Table of Contents

IRB Tier 3 Instructions

Official CMS name of the site is T3_HR_IRB. PHEDEX Configuration is found under lorienmaster.irb.hr:/home/phedex (user: phedex)

Site hosts: lorienmaster.irb.hr (headnode), lorientree01.irb.hr and lorientree02.irb.hr UPDATE: lorientree03.irb.hr and lorientree04.irb.hr have been added in the meantime.

To use the site, first a new local account needs to be created.

Site contacts are srecko.morovic@cernNOSPAMPLEASE.ch and vuko.brigljevic@cernNOSPAMPLEASE.ch.

Various information on setting up the user workflow on the site (for local users), as well as several administration tasks, are described in the following sections.

New CMSSW installation (using CVMFS)

* set up scram using:

```
export SCRAM_ARCH=slc5_amd64_gcc462
source /cvmfs/cms.cern.ch/cmsset_default.(c)sh
```

UPDATE:

```
export SCRAM_ARCH=slc6_amd64_gcc491
```

* other gcc versions are also supported. CVMFS caches approx 20 GB of CMSSW installation data (this can be increased if necessary), so any version is available without separate installation.

* this installation, (as well as two other found in /users/cms and /users/cmssw) are now configured to use local Squid to proxy and cache condition data needed for CMS data processing.

* Old CMSSW installations in /users/cms and /users/cmssw are obsoleted by this and it is possible that they will be deleted in the future (exception is CRAB in /users/cms) to free disk space. At this point it might only require users to create a new project space and recompile the code.

CRAB installation on the site

* After "cmsenv" (UPDATE: also works before "cmsenv" so you can add this to your environment):

```
source /users/cms/CRAB/CRAB_2_9_1/crab.(c)sh
```

UPDATE: for CRAB3 use

```
source /cvmfs/cms.cern.ch/crab3/crab.(c)sh
```

Grid environment (UI) should already set up automatically.

How to submit grid jobs that copy data back to IRB

* Add/edit these lines in crab.cfg (user section)

```
return_data = 0
copy_data = 1
eMail = your.mail@xxx
user_remote_dir=/somedir # or set it (as below) in multicrab.cfg using "USER.user_remote_dir" (th
storage_element=T3_HR_IRB
#do not set storage_path here. Files will end up in LFN /store/user/%username% (PFN: /STORE/se/cm
```

If your user directory is not created in /STORE/se/cms/store/user, please ask site contacts (Srecko, Vuko). This directory must belong to "storm" user and group. There is a periodically running cron script that makes these directories writable for anyone (setfacl), so that analysis output can be deleted.

* Please be careful not to write to someone else's directory. Currently, access rights do not distinguish between different users.

*In multicrab.cfg (these options can also go in crab.cfg where section is in capital letters below)

```
[COMMON]
CMSSW.pset=your_cfg.py
CRAB.scheduler=remoteGlidein
CRAB.use_server=0
CMSSW.lumis_per_job=50 #set your own
USER.user_remote_dir=IRBtest #this sets subdirectory under "storage_path" as above
#USER.check_user_remote_dir=0
```

* Note: instead of "lumis_per_job" (recommended), it is possible to use "CMSSW.number_of_jobs = XX" in section of each dataset. The latter can be dangerous because of limited amount of space present in condor working directory which is on the system partition. Number of available Condor job slots on all three machines is 90, but better queue more jobs.

* Note: manual deletion or moving of files copied to SRM might still not be possible for local users (e.g. if you want to delete data no longer needed). This will be addressed by running a a cron job to set proper permissions.

* Note: in some cases default voms-proxy-* installed might not work for some grid related activities (e.g. using srmcp). In case of problems, it is recommended to use /opt/voms-clients-compat/voms-proxy-init and related tools for creating voms proxies. E.g.

```
/opt/voms-clients-compat/voms-proxy-init -voms cms # to get proxy with access to CMS resources
```

* Crab takes care of creating voms proxiy itself, so use above only in case of problems. Just make sure to have .globus populated with proper CERN certificate and key.

