A New Community Resource for Experiments at Scale: PRObE
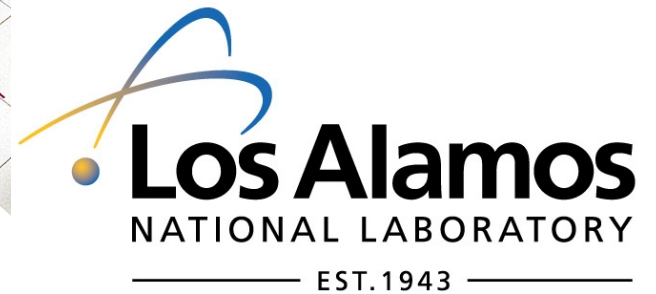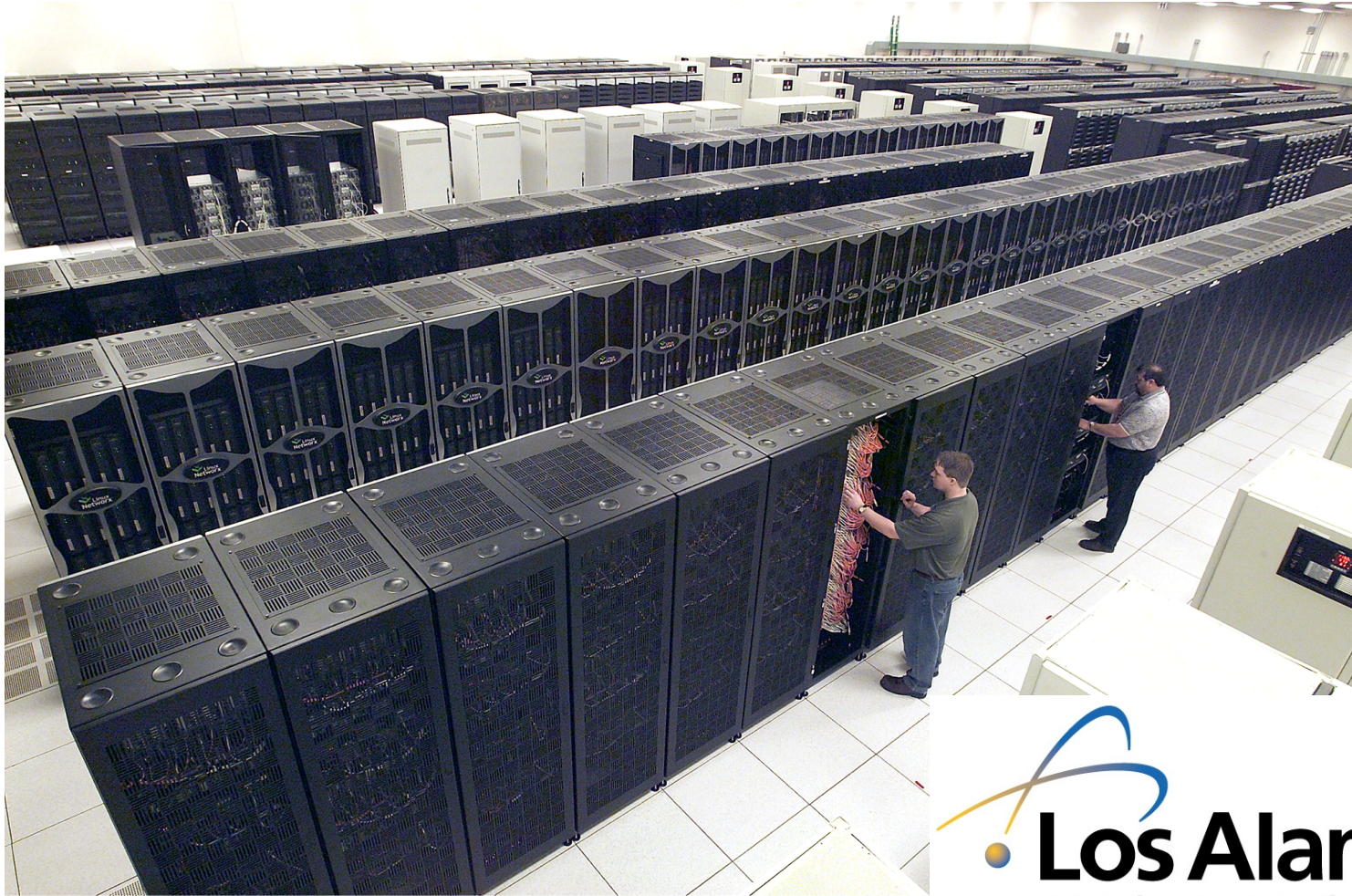
Garth Gibson, Carnegie Mellon University

