ATLAS Computing Model, Tier3 Report, Tier3 Plans

Jim Cochran
Iowa State

Outline

Computing Model: Data formats & data flow model

What's decided and what's not

Expected analysis patters for early data

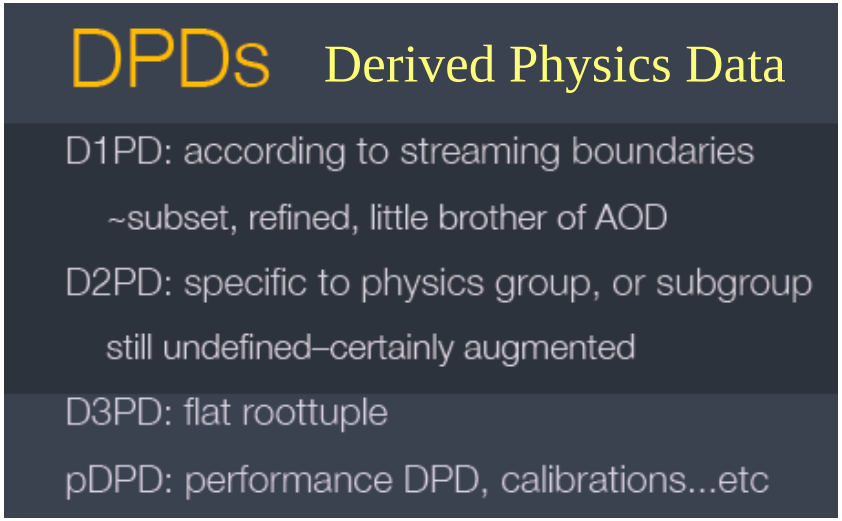
Tier3 Report Summary

Some Results from the US Requirements Study

Tier3 Plans (focus on T3w and T3g)

Data Formats

Format	Size(MB)/evt
RAW - data output from DAQ (streamed on trigger bits)	1.6
ESD - event summary data: reco info + most RAW	0.5
AOD - analysis object data: summary of ESD data	0.15
TAG - event level metadata with pointers to data files	0.001

