



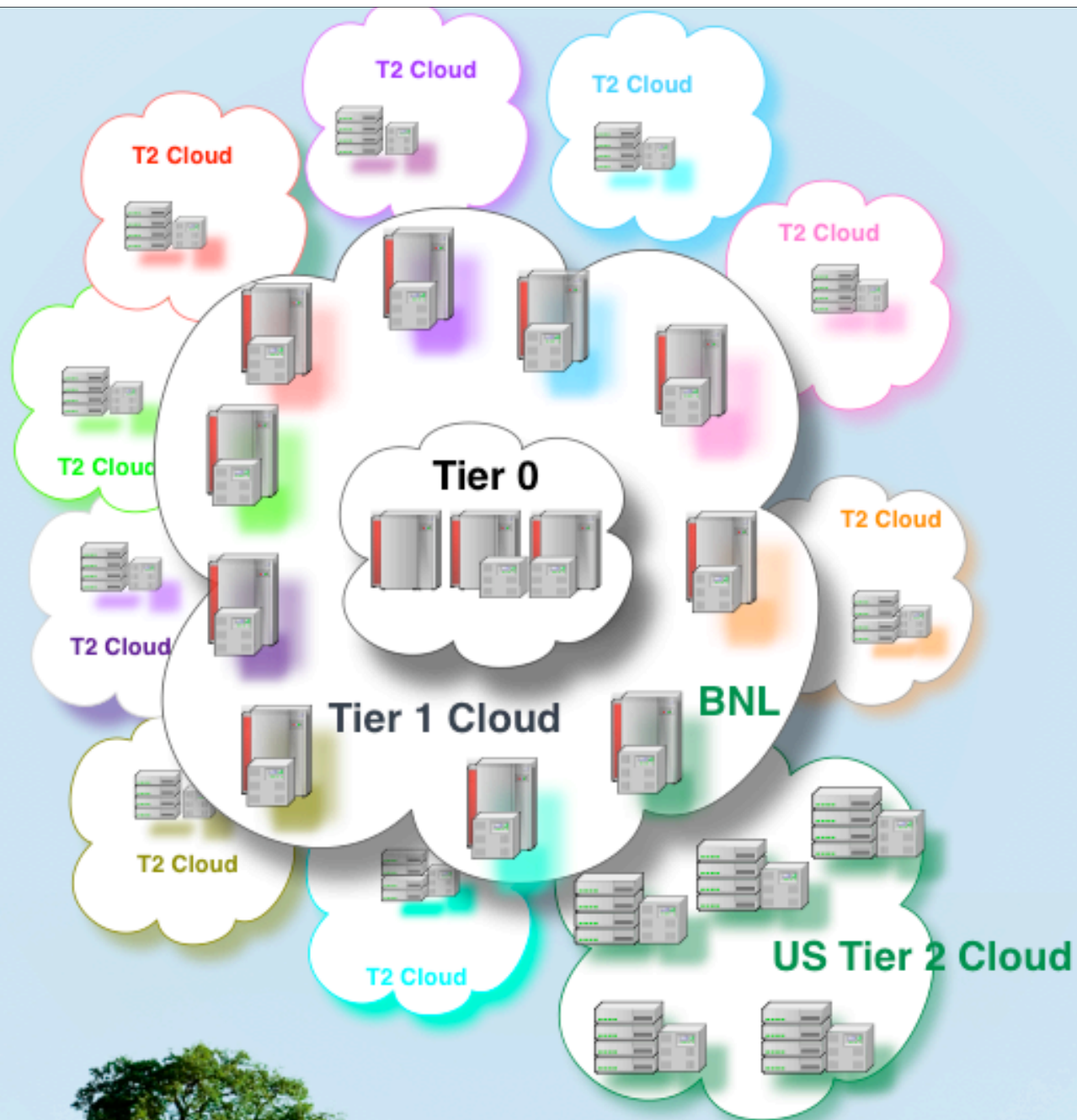
collisions: 1,000,000,000/second
critical rare events: 0.0001/second } $1:10^{-13}$

Have to
analyze it

3 PetaBytes of data/year

keep that up for
2 decades.





the world

the clouds

the ground



Tier 3 Task Force Summary

Chip Brock, Michigan State University

Doug Benjamin, Duke 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

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 detail.
- ▶ 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.~~
- ▶ 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 $X + n*Y$ \$'s where n = number of active physicists.

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 Brooijmans³,
Sergei Chekanov^{4**}, Jim Cochran⁵, Michael Ernst⁶, Amir Farbin⁷,
Marco Mambelli^{8**}, Bruce Mellado⁹, Mark Neubauer¹⁰,
Flera Rizatdinova¹¹, Paul Tipton¹², and Gordon Watts¹³

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

What's a Tier 3 now?

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	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

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	
Disk storage (TB)	
Tape storage (TB)	
Network connectivity	

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 X53 3GHz)
SI2K total units	24k
Disk storage (TB)	4
Network connectivity	1 Gb 100 Mbps net

