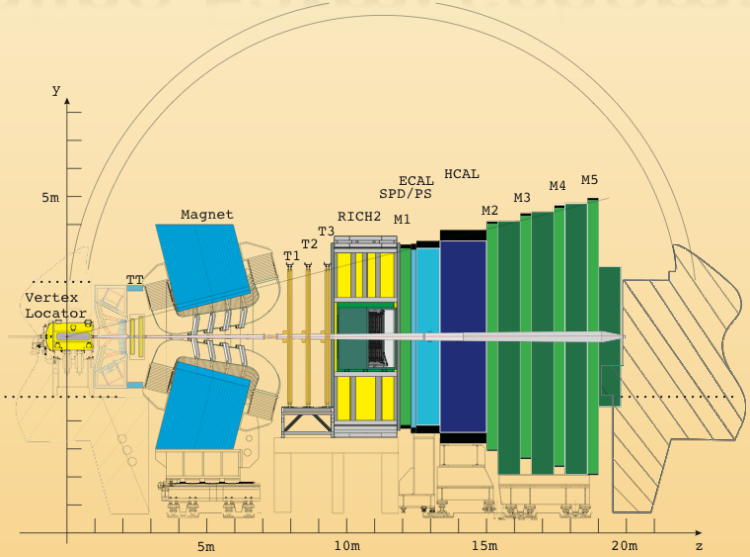
Markus Frank (CERN) & Albert Puig (UB)

Event reconstruction in the LHCb Online Farm

Outline

- An opportunity (Motivation)
- Adopted approach
- Implementation specifics
- Status
- Conclusions

The LHCb Online Farm: schematics



Readout Network

Online cluster
(event selection)

Switch | Switch | Switch

CPU CPU CPU | CPU CPU CPU | CPU CPU CPU

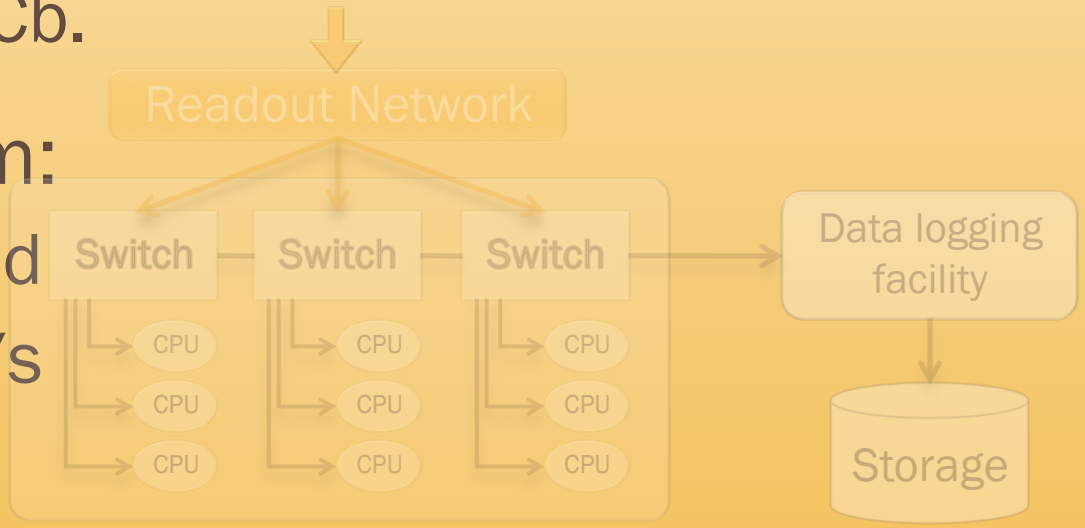
Data logging facility

Storage

LHCb ONLINE

The LHCb Online Farm: numbers

- ~16000 CPU cores foreseen (~1000 boxes)
  - Environmental constraints:
    - 2000 1U boxes space limit
    - 50 x 11 kW cooling/power limit
  - Computing power equivalent to that provided by all Tier 1's to LHCb.
- Storage system:
  - 40 TB installed
  - 400-500 MB/s



An opportunity

- Significant idle time of the farm
  - During LHC winter shutdown (~ months)
  - During beam period, experiment and machine downtime (~ hours)

↓

Could we use it for reconstruction?

