

Tier 3 Task Force Summary

Tier 2/Tier 3 Meeting @ LIGO
March 3, 2009

Chip Brock, Michigan State University

Gustaaf Brooijmans, Columbia,

Sergei Chekanov, Argonne National Laboratory

Jim Cochran, Iowa State University,

Michael Ernst, Brookhaven National Laboratory,

Amir Farbin, University of Texas at Arlington,

Marco Mambelli, University of Chicago

Bruce Melado, University of Wisconsin,

Mark Neubauer, University of Illinois,

Flera Rizatdinova, Oklahoma State University,

Paul Tipton, Yale University,

Gordon Watts, University of Washington,

Chip Brock, Michigan State University

this task force is two things

- ▶ A document
- ▶ A set of comments
 - “observations”
 - “recommendations”

today

- ▶ I will try to be responsive to the Charge, hitting highlights

- ▶ Technical discussions will follow:

Amir Farbin: Modeling the T2/T3 system

Sergei Chekanov: ANL's T3 and a "template"

Doug Benjamin: Duke's creation of a T3

Mark Neubauer: Illinois' creation of a T3 (video)

Jim Shank: What's next?

the document

meant to be complete:
a reference

Tier 3 Task Force, 3/3/09

U.S. ATLAS Tier 3 Task Force

DRAFT 5.5

February 26, 2009

Raymond Brock^{1*}, Gustaaf Brooijmans², Sergei Chekanov^{3**},
Jim Cochran⁴, Michael Ernst⁵, Amir Farbin⁶, Marco Mambelli^{7**},
Bruce Mellado⁸, Mark Neubauer⁹, Flera Rizatdinova¹⁰,
Paul Tipton¹¹, and Gordon Watts¹²

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

charge: 1. Use Cases.

- ▶ Typical workflows for physicists analyzing ATLAS data from their home institutions should be enumerated. This needs to be inclusive, but not in excruciating detailed.
- ▶ It should be defined from within the ATLAS computing/analysis models, the existing sets of T2 centers, and their expected evolutions.
- ▶ ~~If there are particular requirements in early running, related to detector commissioning and/or special low-luminosity considerations, this should be noted.~~
- ▶ ~~If particular ATLAS institutions have subsystem responsibilities not covered by the existing T1/2 deployment, this should be noted.~~

Tier 3 Task Force, 10/3/09 ▶ Is the previous whitepaper relevant?

charge: 2. Characterization of generic T3 configurations.

- ▶ Some T3's may be very significant because of special infrastructure availabilities and some T3's maybe relatively modest.
- ▶ Is there only 1 kind of T3 center, or are their possible functional distinctions which might characterize roles for some T3's that might not be necessary for others?
- ▶ Description of "classes" of T3 centers, if relevant, should be made.
- ▶ Support needs and suggestions for possible support models should be considered.

charge: 3. Funding.

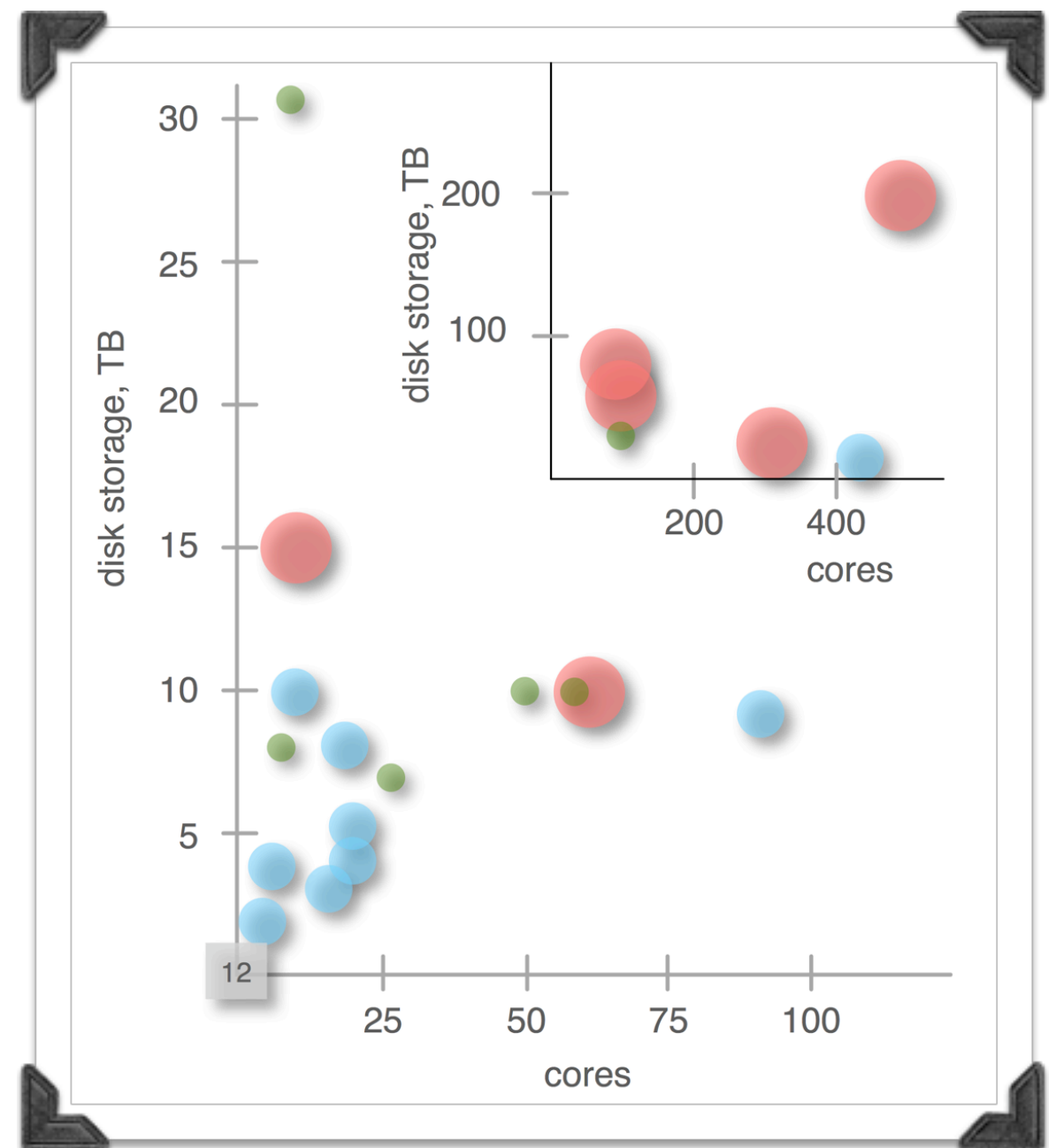
- ▶ This is not part of the US ATLAS Operations budget, so funding must come out of the institutes through core funding or local sources. We would like to make it easier for institutes to secure funding for ATLAS computing--this can only happen if it fits in the DOE and NSF budgets (precedent: the amount of funding groups got for computing equipment in Tevatron experiments) and it must fit in the overall US ATLAS model.
- ▶ For the latter, we have to make the case that the existing T1/2 centers are not enough.
- ▶ Perhaps a recommendation can be justified for an estimated \$ amount needed for a viable Tier 3 cluster -- something like $X + n \cdot Y$ \$'s where n = number of active physicists.

What's a Tier 3 now?

Tier 3s today.

Survey:

all but 2 ATLAS institutes



INSTITUTION	Tufts	LBNL	UT Dallas	U. Wisconsin-Madison	UTA	U. Mass-Amherst	U of Michigan
-------------	-------	------	-----------	----------------------	-----	-----------------	---------------

Do you have T3 cluster (yes/no)	yes, shared with University	yes	yes	yes			
FTE to serve the T3	2.5	1	0.3	1			
HARDWARE:							
Number of computers in the T3 cluster (worker nodes/file servers)	40/1/1		1 gateway, 19 workers	12/5/20			

INSTITUTION	ANL	Columbia	Duke
Number of computers in the T3 cluster (worker nodes/file servers)			2 Gbe to campus net to department switch
SOFTWARE:			
Is your T3 cluster in the GRID? (yes/no)	no, but condor is used for	no	yes

SI2K total units	loc
Disk storage (TB)	no pro
Tape storage (TB)	no pro
Network connectivity	

Cluster Monitoring system (for ex. Ganglia)	no pro
Which method has been used to install the cluster? (PXE, OSCAR,...)	no pro
OTHER:	
any known future purchases	will to 1 with

INSTITUTION	U. of South Carolina
Do you have T3 cluster (yes/no)	not official
FTE to serve the T3	0.05
HARDWARE:	
Number of computers in the T3 cluster (worker nodes/server nodes/file servers)	2,3,1
Number and type of CPU	6 (4-Intel Xeon 2.66GHz Intel Xeon X5 3GHz)
SI2K total units	24k
Disk storage (TB)	4
Network connectivity	1 Gb 100 Mbps net

INSTITUTION	U. of South Carolina	Indiana U	University of Chicago	SMU	OU	Illinois	MSU
				150 Mb(Internet 2)		ICCN Esnet, 1 GigE to servers	campus network + spare capacity of optical network
SOFTWARE:							
Is your T3 cluster in the GRID? (yes/no)	no	yes	no, but have gridftp providing DQ2 endpoint	yes	yes	yes	Yes - OSG
Cluster Monitoring system (for ex. Ganglia)	no	Ganglia	Ganglia, Nagios	Nagios	Ganglia	Ganglia, Grafia, MonaLisa	Ganglia
Which method has been used to install the cluster? (PXE, OSCAR,...)		Rocks	"Cloner" - from ACT	RedHat Kickstart	RHEL5.2 CD	Pacman	Rocks

INSTITUTION	Tufts	LBNL	UT Dallas	U. Wisconsin-Madison	UTA	U. Mass-Amherst	U of Michigan
	Gigabit Ethernet to campus network	Tier 1	nodes, Internet2				
SOFTWARE:							
Is your T3 cluster in the GRID? (yes/no)	No	Yes	Yes	Yes	Yes	No	yes, co-hosted w AGLT2
Cluster Monitoring system (for ex. Ganglia)		Yes	?	Ganglia	Ganglia	Ganglia	Ganglia/Cacti/IT Assistant
Which method has been used to install the cluster? (PXE, OSCAR,...)	RedHat 5.2 with LSF queues		?	PXE	Rocks	manual	PXE
OTHER:							
any known future purchases	8 TB additional storage server	Intend to double the T3 in the next 6 months		no	no	will add 10 dual nodes	next FY small increment

ATLAS

computing/analysis model(s)

2008



ATLAS NOTE

January 14, 2008

Analysis Model Report

Edited by:
D. Costanzo, I. Hinchliffe, S. Menke

Abstract

This report summarizes the feedback and recommendations of the Analysis Model Forum during meetings held in the period June-November 2007. This work was the result of many ATLAS physicists participating in the discussion at different stages and the goal of the recommendations collected in this document is to define the analysis process during initial data taking. A certain degree of flexibility in the analysis model is essential at this stage as the model is expected to consolidate during the Full Dressed Rehearsal exercise and to further evolve during the first years of data taking. Recommendations are provided in section 6 and summarized again in section 7. Recommendations are labeled by a letter refereeing to a section (e.g. E for EDM) and a number in increasing order.

Draft version 4.0



2005

CERN/LHCC/2001-004
CERN/RRB-D 2001-3
Original: English
22 February 2001

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

REPORT OF THE STEERING GROUP^a
OF THE LHC COMPUTING REVIEW

ATL-SOFT-2004-007
CERN-LHCC-2004-037/G-085
V1.2
10 January 2005

2005

ATLAS COMPUTING MODEL

s, D. Barberis, C. Bee, R. Hawking, S. Jarp, R. Jones¹,
Poggioli, G. Poulard, D. Quarrie, T. Wenaus

on behalf of the ATLAS Collaboration

The ATLAS Computing Model is described. The main emphasis is on the initial running of the experiment. The data flow from the output of the detector through processing and analysis stages is analysed, in order to estimate the computing resources, in terms of CPU power, disk and tape storage and network bandwidth, which will be necessary to guarantee speedy access to ATLAS data to all members of the Collaboration. Data Challenges and the commissioning runs are used to prototype the Computing Model and test the infrastructure before the start of LHC operation. The initial planning for the early stages of data-taking is also presented. In this phase, a greater degree of access to the unprocessed or partially processed raw data is envisaged.

¹ Chair and contact person: Roger.Jones@cern.ch

Tier 3 Task Force

2001



ATLAS Computing

Technical Design Report

Issue: 1
Revision: 0
Reference: ATLAS TDR-017, CERN-LHCC-2005-022
Created: 18 March 2005
Last modified: 20 June 2005
Prepared By: ATLAS Computing Group

2000

CERN/LCB 2000-001

Analysis at Regional Centres for LHC Experiments

(MONARC)

PHASE 2 REPORT

24th March 2000

MONARC Members

(KEK), E. Auge (LAL/Orsay), G. Bagliesi (Pisa/INFN),
B. Berti (Milano/INFN), M. Bernabini (CINECA), M. Boschini (CILEA),
J. Caltech/CERN), J. Butler (FNAL), M. Campanella (Milano/INFN),
M. D'Amato (Bari/INFN), M. Dameri (Genova/INFN),
G. Erbacci (CINECA), U. Gasparini (Padova/INFN),
J. P. Galvez (Caltech), A. Ghiselli (CNAF/INFN), J. Gordon (RAL),
K. Holtman (CERN), V. Karimäki (Helsinki),
I. Legrand (Caltech/CERN), M. Lettichouk (Columbia),
P. Lubrano (Perugia/INFN), L. Luminari (Roma1/INFN),
M. Michelotto (Padova/INFN), I. McArthur (Oxford),
H. Newman (Caltech), V. O'Dell (FNAL),
B. Osculati (Genova/INFN), M. Pepe (Perugia/INFN),
R. Pordes (FNAL), F. Pretz (Milano/INFN),
L. Robertson (CERN), S. Rolli (Tufts),
R. D. Schaffer (Orsay), T. Schalk (BaBar),
L. Silvestris (Bari/INFN), G.P. Sirola (Bologna/INFN),
C. Stanescu (Roma3), H. Stockinger (CERN),
C. Vistoli (CNAF/INFN), I. Willers (CERN),
D.O. Williams (CERN).



