BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

HEP Detector Simulation and Analysis

Craig E. Tull, Ph.D.
Staff Scientist/Group Leader
Science Software Systems Group
Computational Research Division
Berkeley Lab Computing Sciences

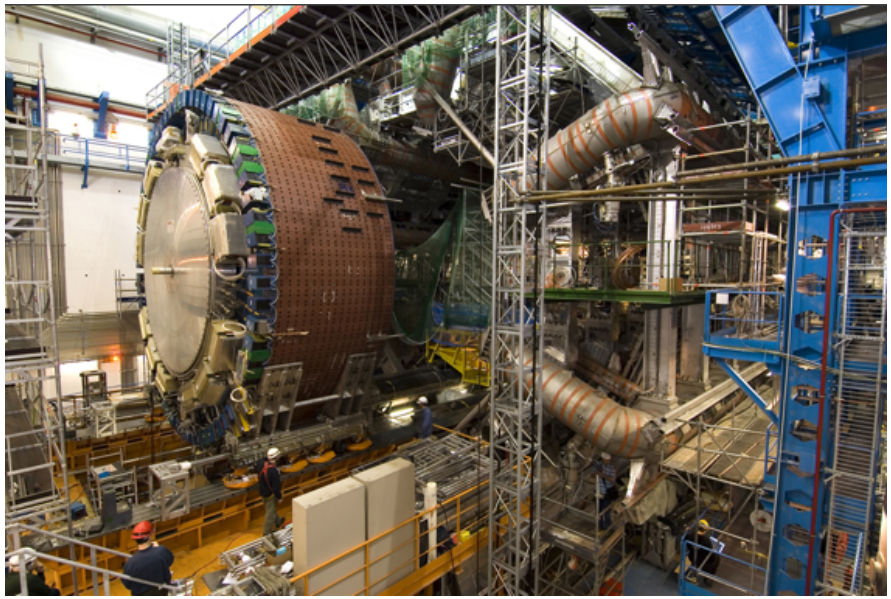
November 12, 2009
Rockville, MD

1. Detector Simulation and Analysis Overview

Summarize the projects in your science area and their scientific objectives for the next 3-5 years

- Current and past users of NERSC:
  - **ATLAS - LHC accelerator at CERN, Geneva (PI: Ian Hinchliff)**
  - **Daya Bay - Nuclear reactor Neutrino detector in China (PI: Kam-Biu Luk)**
  - CDF - Tevatron accelerator at FNAL (PI: Wei Min Yao)
  - BOSS - Baryon Oscillation Spectroscopic Survey
  - JDEM/SNAP - Supernova satellite (PI: Saul Perlmutter)
  - BaBar - PEP-II collider at SLAC, Stanford
  - SNF - SuperNova Factory (PI: Greg Alderige)
- Future Users:
  - Super-B - B physics experiment in Italy (BaBar follow on)
- Nuclear Physics: (Not in the scope of my talk)
  - STAR, ALICE, KamLand, IceCube, Majorana, etc.
- 1200 HEP & NP users of PDSF past and present

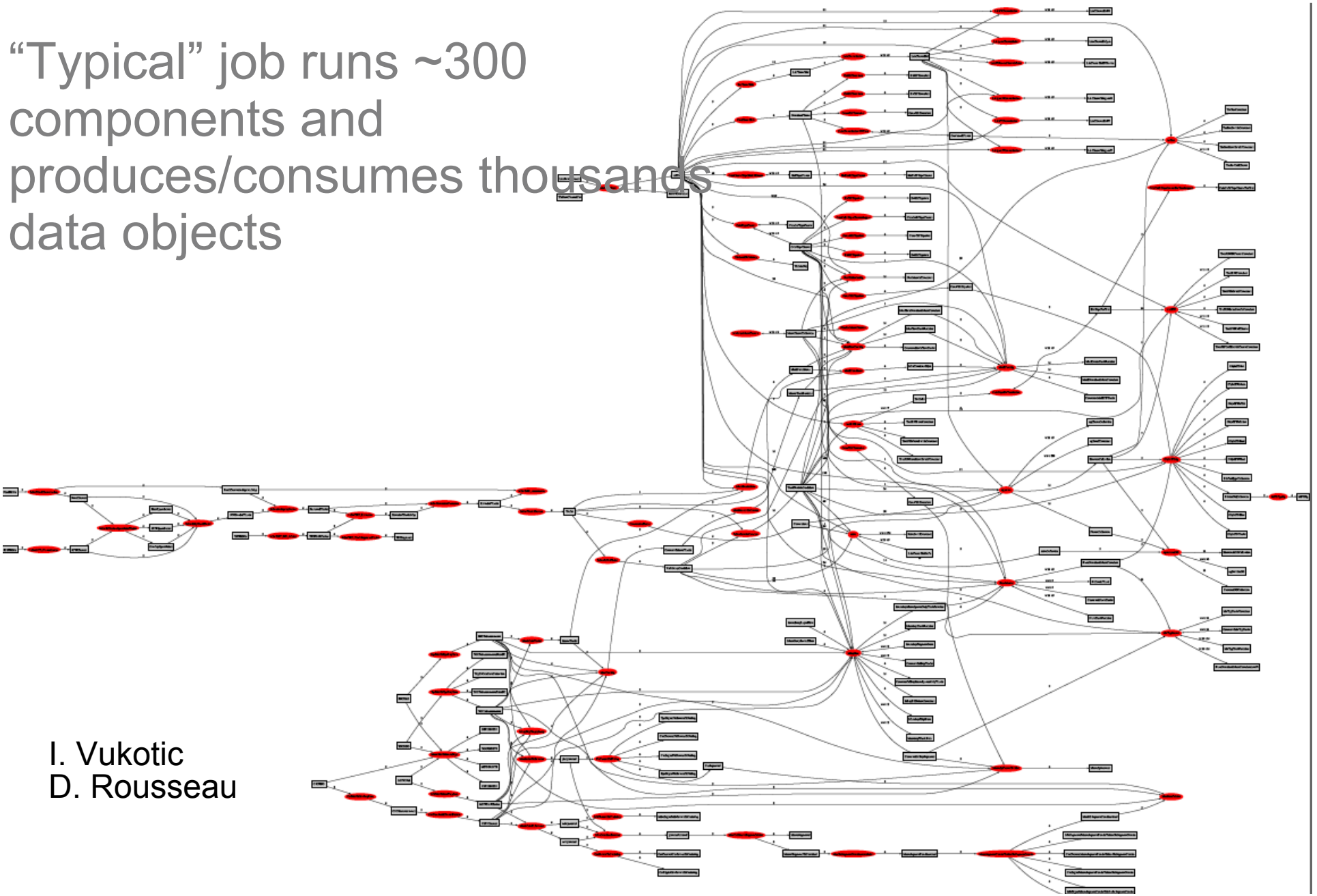
ATLAS - CERN



- Large, complex detector
  - ~10^8 channels
- Long lifetime
  - Project started in 1992, **first data Nov 15, 2009**, last data 2029?
- 320 MB/sec raw data rate
  - ~3-5 PB/year
- Large, geographically dispersed collaboration
  - 2800 people, 169 institutions, 37 countries
  - Most are, or will become, software developers
    - Programming abilities range from Wizard to Neophyte
- Scale and complexity reflected in software
  - 1500 C++ packages, 3000 components, 15,000 C++ classes, 8,100 Python configuration files, 2,100 python modules.
  - Most code is algorithmic (written by physicists). Growth over last 3 years tremendous.
  - Core Software is written by professionals (LBNL 50%).
    - 84 C++ packages, 285 components, 1,000 C++ classes, 800 python modules/scripts.
    - Core software is run in every job. Physics software is pick-and-choose.
  - Provide robustness but plan for evolution
  - Requires enabling technologies
  - Requires management & coherency

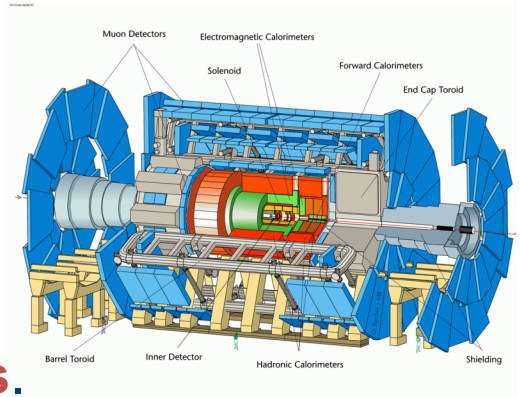
Event Reconstruction Dataflow

"Typical" job runs ~300 components and produces/consumes thousands data objects



I. Vukotic
D. Rousseau

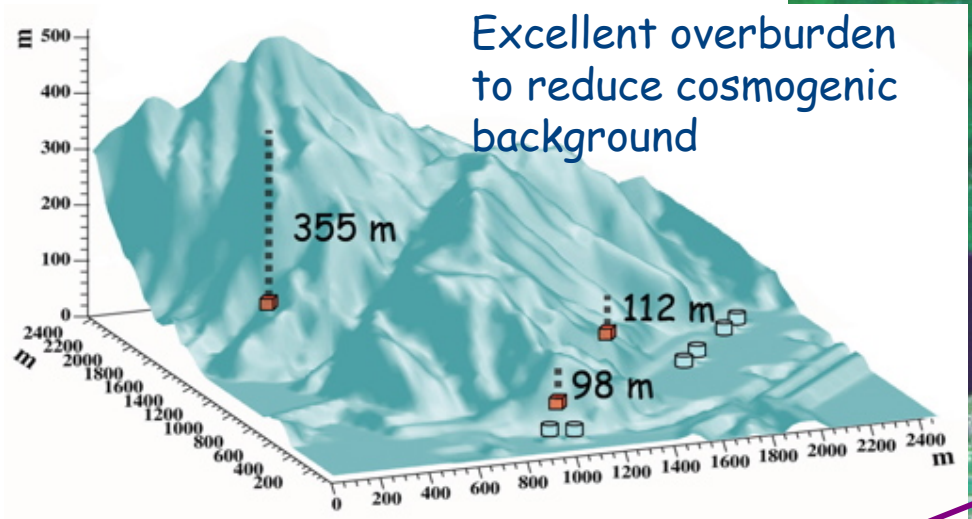
ATLAS Next 5-Years



- After many delays, initial data taking starting in **days**.
  - Lower energy and luminosity. Likely lower data volumes.
  - Initially, exploration of data, evaluation, calibration, and understanding of the detector will take some time.
  - What ATLAS lacks in data volume will more than be made up for in enthusiasm for real beam data.
- At full DAQ rate, ATLAS @ NERSC will simulate and analyze select physics processes.
- The primary purpose of the detector will be studies of the origin of mass at the electroweak scale, therefore the detector has been designed for sensitivity to the largest possible Higgs mass range. The detector will also be used for studies of top quark decays and supersymmetry searches.

