# Organization of DAQ Talks

1. Overall design, status and plan        - Itoh (20min)
2. Front-end readout                            - Igarashi (20min)
3. Event Builder                                  -  Higuchi(20min)
4. Timing distribution and data link      - Nakao (20min)
5. Discussion on the L1 latency            -  Iwasaki + Nakao(10)

# DAQ: Design, Status and Plan

R.Itoh, KEK

SuperKEKB Open Collaboration Meeting
12/12/08

# Outline

1. Requirements
2. Global Design
3. Brief introduction to each component and status
    - Front-end readout and unification effort
    - Timing distribution
    - Readout module and data link
    - Event builder
    - High level trigger and overall data flow
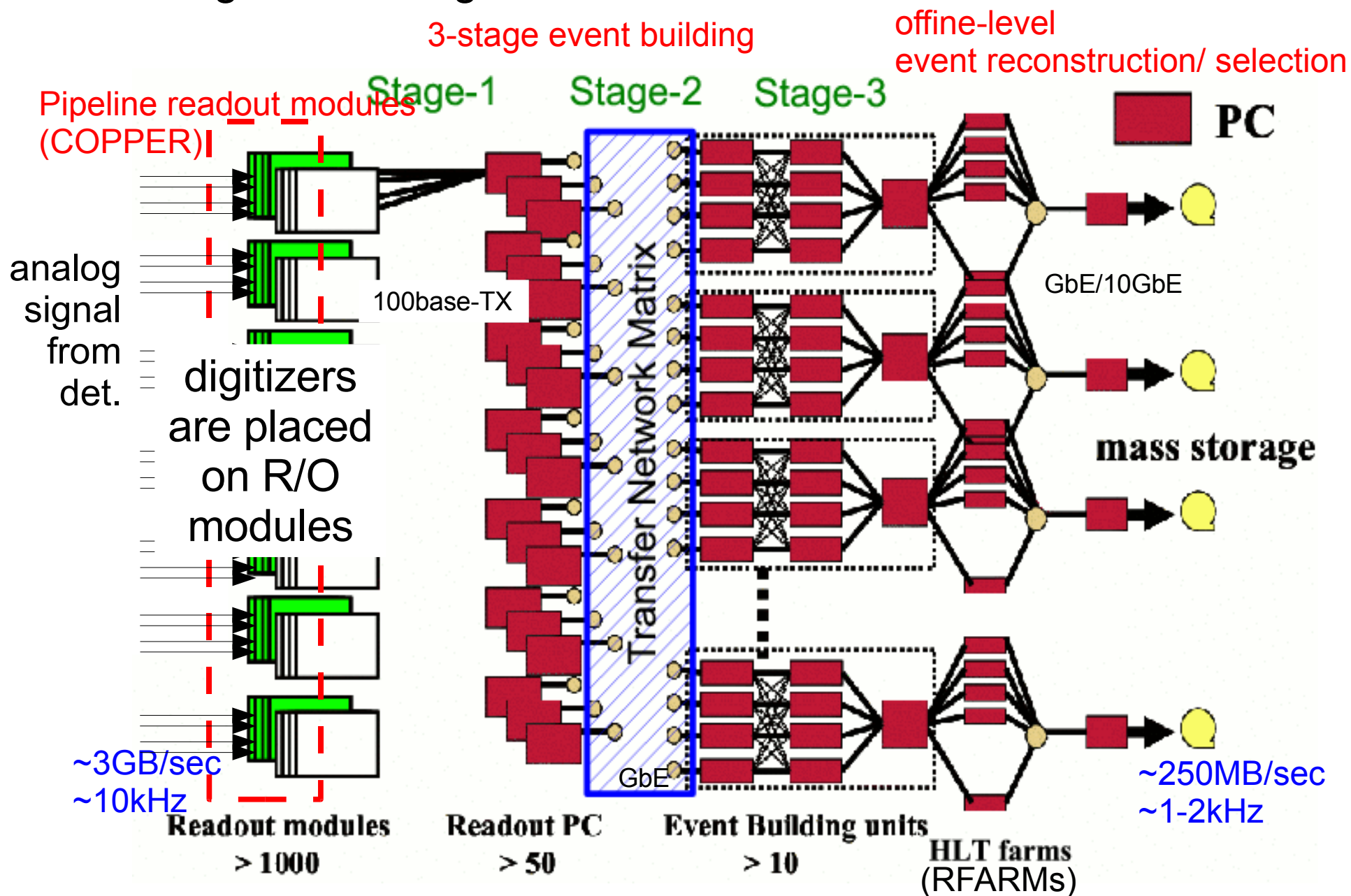    - Slow control
4. Plan
5. Summary

# 1. Requirements

- Expected luminosity @ SuperKEKB ~ $2.0 \times 10^{35}$/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 | 100-300kB/ev |
| flow rate(out) | 20MB/sec | 1-3GB/sec |
| L3+HLT reduction | 1/2 | ~1/10 |
| Storage badwidth | 20MB/sec | 300MB/sec |

- Event size estimation is still quite rough, but should be less in the reality.

# 2. Global Design

"Old design" assuming "smooth" transition from Belle DAQ

3-stage event building

offine-level
event reconstruction/ selection

Pipeline readout modules
(COPPER)

Stage-1    Stage-2    Stage-3

analog
signal
from
det.

digitizers
are placed
on R/O
modules

100base-TX

Transfer Network Matrix

PC

GbE/10GbE

mass storage

~3GB/sec
~10kHz

GbE

~250MB/sec
~1-2kHz

**Readout modules**
**> 1000**

**Readout PC**
**> 50**

**Event Building units**
**> 10**

**HLT farms**
**(RFARMs)**

# Reconsideration of design

1. Most of detector groups are planning to place digitizers near detector.

   *The drastic increase in the number of readout channels (ex. CDC: ~10000(Belle)$\rightarrow$~18000(SuperKEKB)) requires intermediate data merging before transmitted to DAQ placed in E-hut to reduce number of cables.*

