Computing System Modeling (for Tier 3 Task Force)

Amir Farbin
University of Texas, Arlington

Goal

- Make quantitative arguments for Tier 3s.
- Resources argument for Tier 3s:
  - Insufficient resources for Monte Carlo needs
    - Very little room for contingencies...
      - Slow analysis turn around
        - Necessitates organized analysis activity (DPD making) and practical analysis models.
      - No resources allocated for statistical techniques (ie fits, Toy Monte Carlos, Discriminants like boosted decision trees) or advanced techniques (eg Matrix Element Methods)

Task at Hand

- *Goal:* Quantitatively study Analysis Model (AM)/Computing Model (CM) interactions.
  1. *Input AM parameters*: Details of the steps in analysis, like speed, input/output sizes/rates, transfer sizes/rates...
     - A use-case (eg Top analysis) is a processing *chain*
     - A step in the chain is a *transformation*
  2. *Input CM parameters:* Types of facilities, size/allocation of their resources
     - A class of facilities (eg Tier 1s) are a *resource*
  3. *Calculate a figure of merit:* Time it takes to finish a chain.
     - How much bandwidth required between resources.
- *Approach*: everything is a model... but perform a calculation, not a simulation.
  - Steady-state... at least for now.
  - Must study the whole system and the interaction of competing goals: production *vs* different analyses.

The Calculation

1. Specify *Resources*
   - eg T1: 10 x 2200 kSI2K, T2: 30 x 2000 kSI2K, T3: 100 x190 kSI2K
2. Specify *Chains* → Series of *Transformations*
   - *Transforms* calculate how much CPU (kSI2K sec) and Input/Output (KB/s) they need to complete.
3. Collect *Transforms* from *Chains*, assign them to *Queues* at specific *Resources*.
4. Ask *Resources* to assign CPU to *Transforms*.
   - *Production Queues* provide constant throughput.
   - *Analysis Queues* share resources equally between all transforms.
5. Ask *Transforms* to calculate their processing time (CPU and IO).
6. Ask *Chains* to sum up contributions from *Transforms*
7. Ask *Chains* to summarize

Inputs

Year	Tier 2 CPU (kSI2K)	Events Recorded	Events Fully Simulated
2008	21612	8×10^8	3.2×10^8
2009	34441	1.2×10^9	4.8×10^8
2010	60630	2×10^9	6×10^8
2011	92155	2×10^9 (?)	(?)

Step (tt events)	CPU per event (kSI2K sec)
Generation	0.23
Full Simulation	2000
Fast Sim(ATLFAST-II)	100
Fast Sim(G4-Fast)	700
Fast Sim(ATLFAST-IIf)	10
Digitization	29 (*)
Reconstruction	47

Luminosity	* Digitization CPU Factor
1×10^{32}	1
1×10^{33}	2.3
3.5×10^{33}	5.8
1×10^{34}	160

Example Output

- Model:
  - 100% of Tier 1 for Generation+Reprocessing
  - 50% of Tier 2 for Simulation+Reconstruction
  - 2 Chains:
- Calculation:
  - Resources for transformation ~ fraction resources need (excluding IO, for now)
  - All transforms on a resources should finish in about the same time

3. Total # of days for one pass

4. Numbers in parenthesis is for *all* passes

1. The Chain Name

5. Properly account for time spent in Input/Output

2. Transforms in Chain

```
Monte Carlo:
    (Nothing)--> [Generation (Monte Carlo)]--> (Gen)
          Total Time: 209.736 ( 209.74 ) days, IO/CPU Fraction: 0.3
    (Gen)--> [Simulation (Monte Carlo)]--> (Sim)
          Total Time: 185.185 ( 370.37 ) days, IO/CPU Fraction: 0.0
    (Digi)--> [SimReconstruction (Monte Carlo)]--> (SimESDAOD)
          Total Time: 185.186 ( 370.37 ) days, IO/CPU Fraction: 0.0
    (AOD)--> [AOD->D1PD (Monte Carlo)]--> (D1PD)
          Total Time: 163.073 ( 1956.88 ) days, IO/CPU Fraction: 0.01
Chain Total: 209.74 ( 1956.88 ) days, IO/CPU Fraction: 0.03 ( 0.26 )
Flow Volume (TB):  {'Tier2->Tier1': 27.939677238464355, 'Tier1->Tier2': 1.862645149230957}
Flow Rate (MB/sec):  {'Tier2->Tier1': 1.8310470619500168, 'Tier1->Tier2': 0.10778123340503583}

_____________________________________
Reprocessing:
    (RAW)--> [Reconstruction (Reprocessing)]--> (ESDAOD)
          Total Time: 160.738 ( 160.74 ) days, IO/CPU Fraction: 0.0
    (AOD)--> [AOD->D1PD (Reprocessing)]--> (D1PD)
          Total Time: 163.073 ( 326.15 ) days, IO/CPU Fraction: 0.01
Chain Total: 163.07 ( 326.15 ) days, IO/CPU Fraction: 0.01 ( 0.01 )
Flow Volume (TB):  {}
Flow Rate (MB/sec):  {}
```

