

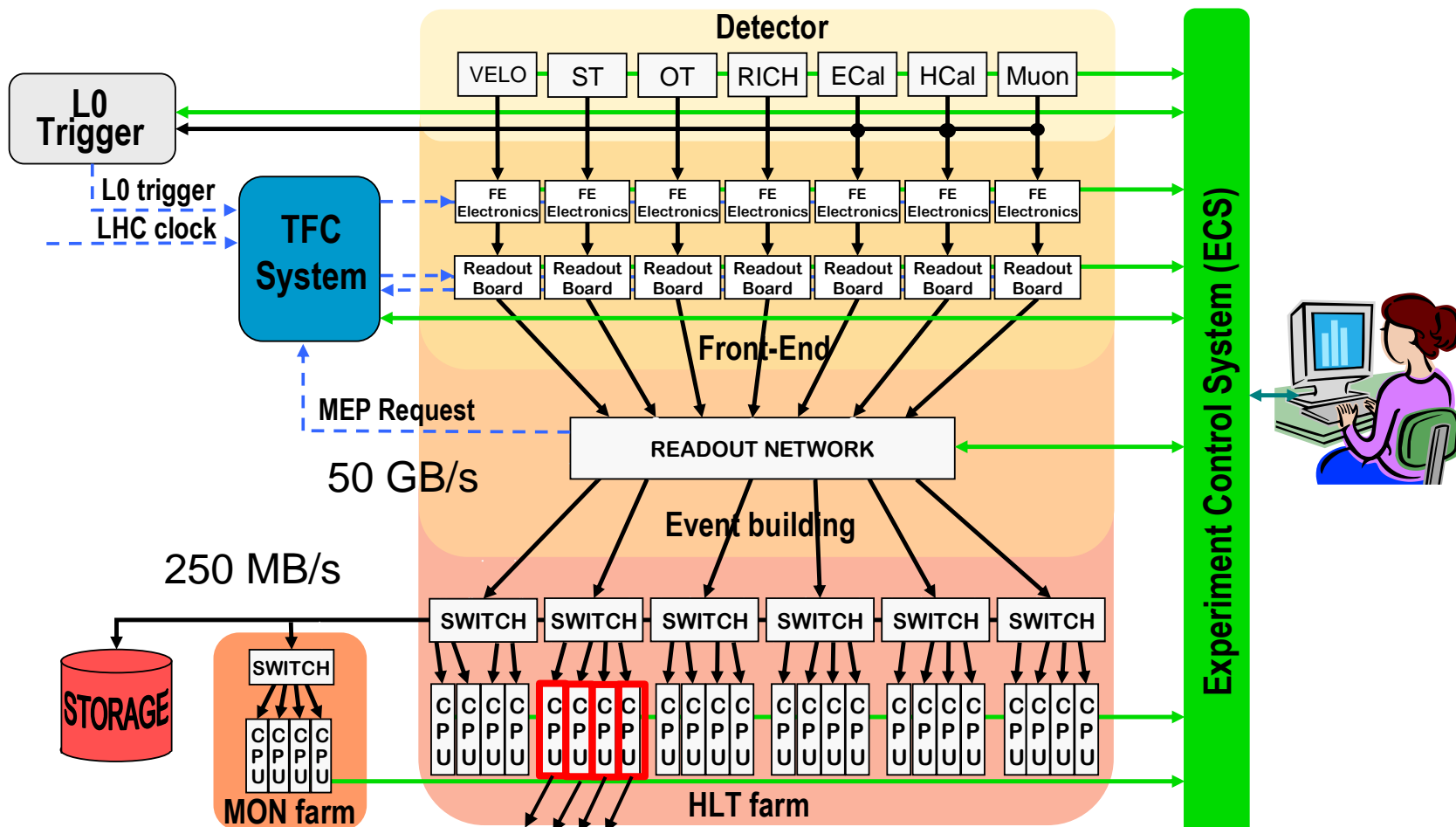
Offline Processing in the Online Computer Farm

