Opportunities for Nuclear Physics to benefit from Grid Computing

DNP
October 23rd, 2008

Frank Würthwein (UCSD)

Fkw's background & biases

- Experimental particle physicist working on:
  - Heavy Flavor Physics 1991-2006
  - Diboson Physics since 2005
    - WZ, ZZ, WW, Higgs to WW at hadron colliders
- Member of Open Science Grid Executive Team since inception.
- Co-lead of CMS Computing Commissioning

Much of what I talk about uses OSG and CMS as examples. Apologies upfront for this bias.

Overview of this talk

- Grid Computing as a team sport
- Open Science Grid as a Service Provider
- Grid as a production infrastructure
  - Architecture
  - Bytes & Cycles
- Science on the Grid
  - Classifying what's ***easy*** versus ***challenging***
  - A few case studies

Open Science Grid
- A team Sport -

- ***Open*** Organization that encourages (and to some extend expects!) ***participation*** by:
  - Roughly 50 Science Communities
  - Close to 60 Clusters
  - Roughly 40 organizations contributing to VDT, the OSG middleware stack.

OSG is many different things
to many different people.

Open Science Grid
- As a Service Provider -

- Tested Middleware Stack & Deployment instructions.
- Deployment & Troubleshooting Support
- Targeted Support:
  - New Science Communities to get started.
  - Established Science Communities to improve their operations.
- Opportunistic Use of Production Infrastructure

However, OSG expects effort on both sides.

No service without participation!

Benefit at Varying Levels

Be a full fledged partner:
Bring resources into OSG.
Bring Science Community.

Use technology and
Production Infrastructure

Use only the
Technology

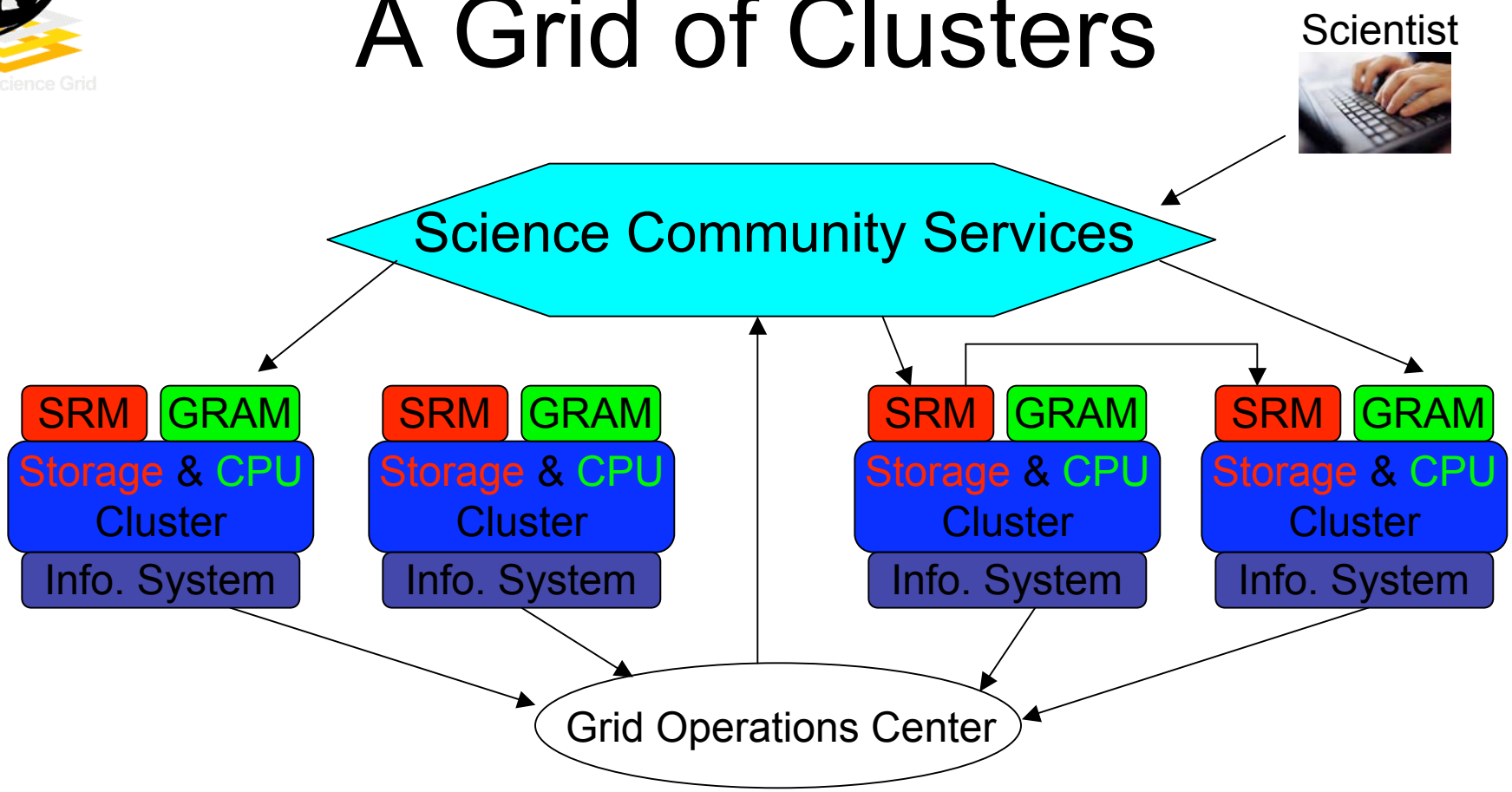
New Participants may enter at any level!

