# WNOD:
# (Worker Nodes On Demand) Running batch jobs in a customized virtual environment

Alessandro Italiano
INFN Tier1-CNAF
EGEE 08, Istanbul(Turkey)

# The problem

*In large computing center farms* **running multi users/ applications batch jobs**, *providing ad-hoc execution environments could force the farm administrators to unpleasant choices, like* **dedicating a subset of the available resources to each user/application** *requiring such an environment, be it a given O/S, the presence or absence of specific program or libraries, or else.* **But this has the important negative side-effect that farm administrators cannot generally optimize resource usage**, *because potentially some resources may remain idle, and a lost CPU cycle is lost forever.*

# Possible solution

*The technology developed in the virtualization area has finally achieved a consolidated level, allowing the use of virtual machines in production environment in a stable and productive manner.*

We therefore propose to run a given batch job on a Virtual Worker Node created on the spot for the sole purpose of running that batch job, and satisfies the batch job environment requests.

# A twofold advantage

- Autonomous execution environment avoids the following abuses.
  - Multiple process forking in a batch job can make a given user job to be more CPU aggressive than others.
  - Inappropriately demonized processes might waste CPU resources, or in some cases generate security concerns.
  - Bugs in a user job code might lead to kernel panic or to memory issues, thus affecting or even killing all the others jobs running on the same machine.

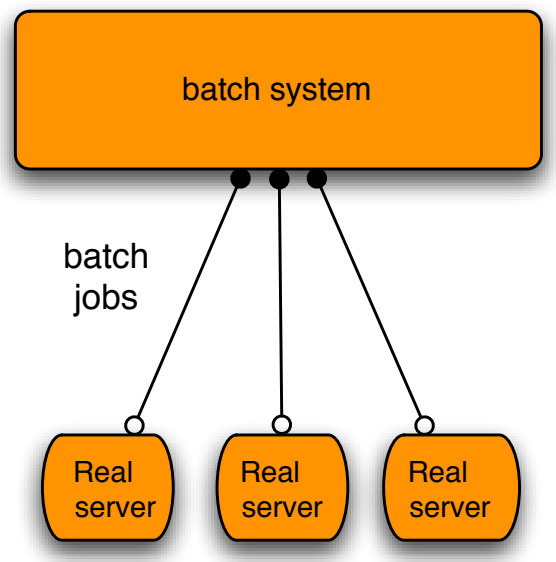
# Benefits

*This solution tries therefore to address the requirements of both farm administrator and of users/applications. In fact, this allows one to* **optimize resource usage** *because it is not necessary to statically dedicate resources anymore. In addition, the solution* **provides autonomous customized virtual environments**, **which can't interact with other batch jobs**, *thus protecting resources and, in the end, users themselves.*

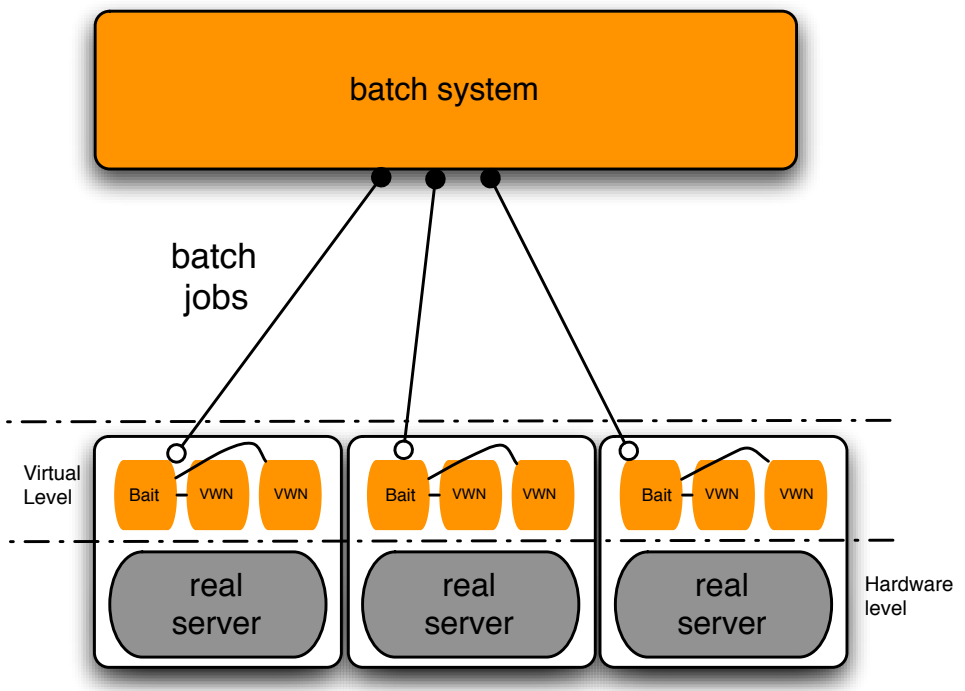
# Technical details:

- Integration with the batch system
- Architecture
- Virtual machine
- WNOD step by step
- Integration into the farm