- Farm is fully LHCb controlled
  - Good internal network connectivity
  - Slow disk access (only fast for a very few nodes via Fiber Channel interface)

Design considerations

- Background information:
  - 1 file (2GB) contains 60.000 events.
  - It takes 1-2s to reconstruct an event.
- Cannot reprocess *à la* Tier-1 (1 file per core)
  - Cannot perform reconstruction in short idle periods:
    - Each file takes 1-2 s/evt * 60k evt ~ 1 day.
  - Insufficient storage or CPUs not used efficiently:
    - Input: 32 TB (16000 files * 2 GB/file)
    - Output: ~44 TB (16000 * 60k evt * 50 kB/evt)
- A different approach is needed
  - Distributed reconstruction architecture.

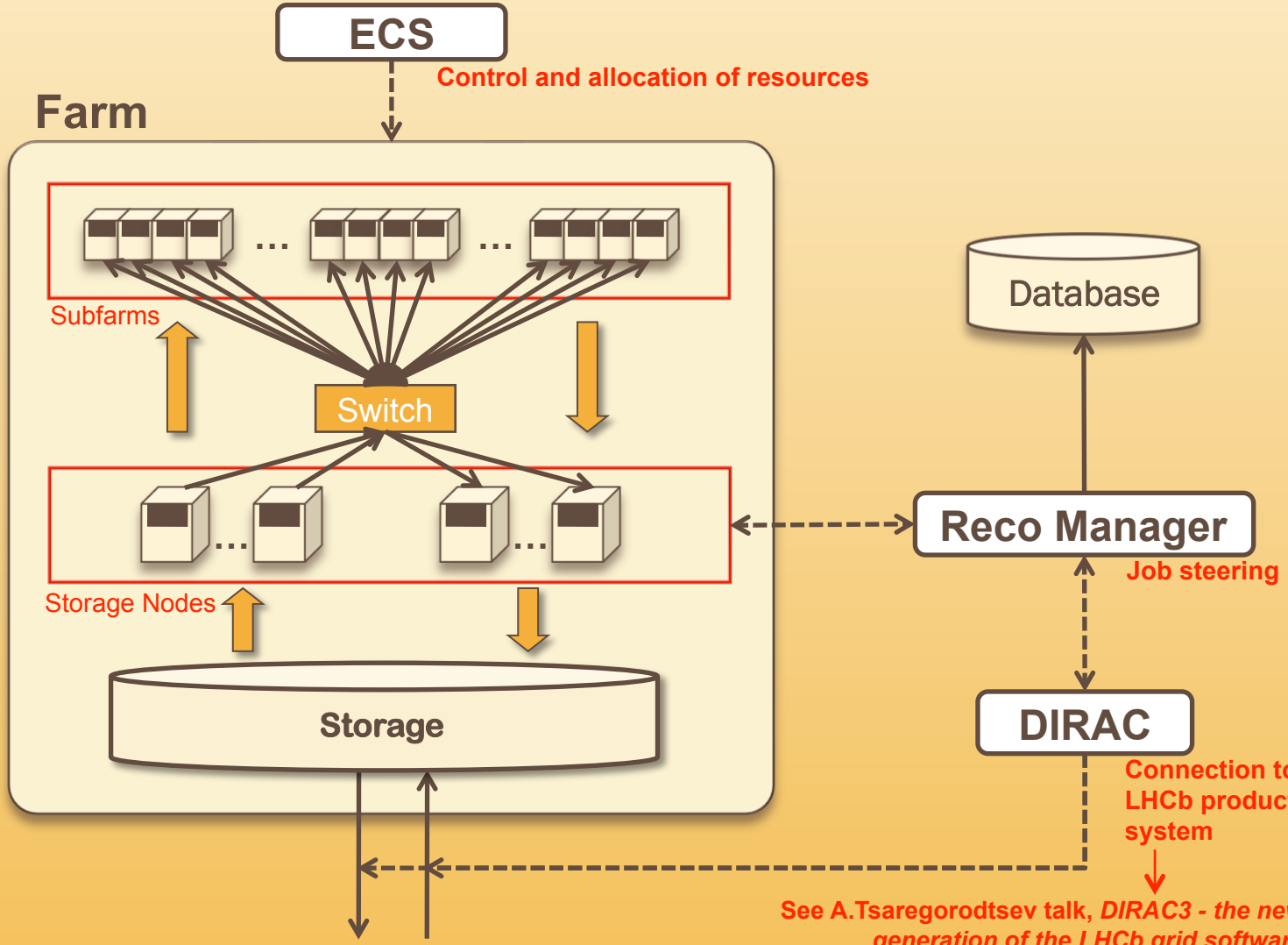
LHCb ONLINE

Adopted approach: parallel reconstruction

- Files are split in events and distributed to many cores, which perform reconstruction:
  - First idea: full parallelization (1 file/16k cores)
    - Reconstruction time: 4-8 s
    - Full speed not reachable (only one file open!)
  - Practical approach: split the farm in slices of subfarms (1 file/n subfarms).
    - Example: 4 concurrent open files yield a reconstruction time of 30s/file.