2. The technologies used for DAQ became somewhat obsolete.

   *The DAQ was designed using early 2000's technology.*
   * *Network: 100base-T        -> GbE/10GbE*
   * *CPU: Single-core@<1GHz -> Multi-core@>3GHz*
   * *Software: "C" based       -> OO with C++ and Python*

3. Recycle Belle's experience as much as possible
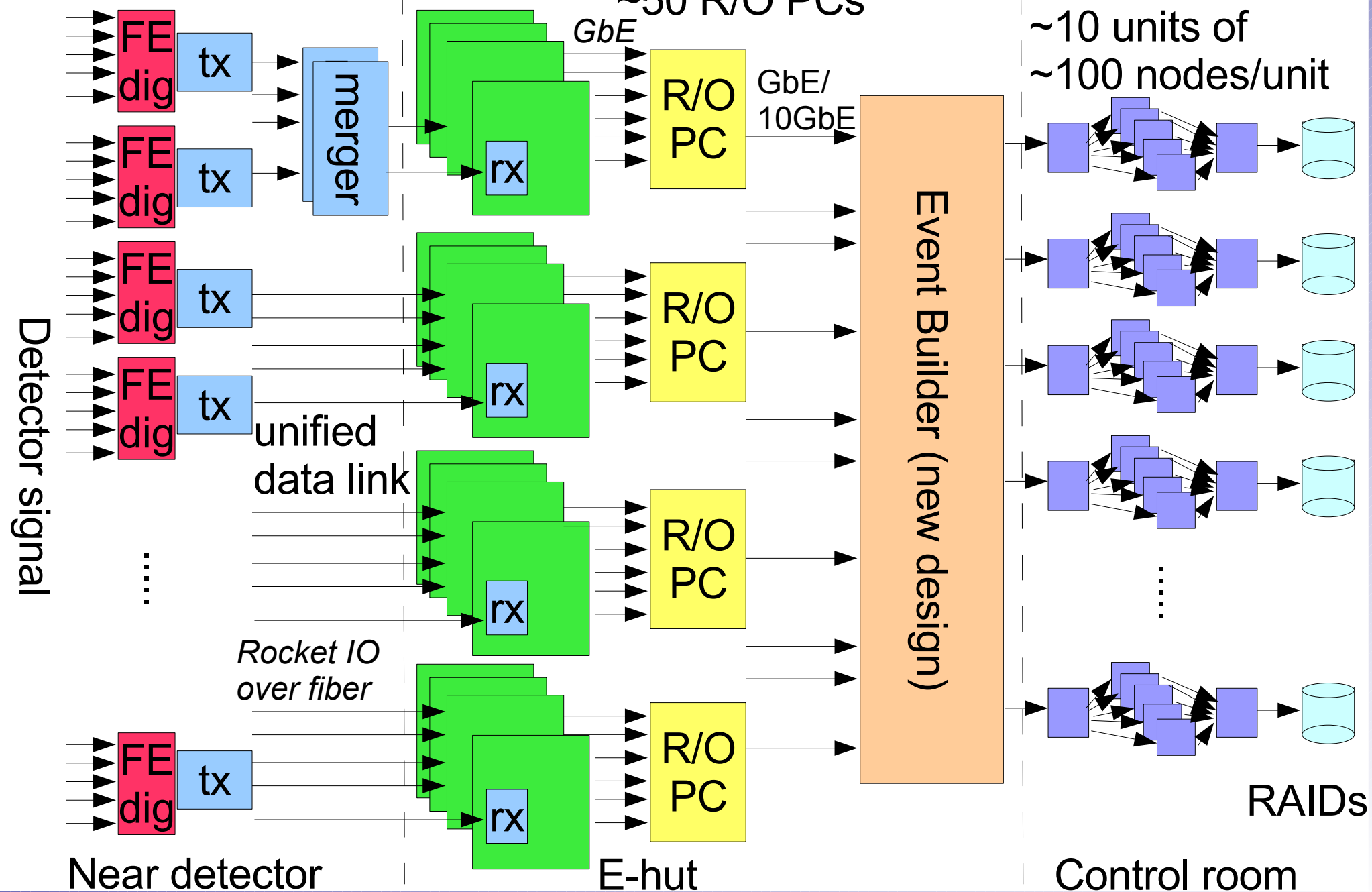   Pipeline readout modules, data flow scheme, slow control......
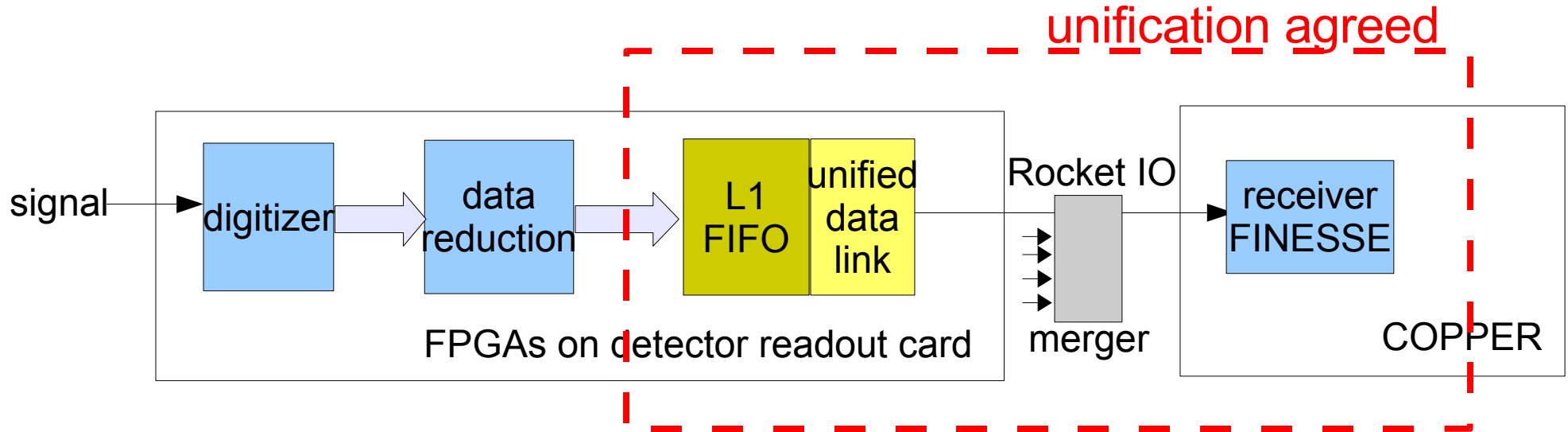
# New Design

~0.1M chns

~500 COPPERs

~50 R/O PCs

* Timing dist. scheme is not included in this figure.

