



# A Distributed Analysis Demonstrator Using pathena Analysis Queues on Tier2 Facilities

**Nurcan Ozturk**

**University of Texas at Arlington**

**US ATLAS Transparent Distributed Facility Workshop**

**University of North Carolina - March 4, 2008**

# Introduction



- **Goal:**
  - Send an analysis job to the analysis queues at Tier2's using pathena as a distributed analysis tool
  - Run on the FDR data for this demo
  - Retrieve and analyze the output
- **How to submit an analysis job:**
  - Setup athena
  - Check out PandaTools package (for pathena)
  - Use HighPtView package as an analysis package
  - Find the FDR data
  - Find out which analysis queue will be used
  - Submit a pathena job
- **Monitor job's status in PanDA monitor**
- **Get the output of pathena job and make plots**

# Setup Athena and Work Area



- Instructions are given to run on acas machines at BNL
- Create a directory (called pathenaDemo) and get the requirements file from:  
<http://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/AtlasSWReleases>
- Make a sub-directory for 13.0.40 (called 13.0.40) under pathenaDemo
- Setup CMT:
  - `source /afs/usatlas.bnl.gov/cernsw/contrib/CMT/v1r20p20070208/mgr/setup.sh`
  - `cmt config`
- Setup athena for release 13.0.40:
  - `source setup.sh -tag=13.0.40,32`
- Check out Tools/Scripts package to setup your work area (easy way of checking out and compiling multiple packages)
  - `cd 13.0.40`
  - `cmt co -r Scripts-00-01-14 Tools/Scripts`
- Setup work area and create run area:
  - `./Tools/Scripts/share/setupWorkArea.py`
  - `cd WorkArea/cmt`
  - `cmt bro cmt config`
  - `cmt bro gmake`
  - `source setup.sh`

# Check Out Necessary Packages



- Check out PandaTools for pathena:
  - `cd` to 13.0.40 directory
  - `cmt co PhysicsAnalysis/DistributedAnalysis/PandaTools`
- Run every time new package(s) checked out:
  - `./Tools/Scripts/share/setupWorkArea.py`

It prints:

```
WorkAreaMgr : INFO
#####
WorkAreaMgr : INFO   Creating a WorkArea CMT package under: [/usatlas/u/nurcan/pathenaDemo/13.0.40]
WorkAreaMgr : INFO   Scanning [/usatlas/u/nurcan/pathenaDemo/13.0.40]
WorkAreaMgr : INFO   Found 2 packages in WorkArea
WorkAreaMgr : INFO   => 0 package(s) in suppression list
WorkAreaMgr : INFO   Generation of WorkArea/cmt/requirements done [OK]
WorkAreaMgr : INFO
#####
```

- Compile PandaTools package from WorkArea:
  - `cd WorkArea/cmt`
  - `cmt bro cmt config`
  - `cmt bro gmake`
  - `source setup.sh`
- Use the HighPtView package from the release 13.0.40 and get the jobOption file into your run area:
  - `cd WorkArea/run`
  - `get_files HighPtViewNtuple_topOptions.py`

# Setup Grid and DQ2, Find FDR Datasets



- Setup Grid:
  - `source /afs/usatlas.bnl.gov/lcg/current/etc/profile.d/grid_env.sh`
- Setup DQ2:
  - `source /afs/usatlas.bnl.gov/Grid/Don-Quijote/dq2_user_client/setup.sh.BNL`
- Look at available FDR datasets at Tier2's from Panda monitor:
  - <http://gridui02.usatlas.bnl.gov:25880/server/pandamon/query?mode=listFDR>
  - Pick up one dataset:  
fdr08\_run1.0003050.StreamEgamma.merge.AOD.o1\_r6\_t1
- One can also list the replicas for a given dataset:
  - `source /afs/usatlas.bnl.gov/Grid/Don-Quijote/DQ2_0_3_client/dq2.sh`
  - `dq2-list-dataset-replicas fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1`  
INCOMPLETE:  
COMPLETE: IJST2,TIER0TAPE,TW-FTT,CYF,DESY-HH,DESYZN,PNPI,JINR,TORON,NAPOLI,  
LIP-LISBON,IFICDISK,LIV,RALPP,ICL,MWT2\_IU,WISC,SLACXRD,BU\_DDM,MCGILL,  
AGLT2\_SRM,SWT2\_CPB,BNLXRDHDD1,INFN-T1\_DATADISK,FZK-LCG2\_DATADISK,  
TRIUMF-LCG2\_DATADISK,NDGF-T1\_DATADISK,PIC\_DATADISK,IN2P3-LPC\_DATADISK,  
RAL-LCG2\_DATADISK, SARA-MATRIX\_DATADISK,TAIWAN-LCG2\_DATADISK,  
IN2P3-CC\_DATADISK, BNL-OSG2\_DATADISK

