Top drift electronics data transmission discussion

D. Autiero
IP2I Lyon

2/8/2021

Full details on the Top Drift charge readout electronics and the digitization system can be found in the Top Drift Electronics CDR review slides:
https://indico.cern.ch/event/1038740/

In this presentation:

- **Some reminders about how data transmission is organized in NP02/Protodune-DP for the same electronics**

→ **NP02 firmware more complicated than the one for continuous streaming foreseen for DUNE**

a) **since it includes the handling of external triggers.**
b) **In particular the firmware version including Huffman lossless compression is complex and resources consuming on the AMC cards (independent compressors implemented for each of the 64 channels of the AMC card)**

The LV1 event builders in NP02 are the machines hosting the network cards receiving the optical links from the uTCA crates and putting together data from all the crates connencted to the machine

- **Discussion of assumptions for possible VD operation mode. Data transmission similar to what being used in NP02 but with a simplified firmware (continuous streaming, no external trigger, no data compression)**

10:45 → 11:00 **Executive Session** 15m
Speakers: Marzio Nessi (CERN) , Steve Herbert Kettell (Brookhaven National Laboratory (US))

11:00 → 11:30 **Top Electronics Overview and requirements** 30m

11:30 → 11:50 **Top Electronics chimneys** 20m
Speaker: Fabien Cavalier (IJCLab)

11:50 → 12:05 **Top analog electronics** 15m
Speaker: Dario Autiero (Centre National de la Recherche Scientifique (FR))

12:05 → 12:20 **Top digital electronics** 15m
Speaker: Vyacheslav Galymov (Centre National de la Recherche Scientifique (FR))

12:20 → 12:35 **CRP adaptor boards** 15m
Speaker: Bo Yu (Brookhaven National Laboratory (US))

12:35 → 12:50 **CRP cabling** 15m
Speaker: Dominique Duchesneau (Centre National de la Recherche Scientifique (FR))

12:50 → 13:05 **Timing distribution** 15m
Speaker: Dario Autiero (Centre National de la Recherche Scientifique (FR))

13:05 → 13:20 **Cold-box tests** 15m
Speaker: Elisabetta Maria Pennacchio (Centre National de la Recherche Scientifique (FR))

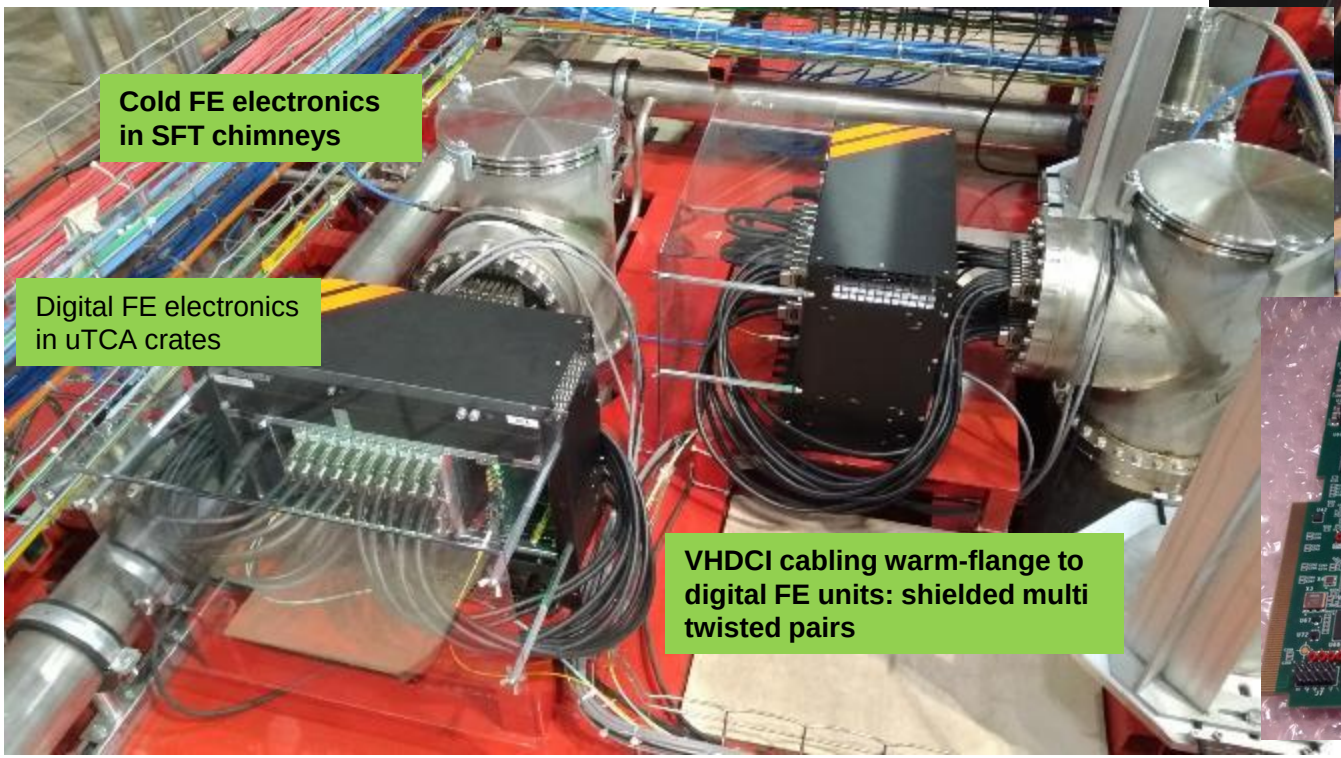
13:20 → 13:35 **Production and QC** 15m
Speaker: Vyacheslav Galymov (Centre National de la Recherche Scientifique (FR))

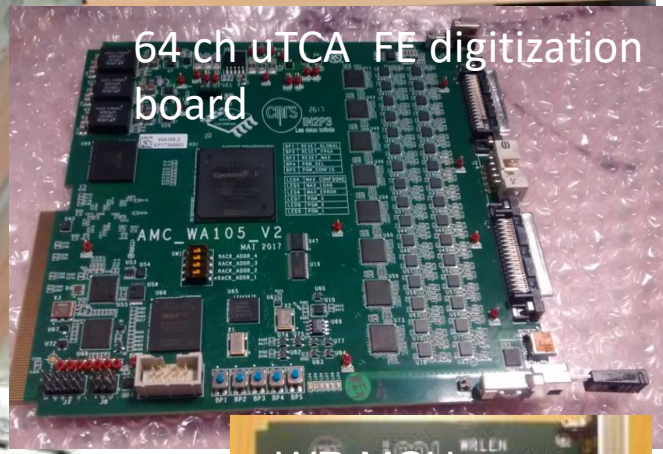
13:35 → 13:50 **Installation** 15m
Speaker: Takuya Hasegawa

13:50 → 14:00 **Summary** 10m
Speakers: Dario Autiero (Centre National de la Recherche Scientifique (FR)) , Takuya Hasegawa

14:00 → 14:30 **Executive Session** 30m
Speakers: Marzio Nessi (CERN) , Steve Herbert Kettell (Brookhaven National Laboratory (US))

ProtoDUNE-DP FE electronics

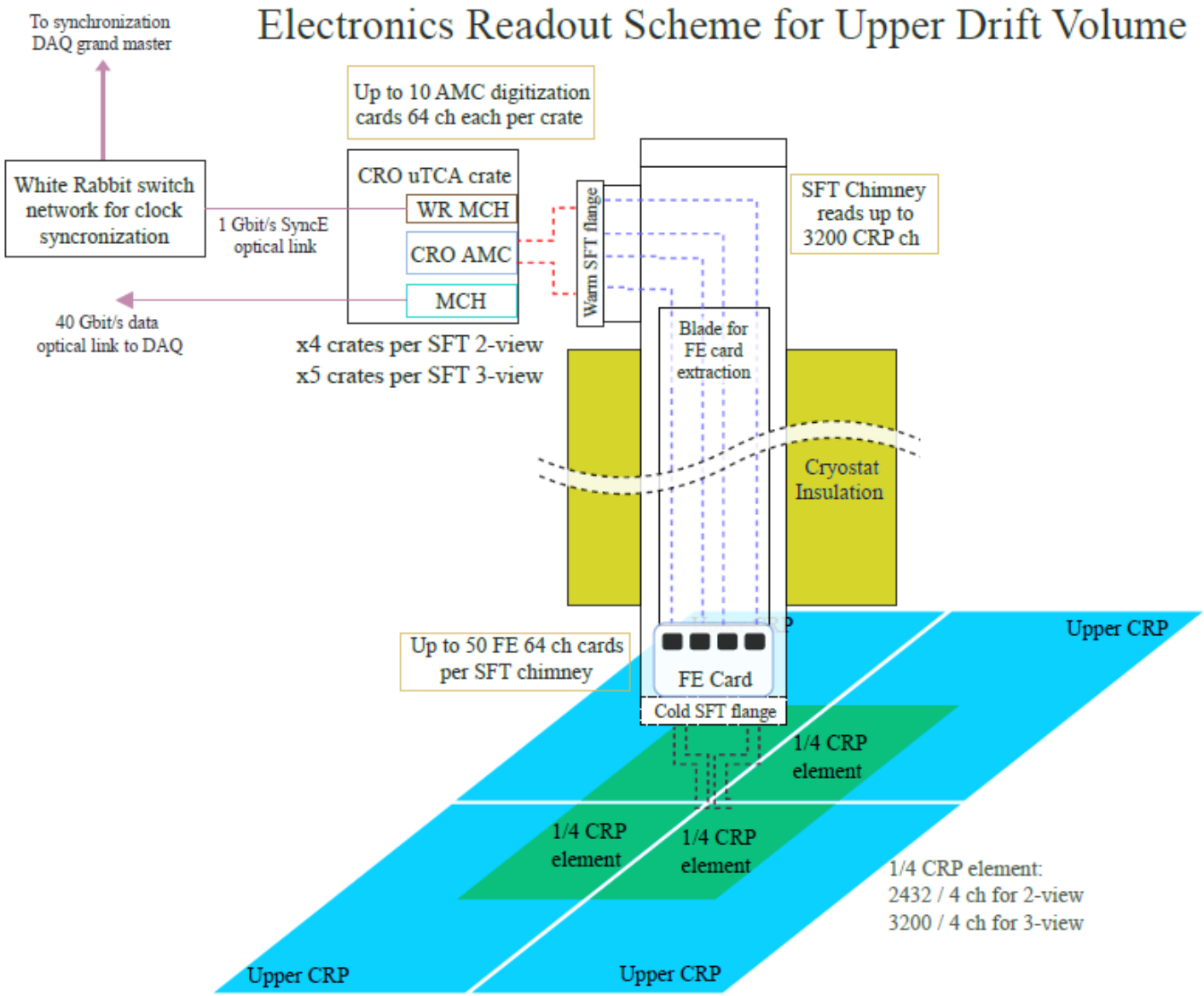




White-Rabbit timing slave node in each uTCA crate →

- 64 channels modularity for both analog and digital FE
- Electronics noise ~600 electrons
- uTCA crates containing 10 (up to 12) FE cards
- 10 Gbit/s (now 40 Gbit/s) connectivity of each uTCA crate
- Reference design for DP far detector module TDR
- Components produced for protoDUNE-DP → 1/20 of a DP FD module

