Panel: 20 Minutes Into Our Future

Near-term technology panel discussion between facility operations, applications developer, and users

Joint Facilities User Forum on Data-Intensive Computing

Argonne NATIONAL LABORATORY

Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory

BERKELEY LAB Lawrence Berkeley National Laboratory

Los Alamos NATIONAL LABORATORY

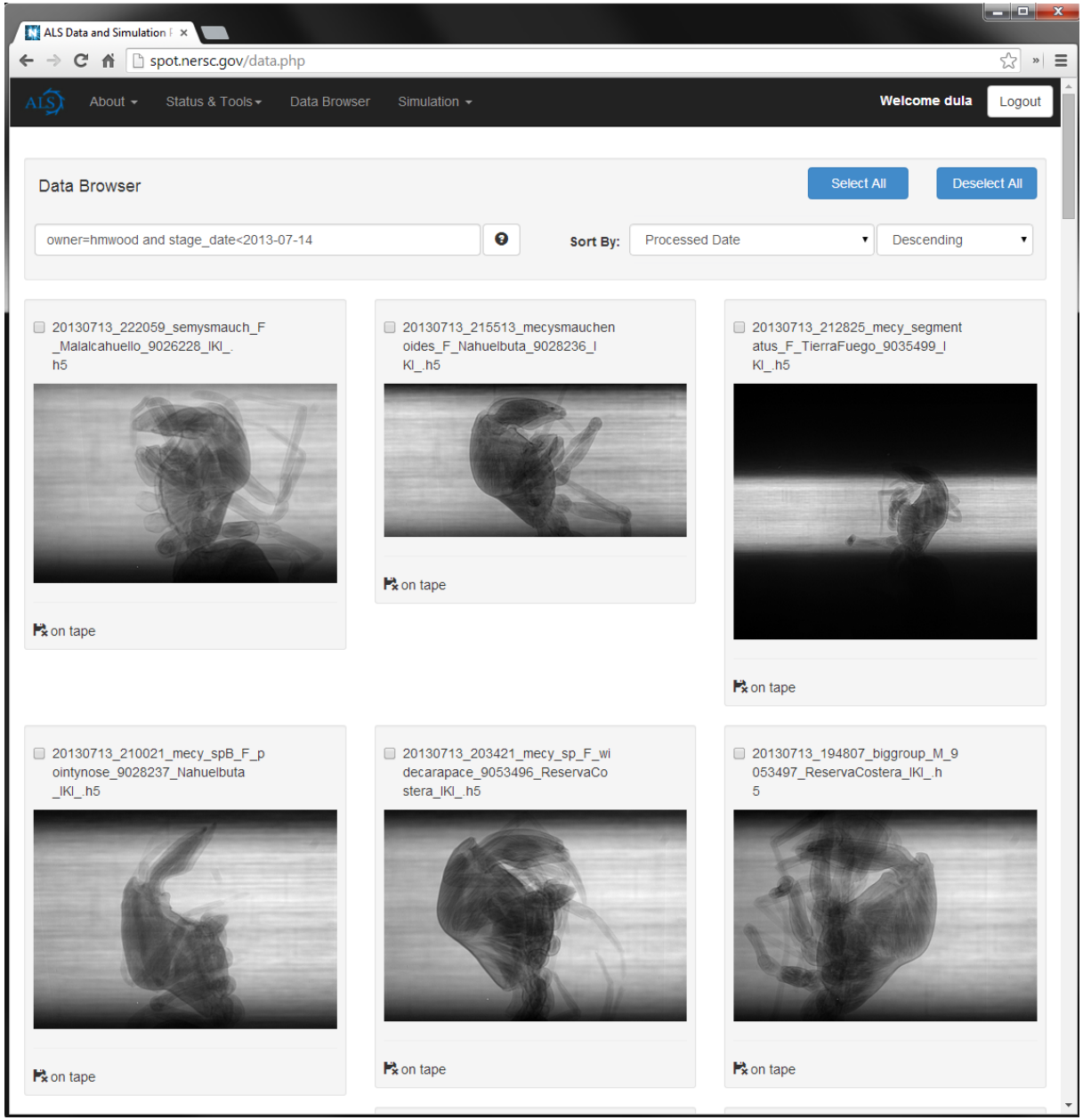
Sandia National Laboratories

Guiding Thoughts of this Panel

- Talking to the Compute, Store, Analyze cycle
  - Users
  - Developers
  - Operators/integrators
- What problems have we solved?
- What problems have we found?
- ***How do we inspire interesting dinner conversation for participants?***

