Multicore jobs: status and plans

&

Experience of ARC CE deployment on LCG T2

- Rod Walker, LMU 10th Dec 2012
- Andrej Filipčič

Multicore background

- AthenaMP processes events in parallel, 1 process per core
  - motivation is high RAM usage of large μ reco
  - fork after 1st event to share RAM
  - can then use all cpu cores
- Secondary motivations
  - save memory on new multicore CPUs
    - scaling of job submission & startup
  - produces fewer and larger output files
  - nice fit for whole-node scheduling/cloud resources
    - run single job and return node/slot when finished
- Use for G4 too
  - no RAM requirement, but other motivations apply
  - must fill dedicated MCORE resources when no reco

Status & Plans

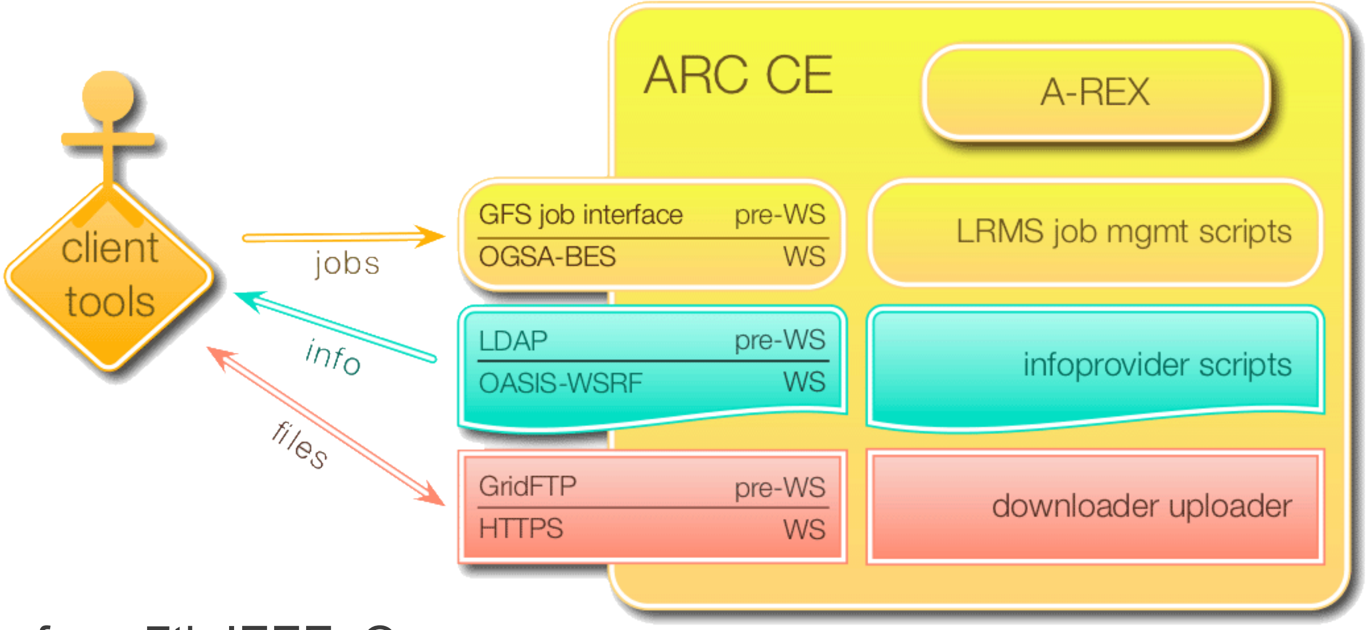
- Requested multicore(8) resources to run Sample A validation
  - around 10 sites provided these MCORE queues
  - validation not successful: even for G4
    - no show stopper, just lowish priority(linked to 25ns story)
    - reco requirement not arisen in 2012 running
  - big change to io part of AthenaMP planned for LS1
- No need for more sites
  - minimize dedicated resources of existing sites
    - some sites free up 8 cores dynamically, so no waste here
- Still need time to test before it does become critical

NorduGrid ARC CE evaluation

- Motivation from HPC cluster(possible usage)
  - no network connectivity from WNs
    - rules out pilot model and data stage-in/out
  - odd Batch system:Loadleveler
- ARC CE famous for low requirements on site
  - supports Loadleveler, no WN IP, data via shared FS
    - good shared FS needed for pileup/analy
- First test on existing LRZ T2

ARC CE description

- Fully-functional Grid CE,with x509/voms auth
- Part of EMI/UMD with strong developer team
- In use at ~20 sites, including those in ND cloud