5. Transforms run simultaneously, so the chain is limited by slowest step.

7. Generation (Tier 1) → Simulation (Tier 2)

8. Reconstruction (Tier 2) → D1PD Making(Tier1)

6. Required network flow: Min (output rate N, input rate of N+1)

- Conclusion: Must dedicate more to Tier 1 resources for D1PD production.

Steps

- Determine how much of Tier 2 resources will be required for analysis.
- Determine the analysis turn around time on Tier 2s (using remaining resources).
- How will tier 3 help?
  - Actually focused on tier 2s will be insufficient.
  - Don’t know the scale of tier 3s.

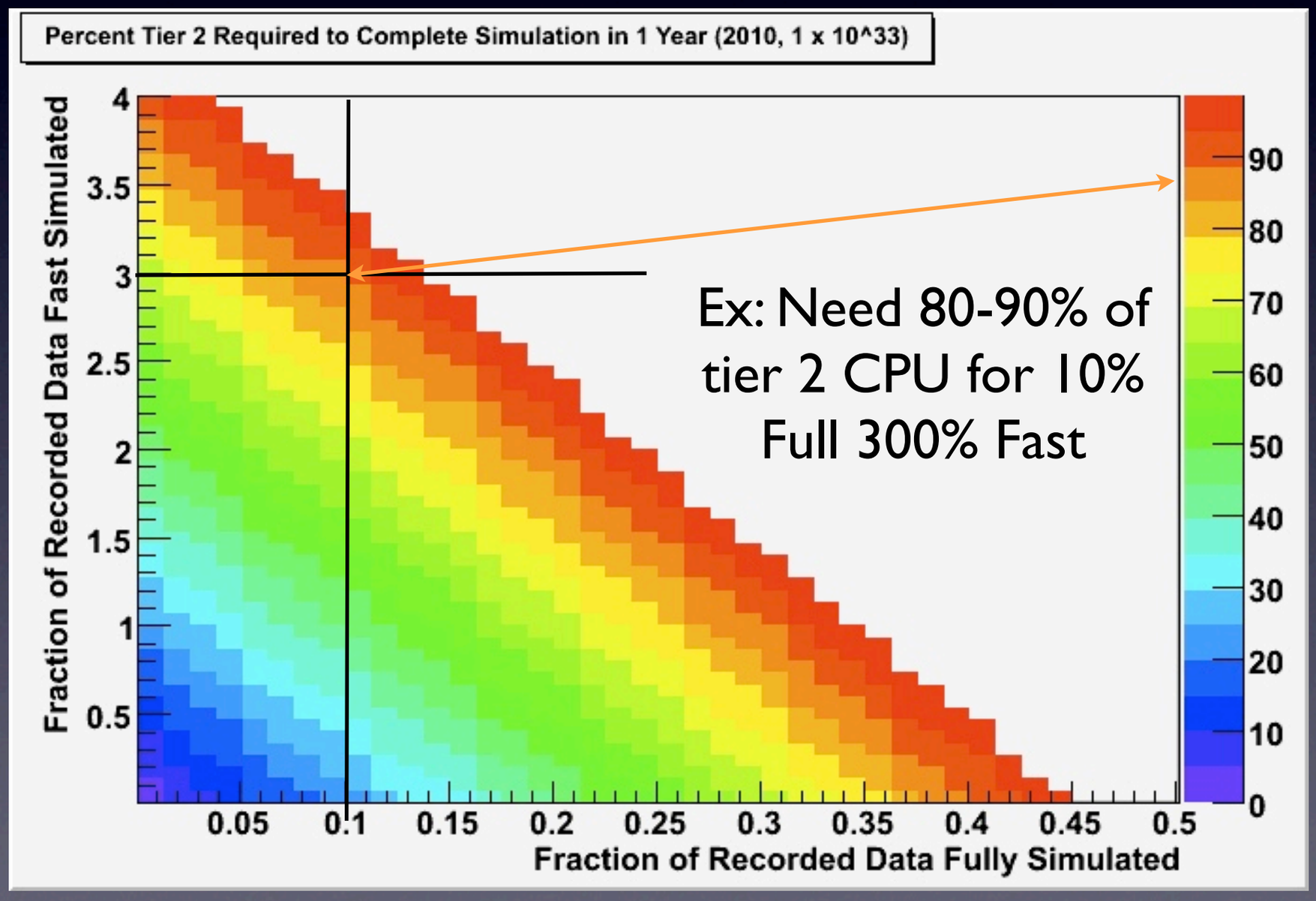
Scenarios

- Analysis on tier 2s get the simulation leftovers...
- So we must first figure out how much of the tier 2s will be available for simulation.
- Illustrative scenarios for 2010:

Calculation	Tier 2 Production Fraction	Sim Fraction	Fast Sim Fraction	Luminosity	Time (days)
1	50%	10%	0%	1×10^{32}	159
2	50%	10%	0%	1×10^{33}	162
3	50%	20%	0%	1×10^{33}	323
4	50%	0%	100%	1×10^{33}	166
5	50%	10%	100%	1×10^{33}	328
6	50%	10%	300%	1×10^{33}	660
7	75%	10%	300%	1×10^{33}	443
8	90%	10%	300%	1×10^{33}	371
9	100%	10%	300%	1×10^{33}	336

Scan

- Calculate fraction of Tier 2 CPU necessary to complete Monte Carlo production in 1 year as function of fraction of recorded data fast/full simulated.



- Note:
  - 2 passes = 2x fraction
  - Different fast sim? Scale y-axis (x 7 for fG4, x 0.1 for fATLFAST-II)

→Assume 80% of Tier 2s for simulation for remainder of calculations.

The Haze

