The ATCA Based Compute Node And Its Application in the Belle II PXD Data Aquisition and Reduction System

Björn Spruck[1], Thomas Geßler[1], Wolfgang Kühn[1], Jens Sören Lange[1], Haichuan Lin[2], Zhen'An Liu[2], David Münchow[1], Hao Xu[3], Jingzhou Zhao[2]

1) University Giessen, Germany

2) Institute for High Energy Physics, Beijing, China

for the Belle II and DEPFET Pixel Collaborations

- Motivation
- Hardware platform
- Implementation
- Test results
- PANDA

DEPFET Active Pixel Detector

中国科学院高能物理研究所
Institute of High Energy Physics Chinese Academy of Sciences

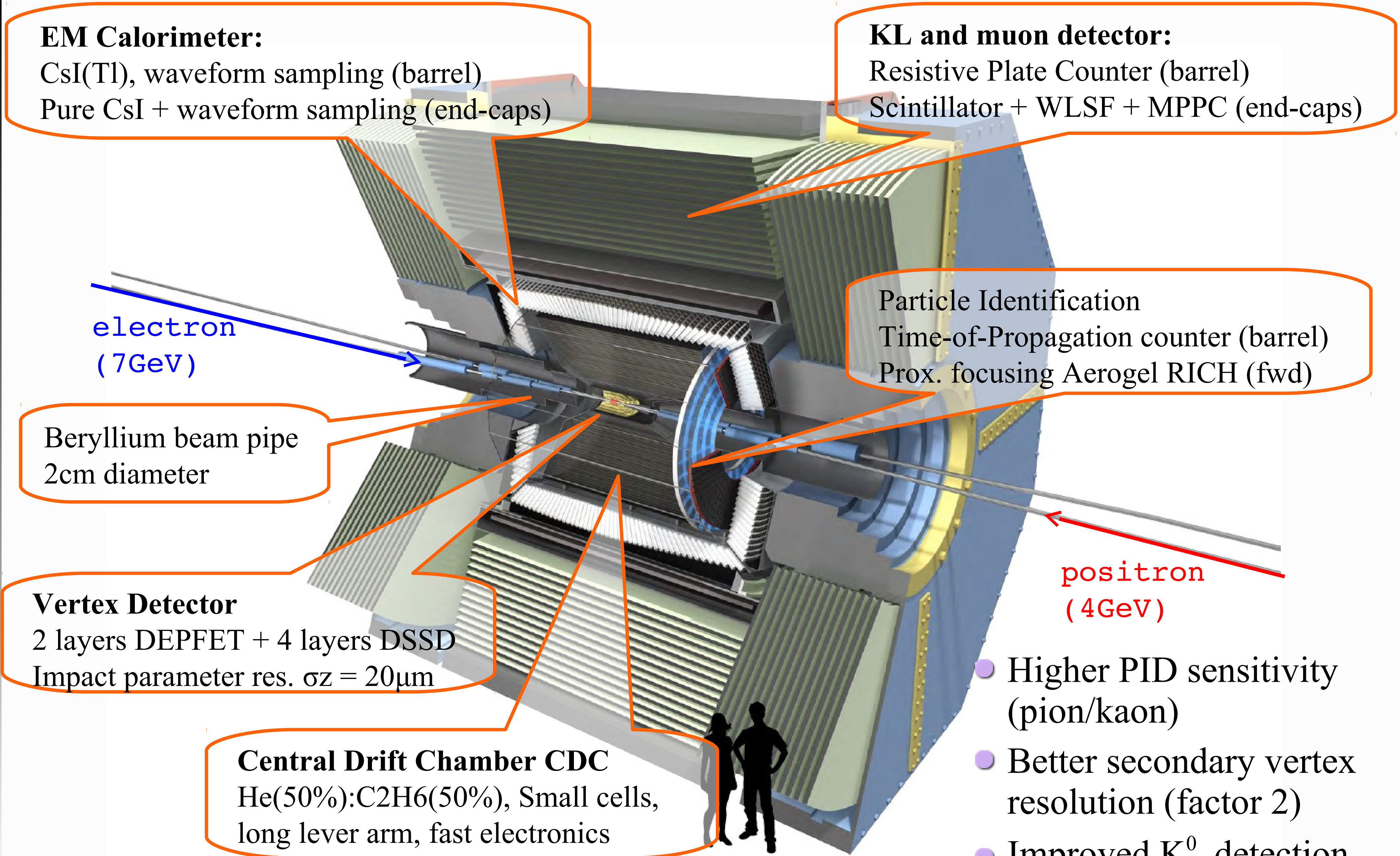
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Belle II

Motivation

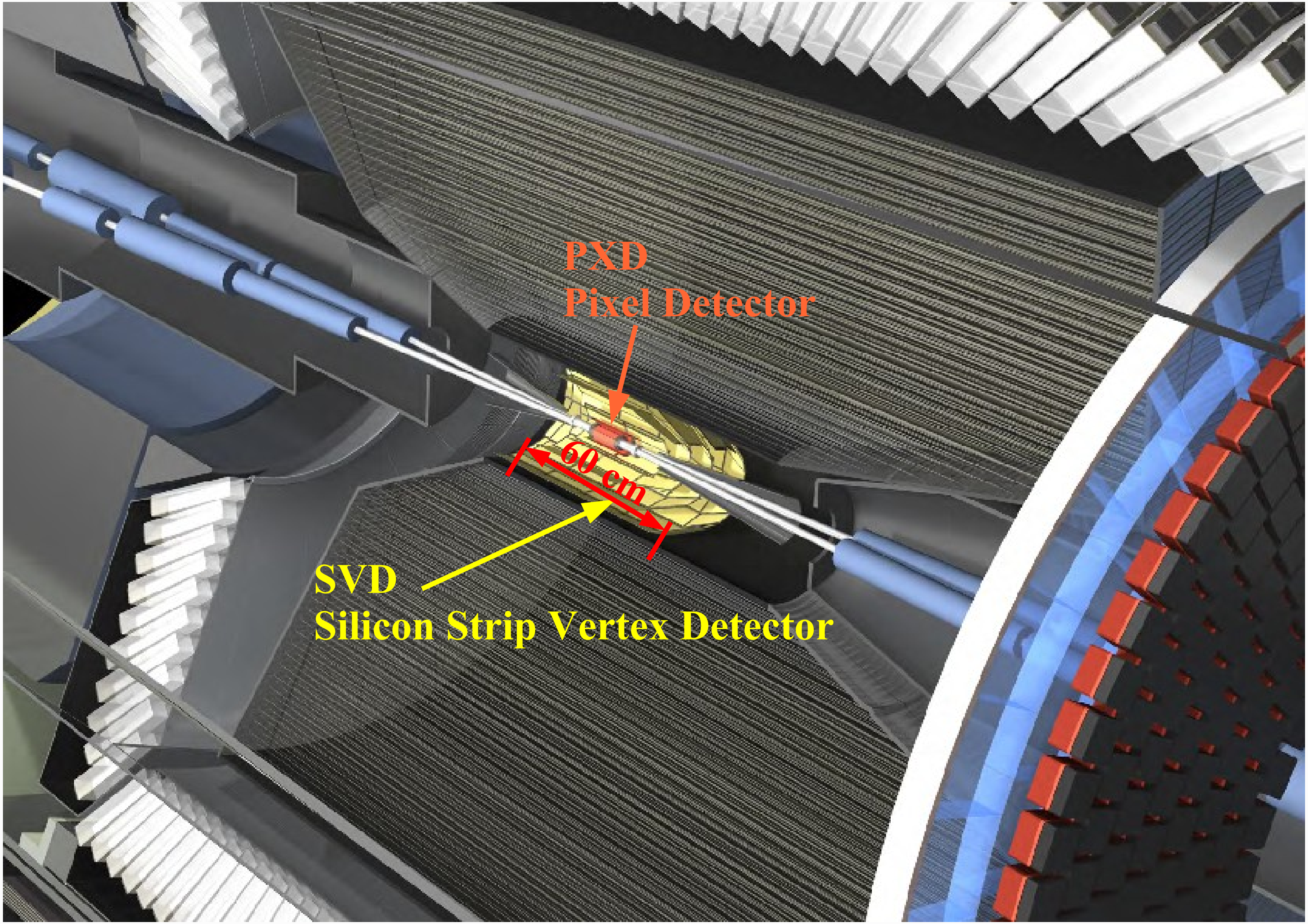
- Belle II @ SuperKEKB Tsukuba/Japan
- e^+e^- asymmetric collider, up to 11 GeV center of mass energy
- Belle and KEKB upgrade
  - Luminosity $L=8\times10^{35}$ cm^{-2} s^{-1}, 40 times of Belle @ KEKB
  - 30 kHz trigger rate
- Physics program
  - Continuation of successful Belle program
  - CP violation at the intensity frontier
  - Decays of B and D mesons
  - Rare and forbidden decays
  - Exotic hadrons
  - New physics beyond standard model
- Start data taking in 2015

Belle II Detector Setup

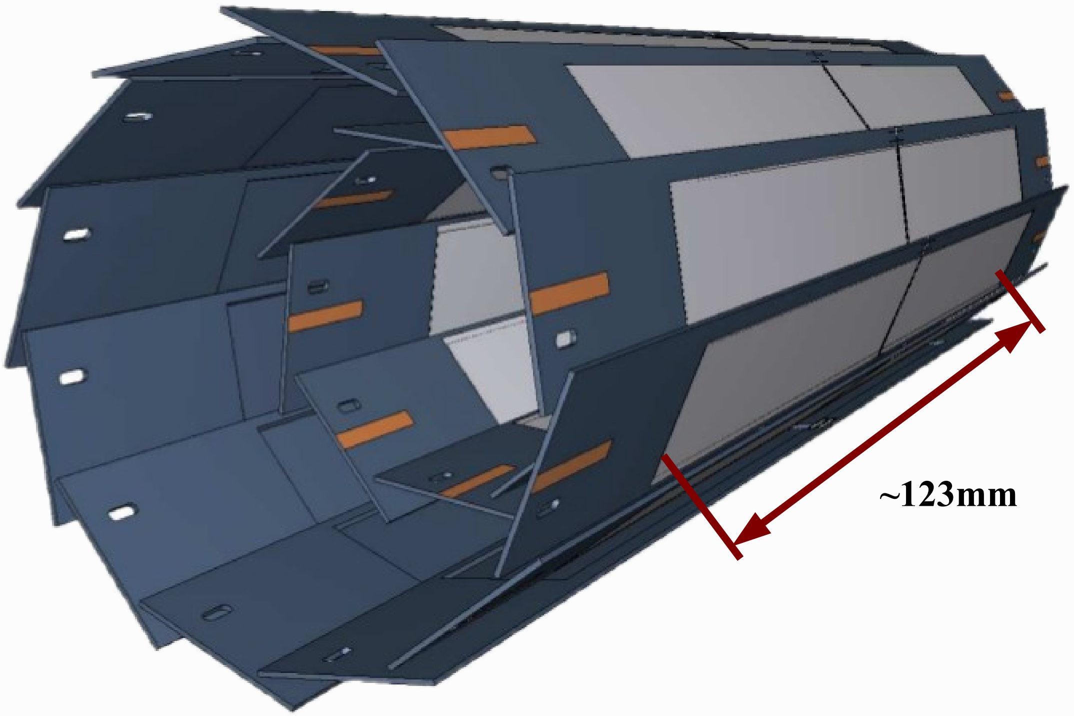


- Higher PID sensitivity (pion/kaon)
- Better secondary vertex resolution (factor 2)
- Improved K^0_s detection