# Name Association Between DDM and Analysis Queue Names



DDM Name	Analysis Queue Name
SWT2_CPB	ANALY_SWT2_CPB
OU	ANALY_OU_OCHEP_SWT2
AGLT2_SRM	ANALY_AGLT2
MWT2_UC *	ANALY_MWT2
SLACXRD	ANALY_SLAC
BU_DDM	ANALY_NET2
WISC	ANALY_GLOW-ATLAS

\* MWT2\_UC and MWT2\_IU share the FDR data, however the analysis queue is setup to use the former.

# Analysis Queues from Panda Monitor



Panda Based Distributed Analysis Dashboard - Microsoft Internet Explorer

Address: <http://gridui05.usatlas.bnl.gov/server/pandamon/query?dash=analysis&reload=yes>

Navigation: [BNL monitor](#) | [Production](#) | [Clouds](#) | [DDM](#) | [PandaMover](#) | [AutoPilot](#) | [Sites](#) | [Analysis](#) | [Physics data](#) | [Usage](#) | [Plots](#) | [ProdDash](#) | [DDMDash](#)

Update Not logged

## Panda Based Distributed Analysis Dashboard

Information and tools for distributed analysis with Panda

[pathena](#) supports user submission to Panda from the (p)athena command line

**Analysis jobs:** [New-style listing of analysis jobs](#). Old-style listing is [here](#) if you prefer it ([tell me why!](#)). To look up a particular Panda job by ID use the quick left or click a PandaID in the job listing.

**Analysis users:** [User list](#) (also linked at top right, or above if you've logged in) shows analysis usage, ordered by most recent. From there you can go to your on the list if you've run a Panda job); if you 'log in' you'll get easier access to your page from a new menu at the top of the page.

**Groups:** [Groups](#) are supported to organize users by role, physics working groups etc. and support collaborative work, accounting rights etc. (Not much used)

**Data access:** See the [physics data](#) page linked above

**Analysis job summary, last 24 hours** ([Node details](#))

Analysis Site	Nodes	Jobs	Latest	Pilots (3hrs)	defined	assigned	waiting	activated	running	holding	transferring	finished	failed to
All	290	0	03-03 13:20	1023	<a href="#">2066</a>	<a href="#">0</a>	<a href="#">0</a>	<a href="#">5805</a>	<a href="#">181</a>	<a href="#">13</a>	<a href="#">0</a>	<a href="#">2375</a>	<a href="#">2485</a> 51%
<a href="#">ANALY_AGLT2</a>	2	6	03-03 13:20	50	<a href="#">2</a>	0	0	0	0	0	0	<a href="#">3</a>	<a href="#">1</a> 25%
<a href="#">ANALY_ALBERTA</a>	1	2	02-28 12:47	0	<a href="#">1</a>	0	0	<a href="#">1</a>	0	0	0	0	0

Task search: Job  Dataset  Task  File

Summaries: Blocks:  days Errors:  days Nodes:  days [Daily usage](#)

Tasks - [search](#)  
[Generic Task Reg](#)  
[EvGen Task Reg](#)  
[CTBsim Task Reg](#)

Windows: start | Panda Based Di... | The University o... | DQ2-DDM Oper... | The University o... | hepmail.uta.edu... | nurcan@atlasg... | Microsoft Power... | 12:20 PM

# Run pathena (1)



- Run pathena with one line command:

```
$ pathena -c "Mode=['FullReco'];DetailLevel=['FullStandardAOD'];  
Branches= ['StacoTauRec']" HighPtViewNtuple_topOptions.py  
--inDS fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1  
--outDS user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3  
--nfiles 1 --site ANALY_SWT2_CPB
```

- HighPtView options:
  - `Mode=['FullReco'];DetailLevel=['FullStandardAOD']; Branches= ['StacoTauRec']"`
- pathena options:
  - Specify input dataset by `--inDS`
  - Specify output dataset by `--outDS`
  - Specify # of files to be run on by `--nfiles 1`
  - Specify the analysis queue name by `--site siteName`
- More pathena options are available at:
  - <https://twiki.cern.ch/twiki/bin/view/Atlas/DAonPanda#synopsis>