12

information is scattered:



information is scattered:



 category | view: ATLAS Meeting | focus on: -- all days -- | -- all sessions -- | details: contribution | manage |     LOCAL: Europe/Zurich  login




ATLAS Week (Where Important Stuff Happens)

from **Monday 01 December 2008 (10:30)**
to **Friday 05 December 2008 (12:20)**
Europe/Zurich
at **CERN (Main Auditorium)**
support: martine.desnyder-ivesdal@cern.ch

[Monday 01 December 2008](#) | [Tuesday 02 December 2008](#) | [Wednesday 03 December 2008](#) | [Thursday 04 December 2008](#) | [Friday 05 December 2008](#) |

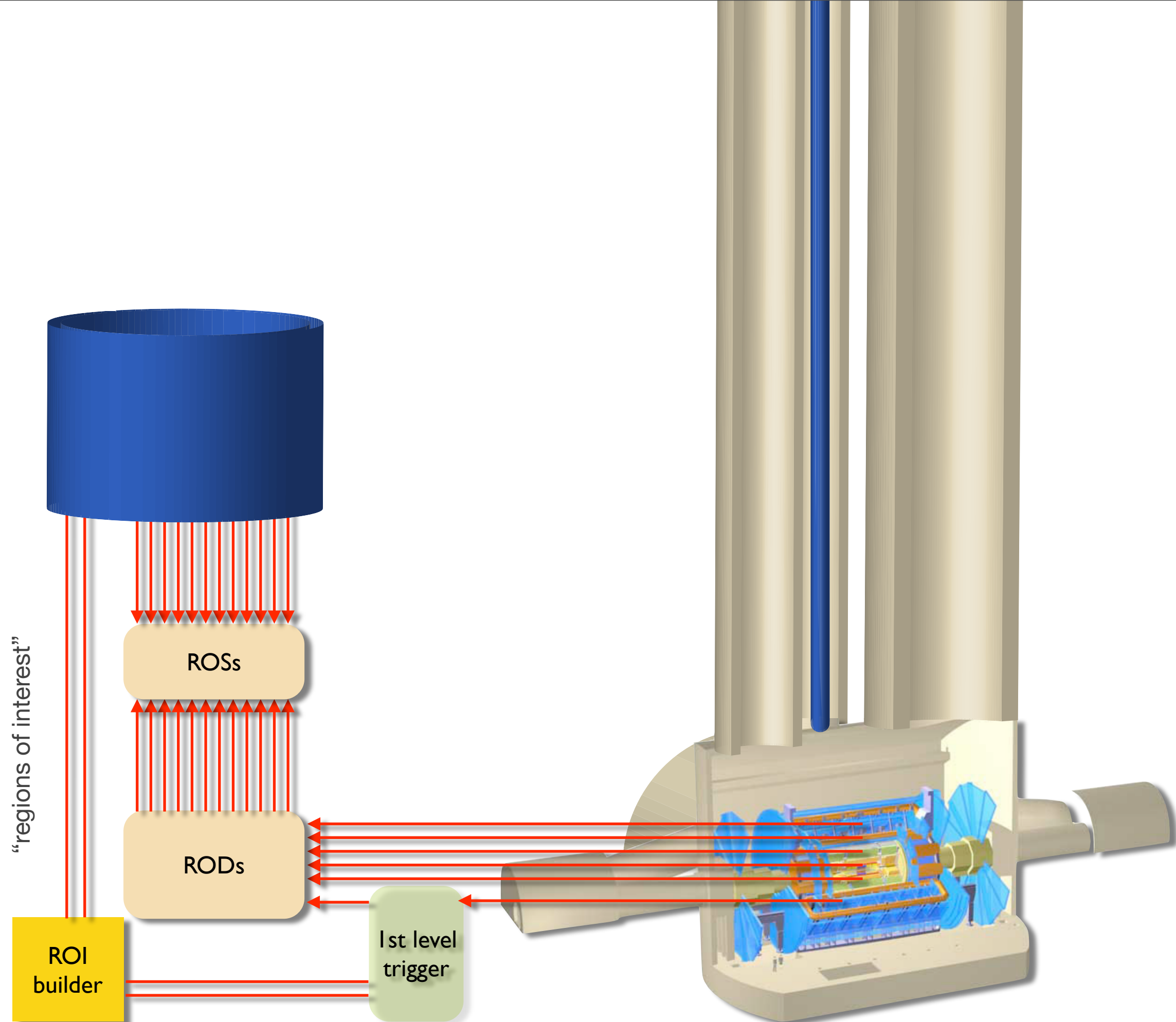
Monday 01 December 2005, 2006, 2007, 2008, 2009 [top↑](#)

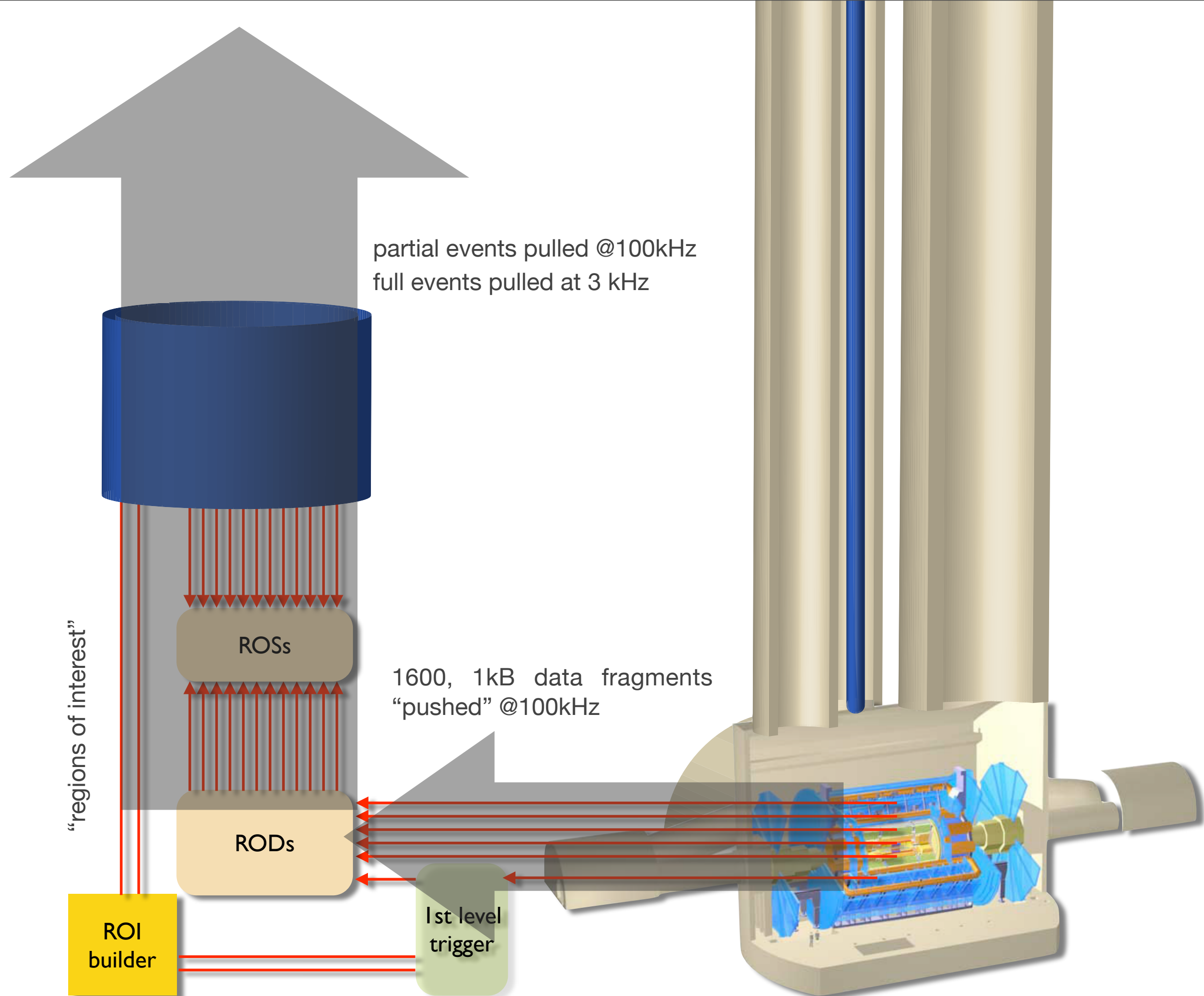
09:00->19:00 Analysis or Computing Model, Policies, and things that might have changed
09:00 Important Computing Slides You'll Want to Treasure... (4h00)  agenda) ([40-4-C01](#))

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

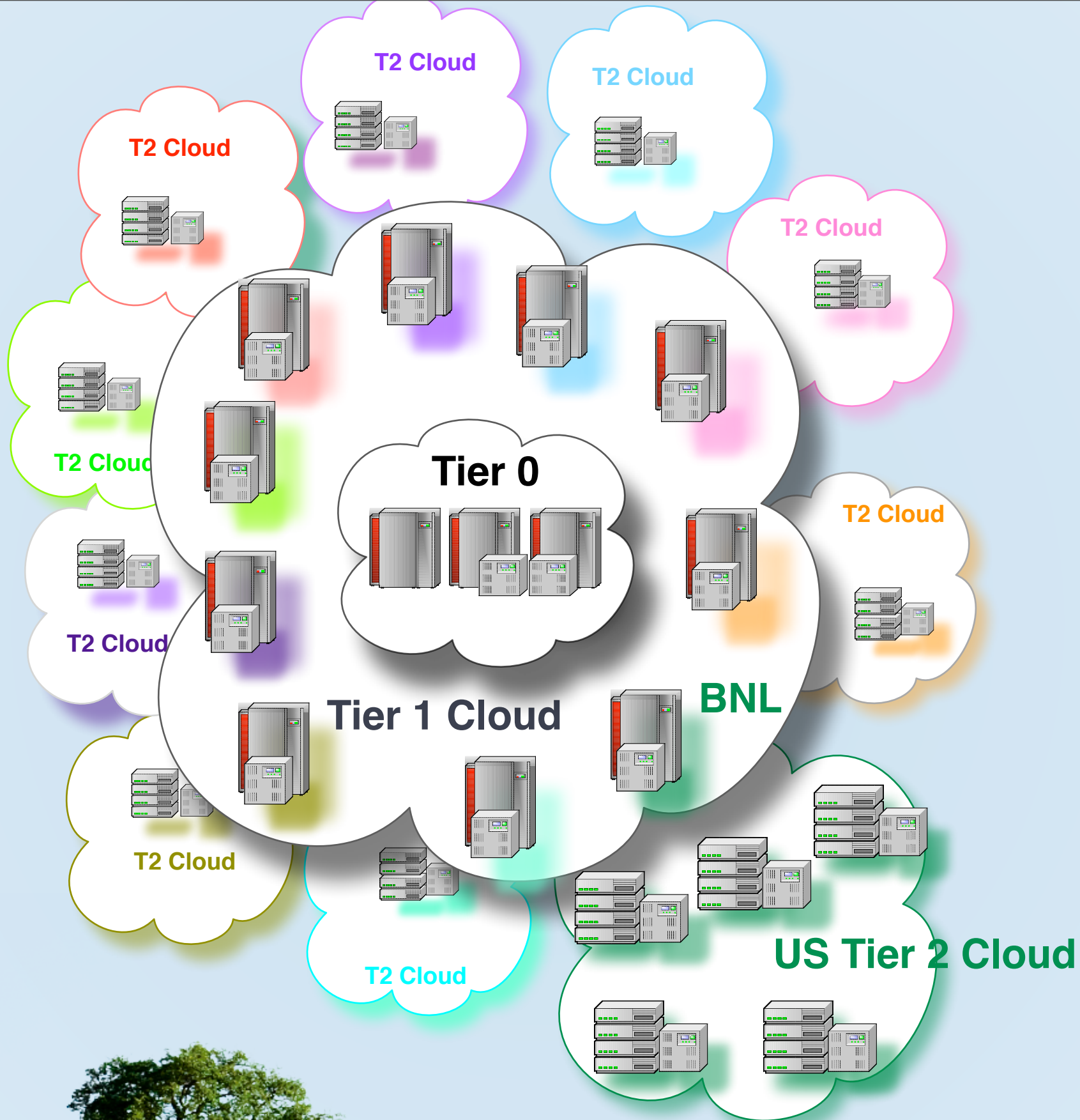
What would a task force be without a plea regarding documentation?



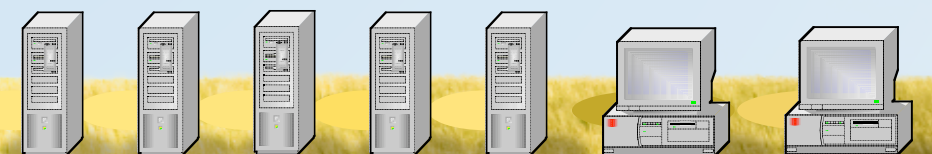


the world

the clouds



the ground



an original naive view of “analysis”

- ▶ AODs reside on T2's

university users submit jobs to the grid to produce roottuples
to bring back home for “analysis”

asynchronous processing of AODs slow, repetitive, resource-hungry

- ▶ This has changed somewhat with Derived Physics Data, DPDs
a part of the production process should include DPD production

DPDs

D1PD: according to streaming boundaries

~subset, refined, little brother of AOD


D2PD: specific to physics group, or subgroup

still undefined—certainly augmented

D3PD: flat roottuple

pDPD: performance DPD, calibrations...etc

as much as 90% of data early, we assumed 20%



D1PD/D2PD
POOL-based

Table 3: Data formats for ATLAS and quantities used in this analysis.