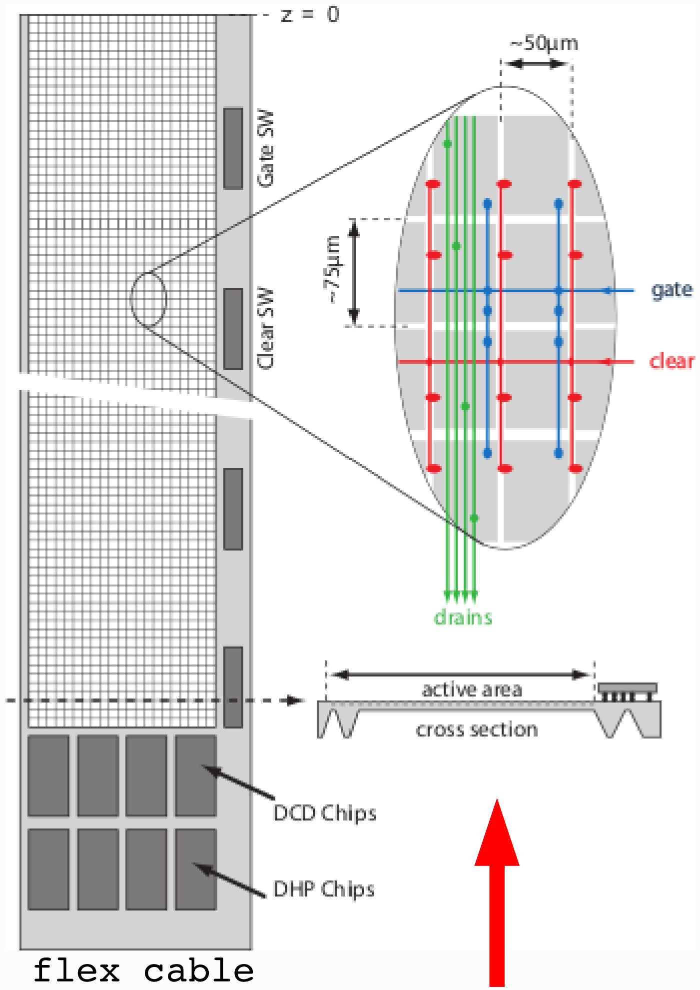
Vertex Detectors



DEPFET Pixel Sensor



Size of an energy drink tin can!

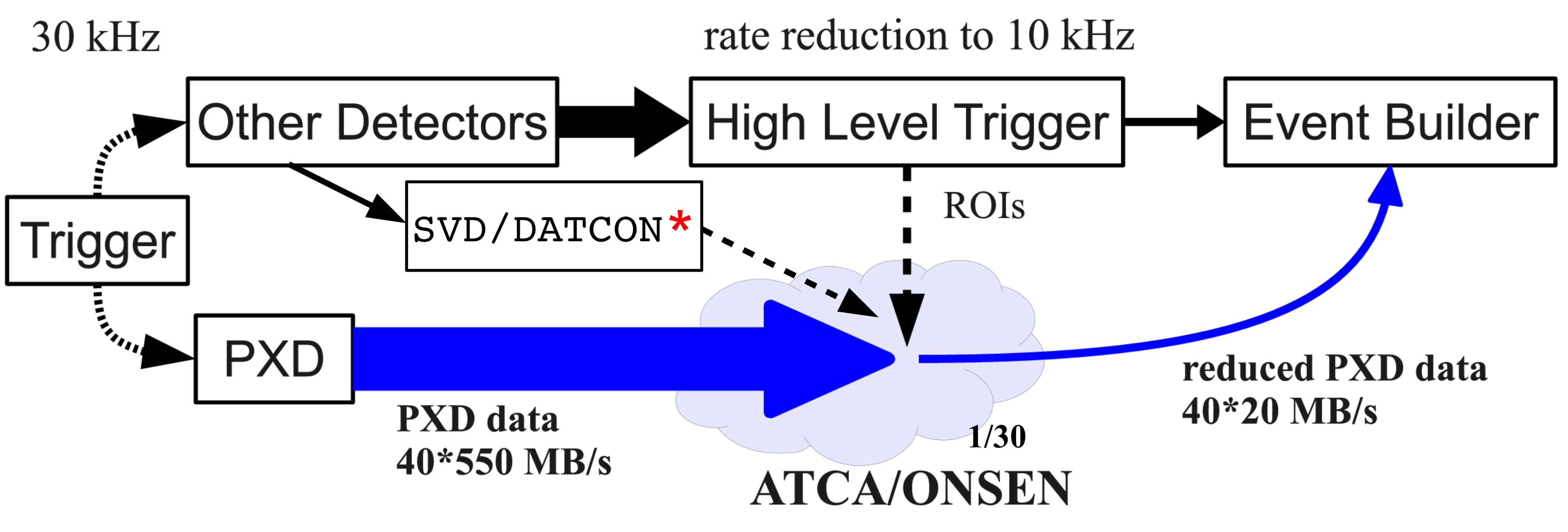


flex cable connection

- 2 layers at 14 and 22mm radii
- 8 inner (12 outer) ladders → 40 modules
- Monolithic all-silicon sensors, no need for support structures
- DEPFET – Depleted Field Effect Transistor
- Pixel size 50µm*50µm (outer: 50µm*75µm)
- Thinned to 75µm thickness to reduce multiple scattering
- Continuous readout; line by line (rolling shutter mode)
- Readout rate 50kHz (full detector)

Pixel Detector DAQ Scheme

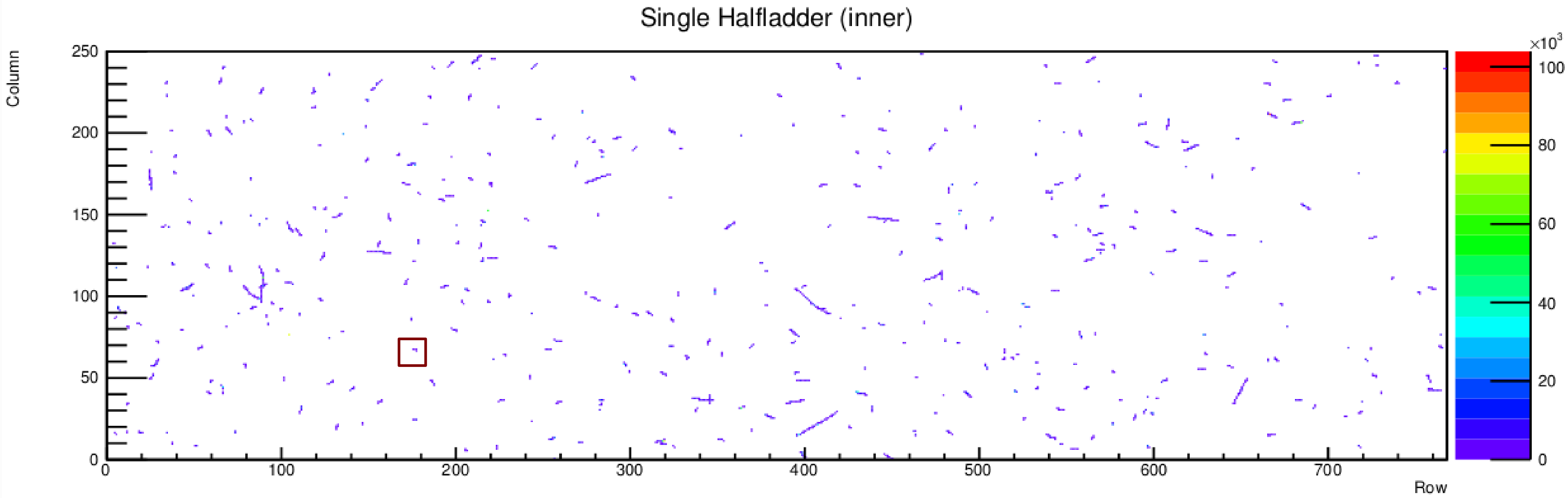
- Expected data rate **≤22 GB/s** for whole PXD detector
  - Corresponds to 3% pixel occupancy, dominated by background
- 10 times more than all other detectors together! Reduction of **factor 30** needed.
- Pixel detector readout differs from other detector sub systems:



- Buffered in ATCA/ONSEN until HLT decision (unordered, latency up to 5 s)
- Selection of pixels inside regions of interest (ROI)
  - ROIs calculated by HLT, additional sources foreseen (SVD)

