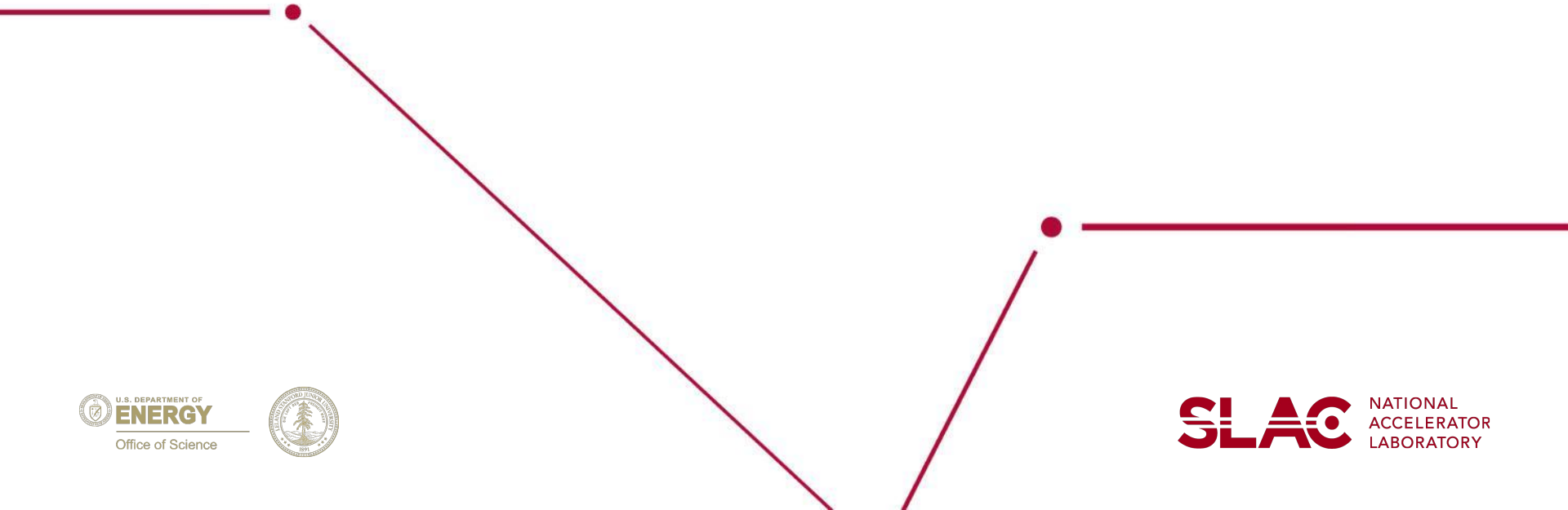
DUNE Single-Phase FD DAQ Overview

Matt Graham, SLAC on behalf of DAQ team

DUNE Calibration Workshop

March 15, 2018



U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

DUNE DAQ Requirements

SLAC

Requirement	Description
Scalability	The DUNE FD DAQ shall be capable of receiving and buffering the full raw data from all four FD modules
Zero deadtime	The DUNE FD DAQ shall operate without deadtime under "normal" operating conditions
Triggering	The DUNE FD DAQ shall provide full-detector triggering functionality as well as self-triggering functionality; the data selection shall maintain high efficiency to physics events while operating within a total bandwidth of 30 PB/year for all operating FD modules
Synchronization	The DUNE FD DAQ will provide synchronization of different FD modules to within 1 μs, and of different subsystems within a module to within 10 ns