# Run pathena (2)



- The following will be printed on the screen:

```
Your identity: /DC=org/DC=doegrids/OU=People/CN=Nurcan Ozturk 155817
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Tue Mar 4 00:50:33 2008
extracting run configuration
ConfigExtractor > No Input
ConfigExtractor > Output=AANT EVAANTupleDump0Stream AANT0
archive sources
archive InstallArea
post sources/jobO
query files in dataset:fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1
submit
```

=====

```
JobID : 8202
Status : 0
> build
  PandaID=8059902
> run
  PandaID=8059903
```

builds the athena environment at the remote site.  
It produces a library dataset.

runs athena and produces the output files

# Monitor Job's Status in PanDA Monitor (1)



Go to "List users" link at the right top corner of PanDA monitor:  
<http://gridui02.usatlas.bnl.gov:25880/server/pandamon/query?ui=users&sort=latest>

Nurcan Ozturk Panda Page - Microsoft Internet Explorer

Address: <http://gridui02.usatlas.bnl.gov:25880/server/pandamon/query?ui=user&name=Nurcan%20Ozturk>

Recent running, activated, waiting, assigned, defined, finished, failed jobs  
 Select analysis, prod, install, test jobs  
 Quick search  
 Job \_\_\_\_\_  
 Dataset \_\_\_\_\_  
 Task \_\_\_\_\_  
 File \_\_\_\_\_

Summaries  
 Blocks: \_\_\_\_\_ days  
 Errors: \_\_\_\_\_ days  
 Nodes: \_\_\_\_\_ days  
[Daily usage](#)

Tasks - search  
[Generic Task Req](#)  
[Ex-Gen Task Req](#)  
[CTBSim Task Req](#)  
[Task list](#)  
[Task browser](#)

Datasets - search  
[Dataset browser](#)  
[Aborted MC datasets](#)  
[Panda subscriptions](#)

**Sites used:** FZK-LCG2 (6) BU\_ATLAS\_Tier2o (94) TRIUMF (68) IU\_ATLAS\_Tier2 (4) LIV (2) LIP-LISBON (2) ANALY\_LONG\_LYON (2) ROMANIA07 (2) RAL (8) PIC (10) MANC (24) DESY-ZN (2) TOKYO (8) OXFORD (2) LPC (10) ANALY\_LONG\_BNL\_ATLAS (142) LAL (2) Australia-ATLAS (2) NIKH

**Job types run:** user (2030) ddm (382) test (1842) cloudtest (220) NULL (16)

**Groups:** all atlas shift usatlas

Usage	(quota)	(quota)	(quota)
Analysis	0 (500)	1 (3000)	9 (9000)
User production	0 (30)	0 (210)	0 (900)
Express	0 (150)	0 (1050)	0 (4500)

Summary of all jobs for the last 1 days in any state at any site taskID Go

**25 jobs.** Click job number to see details.  
**States:** defined:2 assigned:0 waiting:0 activated:20 running:1 transferring:0 holding:1 finished:1 failed:0  
**Users:** Nurcan Ozturk:25  
**Releases:** Atlas-12.0.6:21 Atlas-13.0.40:4  
**Sites:** AGLT2:1 ANALY\_AGLT2:2 ANALY\_SWT2\_CPB:2 ASGC:1 BNL\_ATLAS\_test:13 MONTREAL:1 MWT2\_UC:1 NULL (unassigned):1 OXFORD:1 UC\_ATLAS\_MWT2:1 UTA\_SWT2:1  
**Datasets used by selected jobs**

**Job Sets:**

User:jobID	Created	Latest	Jobs	Pre-run	Running	Holding	Finished	Failed	buildJob	Site
<a href="#">Nurcan Ozturk:8203</a>	03-03 11:29	03-03 12:56	2	1		1			8060506 libDS	ANALY_AGLT2
In:fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1 Out:user.NurcanOzturk.pathenaDemo.StreamEgamma.AGLT2_mar3										
<a href="#">Nurcan Ozturk:8202</a>	03-03 12:51	03-03 12:56	2	1			1		8059902 libDS	ANALY_SWT2_CPB
In:fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1 Out:user.NurcanOzturk.pathenaDemo.StreamEgamma.SWT2_CPB_mar3										

