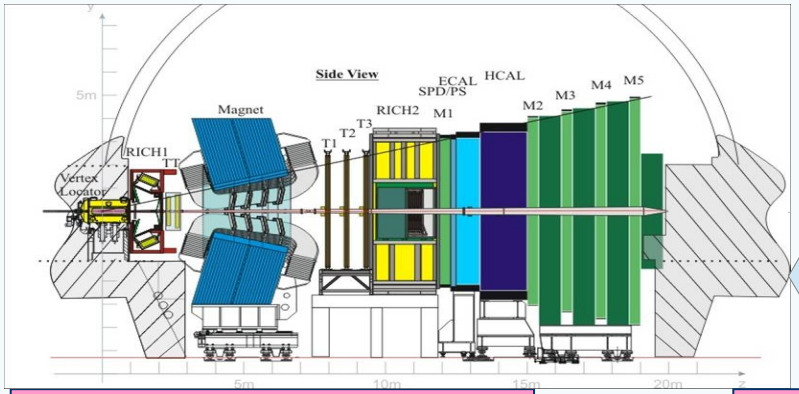
Data Processing Applications in the LHCb Online

- Numerics
- Overview over current situation
  - Basic building block
  - Configurations
- Possible other configurations

Markus Frank PH/LBC

Data Processing Apps > 80.000 Instances



Readout Network

TFC

Data Quality Cluster O(100)
- Process data from disk

CPU Reader

CPU CPU … CPU

Storage Cluster O(100)
- File and Stream handling
- Fork data streams

Storage System

CPU

CPU

Partially replicated data streams:
parasitic, best effort

Switch Switch Switch

HLT CPU Farm

CPU CPU CPU
CPU CPU CPU
CPU CPU CPU

CPU CPU

CPU CPU … CPU

CPU CPU … CPU

Monitoring Cluster O(100)
Low level monitoring

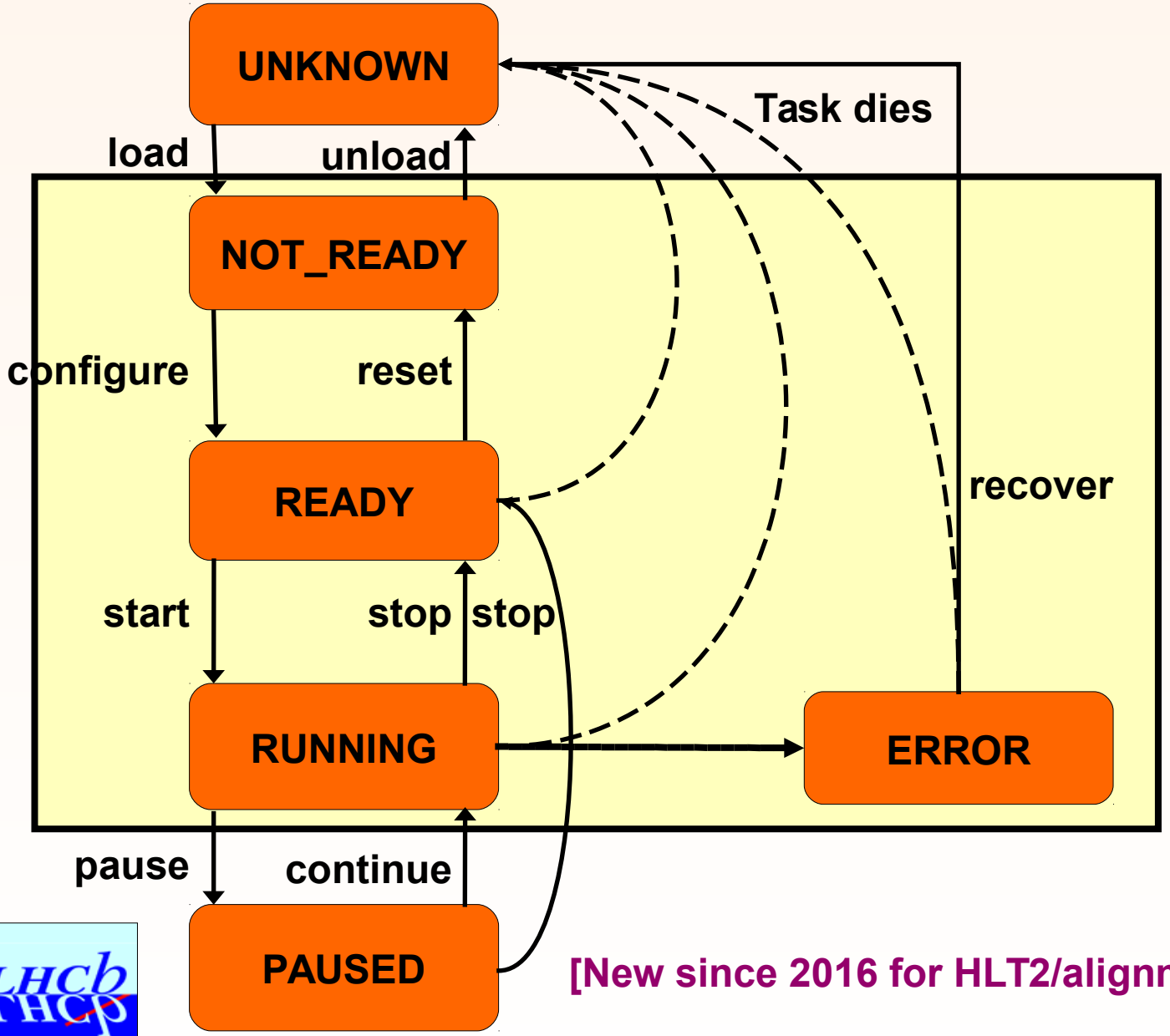
Reconstruction Cluster O(100)
High level monitoring with fully reconstructed events