- Haze = The steady load on our computing systems
- Consists of:
  - *Production*: Reprocessing, Monte Carlo (Simulation), Primary DPD Making
  - *Performance Activity*: Read Perf D1PD, high CPU.
  - *DPD Making*: Large scale data preparation. eg AOD, D1PD → D2PD, D3PD.
  - *Final Analysis*: Repeated iterations over DPD producing results (plots, measurements, etc).
- All of these co-exist on our system, competing for resources.

Analysis Complications

Stage 0

Re-reconstruction/re-calibration- CPU intensive... often necessary.

Stage 1

Algorithmic Analysis: Data Manipulations ESD→AOD→DPD→DPD

- *Skimming*- Keep interesting events
- *Thinning*- Keep interesting objects in events
- *Slimming*- Keep interesting info in objects
- *Augmentation*-
  - 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.

- Framework (ie Athena) based
- Resource intensive
- Large scale (lots of data)
- Organized
- ***Batch***

Primary difference

Stage 2

Interactive Analysis: Analysis Development. Debugging. Making plots/ performing studies on highly reduced data.

Stage 3

Statistical Analysis: Perform fits, produce toy Monte Carlos, calculate significance.

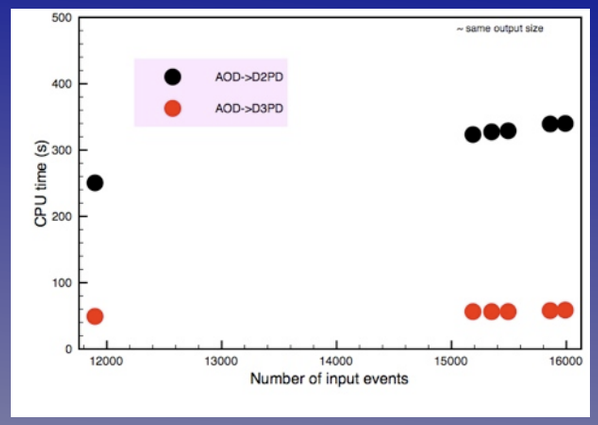
- Often exo-framework
- ***Interactive***

Modeling Analysis Plans

Stage 0: Re-reconstruction/ calibrate

Inputs based on performance DPD contents and reconstruction profiling.

Stage 1: Algorithmic Analysis



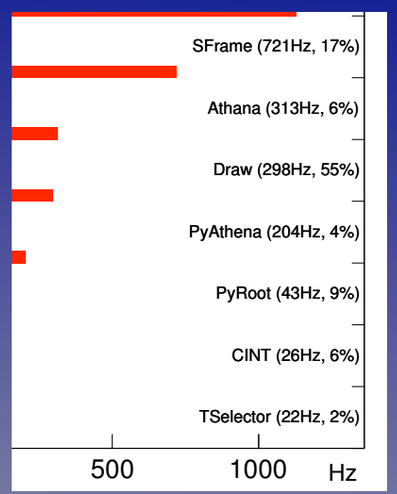
Inputs based on studying existing DPD making jobs in PANDA logs.

