PRObE

Large Scale Systems Testing Facility

http://www.nmc-probe.org/

probe@newmexicoconsortium.org

New Mexico CONSORTIUM

PRObE Parallel Reconfigurable Observational Environment

Motivation

- Systems research community lacked very *large* dedicated resource for repeatable experiments, fault injection, and hardware control
- Research on large compute resources often constrained by imposed software stack, cost of extensive dedicated allocation

Los Alamos decommissions clusters

They mostly go into the chipper…

NSF PRObE to the rescue!

- NSF Funds the New Mexico Consortium (NMC) to bring LANL supercomputers back to life
- **PRObE** –

Parallel **R**econfigurable

Observational **E**nvironment

Los Alamos NATIONAL LABORATORY EST.1943

Carnegie Mellon University

New Mexico CONSORTIUM

THE UNIVERSITY OF UTAH

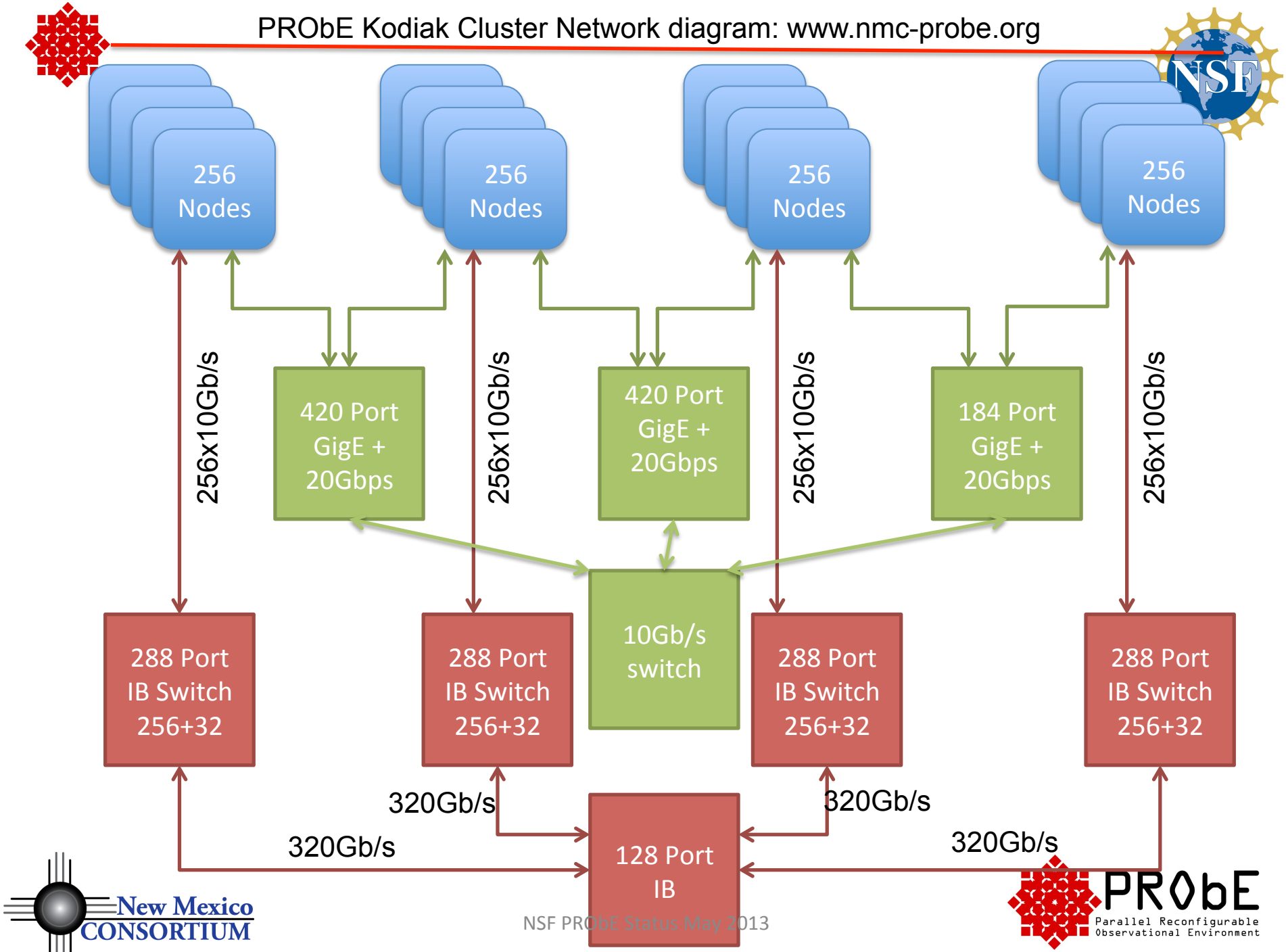
What is PRObE?

