Storage and Data Management

Richard P Mount
SLAC

Michelle Butler
NCSA

Mike Hildreth
Notre Dame

Outline

SLAC

- Input from the Science Frontiers
- Technology Outlook (Michelle Butler)
- Data/Software/Physics Preservation (Mike Hildreth)
- HEP Outlook

Input from the Science Frontiers

Input from the Science Frontiers (1)

SLAC

- Input from Energy Frontier
  - Affordable offline computing (for which storage is the largest cost) does and will largely determined the trigger rate to persistent storage. [pp experiments]
  - Tape is probably underused – current insistence is that all raw data passing the high-level triggers are “golden”
  - Distributed data management is a major continuing development activity and major operational activity.

Input from the Science Frontiers (2)

SLAC

- Input from Intensity Frontier
  - Storage capacity is not currently a major issue.
  - Storage capacity/performance and data management, was a major cost for BaBar – in its day HEP's most data-intensive experiment.
  - Expect that storage capacity/performance and data management will be an issue for Belle-II and future "factory" experiments.
  - Data management software will be a significant need for some IF experiments (e.g. CTA).

Input from the Science Frontiers (3)

SLAC

- Input from Cosmic Frontier
  - 1 PB total today
  - 50 PB total by 2025
  - 400 PB/year in 10 – 20 years (SKA)
  - “Can easily generate many PB from simulations, but no place to store them or analyze them”

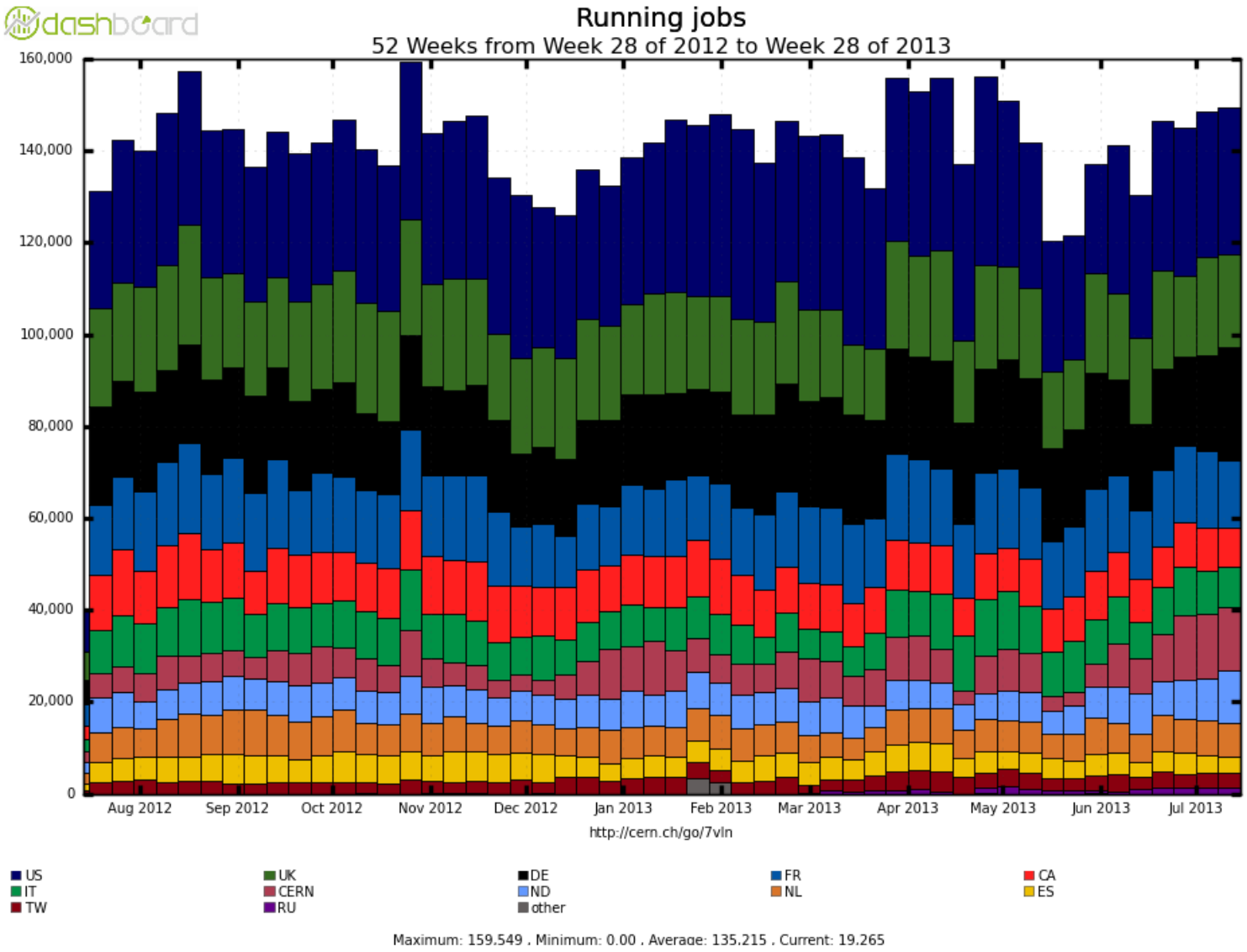
- Input from Accelerator Science
  - Data volume cannot compare with HEP experiments
  - But data rates (remote simulation on supercomputer moved to local analysis facility for “control room feedback”) can exceed those of HEP experiments.