M. Neubauer, A. Shibata

Monte Carlo Production + Reprocessing

Inputs based on production profiling.

Stage 2: Interactive Analysis



Detailed profiling of different analysis styles. *A. Shibata*

Computing System Modeling

A. Farbin

Toy Model of Analysis Activity

Analysis	Tier 1/2	Tier 3 (or Tier 1/2)	# Events	Instances
Dijet	High (ESD, DB)	Low	High	2
Top	Low	High (eg Kin Fits, ...)	High	5
SUSY	Low	Low	High	10
Higgs (rare)	Low	High (eg Vertex Refit)	Low	30

A. Farbin

Optimal?

Computing Model

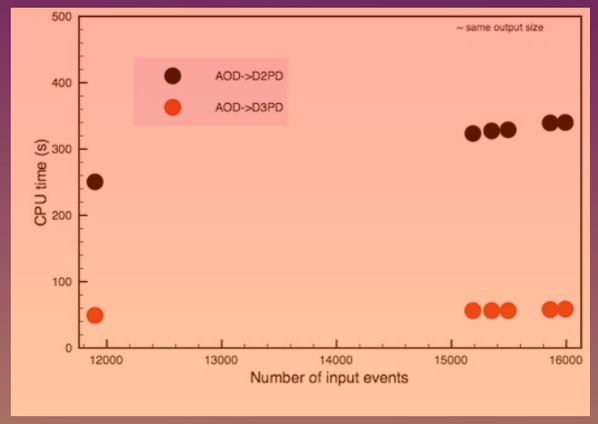
Resources at Tier 1, 2, 3 and analysis facilities.

Modeling Analysis Plans

Stage 0: Re-reconstruction/ calibrate

Inputs based on performance DPD contents and reconstruction profiling.

Stage 1: Algorithmic Analysis



Inputs based on studying existing DPD making jobs in PANDA logs.

M. Neubauer, A. Shibata

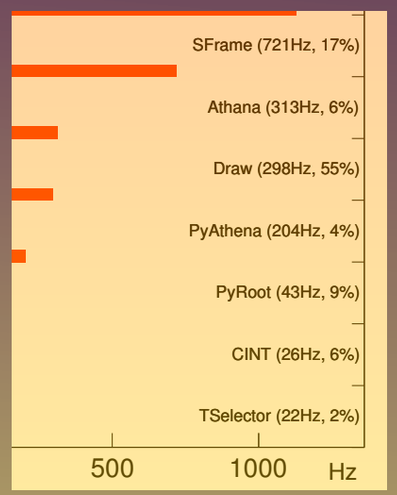
Monte Carlo Production + Reprocessing

Inputs based on production profiling.

Computing System Modeling

A. Farbin

Stage 2: Interactive Analysis



Detailed profiling of different analysis styles. *A. Shibata*

Toy Model of Analysis Activity

Analysis	Tier 1/2	Tier 3 (or Tier 1/2)	# Events	Instances
Dijet	High (ESD, DB)	Low	High	2
Top	Low	High (eg Kin Fits, ...)	High	5
SUSY	Low	Low	High	10
Higgs (rare)	Low	High (eg Vertex Refit)	Low	30

A. Farbin

Optimal?

Computing Model

Resources at Tier 1, 2, 3 and analysis facilities.

Modeling Analysis Plans

Stage 0: Re-reconstruction/ calibrate

Inputs based on performance DPD contents and reconstruction profiling.

Stage 1: Algorithmic Analysis

Inputs based on studying existing DPD making jobs in Athena logs.

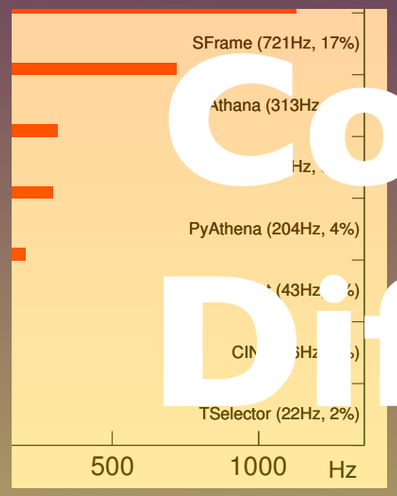
M. Neubauer, A. Shibata

Monte Carlo Production + Reprocessing

Inputs based on production profiling.

