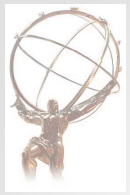
ATLAS Data Formats and Impact on Federated Access
(de facto title - changes to ATLAS analysis model)

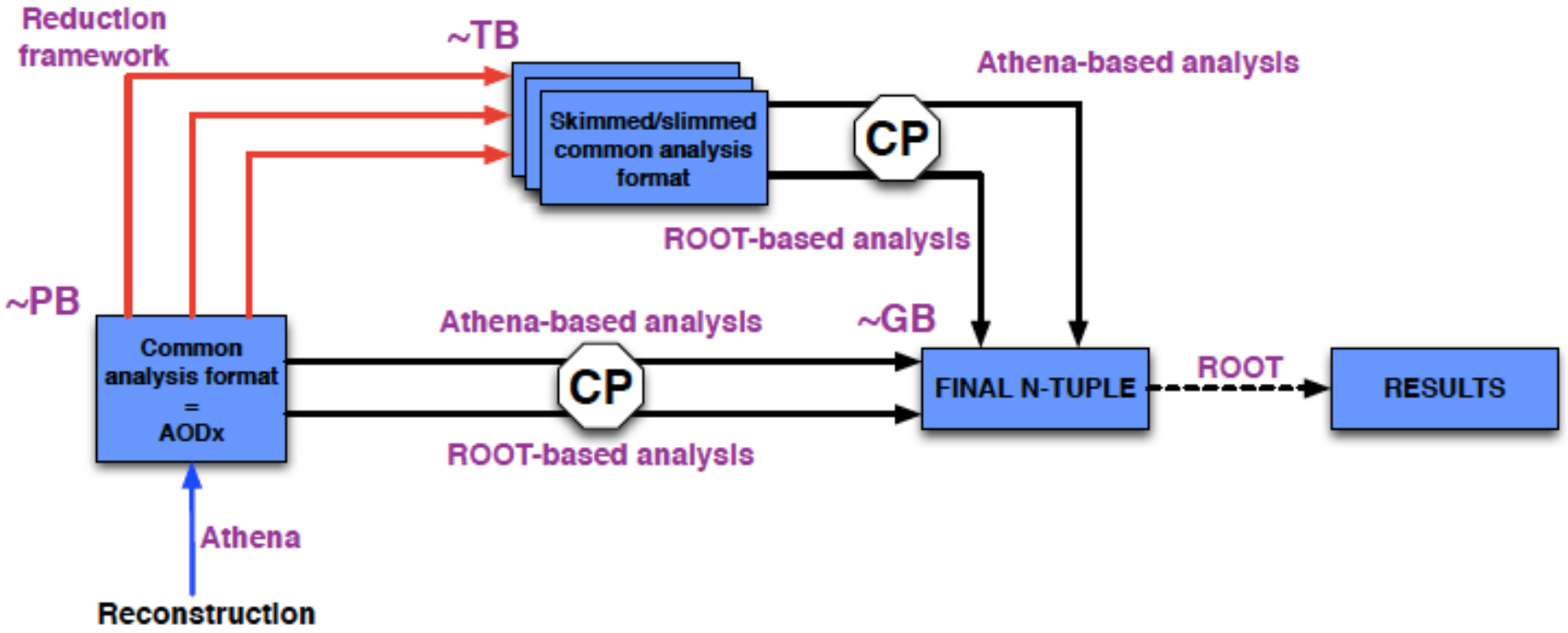
Doug Benjamin
Duke University
Presenting results on behalf of the ATLAS collaboration
especially
Thomas Maier, Johannes Elmsheuser, Günter Duckeck
Ludwig-Maximilians-Universität München



ATLAS analysis model transition (or AOD into xAOD)

- At the end of Run 1 ATLAS commissioned the ATLAS Model Study Group to better prepare itself for Run 2
- Report released at the start of Long Shutdown 1 (LS1)
  - Acknowledged how physics actually interacted with data and made proposals for what needed to change
  - Physics groups using central resources were producing many large flat ntuples to be analyzed exclusively in ROOT
    - Duplicate data – some of the same information in many different but very similar data formats
    - AOD (Analysis Object Data – produced by Event reconstruction in Athena) were not that popular and converted to DPD's (Derived Physics Data) as large flat ntuples (D3PD's).
  - Software Tools for analysis had to be written for two frameworks (ATLAS software framework - Athena and stand alone ROOT)
  - DPD production was not scalable for Run 2 – too resource intensive – new model was needed.

New and improved!!!
Run 2 ATLAS analysis model



Requires a new data product – xAOD and will produce smaller more refined data products with a Central system based on the same format (DxAOD)
- Both useable in Athena-based analysis or a ROOT-based analysis

Introducing a new data format xAOD

- Redesign the event data model data product – AOD to have the efficiencies found in the large flat ntuples
- Quoting the Task Force that redesigned the new EDM

"In principle all ROOT readable classes in the new EDM should look like:

- A "core class" only holding very minimal information, but providing the full interface of the object in question: This would be a class like Electron or Muon. It shall only hold the most important information about the objects itself. For particles this would practically be just their "best" four momentum, and all the (element) links to objects related to them.
- One (or sometimes no) associated auxiliary object: Each core container object links to one (or no) auxiliary object. These will function much like how egDetails and JetMoments was used so far. In fact, most of the information about the objects would be kept in such auxiliary objects, and not in the core objects themselves. All the accessor functions defined in the core classes would interact with the auxiliary store to save and retrieve data."

- Core objects have large flat interface
- To access the objects/containers in a simple ROOT environment, the user will need to go through a thin code layer.
  - Implies User's must use dedicated access libraries in ROOT. No more TTree::SetBranchAddress calls. Must access through event Class.

