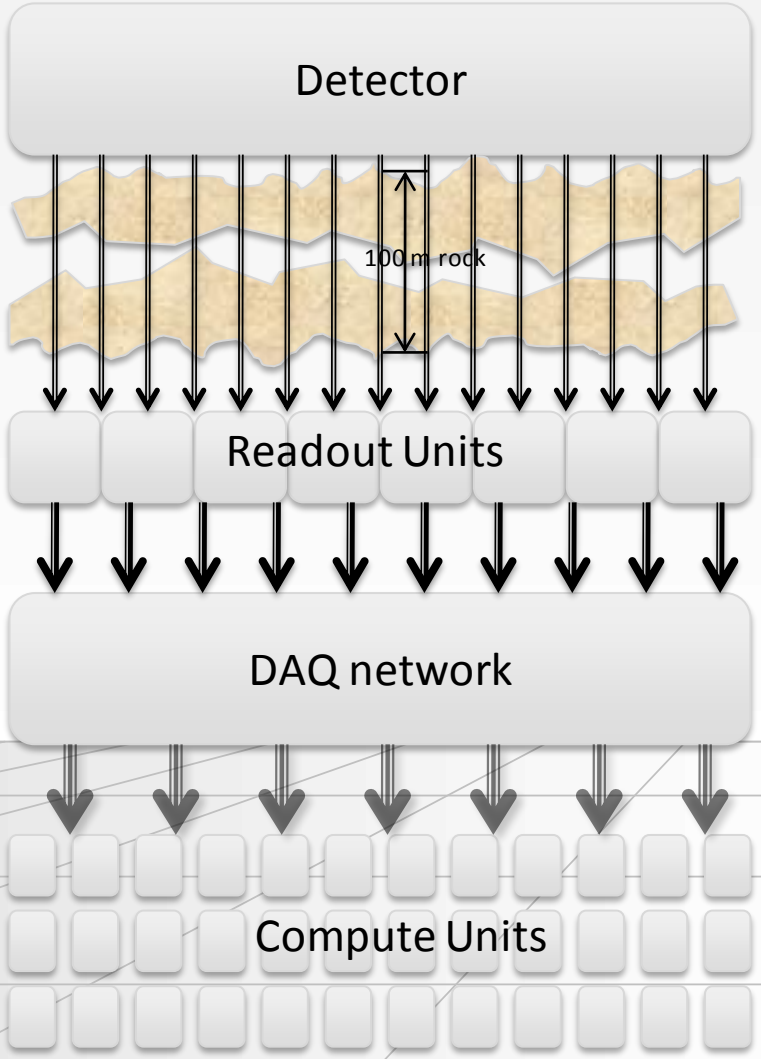
LHCb Upgrade Architecture Review

BE<=>DAQ Interface

Rainer Schwemmer

Sketch of the upgrade DAQ

LHCb Upgrade DAQ



GBT: custom radiation-hard link over MMF, 3.2 – 4.8 Gbit/s (~ 10000 - 12000)

Input into DAQ network (10/40 Gigabit Ethernet or FDR IB) (1000 to 4000)

Output from DAQ network into compute unit clusters (100 Gbit Ethernet / EDR IB) (200 to 400 links)

Key figures

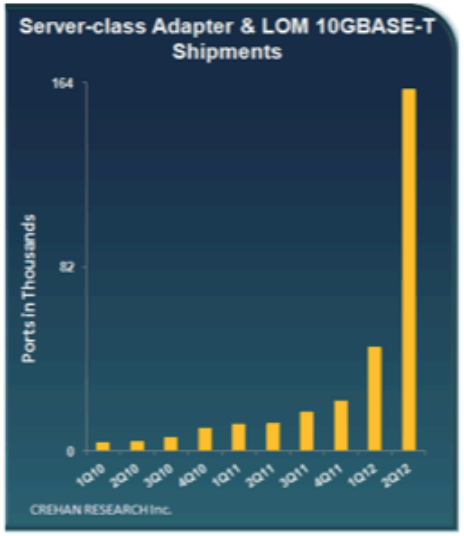
- Minimum required bandwidth: > 32 Tbit/s
  - # of 100 Gigabit/s links > 320, # of 40 Gigabit/s links > 800, # of 10 Gigabit/s links > 3200
- # of compute units: many (depends on CPU power, trigger algorithms, GPU or not, etc…) → design should scale “easily” between 2000 - 10000
- An event (“snapshot of a collision”) is about 100 kB of data → for the DAQ only the bandwidth counts
- # of events processed every second: 10 to 40 millions
- # of events retained after filtering: 20000 to 30000 per second (data reduction of at least a factor 1000)