High Level Trigger
HLT1: O(3000-25000)
HLT2: O(25000-50000)
Infrastructure: O(15000)

Software Architecture Based on Reuse

- Few basic building blocks
- Consequently combine them to functional units
  - Define node configurations
- Standard dataflow framework
  - Processes are specialized by combining components at run-time
  - FSM handling based on DIM
- Gaudi has its own special envelope
  - With time overheads and complications became too large
  - Regained flexibility and performance

Synchronization: FSM States of All DAQ Tasks

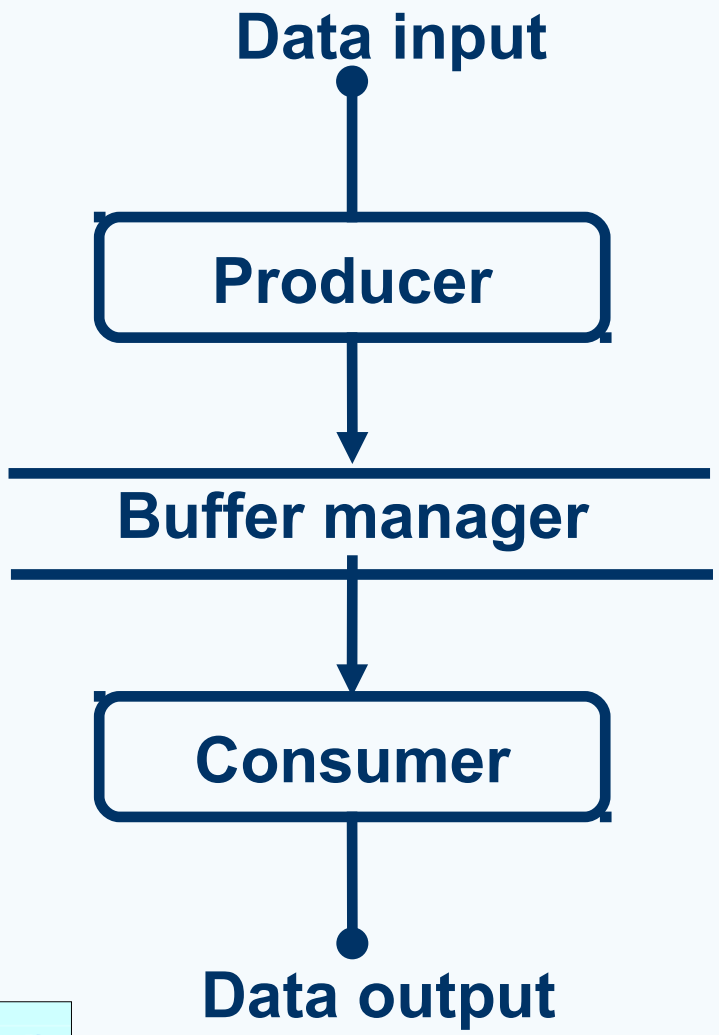


(EDMS 655828)

- **Mandatory implementatio**
  - **Independent of Gaudi**
- **Transitions steered by network messages**
- **State reported by network messages**