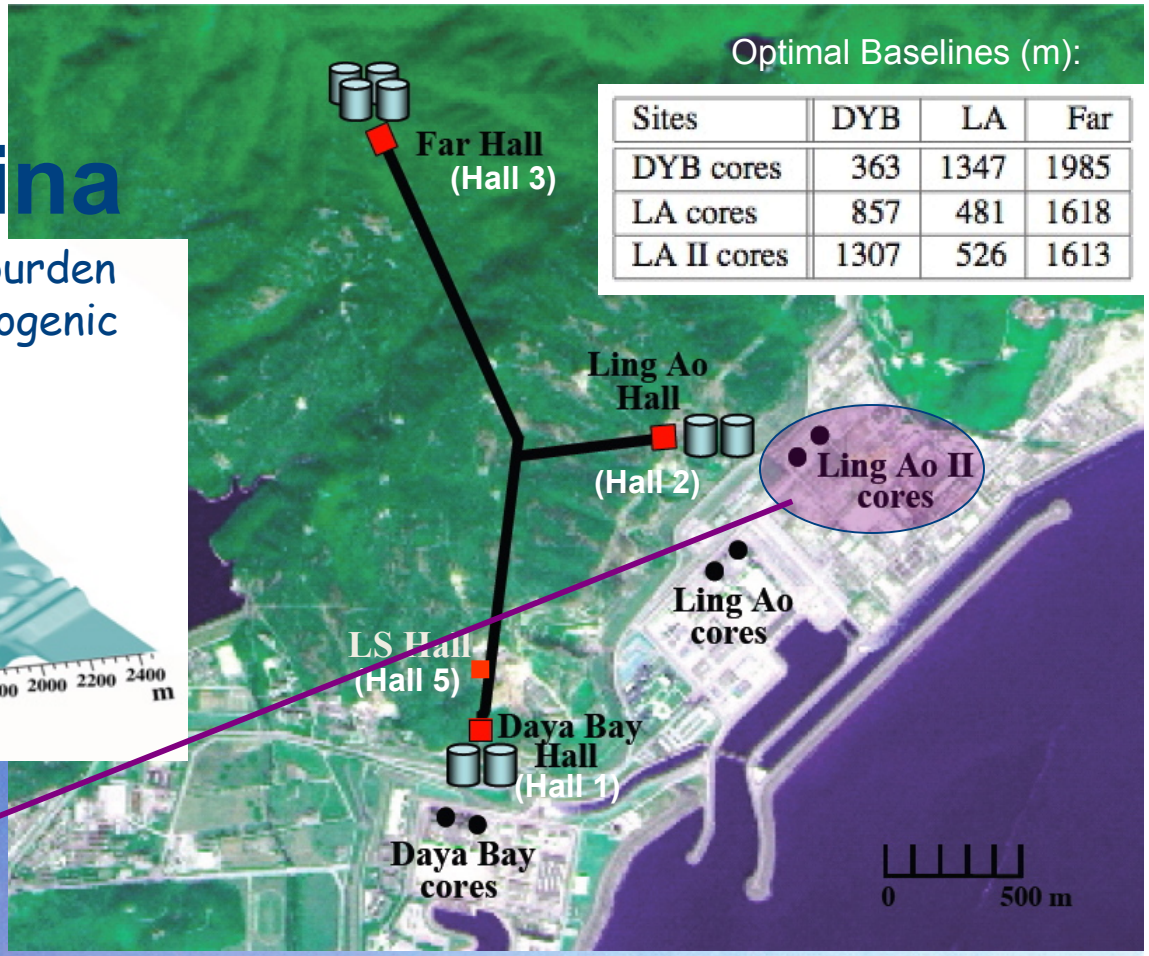
Daya Bay - China

Excellent overburden to reduce cosmogenic background



Optimal Baselines (m):

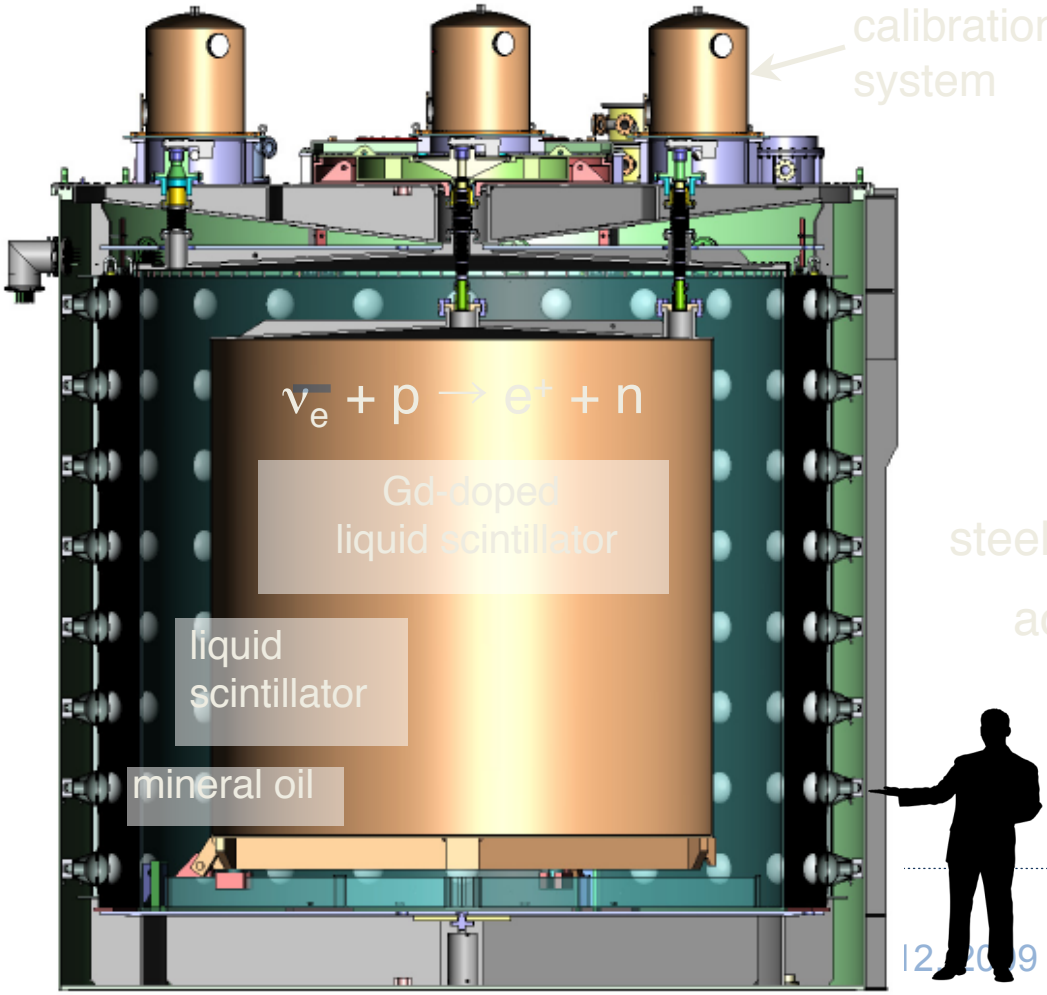
Sites	DYB	LA	Far
DYB cores	363	1347	1985
LA cores	857	481	1618
LA II cores	1307	526	1613

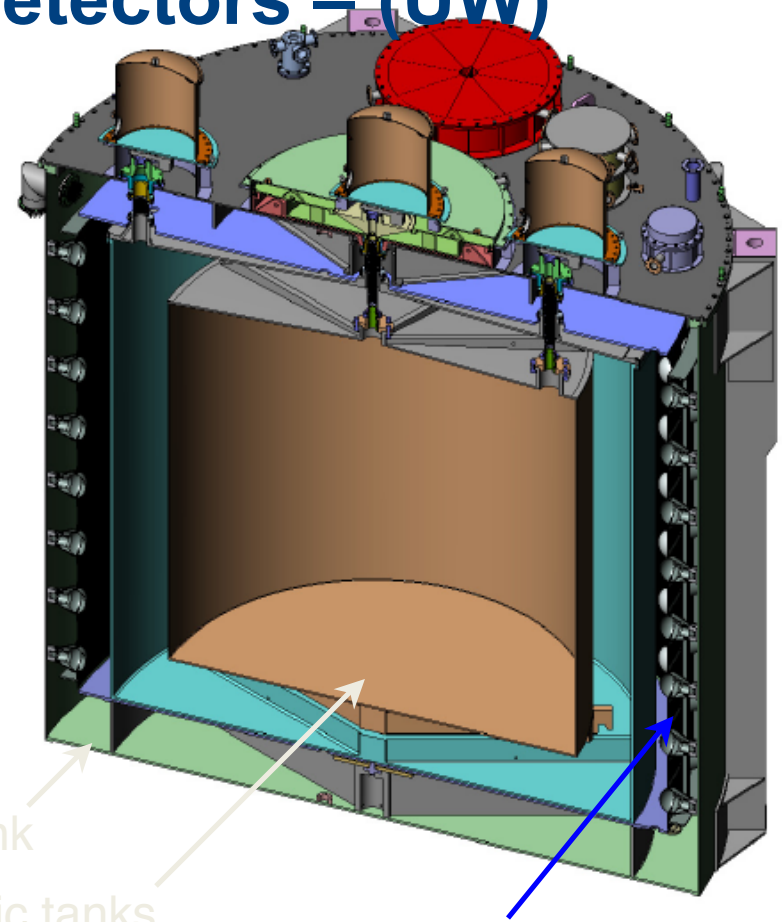


Powerful $\bar{\nu}_e$ source:
Current: 11.6 GW_{th}
2011: 17.4 GW_{th}

1.1 Antineutrino Detectors – (UW)

- 8 "identical", 3-zone detectors
- no position reconstruction, no fiducial cut

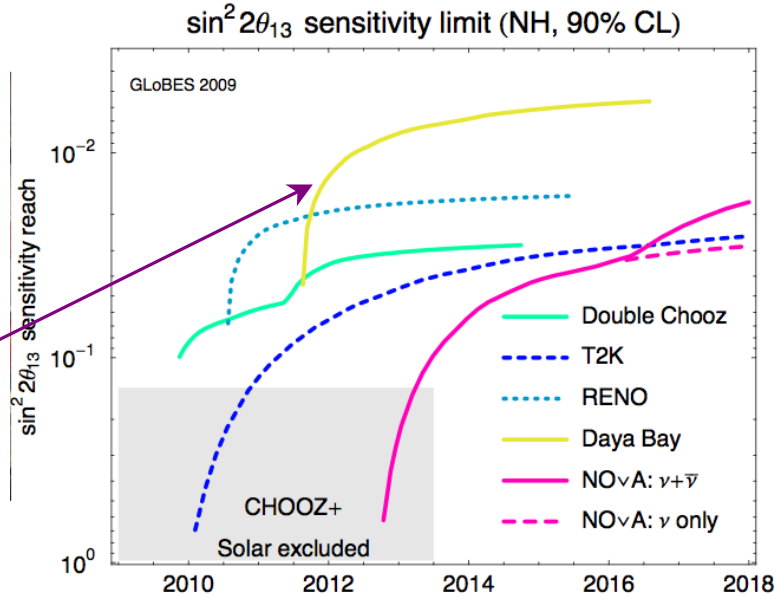




target mass:	20t per detector
detector mass:	~ 110t
photosensors:	192 PMTs
energy resolution:	12%/√E

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

Daya Bay - Schedule



- **Highly aggressive, success-oriented**
  - CD1-CD3b = 14 months
- In our first 6 months of data taking, Daya Bay will have world's best sensitivity to $\sin^2(2\theta_{13})$
  - Physics ready on day 1
- With minimal time for design or development, evaluation, adoption, and extension of state-of-the-art system was our only option.
- Adoption of components and systems from ATLAS, IceCube, MINOS, BaBar allowed us to focus on Daya Bay-specific extensions and developments. Scientists were able to focus on detector design and science instead of software.
- Daya Bay is typical of a small-medium ($34M US scope) future HEP project.

NERSC 2009 Configuration

Large-Scale Computing System
Franklin (NERSC-5): Cray XT4
- 9,740 nodes; 38,960 cores
- ~25 Tflops/s sustained SSP (355 Tflops/s peak)

NERSC-6 planning is underway

Clusters

Bassi (NCSb)
- IBM Power5 (888 cores)

Jacquard (NCSa)
- LNXI Opteron (712 cores)

PDSF (HEP/NP)
- Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)

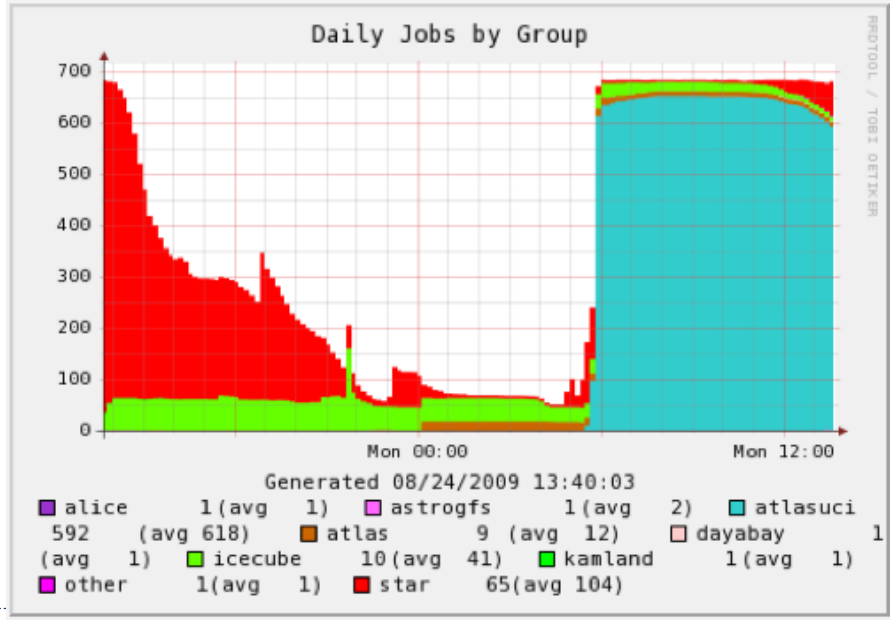
230 TB; 5.5 GB/s

HPSS Archival Storage
- 44 PB capacity
- 10 Sun robots
- 130 TB disk cache

Analytics / Visualization
- Davinci (SGI Altix)

