P. Nilsson (BNL), A. Anisenkov (Budker Inst.), M. Lassnig (CERN), A. Di Girolamo (CERN)

PILOT MOVERS

New site mover architecture in developments

Introduction to Site Movers

What are they and why do we want to change them?

- The pilot site movers are essentially wrapper functions for various copy tools (xrdcp, lcgcp, ..)
- The current implementation is among the oldest parts of the pilot and not all site movers are in use any more (legacy code)
- Difficult and time-consuming to maintain
- Site movers rely on complicated hard-coded logic and complex algorithms to build paths to the replicas
  - (both stage-in and stage-out paths)
- **Take advantage of the fact that AGIS already knows e.g. the full paths to the final storage area for a given queue - no need to use the old error prone practice of puzzling together schedconfig pieces that might not be correct**

Current usage list of site movers

N = Number of PanDA queues

N	Copytool pilot names	Analy stagein	Analy stageout	Prod stagein	Prod stageout	Special stagein	Special stageout
366	lcgcp2	29	85	49	199	2	2
82	xrdcp	14	2	61	5		
66	lsm	9	9	23	23	1	1
64	dccplfc	21		43	Uses dccp stage-in; not using lfc		
64	mv	6	6	26	26		
46	storm	15	1	28	2		
20	rfcplfc	9		11	Uses rfcp for stage-in; not using lfc		
14	cp	3	3	3	3	1	1
7	aria2c			7			
5	rfcpsvcclass			5			
2	gfal-copy					1	1
2	BNLdccp					1	1
2	xcp					1	1
2	curl			2			

Old vs. new AGIS/Schedconfig fields

- Old schedconfig fields are a major limitation
  - Using/maintaining/adding old/new fields have become increasingly more difficult
- Pilot grows every year
  - Many requests are made in a hurry, with little time to restructure/consolidate
- Adding support for new transfer protocols can be very complicated
  - Last major effort related to site movers: support for object stores, which we are still improving on (Wen Guan)
- We decided to improve this situation last summer
  - Alexey is designing and testing the new site mover architecture and have started to implement several site movers; Mario has implemented a rucio site mover
  - Not done yet but well on our way!

Schedconfig attributes (1/3)

List of schedconfig attributes used to configure copytool/site mover in old implementation – **don't panic, we will not go through them here..**

- **COPYPREFIX** - Prefix used to build paths for stage-out. If COPYPREFIXIN is not defined, COPYPREFIX will be used for both stage-in and stage-out. COPYPREFIX can be used to convert SURLs to TURLs. In general, COPYPREFIX = "from_prefix/to_prefix".
- **COPYPREFIXIN** - Prefix used to construct paths for stage-in (only). Can be left unset.
- **COPYSETUP** - Special instructions related to copy tool setup and usage of file stager and direct access; e.g. "=/somepath/setup.sh^srm://lcg-se0.ifh.de/^dcap://lcg-dc0.ifh.de:22125/^False^True=".
- **COPYSETUPIN** - Essentially the same as COPYSETUP but can be used to specify a different stage-in setup than for stage-out. COPYSETUPIN can be left unset.
- **COPYTOOL** - The (inbound) outbound copy tool for (in)output files. Examples: cp, lcg-cp, storm, xcp. If COPYTOOLIN is None, COPYTOOL covers both stage-in and stage-out.
- **COPYTOOLIN** - The inbound copy tool for input files. Examples: cp, lcg-cp, storm, xcp. If set to None, COPYTOOL will be used for both stage-in and stage-out
- **DDM** - A comma-separated list of DDM endpoints used for input data. For a T1 the order should match the setoken order, because this is used to translate the output file space token into the final ddm destination. It is ok for the ddm list to be longer than setokens, e.g. to have an extra input location. For a T2 this is typically just local PRODDISK. The first value in the list has some special meanings. 1) The free space in this endpoint ends up in schedconfig.space(updated via curl). 2) If the site is a T1, then it also ends up in the cloud table space, which is used for task assignment, 3) Bamboo may choose to subscribe EVNT input between clouds, to aid brokerage - the destination space token is this one.

