Argonne NATIONAL LABORATORY

# Duplicated events in digi+reco

A. Vaniachine

Event Index Workshop

June 22-23, 2015

IFIC, Valencia, Spain

U.S. DEPARTMENT OF ENERGY

# Introduction

- Recently AthenaMP began to be used to keep memory usage within practical limits
- Previously, MC digi+reco jobs were executed in three steps within one transform:
  1. HITS to RDO
  2. RDO to ESD
  3. ESD to AOD
- Unexpected memory increases in MC15 were alleviated by adding an extra step:
  1. HITS to RDO
  2. RDO to RDO_TRIG
  3. RDO_TRIG to ESD
  4. ESD to AOD
- Fortunately, validation of the new RDO to RDO_TRIG step omitted routine events count match between input and output files
  - This exposed rare problem when two AthenaMP jobs start at the same time on the same many-core node, which may result in a clash in AthenaMP shared queue names
    - In this case, out of all input events only about 75% are processed, with about a third of those are processed twice (these are called duplicated events)

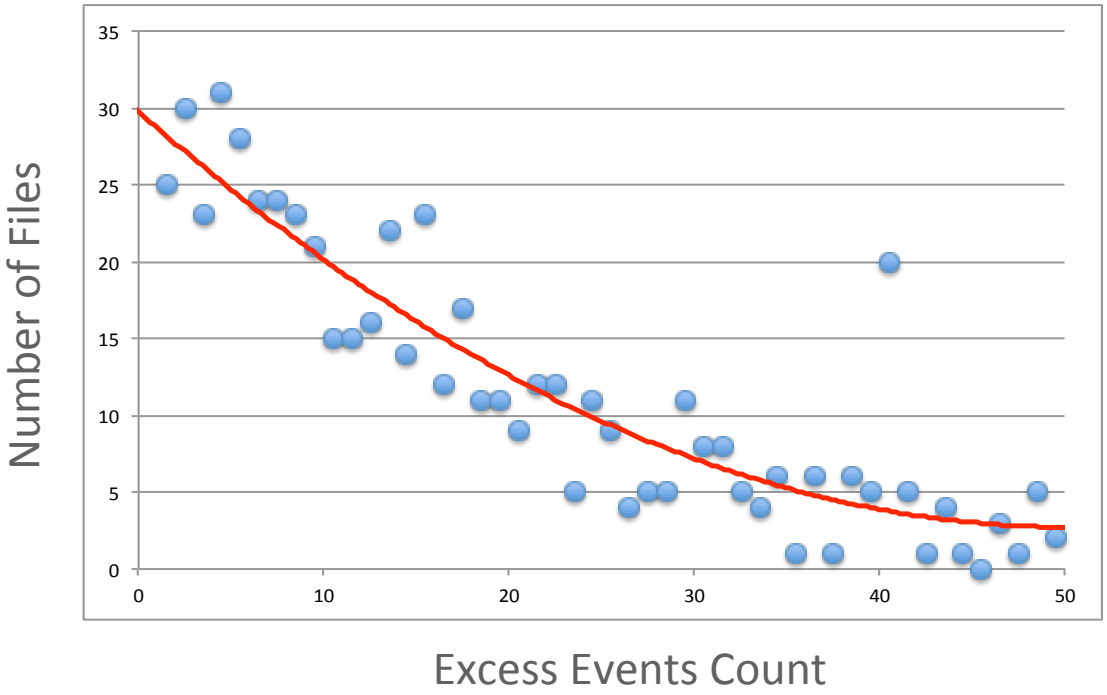
# Duplicated events in digi+reco

- The clash in shared queue names often results in a (matching) mismatch in events counts between the input and output files, e.g.:
  - File AOD.05371597._001522.pool.root.1 has 984 events
    - Deficit of 16 events vs. the expected 1000 events
  - File AOD.05371597._001523.pool.root.1 has 1016 events
    - Matching excess of 16 events vs. the expected 1000 events
- The mismatch helps to identify files with duplicated events


- It is possible that the clash in shared queue names result in matching events count
  - How many files are expected to have duplicated events, but matching events count?

# Files with Matching Events Count

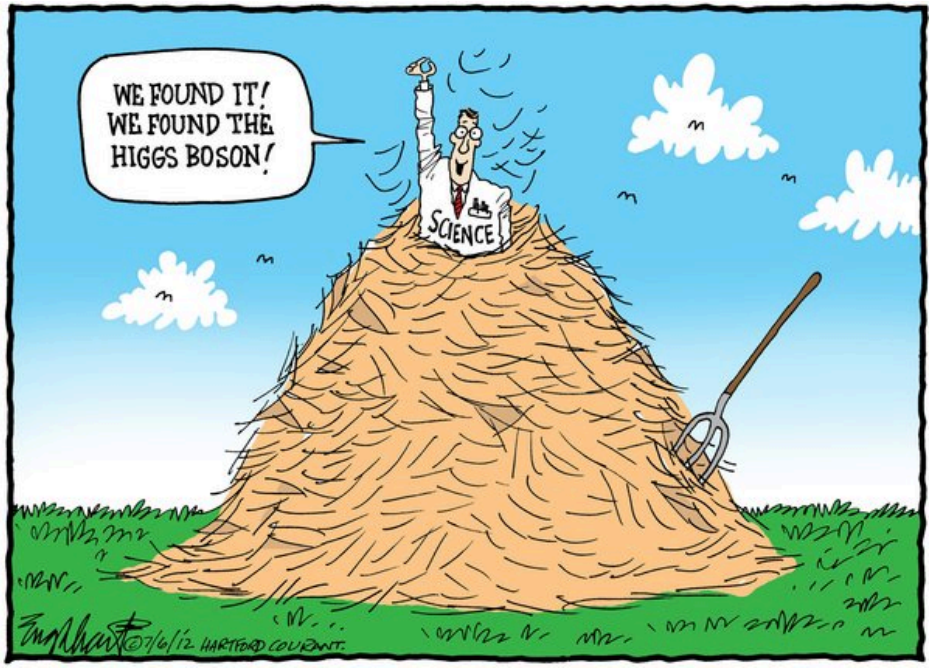
- Extrapolation to zero of the number of files vs. excess events count provides an estimate of 2 x 30 = 60 files for the RDO to RDO_TRIG step only




- The digi+reco job has four steps total, plus about five times more G4 jobs per file:
  - estimated number of the affected files with matching events count is 9 x 60 = 540

# A Needle in a Haystack

- Event Index is the only practical way to confirm the estimate of 0.00026 files with duplicated events but matching events count in more than two million AOD files
  - Preferably, we would remove files with duplicated events from mc15_13TeV samples
    - What is an estimated timescale to identify these files in all mc15_13TeV samples?




http://www.courant.com/opinion/cartoons/hc-higgs-boson-20120705-story.html