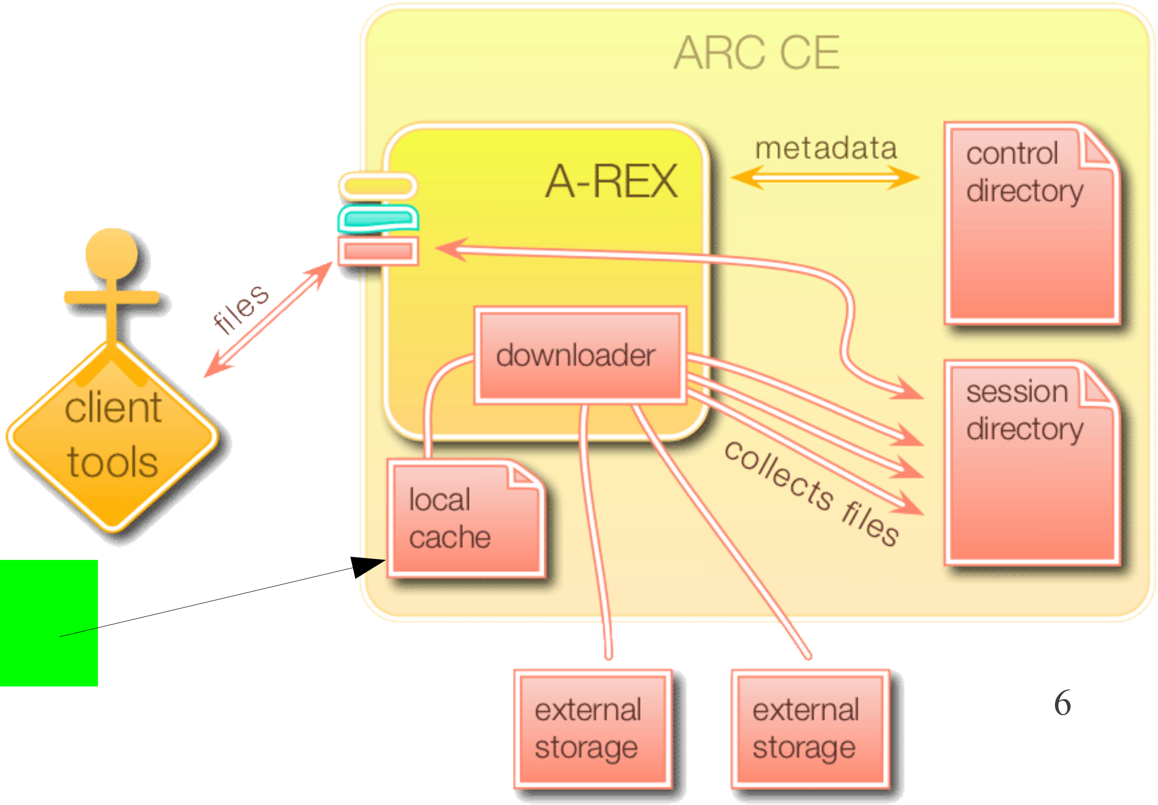


Pictures taken from 7th IEEE, Oxana.

But there is more ...

- Job description language includes input and output data. ARC CE takes care of...
  - input data staged to local cache before LRMS submit(gridftp,https,xrootd,.. from preferred source)
  - output data stored to final destination after job leaves LRMS

- Subscriptions, DDM, FTS,SRM not need to run jobs



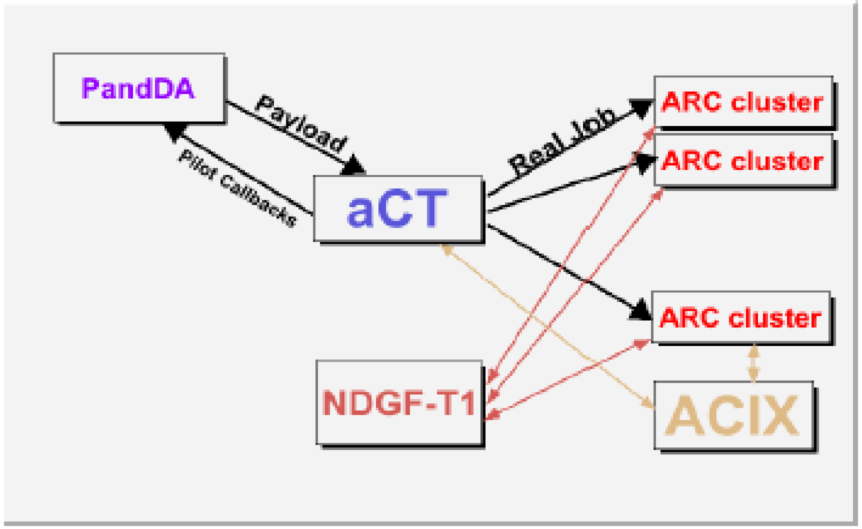
1 or more shared FS areas, accessible to WNs

Panda integration

- ARC CE has real jobs with their data
- Panda pilot model - pull job to WN

🚫

- ARC Control Tower(aCT) mimics pilots
  - submits real jobs to ARC



Some generalizations needed to use in other ATLAS clouds. Ready in Jan`13, if needed. Eg. would not channel all data via ND T1

Taken from CHEP paper

LRZ experience

- Installed ARC CE in morning and had real ATLAS jobs via ND in the afternoon.
- 1 metarpm plus CAs, in epel,emi repos
- NO yaim! Single config file /etc/arc/conf.
- Documentation is good (used Andrej for speed)
- Much nicer experience than some middleware
- Typically running 800 ARC jobs, plus LCG(cream)
  - CSCS has run happily like this for years