Too Complicated. Too Confusing. Difficult to explain.

Stage 2: Interactive Analysis



Detailed profiling of different analysis styles. *A. Shibata*

Toy Model of Analysis Activity

Analysis	Tier 1/2	Tier 3 (Tier 1/2)	# Events	Instances
Dijet	High (ESD, DB)	Low	High	2
Top	Low	High (eg Kin Fits, ...)	High	5
[illegible]	Low	Low	High	10
Higgs (rare)	Low	High (eg Vertex Refit)	Low	30

A. Farbin

Computing System Modeling

A. Farbin

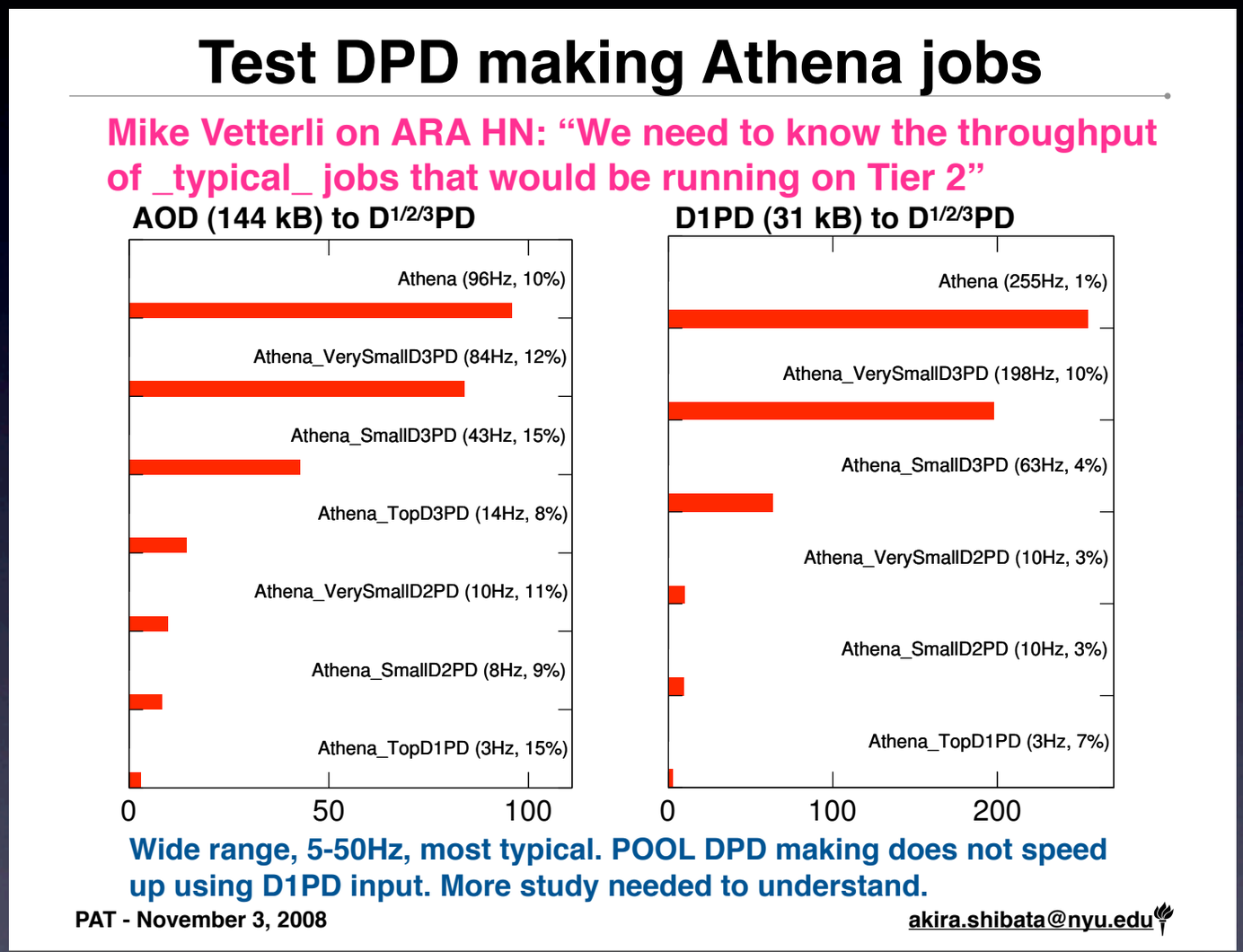
Optimal?

Computing Model

Resources at Tier 1, 2, 3 and analysis facilities.