Gary Grider, Los Alamos National Laboratory

Katharine Chartrand, New Mexico Consortium

Andree Jacobson, New Mexico Consortium

Carnegie Mellon
Parallel Data Laboratory

LANL is "giving us" Lightning

Carnegie Mellon
Parallel Data Laboratory

NSF Funds NMC to Recycle

- ## NSF funds PRObE (2011-2014)
  - ### Parallel Reconfigurable Observational Environment

- ## Large scale clusters for systems researchers
  - ### For dedicated use, long periods of time (days, weeks)
  - ### Allow replacement of any and all software

Carnegie Mellon
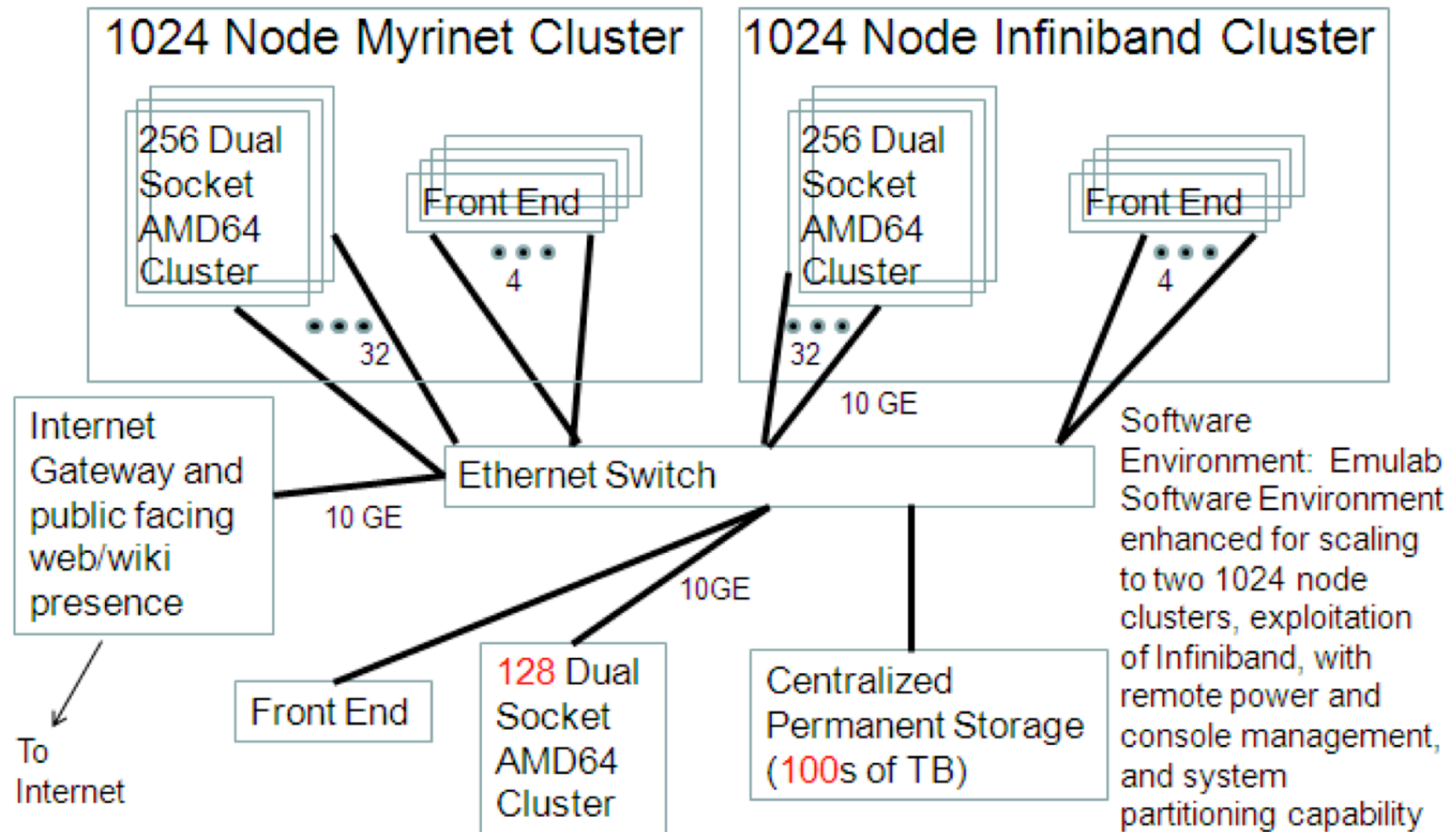**Parallel Data Laboratory**