HLT farms
~10 units of
~100 nodes/unit

Detector signal

FE dig

tx

merger

rx

*GbE*

R/O PC

GbE/ 10GbE

Event Builder (new design)

FE dig

tx

FE dig

tx

FE dig

tx

unified data link

rx

R/O PC

*Rocket IO over fiber*

rx

R/O PC

FE dig

tx

rx

R/O PC

RAIDs

Near detector

E-hut

Control room

# 3. Component Status

## a) Front-end Readout and Data Link



unification agreed

signal → digitizer ⇒ data reduction ⇒ L1 FIFO | unified data link → Rocket IO → merger → receiver FINESSE

FPGAs on detector readout card  — merger —  COPPER

- Effort to use the unified digitizers/ASIC and data transport

* Data transport : unification basically agreed
    data transfer : link is based on Rocket IO
        looks as "remote FINESSE I/F" on readout cards
* Digitizer/ASIC : under negotiation with each detector
        subgroup by G.Varner

Data link -> Detail will be covered in Nakao-san's talk

# Gary Varner's proposal on the unification of front-end readout

| Subdetector | ASIC | ref. ASIC | Location | FPGA link |
|---|---|---|---|---|
| PXD | **TBD** | **TBD** | hybrid/dock | yes |
| SVD3 | APV25 | | E-hut | no |
| new SVD | BSR/KUPID | APV25 | hybrid/dock | yes |
| CDC | BCA | TARGET | in detector | yes |
| PID SiPMT | BCA | TARGET | in detector | yes |
| PID HP-PMT | HPBA | BLAB2 | in detector | yes |
| ECL | N/A | | on detector | yes |
| Scint. KLM | BCA | TARGET | in detector | yes |
| VFV | BCA | TARGET | in detector | yes |

* Key: Wave form sampling everywhere
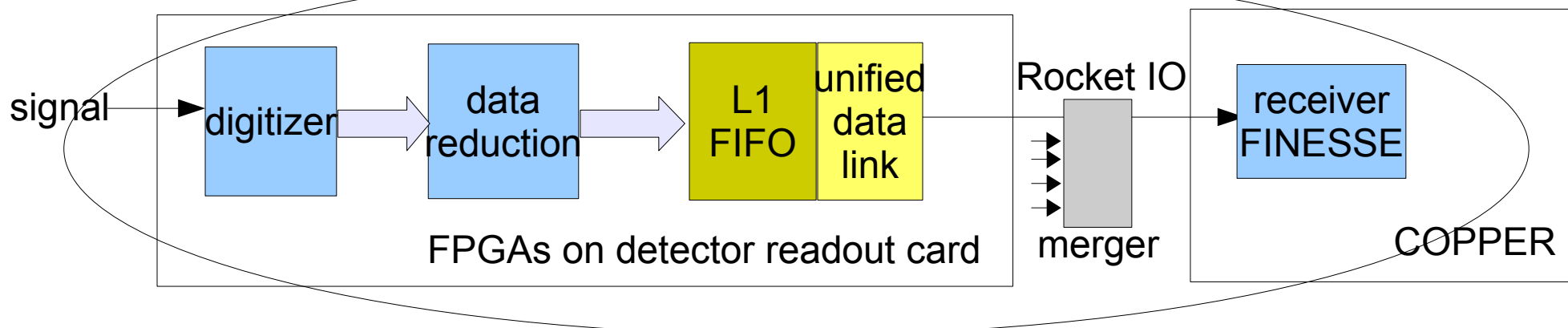* TARGET chip : analog pipelined ADC chip for wave form sampling (BLAB2 is similar)

* Detector groups have their own idea.
    PXL -> considering the application of readout board designed for PANDA (S.Lange)
    CDC -> new design based on FPGA-TDC + slow ADC
        ===> Detail will be covered in Igarashi-san's talk