A few more: total data rate to tape <~ 30 PB/year

The Plan & Expected Data Rates

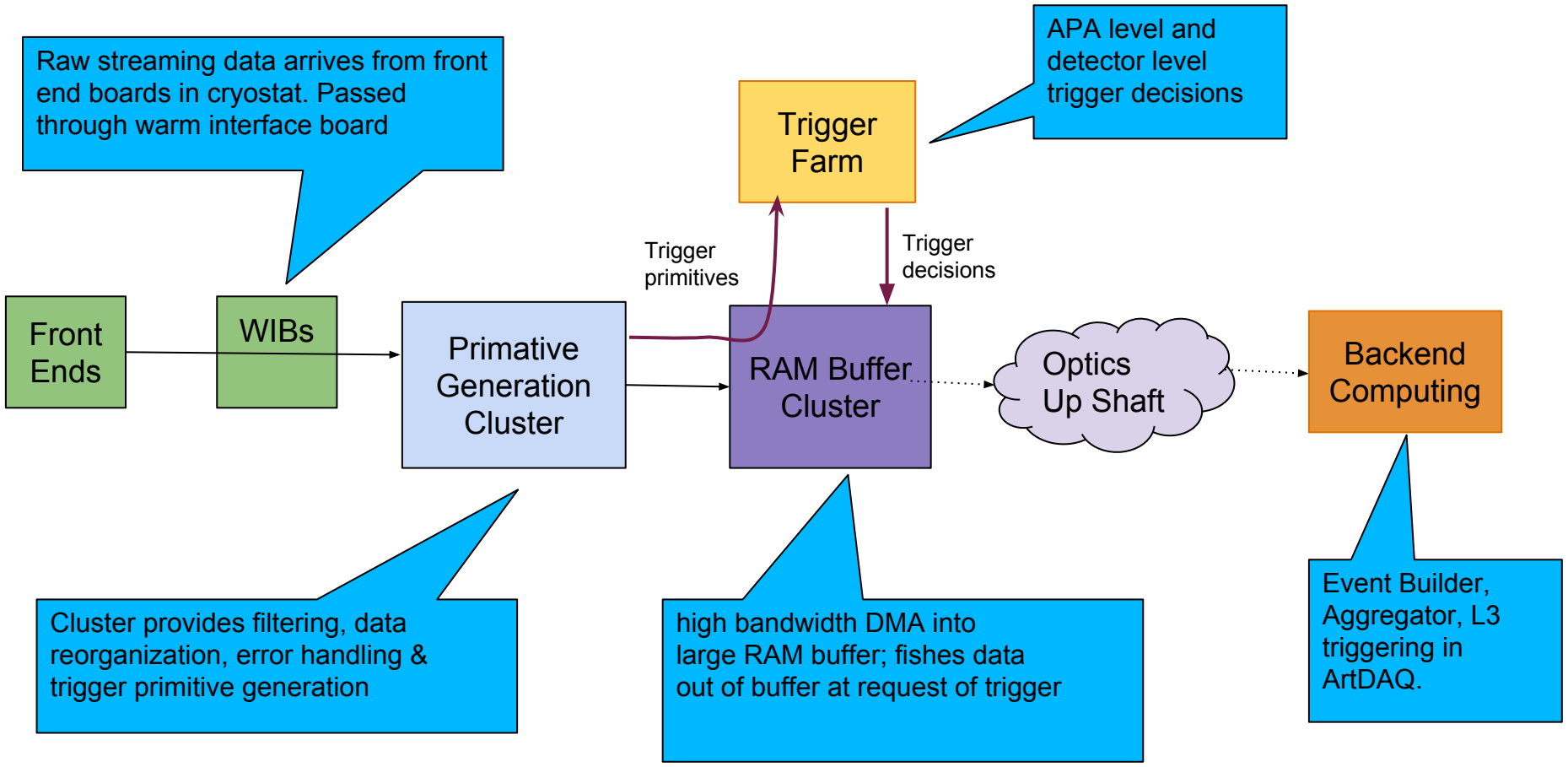
SLAC

"Normal data taking" in one (long) sentence: All data is streamed out of the TPC where it is skimmed for "trigger primatives" (i.e. hits) and buffered; based on those trigger primatives, we decide whether there was an event and if so save ALL of the TPC (150 APAs) for 5.4ms

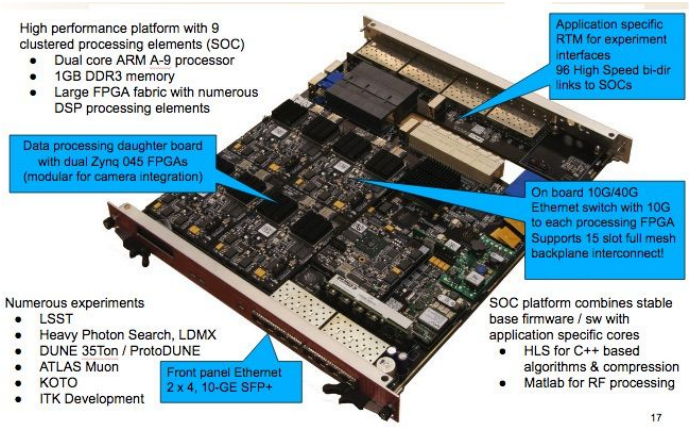
Event Type	Annual Data Volume	Assumptions
Beam interactions	27 TB	800 beam and 800 dirt muons; 10 MeV threshold in coincidence with beam time; include cosmics
Cosmics and atmospherics	10 PB	
Radiologicals	$\leq$1 PB	fake rate of $\leq$100 per year [?]
Front-end calibration	200 TB	Four calibration runs per year, 100 measurements per point
Radioactive source calibration	100 TB	source rate $\leq$10 Hz; single APA readout; lossless readout
Laser calibration	200 TB	1×10^6 total laser pulses, lossy readout
Random triggers	60 TB	45 per day
Trigger primitives	$\leq$6 PB	all three wire planes; 12 bits per primitive word; 4 primitive quantities; ^{39}Ar-dominated

