Tier 3s and Analysis

Amir Farbin
University of Texas, Arlington

Overview

- Resources Available for Analysis
- Analysis Model
  - DPDs
    - Mapping to resources
- Tier 3 resource estimation.
  - Question: can we process D1PDs at tier 3s?
- Managing Tier 3s.

Resources for Analysis

- *Tier 1 + 2*- Distributed resources accessible only through GRID software/system (eg PANDA):
  - *Good*: Potentially a large resource. Automatic book-keeping, recovery, etc. Full access to AODs/DPDs.
  - *Bad*: Large/Sophisticated (Difficulty of use. Reliability. Debugging.) Not interactive. Results need to be collected to another site for interactive analysis. Shared there is competition and we need to manage resource allocation.
- *Central Analysis Facilities* (BAF, CAF, ...)- Large shared resource (reminiscent of analysis computing at LEP, Tevatron, B-factories, ...) I'm not sure of the size.
  - *Good:* Access to ESD (?), AOD, DPDs. Data potentially staged to worker-nodes for even faster access (eg via PROOF). Interactive. Results available immediately for interactive analysis.
  - *Bad*: Shared resource for 100s of physicists → potentially over subscribed. Managed resource allocation. Compute intensive batch processing of DPDs may dominate.
- *Tier 3s*- Local resources.
  - *Good*: Interactive. Personal, not shared.
  - *Bad:* No funding. Difficult to manage/scale properly. Limited resources (ie Disk, CPU). Data must be transferred in.

DA on Tier 1/2s not enough?

- To the user, the defining difference between tier 1/2 and CAF/BAF + tier 3 is interactive access.
- Even if all analysis is done using Distributed Analysis tools on GRID, users need a place to login, in order to
  - Develop code, run test jobs
  - Submit large scale analysis to GRID
  - Gather results from GRID and perform final stages of analysis
  - Just do the type of work which is not conducive to GRID processing (eg Toy MCs, fits)
- Tier 2s in general cannot manage user accounts and interactive usage patterns.
  - Though tier 2s are a shared resource, some allow access to local users... unfair?
- CAF, BAF, ... will provide interactive analysis resources (not defined yet)
  - But these will not be at the scale provided in previous generation experiments... cannot fulfill all analysis resource needs.
  - Better reserved for things we cannot do on the GRID or Tier 3: large-scale interactive AOD/DPD analysis, and ESD analysis, calibrations (are these supposed to be at tier 1 instead?).
- Tier 3 fills in the hole... local resources used for local needs.

Other Realities

- Processing times are currently significantly longer than Computing Model (Simulation: > factor of 5. Reco: factor of 2)
- Event Data sizes are currently larger than expected (~ factor of 2 for AOD)
  - There is hope that AOD sizes can be bought within budgets
- AMF Report: Suggests that D^1PDs occupy same volume as AOD. Can potentially mean less replicas of AODs.
- We also need to store D^2PD/D^3PD.

- To cope with these pressures, the analysis capacity of tier 1/2 has been reduced in CM
  - Ultimate result: Analysis is being pushed from tier 1 → 2 → 3...
- We do not have a tier 3 model (role/size?).
- Tier 3 funding?
  - Technically there is no funding in the US for tier 3s!
  - Tier 3 size is not necessary set by requirements. It is set by what YOU can afford and manage.

Analysis Activity

- *Re-reconstruction/re-calibration*- CPU intensive... often necessary.
- *Algorithmic Analysis*: Data Manipulations ESD→AOD→DPD→DPD
  - *Skimming*- Keep interesting events
  - *Thinning*- Keep interesting objects in events
  - *Slimming*- Keep interesting info in objects
  - *Reduction*-
    - Application of algorithms: combinatorics, overlap-removal, kinematic fitting, sphericity calculation...
    - Encapsulation of the results into higher-level objects
- Basic principle: Data Optimization + CPU intensive algs → more portable input & less CPU in later stages.

- *Tier 1/2 Activity*
  - Framework (ie Athena) based
  - Resource intensive
  - Large scale (lots of data)
  - Organized
  - ***Batch***

Primary difference

- *Interactive Analysis*: Analysis Development. Debugging. Making plots/ performing studies on highly reduced data.
- *Statistical Analysis*: Perform fits, produce toy Monte Carlos, calculate significance.

