# Upgrade plan of Trigger and DAQ for SuperKEKB

R.Itoh, KEK

SuperKEKB proto-collaboration meeting
2/20/2008

# Outline

1. Target performance and requirements
2. Trigger upgrade
3. DAQ upgrade
4. Summary

# 1. Target performance and requirements

- Expected luminosity @ SuperKEKB ~ 2.0 x $10^{35}/cm^2/sec$ (first stage)

- Keep the same L1 trigger policy as that of Belle

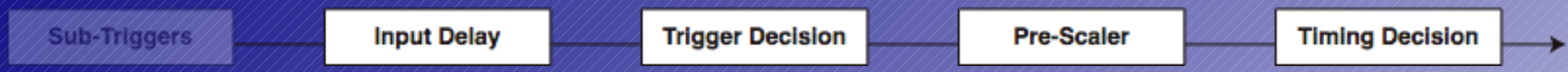
| | Current Belle | Upgraded KEKB |
|---|---|---|
| Typical L1 rate | 0.5kHz | 10kHz |
| (Maximum L1 rate | ~1kHz | ~30kHz ) |
| L1 data size(in) | 40kB/ev | 300kB/ev |
| flow rate(in) | 20MB/sec | 3GB/sec |
| reduction | 1 | 1/3 |
| data size(out) | 40kB/ev | 100kB/ev |
| flow rate(out) | 20MB/sec | 1GB/sec |
| L3+HLT reduction | 1/2 | ~1/10 |
| Storage badwidth | 20MB/sec (including full rec. results by RFARM) | 250MB/sec |

* Powerful data reduction at each step is the key.

# 2. Trigger

Y.Iwasaki

# Level 1 Trigger




- Keep Belle triggering scheme
  - Main triggers : charged tracks, energy sum, and energy clusters
  - Fast decision : latency ∽ a fewμsec
  - Tight but efficient logic : S/N >> 0.1, ε(Y) ∽ 1
- Input
  - Charged track (θ, φ, pt, pz) : order(1000 bits)
  - Energy sum (θ, φ, E) : order(1000 bits)
  - Energy cluster (θ, φ) : order(400 bits)
- Output rate
  - Average L1 rate ∽ 10 kHz, Maximum ∽ 30 kHz
- Present key issues
  - Timing sources (PID ?) and precision ( ∽ 10 nsec ?)
  - Latency ( ∽ 5 μsec or shorter?)

# Timing Sources and Precision

- To make redundant system, we like to receive multiple timing signals
  - Very useful to check the trigger timing if we have multi-sources
- We expect the timing signal from ECL and PID
  - ECL : same precision as it is now [Order(~30ns)]
  - PID : as a replacement of precise timing of present TSC timing

    -> TOP can provide O(10)ns precision timing
- Timing precision
  - For the event reconstruction, ECL precision is enough
  - SVD requires O(10) ns timing for their L0 trigger

# L1 Latency

- We requested DAQ group to give us ~5 μsec for L1 latency
  - To perform complicated triggering algorithm if we have 5μsec
  - 5μsec is present agreement
- SVD may not wait for 5 μsec
  - Limitation by the pipeline depth of APV25 = 3.8us

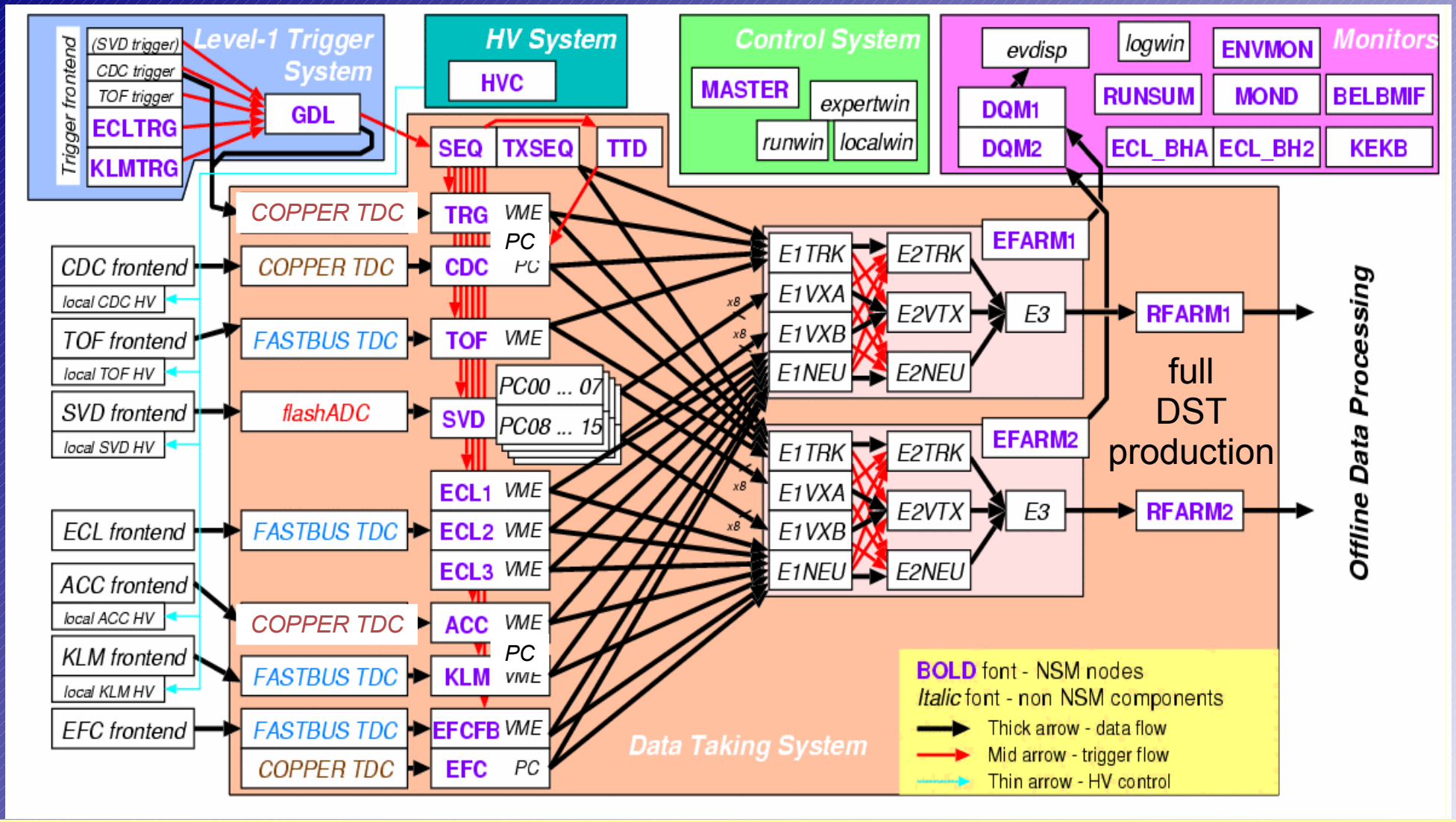
Discussing the shortening of the latency to ~3usec

# # Input Channels

- Correlation between sub-triggers are important
  - L0 signal (CDC+TOF)
  - Bhabha and gamma-gamma (ECL+CDC)
- It's easier to make such a correlation trigger if GDL receive finer sub-trigger information
  - CDC : track phi and theta, pt, and pz
  - ECL : cluster phi and theta, and energy sum
- We start to design GDL to accept ~1000 bit information

# 3. DAQ

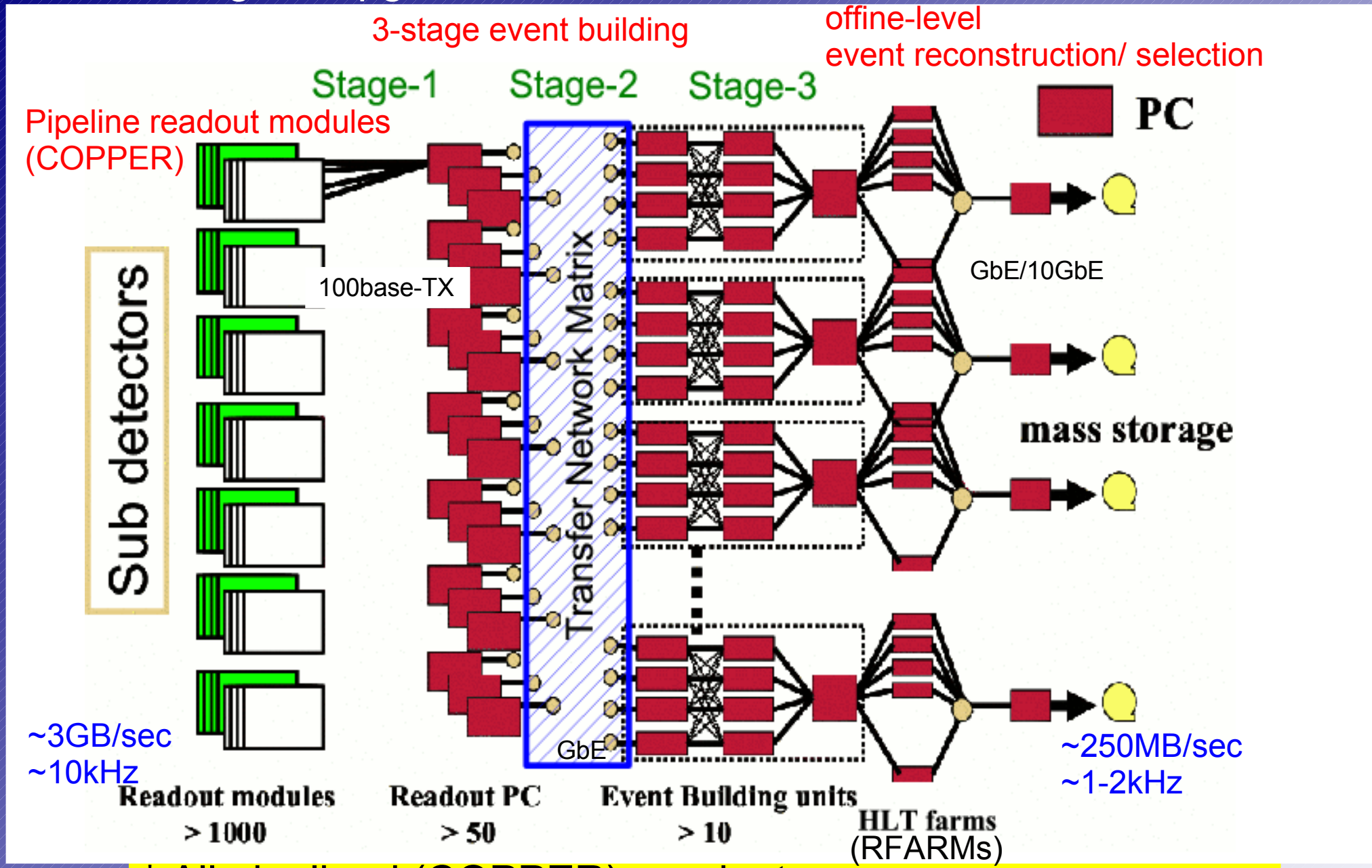
# Current Belle DAQ system




* Readout: Replacement of FASTBUS TDC to COPPERs is in progress.
  (Already replaced: CDC, ACC, TRG and EFC)
* Event builder / HLT(RFARM) : Modularized and operated with 2 units.
  Easily expandable by adding more units.

