



# CRAB: Advanced

J Term III  
15 January 2009

Eric Vaandering

CMS/Fermilab



# Audience

- Assume you are already familiar with using CRAB
  - Lots of developments in the last year
  - Online tutorial for first time users
    - <https://twiki.cern.ch/twiki/bin/view/Main/EricVaanderingCRABPreTutorialJTerm>
- Give you an overview of things you can do with CRAB, some new, some never really explained
  - Using CRABServer & LPC CAF
  - Staging out and publishing data
  - Multiple datasets, RAW data
- CRAB Coming Attractions

# Goals

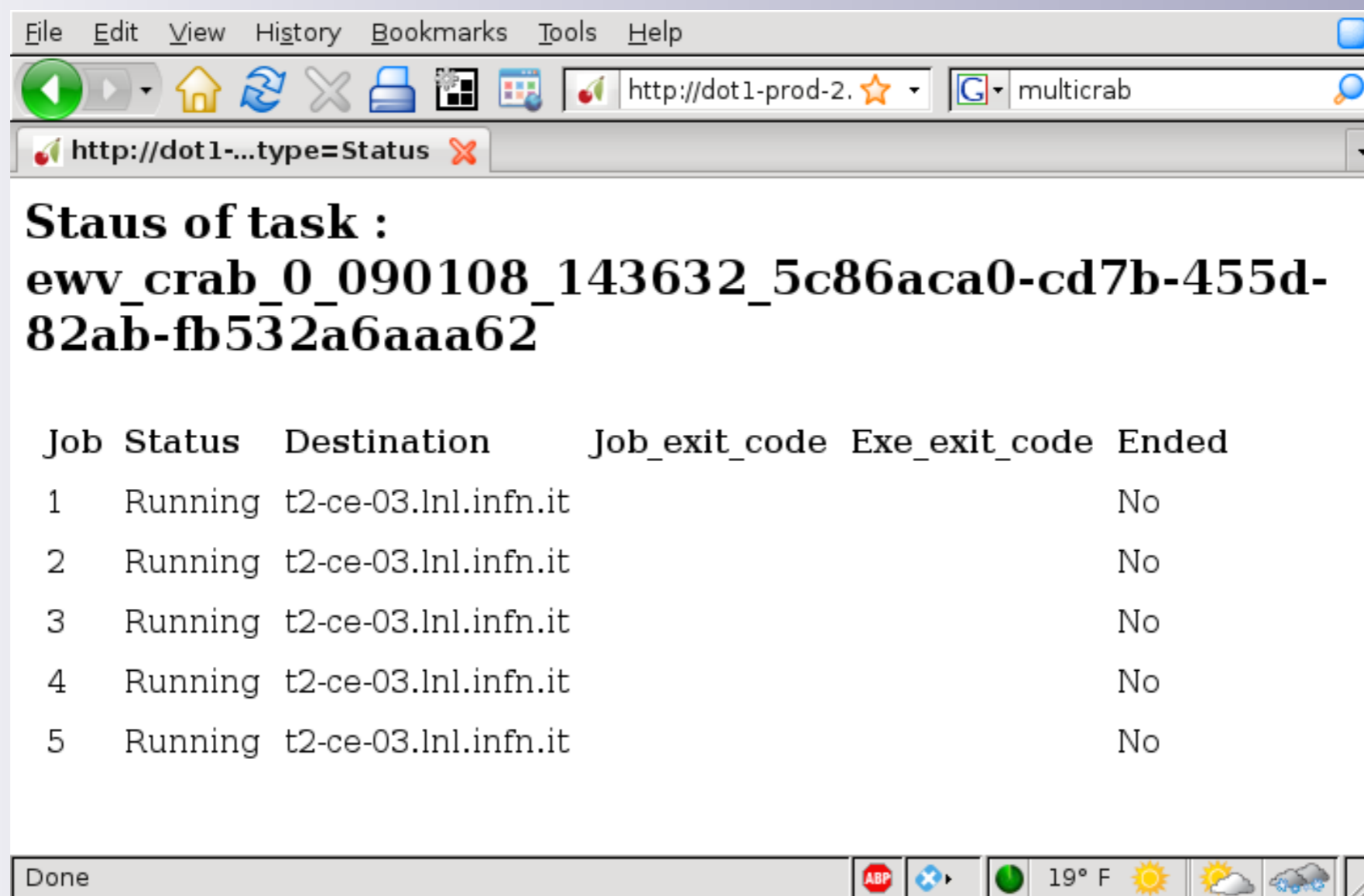


- This talk won't explain in detail how to do anything but the most trivial things
- Instead, a quick overview and a pointer to more information
  - Usually a Workbook page or other Wiki page
- Following the talk (and break) we'll have a hands-on session
  - Encourage you to pick something new from the list I show and try it
  - PAT/SK starts at 11:00

- What is the CRAB Server?
  - A central server (just a few in CMS)
  - Submits jobs, watches, re-submits
  - Collects output for easy retrieval
  - Best for large numbers of jobs (~20 or more)
  - Does a lot of the work of the user
  - Can notify you when your jobs are done

- To first order very simple
  - Add `server_name = bari` to [CRAB] section of `cfg`
  - [https://twiki.cern.ch/twiki/bin/view/CMS/CrabServer#Server\\_available\\_for\\_users](https://twiki.cern.ch/twiki/bin/view/CMS/CrabServer#Server_available_for_users) keeps a list of servers you can use
- All commands (`-create`, `-submit`, `-status`, `-getoutput`, `-kill`) work as usual
  - CS checks status itself every few minutes, so you are getting cached information
- Can be notified when your jobs finish
  - eMail = `user@fnal.gov`
  - `thresholdLevel=100`

