Small Files

and what ATLAS intents to do about them

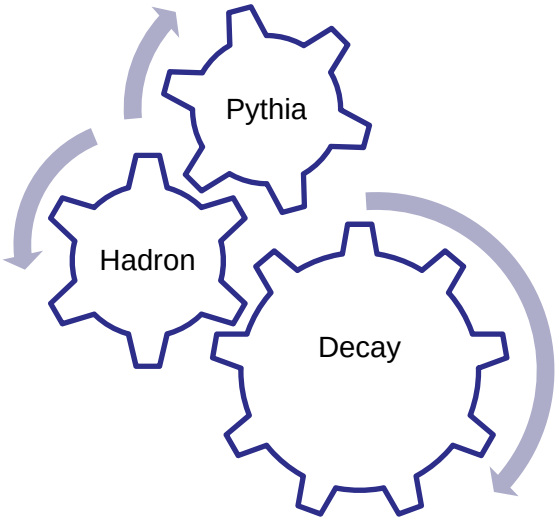
Kors Bos

NIKHEF , Amsterdam

Small files

- We want to work with files of order 5 GB
- Not smaller than 1, not bigger than 10 GB
- Our average file size is order 50 MB !!
- 100 times smaller → 100 times more files
- Bigger files are better for transport
- Bigger files are better for storage
- *Effort to create bigger files*

Production of evgen files



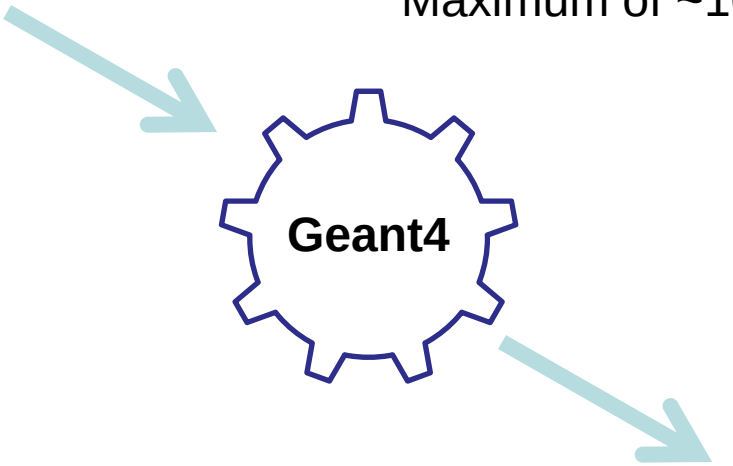
Very fast
Many thousands of events generated
Files of order 100 MB
Needed everywhere where simulation is done
Distributed to all T1's

evgen

Production of HITS files

evgen

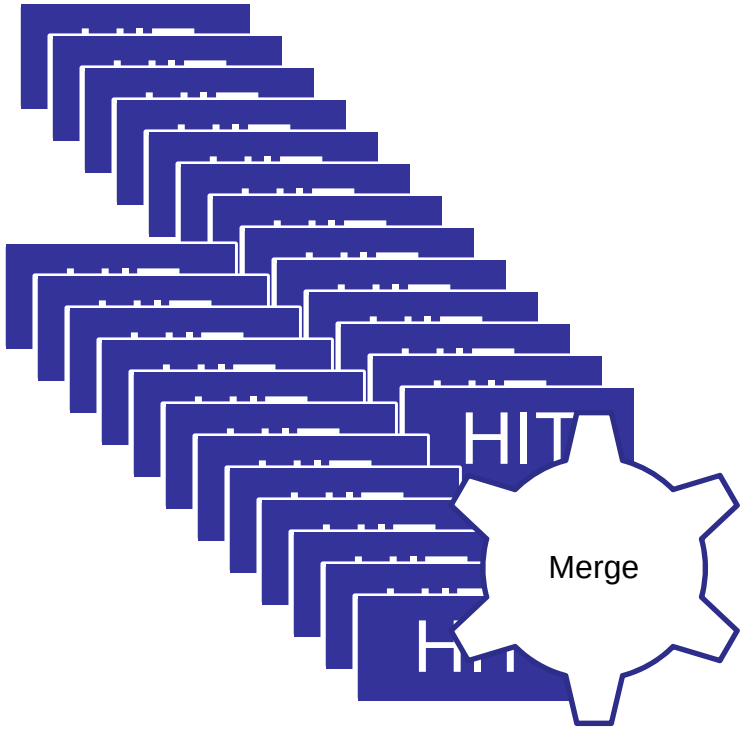
GEANT4 very slow
1 event takes 1400 kSI2K.sec = 700 sec
Cpu time limited to ~24 hours
Maximum of ~100 events