Implications of an Event interface

- During Run 1 the vast majority user analysis jobs on the grid were ROOT based analysis
  - Analysis of activity from storage systems (but not the jobs themselves) revealed that the users read a small fraction of the files (<< 10%)
  - Had little or no monitoring of what the job was accessing
- Common Run 2 Event interface allows for improvements
  - All user analysis jobs must use the interface. It has optimizations as to not incur too much over head
  - Ability to monitor what is being read from the file
  - Implements TTreeCache automatically – important for WAN data access
  - Standard and consistent names in all derived data produces – avoid needless data duplication

Derivations

- Intermediate data formats **(derivations)**
- made centrally
- the purpose of the derivation framework is to provide the offline software tools and structures for doing this in a transparent way.
- Derivations will be made from XAOD input and will have the same general format as XAOD but containing less data (DXAOD).
- Derivations will be made from the full data via four operations:
  - skimming: removing whole events
  - thinning: removing whole objects from within an event, but keeping the rest of the event
  - slimming: removing information from within objects, but keeping the rest of the object
  - augmentation: adding data not found in the input data
- Currently – Physics group – 33 Derivations and Performance groups – 19 Derivations

Some xAOD I/O measurements using local tests and HammerCloud

Thomas Maier, Johannes Elmsheuser, Günter Duckeck

Ludwig-Maximilians-Universität München

22 January 2015/ATLAS Distributed ROOT I/O working group

Contents

- Local tests to LRZ-LMU_LOCALGROUPDISK
- Comparision of nfs, dcap, xrootd/FAX, webdav/davix at LRZ-LMU
- 3 different analysis: slow, medium, fast xAOD MC
- Throughput plots of local network card
- HammerCloud tests of slow and fast analysis using different protocols

xAOD read speed tests I

Comparison of different protocols:

- Use "real-life" tutorial analysis:
  `https://twiki.cern.ch/twiki/bin/view/AtlasComputing/SoftwareTutorialxAODAnalysisInROOT`
- EventLoop based xAOD analysis, reads/processes muons, jets, stores some output
- Discussed ATLAS distributed ROOT I/O meetings

xAOD read speed tests II

- Using DC14 mc14_13TeV r5787_r5853 xAOD input sample:
  mc14_13TeV:mc14_13TeV.110401.PowhegPythia_P2012_ttbar_nonallhad.merge.AOD.e2928_s1982_s2008_r5787_r5853
  which is replicated to 6 sites
- Access mode: default Panda (dcap, copy-to-scratch, xrootd), xrootd, davix access to local SE
- 3 analysis:
  - AnalysisRelease, Base,2.0.22 with ROOT 5.34.24-x86_64-slc6-gcc48-opt
  - (1a) Rather intensive analysis of muons, jets with systemtatics and heavy output writing (also used in HammerCloud tests)
  - (1b) Intensive analysis without systematics
  - (2) Rather fast analysis of only muons and electrons (also in HC)
- Access with `optXaodAccessMode_class` or `optXaodAccessMode_branch`
- `TTreeCache` is enabled with 10 events/10 MB learning in code
- Using in addition: `ROOT_TTREECACHE_PREFILL=1` `ROOT_TTREECACHE_SIZE=1`

Local access to LRZ-LMU I

Reading on local Munich desktop from LRZ-LMU_LOCALGROUPDISK

Analysis (1a), class access mode

- local/nfs: 80 events/s
- dcap: 75 events/s
- FAX: 69 events/s
- Davix: 72 events/s (TTreeCache.PREFILL=1)

Analysis (1a), branch access mode

- local/nfs: 89 events/s
- dcap: 94 events/s
- FAX: 86 events/s (TTreeCache.PREFILL=1)
- Davix: 84 events/s (TTreeCache.PREFILL=1)

Important:

- Davix suffers from missing buffering in start-up - PREFILL=1 cures the problem

Local access to LRZ-LMU II

Reading on local Munich desktop from LRZ-LMU_LOCALGROUPDISK

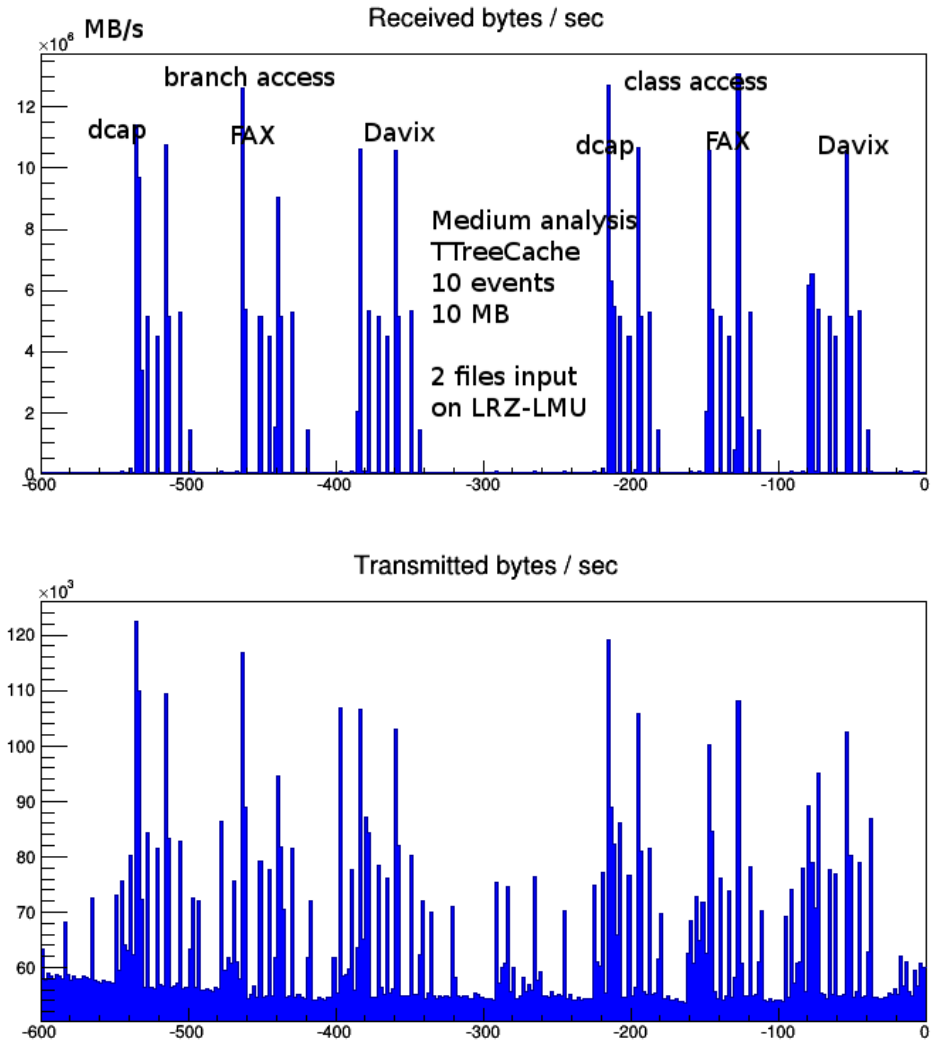
Analysis (1b), class access mode

- local/nfs: 210 events/s
- dcap: 196 events/s
- FAX: 193 events/s
- Davix: 192 events/s (TTreeCache.PREFILL=1)

Analysis (1b), branch access mode

- local/nfs: 230 events/s
- dcap: 228 events/s
- FAX: 211 events/s (TTreeCache.PREFILL=1)
- Davix: 205 events/s (TTreeCache.PREFILL=1)

Local access to LRZ-LMU III - eth0 rate



10-12 MB/s troughput, no large difference between branch/class access, reading spike for every file open, not much protocol difference

Local access to LRZ-LMU IV

Reading on local Munich desktop from LRZ-LMU_LOCALGROUPDISK
Fast Analysis (2), branch access mode:

- local/nfs: 550 events/s (w/o 8-11s init: 1050 events/s)
- dcap: 495 events/s (w/o 8-11s init: 850 events/s)
- FAX: 497 events/s (w/o 8-11s init: 830 events/s)
- Davix: 483 events/s (w/o 8-11s init: 815 events/s)

class access mode:

- local/nfs: 195 events/s (w/o 8-11s init: 200 events/s)
- dcap: 147 events/s (w/o init: 155 events/s)
- FAX: 182 events/s (w/o init: 194 events/s)
- Davix: 190 events/s (w/o init: 204 events/s)

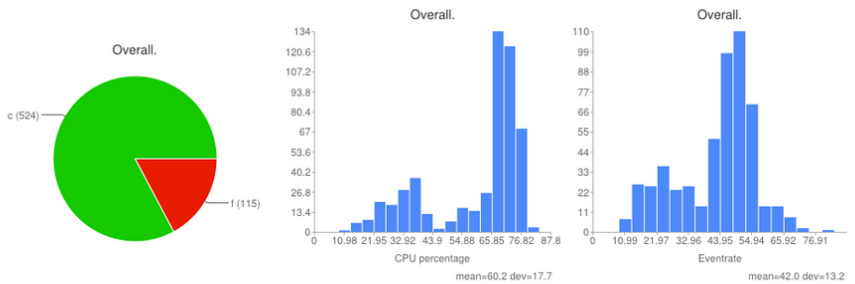
Short/fast analysis with non neglible fraction of init time
large fluctutations of $\pm 50 - 80$ events/s in several repetition
Tried to use xAOD::ReadStats - wrong measurements results ?!

HammerCloud - default Panda setup I

http://hammercloud.cern.ch/hc/app/atlas/test/20051843/

state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20051843	it-hammercloud-submit-atlas-08	US, DE_PANDA, FR_PANDA, 1 more...	15/1/2015 11:55	15/1/2015 14:44	904

Input type: PANDA
Output DS: user.gangarbt.hc20051843.*
Input DS Patterns:
mc14_13TeV:mc14_13TeV.110401.PowhegPythia_P2012_ttbar_nonallhad.merge.AOD.e2928_s1982_s2008_r5787_r5853*
Ganga Job Template: AnalysisReleaseBase/template_stress_v2022.tpl
Athena User Area: AnalysisReleaseBase/jobcontents_v2022.tgz
Athena Option file: AnalysisReleaseBase/runjob.sh
Template: Stress AnalysisReleaseBase 2.0.22 eventLoop Tutorial, default setup
View Test Directory (for debugging)



more plots »

Sites

Show 10 entries Search:

Site	S	R	C	F	Eff	T	Datasets	Queue	Max R	Resubmit	R. Force	Link
ANALY_LRZ	0	0	0	0	0.00	0	1	0	1	yes	no	»
ANALY_IN2P3-CC	42	39	382	112	0.77	576	1	0	1	yes	no	»
ANALY_FZK	0	5	17	1	0.94	31	1	0	1	yes	no	»
ANALY_ECDF_SL6	167	0	1	2	0.33	170	1	0	1	yes	no	»
ANALY_CPPM	0	0	0	0	0.00	0	1	0	1	yes	no	»
ANALY_BNL_SHORT	0	3	124	0	1.00	127	1	0	1	yes	no	»
Site	S	R	C	F	Eff	T	Datasets bulk	Min queue	Max running	Resubmit	R. Force	Link