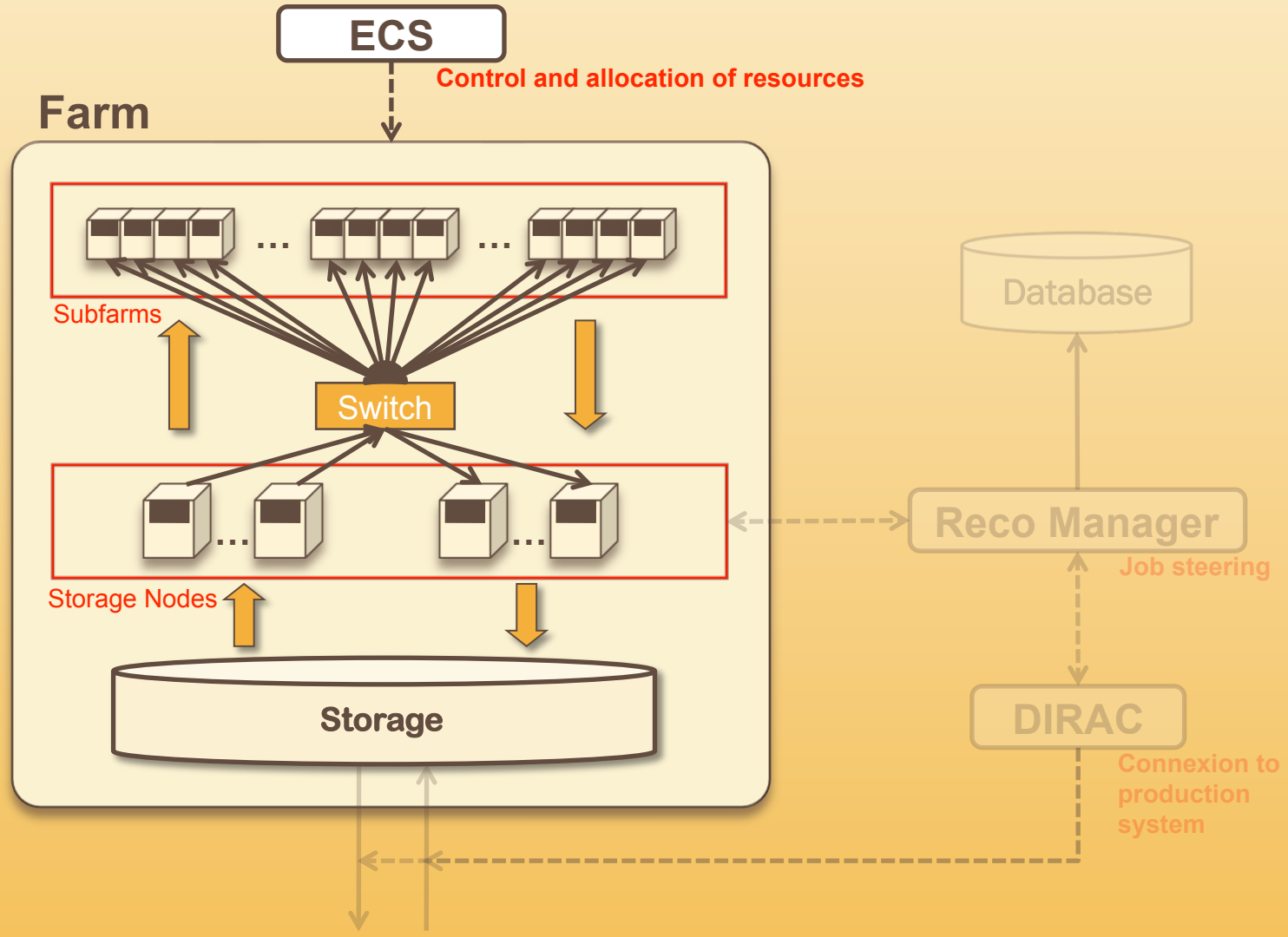
LHCb ONLINE

Implementation



LHCb ONLINE

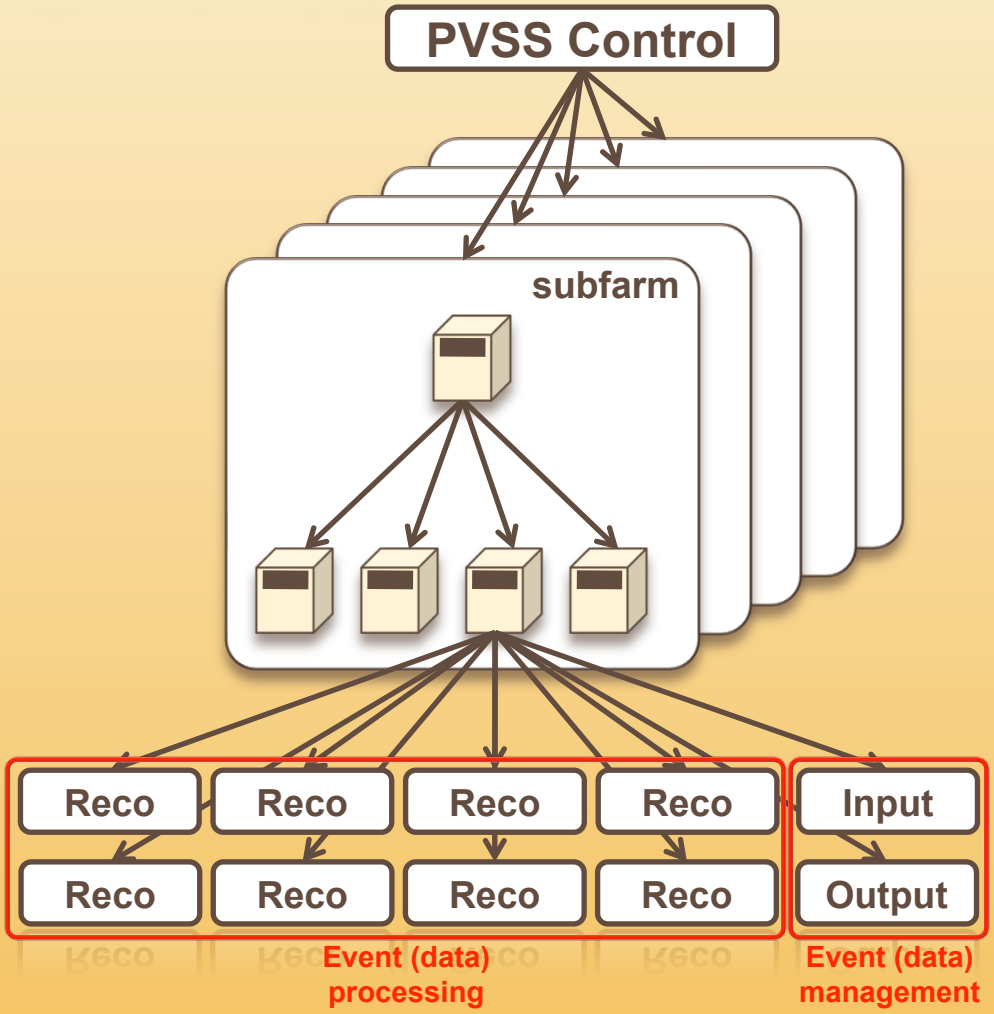
Resource management



Farm and process control

- Control using standard LHCb ECS software:
  - Reuse of existing components for storage and subfarms.
  - New components for reconstruction tree management.
  - See Clara Gaspar's talk (*LHCb Run Control System*).
- Allocate, configure, start/stop resources (storage and subfarms).
- Task initialization slow, so tasks don't restart on file change.
  - Idea: tasks sleep during data-taking, and are only restarted on configuration change.