***this does not include photon detector...expected to be x10 smaller rate

DUNE Single Phase Data Flow

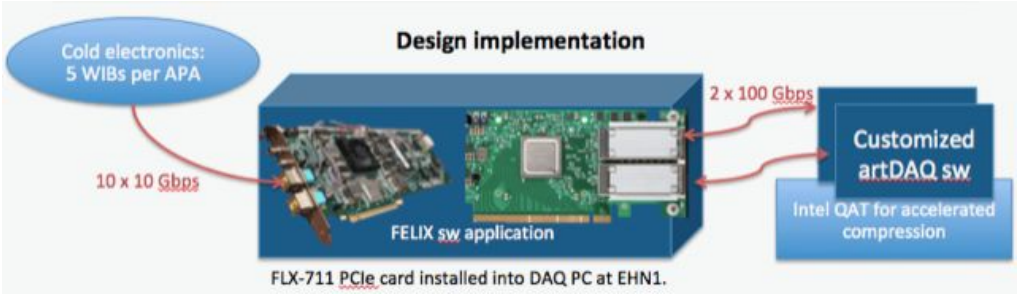


Strawman/Baseline Implementation



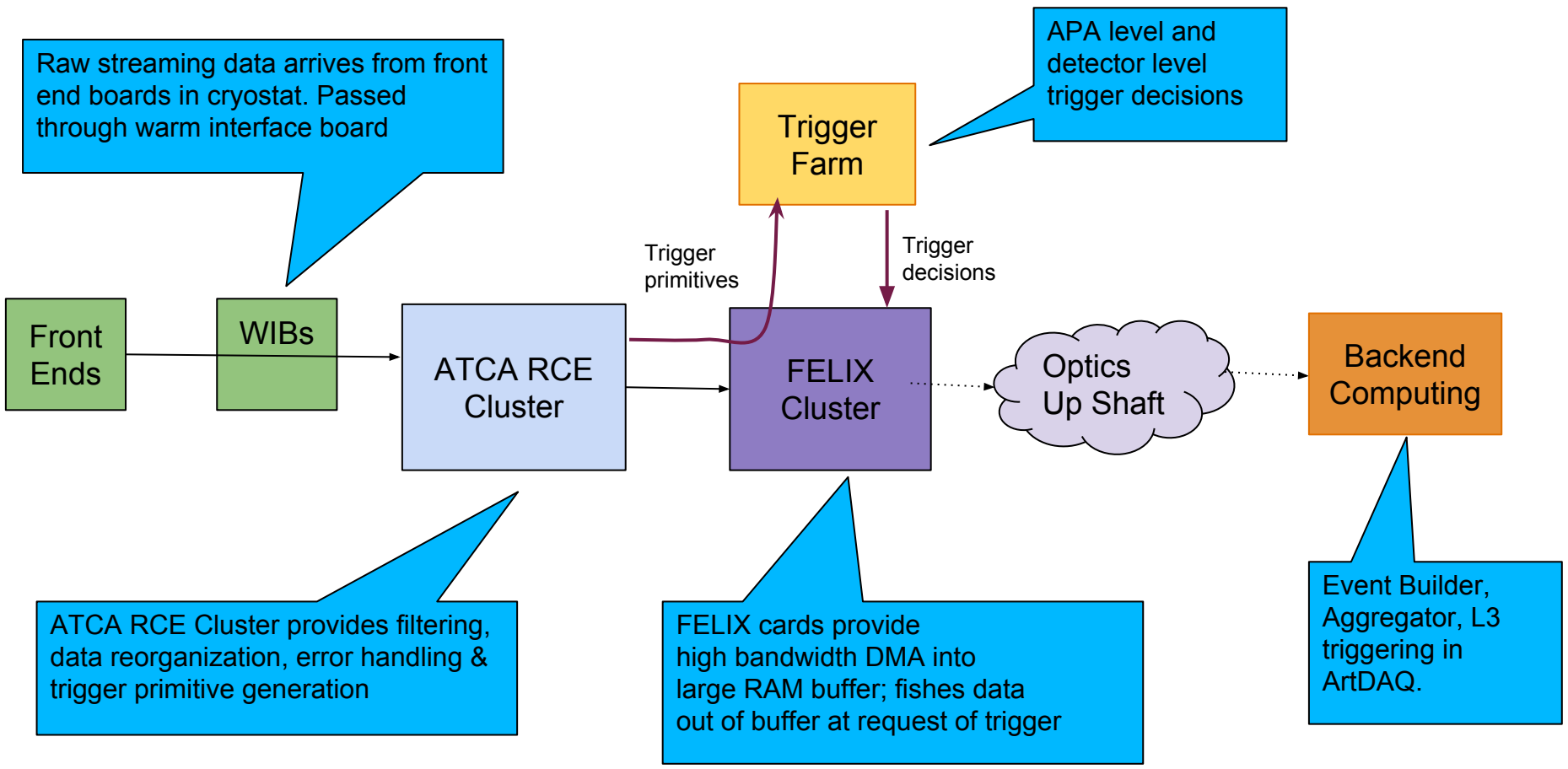
Trigger Primitive Generation: ATCA-based RCE cluster (i.e. FPGA farm)