TOP drift readout: general layout



uTCA AMC digitizer cards (See AMCs presentation at the CDR review)

- FPGA Cyclone V with NIOS virtual processor, ADC AD9257, **64 channels per card sampled at 2.5 MHz (DP), up to 10 Gbit/s Ethernet data flow per card, for vertical drift can reduce sampling to 2MHz (real sampling up to 40MHz)**

- **Time sync at ns level, external triggers handling via White Rabbit network (Dedicated WR slave node in uTCA crate and timing/clock dedicated lines on backplane). Transmission of external triggers timestamps on WR network.**

Working mode in protoDUNE-DP (based on external triggers):

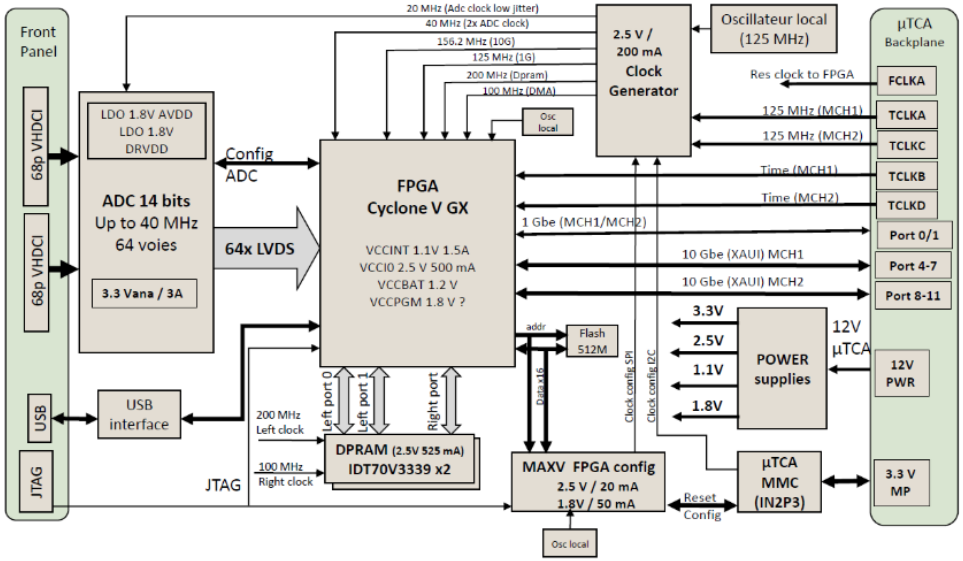
a) No compression mode saturating the 10 Gbit link

b) Lossless compression (Optimized Huffman, up to factor 10 gain), developed, exploited in NP02 in August 2020

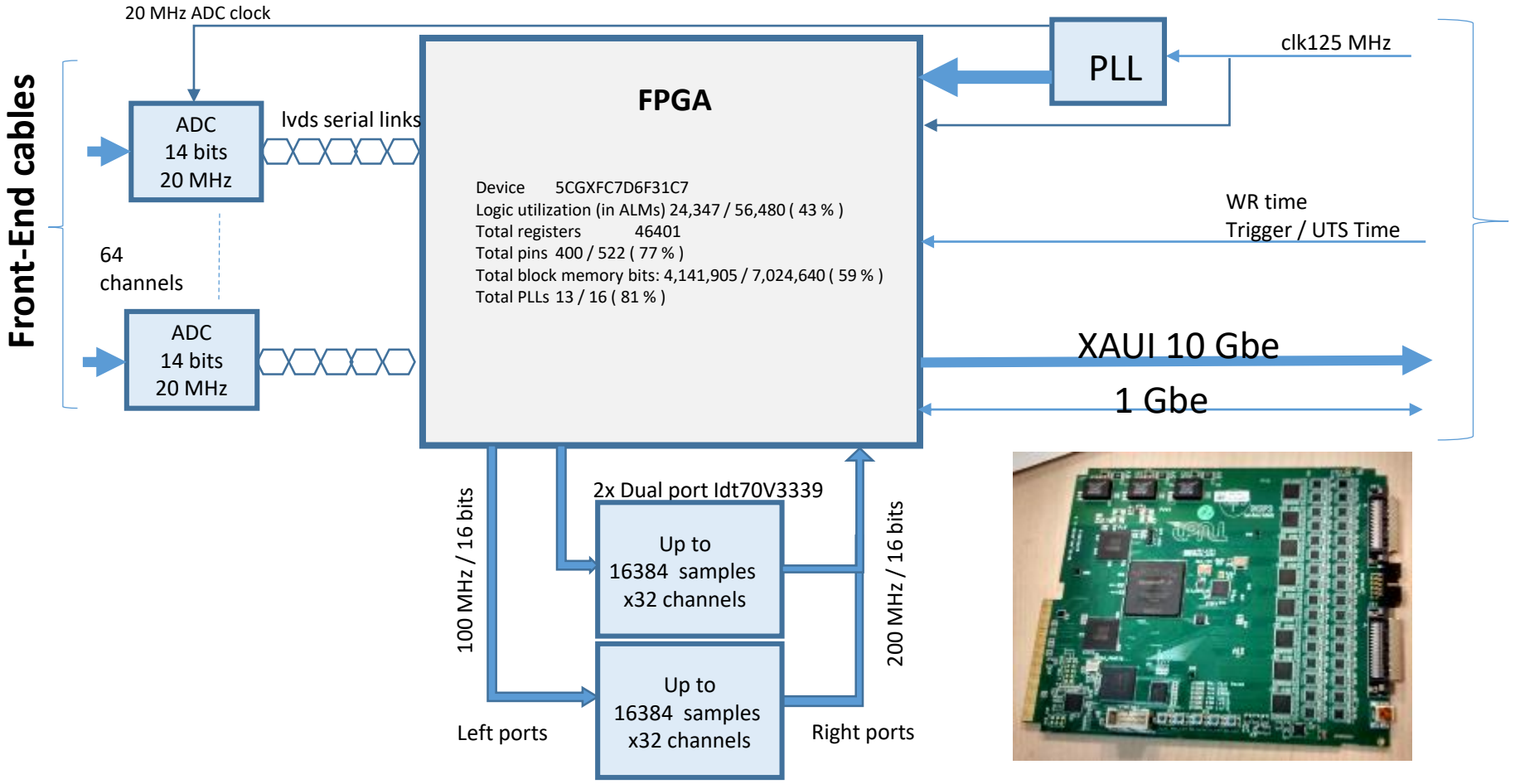
Working mode in DUNE:

Originally foreseen with 10 Gbit uTCA MCH: Continuous streaming + lossless Huffman compression (10 Gbit/s MCH)

New baseline (40 Gbit MCH): Continuous streaming, no compression (40 Gbit/s MCH), possibility for trigger primitives