Analysis Model Inputs

- Use Akira's performance studies



- POOL DPD production isn't sensitive to input file size.
- CPU time strongly correlated to output size.
- More info written out, more data read, more operations performed.

- I assumed different input/ output sizes, so I use these numbers as guidelines...

		Rate (Hz)	Rate (Hz)
Output	Event Size (KB)	Input: AOD	Input: D1PD
None	0	96	255
VerySmall D3PD	0.37	84	198
Small D3PD	0.71	43	63
Top D3PD	4.9	14	N/A
VerySmall D2PD	1	10	10
Small D2PD	18.7	8	10
Top D1PD	31.4	3	3

Simple Model

Data	Task	Organization
D^1PD 25 KB/event 10% of all data (recorded+sim)	*D^2PD Making* No skimming/thinning Augmentation (1.2x) ~ 3 Hz	*Physics Groups...* Nominally ~10 in ATLAS
D^2PD 30 KB/event 10% of all data (recorded+sim)	*D^3PD Making* No skimming Thinning/Slimming ~ 10 Hz	*Physics Sub-Groups...* Nominally ~5 per Physics Group
D^3PD 10 KB/event 10% of all data (recorded+sim)	*Plotting* 10000 Hz	*Individual* Nominally ~10 per Physics Sub-group
Plots 0 KB/event		=Total of 500 Analyzers

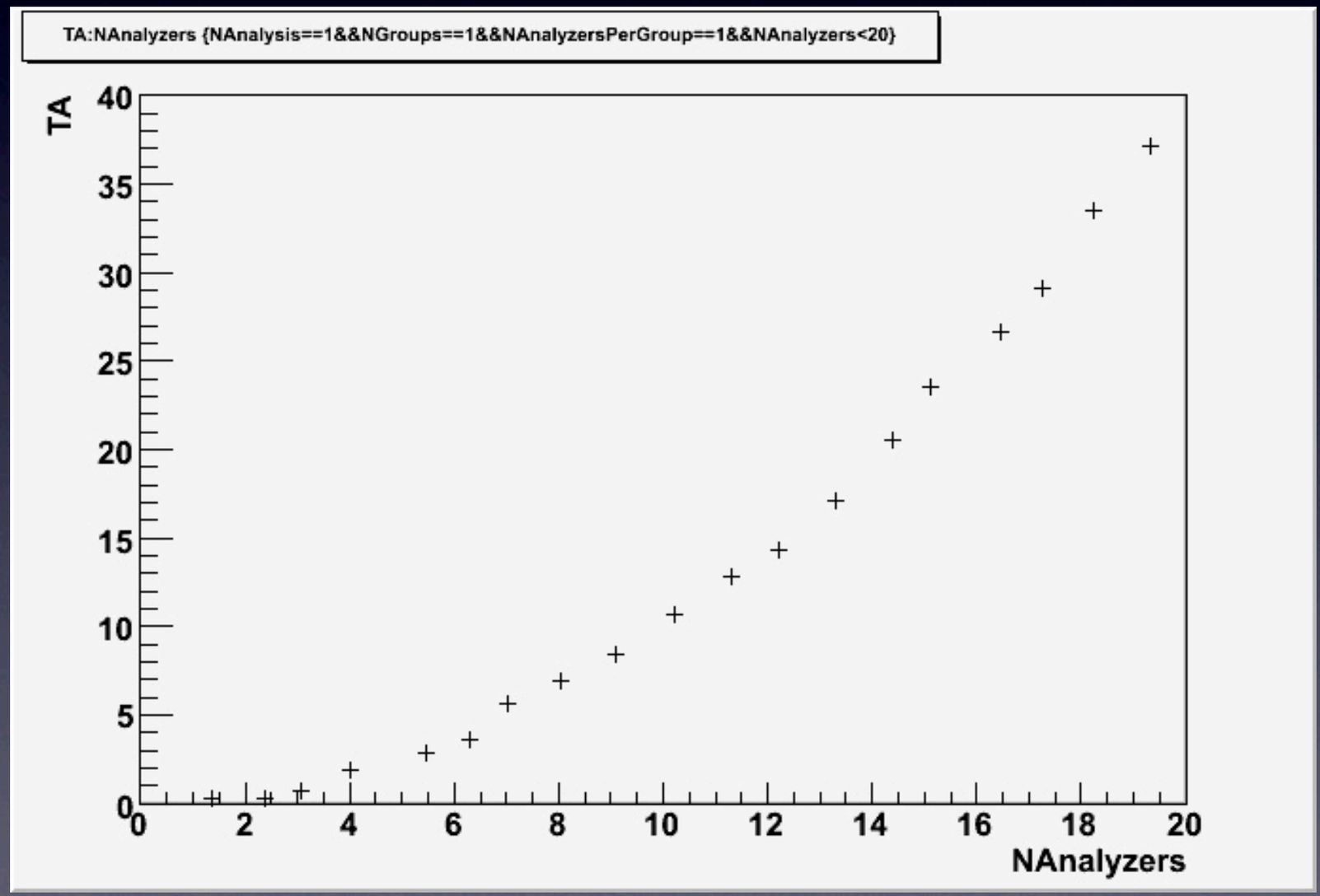
Walk-through ($D^1PD \rightarrow D^2PD$)