The network options

- Two local area network technologies are investigated: Ethernet and InfiniBand
- Speed-grades
  - Ethernet 10 Gbit/s, 40 Gbit/s
    - 100 Gbit/s (only for up-links and interconnects, not from the BE-board)
  - InfiniBand QDR 32 Gbit/s FDR 54 Gbit/s (in the future: EDR ~ 100 Gbit/s)

How to minimise cost in the DAQ

1. Minimise distances between network devices (switches) and end-nodes (TELL40s, compute-units) → location of BE (later)
2. Try to use high-bandwidth LAN devices with shallow / no buffering (like InfiniBand or “cut-through” Ethernet)
3. Minimise the use of optics - use copper links (industry standard, largest market volume)
   – copper = SFP+ twinax cable or 10 GBaseT (twisted pair)



Network costs (based on recent offers 2012)

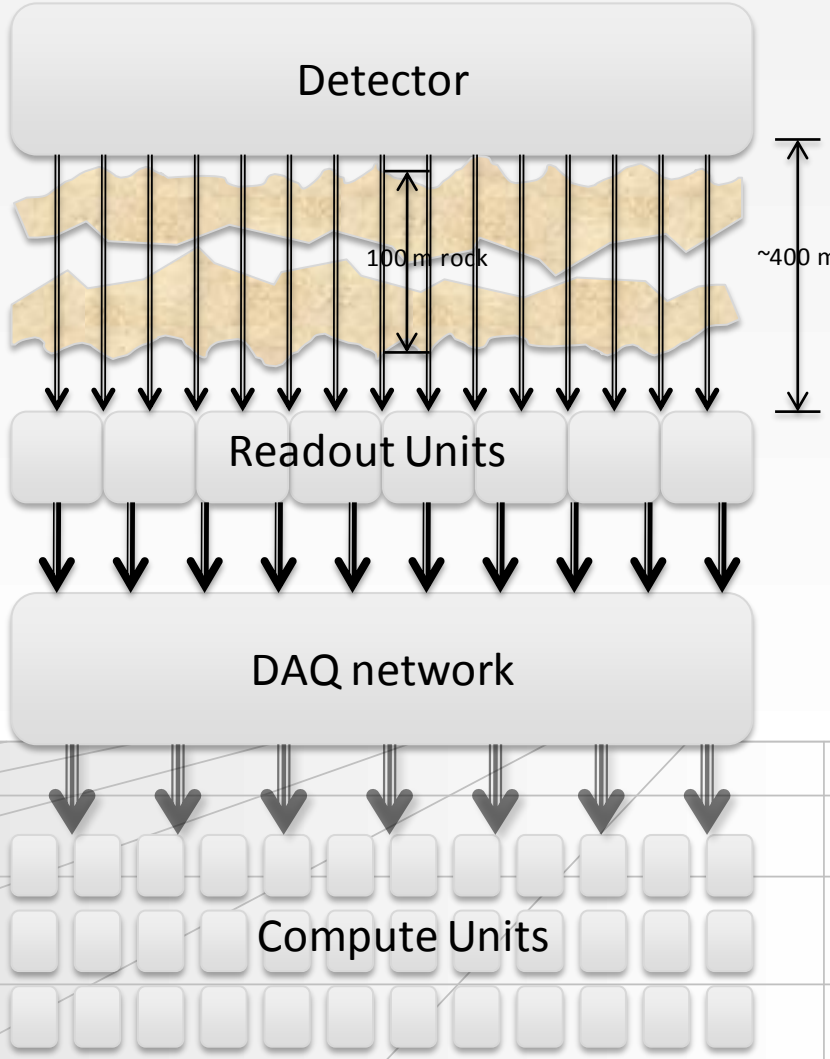
- Cost of switch ports excluding optics based on 2012 offers (not list-prices)
- “Edge” ports are ports with shallow buffers (cut-through) and without high-speed uplinks, typically found in switches build around chip-sets (Broadcom) by OEMs
- “Core” ports are ports with deep buffers (store and forward) and/or including interconnects and chassis overhead costs