Format

- Each panelist giving a 3-5 minute story
- Panel came up with a couple questions to get dialog started
- Open to audience questions



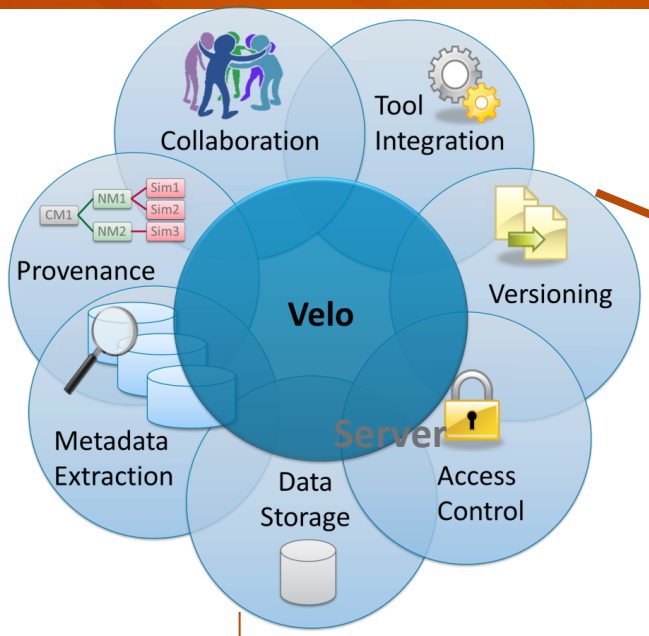
SPOT Suite:

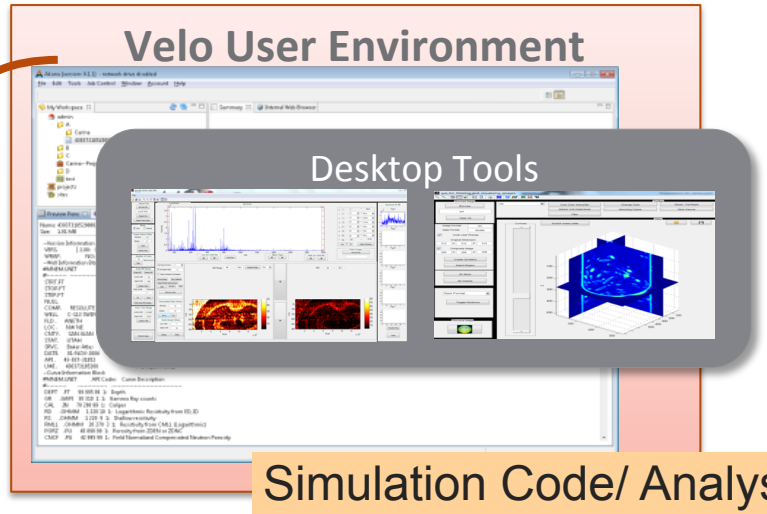
Making supercomputers available for users of light source facilities

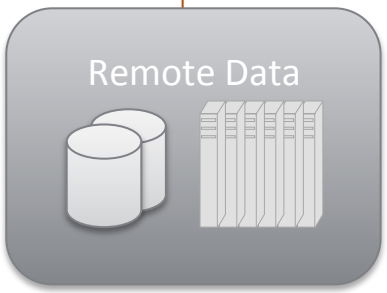
Craig Tull, Jack Deslippe, Simon Patton, Abdelilah Essiari, Alex Hexemer, Dula Parkinson

Velo: Knowledge and Tool Management for Collaborative Scientific Projects

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965







HPC Computing Environment

Simulation Code/ Analysis Tools/ Data/ Metadata / Workflow Relationship Management and Sharing