# Global design of upgraded DAQ

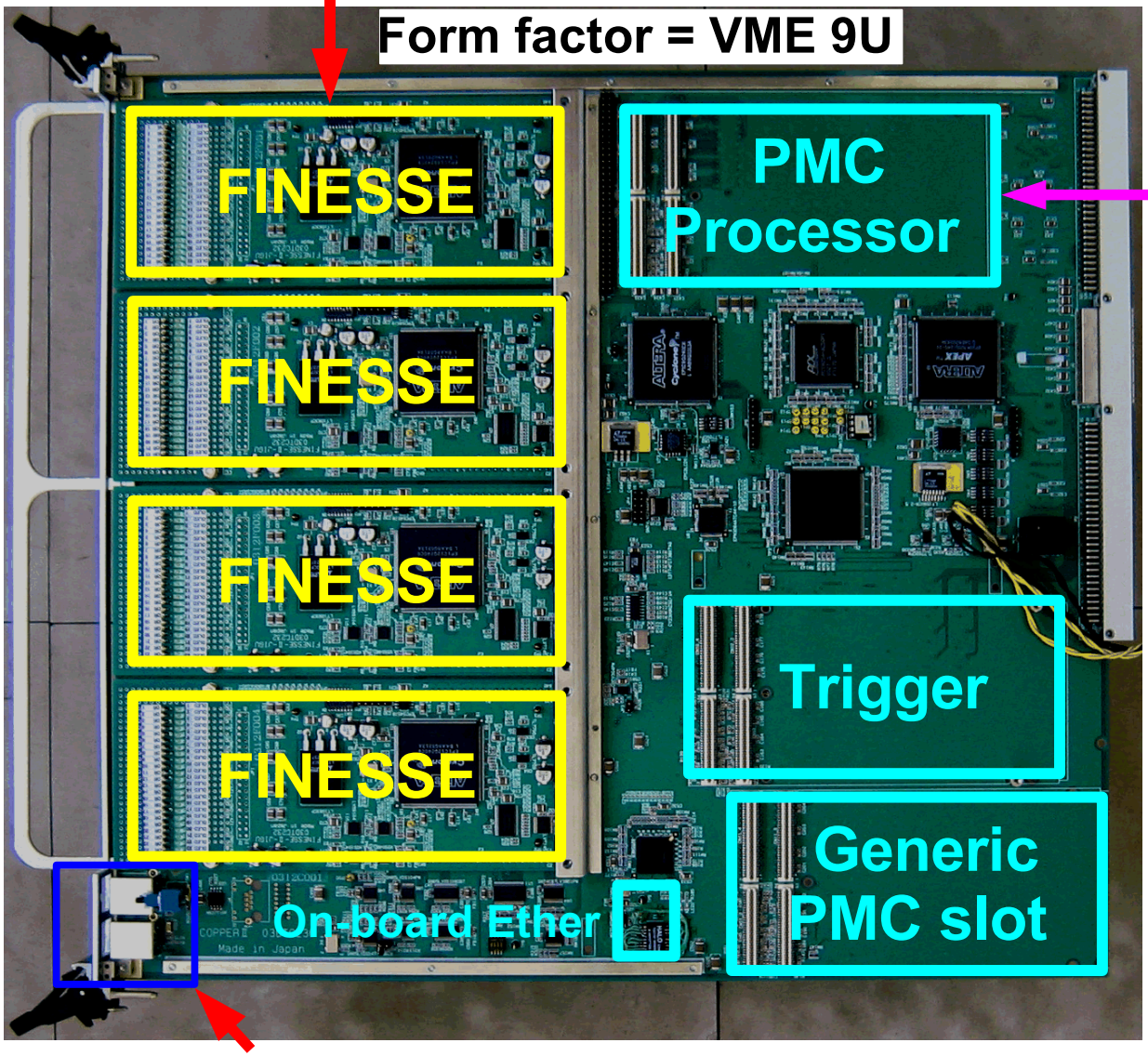
**"Smooth" upgrade from existing DAQ**




* All pipelined (COPPER) readout
  - The COPPERs currently used (+ implemented in FY2008) are all recycled!
* More units of Event Builder+HLT module : O(10)

# Unified Pipeline Readout Module (COPPER)

Digitizer cards (implemented as daughter cards)




two100base-TX ports
(for control and data flow)

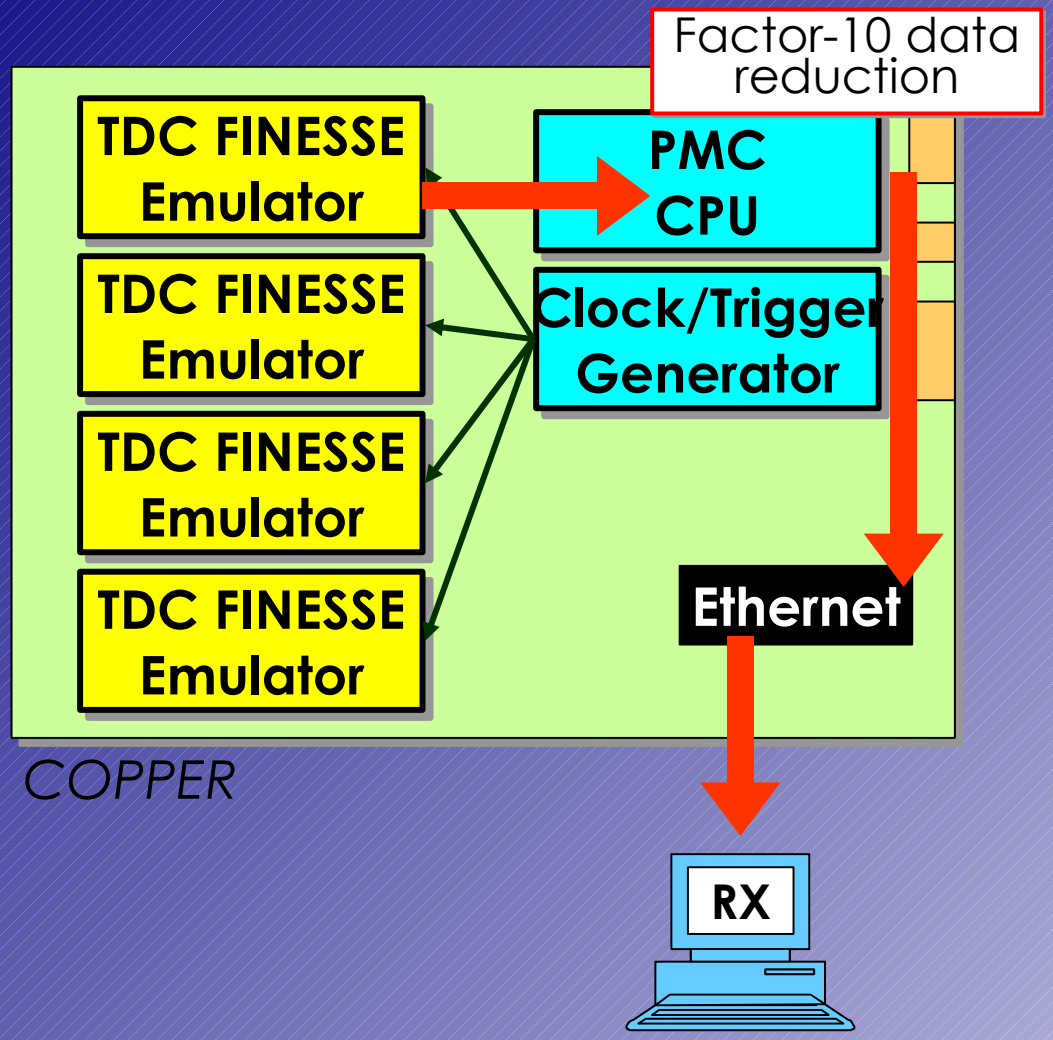
CPU card
(operated by Linux)

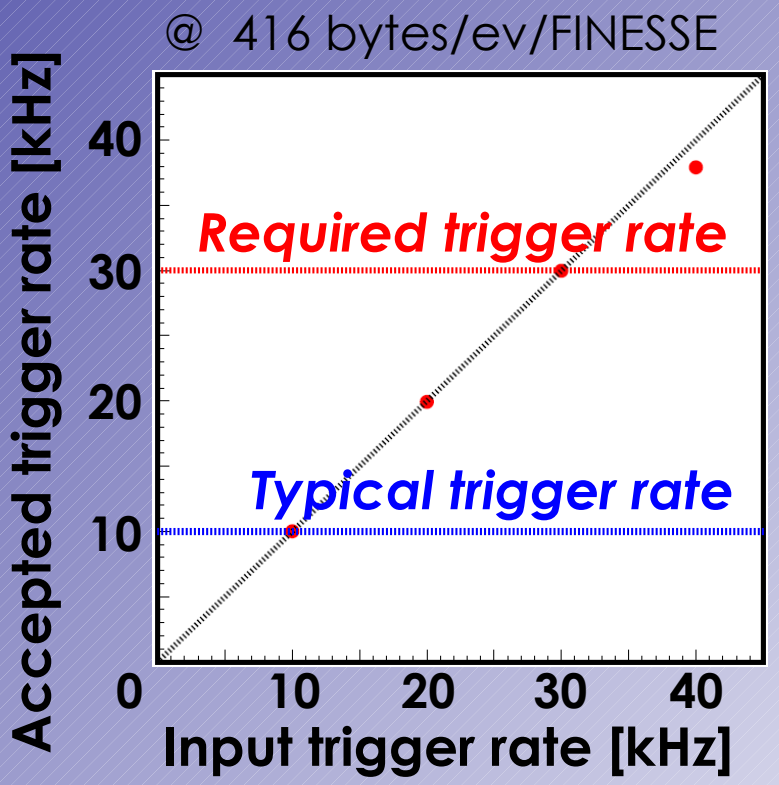
RadiSys
EPC-6315

- **Intel PentiumIII 800 MHz** w/ **256 MB memory**.
- Network booted
- RedHat Linux 9

