PRObE: A Thousand-Node Experimental Cluster for Computer Systems Research

GARTH GIBSON, GARY GRIDER, ANDREE JACOBSON, AND WYATT LLOYD

Garth Gibson is a Professor of Computer Science at Carnegie Mellon University and the co-founder and Chief Scientist at Panasas, Inc. He has an MS and PhD from the University of California at Berkeley and a BMath from the University of Waterloo in Canada. Garth's research is centered on scalable storage systems, and he has had his hands in the creation of the RAID taxonomy, the SCSI command set for object storage (OSD), the PanFS scale-out parallel file system, the IETF NFS v4.1 parallel NFS extensions, and the USENIX Conference on File and Storage Technologies.
garth@cs.cmu.edu

Gary Grider is the Acting Division Leader of the High Performance Computing (HPC) Division at Los Alamos National Laboratory. Gary is responsible for all aspects of high performance computing technologies and deployment at Los Alamos. Gary is also the US Department of Energy Exascale Storage, I/O, and Data Management National Co-coordinator. ggrider@lanl.gov

Andree Jacobson joined the New Mexico Consortium (NMC) in August 2010 as the Computer and Information Systems Manager and the project manager for the $10M NSF-sponsored PRObE project. Prior to NMC, he spent five years as a Computer Science Senior Lecturer at the University of New Mexico (UNM). During his time at UNM, he also spent his summers teaching the highly successful Computer Systems, Clusters, and Networking Summer Institute, now run as a part of the PRObE project.
andree@newmexicoconsortium.org

Wyatt Lloyd is a PhD candidate in Computer Science at Princeton University. His research interests include the distributed systems and networking problems that underlie the architecture of large-scale Web sites, cloud computing, and big data. He received his master's degree in Computer Science from Princeton University, and a bachelor's degree in Computer Science from Penn State University. wyatt.lloyd@gmail.com

If you have ever aspired to create a software system that can harness a thousand computers and perform some impressive feat, you know the dismal prospects of finding such a cluster ready and waiting for you to make magic with it. Today, however, if you are a systems researcher and your promised feat is impressive enough, there is such a resource available online: PRObE. This article is an introduction to and call for proposals for use of the PRObE facilities.

Server computing is increasingly done on clusters containing thousands of computers, each containing dozens of traditional cores, and the exascale supercomputers expected at the end of this decade are anticipated to have more than 100 thousand nodes and more than 100 million cores in total [1, 2]. Unfortunately, most academic researchers have only dozens of nodes with a handful of cores each. One of the best responses today is to rent a virtual datacenter from a cloud provider, such as Amazon or Google. We expect increasing numbers of papers to report experiments run on these virtual datacenters, but virtualization makes some experiments more difficult. Performance repeatability, network topology, and fault injection, for example, are not as well controlled on virtual datacenters as they

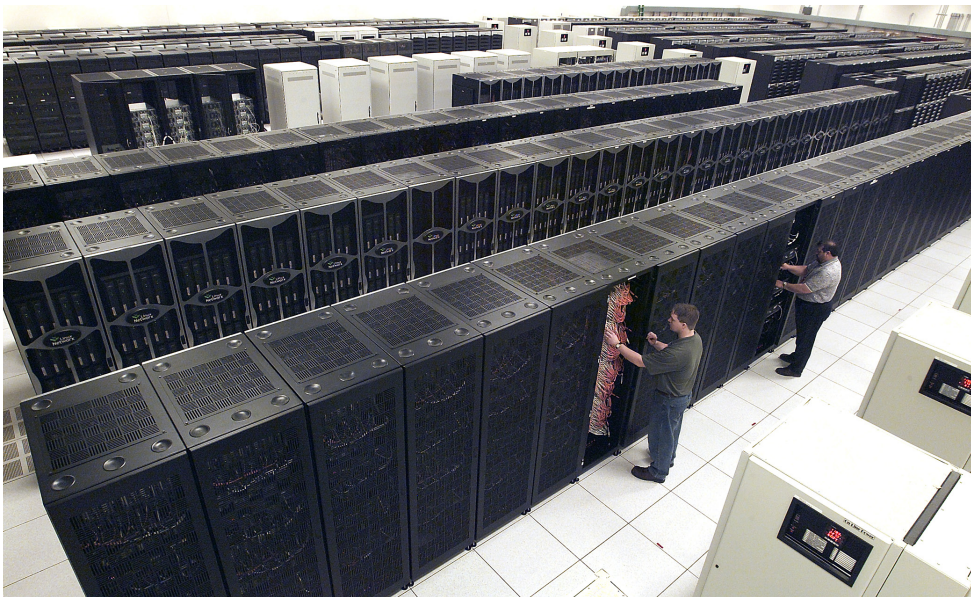

Figure 1: About one quarter of a Los Alamos National Laboratory supercomputer recently decommissioned and, probably, destroyed

