REVIEW OF RECENT DEVELOPMENTS IN ATLAS COMPUTING

ALDEN STRADLING, AMIR FARBIN

UNIVERSITY OF TEXAS AT ARLINGTON

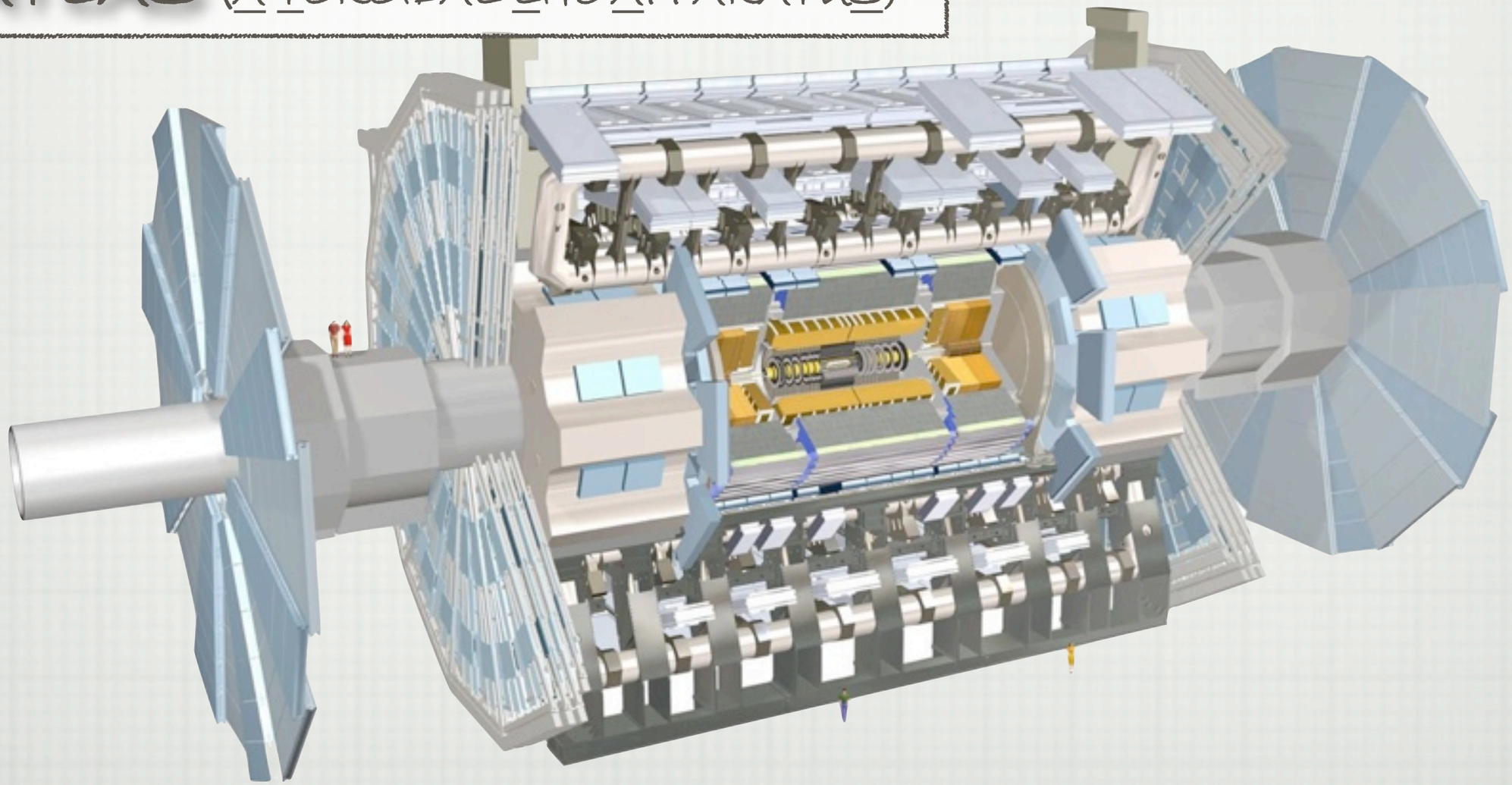
DPF 2009, WAYNE STATE UNIVERSITY

ON BEHALF OF THE ATLAS COLLABORATION

DESCRIBING RECENT DEVELOPMENTS

MEANS CREATING A CONTEXT

ATLAS (A TOROIDAL LHC APPARATUS)

