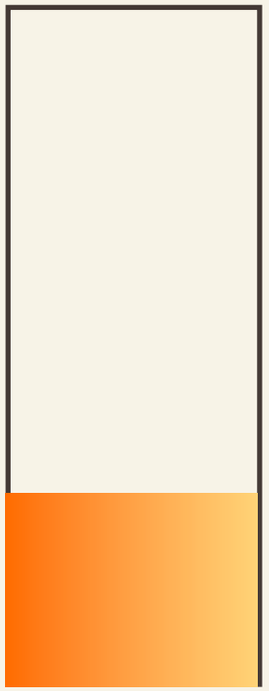
Buffer



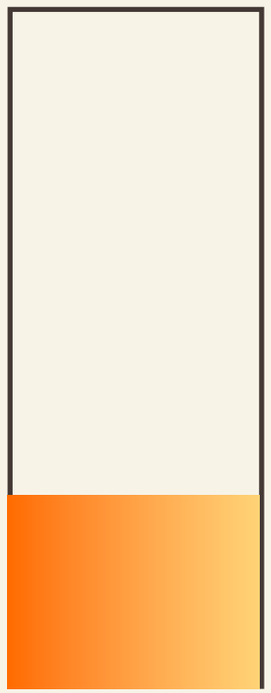
Buffer



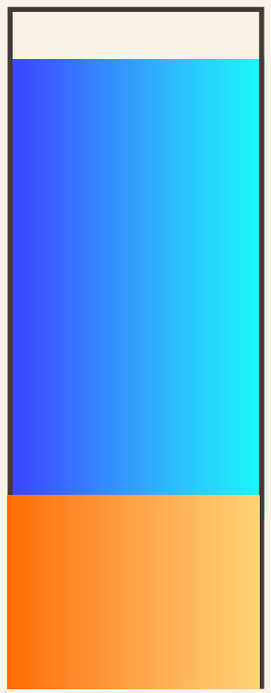
Buffer



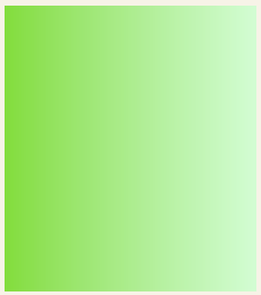