HIT = 2 MB / event
We do simulation runs with 50 events
So HITS files are order 100 MByte
Need to be merged to create bigger files

HITS

HITS Merging

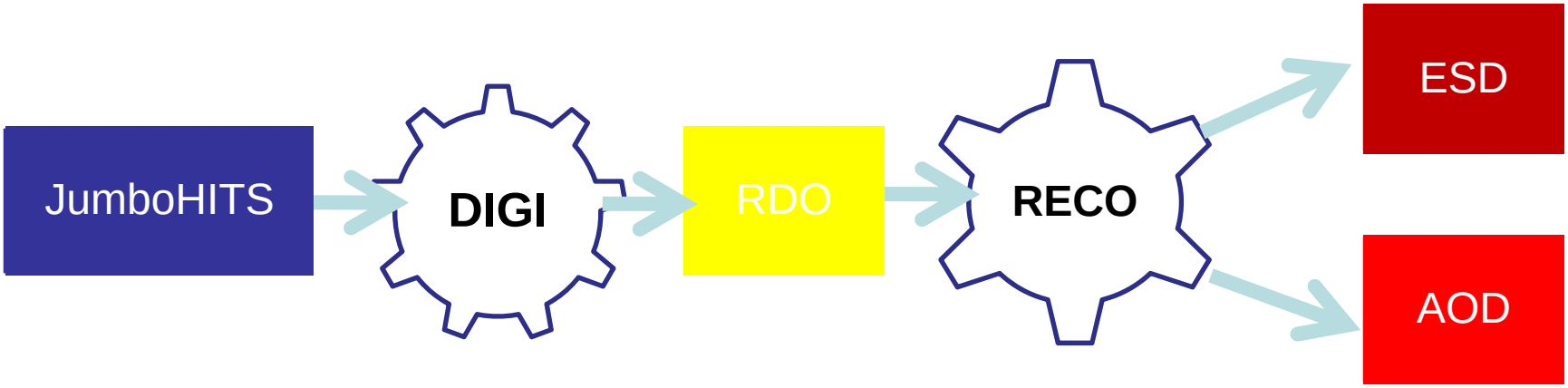


Merge 50 HITs files into one JumboHITS file
JumboHITS file order 5 GByte
and contains 2500 events
Simulation and merging run at the Tier-2's
JumboHITS file uploaded to Tier-1
Reconstruction runs in Tier-1's

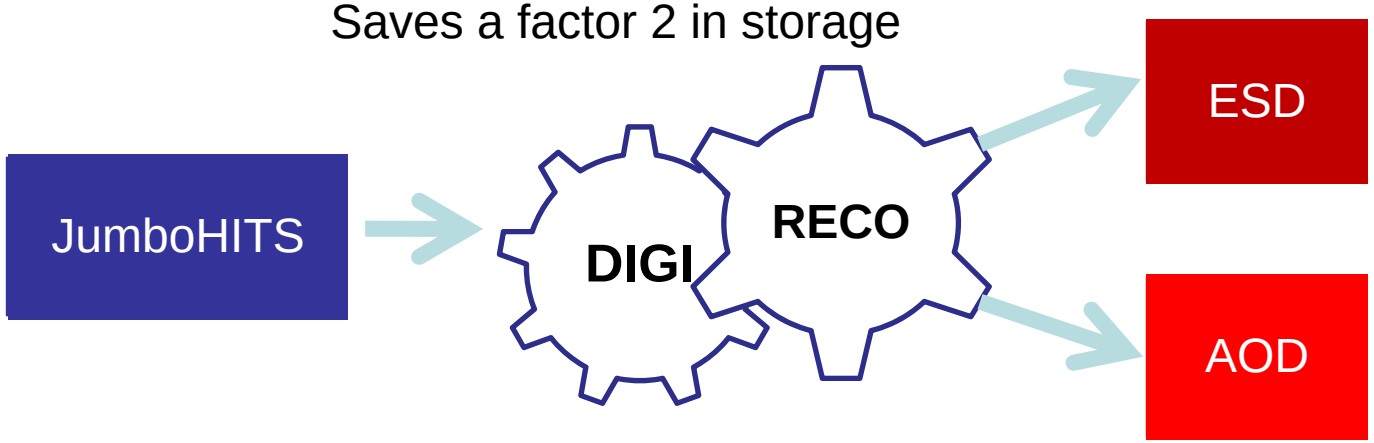
JumboHITS

TRF doesn't exist yet

Digi + Reco



DIGI step is very fast compared to RECO
Currently RDO files stored (same size as HITS)
Easier to re-create them from HITS
Saves a factor 2 in storage