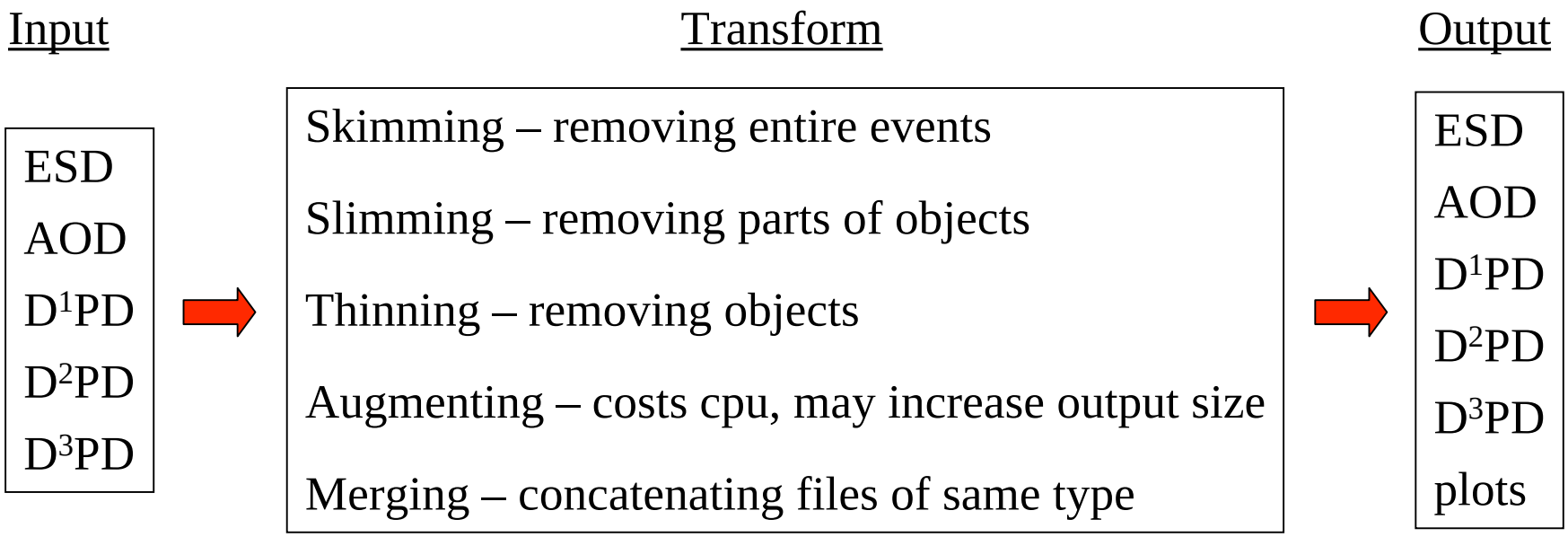


~25 kb

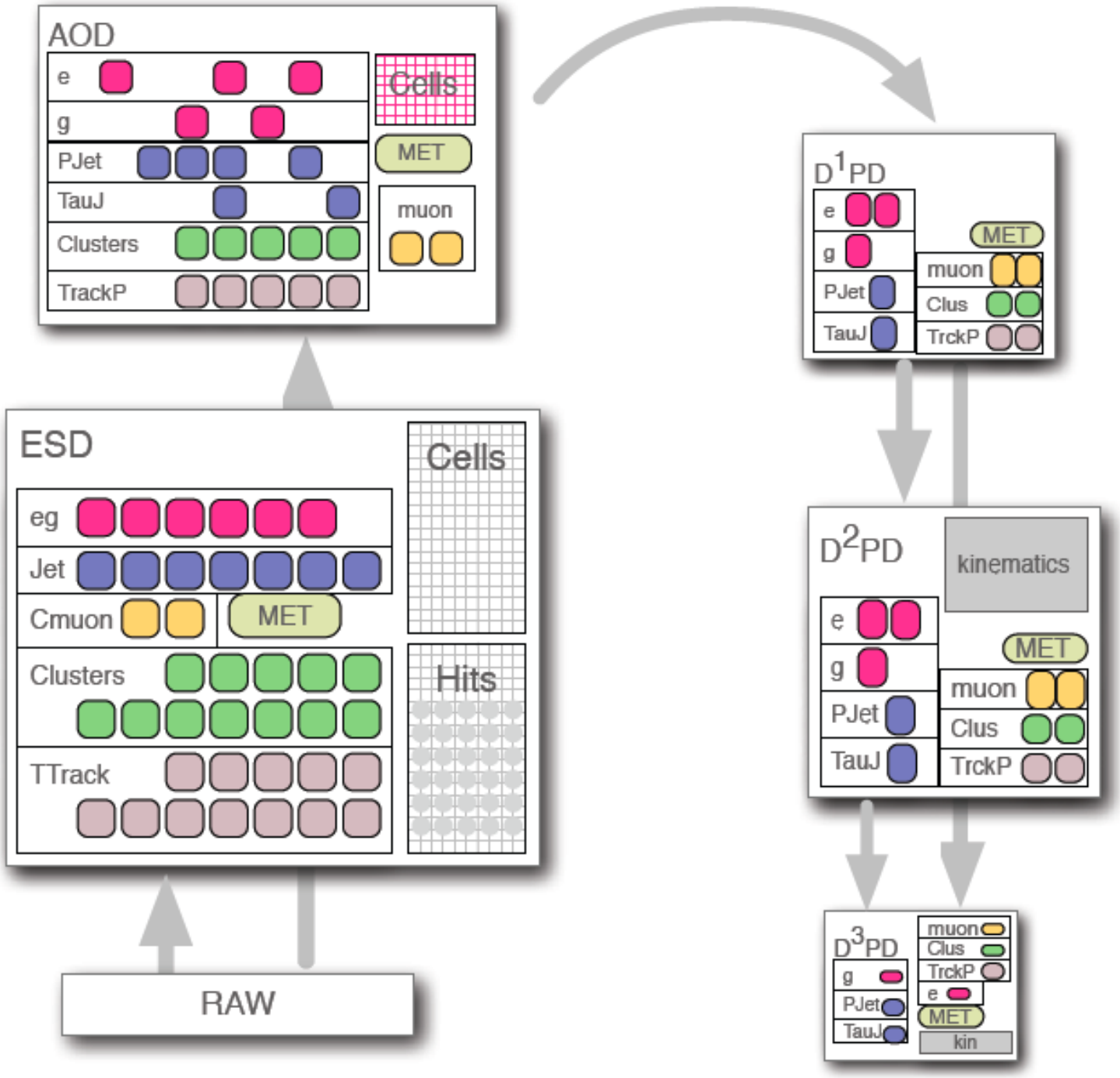
~30 kb

~5 kb

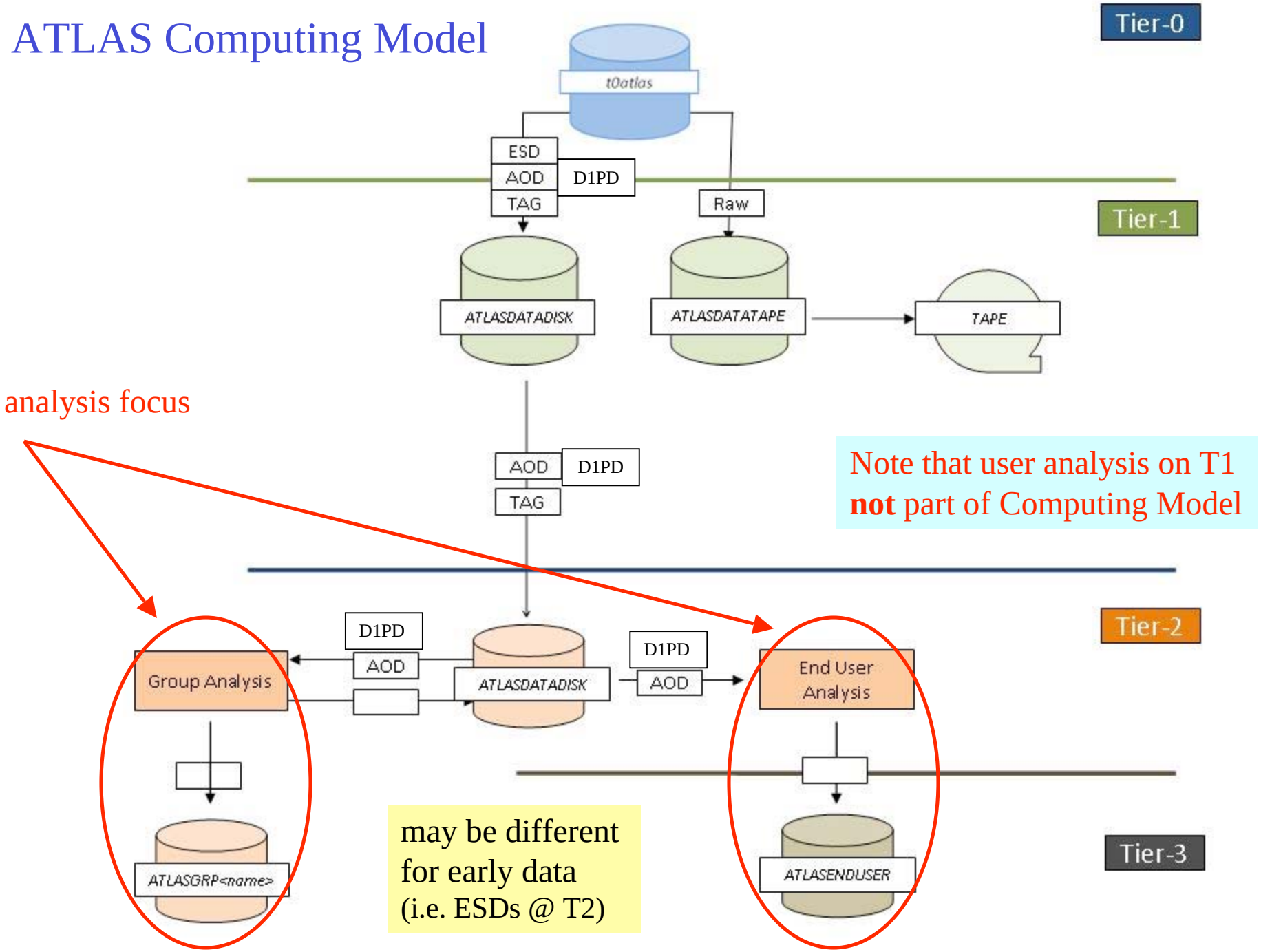
Transforming from one format to another



(of course, transforms typically go from a larger to a smaller format [except for D^2PD])

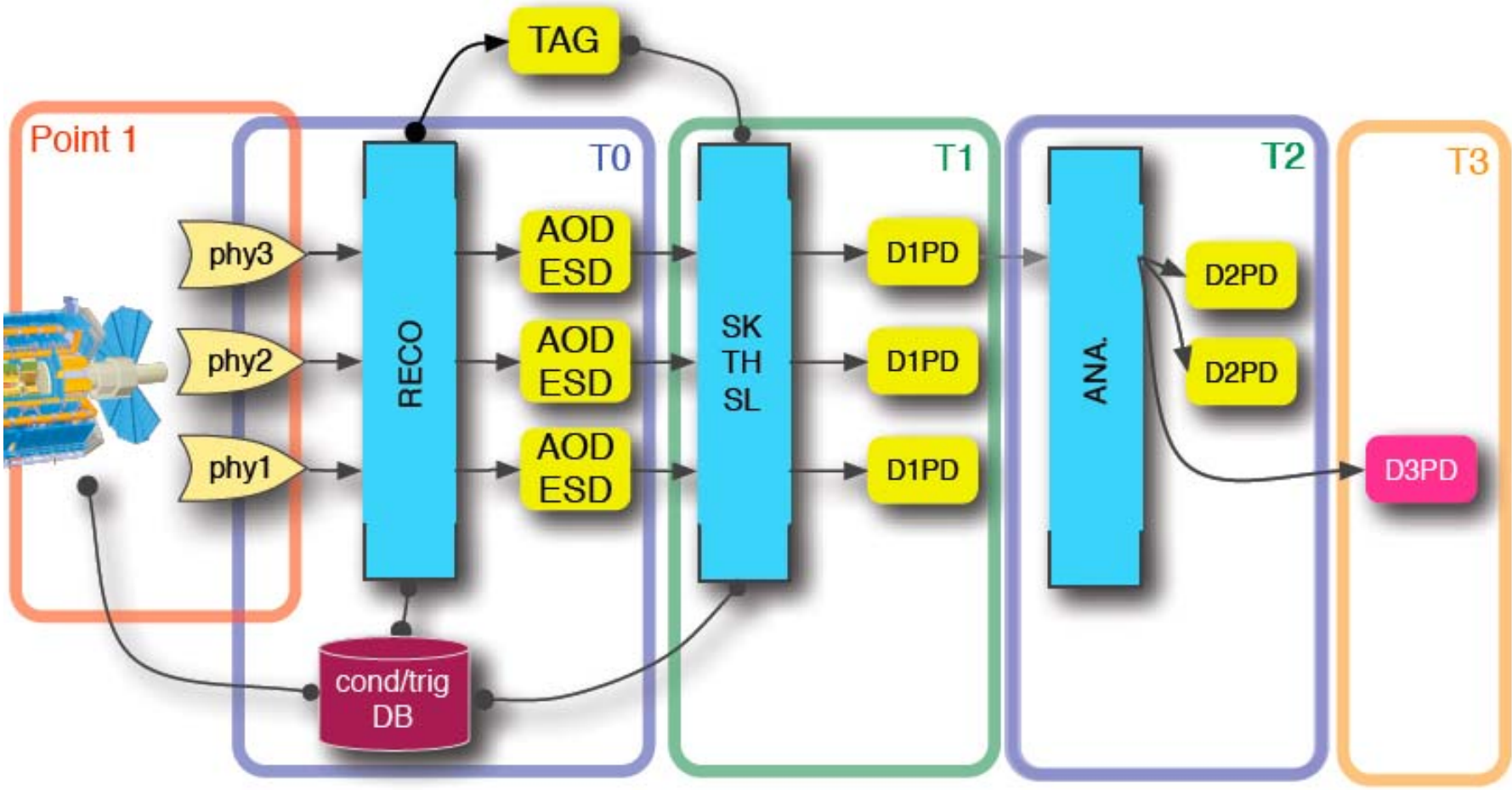

AOD
e
g
PJet
TauJ
Clusters
TrackP
Cells
MET
muon
ESD
eg
Jet
Cmuon
MET
Clusters
TTrack
Cells
Hits
RAW
D1PD
e
g
PJet
TauJ
MET
muon
Clus
TrckP
D2PD
kinematics
e
g
PJet
TauJ
MET
muon
Clus
TrckP
D3PD
g
PJet
TauJ
muon
Clus
TrckP
e
MET
kin

ATLAS Computing Model



One of the many potential reduction chains

(this one considered likely for "mature" phase of experiment)



What's decided and what's not

Decided

Included in LCG pledge: T1: All AOD, 20% ESD, 25% RAW
each T2: 20% AOD (and/or 20% D^1PD ?)