Price per port (USD)	10 GbE	40 GbE	IB FDR (52 Gb)
edge	150	400	300
core	1500	4800	1000

Why worry about BE location?

- Using 10 Gig links from UX to SX requires a lot of links (cost in ports and optics!)
- Faster links (40 G / 100 G) do not have the required range
  - (40G-SR4 and 100G-SR10 are specified over 100 m (OM3) 125 m (OM4)
- Want to use long-range GBT → bring the BE to SX
- Even if the BE cannot use copper (which is already costly) at least between core-network and compute units (farm) we want to maximize the use of copper → network and farm have to be co-located
- The closer everything is the more we can use data-centre technology (volume!)

Link technology – location of back-end

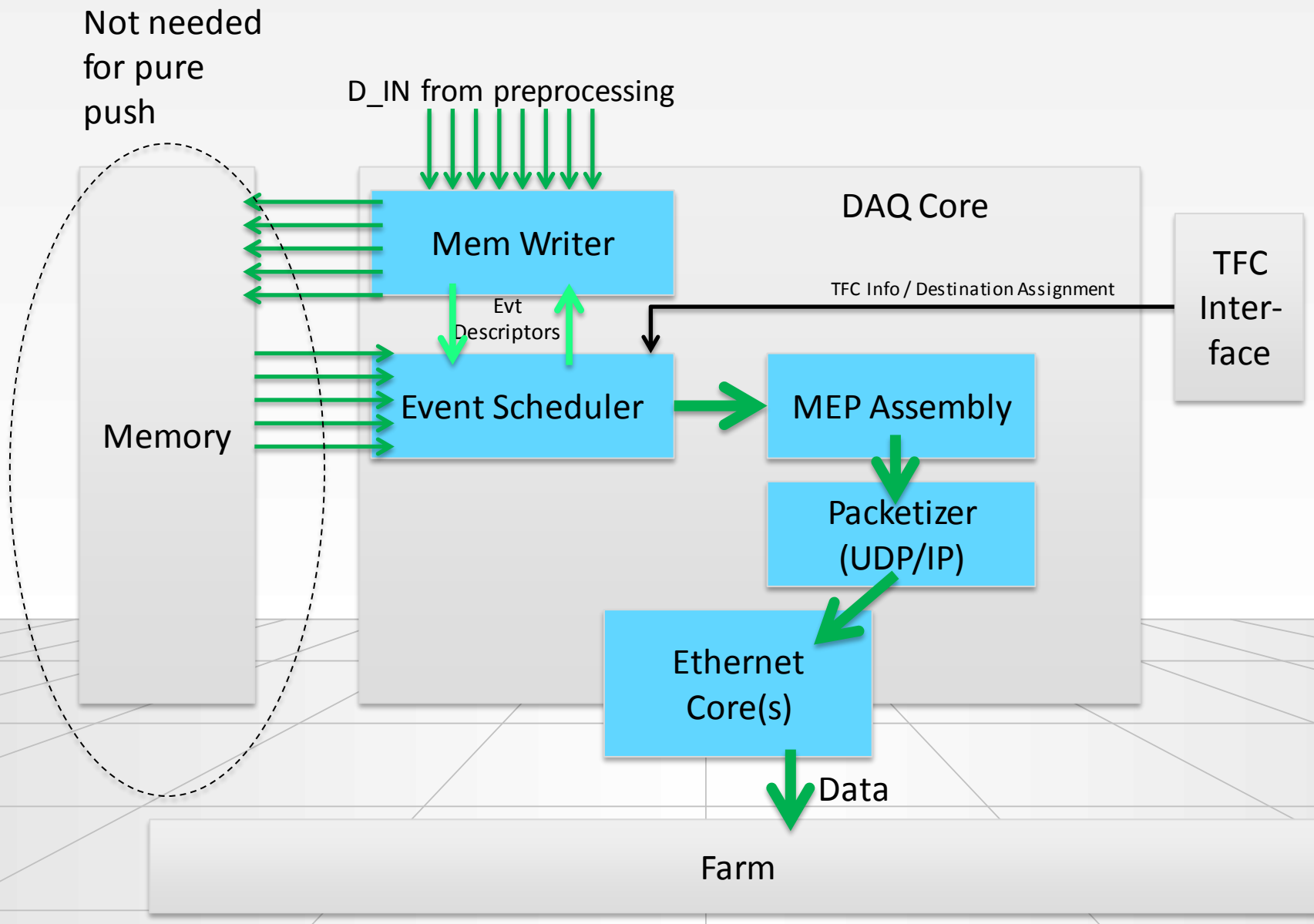


Long distance covered by low-speed links from detector to Readout Units.
☺ Cheap and optical links required anyhow

- GBT: custom radiation-hard link over MMF, 4.8 Gbit/s (about 10000)
- Input into DAQ network (10/40 Gigabit Ethernet or FDR IB) (1000 to 4000)
- Output from DAQ network into compute unit clusters (100 Gbit Ethernet / EDR IB) (200 to 400 links)

BE ←→ DAQ Implementation

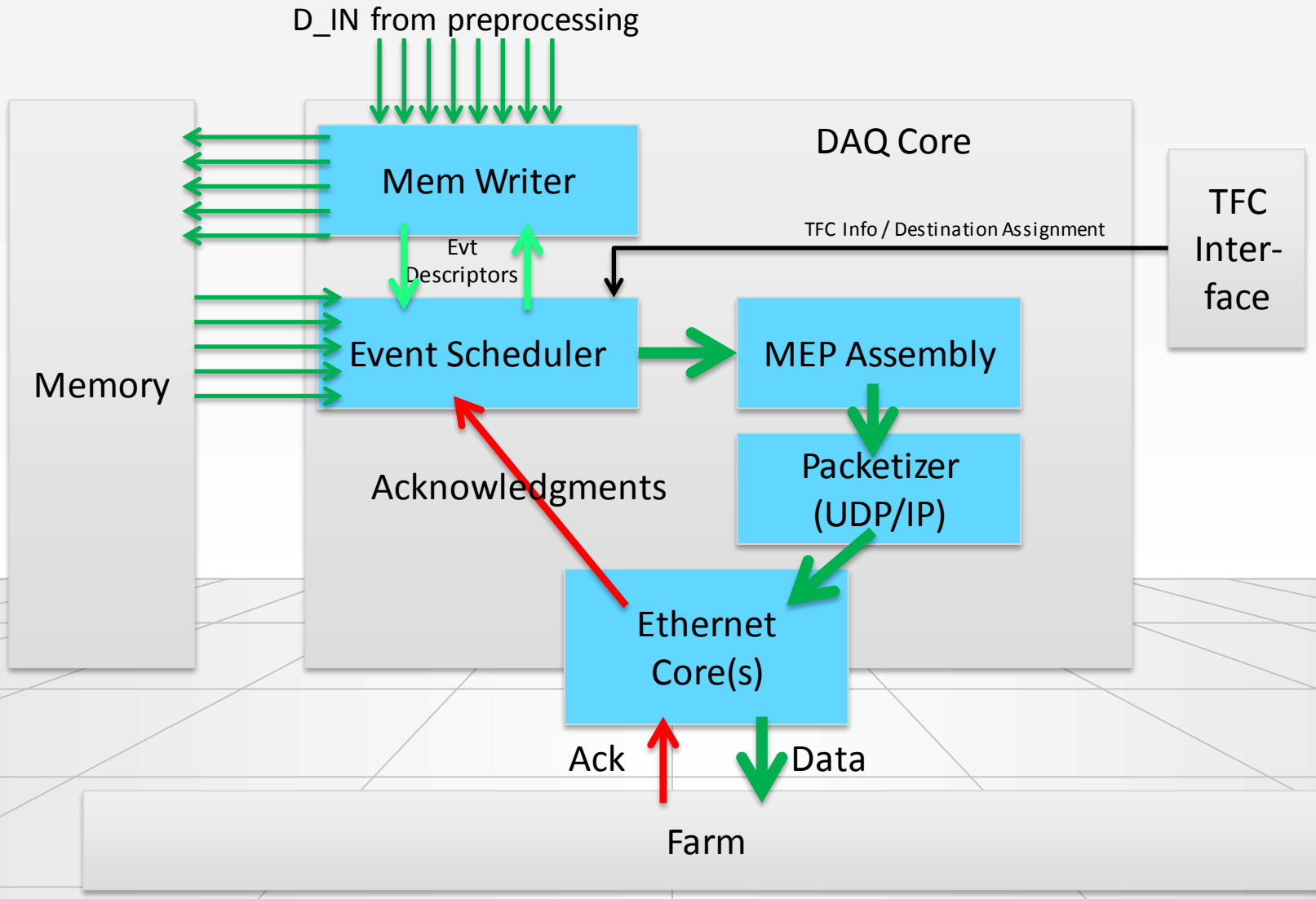
Example 1: FPGA DAQ Core (Ethernet / Push)



