Paul Nilsson

EVENT SERVICE IN THE PILOT

Refactoring, payload integration & interaction, input/output management, object store usage

Refactoring Status

- New pilot version 59a PICARD – currently being rolled out - contains several refactoring efforts
  - glExec
    - Many changes related to this, many bugs discovered after refactoring but now corrected – **DONE**, larger scale tests to begin shortly (by Edward Karavakis)
    - By Fernando Barreiro Megino/Ramon Medrano Llamas/Edward Karavakis
  - Site movers
    - Refactoring wanted due to old code base, also because of experiment specific code such as DQ2 tracing – **IN PROGRESS** (not a high priority)
    - By Wen Guan
    - Wen has also provided new site movers (GFAL2 and for object stores)
    - Other site mover cleanup by Paul Nilsson (removal of several outdated/unused site movers)
  - RunJob modules
    - Refactoring desired to facilitate introduction of HPC and Event Service code (see later presentation) - **DONE**
    - By Paul Nilsson
  - Main pilot module
    - Partly refactored for glExec
    - Separation of bulky job recovery algorithm incl. rewrite (not a high priority) – **TO BE DONE**
    - By Paul Nilsson

Event Service Tests

- Using nightlies releases (19.X.0)
- Job task created by JEDI
  - Task submission cript created by Tadashi Maeno
- Test round consists of two jobs
  - Initial Event Service job
  - Merge job
- Event Service Tasks monitor page
  - http://bigpanda.cern.ch/tasks/?eventservice=1&display_limit=300

Initial Event Service Job

- I.e. the “main” Event Service job
  - Pilot downloads event ranges, which are processed by AthenaMP
  - Output is stored in object store
- Current test job uses EVGEN + TAG input files
  - Stage-in to disappear asap when Event Index is available and AthenaMP can direct access the EVGEN file [should now be possible]
- Pilot downloads job definition as usual
  - Event service job definition contains special identifier key ‘eventService’ (and is set to ‘True’)
- Pilot/monitor asks the Experiment object which subprocess module to be used ..
  - Which means subprocess module can easily be made experiment specific
- .. and how it should be launched
  - I.e. which python options to use
- Default subprocess module is RunJobEvent (see later slide)