# Monitor Job's Status in PanDA Monitor (2)



Quick guide to the Panda monitor - Microsoft Internet Explorer

Panda monitor and browser - Microsoft Internet Explorer

Address: http://gridui05.usatlas.bnl.gov:25880/server/pandamon/query?job=8059902

**BNL monitor** Production Clouds DDM PandaMover AutoPilot Sites Analysis Physics data Usage Plots ProdDash DDMDash

Update

Not logged in. [List users](#)

**Panda monitor** **Panda job information**

Jobs: 8059902

Jobs - search  
Recent running, activated, waiting, assigned, defined, finished, failed jobs  
Select analysis, prod, install, test jobs  
Quick search  
Job 8059902

**Job Sets:**

User:jobID	Created	Latest	Jobs	Pre-run	Running	Holding	Finished	Failed	buildJob	Site
<a href="#">Nurcan Ozturk:8202</a>	03-03 12:51	03-03 12:51	1				1		8059902 libDS	ANALY_SWT2_CPB
Out:user.NurcanOzturk.acas0003_99.lib_008202										

Quick search  
Job 8059902

Dataset  
Task  
File

**Summaries**  
Blocks: days  
Errors: days  
Nodes: days  
[Daily usage](#)

**Tasks - search**  
[Generic Task Req](#)  
[EvGen Task Req](#)  
[CTBsim Task Req](#)  
[Task list](#)

Dataset  
Task  
File

**Summaries**  
Blocks: days  
Errors: days  
Nodes: days  
[Daily usage](#)

**Tasks - search**  
[Generic Task Req](#)  
[EvGen Task Req](#)  
[CTBsim Task Req](#)  
[Task list](#)

**Job 8059902 details**

2 files for job 8059902:

Filename	Type	Status	Dataset
<a href="#">user.NurcanOzturk.acas0003_99.lib_008202.lib.tgz</a>	output	ready	<a href="#">user.NurcanOzturk.acas0003_99.lib_008202</a>
<a href="#">user.NurcanOzturk.acas0003_99.lib_008202.log.tgz</a>	log	ready	<a href="#">user.NurcanOzturk.acas0003_99.lib_008202</a>

Job outputs were registered in the SWT2\_CPB storage element  
[Pilot logs](#)

start

start

Panda monitor and br... hepmail.uta.edu - Pu... nurcan@atlasgw00:~ Microsoft PowerPoint ... 12:05 PM

# Monitor Job's Status in PanDA Monitor (3)



Nurcan Ozturk Panda Page - Microsoft Internet Explorer

Address: http://gridui05.usatlas.bnl.gov:25880/server/pandamon/query?ui=user&name=Nurcan%20Ozturk

Jobs: 4586 total jobs, last at 2008-03-03 01:18

Sites used: FZK-LCG2 (6) BU\_ATLAS\_Tier2o (94) TRIUMF (68) IU\_ATLAS\_Tier2 (4) LIV (2) LIP-LISBON (2) ANALY\_LONG\_LYON (2) ROMANIA07 (2) RAL (8) PIC (10) MANC (24) DESY-ZN (2) TOKYO (8) OXFORD (2) LPC (10) ANALY\_LONG\_BNL\_ATLAS (142) LAL (2) Australia-ATLAS (2) NIKH

Job types run: user (2030) ddm (382) test (1842) cloudtest (220) NULL (16)

Groups: all atlas shift usatlas

Usage	1 day (quota)	7 day (quota)	30 day (quota)
Analysis	0 (500)	1 (3000)	9 (9000)
User production	0 (30)	0 (210)	0 (900)
Express	0 (150)	0 (1050)	0 (4500)

Summary of all jobs for the last 1 days in any state at any site taskID Go

25 jobs. Click job number to see details.

States: defined:1 assigned:0 waiting:0 activated:19 running:1 transferring:0 holding:0 finished:4 failed:0

Users: Nurcan Ozturk:25

Releases: Atlas-12.0.6:21 Atlas-13.0.40:4

Sites: AGLT2:1 ANALY\_AGLT2:2 ANALY\_SWT2\_CPB:2 ASGC:1 BNL\_ATLAS\_test:13 MONTREAL:1 MWT2\_UC:1 NULL (unassigned):1 OXFORD:1 UC\_ATLAS\_MWT2:1 UTA\_SWT2:1

Datasets used by selected jobs

Job Sets:

User:jobID	Created	Latest	Jobs	Pre-run	Running	Holding	Finished	Failed	buildJob	Site
<a href="#">Nurcan Ozturk:8203</a>	03-03 11:29	03-03 12:56	2				2		8060506 libDS	ANALY_AGLT2
In:fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1 Out:user.NurcanOzturk.pathenaDemo_StreamEgamma_AGLT2_mar3										
<a href="#">Nurcan Ozturk:8202</a>	03-03 12:51	03-03 12:56	2				2		8059902 libDS	ANALY_SWT2_CPB
In:fdr08_run1.0003050.StreamEgamma.merge.AOD.o1_r6_t1 Out:user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3										

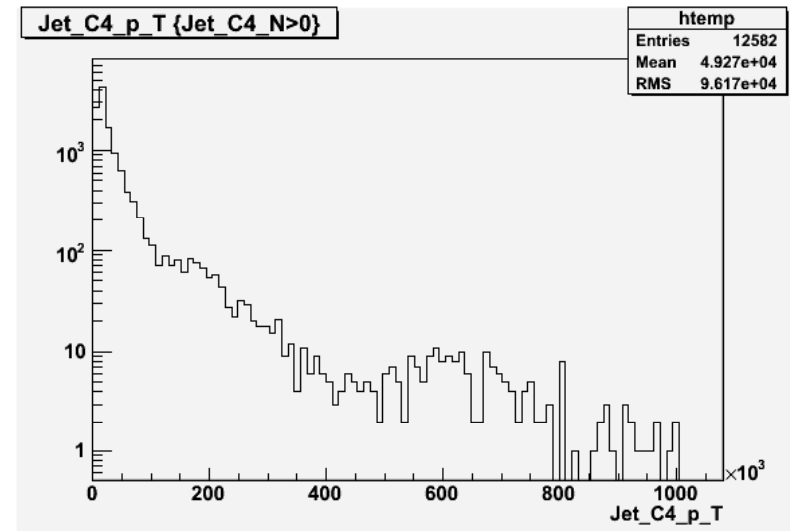
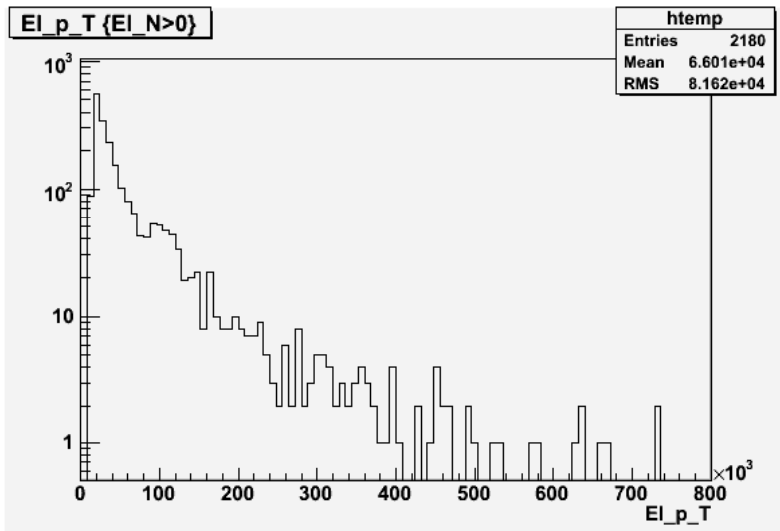
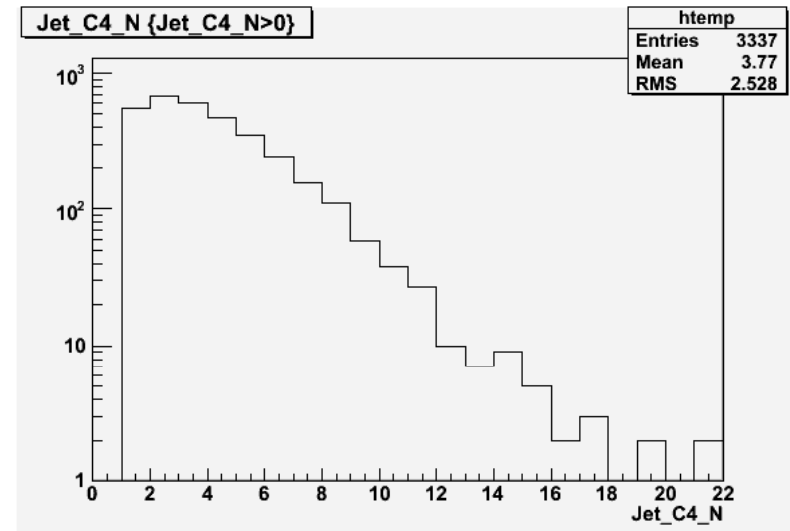
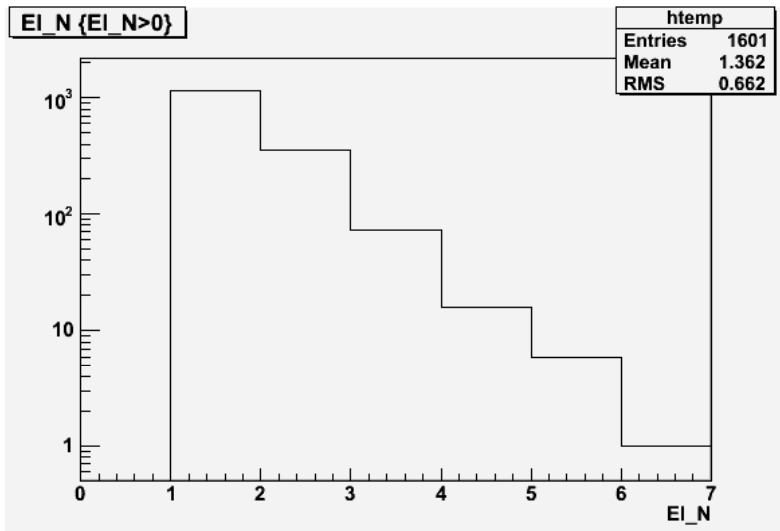
start | Nurcan Ozturk P... | The University o... | DQ2-DDM Oper... | The University o... | hepmail.uta.edu... | nurcan@atlasg... | Microsoft Power... | 12:26 PM