Storage and Data Management – Other Requirements

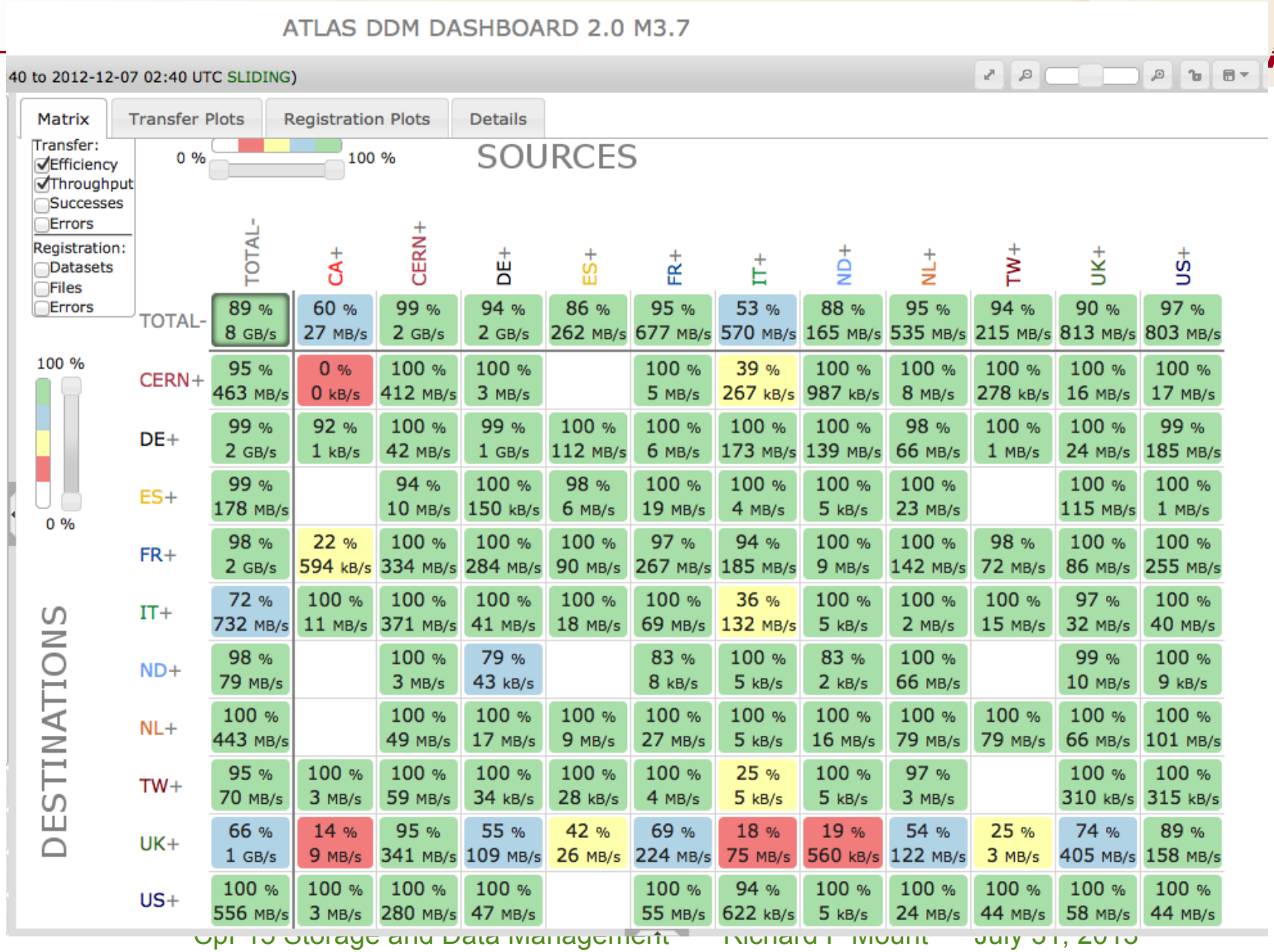
SLAC

- Large and costly experiments require international collaboration
- Resources (storage, CPU) must be geographically distributed
  - Need funding from many nations
  - Need access to shared computing resources
  - Need access to opportunistic computing resources
  - Need access to (and development of) distributed expertise
- Wide area networks, distributed storage, distributed CPU, distributed data management and distributed workflow management are all essential.

