Lattice QCD on a Blue Gene/L

James C. Osborn
Boston University

ILFTN 3
Jefferson Lab
Oct. 2005

IBM Blue Gene/L

- relative of QCDOC
- 1 rack =
  1024 dual core
  700MHz PPC 440d
  (5.7 Tflops peak)
- 3D torus network,
  tree (for global ops)

Known Available Installations

- Argonne National Lab – 1 rack
- BU – 1 rack
- Edinburgh – 1 rack
- MIT – 1 rack
- San Diego Supercomputer Center – 1 rack

Processing Modes

- coprocessor mode (CO)
  - 1st CPU for main computation
  - 2nd CPU used to offload work
  - can be used to assist communications
- virtual node mode (VN)
  - each CPU runs a separate MPI process
  - each CPU handles its own communications

Machine Partitioning

- flexible partitioning (1 rack):
  - 32, 128, 512 and 1024 node partitions
- 32 and 128 node partitions use MESH
- 512 and 1024 node partitions use TORUS
- all partitions can use MPI to send to any other node

Porting Software

- copy source to machine
- configure to use cross compiler
- compile
- run
- optimize (if necessary or desired)

SU(3) matrix-vector product (single)

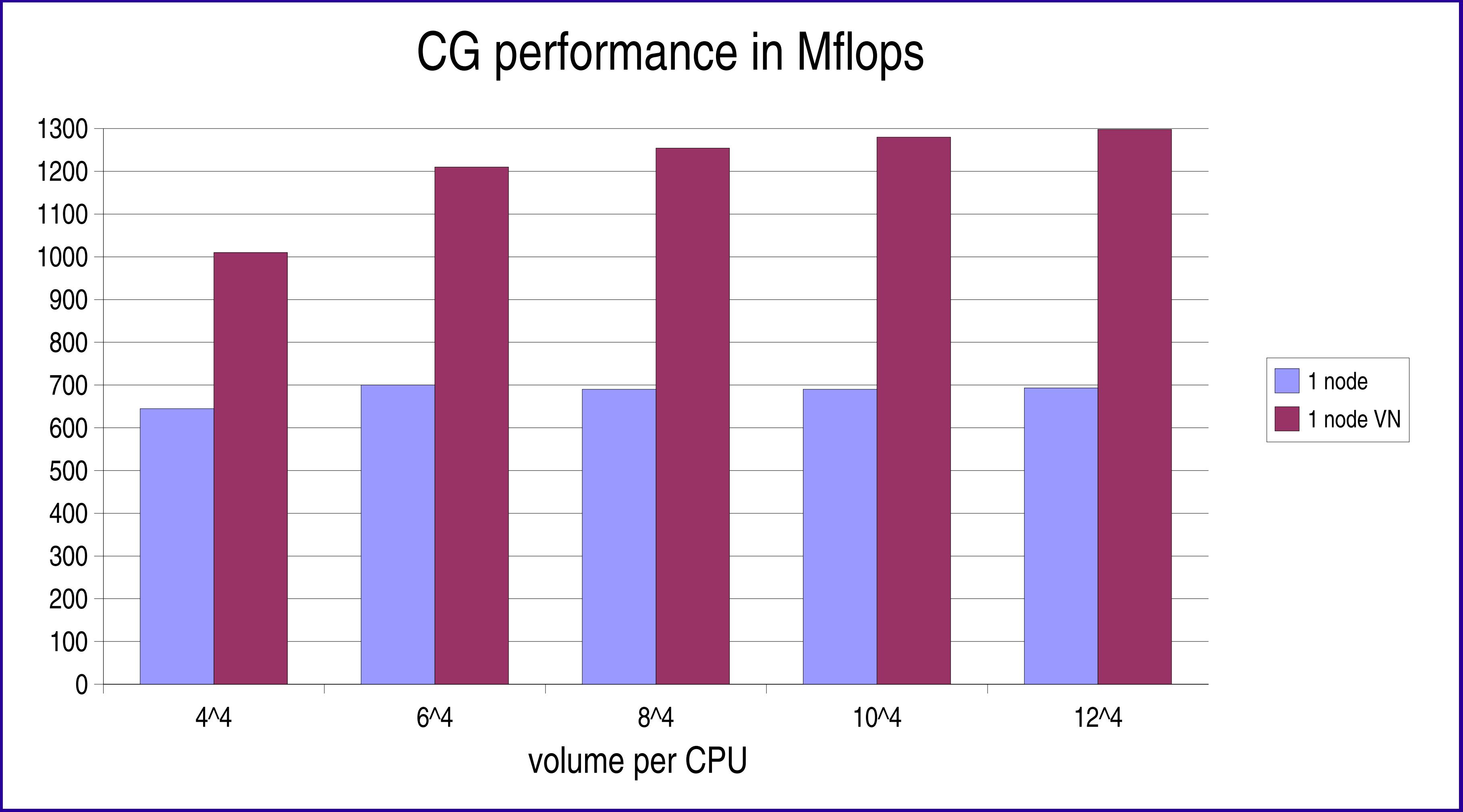
- for(i=0; i<N; i++) v2[i] += m[i] * v1[i];
- speed in Mflops for code compiled with xlc

array size:		10^2	10^3	10^4
naive C	-O2	135	122	122
	-O5	224	183	183
reordered C	-O2	1105	699	701
	-O5	494	494	493
440 asm		987	809	812