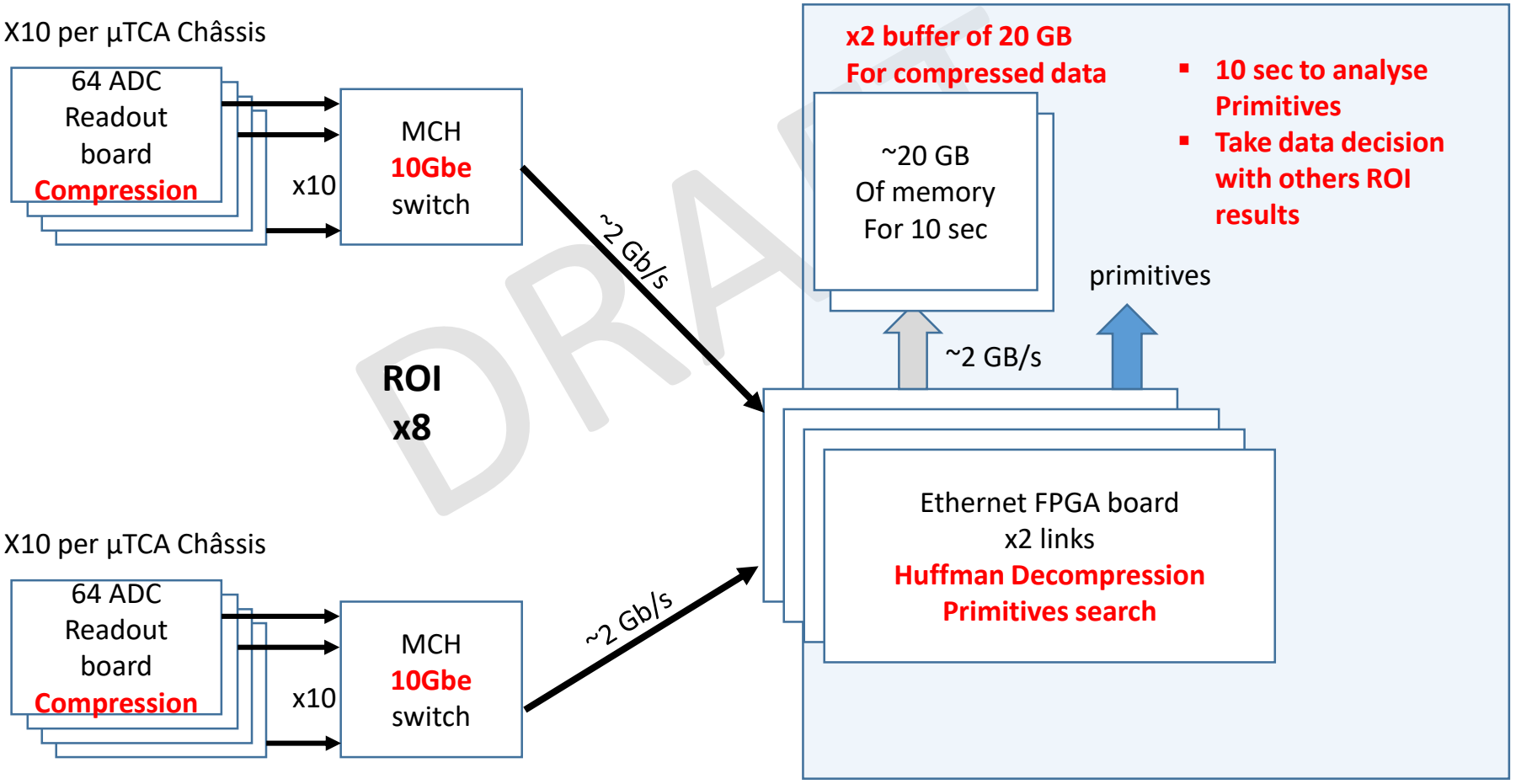


PROTO DUNE ADC front-end board simplified bloc diagram

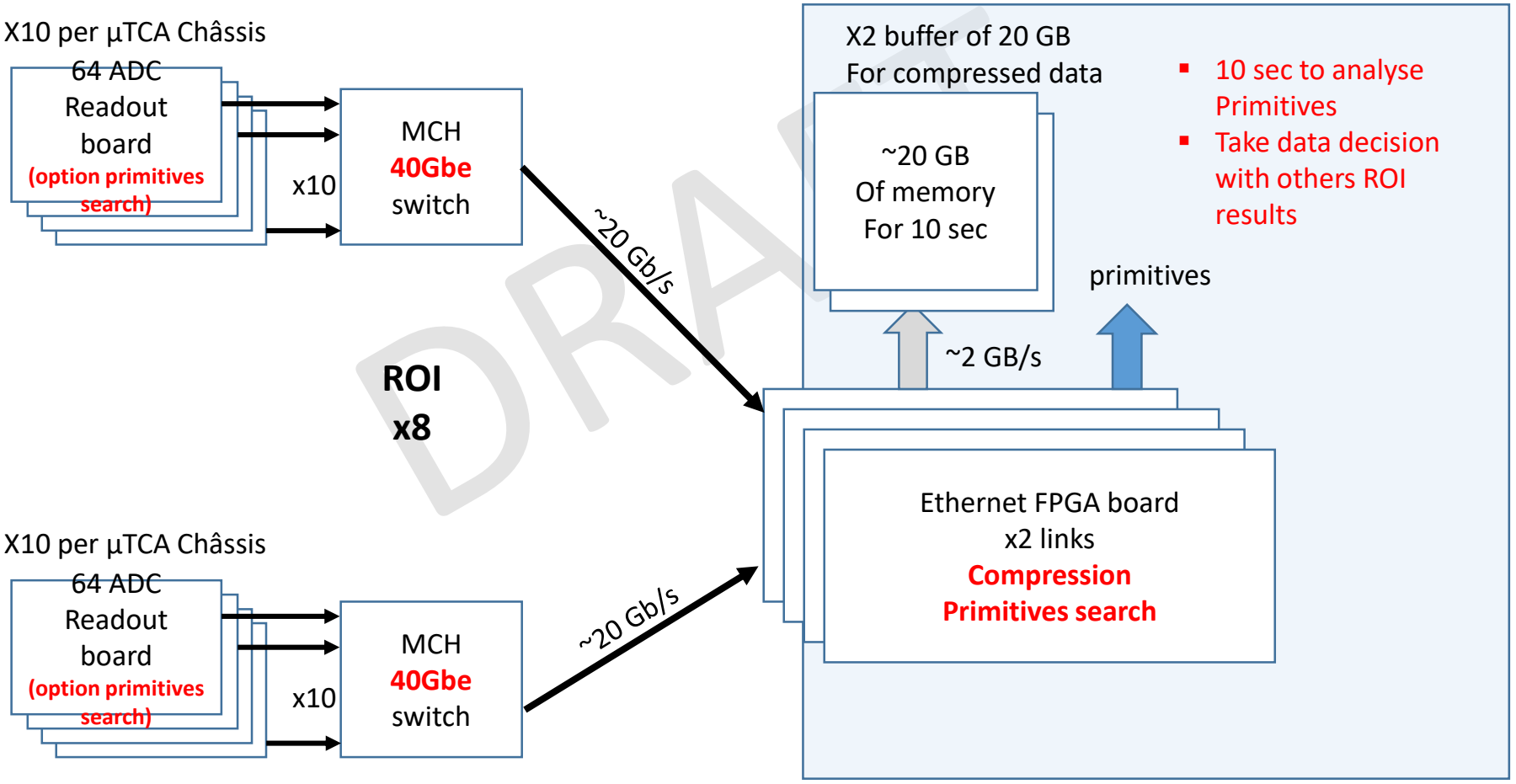


Readout architecture in continuous streaming with 10Gb/s MCH links and hardware compression/decompression.

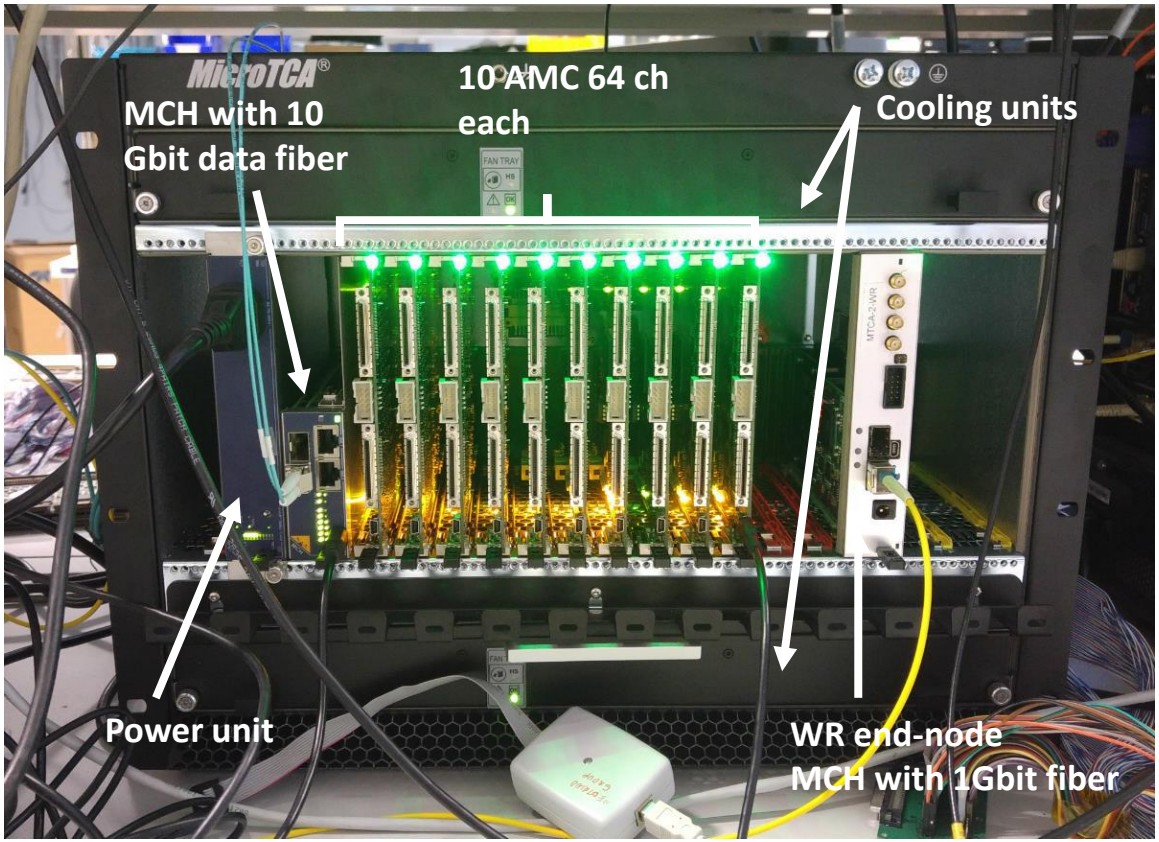
(Original Baseline for DP 10kton module TDR)



Readout architecture with no compression and 40Gb/s MCH links
(New uTCA connectivity Baseline)



uTCA crates (See AMC dedicated presentation at the CDR review)

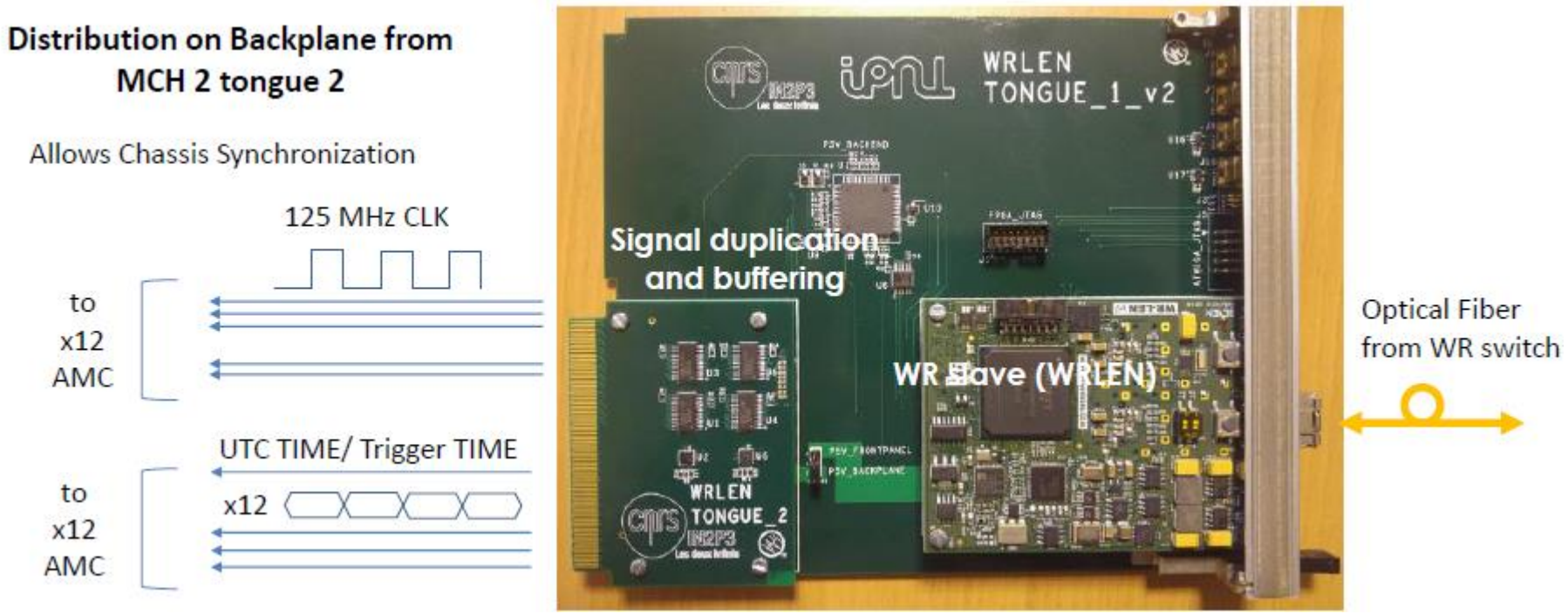


Readout system with 10 AMC cards (640 channels) and 10Gbit MCH (like in NP02)

x5 40 Gb systems deployed for Cold-box tests:
x3 systems in 2021 for shared top-bottom single CRP
x5 systems in 2022 for full top CRP tests

Same uTCA system with 40Gbit MCH

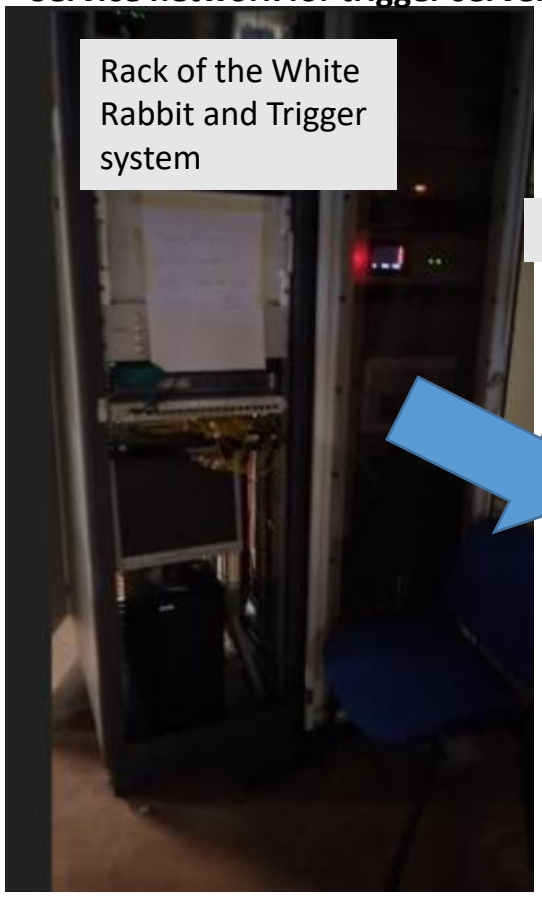
WR-MCH (See interface dedicated presentation at the CDR review)