# Retrieve Results and Make Plots

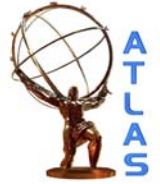


- Use dq2 client tools to retrieve the output dataset:
  - `dq2_get -rv user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3`
- This copies the output files:
  - `user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3._8059903.log.tgz`
  - `user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3.AANT0._00001.root`
- One particular user needed to use “-s OU” to retrieve output dataset from ANALY\_OU\_OCHEP\_SWT2, under investigation why
- Wisconsin site (ANALY\_GLOW-ATLAS) added all ATLAS users to its gridmap file so that all ATLAS users can retrieve the files
- Open the file in root and make some plots:
  - `root`  
`user.NurcanOzturk.pathenaDemo_StreamEgamma_SWT2_CPB_mar3.AANT0._00001.root`
  - `root [1] FullRec0->GetListOfLeaves()->Print();`
  - `root [2] FullRec0->Draw("EI_N", "EI_N>0");`
  - `root [3] FullRec0->Draw("EI_p_T", "EI_N>0");`
  - `root [4] FullRec0->Draw("Jet_C4_N", "Jet_C4_N>0");`
  - `root [5] FullRec0->Draw("Jet_C4_p_T", "Jet_C4_N>0");`

# Some Plots



# Future Developments with pathena



- Near term PanDA activities and plans were presented by Torre Wenaus at the Software&Computing Workshop last week. Among others related to analysis:
  - Automatic redirection of analysis jobs within a cloud
  - Namely, no need to specify site - pathena will choose the best site based on data availability and available CPU's

# References



- Athena software releases and how to use them:
  - <http://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/AtlasSWReleases>
- FDR datasets available at Tier2's:
  - <http://gridui02.usatlas.bnl.gov:25880/server/pandamon/query?mode=listFDR>
- pathena wiki page “Distributed Analysis on Panda”:
  - <https://twiki.cern.ch/twiki/bin/view/Atlas/DAonPanda>
- How to submit same pathena job on multiple datasets:
  - [https://twiki.cern.ch/twiki/bin/view/Atlas/DAonPanda#example\\_6\\_re\\_submit\\_the\\_same\\_ana](https://twiki.cern.ch/twiki/bin/view/Atlas/DAonPanda#example_6_re_submit_the_same_ana)
- HighPtView wiki page:
  - <https://twiki.cern.ch/twiki/bin/view/Atlas/HighPtView>
- Wiki pages by Akira Shibata:
  - <https://twiki.cern.ch/twiki/bin/view/Atlas/TopFDR>
  - <https://twiki.cern.ch/twiki/bin/view/Atlas/TopFdrPanda>