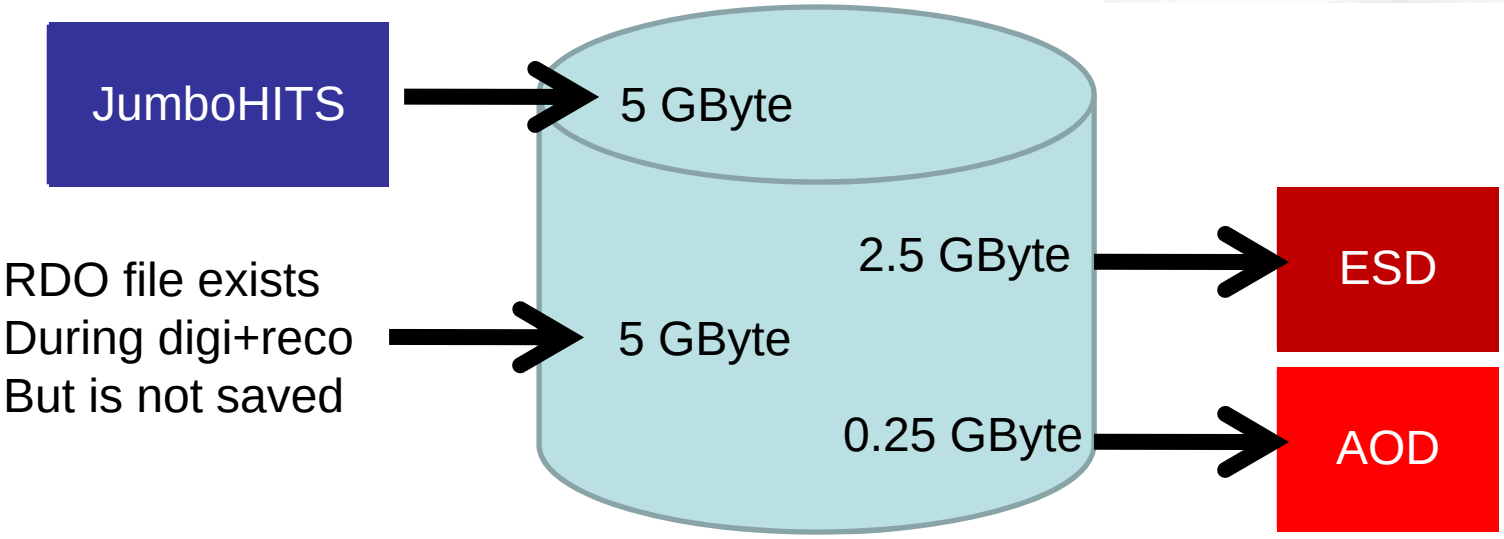
ESD = 1 MB/event
2500 events
File size 2.5 GB