- *Tier 3 Activity*
  - Often exo-framework
  - ***Interactive***

Evolution of DPDs

POOL-Based:
Analyze in Athena or ROOT via ARA

ESD

AOD

D¹PD (primary)

- Subset of the AOD through skimming, thinning, slimming
- Centrally produced
- Defined by physics/ combined-reco groups
- 10-20 different D¹PDs w/ total volume = AOD

D²PD (2ndary)

- Similar to D¹PD
- More analysis specific
- Store UserData- results of complex analysis algorithms: CompositeParticles, EventViews, ParticleViews, etc

"Flat":
Analyze in plain ROOT

D³PD (tertiary)

- Maybe similar in content to D¹PD or D²PD
- But most likely highly reduced
- Just the few quantities necessary to quickly make the final plots for your analysis
- Complex analysis on D³PD is discouraged
  - Format is proprietary

Note: AOD → D¹PD → D²PD → D³PD chain is not necessary. Steps may be skipped.

Analysis Activity Placement

- Centalized/Organized
- Requires AOD access
- Possible at BAF if AOD available... but better on Tier 1/2

- At Tier 3 if input/out fits and sufficient CPU to have reasonable processing time.

- Great for first-data and algorithm developement/ tuning.
- Not practical for lots of data.

CAF/BAF

AOD→D¹PD 1

AOD→Plots 3

AOD→D²PD/D³PD 2

D¹PD→D²PD/D³PD 5

D¹PD→Plots 4

D²PD→D³PD 6

D²PD/D³PD→Plots 7

- Tasks which require more info than D¹PD (eg re-reconstruction, jet-finding, calibration)
- Group-wide/ personal

Tier 1/2

- Where the analysis heavy-lifting occurs.

- Likely to be highly iterative

Tier 3

Physicist's Analysis Model

- The defined (by us) and adopted (by users) Analysis Models will always be at odds.
  - We anticipate their issues and try to scale.
  - Users vote with their feet... we adjust.
- Elements worth considering when trying to estimate analysis resources:
  - Raw numbers: DPD size, CPU use per event, ...
  - Inefficiency:
    - Physicists will likely first try to do things locally (eg BAF or Tier 3)... hunt for bigger resources when they have saturated their existing ones.
    - Hard to anticipate all requirements:
      - Preference for delaying decisions as much as possible. Only throw out info when run out of space.
      - Analysis will start simple... quickly become complex... procedures/ implementation won't scale well... redesign personal Analysis Model.
- AMF Report is a good example...

"AMF Report" Analysis Model

- 10-20 Centrally produced "Primary" DPDs (aka D^1PD) with total volume = AOD.
  - D^1PD: AOD volume reduction based on **simple** requirements.
    - No strict decisions like electron ID or overlap-removal.
    - Mostly reduction by throwing out containers of not need objects (eg tracks or clusters) and skimming.
  - In response to:
    - Strict selections made before DPD making for CSC analyses make iterations difficult.
      - Inability to quickly remake & download DPD... GRID DA is inconsistent with the *natural* iterative cycle of analysis.
      - Dislike + lack of familiarity with athena-based analysis tools.
  - Coupled with AthenaROOTAccess, users can go from AOD/D^1PDs directly to histograms/results.
- Why D^2PDs and D^3PDs?
  - D^1PDs will in general be large and very little analysis will go into them... so analysis on D^1PDs may be resource intensive.
  - Must store the output of D^1PD analysis as D^2PDs and D^3PDs so you can iterate faster.

The benefits of D1PDs

- First pass Event Selection (Skim) and Event Data content organized at physics/ performance group level.
  - Great means of organization of analysis activity.
  - Centralized D1PD production is a good place for re-reconstruction/re-calibration.
- Faster per event processing of D1PDs vs AODs.
  - Observed/reported by people... though the why and how is not understood (points to a problem with athena).
- Each D1PD stream has 10-20 times smaller volume than AOD
  - Smaller files may help with data staging to worker nodes at the tier 2s.
  - D1PD streams may fit at tier 3s
    - Can off-load some of the tier 2 load to tier 3s.
    - People prefer working locally.
- Great means of directly looking at large amounts of data locally. Likely very important for very early data.

Another View of D1PDs

- Why create D1PD?
  - If you are on a tier 2, the D1PD provides little added value:
    - All AOD is present on tier 2s.
    - Skimming can be achieved using TAG... don't read any events you don't want.
    - No per event speed improvements: athena only reads in containers you request.
    - So why have the D1PDs space compete with AOD?
- Answer: Since the D1PD volume is 10-20 times smaller than the AOD, you can transfer it to a tier 3 and do all your analysis using ROOT + AthenaROOTAccess.
  - Essentially the D1PD is a mechanism of moving what we imagined was AOD analysis to tier 3s.
  - The promise that you may escape using Athena and GRID is very appealing...
  - Huge implications on Tier 3. Big question: can D1PD analysis be done at Tier 3?
- In the long-run, it may not be worth letting DPD compete with AOD for space.