are on physical datacenters. Moreover, debugging performance at scale is hard enough when all the hardware and software is known and controllable; learning from and perfecting innovative systems software in virtual datacenters is even harder. The systems research community needs access to larger-scale physical clusters, especially for the training of the next generation of computer systems scientists.

PRObE (Parallel Reconfigurable Observational Environment) is a systems research community resource for providing physical access to a thousand-node cluster. Made available by National Science Foundation operating support, equipment donations from the Los Alamos National Laboratory, and the facilities of the New Mexico Consortium, PRObE offers multiple clusters totaling more than 1,500 computers, with one cluster of more than 1,000 computers. The equipment in PRObE comes from computers at Los Alamos National Laboratory, such as shown in Figure 1, which have been decommissioned to make room for faster, more cost- and energy-efficient replacement computers. Researchers using PRObE have complete remote control of the hardware and software while running experiments and can inject both hardware and software failures as they see fit. Any operating system can be deployed on the systems using Emulab for physical cluster allocation [3].

PRObE is operational today. One of the first uses of PRObE's largest cluster was published in the 2013 Networked Systems Design and Implementation (NSDI '13) conference in a paper that validated the scalability of a geo-replicated storage system with causal consistency called Eiger [4]. Eiger's predecessor, called COPS, had been validated on only 16 nodes, whereas Eiger's use of PRObE allowed validation on up to 128 nodes (which, through replication, actually used 384 machines). Because a key contribution of Eiger is to scale to a large number of nodes per datacenter, while providing causal consistency and low latency with a rich data model and write-only transactions, having a large testbed was essential to the experiment. To quote the paper, "This experiment was run on PRObE's Kodiak testbed [results shown in Figure 2], which provides an Emulab with exclusive access to hundreds of machines. Each machine has 2 AMD Opteron 252 CPUS, 8GB RAM, and an InfiniBand high-speed interface." The Eiger paper is a fine example of the purpose of PRObE: enabling innovative systems to be tested at scale after they have been developed and explored on smaller private facilities.

To become a user of PRObE resources, follow these steps. First, all users of PRObE agree to publish, or otherwise make public, the results of PRObE use and give credit to the funders and providers of PRObE. Second, PRObE is an NSF-funded facility, so the organizations that request its use must be eligible to receive NSF funding. These constraints are explained in a user agreement on the PRObE Web site [5].

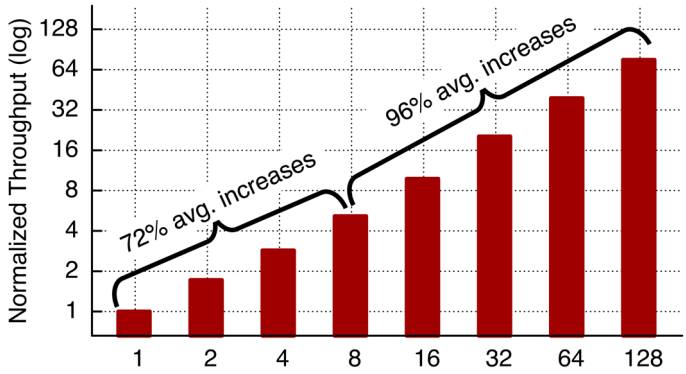

Figure 2: This figure shows the normalized throughput of multiple N-server clusters running the Facebook TAO workload [4]. Throughput approaches linear for up to 128 machines per cluster, using a total of 384 machines on PRObE's Kodiak cluster.

