OSG Storage Forum, Chicago, Sept. 21-22, 2010

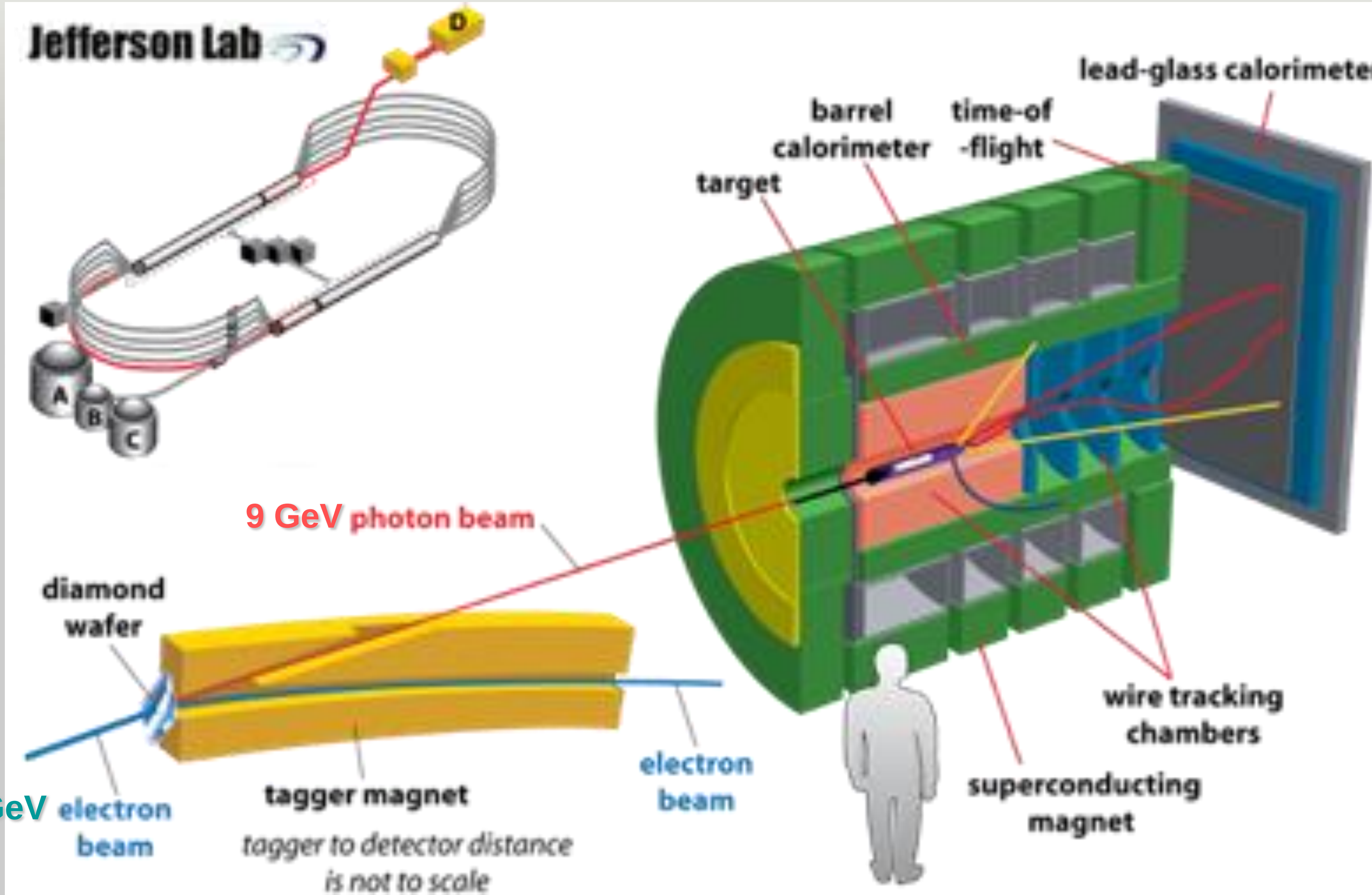
Gluex Storage Resources

Richard Jones, University of Connecticut

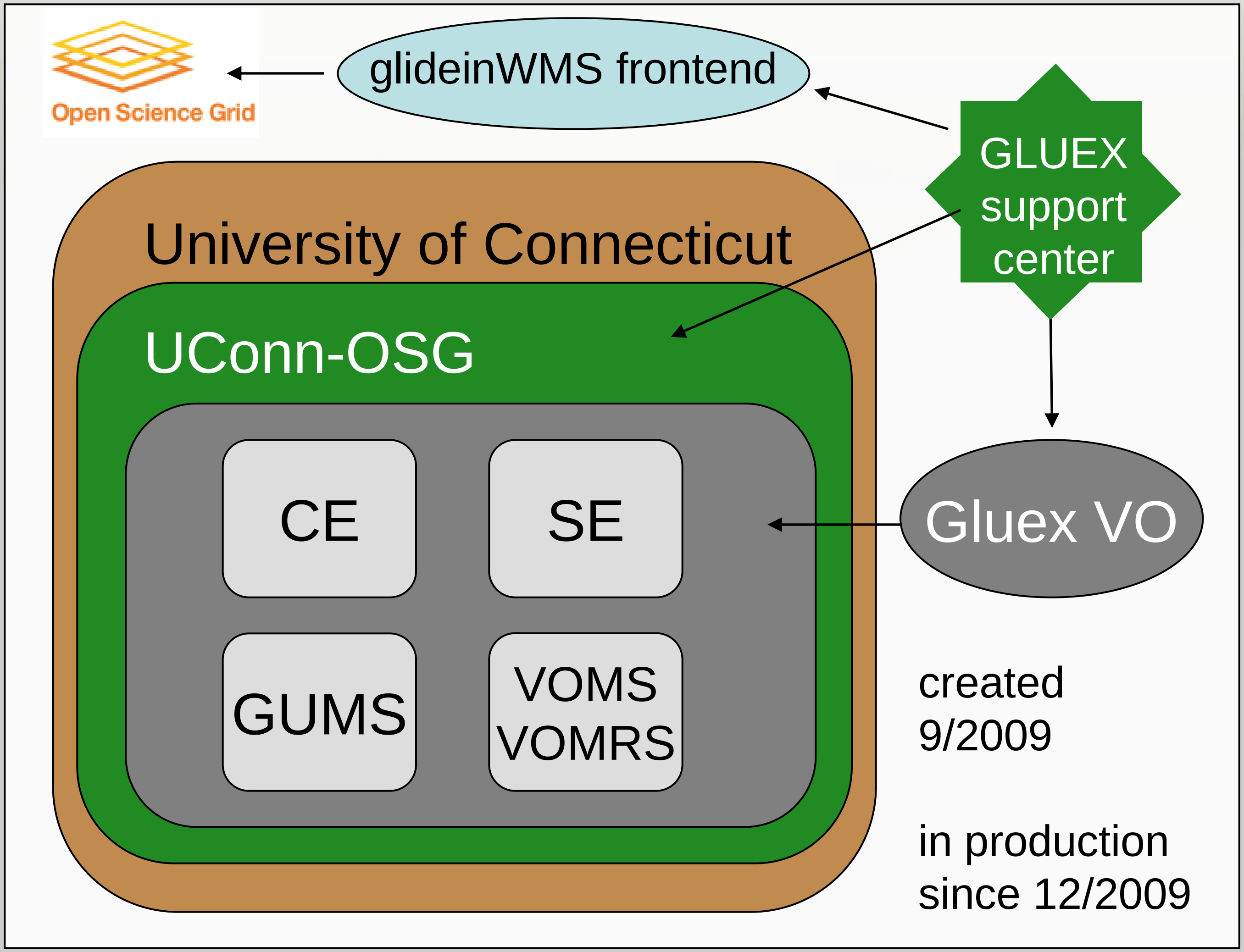
Outline

- who is Gluex?
- data storage and delivery needs
- existing resources, experience
- plans and outlook

Gluex – search for hybrid mesons



Gluex VO – the collaboration



member institutions

- University of Athens
- Carnegie Mellon Univ.
- Catholic University
- Christopher Newport Univ.
- University of Connecticut
- Florida International Univ.
- Florida State University
- University of Glasgow
- IHEP Protvino
- Indiana University
- Jefferson Lab
- U. of Massachusetts
- North Carolina A&T State
- U. of North Carolina
- Santa Maria University
- University of Regina