Hardware Plan

- ## Fall 2011: Sitka (2048 cores) -- allocated
  - 1024 Nodes, Dual Socket, Single Core AMD Opteron; 2 GB per core; Myrinet

- ## Fall 2012: Kodiak (2048 cores) -- identified
  - 1024 Nodes, Dual Socket, Single Core AMD Opteron; 4 GB per core; SDR Infiniband

- ## Fall 2013: Nome (1600 cores)
  - 200 Node, Quad Socket, Dual Core AMD Opteron; 2 GB per core; DDR Infiniband

- ## Plus
  - Ethernet & Fat-tree high-speed interconnect

Hardware Plan II

- Small (128 nodes) staging clusters, and
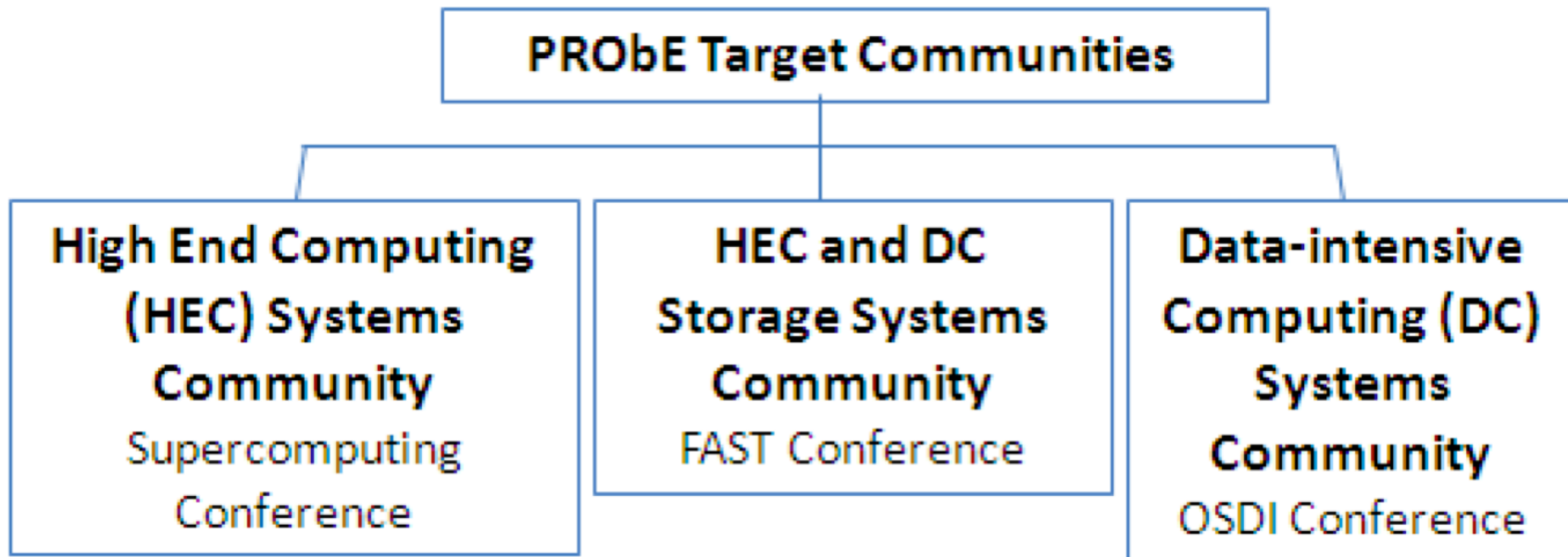
- Smaller (buy new) higher-core-count clusters
  - Summer 2011: Susitna (1728 cores) -- tbd
    - 36 Nodes, Quad Socket, 12 core AMD (?); 1-2GB RAM per core; EDR Infiniband high-speed interconnect
  - Summer 2013: Matanuska (3456 cores)
    - 36 Nodes, Quad Socket, 24 core AMD (?); 1-2GB RAM per core; 100 GigaBit Ethernet (or similar)