Showing 1 to 6 of 6 entries

First Previous 1 Next Last

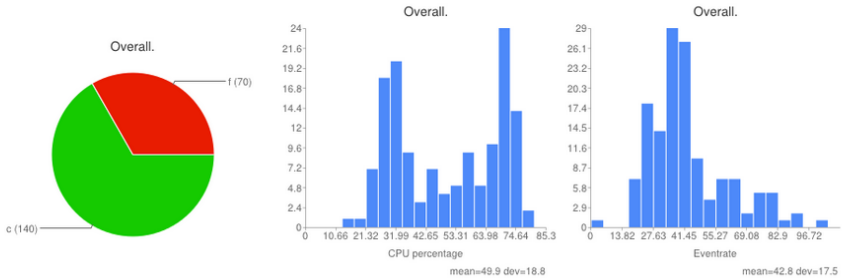
Analysis (1a),LRZ was missing the dataset at the time of the test

HammerCloud - FAX local replica I

http://hammercloud.cern.ch/hc/app/atlas/test/20051868/

state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20051868	it-hammercloud-submit-atlas-08	US, DE_PANDA, FR_PANDA, 1 more...	16/1/2015 9:35	16/1/2015 11:35	381

Input type: PANDA
Output DS: user.gangarbt.hc20051868.*
Input DS Patterns:
mc14_13TeV:mc14_13TeV.110401.PowhegPythia_P2012_ttbar_nonallhad.merge.AOD.e2928_s1982_s2008_r5787_r5853*
Ganga Job Template: AnalysisReleaseBase/template_stress_v2022_fax.tpl
Athena User Area: AnalysisReleaseBase/jobcontents_v2022.tgz
Athena Option file: AnalysisReleaseBase/runjob.sh
Template: FAX direct Stress AnalysisReleaseBase Base 2.0.22 eventLoop Tutorial, FAX setup
View Test Directory (for debugging)



more plots »

Sites

Show 10 entries Search:

Site	S	R	C	F	Eff	T	Datasets	Queue	Max R	Resubmit	R. Force	Link
ANALY_BNL_SHORT	17	19	36	7	0.84	85	1	0	1	yes	no	»
ANALY_CPPM	0	0	0	0	0.00	0	1	0	1	yes	no	»
ANALY_ECDF_SL6	0	26	74	10	0.88	170	1	0	1	yes	no	»
ANALY_FZK	2	9	18	0	1.00	31	1	0	1	yes	no	»
ANALY_IN2P3-CC	0	23	11	50	0.18	84	1	0	1	yes	no	»
ANALY_LRZ	0	7	1	3	0.25	11	1	0	1	yes	no	»
Site	S	R	C	F	Eff	T	Datasets bulk	Min queue	Max running	Resubmit	R. Force	Link

Analysis (1a)

HammerCloud - Fast Analysis

Analysis (2), Example event rates:

- LRZ-LMU (dcap): 589 Hz
- ECDF (xrootd): 576 Hz
- IN2P3-CC (xrootd): 433.28 Hz
- FZK (dcap): 551 Hz
- BNL (copy-to-scratch): 834.75 Hz (without copy time!)