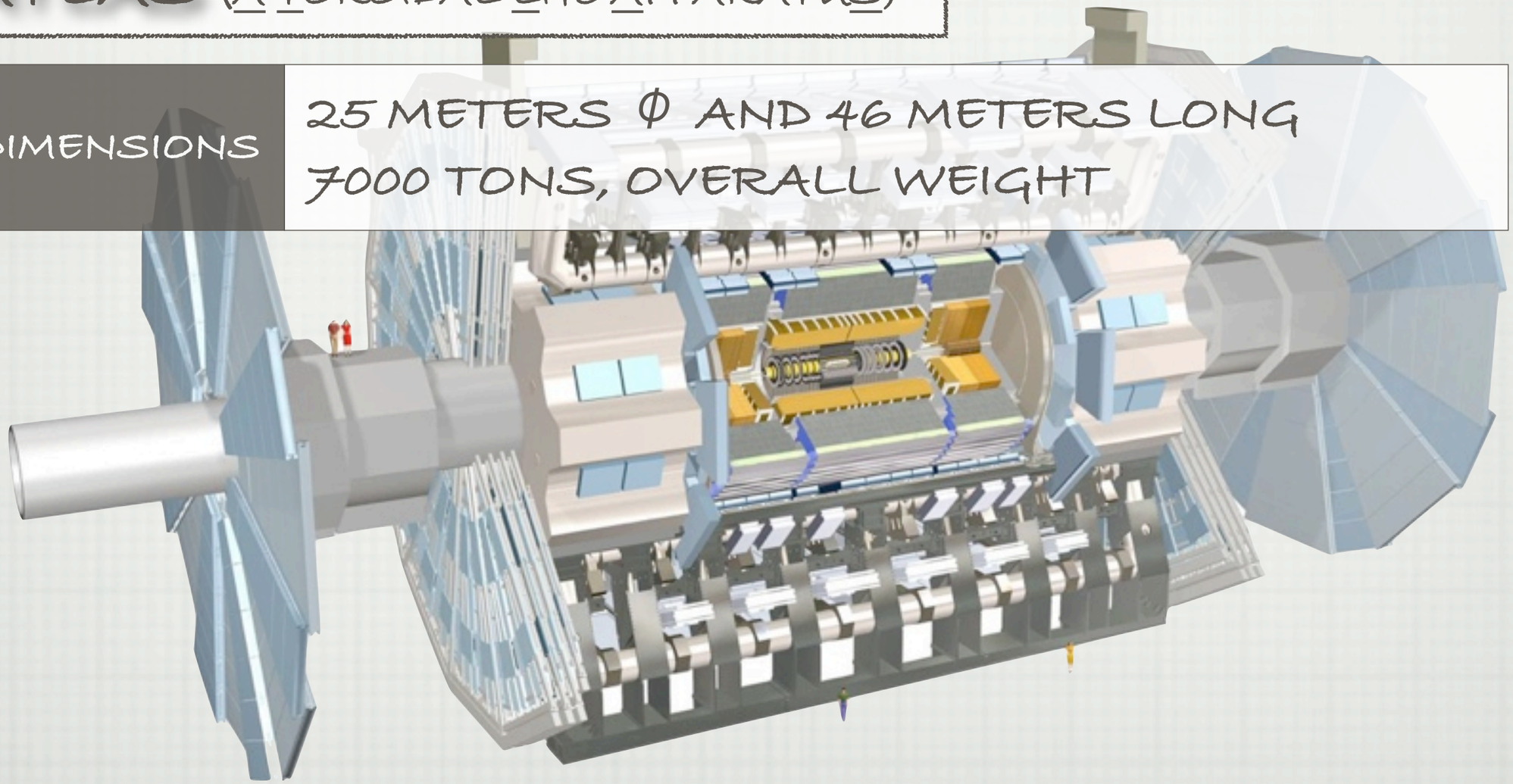


DESCRIBING RECENT DEVELOPMENTS

MEANS CREATING A CONTEXT

ATLAS (A TOROIDAL LHC APPARATUS)