Buffer

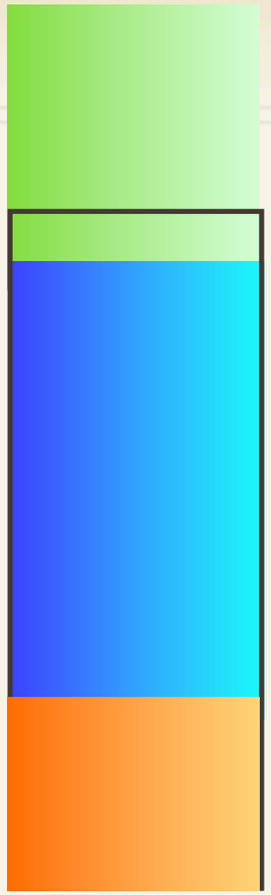


Buffer

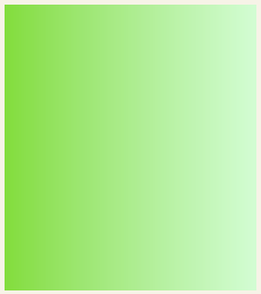


Buffer

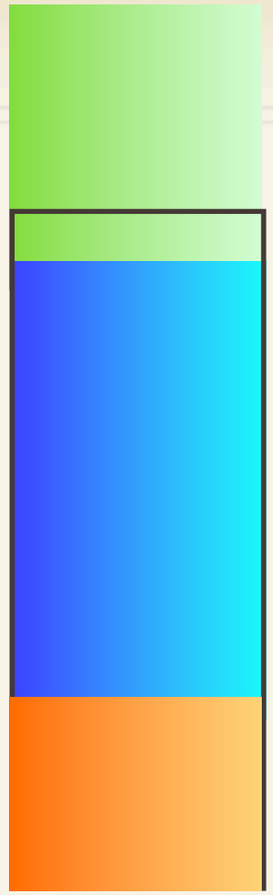




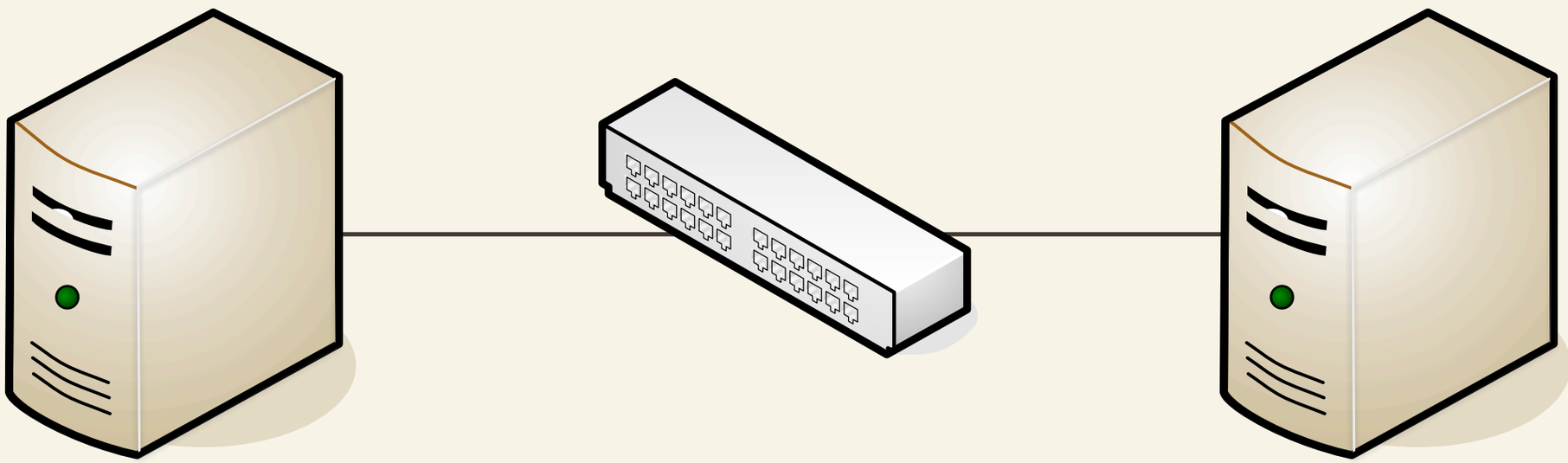