High-bandwidth RAM Buffer: FELIX cards on commodity PCs

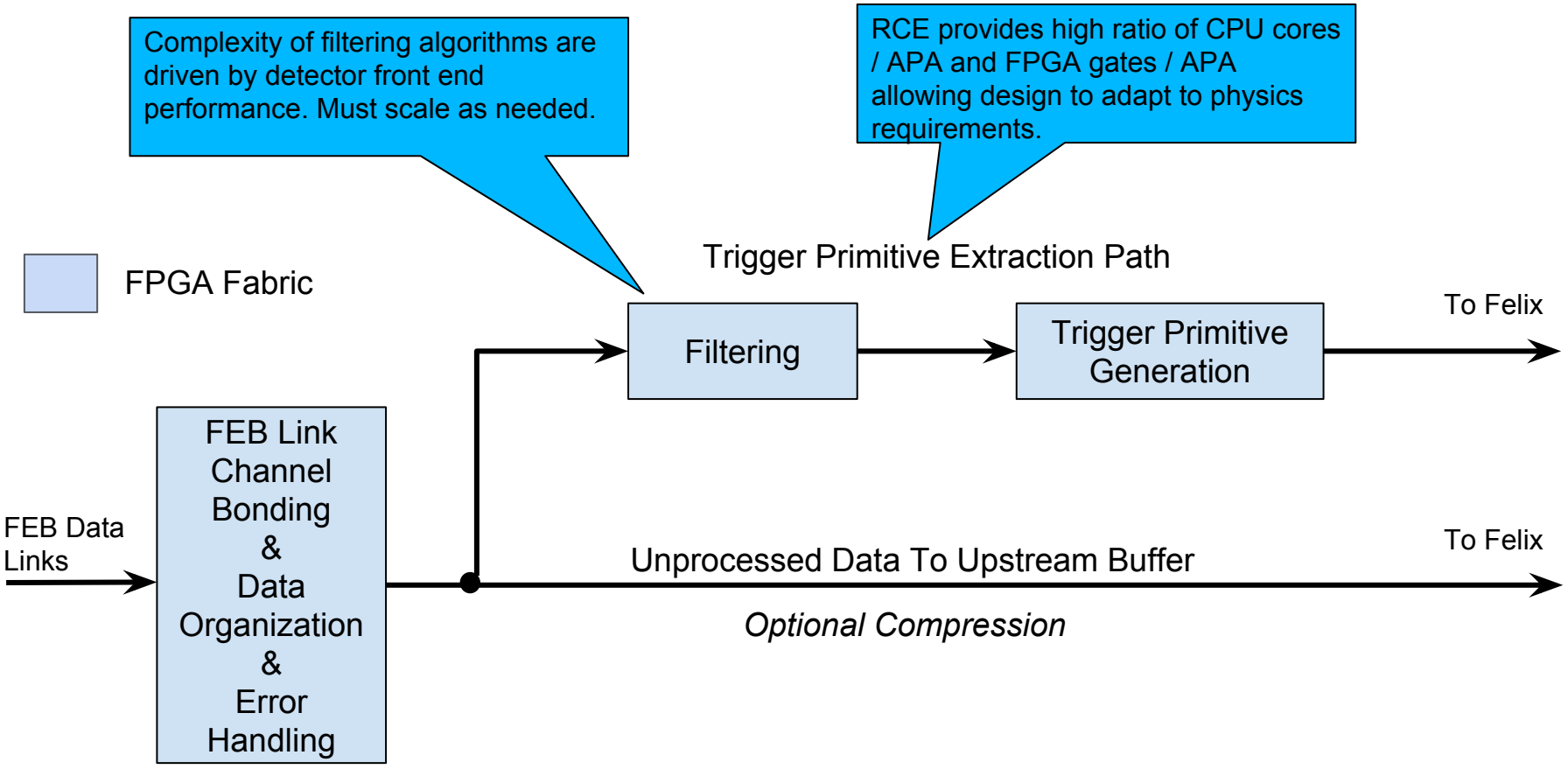


Trigger farm & backend computing: commodity PCs & switchs

DUNE Single Phase Data Flow, again



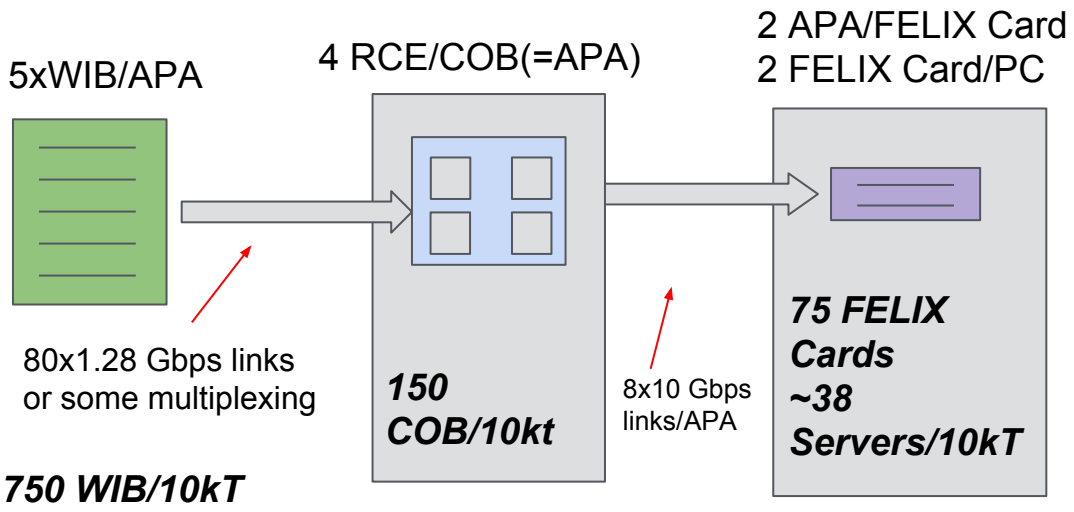
DUNE Data Flow In ATCA RCEs



Baseline data flow in RCE doesn't touch the PS; data is received and sent out through the fabric using the high-speed IO links

The target is to process 640 wire channels/RCE → 1 APA/COB

Yet another data flow diagram



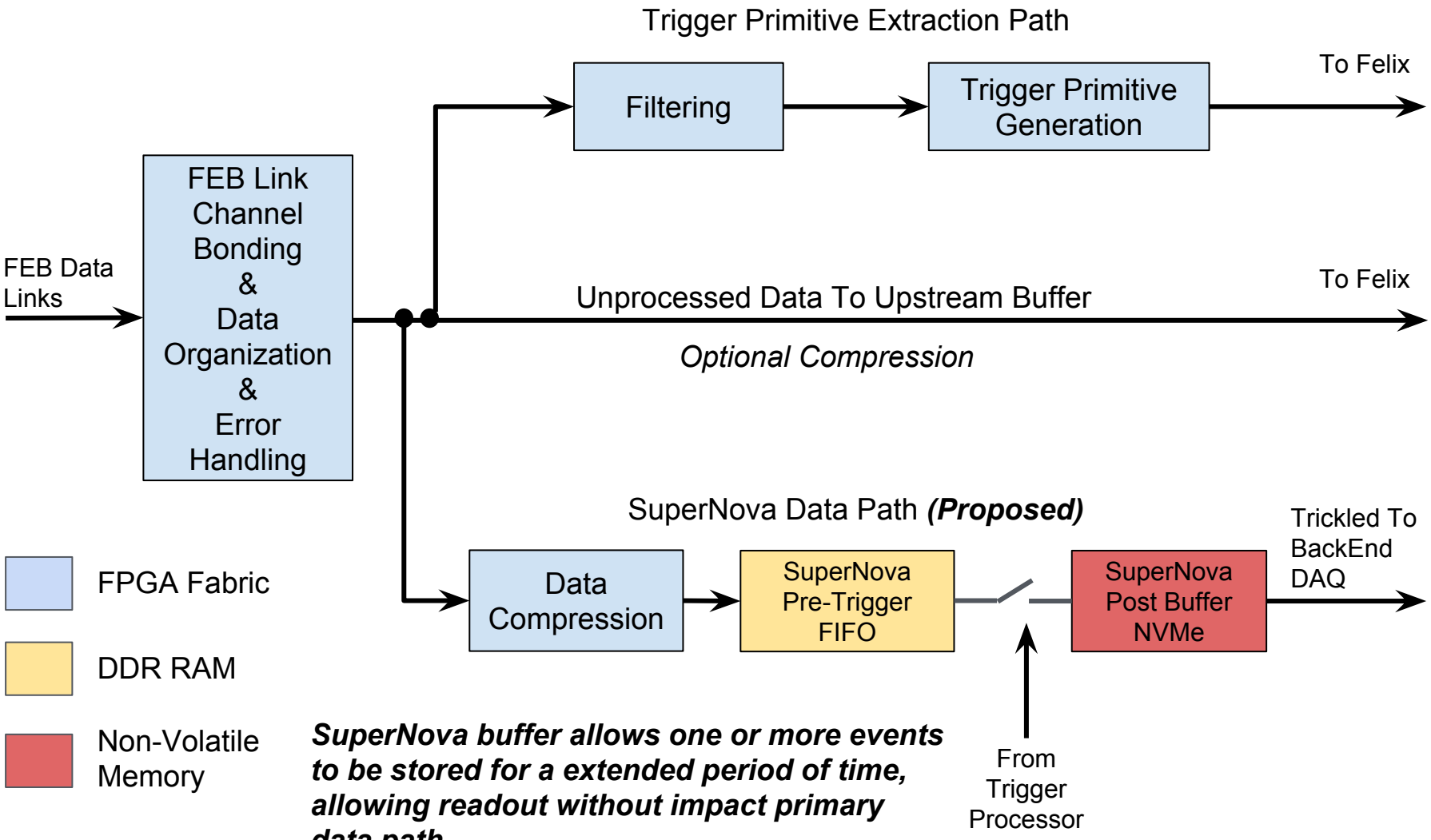
- FE+WIB → RCE: all raw data into RTM with some custom format (e.g. COLDATA); 8B/10B (probably) at 1.28 Gbps
  - Numerology is important! 5 WIBs vs 4 DPMs/APA; multiplexing at WIB (2xFEMB links e.g.) reduces flexibility
- RCE → FELIX: all raw data *out of* the RTM some custom format (GBT etc) of multiplexed data ~10
- FELIX → Backend Computing: triggered raw data over ethernet on switched network
- Trigger Path: RCE-extracted primitives go to RCE → FELIX → trigger farm on separate stream

What about supernova?
(or, if there is “normal data taking”, is there “abnormal data taking”?