# Joint development proposal with IHEP(China) lead by Z.Liu



**signal** → digitizer → data reduction → L1 FIFO / unified data link → Rocket IO → merger → receiver FINESSE

FPGAs on detector readout card

COPPER

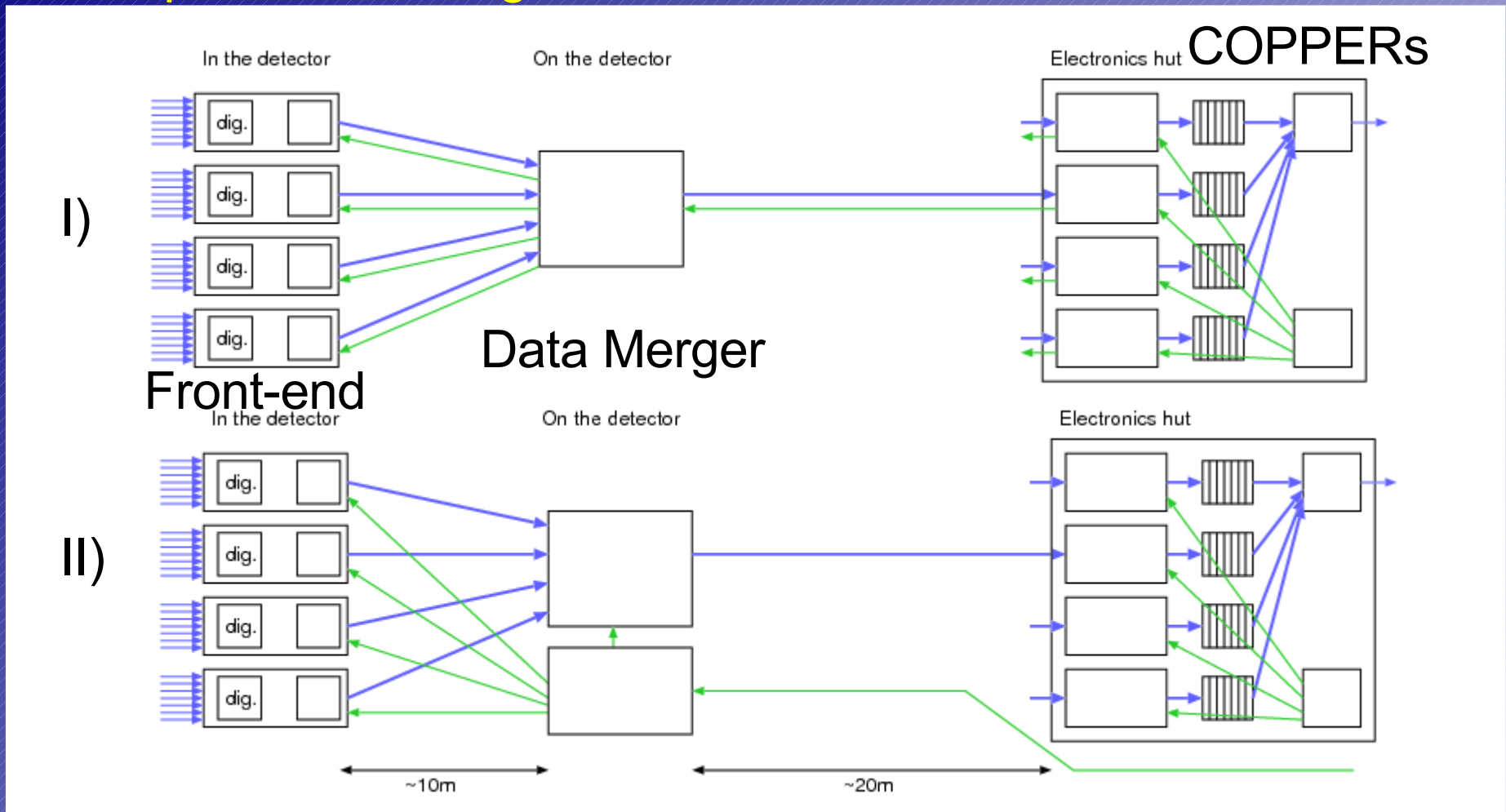* IHEP (+ Gary) is responsible for the entire scheme.
   The responsibility includes:
   - Unification of readout ASIC and digitizer which includes
     the <u>negotiation with each detector group</u>.
   - Total responsibility on data link (even if unification fails.....):
     + L1 FIFO management
     + Design of data transfer protocol
     + FPGA implementation of TX and RX
     + Data merger
     + Receiver FINESSE development
=> Final design of readout and data link with prototype
   by the end of JFY2009.

# b) Timing distribution
## Two options are being considered.



I)

**COPPERs**

In the detector | On the detector | Electronics hut

**Front-end**

**Data Merger**

II)

In the detector | On the detector | Electronics hut

~10m          ~20m

* Signals: clock, trigger, etc.
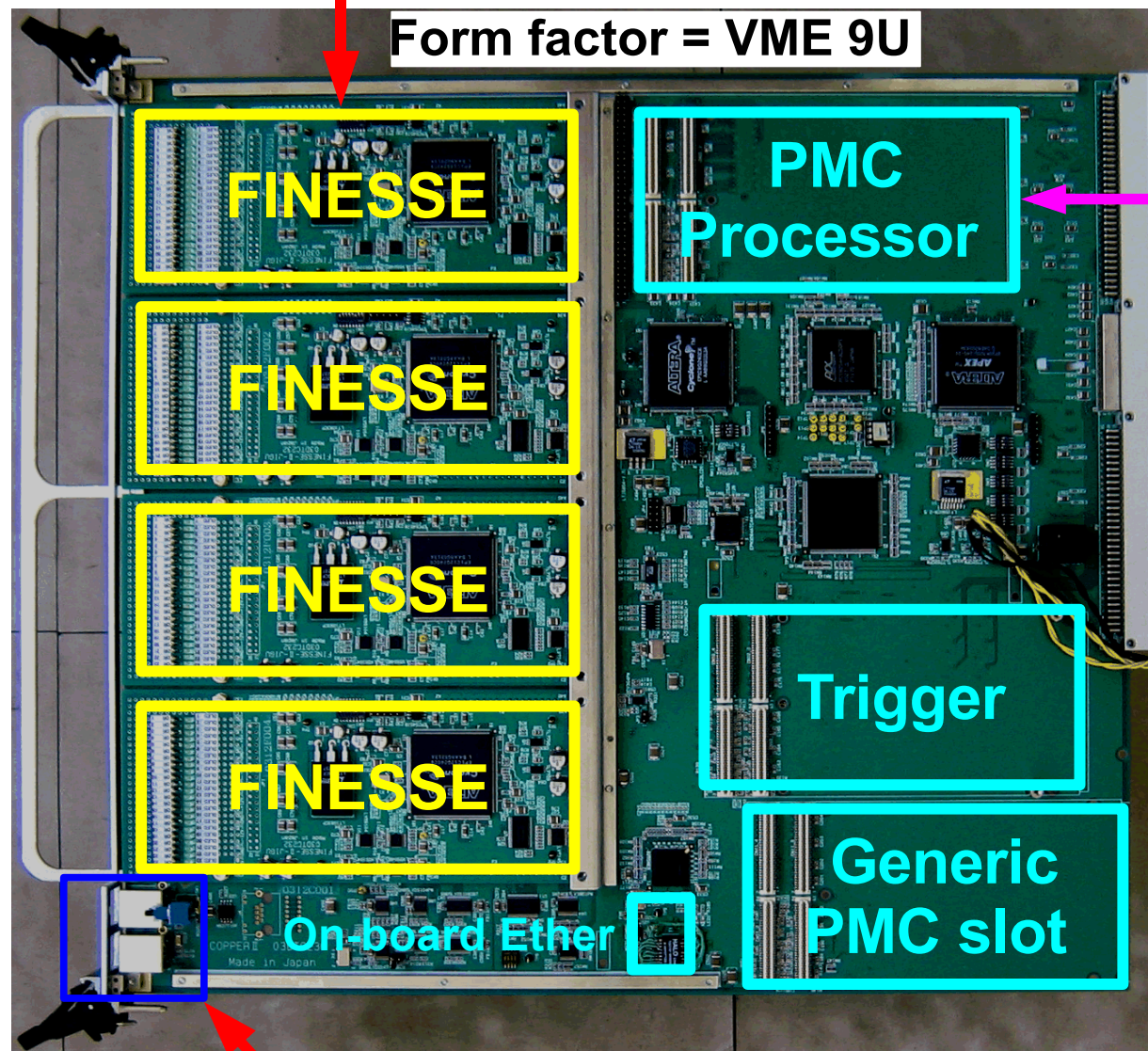* LVDS on STP
* No SERDES for the signal transmission (latency problem).
* Trigger handshake : synchronous or asynchronous