Schedconfig attributes (2/3)

- **SE** - space token and full endpoint of default output destination. The path is extracted from the matching token in seprodpath If output files do not have spacetoken set, then the token in this entry is used. This applies to direct storage from T1 jobs, but not to subscriptions from T2(see setokens).
- **SEIN** - Comma separated list of SEs, same length as DDM and SEPATH/SEPRODPATH for input transfers
- **SEPATH** - Same as SEPRODPATH but for analysis sites. We suggest to keep SEPATH and SEPRODPATH identical.
- **SEPRODPATH** - list of destination paths, to append to endpoint in 'se'. Compressed notation is supported /blah/[tok1,tok2]/more/. Must be the same length as setokens. To be Rucio compliant, it should end with /rucio for DISK, e.g. /bla/blabla/home/atlas/[atlasscratchdisk/rucio,atlaslocalgroupdisk/rucio], while it should not have /rucio for TAPE, e.g. of a DISK-TAPE mixed one /pnfs/gridka.de/atlas/[disk-only/atlasdatadisk/rucio,atlasdatatape,atlasmctape]
- **SETOKENS** - List of destination space tokens. The first token is the default destination(for subscription from T2) if an output file does not have spacetoken set. Must match length of SEPRODPATH

Schedconfig attributes (3/3)

- **ENVSETUP** - Command that is run (can be variable setting or script) to set up copy for (in)out transfers. If ENVSETUPIN is None, this covers all copies in and out of the queue.
- **ENVSETUPIN** - Command that is run (can be variable setting or script) to set up copy for out transfers.

Plus special settings for FAX cases:

- **copyprefixin_fax_direct**
- **copyprefixin_fax_xrdcp**
- **copysetup_fax_direc**:
- **copysetup fax xrdcp**
- **copysetupin fax direct**
- **copysetupin fax xrdcp**
- **faxredirector - to specify fax endpoint**

Key points of old implementation

- Many schedconfig attributes need to be checked and properly set by site-admins/operations to enable required site mover to be used
- Several of the attributes contain compound and very complex values that require careful manual validation and make operations really difficult
- Adding support for a new protocol/copytool becomes a challenge
- Mirror/duplicate of schedconfig settings for FAX cases
  - (special FAX settings mentioned in previous slide)
- Validating affected schedconfig attributes on AGIS side becomes difficult since several settings contain raw/compound/non-structured values (blob data)

Key points of new implementation

- Fetches structured protocol + copytools definition from AGIS in RESTful way
  - schedconfig json export - CVMFS stored, or directly from AGIS, or from PanDA server cache if required
- Allows AGIS to validate and manage settings via WebUI
- Shares SE protocol declaration used by PanDA/Pilot with DDM
- Footnote: pilot should nevertheless not have strong dependency on AGIS
  - This is just a matter of proper configuration; JSON does not have to come from AGIS..

Site mover settings in AGIS JSON

- AGIS defines 3 activities and 4 attributes for each activity
  - Activities: pr = pilot read (stage-in), pw=pilot write (stage-out of data), pl=pilot log (separate configuration of stage-out log file):

1. copytool (name of copy tool)
2. ddmendpoint
3. path (local SE path to be used)
4. se (SE protocol endpoint)

Example: Basic site mover settings

```
{
        "copytool": "lcgcp",
        "ddm": "LRZ-LMU_SCRATCHDISK",
        "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atlasscratchdisk/rucio/",
        "se": "srm://lcg-lrz-srm.grid.lrz.de:8443/srm/managerv2?SFN="
},
```

http://atlas-agis-api.cern.ch/request/pandaqueue/query/list/?json&preset=schedconf.all&panda_queue=ANALY_LRZ_TEST

Example: Lists of protocols

- Ordered lists of protocols can be specified
  - so that in case of failover the next protocol could be used for reading
- Additionally, “copytools” attribute stores extra settings specific for site movers:

```
"copytools": {
    "dccp": {
     "setup": "/cvmfs/atlas.cern.ch/repo/ATLASLocalRootBase/x86_64/emi/current/setup.sh"
    },
    "lcgcp": {
     "setup": ""
    },
    "xrdcp": {
     "setup": "$VO_ATLAS_SW_DIR/local/xrootdsetup.sh"
    }
  },
```