Format	Target Range	Current	Used	1 Year Dataset
RAW	1.6 MB	0.7 MB	1.6 MB	1600 TB
ESD	0.5 MB		0.5 MB	500 TB
MC ESD	0.5 MB		0.5 MB	500 TB
AOD	0.1 MB	0.17 MB	0.150 MB	100 TB
TAG	1 kB		1 kB	1 TB

that's a lot of data

Table 6: DPD formats and size estimates. N.B. The DPD current amounts are from [15] and are approximations to FDR $t\bar{t}$ data and are just presented as a snapshot and not to be taken literally.

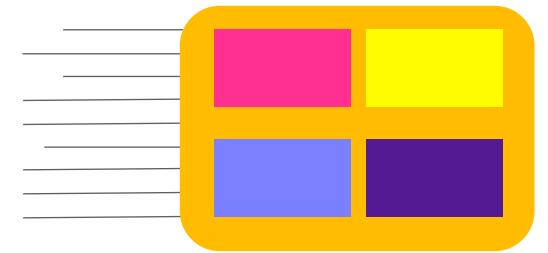
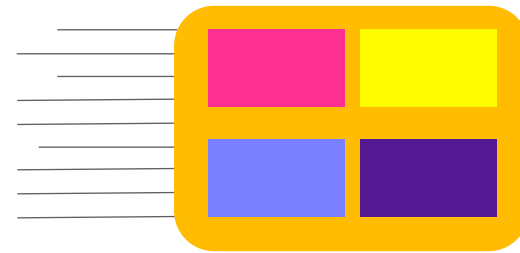
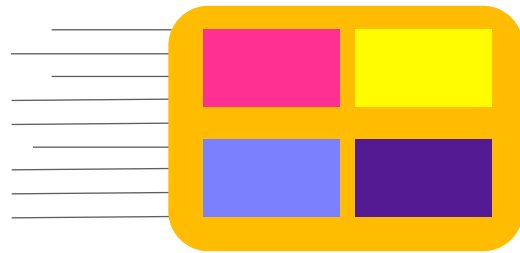
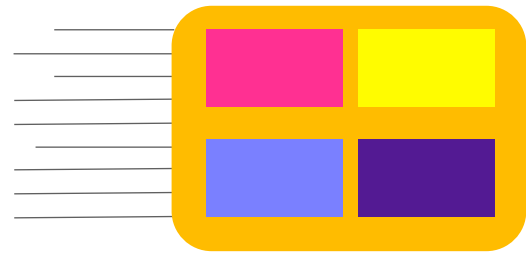
Format	Target Range	Current	Used	1 Year Dataset
D ¹ PD	1/4× AOD	31 kB	25 kB	25 TB
D ² PD	1.1× D ¹ PD	18 kB	30 kB	30 TB
D ³ PD	1/3× D ¹ PD	5 kB	6 kB	6 TB
pDPD	?	NA	?	?

that's a lot of formats

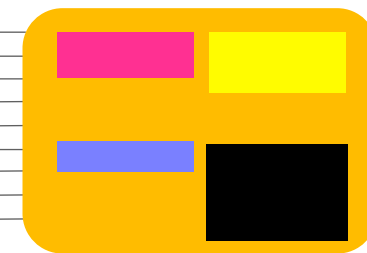
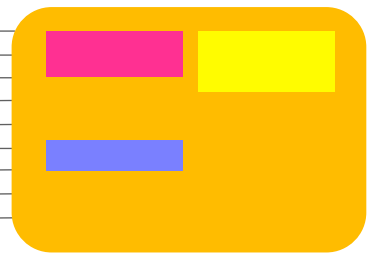
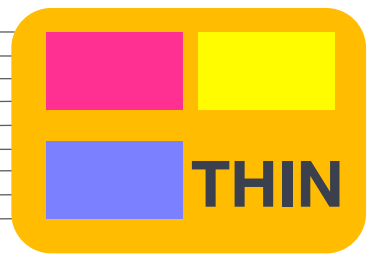
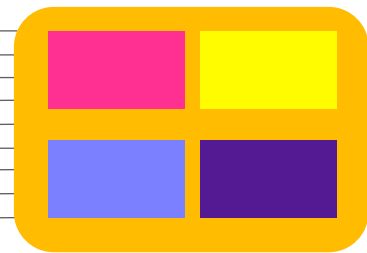
ATLAS data come in all shapes and sizes

where are they made? where are they stored? Not determined yet.

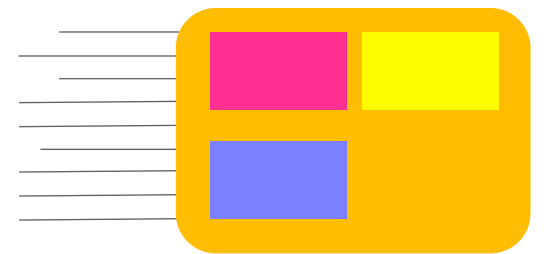
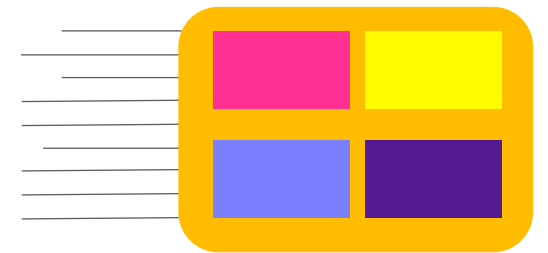
skimthinslimaug



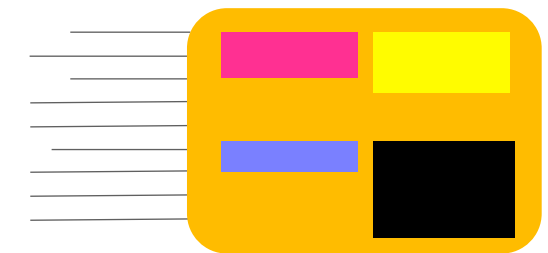
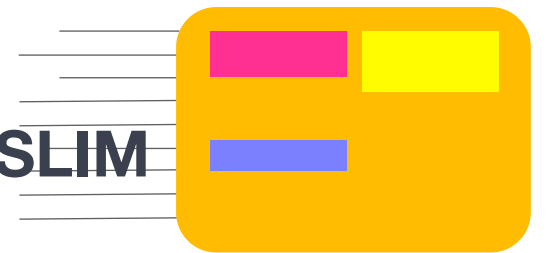
RECO