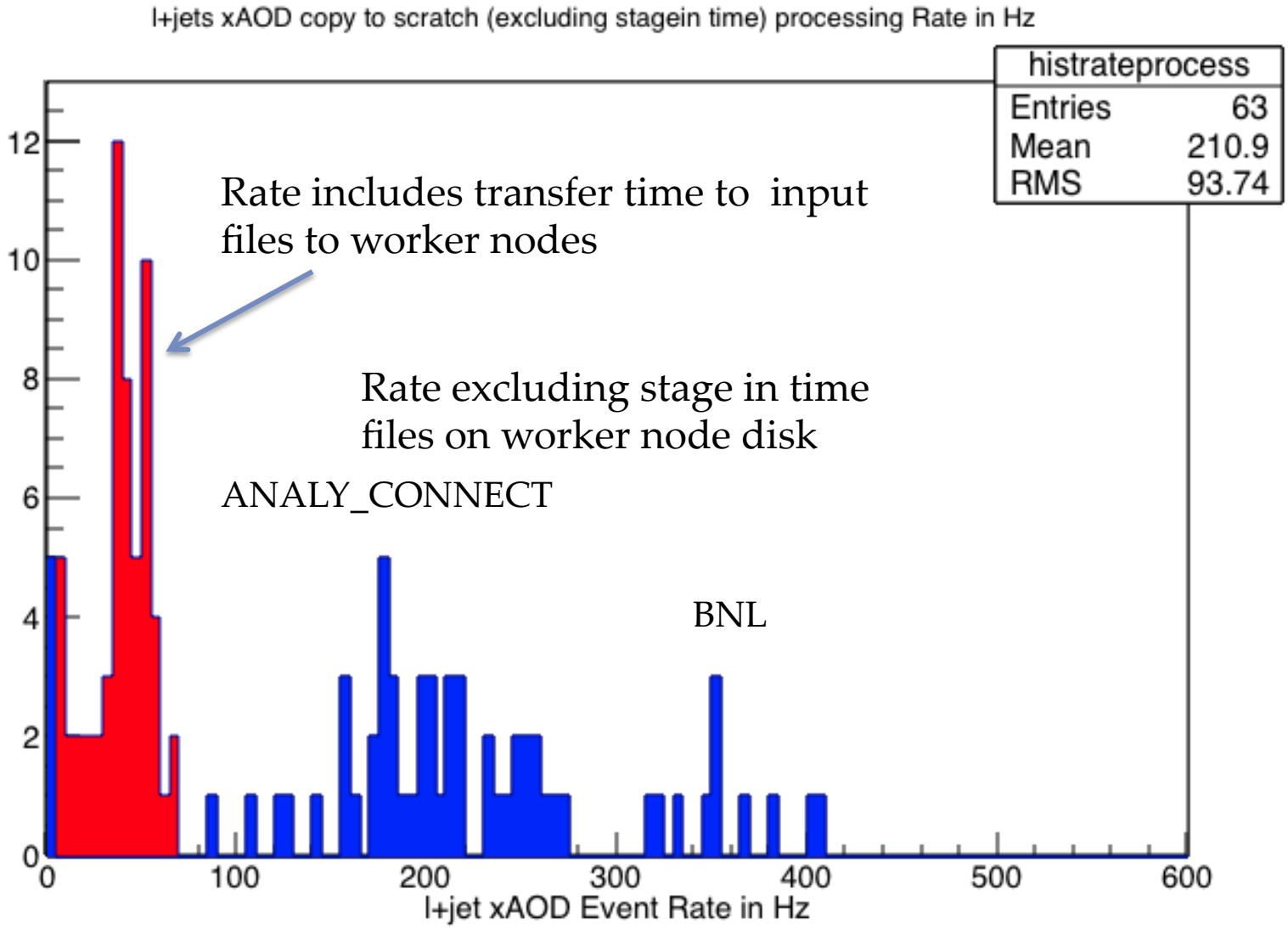
Summary and Conclusions

- Different setups to test default/dcap/FAX/webdav access locally and Panda
- All protocols show similar performance in local and Panda test within the uncertainties
- Event rate heavily dependent on analysis type: 80-1000 event/s
- Large difference in branch vs. class access mode for fast analysis
- Local performance of specific analysis comparable to Panda job results, but some performance loss due to busy sites
- Panda errors due to local site problems
- xAOD::ReadStats with no reliable results

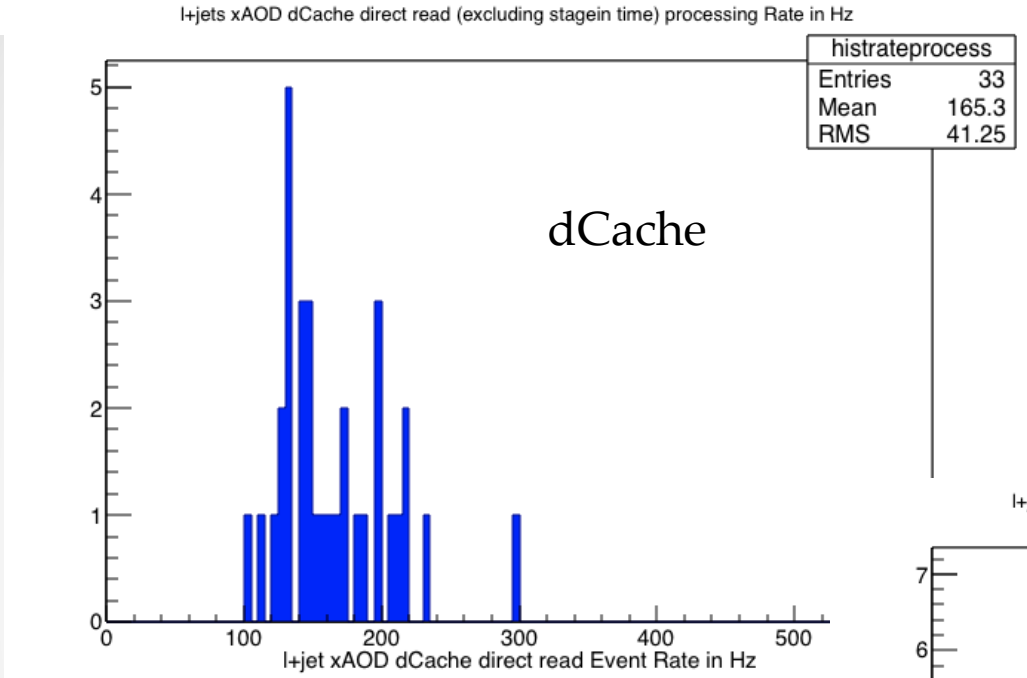
Code and datasets

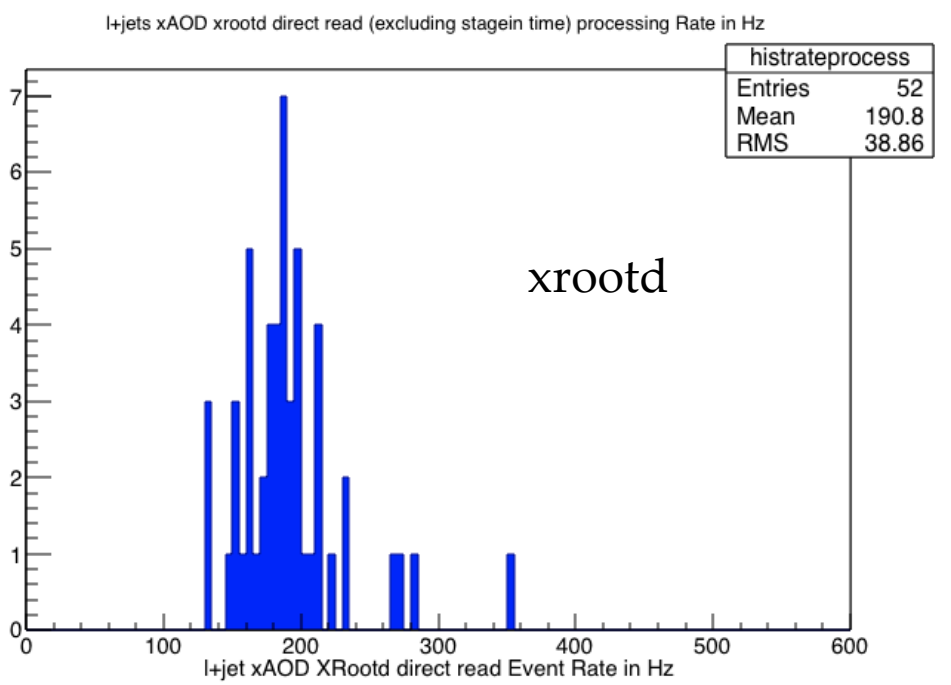
- Top Physics group xAOD/DxAOD analysis code
  - Compiled C++ Event loop
  - Branch access
  - using common xAOD analysis libraries
  - Contains full suite of correction and systematic effects
  - Produces flat ntuple output
- Full xAOD – Top ttbar Monte Carlo file 6272 variables
- Recent Top lepton + jets derivation (smaller DxAOD) – 2567 variables
  - Deriviation used the full xAOD Top ttbar Monte Carlo data files as input.
- Two analysis algorithms used
  - Lepton + jets search
  - Dilepton search