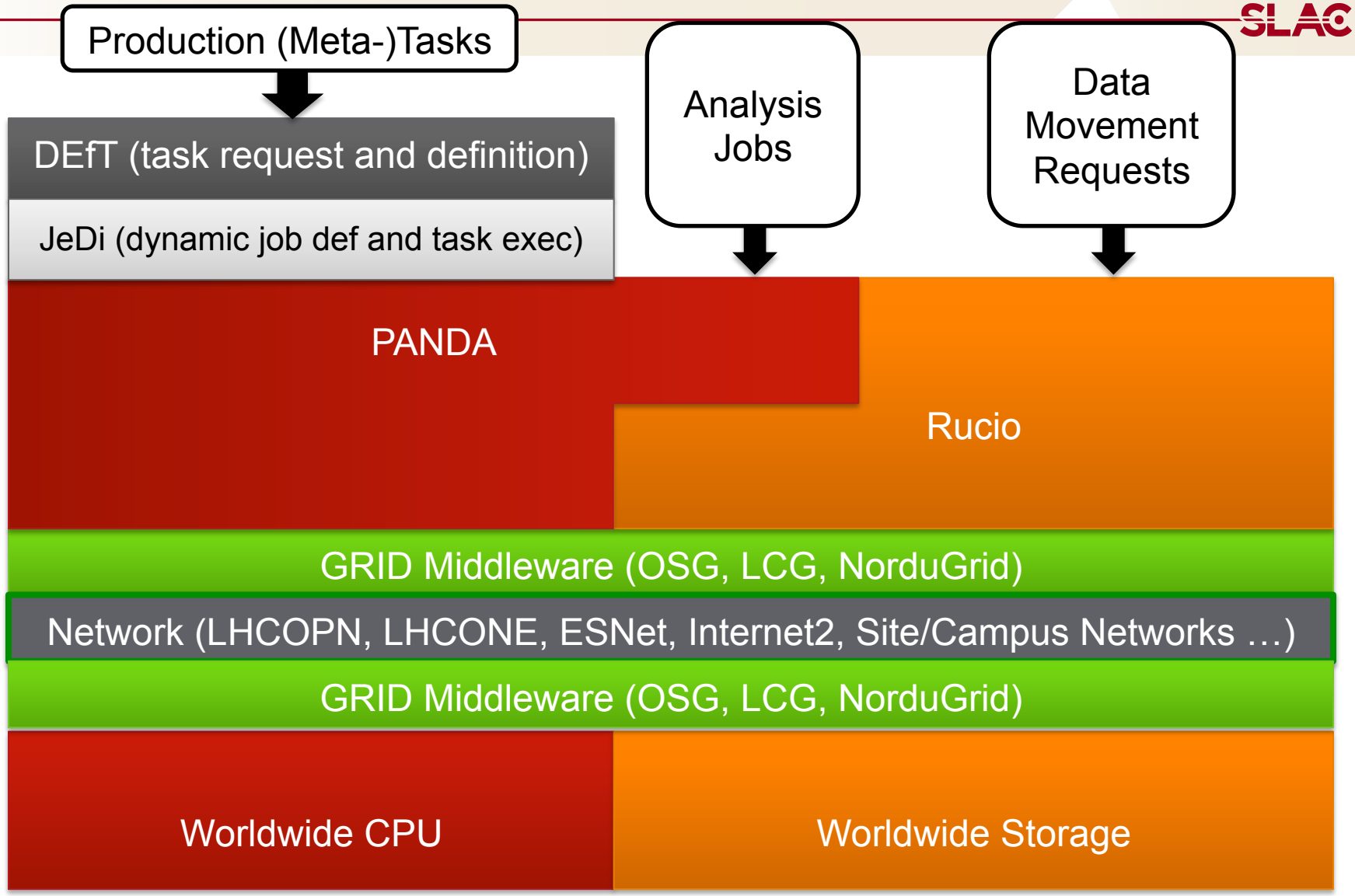
ATLAS Production and Analysis



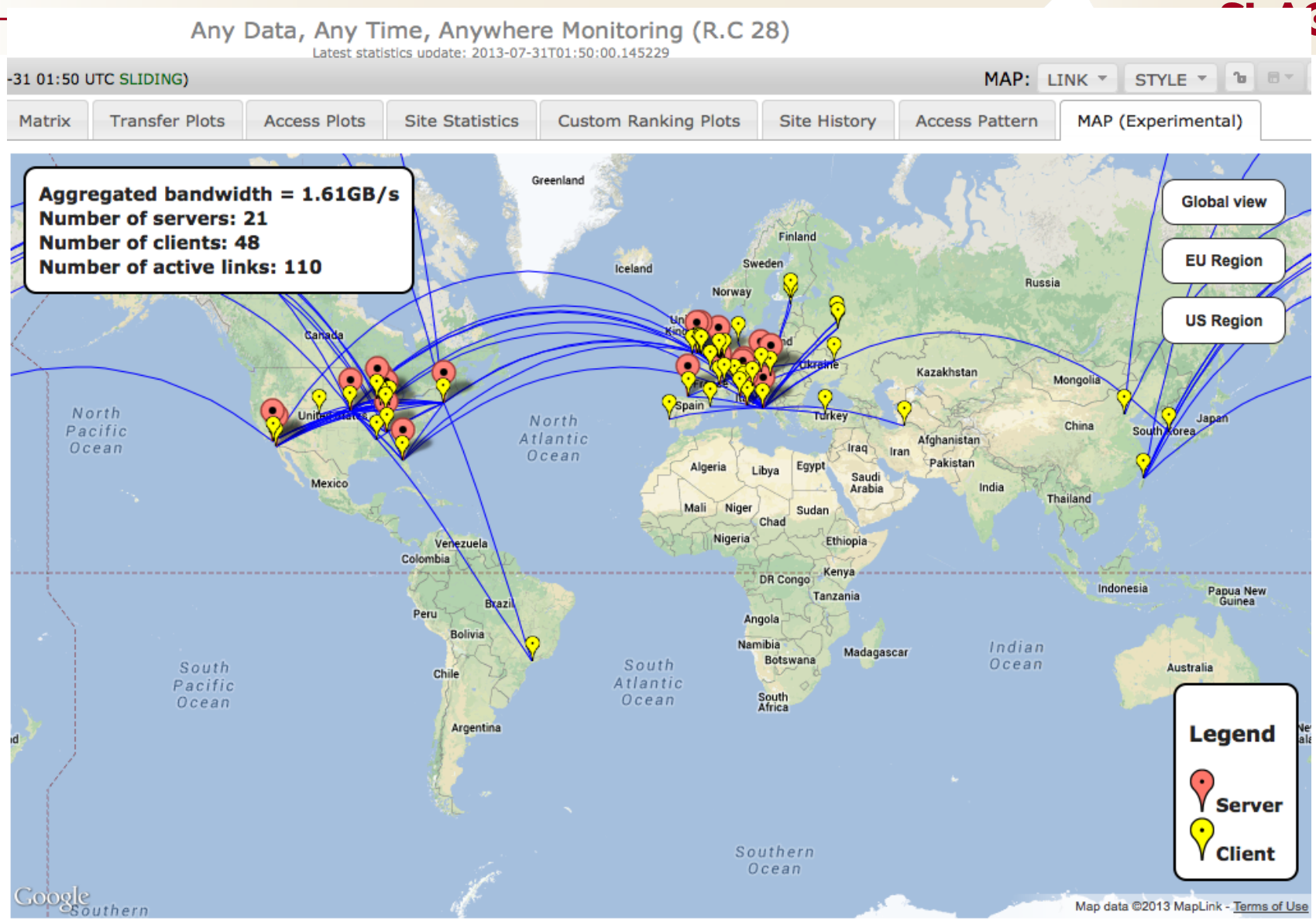
ATLAS Data Distribution



ATLAS Distributed Computing 2014



Any Data Anywhere Anytime – Remote Access without Borders



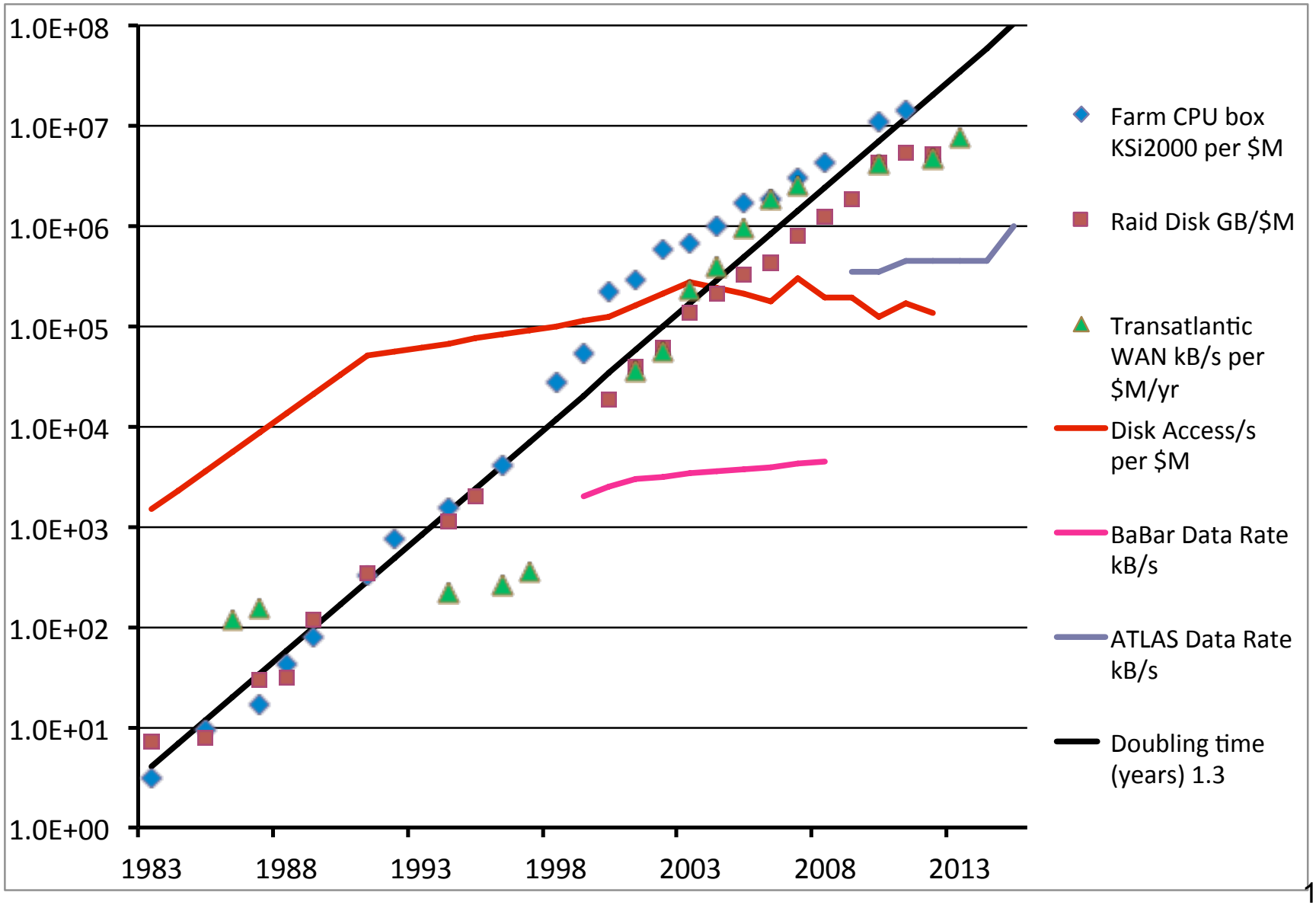
Summary of Input

SLAC

- Energy Frontier, and some Cosmic Frontier and Intensity Frontier experiments need as much storage, data access and computing as they can get (at tolerable cost).
- For these cases “requirements” have little meaning – the role of storage is determined by including it (usually implicitly) in an optimization of all major aspects of the experiment to maximize science/$.
- Distributed data management software is one of the major (costly) aspects of an experiment that can have a huge impact on the experiment’s effectiveness.

Technology Outlook

The Past: Exponential growth of CPU, Storage, Networks



Storage Futures