DIMENSIONS	25 METERS Ø AND 46 METERS LONG 7000 TONS, OVERALL WEIGHT

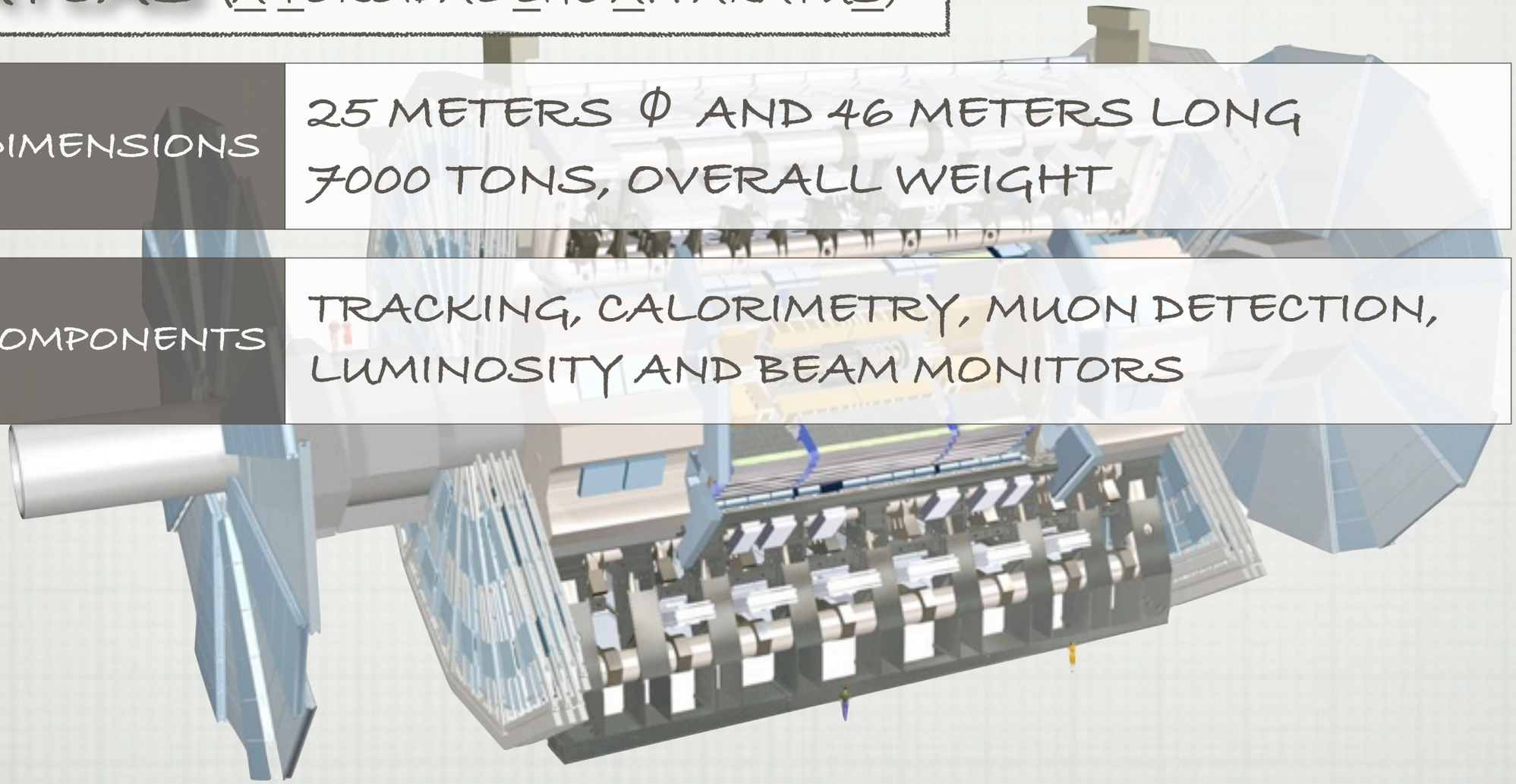


DESCRIBING RECENT DEVELOPMENTS

MEANS CREATING A CONTEXT

ATLAS (A TOROIDAL LHC APPARATUS)

DIMENSIONS	25 METERS Ø AND 46 METERS LONG 7000 TONS, OVERALL WEIGHT
COMPONENTS	TRACKING, CALORIMETRY, MUON DETECTION, LUMINOSITY AND BEAM MONITORS



DESCRIBING RECENT DEVELOPMENTS

MEANS CREATING A CONTEXT

ATLAS (A TOROIDAL LHC APPARATUS)

DIMENSIONS	25 METERS Ø AND 46 METERS LONG 7000 TONS, OVERALL WEIGHT
COMPONENTS	TRACKING, CALORIMETRY, MUON DETECTION, LUMINOSITY AND BEAM MONITORS
DATA CHANNELS	TRACKING – ~90 MILLION CALORIMETRY – 700 K MUONS – ~ 1.25 MILLION

TOTAL
~100 MILLION

DESCRIBING RECENT DEVELOPMENTS

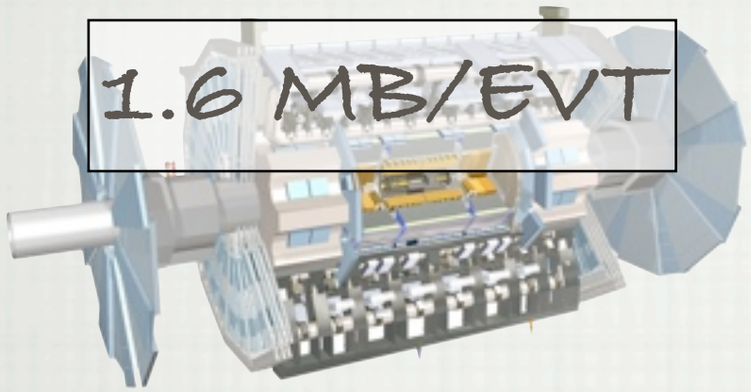
MEANS CREATING A CONTEXT

ATLAS (A TOROIDAL LHC APPARATUS)

DIMENSIONS	25 METERS Ø AND 46 METERS LONG 7000 TONS, OVERALL WEIGHT	
COMPONENTS	TRACKING, CALORIMETRY, MUON DETECTION, LUMINOSITY AND BEAM MONITORS	
DATA CHANNELS	TRACKING – ~90 MILLION CALORIMETRY – 700 K MUONS – ~ 1.25 MILLION	TOTAL ~100 MILLION
EVENT RATE	40 MHZ (25 NS EVENT SPACING)	

KEEPING EVENT RATES DOWN

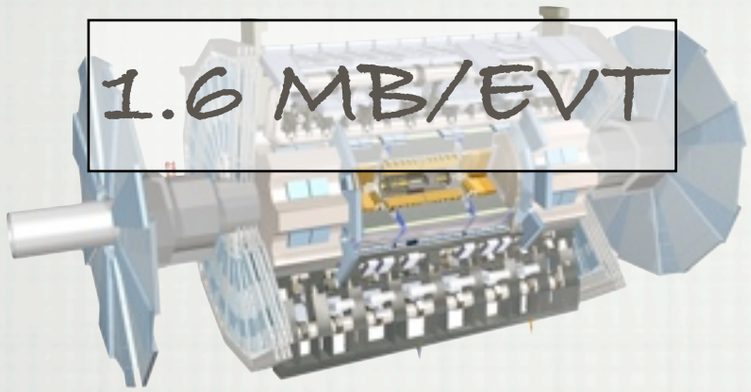
(OR: IF YOU CAN'T STORE IT, WHAT'S THE POINT?)



40 MHZ → 61 TB/S

KEEPING EVENT RATES DOWN

(OR: IF YOU CAN'T STORE IT, WHAT'S THE POINT?)