Buffer



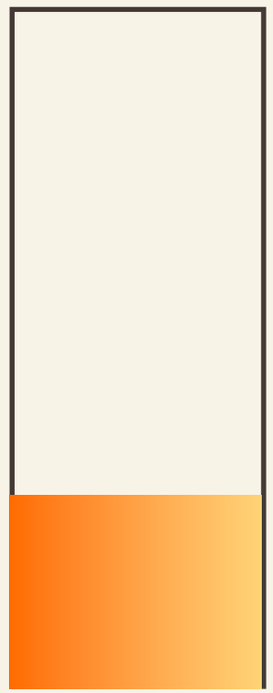
Oh, I have to wait

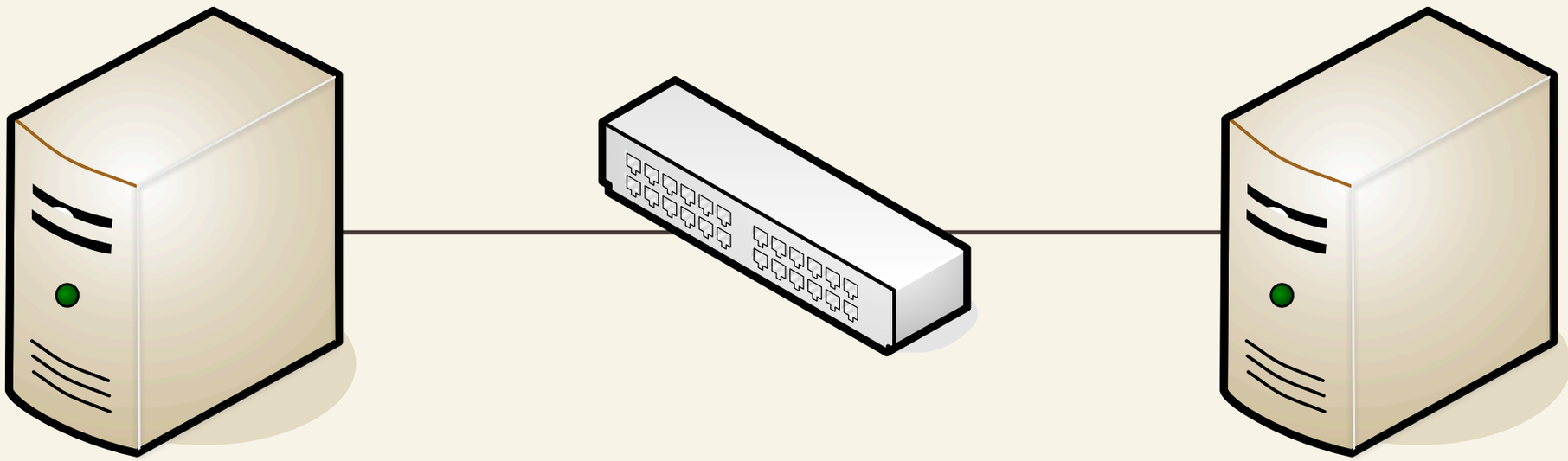


Buffer

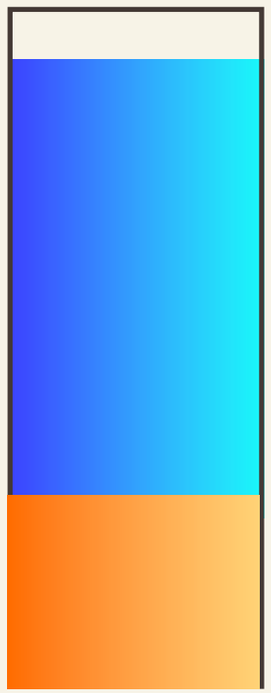