Michelle Butler

NCSA

Storage futures

- Disk and tape current technologies
  - Disk is always trying to stomp out tape
  - Tape continues to hang in
    - Due to sheer size of data continues to grow
    - Power and cooling now come into play
    - Archives such as the BlueWaters machine at 380PB in 5 years is more economical to provide on tape with RAIT to protect for disaster recovery (DR)
    - Has a very bright future at performance and data doubling for many years out (10 years at least)
    - Every 18 months was new drive/technology, but now without much competition it's every 3 years

Storage Futures (2)

- Disk has had a very good run
  - SATA (or nearline NL) disks
    - At 7200RPM at current 4TB drives have just in the last 3 years gone from 1TB, 2TB , 3TB, now 4TB.
    - Slowing down, not able to get many more large jumps in TB.
    - Can be cheap at the BestBuy level, or a little more for enterprize class drives. Still fail often and need RAID6 type protection.
  - SAS drives –
    - At 10K and 15K RPM with 2.5 and 3.5" drives
    - At 600GB to 900GB drives now. These are blazing fast, but are expensive. Less likely to fail and can go with a RAID 5, but most use cases are for DB or something with small I/O requirements due to the cost.

Storage Futures (3)

- Disk capacities are slowing down due to fast approaching the physical limits of the magnetic materials.
- The technology requires changing, but what will stick in the commodity market is unsure.
- What will be cheap enough for commodity storage in the next 5 years beyond the 4TB maybe 6TB disk drive is not something that can be seen at this time.
- Solid state and memory type devices are in production, but none are really large enough to hold PB of storage
  - Still require large amounts of disk storage behind with data movement/management schemes to move data from fast expensive technologies to slower cheaper media.

Middleware

- Globus Online for data movement which is an enhanced gridftp interactive application with retry and notification.
- GO storage is using FTP storage systems set up around the country and a seamless storage fabric can be used by users.
- File systems “managed” in the next year Lustre and GPFS (now)
  - Inode stays in the file system while data can be “moved” to offsite FTP archive based on policy. As data requested it will be moved back to online disk system.

Disks – from Per Brashers/DDN

SLAC

- The area of a “bit” in current products is close to the limit where what is written will remain magnetically stable.
- New technologies to make the “bits” more stable are on the horizon:
  - “Shingled Recording” Not easily re-writable

  - Heat Assisted Magnetic Recording (HAMR)
  - [Laid-out-in-advance] Bit Patterned Recording
- None of these looks good for the near future.