* Alternative copying mode (use ONLY if above doesn't work):

```
[USER]
return_data = 0
copy_data = 1
eMail = your.mail@xxx
storage_element = lorienmaster.irb.hr
storage_path = /srm/managerv2?SFN=/STORE/se/cms/store/user/username  #set your user dir
storage_port = 8444
#srm_version = srmv2 #optional
```

UPDATE: For using CRAB3, in your configuration python script set storageSite to T3_HR_IRB (you don't need to have an account!):

```
config.Site.storageSite     = 'T3_HR_IRB'
.
.
```

```
.
config.Data.outLFNDirBase  = '/store/user/%username%/%directory%'
```

To check and/or copy your files from T3_HR_IRB, use lcg-ls and lcg-cp, for example:

```
lcg-ls -v -b -l -T srmv2 --vo cms srm://lorienmaster.irb.hr:8444/srm/managerv2\?SFN=/STORE/se/cms
```

How to run CRAB jobs directly on T3_HR_IRB:

To run jobs on Condor batch directly on the site (on three machines that ara available), change scheduler to:

```
CRAB.scheduler=condor
```

* Jobs must be submitted directly from any of the three site hosts. Also, any samples to be processed must be available on the site (transferred using PHEDEX, which site admins/executives can do). Do not use "condor_g" scheduler, only plain "condor" * Note: use first method for copying only (alternative ignores subdirectory setting)

List of datasets currently replicated on T3_HR_IRB

Query it here:

https://cmsweb.cern.ch/das/request?view=list&limit=10&instance=cms_dbs_prod_global&input=dataset+site%3DT3_

Custom analysis datasets (add your analysis dataset here):

Dataset name	size	status	DBS instance

Manualy copying data from other sites (examples)

#using lcg-cp:

```
lcg-cp -v -b -D srmv2  srm://cmssrm.hep.wisc.edu:8443/srm/v2/server\?SFN=/hdfs/store/user/smorovi
13Jul2012-v1/patTuple_10_1_QVr.root  srm://lorienmaster.irb.hr:8444/srm/managerv2\?SFN=/STORE/se/
```

#using srm-cp

```
srmcp -retry_num=0 file:////tmp/testfile srm://lorienmaster.irb.hr:8443/srm/v2/server\?SFN=/STORE
```

* Note: srm-to-srm needs -pushmode switch. Also, srmcp only recognizes certificates created by /opt/voms-clients-compat/* tools (not default ones installed).

#xrootd

Xrootd service is presently not installed. This is on a TODO list.

UPDATE: xrootd service is installed and working. After "cmsenv" and creating proxies (voms-proxy-init -voms cms) simply use, for example xrdcp for copying.

Local DBS for publishing locally processed datasets to private DB (updated)

It is possible to publish processed data to CMS DBS analysis instance. This is not yet tried, but detailed here (for another T3 site)
http://wiki.crc.nd.edu/wiki/index.php/Using_CRAB#Running_jobs_on_the_Local_Condor_Queue

Name of your analysis dataset can be arbitrary, however I recommend this convention:

```
T3HRIRB_VERSION
```

User analysis datasets can be published to cms_dbs_ph_analysis_01_writer or cms_dbs_ph_analysis_02_writer DBS instance. For the former, use this in crab.cfg:

```
return_data = 0
copy_data = 1
eMail = x.y@cern.ch
storage_element = T3_HR_IRB
publish_data=1
publish_data_name = T3HRIRB_V00
dbs_url_for_publication = https://cmsdbsprod.cern.ch:8443/cms_dbs_ph_analysis_01_writer/servlet/D
```

After all jobs are fully processes, do:

```
(multi)crab -getoutput
(multi)crab -publish
```

This will produce dataset in the form:

```
/WprimeToWZToLLLNu_M-200_TuneZ2star_8TeV-pythia6-tauola/smorovic-T3HRIRB_V00_MCPatTrilepton-W07-0
```

Your dataset can be found on DAS: https://cmsweb.cern.ch/das after selecting DBS instance "cms_dbs_ph_analysis_01_writer" in a drop down menu and performing the search (e.g. "dataset dataset=YOUR_DATASET")

Now you can process this dataset in CRAB similarly to the previous steps. Be careful to specify the name of the output file produced (e.g. name of the TTree root file produced by a WZAnalyzer module) and a separate output subdirectory.

See also CRAB FAQ: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCrabFaq

Submitting jobs to CONDOR

A brief description is given on how to submit jobs to condor. This can be used for submitting any type of executables, including CMSSW programs.

Note: you are strongly encouraged to use condor to run any CPU and/or memory intensive job that will take longer than a few minutes to run.

Note: If you run CMSSW on local datasets using CRAB as explained above, you are actually implicitly using CONDOR.

Create a job description file `job_desc.txt` with this content:

```
executable  =  your_executable
universe    =  vanilla
log         =  Your_Log_File
Output   = Your_output_file
error       =  Your_error_file
initialdir  =  your_initial_directory
```

```
queue
```

If you will need the environment from which you submit the job, add the following line to the job description file (before the last line, "queue" should always be the final line):

```
getenv      =  True
```

The values of the variables `executable`, `log` and `initialdir` should be adapted to your job. You can then submit the job with the command:

```
condor_submit job_desc.txt
```

You can check the status of your job with:

```
condor_q
```

And you can check the status of all condor queues with

```
condor_status
```

All these commands can be issued from any of the 3 hosts in the "lorien forest".

