PRObE: A New Community Resource for Experiments at Scale

http://newmexicoconsortium.org/probe

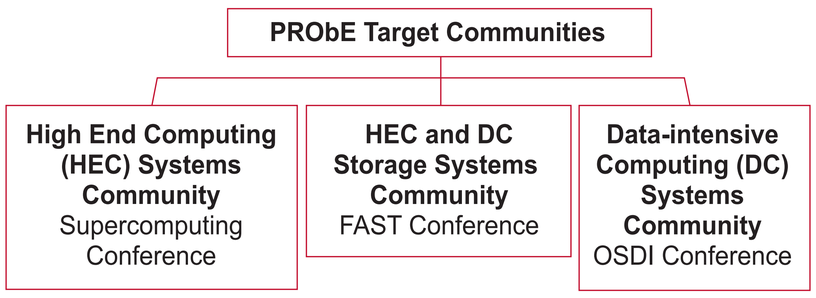
Garth Gibson, Carnegie Mellon University; Gary Grider, Los Alamos National Laboratory; Katharine Chartrand, New Mexico Consortium; Andree Jacobson, New Mexico Consortium

Backstory

- **August 2006, FSIO workshop challenge to panel**
  - **How can gov't help academics do better at-scale work**
  - **Answer: tools to develop experience at scale**
- **Early 2008, LANL says "still a problem?"**
  - **Large virtualized resources available for user-level experience, but at level of bare metal**
  - **LANL says "and if we gave you a huge cluster…."**
- **Late 2008, NSF gets a proposal**

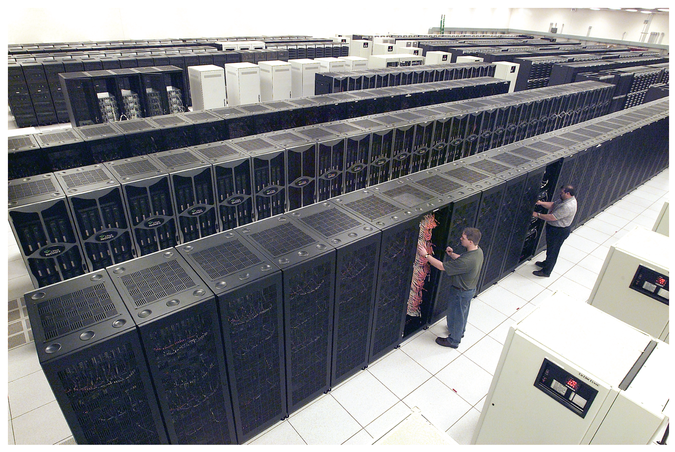
NSF Funds NMC to Recycle

- **Large scale clusters for systems researchers**
  - **For dedicated use, long periods of time (days, weeks)**
- **PRObE (Parallel Reconfigurable Observational Environment) funded by NSF (2011-2014), in Los Alamos, NM**



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**

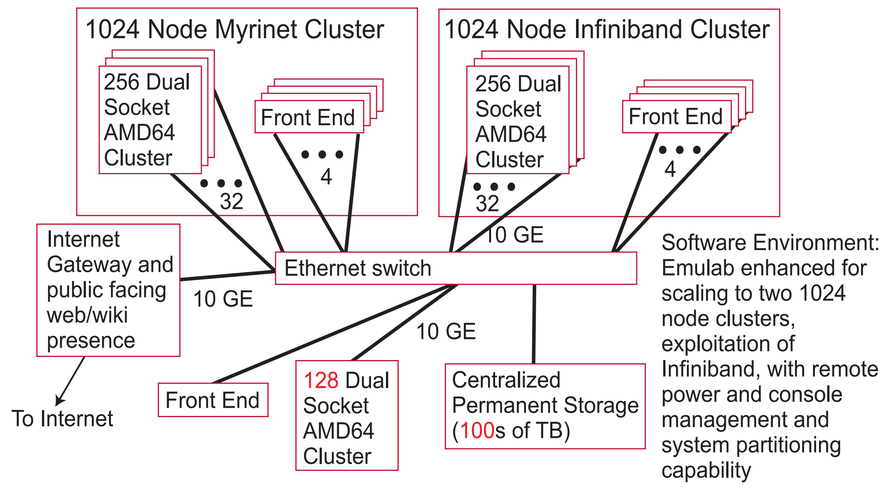


Los Alamos contributing Lightning-class supercomputer clusters

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)**

Two Large Machine Hardware Environment



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

Software

- **Researchers can use any software they want on the clusters**
- **Emulab (www.emulab.org), a well known tool for managing clusters of hardware for research**
  - **On staging clusters, also on large clusters**
  - **Enhanced by Univ. of Utah for PRObE hardware, scale, networks, partitioning policies, remote power/console, failure injection, deep instrumentation**
- **PRObE provides hardware support**

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**

Carnegie Mellon University

Los Alamos NATIONAL LABORATORY EST.1943

New Mexico CONSORTIUM

NSF