40 MHZ

61 TB/S

UNMANAGEABLE!

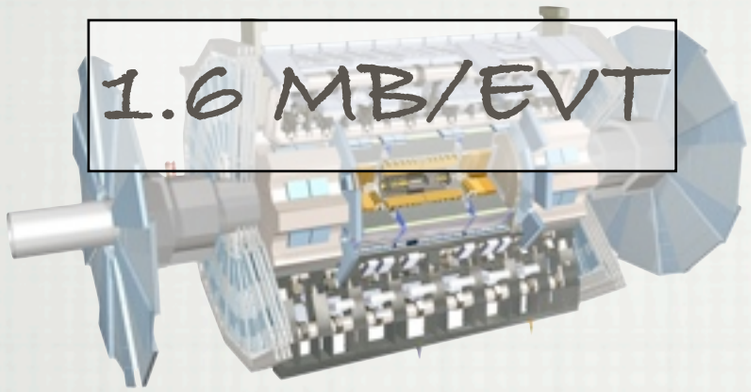
LEVEL-1 TRIGGER

75 KHZ

117 GB/S

KEEPING EVENT RATES DOWN

(OR: IF YOU CAN'T STORE IT, WHAT'S THE POINT?)

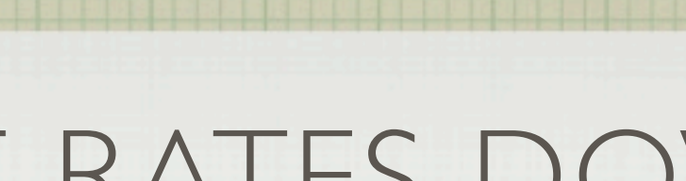


40 MHZ

61 TB/S

UNMANAGEABLE!

LEVEL-1 TRIGGER



75 KHZ

117 GB/S

STILL TOO MUCH

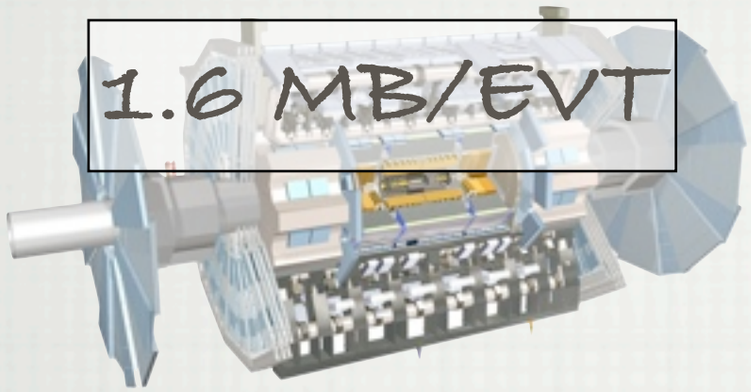
LEVEL-2 TRIGGER

2 KHZ

3 GB/S

KEEPING EVENT RATES DOWN

(OR: IF YOU CAN'T STORE IT, WHAT'S THE POINT?)



40 MHZ

61 TB/S

UNMANAGEABLE!

LEVEL-1 TRIGGER

75 KHZ

117 GB/S

STILL TOO MUCH

LEVEL-2 TRIGGER

2 KHZ

3 GB/S

NEARLY...

EVENT FILTER

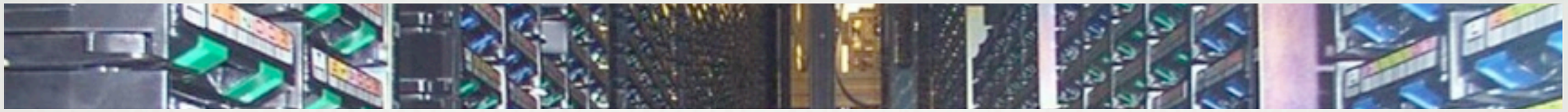
200 HZ

FINALLY.

320 MB/S

SQUEAKING BY

(THOSE EVENTS ADD UP FAST)



- 320 MB/S IS 3,200 TB/YEAR FOR ATLAS
  - THAT'S JUST FOR THE RAW DATA. LIMITING FACTOR
  - 303 DAYS TO TRANSFER VIA GIGABIT AT PEAK RATES
- DERIVED FORMATS ARE CREATED
  - ESD – EVENT SUMMARY DATA. STORED IN ROOT FILES AS PERSISTIFIED C++ OBJECTS. LARGE (500 KB). INTENDED FOR DETECTOR STUDIES
  - AOD – ANALYSIS OBJECT DATA. ALSO PERSISTIFIED C++ OBJECTS, BUT REDUCED TO QUANTITIES SUITABLE FOR PHYSICS ANALYSIS. MUCH SMALLER – 100 KB PER EVENT

TO MAKE MATTERS WORSE

(RESOURCES GET EVEN TIGHTER)

- THOSE NUMBERS DON'T TAKE INTO ACCOUNT THE NEED FOR REPROCESSING.
  - RECALIBRATION OF THE DATA MUST BE DONE REPEATEDLY
  - SEVERAL DATA STREAMS ARE CREATED FOR VARIOUS TASKS
- THERE'S SIMULATION, AS WELL
  - PRESENTLY PLANNING TO FULL SIM ONLY 10% OF THE ATLAS DATA
  - TO COMPENSATE, FAST (PARAMETRIZED) SIMULATION OF 3X THE DATA
  - STILL A GARGANTUAN TASK – SINGLE EVENTS CAN TAKE HOURS
  - SEVERAL STAGES – REQUIRES MORE STORAGE THAN SIMPLE DATA