Access to remote storage and computing

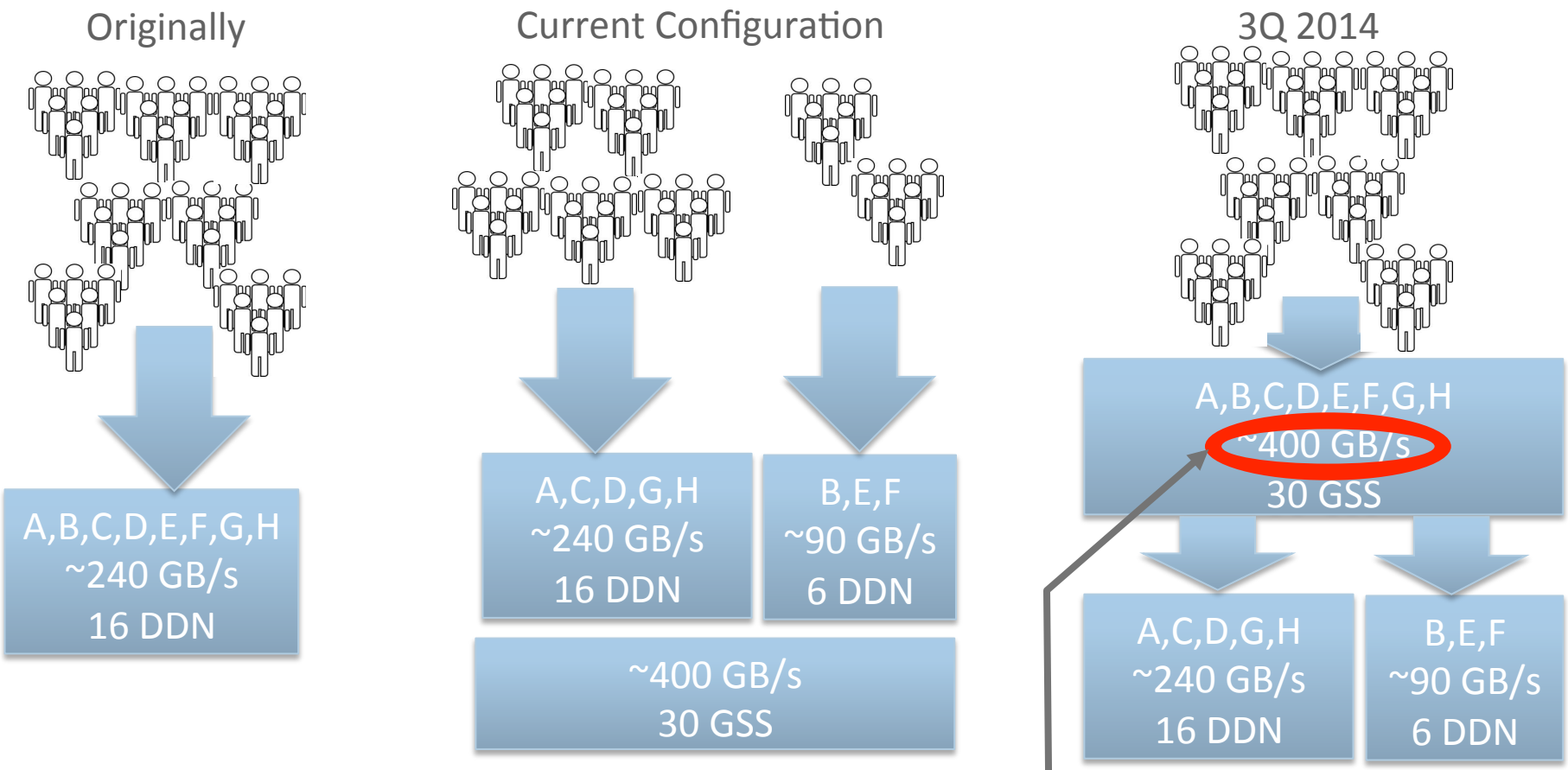
Tool Integration (connect modeling, analysis, and visualization tools to underlying data)

Simulation management, including UQ ensembles

Burst buffers, are applications ready for them?