GE-FANUC
PSL-09
- Intel PentiumM 1.4GHz

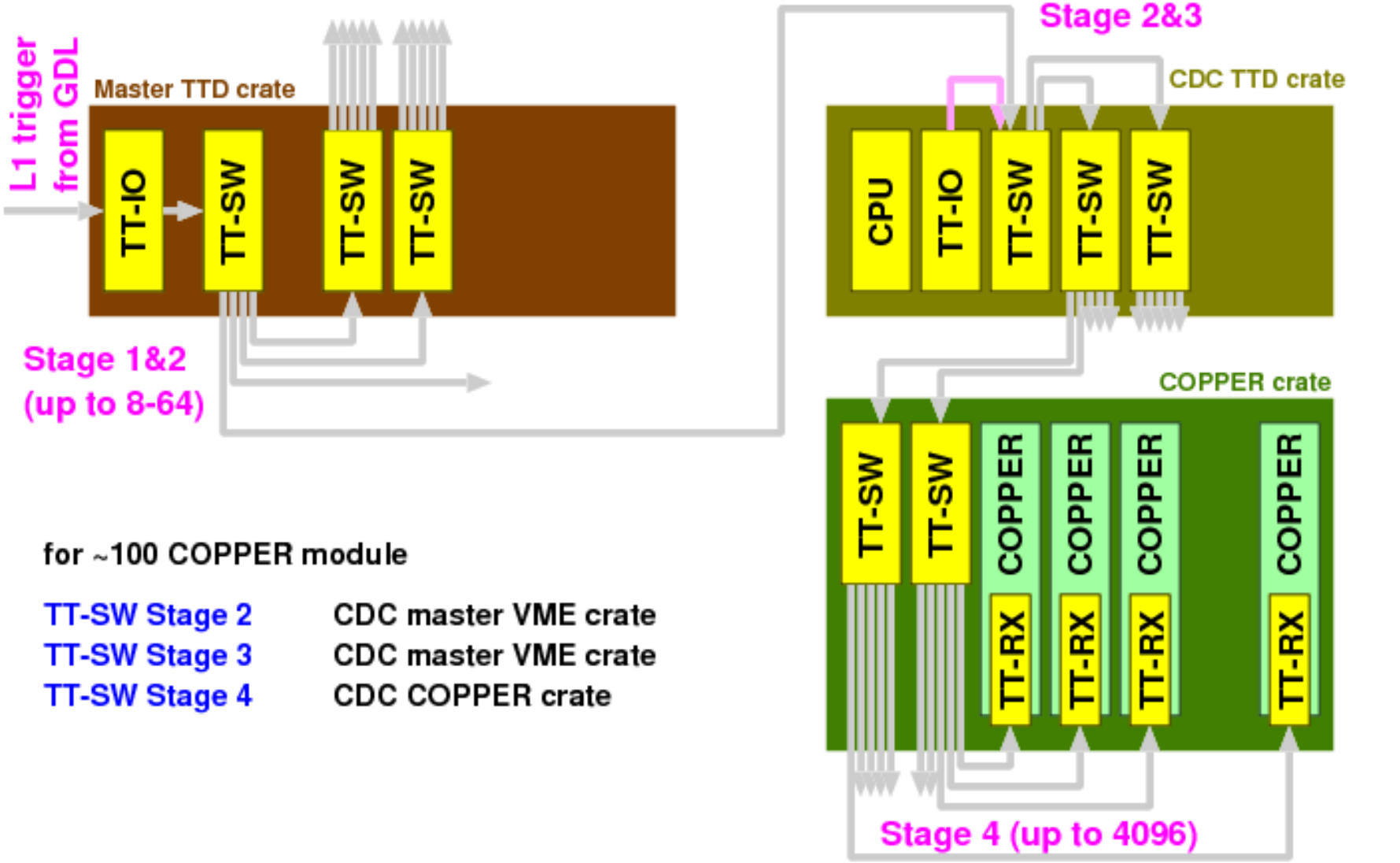
# High rate test of COPPER using a test bench




**The COPPER works > 30 kHz input rate**

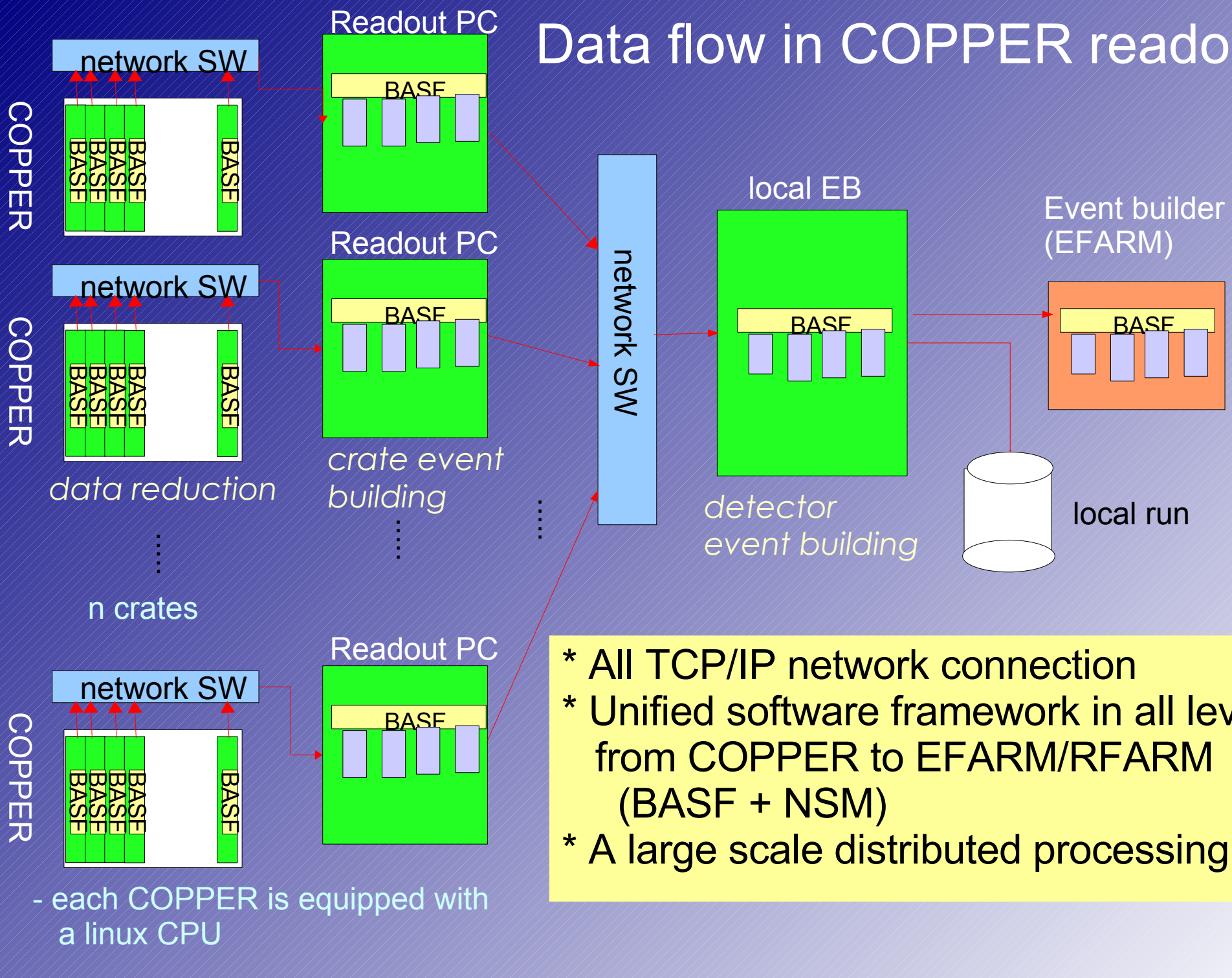
# Timing Distribution System for COPPERs




for ~100 COPPER module

| | |
|---|---|
| TT-SW Stage 2 | CDC master VME crate |
| TT-SW Stage 3 | CDC master VME crate |
| TT-SW Stage 4 | CDC COPPER crate |

# Data flow in COPPER readout




* All TCP/IP network connection
* Unified software framework in all levels from COPPER to EFARM/RFARM (BASF + NSM)
* A large scale distributed processing

# Digitizers and Front-end readout

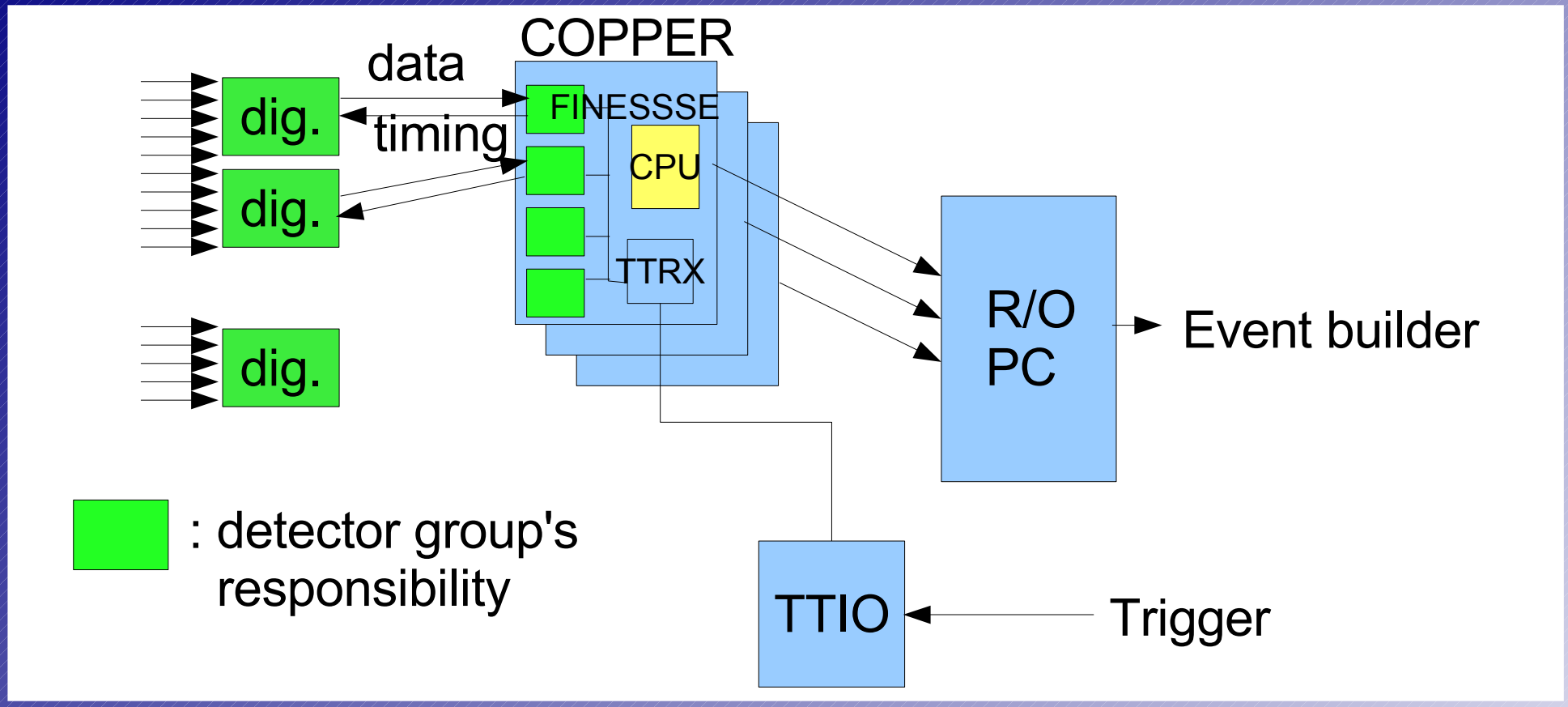
- Belle's signal digitizing : Unified scheme (except for SVD)
  = Q-to-T + multihit TDC
  all placed in E-hut