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

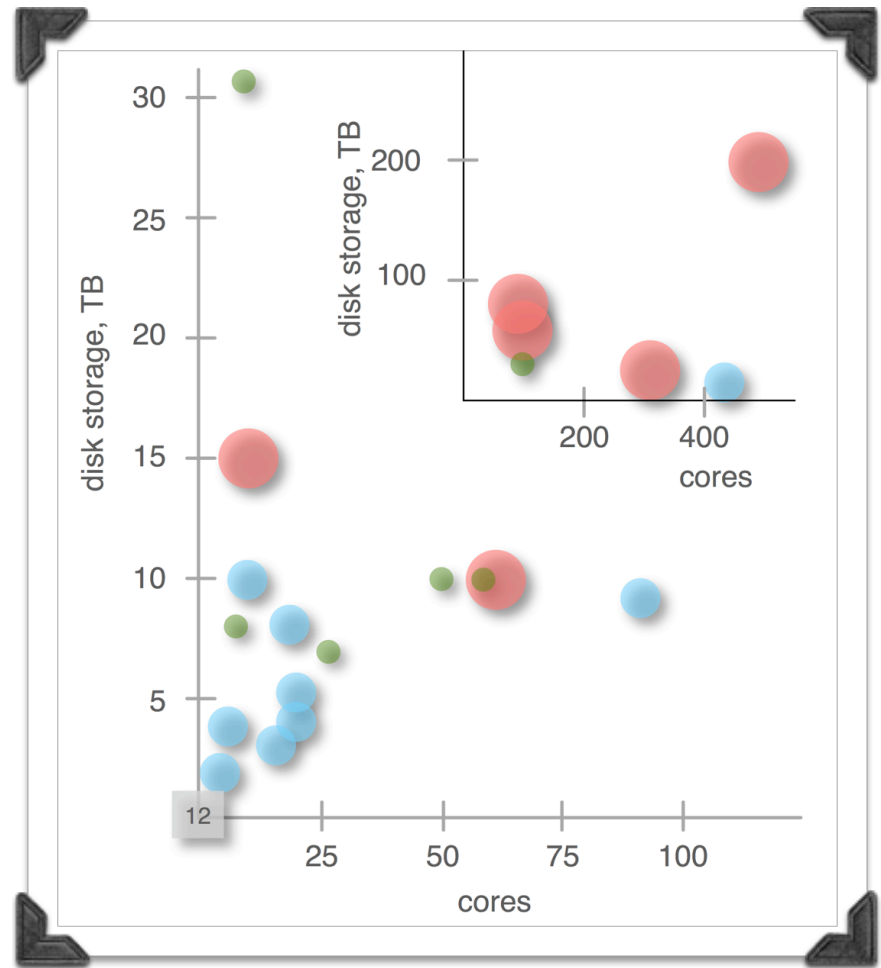
Tier 3s today.

Survey:

33 ATLAS university institutes

dot size/color: network connectivity:

100Mbps, 1Gbps, 10Gbps




information is scattered:



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

contribution | manage |



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... (4n00) ([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?

computing & analysis models

tied to the data formats

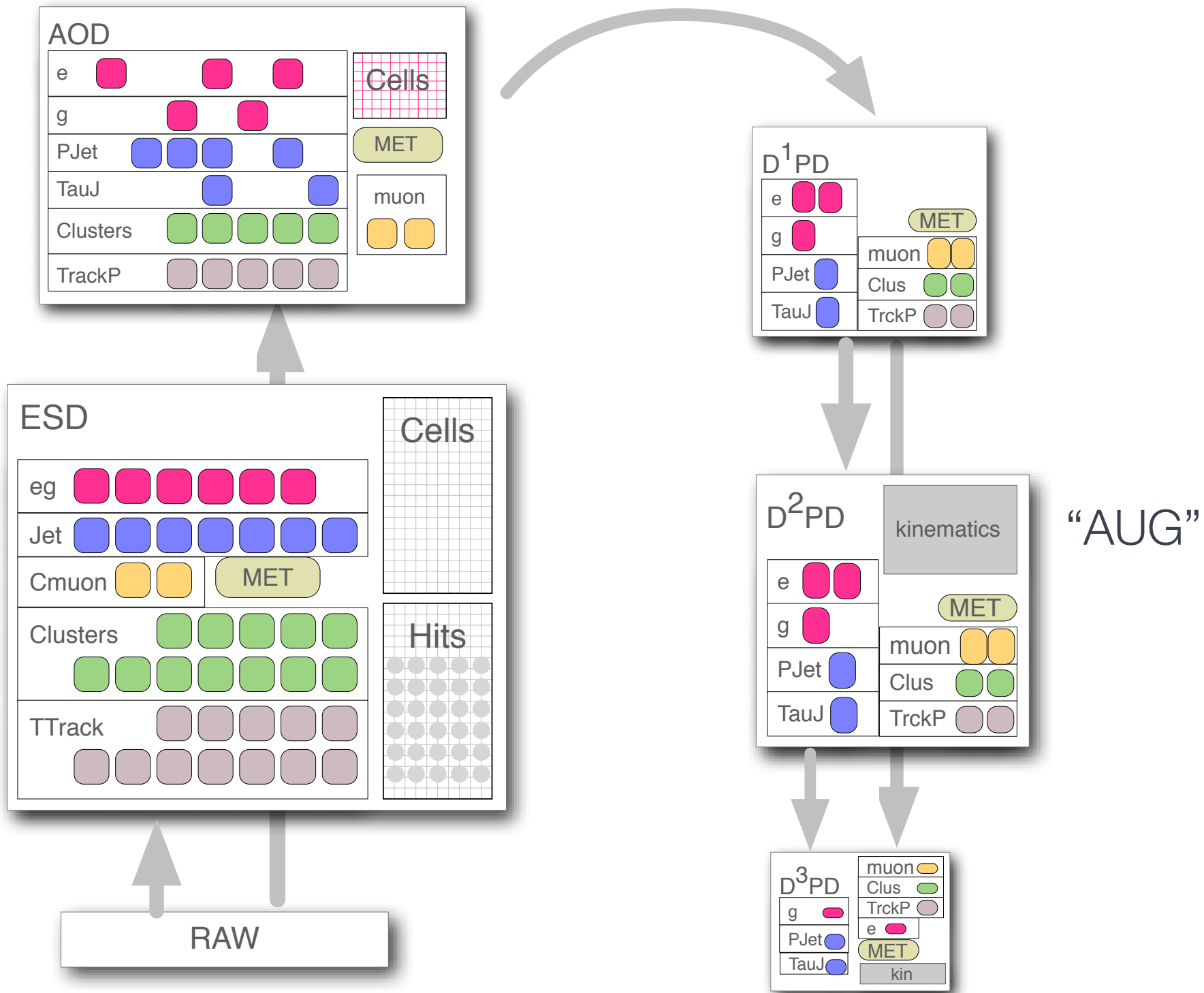


Table 3: Data formats for ATLAS and quantities used in this analysis.