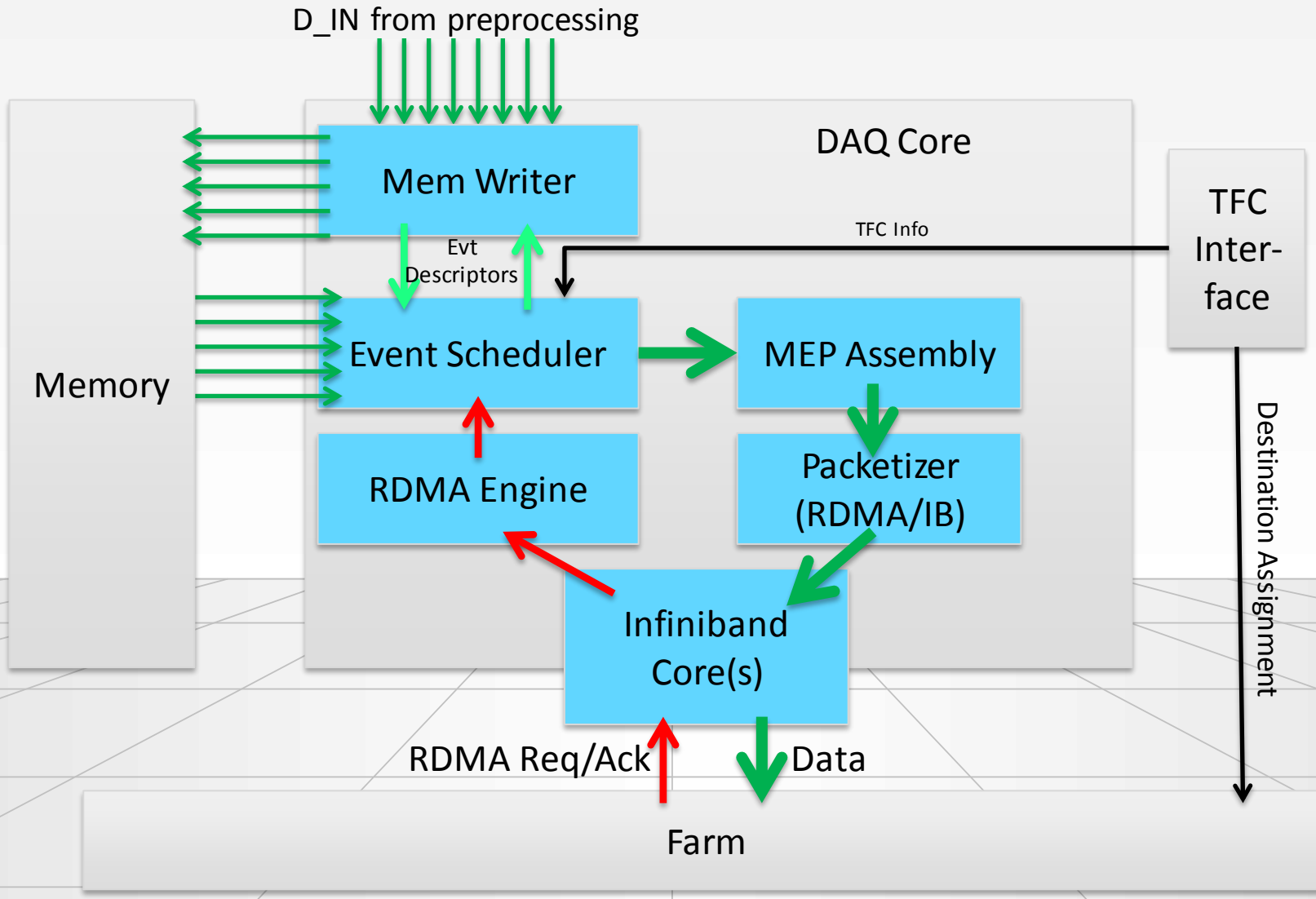
Buffering
Why do we need it?