SuperKEKB : different digitizers for different detectors

SVD: APV25 readout placed in E-hut
CDC: ASD-chip based readout located near detector
PID: Special ASIC located near detector
ECL: Wave-form sampling digitizer located near detector
KLM: not yet decided

- COPPER is used as a “unified readout module” between digitizers and event builder units.

# COPPER as Read Out Module (ROM)



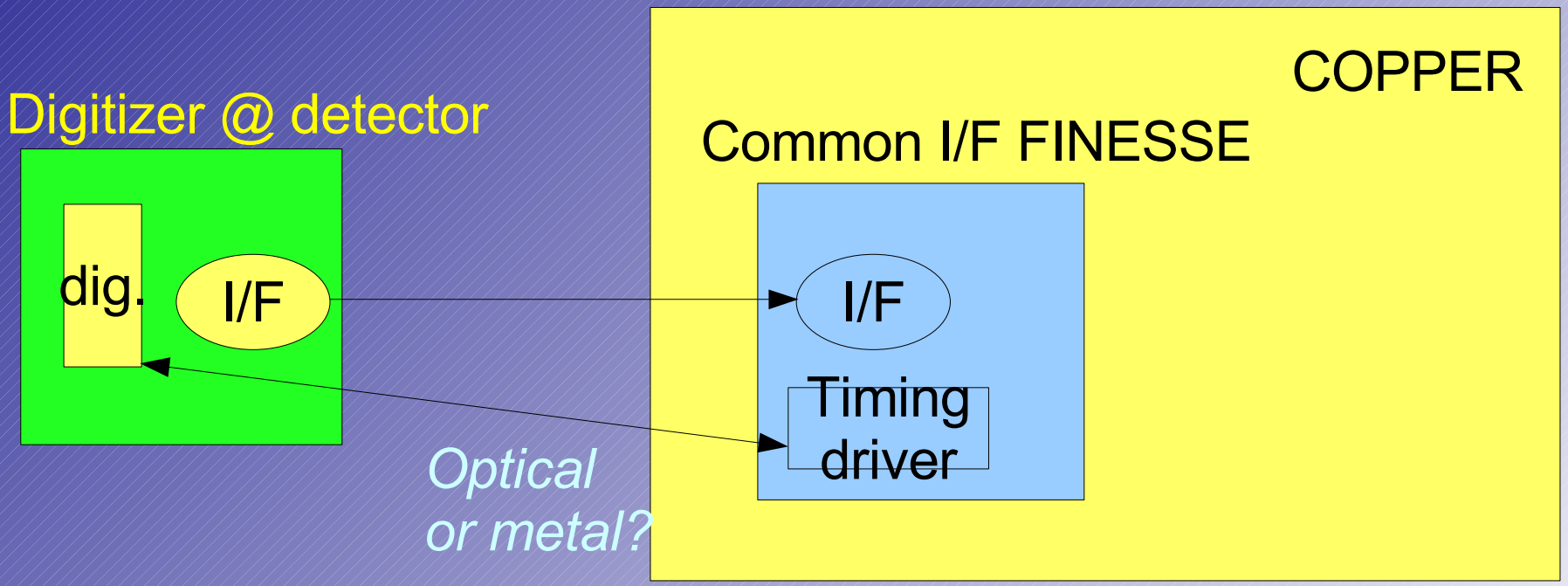

* We need COPPER CPUs for the data reduction at the early stage.
  <- provides a huge computing power by the distributed computing which is essential to meet the requirements in the data reduction.

* We already have the scheme to handle both control and data flow with this scheme and the development cost can be minimized.

# Idea of common interface FINESSE

- It is desired to prepare a common interface FINESSE to connect external digitizers.

- Need to develop
  * timing distribution scheme (over long cables)
  * standard data transfer protocol




- CDC's new electronics could be the test case.
- We started to think about this common interface FINESSE.

# Timing distribution: Issues

- Timing distribution : TTD system <- serial bus connection
  which can distribute timing to COPPERs inside E-hut only.
  -> Distribution to outside E-hut -> too far.....
  max. 10m. with current sys.
  => need to think about optical / parallel metal options

- Trigger-busy handshake -> somewhat a large dead time
  (a few μsec/trigger = a few % at 10kHz )
  => something like “free-run mode” is possible?
  We started to think about this possibility.
  -> requires a development of new event building software.

- Timing jitter
  * TOP requires O(20ps!!) precision
  -> apparently not possible.
  Idea: measurement of RF clock phase -> offline processing

# Lifetime of COPPER

- COPPER was designed in 2003 and manufactured for ~5 years.
- Some of its design are being obsolete -> worries to keep using.
  * need to prepare for the discontinuation of some parts.
    - FPGA chip, CPU board, etc.



We will keep using current COPPER with some revisions.
- "Refresh" its design using the latest parts.
  * new FPGA chips, GbE interface, etc.
  * new CPU cards
    Radisys EPC-6315 (PIII@800MHz)
    -> GE Fanuc PSL09(Pentium-M@1.4GHz)
- New additional COPPERs will be produced with this design.

We will keep using existing COPPERs as long as possible.
When one such COPPER is dead, it will be replaced with this newly developed one.

# Event builder (EFARM)

- "Switchless" event building farm was adopeted in 2001 in Belle DAQ.
- It is scalable (confirmed at least up to 6 units) and is planned to be used in SuperKEKB.
- In the current system, an event builder unit consists of 8 nodes to cope with the required data flow of ~20MB/unit (designed in 2001, assuming (700MHz Xeon x 4)/node).
- Recent PC servers/network are fast enough to handle whole the data flow in a event builder unit : ex. (3GHz quad core x 2)/node.

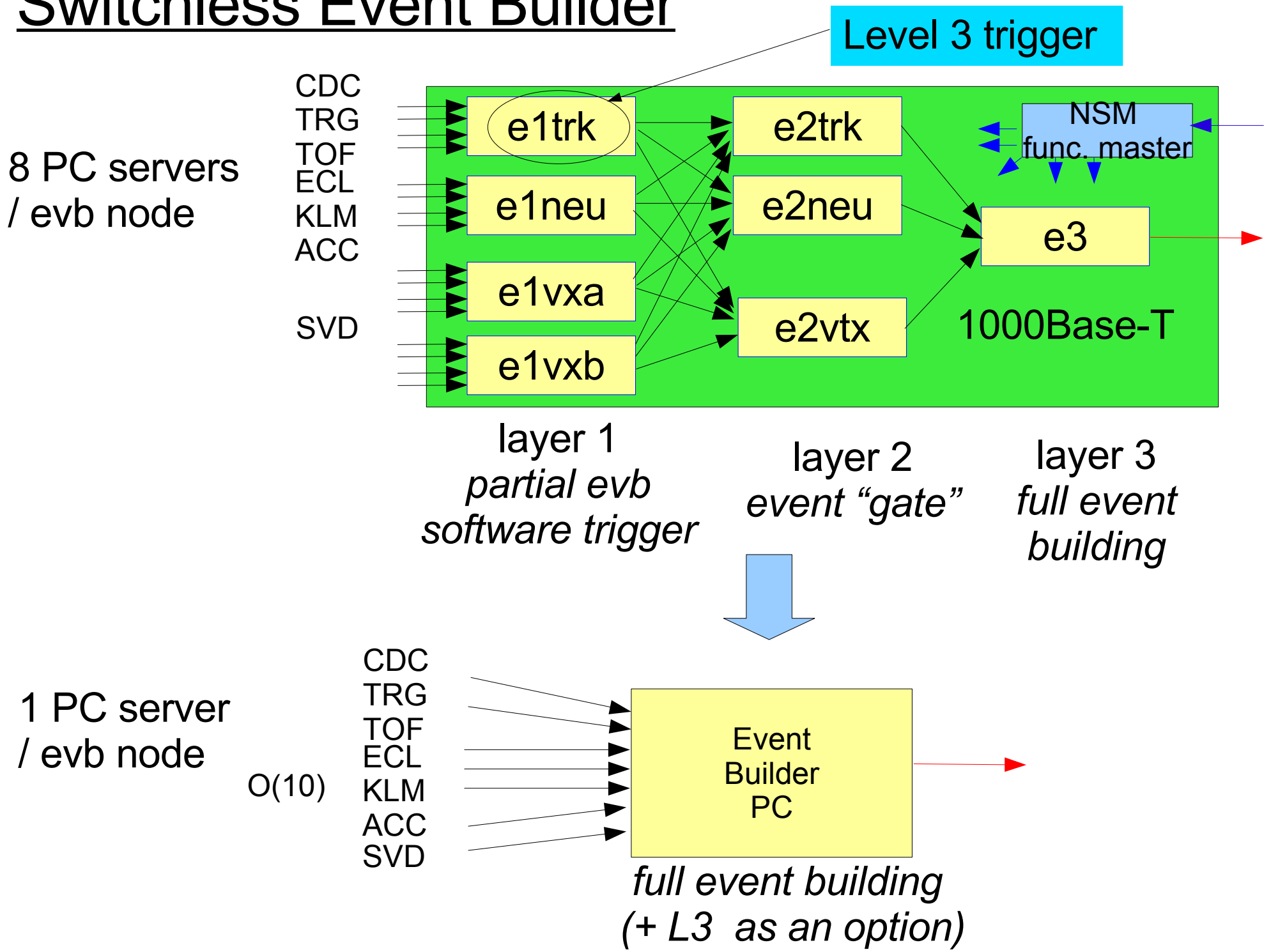
Is it possible to implement all the event building function in a single node?