CENTRAL PLANNING

EGGS AND BASKETS

- CENTRAL FACILITY IS TEMPTING
  - SHORT PATHS, FAST NETWORKS
  - NO NEED FOR DUPLICATION OF SERVICES AT MULTIPLE SITES
- BUT IT'S ALSO PROBABLY IMPRACTICAL
  - FOR VARIOUS REASONS WHICH I WON'T GET INTO, BUT WHICH ARE NOT ALL TECHNICAL

GRID MODEL

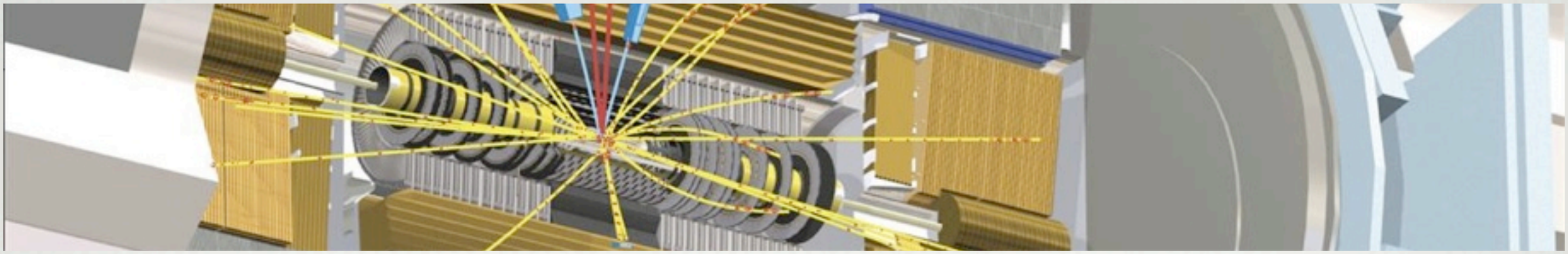
COMPUTING ON DEMAND

- DISTRIBUTE THE LOAD
  - SHARE A SECURITY FRAMEWORK, TRACK DATA ACROSS RESOURCES
  - PERMIT USERS TO RUN TASKS WITHOUT CARING ABOUT SPECIFICS
  - SHARE RESOURCES WITH OTHER COLLABORATIONS AND FIELDS
  - TAKE ADVANTAGE OF LULLS IN COLLEAGUES' DATA PROCESSING
  - REDUNDANCY IS INHERENT TO THE SYSTEM
- BUT... IT'S EXTREMELY DIFFICULT
  - FLEXIBLE ENOUGH TO USE, SECURE ENOUGH TO INSTALL
  - DATA MANAGEMENT IS A TREMENDOUS PROBLEM

DRIVING PHYSICISTS TO TIERS

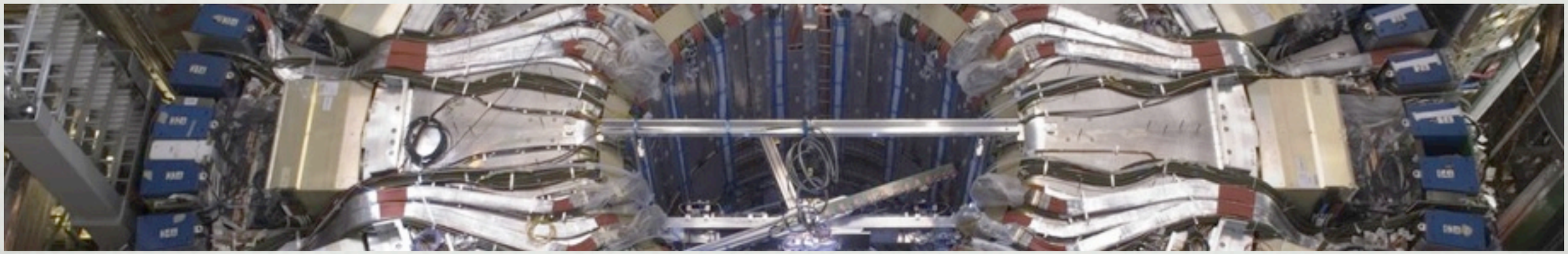
A COMBINED APPROACH – CENTRAL TO DISTRIBUTED

- TIER-0 – THE CENTRAL FACILITY AT CERN
  - HANDLES OUTPUT FROM THE EXPERIMENT (320 MB/S)
  - RECONSTRUCTS RAW DATA AND CREATES ESD, AOD, REPLICATES THEM TO TIER-1 FACILITIES (1020 MB/S)
- TIER-1 – REGIONAL FACILITIES
  - 10 SITES IN NORTH AMERICA, EUROPE AND TAIWAN
  - STORAGE AND MANAGED DATA REPROCESSING/ANALYSIS
  - MAINTAIN COPIES OF AOD DATA ON REGIONAL TIER-2 CLOUD
- TIER-2 – ANALYSIS AND SIMULATION
  - HANDLES THE SIMULATION LOAD FOR THE COLLABORATION
  - ACCEPTS GRID-BASED USER ANALYSIS JOBS
- TIER-3 - LOCAL INTERACTIVE PROCESSING



ENOUGH HISTORY.

HOW IS ALL THIS WORKING OUT?