# Integration with the batch system

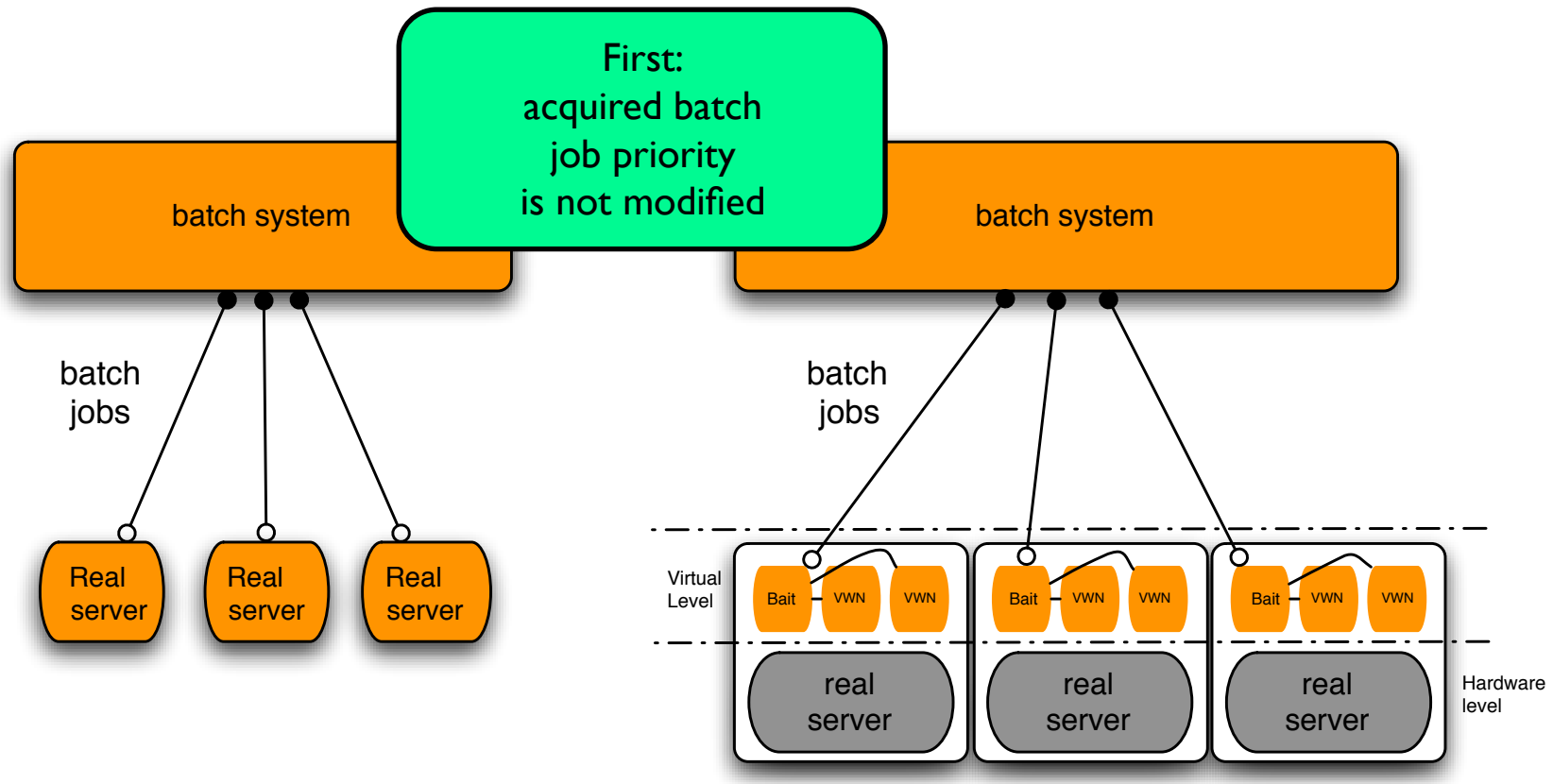



Standard batch system




Batch system for WNOD

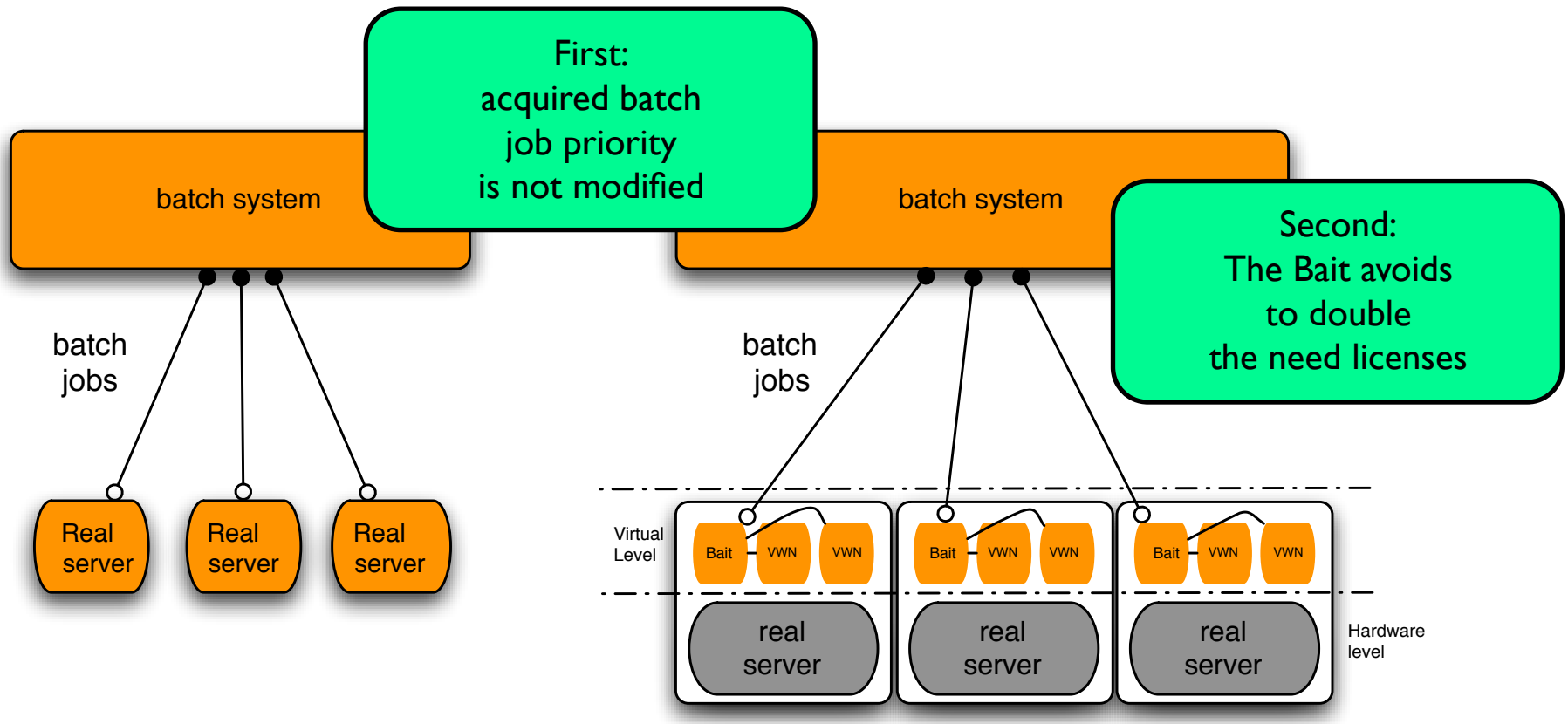
# Integration with the batch system




Standard batch system

Batch system for WNOD

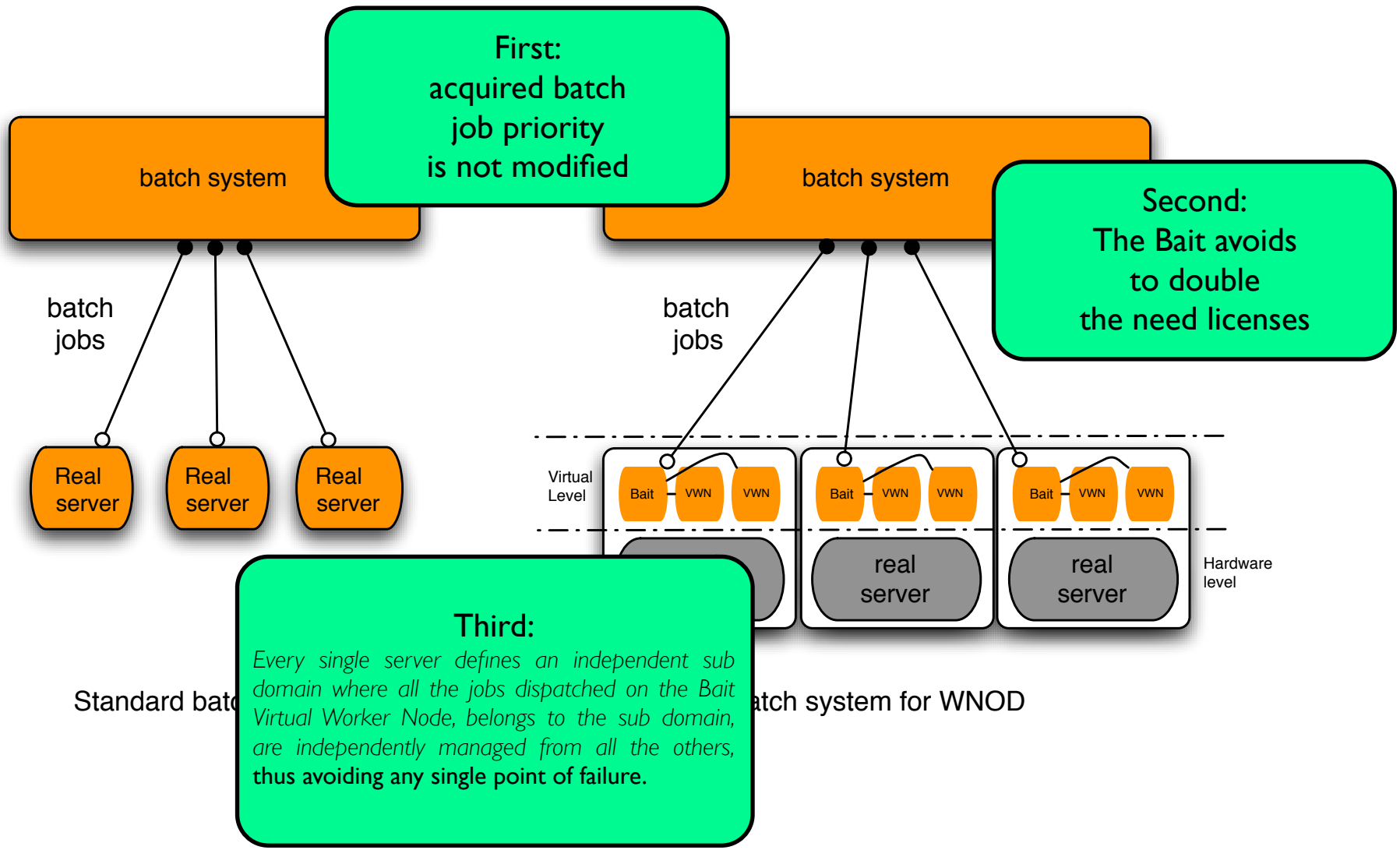