PDSF Fair Share - truly shared resource

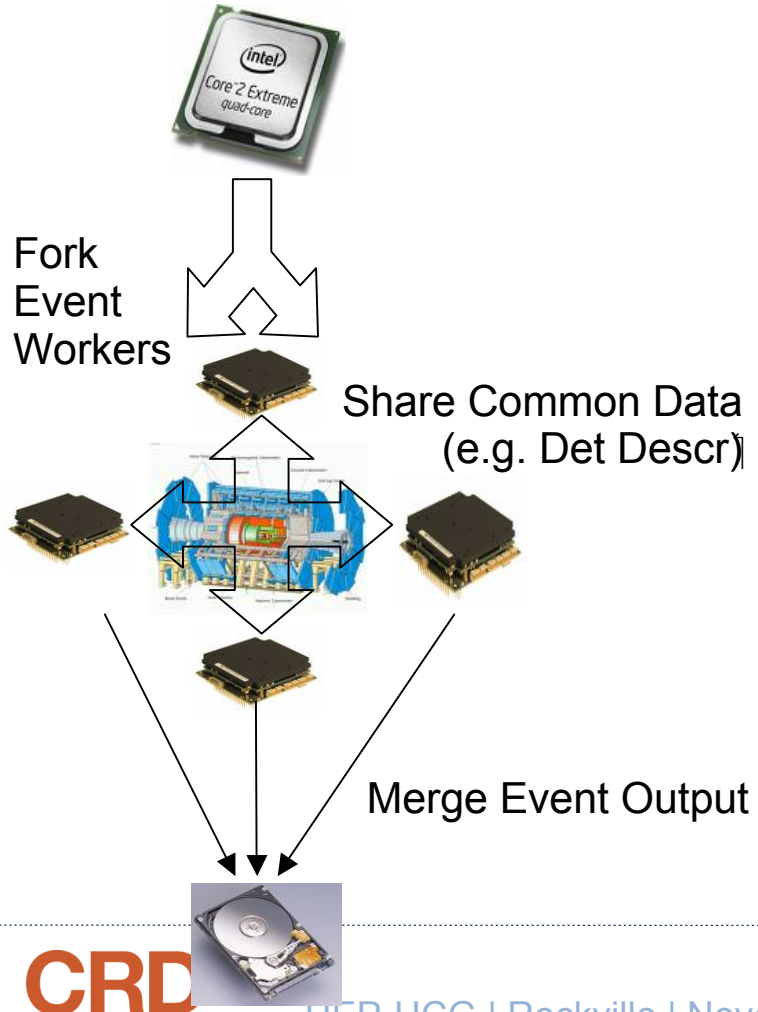
- 10-15 active groups, ~350/1100 active users
- Contributors to PDSF (STAR, ATLAS, Daya Bay, KamLand, …) get guaranteed access to their fair share of the resources.
- Non-contributers can get access to spare cycles.
  - This normally amounts to a small sliver of CPU.
  - There are opportunities for agile projects.
- CPUs have 3 year life span
- Disks have 3-5 year life span
- At current size ~$200k would replace retiring resources.



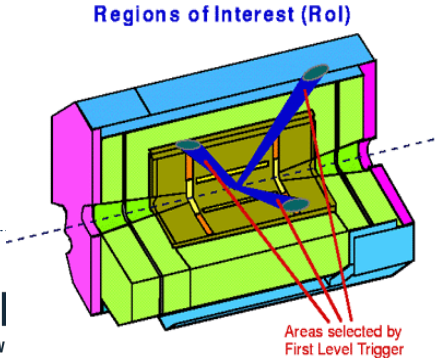
2. Current HPC Requirements

- Architectures
  - Primarily loosely coupled Linux (SLC common), some Mac OS X and even windows for desktop development. **Must** synchronize with wider collaborations' supported platforms.
    - Use of CHOS & CERNVM critical
- Compute/memory load
  - RAM varies tremendously, but up to >2 GB/core
  - ATLAS & Daya Bay - 100s of PDSF cores running 24/7 growing over time, constrained by budget realities. (~2-3 M CPU-hr)
- Data read/written
  - Heavy use of large disk and HPSS for data storage.
    - Daya Bay ~150 TB/year raw, simulated, processed data
    - ATLAS @ NERSC currently 60 TB, increasing 50 TB/year
  - Heavily data-intensive. I/O impacted by both raw data volumes and software flexibility and configurations.

Effective Multicore

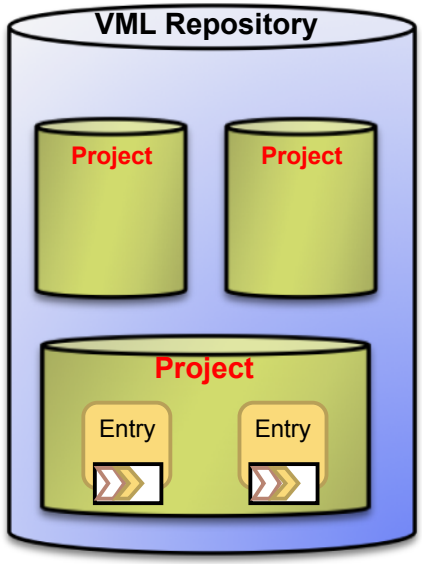


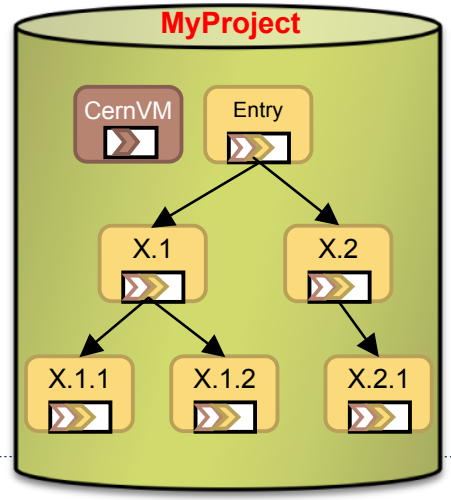
* **AthenaMP (2-8 cores)**
  - Lightweight, process-based, event parallelism
  - Memory-optimized: use fork() to share memory automatically
* **Many-core (>16) challenges**
  - Reduce memory footprint, maximize sharing (memory bandwidth)
  - Optimize disk I/O, especially event merging
  - Restructure reconstruction algorithms so that they can be parallelized



Virtual Machine Logbook (VML)

- A tool to organize and share virtual machines
  - Space-optimized to improve start-up times, and save disk space and network bandwidth
    - One full entry containing ATLAS reconstruction job <1GB
      - Compare to 7GB distribution kits
    - Differential, and “domain-optimized” entries can be as small as 10MB
- Builds on CERNVM and libVirt projects
  - goal: technology independence





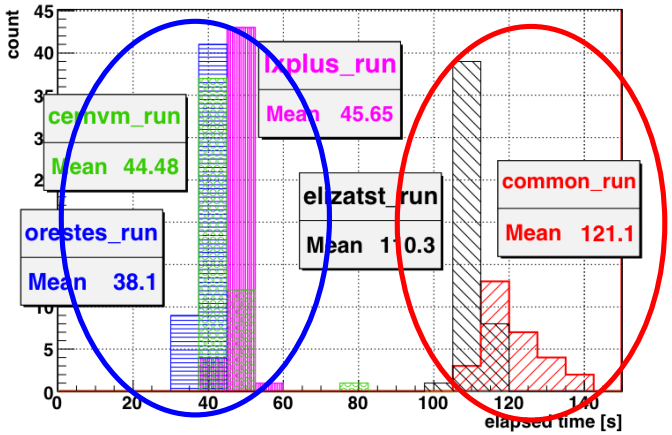
Testing Performance of CernVM FS and GPFS on PDSF

- PDSF users complain GPFS is slow to run/develop ATLAS
- CernVM Filesystem can be installed on PDSF to server ATLAS software
- We try to compare CernVM FS with GPFS and see which is better
- Note: In this slide we are only working with the CernVM filesystem, not the virtual machine.

Test I: CVMFS vs GPFS

Box1	VS	Box2
2 GHz CPU VMware+CernVM ATLAS on CVMFS		A Node in PDSF 2 GHz CPU ATLAS on GPFS

Time Needed for Executing ATLAS Job



CVMFS GPFS

Chart above shows CVMFS is 3 times faster than GPFS when running ATLAS Jobs.
However, above tests are not done with identical conditions, so we need Test II.

Test 2: CVMFS vs GPFS in the same box

Test Box: A Node in PDSF		
ATLAS on FUSE+CVMFS	VS	ATLAS on GPFS

The above setup will give more accurate results.
If the result is positive, we can deploy CVMFS across PDSF nodes to improve ATLAS performance.

Experts from PDSF have helped us to setup FUSE/CVMFS on one PDSF node.
Tests are in progress.

2. Current HPC Requirements

- Necessary software, services or infrastructure
  - PDSF software provided as "modules" in conjunction with "chos"
  - Some Grid services and infrastructure are needed for eg. STAR, ATLAS. Reliant on Open Science Grid.
  - All projects need global & international accessibility (both ways).
  - Support for Virtualization (CERNVM) will be needed in future.
    - ATLAS "Tier 3 in a box" and Daya Bay "NuWa in a box"
  - Heavy reliance on open-source software, with little use of commercial packages. (IDL for Astro is a notable exception, Objectivity for BaBar was another, as is Oracle at CERN)
  - Python is widely used for interactivity and configuration.
  - MySQL & other RDBs used for many purposes.
  - Most of our codes are run in both Batch and Interactive modes. Therefore interactive nodes for more than compiling are critical.

2. Current HPC Requirements

- Current primary codes and their methods or algorithms
  - Gaudi, Athena, NuWa: C++ simulation and analysis frameworks
  - GEANT4: C++ geometry, detector, material, particle simulation engine
  - ROOT: C++ analysis toolkit and framework
  - XRootD: distributed I/O and communication
  - dCache: distributed file system (not in use at NERSC)
  - **Methods (Algorithms) are extremely varied. From simple calibration calculations, to Kalman filters, neural nets, Baysian statistics, 3-D sparse pattern matching, etc.**
  - **Almost all are "event" independent => Inter-process communication is not needed for most detectors.**
  - Legacy PAW & GEANT3 & CERNLIB still in use

2. Current HPC Requirements

- Known limitations/obstacles/bottlenecks
  - File system performance is critical, as is Networking.
  - Currently these codes do not run on the big MPP resources at NERSC. Due to both portability, and to DOE/NERSC policy.
    - With the advent of virtualization, policy is the only obstacle.
  - ATLAS runs effectively on Multi-core, but grappling with Many-core issues.
- Anything else?
  - Huge collaborations require formalized infrastructure of their own
  - HEP experiments frequently ask for, and collaborate with NERSC to stand up collaboration-wide services.
  - Science Data Gateways are an excellent example.
  - This is above and beyond CPU-hrs and TBs, but can have a profound effect on science productivity.