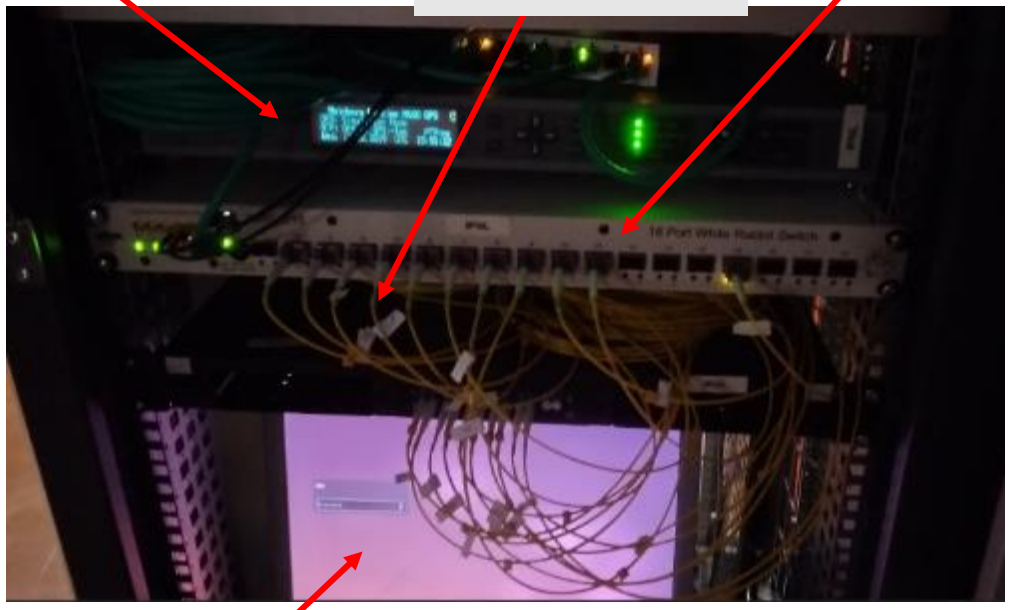


- Simple board on which the WLREN (commercial WR end-node) is plugged in
- Occupies the second MCH slot (12) and provides power to the WRLEN via standard uTCA facilities
- Delivers via the backplane a pair of WRclock (125 MHz) and WR_DATA for sync to each AMC
- Control signals timing signals available on front-face connectors
- WR network can be used also to transmit the WR time-stamped trigger data to all digitizing units (trigger mode in NP02)

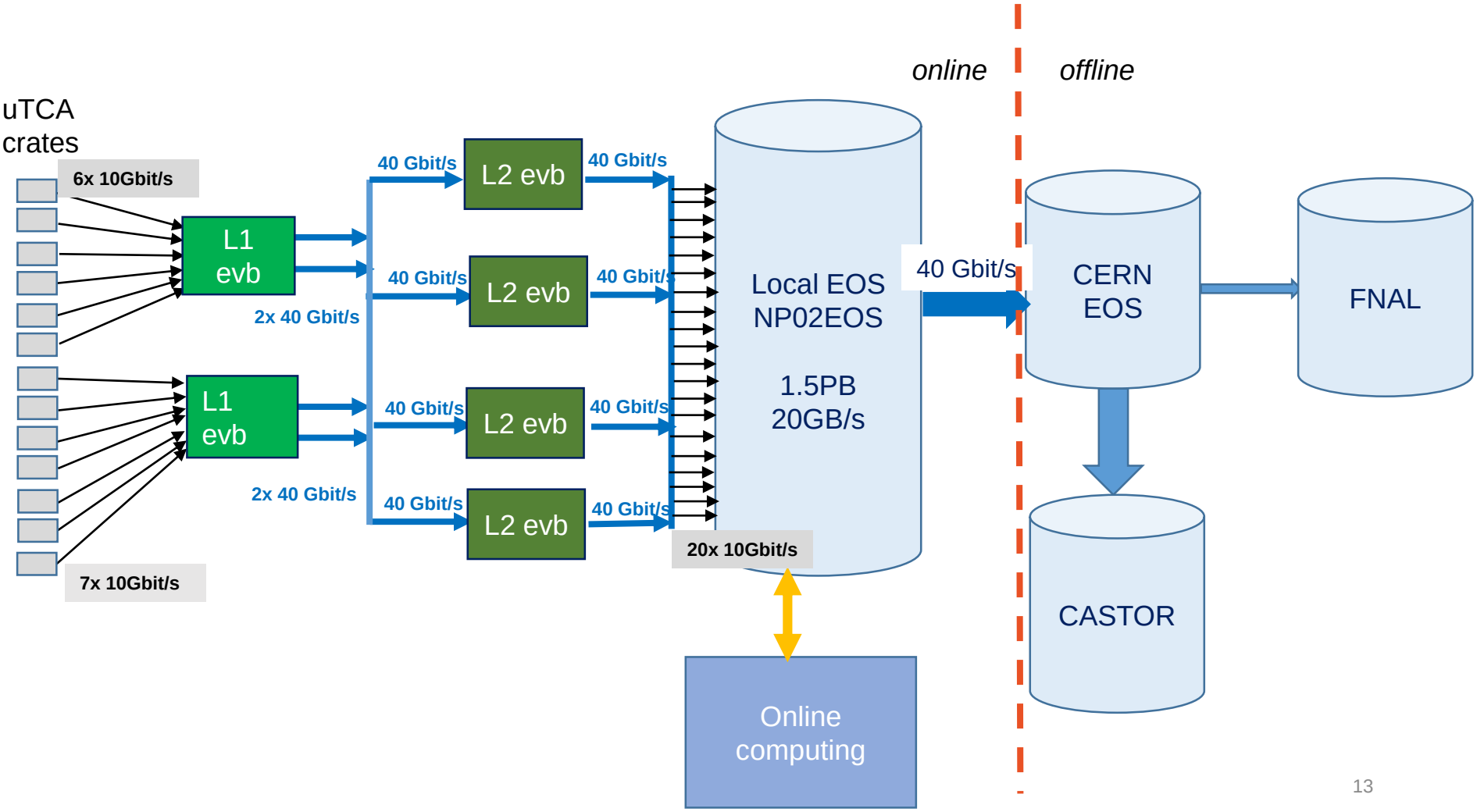
ProtoDUNE-DP Timing System (similar to the system which was operating on the 3x1x1):

- **GPSDO GPS disciplined oscillator (generates 1PPS, 10 MHz, NTP timing)**
- **White Rabbit Grand Master (connects to slave node in the uTCA crates and to the timestamping card in the trigger server**
- **Trigger server with WR FMC-DIO for external triggers time-stamping (Light, Cosmic Counters, Beam, Calibration), new machine, new network interfaces and switches**
- **Private fast trigger network to the two LV1 event builders (dedicated fiber + switch)**
- **Service network for trigger server, GPSDO and WR Grand Master with local switch and cable to DAQ room**