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 – $40\text{MHz} > 1\text{MHz}$
 - Second Level (HLT) – Software based – $1\text{MHz} > 5\text{KHz}$
 - ~1500 Linux PCs
 - 16000+ cores
- ▶ Outside data taking period:
 - Little or no usage of the HLT Computer system
 - Low efficiency

LHCb DAQ System



— Event data
 - - - Timing and Fast Control Signals
 — Control and Monitoring data

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

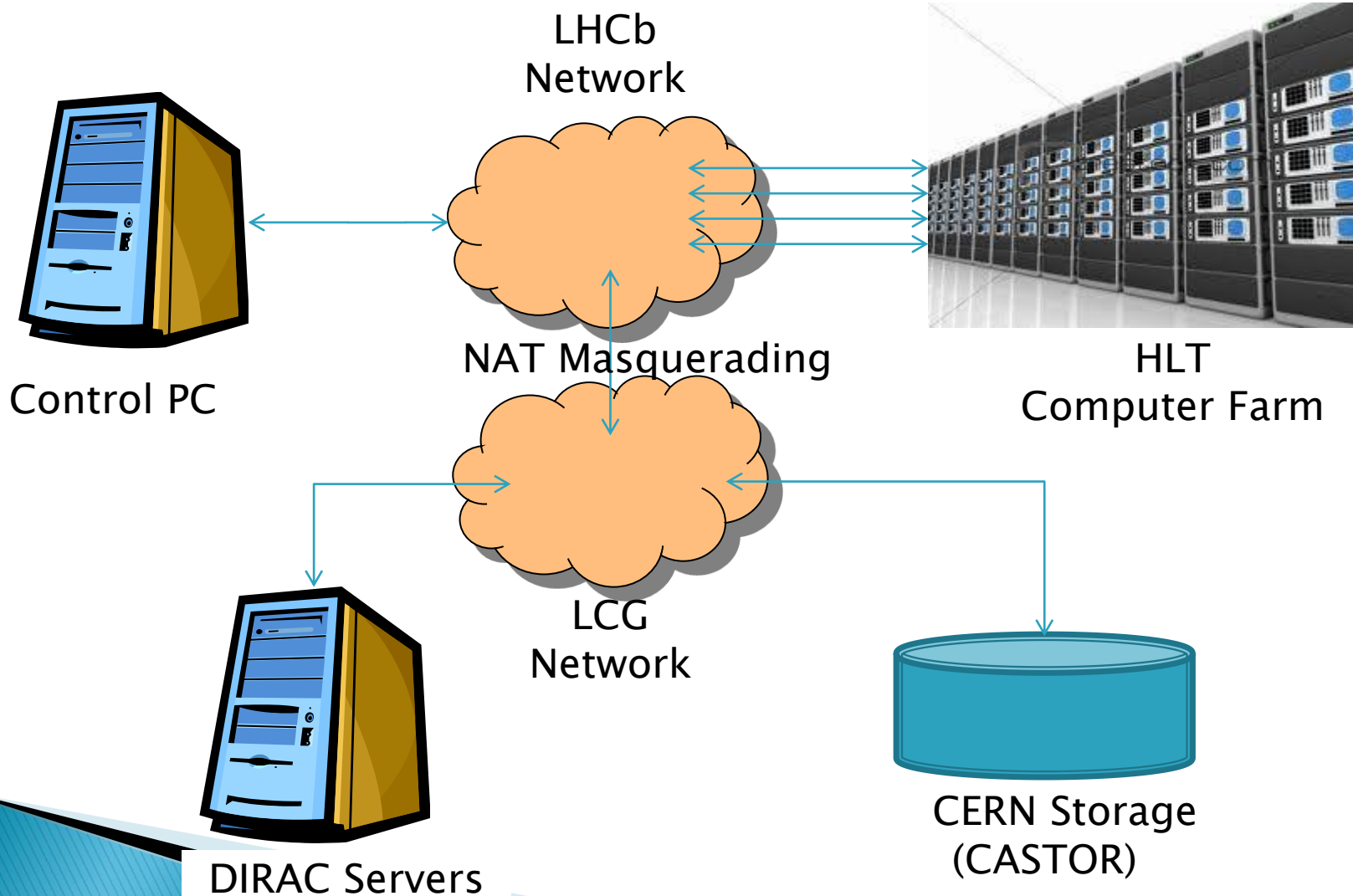