(*) Talk by Michael Schnell

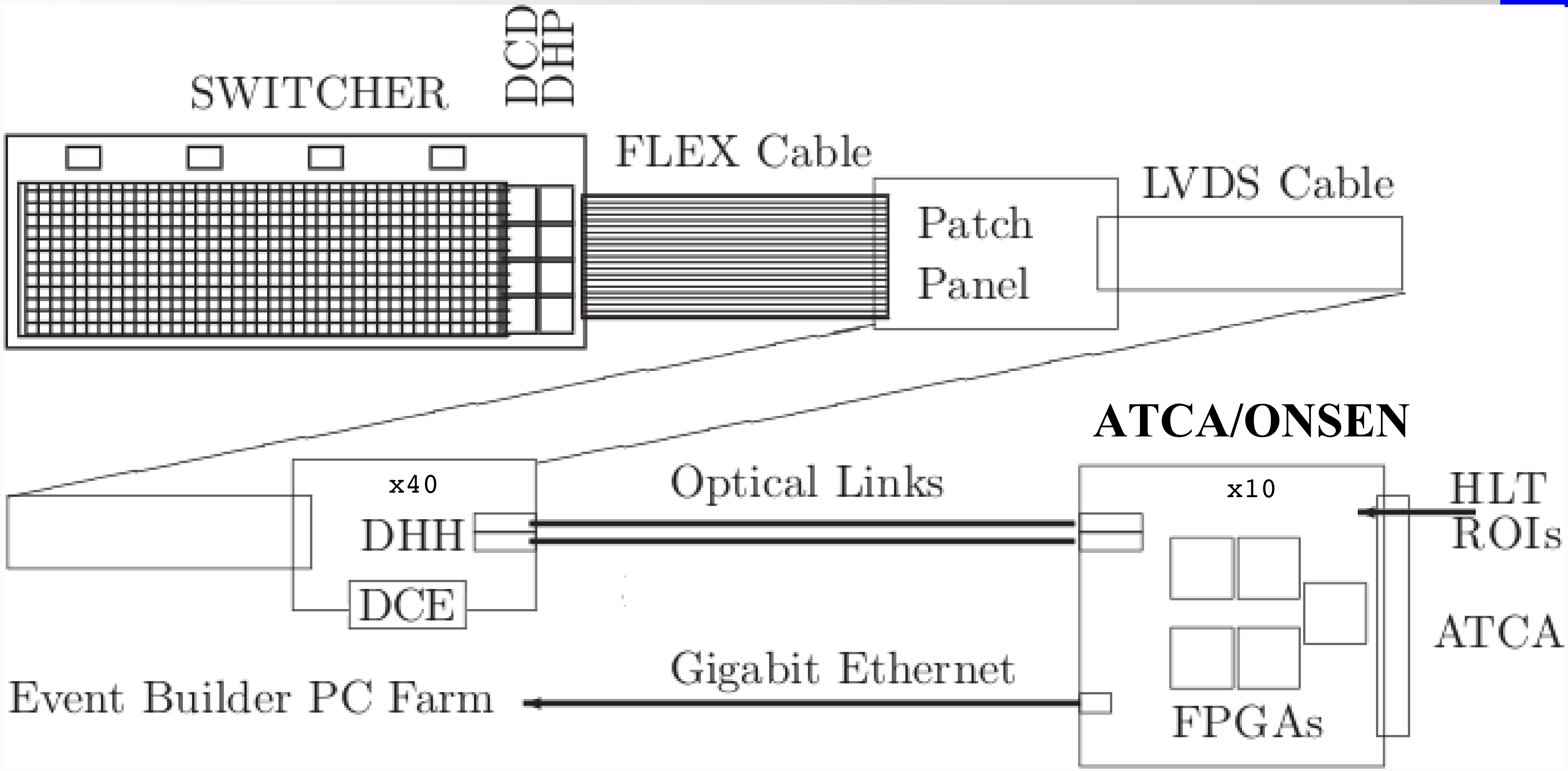
Simulation Data – Beam Background Only



- No way to decide what is relevant from PXD alone
- Regions of Interest from tracking needed

(Simulation by Andreas Moll)

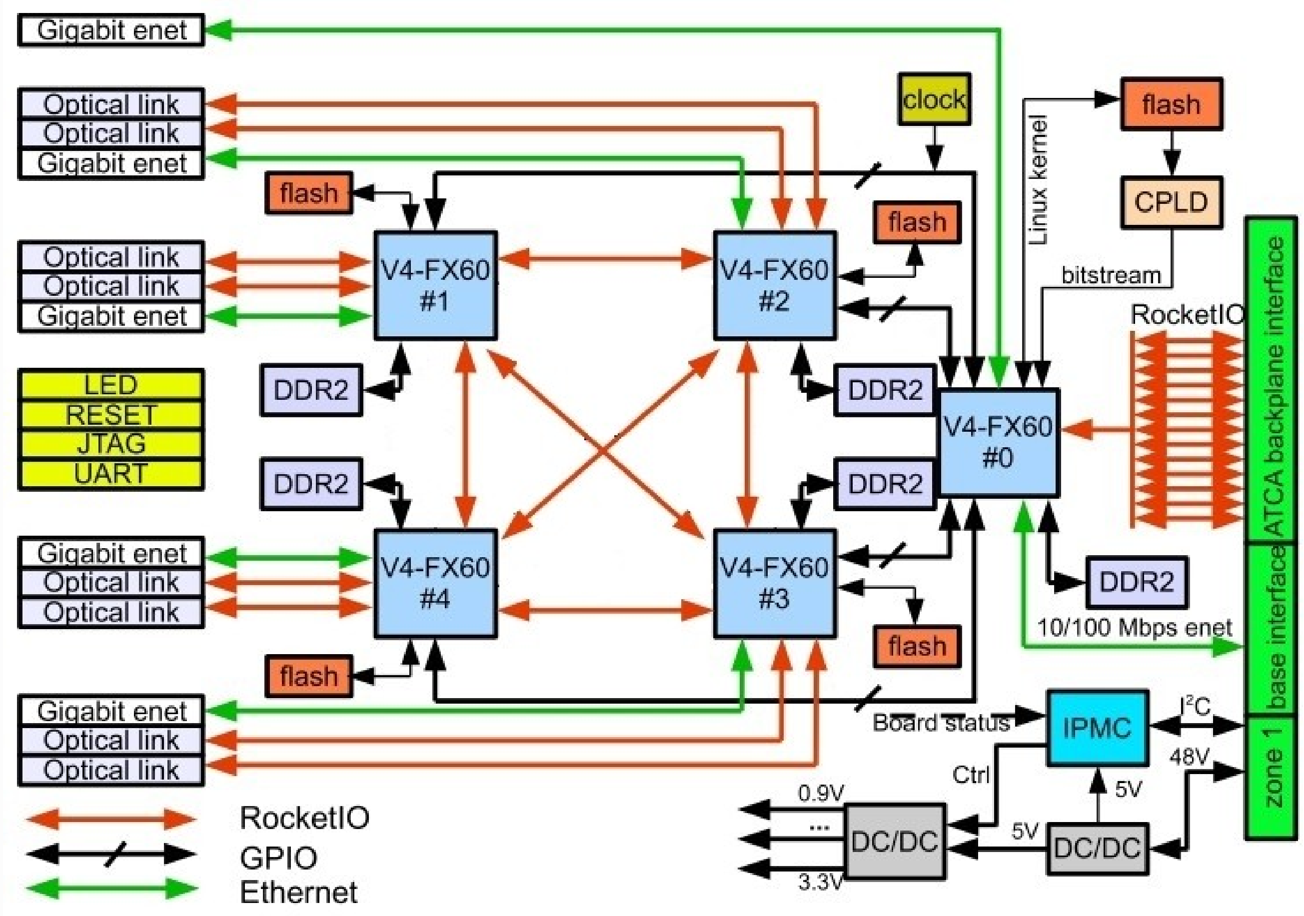
PXD Readout Chain



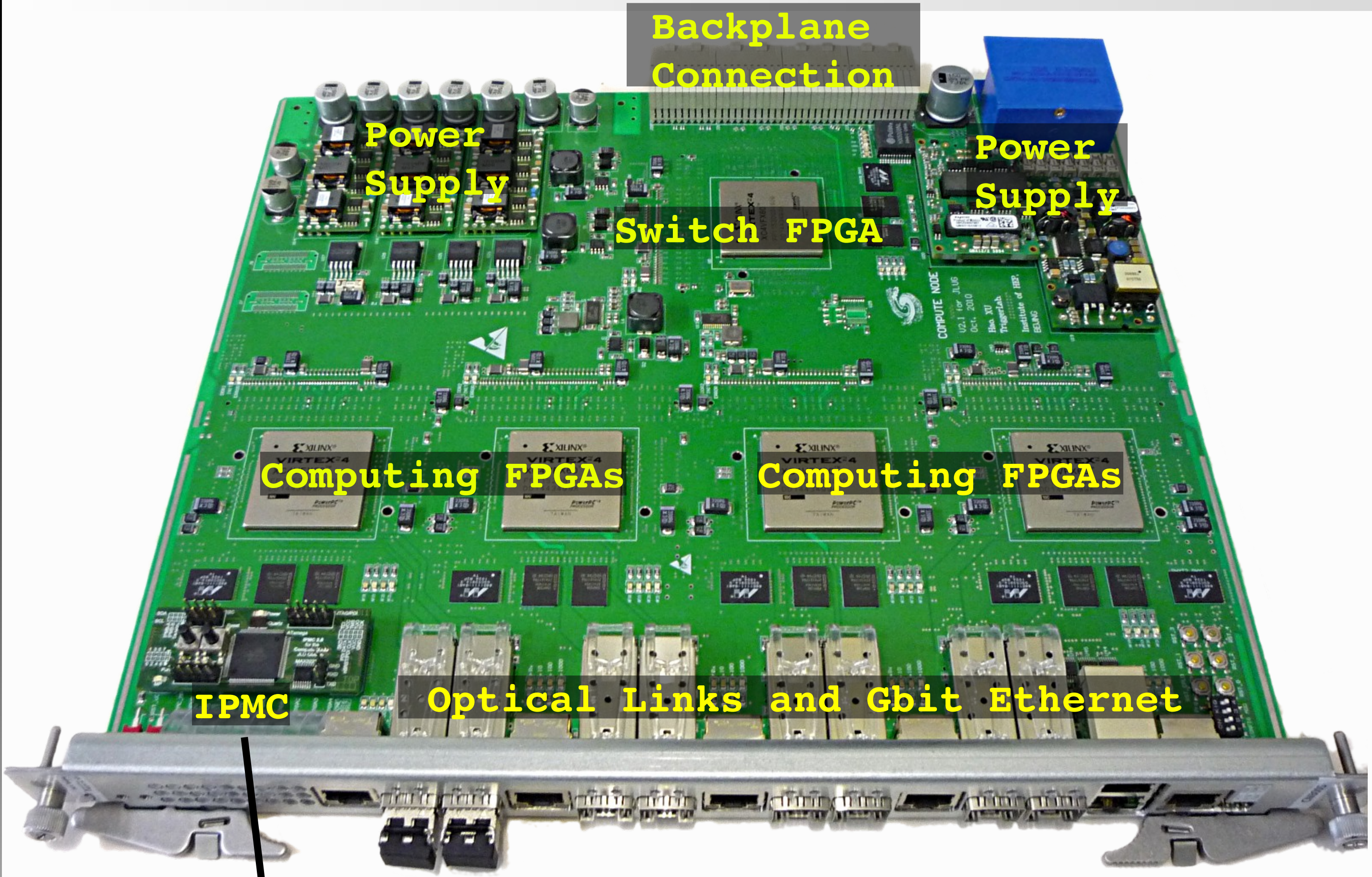
- DHH & DHHC (I. Konorov, TU Munich): Framing, Loadbalancing, ATCA/uTCA, Virtex 6
- Clusterfinder (DCE) “addon”, ASIC -> (A. Wassatsch, MPI Munich)