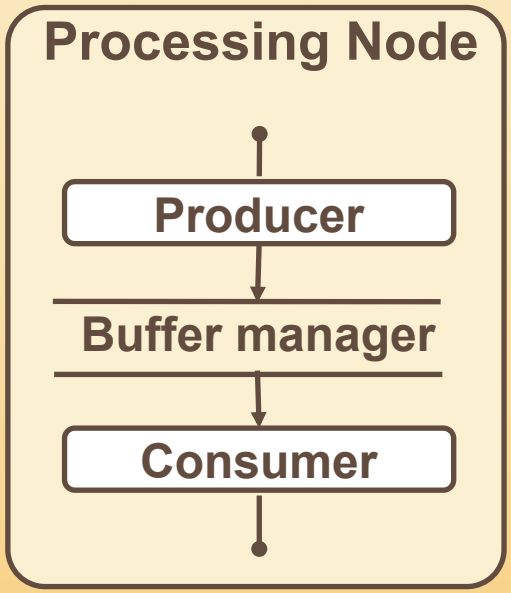
Farm Control Tree



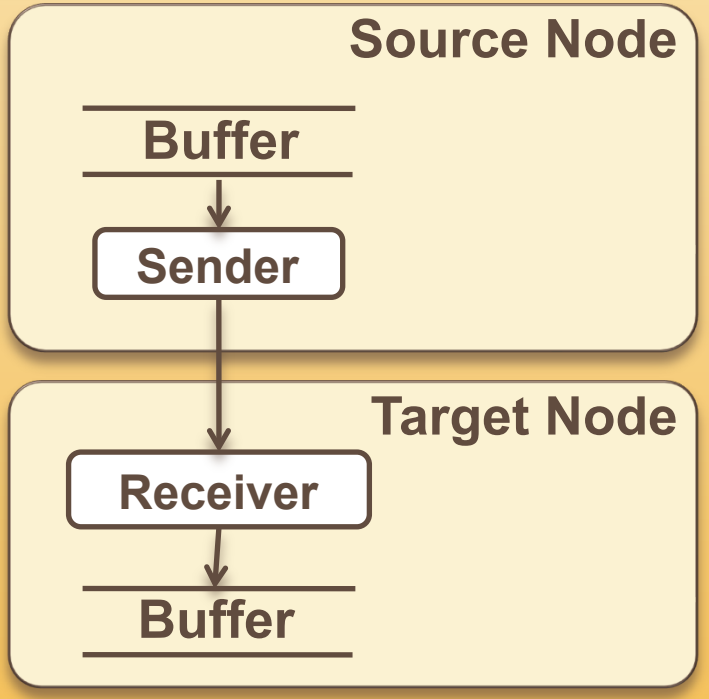
x 50 subfarms
- 1 control PC each
- 4 PC each