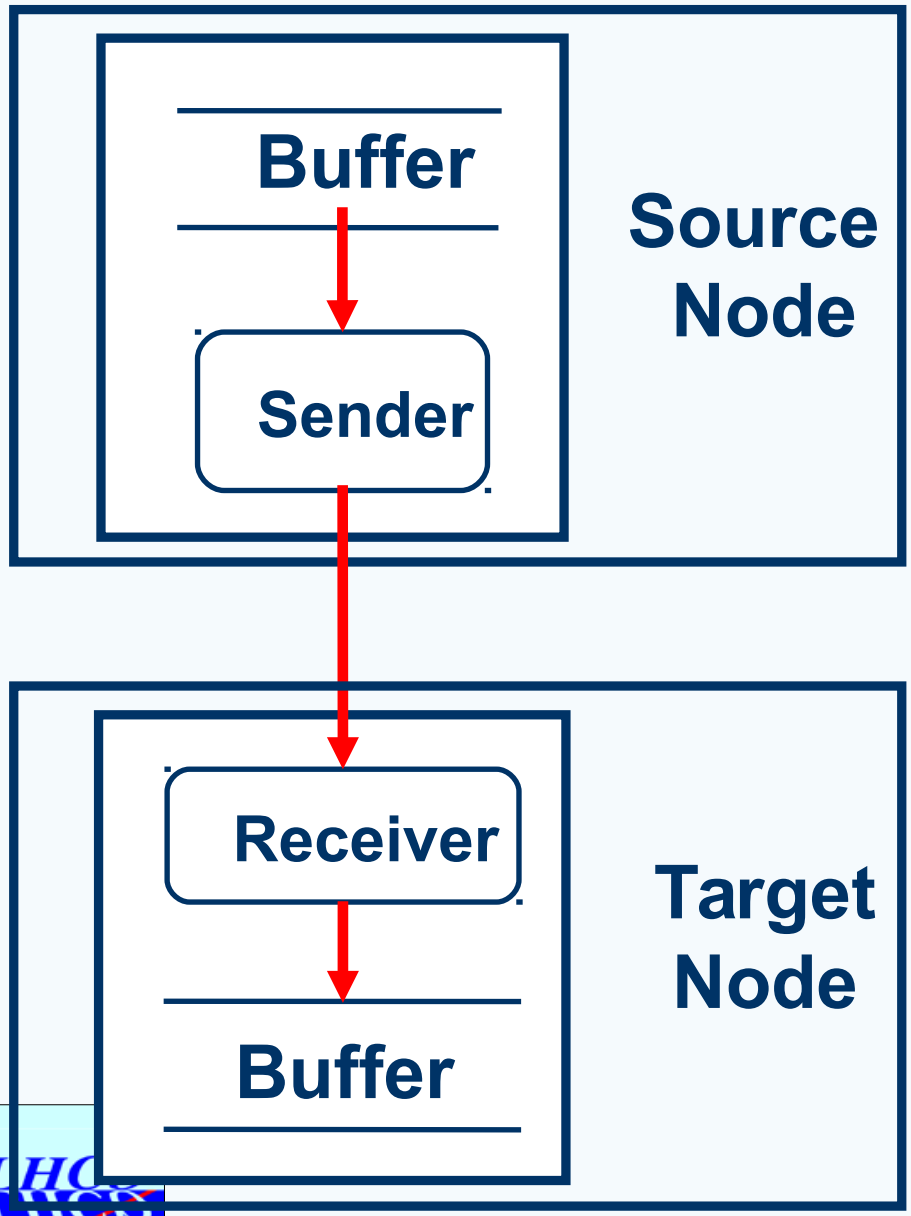
[New since 2016 for HLT2/alignment/calibration]

Data Processing Block



- **Producers deposit events in buffer manager**
  - Partition ID
  - Event type
  - Trigger mask
- **Consumers receive events by**
  - Partition ID
  - Event type
  - Trigger mask (OR accepted) and VETO mask
  - May queue different requests simultaneously
- **3 Consumer classes**
  - BM_ALL: Request to receive all events according to request definition.
  - BM_ONE: Out of a group of consumers with identical request definition one event is received by exactly one consumer.
  - BM_NOTALL: Request to receive some of the events according to request definition and buffer occupancy.