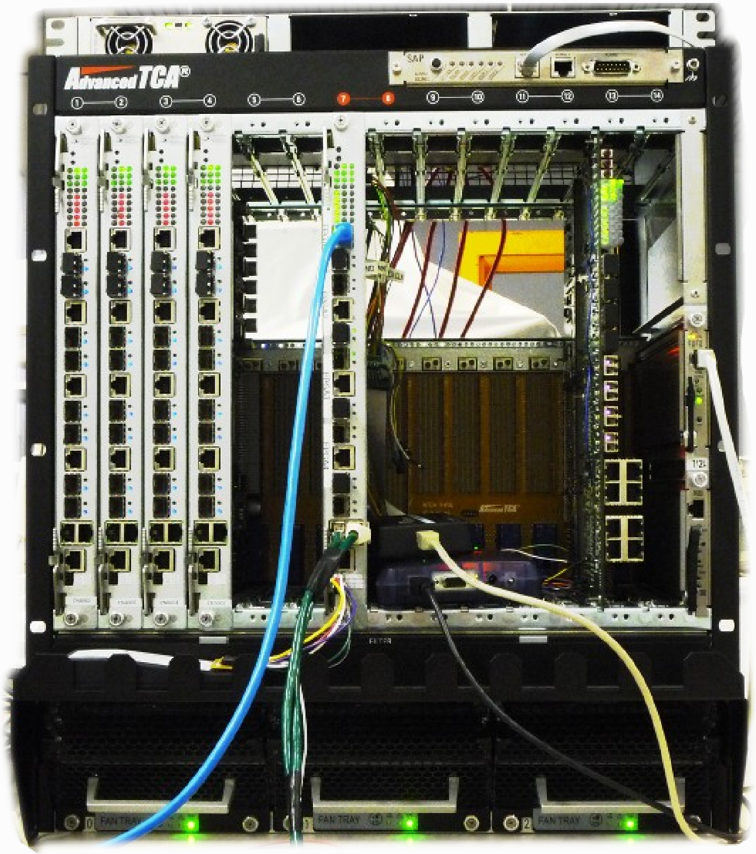
The Compute Node (CN)

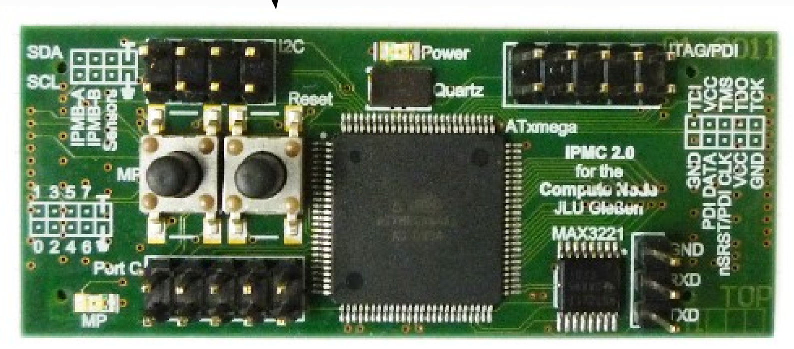
- High Performance Computing
- 5 Virtex-4 FX60 FPGA
  - (new: Virtex-5 FX70T)
- 5×2GB DDR2 RAM
  - (new: 2*2GB/FPGA)
- Interconnected by RocketIO
- 8 Optical Link (3Gbps each)
  - (new: 6.5Gbps)
- 5× Gigabit Ethernet
- 16× RocketIO to backplane (full mesh)
- Embedded PowerPC in each FPGA for slow control
- ATCA compliant (Advanced Telecommunications Computing Architecture)



Compute Node, 2nd Revision



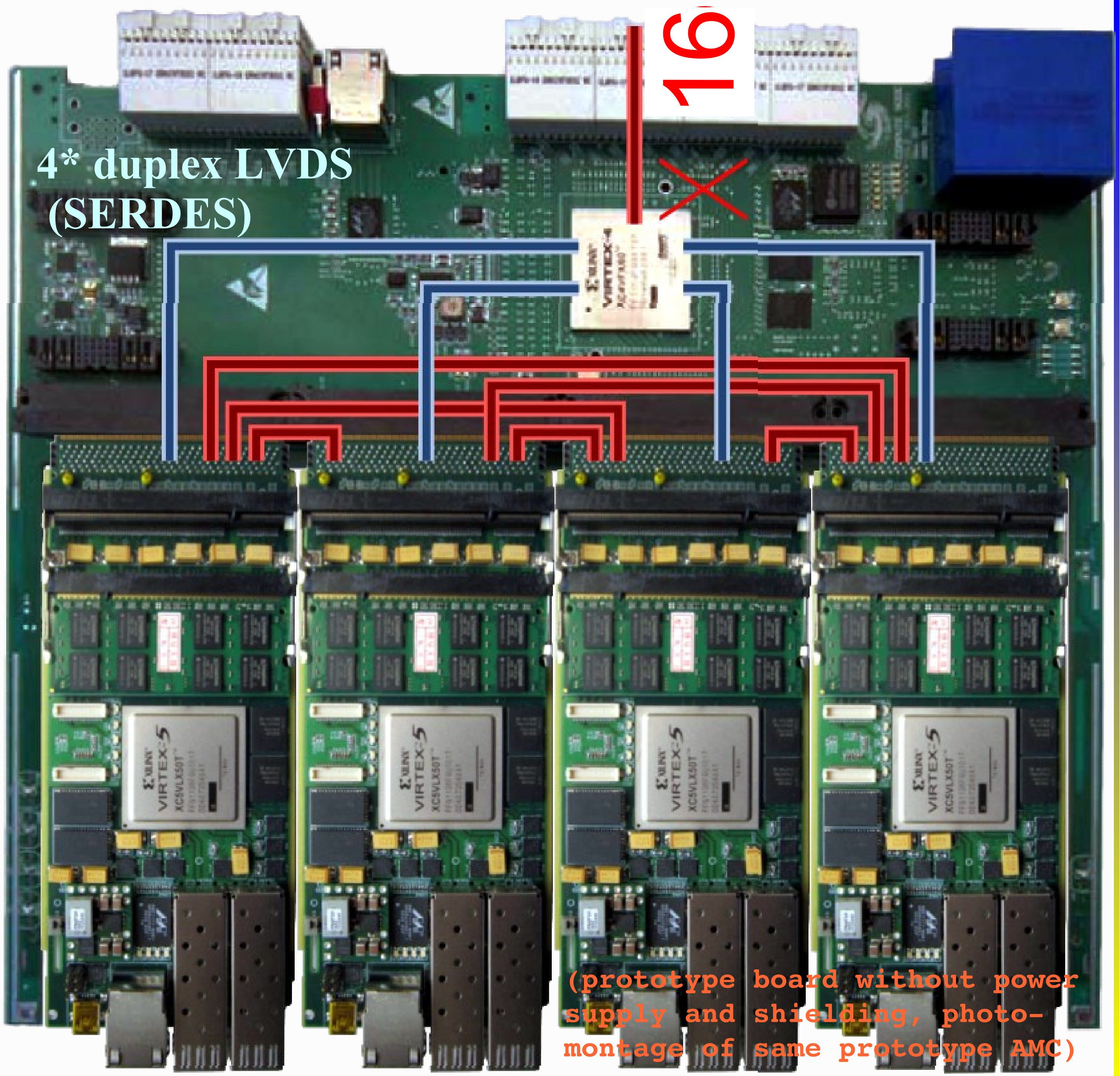




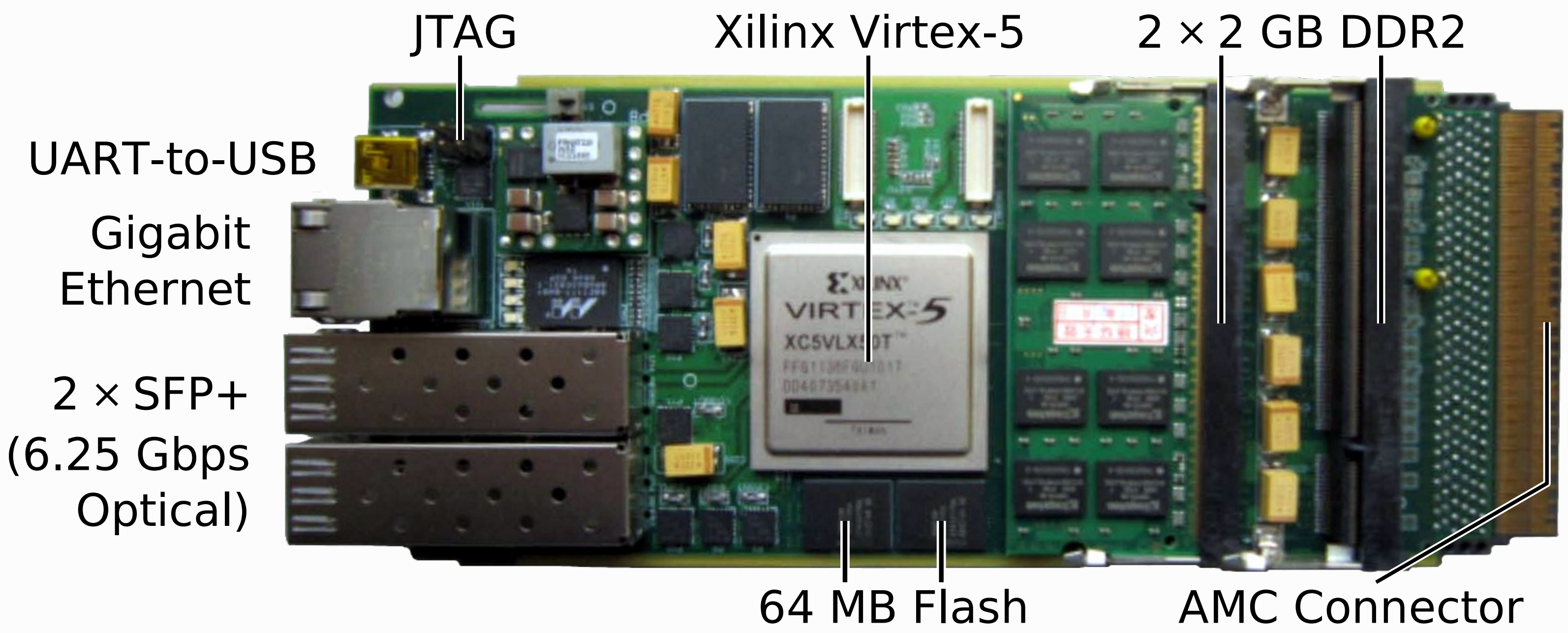
- IPMI remote control
  - Small piggy-back board
  - Functions: power on/off, power negotiation, health monitoring, board reset, bitstream selection, global addressing

Compute Node 3rd Revision

- 1 xTCA carrier + 4 AMCs (Advanced Mezzanine Cards)
- Computing and switching part are separated
- Connectivity similar to version 2
- Full mesh on the carrier and full mesh on the backplane
- Each AMC has
  - 4* LVDS to MB
  - 2* LVDS to AMC
  - 1* RocketIO to AMC
  - 3* RocketIO to RTM
  - 2* LVDS to RTM