Worldwide Data Management

SLAC

- There are no commercial or widely used open source offerings meeting HEP needs at this time. This is, in part, due to the requirement for the highly efficient integration of tens to hundreds of autonomous computer centers.
- Other scientific fields have increasingly similar needs.
- Our current approaches are almost too labor intensive even for 3000-physicist collaborations.
- Operational cost must be greatly reduced.

Data/Software/Physics Preservation

Data/Software/Physics Preservation

Mike Hildreth

Université de Notre Dame du Lac & Fermilab

Overview:

- "Data Preservation" isn't just about data…
  - keeping the data on disk/tape, migrating it to future storage media, etc., is the easy part
    - "solved" problem
  - re-processing the data to reproduce an old analysis or to produce new results is much harder
- Common Issue:
  - mentioned by essentially every group here with large projected datasets
    - here, moderate differences in use cases
  - common to many other fields
    - although, use cases vary dramatically
- Energy Frontier experiments are leaders in this effort
  - we have the resources to work with other areas to provide common solutions to some of these problems
  - These efforts should be coordinated to maximize impact, minimize effort

Motivations

- "Self Preservation" is probably the most important
  - decade- or decades-long experiments will need to be able to look back at their own data. Even looking back two or three years can be very difficult without proper planning
  - solving this problem gets one most of the way to a "knowledge preservation" infrastructure
    - documentation, software, and the ability to run executables must be maintained
- "Outreach"
  - Inspire supplementary data, tools like Rivet, HEPDB, etc., constitute one set of "knowledge preservation" tools
    - "outreach" to theorists, colleagues
  - datasets, and the associated instructions for true outreach to non-specialists can also serve as vehicles for knowledge preservation
- Mandates
  - could be coming from funding agencies – should be prepared!

(Some) Current Efforts

- Rising tide of interdisciplinary interest/need for infrastructure
- Many individual archives in Astrophysics
  - currently “small” on the scale of Pb
- DPHEP and associated work
  - DESY self-validating software archive
  - “global” discussion of needs/efforts
    - focus on common solutions
- DASPOS
  - “US”-based effort to understand needs and build some common infrastructure (metadata, databases, etc.) for broad use
- Research Data Alliance
  - Global discussion of data/software/knowledge preservation problems
- more…

(Preliminary) Conclusions

- Primary recommendations of report:
  - "Knowledge Preservation" is a pressing problem for many experiments, especially those with long time scales
  - Frontiers should communicate needs so that possibilities for common solutions can be evaluated
    - (already ongoing with DPHEP/DASPOS efforts)
  - Common solutions can and should be developed
  - More resources are needed to realize system-wide infrastructure
    - especially if mandates from funding agencies are forthcoming

HEP Outlook

Impact of Disk Technology Evolution

SLAC

- The marked slowdown in disk-capacity-per-unit price evolution will reduce the benefits of retiring older equipment that is still working well.
  - This argues for buying well-engineered disk storage and keeping it for as much as 8 years.
  - As a result there will be additional pressure on space, power and cooling.
  - This will also be true for non-HEP applications so there is some hope that market forces will drive the availability of well-engineered, dense low-power hardware.

Solid-State Storage

SLAC

- Not necessarily Solid State ***Disks***
- Will play an increasingly important role in hiding the abysmal sparse/random access performance of rotating disks.
- Will slowly become more affordable with respect to rotating disks (but will not kill disks).

Tape

SLAC

- The death of tape will recede further into the future.
- Will be cheaper than rotating disk storage by about the same factor as now.
- Has properties (price, error rates, failure modes) that will continue to meet needs in HEP and the commercial marketplace.

Solid-State/Disk/Tape Data Management

SLAC

- The management of data flow within this optimized hierarchy will probably require HEP-specific software development.
- Generic automated cache-management software is unlikely to be sufficiently application aware.

Overall Optimization of Computing

SLAC

- HEP should be prepared for significant shifts in the relative unit costs of storage (all levels), CPU and networking, leading to new optimizations that are likely to require significant advance effort on software.
- One of the most important optimizations is allowing derived datasets to be instantiated or virtual:
  - Without any need to change how physicists interact with the data management system.
  - This has the major additional benefit of requiring fully automated capture and use of provenance information.