▶ Requirements

- 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



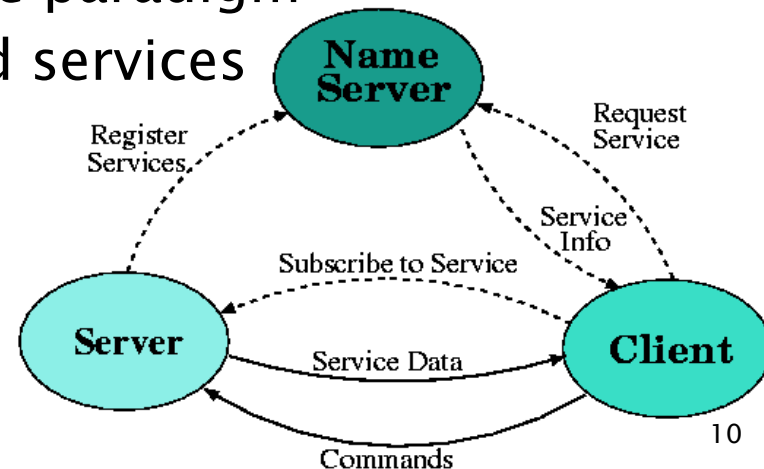
DIRAC on the Online Infrastructure

▶ 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

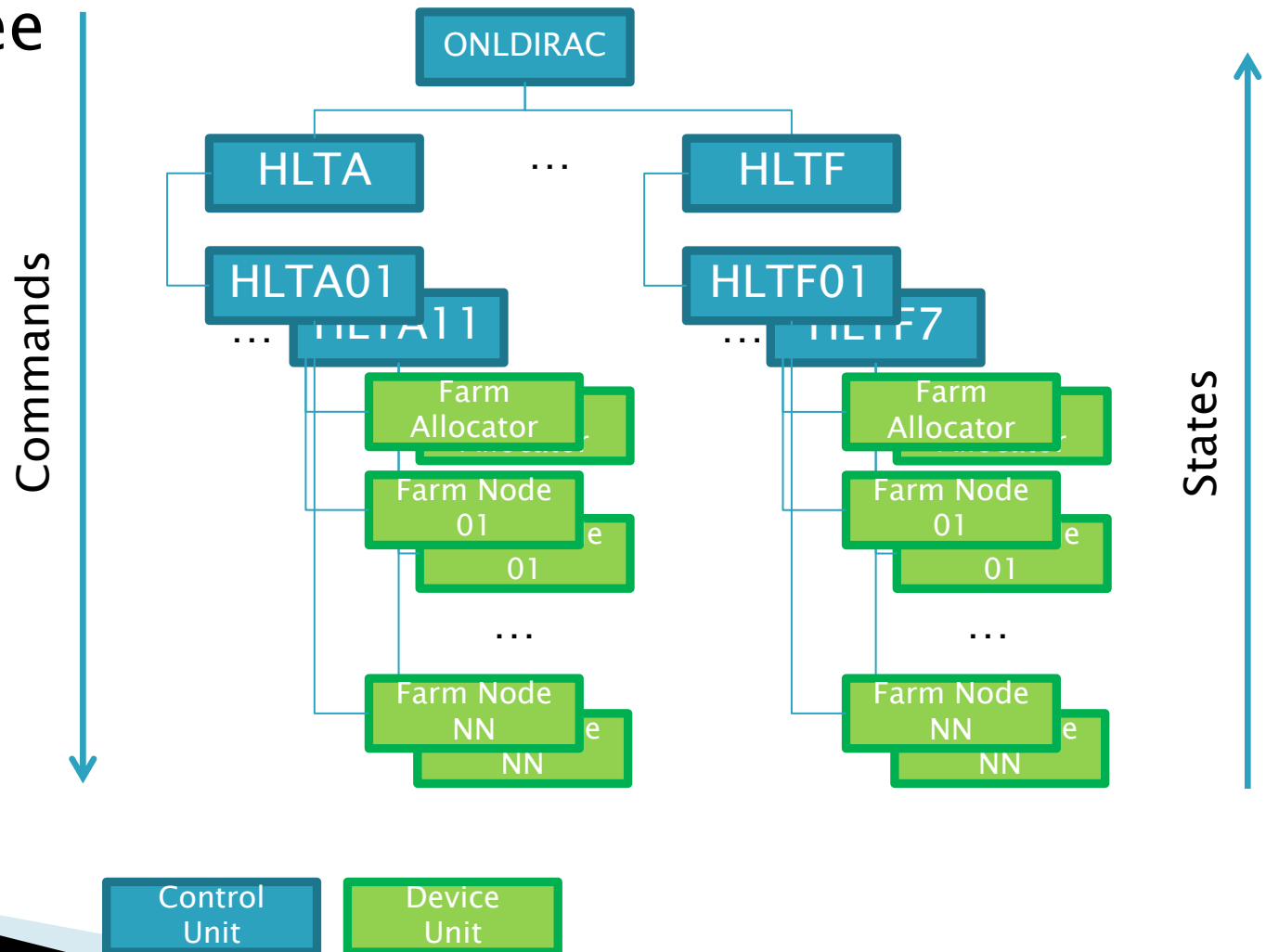
- ▶ PVSS Control Manager
 - Connects/disconnects monitoring of the worker nodes
 - Manages startup of DIRAC Agents
 - Balances the load on the allocated farms
 - Monitors the connection of the required DIM services