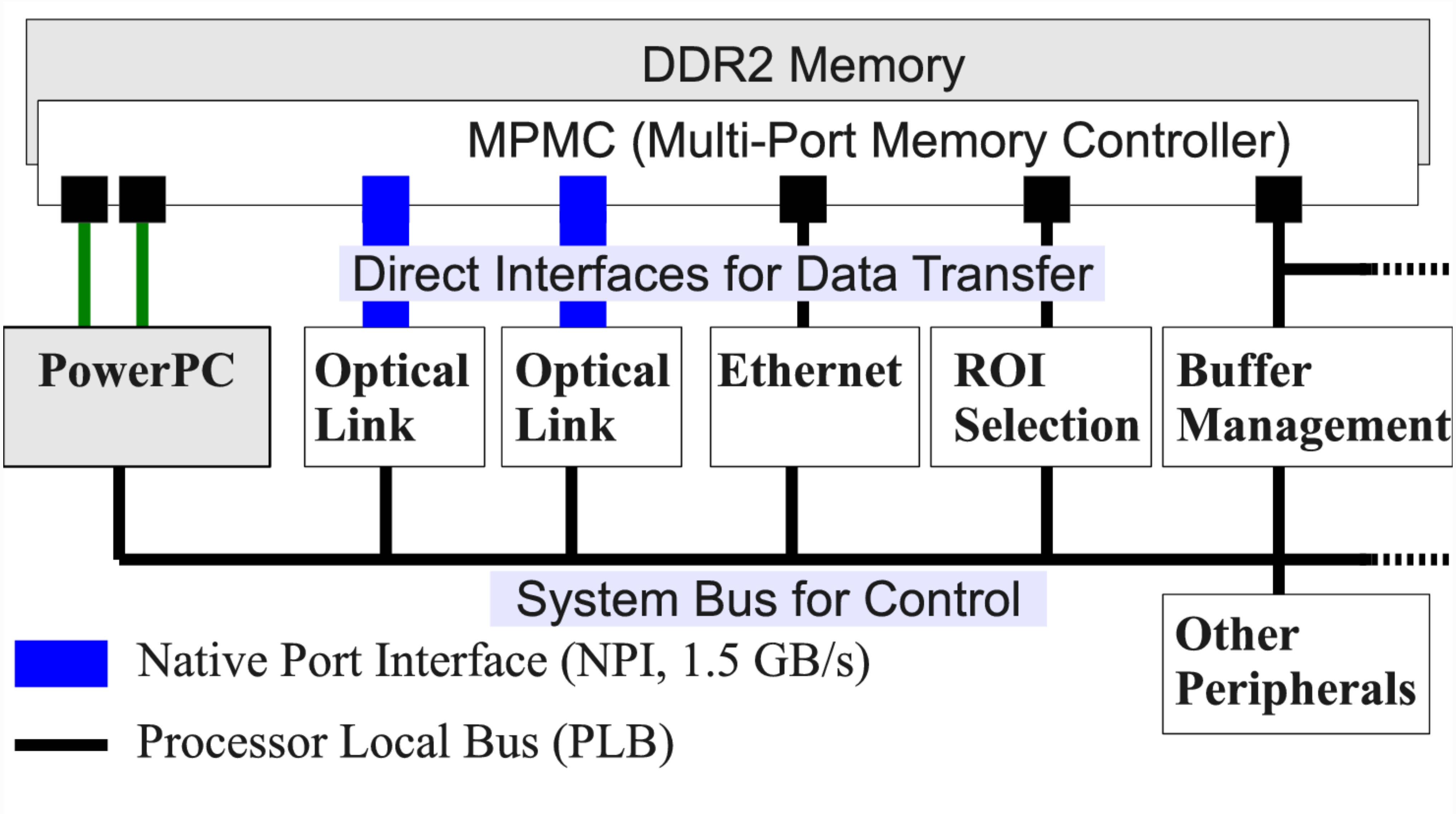


Advantages of 3rd Version



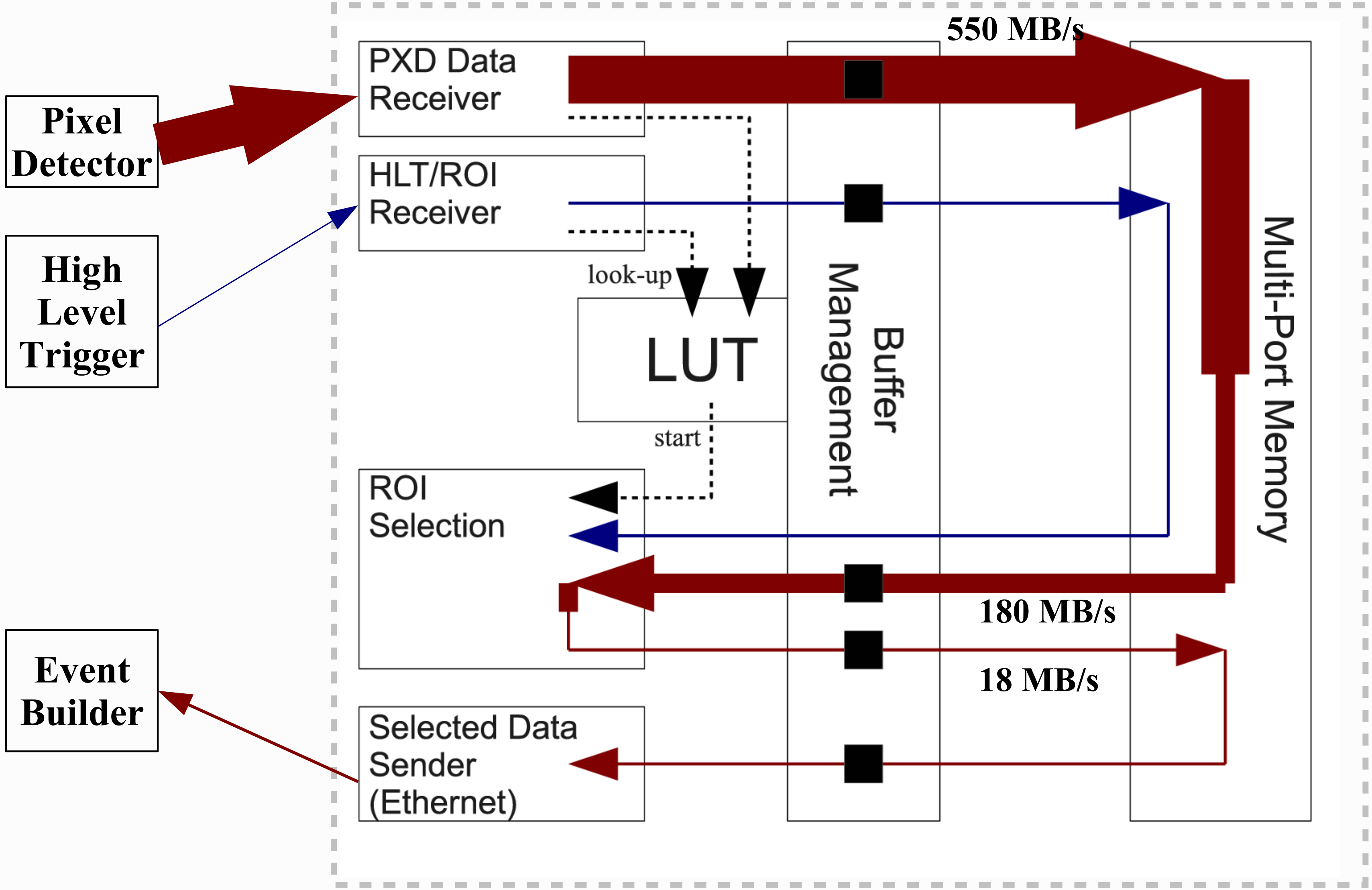
- AMC can be updated independently, no redesign of whole board required
- Different AMCs can be used on same carrier board
  - Adapt to special purpose
- Example:
  - AMC with 4 optical links (same board design)
  - More powerful or cheaper FPGAs
- AMCs can also work standalone in μTCA shelf.

FPGA System Setup

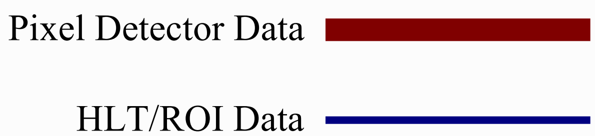


- Operating System: Linux (kernel 3.3.+)
- Error handling by interrupts in a kernel driver

