International Resources for Lattice QCD

Steven Gottlieb, Indiana University

sg at indiana.edu

Outline

- Predictions from Last May
- International Resources: Peak Speeds
- Performance Estimates
- International Resources: Actual Speeds
- Projects Planned

May 2005 Resource Estimate

- UKQCD: 5 TF QCDOC
- Japan:
  - Tsukuba: 2048 node cluster (early 2006) growing to 3072 or more nodes the next year. (2.8 GHz Xeon CPU; 12–17 TF peak)
  - KEK: New 20 TF (peak speed) system to replace Hitachi SR800F1, early 2006
  - Earth Simulator: Some fraction of this 40 TF machine will be used for LGT
- Germany:
  - Bielefeld: 5TF APEnext, end of this year
  - DESY: 2.4 TF APEnext, after July 2005
- Italy: 10.4 TF (peak) APEnext, July–Dec. 2005

International Resources: Peak Speeds

- Organize list by architecture
- Only list resources outside of US
- Not all systems in place yet. Details of usage, installation later.

Type	speed	quantity	peak speed
apeNEXT	819 GF/rack	24 racks	19.7 TF
BlueGene/L	5.73 TF/rack	8 racks	45.8 TF
BlueGene/L	5.73 TF/rack	10 racks	57.3 TF
BlueGene/L	5.73 TF/rack	1 rack	5.7 TF
Hitachi	134 GF/node	16	2.1 TF
PACS-CS	5.6 GF/node	2560 nodes	14.3 TF
QCDOC	819 GF/rack	12 racks	9.8 TF
SGI Tollhouse	10.6 GF/core	3328*2 cores	70.0 TF

Performance Estimates: apeNEXT

Source: F. Belletti *et al.*, Comp. Sci. & Eng., vol. 8 No. 1, 18 (2006)

- Double precision CPU
- 200 MHz $\times$ 8 Flops/cycle $=$ 1.6 GFlops
- Current machines running at 160 MHz
- Memory Bandwidth: 3.2 GBytes/s
- Network Bandwidth: 200 Mbytes/s per link (each direction); 7 links
- Wilson Dslash: 54% of peak, other operations 29–41%
- Performance Estimate: 50% of peak overall

Performance Estimates: BlueGene/L

Sources: Doi, Hashimoto, Kreig talks at
Boston U. BlueGene/L Workshop

- Double precision CPU
- 700 MHz $\times$ 4 Flops/cycle $=$ 2.6 GFlops
- Dual CPU nodes $=$ 5.6 GFlops/node
- Wilson $24^3 \times 48$ on 1/2 rack: 29% of peak (Hashimoto)
  - 22–29% of peak on 1 rack (Doi)
  - 18% using inline assembly and MPI
  - <10% without assembly
  - Dlash_ eo $16^3 \times 32$, g++, inline asm: 15% (Kreig)
  - Dlash_ eo $16^3 \times 32$, g++, intrinsincs: 12%
- Domain Wall $24^3 \times 48 \times 16$ on 1/2 rack: 22% (Hashimoto)
- Performance Estimate: 25% of peak overall

Performance Estimates: PACS-CS

Source: talk by A. Ukawa at 4th ILFT Network Workshop

- Xeon 2.8 GHz: 5.6 GFlop peak (?)
- Memory Bandwidth: 6.4 GB/s
- Network: 3D hypercrossbar; dual Gigabit in each direction
- $8^3 \times 64$ Single Node Wilson-Clover Dirac operator
  - C with SSE3 assembler: 33%
  - C with Intel intrinsic: 34%
  - Fortran: 26%
- For BiCGStabL2, estimate computation at 72% of time, rest network
- Performance Estimate: 23% of peak overall

Performance Estimates: QCDOC

Source: SciDAC documents, talk by Chulwoo Jung at Boston U. BlueGene/L Workshop

- Double precision CPU
- 400 MHz $\times$ 2 Flops/cycle $=$ 800 MFlop/node
- 6D hypertorus network
- Asqtad: 30–55% of peak for various parts of code, 35–40% for CG
- Avg of Domain Wall and Asqtad: 43% of peak

Performance Estimate.: Other

- Hitachi SR11000 K1
  - Installed at KEK
  - Have not seen any benchmarks
  - It is fairly small anyway
- SGI Tollhouse
  - To be installed in Leibniz-Rechenzentrum (LRZ) Munich
  - Not restricted to LGT
  - Itanium2 based system, to be installed 2Q06
  - Dual core Montecito CPUs, NUMAlink 4 network for shared memory
  - No benchmarks

International Resources: Actual Speeds

Final estimates consider multiple users for Julich, KEK, Munich, Edinburgh

Location	type	size	peak	est. perf.	total
Paris-Sud	apeNEXT	1 racks	0.8 TF	0.4 TF	0.4
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 $\times 1/2$?	10–15
Munich	SGI Tollhouse	3328 nodes	70 TF	14 TF?? $\times$?	
Rome	apeNEXT	12 (8) racks	9.8 TF	4.9 TF	5
KEK	BlueGene/L	10 racks	57.3 TF	14.3 TF	
Tsukuba	PACS-CS	2560 nodes	14.3 TF	3.3 TF	14–18
KEK	Hitachi		2.1 TF	1 TF ?	
Edinburgh	QCDOC	12 racks	9.8 TF	4.2 TF	4–5
Edinburgh	BlueGene/L	1 racks	5.7 TF	1.4 TF $\times$?	

Projects Planned

- KEK BlueGene/L: dynamical overlap; $16^3 \times 32$ and $24^3 \times 48$, $N_f = 2$ then 2+1 [Hashimoto]
- PACS-CS: Wilson-clover N_f =2+1; $m_{ud}/m_s = 0.2$ using domain-decomposed HMC [Ukawa]
- Julich BlueGene/L: Dynamical Overlap, Twisted Mass [Kreig, Montvay]
- DESY, Paris, Rome apeNEXT: Dynamical Twisted Mass, $N_f = 2$, 2+1+1 [Montvay]
- Edinburgh QCDOC: Dynamical domain wall, improved staggered

Concluding Remarks

From last May:

- Investments by Italy, Germany and, particularly, Japan rival what is currently proposed by DOE and could result in systems that exceed the capability of our own.

April, 2006:

- It seems clear that the concerns of last May have come to pass
- Germany and Japan, which have substantial lattice communities, but smaller than that in the US, have surpassed us
- With current hardware plans, they are likely to remain ahead for some time
- Given our ambitions and requests that considerably exceed possible allocations, it will be a challenge to realize the full physics potential

A Multiyear Program

- It will take several years to decrease the lattice spacing and to approach the chiral limit for each lattice spacing
- A capability of 5 TF is sufficient for all but the two most demanding runs below

a(fm)	m_l/m_s	Lattice	Traj.	TF-Yr
0.09	0.1	$40^3 \times 96$	3,000	0.54
0.09	0.05	$56^3 \times 96$	4,200	6.05
0.06	0.4	$48^3 \times 144$	3,000	0.45
0.06	0.2	$48^3 \times 144$	3,750	1.68
0.06	0.1	$60^3 \times 144$	4,500	7.98
0.06	0.05	$84^3 \times 144$	6,300	93.20
0.045	0.4	$56^3 \times 192$	4,000	2.25
0.045	0.2	$56^3 \times 192$	5,000	7.52
0.045	0.1	$80^3 \times 192$	6,000	54.80