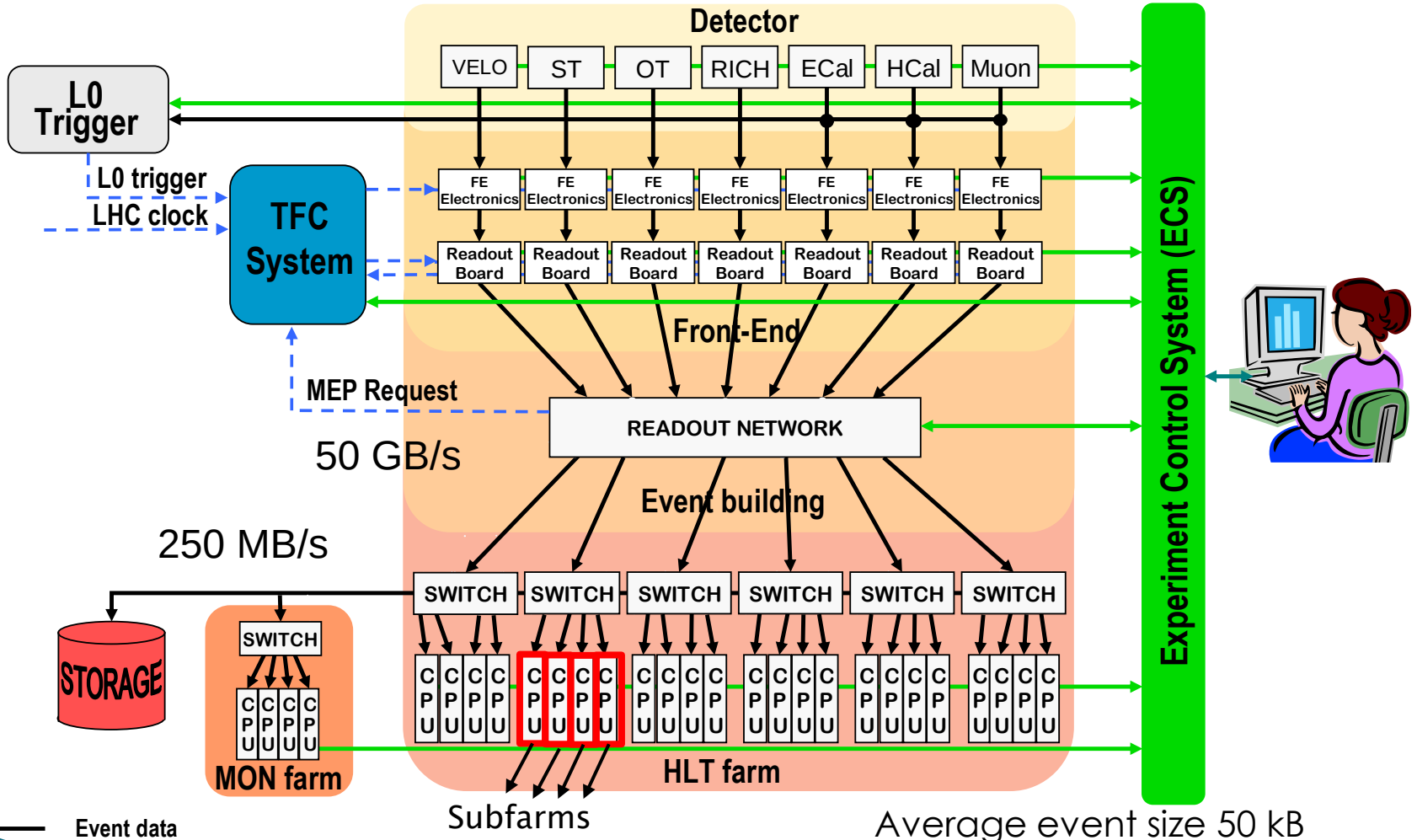
Offline Processing in the Online Computer Farm

CHEP 2012 – New York City

LHCb DAQ System

- LHC Delivers bunch crossing at 40MHz
- LHCb reduces the rate with a two level trigger system:
  - First Level (L0) – Hardware based – 40MHz >1MHz
  - Second Level (HLT) – Software based – 1MHz >5KHz
    - ~1500 Linux PCs
    - 16000+ cores
- Outside data taking period:
  - Little or no usage of the HLT Computer system
    - Low efficiency

