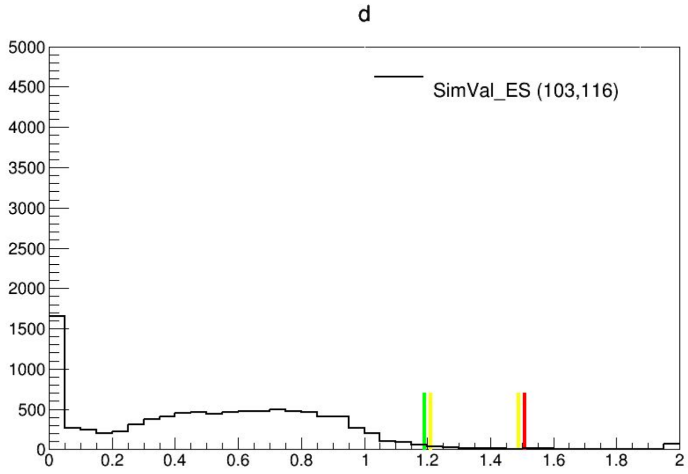
Event Service

Where are we, but in particular:
Where are we going?

ES validation

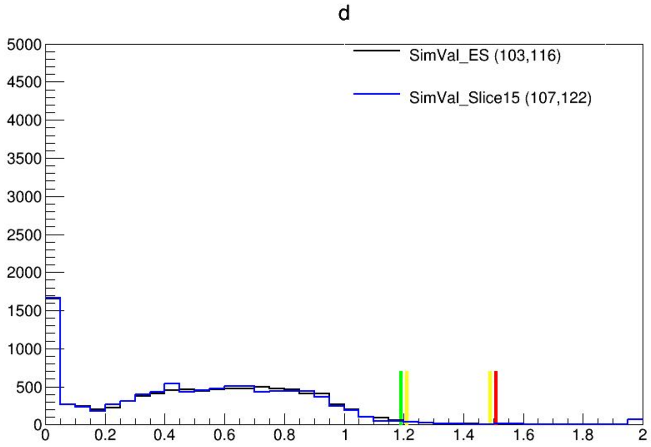
- In Valencia we saw:

two simulation jobs, (one ES and one normal) with the same seed : the distribution of Chi2s comparing final distributions were not always compatible:



So we decided, let's **take ES out of the equation :** do the same for two normal simulation jobs **:**

We see the same thing:



Which resources, and how?

- We are (almost) there for “well known” Grid resources
  - Opportunistic grid resources: BNL_Local, ATLAS_OPP_OSG*, CERN-P1
  - Some “standard” resources
  - Plus starting with some HPC
- Main changes:
  - We are able to write and read to/from OS (few around the world)
  - We are now able to use normal GridStorages
  - Automation: we are now switching tasks automatically to ES
    - Not so painless (thanks MCprod guys!)
- … and now: HPCs

Event Service on HPC: SuperMUC

- Allocation is 300 nodes(4800 physical cores) in preempt mode
    - Jobs are killed when any true HPC job needs them
    - Island structure of cluster means many jobs not preempted
- Lack of WN network means
    - events assigned to job before submission(500)
    - Stage-out must happen from ARC CE, to S3 objectstore(CERN)
        - Normally finishing jobs zip events and create metadata at the end
        - Preempted jobs have this done by the ARC CE: using same code
            - Workdir is on shared-FS and preserved on preemption
- Outstanding issue is accounting
    - walltime not filled correctly due to curious way to run with SLES11/Loadleveler/preempt

Event Service on HPCs in the US

- Complete Harvester vetting in ManyToOne mode **without Event Service** (Oct 2017)
- Distributed Harvester to other HPC sites (Oct-Nov 2017)
- Then begin testing Yoda with Harvester in OneToMany mode on Theta using a second panda queue (Dec - Mar 2018)
  - Using Jumbo Jobs
  - Verify events are running to end of wall-clock with no big gaps in processing
  - Verify outputs are being tarred and uploaded to BNL properly
  - Verify log files are being uploaded with the proper content
- Verify merging steps are succeeding (Apr 2018)
- Distribute to NERSC & OLCF (May - Jun 2018)

Storage

- The OS seems to be the natural technology for Event Service (and ES like) workflows.
    - Our (ATLAS) experience is growing up while we use them
    - A lot of monitoring implemented and available, https://twiki.cern.ch/twiki/bin/view/PanDA/EventServiceStorageWorkflow
    - Deletion is still an open point
    - Belated cry that we need pro-active OS experts to tell ES folks “how to do things better”
- To make sure we were able to commission ES itself we decided to use also normal storages:
    - For non-opportunistic resources, no real reason to upload a chunk each 10mins.
    - Now we are flexible, we can set upload freq, we register the events(files) in Rucio
    - Still some open points, e.g. the merge jobs

… next

- We are enabling “standard” resources to run EventService tasks
    - This will solve the tails problems
    - We might evolve and write the full file instead of the tarred one if a slot is able to reach the end of the full event range
- We want to evaluate the EventService for other workflows (later)
    - Understand the benefit, opening the
- We can start thinking about consolidating the EventService “bit and pieces” for all the workflows
    - (almost all, i.e. the single step trans)
    - Not going to be easy or short term, many things need to be considered (e.g. just to say one, our system was designed over “files”). Long term, to be carefully understood.