Gluex VO – data storage, delivery needs

- raw data: 10 kB/event, 20 kHz event rate = 2 TB / year
  - archived on Jlab site (tape library)
  - reconstruction -> DST with 5% of raw events, 20 kB/event

  200 TB/year, 5 years = 1 PB total for export offsite

- Monte Carlo: 20 kB/event, 100 kB/s on a 2.5GHz core2
  - minimum-biased event sample most challenging
  - ideally should approach raw data statistics
  - simulate, reconstruct, keep only MC DST

  100 TB/year, 5 years = 500 TB for provision offsite

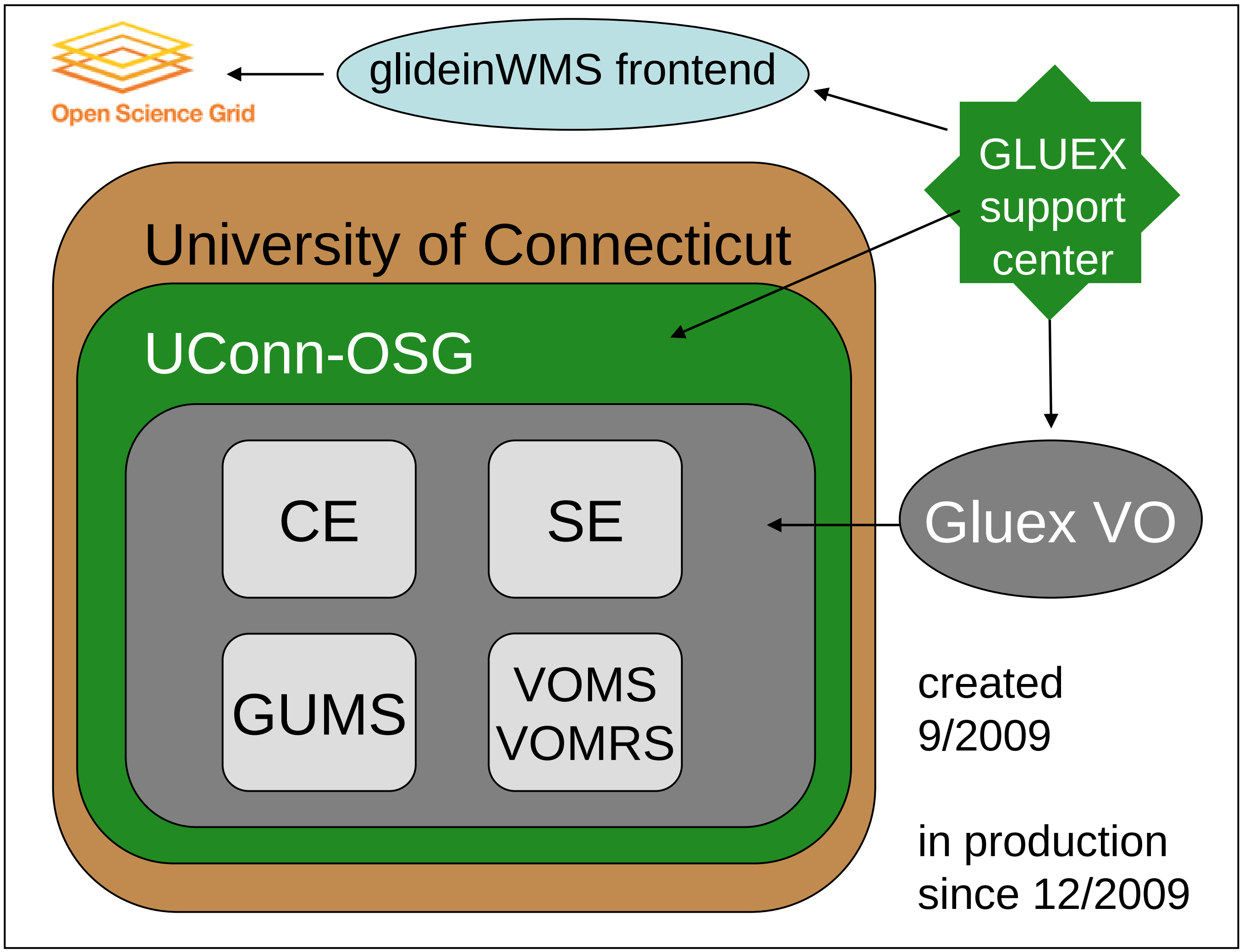
  - production targeted for OSG (min.bias sample: ***30M hours***)