LHCb DAQ System



Average event size 50 kB
Average rate into farm 1 MHz
Average rate to tape 5 kHz

LHCbDIRAC

- DIRAC System (Distributed Infrastructure with Remote Agent Control)
  - specialized system for data production, reconstruction and analysis
  - produced by HEP experiments
  - follows the Service Oriented Architecture
    - 4 categories of components:
      - Resources – provide access to computing and storage facilities
      - Agents – independent processes to fulfill one or several system functions
      - Services – help to carry out workload and data management tasks
      - Interfaces – programming interfaces (APIs)

LHCbDIRAC

- DIRAC Agents:
  - light and easy to deploy software components
  - run in different environments
  - watch for changes in the services and react:
    - job submission
    - result retrieval
  - can run as part of a job executed on a Worker Node
    - Called “Pilot Agents”

LHCbDIRAC

- Workload Management System (WMS)
  - Supports the data production
  - Task Queue
- Pilot Agents
  - Deployed close to the computing resources
  - Presented in a uniform way to the WMS
  - Check operational environment sanity
  - Pull the workload from the Task Queue
  - Process the data
  - Upload the data

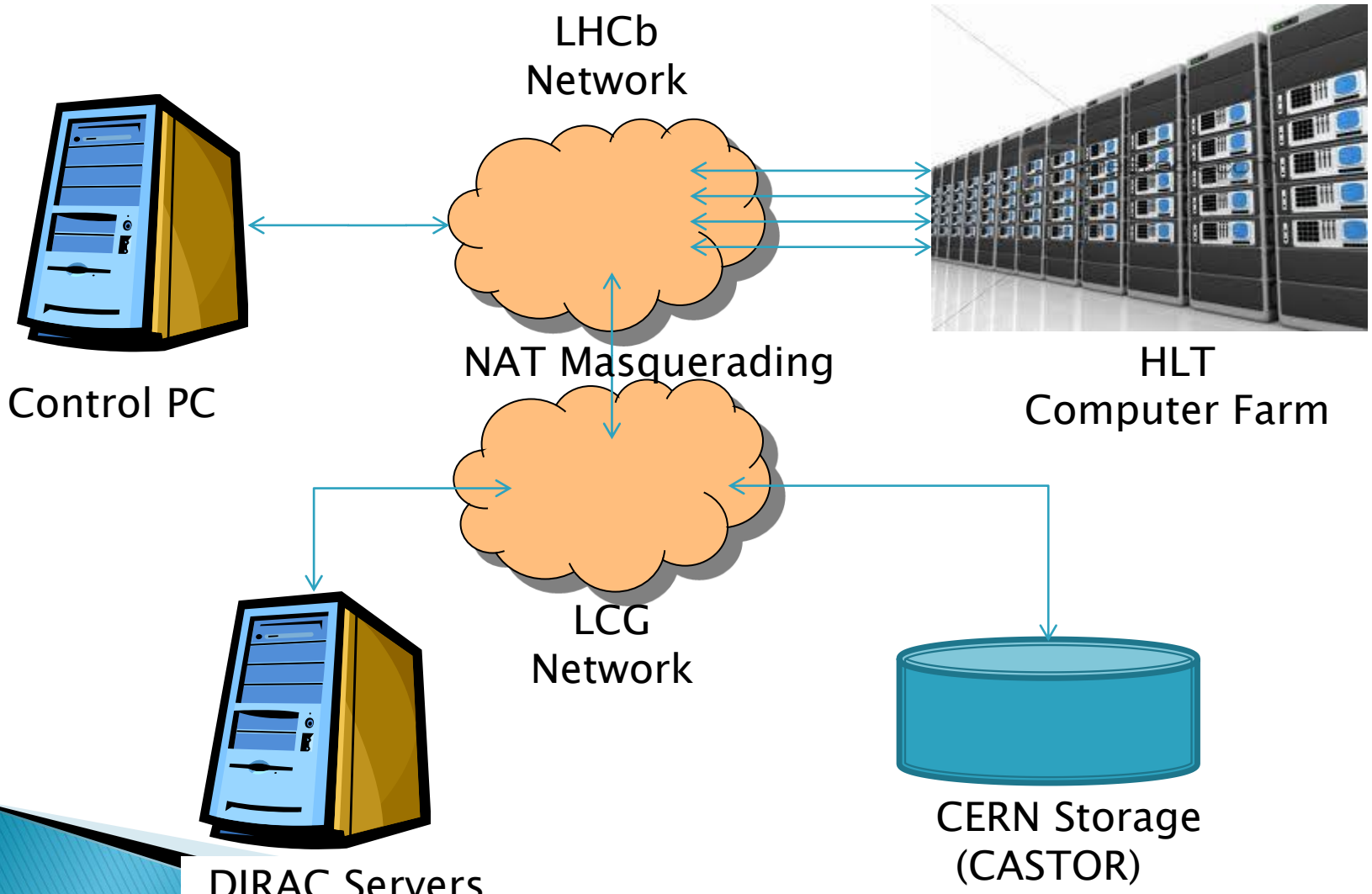
DIRAC on the Online Infrastructure

- Requirements
  - Allocate HLT subfarms to process data
    - Interface Experiment Control System
    - Does not interfere with normal data acquisition
  - Manage the start/stop of offline data processing