2 copies of AODs (data+MC) are distributed over US T2s (probably same for D^1PDs)

1 copy of ESD (data only) distributed over US T2s (probably only 2009-2010)
(may be able to use perfDPDs in some cases)

D^1PDs initially produced from AODs as part of T0 production, replicated to T1s, T2s
D^1PDs will be remade from AODs as necessary on the T1

Not Decided

Final content of D^1PDs

Streaming strategy for D^1PDs (3 options under consideration - very active area of discussion)

Too early to make decisions about D^2PDs

Expected analysis patterns for early data

Assume bulk of user activity will happen on T2s and T3s
(define user accessible area of T1 as a T3)

Assume final stage of analysis (plots) happens on T3s (T2s are not interactive)

Two primary modes:

(1) user/group runs job on T2s to make reduced dataset (usually D^3PD)
(potential inputs: ESD,AOD,D^1PD)
reduced dataset is then transferred to user's T3 for further analysis

(2) user/group copies input files to user's T3 (potential inputs: ESD,AOD,D^1PD)
On T3 user/group either generates reduced dataset for further analysis or performs final analysis on input data set

Choice depends strongly on capabilities of T3, size of input data sets, etc.

Tier 3 Task force Report

slides shamelessly stolen from Chip Brock's talk at last week's LBL Jamboree

this task force is two things

- A large document

  intentionally written for multiple audiences:

  geeky ATLAS people, sure; ATLAS physicists who are only just contemplating computing at home; technical, non-physicists, and certainly, agency folks

- A set of comments

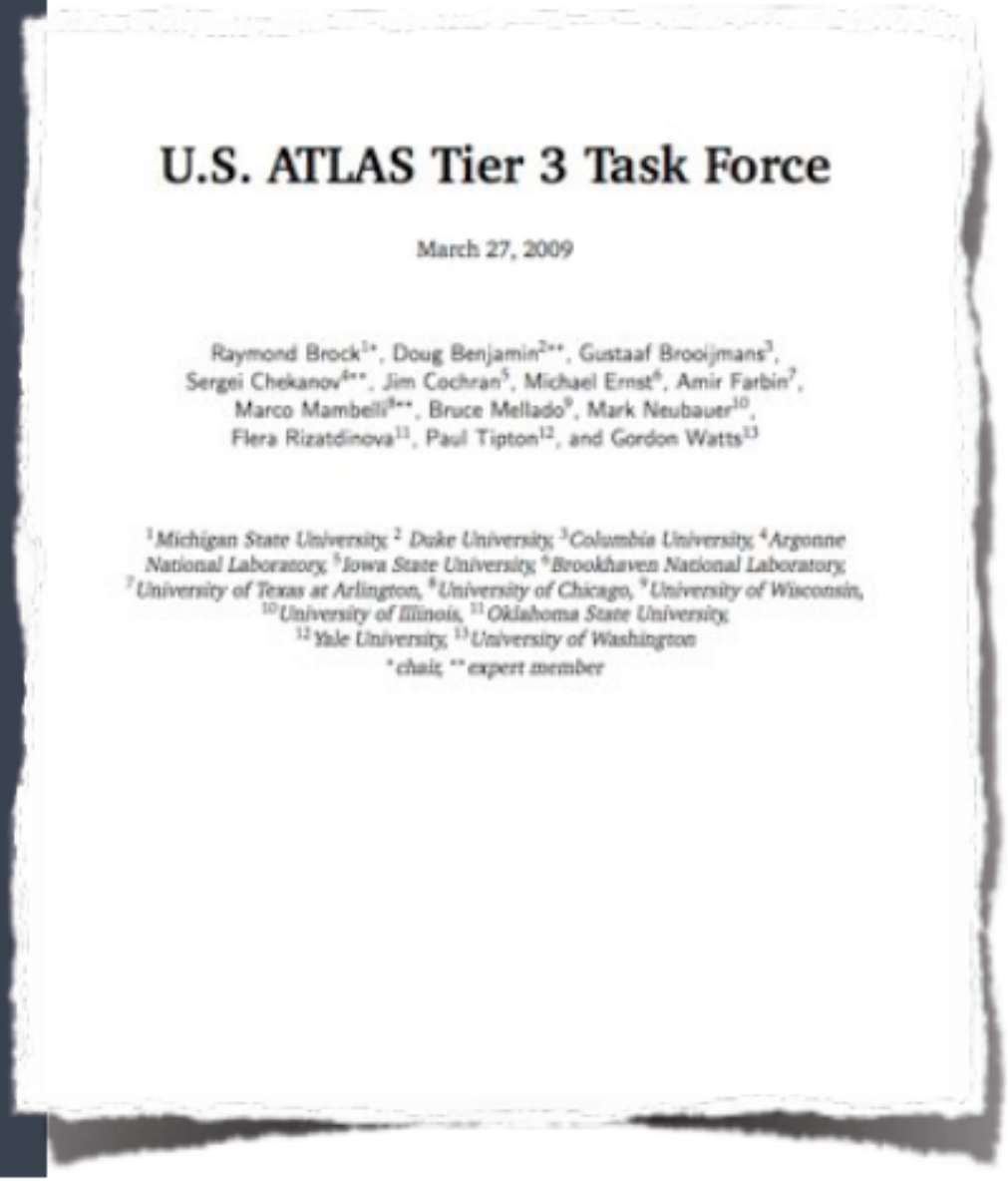
  "observations"

  "recommendations"

the document

meant to be complete:

a reference



U.S. ATLAS Tier 3 Task Force

March 27, 2009

Raymond Brock[1*], Doug Benjamin[2**], Gustaaf Brooijmans[3], Sergei Chekanov[4**], Jim Cochran[5], Michael Ernst[6], Amir Farbin[7], Marco Mambelli[8**], Bruce Mellado[9], Mark Neubauer[10], Flera Rizatdinova[11], Paul Tipton[12], and Gordon Watts[13]

[1] *Michigan State University,* [2] *Duke University,* [3] *Columbia University,* [4] *Argonne National Laboratory,* [5] *Iowa State University,* [6] *Brookhaven National Laboratory,* [7] *University of Texas at Arlington,* [8] *University of Chicago,* [9] *University of Wisconsin,* [10] *University of Illinois,* [11] *Oklahoma State University,* [12] *Yale University,* [13] *University of Washington*

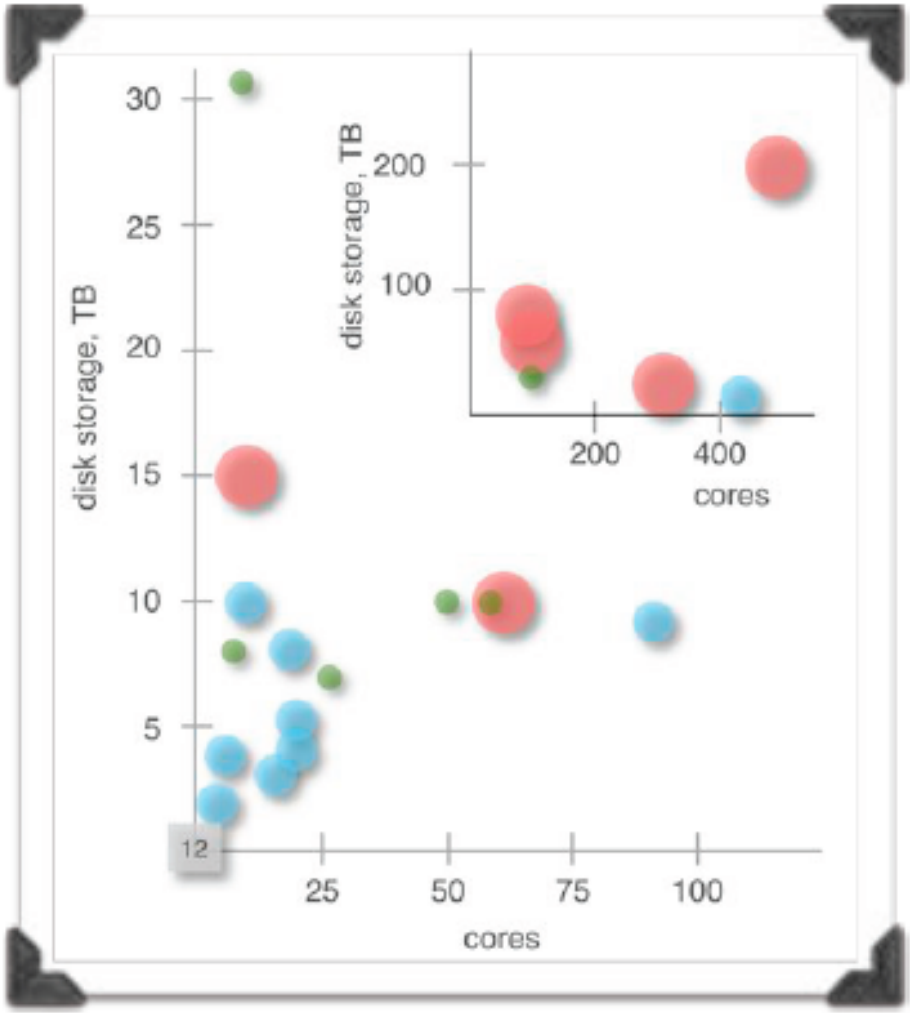
*chair, **expert member

Tier 3s today.

