Job preparation with CRAB

Stefano Lacaprara

Department of Physics
INFN and University of Padova

FNAL Physics Week, Software-Computing parallel session, 13 april 2005

Executable and Libraries

- User declare which executable want to run
- CRAB works in an environment where `eval ‘scram run -csh‘` is required
- so, executable is already in the shell `$PATH`: same for libraries in `$LD_LIBRARY_PATH`
- `which executable` exe location
- `ldd ‘which executable‘` which libraries are used
- If executable in user working are, put it into tgz
- If libraries in user working are, put it into tgz
- Issues:
  - Plugin libraries not dealt with: eventually pack all libraries in private area
  - Assume UI and WN has same OS: still open (possible to recompile, but maybe nota smart idea)


.orcarc

- Totally under user responsibility
- CRAB just pick up whatever user provides
- Only changes are in catalogs and/or input data
  - Remove user `InputFileCatalogURL` and add the proper catalogs as retrieved from PubDB
  - Replace user `InputCollections` with the correct one to access dataset/owner as declared in crab.cfg
  - Replace/add `FirstEvent, MaxEvents` to accomplish job splitting
  - Splitting by Run (`InputCollections = /System/Owner/Dataset/EvC_RunXXX`) not yet in place

Geometry

- Accessed via XML/DDD
- Dealt with by runtime configuration (scram)
- Release Geometry is available (as xml file set) on remote site
- If user want to change something, he just have to put the modified xml in the proper place in his working area on the UI (`ORCA_X_Y_Z/src/Data/...`)
- If anything is present in UI `Data` directory, CRAB just pack it and ship it with the job to WN
- CRAB reproduce exactly the UI working area on the WN, the scram deals with environment configuration
- Same for any other file (such as HLT xml configuration) which can be placed in `Data`

Ancillary files

- any other files the user code could need
- User must declare them as *to be shipped*, and CRAB ship them to WN
- On the WN the files will be placed on the directory where the executable runs

Output files

- User must declare them as such, specifying file names
- Up to user to define these file names in his `.orcarc` or hardcoded or anywhere
- CRAB post-processing stage will change the name according to splitting *e.g.* `MyOutput.root` will become `MyOutput_1.root`
- According to user requirement, the output files will be returned back via output sandbox, or copied to a `gsiftp` server or registered in a SE
- Any number of output files are allowed: names must be known in advance
- Requirement to `tgz` output files, would solve also unknown–in–advance file names
- Output merging can happen on UI (if output returned there)

New EDM

- Geometry:
- Sources: XML (via DDD) &/—— Alignment section of Condition DB
- How will this accessed? Where?
- Will the DB(s) available at remote sites (as today happen for Geometry xmls)
- Configuration needed to access the DB: who? Scram? Site specific?
- If not available locally: should ship with the job? What? Full DB, section, cache, pre–selected info? How to know which part of DB will be needed by job?
- Would there be a global/regional/T0/T1/T2 (*ad lib*) DB access guaranteed?
- Conditions: Same comments and issues as above

- ParameterSet:
- Today algorithms configuration parameters are:
  - default is hard–coded in algos
  - can be changed in user code
  - can be defined in user .orcarc
- What about new EDM?
- *ParameterSet will be handled by an external service*
- Something like the CondDB? If yes, same issue as above, plus other in case user job want to write actual job configuration in the "service" for Job provenance

- Input Data
- Today, COBRA need catalog(s) plus Owner/Dataset/Collection and then find the files (first Meta, then EVD)
- EDM? Same/similar? Input will be list of files? ...?
- All this is for batch jobs
- What about interactive jobs? Should I care about them or they will only run on laptop/T3?