  - Balance workload on the allocated nodes
  - Easy User Interface

DIRAC on the Online Infrastructure

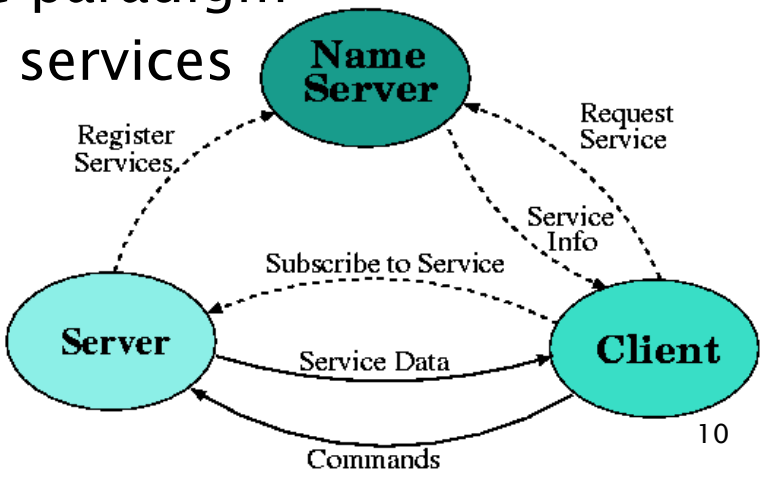
- Infrastructure
  - Composed of a Linux PVSS PC
- LHCb has a private network
  - HLT Worker Nodes have private addresses
  - HLT Worker Nodes not accessible from outside LHCb
  - Masquerade NAT deployed on the nodes which have access to the LHC Computing Grid (LCG) network
  - With masquerade NAT HLT Nodes are able to access data from DIRAC

DIRAC on the Online Infrastructure



DIRAC on the Online Infrastructure

- PVSS
  - Main SCADA System at CERN and LHCb
  - Provide the UI
  - Provides global interface to communications layer
- Communication Layer
  - Provided by DIM (Distributed Information Management)
    - Based on the publish/subscribe paradigm
    - Servers publish commands and services
    - Clients subscribe them



DIRAC on the Online Infrastructure

- Farm Monitoring and Control (FMC)
  - Tools to monitor and manage several parameters of the Worker Nodes
  - Task Manager Server
    - Runs on each of the HLT nodes
    - Publishes commands and services via DIM
    - Starts/stops processes on nodes remotely
    - Attributes to each started process a unique identifier (UTGID)

DIRAC on the Online Infrastructure

- DIRAC Script
  - Launches a Pilot Agent
    - Queries DIRAC WMS for tasks
    - Downloads data to local disk and processes it locally
    - Uploads data to CERN Storage (CASTOR)
    - During execution sends information to DIRAC