3. HPC Usage and Methods for the Next 3-5 Years

- Upcoming changes to codes/methods/approaches
  - Continually evolving, but manycore and virtualization are the main projected changes. Though the advent of GPU or other non-standard architectures may change things.
- Changes to Compute/memory load
  - As data accrues, CPU required for full passes increases.
  - New computational techniques could increase both - but dependent on capabilities.
- Changes to Data read/written
  - ATLAS and Daya Bay are ramping up in the next 12 months, CDF is ramping down.
  - Exploration of non-ROOT I/O may alter patterns.
- Changes to necessary software, services or infrastructure
  - Impossible to predict, IMHO

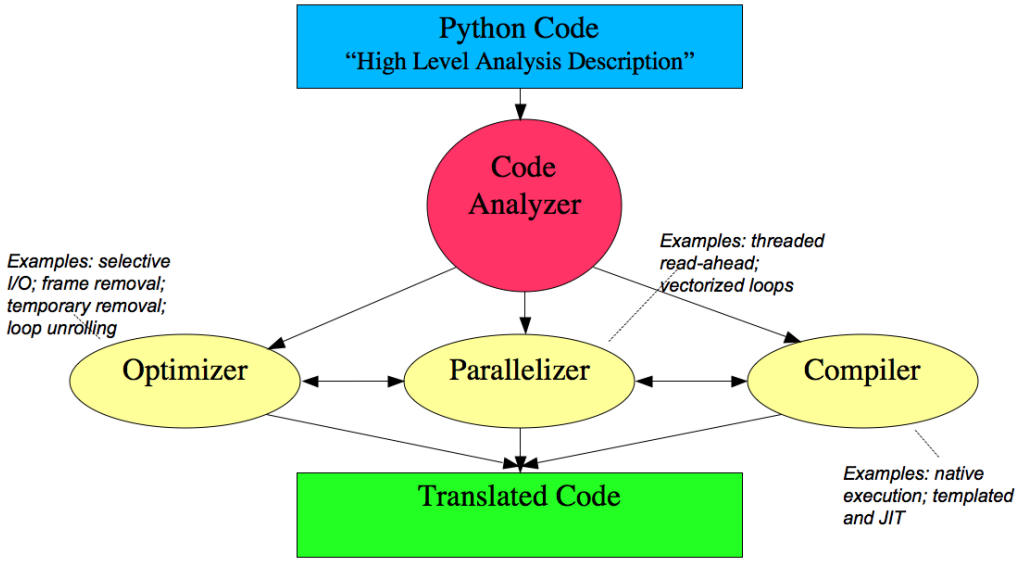
3. HPC Usage and Methods for the Next 3-5 Years

- Anticipated limitations/obstacles/bottlenecks on 10K-1000K PE system.
  - Balance of CPU vs Disk I/O would be my main concern.
- Strategy for dealing with multi-core/many-core architectures
  - PyROOT optimization (VIPER)
    - JIT Python compilation, on the spot parallelization
  - Efficient Multicore exploitation
    - CRD leading ATLAS (and LHC) in this R&D work
      - athenaMP: process-based task farm, exploit Linux COW
        - » ~20% extra shared-memory, ~20% more jobs in multicore farms
        - » athenaMP ~production quality. Starting R&D for many core

PyROOT & Viper: Python Analysis Optimization

- **Today: PyRoot**
  - Very popular ROOT shell
  - Python/C++ API, reflection-based
  - Developed and actively maintained by one LBL ATLAS collaborator
  - Known to be used in >35 projects

- **Tomorrow: Viper**
  - Python code is translated into lower level, more static code
  - Generate code optimized for target platform (e.g. *multicore)*



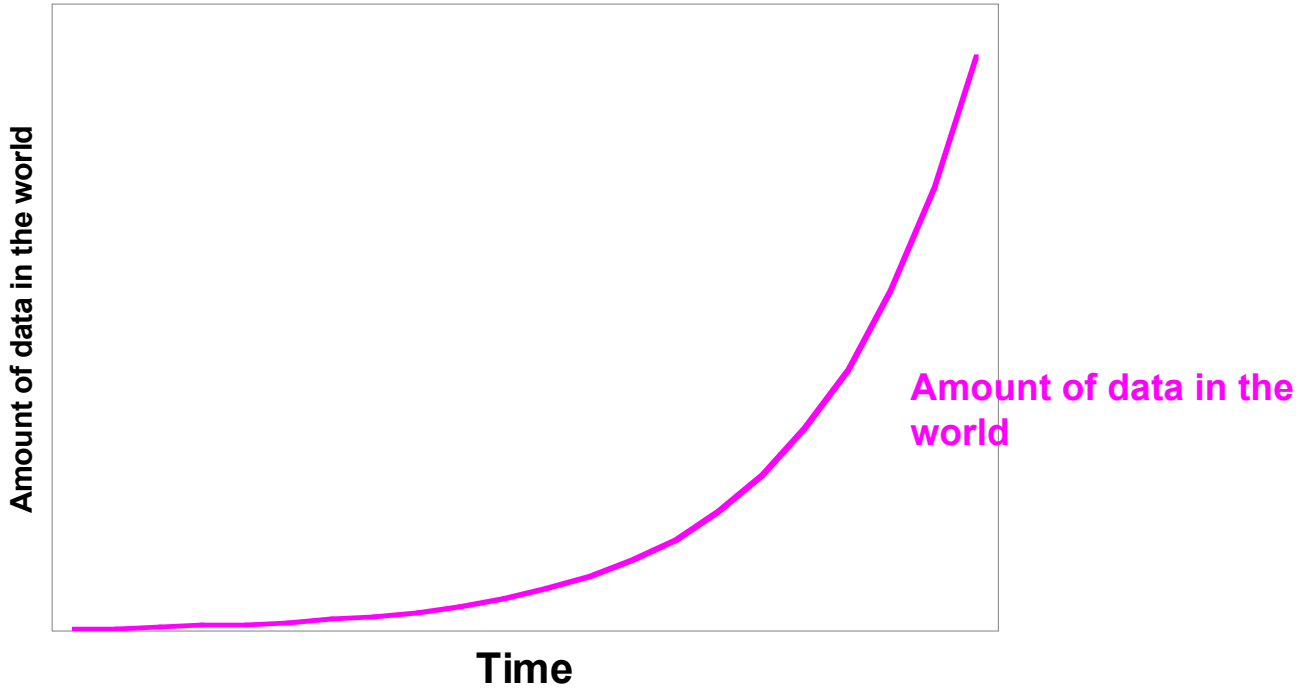
4. Summary

- Recommendations on NERSC architecture, system configuration and associated service requirements needed for your science:
  - Explicit support on the larger machines for the kind of "event" based data analysis and simulation currently done on PDSF. This would require support for HEP VMs.
  - NERSC matching in kind to counter aging-out of PDSF HW.
  - Cloud Computing (eg. Magellan) might very nicely map onto this solution space if concerns of data access can be addressed.
- What significant scientific progress could you achieve over the next 5 years with access to ~50X NERSC resources?
  - ATLAS (CMS, Super-B, …) data are very rich. 50X resources at NERSC would provide scientists with greater opportunities for data exploration, and scientific discovery.
  - At this time, NERSC is not a power-player in LHC data analysis. Such a resource would draw many US scientists.
  - N.B. There are many HEP experiments I have not mentioned. These currently rely primarily upon local or project resources.

Some Closing Thoughts

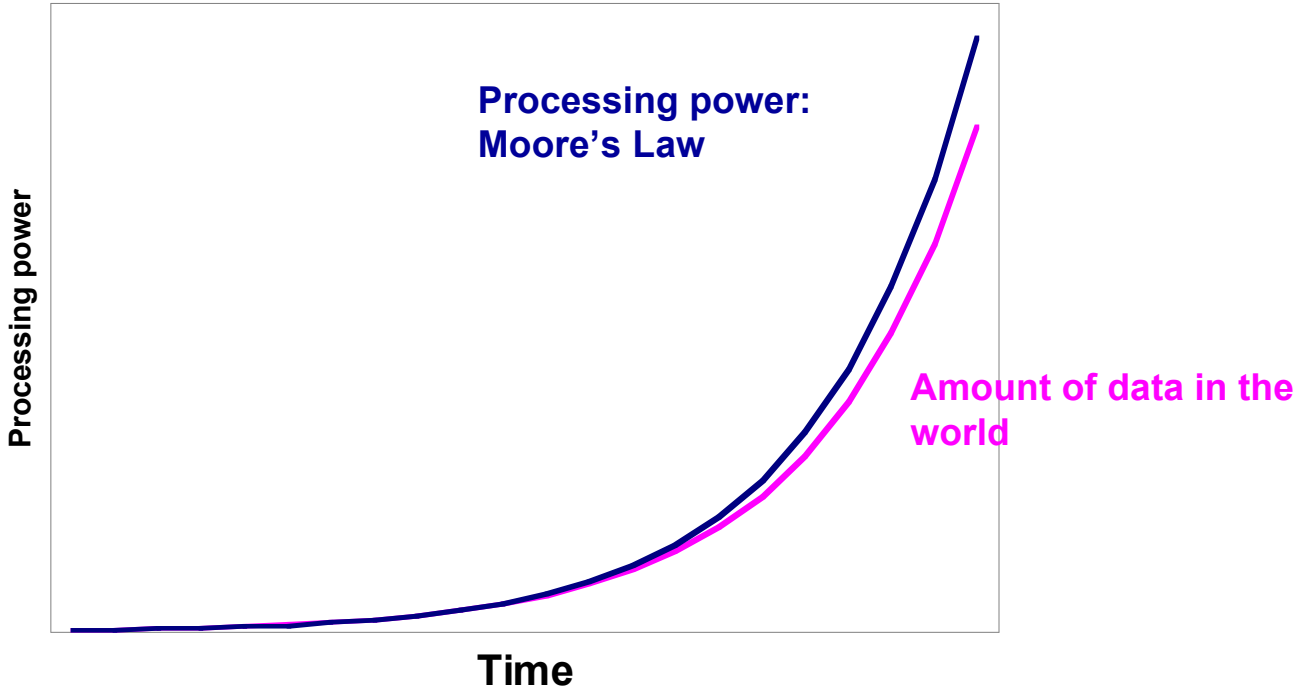
- Computing is the 3rd critical technology for HEP.
  - in addition to Accelerator and Detector Technology
  - In my opinion, computing technology/science is in ascendancy.
- As accelerators and detectors become larger and more expensive, the imperative for extracting maximal scientific discovery from each data set grows.
  - As the volume and richness of data sets increase, human cognitive scaling can't solve this.
  - New machine driven searches for anomalies or domains of interest will surely help. (eg. knowledge discovery / data mining techniques)
  - These kind of techniques could vastly increase computing requirements per petabyte of experimental data.

Solving the Petascale Challenge:
What is the rate-limiting step in data understanding?

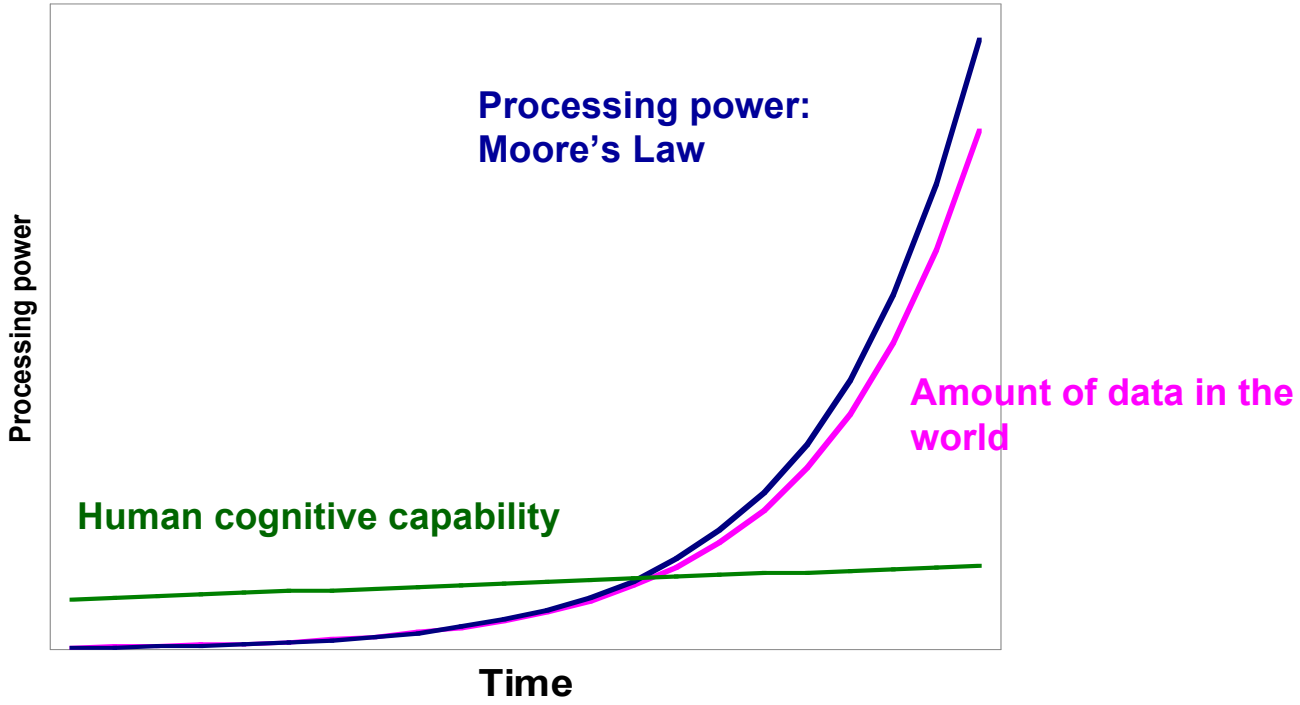


C.Aragon

Solving the Petascale Challenge:
What is the rate-limiting step in data understanding?