Grid as a Production Infrastructure

Architecture

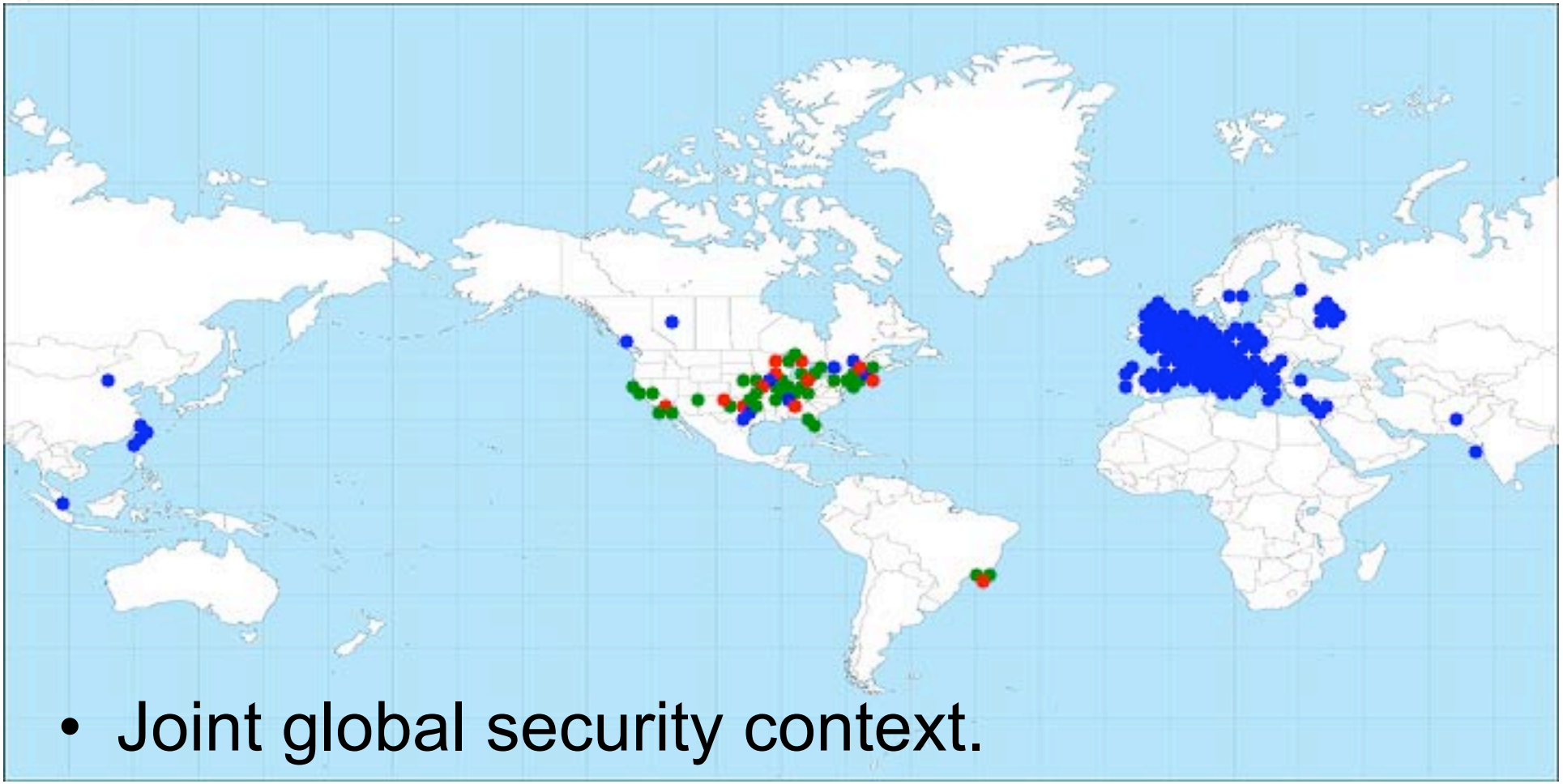
Bytes & Cycles

A Grid of Clusters



- Each Cluster is autonomously managed.
- Common Access Protocols
- Science Communities operate across only those clusters they care for.

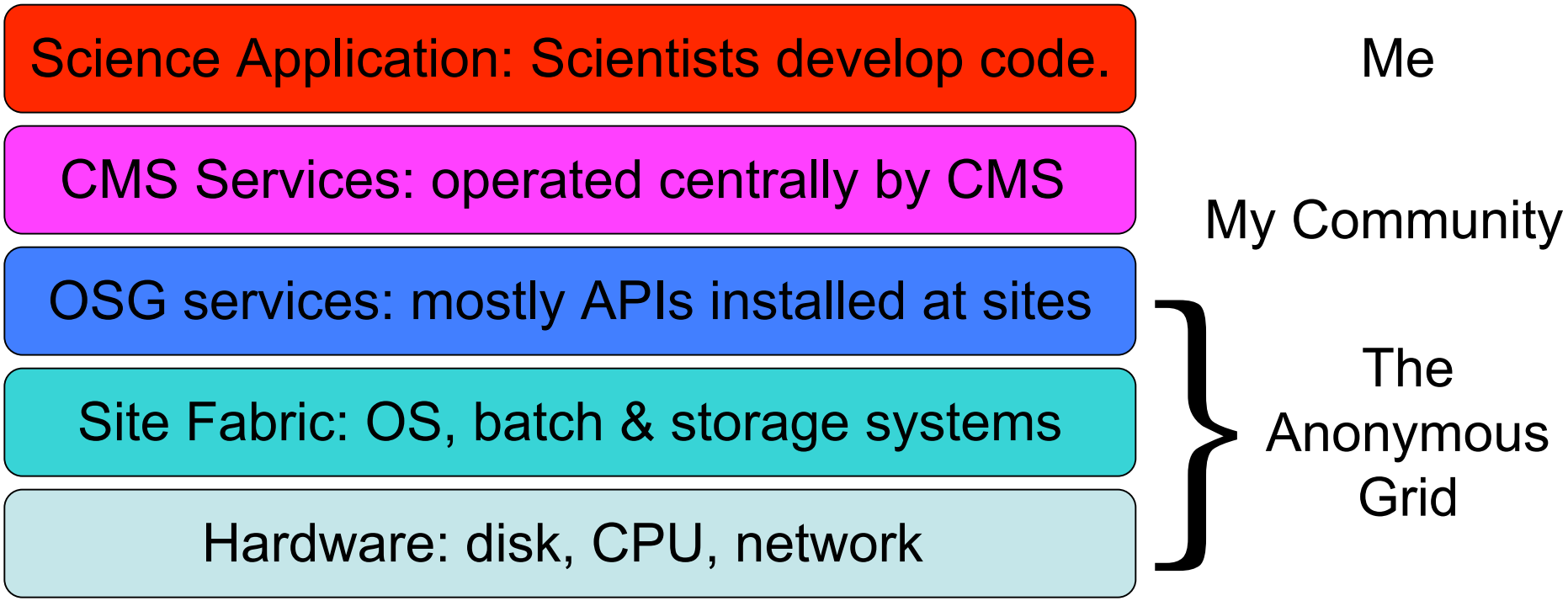
A Grid of Grids



- Joint global security context.
- Interoperable Access Protocols …

  … but regional differences in implementation.