- A single person running a ~3Hz $D^1PD \rightarrow D^2PD$ making job on Tier 2s.
- The total T2 CPU in 2010 is 60630 kSI2k. So 20% for analysis is ~12000 kSI2K.
- Total CPU for $D^1PD \rightarrow D^2PD$ = 12000/3 kSI2K = 4000 kSI2K
  - This is because we are running $D^1PD \rightarrow D^2PD$, $D^2PD \rightarrow D^3PD$, and $D^3PD \rightarrow$ Plots all at the same time, and they all get the same amount of resources.
- Total Events= 2×10^9 Events recorded + 2×10^8 Event simulated + 8×10^9 Fast simulated * [fraction in D^1PD = 0.1] = 8.2×10^8 Events
  - (Note that I'm assuming we are going to run over the fast simulation data, which Chip assumes is 3x recorded data... If you want to ignore fast sim, just reduce all times by 1/4)
- 3 Hz = 1.4 / 3 kSI2K sec = 0.47 kSI2K per event
  - $\Rightarrow$ Total required CPU: 3.8×10^8 kSI2K sec
- Total time= 3.8×10^8 kSI2K sec / 4000 kSI2K = 95000 secs. = 27 hours ... let's say 1 day.
- So 10 people making D^2PDs will take 10 days.

Ex: No Organization

- Assume everyone does every step.
- So N groups = N subgroups = N plotters = N Analyzers