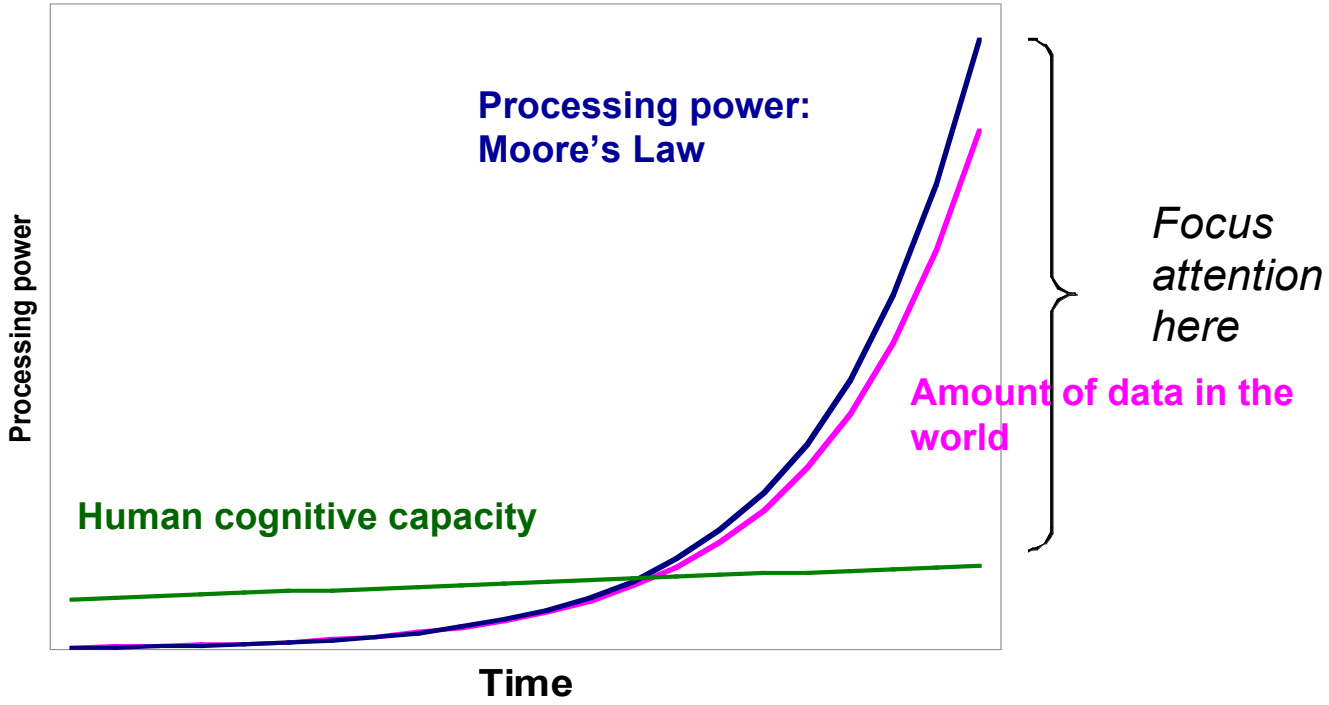


Solving the Petascale Challenge:
What is the rate-limiting step in data understanding?



Idea from "Less is More" by Bill Buxton (2001)

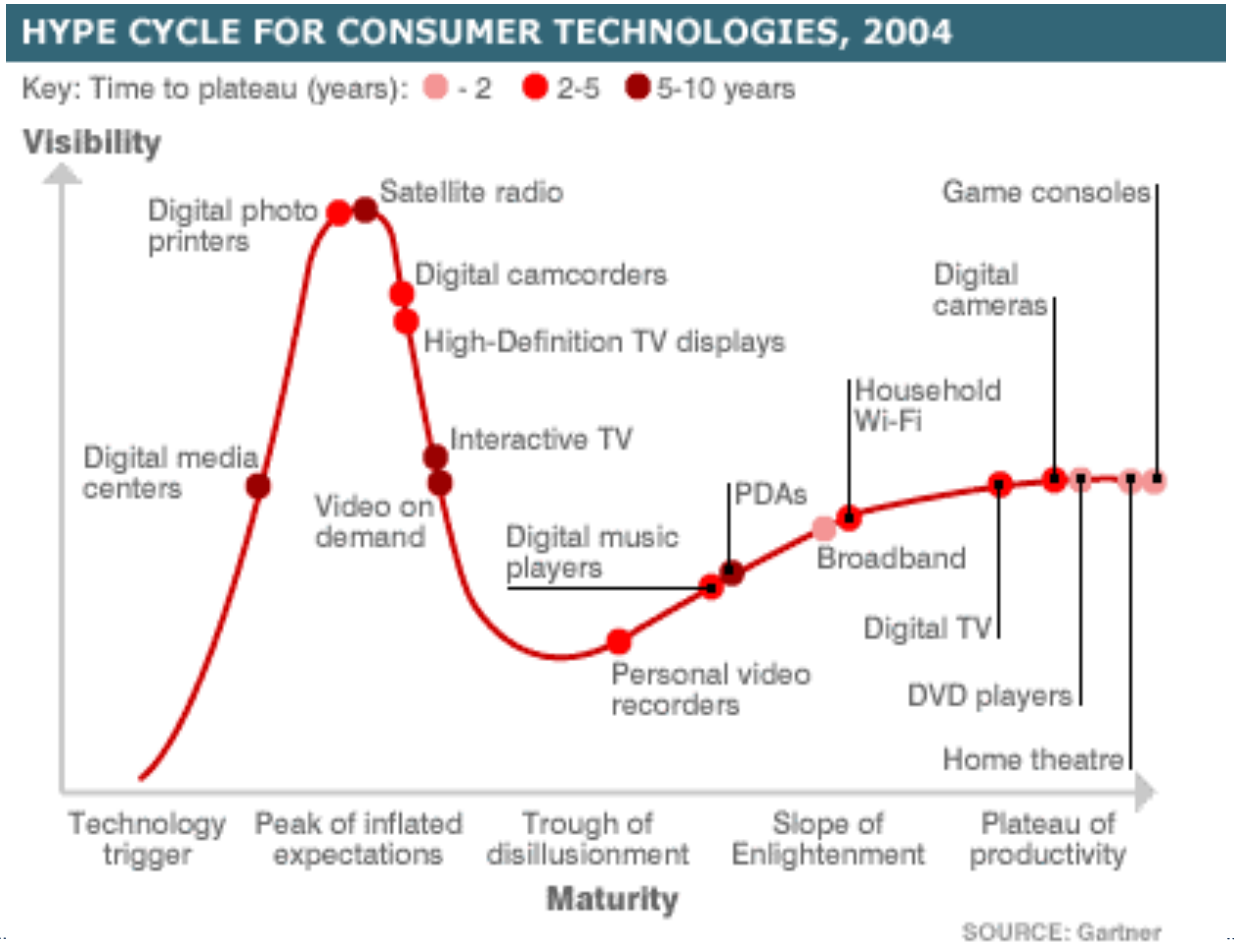
Solving the Petascale Challenge:
What is the rate-limiting step in data understanding?



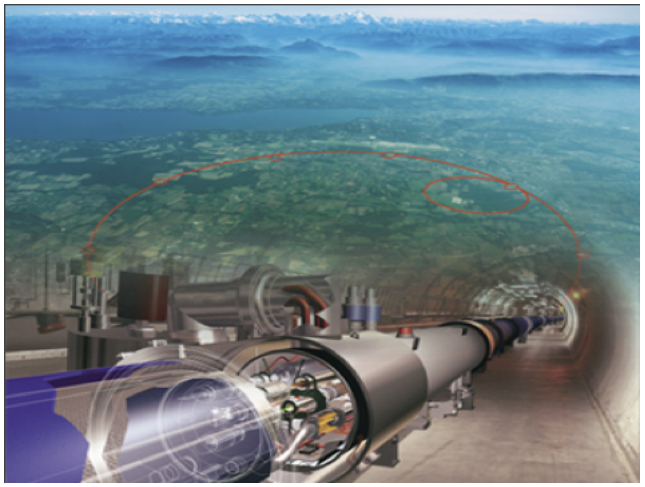
Idea from "Less is More" by Bill Buxton (2001)

THANK YOU
(XIE-XIE 谢谢)

Gardner's Hype Cycle



ATLAS Current Work and Future Directions



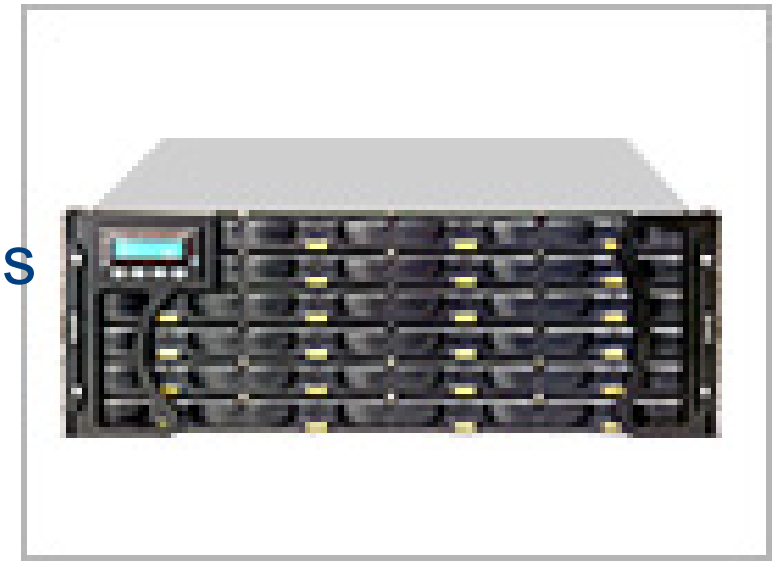
- PyROOT optimization
  - JIT Python compilation, on the spot parallelization
- Efficient Multicore exploitation
  - CRD leading ATLAS (and LHC) in this R&D work
    - athenaMP: process-based task farm, exploit Linux COW
      - ~20% extra shared-memory, ~20% more jobs in multicore farms
      - athenaMP ~production quality. Starting R&D for many core
- Virtualization for Analysis and Production
  - Early involvement with CernVM project
    - Several plugins (mostly ATLAS-oriented) contributed
    - Virtual Machine Logbook: tech-independent tool to organize/share disk-optimized VM snapshots

Virtualization

- Home grown, based on chroot (2003)
- Support for 32 and 64 bits (as long as software is not using uname to get the “bitness”)
- Support for SL4 (32 and 64 bit), SL3, RH8
- Essential in preventing resource fragmentation.
- Evaluating new solutions

PDSF Filesystems

- GPFS storage
  - Home directories
  - Group software, applications
  - Data storage
- Local drives on compute nodes (4x750GB) - xrootd planned