Format	Target Range	Current	Used	1 Year Dataset
RAW	1.6 MB		1.6 MB	1600 TB
ESD	0.5 MB	0.7 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

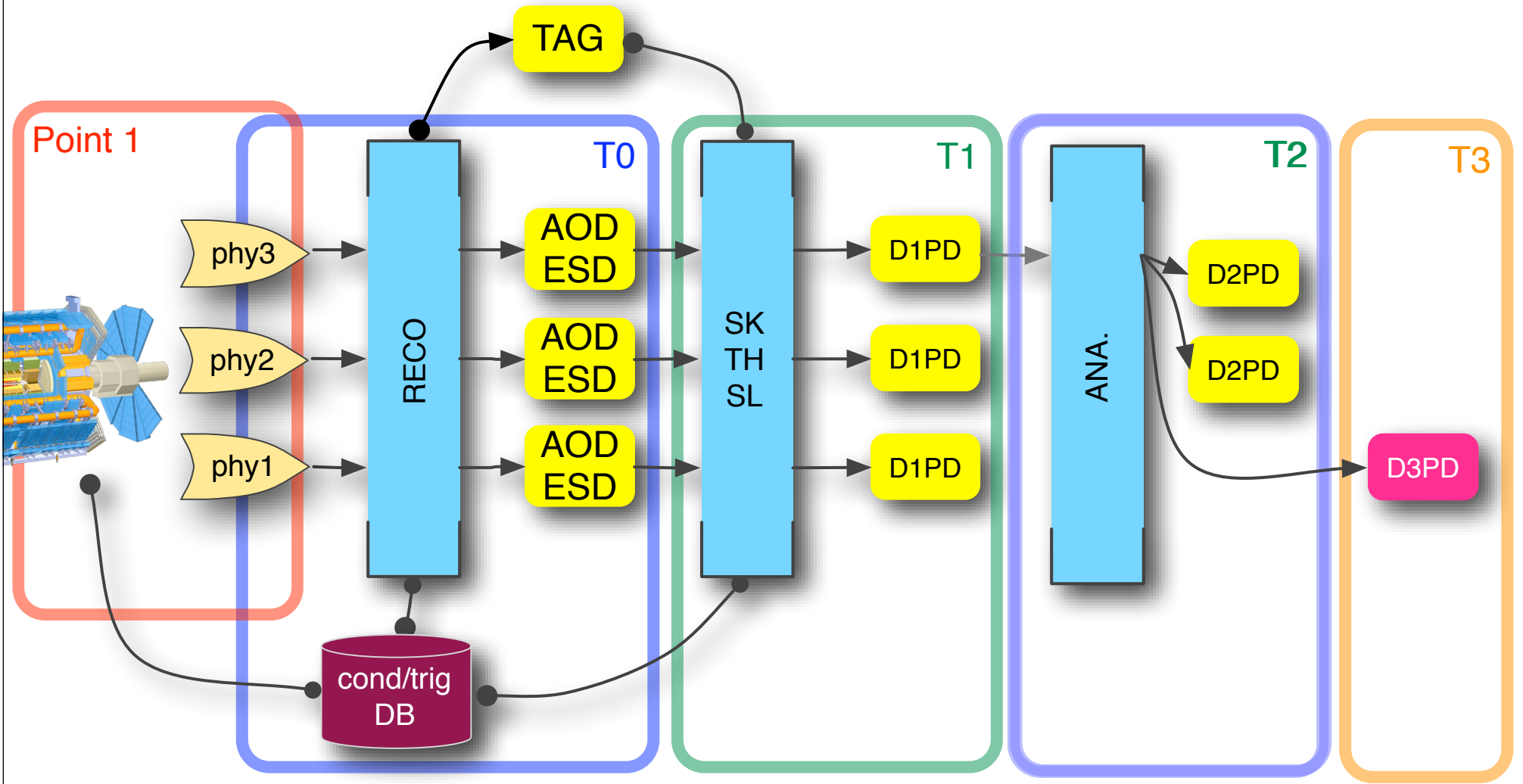
Table 6: DPD formats and size estimates. N.B. The DPD current amounts are from [15] and are approximations to FDR $t\bar{t}$ data and are just presented as a snapshot and not to be taken literally.

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 wholly determined yet.