- NSF systems research community resources
- Days to weeks of dedicated large scale allocations
  - Physical and remote access
- Complete control of hardware and all software
- Enables fault injection, even destructive possible
- Targets parallel and data intensive usage
- Managed by community's own Emulab (www.emulab.net)

portal.nmc-probe.org

- *Marmot – 128 nodes / 256 cores (CMU)*
  - Dual socket, Single core AMD Opteron, 16GB/node
  - SDR Infiniband, 2TB disk / node
- *Denali – 64+ nodes / 128+ cores (NMC)*
  - Dual Socket, Single core AMD Opteron, 8GB/node
  - SDR Infiniband, 2x1TB disk / node
- *Kodiak – 1024 nodes / 2048 cores (NMC)*
  - Dual Socket, Single Core AMD Opteron, 8GB/node
  - SDR Infiniband, 2x1TB disk / node (1.8PiB total)

PRObE Kodiak Cluster Network diagram: www.nmc-probe.org



New Mexico CONSORTIUM

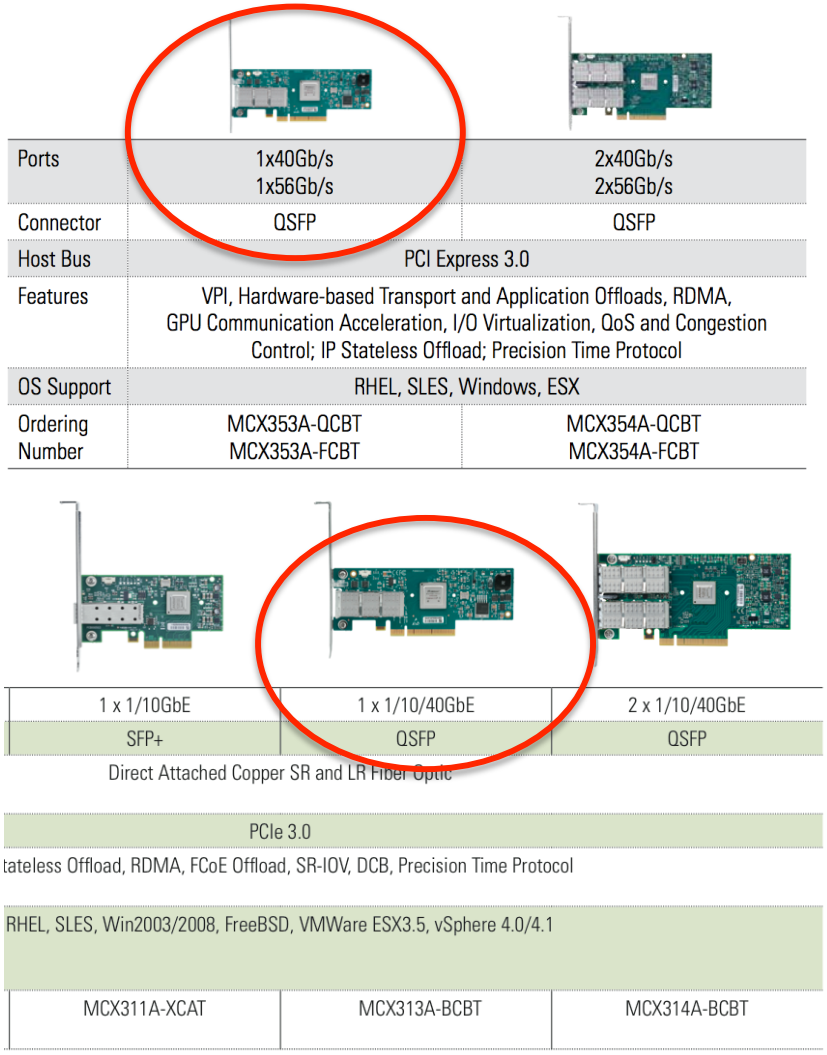
PRObE Parallel Reconfigurable Observational Environment

Coming soon: 34 Susitna Nodes

- 4 x AMD6272 = 64 core
  128 GB, 40GE, FDR10 IB,
  1TB OS + 2 x 3TB disk

Key Features

1. Four AMD Opteron™ 6000 series (6300 ready) processors (Socket G34) 16/12/8/4-Core ready; HT3.0 Link support
2. AMD SR5690/SR5670+SP5100 Chipset
3. Up to 1TB DDR3 1600MHz ECC Registered DIMM; 32x DIMM sockets
4. Intel® 82576 Dual-Port Gigabit Ethernet Controller
5. 6x SATA2 (3.0 Gbps) ports via AMD SP5100 Controller; RAID 0, 1, 10
6. 2x PCI-e 2.0 x16 slots, 1x PCI-e 2.0 x8 slots, and 1x UIO slot or 1x PCI-E x8 slot
7. Integrated Matrox G200eW Graphics
8. Integrated IPMI 2.0 with Dedicated LAN



Ports	1x40Gb/s 1x56Gb/s	2x40Gb/s 2x56Gb/s
Connector	QSFP	QSFP
Host Bus	PCI Express 3.0	
Features	VPI, Hardware-based Transport and Application Offloads, RDMA, GPU Communication Acceleration, I/O Virtualization, QoS and Congestion Control; IP Stateless Offload; Precision Time Protocol	
OS Support	RHEL, SLES, Windows, ESX	
Ordering Number	MCX353A-QCBT MCX353A-FCBT	MCX354A-QCBT MCX354A-FCBT