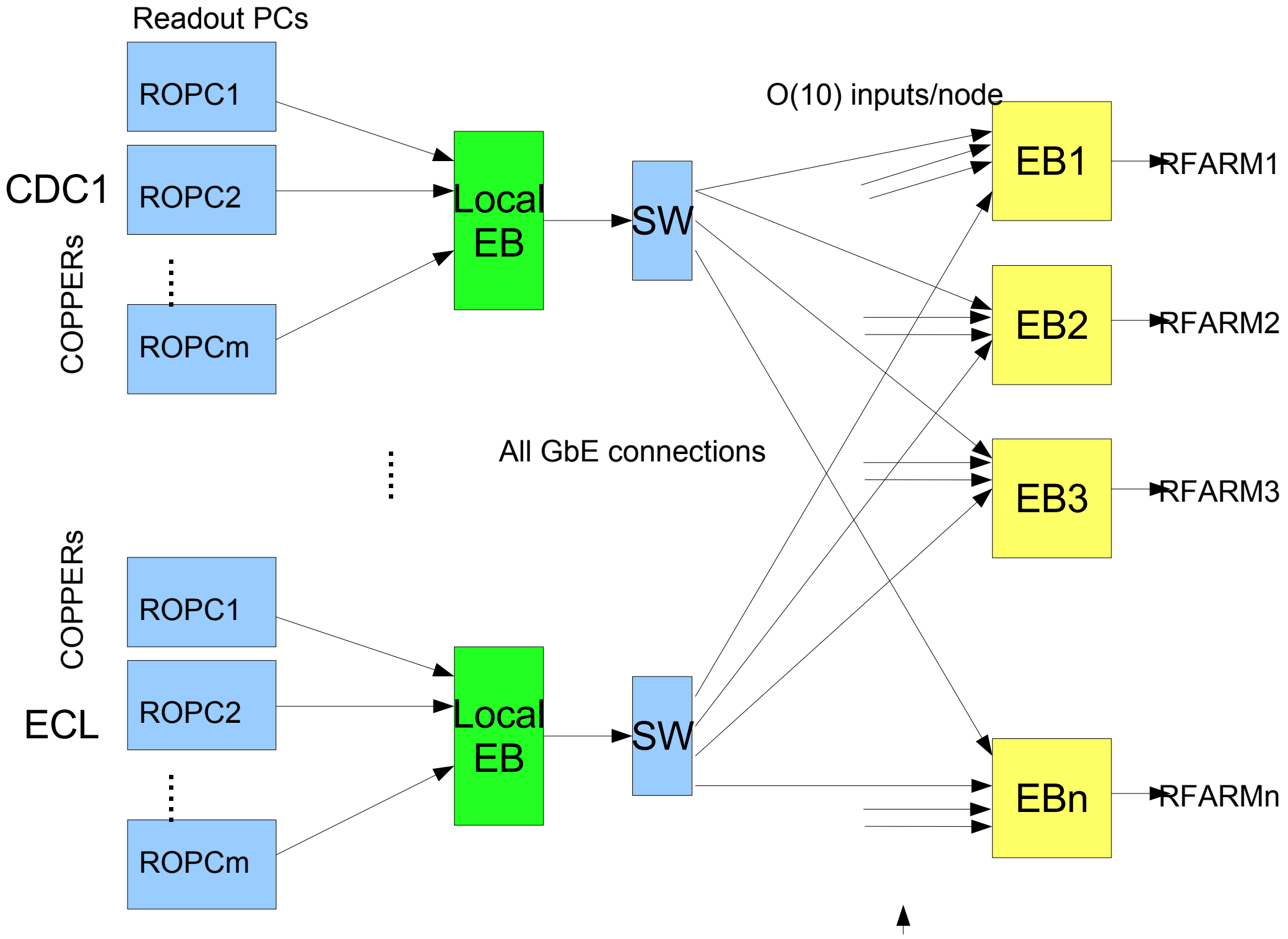
* A local event building is supposed to be done for each detector.

===> A PC server with O(10) GbE network ports as one event builder unit.

- New event building software has to be developed to cope with "free running" data flow w/o event by event synchronization.

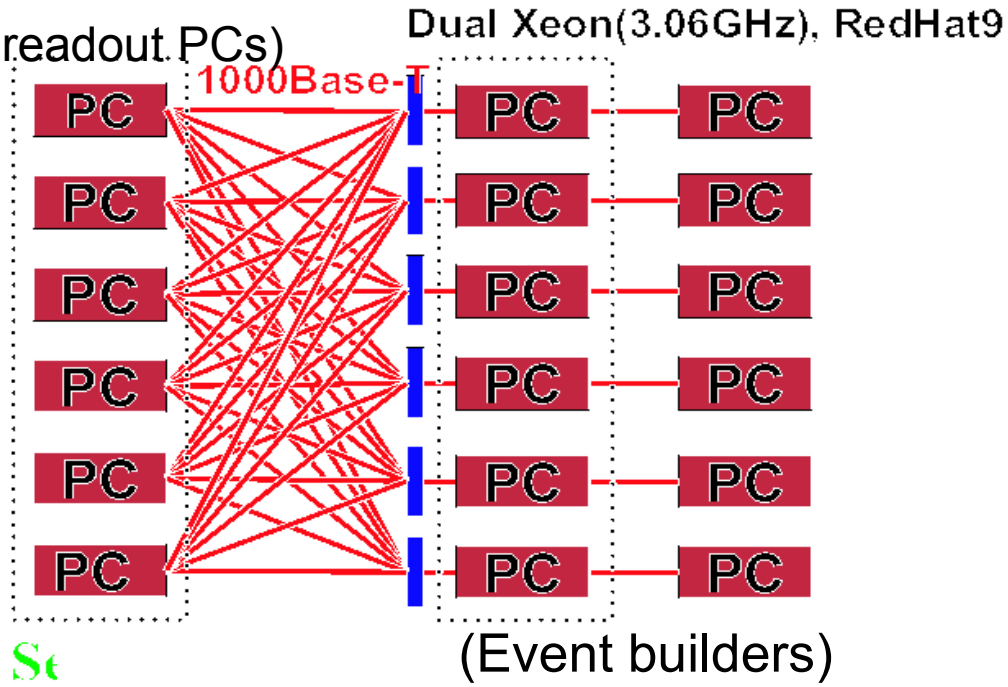
# Switchless Event Builder

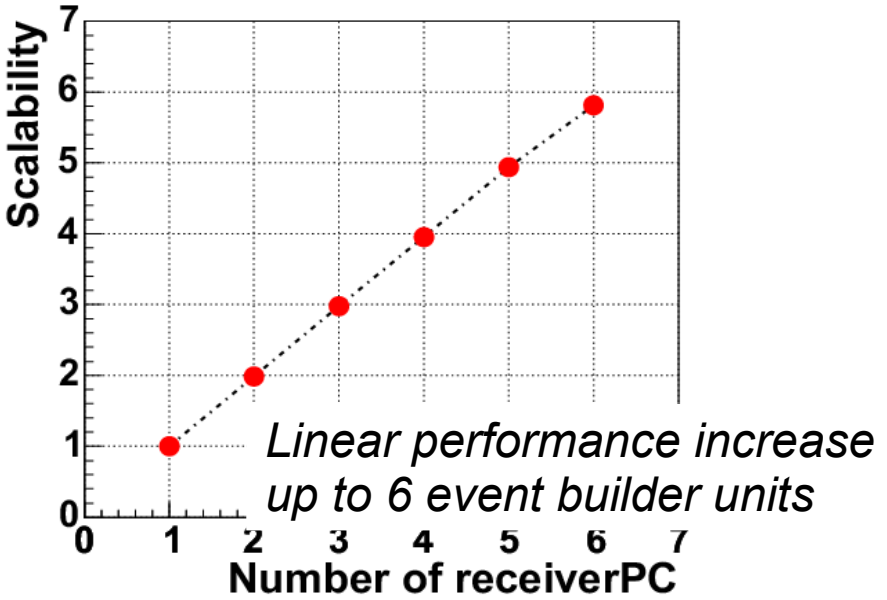


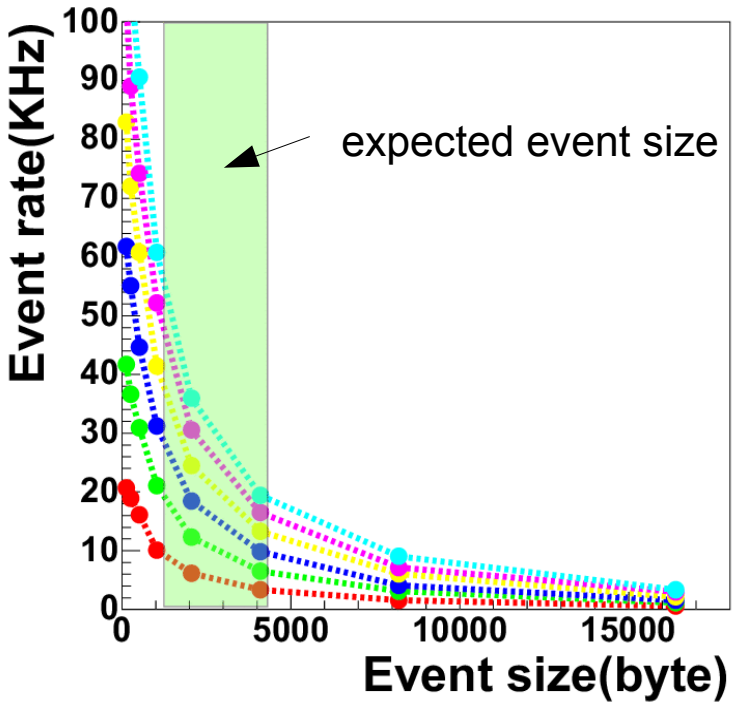

Readout PCs
ROPC1
ROPC2
ROPCm
CDC1
COPPERs
Local EB
SW
O(10) inputs/node
EB1
EB2
EB3
EBn
RFARM1
RFARM2
RFARM3
RFARMn
All GbE connections
ROPC1
ROPC2
ROPCm
ECL
COPPERs
Local EB
SW