This should be enough to get you started and should probably satisfy most of your needs. You can find more detailed instructions and options in the CONDOR manual.

There is also a CONDOR monitoring page where you can get statistics and history plots of CONDOR jobs.

Site Administration

Restarting gluster after shutdown

As of now, gluster will not be able to restart cleanly after shutdown, and you need to the following: (OBSOLETE - 19.08.2016)

It is possible that /STORE is not mounted on some hosts after rebooting. If that is the case, simply mount it (as root):

```
mount /STORE
```

Adding a new user (OBSOLETE after upgrade to)

This procedure is currently not automated, however the job could be simplified by a script.

On lorientree05, as "root" go to the following directory:

```
cd /etc/openldap/inputs
cp usertemplate.ldif %username%.ldif
```

choosing some new name for the %username%. modify all instances of "username" and "User Name" in the file to reflect credentials of the new user. Set up a new UID number.

The UID must not overlap any existing UID. After picking some number, for example 12345, (try to use numbers over 10000), Check that the following commands don't find anything:

```
grep "12345" /etc/passwd
```

```
ldapsearch -x | grep 12345
```

*Note: Setting the password hash is no longer necessary in the ldif file. If for some reason you need this, a SHA hash can be generated using the slappasswd command.

NOTE (17-03-2016): the hostname from which home directories in the automounter map has to be lorientree05.zef.irb.hr, and not lorientree05.irb.hr!!! Mounting will not work otherwise.

After completing the ldif file, upload it to the LDAP server:

```
ldapadd -D"cn=root,dc=irb,dc=hr" -W -x -f newusername.ldif
#Type password for ldap server. It can be found under "rootpw" entry in /etc/openldap/slapd.conf
```

If the file is inconsistent or you forgot some entries, this can fail. If successful, restart the NSCD daemon to refresh the name cache (It might take a few minutes for username to be picked up by the system):

```
/etc/init.d/nscd restart
```

Add a home directory and storage directory

```
mkdir /home/%username%
chown username:users /home/%username%
mkdir /STORE/se/cms/store/user/%username%
chown storm:storm /STORE/se/cms/store/user/%username%
```

The last command above needs to be executed from lorienmaster.

It might be necessary to "reload" autofs (on any of the Site hosts), but possibly it is not needed.

```
/etc/init.d/autofs reload
```

Finally, add the user to Kerberos DB and set password:

```
kadmin.local
addprinc %username%@IRB.HR
#now set password
q #quit
```

Password can also be changed by the user with "kpasswd" (before that type "kdestroy").

If the user wants to use AFS, s/he needs to init a CERN token:

```
kinit %username%@CERN.CH
aklog #converts it to legacy krb4 token for AFS
```

Modifying an LDAP entry for an existing user (OBSOLETE after upgrade to)

This procedure is currently not automated. However, the job could be simplified by a script.

On lorientree05, as "root" go to the following directory:

```
cd /etc/openldap/inputs
```

Create a text file which specifies which entry and for what user needs to be modified. Here is an example of a recently done change that modifies the default shell for user `ceci` to Bash

```
dn: uid=ceci,ou=People,dc=irb,dc=hr
changetype:modify
replace: loginShell
loginShell: /bin/bash
```

After completing the modification description file (in this case called `modif-shell-ceci`), upload it to the LDAP server:

```
ldapmodify -D"cn=root,dc=irb,dc=hr" -W -x -f modif-shell-ceci
#Type password for ldap server. It can be found under "rootpw" entry in /etc/openldap/slapd.conf
```

If successful, restart the NSCD daemon to refresh the name cache (It might take a few minutes for username to be picked up by the system):

```
/etc/init.d/nscd restart
```

Note: Not sure if restarting the NSCD deamon is really needed but it is done here just to be on the safe side.

Reconfiguring the number of condor queues on each host

To see the number of available queues type on lorientree05:

```
condor_status
```

To reconfigure it on a given host, login to that host and edit the file /etc/condor/condor_config.

The line with the number of slots is

```
NUM_SLOTS = 20
```

After changing the value let condor reload its configuration:

```
systemctl reload condor
```

(restart instead of reload will also work)

How many slots should you open on each host? That depends on the expected average RAM usage of your jobs. We should try to avoid getting into a situation where jobs have to page, because that brings the cluster efficiency down and in addition makes interactive work very slow.