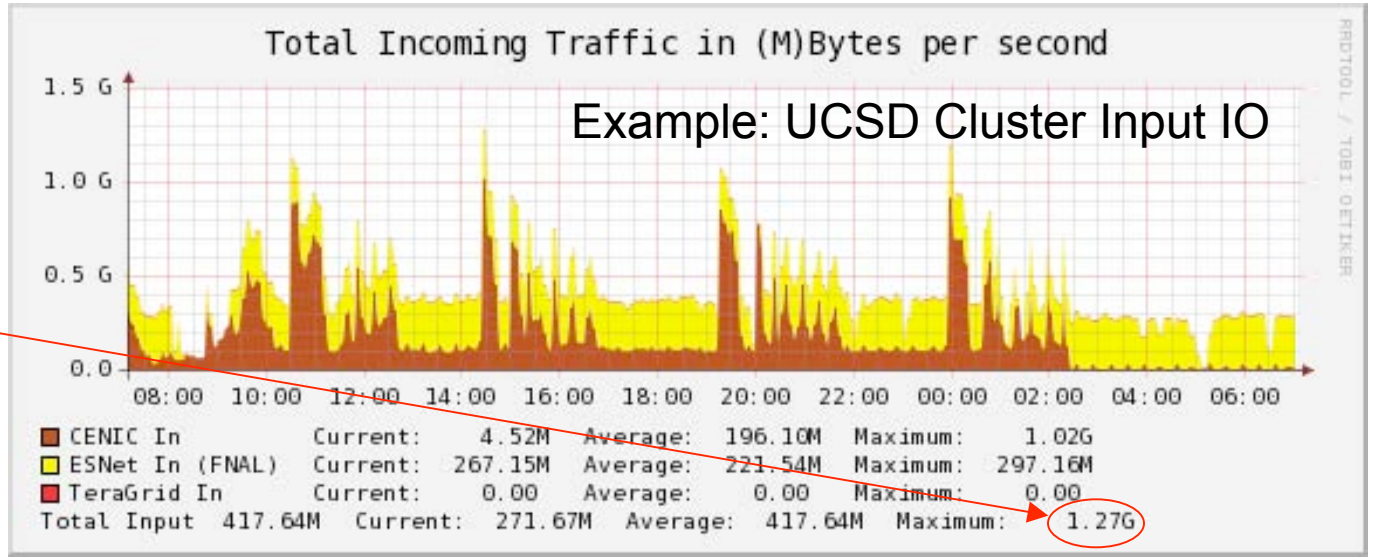
Layering of Software & Responsibilities



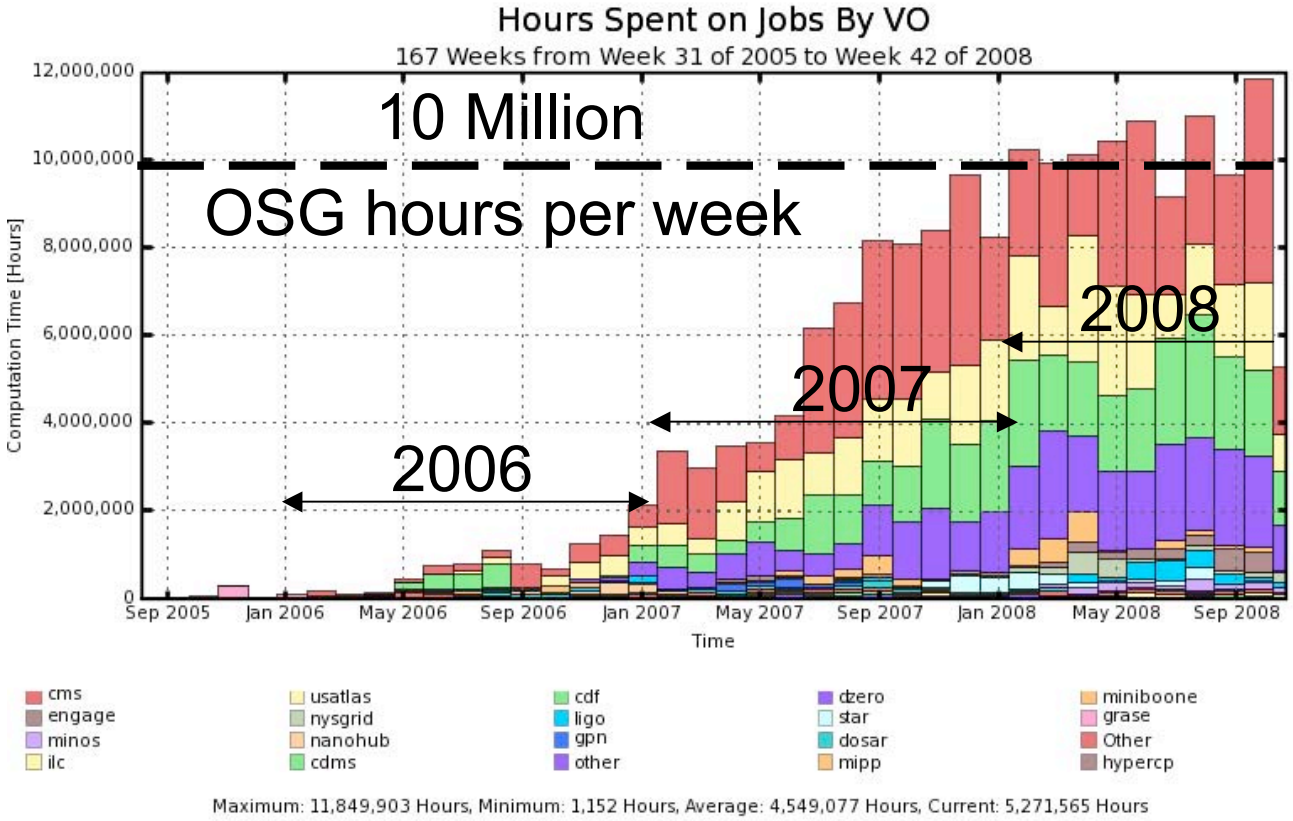
Building a layer of middleware that is specific to, and supported by your science community is an essential ingredient for success !

OSG can help by sharing of ideas, experience & successful software modules.
OSG can not take on the responsibility for your community long term.