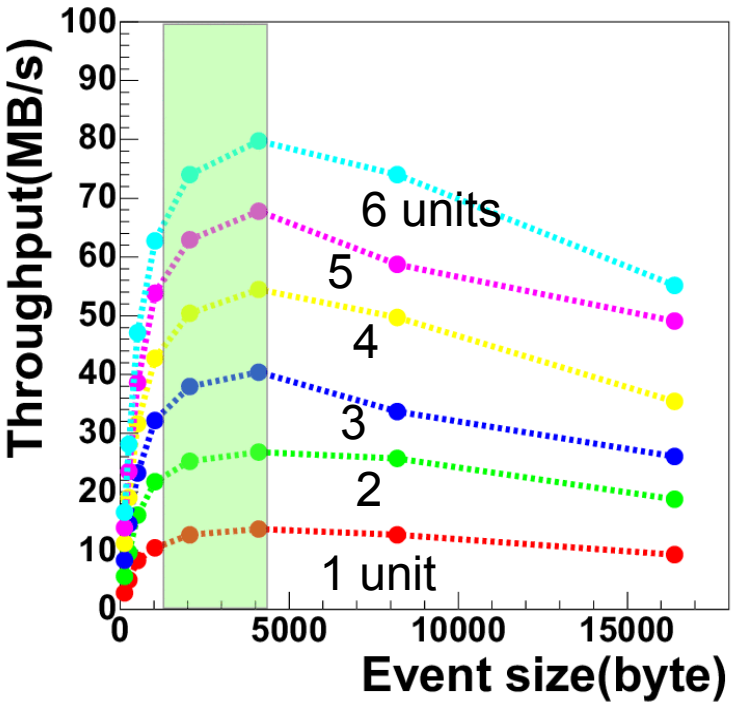
# Scalability test of modular event builder+HLT

*Test bench*



# Transfer Network Matrix + Event Builder 1 and 2

# RFARM (High Level Trigger:HLT)

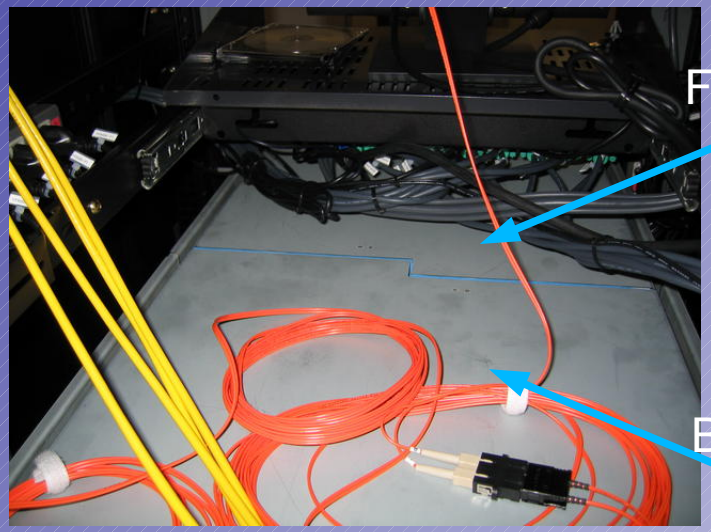
- HLT farms (called RFARMs) are placed after event builder units. One unit consists of ~100 CPUs (3GHz Xeon) and can process data flow corresponding to L of ~$1.5 \times 10^{34}$.
  -> O(10) units of RFARMs are enough to process L of $2 \times 10^{35}$.

- The full event reconstruction using the same offline code is performed in real time. It enables the use of physics event skim (like hadronic event selection) as the HLT software which can be a powerful data reduction tool.

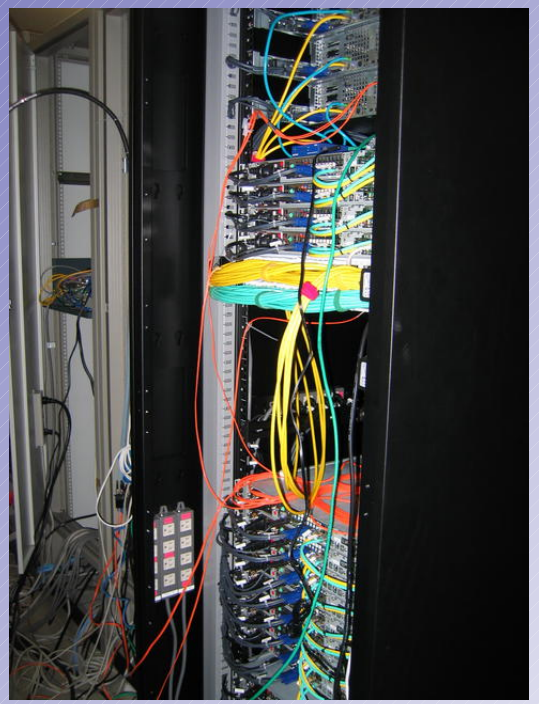
- Multiple output data streams have to be managed. Belle is already managing 2 streams by a simple merging, however, O(10) merging might not bet straightforward.
  -> better to have a kind of “Event Database” as a backend storage (i.e. CASTOR @ LHC)

# Belle's RFARM


40 x Dual Xeon servers (3.4GHz)