Survey:

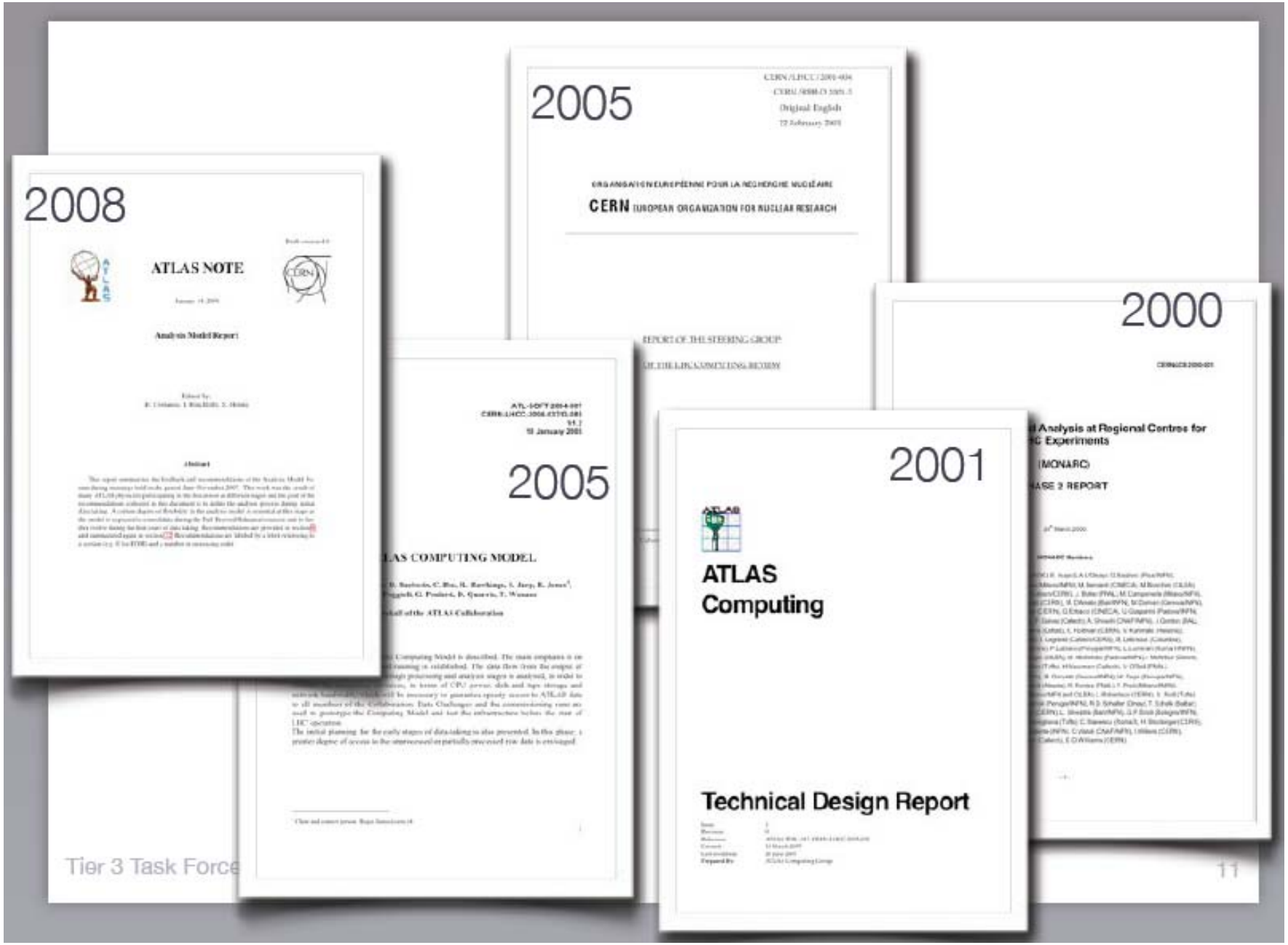
33 ATLAS university institutes

dot size/color: network connectivity:
100Mbps, 1Gbps, 10Gbps



To understand the role of the Tier3s, we first needed to understand the role of the Tier0, Tier 1s, and Tier2s

This lead to a careful review(search?) of existing documentation (often only in talks!) …


2008
ATLAS NOTE
Analysis Model Report
2005
ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH
REPORT OF THE STEERING GROUP
2005
LAS COMPUTING MODEL
2000
Analysis at Regional Centres for
C Experiments
(MONARC)
ASE 2 REPORT
2001
ATLAS
Computing
Technical Design Report
Tier 3 Task Force
11

information is scattered:

Recommendation 9: ATLAS computing and analysis policies, existing resource amounts, targeted resource quantities, data format targets, times for data reduction, etc.: basically all parameters and rules should be in one place. A policy should be considered "official" only when updated at a single twiki page. One repository should define official reality and should be updated when that reality changes. (page 9)

Recommendation 9

What would a task force be without a plea regarding documentation?

tried to identify various workflows

1. Steady State Dataset Distribution
2. Dataset creation
3. Monte Carlo Production
4. “Chaotic” User Analysis (“Chaotic User” Analysis?)
5. Intensive Computing Tasks

1. Steady State Dataset Distribution
2. Dataset creation
3. Monte Carlo Production
4. **"Chaotic" User Analysis**
5. Intensive Computing Tasks

"analysis" is not a single thing

in modern HEP experiments:

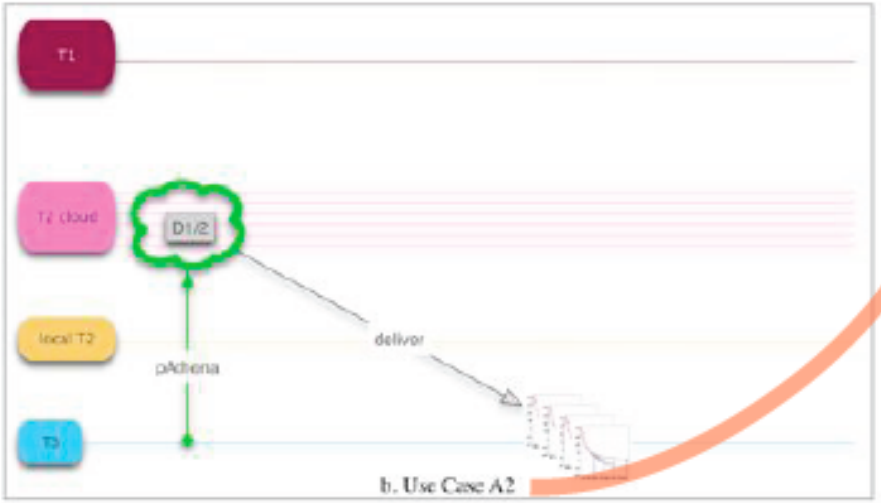
repetitive skimming, selection

human-intensive data-handling

because file transfers fail, networks fail, mistakes are made

Table 11: The Chaotic Analysis Use Cases.

	data in:	data out:	from:	to:	by:	trans:	who:
A1	ESD	hist	T1	T3	T1,T2	SK, AU	analyzer
A2	D^2PD	hist	T2CL	T3	T2CL	SK	analyzer
A3	D^3PD	hist, txt	T3	T3	T3	AU, CH	analyzer
A4	D^3PD	hist, txt	T3	T3	T2CL	AU	analyzer
A5	AOD	hist	T2CL	T3	T2CL	SK	analyzer



N.B.