Some recommendations: in general you can use:

```
nr of slots = (available_ram - safety_margin) / average_job_memory_usage
```

For the safety margin:

- 20GB on master and lt05 to leave enough memory available for interactive work and the server processes running on these 2 hosts.
- 5 GB on all other hosts

By default we assume that the average job uses up to 2GB of memory. But there are cases, such as producing VH Shapes with the AnalysisTools ShapeMaker program, that use significantly more and require to reduce the number of slots.

Available RAM on lorien hosts:

- 32 GB on lt03 and lt04
- 64 GB on all other

(Re)starting PHEDEX scripts

```
su - storm #as root
cd ~
source stopallkill
source cleanall #wipes all logs and previous state (optional)
#check that no phedex perl scripts are running
ps aux | grep phedex
source startall
```

NOTE 2018-02-02: originally, this was done from the phedex user, but some access problems showed at some point and we switched to the same user running the processes (Information from Srecko).

Completing PHEDEX transfers which complain about duplicate files

In some cases, it can happen that the transfer job fails for some reason (for example, overloaded server so the checksum script times out), while the file remains on the disk. Then Phedex will retry the transfer, but complain about duplicate file (noticeable in /var/log/storm/storm-backend.log, or in error log on Phedex web).

The simplest trick is to (log in as root) rename the dataset directory to a temporary name, to allow those transfers to complete. Then after transfer is 100%, move files back from the temporary into correcty location (possibly keep the latter copy of duplicates because it passed the checksum). It is also possible to delete the offending files, but has to be done for each of them (a lot of work). It is generally recommended not to overload the "forest" machines while Phedex transfers are ongoing to avoid these problems. Alternatively, checksum script could be modified to detect the stall and do some proper action (or catch the kill signal and delete the file before terminating?), so there is a TODO item for this.

File system troubleshooting

ALL COMMANDS IN THIS TROUBLESHOOTING SECTION NEED TO BE EXECUTED AS ROOT.

Home directories in /users (or part of it) not visible

In case some of the home directories under /users/ become invisible, mostly on lorientree01 and lorientree02, you can try to restart the NFS service on lorientree05.

```
/sbin/service nfs restart
```

This will probably solve the problem (though give it a few seconds to become effective). In case it does not, you can also try to reload the automounter (autofs) maps on the 2 tree machines:

```
[root@lorientree01 ~]# /sbin/service autofs reload
Reloading maps
```

Parts of gluster distributed filesystem /STORE invisible

It can happen that parts of the distributed gluster file system in /STORE become invisible. You may notice that some files are invisible or that the total visible /STORE file system is smaller than what it should be, simply with the df command. e.g.:

```
[root@lorienmaster ~]# df -h
Filesystem            Size  Used Avail Use% Mounted on
/dev/mapper/VolGroup00-LogVol00
                      162G   57G   97G  38% /
/dev/mapper/VolGroup01-LogVol02
                      7.8T  3.2T  4.2T  44% /home
/dev/sda1              99M   27M   68M  28% /boot
tmpfs                  32G  612K   32G   1% /dev/shm
/dev/sdc1              25T   24T  831G  97% /export/brick1
AFS                   8.6G     0  8.6G   0% /afs
glusterfs#lorienmaster.zef.irb.hr:/gv0
                       71T   55T   17T  78% /STORE
cvmfs2                 20G   16G  3.8G  81% /cvmfs/cms.cern.ch
```

The size of the /STORE file system should be 79 TB (NEW: the total capacity is 96 TB after the integration of 2 new servers in May 2015). So in the example above, part of it is missing. We can find out which "brick" (element of the gluster fs) is missing by checking the status of the gluster volume:

```
[root@lorienmaster ~]# gluster volume status
Status of volume: gv0
Gluster process                                         Port    Online  Pid
------------------------------------------------------------------------------
Brick lorientree02.zef.irb.hr:/export/brick1            24009   Y       25405
Brick lorientree01.zef.irb.hr:/export/brick1            24009   Y       1519
Brick lorienmaster.zef.irb.hr:/export/brick1            24010   N       20910
NFS Server on localhost                                 38467   Y       20916
NFS Server on lorientree01.zef.irb.hr                   38467   Y       1524
NFS Server on lorientree02.zef.irb.hr                   38467   Y       25410
```

In this case, we see that the brick on lorienmaster is not online. We can restart it as follows:

```
[root@lorienmaster ~]# gluster volume stop gv0
Stopping volume will make its data inaccessible. Do you want to continue? (y/n) y
Stopping volume gv0 has been successful
[root@lorienmaster ~]# gluster volume start gv0
Starting volume gv0 has been successful
```

If you see that another host is not online, you need to login to that host and execute those commands there.