DATA PROCESSING

DATA MANAGEMENT

TILTING AT WINDMILLS. LONG STORY.

- IN ATLAS, THE PROJECT IS CALLED DON QUIJOTE (DQ2)
  - LONG-TIME PRIORITY OF THE GRID DEVELOPMENT EFFORT
  - IN CHARGE OF ALL DATA MOVEMENT AND TRACKING ON THE ATLAS GRID
- SKIM OVER THE PROBLEMS
  - UPDATED CATALOG, UNIQUENESS
  - RELIABLE TRANSFERS AND FAULT TOLERANCE
  - USABILITY AND USER INTERFACE (NOT POSIX)

CENTRAL PRODUCTION SYSTEM

KEEPING THINGS MOVING WITH PanDA

- MOST SIMULATED DATA ARE HAD IN COMMON
  - EVENT PRODUCTION IS DONE BY RECIPE – CAREFULLY CHECKED
  - LARGE OFFICIAL RUNS ARE CREATED AND MAINTAINED IN THE CENTRAL PRODUCTION SYSTEM (PRODSYS).
- PRODSYS SHIFTERS KEEP THE SYSTEM FULL, VALIDATE RESULTS, AND TROUBLESHOOT
  - ALL ATLAS PRODUCTION NOW RUN WITH THE PANDA PRODUCTION SYSTEM

PANDA, DESCRIBED

- PULL-BASED SYSTEM
  - JOBS ARE DEFINED IN A CENTRAL DATABASE
  - SITES RUN GENERIC "PILOT JOBS" THROUGH THEIR QUEUES
  - PILOT JOBS CHECK LOCAL CONDITIONS ON THE RESOURCE AND ASK THE DATABASE FOR A JOB THAT MATCHES
  - PILOT THEN RUNS THE JOB ON THE RESOURCE IN A SECURE ENVIRONMENT*, AND CHECKS THE OUTPUT INTO DATA MANAGEMENT (DDM)
- CENTRAL MANAGEMENT NECESSARY
  - ALLOWS QUOTAS AND FAIR SHARES TO BE MANAGED BY ATLAS, RATHER THAN BOTHERING LOCAL SITE ADMINS WITH COMPLEX AND TIME-CONSUMING IMPLEMENTATIONS

* FORTHCOMING

USER ANALYSIS

THE OPPOSITE OF A MANAGED SYSTEM

- USER ANALYSIS IS CHAOTIC, BY DEFINITION
  - RESEARCHERS EXPLORE IN ALL DIRECTIONS – UNPREDICTABLE
  - IMPATIENT, ESPECIALLY AROUND CONFERENCES
  - NO PLANNED, SEQUENTIAL DATA ACCESS PATTERN MAKES TAPE STORAGE LESS PRACTICAL
  - THOUSANDS OF INDIVIDUALS WITH DIFFERENT REQUIREMENTS
- USERS NEED SUPPORT
  - EXISTS ALREADY – THE ATLAS DAST (DISTRIBUTED ANALYSIS SUPPORT TEAM) HANDLES USER DIFFICULTIES AND MONITORS SITE CONDITIONS IN CLOSE COORDINATION WITH THE DIST. ANALYSIS DEVELOPERS

EVOLUTION OF TOOLS

USER ANALYSIS IN FLUX

- TWO INTERFACES, CONVERGING
  - GANGA
    - USES EUROPEAN AND NORDIC GRID SITES, CROSS-COLLABORATION
    - CAN EMPLOY A VARIETY OF BACKENDS: LOCAL BATCH SYSTEMS OF VARIOUS SORTS, AND ALL OF THE ATLAS GRID FLAVORS (INCLUDING PANDA)
  - PANDA
    - USES THE SAME BACKEND AS THE PRODUCTION SYSTEM – STABLE AND WELL-MAINTAINED
    - ATLAS-SPECIFIC – LESS DEPENDENCY, FASTER TURNAROUND

INTERACTIVITY AND TIER-3

WHAT ANALYSTS WANT

NOT NECESSARILY WHAT IS POSSIBLE…

- ☐ IN A PERFECT WORLD, USERS WOULD:
  - ☐ DO AN INSTANT ANALYSIS
  - ☐ OVER ANY DATA
  - ☐ IN WHATEVER STAGE OF PROCESSING
  - ☐ ON THEIR LAPTOP

- ☐ IN <u>OUR</u> WORLD:
  - ☐ RAW DATA ARE ALL BUT INACCESSIBLE
  - ☐ DETAILED DETECTOR DATA ARE HUGE (ESD)
  - ☐ EVEN THE FULL AOD WILL NOT BE AVAILABLE AT A SINGLE TIER-2

PROPOSED TIER-3 CLUSTERS

- SCALE: 1-10 PEOPLE (A LOCAL HEP GROUP)
  - CONSUMES MOSTLY USER ANALYSIS DPD FILES (MORE ON THAT LATER) FROM TIER-2
  - COMMODITY HARDWARE, PERHAPS EMPLOYING VIRTUAL MACHINES FOR CONFIGURATION AND SOFTWARE DELIVERY
  - GRID SOFTWARE (AND PERHAPS SERVICES) INSTALLED LOCALLY
- GOALS:
  - ADEQUATE PERFORMANCE AT REASONABLE PRICE
  - EASY TO MAINTAIN

DATA FORMATS AND PRACTICALITY

EVENT DATA MODEL (EDM)

(OR: THE COMPROMISES, AS THEY STAND)

RDO
RAW DATA OBJECT

RAW CHANNELS
1.6 MB/EVENT

ESD
EVENT SUMMARY DATA

RECON. OUTPUT
500 KB/EVENT
CELLS, HITS, TRACKS, CLUSTERS, ETC. FOR CALIBRATION

AOD
ANALYSIS OBJECT DATA

ANALYSIS DATA
100 KB/EVENT
'LIGHTWEIGHT' TRACKS, CLUSTERS, ELECTRONS, JETS, MUONS...