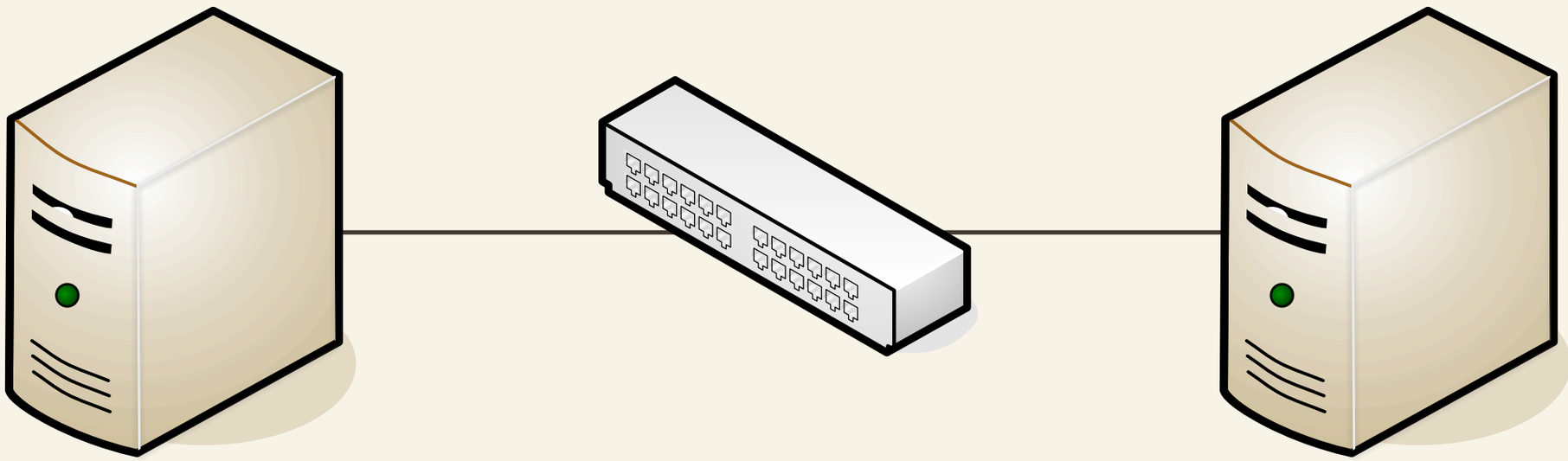
Buffer





Buffer

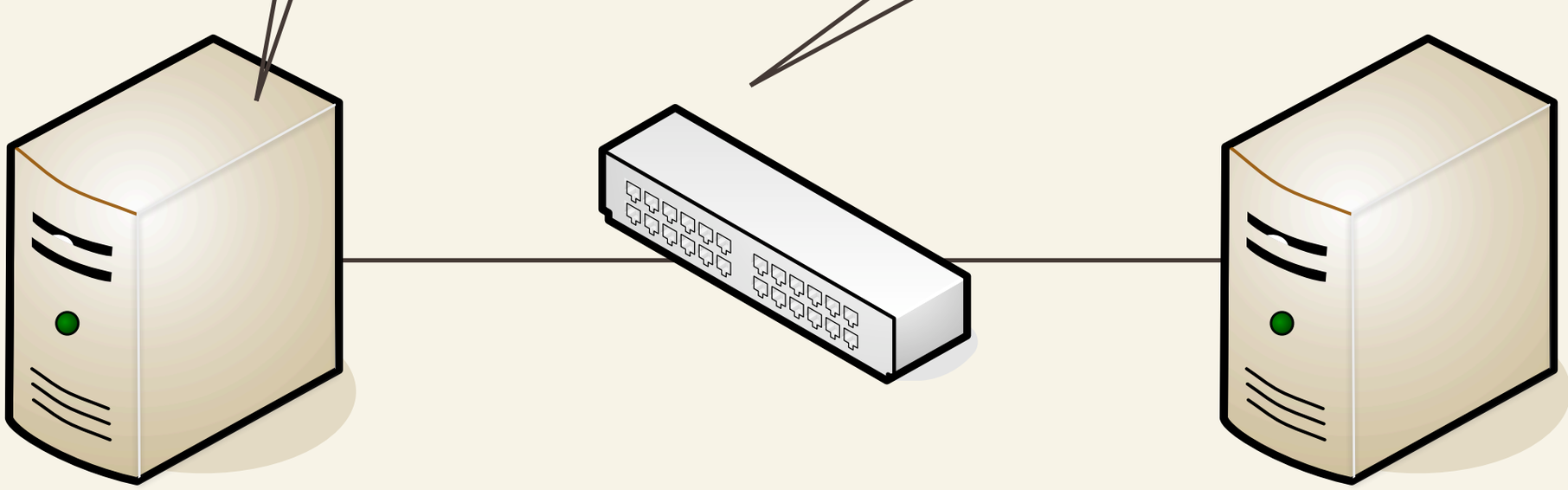




Buffer

Data is
successfully sent.

Oh, I have to wait



Where is unsent data?