Data Flow for the PXD Data

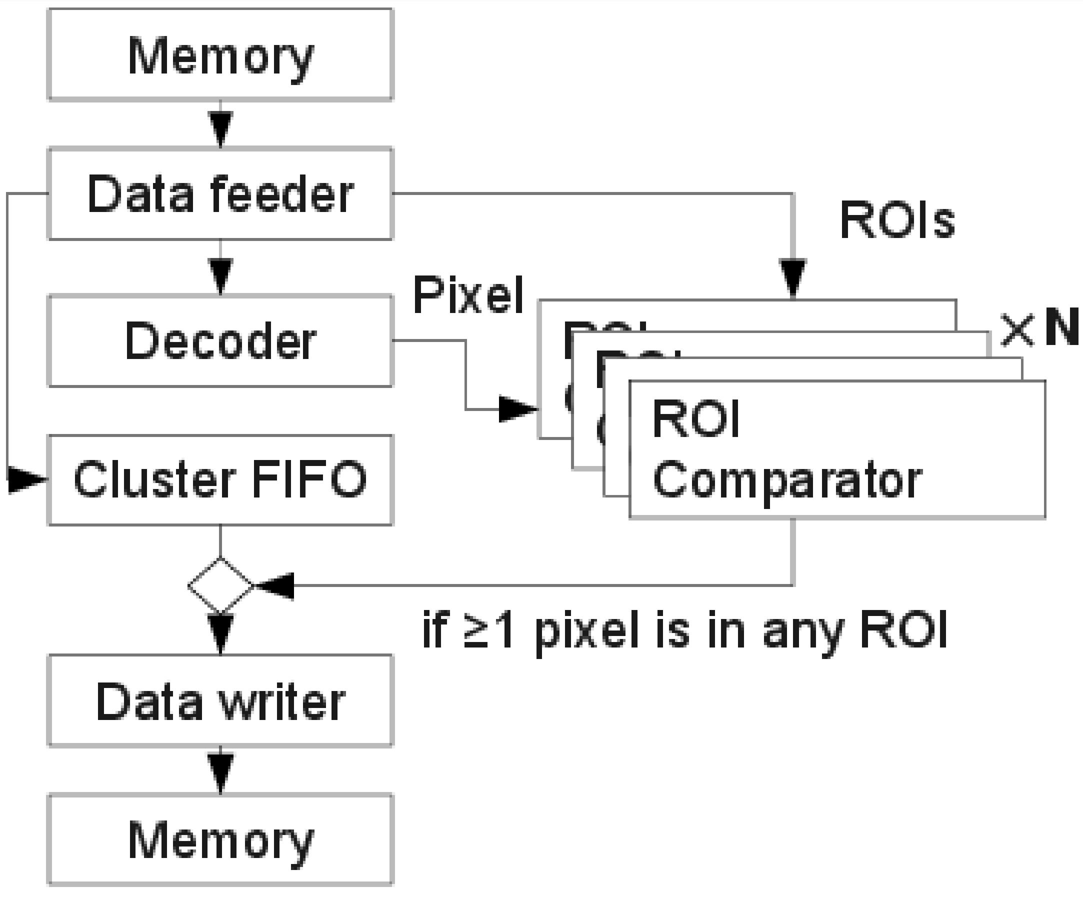


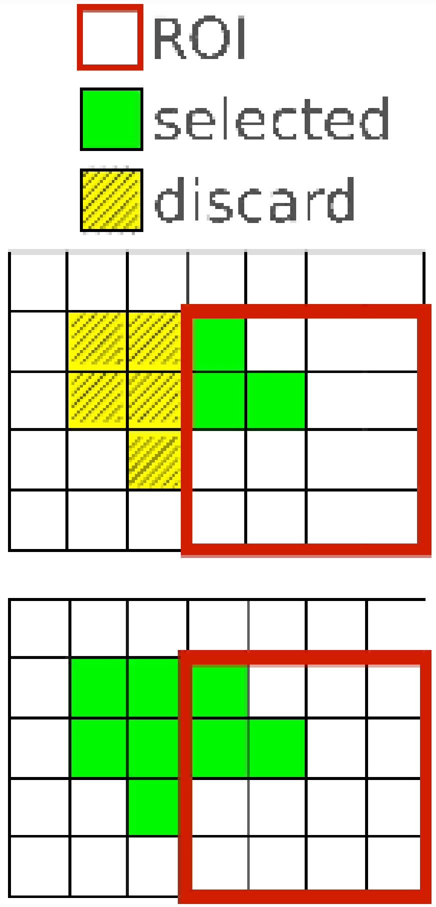
- All parts implemented in VHDL



Region of Interest Selection

- Selecting pixel/cluster data inside rectangular regions
- One pixel is processed against all ROIs in parallel
- Two modes:
  - Pixel mode: Each pixel is checked separately
  - Cluster mode: If at least one pixel is inside a region, keep full cluster



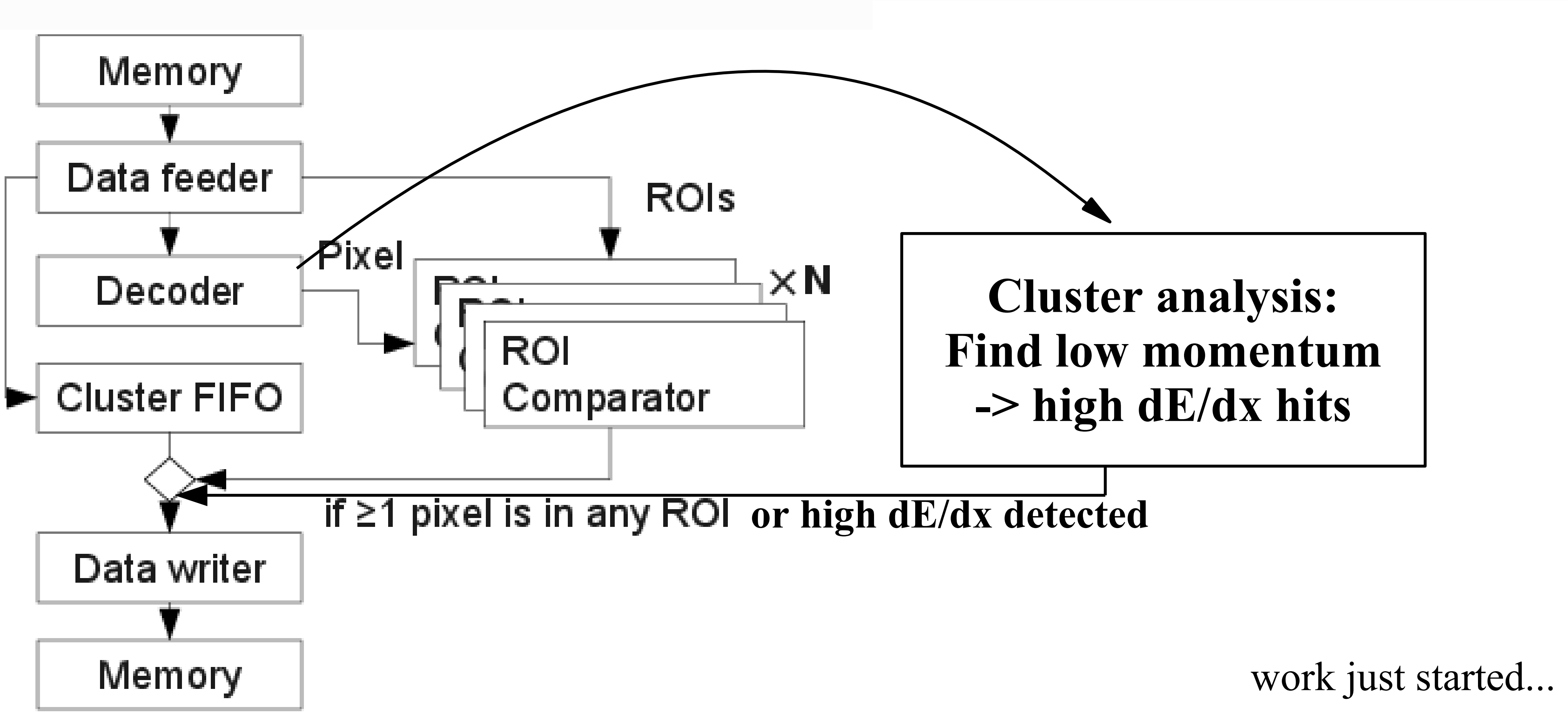


Pixel based selection

Cluster based selection

Cluster analysis - dE/dx hit rescue

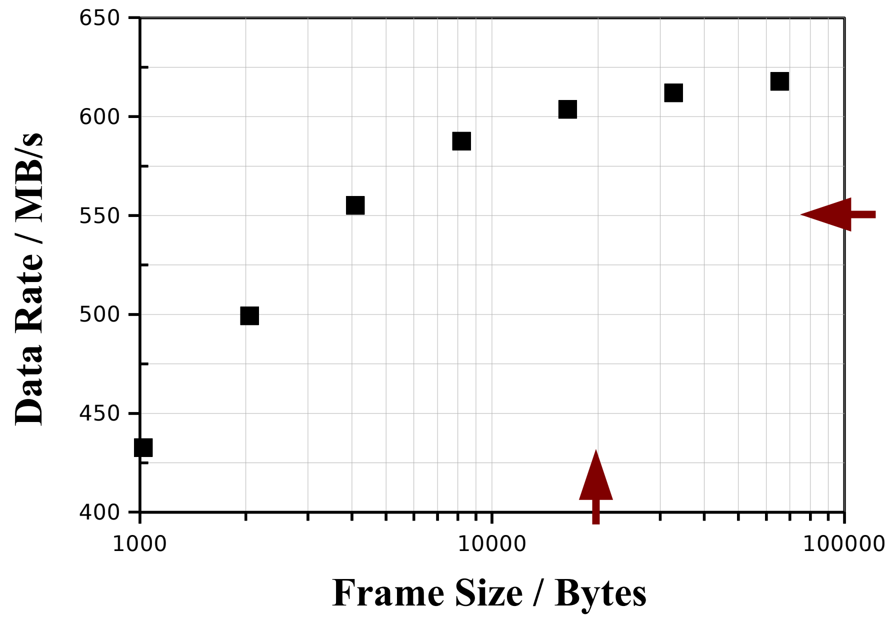
- Low momentum pions important for several Ds* decays
- Large energy loss (dE/dx), will not reach CDC driftchamber
  - High level trigger will miss them.
- Rescue pixel information by cluster analysis
- Cluster charge, size, shape etc.

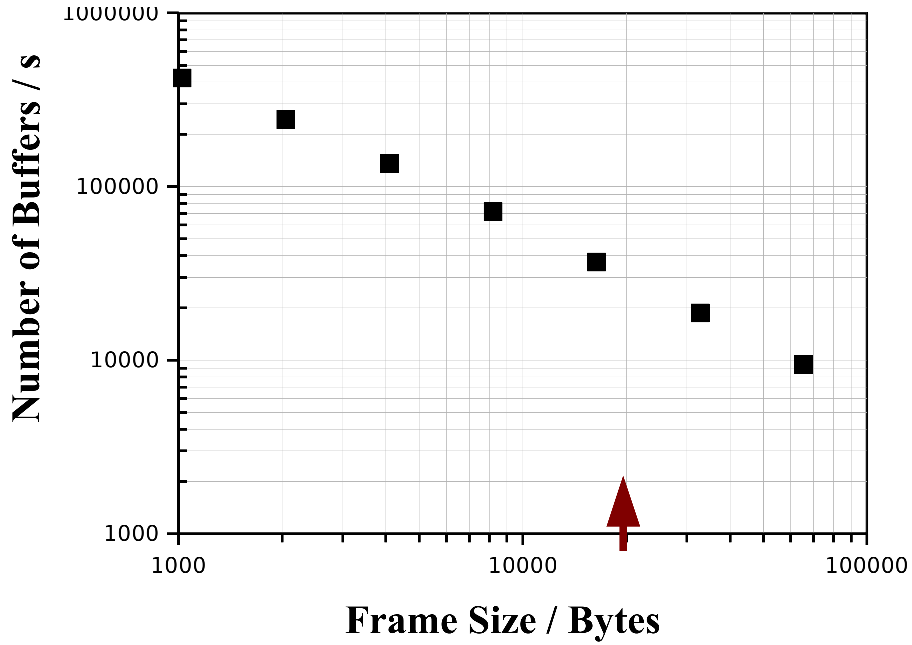


work just started...

Optical Links

- Tested with different speeds (2, 3, 6.25 Gbit/s) between CN2 and CN2/CN3 with Aurora protocol
- Realistic test:
  - Memory → Optical Link on FPGA 1
  - Optical Link → Memory on FPGA 2
  - Including Aurora
  - Including buffer manager
- At 20 kB (expected PXD event size per module) data rate is >600 MB/s, well above requirements
- Corresponds to a event rate of 30 kHz





Ethernet

- (Linux) Software Stack is slow, even so some “offload” is done in hardware
- For high data rates, FPGA implementation mandatory
- SiTCP(1)
  - >100 MB/s achieved (reading data from memory)
  - Limit: Only one TCP connection, not usable with an event builder farm
- UDP implementation
  - >120 MB/s send out (test packets), 60 MB/s data from memory
  - Copper and optical Gbit Ethernet (by optical link)
  - No package drop detection (yet)
  - Evaluation of packet losses and receiving in hardware is ongoing
  - Optional input/output tunnel to OS (for TCP/IP, slowcontrol)
- Requirements of 32 MB/s are met.

work by Grzegorz Korcyl
University Krakow

(1) T. Uchida, Hardware-Based TCP Processor for Gigabit Ethernet, IEEE Trans. Nucl. Sci., vol. 55, no. 3, 2008.