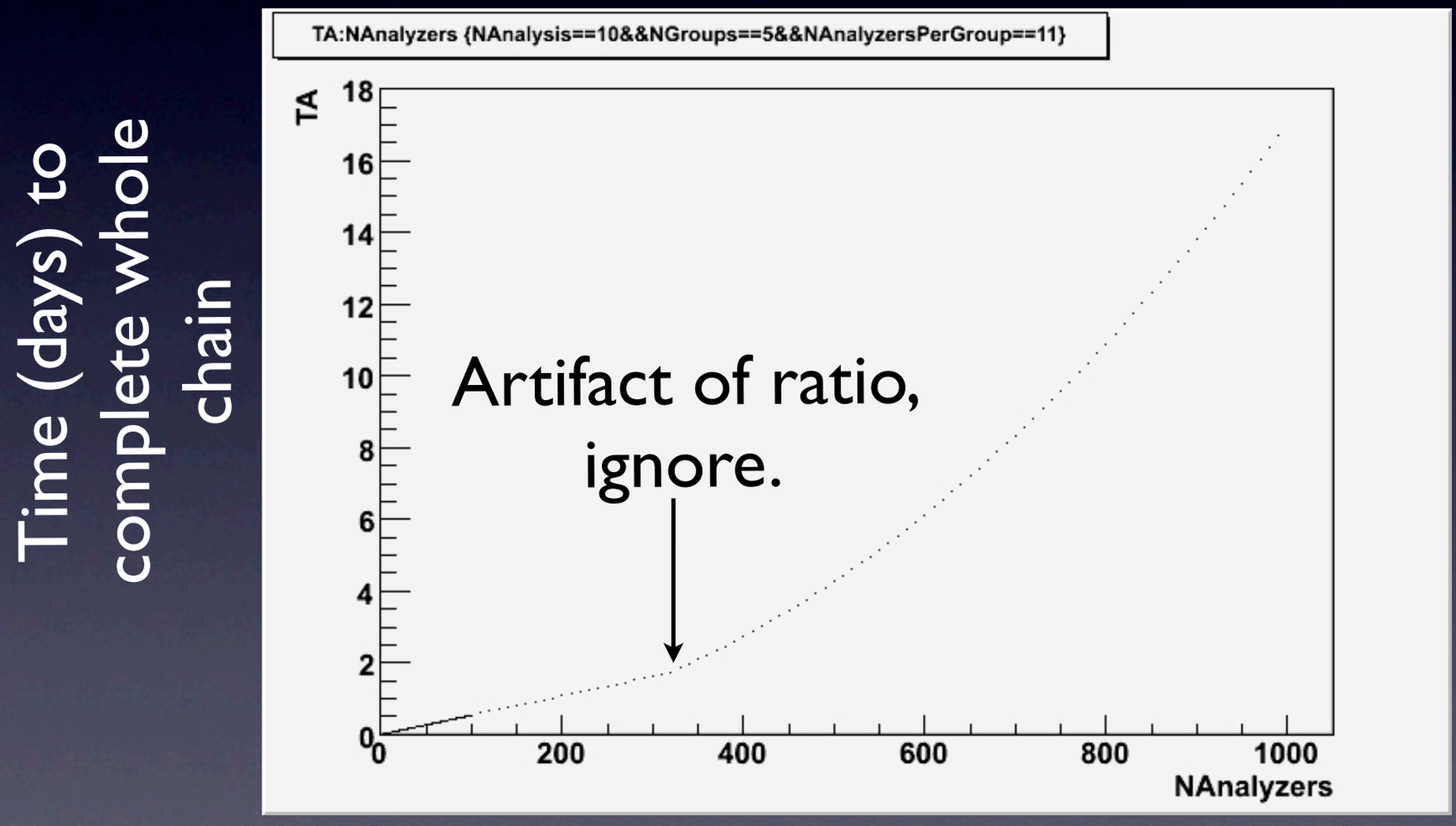
Time (days) to complete whole chain



➙Takes 10 Simultaneous Analyzers 12 days for one pass!

Organized Analysis

- Nominal Physics groups:Sub-groups:Plotters=10:5:10= 500 Analyzers
- Keep same ratio, change number of analyzers



➙Takes ~800 Simultaneous Analyzers ~10 days for one pass.

- But D3PD making and Plotting passes can be repeated quickly.
- Note: New D1PDs of all data once a month.

Details

```
Generic D1PD Analysis:
    (D1PD)--> [D1PD->D2PD (Generic D1PD Analysis)]--> (D2PD)
      NEvents:  820000000.0  CPU Needed:  3472700000.0  CPU Provided:  4042.0
      In:  25.0 ( 25.0 ) Out:  30.0 ( 30.0 )
      IO Needed:  178.571428571  IO Provided:  27940092.1659
          Total Time: 10.131 ( 121.57 ) days, IO/CPU Fraction: 0.02
    (D2PD)--> [D2PD->D3PD (Generic D1PD Analysis)]--> (D3PD)
      NEvents:  820000000.0  CPU Needed:  5912200000.0  CPU Provided:  4042.0
      In:  30.0 ( 30.0 ) Out:  9.9 ( 9.9 )
      IO Needed:  372.932049724  IO Provided:  27940092.1659
          Total Time: 17.607 ( 211.28 ) days, IO/CPU Fraction: 0.04
    (D3PD)--> [D3PD->Plots (Generic D1PD Analysis)]--> (Plots)
      NEvents:  820000000.0  CPU Needed:  57400000.0  CPU Provided:  4042.0
      In:  9.9 ( 9.9 ) Out:  0.0 ( 0.0 )
      IO Needed:  9082.56880734  IO Provided:  27940092.1659
          Total Time: 0.284 ( 34.06 ) days, IO/CPU Fraction: 0.73
Chain Max: 17.61 ( 211.28 ) days, Chain Total: 28.02 ( 366.91 ) days, IO/CPU Fraction: 0.07
( 0.04 )
Flow Volume (TB):  {}
Flow Rate (MB/sec):  {}
```

Summary

- Biggest argument for Tier 3 is contingency.
  - We will always be wanting for more full simulation.
  - Make up the rest with fast MC... which isn't free.
  - Tier 2s MC production capacity (assuming 80%) only allows for 1 pass/year at 10% full 300% fast.
- Most DPD production activity will need to happen on Tier 2s.
  - DPD analysis (eg making plots from D3PD) is best on Tier 3.
  - Moving such activity from Tier 2 to Tier 3 provides more DPD making capacity to Tier 2s.

Final Remarks

- Predicting Analysis activity is nearly impossible...
  - If you think my assumptions are too pessimistic, consider all of the difficult use cases that I didn't put in.
    - No matter how you work it, analysis resources will be scarce.
  - My model already accounts for complicated effects like ROOT I/O limit, disk-overloading, site network limits, transient→persistent time... but I need a lot of inputs to study effect of these features.