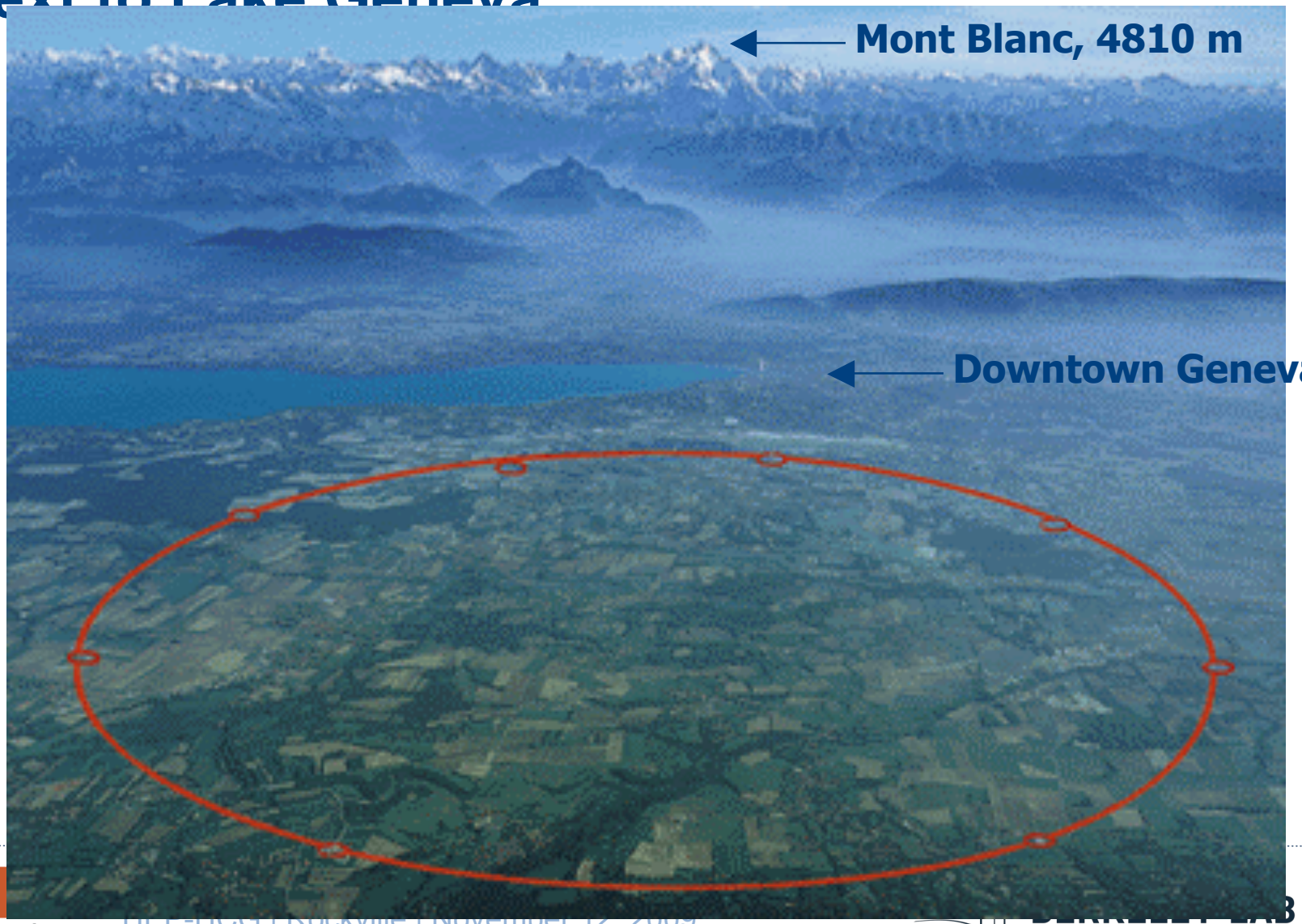
- NFS being phased out (Only used for node installation and maintenance)
- AFS at NERSC is only used by PDSF and we provide client access only

CHROOT vs VM

- Can easily save/load system snapshots, support live migration
- Security: a malicious root user in CHROOT can be harmful, so that in CHROOT one can never give out root. A root user in a VM can't take full control of the host.
- Virtualized Guest systems can have their own linux kernel, while CHROOT have only one kernel.
    - e.g. kernel 2.6. 31 will tell you where is "heap" or "stack" in /proc/pid/smaps, but kernel 2.6.9 won't. So some memory diagnose tool of ATLAS can't be used.
- It's application developer's task to maintain/patch the guest system. IT people at computer centers can focus on improving the overall performance of the cluster, instead of patching/installing packages for each user every other day.
    - Of course the maintainers and users of the cluster will need to work on a set of pre-defined standards.
- Container based virtualization (Like CHROOT or Linux VServer) can perform better with I/O bound applications

● CHROOT is better ● VM is better

CERN site:
Next to Lake Geneva



D.Foster

LHC data (simplified)

Per experiment:

- 40 million collisions per second
- After filtering, 100 collisions of interest per second
- A Megabyte of digitised information for each collision = recording rate of 100 Megabytes/sec
- 1 billion collisions recorded = 1 Petabyte/year

With four experiments, processed data we will accumulate 15 PetaBytes of new data each year

= 1% of

1 Megabyte (1MB)
A digital photo

1 Gigabyte (1GB)
= 1000MB
A DVD movie

1 Terabyte (1TB)
= 1000GB
World annual book production

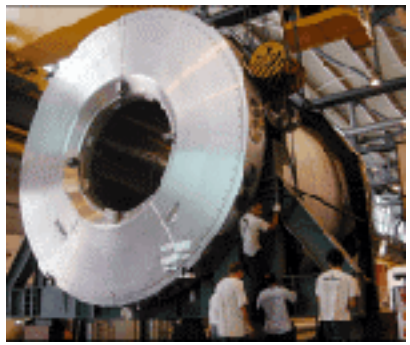
1 Petabyte (1PB)
= 1000TB
10% of the annual production by LHC experiments