Defining Tier 3s

- Considering:
  - Analysis Model: A primary benefit of D1PDs is that users can in principle move AOD analysis to tier 3s.
  - Computing Model: Resource constraints result in reduction in analysis resources... to be picked up on tier 3s.
- Ask Questions:
  - How much resources do you need to process D1PDs at tier 3s?
  - Disk: Does the data fit?
  - CPU: Is there enough CPU to process it within reasonable time frame?
  - DDM/Network:
    - Can the D1PD be brought to tier 3 within appropriate time frame?
    - Can the DDM handle the traffic?

D¹PDs at Tier 3s (Disk)

FDR Feedback

Size (in kB/event) considering one FDR08 run at a low luminosity (10^{+31}/cm^2/s): run 3050 (0.036/pb, 30 lumi-blocks of 2mn each). The following numbers are based on this tag: o1_r6_t1.

	Egamma	Muon	Jet	Signal (top)	Atlfast
AOD	162.2	172.9	163.5	390.4	40.5
	6785	3491	17407	545700	
	1.1GB	0.603GB	2.85 GB	212 GB	
	30GB/pb	16.2GB/pb	79.2GB/pb		
D2PD	17.5	14.3	18.4	31.8	2.4
	0.117GB	0.051GB	0.311GB	18GB	
	3.25GB/pb	1.41GB/pb	8.64GB/pb		

- D¹PDs are 5-10% of AOD. Nominally AODs are 100TB/year → 5-10TB per D¹PD
  - D¹PD ~ 1/3 of present AOD contents to get full functionality of some present AOD analyses.
  - Suggests 15-30% skim of data in each D¹PD.

FDR Feedback

☞ A realistic approach that we will check once more with the coming FDR2 exercise:

- run on the AOD datasets on the GRID

- store the produced D2PD on the Tier3 and access them for analysis refinement.

- From Nabil Ghodbane Top DPD
  - Current D¹PD already too big for tier 3... yet not enough info for full analysis.
  - D²PD: 13.3 GB/pb. 1.3 TB for 100/pb.

☞ Unknown issues: so far, with the current EDM, reaching a significant fraction of an AOD size requires dropping trigger, cluster information, etc...(This is the case for TopPhysDPDMaker, the Top WG D1PD tool)

How will this evolve with trigger EDM ?

Question: Will in the end D1PD be simply dropped for D2PD production directly from AODs? (is CPU cheaper that DISK ?)

CPU For D1PD Processing

- CPU Usage scales with number of events and how much data is read (not coupled to size of AOD/DPD)
- Difficult to estimate CPU needs... depends on task.
  - Some calibration/re-calibration tasks... eg jet finding.
  - Selection, matching, overlap-removal, combinatorics, kinematic fitting, variable calculation, ...
  - My old estimates (2 years ago) suggested O(100 ms)/event for realistic analysis... 20 ms/event cutting all inefficiencies.
  - Currently estimates from Top (Nabil) and SUSY (Renaud Bruneliere) of simple analyses (w/ selection, overlap removal) point at ~ 2-3 ms/event.
  - For today, I'll guess 10 ms/event.
- Basic (Trivial) numbers (mistake in table):

Percent of 1 year's data processed:

Processing time (ms)	Rate (Hz)	Ignore!	% per night per cores	% per night per 10 cores	% per night per 100 cores
1	1000	0.288%	2.880%	28.800%	All
2	500	0.144%	1.440%	14.400%	All
10	100	0.029%	0.288%	2.880%	28.800%
100	10	0.003%	0.029%	0.288%	2.880%

Processing time depending vs % of AOD read.

% AOD Read	Rate (Hz)	Processing time (ms)
1	5000	0.2
10	500	2
20	250	4
30	166.67	6
50	100	10
100	50	20

How many CPUs?

- Consider the percentage of 1 year's data (10^9 events) processed by N cores.

- No CPU usage, pure I/O. Reading all D^1PD contents (30 Kb/event). Perfect hardware.

Cores	1	25	100	1000
1 Hour	0.11%	2.72%	10.88%	All
Overnight	1.31%	32.63%	All	All
1 Week	18.27%	All	All	All
1 Month	78.31%	All	All	All

- 25 cores→1 analysis iteration by afternoon.