1 x 1/10GbE	1 x 1/10/40GbE	2 x 1/10/40GbE
SFP+	QSFP	QSFP

Direct Attached Copper SR and LR Fiber Optic

PCIe 3.0

tateless Offload, RDMA, FCoE Offload, SR-IOV, DCB, Precision Time Protocol

RHEL, SLES, Win2003/2008, FreeBSD, VMWare ESX3.5, vSphere 4.0/4.1

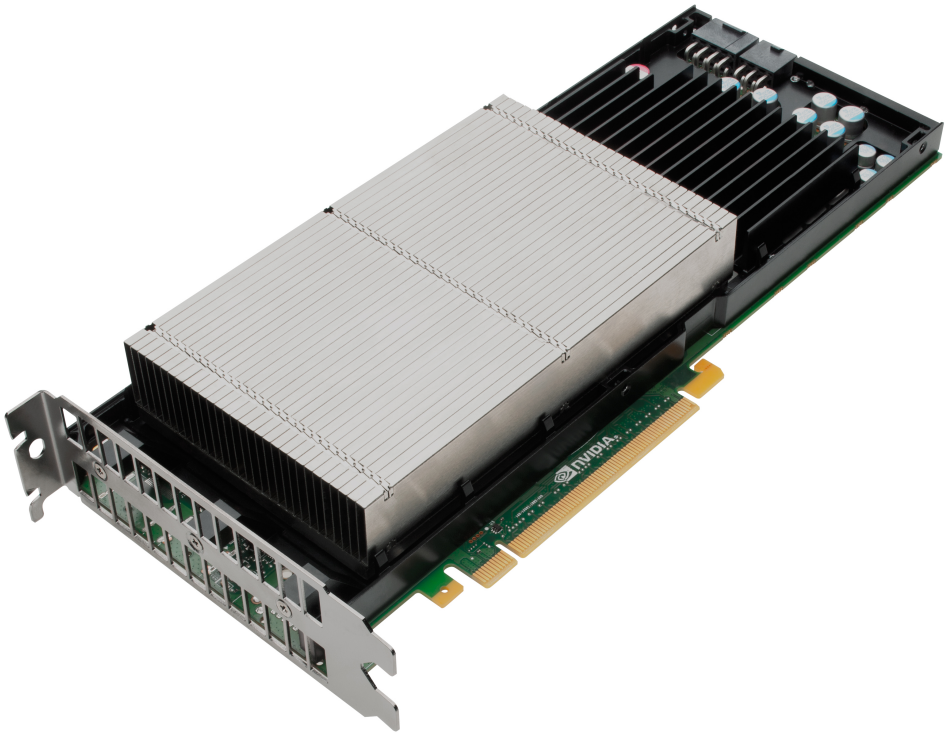
MCX311A-XCAT	MCX313A-BCBT	MCX314A-BCBT

New Mexico CONSORTIUM

PRObE Parallel Reconfigurable Observational Environment

Nvidia Donation: 34 x K20

- 1.2 TFLOPS double-precision, 2500 cudacores, 5GB @ 200 GB/s, 225W, 7B transistors



The innovative design of the Kepler compute architecture includes:

- **SMX** (streaming multiprocessor) design that delivers up to 3x more performance per watt compared to the SM in Fermi. It also delivers 1 petaflop of computing in just 10 server racks.
- **Dynamic Parallelism** capability that enables GPU threads to automatically spawn new threads. By adapting to the data without going back to the CPU, it greatly simplifies parallel programming and enables GPU acceleration of a broader set of popular algorithms, like adaptive mesh refinement (AMR), fast multipole method (FMM), and multigrid methods.
- **Hyper-Q** feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single Kepler GPU, dramatically increasing GPU utilization, slashing CPU idle times, and advancing programmability. Ideal for cluster applications that use MPI.

Current Usage Status

- More than 1200 computers available today
- Kodiak PIs: Calton Pu (GaTech), Mike Dahlin (UT Austin), Robbert van Renesse (Cornell), Mike Freedman (Princeton), Rob Ricci (Utah), Dave Andersen (CMU), Christos Christodoulou (UNM), Kevin Harms (Argonne), Michael Lang (LANL)
- Staging cluster PIs: Greg Ganger (CMU), Daniel Freedman (Technion), Jonathan Appavoo (BU), Patrick Bridges (UNM), Jun Wang (UCF)
- Publications starting to appear, e.g.:
  - Wyatt Lloyd, et. al., "Stronger Semantics for Low-Latency Geo-Replicated Storage," NSDI'13. Uses Kodiak to show scaling.
  - Garth Gibson, et. al., "PRObE: A Thousand-Node Experimental Cluster for Computer Systems Research,"USENIX login, June 2013. Advertising.

Contacts

- Website
  - http://www.nmc-probe.org/ Info & case studies
  - http://portal.nmc-probe.org/ Emulab portal
  - Email: probe@newmexicoconsortium.org
  - Groups: probe-users@googlegroups.com

Please use PRObE

Your participation is greatly encouraged!