Computational Requirements

Don Holmgren

LQCD Project Progress Review

May 25-26, 2006

Fermilab

Lattice QCD Code Requirements

- Large sustained floating point delivery
  - O(100 Gflops) – O(1 Tflops) on O(24 hour) jobs
- Large memory bandwidth
  - 1.82 Bytes/Flop single precision
  - 3.63 Bytes/Flop double precision
  - Not cache friendly
- Large network bandwidth (asqtad example)
  - MB/sec required = 0.545 x (Mflops Sustained) / L
    - L=local lattice size (L^4)
  - O(100 MB/sec) for O(Gflops)
  - Communications limit scaling for large node counts

Types of LQCD Calculations

- Vacuum Gauge Configuration Generation
  - Long stochastic evolution that can only be done in a few streams in which lattice spacing and quark mass are varied
  - Computational needs increase as lattice spacing and quark masses decrease
  - High end capacity computing, O(Tflops) required
  - Very low I/O requirements – store a new configuration of O(1 Gbyte) once an hour
  - Precious outputs – investments of 10's of Tflops-yrs

Types of LQCD Calculations

- Analysis using vacuum gauge configurations
  - Small jobs (16-128 nodes on a cluster) run using archived gauge configurations, generating propagators
  - Smaller jobs (4-8 nodes on a cluster for asqtad) run against stored propagators (correlation calculations)
  - Jobs are independent (per gauge configuration), so hundreds can be run in parallel
  - Capacity computing, O(100 Gflops) per job
  - Higher I/O requirements – consume O(1 GB) file, generate several propagators of size 2-17 GB, temporarily store propagators for later correlation calculations

Further Aspects of Analysis

- Multiple groups analyze the same configurations for different physics projects. This requires the coordination of configuration generation across the entire U.S. community, and the coordination of analysis campaigns across multiple groups.
  - For example, multiple groups may want to use the same propagators; coordination saves regeneration.
- Performance requirements for analysis vary widely. Some projects require performance similar to configuration generation, while others require much lower performance and are ideal for clusters.
- Multiple computers are valuable for analysis.

Recent Allocations

- The Scientific Program Committee allocates time according to scientific merit.
- Requests for time in the most recent allocation process (ended April 6/7, 2006, allocations to run July 1 – June 30) exceeded what is available.
- Normalized requests in millions of QCDOC and JLab “4G” node hours (note: QCDOC node = 0.35 Gflops,
  4G node = 1.2 Gflops = 3.4 QCDOC nodes)

	QCDOC	Clusters
available	87M	17.5M
requested	160M	50M

I/O and Storage Requirements (FY2006)

- Gauge configurations
  - Approximately 10 Tbytes, archived on tape and with subsets resident on disk
  - Stored at multiple sites for protection
  - Individual files are O(100MB) – O(2 GB)
  - Few are written per large job
- Propagators
  - Approximately 150 Tbytes in total, limited lifetime
  - Individual files are O(100MB) – O(17 GB)
  - Several are read/written per small job

International Resources – FY2006

Current capabilities (Estimates of non-US resources from Steve Gottlieb):

Country	**Sustained Performance**
Germany	10-15 TF
Italy	5 TF
Japan	14-18 TF
UK	4-5 TF
United States	
LQCD Project	8.6 TF
National Centers	1.0 TF
Dedicated Machines	4.8 TF
US Total	14.4 TF

Backup Slide

International Resources

Current capabilities (Estimates from Steve Gottlieb's talk):

Location	Type	Size	Peak	Est. Perf.	Total
Paris-Sud	apeNEXT	1 rack	0.8 TF	0.4 TF	0.4 TF
Bielefeld	apeNEXT	6 (3) racks	4.9 TF	2.5 TF	
DESY (Zeuthen)	apeNEXT	3 racks	2.5 TF	1.2 TF	
Julich	BlueGene/L	8 racks	45.8 TF	11.5 TF x ½?	10-15 TF
Munich	SGI Tollhouse	3328 nodes	70 TF	14 TF?? x?	
Rome	ApeNEXT	12 (8) racks	9.8 TF	4.9 TF	5 TF
KEK	BlueGene/L	10 racks	57.3 TF	14.3 TF	
Tsukuba	PACS-CS	2560 nodes	14.3 TF	3.3 TF	14-18 TF
KEK	Hitachi		2.1 TF	1 TF?	
Edinburgh	QCDOC	12 racks	9.8 TF	4.2 TF	
Edinburgh	BlueGene/L	1 racks	5.7 TF	1.4 TF x ?	4-5 TF