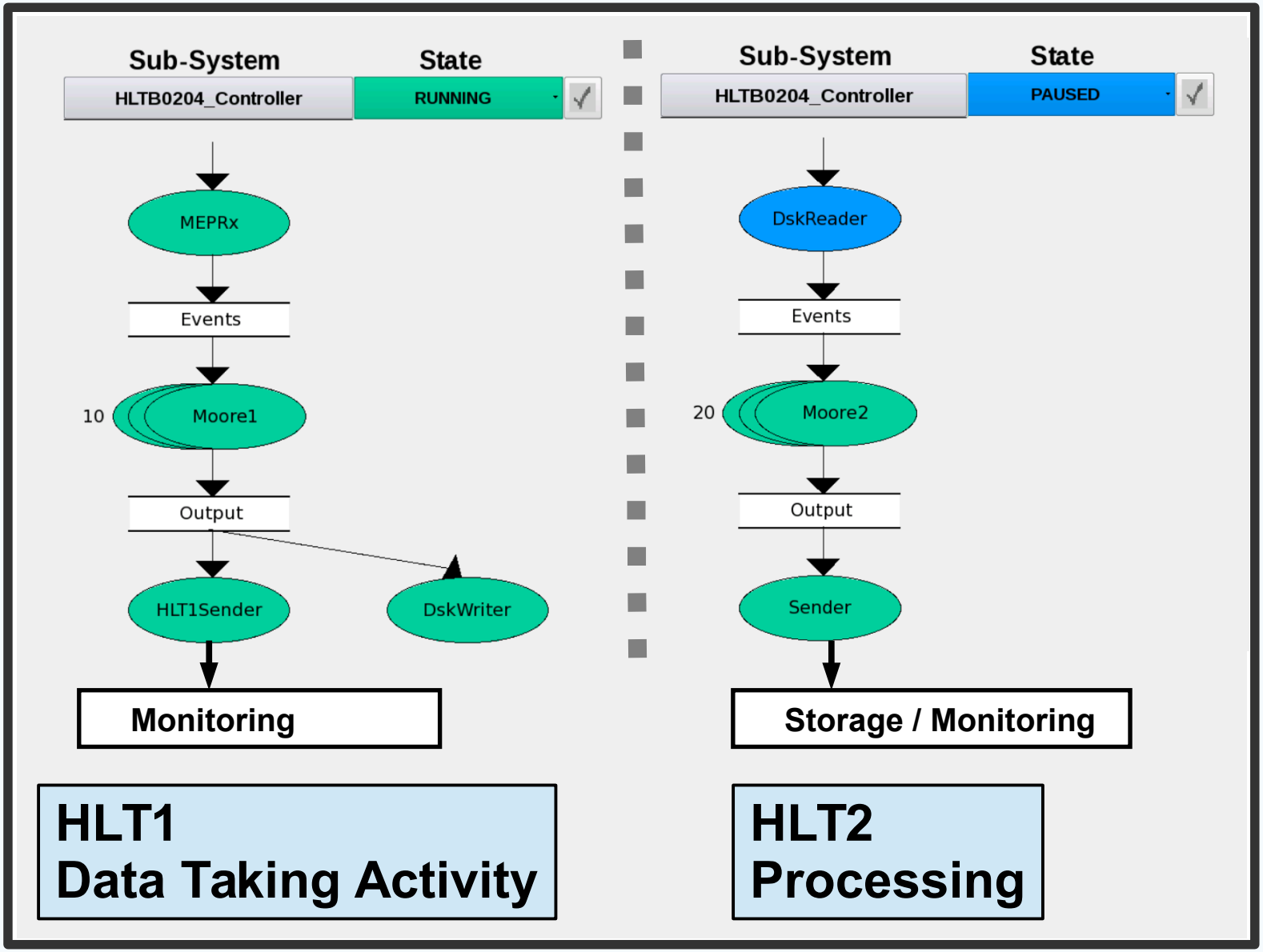
Data Transfer Block



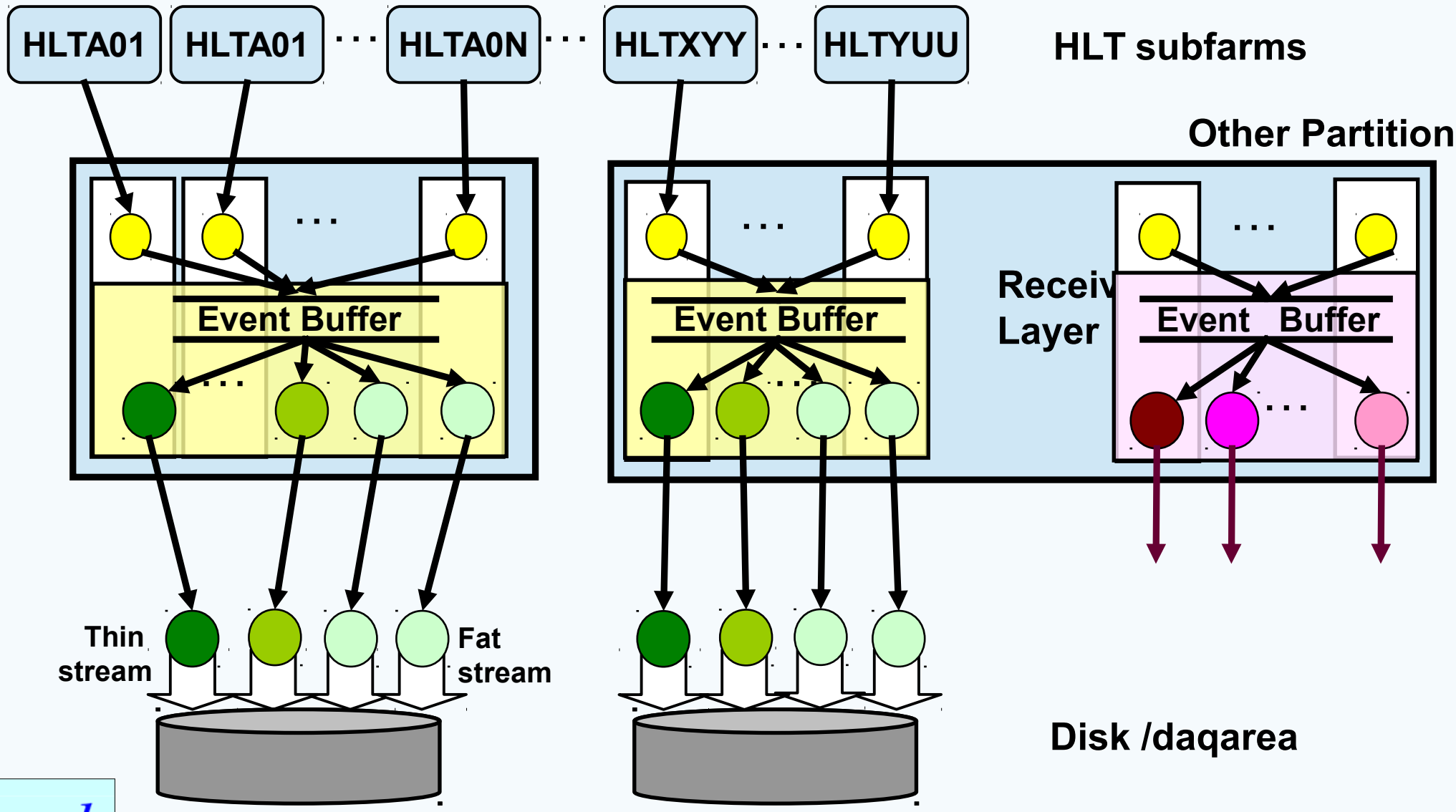
- Reversed data processing block
- Sender tasks accesses events from buffer manager on the source node
  - Consumer process
  - Send data to target process
  - Example: Data Sender on HLT farm node
- Receiver task reads data sent and declares data to buffer manager on the target node
  - Producer process
  - Example: Receiving process on the Storage System