=> Detail will be covered in Nakao-san's talk

## c) Readout Module (COPPER)



Digitizer cards (implemented as daughter cards)

Form factor = VME 9U

FINESSE

FINESSE

FINESSE

FINESSE

PMC Processor

Trigger

Generic PMC slot

On-board Ether

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

# COPPER revision

- 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.
    => Prototype was already made and now is being tested.
  -new CPU cards
          Radisys EPC-6315 (PIII@800MHz)
            -> GE Fanuc PSL09(Pentium-M@1.4GHz)
     => Talking to some companies about the custom production of
          CPU cards with Atom@1.6GHz chip.

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.

## d) Event Builder

Previous design was based on Belle's switchless event building.
  - Multi-step event building to provide enough CPU power for both event building and L3 trigger running on it by a cluster of "slow" CPUs + 100base-T network. (designed in 2001).

* CPU power and network speed drastically increased.
* No L3 software on event builder any more.
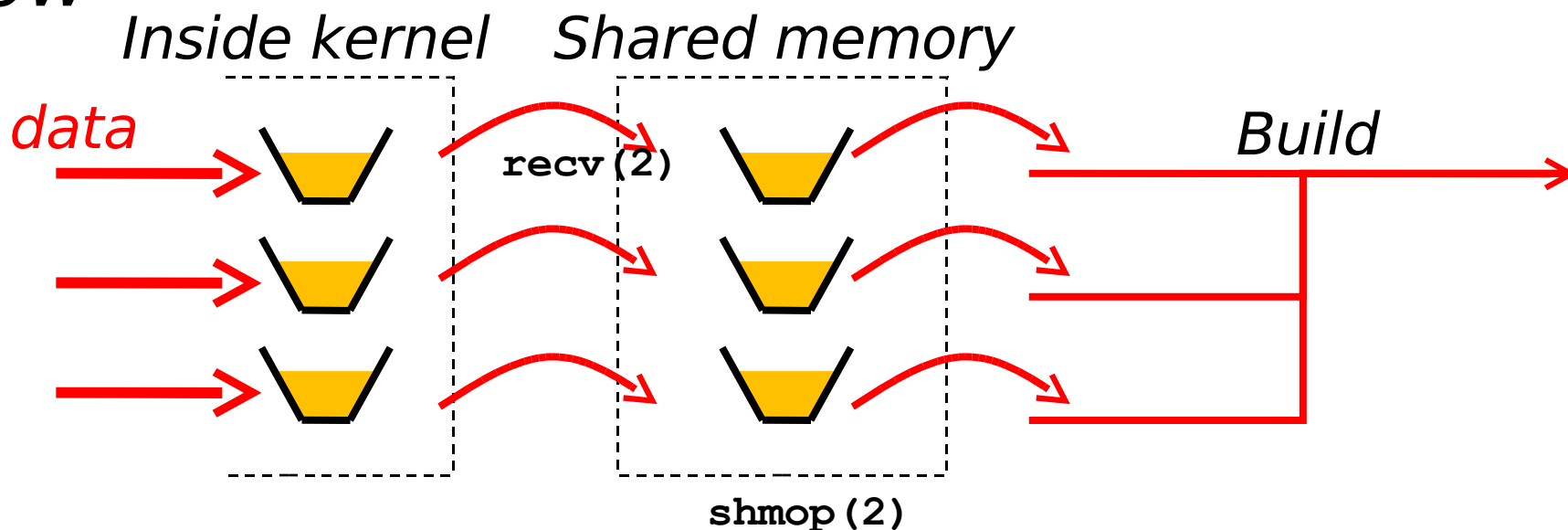      => 2 level trigger scheme with hardware L1 + HLT software