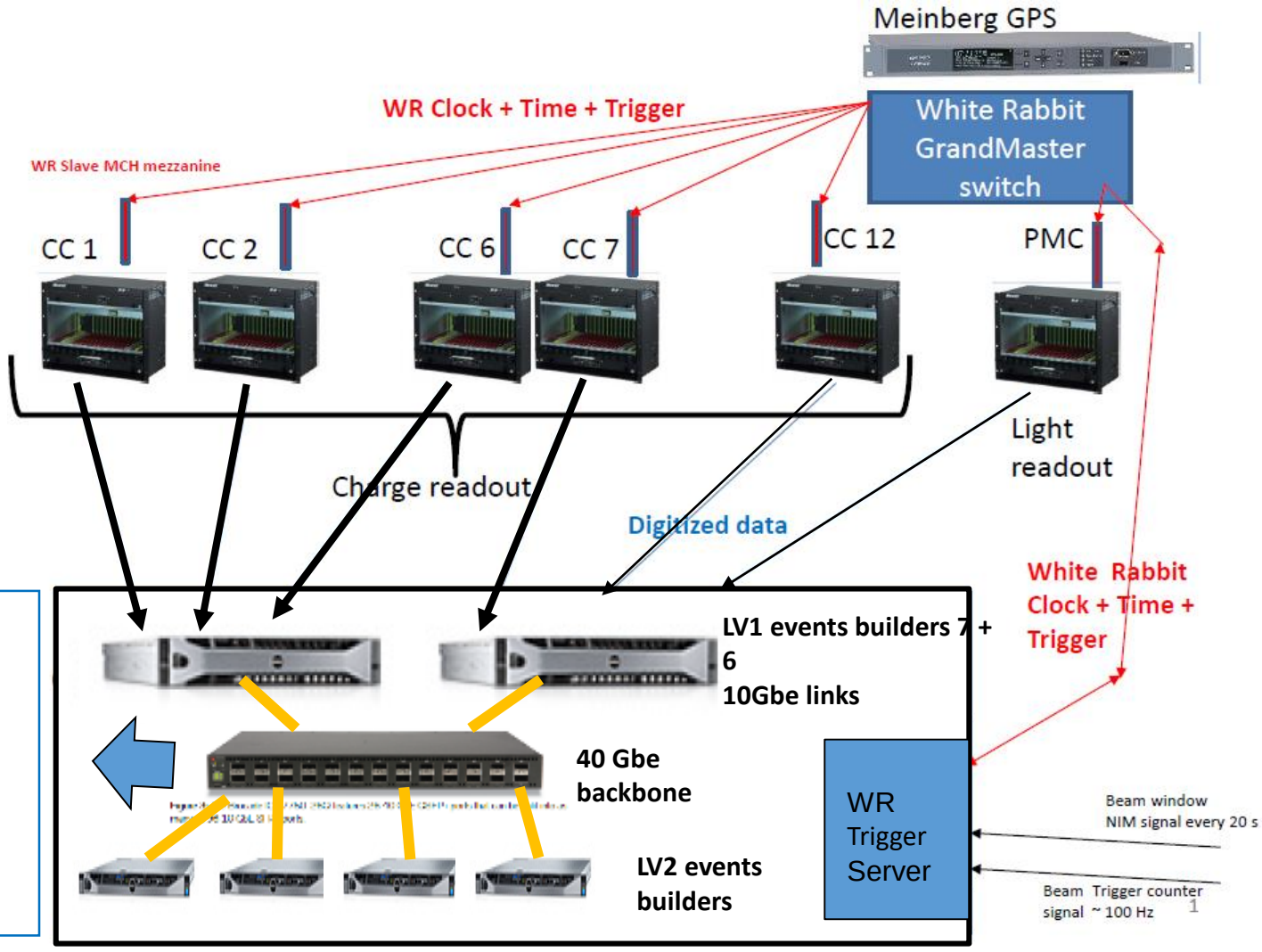
NP02 DAQ/network infrastructure



13

Global uTCA DAQ architecture for ProtoDUNE-DP

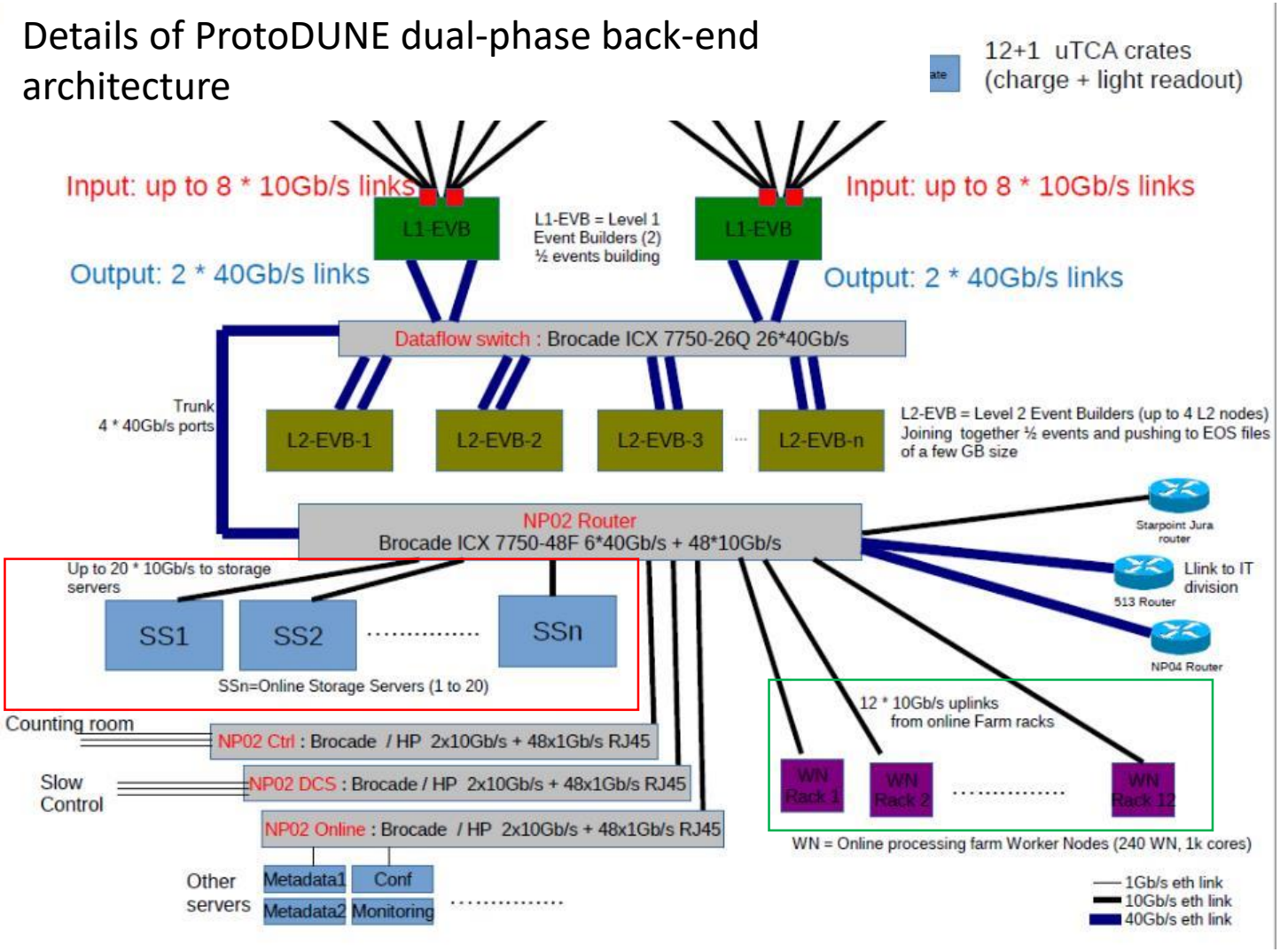
DAQ front-end integrated with **White Rabbit (WR) Time and Trigger distribution network:**
White Rabbit slaves MCH nodes in uTCA crates + WR distribution system (GPSDO time source, WR Grand Master switch, trigger time-stamping system)



Online Storage and processing Farm

20 GB/s disk bandwidth

Details of ProtoDUNE dual-phase back-end architecture



DAQ back-end equipment in the DAQ room (support for 4 active CRPs readout):

- **High bandwidth (20GByte/s) distributed EOS file system for the online storage facility**

→ Storage servers: 20 machines + 5 spares (DELL R510, 72 TB per machine): up to 1.44 PB total disk space for 20 machines, 10 Gbit/s connectivity for each storage server.

- **Online storage and processing facility network architecture:**

→ Backend network infrastructure 40 Gbit/s DAQ switch (Brocade ICX7750-26Q) + 40/10 Gbit/s router (Brocade ICX 7750-48F)

→ Dedicated 10 Gbit multi-fibers network to uTCa crates

→ Dedicated trigger network (x2 LV1 event builders + trigger server)

→ x2 40 Gbit/s link to IT division

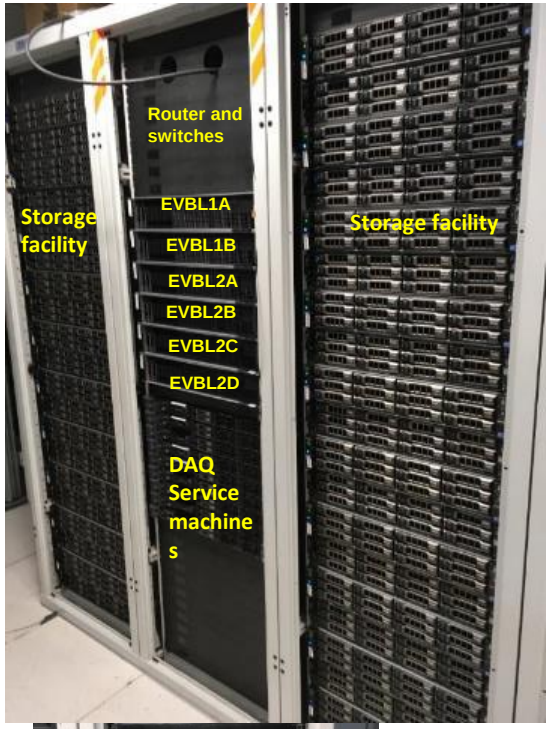
- **DAQ cluster and event builders:**

→ DAQ back-end: 2 LV1 event builders (DELL R730 384 GB RAM) + 4 LV2 event builders (DELL R730 192 GB RAM)

→ DAQ cluster service machines: 9 Poweredge R610 service units: 2 EOS metadata servers, configuration server, online processing server, batch management server, control server, …

- **Online computing farm (room above the DAQ room):**

→ 40 servers Poweredge C6200 (450 cores)



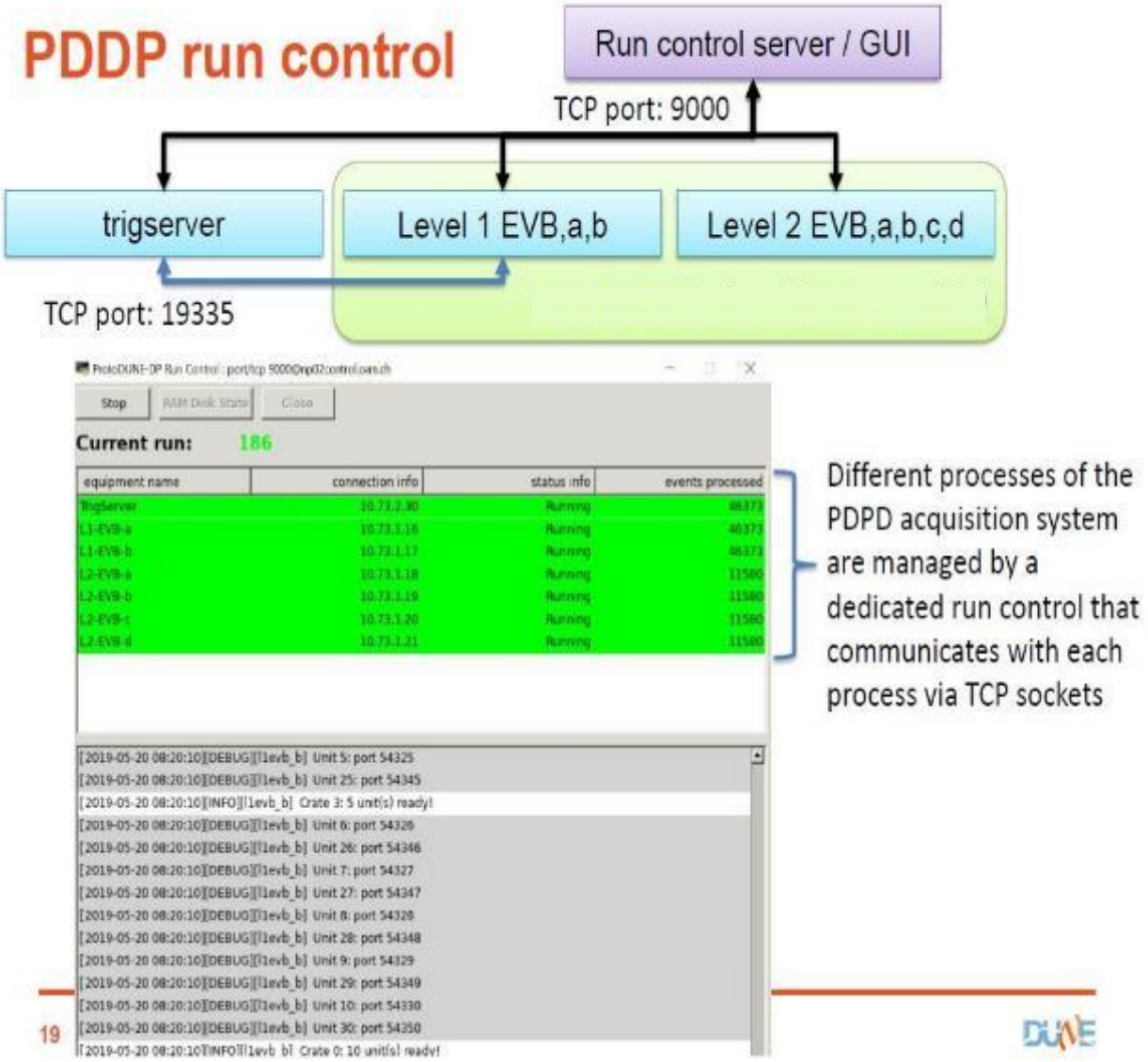
Brief reminder of the NP02 DAQ system architecture:

The NP02 DAQ system has been presented in details numerous times in the four last years at the DUNE and ProtoDUNE-DP meetings. It is a Ethernet network based DAQ system which can acquire data at very high bandwidth (up to 20 GB/s). **The front-end digitization units** (AMCs) are contained in uTCA crates. Each charge readout crate, located in front of the corresponding signal feedthrough chimney on the cryostat roof can include up to 10 AMC reading each 64 channels for a total of 640 channels per crate digitized at 2.5 MHz. A uTCA crate is a 10 Gbit/s network system connecting the AMC with its own switch included in the crate controller (MCH). The MCH of each crate is connected with a dedicated 10 Gbit/s to a **Level 1 (L1) event builder**. Two L1 event builders are used to read several crates corresponding to a detector half. The L1 event builders are connected via several links on a high speed network at 40 Gbit/s to the Level 2 (L2) event builders and to a high bandwidth distributed storage system (EOS), the network infrastructure ensures total 20 GB/s bandwidth. Each L1 event builder puts together the data, corresponding to the drift window starting with the trigger timestamp, acquired from the connected crates to build an event half on its ramdisk. The L1 ramdisks are visible via the network to the **L2 event builders** who assemble together the two event halves in the final even format and assemble on their own ramdisks the events in 3 GB files which are then pushed through the EOS high bandwidth storage system which can absorb up to 20 GB/s data writing on disk. Four L2 event builders work in parallel by sharing evenly the events produced by the L1 event builders and producing the final 3 GB data files to be written on disk. **Automatic file transfer systems** transfer the data from the local EOS system to the CERN IT division and Fermilab. **A trigger server** handles the white-rabbit timestamping of external trigger signals (beam counters, cosmic counters, PMTs trigger, calibration triggers) and the transmission of these timestamps to the AMCs via the white-rabbit network and of the trigger information to the L1 event builders via a dedicated Ethernet network. **The white-rabbit network** ensures also the timing and synchronization of the AMC digitization units. **The run control interface** ensures the control and monitoring of all the components of the system in order to start and stop runs and transfer the data to the local EOS and to the final storage for offline exploitation.