Prototype System For $\overline{\text{P}}$ANDA

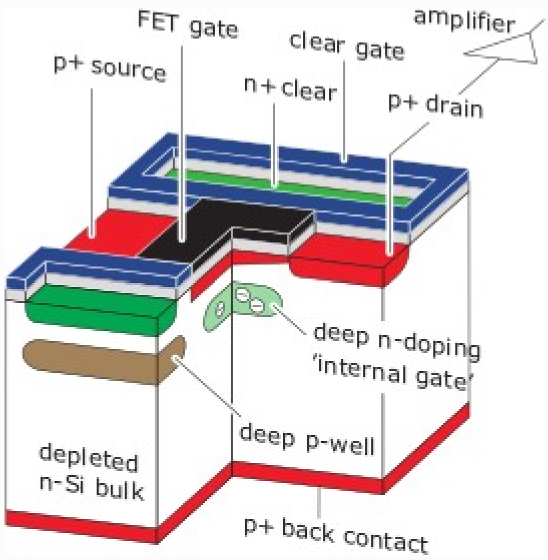
- Compute Node is a flexible system, foreseen for $\overline{\text{P}}$ANDA DAQ
  - Several CN v2 have been distributed within the collaboration
- Gain experience with CNs for $\overline{\text{P}}$ANDA with a smaller system
- Belle II DEPFET pixel detector delivers 22GB/s data (already zero suppressed)
  - Data processed on 10 Compute Nodes (1 ATCA crate)
  - 10% of PANDA
- PANDA is triggerless
  - Not always FIFO stream processing
  - Matching: Data to be buffered until a second data input is ready
  - Random access (and bookkeeping) of memory resources necessary
  - Data switching & event building

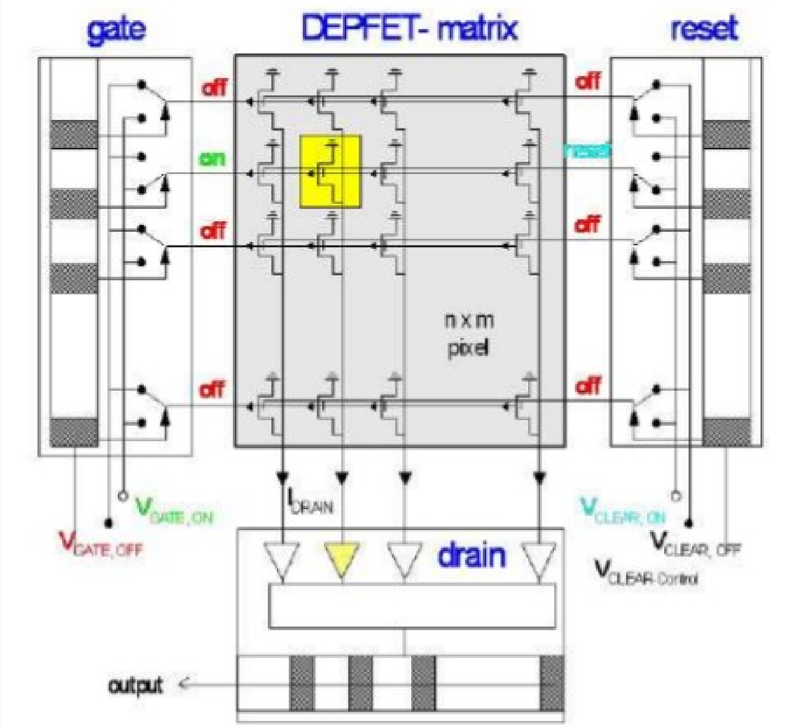
Summary and Outlook

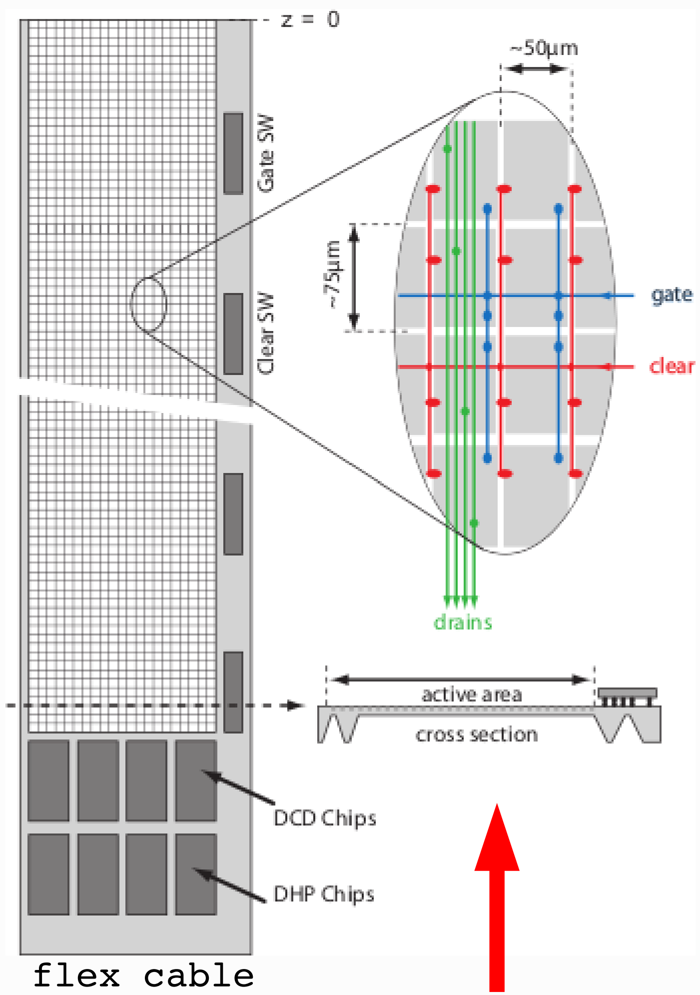
- Belle II DEPFET pixel detector requires data reduction by a factor of 30
  - Selecting only data which corresponds to a track found by HLT or SVD
  - Using ATCA compliant Compute Node
- Implementation
  - Complete system implemented in VHDL
    - Optical link receiving, buffer management, event look-up tables, region-of-interest selection, Ethernet
  - Initialization, slow control and error handling on PowerPC
- Prototype working (input, buffering, processing, output)
- Critical requirements fulfilled (data throughput)
- Ongoing work:
  - Transition to μTCA based Compute Node revision 3, Virtex 4 → Virtex 5
  - Ethernet (speed improvements)
  - Full readout chain test (2013)