TAG

TAG DATA
1 KB/EVENT
EVENT SUMMARY FOR SELECTION: TRIGGER, JETS, 4 BEST ELECTRONS

$D^{N}PD$
DERIVED PHYSICS DATA

$D^{(1,2,3)}PD$ FILES
10-20 KB/EVENT
SPECIFIC VARIABLES FOR ANALYSES AND GROUPS

WAIT A MINUTE...

- THOSE "DERIVED" FORMATS IMPLY THAT ~2000 PHYSICISTS AGREED...
  - THAT THEY COULD ALL USE THE SAME SUBSET OF DATA!
  - EITHER NOT SPACE-OPTIMIZED, OR A LOT OF PEOPLE LEFT OUT
- THERE'S TENSION HERE BETWEEN:
  - HARSH SPACE AND CPU REQUIREMENTS
  - SCIENTIST'S DESIRE TO RECORD EVERYTHING AND KEEP IT FOREVER

ORIGIN OF THE SPECIES

OF DATA FORMATS

- ARCHIVES OF THE ATLAS SOFTWARE ARE LITTERED WITH DEFUNCT DATA FORMATS
  - SOME EFFORTS CREATED COMMON FORMATS BY PHYSICS GROUP – BECAUSE GROUPS OFTEN NEED A VARIETY OF CHANNELS. THESE EFFORTS SEEM TO BE LANGUISHING
  - ANYTHING THAT IS ACTUALLY USED, HOWEVER, IS VERY HARD TO PRY FROM USERS' FINGERS.
- SURVIVORS: THE ONES PEOPLE USED
  - OF THE $D^{N}PD$ FORMATS, ONLY THE $D^{3}PD$ AND THE "PERFORMANCE" DPD SEEM TO BE GETTING ANY TRACTION
  - PERFORMANCE DPD CONTAINS THE NECESSARY CALIBRATION DATA
  - UPCOMING "PHYSICS" DPD MAY ADDRESS PRESENT SPACE CONCERNS

MOVING FROM GROUPS TO SIGNATURES

John Hancock

- A SUSY OR HIGGS ANALYSIS DEPENDS ON A VARIETY OF DATA TYPES
  - MUON, ELECTRON, MISSING E_T, TRACKING, CALORIMETRY
  - MOST PARTS OF THE AOD CAN'T BE RELIABLY "THINNED" AWAY, BECAUSE AT LEAST ONE ANALYSIS WILL NEED THAT ONE
- CREATING CHANNEL-SPECIFIC "PHYSICS" DPD FILES CAN ALLOW FOCUSED THINNING OF THE AOD, AND PERHAPS SAVE SOME SPACE
  - USERS CAN PICK AND CHOOSE AS NECESSARY FROM THE DPD SETS
  - WE'LL SEE HOW THE UPTAKE ON THIS IDEA GOES