Can we save all channels, non-zero-suppressed for X seconds (X → determined by SN group) for:

a) a price that doesn’t blow the budget
b) does not interfere with normal data taking

DUNE Data Flow In ATCA RCEs with SN Buffer



SuperNova buffer allows one or more events to be stored for a extended period of time, allowing readout without impact primary data path

Supernova Buffering In Two Stages (details)

- Pre trigger buffer stores data in a ring buffer waiting for a supernova trigger
  - 640 channels per RCE (1x APA per COB)
  - 2 MHz @ 12-bits ADC sampling rate
  - Raw Bandwidth: 15.36 Gbps (1.92 GB/s)
    - 640 x 2MHz x 12b
  - Each DPM has 16 GB RAM:
    - 9.6 TB DDR4 RAM for all system across 150x COBs
  - Total Memory for supernova “pre-buffering”: 15 GB
    - PL 8 GB + PS 7 GB (1GB for Kernel & OS)
  - Without compression: 7.8 seconds pre-trigger buffer
    - Assuming 12-bit packing to remove 4-bit overhead when packing into bytes

- Post trigger buffer stores data in flash based SSD before backend DAQ
  - Write sequence occurs once per supernova trigger: Low write wearing over experiment lifetime
  - Low bandwidth background readout post trigger: Does not impact normal data taking
  - 512GB/DPM = 266 second post-trigger buffer
  - Samsung NVME SSD 960 PRO: Sequential write up to 2.1GB/s
    - SSD write bandwidth matches well with 640 channels of uncompressed data

NOTE: **Simple, low footprint compression in firmware will expand pre and post buffer storage!**

Questions from DAQ to Calibration Group (from Georgia)

SLAC

1. How many calib. subsystems will we have, and which ones.
2. Radiological calib: How many radiological (EXT) triggers do we need
3. Ar flow calib: how many windows and for how long (EXT) do we need
4. External radiological source and neutron gun; how would they work? how many events and for how long?
5. What is our E>100 MeV visible threshold for triggering? How does this get implemented? Calib will do studies on this. Also they will study what else is in there in cosmogenics; ensure we record it.
6. PD diffusers on cathode, mostly calib during commissioning and shut downs and down times

How long are calib runs? During normal data taking or separate runs? How many events? What is duration of each event? (How long they record info associated with every LED). Read out entire detector or parts of it?

How does triggering work? do we form a trigger on primitives while PD sends the diffuser pulse? Or do we want to read “unbiased” primitives.

7. LASER: How long to scan whole detector?

Do they want to send trigger to DAQ? What data is needed? Crossing tracks?

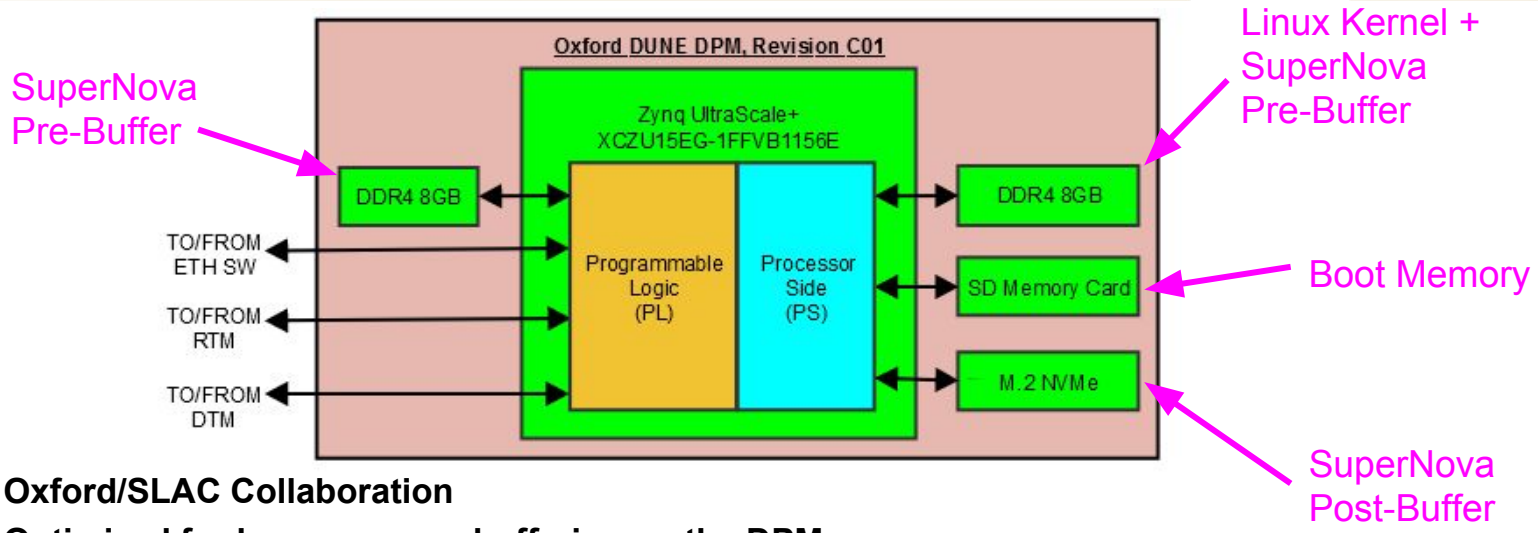
How do we synchronize: associate laser track with the right data from the DAQ? Can we run on the same clock? It would be good to have that.

