THE ALICE WORKLOAD MANAGEMENT SYSTEM: STATUS BEFORE THE REAL DATA TAKING
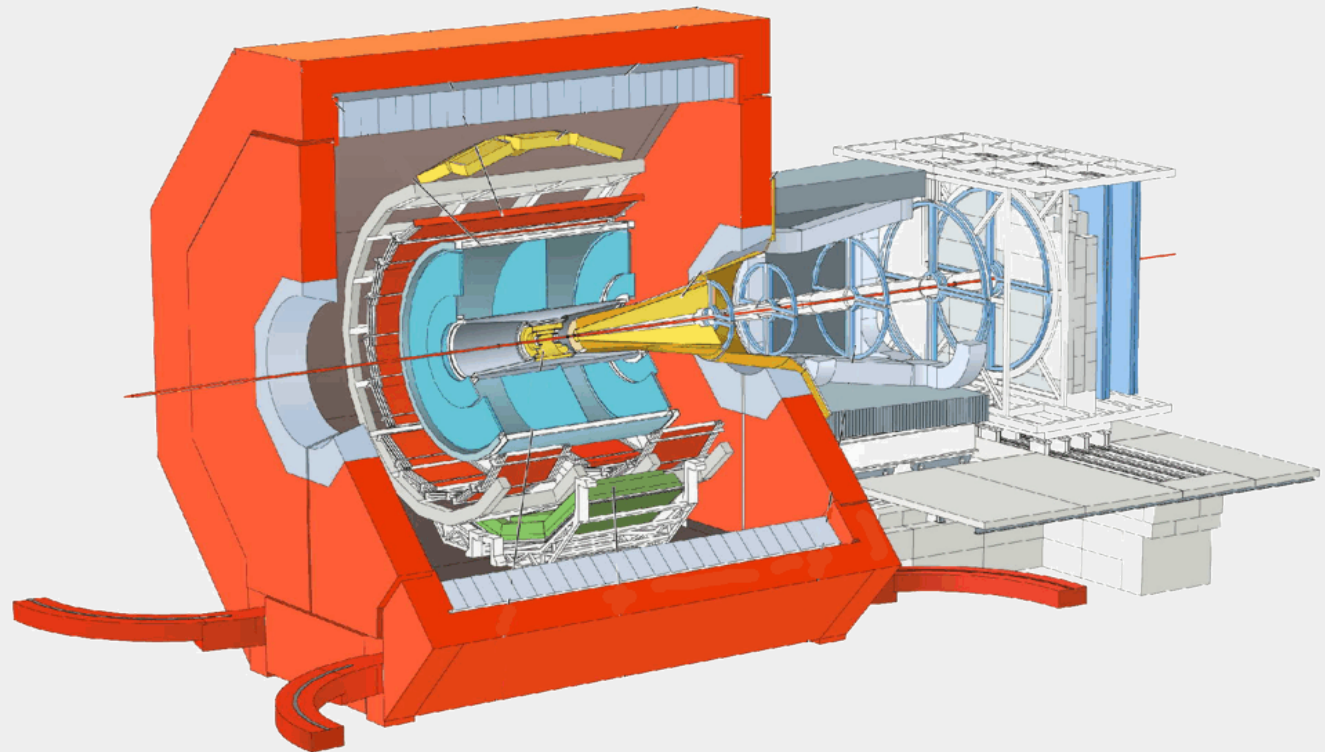
Stefano Bagnasco

INFN Torino

On behalf of the ALICE AliEn team

- Job Management model in ALICE
- Issues with gLite 3.1 WMS
- Experience with CREAM
- Conclusions

- Task Queue and Optimizers
    - Central DB of jobs to be executed
    - Optimizers split and arrange jobs according to input data, priority policies and/or user defined criteria
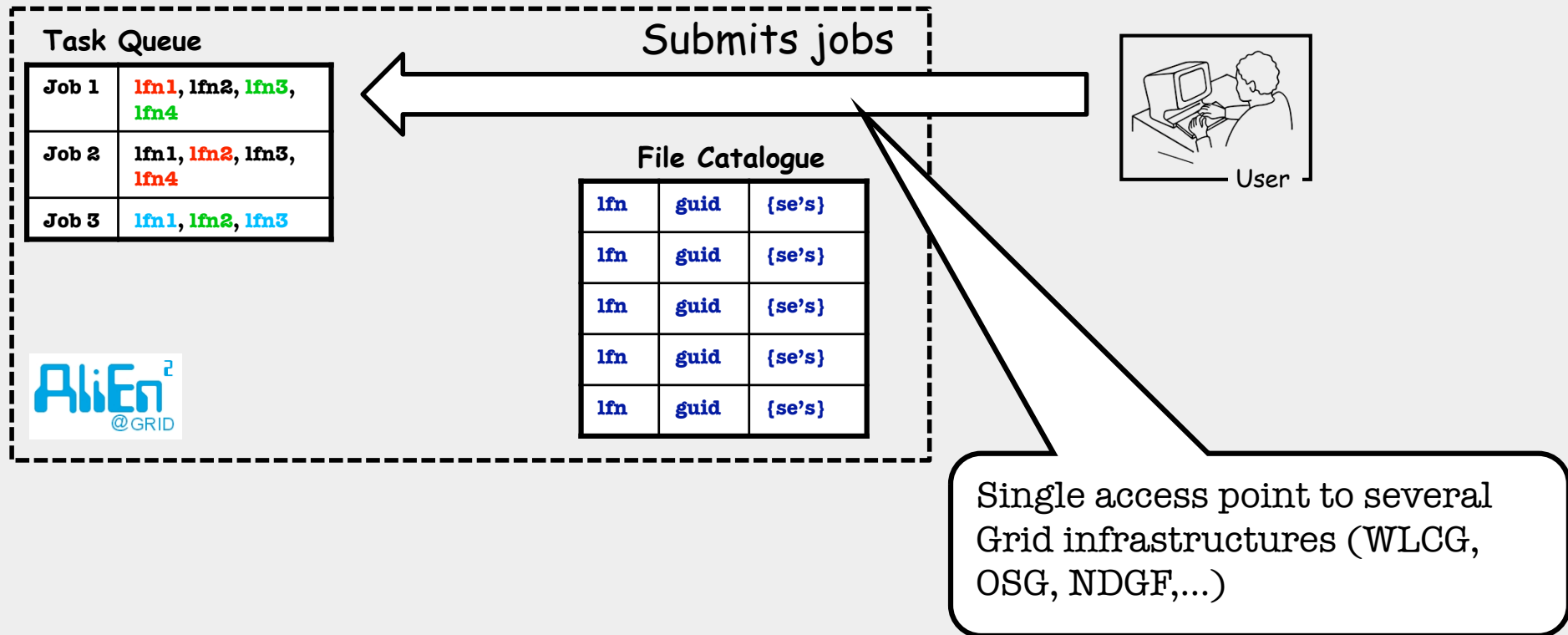
- Site VO-Box
    - Thin interfaces to underlying Grid site services
    - Submits JobAgents to site
    - Takes care of proxy management