RESOURCE ESTIMATION

RESOURCE AVAILABILITY

BECAUSE IT'S GOOD TO PLAN AHEAD

"No battle plan survives contact with the enemy"
Helmuth von Moltke

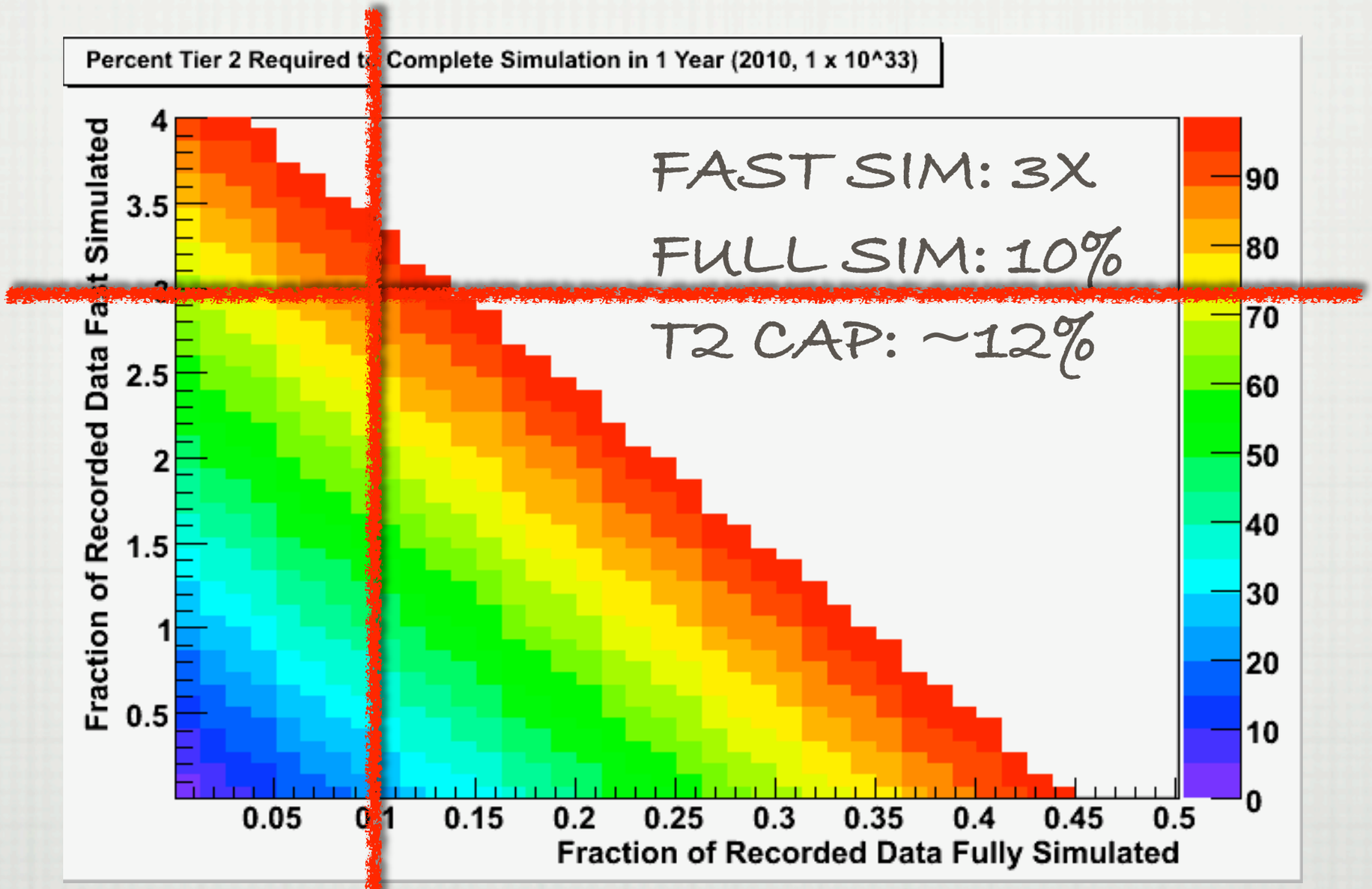
- CLEARLY, THE REQUIREMENTS FOR ATLAS COMPUTING ARE A MOVING TARGET
  - FORMAT SIZES AND REQUIREMENTS CHANGE (USUALLY UPWARD), AS DO COMPUTING CAPABILITIES. OTHER EXPERIMENTS KNOW WELL.
  - SCIENTISTS FIND UNFORESEEN AND COSTLY NEW WAYS TO USE THE RESOURCES YOU BOUGHT FOR THE CALCULATIONS THEY NEED
- SINCE THE REQUIREMENTS CHANGE GREATLY IN THE FACE OF REALITY, MUST BE SURE TO MODEL THE SYSTEM FLEXIBLY

EVALUATING THE MODELS

- DONE BY AMIR FARBIN, UT ARLINGTON
- CALCULATING FIGURES OF MERIT IN ATLAS COMPUTING TASK CHAINS (GEN, SIM, RECO)
  - LOTS OF COMPETING OPTIONS! ATHENA VS. ROOT, POOL VS. "FLAT", FRAMEWORKS VS. MAKECLASS, PYTHON VS. C++, VARIOUS DATA FORMATS
  - ANY AND ALL OF THESE WILL BE NECESSARY TO ADDRESS SPECIFIC TASKS AND PROBLEMS. YOU CAN'T PICK A WINNER IN ADVANCE
- SIMULATION MUST BE SIMPLE, WITH FLEXIBLE INPUTS
  - RUN OVER A RANGE OF SCENARIOS QUICKLY, LOOK FOR OPTIMIZATIONS

TIER 2: FAST VS. FULL SIMULATION

AS AN EXAMPLE



EXPLORING CHANGES TO THE ATLAS ANALYSIS MODEL

- ☐ TIMING
  - ☐ LOOK AT THE EFFECTS OF TRANSFER SPEED AND LATENCY
  - ☐ LONGER VS. SHORTER JOBS, LARGER VS. SMALLER FILES
  - ☐ CALCULATE COST OF USING ONE TOOL VS. ANOTHER
- ☐ FAILURE RATES
  - ☐ EXPLORE THE EFFECTS, DECIDE WHICH ONES TO TREAT FIRST
  - ☐ BUILD IN EFFICIENCY FACTORS, ADJUST TO OBSERVATION WHEN THE SYSTEM IS RUNNING
  - ☐ A MATURE SYSTEM WILL ALLOW MANAGEMENT TO TUNE AND OPTIMIZE ATLAS COMPUTING – BOTH THE MODEL AND DAY-TO-DAY POLICY

SUMMARY

AND THANKS FOR YOUR ATTENTION

- ☐ ATLAS COMPUTING MANAGES PROCESSING AND DISTRIBUTION OF MORE THAN 20 TB/DAY
- ☐ GRID AND CENTRAL RESOURCES MAKE IT POSSIBLE TO GET THESE DATA TO THE USER
- ☐ THE ANALYSIS AND DATA MODELS ARE STILL EVOLVING, AND WILL CONTINUE TO DO SO
- ☐ LOCAL TIER-3 RESOURCES WILL PLAY A LARGE ROLE
- ☐ PLANNING TO ACCOMMODATE FUTURE TASKS MUST BE AGILE AND ACCURATE, AND TOOLS ARE UNDERWAY TO DO SO

FIN.