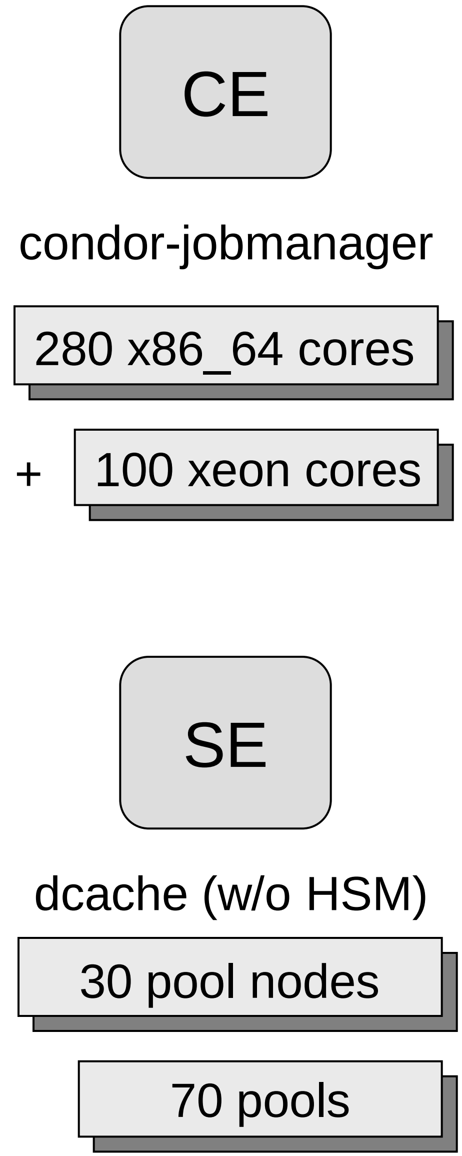
Gluex VO – data storage, delivery needs

- analysis: cuts to select exclusive final states
  - reduction jobs go to where data resides (major sites)
  - micro-DSTs (root trees, few TB each) per analysis
  - Monte Carlo (not min.bias) needed on-demand

- PWA fits: performed on dedicated GPU hardware
  - should be interactive
  - requires real-time access to micro-DSTs
  - may move toward scheduled GPU resources

Gluex VO: existing resources





Gluex VO: why dcache?

- **experience before dcache (pre 2004)**
  - **pvfs** – parallel virtual filesystem (*R. Ross, Clemson*)
  - 10K files (2-3 TB) splintered across 15 nodes
  - performance ok (could saturate network)
  - administration painful: **1 server down => file system hangs**
  - **kernel integration: encumbers OS upgrade scheduling**
  - metadata uncopyable, files unrecoverable if corrupted
  - **zero redundancy, frequent data loss**

- **dcache seemed to answer many of these problems**
  - layered on top of an ordinary unix filesystem
  - uses the built-in kernel nfs support (no custom kernel modules)
  - metadata stored in a standard database
  - filesystem robust against single pool node failures

Gluex VO: why dcache?

- **experience with dcache (2004-2009, pre-OSG)**
  - peak performance somewhat worse than pvfs (factor 2-3)
  - net throughput with parallel jobs was about equal to pvfs
  - overall experience was <u>much, much better</u>
    - rare data loss (3-4 times in 5 years, human error)
    - robust hands-off operation for weeks at a time (~1TB i/o per wk)
    - stable across OS upgrades
- **recent experience (with operation as a OSG SE)**
  - requires considerable work to keep it running
  - suffers from an authentication bottleneck (GUMS timeouts)
  - seeing out-of-heap-memory errors under heavy load
  - SRM response seems sluggish (30s for a short ls)
  - first time full authentication layers are exercised

Gluex Storage: Plans and Outlook

Why dcache might work for us:

1. the right mix of protocols: SRM/gridftp, xrootd, plain http-get
2. flexible configuration with control over replica management
3. nfs namespace introspection
4. ongoing development, large user base
5. no kernel-space code

Why dcache might not work for us:

1. authentication/authorization performance
2. SRM transaction overhead
3. lack of a comprehensive “fsck” tool
4. pain of administration (robustness under realistic conditions)

Next on our list to evaluate: hadoop