# Integration with the batch system



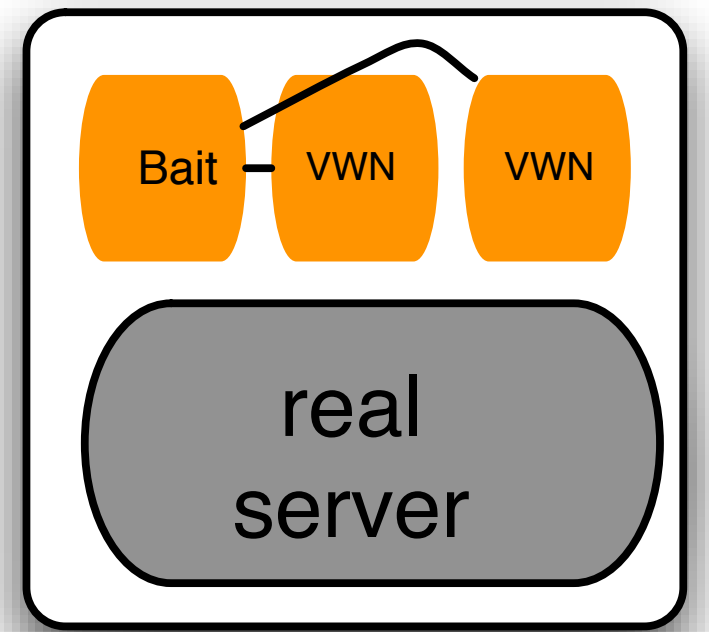

Standard batch system

Batch system for WNOD

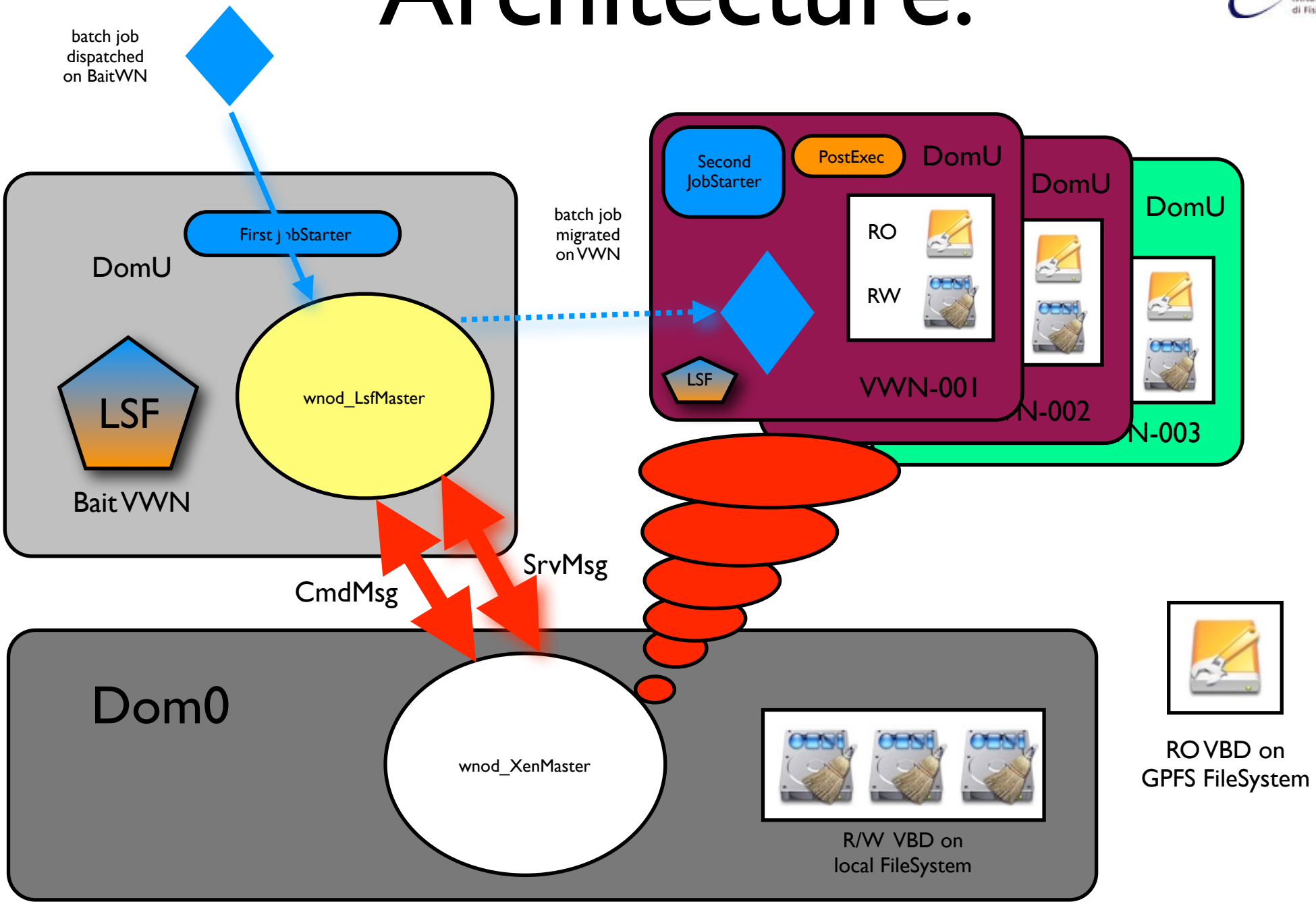
# Integration with the batch system



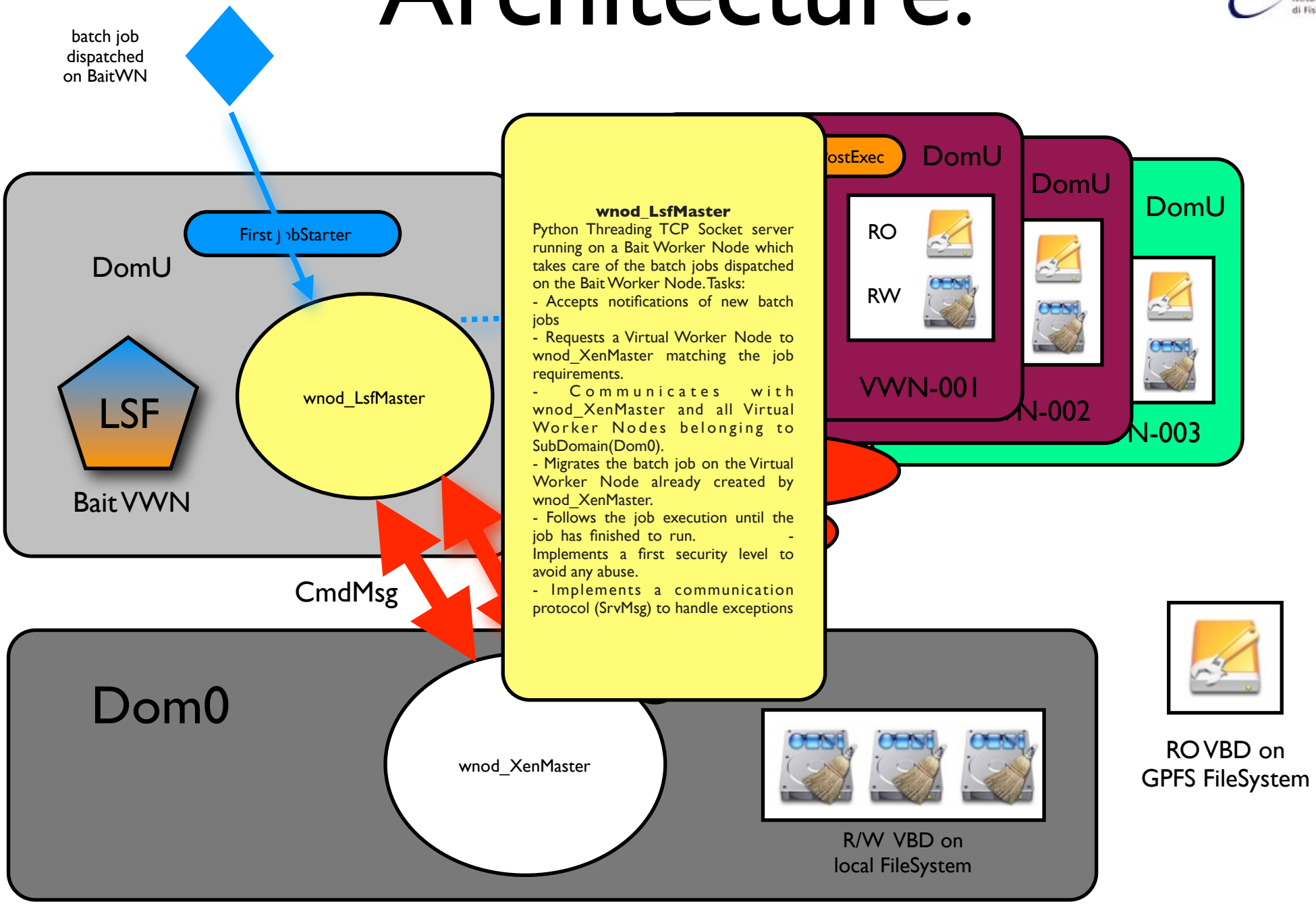
# Architecture:



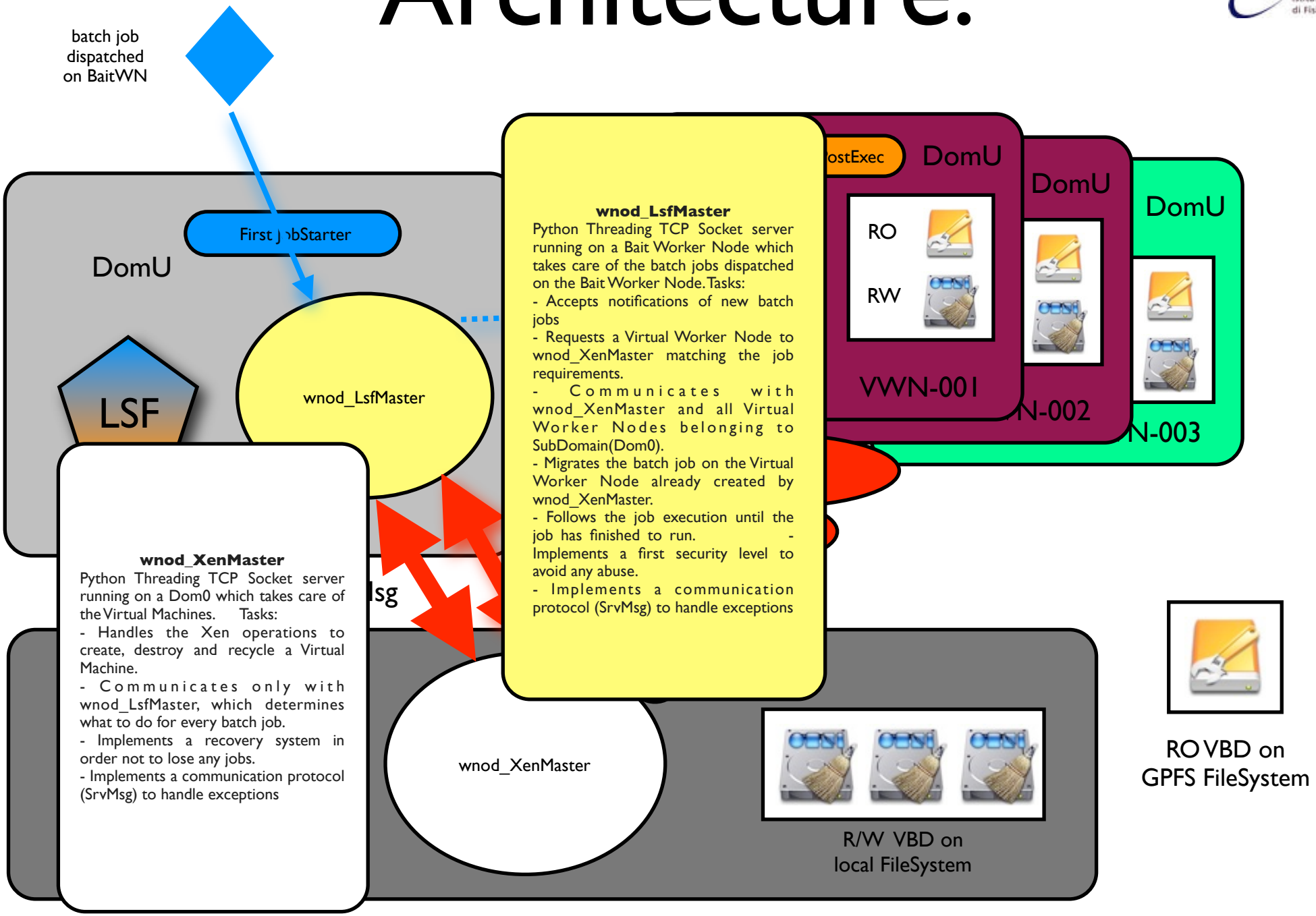
# Architecture:



# Architecture:

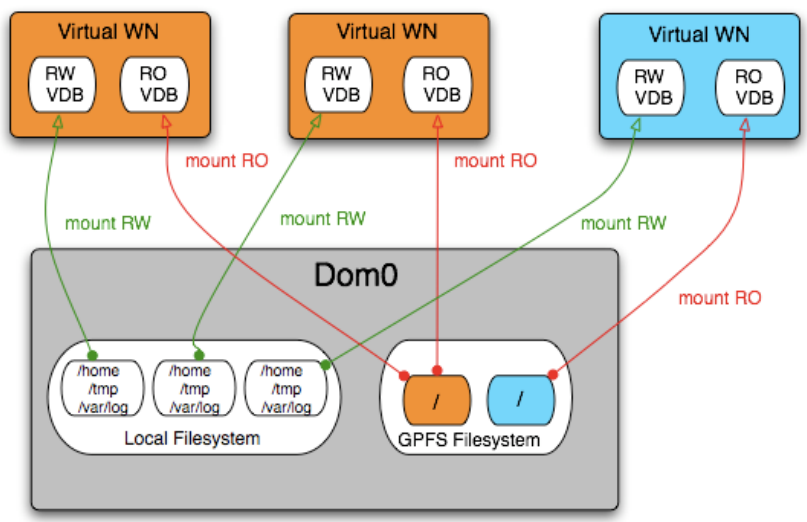


# Architecture:



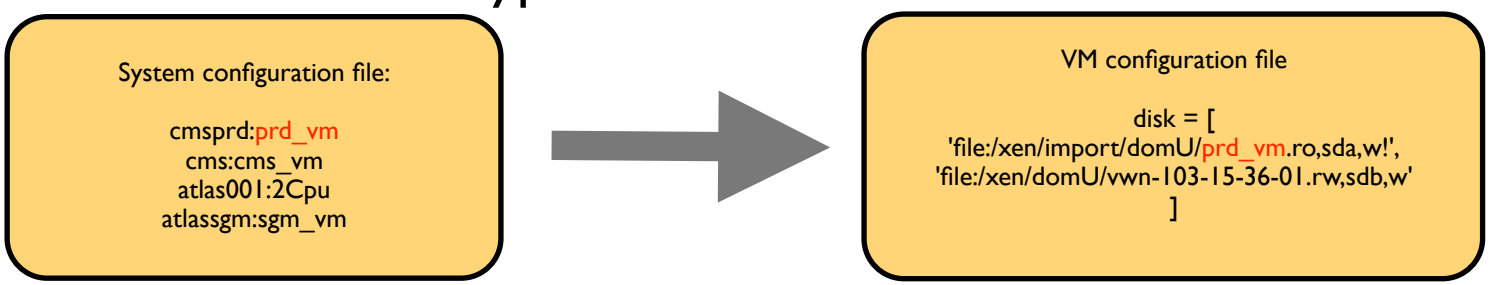
# Virtual Worker Node

- We use Xen 3.2
- Guests are Paravirtualized
- Every VM has two file-backed VBDs
  - one R/O with specific OS and SW
  - one R/W which provides writing file system