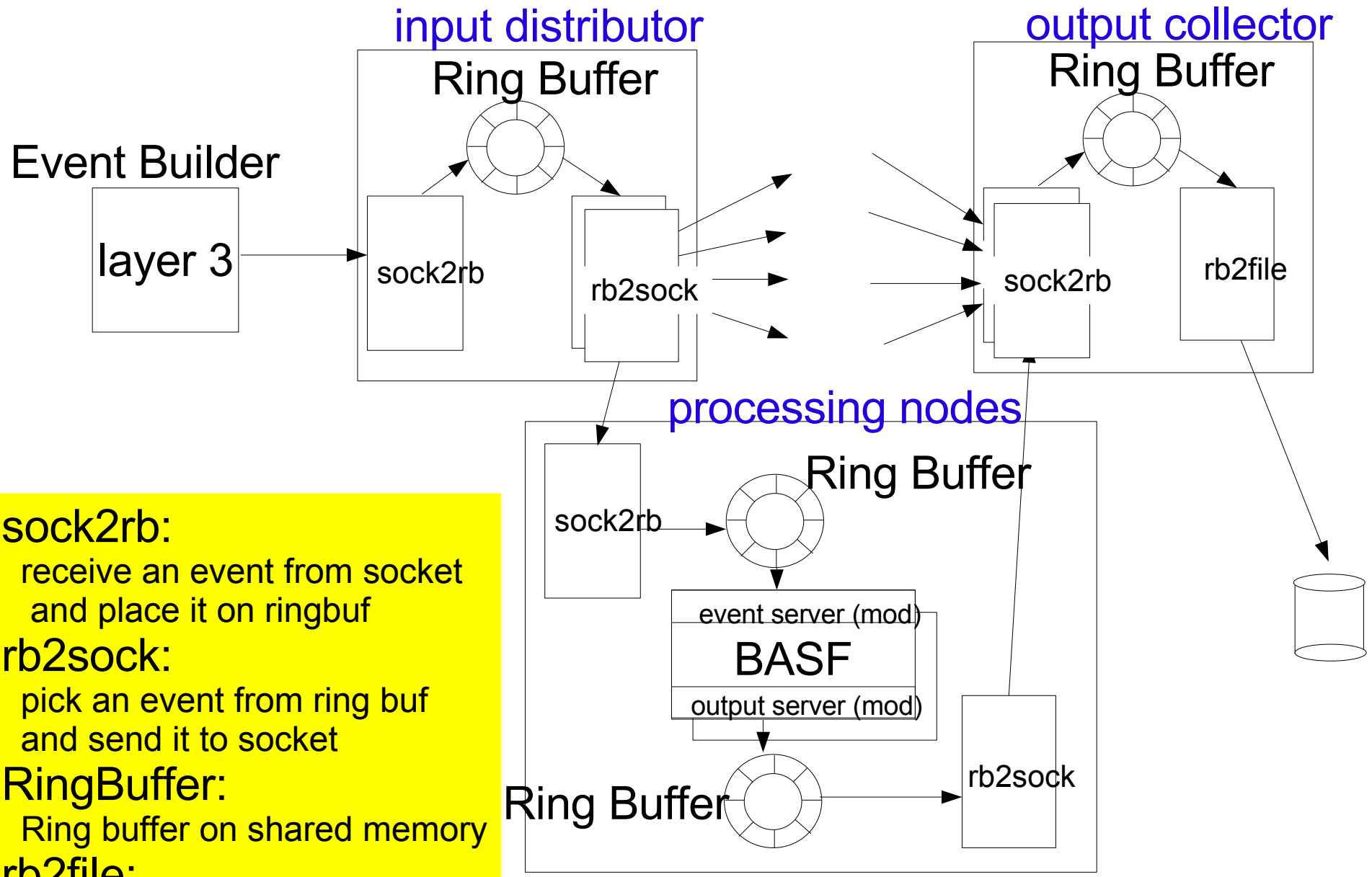



Each server houses dual Xeon in 1/2U
→ **4 Xeon CPUs in 1 U**



Backside

# RFARM Data flow




**sock2rb:**
receive an event from socket and place it on ringbuf

**rb2sock:**
pick an event from ring buf and send it to socket

**RingBuffer:**
Ring buffer on shared memory

**rb2file:**
save data in a file

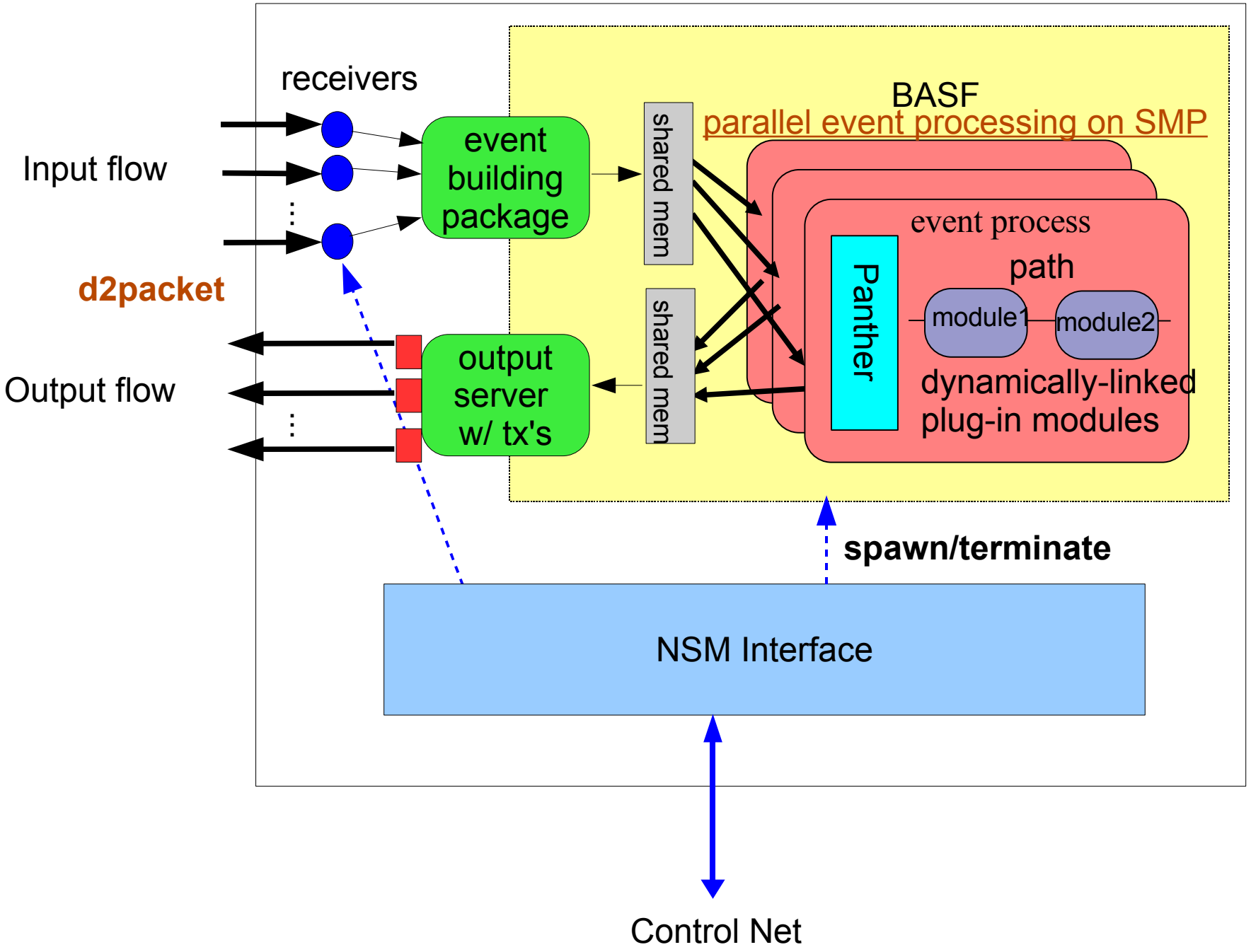
**~100 nodes**

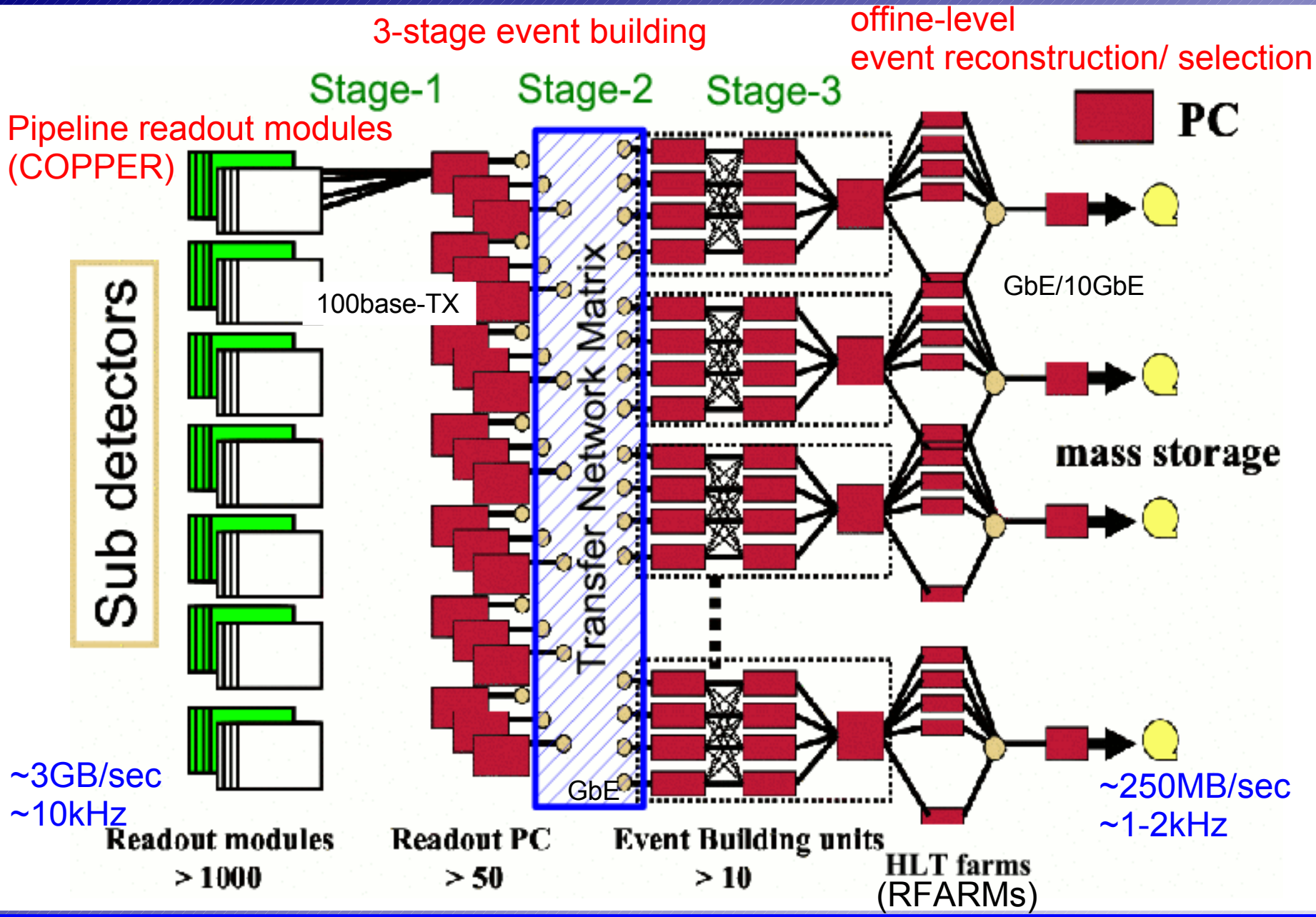
# DAQ Software

Concepts:
1) Obtain a huge and versatile processing power for the data reduction by the **widely distributed and parallel computing** technology.
2) Easy and common programming environment compatible with offline software.

For the realization:
* Linux operated CPUs are implemeted on all DAQ nodes from COPPERs ,Event bulder nodes and RFARMs.
  COPPERs : O(1000)
  Event builders : O(10-100)
  RFARMs : O(1000)
* Unified software environment exactly the same as that of offline by utilizing the same software framework (BASF).
* Stable and well organized system(slow) control (NSM)

# Unified DAQ Software Framework



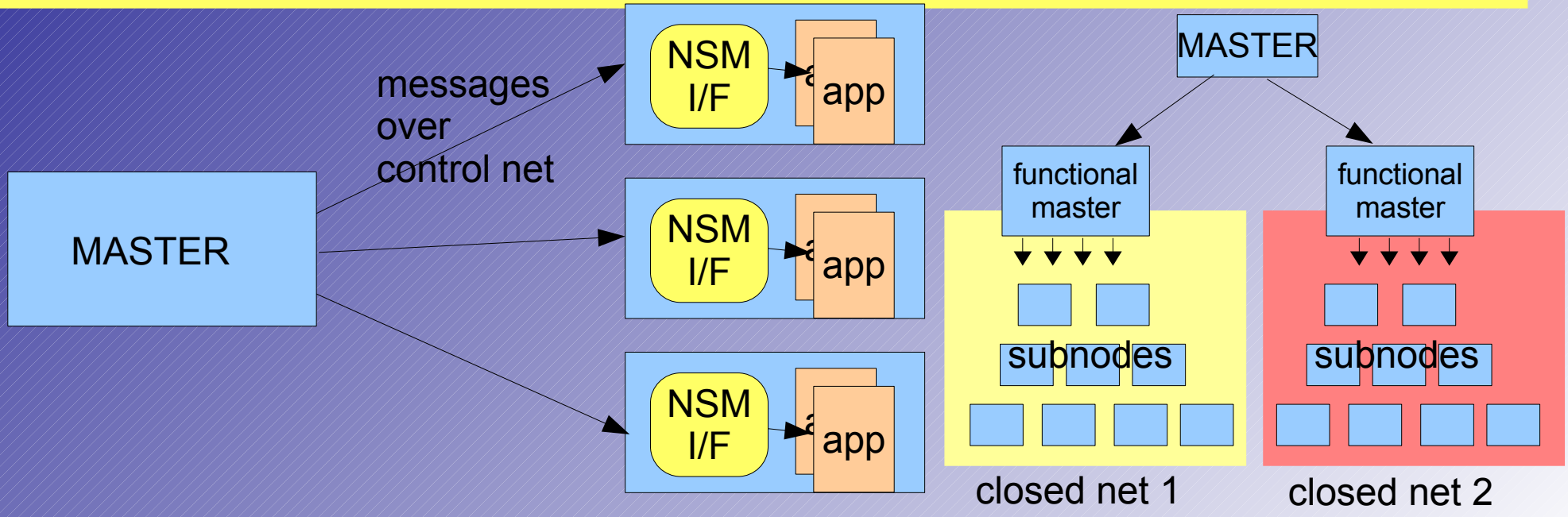



All the processing nodes are operated by the same unified software framework! But cooperative processing of a few thousand CPUs is quite tough!

# System Control

**NSM (Network Shared Memory) is used.**

* Capable of
  - shared memory handling over network (UDP broadcast based)
  - message passing between nodes (TCP based)
    ← asynchronous handling by hooked-up action functions
* DAQ control is done through message passing from one MASTER node to many client nodes.
* Support for hierarchical network structure through functional master.



# DAQ upgrade schedule

| | |
|---|---|
| FY2007 : | * TRG-COPPER upgrade<br>* Test of new CPU card (PSL09) in CDC (1-crate) |
| FY2008 : | * KLM-COPPER upgrade<br>* TOF-COPPER upgrade (not decided yet)<br>* SEQ upgrade<br>* COPPER/TTRX revision<br>* R&D on standard interface FINESSE |
| FY2009 : | * Purchase revised COPPERs (~10-20 for test)<br>* Interface FINESSE prototype<br>* Update of E/RFARM software |
| FY2010 : | * Massive COPPER purchase I (~100)<br>* Interface FINESSE mass production<br>* Prototype of new E/RFARM |
| FY2011: | * Massive COPPER purchase II (~100)<br>* Purchase PC servers for E/RFARMs (~5 sets)<br>* Start integration with detector front-end |

**early FY2012**
**Cosmic ray run**

# 4. Summary

* Present design of the Belle Trigger/DAQ system is already aiming at SuperKEKB upgrade.
  - The same L1 trigger strategy is expected to work at SuperKEKB.
  - COPPER based pipelined readout system which is tolerable up to 30kHz.
  - Modular event builder + HLT which can manage a gradual luminosity increase up to $10^{36}$ $cm^{-2}sec^{-1}$ just by adding more units.
  - Unified DAQ software which allows the widely distributed processing for the data reduction.