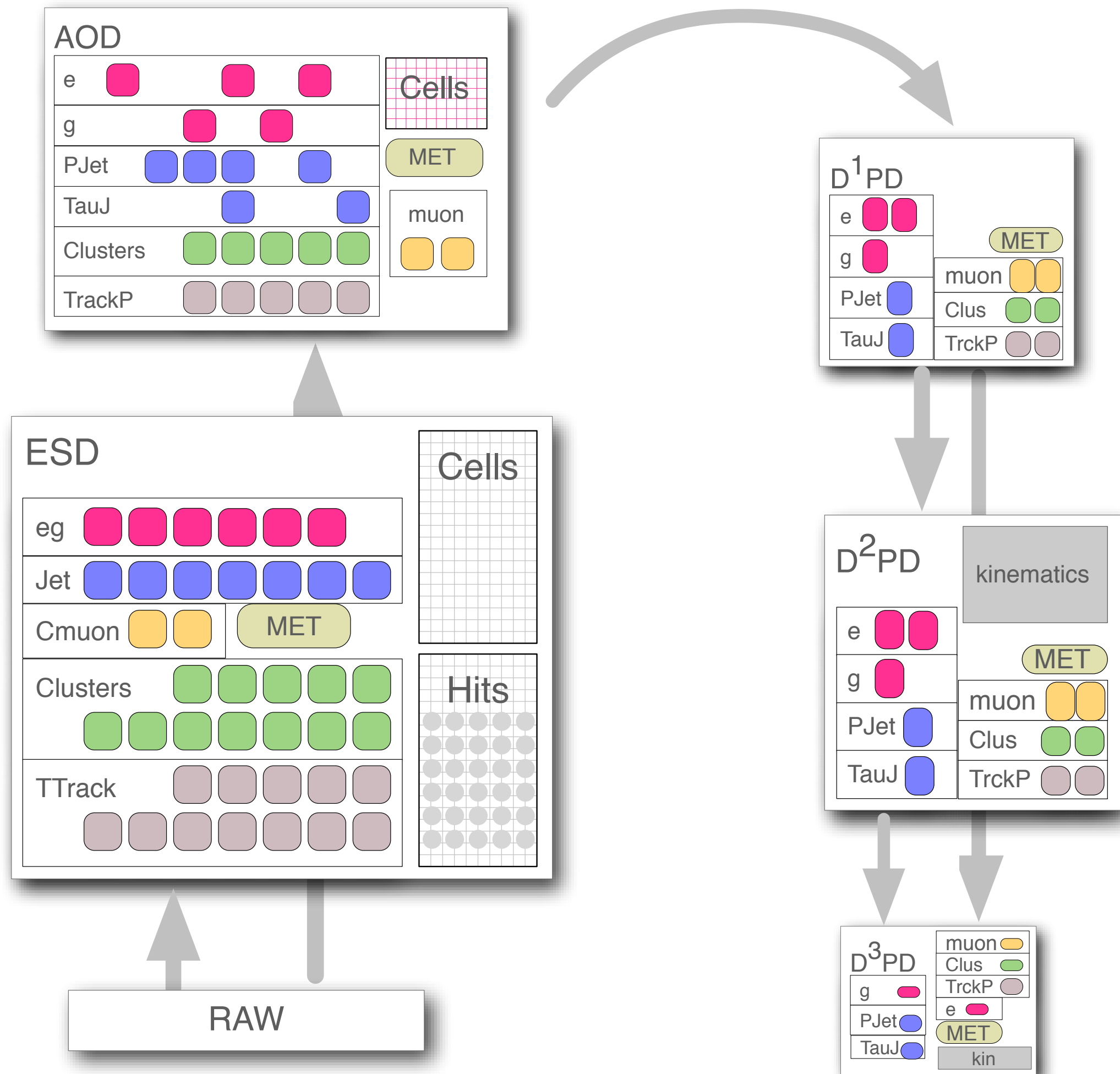
SKIM

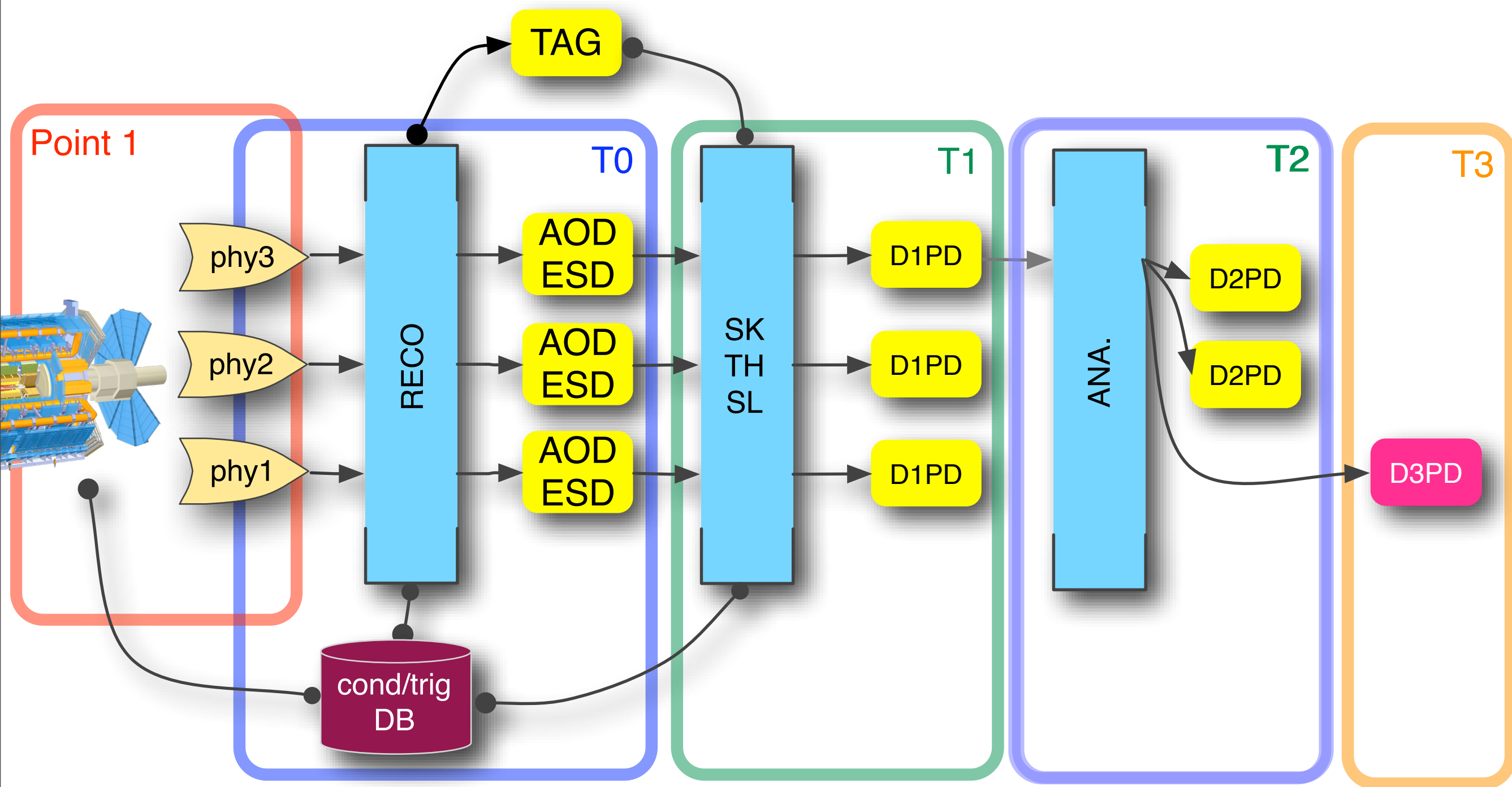


SLIM



AUGment





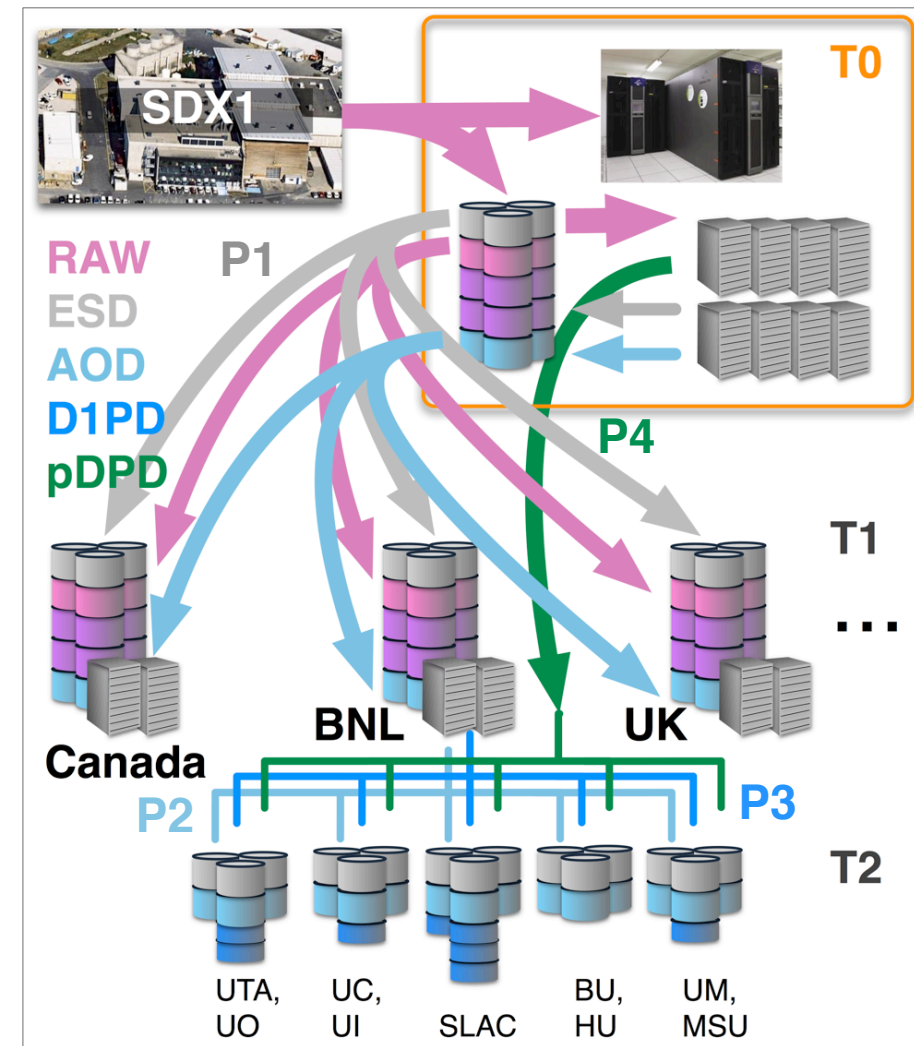
workflow

- ▶ Steady State Dataset Distribution
- ▶ Dataset creation
- ▶ Monte Carlo Production
- ▶ “Chaotic” User Analysis (“Chaotic User” Analysis?)
- ▶ Intensive Computing Tasks

Steady State Data Distribution

Table 8: The Steady State Data Distribution Use Cases. In most cases, this is a Copy operation involving Primary formats.

	data in:	data out:	from:	to:	by:	trans:	who:
P1	ESD	ESD	T0	T1	T0	C	all groups
P2	AOD	AOD	T0	T1	T0	C	all groups
P3	AOD	D1	T1	T1,T2	T1	SK, SL, TH	all groups
P4	ESD	pDPD	T0,T1	T2,T3	T0,T1	SK, SL, TH, AU	all groups

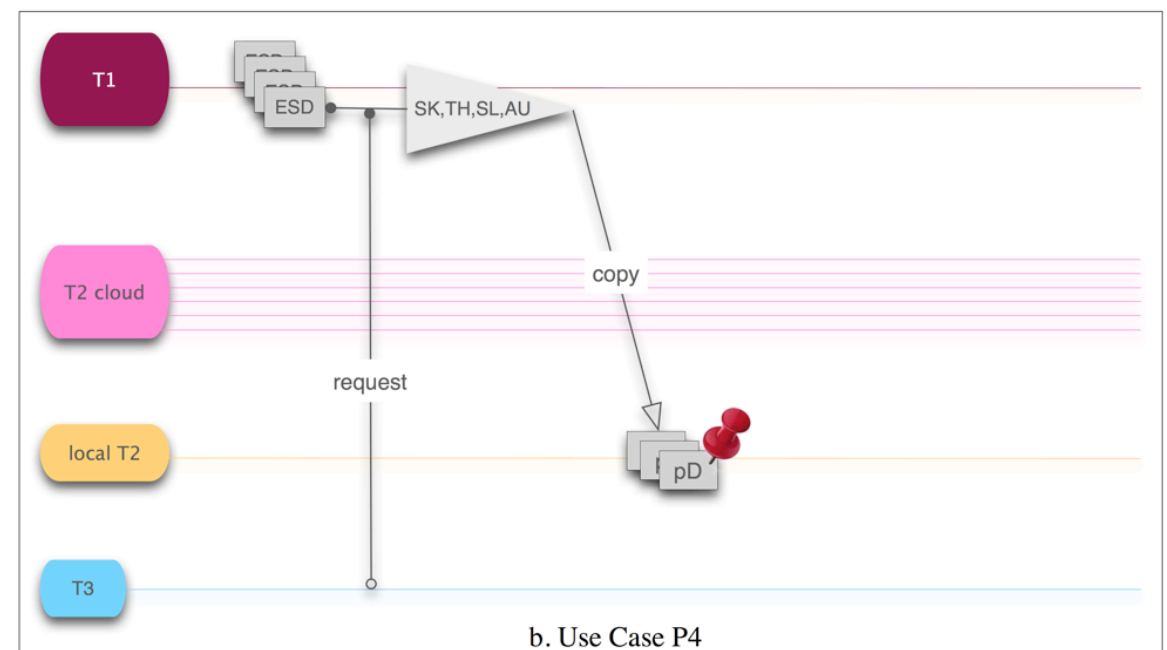
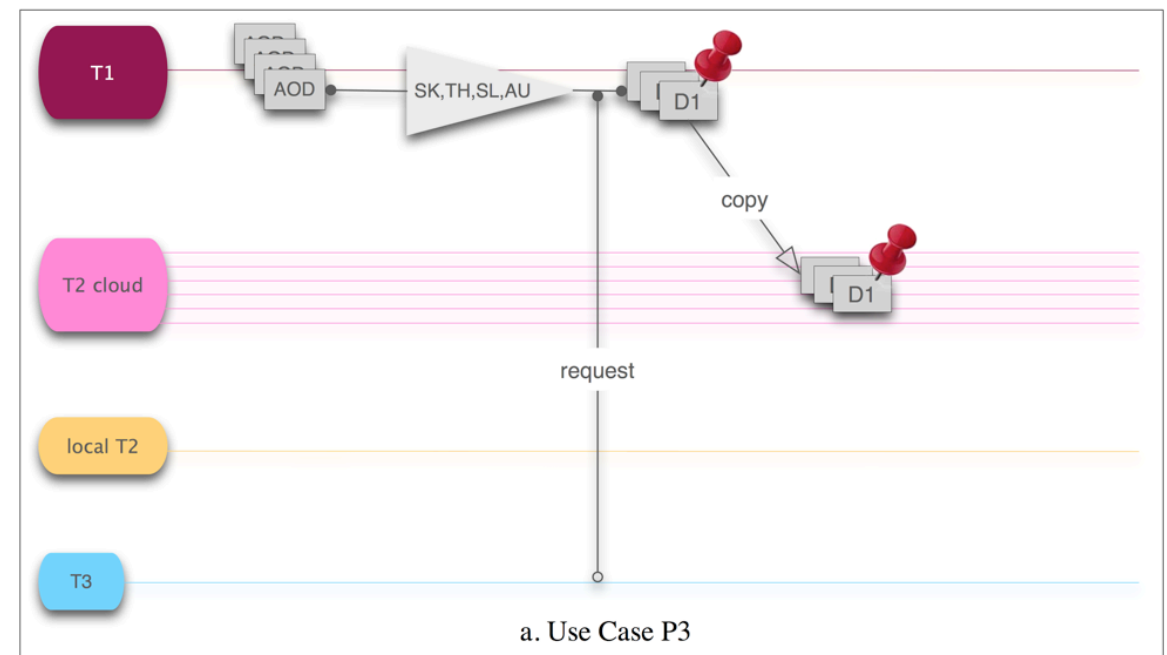


Steady State Data Distribution

Tier 3 Task Force, 3/3/09

Table 8: The Steady State Data Distribution Use Cases. In most cases, this is a Copy operation involving Primary formats.

	data in:	data out:	from:	to:	by:	trans:	who:
P1	ESD	ESD	T0	T1	T0	C	all groups
P2	AOD	AOD	T0	T1	T0	C	
P3	AOD	AOD	T1	T2	T1	C	
P3	AOD	D1	T1	T1,T2	T1	SK, SL, TH	all groups
P4	ESD	pDPD	T0,T1	T2,T3	T0,T1	SK, SL, TH, AU	all groups



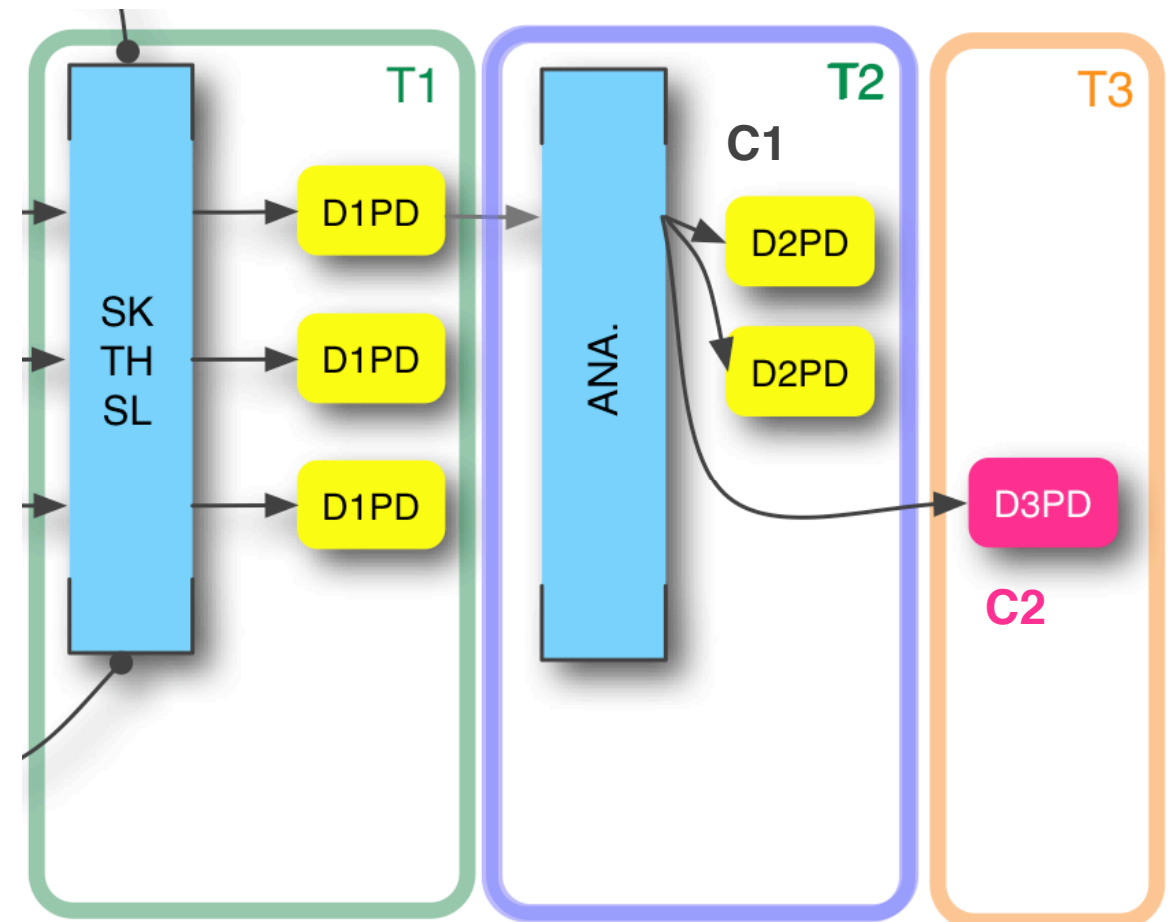
dataset creation

D1PD → D2PD:

not entirely determined

Table 9: The Steady State Data Format Creation Use Cases. In addition, a Fixing use case has been included.