See poster presentation No. 138:
"Data Stream handling in the LHCb experiment"

HLT Worker Node Architecture

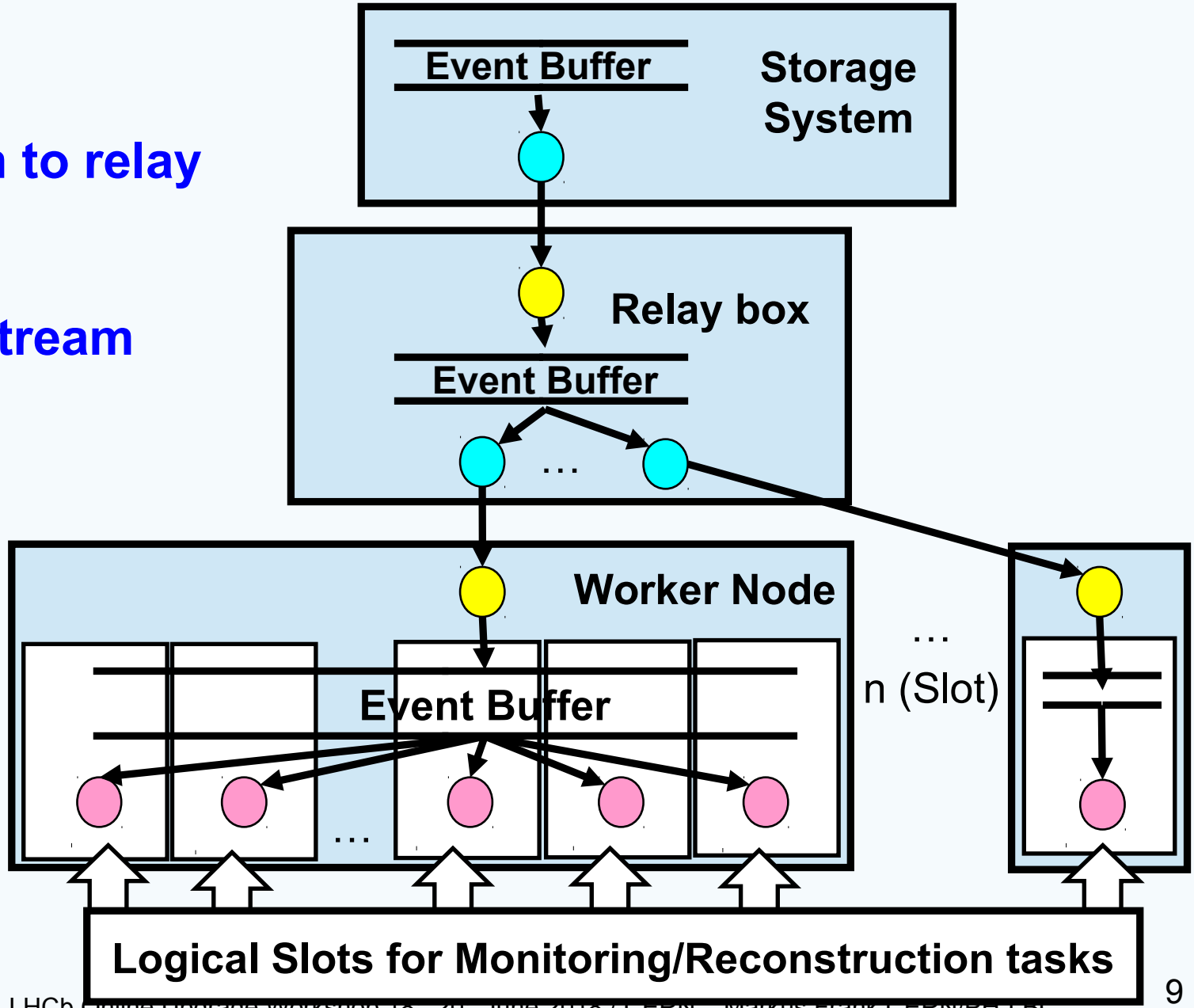


The Process Architecture: Storage



The Process Architecture: Monitoring and Reconstruction