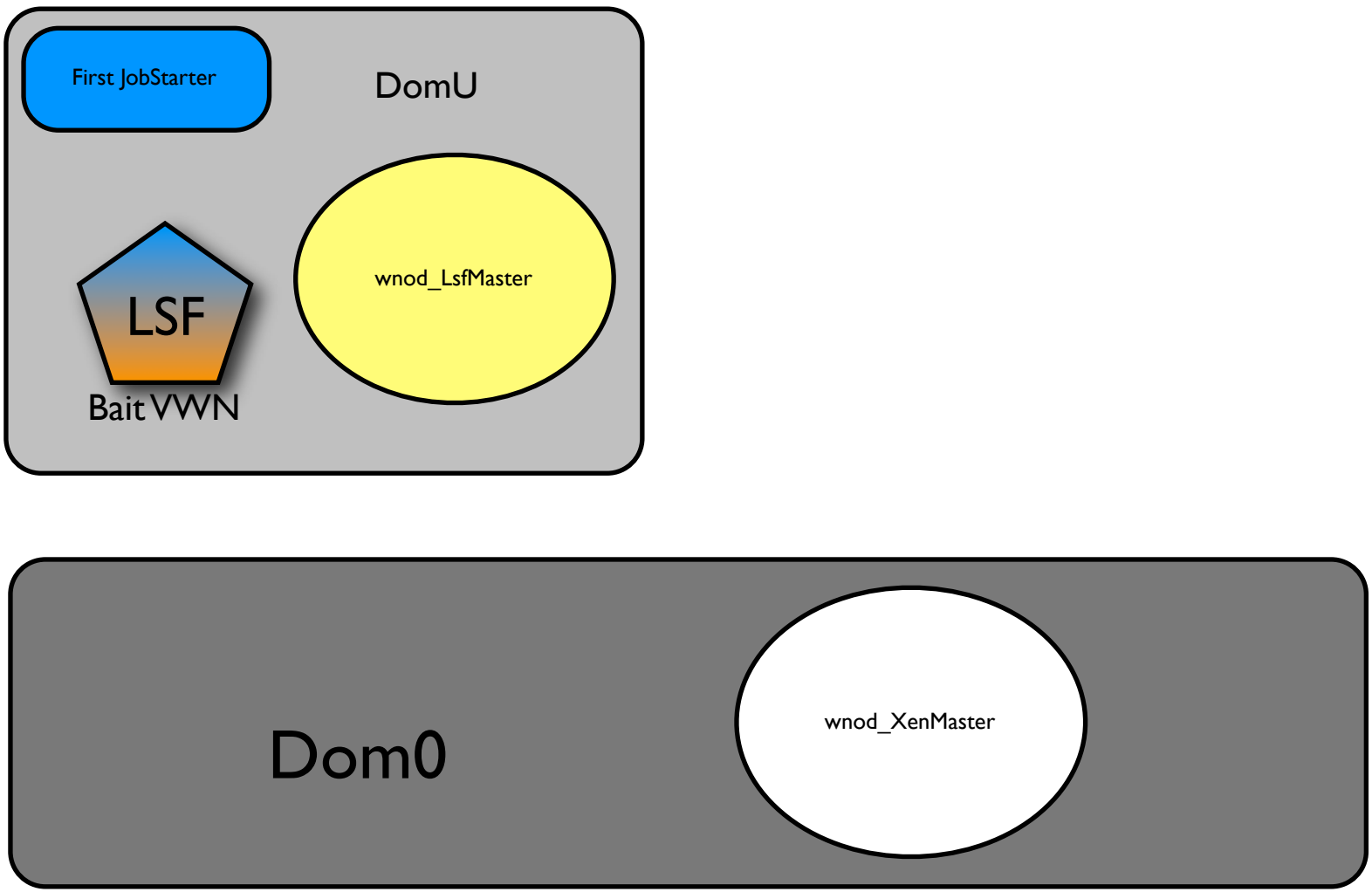
# Virtual Worker Nodes: Implementation details

- R/O VDB is set in a static way
- The system selects the VM type statically. It uses unix user details to find out which VM type the user can use.

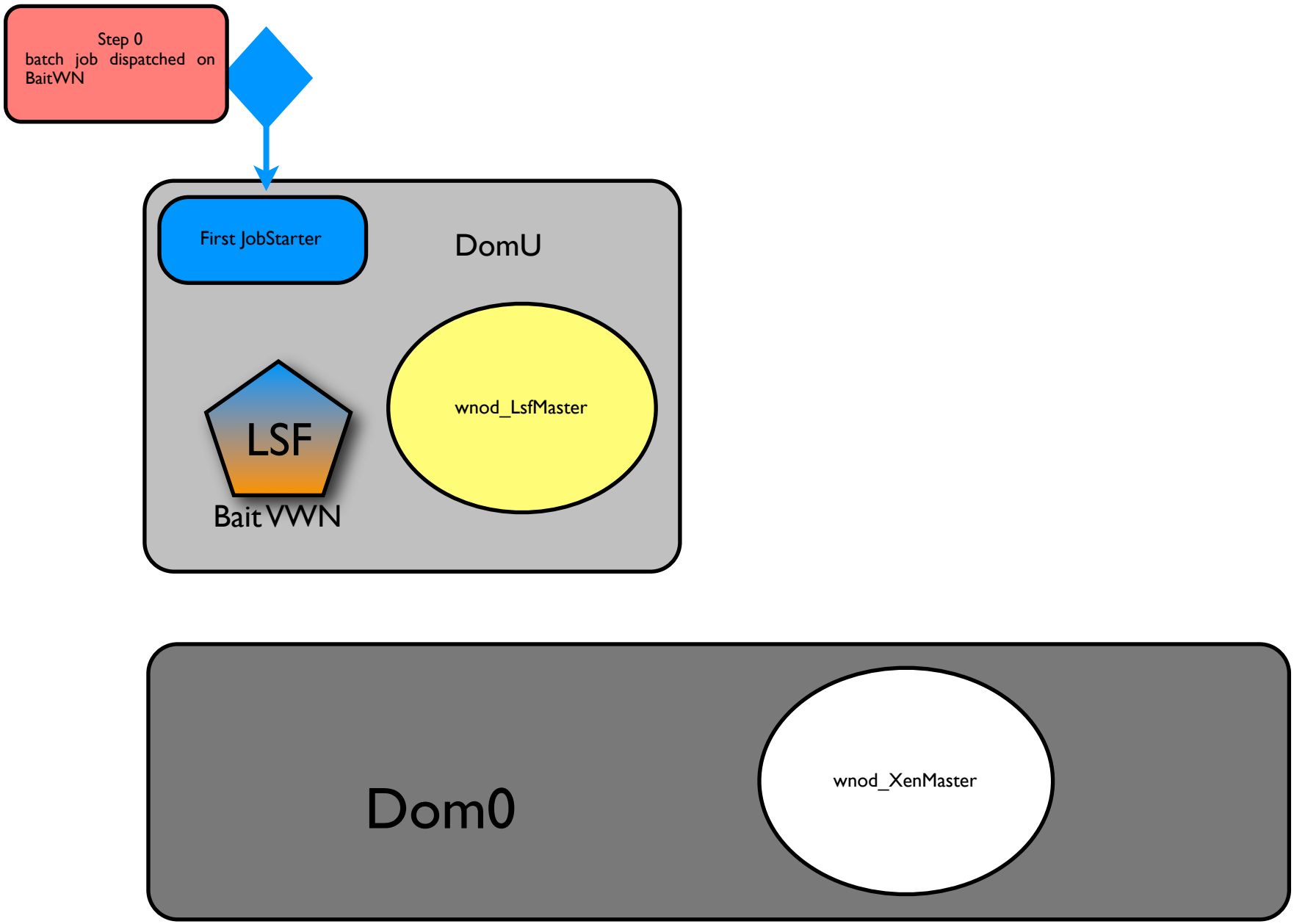



- At the end of each job, the VM will not be destroyed in order to try to reuse it.
- Applications and OS updates will be installed through a dedicated VM. In the update process the R/O VBD will be mounted in R/W mode.