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

The screenshot displays the DIRAC web interface. At the top, the title bar reads "HLTE: TOP (ONLDIRAC - ONLDIRAC; #1) (on onldirac01)". The main content area is divided into two sections. The upper section, titled "System", shows the overall state as "RUNNING". Below this is a table of sub-systems:

Sub-System	State
HLTE01	NOT_ALLOCATED
HLTE02	RUNNING
HLTE03	RUNNING
HLTE04	RUNNING
HLTE06	NOT_ALLOCATED
HLTE07	NOT_ALLOCATED
HLTE08	NOT_ALLOCATED
HLTE09	NOT_ALLOCATED
HLTE10	NOT_ALLOCATED
HLTE11	NOT_ALLOCATED

Below the sub-system table is a grid of 15 columns and 15 rows of circular icons representing agents. Some icons are highlighted in green, indicating they are running. Callouts for these agents show their IDs and dates: HLTE01 (0/81 0/27/27), HLTE02 (67/81 27/27/27), HLTE03 (66/79 27/27/27), HLTE04 (68/81 27/27/27), HLTE06 (0/81 0/27/27), and HLTE07. A large black arrow points from the text "Agents running on sub-farm nodes" to the green-highlighted agent icons.

The lower section of the interface is titled "JobMonitoring" and contains a table of jobs:

JobId	Status	MinorStatus	ApplicationStatus	Site	JobName	LastUpdate [UTC]	LastSignOfLife [...]	SubmissionTim...	Owner
31131997	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:21	rrgracian
31131993	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:21	rrgracian
31131981	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:21	rrgracian
31131980	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:21	rrgracian
31131975	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:21	rrgracian
31131968	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:20	rrgracian
31131966	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:20	rrgracian
31131963	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:29	2012-03-26 11:28	2012-03-26 10:20	rrgracian
31131956	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131955	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131954	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:28	2012-03-26 11:28	2012-03-26 10:20	rrgracian
31131953	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131951	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131947	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:28	2012-03-26 10:20	rrgracian
31131944	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131934	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian
31131927	Running	Application	Gauss v41r2 step 1	DIRAC.ONLINE...	00017281_000...	2012-03-26 10:27	2012-03-26 11:27	2012-03-26 10:20	rrgracian

Arrows from the text "Agents monitoring on DIRAC" point to the "JobMonitoring" table. A "Message" box is visible at the bottom left of the interface.

FSM Control

Agents running on sub-farm nodes

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