Event Rate lepton+jets xAOD



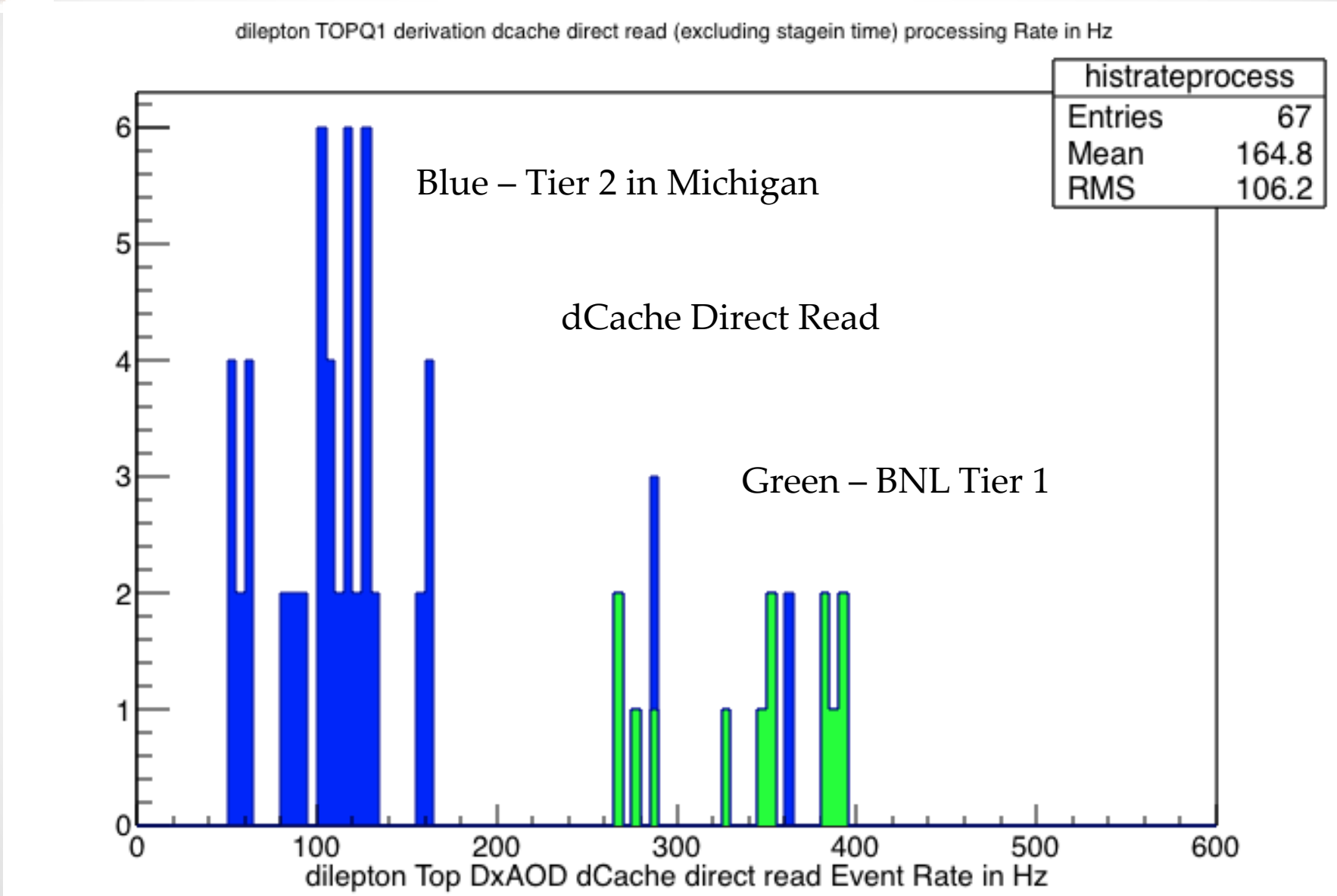
lepton+jets xAOD direct read



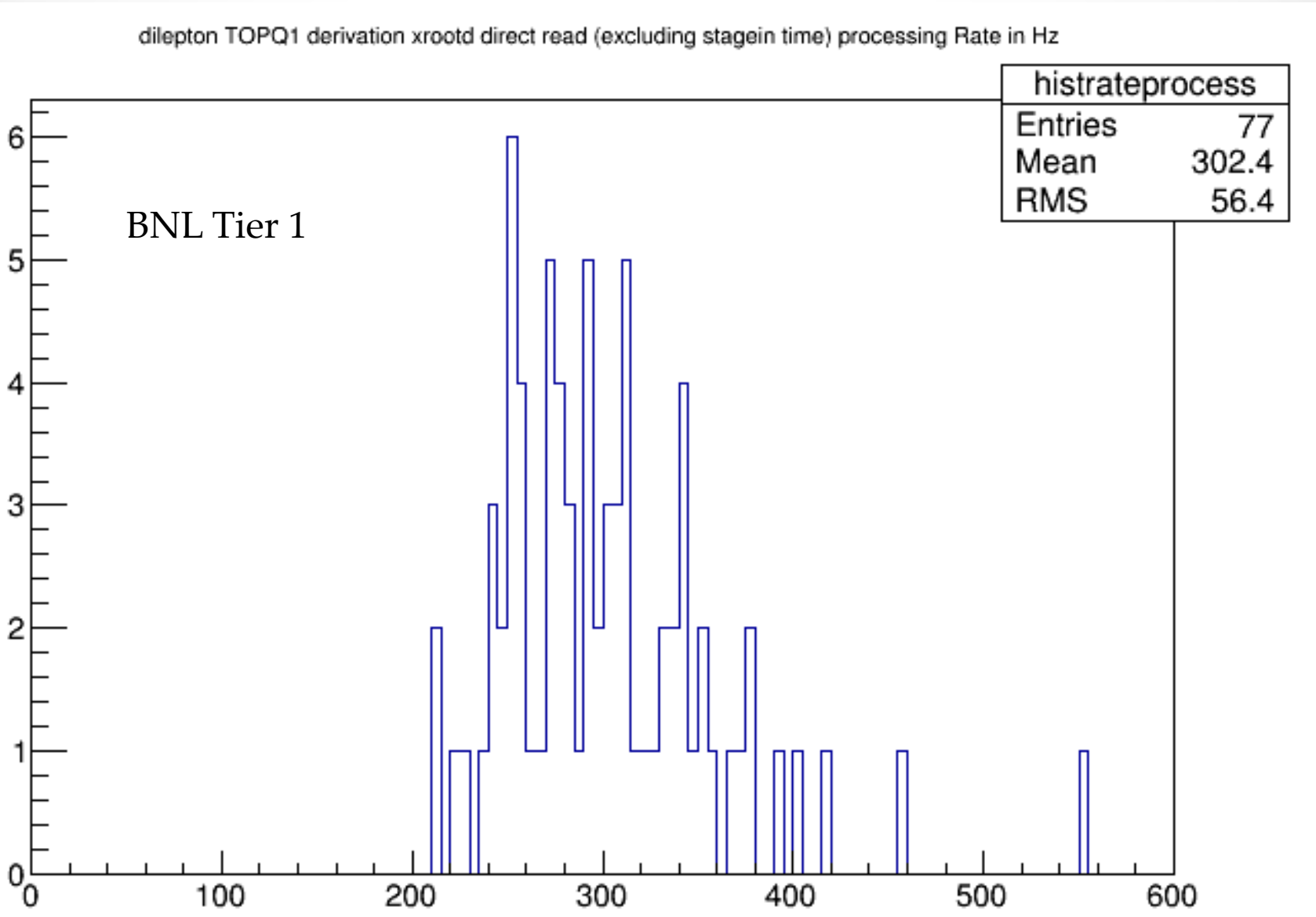