1 Exabyte (1EB)
= 1000 PB
World annual information production

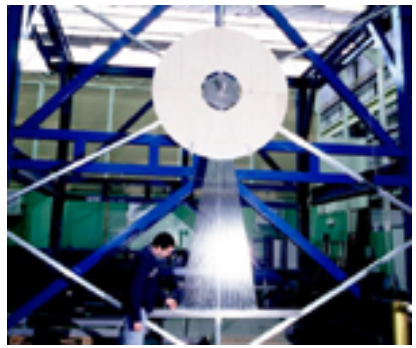
CMS

LHCb

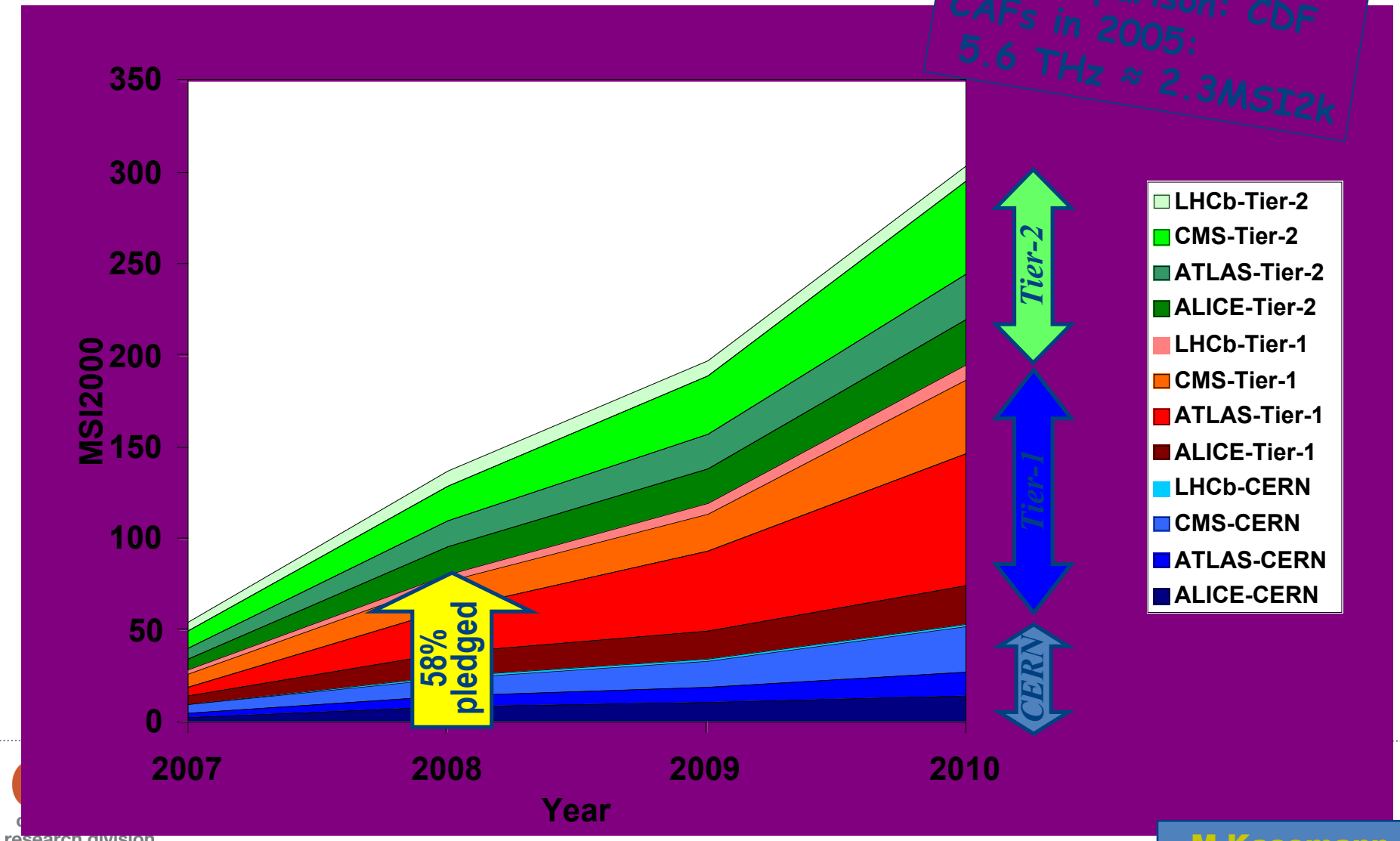
ATLAS



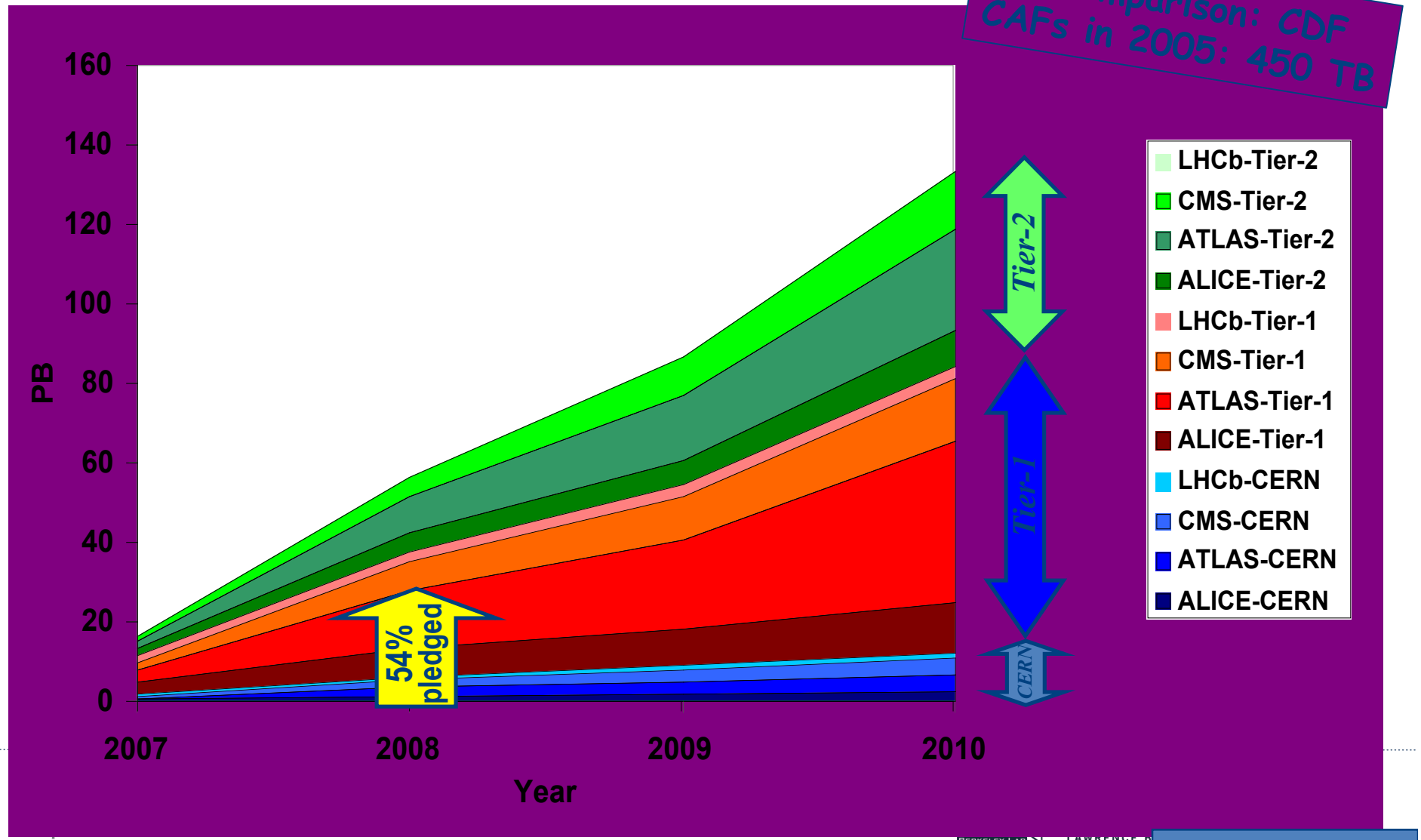
ALICE



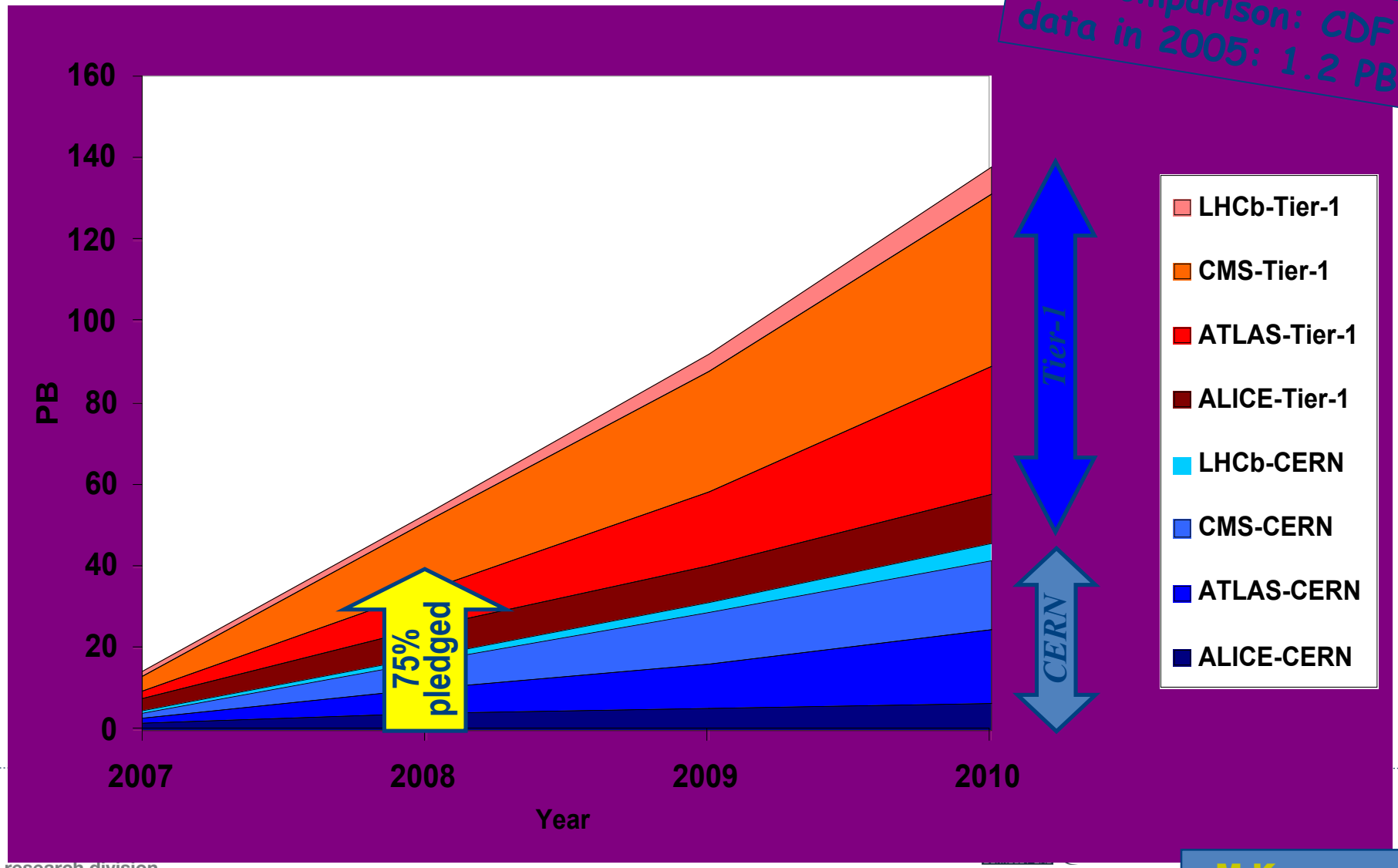
LHC Experiments: CPU Requirements



LHC Experiments: Disk Requirements

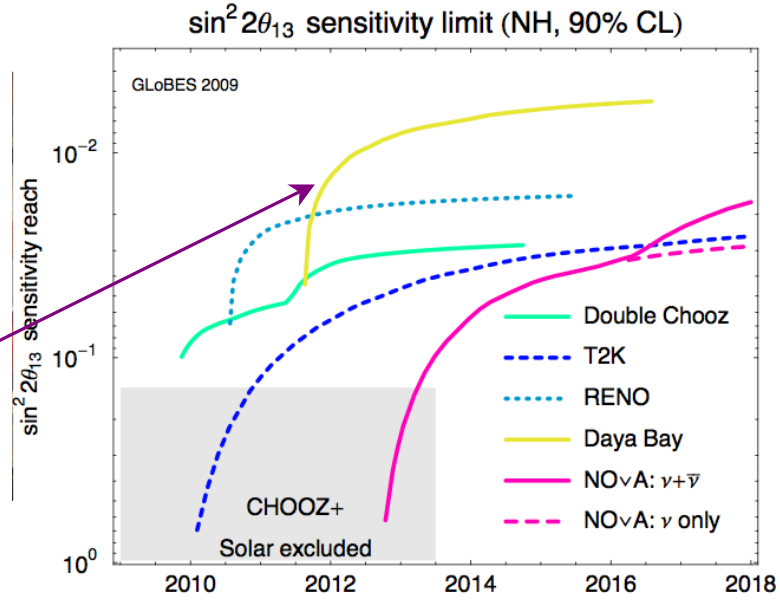


LHC Experiments: Tape Requirements



Daya Bay - Schedule

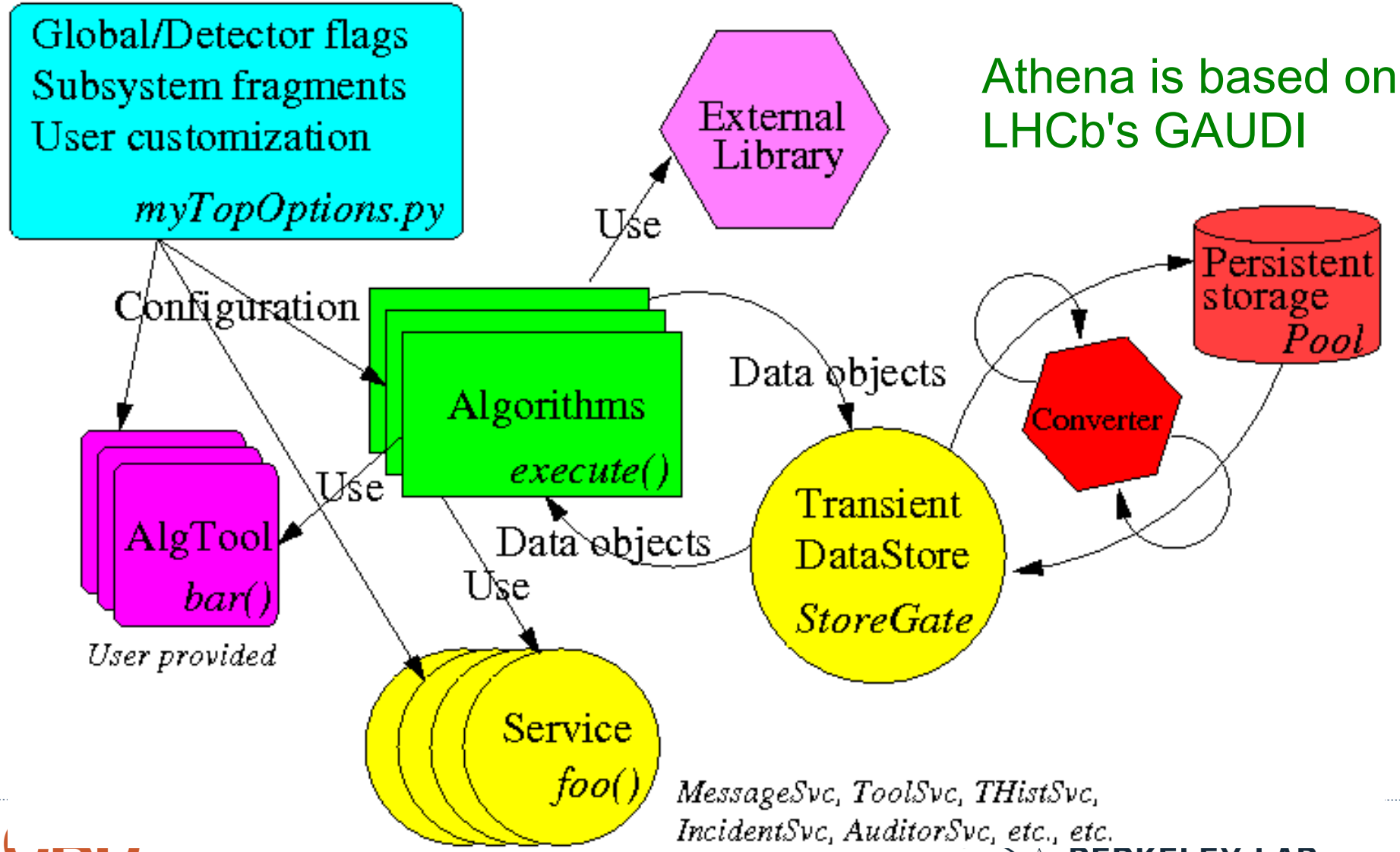
- In our first 6 months of data taking, Daya Bay will have world's best sensitivity to $\sin^2(2\theta_{13})$



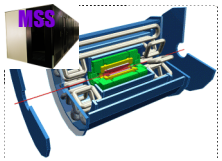
Milestone Description	Current Forecast
Beneficial Occupancy of Surface Assy Bldg (SAB)	Mar-09
Beneficial Occupancy of Halls 1 & 5	Oct-09
DB Near Hall Physics Ready	Aug-10
US CD-4a Approval Request	Aug-10
Beneficial Occupancy of Halls 2 & 3	Jul-10
Far Hall Physics Ready	Nov-11
US CD-4b Approval Request	Nov-11

* Delta is from Orig Baseline to Current Forecast

Athena in a nutshell



Athena is based on LHCb's GAUDI



The Computing Model

PC (2004) = ~1 kSpecInt2k

~Pb/sec

Event Builder

10 GB/sec

Event Filter
~7.5MSI2k

320 MB/sec

•Some data for calibration and monitoring to institutes

•Calibrations flow back

~ 75MB/s/T1 for ATLAS

Tier 0

T0 ~5MSI2k

~5 Pb/year

No simulation

Tier 1

US Regional Centre

Dutch Regional Centre

French Regional Centre

UK Regional Centre (RAL)