Carnegie Mellon
Parallel Data Laboratory

Two Large Machine Hardware Environment

1024 Node Myrinet Cluster

256 Dual Socket AMD64 Cluster

Front End

• • •
4

• • •
32

1024 Node Infiniband Cluster

256 Dual Socket AMD64 Cluster

Front End

• • •
4

• • •
32

10 GE

Internet Gateway and public facing web/wiki presence

To Internet

Ethernet Switch

10 GE

Front End

10GE

128 Dual Socket AMD64 Cluster

Centralized Permanent Storage (100s of TB)

Software Environment: Emulab Software Environment enhanced for scaling to two 1024 node clusters, exploitation of Infiniband, with remote power and console management, and system partitioning capability
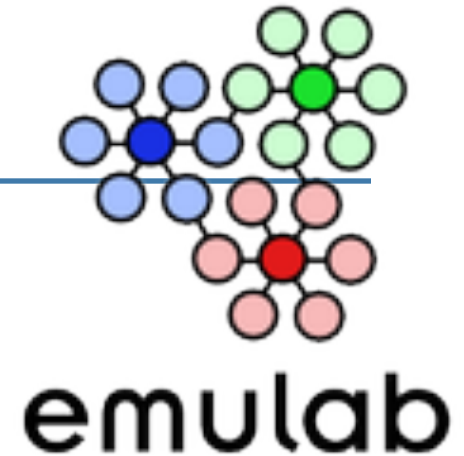
Parallel Data Laboratory

For Systems Research Users

- NSF "who can apply" rules
  - Includes international and corporate research projects ("best" in partnership with US university)

PRObE Target Communities

High End Computing (HEC) Systems Community
Supercomputing Conference

HEC and DC Storage Systems Community
FAST Conference

Data-intensive Computing (DC) Systems Community
OSDI Conference

Software

- First, "none" is allowed
  - Researchers can put any software they want onto the clusters

- Second, a well known tool managing clusters of hardware for research
  - Emulab (www.emulab.org), Flux Group, U. Utah
  - On staging clusters, also on large clusters
  - Enhanced for PRObE hardware, scale, networks, resource partitioning policies, remote power and console, failure injection, deep instrumentation

- PRObE provides hardware support (spares)

Allocation

- Competitive (target a few pages per proposal)
  - Justified for research needing PRObE resources
  - Not for cycles – for systems research
  - Results must be published & credit given
- Low threshold to get onto staging clusters
  - Emulab procedures wherever appropriate
- Allocation by community importance/merit
  - Committee recommends order & duration of use
  - Allocation opportunity tokens used to incent usage
    - Prompt return of resources, other contributions
    - Unused time offered to pending projects