Steady state
~ 3 Gbps
Peaks above
10 Gbps

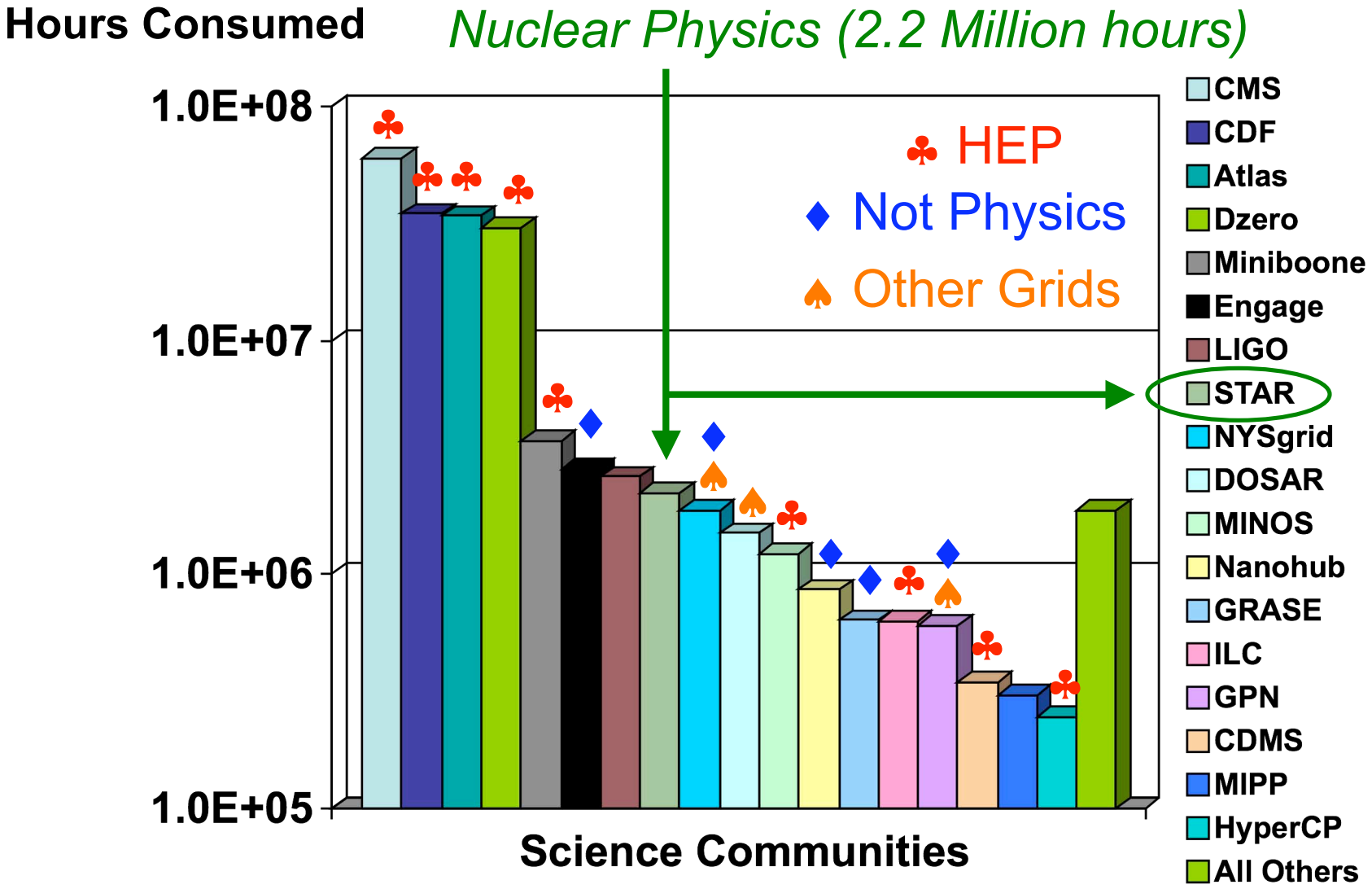


Bytes & Cycles

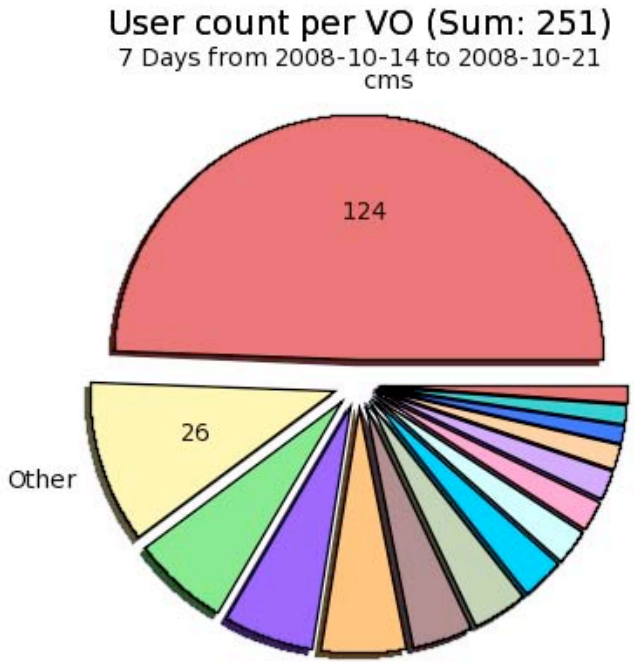


Rapid Rise in 2007 due to LHC Computing buildup. Expect x10 by 2012.

Cycles by Science Community



Number of individual Scientists Active last week.



251 Scientists across more than 20 scientific communities ran jobs on Open Science Grid last week.
About half of them are from CMS.

Note: Some communities run all jobs as only one user.
E.g. Atlas had 69 users accounted as one within last week.