intensive calculations

Matrix Element calculations

many cpu-centuries of computation

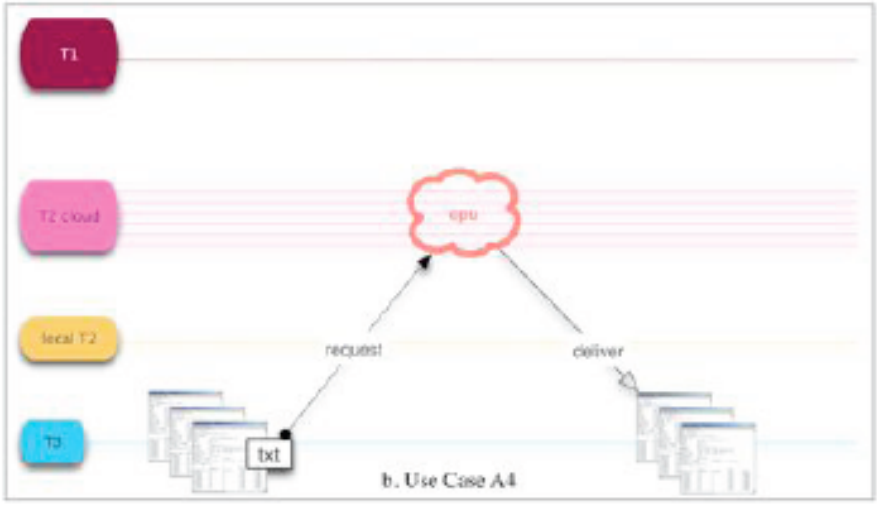
grid has failed DØ for these

Multivariate combinations

COLLIE

Ensemble simulation

~O in  ~O out



this is important:

Nobody had ever dreamed of these sorts of analysis tasks before this century

What kinds of surprises will the ATLAS era see?

history is our only source of data

history=tevatron

- DØ and CDF had to re-invent their computing models many times

- emerging technologies

  made unanticipated, clever analyses possible

- unanticipated, clever analyses

  made extending technologies essential

neither of these are necessarily consistent with tight resource planning

- the world changed many times in the lifetime of the Tevatron

1. *ubiquity of OO coding*
2. *emergence of inexpensive, commodity computer clusters*
3. *availability of distributed disk servers and management systems*
4. *development of high-speed networking and switching technologies*
5. *the Web, from cute to essential*

planning computing is hard

Scientific and Computing administrators

argue for funds against a plan

Scientists–the users–have one thing in mind

and they are often not so great about sticking to a plan

Physics analysis moves faster than plans.

prediction is hard

"I believe OS/2 is destined to be the most important operating system, and possibly program, of all time."

Bill Gates, OS/2 Programmers Guide, November 1987

	1997 projections	2006 actual
Peak (average) data rate (Hz)	50 (20)	100(35)
Events collected	600M/year	1500M/year
Raw Data Size (kB.event)	250	250
Reconstructed Data size(kB/event)	100	80
User format (kB/event)	1	40
Tape Storage	280 TB/year	1.6 PB on tape
Tape reads/writes (weekly)		30 TB/7TB
Analysis/cache disk	7 TB/year	220 TB
Reconstruction time (GHz-s/event)	2.0	50
User analysis times (GHz-s/event)	?	1
User analysis weekly reads	?	3B events
Primary reconstruction farm size (THz)	0.6	2.4 THz
Central analysis farm size (GHz)	0.6	2.2 THz
Remote resources (GHz)	?	~ 2.5THz
	after Run 1	after Run 2a

...the scale of the software development effort for Run II is quite comparable to that of Run I. In Run II the system will again include multiple platforms of at least three currently supported flavors of UNIX and very likely some version of the NT operating system as well by the end of Run II. "Run II Computing and Software Plan for the DØ Experiment," 1997.

flexible and nimble

we have to plan for revolutions

Observation 1 *Challenges to efficient LHC physics analysis are likely to be greater than imagined and so "flexible" and "nimble" should continue to be the guiding principles in the design of computing infrastructure.*

Observation 2 *Physicists often reduce dataset sizes in order to bring as much data, as near to their desktop as is feasible, as often as is required.*

We could argue about whether this is according to the liturgy...but it will happen, one way or the other.

observations

All of this argues for the deepest possible computing architecture.

“analysis”

- is not remote
- it’s interactive...because things don’t always work

Tier 2's are the heroes of ATLAS

- But:

  Are they physicist-innovation-capable?

  Can they really handle the sort of human-intense load that will be likely?

  Will physicists still try to move data near to them?

- Will they be available?

Tier 2 resources

- 50%,

  centrally managed for simulation
- 50%

  for national analyses
- How much full simulation?

  30%→20%→10%

US Pledge to wLCG	2007	2008	2009	2010	2011
CPU (kSI2k)	2,560	4,844	7,337	12,765	18,194
Disk (TB)	1,000	3,136	5,822	11,637	16,509
Tape (TB)	603	1,715	3,277	6,286	9,820

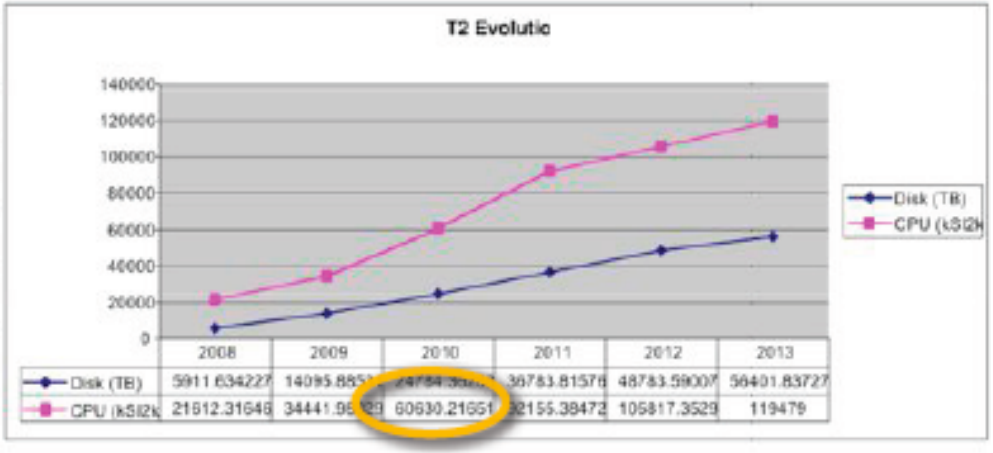
Sample	Generation	Simulation	Digitization	Reconstruction
Minimum Bias	0.0267	551.	19.6	8.06
$t\bar{t}$ Production	0.226	1990	29.1	47.4
Jets	0.0457	2640	29.2	78.4
Photon and jets	0.0431	2850	25.3	44.7
$W^{\pm} \to e^{\pm}\nu_e$	0.0788	1150	23.5	8.07
$W^{\pm} \to \mu^{\pm}\nu_{\mu}$	0.0768	1030	23.1	13.6
Heavy ion	2.08	56,000	267	-

Table 18. in kSI2k-s, without pileup

K. Assamagan, et al., ATLAS Monte Carlo Project, 2009.

Benchmark: $10fb^{-1}$ → 2010 → $2x10^{33}$ → 3.5

quantity	value used	high	low	comments
LHC year	2010	2011	n.a.	assume 2008 start
Ins. $\mathcal{L}$ $cm^{-2}s^{-1}$	2×10^{33}	3.5×10^{33}	10^{33}	Garoby, LHCC 08
annual $\int \mathcal{L}dt$ fb^{-1}	10	?	?	rounded from 12
annual dataset	2×10^9 events	?	?	[7]
sim. time	1990 kSI2K s ($t\bar{t}$)	2950 kSI2K s γj	1030 kSI2K s $W \to \mu$	[16]
dig. time	29.1 kSI2K s ($t\bar{t}$)	29.2 kSI2K s j	23.1kSI2K s $W \to \mu$	[16]
recn. time	47.4 kSI2K s ($t\bar{t}$)	78.4 kSI2K s j	8.07 kSI2K s $W \to e$	[16]
digitization pileup factor	3.5	5.8	2.3	[16]
fraction of full dataset for full sim	0.1	0.2	na.	
factor rel. to full sim. for $t\bar{t}$	0.05 (ATLFAST-II)	0.38 (fG4)	0.004 (ATLFAST-IIF)	[16]
$D^1PD \to D^2PD$	0.5 kSI2K s	?	?	[15]
$D^2PD \to D^3PD$	0.5 kSI2K s	?	?	[15]
disk R/W	100 MBps	200 MBps	10 MBps	S. McKee private
sustained network	50 MBps	100 MBps	10 MBps	S. McKee private
fraction of data in pDPD	20%			
# primary DPD	10			
# subgroups	5			
average CPU	1.4 kSI2K units	2	NA	
total ATLAS Tier 2 computing	60.63MSI2k			[11]



modeled it.