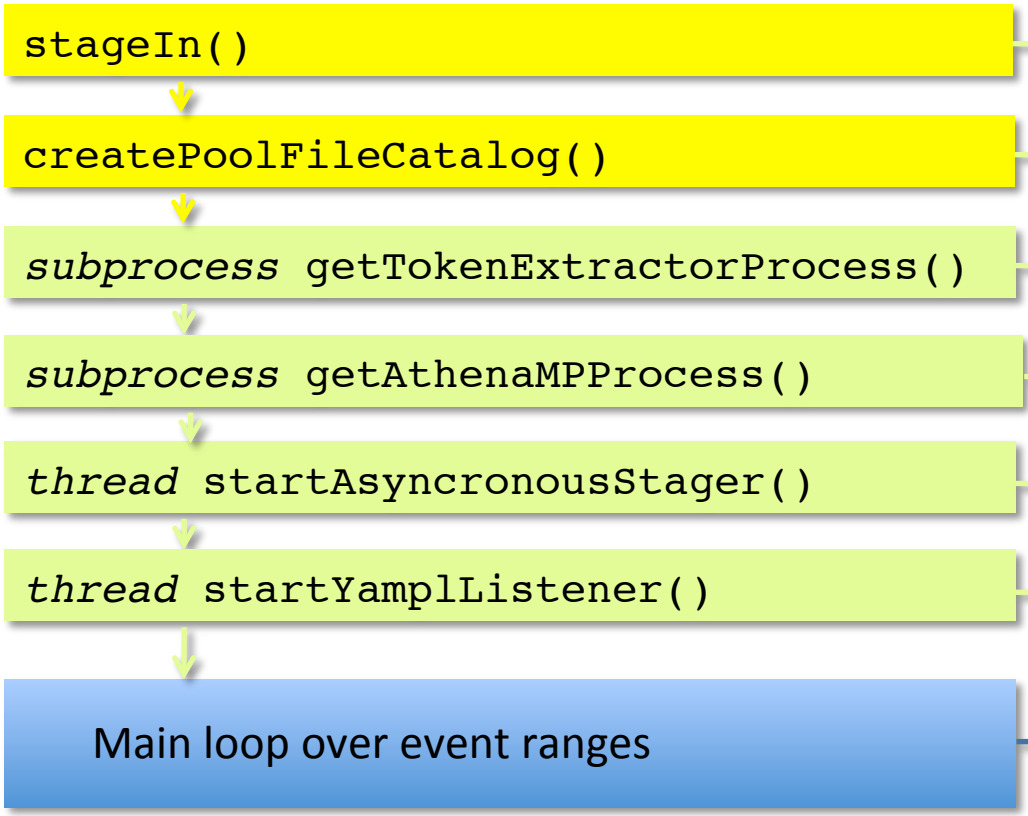
Merge Job

- Merge job treated as “normal” PanDA job, i.e. pilot uses RunJob module to execute payload
- Triggered when initial Event Service job finishes
  - Initial job goes to “cancelled” state
- Job definition contains specialHandling=‘esMerge’
  - When pilot sees this, it switches the stage-in copy tool to “objectstore”
    - Pilot stages in files from object store
    - Actual object store copy tool selection depends on specified protocol (e.g. xrootd or s3) in schedconfig.objectstore
- Merge trf merges input files and produced a single output
  - Transferred to “normal” SE

RunJobEvent

- Subprocess module that knows how to
  - Download/update event ranges from server
  - Run Token Extractor
    - Currently using TAG file staged-in by the pilot
    - New TE version available, pilot change needed
  - Run AthenaMP and communicate with it using Yampl
    - Currently pilot stages in input file needed to launch AthenaMP
    - Each event range sent to AthenaMP sequentially
  - Wait for output files to be produced
    - AthenaMP informs when a new file is ready for stage-out
    - Thread does asynchronous stage-out to object store until file queue is empty
  - Communicate with main pilot process (for heartbeat etc)
    - Currently via TCP message but is likely to be replaced by alternative method soon (yampl)
- Subclass that inherits from newly developed RunJob class
  - Base class used for running “normal” jobs
  - To be developed: RunJobHPCEvent subclass, will inherit from RunJobEvent

RunJobEvent Overview



Transfer input files (for the time being)

Trf needs a PFC to start (based on input files)

Launches the Token Extractor

Launches AthenaMP

Stage-out output files as soon as they are available

Listens for messages from AthenaMP

Download event ranges, send/ receive instructions from AthenaMP

Pilot-AthenaMP Communications

- Message server run by pilot using Yampl
  - "Pilot" really means RunJobEvent here
- Event ranges are sent sequentially to AthenaMP
  - E.g. [{u'eventRangeID': u'4017466-2244492272-56108538-1001-3', u'LFN': u'EVNT.01461041._000001.pool.root.1', u'lastEvent': 1001, u'startEvent': 1001, u'GUID': u'BABC9918-743B-C742-9049-FC3DCC8DD774'}]
  - AthenaMP starts a worker that begins to process the event range
- At the end of the event range loop, 'No more events' message is sent to AthenaMP
- When an AthenaMP worker is ready with processing an event range, it sends the location of the output file as well as accounting info (CPU and wall time)
  - "<file_path>,<event_range_id>,CPU:<number_in_sec>,WALL:<number_in_sec>"
  - E.g. /tmp/Panda_Pilot_29177_1408268256/ PandaJob_2244492272_1408268264/athenaMP-workers-AtlasG4Tf-sim/ worker_3/panda.jeditest.HITS.ad4aeafe-8d4e-4d5b-b7bc-6525549d86bd. 000001.HITS.pool.root. 1.4017466-2244492272-56108538-1001-3,4017466-2244492272-56108538-1001 -3,CPU:599,WALL:600

Object Stores

- New object store site movers developed by Wen Guan (Wisconsin)
  - objectstoreSiteMover base class for xrootdObjectstoreSiteMover (used at CERN) and S3ObjectstoreSiteMover (used at BNL)
  - Object store site mover selecion based on protocol listed in schedconfig.objectstore
  - For BNL object store secret keys are stored on the PanDA server; site mover downloads when it needs them (once per job)
- Checksums on Object Stores
  - Not provided/stored by default
  - Currently pilot can verify local vs remote checksum for the initial ES job since xrdcp returns the checksum, but then the info is forgotten (no file catalog)
  - At the moment pilot skips checksum verification until it has been added to the object stores
  - The whole point with an object store is to store objects and not only files, so it should be possible to store metadata like checksums
    - Wen found a way to store the checksums on the BNL object store
    - For the CERN object store, Andreas Peters will upgrade to a new XRootD version and configure their server for checksums