All jobs run at BNL

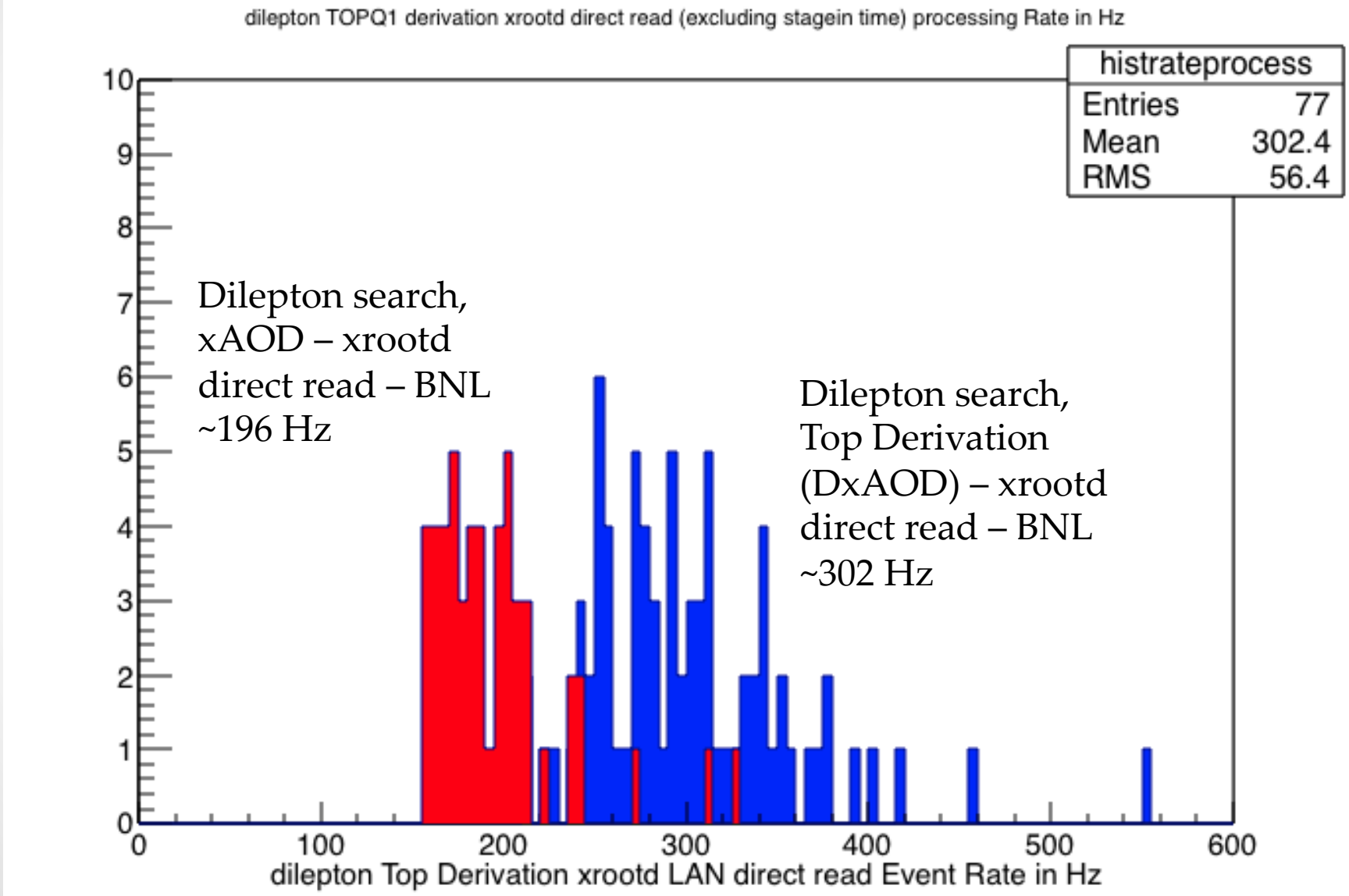
Event Rate (Hz) Dileptons – Top Derivation data