More detailed information can be retrieved from the slides shown for instance at the May 2019 DUNE collaboration meeting:

1) Digital front-end: https://indico.fnal.gov/event/18681/session/7/contribution/149/material/slides/0.pdf
2) Analog front-end and timing/trigger system: https://indico.fnal.gov/event/18681/session/7/contribution/150/material/slides/0.pptx
3) Back-end and online storage/computing: https://indico.fnal.gov/event/18681/session/7/contribution/151/material/slides/0.pdf

- See also the NP02 shifters DAQ documentation: https://twiki.cern.ch/twiki/pub/CENF/DUNEProtDPOps/DAQforshifter_v2r2.pdf



Developed online software:

1) Data acquisition processes (LARGUI)
running on the LV1 even builders from the front-end DAQ AMC in the uTCA crates

2) Event building and EOS data writing software

3) Run control software

4) Software for the management and synchronization of the different components of the back-end system and online computing

→ **Stable operation for several months without problems**

Data transmission packets:
(current implementation in trigger mode in NP02, drift window of 10k samples)

- Each AMC card has individual connectivity at 10Gb to the uTCA crate backplane and it needs to transmit the data of 64 channels (10,000 samples taken at 2.5 MSPS for NP02)

- Using JUMBO frames : MTU is 9000

- 3 packets per channel are required in order to transmit the 10,000 samples of each channel in no compression mode
  - Two long packets of 8006 B and one short packet of 4006 B, where 6 B is the header
  - Currently 12 bit ADC data are sent as 16 bit words: 8000 B (4000 ADC samples) + 8000 B (4000 ADC samples) + 4000 B (2000 ADC samples) → 20 kB for 10,000 samples

- The actual packet sizes have 42 B overhead for standard headers (UDP header + Eth header)
  - So 2x 8048 B + 1x 4048 B per channel

Data packets header

- Each data packet contains a 6B header:
  - 12 bit keyword to signal data packet (== 0xDEF)
  - 6 bit counter for total number of packets per channel (== 3)
  - 6 bit counter for current packet number (1,2,3)
  - 6 bit counter for channel to which this packet belongs (0 … 63) ← in total each card has 64 channel
  - 16 bit total packet counter: goes up to 1920 per crate with 10 AMCs

- The counters in the packets headers are used by L1 event builder to detect packet losses and sort data into relevant container for each crate/card/ch
  - A unique UDP port is assigned for Crate / Card so L1 evb knows whose AMC data these are
  - AMCs also know Crate No. and its own Slot No and have individual ip addresses set on the basis of the Crate and Slot no

  → so there is no technical roadblock this info cannot be part of the header

UDP ports assignments

Ports on even builder	Ports on each AMC card	
EB IP: port num	**Card IP: port num**	**Function**
4660		Multicast port
65000	65000	Configuration packets using TFTP protocol
	64000	UDP data request packet
54321		Port for card #1 to send event data
…	…	…
54341		Port for card #20 to send event data

Assigned at init stage

Each AMC card is told which UDP port to send data by the L1 event builder

Proto DUNE UDP DATA packet Format

IP/ UDP Header + Data Header
48 bytes
(64 bits aligned)

Source IP[2] = 32+ChassisNum
Source IP[3]= AMC slot + 12
Source MAC[4] = chassisNum+1
Source MAC[5] = AMC slot + 12

UDP Lenght	UDP lenght	0xD (4 bits)	0xE (4 bits)	0xF (4 bits)	Total Packet Number (6 bits)	Current Packet Number (6 bits)	ADCchannel (6 bits)	Global Packet Counter (MSB)	Global Packet Counter(LSB)

ADCs or Encoded Huffman Sequence
(10000 samples for protoDUNE)

No compression => 2 packets of 8000 bytes (4000 samples)
+ 1 packet of 4000 bytes (2000 samples)

Compression => 1 packets of 8048 bytes max
Size of Huffman Sequence is the UDP length – 6 bytes

L1 EVB/AMCs Communication protocol in NP02 (external trigger mode)

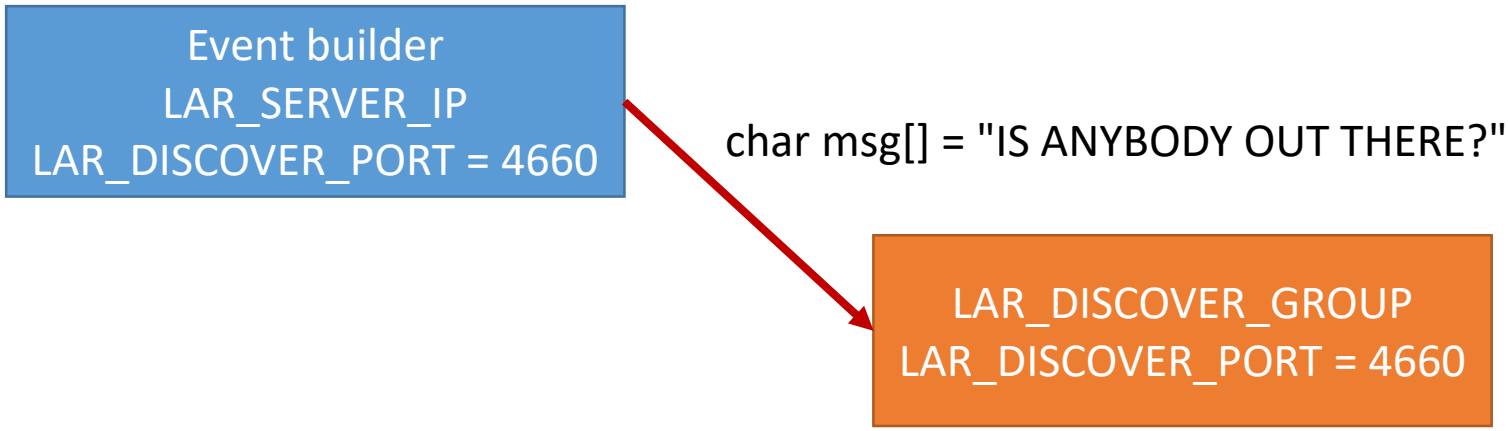
L1 EVB ←→ AMC Cards communication steps:

1. Multicast broadcast (AMCs discover)
2. AMCs initialization using TFTP protocol
3. Event acquisition cycle

- **AMC cards configured via TFTP protocol:**
  - Given to the AMCs the L1 event builder port numbers to send data to
  - Can also ask the AMC card for its status and configure delays for packet transmission pipelining
- **During a data acquisition cycle:**
  - When run starts / stops L1 event builder notifies AMCs accordingly
  - Trigger packet from WR trigger server received by L1 event builder causes it to send data requests to AMCs (REQ)
  - AMCs answer conforming they have data (ACK) and start sending event data to L1 event builder
  - The transmission is assumed terminated when the L1 event builder receives the last expected data packet according to the packet counter contained in AMC packet headers

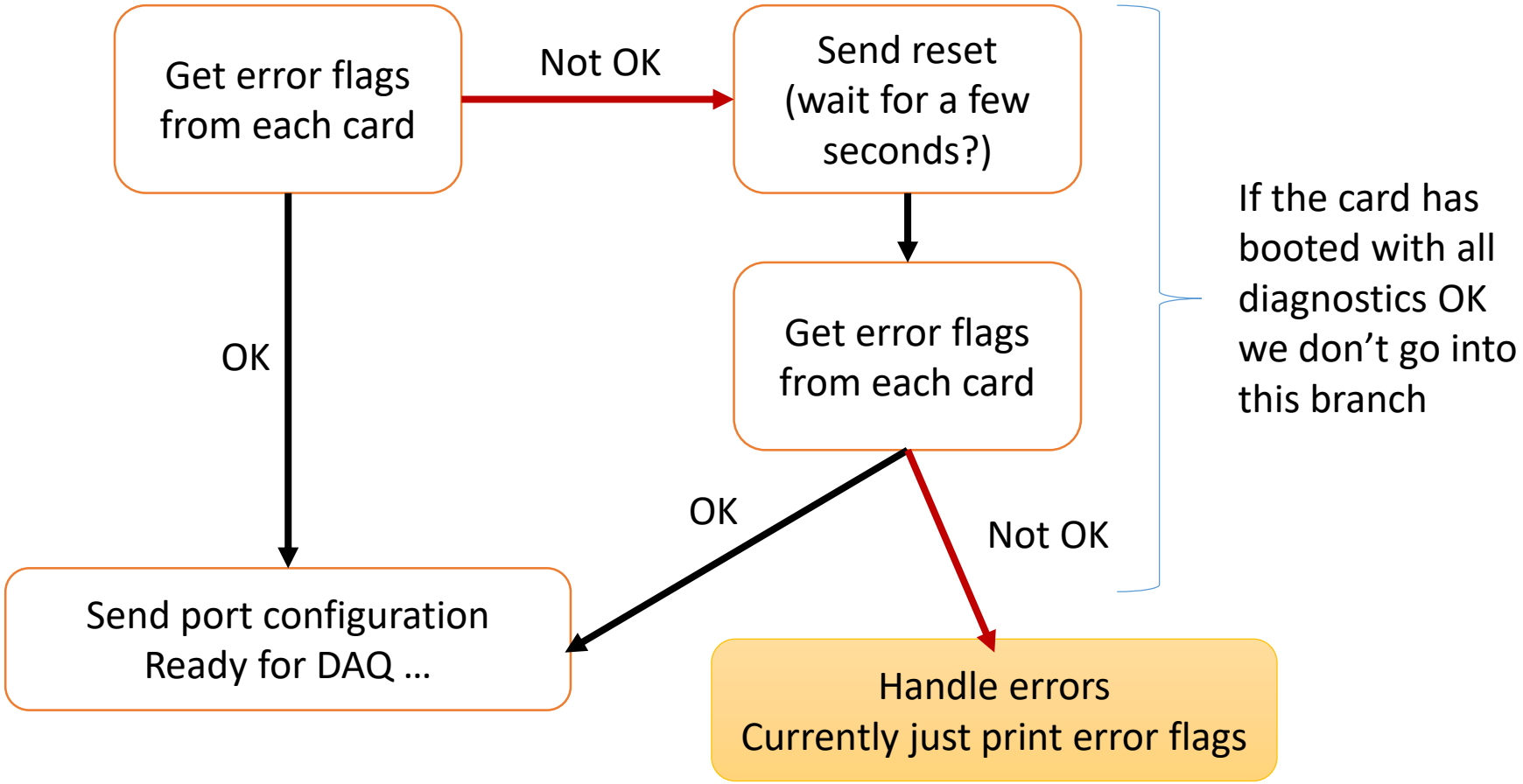
Multicast broadcast (run control discovers AMCs presence)