Amir Farbin...heroic calculation

Tier 2 simulation for one year

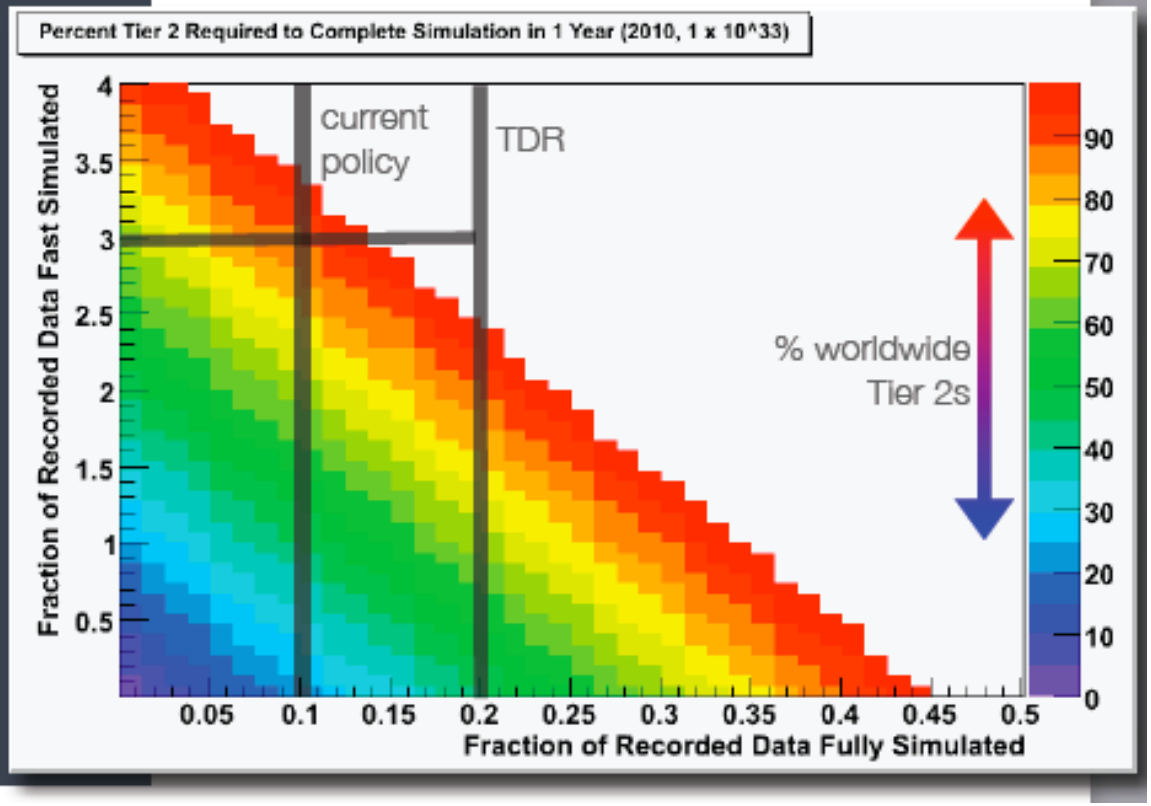
- horizontal axis:

  fraction fully simulated

- vertical axis:

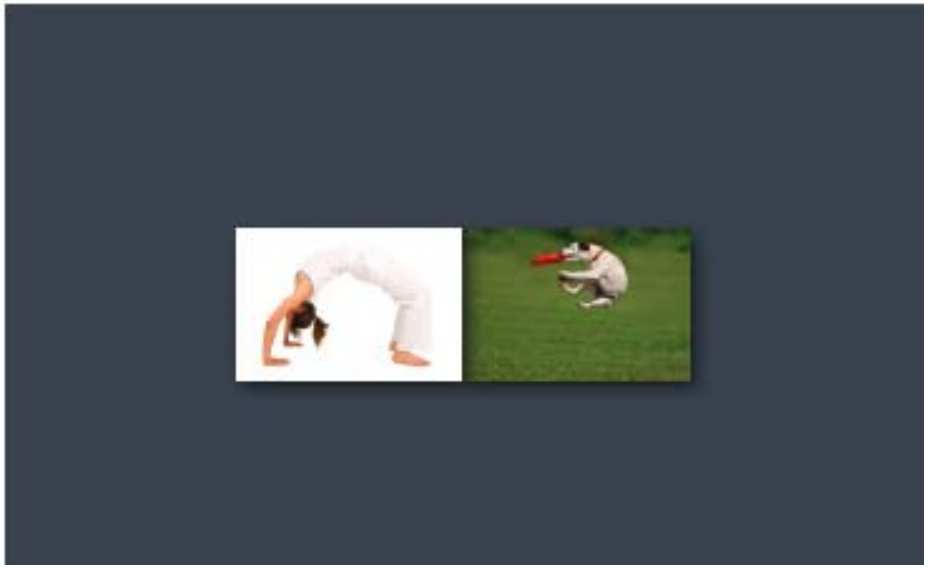
  fraction fast-simulated

  (ATLFAST-II...from Assamagan)



Observation 4 *The Tier 2 systems' responsibilities are tremendously significant. Should we discover an underestimate in CPU, storage, or network needs of ATLAS as a whole, the analysis needs of U.S. university physics community will be adversely affected.*

Observation 5 *Is there any reason to think that the first 20 years of the ATLAS computing experience will be any less astonishing? Is it wise to design tightly to current expectations, as if the future will be a continuous extrapolation of the present? If history is at all a reliable guide, it argues for the most flexible, most modular, and least rigidly structured systems consistent with 2008 technology and budgets.*



recommendations

5 Primary Recommendations

Minimum necessary requirements

Recommendation 1: With past history as a guide and with prudent concern for the challenge and uncertainties of ATLAS analysis, the *structured* U.S. ATLAS computing infrastructure should be deeper than the Tier 2 centers. A flexible and nimble infrastructure would include strategically extending some data production, Monte Carlo simulation, and analysis into the U.S. ATLAS Tier 3 sector. (page 70)

Recommendation 1

Recommendation 2: The strategy for building a flexible U.S. ATLAS Tier 3 system should be built around a mix of 4 possible Tier 3 architectures: T3gs, T3g, T3w, and T3af. Each is based on a separate architecture and each would correspond to a group's infrastructure capabilities. Each leverages specific analysis advantages and/or potential ATLAS-wide failover recovery. They are specifically defined in Section 7.1.2. (page 72)

Recommendation 2

4 Specific classes of Tier 3s

a vocabulary, a set of identifiable targets for groups' evolution

The "Tier 3 Quartet"

1. "**T3gs**": a center with full **g**rid **s**ervices
   likely a significant center with infrastructure in place
   local resource control, but production-capable - T2 failover capability
2. "**T3g**": a cluster with **g**rid connectivity
   "tower cluster", no cooling/power infrastructure (ANL Model)
   or a rack-based model (Duke Model)
3. "**T3w**": individual, personal **w**orkstations
   RootTuple analyses, grid submission
4. "**T3af**": within the confines of a an **a**nalysis **f**acility
   like the "CDF model" at Fermilab: fair-share computing in exchange for contribution

1. **"T3gs"**
2. "T3g"
3. "T3w"
4. "T3af"

Tier 3 with "**g**rid **s**ervices"