We can now recheck the status.

```
[root@lorienmaster ~]# gluster volume status
Status of volume: gv0
Gluster process                                         Port    Online  Pid
------------------------------------------------------------------------------
Brick lorientree02.zef.irb.hr:/export/brick1            24009   Y       21159
Brick lorientree01.zef.irb.hr:/export/brick1            24009   Y       9833
Brick lorienmaster.zef.irb.hr:/export/brick1            24010   Y       25150
NFS Server on localhost                                 38467   Y       25156
NFS Server on lorientree01.zef.irb.hr                   38467   Y       9838
NFS Server on lorientree02.zef.irb.hr                   38467   Y       21164
```

Everything is now fine, and we can now see that /STORE recovered its full size:

```
[root@lorienmaster ~]# df -h /STORE
Filesystem            Size  Used Avail Use% Mounted on
glusterfs#lorienmaster.zef.irb.hr:/gv0
                       96T   79T   17T  83% /STORE
```

NOTE on 23.02.2016.: We added 2 new servers in May 2015, and splitted bricks to have all equal size bricks, the status is now:

```
[root@lorienmaster ~]# gluster volume status
Status of volume: gv0
Gluster process                                         Port    Online  Pid
------------------------------------------------------------------------------
Brick lorientree03:/export/sdb1/brick1                  49152   Y       2923
Brick lorientree03:/export/sdb2/brick2                  49153   Y       2922
Brick lorientree03:/export/sdb3/brick3                  49154   Y       2932
Brick lorientree04:/export/sdb1/brick1                  49152   Y       2953
Brick lorientree04:/export/sdb2/brick2                  49153   Y       2959
Brick lorientree04:/export/sdb3/brick3                  49154   Y       2964
Brick lorientree02:/export/sdb1/brick1                  49152   Y       12898
Brick lorientree02:/export/sdb2/brick2                  49153   Y       12906
Brick lorientree01:/export/sdb2/brick2                  49156   Y       22128
Brick lorientree01:/export/sdb3/brick3                  49157   Y       22129
Brick lorienmaster:/export/sdb2/brick2                  49152   Y       3076
Brick lorienmaster:/export/sdb3/brick3                  49153   Y       3075
Brick lorientree02:/export/sdb3/brick3                  49154   Y       12911
Brick lorientree01:/export/sdb1/brick1                  49158   Y       22127
NFS Server on localhost                                 2419    Y       3089
NFS Server on lorientree01                              2419    Y       3372
NFS Server on lorientree02                              2419    Y       12933
NFS Server on 192.168.111.63                            2419    Y       2939
NFS Server on lorientree04                              2419    Y       2971

Task Status of Volume gv0
------------------------------------------------------------------------------
Task                 : Rebalance
ID                   : 697ef0ff-91c5-4e82-9f14-25a6b956944d
Status               : completed
```

/STORE not accessible on lorienmaster

It happened the whole /STORE file system was invisible on lorienmaster, due to a problem with the mount point. What you would then see is:

```
[lorienmaster] /users/tsusa/wbbAna/compiled_code/code> df -h /STORE
df: `/STORE': Transport endpoint is not connected
```

For now, we were only able to solve it by rebooting lorienmaster, which is obviously highly unsatisfactory. This problem has usually been observed when many processes were accessing files on /STORE. It has however been noted that /STORE is completely visible on the other two hosts.

Condor troubleshooting

Sometimes condor schedd complains about a directory in /tmp to which it does not have write permissions. Maybe you have to delete this file to get Condor working again, so delete this as necessary (the exact error is logged in one of Condor log files --> /var/log/condor/).

If you see that Condor restarts jobs after restarting it: To wipe them out, run (possibly as root):

```
condor_rm -all -forcex
```

Useful links

https://goc.egi.eu/portal/index.php?Page_Type=Site&id=475

https://mon.egi.cro-ngi.hr/nagios/cgi-bin/status.cgi?host=lorienmaster.irb.hr

https://mon.egi.cro-ngi.hr/nagios/cgi-bin/extinfo.cgi?type=1&host=lorienmaster.irb.hr

https://cmsweb.cern.ch/phedex/prod

Log of changes

- 19.12.2014: add PREEMPT and CLAIM_WORKLIFE options to /etc/condor/condor_config.local
- 06.06.2016: /STORE was not mounted after rebooting on lorientree0{1,2,4}: simply mount it "mount /STORE"

-- VukoBrigljevic - 26 May 2014

This topic: Main > IRBTier3Instructions
Topic revision: r14 - 2021-02-24 - VukoBrigljevic