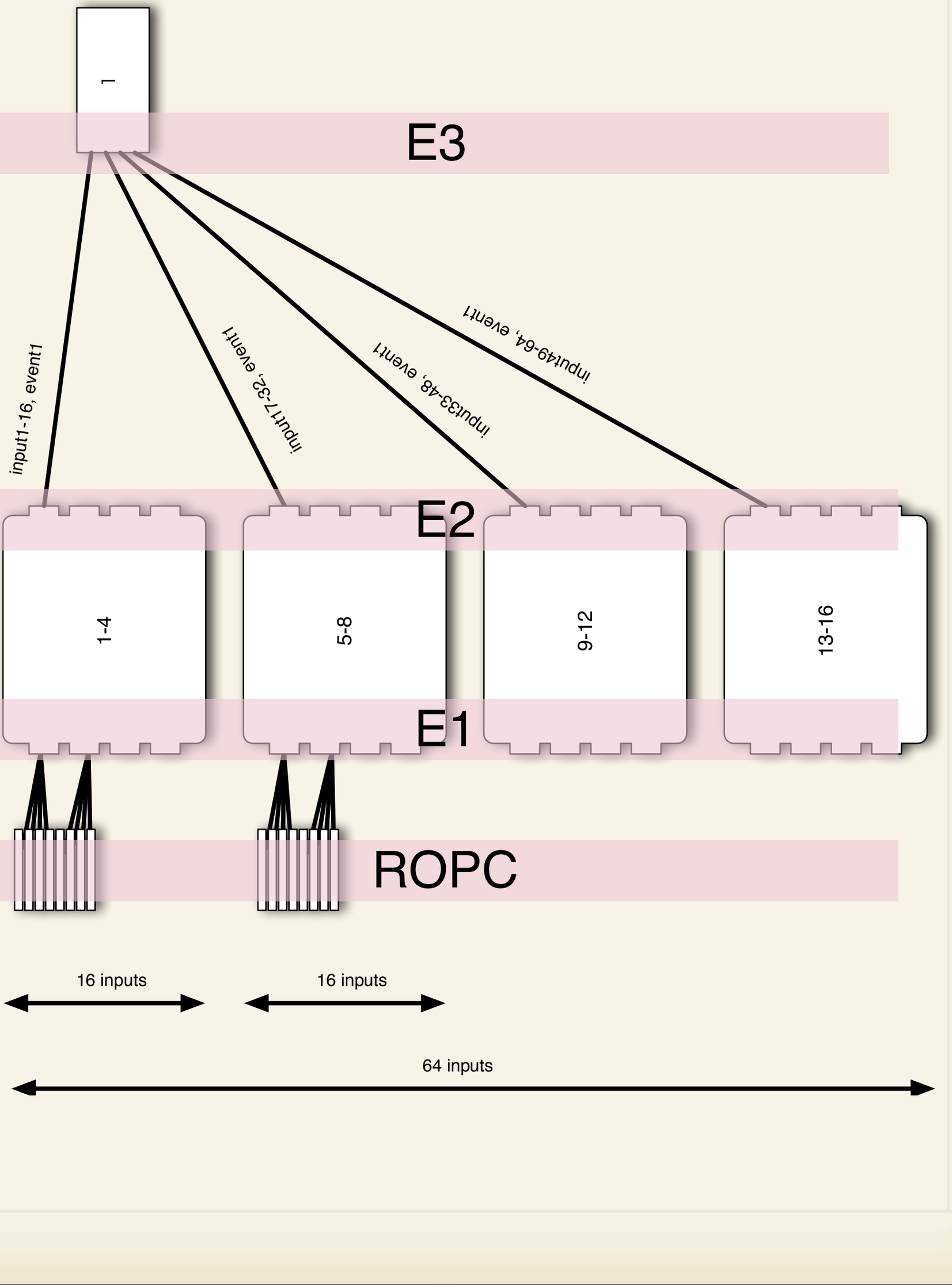
- The hardware buffer in the switch
- Whats happen when it overflows?
- Switch just drops packets silently.
- Switch doesn't send PAUSE to sender.