Science on the grid

From Easy to hard

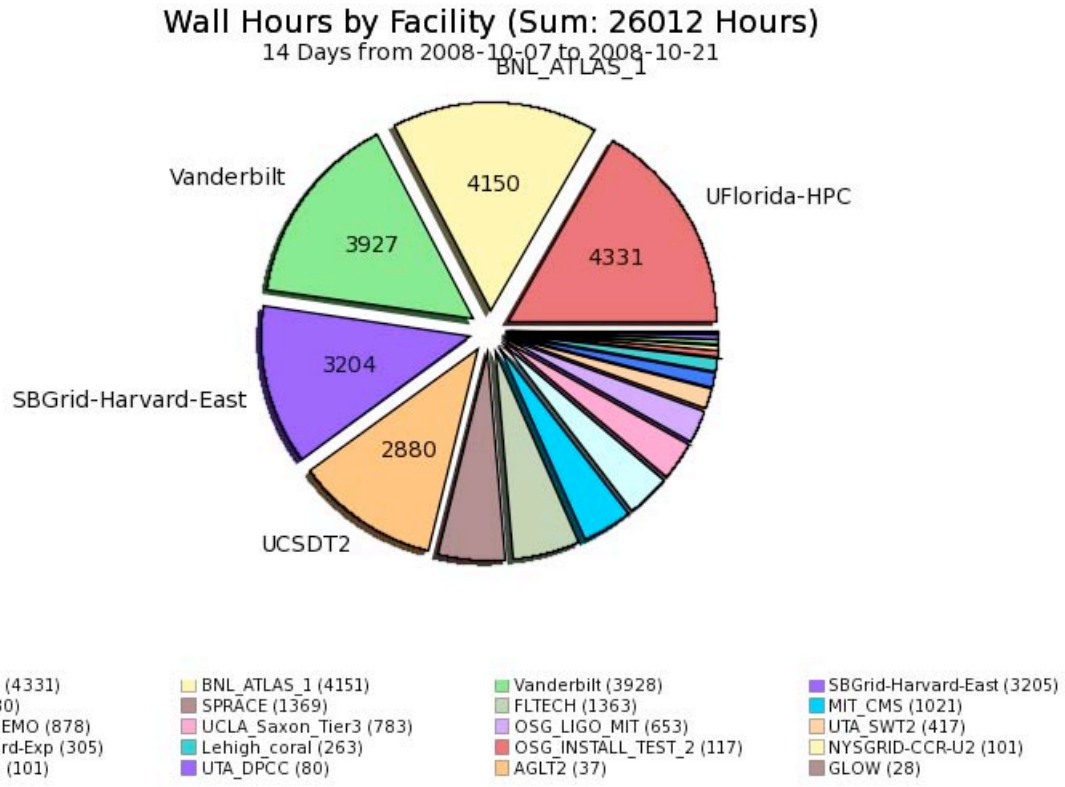
Some case studies

Easy Applications

- Trivially Parallel Simulation
- “zero” inputs
  - Statically linked exe with some scripts & random numbers.
- Small output for large CPU consumption
  - 10 hours per job
  - Less than 1GByte output file
- Stage out directly to large storage @ home
  - Requires storage @ home to scale to operations!

Expect success rate to be limited primarily by scalability of Science Community infrastructure @ home

Example Genomics analysis

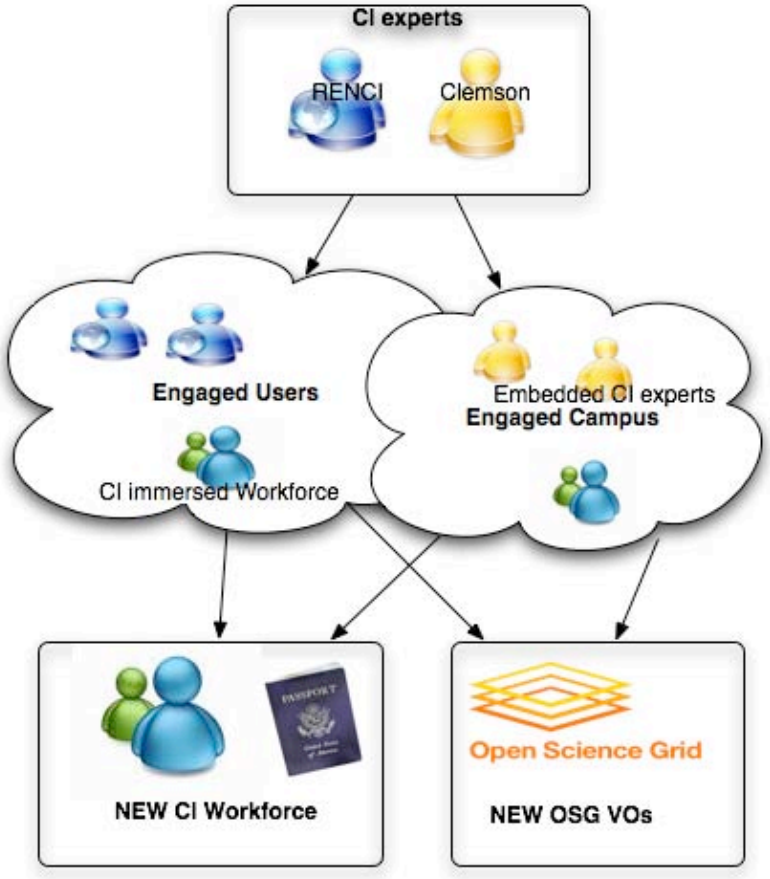


A new application started on OSG 3 days ago.
It immediately spread out its activity across many sites,
in the US and Brazil. This was possible by having a campus grid
(GLOW at UW Madison) fully integrated into OSG.

Campuses and CI-TEAM

CLEMSON SCHOOL OF COMPUTING

CI-TEAM is a NSF award to outreach to campuses, help them build their cyberinfrastructure and make use of it. As well as help campus users run their applications on the national OSG infrastructure. *"Embedded Immersive Engagement for Cyberinfrastructure, EIE-4CI"*



- **Provide help to build cyberinfrastructure on campus (compute resources).**
- **Provide help to make your application run on "the Grid"**
- **Train experts on your campus.**