Backups

DEPFET Pixel sensor

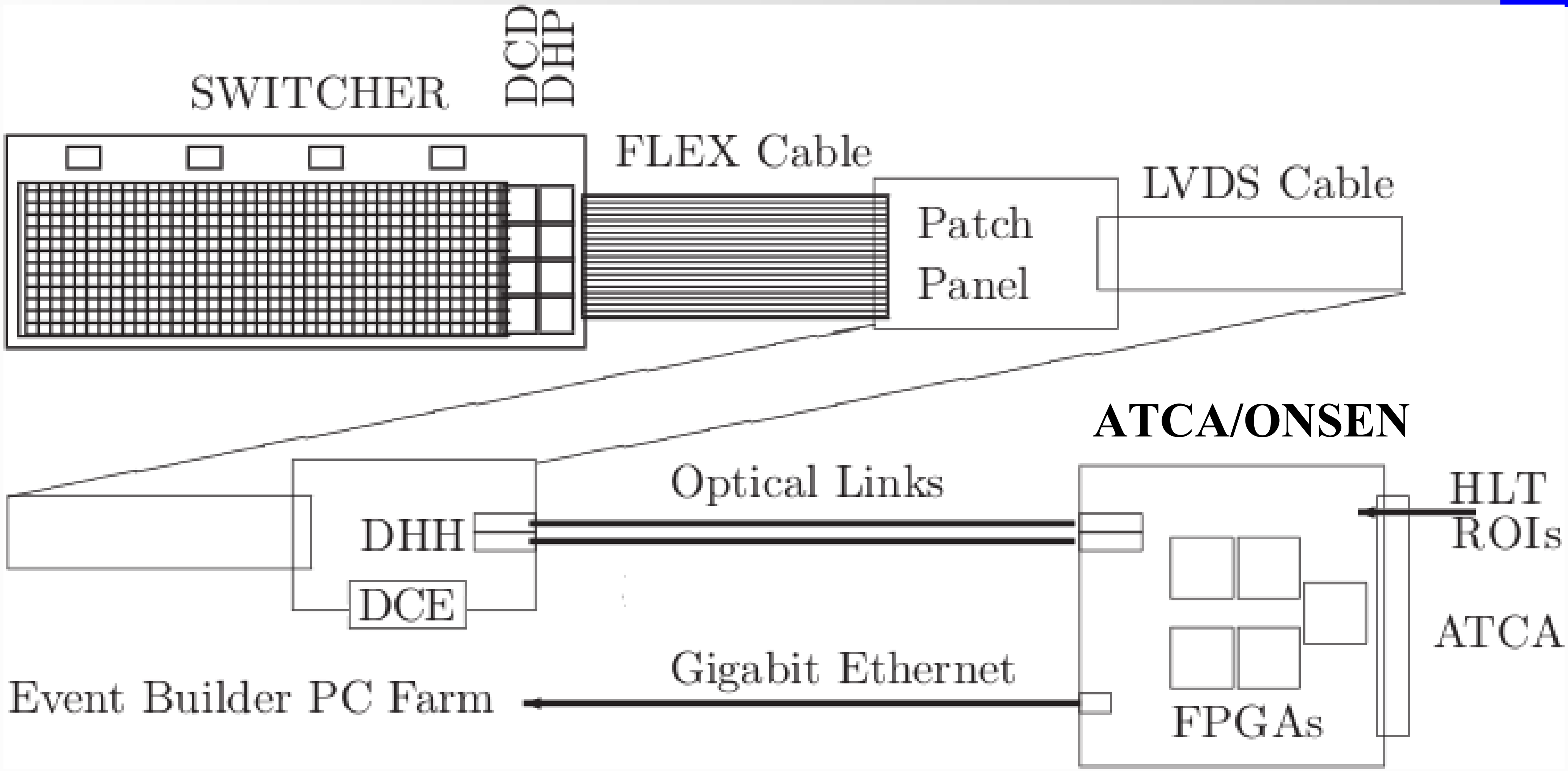






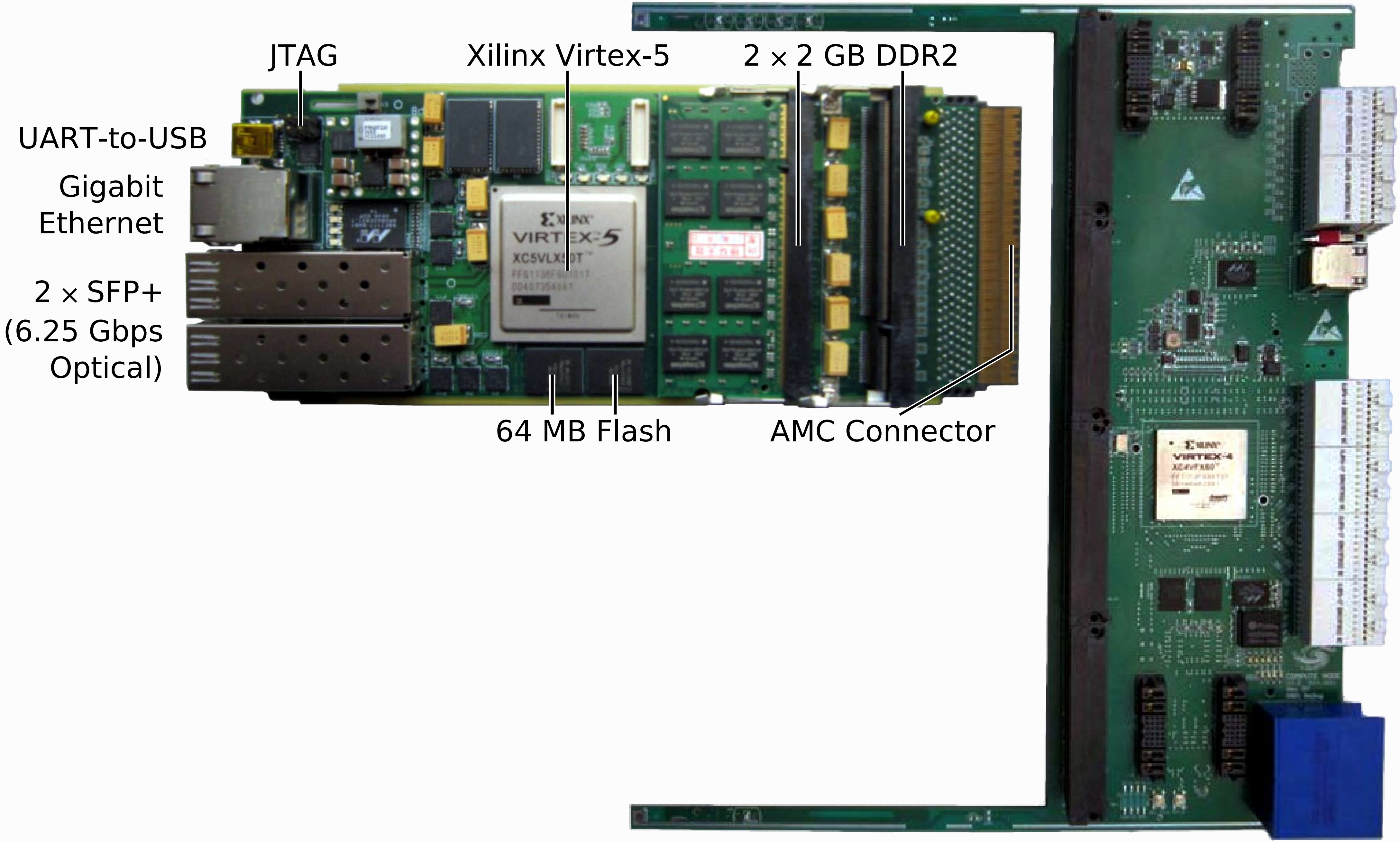
- monolithic all-silicon sensors, no need for support structures
- pixel size 50um*50um (outer: 50um*75um)
- thickness 75um
- continuous readout; line by line (rolling shutter mode)
- readout rate 50kHz (full detector), deadtime 20us
- designed for 10MRad, >5 years operation.
- 2*90W power
- CO_2 two phase evaporation cooling at -15°C and 23bar

PXD Readout Chain



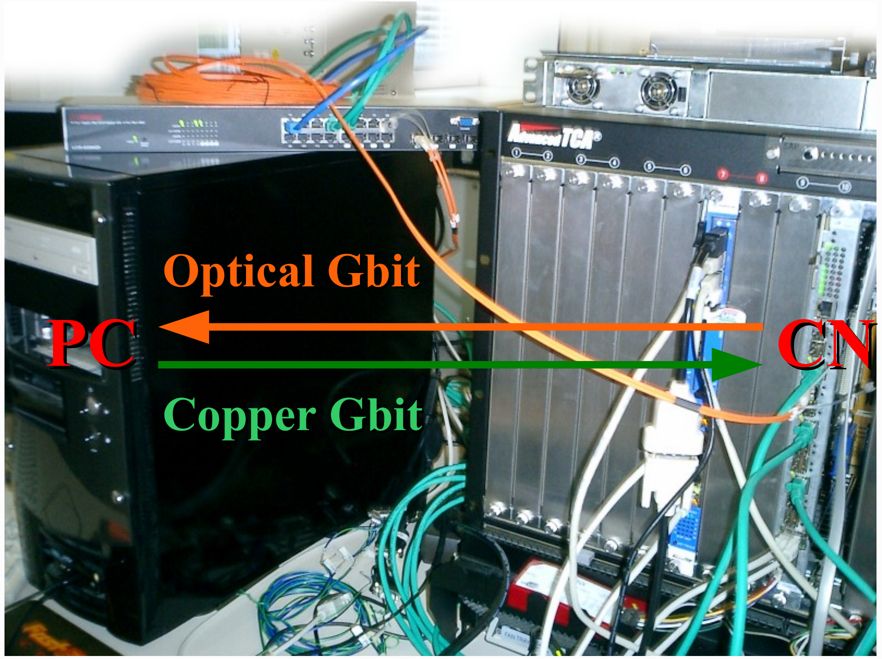
- PXD: (8+12)*2=40 half ladders
- Each half ladder readout by two optical links
- ATCA/ONSEN sends data to Event Builder #2 (Gbit Ethernet)

Compute Node 3th generation

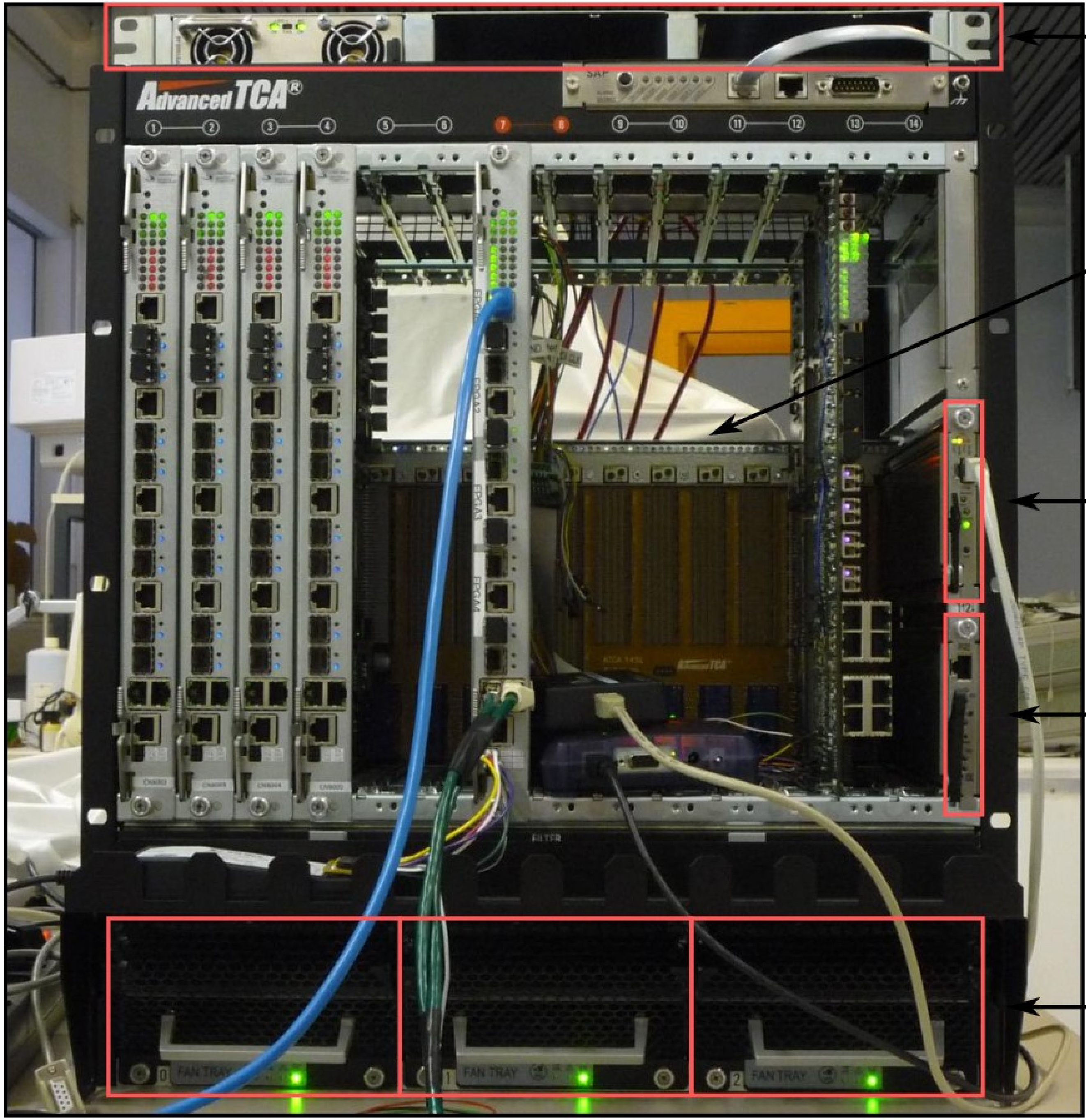


Prototype Test Setup

- PC as data source (random data and ROIs) and receiver
- Send in data and delayed high level trigger by Ethernet (PowerPC Linux stack)
- Anything else in VHDL/hardware
  - Processing core
  - Memory and buffer management
  - UDP sender
- PC crosschecks results
- No error has been observed in more than 4×10^{10} processed pixels



ATCA Shelf



3 kW Power Supply

Full Mesh Backplane

Primary
Shelf Manager

Backup
Shelf Manager

Fan Trays

Possible Extensions: Slow Pions

- Improvements: Low pT tracks
  - Pions from $D^*(2010)^\pm \rightarrow D^0\pi^\pm$ have low momentum (pT<80MeV) but are important for B-tagging
  - below minimal ionizing, large dE/dx
  - Track loses energy, might not be seen by HLT (no CDC track)
  - Find pixels/clusters with a very high energy deposition and keep them as additional ROIs
  - But: Keep <10% of pixel data; Efficiency ↔ Purity

- Single photon / single pixel rejection
  - Isolated fired pixels
  - Saves bandwidth if removed before stored to RAM
  - Isolated pixels possible for real tracks, too. Simulation needed.

Background in Pixel detector

Belle II

- Occupancy of PXD dominated by background
  - physics <1% occupancy
  - background up to 2%
- Beam related background by
  - Synchrotron radiation
  - Beam - gas reactions
  - Touschek effect:
    - Intra-beam scattering because of high particle density
    - exchange of transversal to longitudinal momentum
    - loss of beam particles
    - simulation of full storage ring needed!
- Interaction background
  - Radiative QED