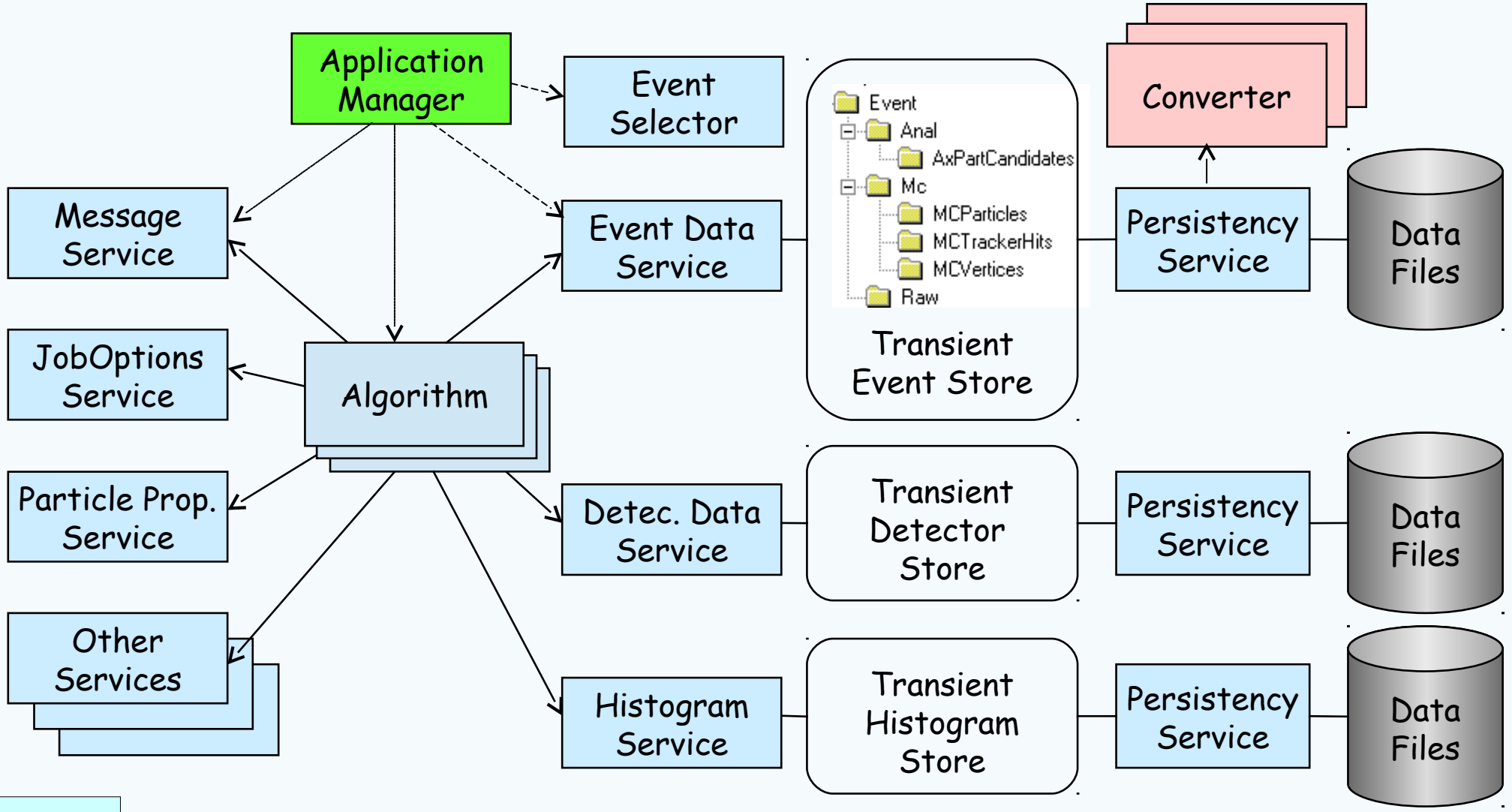
- **Storage**
  - Fork stream to relay
- **Relay box**
  - Distribute stream to workers
  - mona0801 mona0901
- **Worker**
  - Provide clients with data