Take-aways (and more questions)

- Normal data taking will give ~5.4ms snapshots of the ***entire*** detector...so even things we don't trigger on (Ar39) we will see at some accidental rate
  - is that rate enough for the calibration needed?
  - do we need to have a lower threshold trigger that's prescaled?
- It's possible to save a long stretch of full detector (10-100s?), lossless data for a supernova burst trigger for a incrementally small price. We should do this. (we will do this)
  - imagine this could be useful for calibrations, BUT we can't take this sort of data too often or we'll burn out SSDs; also, need to consider the data-to-tape @ FNAL (does this data need to go to FNAL?)

Backup Slides For
Reliability & Value Engineering & Random Details

DPM Redesign for DUNE



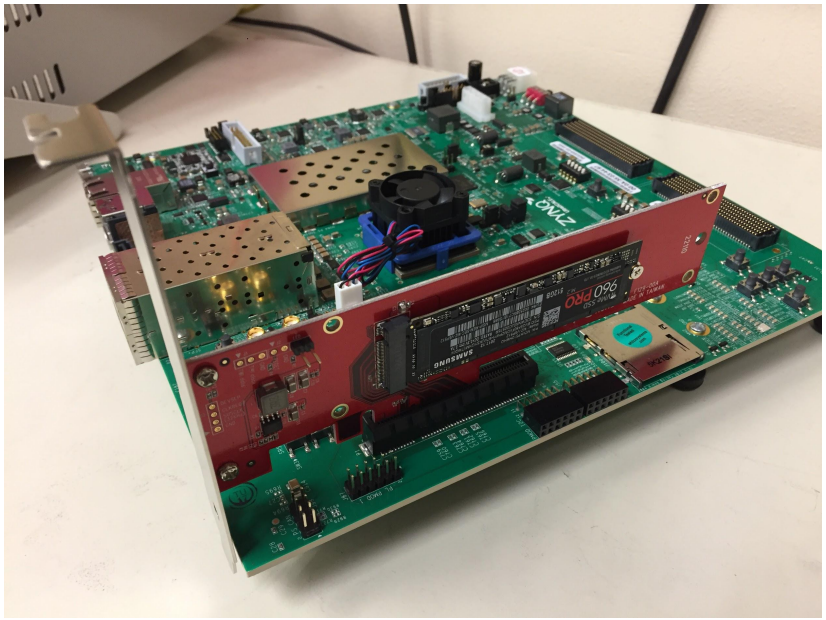
- **Oxford/SLAC Collaboration**
- **Optimized for large memory buffering on the DPM**
- **Only 24 GT channels on this FPGA**
  - **20 of 24 GTs for the FEBs:**
    - **80 links/COB @ 1.28 Gbps (8B/10B)**
  - **2 of 24 GTs for the ETH SW:**
    - **two separate 10 GbE (10Gbps/lane, 64B/66B) to ETH SW**
  - **2 of 24 GTs for the Felix:**
    - **2 RX lanes and up to 22 TX lanes**
      - **Able to support redundant Felix connections**
    - **20 Gb/s @ 2 lanes (10Gbps/lane, 64B/66B)**

Unused FEB TX lanes can be used to increase bandwidth to Felix

Zynq Ultrascale+ and M.2 SDD Performance

SLAC

- Benchmarked read/write bandwidth into Samsung NVMe SSD 960 PRO with the ZYNQ PS PCIe root complex interface
- M.2 SDD mounted and formated as EXT4 hard drive on ArchLinux
- Measuring ~1.6GB/s for read/writing dummy data generated by the CPU
  - Limited by the Zynq's PCIe GEN2 x 4 lane interface (Theoretical limit: 2.0Gb/s)
    - Not limited by M.2 SDD's controller
- Because the input bandwidth is 1.92GB/s > 1.6 GB/s SDD write speed, we would be able to buffer for 37 seconds in DDR before 100% back pressure
- Need small amount (20%) of compression before the SSD to prevent bottlenecking at the SDD
  - Low footprint compression approaches (lookup table) are being investigated
- This is a very simple test with only one process
  - Need to do stress testing of other interfaces in parallel of SDD to confirm rate is still 1.6GB/s