Full event building can be performed by
a high speed switch + a set of  event building PCs
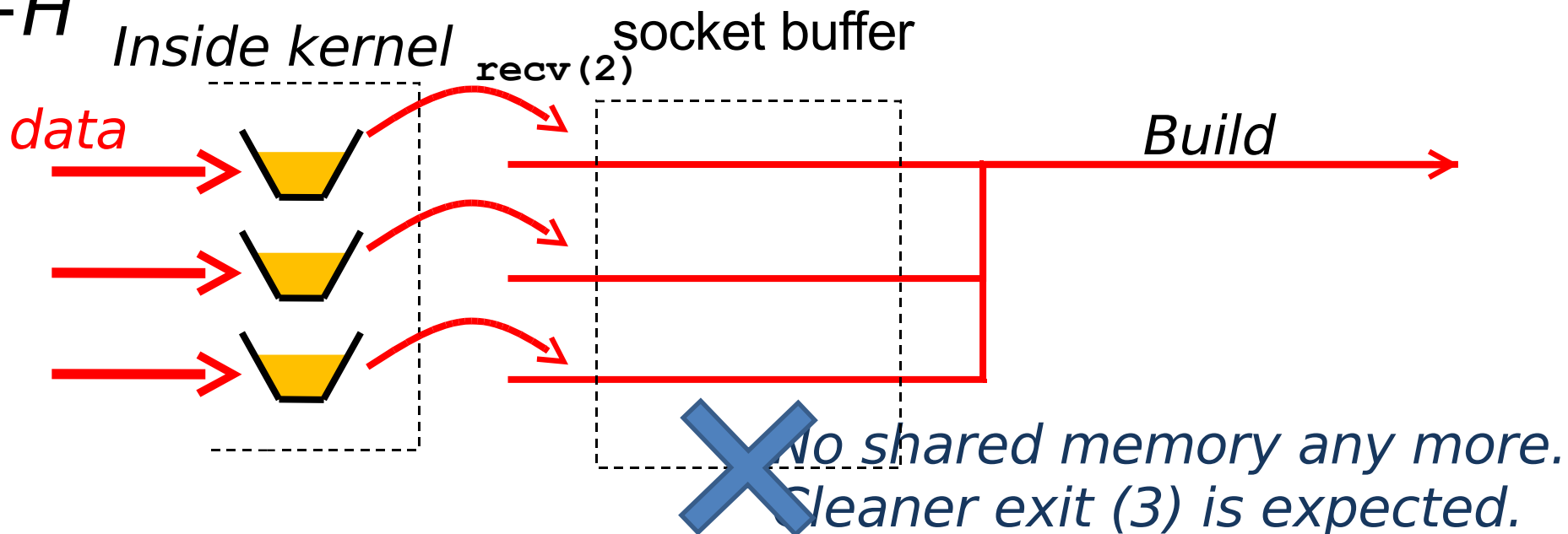connected to HLT units.
(<= standard in HEP experiment to date)

=> Detail will be covered in Higuchi-san's talk

# Idea of event building by Higuchi and Yamagata

- *Now*

*Inside kernel*   *Shared memory*

*data*

`recv(2)`

*Build*

`shmop(2)`

- *Y+H*

*Inside kernel*   socket buffer

*data*

`recv(2)`

*Build*

*No shared memory any more.*
*Cleaner exit (3) is expected.*

## e) HLT and data flow software

- HLT is supposed to perform full event reconstruction and
  and physics-level event selection as software trigger.
    -> Hadronic event + Low mult. event + scaled Bhabha/mm
        = Belle's "skim" level selection
    -> Full compatibility with offline reconstruction software
        is required

- Required CPU power (estimated based on Belle's RFARM)
    -> 100 chips of Xeon@3GHz for 1.0E34 luminosity.
        Data flow / unit ~ 20MB/sec

- Event-by-event parallel processing mechanism on a PC cluster
  is required to handle the data stream.

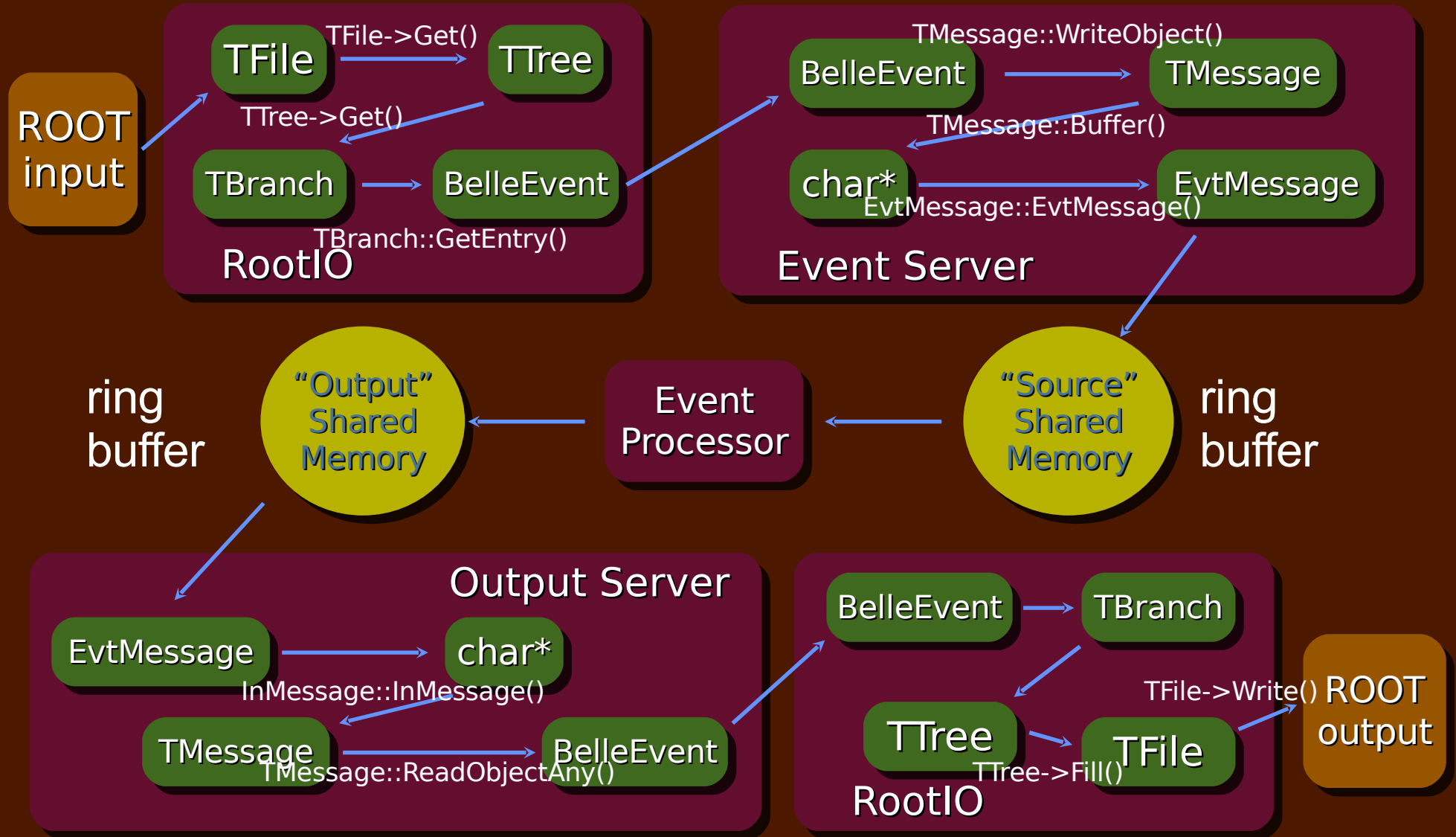Application of SuperKEKB analysis framework for HLT