- In addition to the usual `crab -status`, you can directly check on your jobs with the server
- `crab -printId` gives you the task name and URL to use



**Staus of task :**  
**ewv\_crab\_0\_090108\_143632\_5c86aca0-cd7b-455d-82ab-fb532a6aaa62**

Job	Status	Destination	Job_exit_code	Exe_exit_code	Ended
1	Running	t2-ce-03.lnl.infn.it			No
2	Running	t2-ce-03.lnl.infn.it			No
3	Running	t2-ce-03.lnl.infn.it			No
4	Running	t2-ce-03.lnl.infn.it			No
5	Running	t2-ce-03.lnl.infn.it			No

- CRAB will no longer submit jobs to CMS Tier1 sites
  - Tier0 blocked some time ago
  - Tier1 resources are committed to central processing for CMS
  - Nothing left over for chaotic user analysis
  - If your favorite dataset is not at a Tier2, then you have to request it be copied to one
  - Tier2's are associated with Analysis Groups  
<http://indico.cern.ch/materialDisplay.py?contribId=28&sessionId=22&materialId=slides&confId=41026>
- Not quite an absolute ban, as we'll see later

- One way around the block on Tier1 sites
- From cmslpcXX you can submit to the local condor queue
  - No need to write your own job splitting
  - Provides the same interface you are used to
  - Data must be at FNAL, but LPC CAF has access to all the same data as the FNAL Tier1
- Just change to “scheduler = condor” in your crab.cfg
- One time setup with ~/.profile, for details see <https://twiki.cern.ch/twiki/bin/view/Main/CRABonLPCCAF>



- Every user will have access to /store/user space at a Tier2
- CRAB has a nice interface to this
  - Don't need to know details of directory hierarchy at a remote site
  - [USER]  
copy\_data = 1  
storage\_element = T1\_US\_FNAL\_Buffer  
user\_remote\_dir = myTopAnalysis
  - Will look up your HN username from SiteDB
  - You can get storage space at your “local” Tier2
    - [http://www.uscms.org/uscms\\_at\\_work/software\\_computing/tier2/store\\_user.shtml](http://www.uscms.org/uscms_at_work/software_computing/tier2/store_user.shtml)

- The previous way of staging out to dCache resilient is still supported
- Not put onto tape (unlike /store/user at FNAL)
- crab.cfg settings have changed

```
[USER]
copy_data = 1
storage_element = cmsrm.fnal.gov
storage_path = /srm/managerv2?SFN=/resilient/USERNAME/OPTDIR/
user_remote_dir = SUBDIRECTORY
```
- For more options, see [https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCrabHowTo#Store\\_output\\_with\\_CRAB\\_2\\_4\\_serie](https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCrabHowTo#Store_output_with_CRAB_2_4_serie)

# Publishing Data

- You have to publish into a local DBS, not Global Production DBS
- Before submitting jobs, you have to prepare your crab.cfg config file to eventually publish
- [USER]  
copy\_data = 1  
storage\_element = CMS\_site\_name  
(i.e T2\_IT\_legnaro)  
publish\_data=1  
publish\_data\_name = data\_name\_to\_publish  
(i.e myprocessingCMSSW\_1\_6\_8)  
dbs\_url\_for\_publication = your\_local\_dbs\_url  
(i.e https://cmsdbsprod.cern.ch:8443/cms\_dbs\_prod\_local\_09\_writer/servlet/DBSServlet)
- crab -publish when your output is retrieved
- More info:  
<https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCrabForPublication>

- CMSSW supports running a job on two groups of files simultaneously
  - 2<sup>nd</sup> dataset must be parent of the dataset
  - Used to access event products discarded in later processing
  - e.g. RECO+RAW it was produced from
- CRAB supports this too
  - Both datasets must be at the same site
  - Set `use_parent = True` in the [CMSSW] section of `crab.cfg`

- Run CRAB multiple times on different datasets (other use cases too)
- Typical config file:

```
[MULTICRAB]
cfg=crab.cfg
[COMMON]
CMSSW.total_number_of_events=-1
[Wmunu]
CMSSW.datasetpath=/Wmunu/CSA08_CSA08_S156_v1/GEN-SIM-RECO
CMSSW.events_per_job=1000
[Zmunu]
CMSSW.datasetpath=/Zmunu/CSA08_CSA08_S156_v1/GEN-SIM-RECO
CMSSW.events_per_job=5000
```
- Creates a crab job in Wmunu/Zmunu directories
- Same commands as crab (-create, submit, etc.)
- More info:  
<https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideMultiCrab>

- -extend: submit jobs on new data in a dataset (open datasets)
- -copyData: retrieve data from SE to working directory
- -match: shows if resources are available for your job (after white/blacklisting)
- -postMortem: why did your job abort?
- `$CRABDIR/python/full_crab.cfg` shows a config file with ALL available options

- CondorG + CRABServer: quick, reliable access to US resources
- scheduler=glidein: pilot jobs pull your job to a site where CPU available and releases verified
- CRABServer+glidein
- Submission to CRABServer without LCG middleware (easily installed on your laptop)
- CRABServer for access to Tier1 data
- Support for Analysis Datasets
  - List of good run/lumis or run/events with a dataset