- ***http://www.eie4ci.org***

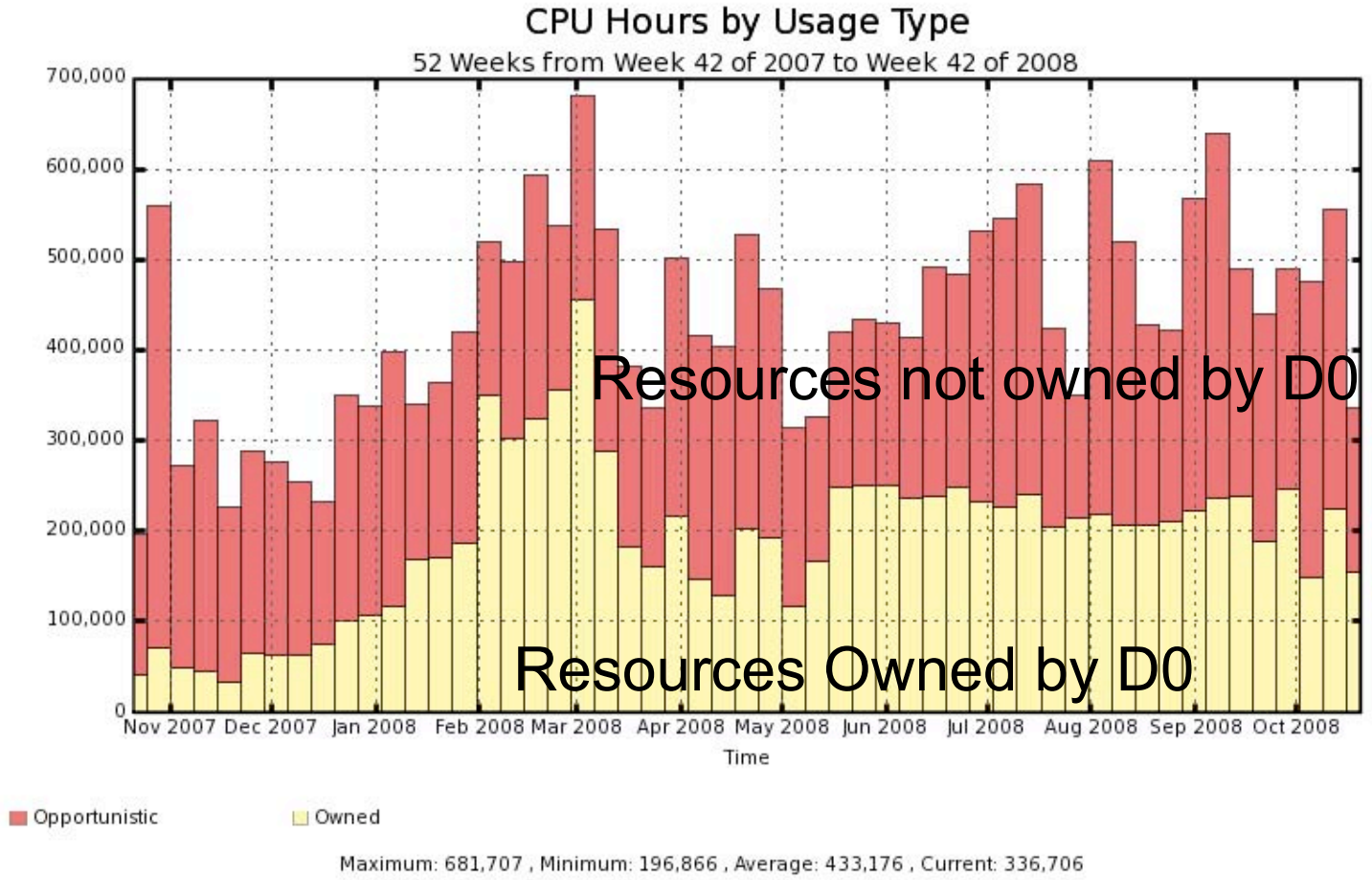
Thanks to S.Goasguen for this slide.

Medium Difficulty

- Trivially Parallel CPU intensive data processing
- Stage-in moderate size fixed dataset
  - 1TB of data across 1000 files
- Rerun many CPU intensive jobs on the same data, e.g. parameter sweeps.
  - Runtime of 10 hours for < 1GByte output file.
- Stage-out directly to large storage @ home
  - Requires storage @ home to scale to operations

Expect success rate to be limited primarily by scalability of Science Community infrastructure @ home. Not all clusters on OSG offer TB size space for all people.

Example: D0 within last year



D0 does roughly half of its total activity on OSG with cycles it does not own. => Scavenger !

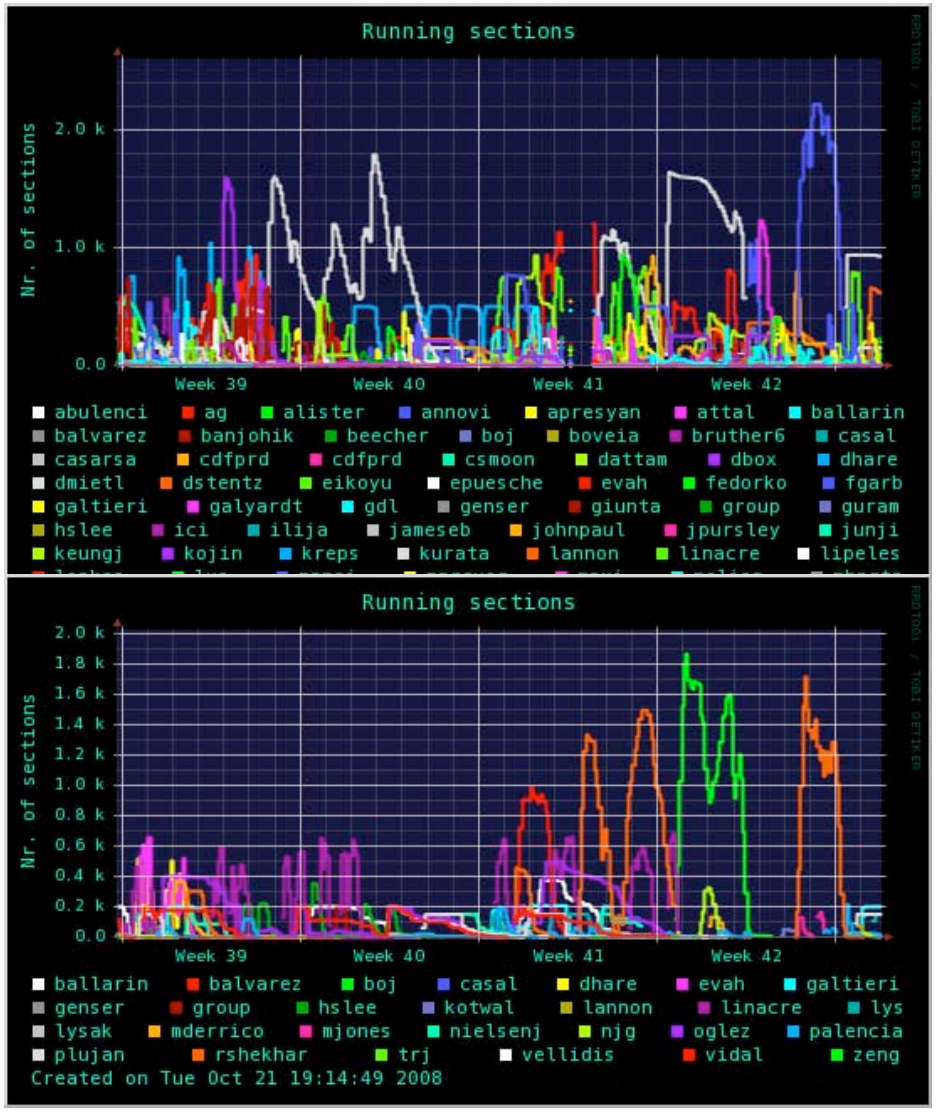
Hard but not impossible

- Trivially Parallel large IO data processing.
- Stage-in 10's of TB across 10's of thousands of files.
- Run many moderately CPU intensive jobs
  - 1 hour jobs that open many files
  - Output is input for another job, and thus needs to be stored in storage local to cluster.