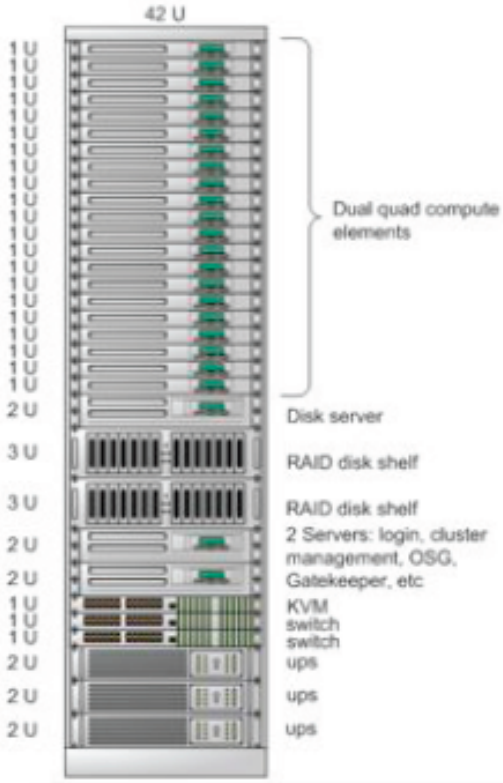
a campus-based, significant cluster

requiring AC/power infrastructure

Characterized a strawman

~$80k

University of Illinois building one



168 processors
>250kSI2k

24TB usable

component	typical model	quantity	unit cost, k$
UPS	DELL	3	1.0
switch	DELL PowerConnect 48GbE, portmanaged	2	1.5
servers	DELL PE2950 E5440 processor, 2.83GHz, 32GB RAM, 250GB drive	3	4.2
compute elements	DELL PE1950 E5440 processor, 2.83GHz, 16GB RAM, 250GB drive	21	2.4
storage elements	DELL MD1000	2 (24TB, usable)	5.4
KVM	Belkin	1	1.3
rack			1
total cost			$82.1k

T3gs use cases, enhanced

- Production: Physics Group D2PD from cached D1PD

  assume a full stream

  few days to produce

- Monte Carlo Production: in support of a physics group

  ttbar–sample appropriate to the 10fb benchmark

  sample-sized, signal + background, ATLFAST-II

  few days

1. "T3gs"
2. **"T3g"**
3. "T3w"
4. "T3af"

Tier 3 with "**g**rid" connectivity

a campus-based,

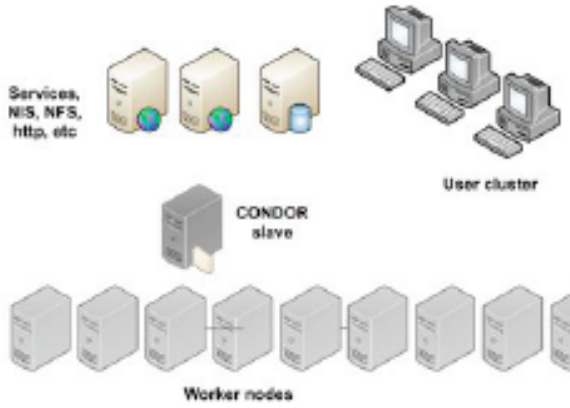
tower or rack-based cluster

minimal services required

Characterized a strawman

~$25k

ANL and Duke are building them



80 processors
>100SI2k

20TB

component	typical model	quantity	unit cost, k$
switch	Cisco 1GB	1	2.5
worker towers	Intel-based E5410 2.33GHz, 2 TB storage 8GB RAM	10	2.0
server elements	DELL PE1950 E5440 processor, 2.83MHz, 16GB RAM, 250GB drive	4	0.5
total cost			$24.5k

the data

- In a world where even roottuples will be TB's

  access to the data is crucial at a Tier 3gs and T3g

Recommendation 3: In order to support a Tier 3 subscription service, without a significant support load or the need to expose itself to the ATLAS data catalog, a particular DQ2 relationship must be established with a named Tier 2 center, or some site which can support the DQ2 site services on its behalf. This breaks the "ubiquity" of Tier 2s — here, a particular Tier 3 would have a particular relationship with a named Tier 2. This dual-capability (limited exposure of a site's file catalog and a subscription-like functionality) has been colloquially referred to as "outsourcing" DQ2 site services.

Recommendation 3

must be able to subscribe to large datasets

cannot move TBs by hand...

Recommendation 4: U.S. ATLAS should establish a U.S. ATLAS Tier 3 Professional, a system administration staff position tasked to 1) assist in person the creation of any Tier 3 system; 2) act as a named on-call resource for local administrators; and 3) to lead and moderate an active, mutually supportive user group. (page 85)

Recommendation 4

Support is a serious issue for many

but worth the investment if it makes T3g's possible

Recommendation 5: In order to qualify for the above U.S. ATLAS Tier 3 support, U.S. ATLAS Tier 3 institutions must agree to 1) supply a named individual responsible on campus for their system and 2) adhere to a minimal set of software and hardware requirements as determined by the U.S. ATLAS Tier 3 Professional. (page 85)

Recommendation 5

quid pro quo

to keep the support personnel sane

2 Technical Recommendations

Service modifications to Panda

Focus on point-to-point communications

Recommendation 6: We recommend that the recent addition of `pAthena` local control-functionality be maintained, and possibly extended to allow for more convenient control and access/monitoring of the Tier 3 site configuration by local administrators. (page 87)

Recommendation 6

With a switch - same interface for local and T1/2 pAthena services

Recommendation 7: Sustained bandwidth of approximately 20MBps is probably required for moving TB sized files between Tier 2 and Tier 3 locations and it should be the goal that every campus or lab group establish such capability within a few years. This requires a high level of cooperation and planning among U.S. ATLAS computing, national network administrators, and campus administrators. Note: it might be useful and prudent to tune bandwidth between *particular* Tier 3 locations and *particular* Tier 2 centers rather than to set a national standard which might be difficult to meet. Note that the Resource Allocation Committee will have authority over the large-scale movement of data and any large scale caching of Tier 3 generated files into the Tier 1 or Tier 2 clouds.

Recommendation 7

Rough goal:

*1-2TB transfers **point-to-point** in a ~day*

EPISODIC!

Partnership recommendation

Recommendation 8: Enhancement of U.S. ATLAS institutions' Tier 3 capabilities is essential and should be built around the short and long-term analysis strategies of each U.S. group. This enhancement should be proposal-based and target specific goals. In order to leverage local support, we recommend that U.S. ATLAS leadership create a named partnership or collaborative program for universities which undertake to match contributions with NSF and DOE toward identifiable U.S. ATLAS computing on their campuses. Public recognition of this collaboration should express U.S. ATLAS's gratitude for their administration's support and offer occasional educational and informational opportunities for university administrative partners such as annual meetings, mailings, video conferences, hosted CERN visits, and so on. (page 86)

Recommendation 8

Involve universities in a public fashion

more depth will enhance

- Tevatron experience suggests:

  "planning" is a process–the ground shifts

  "analysis" is a highly-interactive activity "above" flattened roottuples

  physicists' innovation is a critical scientific and competitive advantage

- We have tried to indicate that

  the "analysis fraction" of Tier 2 resources may be in some jeopardy

The Tier 3 quartet:

- Could leverage fail-over production and MC contributions

  for targeted physicists' tasks

  allow university groups opportunities for important, local responsibilities

- Would create a common worldview in US ATLAS

  a common vocabulary and glossary: "T3gs" "T3g" "T3w" T3af"