AOD = 0.1 MB/event
2500 events
File size 250 MByte

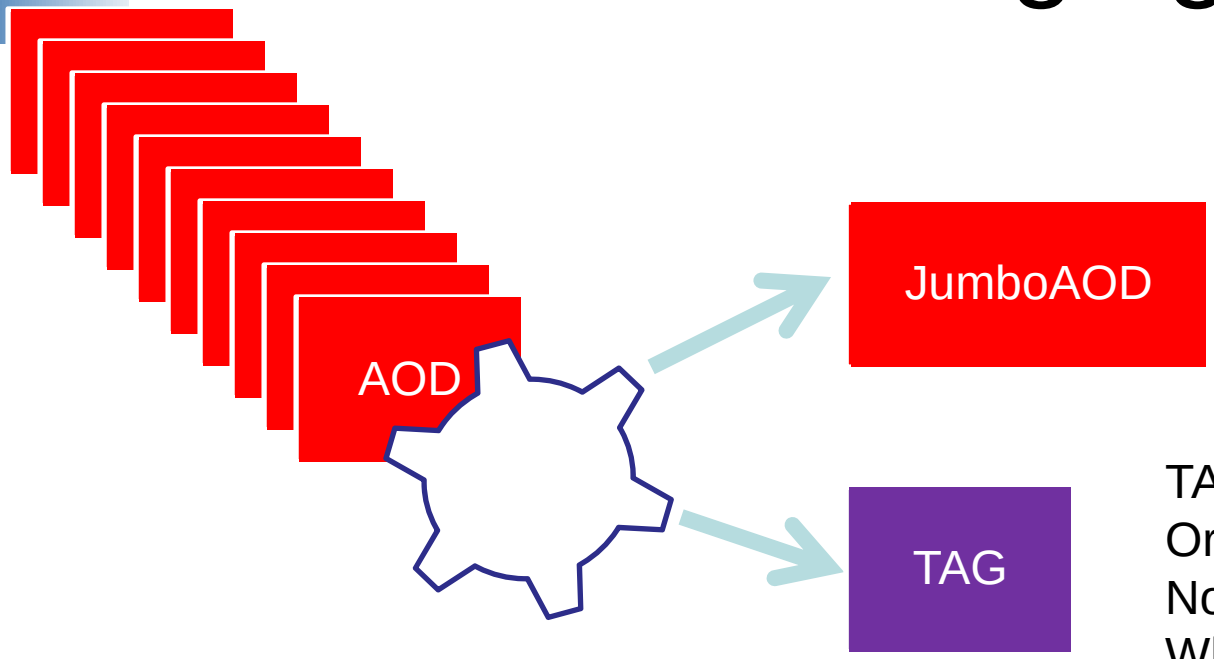
Works ! But still testing

Local disk limitation

Local disk is limited
Boards now come with 4,8, .. cpu's
Cpu's have multiple cores
But disk size per board has not increased
GDB: count on not more than 15 Gbyte/core



AOD merging



10 AOD files input for TAG creation
In same step JumboAOD could be created
Filesisze JumboAOD 2.5 Gbyte
And contains 25000 events
TRF does exist but is broken at this moment
Back navigation and lumi blocks are issues

TAG files are also small
Order 100 Mbyte
Not merged until decided
What to do with TAGs
Could be tar-ed before transpor

Other small files

- DPDs
  - Same format as AOD, same trf to merge
- CBNTAA, SAN, HighPT not in rlse 13
- Logfiles
  - Dedicated ftp (3) servers
  - Tarred after use and stored with the data
- User files
  - Worry!! Don’t know what to do yet
- And last but not least RAW data

RAW and Deirved Data at the T0

Inclusive Streaming ("86400s/day")

Phys. Stream	Jet	Electron	Muon	Tau	Photon
Events/LB	3600	3000	2400	2400	600
RAW Files/LB	5	5	5	5	5
RAW File Size	1152 MB	960 MB	768 MB	768 MB	192 MB
RAW Files/Day	36000				
Merged RAW Files/LB	1	1	1	1	1
Merged RAW File Size	5760 MB	4800 MB	3840 MB	3840 MB	960 MB
Merged RAW Files/Day	7200				
ESD Files/LB	1	1	1	1	1
ESD File Size	3600 MB	3000 MB	2400 MB	2400 MB	600 MB
ESD Files/Day	7200				
AOD Files/LB	3	2	2	2	1
AOD File Size	120 MB	150 MB	120 MB	120 MB	60 MB
AOD Files/Day	14400				
Merged AOD Files/Run	3	2	2	2	1
Merged AOD File Size	3600 MB	4500 MB	3600 MB	3600 MB	1800 MB
Merged AOD Files/Day	480				
Reco Job Length [kSI2k]	15 h	12.5 h	10 h	10 h	2.5 h

RAW File Merging

- Problem: combination of (1min LBs, O(5) physics streams, O(5) SFOs) results in relatively small RAW files
  - About 800MB on average
  - Mass storage (i.e. tape) systems and data export (DDM) prefer large files
- In past meetings we have already suggested and discussed several RAW file handling and merging scenarios
- Systematic, comprehensive tests and measurements (also by Tier-1s) were planned, but have not taken place so far
  - Schedule conflicts with M* weeks and throughput/functional tests, etc.
- There is evidence from past tests that CERN/CASTOR (Tier-0 setup) and DDM Tier-0 → Tier-1 export are able to cope with small RAW files
- Original plan was to dedicate FDR to deciding on RAW file handling scenario
  - Use small, unmerged RAW files in FDR-1
- Last year's Tier-1 Jamboree: unanimous request of all Tier-1s to go for RAW file merging straight away

Possible RAW File Merging

- **Issues (difficult to reconcile…)**
  - CM requirement: archival of RAW data on tape a.s.a.p. after arrival at CASTOR
  - Merging adds at least another 320 MB/s Tier-0 internal writing load
  - Asymmetry of files at CERN and at Tier-1s
    - Extra book-keeping (mapping of small ↔ merged files)
  - "Real" merging processing (on BS level) requires
    - Appropriate software; CPU
    - Careful validation of the merged file
  - Can original, small files eventually be discarded?
- **Suggestion: "Minimal asymmetric" scenario**
  - Archive small RAW files on tape a.s.a.p. after arrival at CASTOR
  - Register small RAW files with DQ2 (location: CERN)
  - Do tar'ring of RAW files in sequence with the reconstruction job
    - Adds "minimal" 320 MB/s writing load
  - Put merged RAW files on a temporary CASTOR disk buffer
  - Create merged RAW datasets, register with DQ2
  - Export merged RAW datasets to Tier-1s
    - NB: Inevitable latency of 24h-48h
  - After successful export: delete CERN copies from CASTOR and DQ2 catalogues
  - Will be during CCRC-1