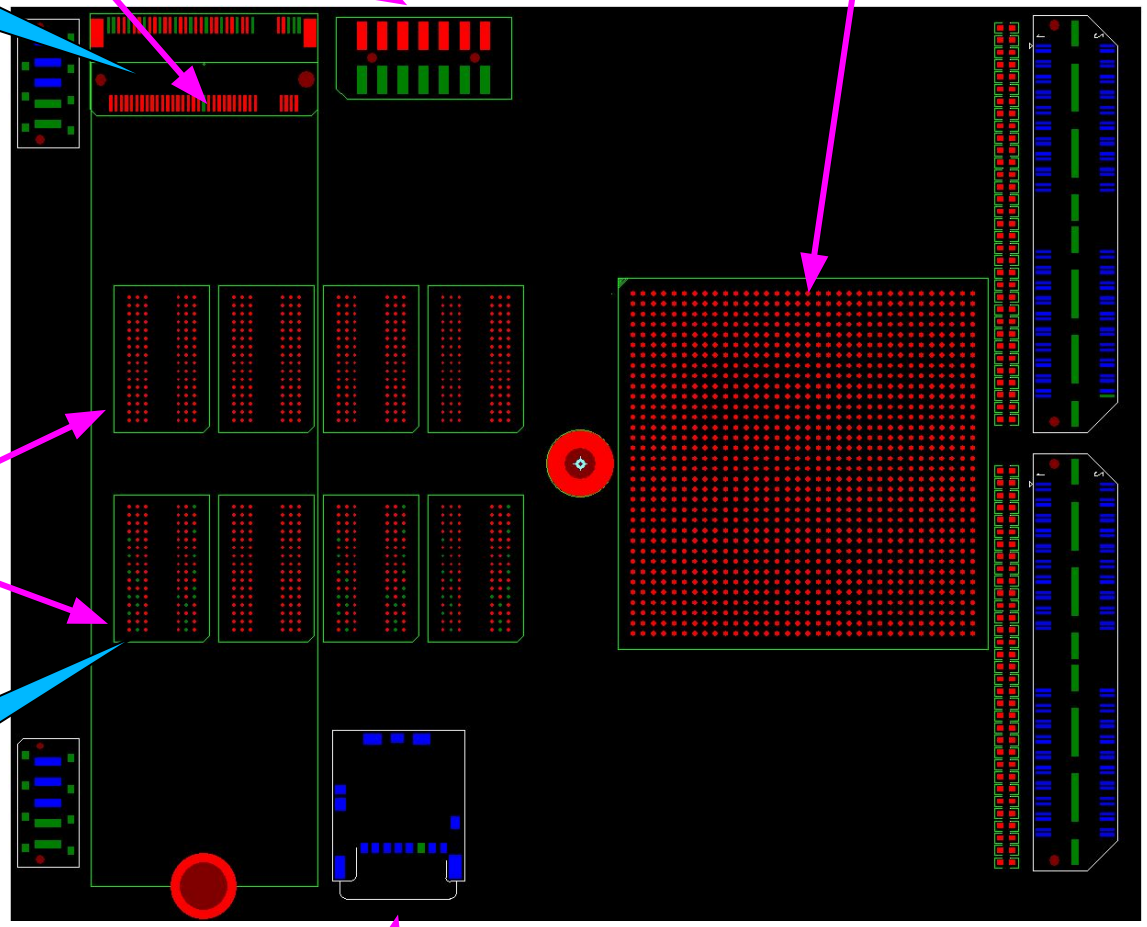


Oxford Design: Revision C01

SLAC

Distributed NVRAM modules provides good match for data path rate in each RCE.

- **ZYNQ: XCZU15EG-1FFVB1156E**
- **PL DDR4: 8 GB on DPM**
- **PS DDR4 8 GB on DPM**
- **M.2 NMVe: 512 GB on DPM**
  - **Located above the DPM's DDR ICs**
- **Dimensions: 85.09 mm x 110 mm**
  - **Increased by 1.27mm for NMVe**



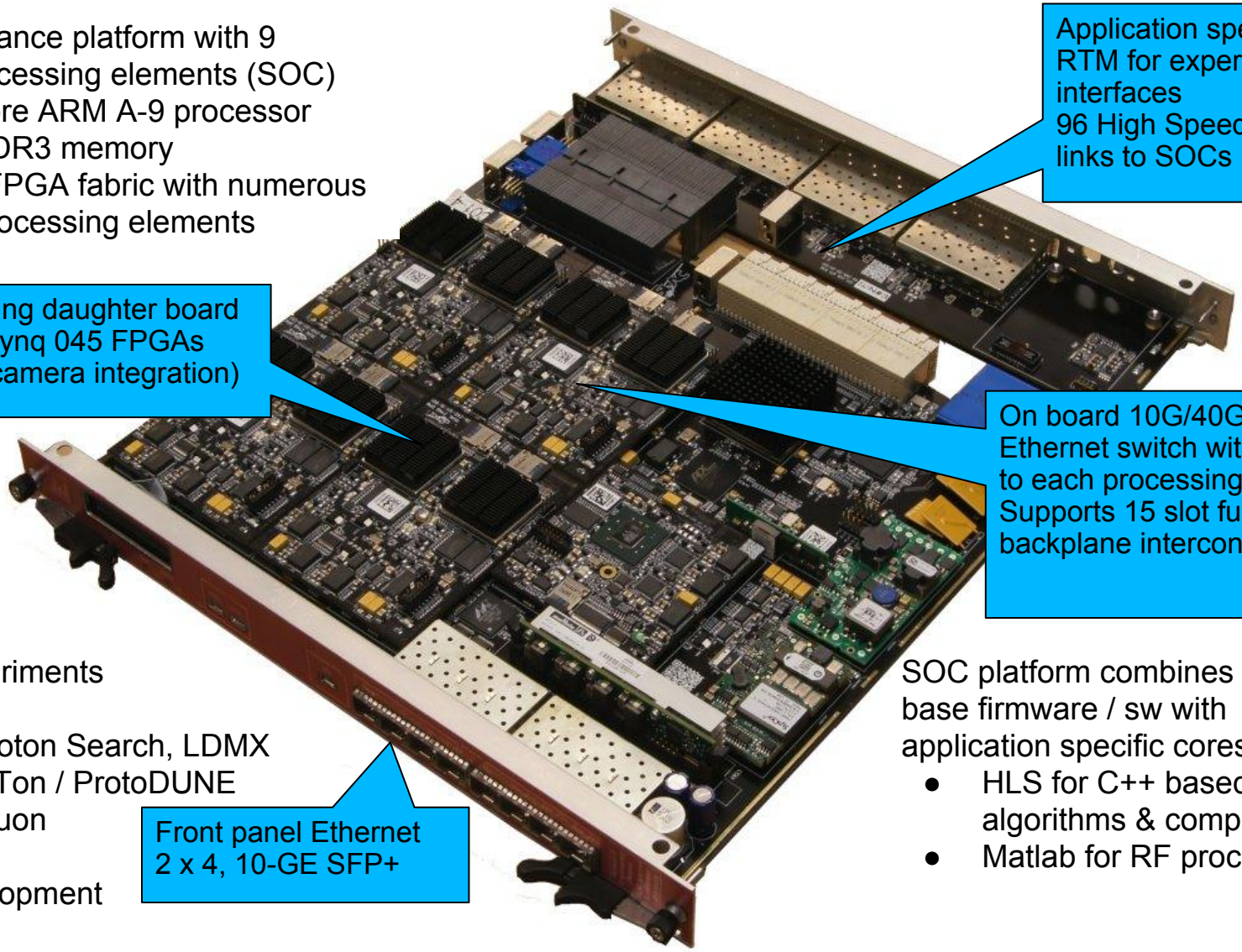
Inexpensive DDR memory provides pre-trigger buffering. Already in place to support PS/PL operation. (minimum size of 64-bit memory if)

The RCE Platform
High Density / High Performance Processing In ATCA

TID-AIR SLAC

High performance platform with 9 clustered processing elements (SOC)

- Dual core ARM A-9 processor
- 1GB DDR3 memory
- Large FPGA fabric with numerous DSP processing elements



Application specific RTM for experiment interfaces
96 High Speed bi-dir links to SOCs

Data processing daughter board with dual Zynq 045 FPGAs (modular for camera integration)

On board 10G/40G Ethernet switch with 10G to each processing FPGA
Supports 15 slot full mesh backplane interconnect!