# Prototype of analysis framework : "roobasf"

## *Requirements*

- Software bus (pipeline) is kept.
    * *Compatibility with modules written for B.A.S.F.*

- Object persistency
    * *ROOT I/O as the persistency core.*
    * *Panther I/O is kept as a legacy interface.*

- More versatile parallel/pipeline processing scheme
    * *Transparent implementation which utilizes both multicore CPUs and network clusters: Manage ~100 nodes by single framework*
    * *Dynamic optimization of resource allocation*
    * *Module pipelining over network*

- Integrated database and GRID interface for file management

- Dynamic customization of framework
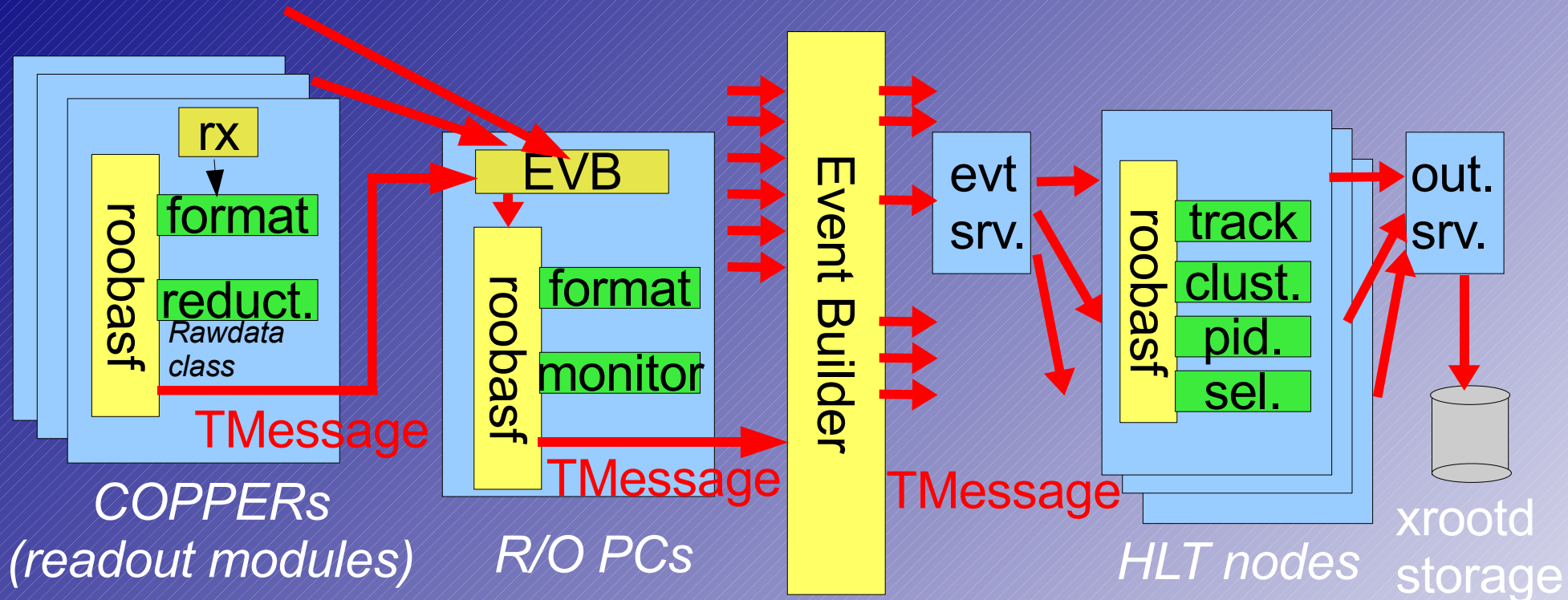    * *replaceable I/O package, user interface .......*

# Parallel processing in roobasf

S.H.Lee

**RootIO**
- ROOT input
- TFile —TFile->Get()→ TTree
- TBranch → BelleEvent
- TTree->Get()
- TBranch::GetEntry()

**Event Server**
- BelleEvent —TMessage::WriteObject()→ TMessage
- char* → EvtMessage
- TMessage::Buffer()
- EvtMessage::EvtMessage()

ring buffer — "Output" Shared Memory ← Event Processor ← "Source" Shared Memory — ring buffer

**Output Server**
- EvtMessage → char*
- TMessage → BelleEvent
- InMessage::InMessage()
- TMessage::ReadObjectAny()

**RootIO**
- BelleEvent → TBranch
- TTree → TFile
- TFile->Write() → ROOT output
- TTree->Fill()

=> Can be used as a framework for HLT

# Object-oriented data flow and processing in SuperKEKB DAQ

rx

roobasf
format
reduct.
Rawdata class

TMessage

**COPPERs (readout modules)**

EVB

roobasf
format
monitor

TMessage

**R/O PCs**

Event Builder

TMessage

evt srv.

roobasf
track
clust.
pid.
sel.

out. srv.

**HLT nodes**

xrootd storage

- The same software framework from COPPER to HLT (and offline).
- ROOT:TMessage based data flow throughout DAQ
  -> class to contain streamed object
- ROOT IO based object persistency at storage
  * use of "xrootd" for remote storage if enough bandwidth can be guaranteed.
- Real-time collection of histograms from any node for monitoring.