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

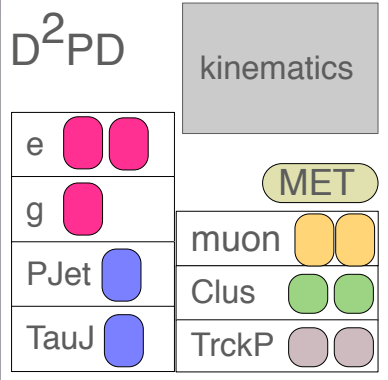
1. Steady State Dataset Distribution

2. Dataset creation

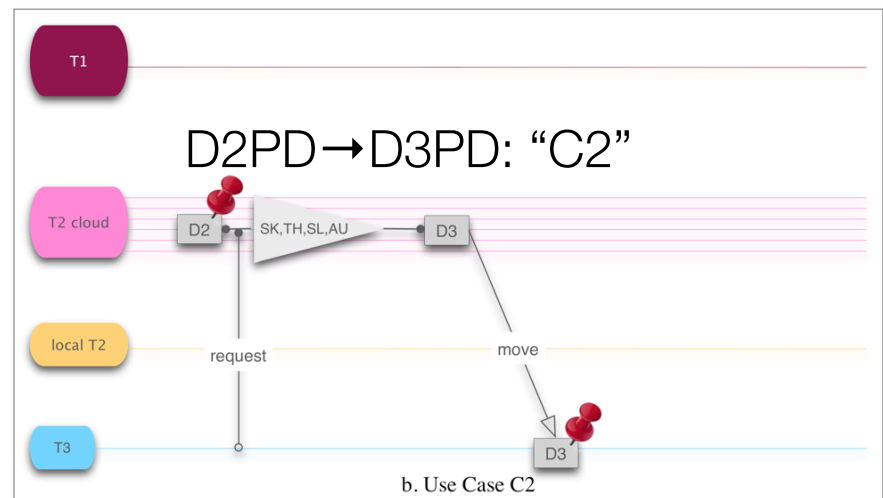
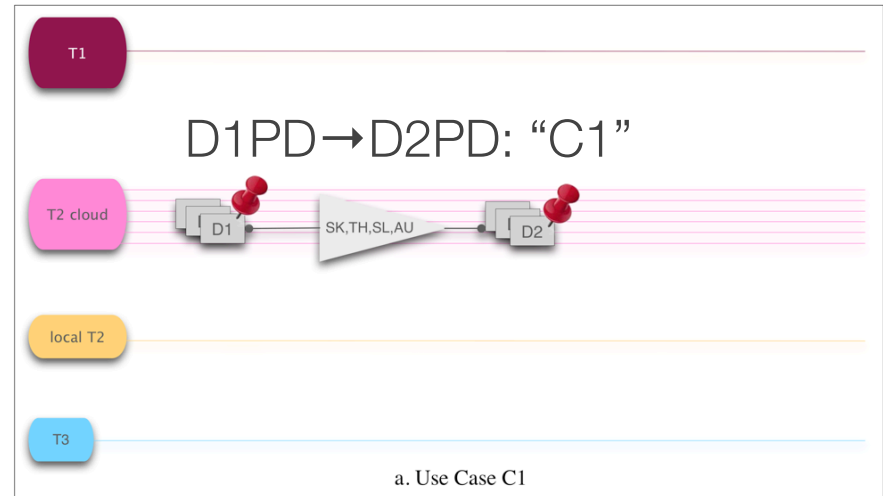
3. Monte Carlo Production

4. "Chaotic" User Analysis

Intensive Computing Tasks



Tier 3 Task Force



1. Steady State Dataset Distribution
2. Dataset creation

3. Monte Carlo Production

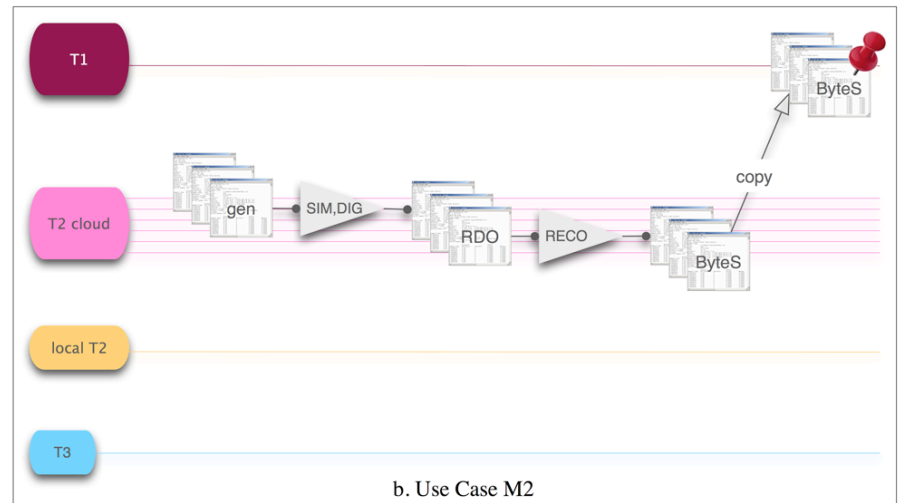
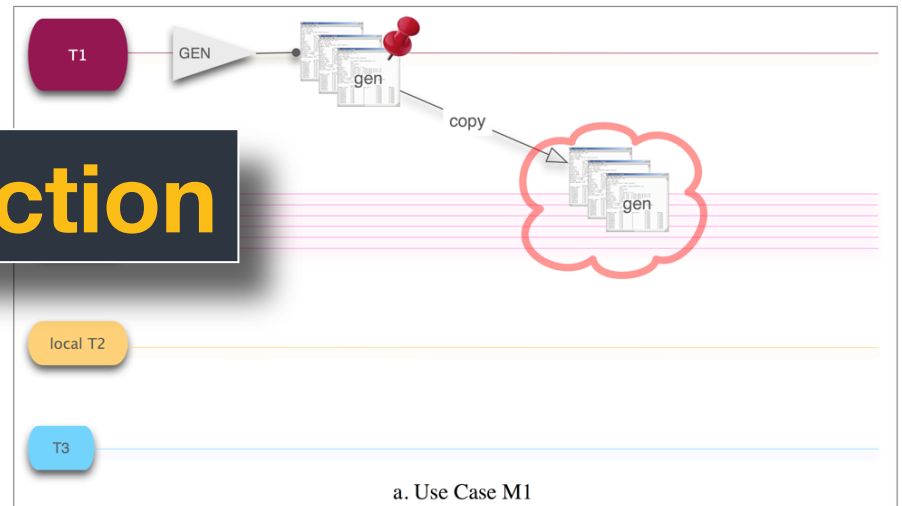
4. "Chaotic" User Analysis
- Intensive Computing Tasks