Expect to think through your application workflow Carefully. You might need to own the cluster to control 10's of TB storage in conjunction with CPU.

Example: CDF last month

CDF operations on Fermigrid



CDF deploys its own resources at FNAL via the FNAL campus grid.

Fermigrid uses OSG Technology & provides spare cycles to scientists on OSG.

CDF uses the same software also to submit to the rest of OSG.

Easy apps -> OSG
Difficult apps -> stay at FNAL

← CDF operations on OSG outside FNAL.

To avoid unnecessary difficulties

- Avoid small files
  - Storage on OSG has transaction overheads per file.
- Stick to less than 10GByte total footprint per job.
  - That's all you get as local disk per core at many sites.
- Stick to run times of 1-10 hours.
  - Grid has scheduling overheads per job.
  - Distributed systems are inherently error prone. The longer you run, the more likely you run into trouble.
- Stick to Memory footprint < 1GByte per job.
  - Most sites will have at least this much per core.

MPI and alike

Still an R&D project.

Some MPI has been run, but it's not the strength of OSG !!!

CMS, a case study

Globally Distributed
Data Intensive Computing

The CMS Experiment

- **80 Million electronic channels**
  x 4 bytes
  x 40MHz

  ~ **10 Petabytes/sec** of information
  x 1/1000 zero-suppression
  x 1/30,000 online event filtering

  ~ 300 Megabytes/sec raw data to tape
  1 to 10 Petabytes of total data per year

- 2000 Scientists (1200 Ph.D. in physics)
  - ~ 180 Institutions
  - ~ 40 countries

- 12,500 tons, 21m long, 16m diameter

Computing Activities

- 80 Million Electronic Channels
  - Iteratively improve calibrations
- 100's to 1000's of particles per collision
  - Iteratively improve reconstruction

=> Expect few years of analysis to reach asymptotic limit of detector understanding.

- Develop physics analysis strategies.
- Derive physics results.

=> Expect different levels of derived format data that is reprocessed, replicated, archived, retrieved many times over by many people and groups.

CMS Computing Top Down

Year	2008	2010	2012
tape [PB]	15	38	67
disk [PB]	15	27	44
CPU [MSi2k]	30	140	220

Chose a technological solution that allows
computing resources as distributed as human resources.

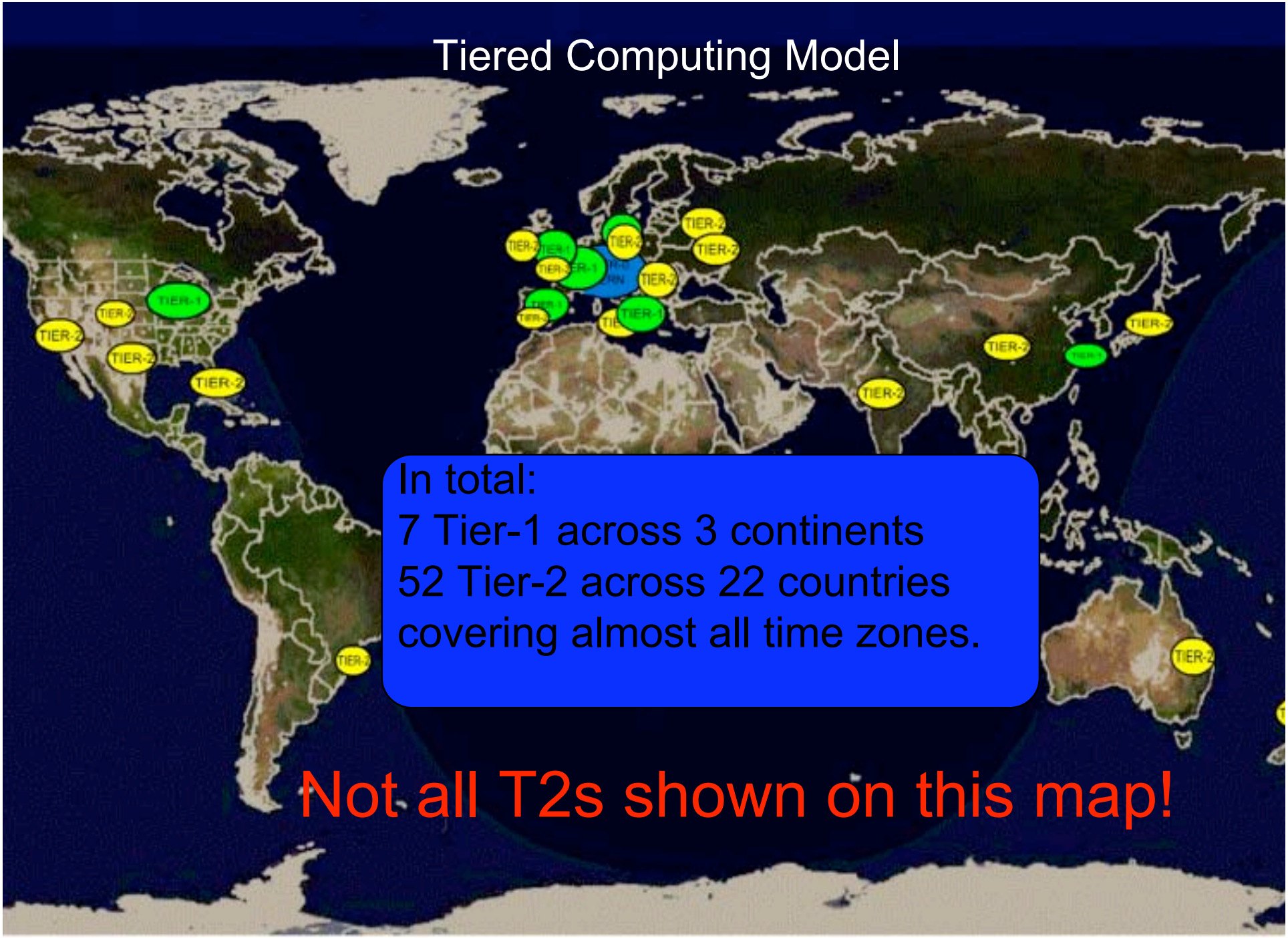
⇒ **Resource use controlled part "central" part "local".**

⇒ **Grid technology is the underpinning of our computing.**

⇒ Best match for organizational structures we are dealing with.