Numerous experiments

- LSST
- Heavy Photon Search, LDMX
- DUNE 35Ton / ProtoDUNE
- ATLAS Muon
- KOTO
- ITK Development

Front panel Ethernet
2 x 4, 10-GE SFP+

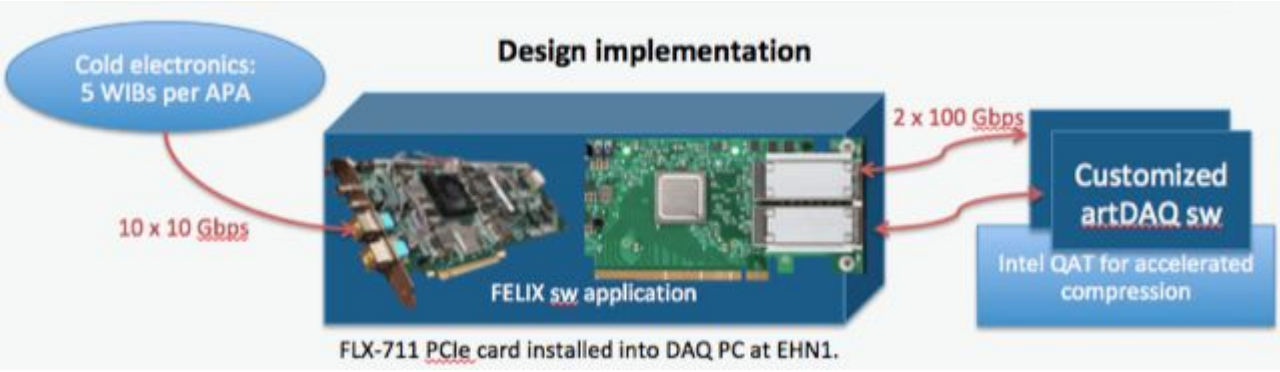
SOC platform combines stable base firmware / sw with application specific cores

- HLS for C++ based algorithms & compression
- Matlab for RF processing

The Felix Platform

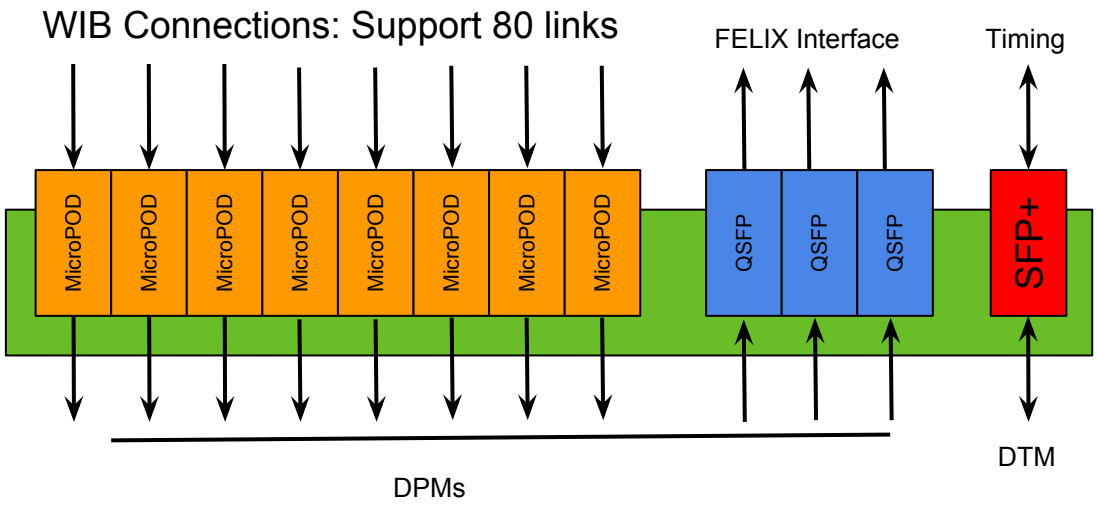
FLX-711 from BNL

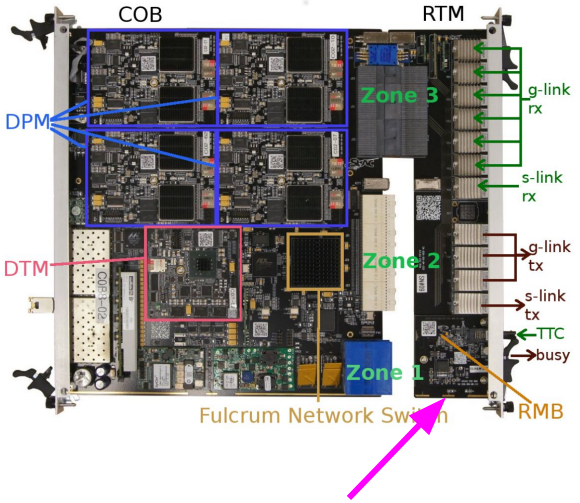
- → FELIX phase 1 prototype
- → TTC input ADN2814 + Si5338
- → Xilinx Kintex **Ultrascale** XCKU115
- → 48 duplex optical links (based on MiniPODs)
- → PCIe Gen3 x16



FLX-711 PCIe card installed into DAQ PC at EHN1.

RTM Block Diagram





Broadcom MicroPOD transmitter/receiver supports 10.3125 Gbps per lane:
http://www.fit-foxconn.com/Images/Products/Spec/AFBR-77D1SZ_20160510175052121.pdf

Experience with high density fiber optic RTMs

This is simply an example showing it's feasible to fit the connectors onto an ATCA RTM...not tied to these specific parts