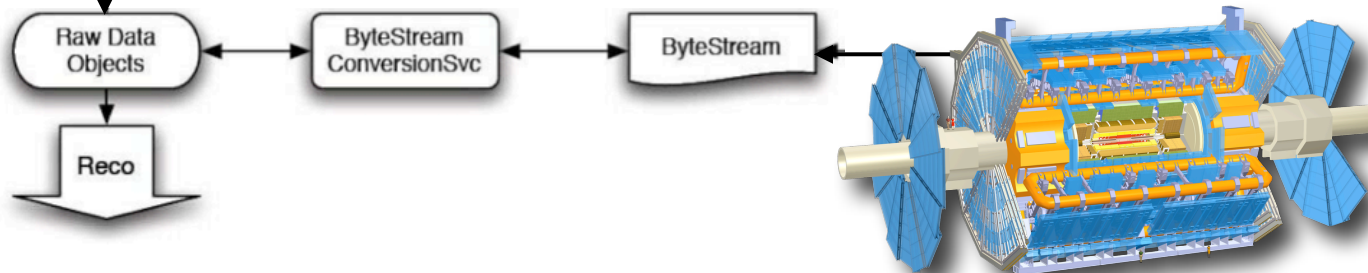
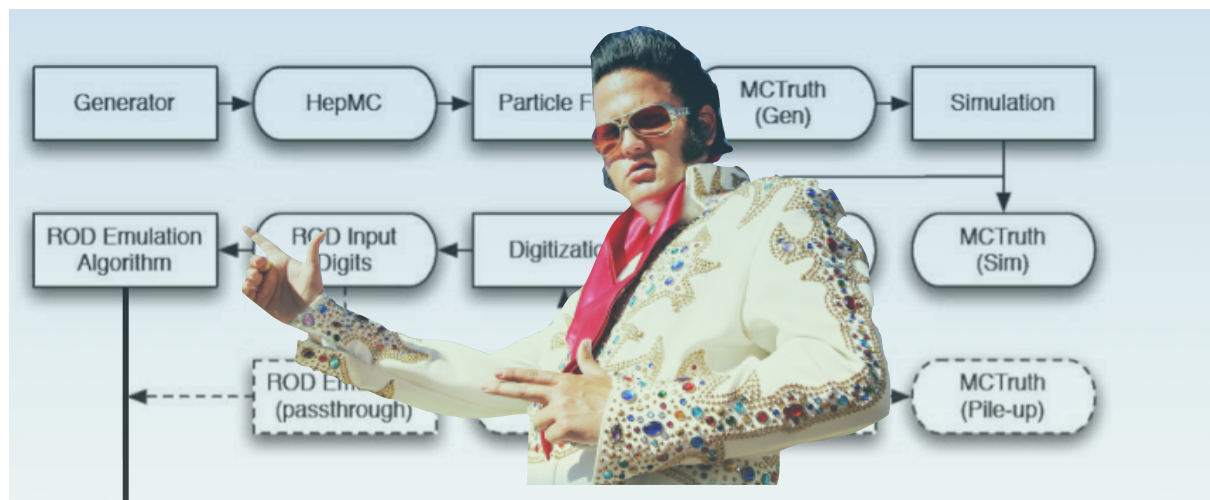
Generation: T1

Simulation: T2

Digitization: T2

Reconstruction: T2





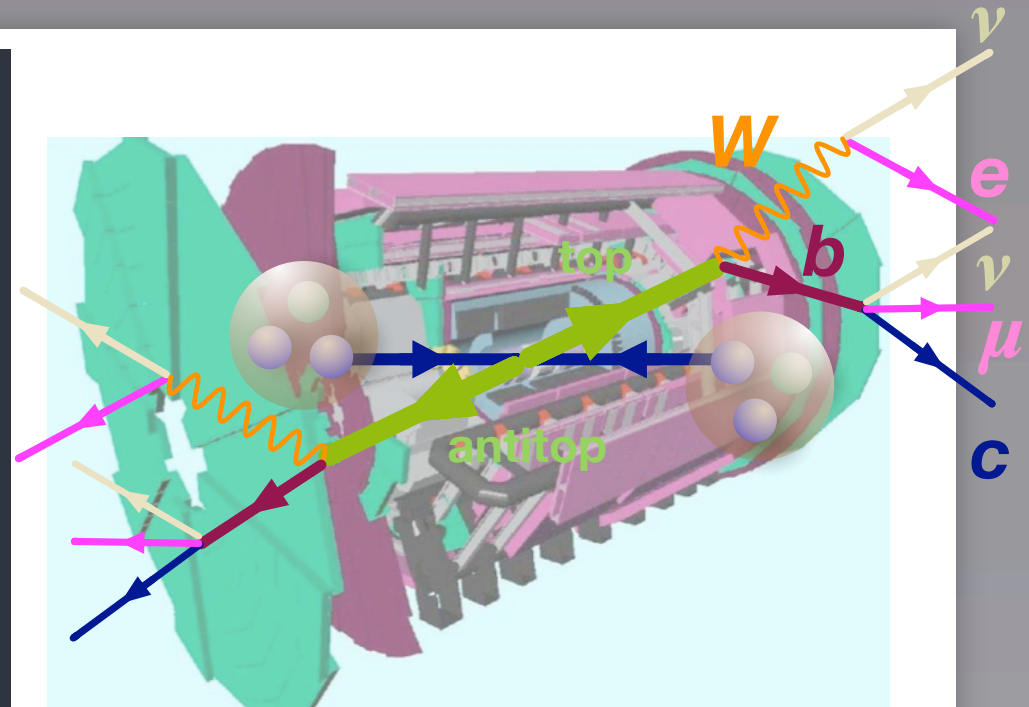
MC Impersonations look like data

simulation:
computationally
expensive

Generation

Simulation

Digitization



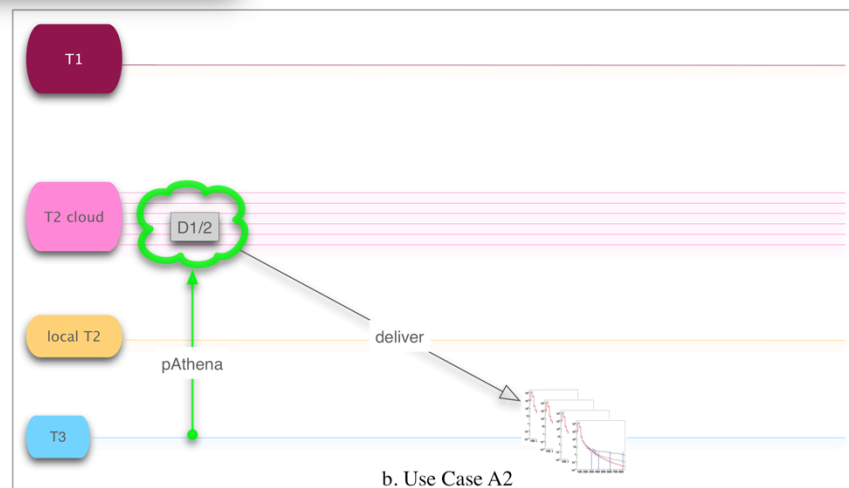
Sample	Generation	Simulation	Digitization
Minimum Bias	0.0267	551.	19.6
$t\bar{t}$ Production	0.226	1990	29.1
Jets	0.0457	2640	29.2
Photon and jets	0.0431	2850	25.3
$W^\pm \rightarrow e^\pm \nu_e$	0.0788	1150	23.5
$W^\pm \rightarrow \mu^\pm \nu_\mu$	0.0768	1030	23.1
Heavy ion	2.08	56,000	267

kSI2k-s !