- **BB Type 1 (BB as integrated part of compute system)**
  - SSDs are available via I/O nodes within a compute system
  - Options exist for using the scheduler to setup BB's for each job executed on the system
    - Node local storage (private)
    - Shared storage across job's compute nodes
  - Do I understand my I/O workload enough to know I can take advantage of the BB?
  - How much BB do I need?
  - To POSIX I/O, or not to POSIX I/O?
  - Do I also want access to the parallel file system (PFS)?
  - Should I just use the BB and have it all drain to the PFS?
  - Should I archive data from the BB?
- **BB Type 2 (BB external to the compute system, shared storage)**
  - Do I prefer having BB available across compute systems?
  - Are there any user features in software that improve performance or is this solution transparent?

Our storage upgrade path



Per GPFS Development, this is the fastest GPFS filesystem in the world

- IBM GPFS Storage Server (GSS)
  - Uses GPFS "Native RAID" (Software RAID)
  - No controllers; x3650 servers with attached JBODs

3D Map of the Universe

Stephen Bailey
LBNL

Current Generation

– 2.5M galaxies, 2600 jobs, 50k CPU-hours

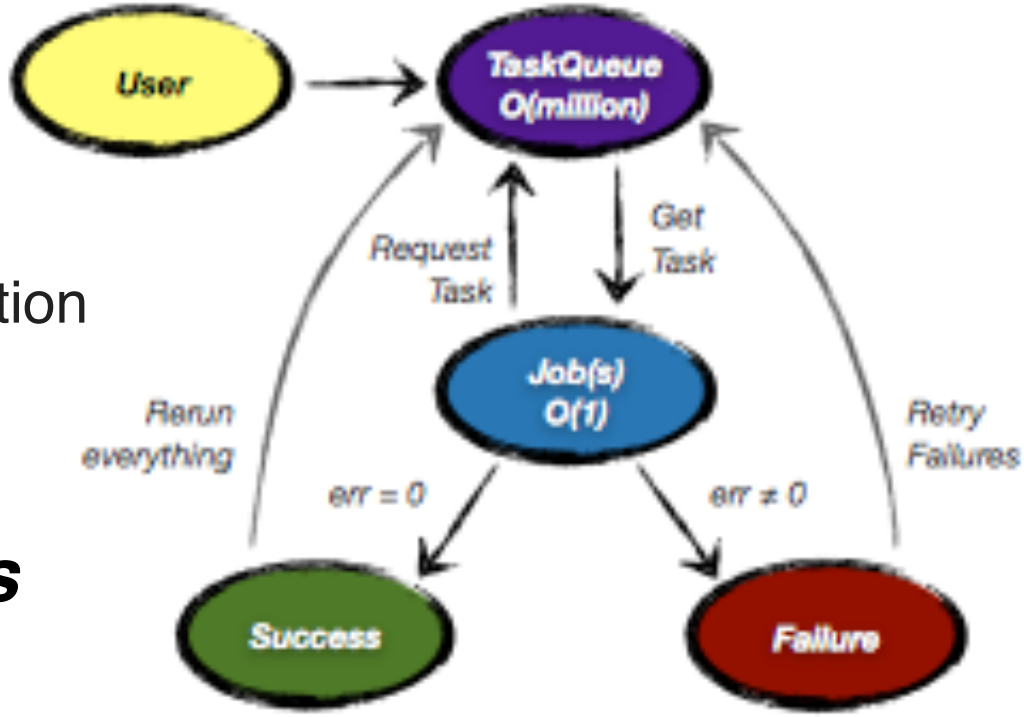
Next Generation

– 25M galaxies, 100k (?) jobs, 10M CPU-hours

How to adapt this workflow to NERSC-level supercomputers?

Common problem

- LBL Director's R&D grant
- Study options
- Develop ***cross-project*** solution



https://bitbucket.org/berkeleylab/qdo

Decouple *task queue* from worker *batch jobs*

- Scale # workers up/down
- Add tasks to queue after jobs have started
- Queue lives on even if jobs go up in flames
- Manage tasks in aggregate
- Lightweight, flexible

Data-intensive computing
Support workflow tools analogous to current support for compilers & numerical libraries

Questions…