* Further efforts will be concentrated on
  1) Innovative trigger decision logic,
  2) the development of digitizers and their interface to the readout module (COPPER), and
  3) to pursue the possibility of "free-running" DAQ. (timing distribution/handshake, event building.....)

We need your help!

# Backup Slides

# Trigger Issues: Summary

1. Do we need a precise timing for detector readout trigger?
   i.e. Can we live with CsI trigger timing (~30ns precision) only?
   * SVD APV25 readout requires the timing precision of ~10ns.
   -> TOP is the only device which can provide such a precision timing signal, but further studies are necessary.
   * For now, we don't think to have additional timing devices like TSC.

3. Trigger latency of current design is 5 μsec.
   -> However, SVD's APV25 readout :
   ~3.8usec pipeline depth @ 40MHz readout.
   => They request to shorten the latency.
   * The effect of shortening in readout/DAQ is now being checked.

Possible “typical” front-end electronics placed near detector:

- preamp + shaping + digitization
- gather data from O(100) channels/module
- send the data to (local) event builder directly thru. optical fibre.

What can be the problem in this approach?

- How to distribute trigger/readout timing.
- Pipeline management
- Lack of intelligence for data formatting/reduction

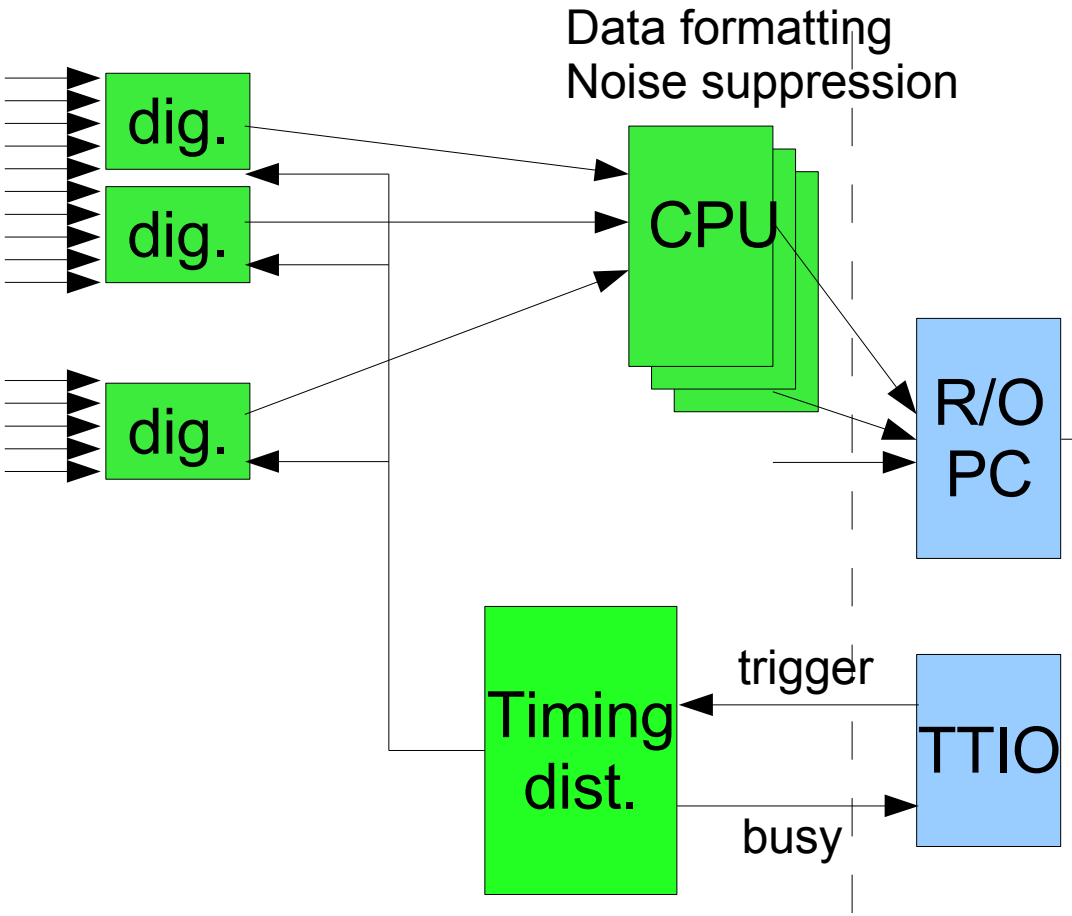
2 approaches

a) Direct connection to readout PC from digitizer modules w/o mediating COPPER.

b) Use COPPER as “readout module”

a) Direct connection to Readout PC

- The boundary between DAQ and detector electronics is just before readout PC.




- CPU is required for every ~50 channels.
  =>ex.
  ~400 CPUs for CDC
  where to place?
- Run control?

: DIY by each detector group

* Need to develop many items including both hardware and software
* May have some advantage in the total cost (w/o considering the man power and work load)

# Use of “SiTCP” for the interface between COPPERs and digitizers

- Current design requires event-by-event trigger handshake.

- SiTCP data flow: the event data are buffered in SiTCP and sent asynchronously.
  -> cannot control the data flow by the handshake.

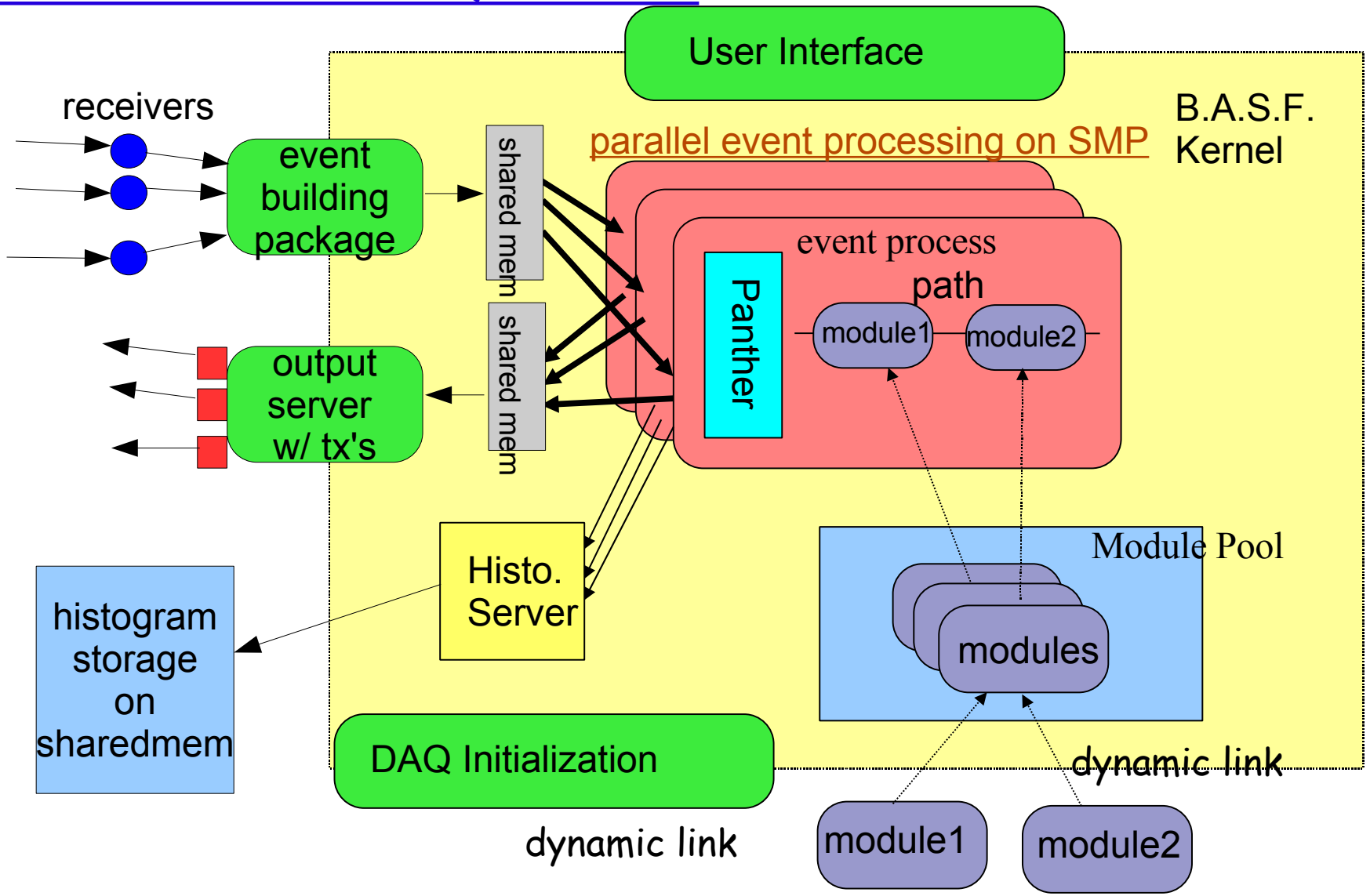
Any Solution?

1. Use high speed serial bus (or parallel link) over optical fibre like S-link instead of SiTCP.
   -> Already used by many other experiments – BaBar, LHC...
   -> SVD : parallel link

2. Event-by-event buffer flushing on SiTCP
   -> performance degrades drastically and not realistic.

3. Abandon event-by-event handshake.

# B.A.S.F. with DAQ mods.




* Data processing codes on every DAQ node can be written as "modules" in the same manner as that for offline.

# Q: 1/10 reduction at HLT is realistic?

* Only 25% of recorded data are used for the physics analysis after Level 3 event selection.
  - Reduction factor of Level 3 selection is ~2.

| | |
|---|---|
| Hadronic Event Selection | : 14.2% |
| $\tau\tau$/2photon | : 9.6% |
| Monitor events (ee,$\mu\mu$...) | : ~1% |
| Total | : ~25% |

O(10) reduction is possible
without loosing events with our interest
*if we implement physics event selections*
*as Level 4 selection in HLT.*

HLT(RFARM): offline-level full event reconstruction is supposed to be performed (as done on Belle RFARM)