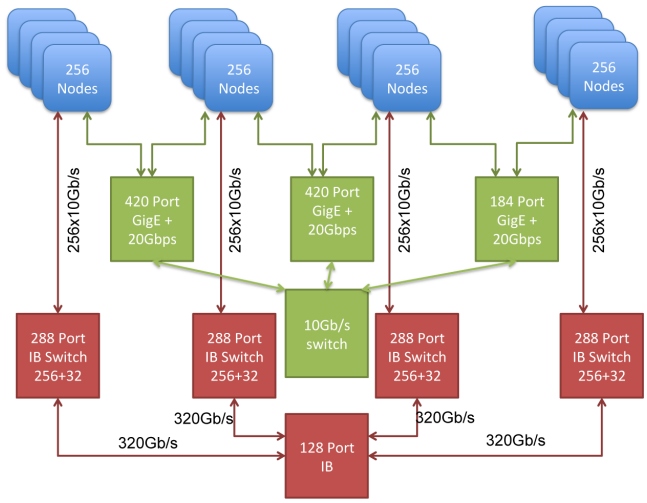

Figure 3: Block diagram of PRObE's Kodiak cluster

A new PRObE user is also an Emulab user. Emulab has been providing physical machine allocation and management in smaller clusters for more than a decade, and much of the systems research community already has experience with it. A new user logs in to a head node, launches a single node experiment with an existing base OS image, logs in to that node to customize the OS image as needed, instructs Emulab to record the customized image, then launches a multi-node allocation naming the customized image. Storage on the nodes is replaced with every launch but is fully available for experiments. Shared storage for images, inputs, and logging/monitoring results is available from Emulab head nodes and an NFS service.

PRObE's largest cluster, Kodiak, is intended to be allocated in its entirety to one project for days to weeks. New users should first log in to one of the smaller (~100 nodes) staging clusters, Denali or Marmot, to port their systems and demonstrate success on a small-scale experiment. Users then propose to use the

Machine	Nodes	Cores	Memory/node	Disk/node	Network/node
Marmot	128	256	16 GB	1 x 2 TB	GE, SDR Infiniband
Denali	64+	128+	8 GB	2 x 1 TB	GE, SDR Infiniband
Kodiak	1024	2048	8 GB	2 x 1 TB	GE, SDR Infiniband

Table 1: Currently available PRObE cluster capabilities

large cluster, with evidence of their readiness to be effective and an explanation of their project's goals and anticipated results. PRObE leadership and a community selection committee, when needed, will prioritize and arbitrate the use of the largest cluster.

The Parallel Reconfigurable Observational Environment (PRObE) is a collaboration between the National Science Foundation (NSF), under awards CNS-1042537 and CNS-1042543, New Mexico Consortium (NMC), Los Alamos National Laboratory (LANL), Carnegie Mellon University (CMU), and the University of Utah (Utah). PRObE facilities are available now and will be available for at least two years. For more information, visit the PRObE Web site at www.nmc-probe.org.

References

[1] Peter Kogge, Keren Bergman, Shekhar Borkar, Dan Campbell, William Carlson, William Dally, Monty Denneau, Paul Franzon, William Harrod, Kerry Hill, Jon Hiller, Sherman Karp, Stephen Keckler, Dean Klein, Robert Lucas, Mark Richards, Al Scarpelli, Steven Scott, Allan Snavely, Thomas Sterling, R. Stanley Williams, Katherine Yelick, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," DARPA IPTO AFRL FA8650-07-C-7724, 2008: citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.6676.

[2] John Shalf, "Exascale Computing Hardware Challenges," keynote talk at 5th Petascale Data Storage Workshop, New Orleans, LA, November 2010: slides available at http://www.pdsw.org/pdsw10/resources/slides/PDSW_SC2010_ExascaleChallenges-Shalf.pdf.

[3] Brian White, Jay Lepreau, Leigh Stoller, Robert Ricci, Shashi Guruprasad, Mac Newbold, Mike Hibler, Chad Barb, Abhijeet Joglekar, "An Integrated Experimental Environment for Distributed Systems and Networks," USENIX Symposium on Operating Systems Design and Implementation (OSDI '02), Dec. 2002: https://www.cs.utah.edu/flux/papers/netbed-osdi02.pdf.

[4] Wyatt Lloyd, Michael J. Freedman, Michael Kaminsky, David G. Andersen, "Stronger Semantics for Low-Latency Geo-Replicated Storage," Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13), Lombard, IL, April 2013: https://www.usenix.org/conference/nsdi13.

[5] PRObE User Agreement version 1, Feb. 1, 2013: www.nmc-probe.org/policies/user-agreement.