x 8 cores/PC
- 1 Reco task/core
- Data management tasks

The building blocks

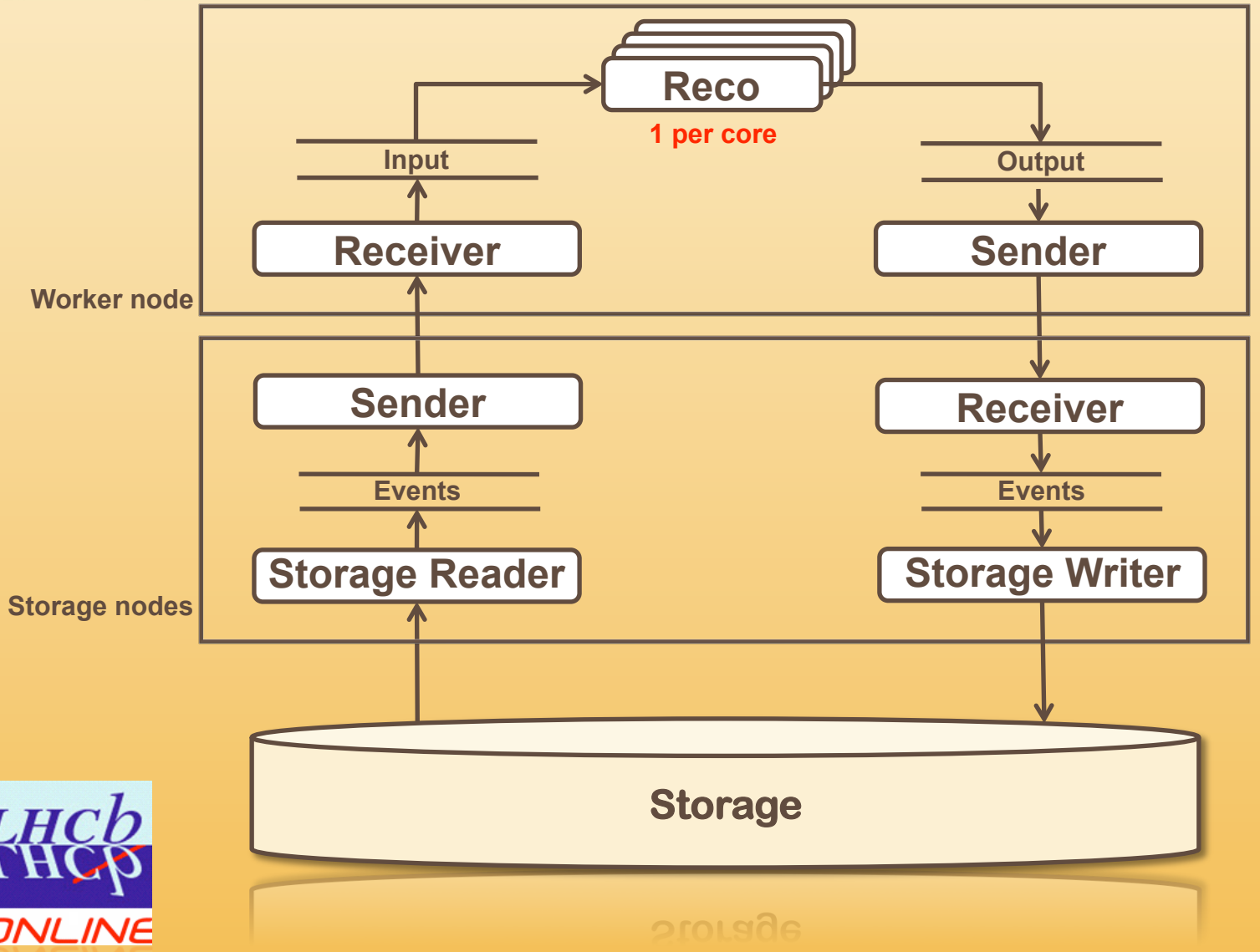


- Data processing block
  - Producers put events in a buffer manager (MBM)
  - Consumers receive events from the MBM