- JobAgent
    - Runs on WNs, downloads payload from the TQ and executes it
    - JAs create a "virtual grid" on top of existing Grid infrastructures

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

Optimizer

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

User

- Split master jobs into subjobs according to data locality and user-defined criteria
- Sort jobs according to priority policies
- Enforce fair share and othe rpolicies

- ## Task Queue and Optimizers

  - Central DB of jobs to be executed
  - Optimizers split and arrange jobs according to input data, priority policies and/or user defined criteria

- ## Site VO-Box

  - Thin interfaces to underlying Grid site services
  - Submits JobAgents to site
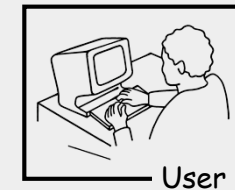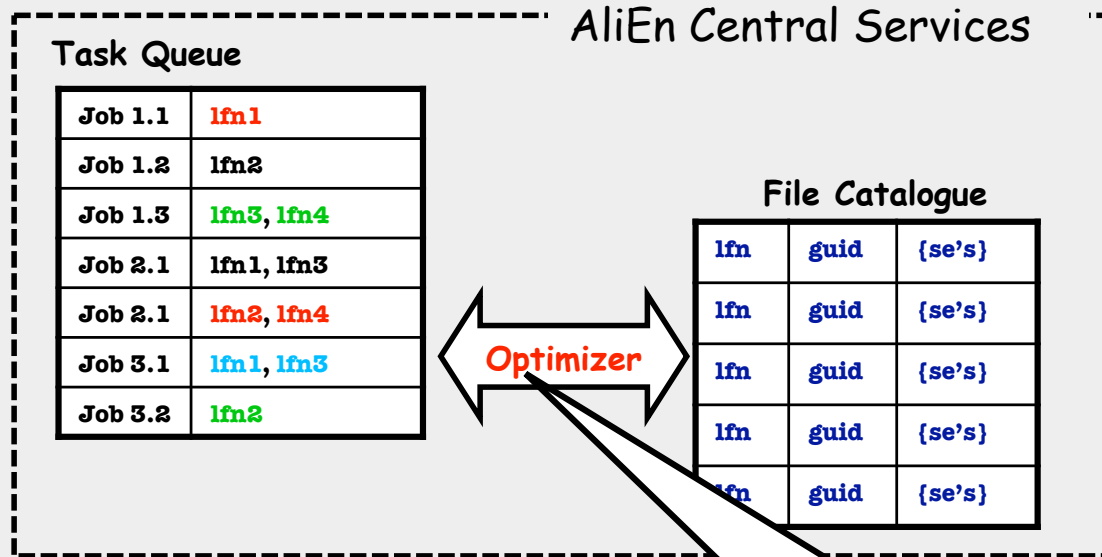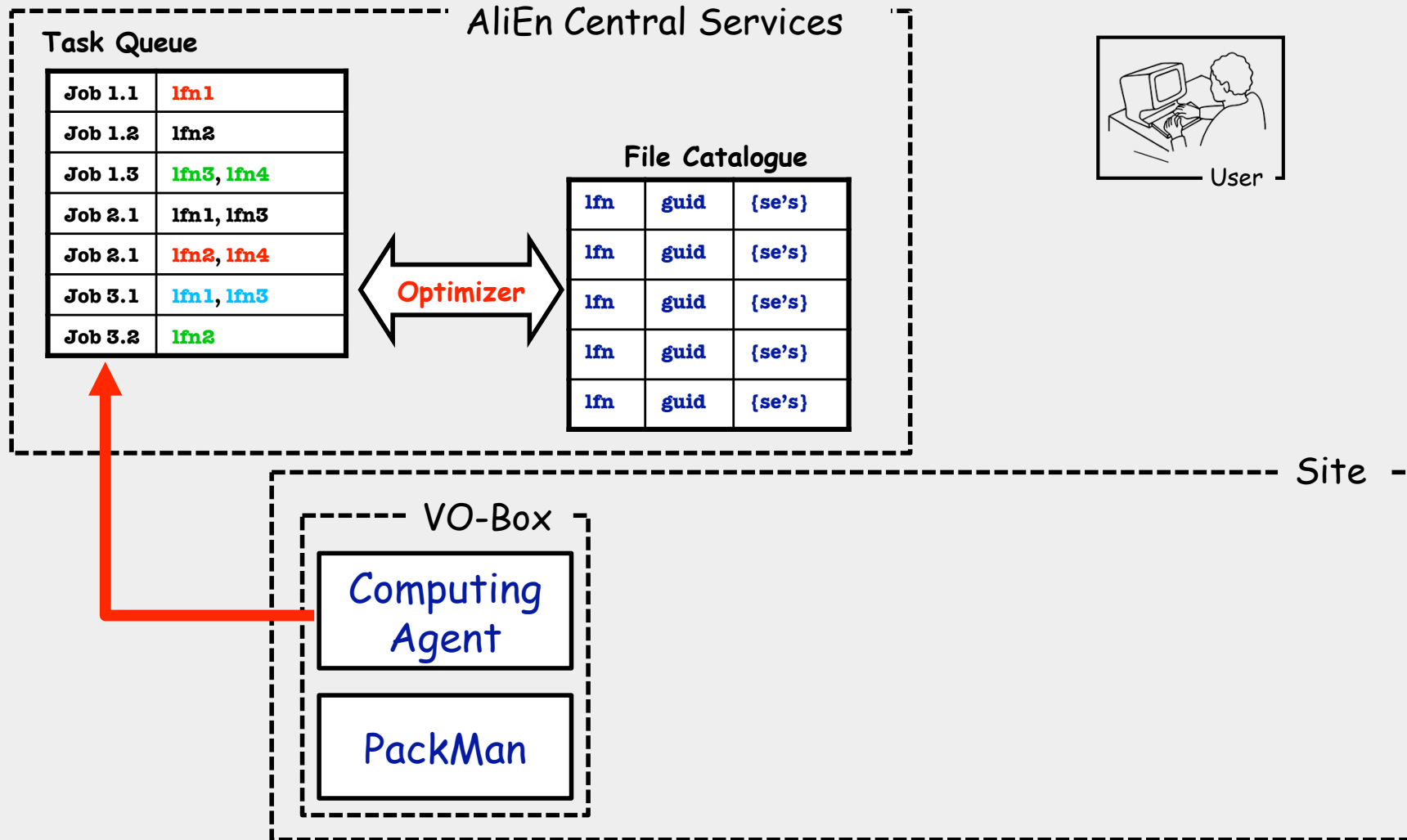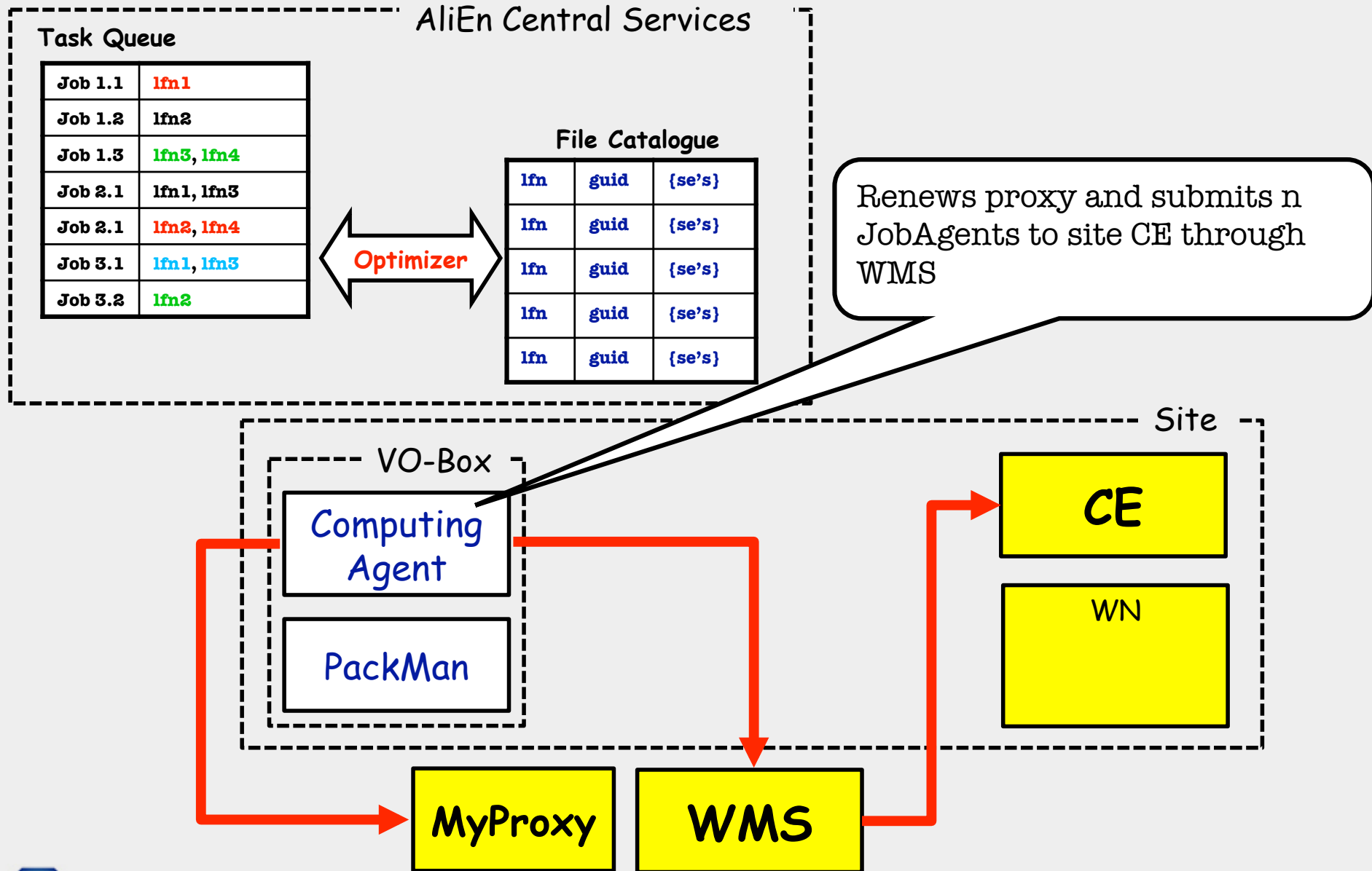 - Takes care of proxy management