Example: ANALY_LRZ_TEST

```
•    "ANALY_LRZ_TEST":  {
•    "aprotocols": {
•        "pl": [
•         {
•           "copytool": "lcgcp",
•           "ddm": "LRZ-LMU_SCRATCHDISK",
•           "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atlasscratchdisk/rucio/",
•           "se":  "srm://lcg-lrz-srm.grid.lrz.de:8443/srm/managerv2?SFN="
•         },
•         {
•           "copytool": "xrdcp",
•           "ddm": "LRZ-LMU_SCRATCHDISK",
•           "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atlasscratchdisk/rucio/",
•           "se":  "root://lcg-lrz-rootd.grid.lrz.de:1094/"
•         }
•        ],
•        "pr": [
•         {
•           "copytool": "dccp",
•           "ddm": "LRZ-LMU_SCRATCHDISK",
•           "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atlasscratchdisk/rucio/",
•           "se":  "dcap://lcg-lrz-dcap.grid.lrz.de:22125"
•         }
•        ],
```

```
•        "pw": [
•            {
•             "copytool": "lcgcp",
•             "ddm": "LRZ-LMU_SCRATCHDISK",
•             "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atl  ..
•             "se":  "srm://lcg-lrz-srm.grid.lrz.de:8443/srm/man ..
•            },
•            {
•             "copytool": "xrdcp",
•             "ddm": "LRZ-LMU_SCRATCHDISK",
•             "path": "/pnfs/lrz-muenchen.de/data/atlas/dq2/atlass ..
•             "se":  "root://lcg-lrz-rootd.grid.lrz.de:1094/"
•            }
•           ]
•          },
```

Enabling the new site movers

- To enable site mover for given PandaQueue, site-admins mainly need to attach SE protocols and copytool details in AGIS WebUI
  - AGIS will centrally populate new copytool/protocol settings from old schedconfig settings
  - Test site (currently) needs to enable special use_newmover schedconfig attribute
  - WebUI currently in development

Status of new site movers

Already developed site movers within new architecture

- xrdcp
  - Tested at ANALY_CERN_TEST and CERN-PROD-preprod
- rucio
  - Tested at CERN and HEPHY-UIBK
- dccp
  - To be tested: preparing HC tests for ANALY_LRZ_TEST
- lcgcp
  - To be tested: preparing HC tests for ANALY_LRZ_TEST

Rucio site mover (1/2)

- New site mover available
  - uses CVMFS rucio upload/download
- Embedded within new site mover architecture
- Supported protocols: gfal, gsiftp, posix, srm, webdav, xroot, ngarc, s3 (experimental)
- Need as input only scope + name
  - does not require other pilot-side DDM lookups
- Supports parallel downloads

Rucio site mover (2/2)

- Mandatory improvements
  - Rucio site mover handles tracing, but so does the Pilot (need to sanitize)
- Good-to-have improvements
  - Pilot also does replica lookup upfront - only necessary if Rucio site mover fails. Otherwise we will have double lookups on Rucio Servers for the same data
  - Unnecessary movement of output files between directories; already fixed in next rucio-clients release

HammerCloud testing

- Tests now running at ANALY_FZK_TEST and HEPHY-UIBK
  - Testing dccp and lcgcp site movers on FZK
  - Testing rucio site mover on UIBK (want to expand tests to more sites)
  - Currently experiencing some problems with runAthena – site movers seem to work fine
- The plan is to add more transformations to be tested from HC side (as many as possible)
  - and in parallel include more panda queues
- Once all tests work we will ask for volunteer sites
  - Switching to new site movers will be gradual

What will be implemented next?

- Extend base site mover architecture
  - to cover alternative stage out workflow
  - to cover Object Store workflow
- Object store site movers implementation
  - Or handled by rucio site mover
- AGIS needs to finalize and deliver a WebUI to easily configure site mover settings required for new architecture