DIRAC on the Online Infrastructure

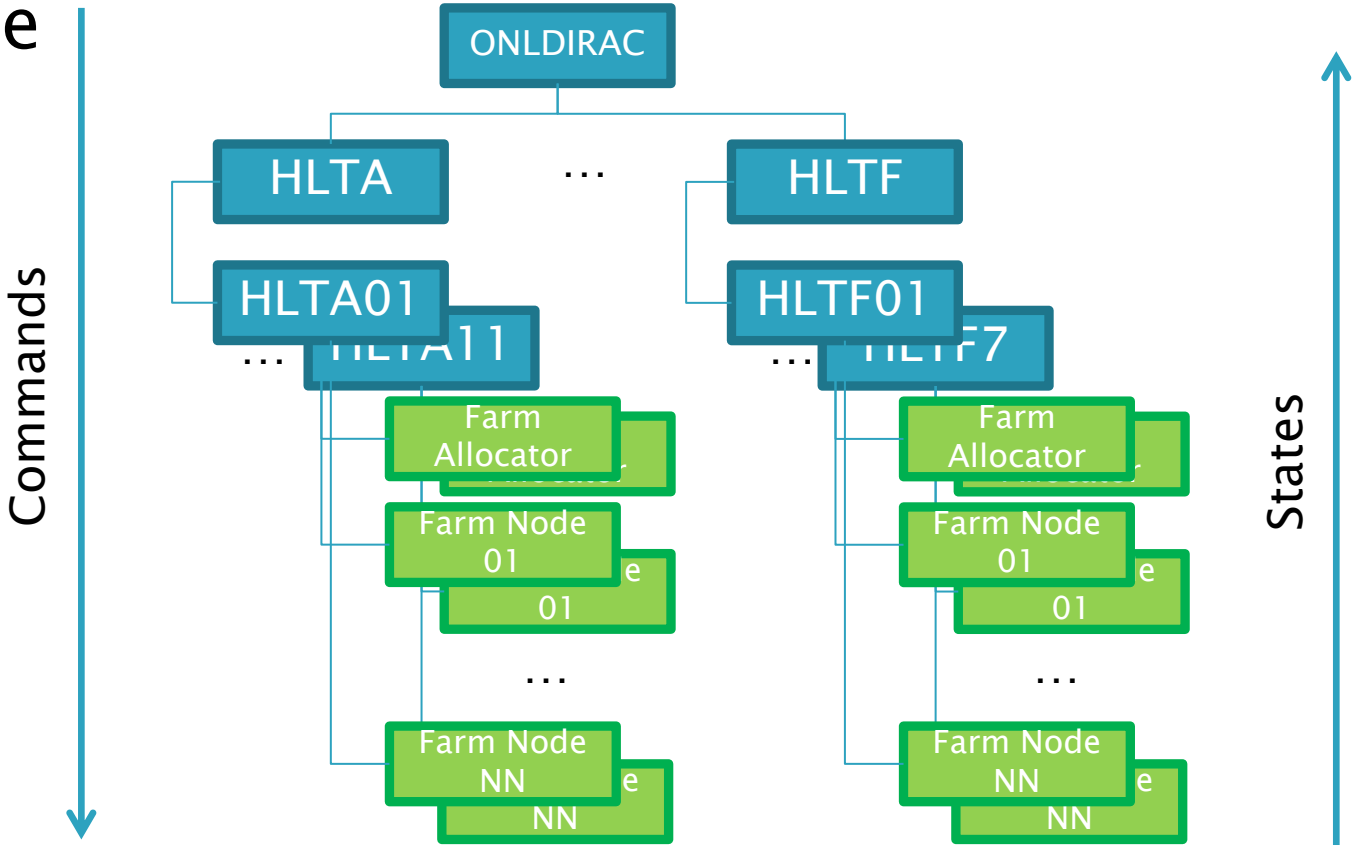
- PVSS Control Manager
  - Connects/disconnects monitoring of the worker nodes
  - Manages startup of DIRAC Agents
  - Balances the load on the allocated farms
  - Monitors the connection of the required DIM services

DIRAC on the Online Infrastructure

- Finite State Machine (FSM)
  - Control System interfaced by a FSM
  - Interfaces PVSS
  - Composed of:
    - Device Units (DUs) – model real devices in the control tree
    - Control Units (CUs) – group DUs and CUs in logical useful segments

DIRAC on the Online Infrastructure

- FSM Tree



DIRAC on the Online Infrastructure

- How it works
  - Allocate subfarm(s)
  - Set Nodes to 'RUN' (GOTO_RUN)
  - DIRAC Script is launched on the worker nodes
    - Delay between launches (DB connections management)
  - Scripts are launched according to pre-defined rules for load balancing
  - Variable delay between process launches
    - No jobs -> longer delays