# f) Slow control

**Keep using NSM (Network Shared Memory) (by Nakao-san)**

* 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.
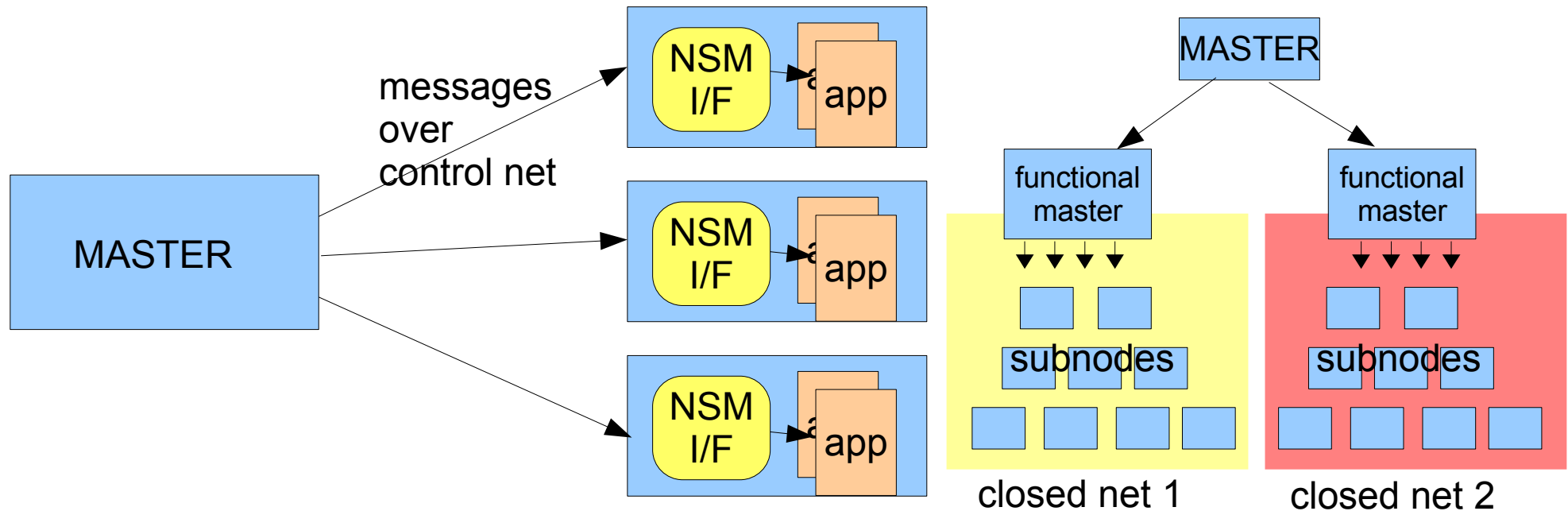
# 4. Plan

- DAQ is the infrastructure for all detectors.
  -> It has to be prepared as early as possible

FY2008:   Complete Belle-based upgrade
          Start R&D on unified data link
          Revision of COPPER = COPPER3(done)

FY2009:   Prototype of data link (tx + rx)
          Prototype of timing distribution
          Final decision on the FFE unification (by the end of this FY)
          Development of new CPU card for COPPER
          Event builder prototype
          HLT framework + data flow software prototype
          Radiation test of components placed near detector
          E-hut: Planning of new layout + detector side layout, power/cooling estimation

FY2010:   Massive purchase of COPPER3+CPU (1) (~100)
          Finalize unified data-link. Delivery of FPGA core to subdetector group
          Finalize timing distribution scheme
          Test batch production of unified data receiver FINESSE (~10-20)
          Event builder full scale prototype
          HLT + data flow : partial system test (COPPER->EVB->HLT)
          E-hut: Clean-up, installation of new infrastructure (power, network, cooling, etc.)
                 installation of new racks/crates (1)

FY2011:   Massive purchase of COPPER3+CPU (2) (~100)
          Mass production of unified receiver FINESSE (1)
          Massive purchase of Event builder/HLT PCs  (1) (~ 5 units)
          Start global system test (readout->COPPER->EVB->HLT)
          E-hut: Installation of new racks/crates (2)

FY2012:   Massive purchase of COPPER3+CPU (3) (~50)
          Mass production of unified receiver FINESSE (2)
          Massive purchase of event builder/HLT PCs  (2) (~ 5 units)
          Global system test w/ trigger + Cosmic with available subdetectors

FY2013:   Tuning in cosmic
          - Summer : commissioning(?)

# Short-term mile stones

- Unified data link :
    * prototype in JFY2009
    * delivery of  sample tx FPGA core + rx FINESSE
      to each detector subgroup :  JFY2010

- Decision on the front-end unification: by the end of JFY2009

# On L1 trigger latency

* In the previous discussion, we tentatively decided to take the L1 trigger latency of 3.5usec to have a enough pipeline length in SVD's APV25 readout @ 40MHz.

* Trigger group is now designing two step sub-trigger signal collection using Rocket IO based on the serialization of signal.

* The latency of encoding/decoding the serialized data is about a few usec.

* The use of two step Rocket IO transmission introduces relatively large latency.

* Trigger group requested to operate APV25 at a lower clock (30MHz) to enlong the L1 latency to ~5usec.

→ Discussion by Iwasaki-san and Nakao-san

# 5. Summary

- Global design of SuperKEKB DAQ has been updated so that
    1) digitizers placed outside COPPERs can be read out using
       the unified data link with the extended timing distribution scheme,
    2) up-to-date technologies can be used in data flow
       (event builder, HLT and software framework), and
    3) Belle's experience can be recycled as much as possible.

- Unification of front-end readout is being considered and under
  the negotiation with each detector subgroup. Meanwhile, new
  CDC readout scheme is proposed by CDC and KEK electronics
  groups.
            -> final decision by the end of JFY2009.

- A new collaboration with IHEP-China has been started to build
  the unified data link and to seek for the possibility of readout
  unification.

TRG/DAQ groups are now planning to have the
**Annual Trigger/DAQ workshop**
sometime in spring
(maybe in April or May)

* Decision on L1 trigger latency
* Detailed survey of detector readout system which includes:
    - Requirement to the timing distribution
    - No. of channels
    - Data size
    - Compatibility with DAQ's picture

Hope to have the workshop at nice place
                but depends on budget situation as usual.
    -> Worst case : at KEK, though