"Discover" all the cards present on the network by sending out periodic (but not during a run is on-going) multicast broadcasts

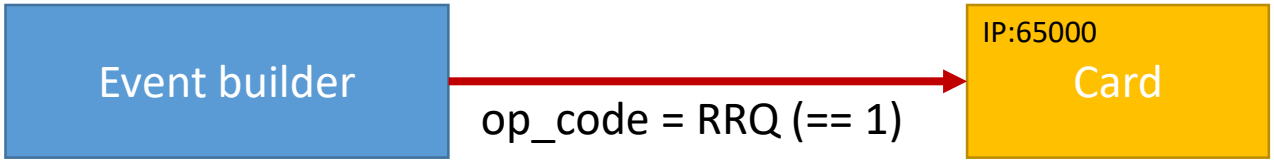


The cards also answer to arping, so one can quickly check if the card has booted and alive

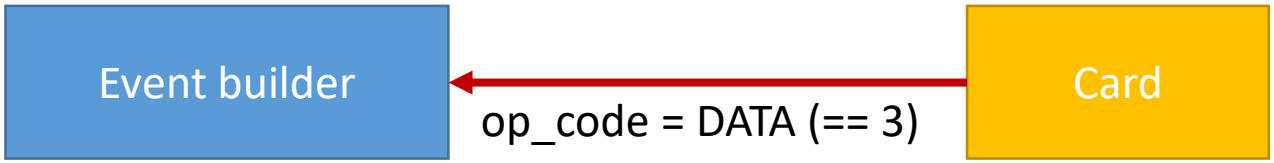
DAQ run initialization by L1 event builder + run control: current scheme in NP02



TFPT packet to read AMCs status (error flags)



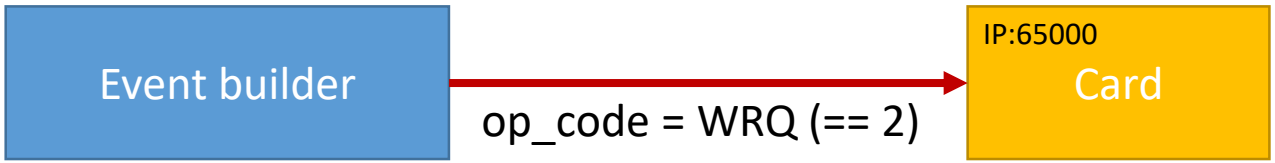
TFTP packet:
uint16_t op_code
uint32_t data[4] ← error flags (all 0's == OK)



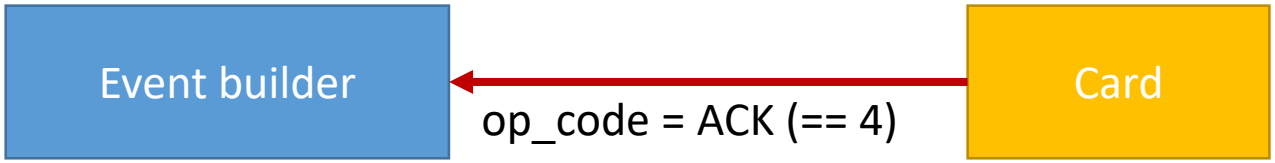
ADC error flags ch 0…31: adcErrFlags[0] = data[0];
ADC error flags ch 32…63: adcErrFlags[1] = data[1];
General error flags (WR, AMC): genErrFlags = data[2];
Spare error flags: sprErrFlags = data[3];

Unit status OK if this statement is true:
bool unitStatus = ((adcErrFlags[0] | adcErrFlags[1] | genErrFlags | sprErrFlags) == 0);

TFTP write packet (AMC configuration packet) + ACK



TFTP packet:
uint16_t op_code
uint32_t data[4] ← run/stop + port # to send event data

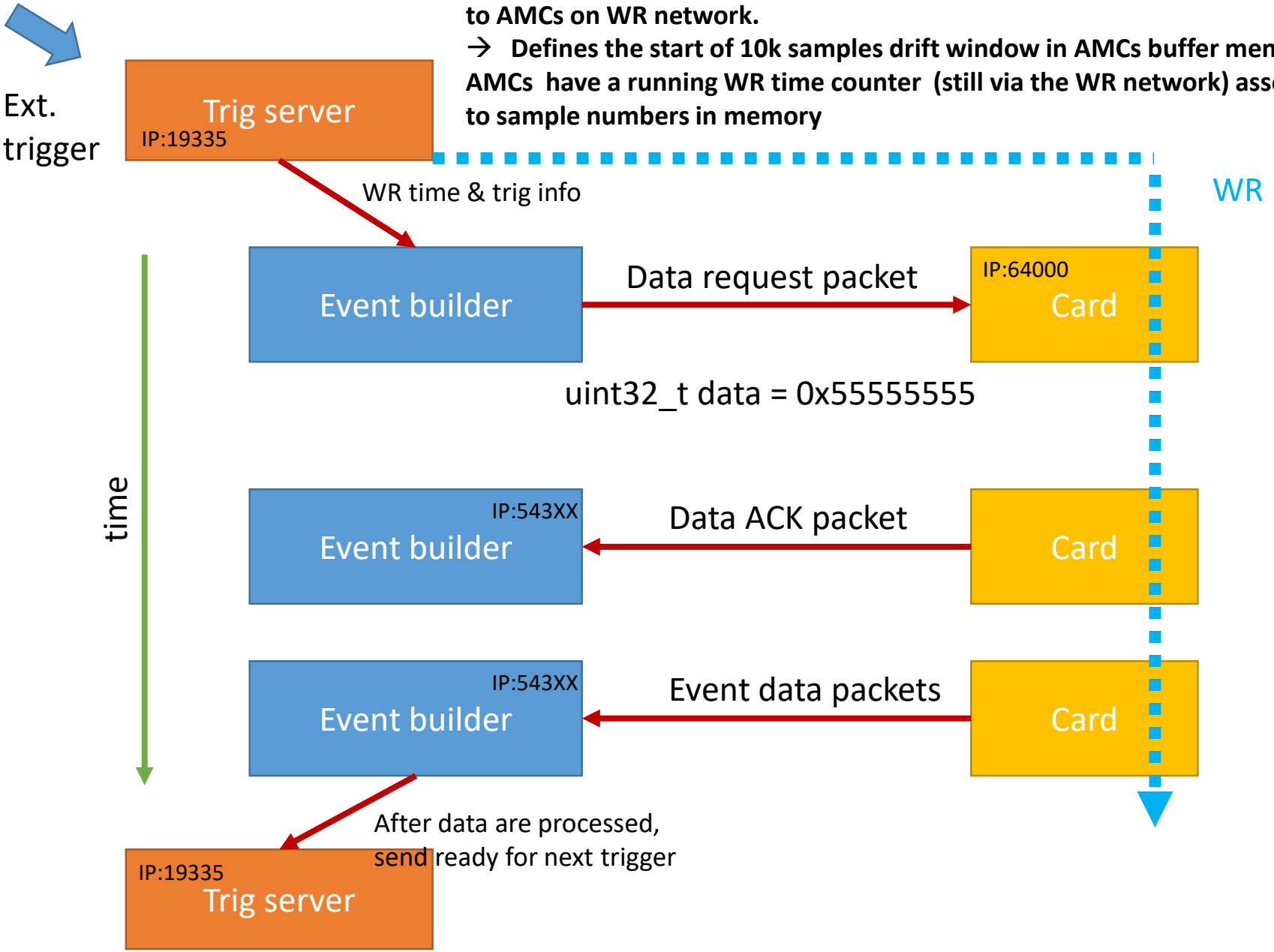


TFTP packet:
uint16_t op_code
uint32_t data[4]

Only op-code flag is checked to be ACK.
The data are not checked, so could be anything

Drift window acquisition cycle in external trigger mode in NP02

External trigger signal timestamp taken in WR time-base and then transmitted to AMCs on WR network.
→ Defines the start of 10k samples drift window in AMCs buffer memory
AMCs have a running WR time counter (still via the WR network) associated to sample numbers in memory



Data ACK packet from AMC to L1 evb Req packet in data acquisition cycle (single packet per drift window)

uint32_t data[3]