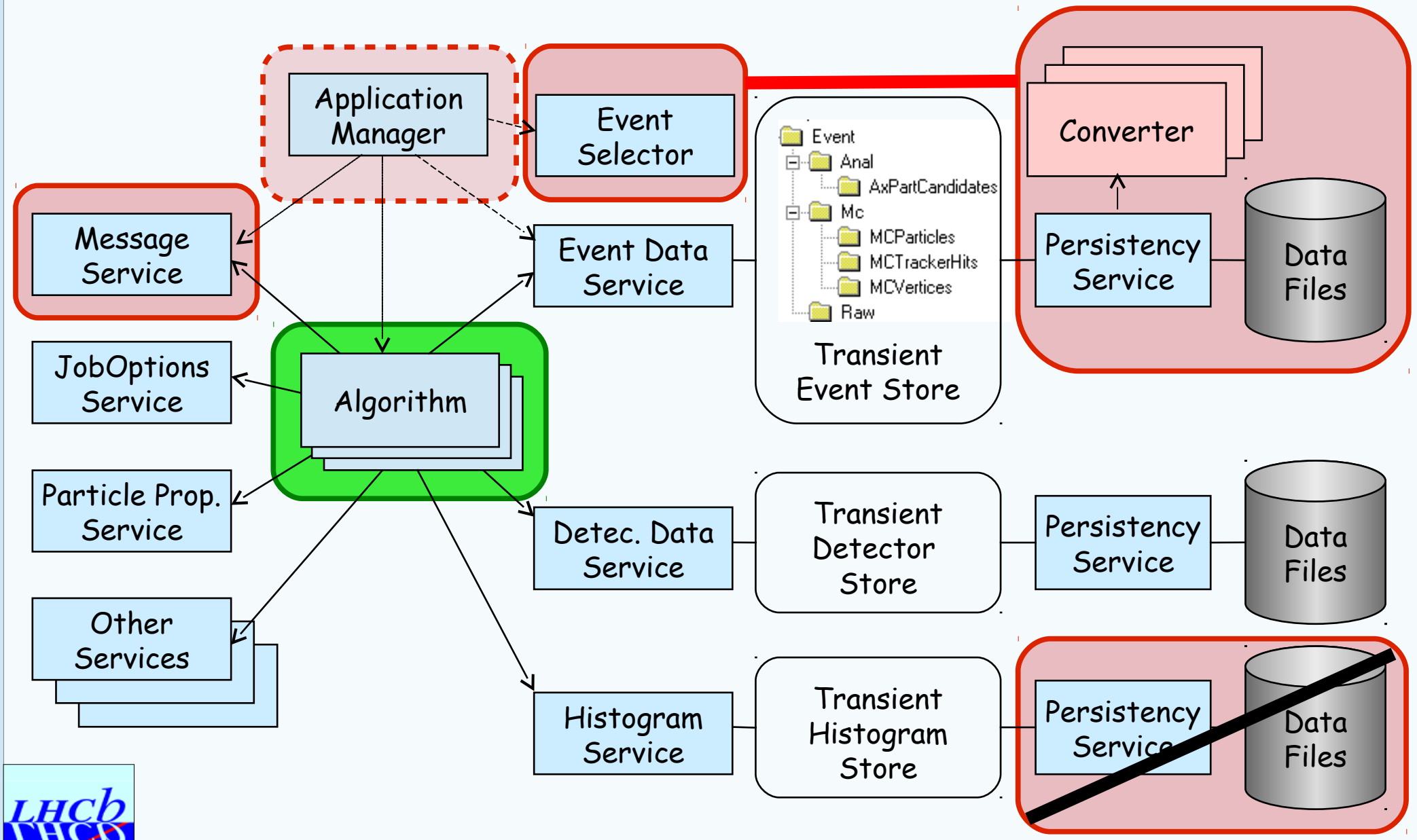
Synchronization: Current Implementation

- Currently an “envelope implementation” is used
  - All Gaudi calls are intrusively wrapped in a FSM driven object
- Future
  - Multi-threaded Gaudi: One instance per physical slot
  - Envelope won’t be. Replacement not (yet) defined Some thoughts were going on
  - Future of forking is also not very clear
    - Offliners do not like it
    - All depends on process startup time