Carnegie Mellon
Parallel Data Laboratory

PRObE Decision Making

- Committees usually about 6, selected by standard academic procedures (via BOFs)
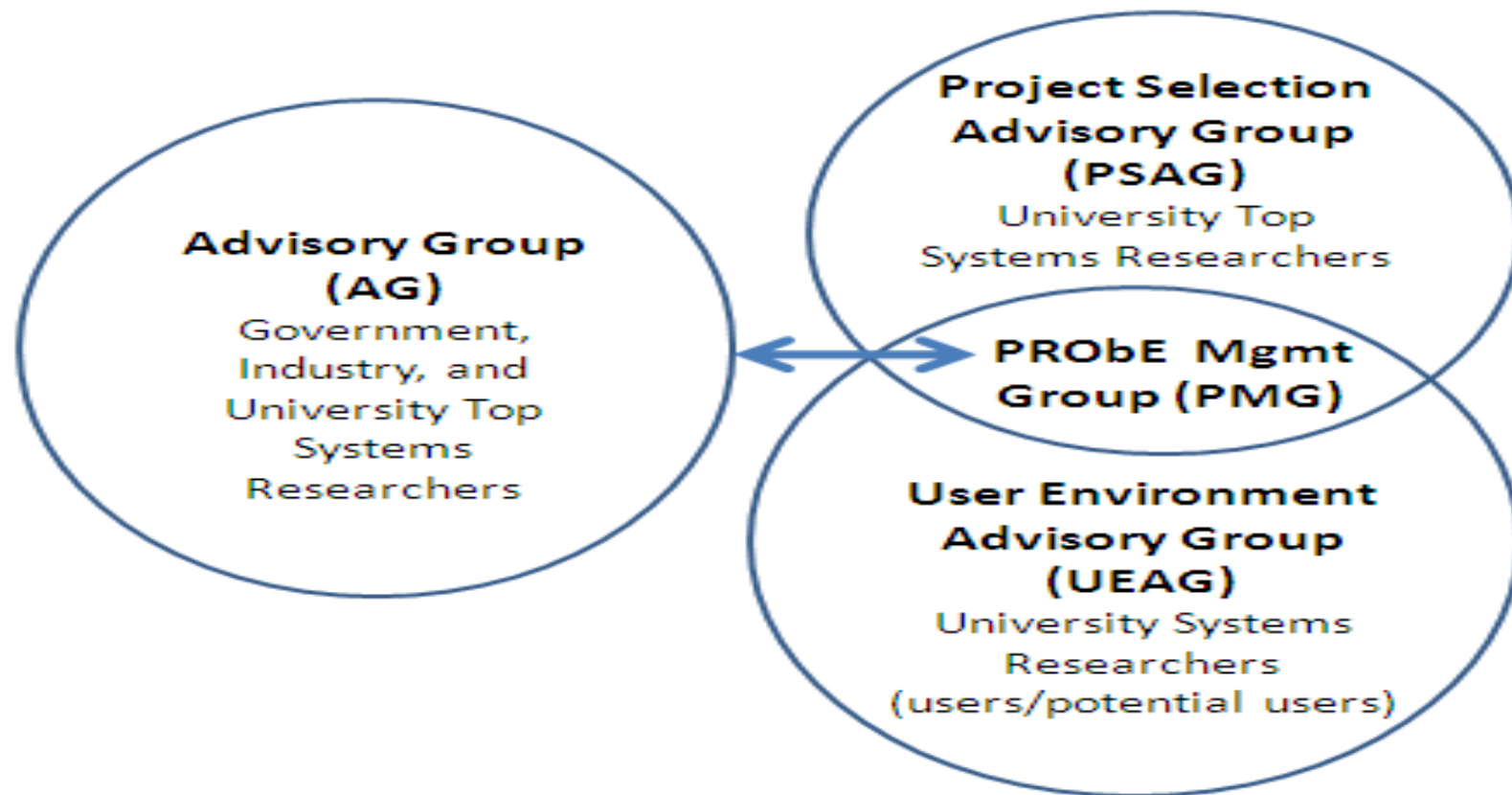
Next Steps

- Identify interested researchers & research
  - Seek candidates to steer (advisory committee)
  - Seek candidates to select program (project selection committee)
  - Seek candidates to shape experience (user environment advisory committee)
- Seek advice on anything else

- probe@newmexicoconsortium.org
- http://newmexicoconsortium.org/probe