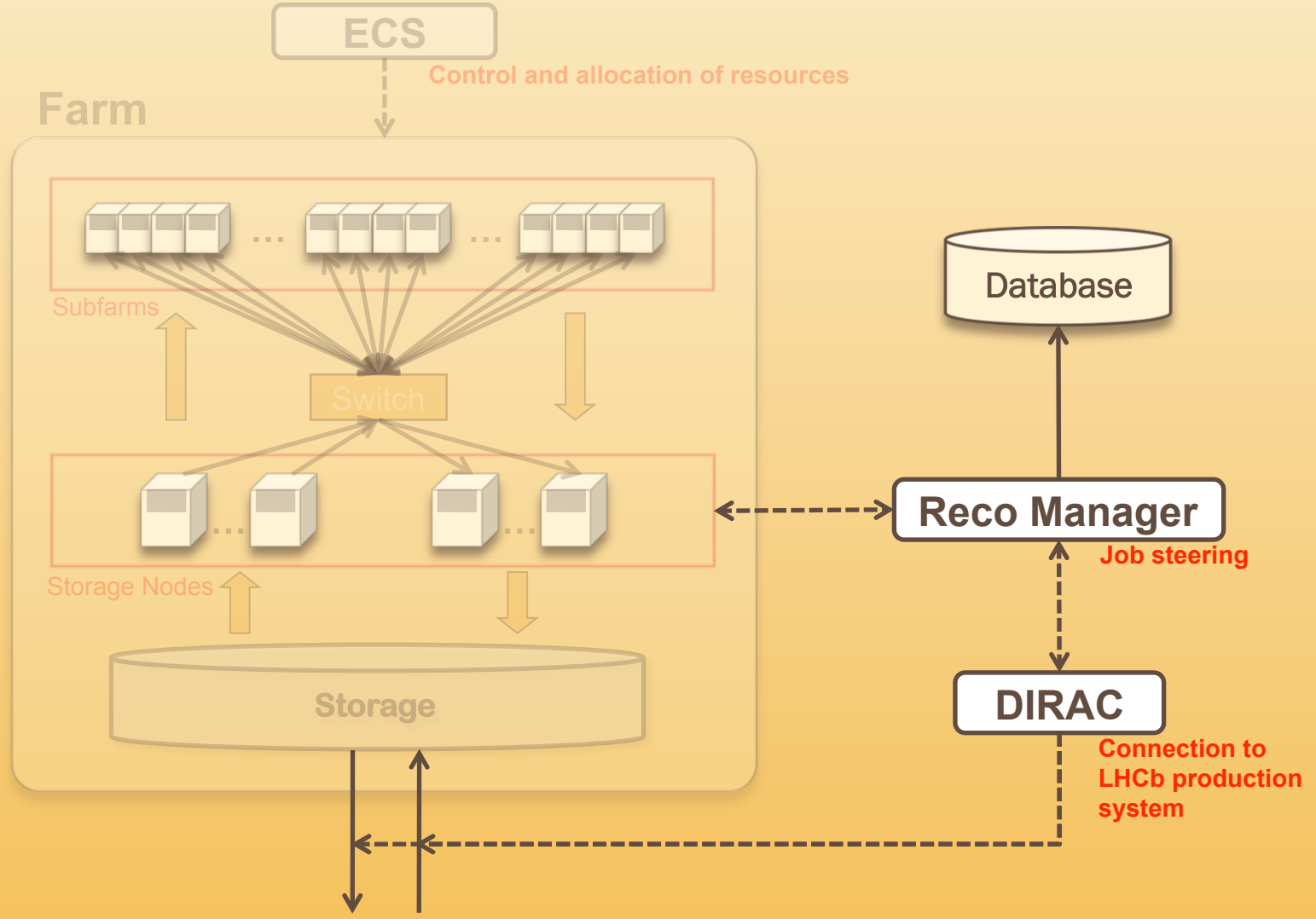


- Data transfer block
  - Senders access events from MBM
  - Receivers get data and declare it to MBM