1. Steady State Dataset Distribution
2. Dataset creation
3. Monte Carlo Production

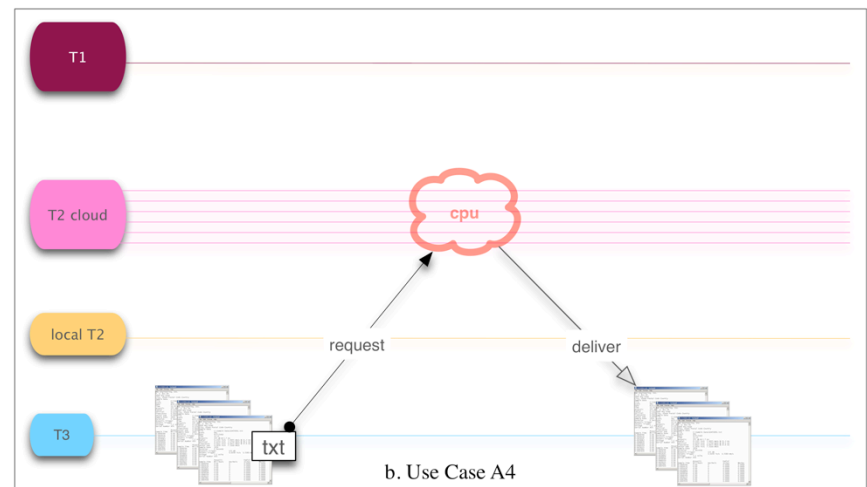
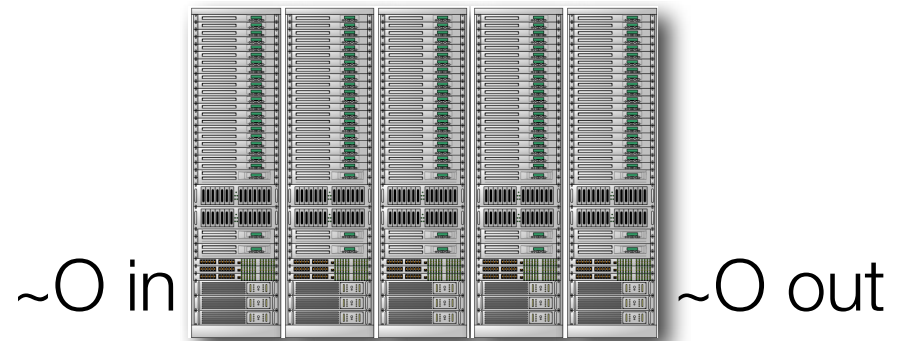
4. “Chaotic” User Analysis

“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



N.B. intensive calculations

Matrix Element calculations
many cpu-centuries of computation
grid has failed DØ for these
Multivariate combinations
COLLIE
Ensemble simulation



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?

How to predict this?

history is our only source of data

history = Fermilab tevatron

DØ and CDF: re-invented 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*

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



flexible and nimble

we have to plan for revolutions

“analysis”

is not remote

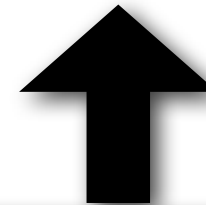
it's interactive...because things don't always work

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.

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

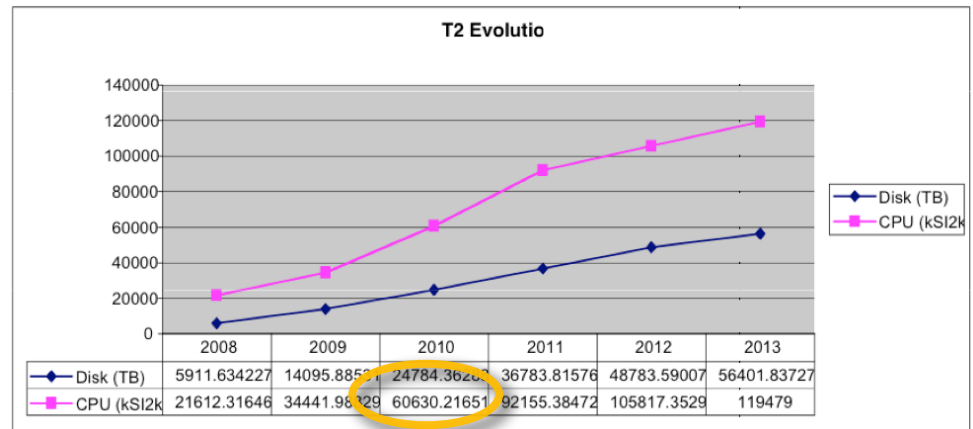
Sample	Generation	Simulation	Digitization	Reconstruction
Minimum Bias	0.0267	551.	19.6	8.06
$t\bar{t}$ Production	0.226	1990	29.1	47.4
Jets	0.0457	2640	29.2	78.4
Photon and jets	0.0431	2850	25.3	44.7
$W^\pm \rightarrow e^\pm \nu_e$	0.0788	1150	23.5	8.07
$W^\pm \rightarrow \mu^\pm \nu_\mu$	0.0768	1030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

Benchmark: $10\text{fb}^{-1} \rightarrow 2010 \rightarrow 2 \times 10^{33} \rightarrow 3.5$

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-IIF)	[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...heroic calculation