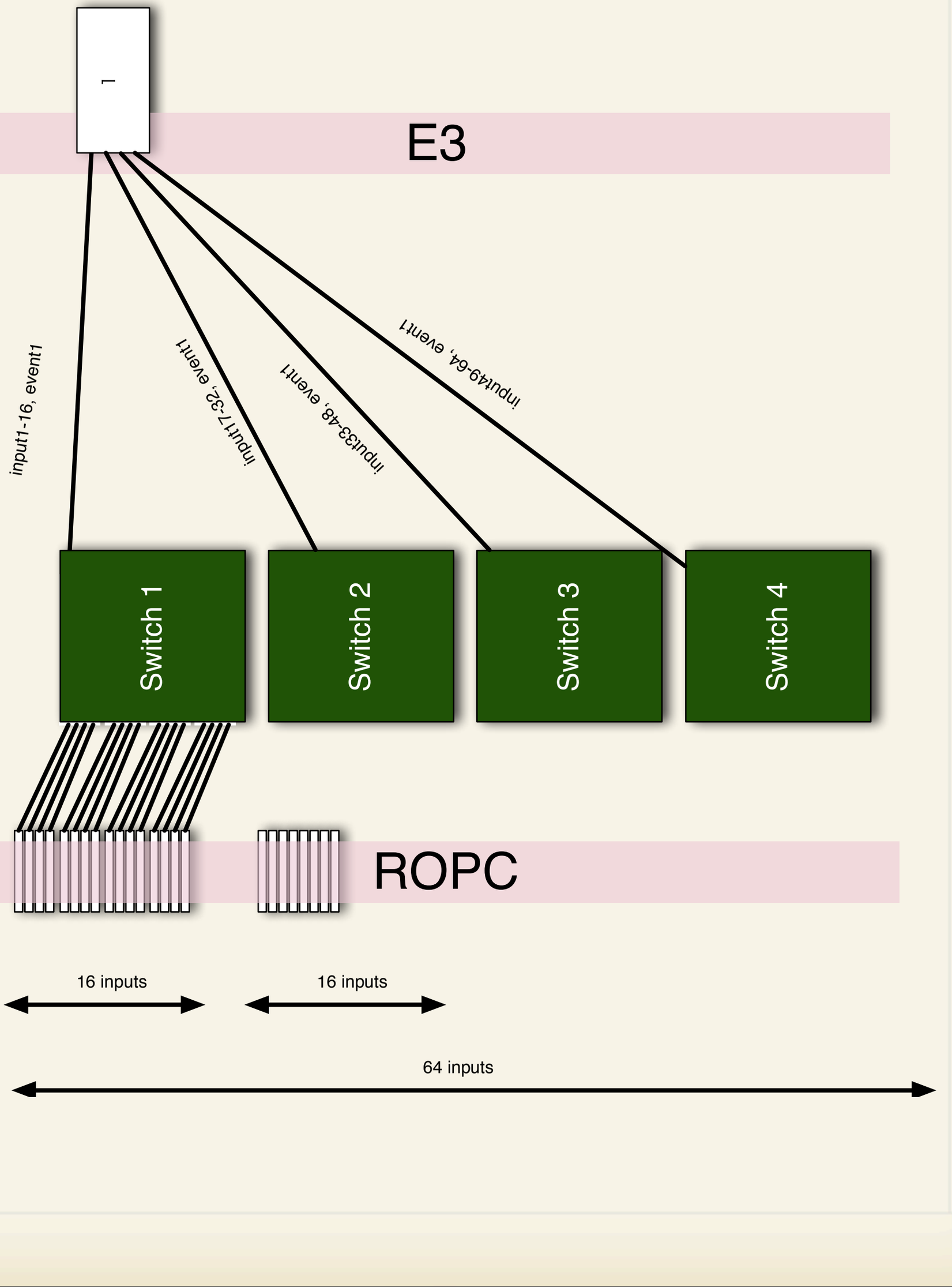
Depth of the buffer

- Most of the switch has only 30-60kbytes as the output queue for every interface
- Short to hold one full-built event.
- Only layer1 or layer2 can use.

Fix by software

- Like LHC experiments
- PULL style data transmission
  - Every time receiver calls sender to receive data.


1
E3
input1-16, event1
input17-32, event1
input33-48, event1
input49-64, event1
E2
1-4
5-8
9-12
13-16
E1
ROPC
16 inputs
16 inputs
64 inputs


1
E3
input1-16, event1
input17-32, event1
input33-48, event1
input49-64, event1
Switch 1
Switch 2
Switch 3
Switch 4
ROPC
16 inputs
16 inputs
64 inputs

In the case of that,

- 4 x Switch
  - # of port > 32
    - wirerate, non-blocking
    - depth of the output queue > 50kB
- 16PCs as E3.

Strategy

- Confirm PC has sufficient performance 4x4 barrel shifter of GbE.
  - 2x2 seems to be OK with 3G C2D.
- Confirm we can avoid the packet loss via network switch under the Belle II data rate.

Summary

- Out candidates
  - Barrel shifter with only PC, no switch
    - Apply switch at Layer1 and 2
  - Now testing whether PC has sufficient power for 4x4 Barrel shifter of GbE.