- ## JobAgent

  - Runs on WNs, downloads payload from the TQ and executes it
  - JAs create a "virtual grid" on top of existing Grid infrastructures

ALICE JOB MANAGEMENT III

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

Optimizer

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

Renews proxy and submits n JobAgents to site CE through WMS

Site

VO-Box

Computing Agent

PackMan

CE

WN

MyProxy

WMS

ALICE JOB MANAGEMENT V

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

Optimizer

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

Complex configuration:

- Multiple CE per site for redundancy and aggregation
- Multiple WMS in failover mechnism
- WMS groups for load-balancing

Site

VO-Box

Computing Agent

PackMan

CE

WN

MyProxy

WMS

- ## Task Queue and Optimizers
  - Central DB of jobs to be executed
  - Optimizers split and arrange jobs according to input data, priority policies and/or user defined criteria

- ## Site VO-Box
  - Thin interfaces to underlying Grid site services
  - Submits JobAgents to site
  - Takes care of proxy management

- ## JobAgent
  - Runs on WNs, downloads payload from the TQ and executes it
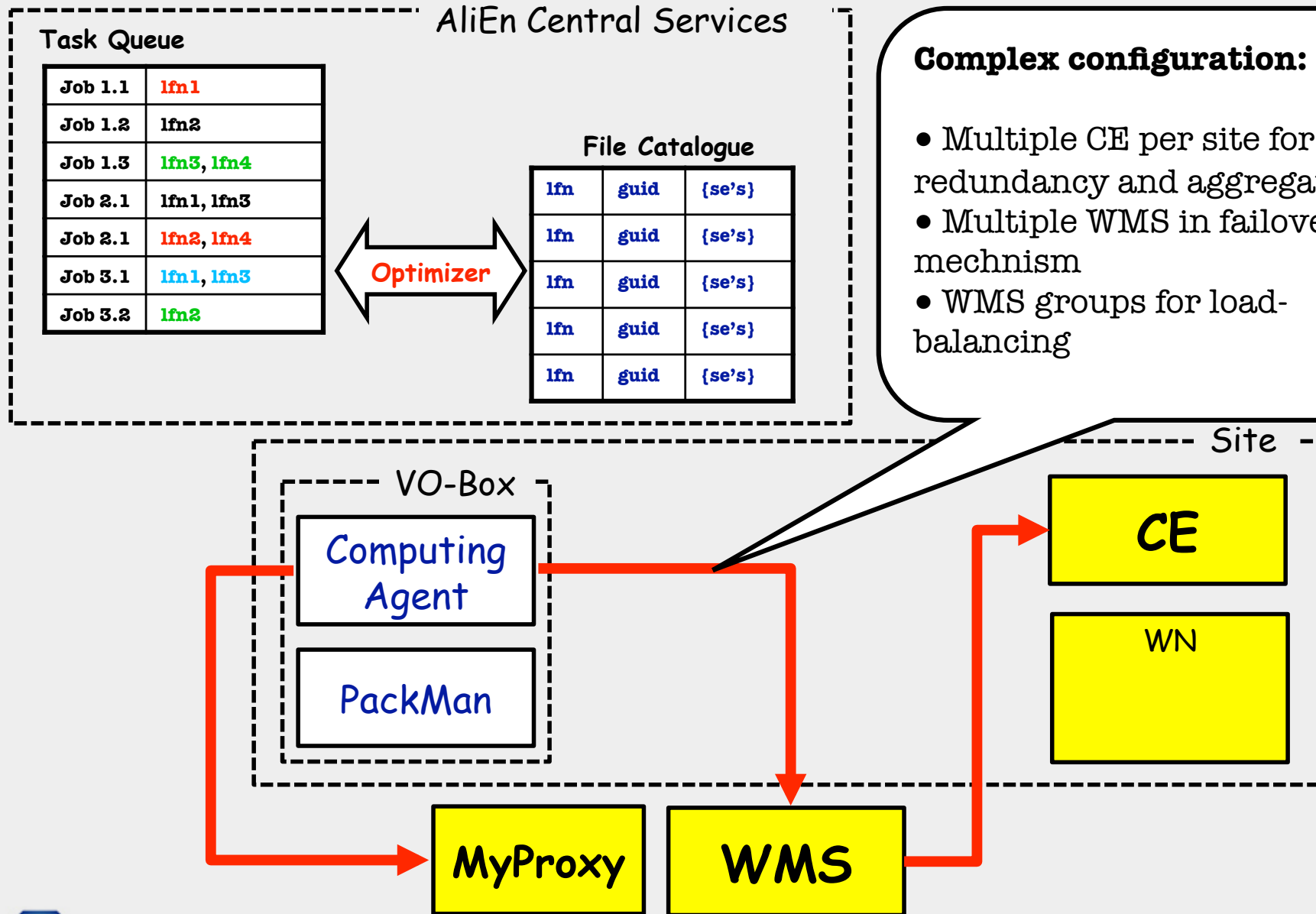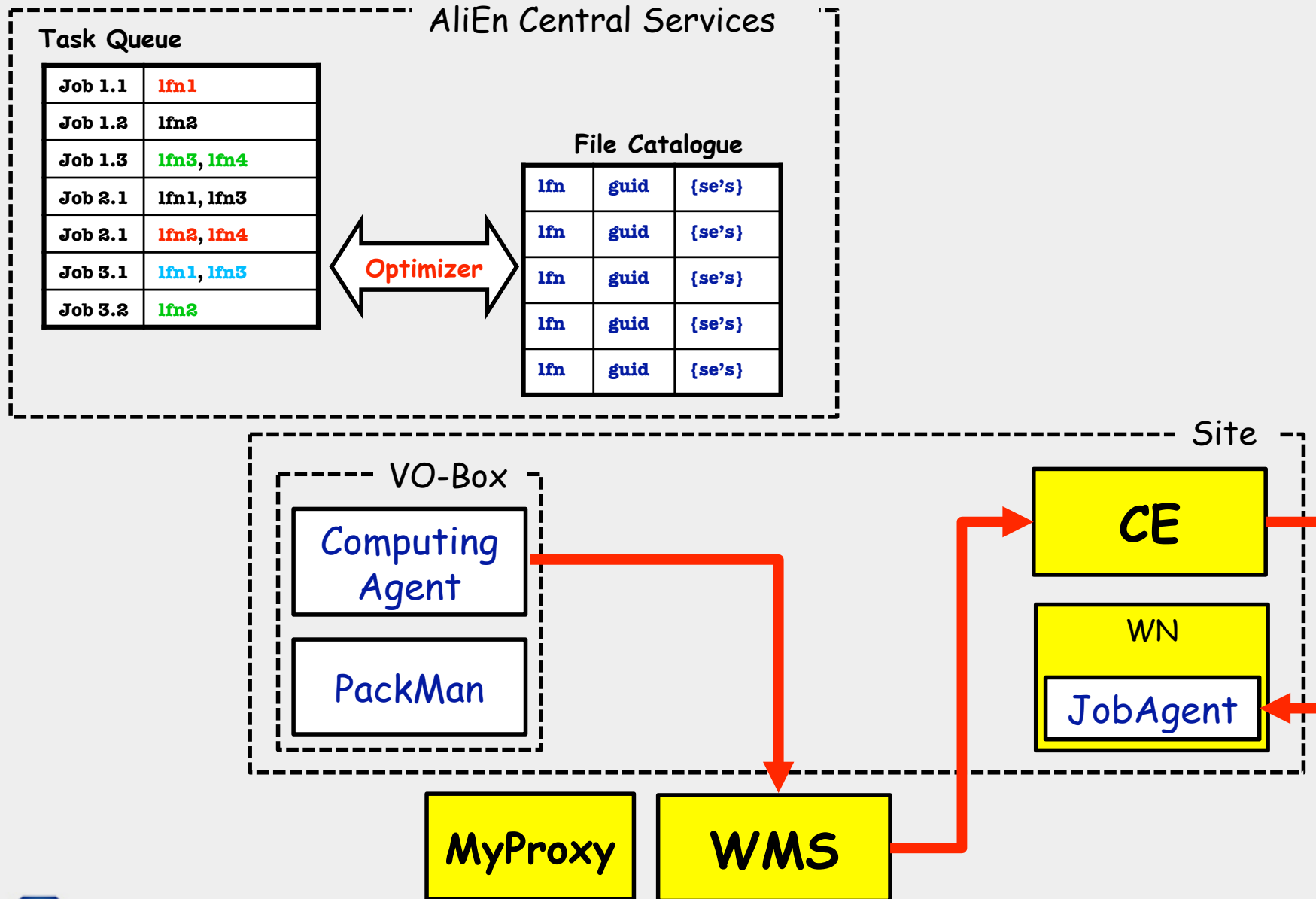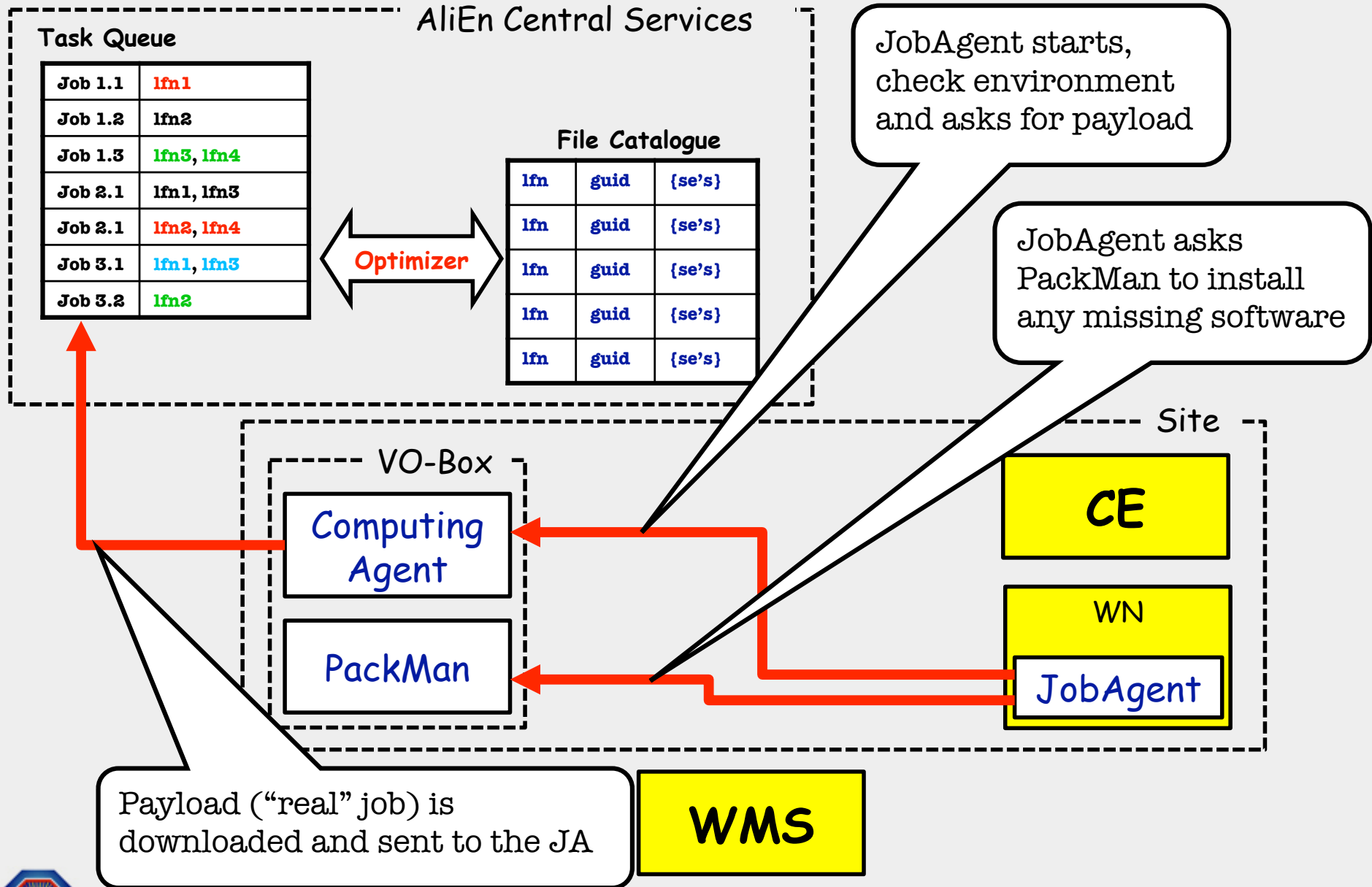 - JAs create a "virtual grid" on top of existing Grid infrastructures