Cores	1	25	100	1000
1 Hour	0.02%	0.52%	2.08%	20.81%
Overnight	0.25%	6.24%	24.97%	All
1 Week	3.50%	87.40%	All	All
1 Month	14.98%	All	All	All

- What we see today for simple analysis:
  - Reading 10% of AOD, writing 1%.
  - 2 ms/event processing.

- 10 (wrong in talk) ms/event, writing D^2PDs
- 25 cores→1 analysis iteration in 1 day.

Cores	1	25	100	1000
1 Hour	0.12%	2.90%	11.61%	All
Overnight	1.39%	34.84%	All	All
1 Week	19.51%	All	All	All
1 Month	83.61%	All	All	All

Moving D1PDs to Tier 3s

- *Network*: AMF calls for D1PDs produced at tier 1/2's, once a month. (or 4x a year?)
  - Full D1PD stream is 5-10 TB.
  - Assuming each site will download one D1PD and O(100) Tier 3s (ie ~3 Tier 3s per Tier 2), can DDM/network transfer 500-1000 TB to tier 3s once a month?
    - Apparently the answer is yes!

Practicalities

- *Disk*:
  - One 5-10TB D^1PD stream is not sufficient.
  - Need data and MC, and multiple versions.
  - Since processing takes time, also need to make/store D^2PD/D^3PDs.
  - Easy to argue that a tier 3 needs at least 20-40TB/year to be able to process D^1PDs.
  - Many worry that tier 3s cannot handle managing this much disk.
- *CPU*:
  - 25 Core Tier 3 takes ~ 1 days for one iteration of D^1PD analysis (at 100 ms/event).
  - Clearly D^1PD analysis shouldn't simply make plots... iterations must be faster. So must make D^2PD, D^3PDs.
  - Need to process D^2PD, D^3PDs too! This isn't instant... but a few cores are sufficient.
- My numbers are very simplistic. They will not scale linearly...
- Bottom line: Tier 3 which wish to process D^1PDs need lots of resources.
  - Non-trivial to setup and efficiently operate a tier 3 of this scale.

What does a Tier 3 looks like?

- 2 types of tier 3:
  - The GRID tier 3:
    - Begins by providing GRID services in order to opportunistically offload pressure from tier 1/2.
    - Eventually provides interactive access to local users, with work areas, batch queues or PROOF... and support.
    - This could be a GRID side that give local access, or a departmental cluster that adds GRID services.
  - The local tier 3:
    - Focus on interactive access to local users... may start very small.
    - Over time grows from a few machines to size that may be worth offering to the GRID.
    - This could be a few multi-core desktops + disks

ATLAS Tier 3

- Many ATLAS tier 3 sites today are the first type... run GRID services, needs to implement interactive access.
- Unfortunately providing interactive access usually requires support people at local institution.
- Need to partition off interactive machines, setup accounts, provide home/data disks, install software, make sure experiment software runs, setup queues or PROOF.
- At UTA we are exploring how to turn our DPCC cluster (old D0 farm) into a real tier 3:
  - Disk aggregation with xrootd
  - User accounts, queues, proper environment setup
  - PROOF
- We are also exploring using Virtual Machines to simplify setting up tier 3s.
  - No need to install OS, etc... just run VMs.
  - Opportunistic use of University resources (eg Windows machines in labs)... best for simulation.
- Goal is to explore possibilities and the provide recommendation on configuration and detailed instructions for others to follow.
- Exploring how to remotely provide interactive support to local users at other sites.

Comment on Simulation

- Many Tier 3s want (or already do) to contribute to simulation production.
- Running simulation on tier 3s and opportunistic resources is easier than running analysis (less data required).
- If analysis resources at tier 2s are scarce, we may want to rely on tier 3 and opportunistic resources for simulation.
  - Better for analysis because the data is already at tier 2s.

Final Remarks

- If the analysis model and computing constraints are pushing for processing D^1PDs at tier 3s, then tier 3s must be rather large.
  - You'll need funds for equipment... and need a person to manage it.
  - Big Universities and Labs may have the means... but small ones need to plan.
    - You can't just stick machines into racks.
- No matter what, D^1PD analysis will take time, so we will need D^2PDs and D^3PDs.... and a place to analyze them (ie tier 3s).
- Alternative is processing D^1PDs into D^2PDs on tier 1/2s.
  - Must make sure there are sufficient resources on tier 1/2.
- We shouldn't forget the role of tier 3s which will not be fulfilled elsewhere:
  - Interactive analysis... presumably of D^2PDs and D^3PDs.
  - Toy MCs, fits...
- Setting up and maintaining Tier 3s will be non-trivial. We need to organize and consider technologies like VMs.