- Simplest case:
  - Push, no resend -> current scheme
  - Problems: See current network issues
    - Current DAQ edge switches: 1.5GB memory / 48 Ports = 31.25 MB/Port
    - “Theoretically” necessary: 3 * Event size * MEP factor = 2 MB
    - Still drops packets due to full buffers
  - In the future: Switches with large buffers tend to be very expensive (mostly telco-market)
- Need at least Buffer for scheduled sending + re-send
  - Needs to buffer Bandwidth x 1 RTT + several MEPs + contingency O(10 MBs per link)
- Buffering also needed for advanced flow-control protocols in Ethernet (QCN, etc…)
- Safest traffic-shaping method: pull protocol
  - Data is buffered on the BE board, the sinks collect it when they are ready for processing
  - Full control over output buffer depth (depends on protocol)
  - O(50 MBs per link)

Example 2: FPGA DAQ Core (Ethernet / Push – with scheduling and acknowledgment)



Example 3: FPGA DAQ Core (Pull)



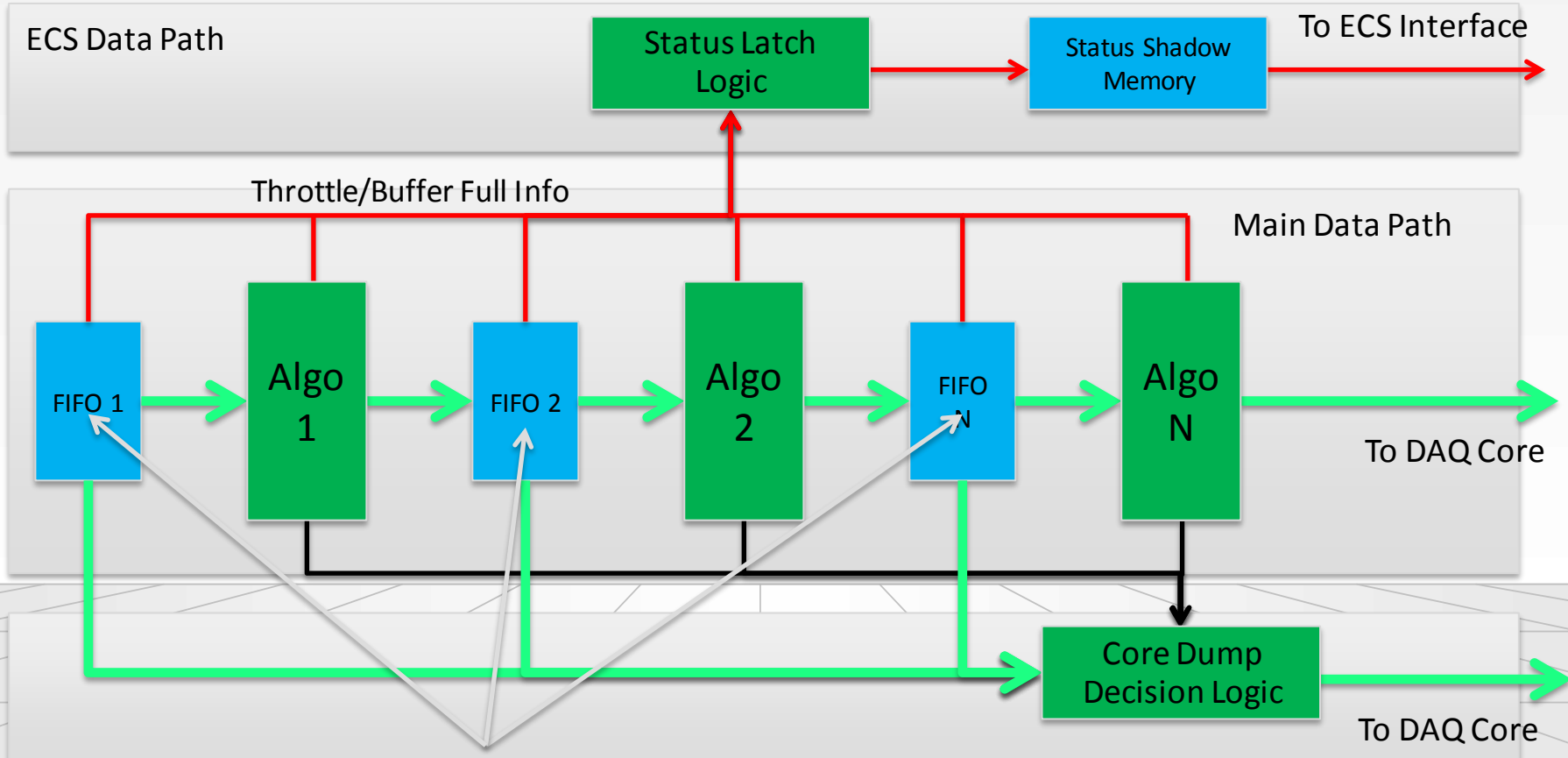
FPGA Interface

- Core interface consists of 3 components
  - Streaming Data Interface for receiving the data from pre-processing (Avalon-ST)
  - Memory I/O Interface: Write once, Read once + ε (Avalon MM)
  - TFC Input data (Avalon ??)
- Additional Interface for ECS
  - To be specified

Debugging

- Essential:
  - Counters and status registers of all sub algorithms on the board