Straight cream replacement

- Submit normal prod and analy pilots via ARC CE
  - uses existing Panda, ddm, fts, srm
  - pilots submitted via CondorG from APF2
    - HTCondor supports ARC
      - tested JDL: being added now to APF2
- Why replace?
  - better 1 ARC than 2 shaky creams(SGE maybe worse than PBS)
    - despite good support from Portugal – cream is complex
  - single ARC CE scale tested to 4500 cores
- Accounting in APEL?
  - works now by piggy back on Cream LRMS log parsing.
  - standalone method being worked on.

Lower the bar for non-grid sites

- Significant resources at off-gridT3s
- One reason for being 'off' is intrusive, high-maintenance middleware
  - SRM + Cream CE, bdii, apel
  - GGUS tickets, exclusion
- Just ARC CE, cvmfs, NFS area
  - GDP will use any free cpu

What about user analysis?

- It is done in ND cloud.
  - can be slow for many short jobs with lots of input
- I think it is reasonably efficient from performant shared FS
  - think storm sites(GPFS,Lustre) do well in HC tests
- Until we know more, analysis can stay using LCG pilot via ARC
  - data pre-placed(mostly static) in site SRM

Conclusion

- I wish I'd tried it sooner
- not closely compared performance c.f. LCG pilots
  - e.g. get more pileup at T2s?Follow priority better.
  - shared FS has role of dCache/DPM – NFS4.1?
- could replace creams for all ATLAS needs
  - submit normal pilots, plus get aCT option
- maybe off-grid T3s would install it
- welcome a few beta testers if there is interest

Xrsl job def snippet

```
&("inputfiles" = ("EVNT.01027462._000988.pool.root.1" "lfc://
prod-lfc-atlas.cern.ch/:guid=D033C1BE-0720-D84C-9660-8D27E6C9D455" ) ("DBRelease-19.4.1.tar.gz"
"lfc://prod-lfc-atlas.cern.ch/:guid=cbb911fe-813b-471a-880d-bff9d0f79d78" ) ("pilot.tgz" "lfc://
prod-lfc-atlas.cern.ch//grid/atlas/dq2/user/user.andrejfilipcic.production/pilot3-SULU52a1.tgz" ) ("NGpilot"
"lfc://prod-lfc-atlas.cern.ch//grid/atlas/dq2/user/user.andrejfilipcic.production/NGpilot.14" ) )("middleware"
= "nordugrid-arc-2.0.0" )("runtimeenvironment" = "APPS/HEP/ATLAS-17.2.2.6-X86_64-SLC5-GCC43-
OPT" )("walltime" = "112740" )("cputime" = "112740" )("stderr" = "log.01081766._034744.job.log.1" )
("outputfiles" = ("gmlog" "" ) ("jobSmallFiles.tgz" "" ) ("HITS.01081766._034744.pool.root.1" "srm://
srm.ndgf.org
;spacetoken=ATLASPRODDISK/atlas/disk/atlasproddisk/mc12_8TeV/HITS/e1600_s1499/mc12_8TeV.16
7320.AlpgenJimmy_Auto_AUET2CTEQ6L1_VBF_ZeeNp0.simul.HITS.e1600_s1499_tid01081766_00/HI
TS.01081766._034744.pool.root.1" ) ("log.01081766._034744.job.log.tgz.1" "srm://srm.ndgf.org
;spacetoken=ATLASPRODDISK/atlas/disk/atlasproddisk/mc12_8TeV/log/e1600_s1499/mc12_8TeV.1673
20.AlpgenJimmy_Auto_AUET2CTEQ6L1_VBF_ZeeNp0.simul.log.e1600_s1499_tid01081766_00/log.01
081766._034744.job.log.tgz.1" ) ("log.01081766._034744.job.log.1" "" ) )("clientsoftware" = "nordugrid-
arc-0.8.3.1" )("hostname" = "f9pc00.ijs.si" )("clientxrsl" = "&(""jobname"" =
""mc12_8TeV.167320.AlpgenJimmy_Auto_AUET2CTEQ6L1_VBF_ZeeNp0.simul.e1600_s1499_tid01081
766._034744.job"" )(""memory"" = ""2000"" )(""disk"" = ""500"" )(""walltime"" = ""1879"" )(""cputime"" =
""1879"" )(""runtimeenvironment"" = ""APPS/HEP/ATLAS-17.2.2.6-X86_64-SLC5-GCC43-OPT"" )
(""executable"" = ""NGpilot"" )(""arguments"" = ""17.2.2.6-X86_64-SLC5-GCC43-OPT""
""logGUID=ce6a950a-0ef9-4871-933f-cf936f63086a&cmtConfig=x86_64-slc5-gcc43-
opt&dispatchDBlockTokenForOut=NULL%2CNULL&destinationDBlockToken....= ….
```