Putting everything together…

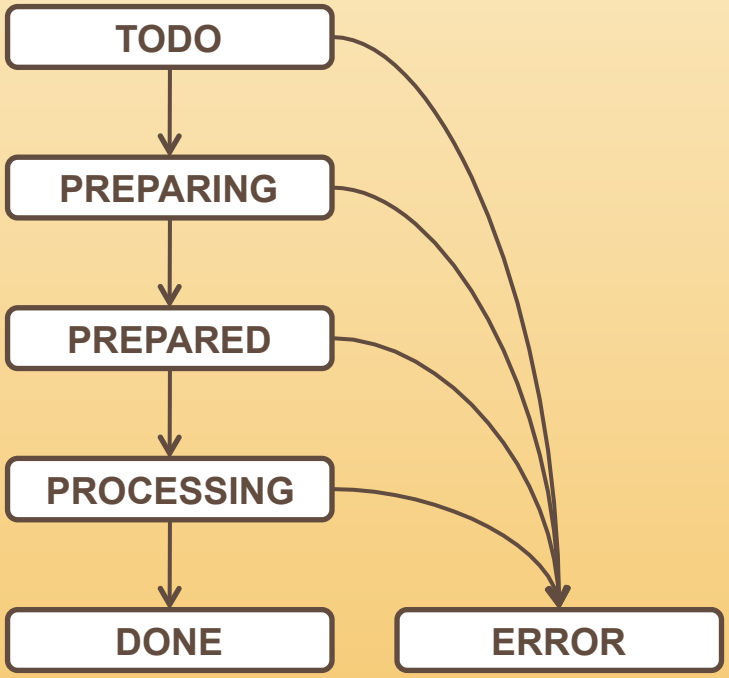


Reconstruction management



Task management

- Granularity to file level.
  - Individual event flow handled automatically by allocated resource slices.
- Reconstruction: specific actions in specific order.
  - Each file is treated as a Finite State Machine (FSM)
- Reconstruction information stored in a database.
  - System status
  - Protection against crashes



Job steering: the Reco Manager

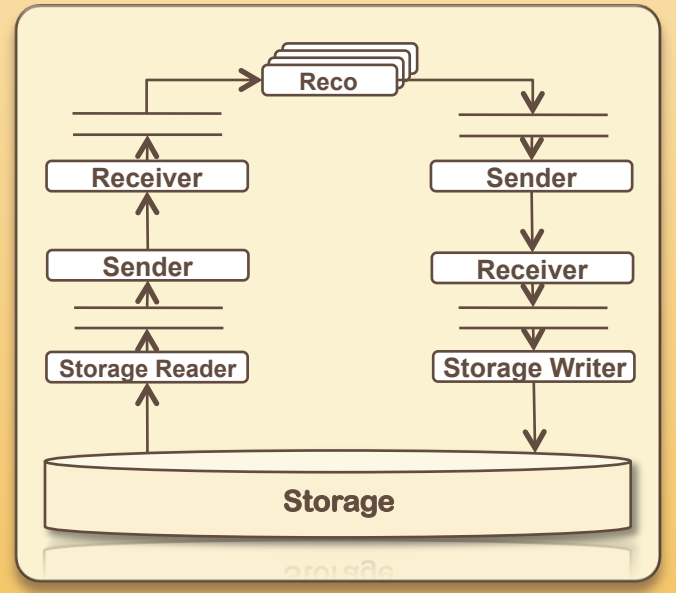
- Job steering done by a Reco Manager:
  - Holds the each FSM instance and moves it through all the states based on the feedback from the static resources.
  - Sends commands to the readers and writers: files to read and filenames to write to.
  - Interacts with the database.

Connection to the LHCb Production System

- The Online Farm will be treated like a CE connected to the LHCb Production system.
- Reconstruction is formulated as DIRAC jobs, and managed by DIRAC WMS Agents.
- DIRAC interacts with the Reco Manager through a thin client, not directly with the DB.
- Data transfer in and out of the Online farm managed by DIRAC DMS Agents.

Status and tests

- Resource handling and Reco Manager implemented.
- Integration in LHCb Production system recently decided, not implemented yet.
- Current performance constrained by hardware.
  - Reading from disk: ~130 MB/s.
    - FC saturated with 3 readers.
    - Reader saturates CPU at 45MB/s.
  - Test with dummy reconstruction
    - Just copy data from input to output
    - Stable throughput 105 MB/s
    - Constrained by Gigabit network
      - Upgrade to 10 Gigabit planned



LHCb ONLINE

Future plans

- Software
  - Pending implementation of thin client for interfacing with DIRAC.
- Hardware
  - Network upgrade to 10 Gbit in storage nodes before summer.
  - More subfarms and PCs to be installed.
    - From current ~4800 cores to the planned 16000.

Conclusions

- The LHCb Online cluster needs huge resources for data event selection of LHC collisions.
- These resources have much idle time (50% of the time).
- They can be used on idle periods by applying a parallelized architecture to data reprocessing.
- A working system is already in place, pending integration in the LHCb Production system.
- Planned hardware upgrades to meet DAQ requirements should overcome current bandwidth constraints.