	data in:	data out:	from:	to:	by:	trans:	who:
C1	D ¹ PD	D ² PD	T2	T2CL	T2CL	SK,SL, TH, AU	all subgroups
C2	D ² PD	D ³ PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular subgroups
F	D ¹ PD	D ² PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular groups



dataset creation

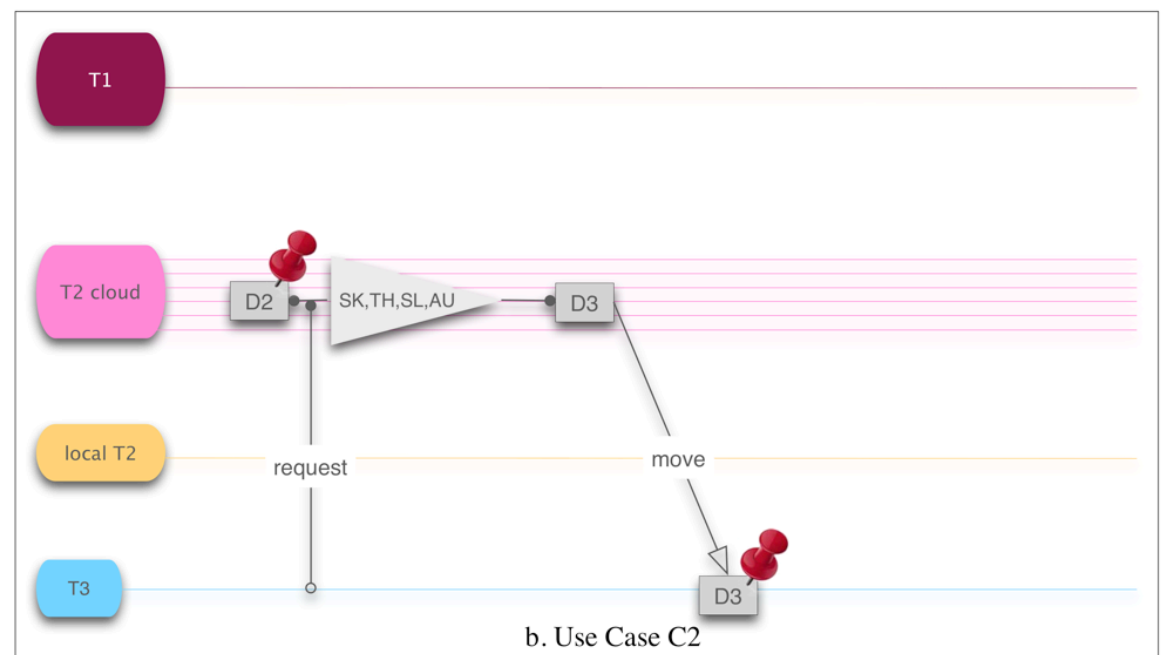
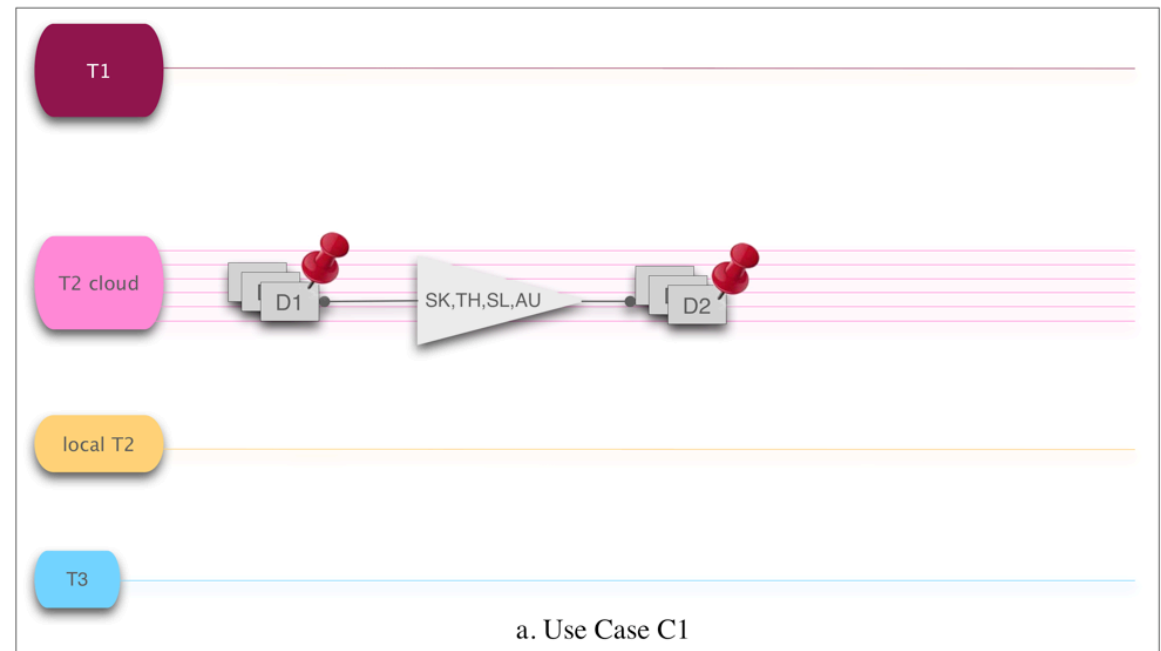
D1PD→D2PD:

not entirely determined

Tier 3 Task Force, 3/3/09

Table 9: The Steady State Data Format Creation Use Cases. In addition, a Fixing use case has been included.

	data in:	data out:	from:	to:	by:	trans:	who:
C1	D ¹ PD	D ² PD	T2	T2CL	T2CL	SK,SL, TH, AU	all subgroups
C2	D ² PD	D ³ PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular subgroups
F	D ¹ PD	D ² PD	T2CL	T2CL	T2CL	SK,SL, TH, AU	particular groups



Monte Carlo production

Generation: T1

Simulation: T2

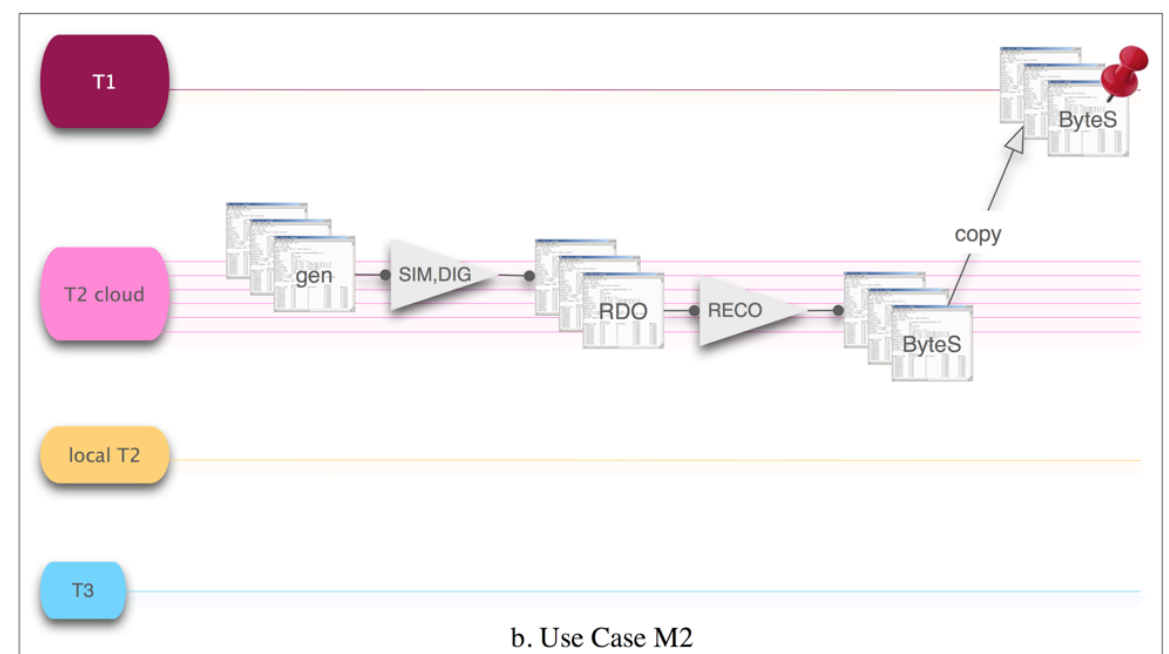
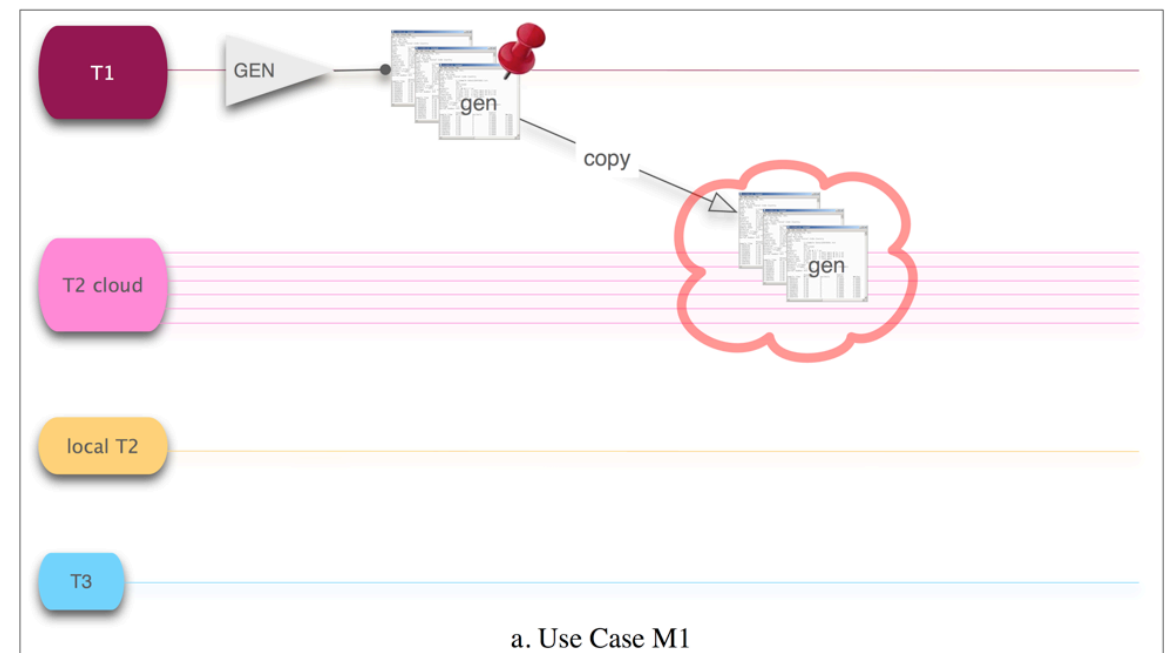
Digitization: T2

Reconstruction: T2

Tier 3 Task Force, 3/3/09

Table 10: The Monte Carlo Production Use Case.

	data in:	data out:	from:	to:	by:	trans:	who:
M1		sp	T1	T2	T1	AU, C	RAC
M2	sp	RDO	T2	T1	T1	AU,C	grid



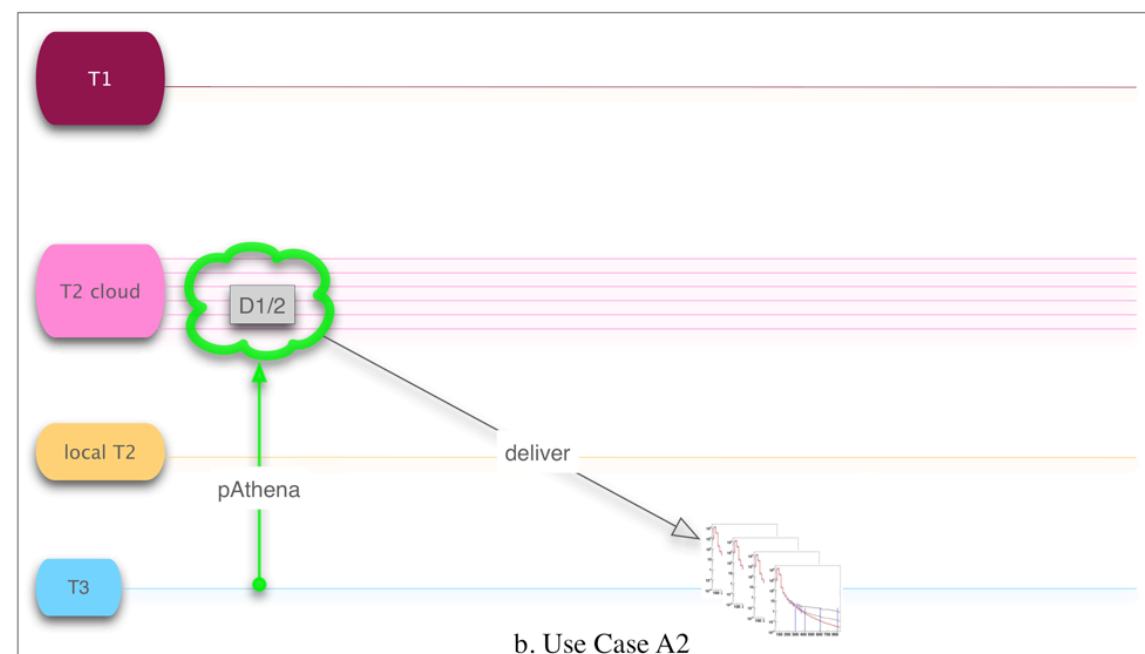
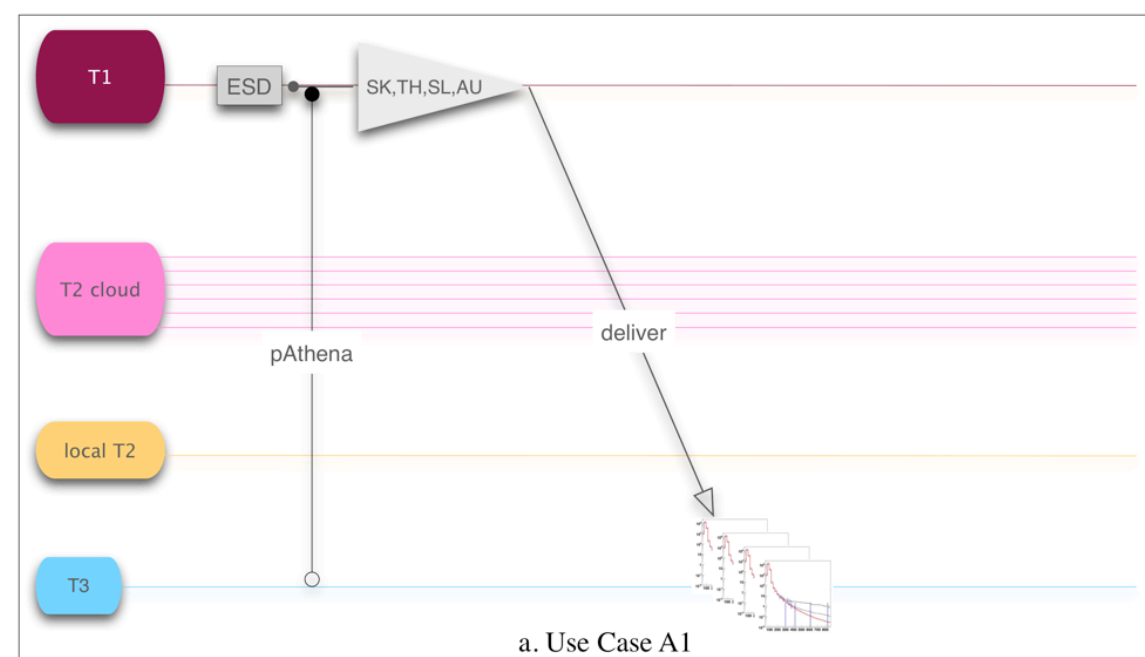
Chaotic User Analysis