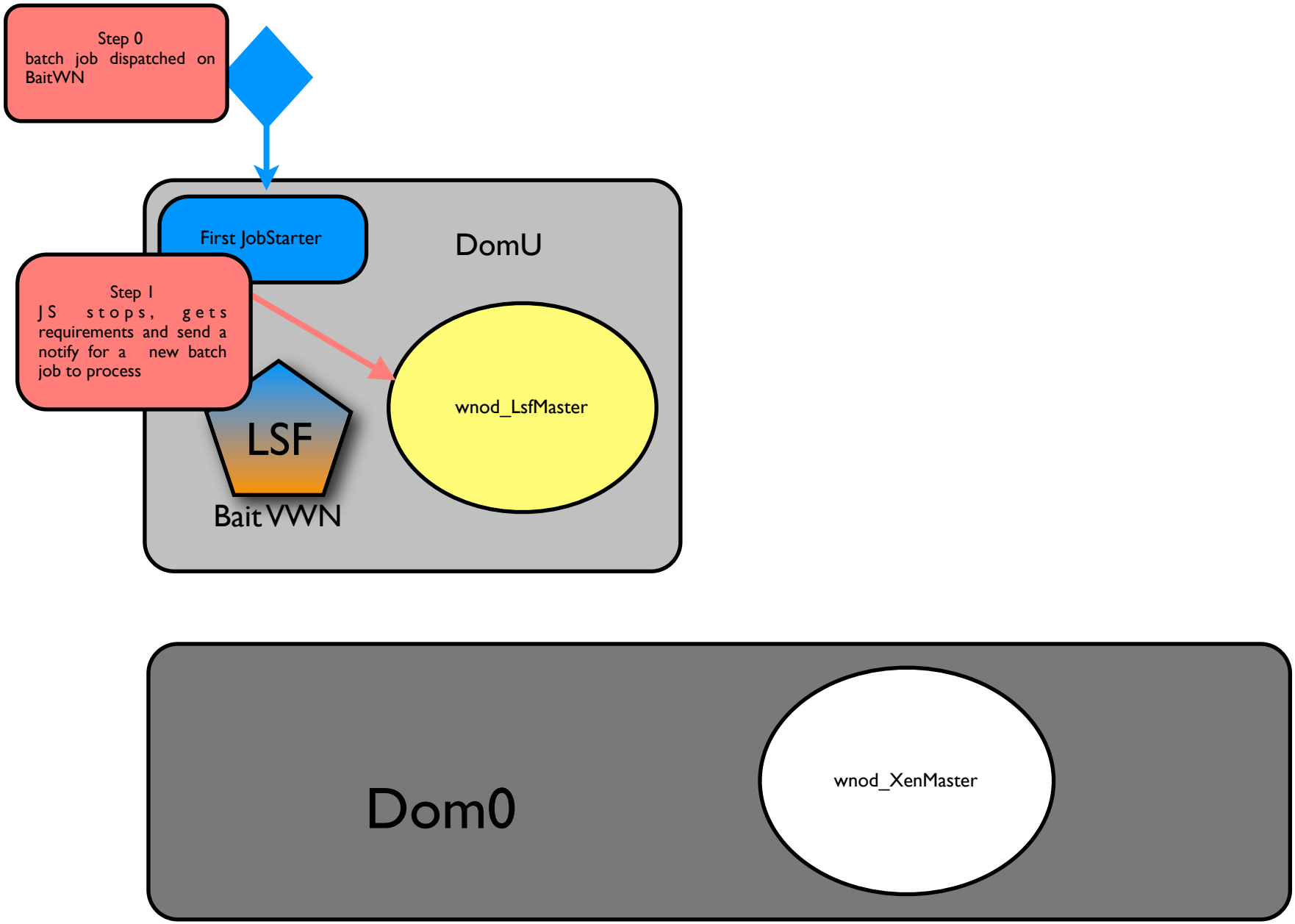
# WNOD step by step



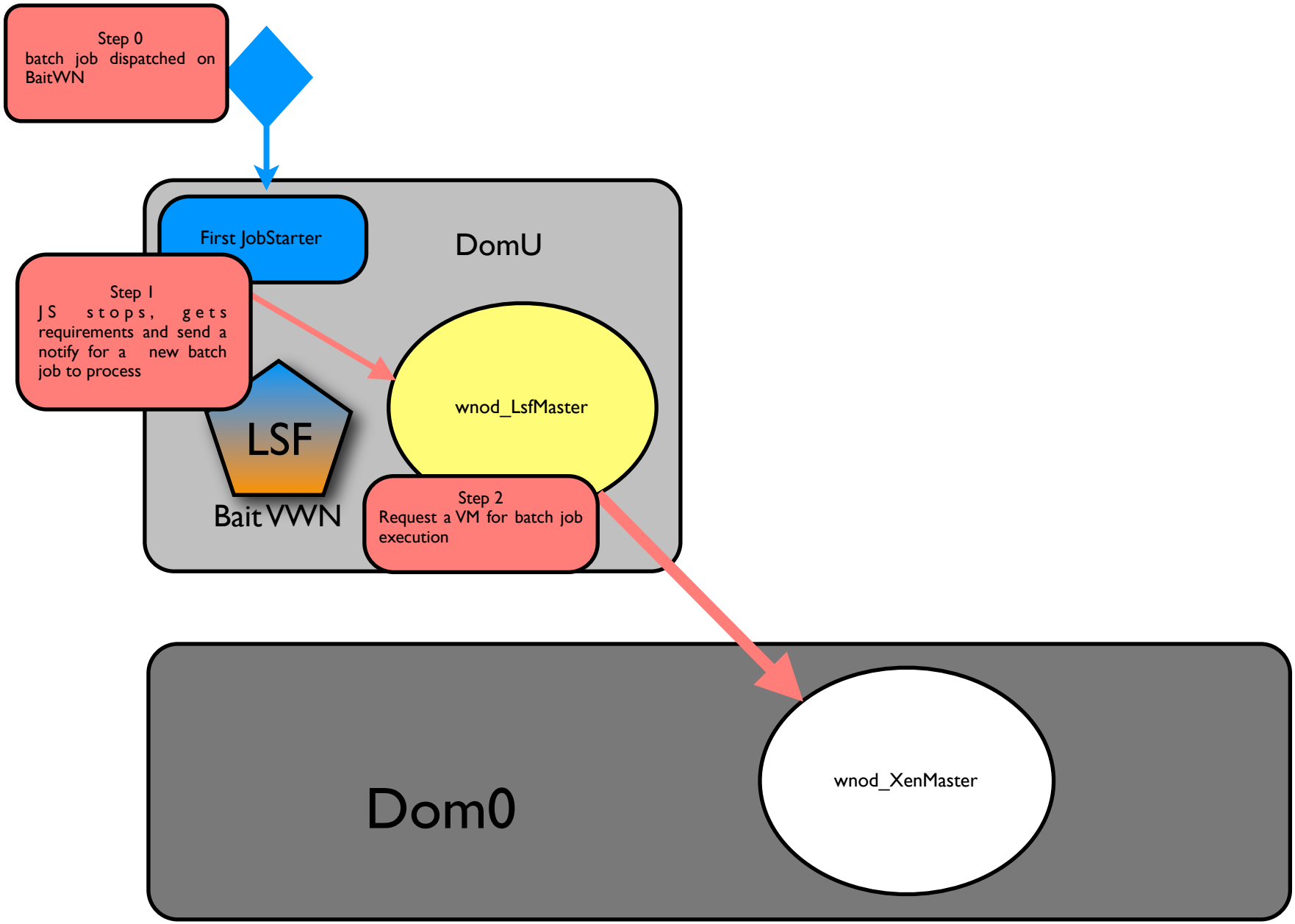
# WNOD step by step



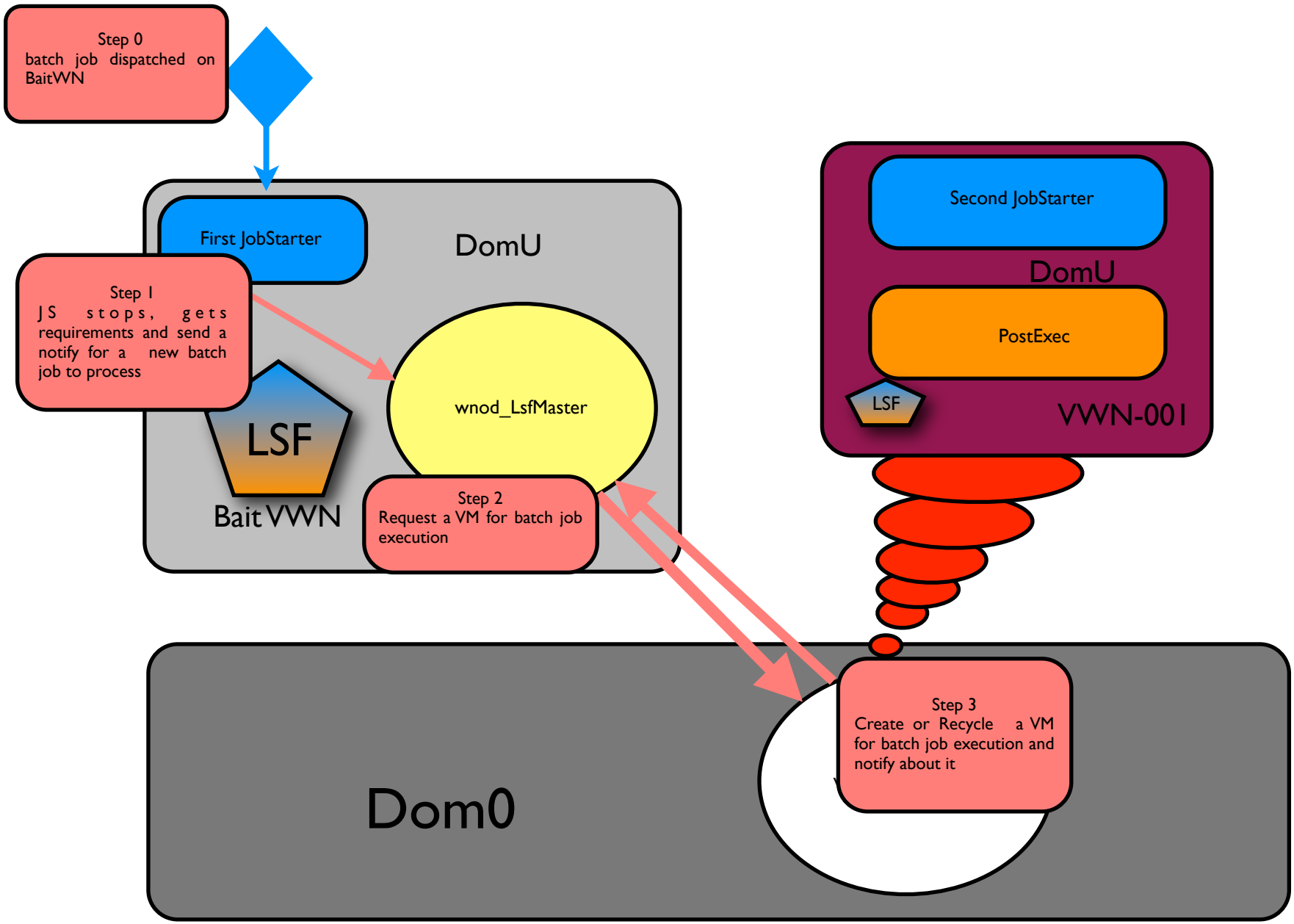
# WNOD step by step



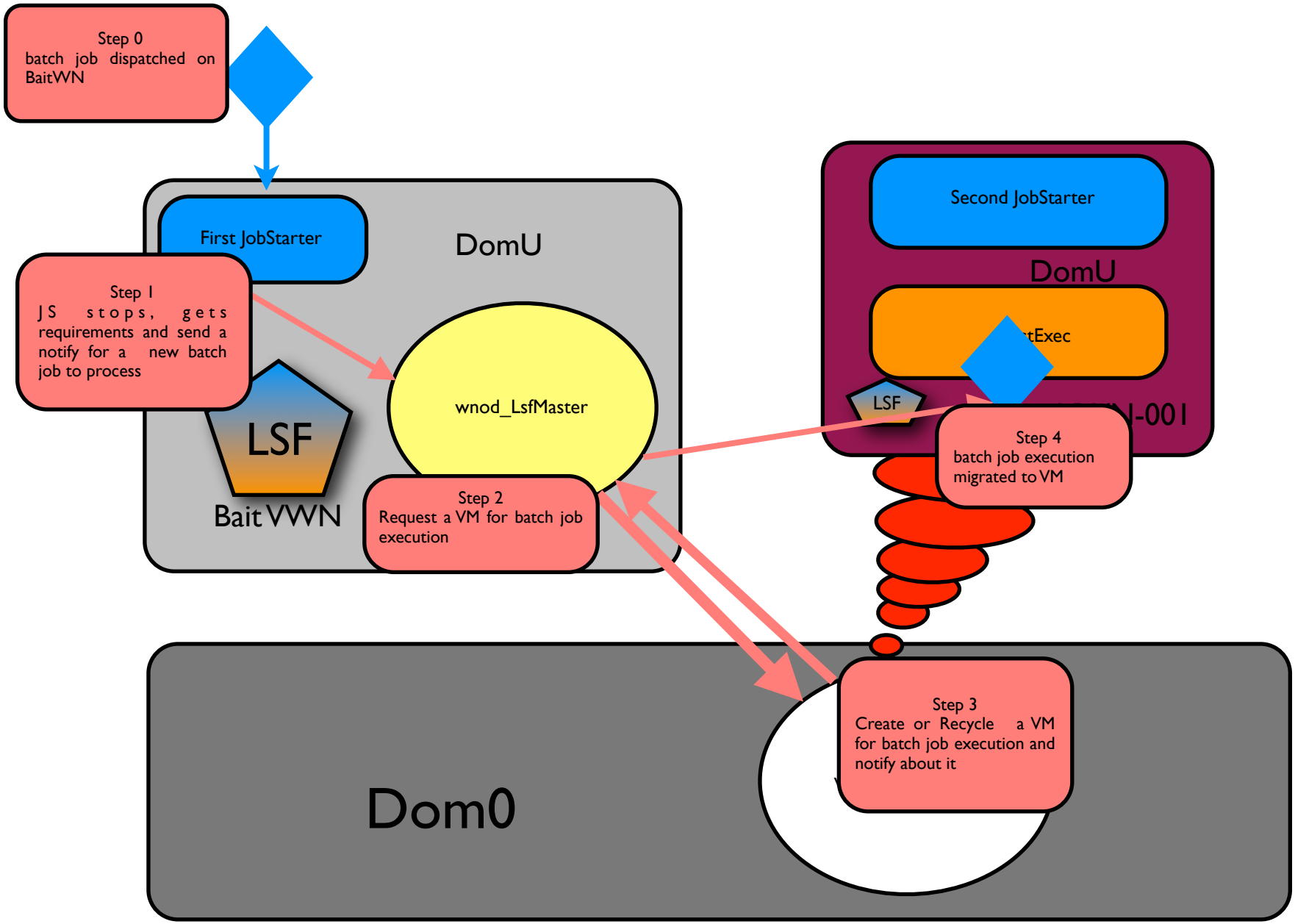
# WNOD step by step



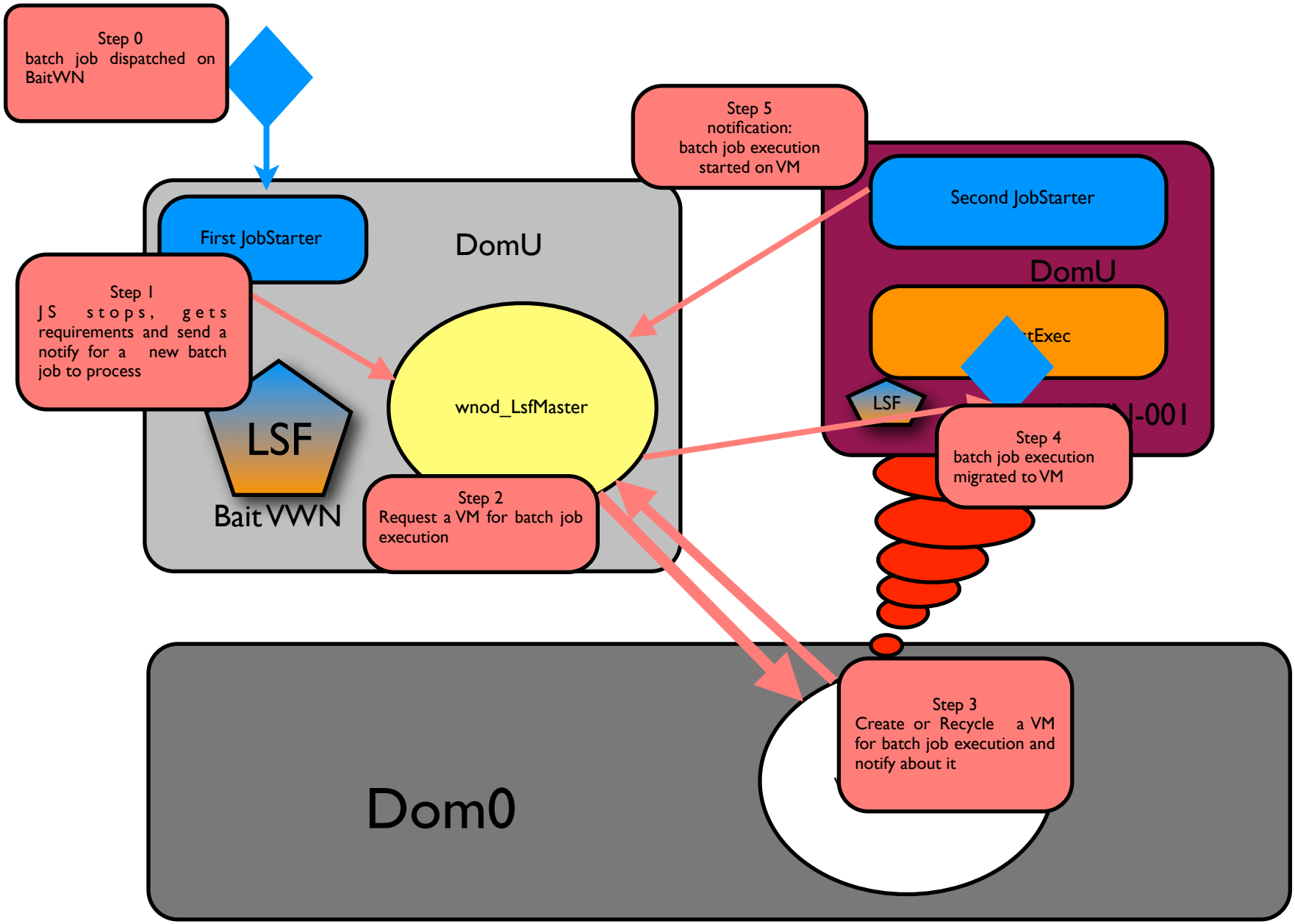
# WNOD step by step



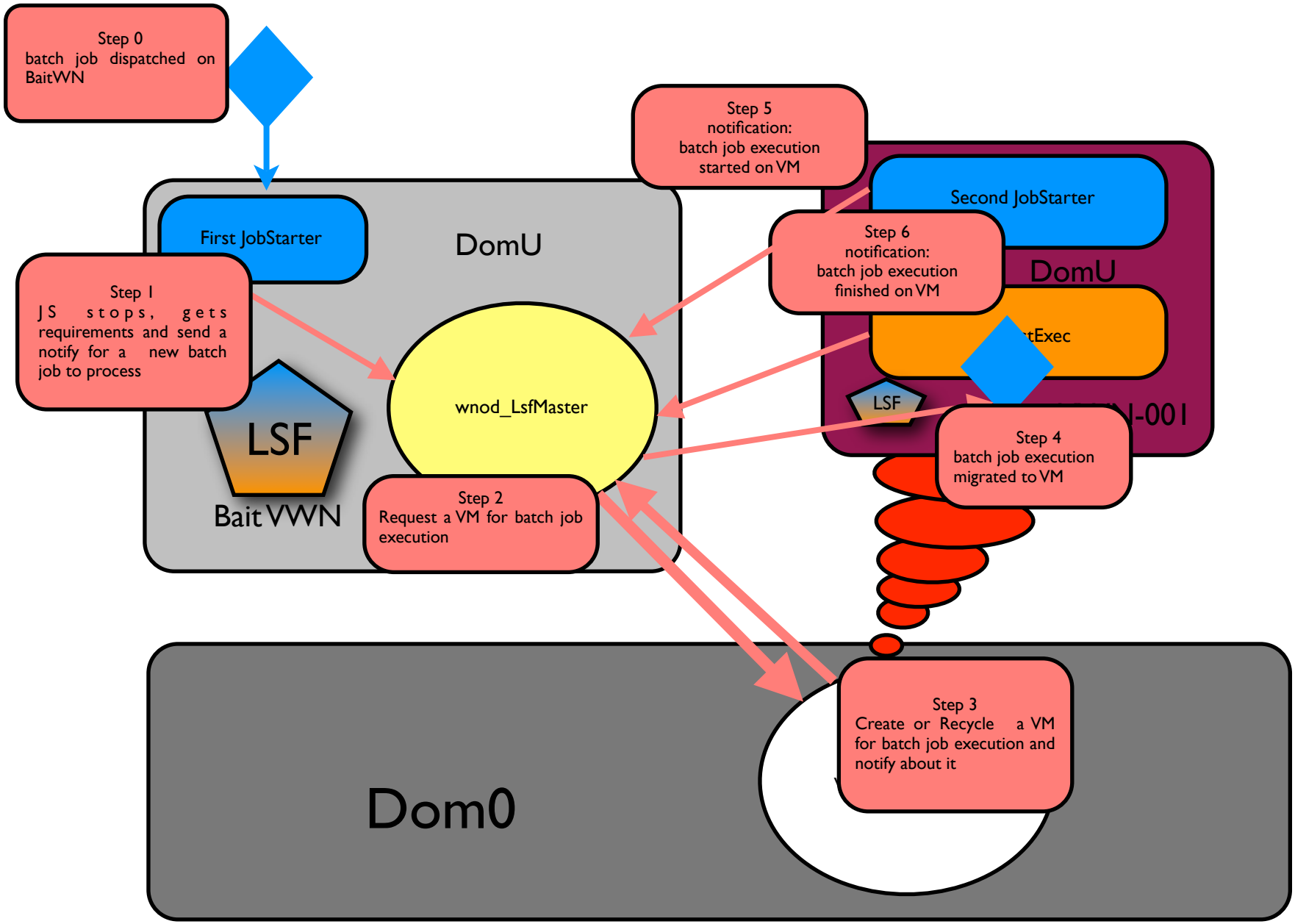
# WNOD step by step



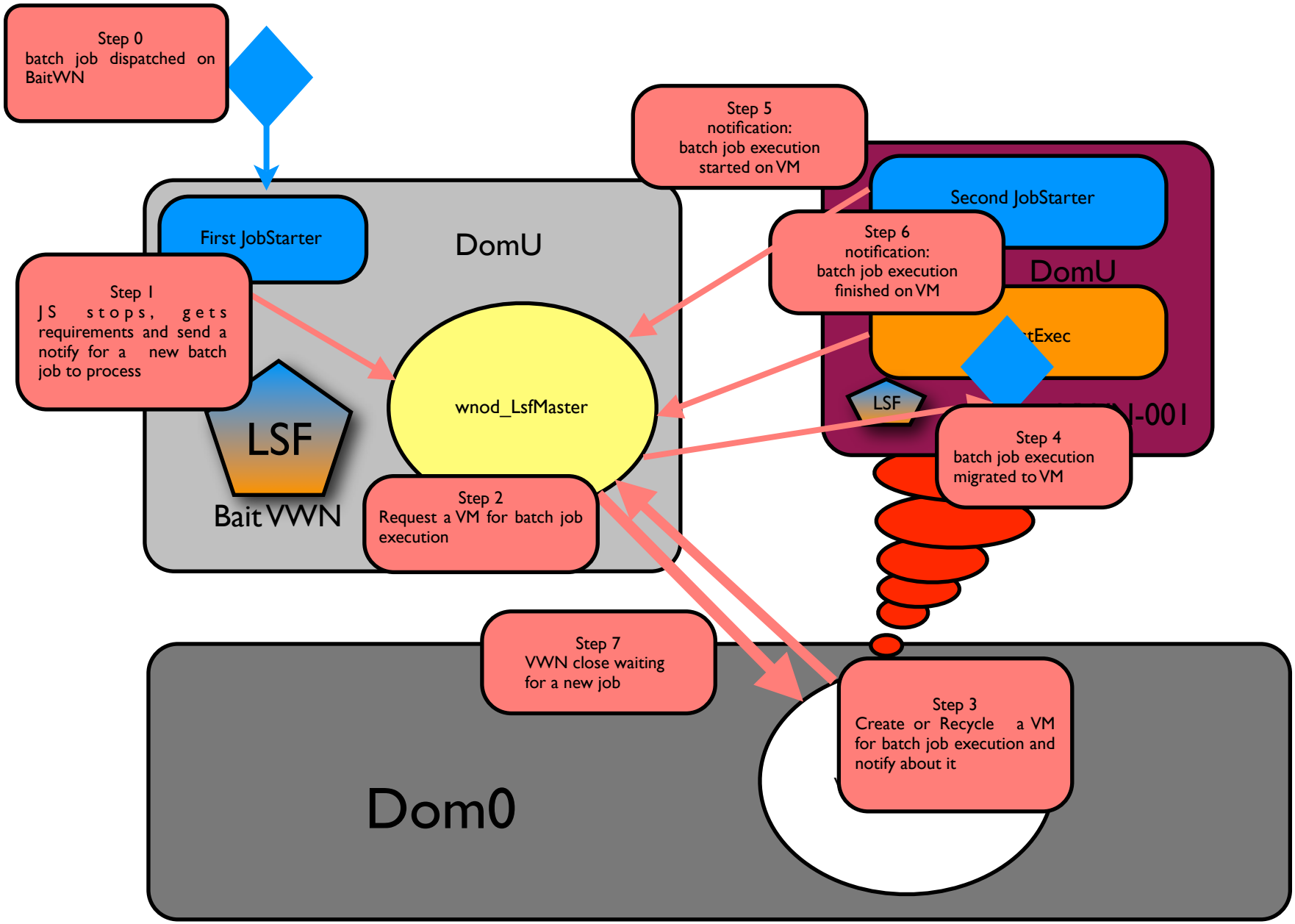
# WNOD step by step



# WNOD step by step



# WNOD step by step



# Integration with the farm

**The integration with the farm is not a issue**. *All farm administrator tools like monitoring, accounting and automated ICM will work as usual; quattor in fact can give a big contribution to the deployment.*
*Since this solution introduce a new element in the cluster, a specific tool to monitor the VMs status and their location in the cluster should improve farm administration activity.*

# Current status

- Process for batch job execution on a VM has been coded.
  - synchronization between batch job migration and VWN availability has been the most important tackled issue.
- We are trying to find out all the corner cases and code them.
- First tests are comforting.