data[0] = 0x56565656 ← ACK flag

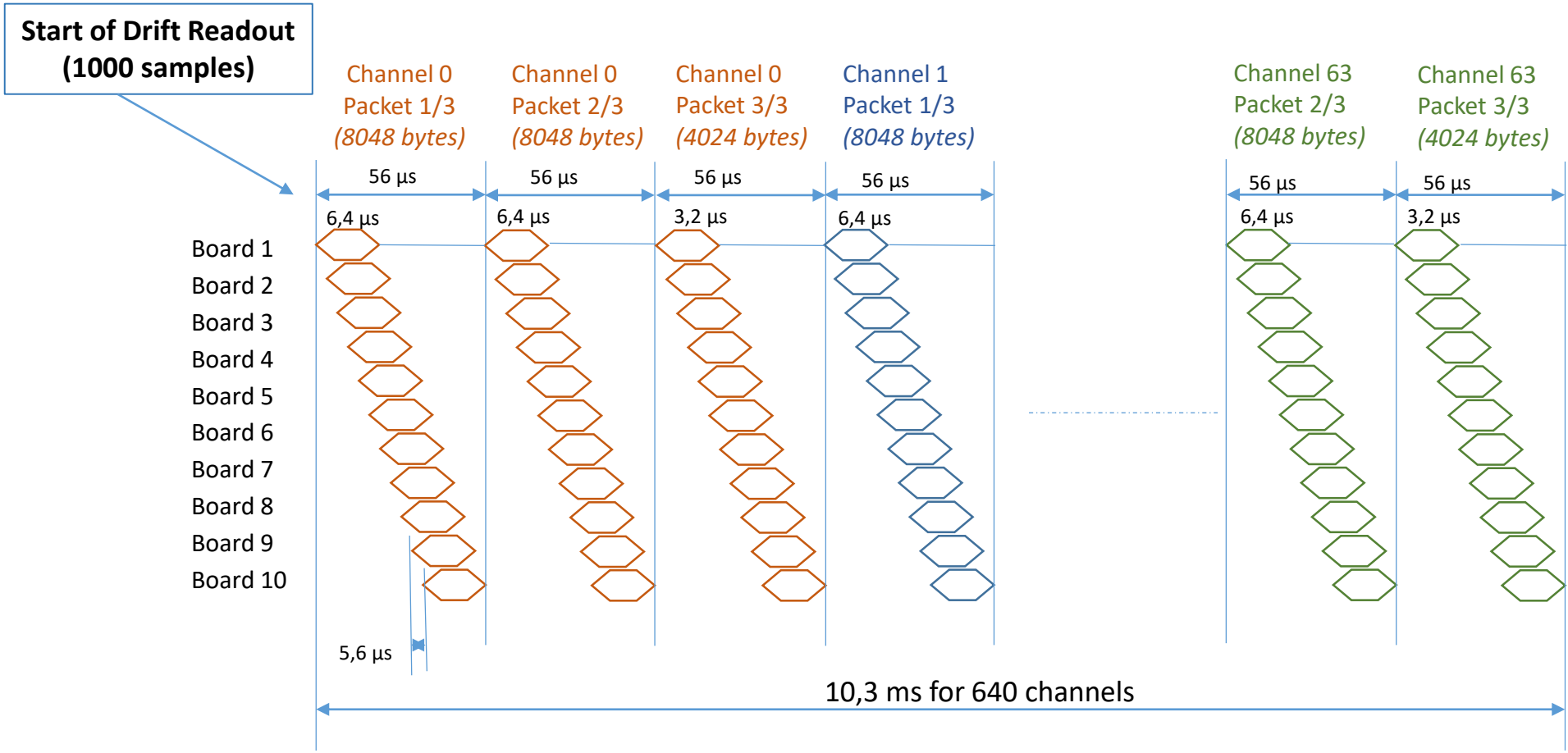
data[1] = UTC sec

data[2] = nsec

} WR timestamp

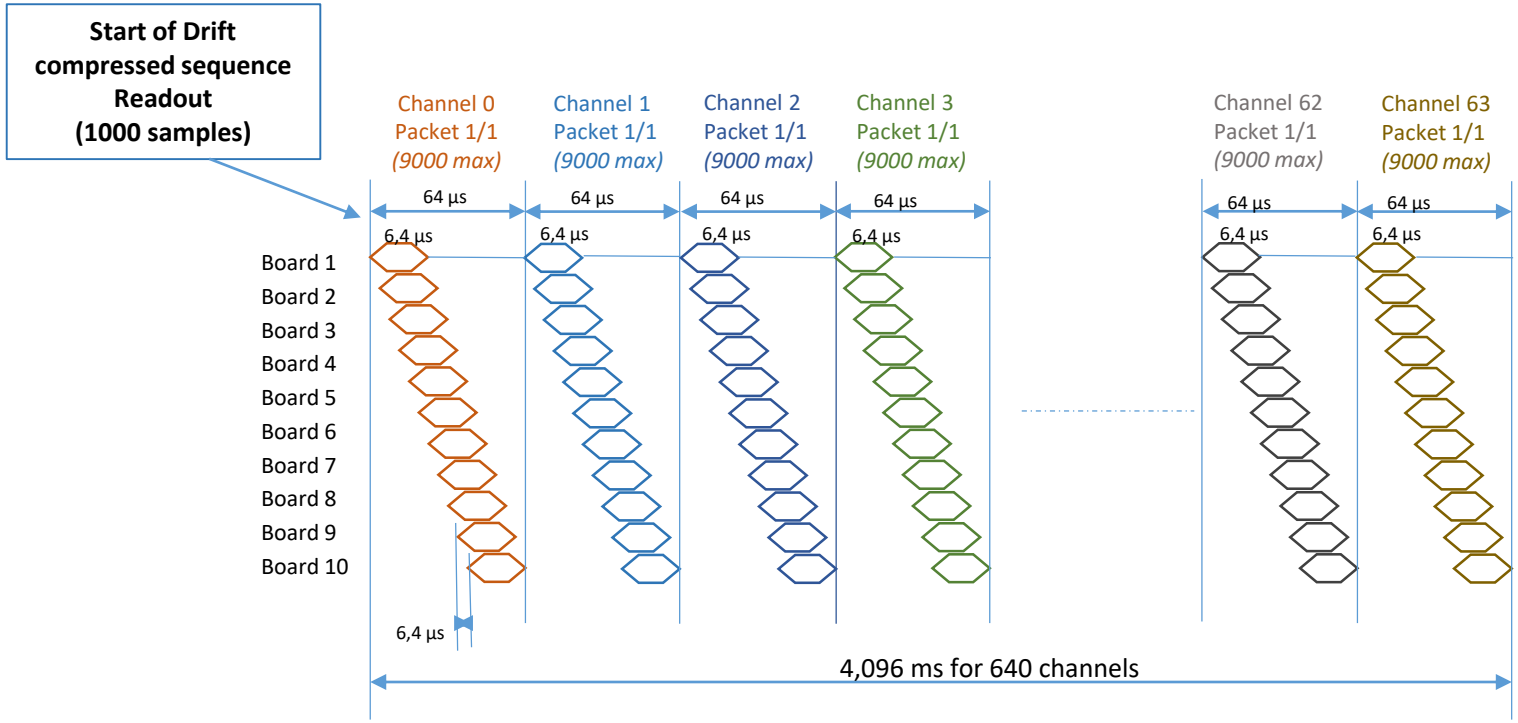
Pipelined transmission exploited in NP02 (3 packets/channel, drift of 10000 samples, no compression)

- **In order to maximally occupy the 10 Gbit link in Protodune-DP a pipelined transmission firmware was set up:**
- **Allows operating close to saturation of 10 Gbit / s MCH link for uncompressed data flow, max trigger rate ~90 Hz**
- **Cards send data packets every 5.6 us sequentially: AMC1 Pkt1, AMC2 Pkt1, …, AMC10 Pkt1, AMC1 Pkt2, etc.**



Pipelined transmission exploited in NP02
(1 packet/channel, drift of 10000 samples, Huffman compression)

Fully deployed and tested in protoDUNE is summer 2020

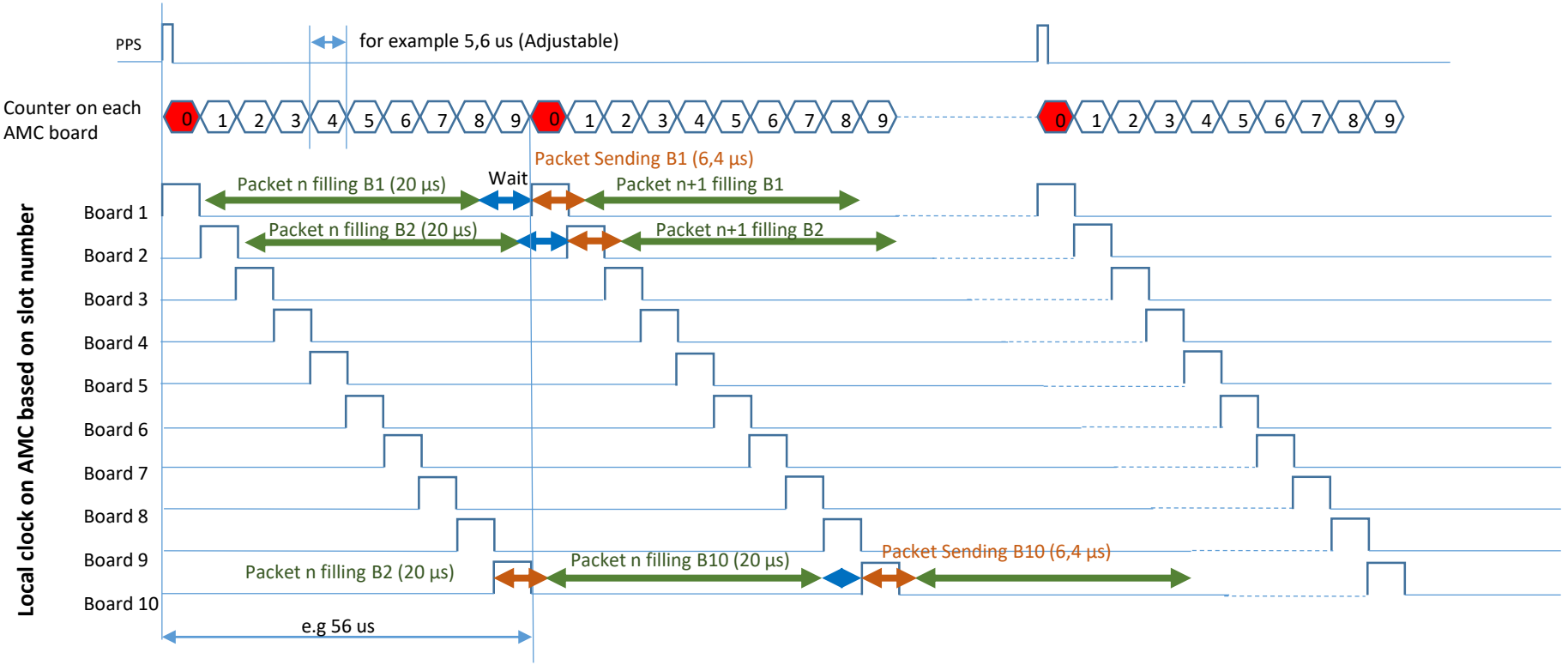


Transmission Duration for a compressed sequence of 10000 samples of full Chassis
(MAX 4,096 ms / MIN 0,8 ms depending of compression ratio)

Packet Transmission Scheduling (based on WR clock)

(Only one packet needed for 10000 samples drift and compression ratio up to 2:1, 10:1 expected)

Packet Scheduling and transmission smoothing



- The Time to fill a packet is given by the dual port bandwidth (3.2Gb/s) => 20 µs for 4000 samples
- Period is set based on the average packet length to send (8048x2 + 4024) => 6707 bytes => 5,3 µs @ 10 Gb/s
- Each board sends a packet every 56 µs
- The MCH receives 10 packets every 56 µs but not in the same time (smoothing).
- Useful to reduce timing constraint on back-end side.

Vertical Drift continuous streaming: examples of basic assumptions

- 2MHz sampling, 12 bit, no compression, 40 Gbit/s MCH data links
- Continuous data streaming (no external trigger). It means that we work similarly as in NP02 but with a deterministic internal trigger rate
- Continuous streaming organized in drift frames (continuous fake internal trigger) of n (10k or another number best fitting packet size ? See example below)
- Samples sent in UDP data packets covering the frame length with similar format as in NP02. Frame length in terms of number of samples to be discussed.
- Exchange of control packets in between frames (should we implement this handshaking protocol?, similar format as in NP02 ?)
- Packets transmission scheduling scheme possible, as implemented in NP02
- AMCs know WR timing and DAQ timestamp counter which can be reported at the beginning of a data frame (timestamp of first sample of the frame) or even in each data packet (timestamp of first sample of the packet). Details under discussion
- Using 9000 as MTU reference for JUMBO frames and assuming ~10 B for packet header (can contain timestamp info, crate / card / ch) + 42 B reserved for Ethernet protocol → max data bytes per packet 8948 B ~ 5965 x 12 bit ADC samples
- If it is preferable not to realize data (i.e., pack 12 bits into 16 bits words). Sending data in 16 bit words (only 12 bits actual ADC value) one can put up **to 4474 ch samples per packet**

→ This would mean a raw data rate of 20.5 Gbit/s which is largely below 40 Gbit / s bandwidth limit