  - Operational experience shows that it is often useless to have counters if not synchronized => snapshot trigger + latched please
  - If your algorithm is blocking the data train, we need to know!
- Necessary for post mortem analysis
  - Need history of input data If we want to know why something is throttling
  - Core dump like functionality:
  - Upon configurable trigger, the DAQ is blocked and all current data in I/O buffers of sub algorithms + N available previous events are sent to the DAQ core
  - Data is sent as high priority packet to HLT which can not be rejected
  - Obviously disabled during production phases

Debugging


ECS Data Path
Status Latch Logic
Status Shadow Memory
To ECS Interface
Throttle/Buffer Full Info
Main Data Path
FIFO 1
Algo 1
FIFO 2
Algo 2
FIFO N
Algo N
To DAQ Core
Core Dump Decision Logic
To DAQ Core
Use Ring Buffers for History analysis

Summary

- DAQ will use exclusively COTS technology (Ethernet or InfiniBand)
- Minimising distances between BE, network and farm allows minimising of cost
- Buffering on the BE allows using cheaper network equipment
- DAQ FPGA core should hide the interconnect technology and protocol from the rest of the firmware

Backup slides

Example of IB based read-out

- TELL40 using IB FPGA core (Bologna, CERN, …) →
  - 2 FDR IB ports (52 Gb/s)
  - need 16 serializers for about 100 Gbit/s output (can absorb 30 versatile links)