“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

Tier 3 Task Force, 3/3/09

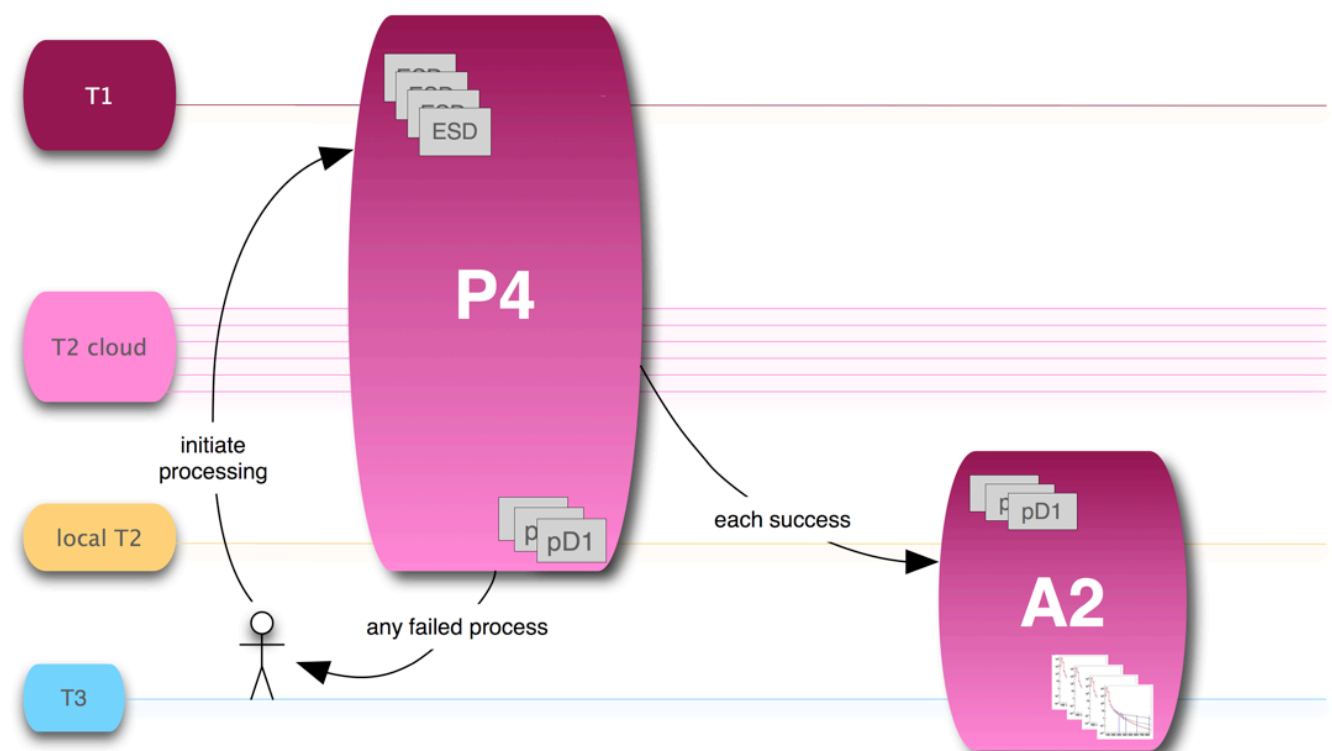
Table 11: The Chaotic Analysis Use Cases.

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



use cases

combinations of the
previous
transformations



intensive calculations

Matrix Element calculations

many cpu-centuries of computation

grid has failed DØ for these

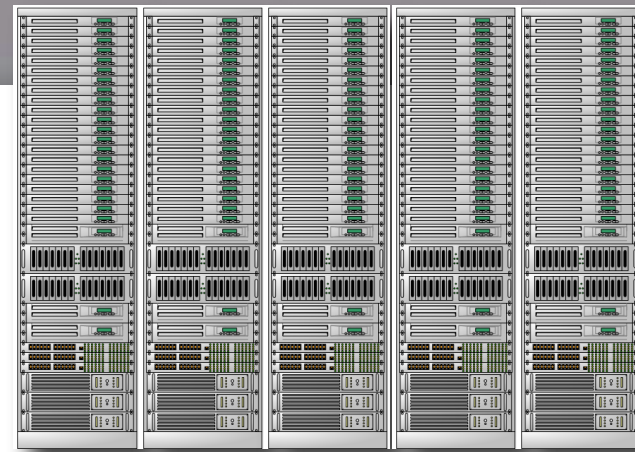
Multivariate combinations

COLLIE

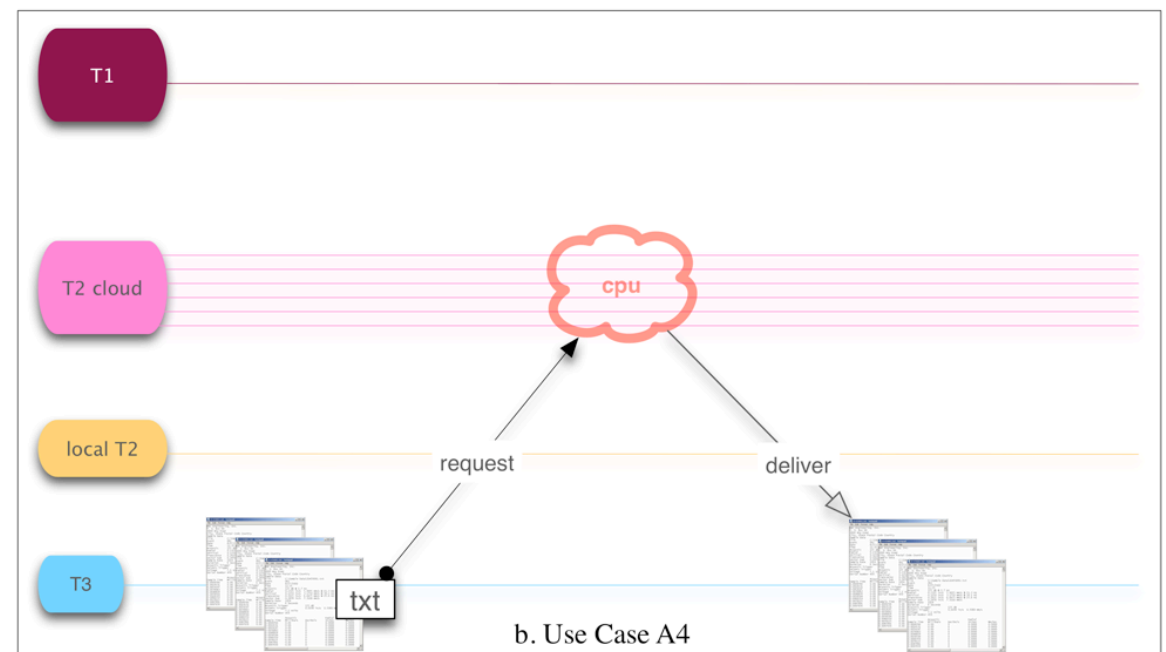
Ensemble simulation

Tier 3 Task Force, 3/3/09

~ 0 in



~ 0 out



About these intensive computational methods:

this is important:

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

About these intensive computational methods:

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

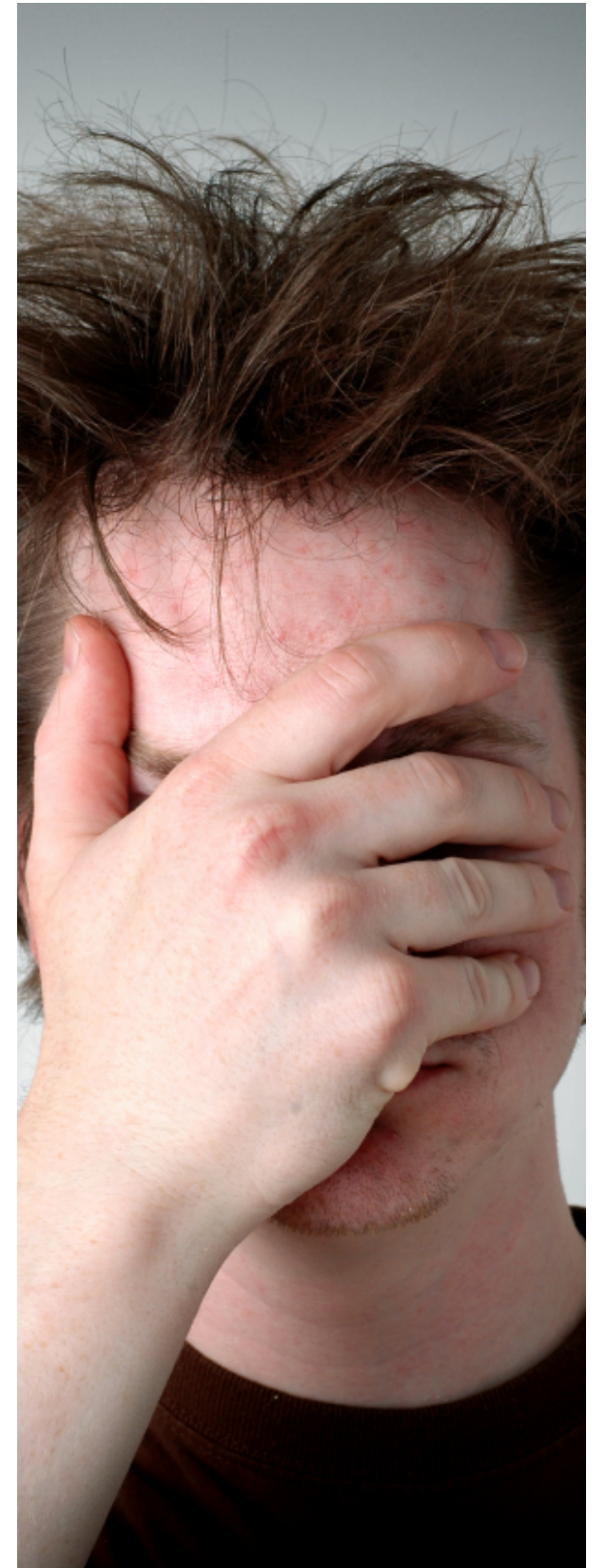
Scientific and Computing administrators

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

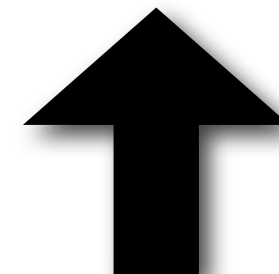
we have to plan for revolutions

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

DØ “tiers”

“Central Analysis Backend” clusters

submission facilitated by common,
integrated tools...including parallel
processing

ATLAS
analog:

Reconstruction Farm:
~400 nodes

~T1/0?