~2MSI2k/T1

~2 Pb/year/T1

≥622Mb/s links

10 Tier-1s reprocess house simulation

Group Analysis

≥622Mb/s links

Tier 2

Northern Tier
~200kSI2k

Tier2 Centre
~200kSI2k

~200 TB/year/T2

Lancaster
~0.25TIPS

Sheffield

Physics data cache

100 - 1000
MB/s links

Desktop

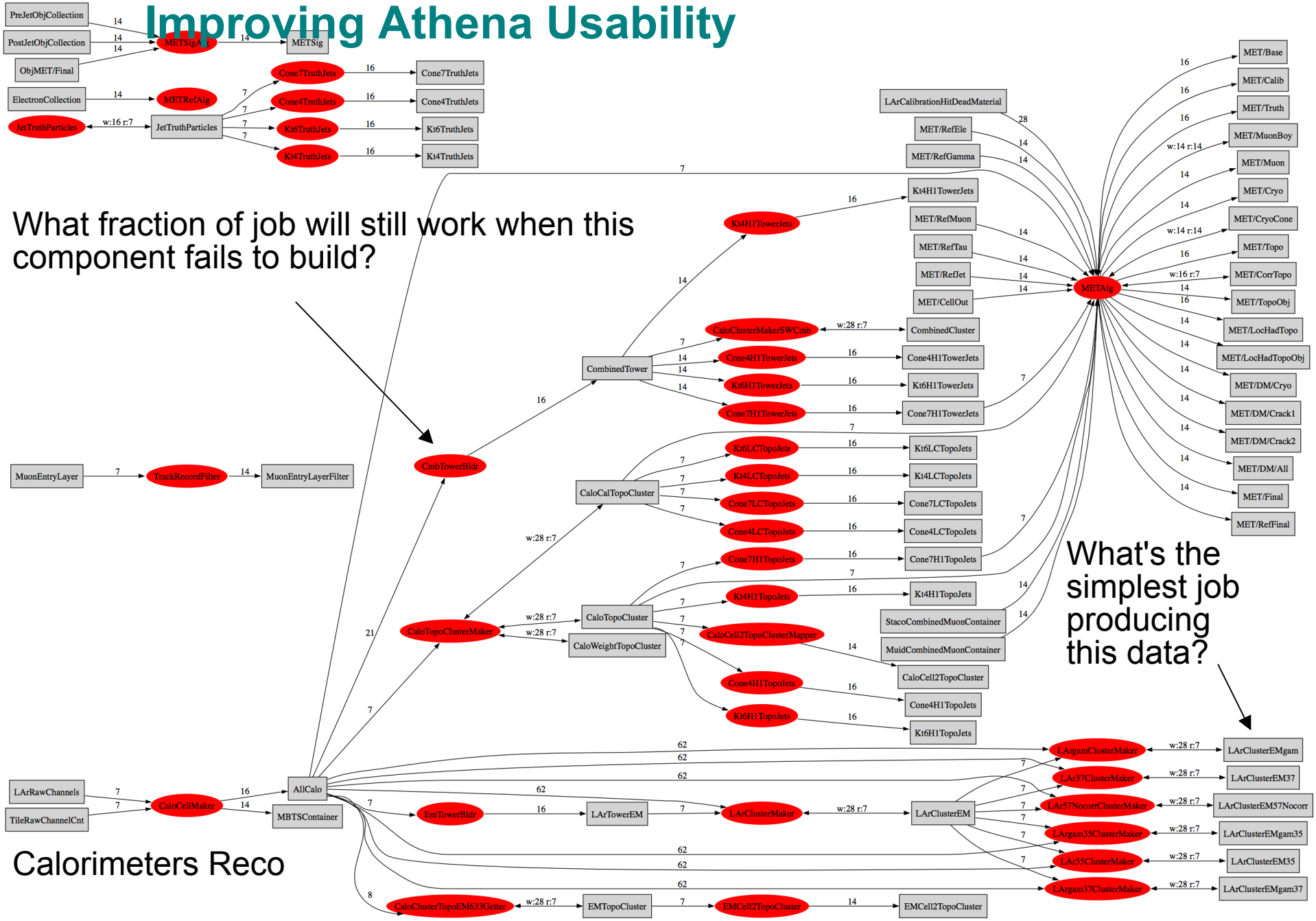
Workstations

Each Tier 2 has ~20 physicists working on one or more channels

Each Tier 2 should have the full AOD, TAG & relevant Physics Group summary data

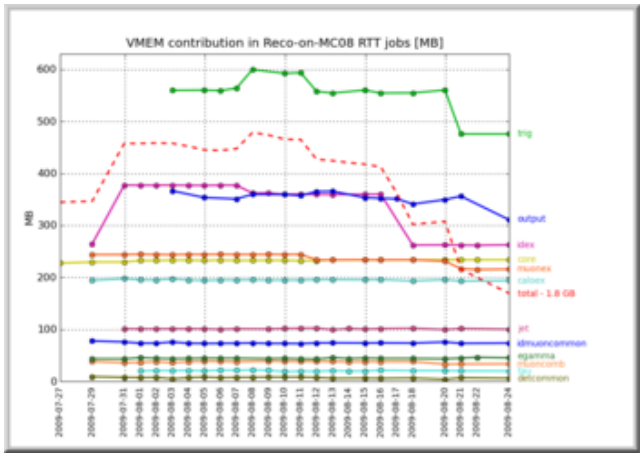
Tier 2 do bulk of simulation

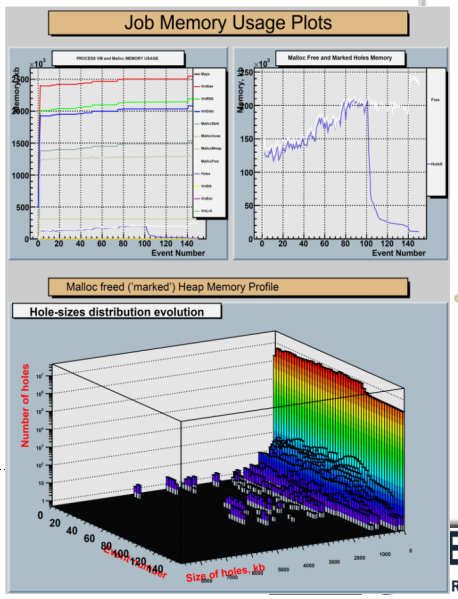
Improving Athena Usability

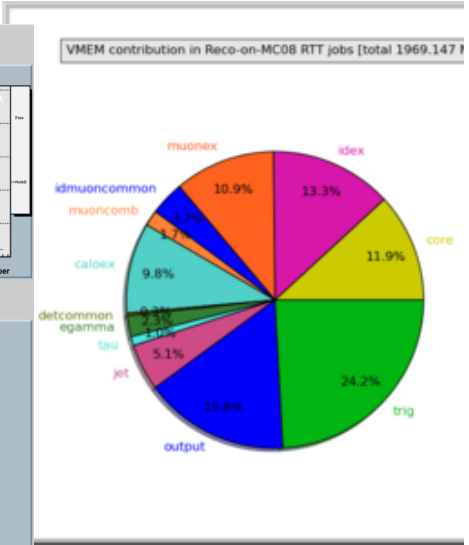


Performance Optimization

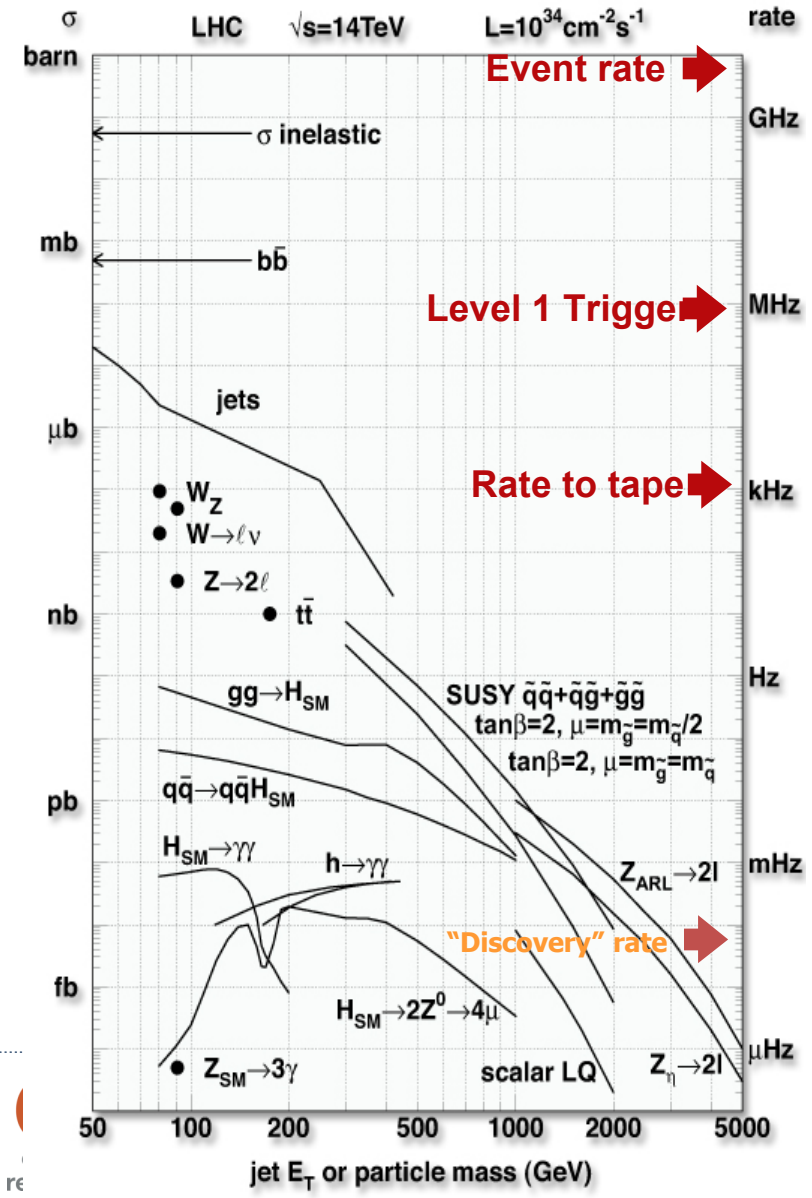
- LHC apps 1GB+ VMEM
  - Off-the-shelf memory profiling tools scale badly (slow)
- Hephaestus
  - general purpose tool to track memory allocations
  - ~50% CPU cost, scales well to 2.5 GB applications
  - powerful valgrind GUI
- Perfmon
  - Athena auditing mechanism
  - History of performance, by domain and by component

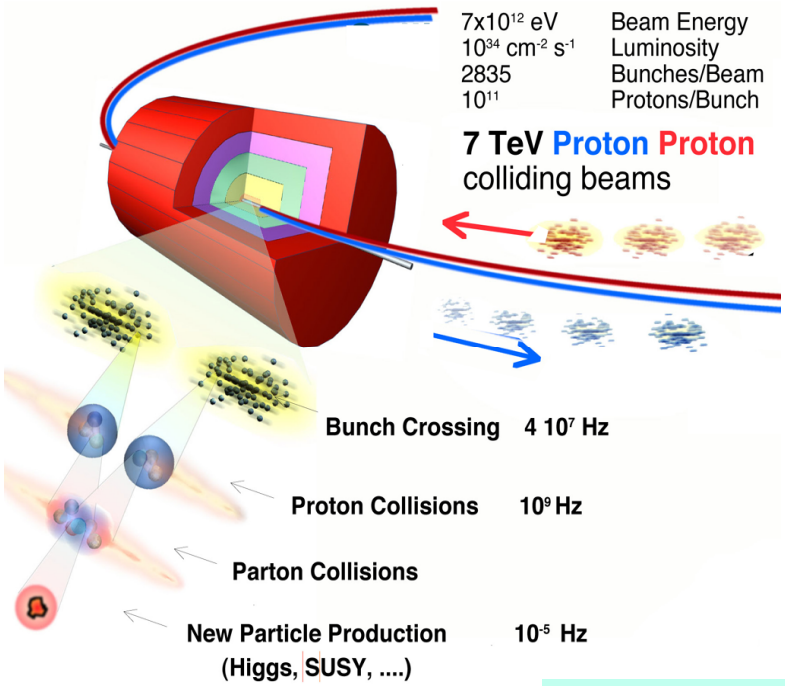






p-p collisions at the Large Hadron Collider





Crossing rate	40 MHz
Event Rates:	~10^9 Hz
Max LV1 Trigger	100 kHz
Event size	~1 Mbyte
Readout network	1 Terabit/s
Filter Farm	~10^7 Si2K
Trigger levels	2
Online rejection	99.9997% (100 Hz from 50 MHz)
System dead time	~ %
Event Selection:	~1/10^{13}

Luminosity
Low 2x10^{33} cm^{-2} s^{-1}
High 10^{34} cm^{-2} s^{-1}

Large-Scale Science (HEP & NP example)

- Large, distributed collaborations are the norm
  - ~2000 scientists, from ~150 institutions in ~50 countries
  - Scientists require equal access to data and resources
- Very long time duration of projects & software
  - Detectors take 5-10 years to design & build. Operational lifetimes of 5-20 years
  - 10 to 30 year Project lifetimes - Software must work early and continuously
- Commodity computing (Intel, Linux)
- Trivial parallelism/Partitioning of calculations
- Data Intensive (100's TB => 1,000's TB)
- The World is Networked and resources are distributed
- Scientists are developers and not just users
  - Many skill levels from Wizard to Neophyte
- Issues of scaling are sociological as well as technical - interfaces are critical