Gaudi Architecture: Object Diagram

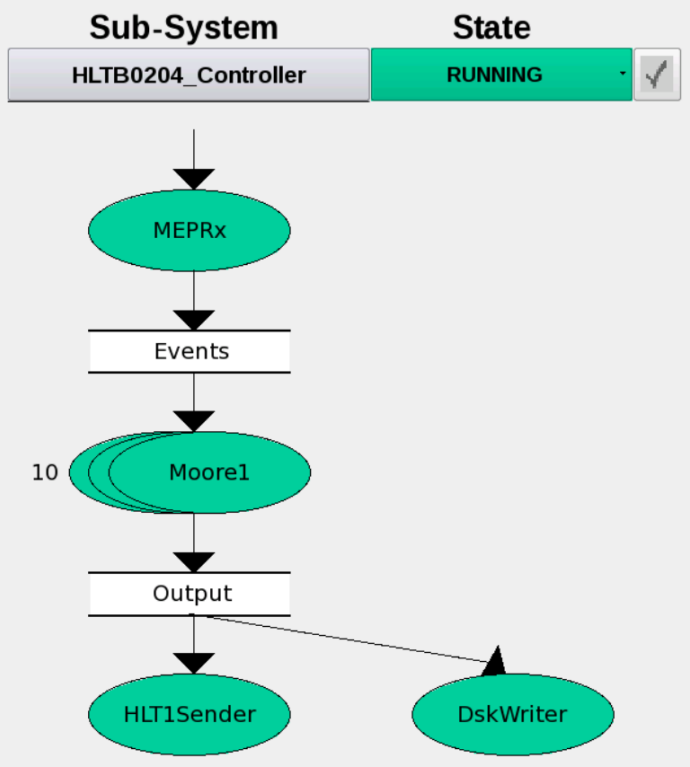


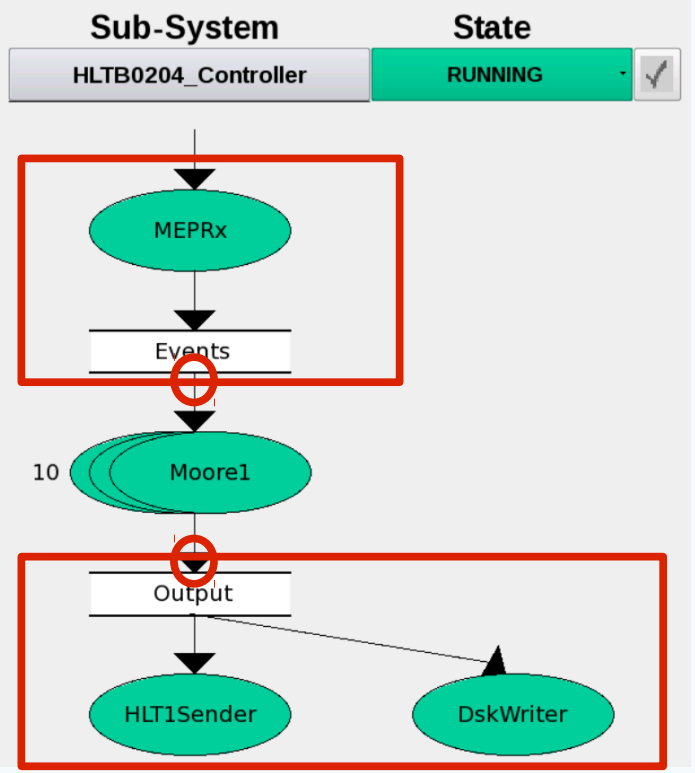
Gaudi Architecture: Object Diagram



Possible alternative configurations

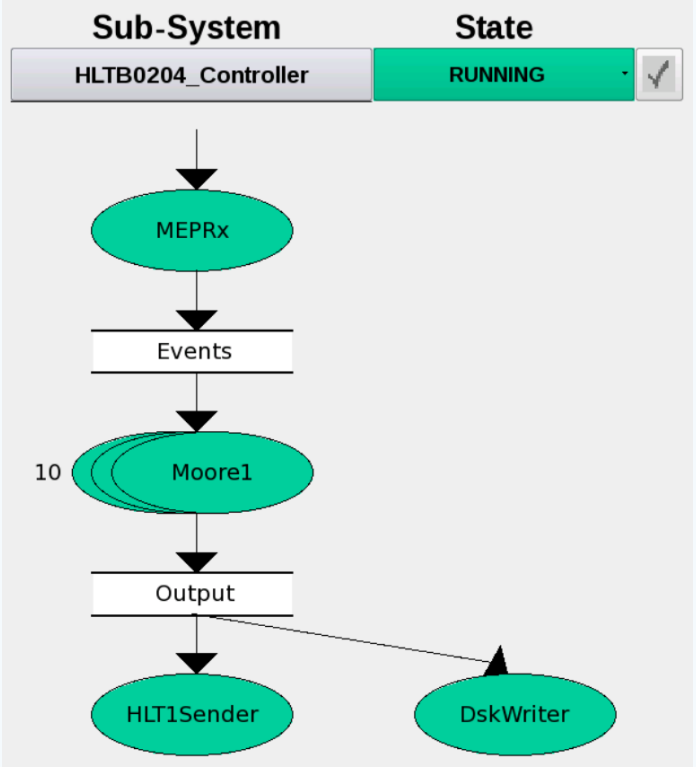
- In-process vs. out-of-process
- Buffer manager is abstract: buffer + input + output

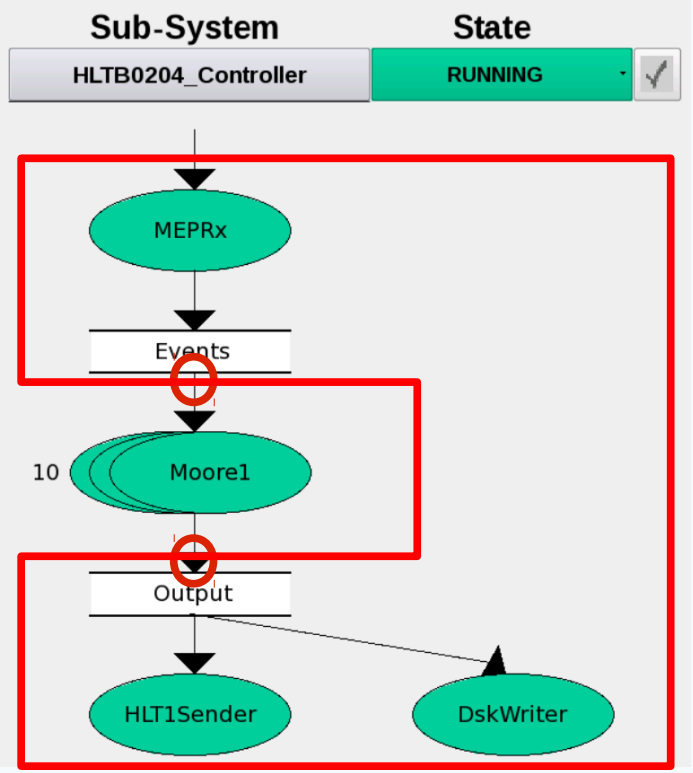




Possible alternative configurations

- Shared memory MBM vs. Unix sockets vs Fifos...
- But: Should modify Moore the least possible
  - Avoid interference





Conclusions

- Need to define “working points”
  - Process architecture
  - I/O mechanism with Moore(s)
  - Data exchange format
    => Direct influence on number of memcpy
  - Special cases: TAE (?), …

- Discussion