“CAB” clusters:
1252 nodes
in 1805 & 3292 cores
400 TB “SAM Cache”
80 TB users
batch only

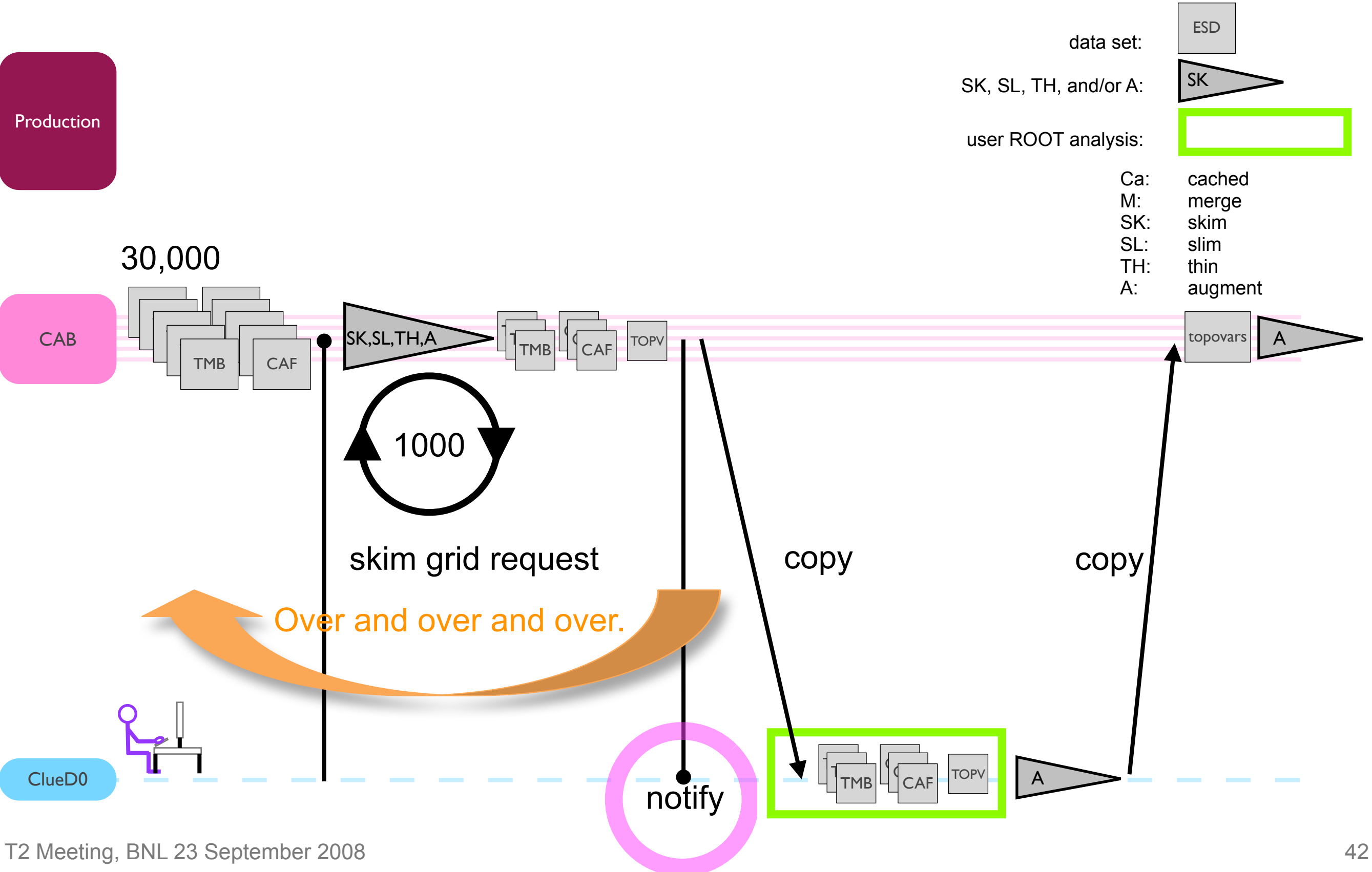
~T2’s?



“ClueD0” desktop cluster:
~500 machines
160 TB served storage
interactive & batch system

~T3’s?

DØ Single Top “use case”



single analyses are intense

► A DØ analysis

about once per month

before systematic error studies

before “editorial board” demands

just one analysis

source	files	events	jobs
data	96k	1600M	2400
QCD background	96k	1600M	2400
signal MC	25.6k	200M	2400
bckgnd MC	12k	120M	560
total	240k	3B	8000

Observation 7 *Full-scale, precision analyses will be a huge load on the Tier 2 structure from the perspective of computation and file-access. Monitoring and resubmitting failed jobs will surely continue to be a serious complication for analyzers. If history is a guide, current predictions of how this maps to the ATLAS analysis future are sure to be underestimated.*

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
$W^\pm \rightarrow e^\pm \nu_e$	0.0788	1150	23.5	8.07
$W^\pm \rightarrow \mu^\pm \nu_\mu$	0.0768	1030	23.1	13.6
Heavy ion	2.08	56,000	267	-

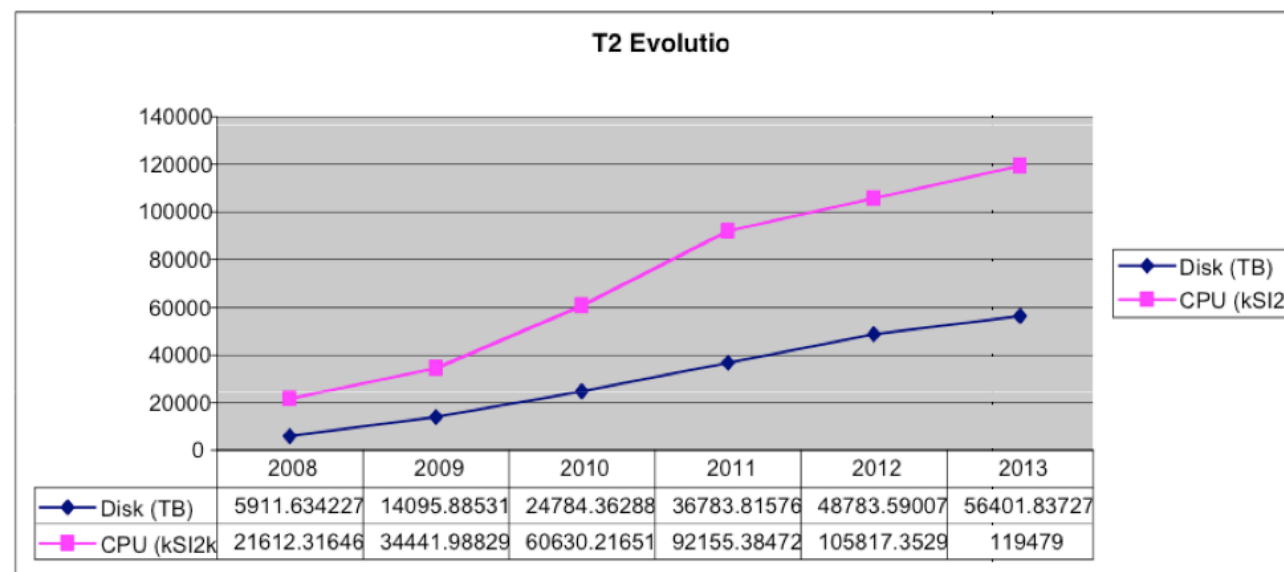
Table 18. in kSI2k-s, without pileup

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

Tier 2/3 Modeling

Benchmark:

quantity	value used	high	low	comments
LHC year	2010	2011	n.a.	assume 2008 start
Ins. $\mathcal{L} \text{ cm}^{-2}\text{s}^{-1}$	2×10^{33}	3.5×10^{33}	10^{33}	Garoby, LHCC 08
annual $\int \mathcal{L} dt \text{ fb}^{-1}$	10	?	?	rounded from 12
annual dataset	2×10^9 events	?	?	[7]
sim. time	1990 kSI2K s ($t\bar{t}$)	2850 kSI2K s γj	1030 kSI2K s $W \rightarrow \mu$	[16]
dig. time	29.1 kSI2K s ($t\bar{t}$)	29.2 kSI2K s j	23.1 kSI2K s $W \rightarrow \mu$	[16]
reco. time	47.4 kSI2K s ($t\bar{t}$)	78.4 kSI2K s j	8.07 kSI2K s $W \rightarrow e$	[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-IIIF)	[16]
$D^1\text{PD} \rightarrow D^2\text{PD}$	0.5 kSI2K s	?	?	[15]
$D^2\text{PD} \rightarrow D^3\text{PD}$	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...later

Tier 2 simulation for one year

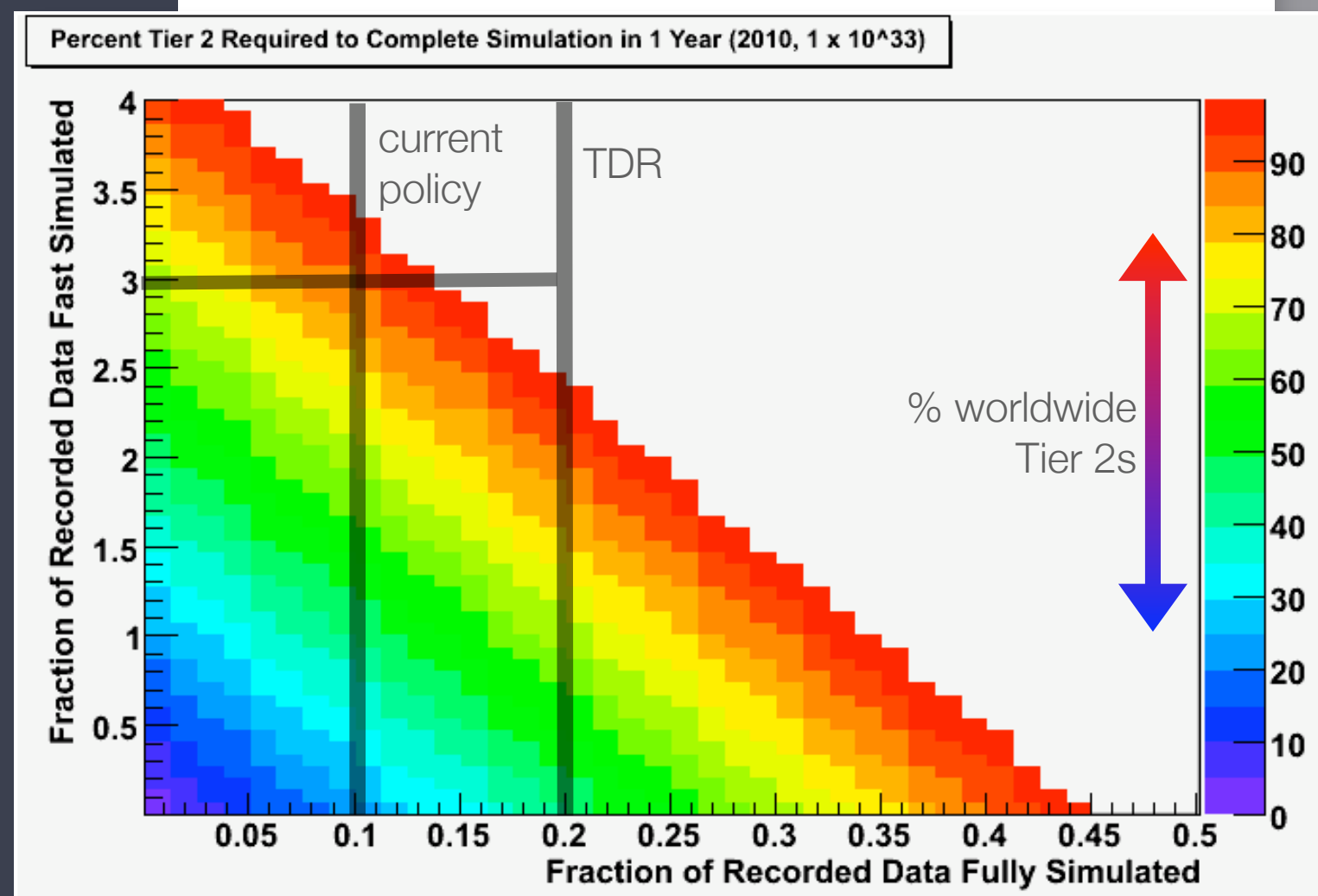
- horizontal axis:

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.*

Observation 8 *Should ATLAS-wide production needs be more than the Tier 2 centers can provide, the only flexibility is to "eat" away at the 50% of the Tier 2 resources nominally reserved for U.S. user analysis. One has to ask what the likelihood is of such an outcome and whether U.S. ATLAS analysis could survive the effects of such a result.*

sobering

could this be wrong? sure.

can we risk ignoring it?

recommendations

5 Primary Recommendations

Minimum necessary requirements

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

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

see Mark's talk

T3gs use cases, enhanced

- Production: Physics Group D2PD from cached D1PD

assume a full stream (1/10)

