



CRAB status

Stefano Lacaprara

Laboratori Nazionali di Legnaro

Computing integration meeting 05-Oct-2006



Data Discovery

- **Data Discovery**
- Use Global DBS/DLS: it's also possible to change the instance of boths (for experts only)
- **New Data Discovery page available at:**
<http://cmslcgco01.cern.ch:8001/> (by Valentin Kuznetsov)
- Data discovery based on SE (not CE), so all CE “close” to a given SE are selected: if a site has many CEs - eg CERN - all are eligible.



Splitting

- **Splitting**
- Much improved wrt previous version:
- Deals correctly with dataset whose blocks are distributed among different sites;
- Possible to access just a fraction of a file (new)
- **cfg parameters**
 - ▶ `total_number_of_events`
 - ▶ `events_per_job`
 - ▶ `number_of_jobs`
- With obvious meaning: of course only (and exactly) two parameters can be used together.
- **Warning** Due to the intrinsic granularity of a dataset (a job cannot cross a file-block boundary) it is possible that the actual number of jobs or events per jobs is not exactly the one required. Anyhow, all required events will be processed.



ORCA support

- **ORCA is not supported with this version of CRAB**
- If you want to use ORCA and corresponding data please use CRAB_1_2_1
- on afs please use
`afs/cern.ch/cms/ccs/wm/scripts/Crab/crab_orca.csh | .sh`
- **Warning** Data Discovery and location for ORCA data is working as before, so you do have access to that data.
- The same story apply also for FAMOS jobs.



BOSS4

- **First version based on BOSS 4**
- No change visible to end-user, but a speedup of creation time (depends very much on file-system used, afs is apparently the worst)
- Many changes internal in CRAB code, that's the reason why we decide to drop ORCA for this version.
- Needed for gLite bulk submission (see after)
- Allows to build RPM (see after)



Other

- As usual, incremental bug fixes
- **Warning:** it is not possible to `-create` only a fraction of a task. So just use `-create` and **not** `-create N` (an error might occurs in latter case)
- improvement of output which has been both reduced and made (hopefully) more readable.
- distribution is still tar-ball based, we are working on a rpm-based one. Once available both distribution will probably co-exist for a while.



Future

- This is the first version which should be used to access large dataset with CMSSW
- We do expect problems with CRAB, with data discovery, with sites, sw installation, etc etc . . .
- So first priority for future (and present) is **user support** and bug fixing
- We foresee a update to 130 with solution of all bugs found in the meanwhile for end of next week
- **As usual, user support takes a lot of developers time, so this could reduce the time to further development**



Future (II)

- Meanwhile, we are also working on:
 - ▶ Performance improvement: especially speedup of creation and optimization of status and getoutput (still rather slow)
 - ▶ Integration with Dashboard: must check that we do send correctly and send all the information needed
 - ▶ Framework Job Report. We have to force it for all jobs, and use its infos such as:
 - ★ I/O report to be sent to Dashboard;
 - ★ correct exit-status: the one reported now is truncated to 32 bits by the shell
 - ▶ copy to SE (srmcp will be used first, then lcg-cp if fails)
- Publication of CRAB output to (local scope) DBS/DLS: unlikely for next week.



Future: gLite

- gLite and bulk submission

- ▶ Almost ready thanks to BOSS4
- ▶ Need some mods in CRAB: well advanced! It might possibly be ready for next week
- ▶ **but** I'm not much in favour to start using gLite just before (or better, during) CSA06, where we should use, in principle, tested and “mature” stuff.
- ▶ IMHO, edg-submission should stay as baseline: will try to have gLite+bulk but as “brave user mode”.