DIRAC on the Online Infrastructure

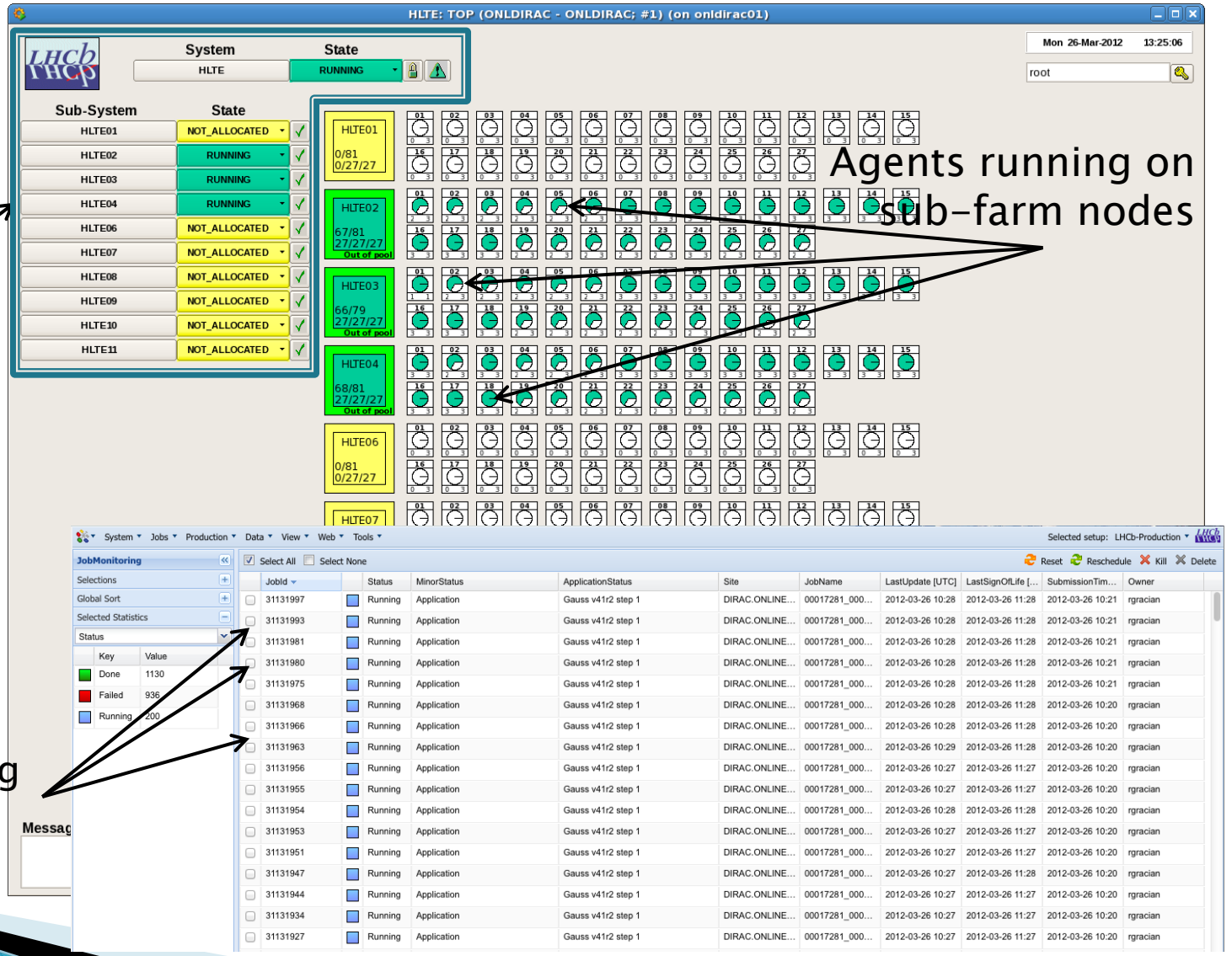
- Granularity:
  - Subfarm – a whole subfarm needs to be allocated
  - Can define only some nodes on the farm to process data
- CPU checks
  - Check what type of CPU is available on the node
    - Set max number of processes accordingly
  - Set max number of nodes independently
- Information exchange with DIRAC
  - Processing state information available only on the DIRAC system
  - Job availability evaluated by agent process duration

DIRAC on the Online Infrastructure

- User Interface (UI)
  - Developed in PVSS
  - Coherent look and feel with the LHCb Control Software
  - Use of synoptic widgets
  - Provides simple statistics:
    - Number of allocated farms/nodes
    - Number of agents running
    - Agents running time

DIRAC on the Online Infrastructure

- UI



FSM Control

Agents monitoring on DIRAC

Conclusion

- Efficiency of online farm usage improved
- Interfaced with DAQ
  - Does not interfere with Data Acquisition needs
- Resource balancing
  - Processing is balanced according to pre-defined rules
- Easy adoption
  - Maintains a coherent Look and Feel with other LHCb control software

Backup Slides

DIRAC on the Online Infrastructure

- State Diagrams