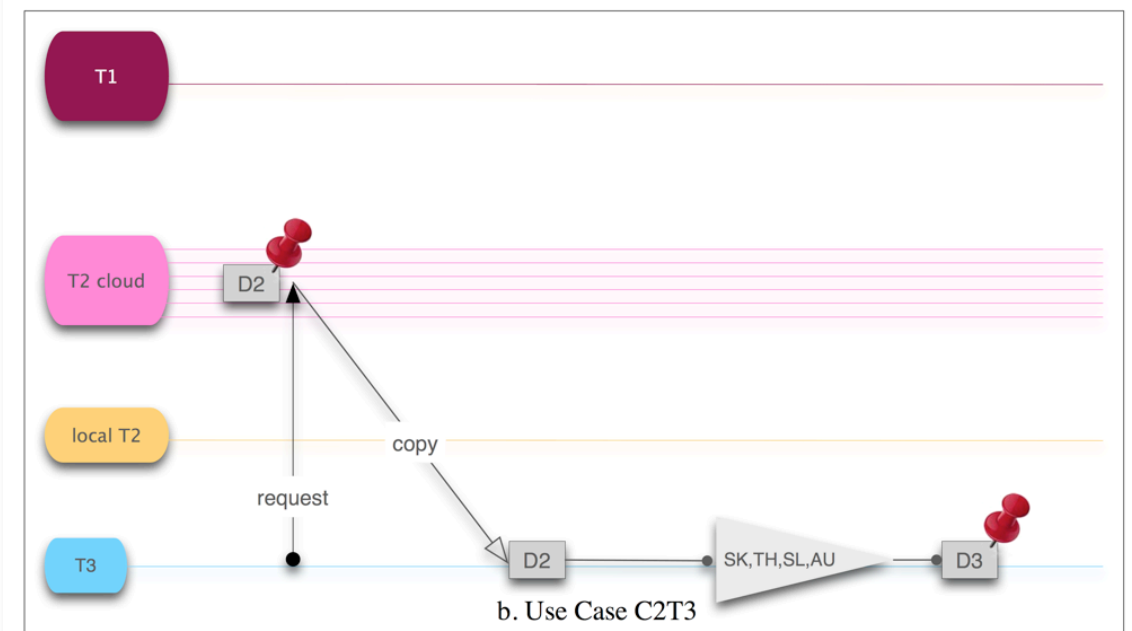
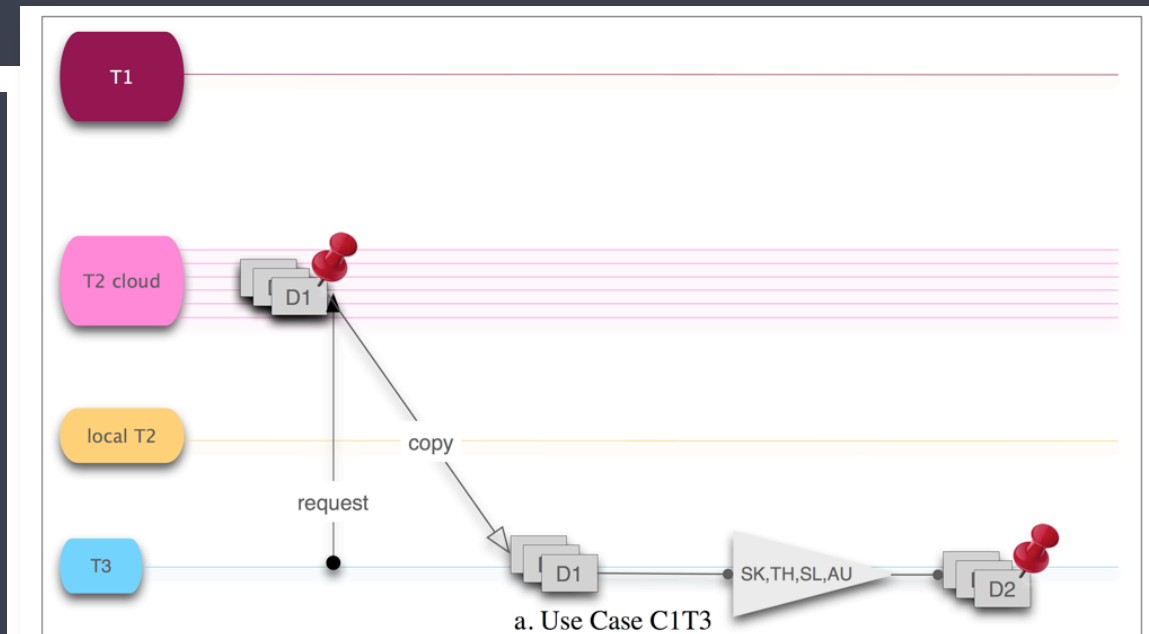
few days to produce

- Monte Carlo Production: in support of a physics group

ttbar-sample appropriate to the 10fb benchmark

sample-sized, signal + background, ATLFast-II

few days



T3g

Tier 3 with “**g**rid” connectivity

a campus-based, tower cluster

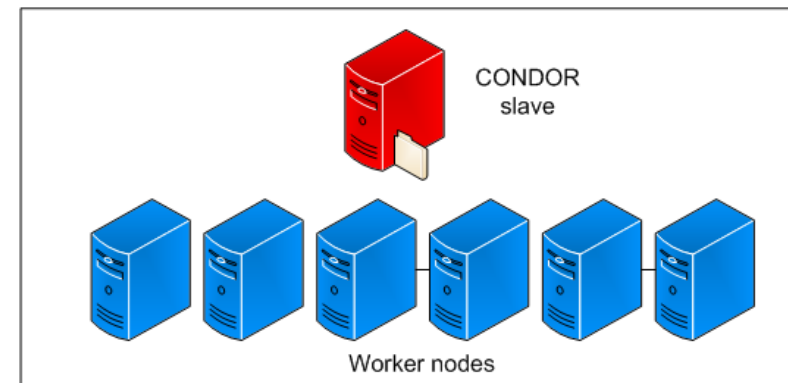
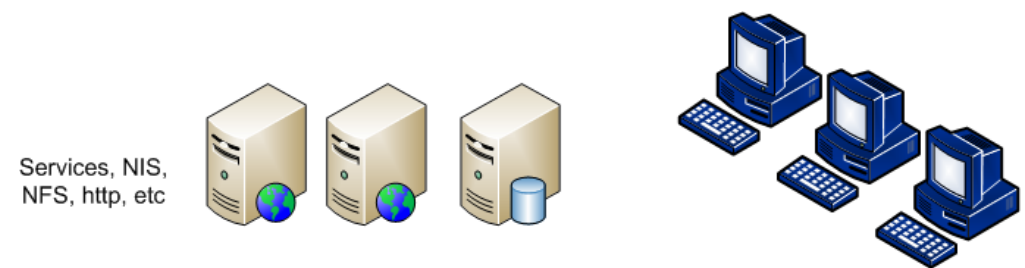
office-based

Characterized a strawman

~\$25k

ANL and Duke are building them

see Sergei and Doug’s talks



80 processors
>100kSI2k

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. (page [82](#))

Recommendation 3

must be able to subscribe to large datasets

cannot move TBs by hand...

see Marco's talk

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

Two other T3 classes

- ▶ T3_w

Tier 3 **W**orkstation

unclustered workstations...OSG, DQ2 client, root, etc

- ▶ T3_{af}

Tier 3 system built into lab or university **a**nalysis **f**acility

special arrangement of purchasing through the AF

the CDF Model—fair-share computing privileges in exchange for contribution

zero special data-access privileges

2 Technical Recommendations

Service modifications to Panda

Focus on point-to-point communications

Recommendation 6: Currently, the submission of pAthena jobs to an internal cluster, exposes that cluster to receipt of pAthena job tokens (aka., Panda pilots) which can cause spurious load and can be used by any user in the collaboration. This would need to be changed to be able to switch off this consequence and decouple such sites from central services. (page [82](#))

Recommendation 6

With a switch - same scripts for local and grid pAthena submission

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. (page [121](#))

Recommendation 7

Rough goal:

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



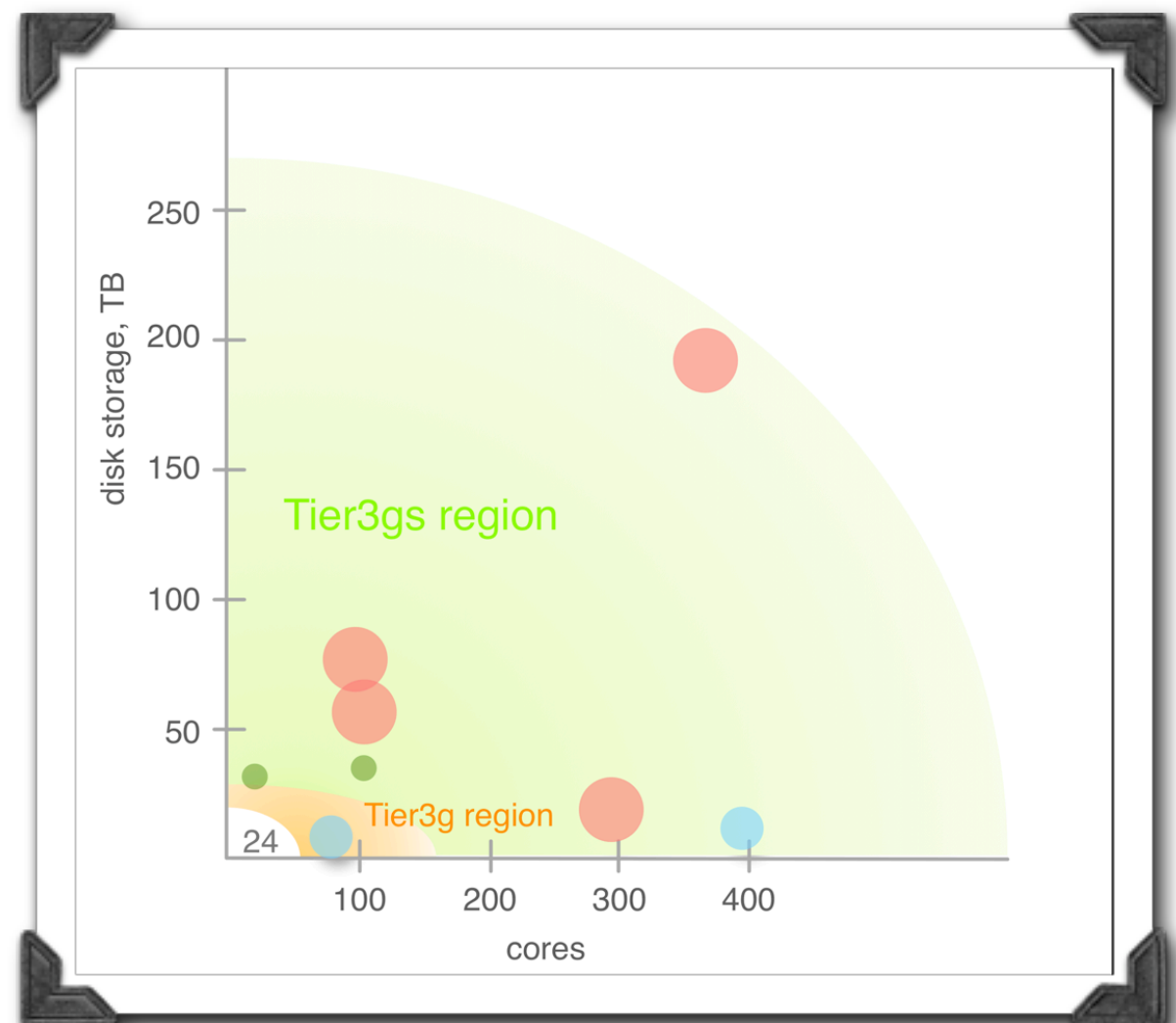
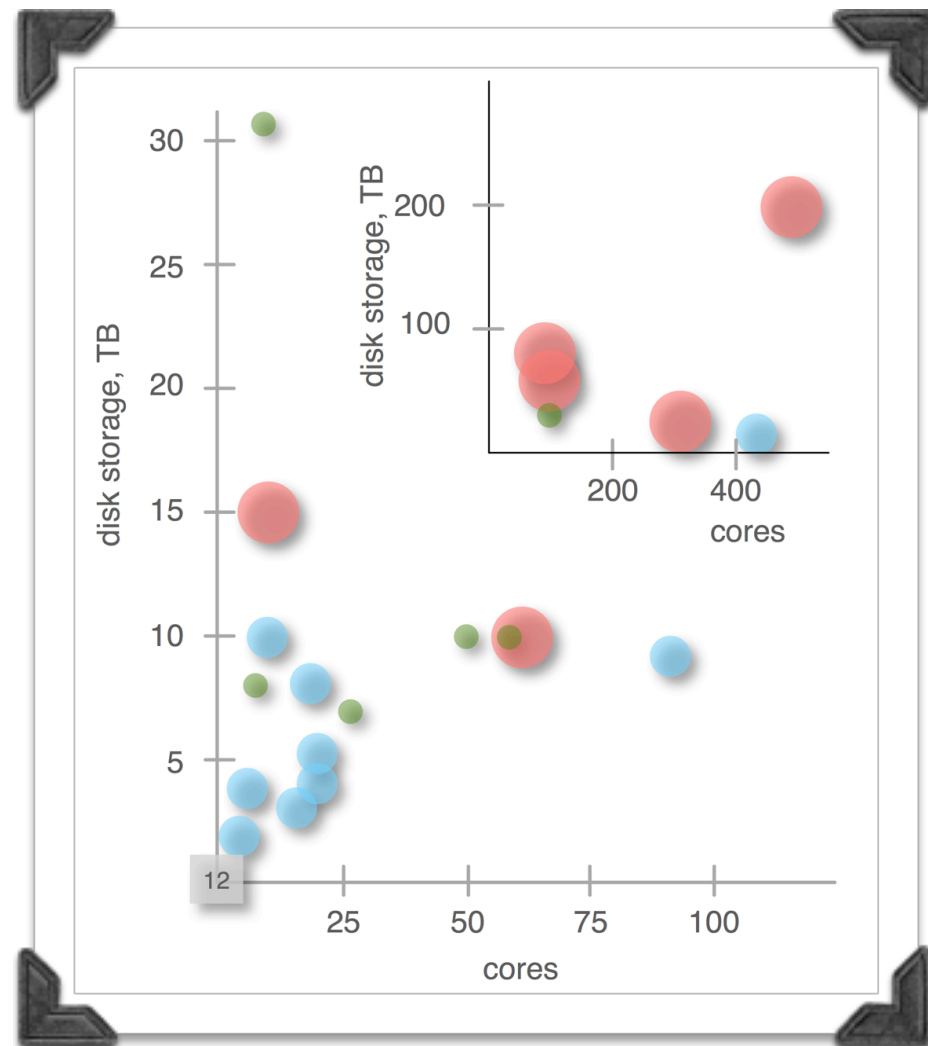
Partnership recommendation

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

conclusions



evolution

more depth will enhance



- ▶ We have tried to indicate that 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

Two critical issues

- ▶ Support model

personal, regular, common

- ▶ Access to the data for 2010-2011 milestones

target point-to-point minimal connectivity

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

One important side issue

- ▶ HEP active and “apparent”
in departments and on campuses
is critical to the success of the LHC mission

what's next:

- ▶ some tinkering with the document and eventual dissemination
- ▶ is it accepted by Mike, Howard, and Jim?
if not...!
if so...?
- ▶ Much time for discussion today.