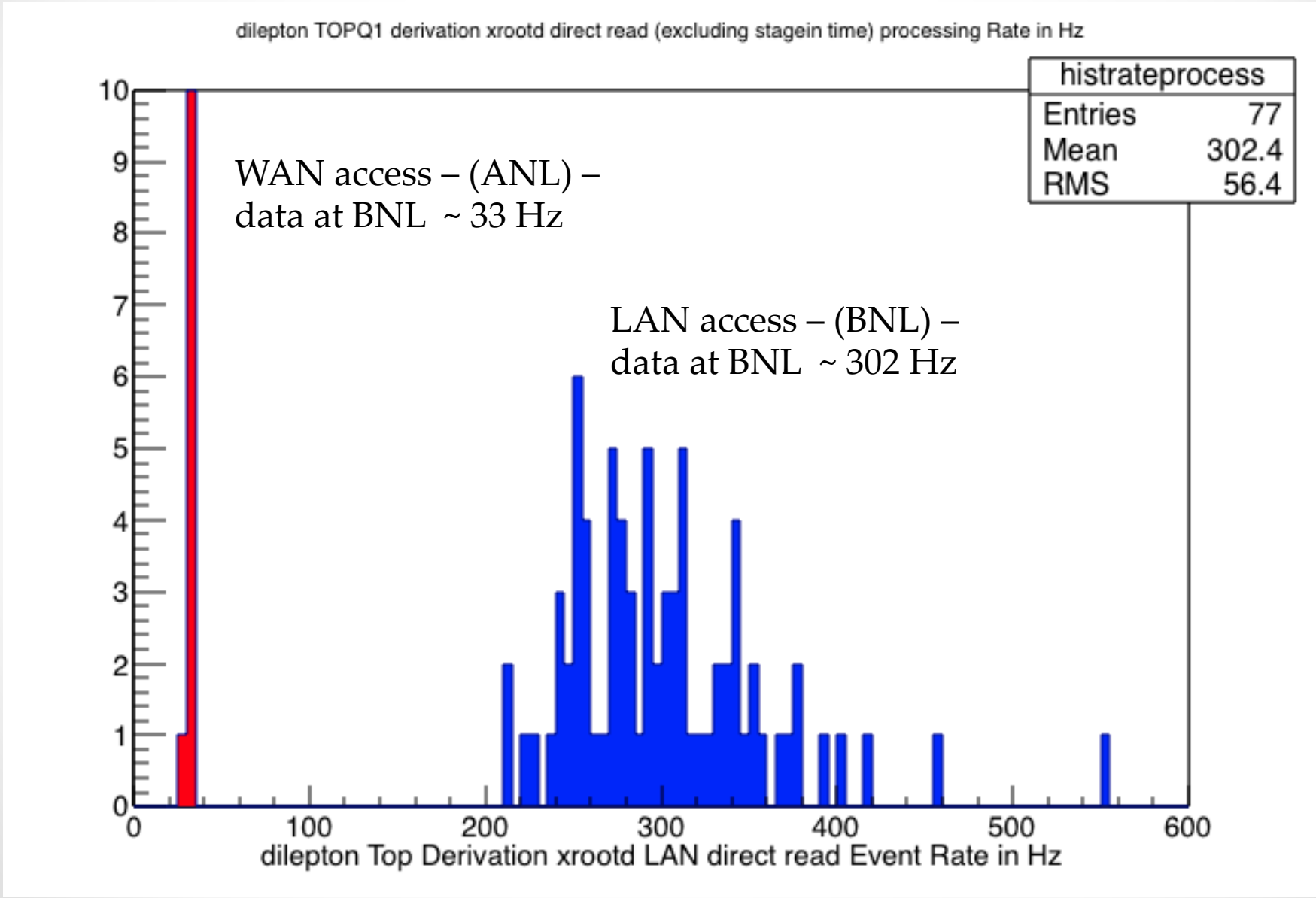
Event Rate (Hz) Dilepton – Top Derivation data – xrootd direct access



Event Rate Comparison of xAOD to DxAOD

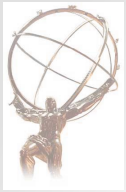


Comparison LAN vs WAN access



Next Steps

- ATLAS physics groups are actively working on producing new smaller Derived data products (DxAOD's) that are better tuned to their needs
  - Events are selected based on ID criteria (SKIMMED)
  - Variables are retained for the the individual analysis needs (SLIMMED)
  - The Slimming of variables is further enhanced (ie SMART SLIMMING) – by not retaining all variables in a given grouping but only some – This varies across physics groups .
- Repeat some of the measurements after SMART SLIMMING
- Repeat after migration to ROOT 6 for analysis – in March or so.



Conclusions

- Based on Run 1 experience ATLAS has revised its analysis model to better use computing resources.
- Through the revision of its Event Data Model (EDM), the ATLAS collaboration is better prepared for WAN data access.
- Further work remains.
- Using real world examples and some idealized test cases, the performance across the network has been measured.
- Some of the measurement infrastructure needs to be fixed to provide meaningful measurements.
- Dedicated resources seem to perform fine
- Opportunistic (Connect queue) and WAN measurements (BNL to ANL) are under performing and future work is really needed for Run 2.