⇒ Extends naturally to "opportunistic resources" thus ~ doubling resources in US.

Tiered Computing Model



Tiered Computing Model

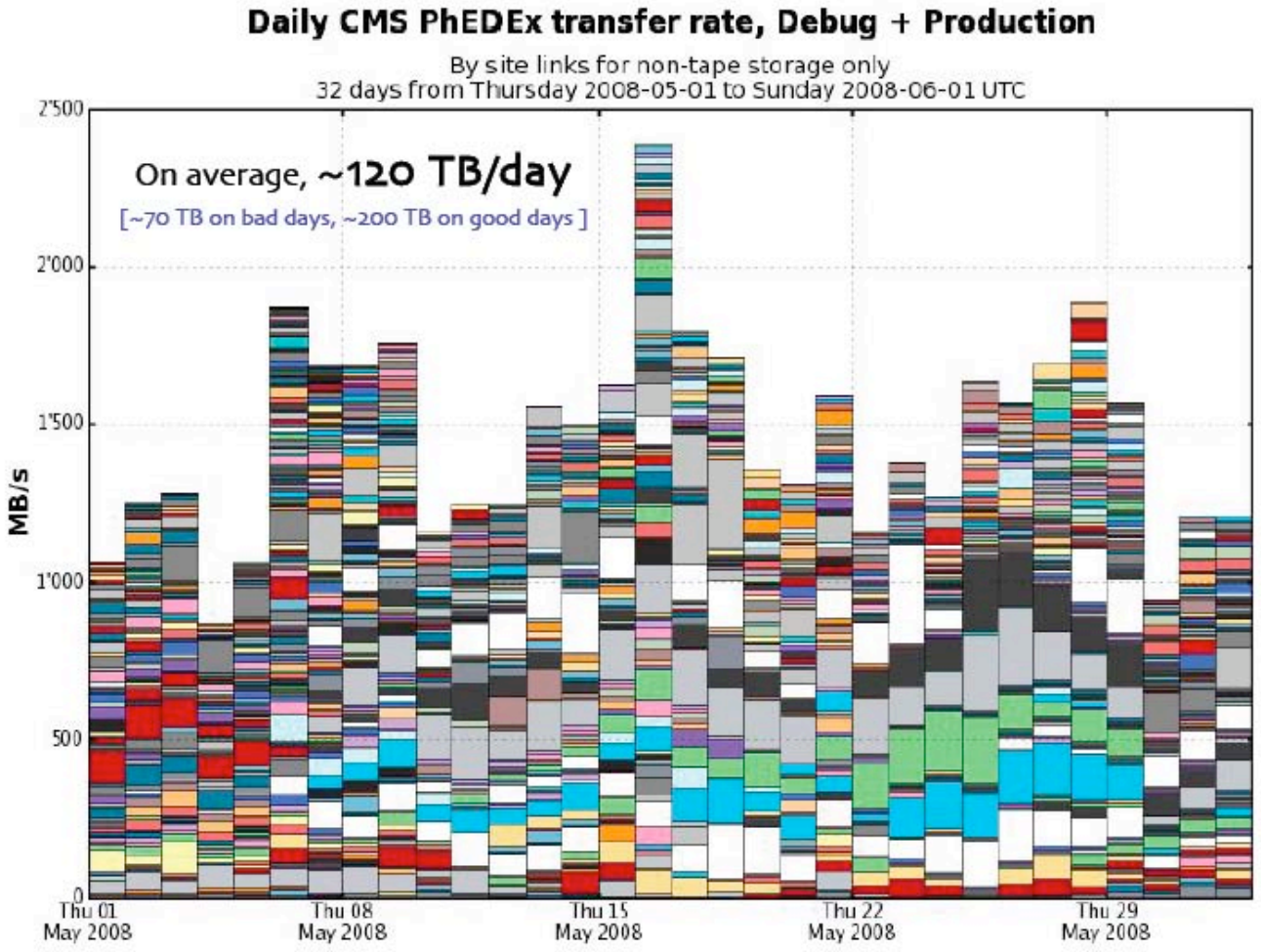
- T0 @ CERN: Prompt reprocessing and archival storage.
- 7x T1: “Custodial live archive” of the data & all “primary” reprocessing.
  - 100% centrally organized
- ~ 50x T2: MC production & physics analysis. Official, group, and user data hosting.
  - ~ 50% central, 50% “local” control
- ~ 100x T3: Resources at home institutions
  - 100% local control.
  - Used centrally only on opportunistic basis.
  - Operations support primarily by OSG,EGEE,Nordugrid.

Aside on Space Allocation Policy

- Each physicists has 1TB allocated at a site in the country they work for.
- Each official physics group has space allocated at multiple sites throughout the world.
- CMS centrally controls some space at all sites for hosting primary data, as well as stage-out space during processing.
- Each T2 serves some “local” community(s), providing them with disk space and CPU power, in addition to providing space for the activities above.

CMS sites manage complex set of space allocation policies. This is done via grid tools but under local control.

Global Data Transfer



Impressive list of few hundreds of links...

End-to-End Analysis: Concepts

- Develop Application on laptop, execute on global grid.
- CMS official releases are installed at all sites.
- Data is placed, and jobs move to data.
- Anybody can request data placement, only destination site manager can approve.
  - All writes are done by local agent, under site's control.
- Site controls logical to physical filename mapping.
- Site provides job output stage-out space
  - Writes done by user to their user space at their site.

Things I did not talk about.

- STAR and CMS-Heavy Ions are the most significant users of OSG resources within Nuclear Physics.
- PHENIX is a significant user of grid tools on their own resources (srm, gridftp)
- ALICE is presently starting to operate on OSG.
  - The need for VOBox is a hurdle for ALICE.
- There are most likely other Nuclear Physics uses of grid technology and infrastructures that I am unaware of.

Summary & Conclusion

- Grid infrastructures and tools are in production at LHC scale.
  - 25TB/day data movements are routine between CMS sites on OSG.
  - 10 Million CPU hours per week are routine on OSG.
- Variety of other science communities are benefiting, in many different ways, from LHC investment into grid computing.
  - By adopting grid tools on their own infrastructures
  - By using OSG to access their own resources.
  - By using other people's resources on OSG.

Looking for more Nuclear Physics Communities to join this party.