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

Optimizer

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

Site

VO-Box

Computing Agent

PackMan

CE

WN

JobAgent

MyProxy

WMS

ALICE JOB MANAGEMENT VII

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

← Optimizer →

JobAgent starts, check environment and asks for payload

JobAgent asks PackMan to install any missing software

Site

VO-Box

Computing Agent

PackMan

CE

WN

JobAgent

Payload ("real" job) is downloaded and sent to the JA

WMS

AliEn Central Services

Task Queue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

Optimizer

File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

Repeat *n* times; communication goes on for the whole life of the JobAgent

Site

VO-Box

Computing Agent

PackMan

CE

WN

JobAgent

MyProxy
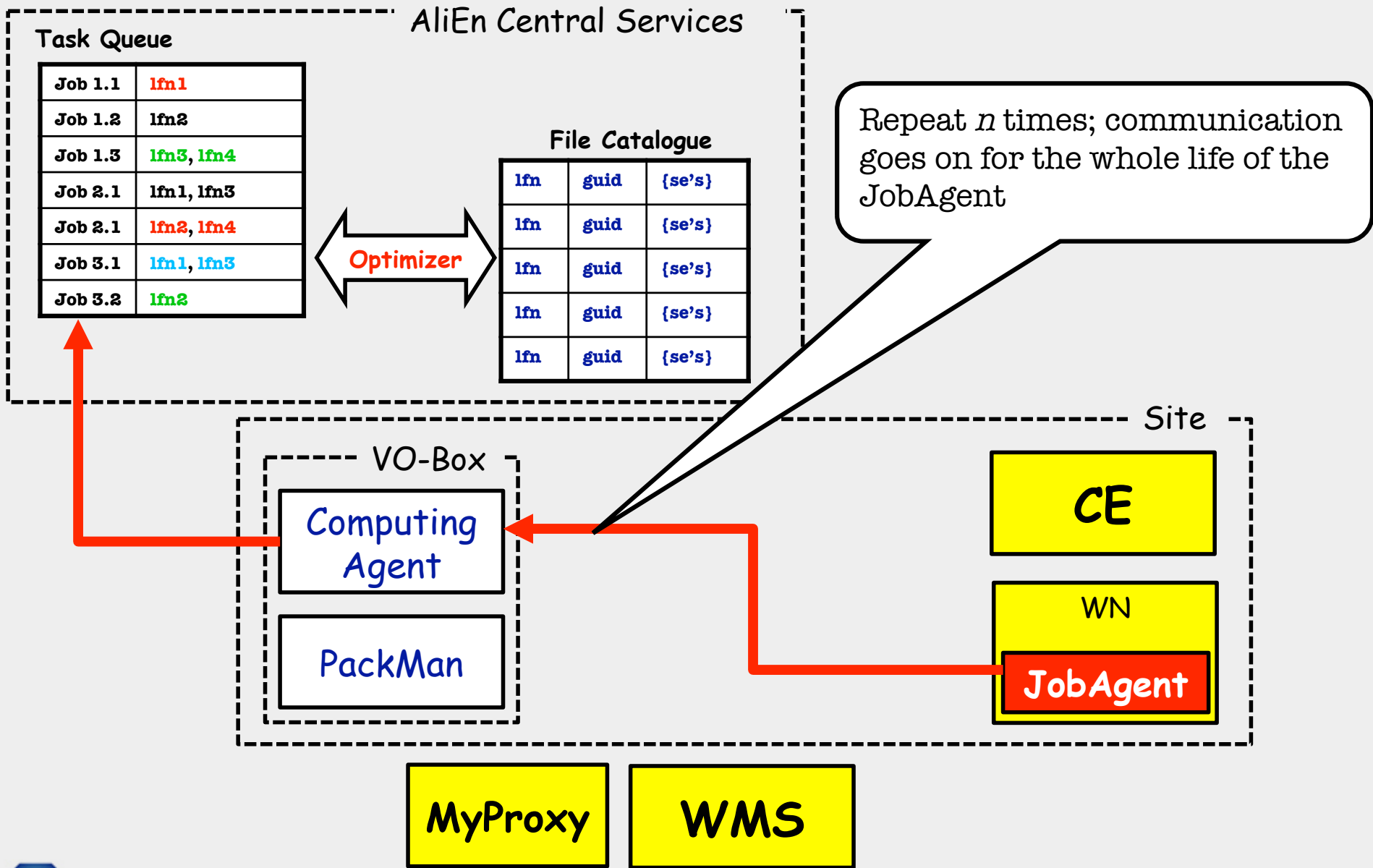
WMS

- **Problem**: huge backlogs in the internal queue
  - Reason still unclear
  - Several conditions can trigger the problem
  - These jobs are "invisible" from the IS

- No direct access to WMS internal status
  - This problem is impossible to catch from the VO-Box
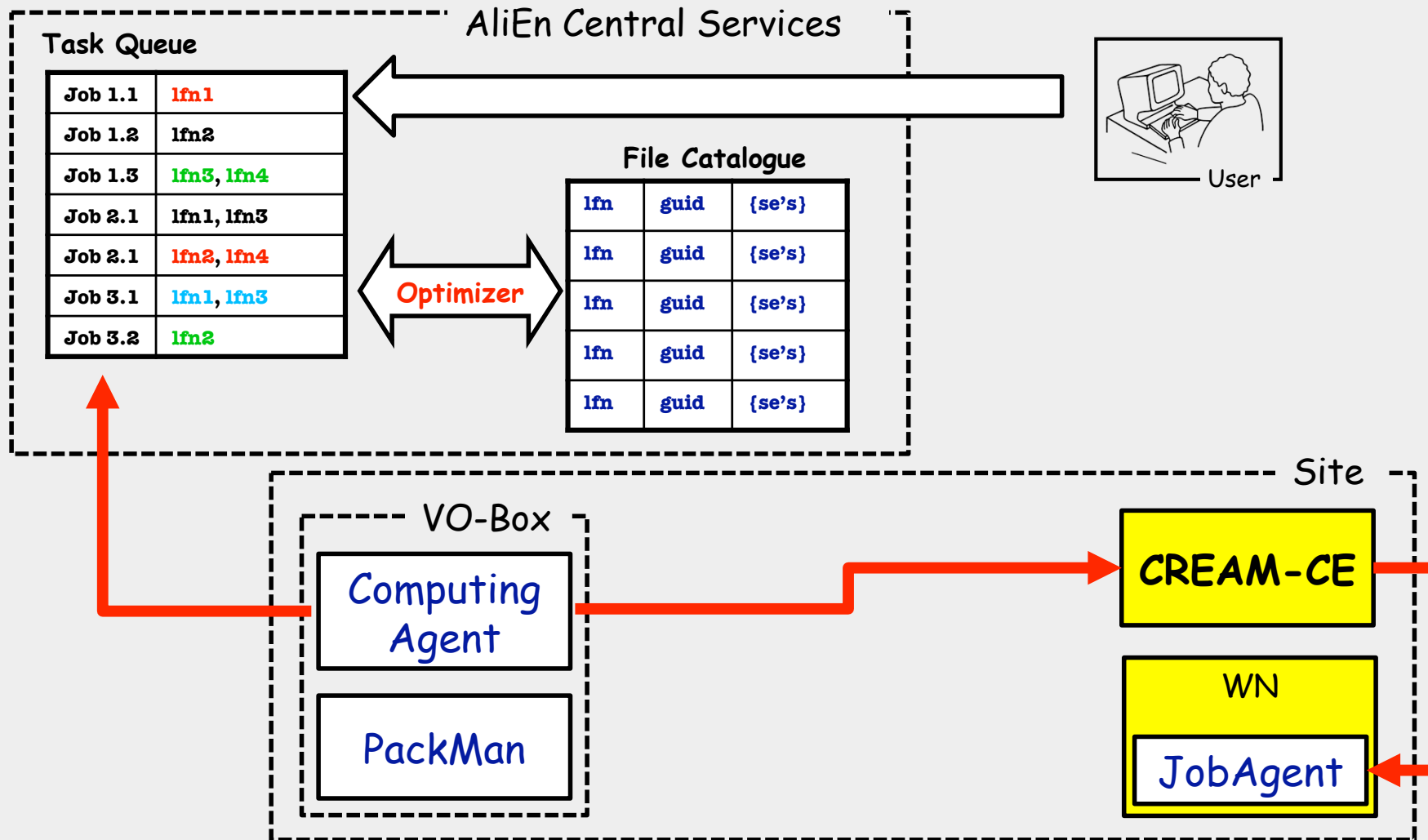 - Only through monitoring tools and site manager intervention