Tier 2 simulation for one year

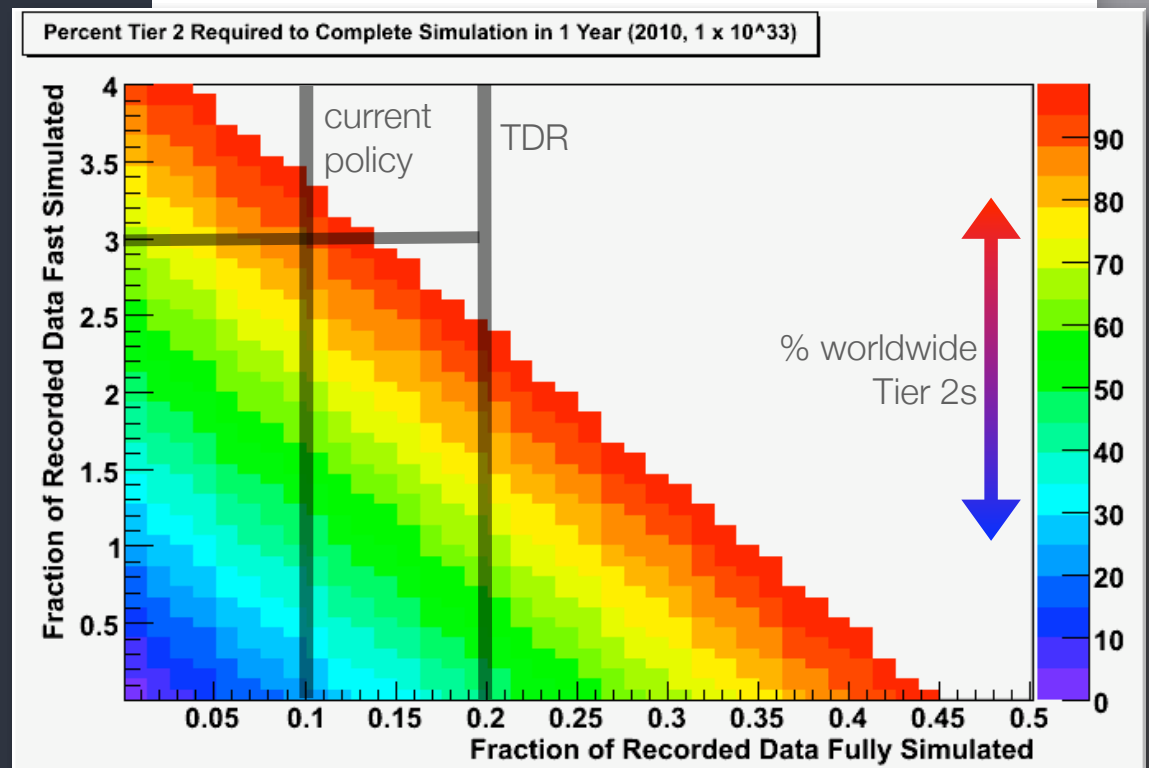
- ▶ horizontal axis:

fraction fully simulated

- ▶ vertical axis:

fraction fast-simulated

(ATLFAST-II...from Assamagan)



look.
scientific computing
planning
is hard

Administrators

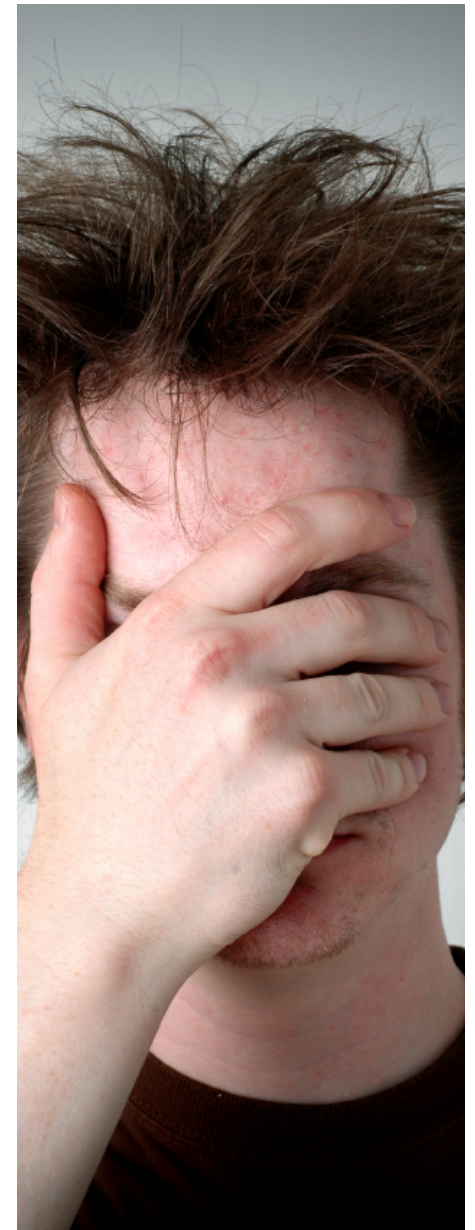
argue for funds against a plan

Users—have one thing in mind

not great about sticking to a plan

Physics analysis moves

faster than the best computing plans.



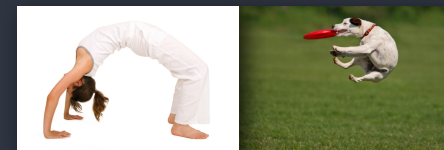
a U.S. computing model
totally reliant on
Tier 2s seems like a risk:

1. The Tier 2s may become overloaded.
2. History tells us to expect the unexpected.
3. *...stuff will happen.*



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.*

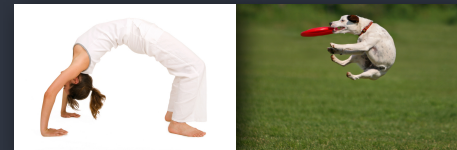


5 Primary Recommendations

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

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

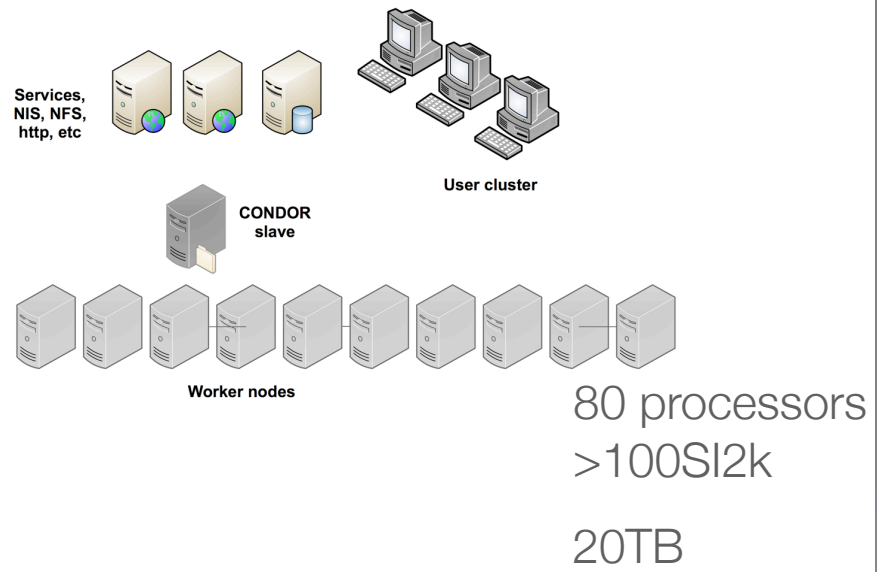
sample-sized, signal + background, ATLFast-II

few days

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

Tier 3 with "grid" connectivity
*a campus-based,
 tower or rack-based cluster
 minimal services required*

Characterized a strawman
 ~\$25k
ANL and Duke are building them



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!



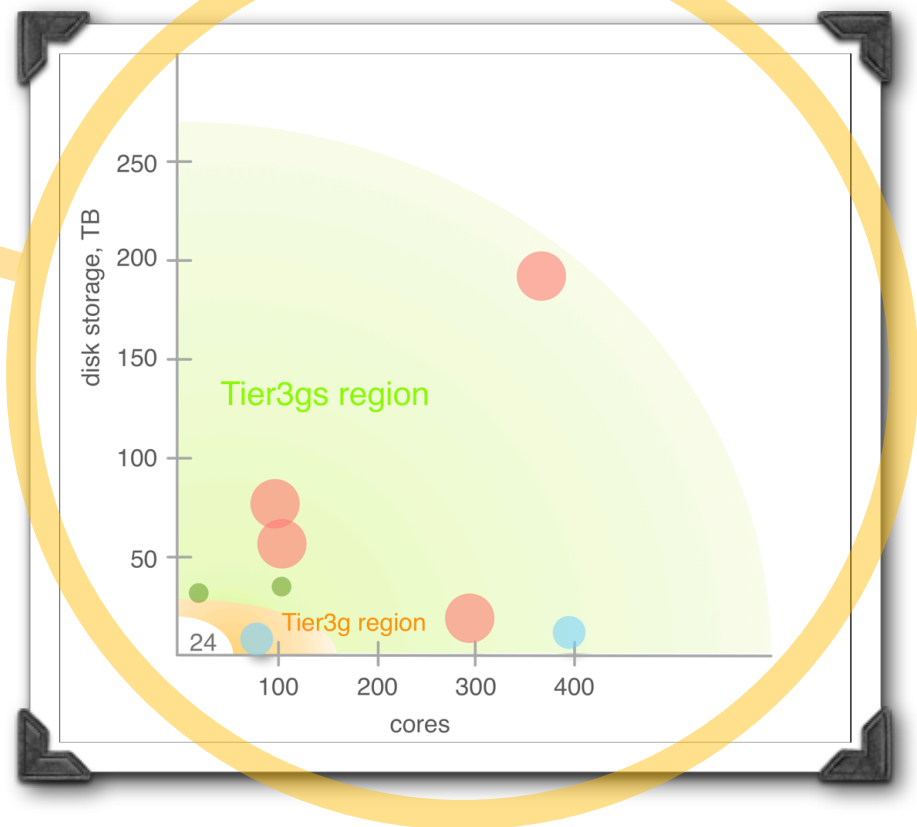
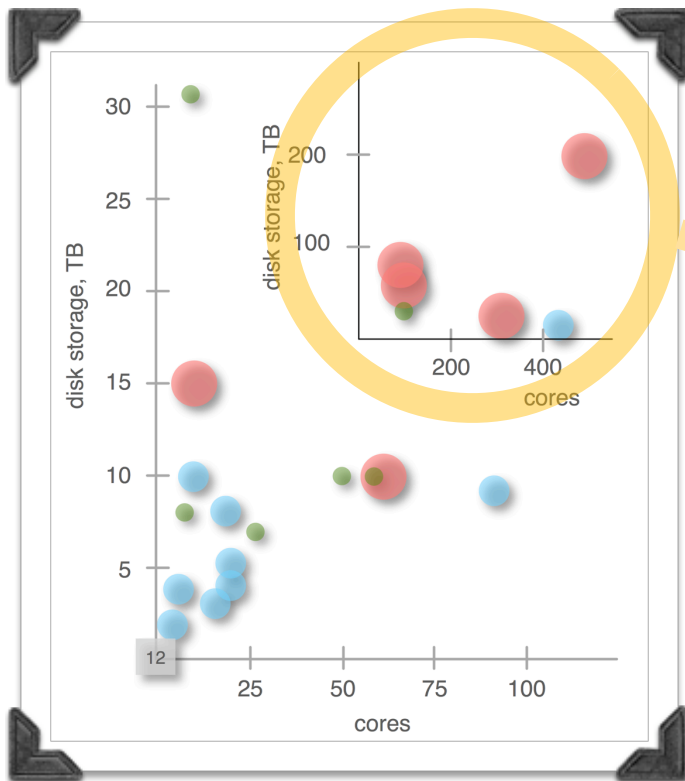
Partnership recommendation

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

conclusions



evolution

more depth will enhance



- ▶ Tevatron experience suggests:
 - “planning” is a process—the ground shifts
 - “analysis” is a highly-interactive activity “above” flattened rooftops
 - 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: deserve focused attention:

- ▶ Support model

personal, regular, common

- ▶ Access to the data for 2011-2012 milestones

target point-to-point minimal bandwidth—**Internet2 is raring to go.**

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