  all stakeholders would know what each implies

  an understood, manageable procurement strategy

Three critical issues

- Support model

  personal, regular, common

- Access to the data for 2011-2012 milestones

  target point-to-point minimal connectivity

  40 institutions...that's probably 40 different evaluations

- DQ2 flexibility

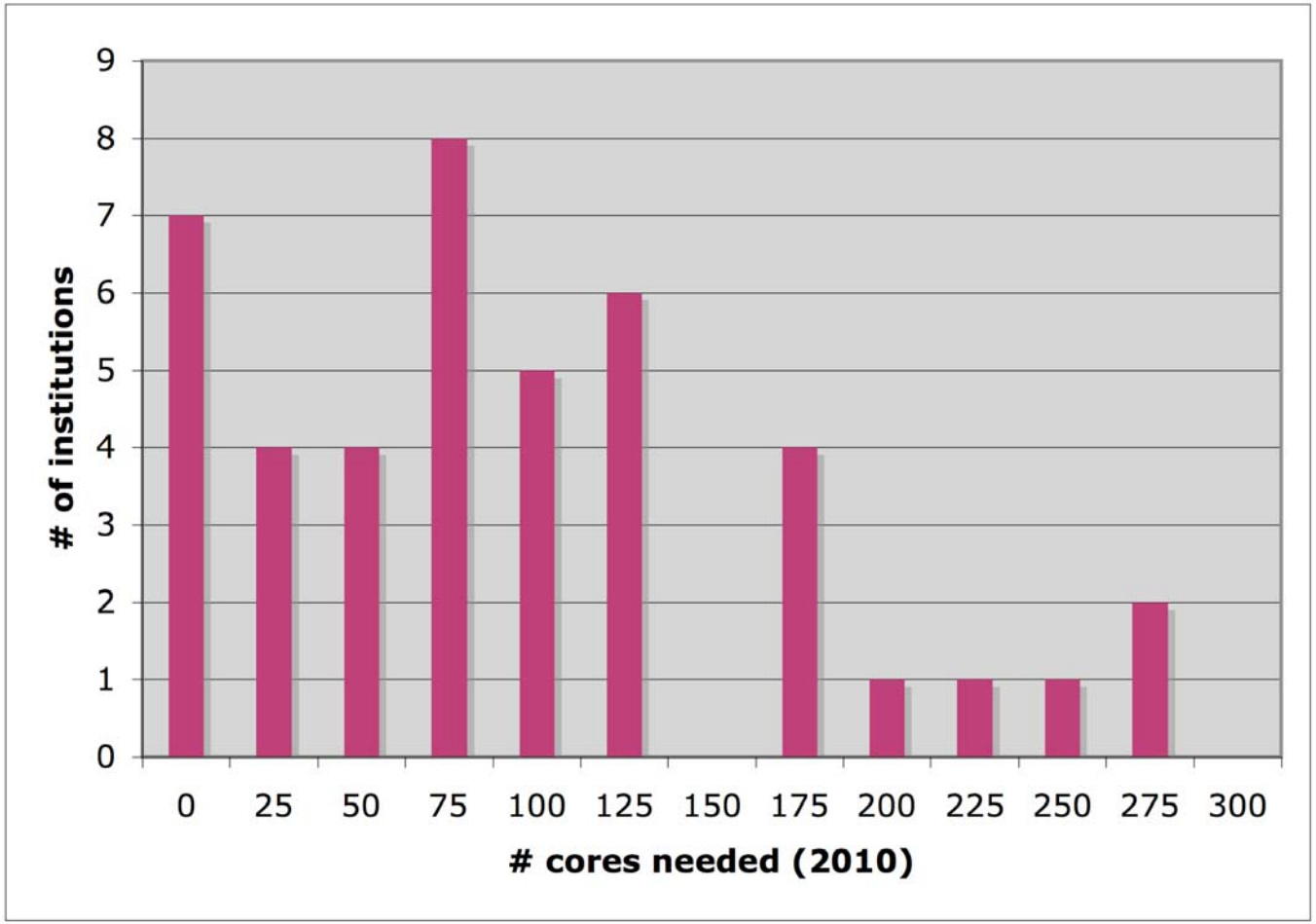
  called now "outsourcing" DQ2 to some Tier 2 or Tier 3

  for catalog support and data subscription

Some results from the US requirements study

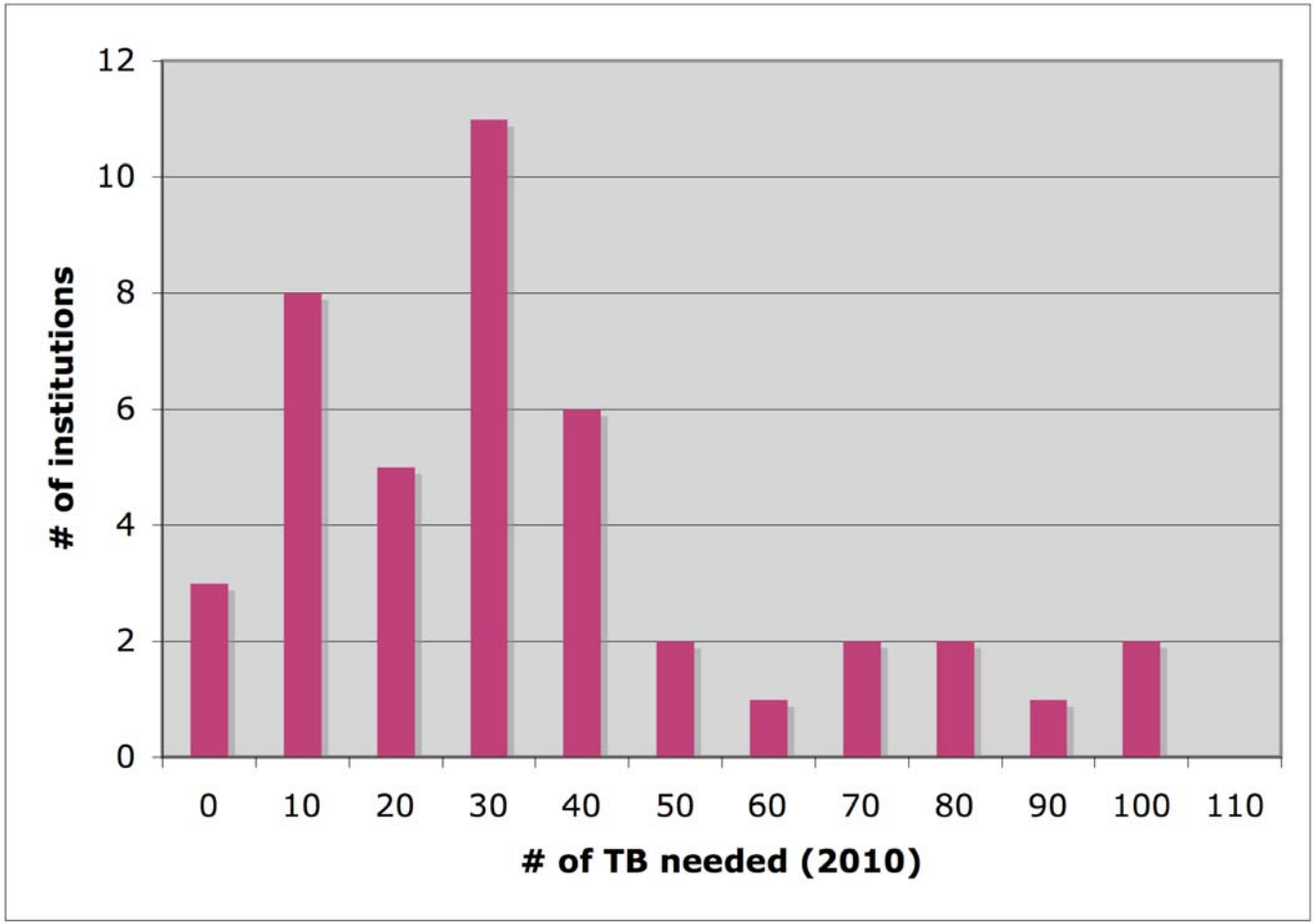
Characterization of Existing T3s: # of cores needed (2010)

(not including Harvard)



Characterization of Existing T3s: # of TB needed (2010)

(not including Harvard)



Scale of "missing" T3 resources

Summing "missing" T3 resources (for both institutions with & without existing T3 hardware):

	# cores	TB
2009	279	101
2010	1590	572

Assuming 2 kSI2k/core

This sets the scale for a potential "single" US analysis facility (T3af)

Or, if we sort by geographic region:

	# cores 2009	# cores 2010	TB 2009	TB 2010
Western	69	454	25	163
Midwestern	23	176	8	63
Eastern	187	960	67	346

Summary of Requirements Study

Assuming chosen model is some reflection of reality:

- groups/users **must** cooperate (common DPDmakers, DPDs) to make effective use of T2s (this seems to be especially true wrt T2 disk space)

- not clear how long we will be able to keep the full ESDs on T2 disk (expect this to be crucially important to early running)

- we need more focus on final stage analysis (T3s)

- T3 support will be an issue

- most existing T3s have insufficient resources for planned activities

- sites without existing T3 may consider contributing to T3af, joining with an existing T3, or building their own new T3

Proposed plan for T3 integration (in US)

From USATLAS meeting @ OSG All Hands meeting (March 3, 2009)

3 steps/phases:

1 - bring all (small) T3 sites to the minimal level to do physics [T3w]
(dq2 client - no subscriptions)

2 - incorporate subscriptions for those T3 sites that want [T3g]
(& can support) them - many options under investigation

3 - incorporate access to conditions db for those T3 sites that want
(and can support) it [T3g]

Not clear which of “2” or “3” will be solved first

Plan for step 1

Further poll existing T3s wrt software, network (supply utility), plans

Prepare “T3 HOWTO” following ANL T3 model:

Doug will prepare T3w HOWTO (will be available by end of April)

Create a validation procedure

Form T3 technical task force to aid in support (coordinated by Doug) - should include expertise from ANL

Pass T3w HOWTO on to T3 community (Twiki)

To insulate T1/T2 system support, T3s encouraged to help each other (USATLAS hypernews: “Tier3 support”, Twiki, biweekly meetings)
Doug and jc will represent T3 community to T1/T2 management

Plan for step 1 (continued)

T3w development will proceed in 2 steps:

(1) Interactive T3w (minimal needed to do physics)

(2) Interactive + local batch T3w

May 18-19: T3 meeting at ANL

Plan for visits by Doug to sites (as needed)