40k jobs in "waiting" status
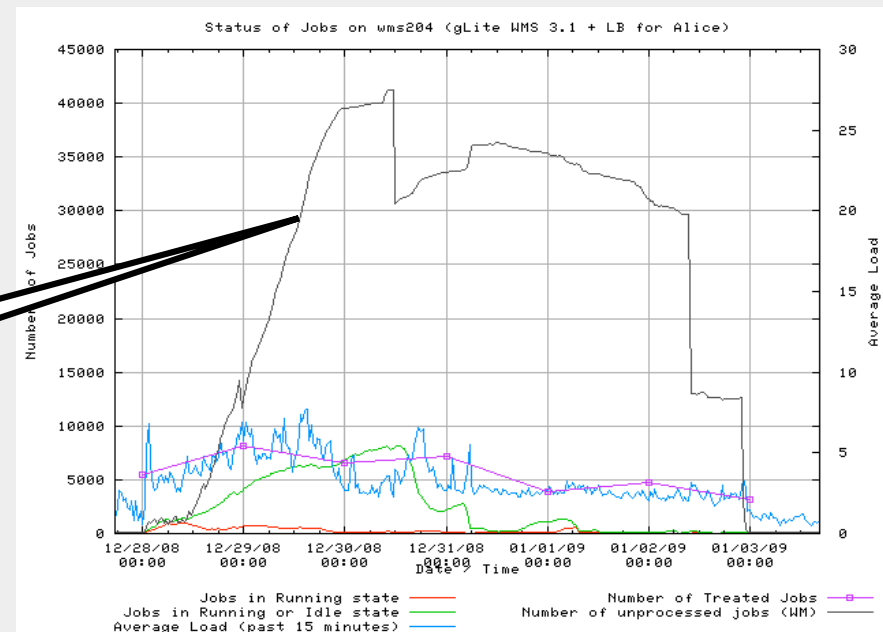
SUBMISSION THROUGH CREAM

- CREAM CE allows direct submission
  - Would have been possible also with older CEs, maybe using Condor
  - Only extra service needed is a GridFTP server on the VO-Box
  - Second VO-Box needed for parallel submission

- First tests in FZK (summer 2008)
  - Very good results
  - Decision to push towards CREAM

- Test phase with 5 sites
  - 67 kJobs in 1 month

- CE installation not always straightforward
  - Many sites reported problems and inconsitencies in documentation

- Excellent support from CREAM developers
  - Thanks to Massimo Sgaravatto

- Afterwards CREAM CE is very stable

- Missing feature: multi-CE submission
  - Simple site-wide load balance

CREAM VS. WMS SUBMISSION

Active jobs in Torino

Active jobs in Torino-CREAM
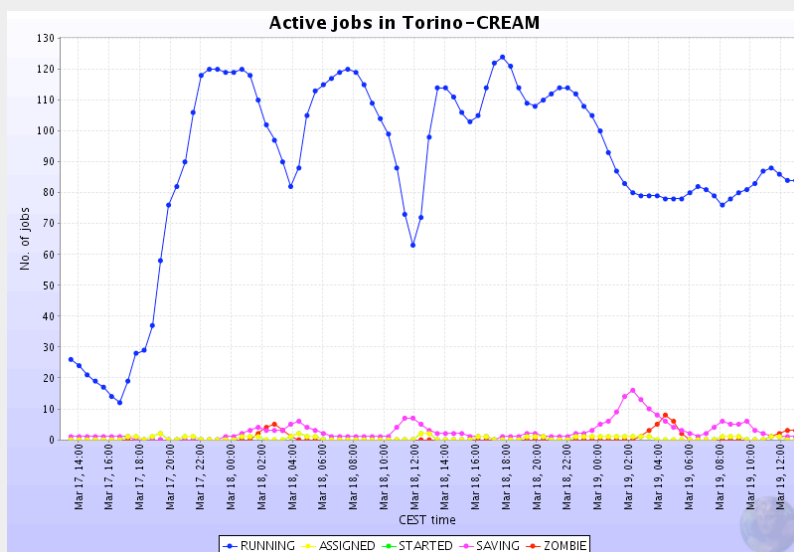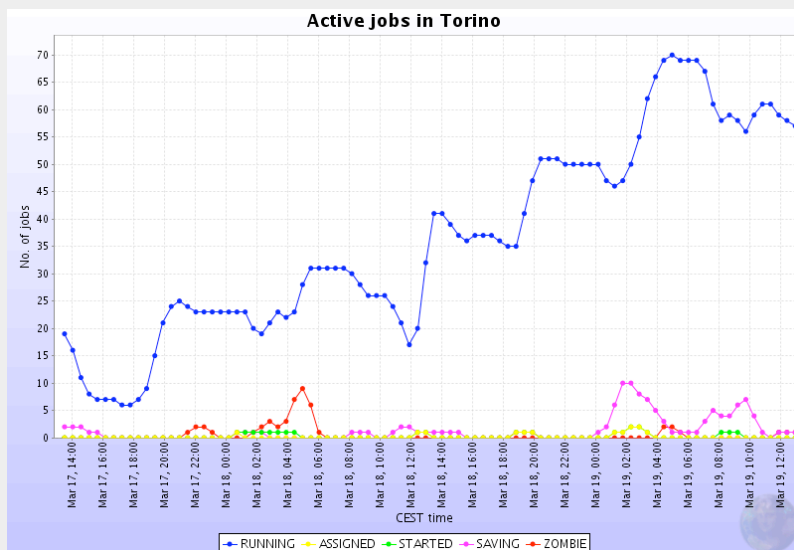
- CREAM and "normal" CEs competing for resources
  - Same WNs, same PBS server physically hosted on the old CE
  - Two WMSs in load-balancing configuration
  - Separate VO-Boxes (similar machines)

- Lower latency translates into steeper ramps

- Factor 2 à regime
  - Please note different vertical scale

- The overall ALICE job management model is mature and proved itself quite stable and scalable

- The VO-Box + JobAgent mechanism provides a good user and central services insulation layer

- Experience with the gLite 3.1 WMS was disappointing

- Direct submission to the CREAM CE fits well into the schema, and is currently proving itself reliable and very efficient

- Questions, comments?