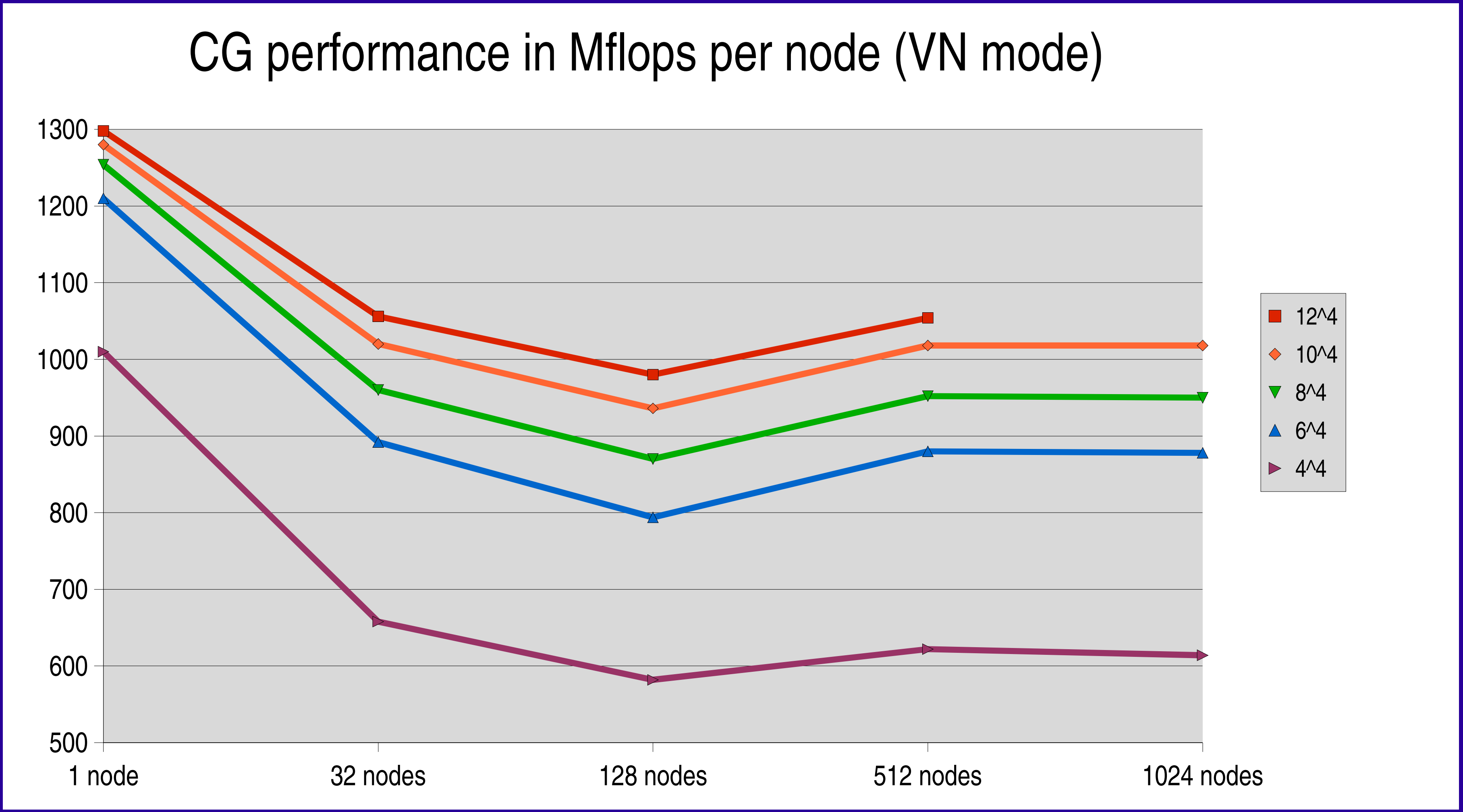
LQCD example

- MILC improved staggered operator (a^2-tad)
- CG inverter written in SciDAC QDP/C (single precision)
- uses pure C QLA with one 440 asm routine
- all code compiled with xlc

staggered a^2-tad CG inverter



staggered a^2-tad CG inverter



Grid Layout

- generally want data layout to match torus
- 512 VN nodes = 8 x 8 x 8 x 2

lattice size	node layout	Mflops/node
64x64x64x16	8x8x8x2	952
32x32x32x128	4x4x4x16	808
128x32x32x32	8x8x8x2	920

Chroma on BG/L

- large LQCD application using SciDAC QDP++
- must use g++ (xlC gives internal compiler error)
- existing optimizations for QCDOC
  - Peter Boyle's Wilson dslash
    - single (unsupported) and double precision
  - BAGEL generated linear algebra
    - currently only double precision

Chroma Domain Wall inverter

- 8^4 local volume per CPU
- results in Mflops per node

number of nodes:	1	32	32VN	512VN
single precision				
plain	49	47		
with PAB dslash	340	300	486	486
double precision				
with PAB dslash and BAGEL	395	315	350	

Caveat

- both xlc and gcc can produce wrong code at high optimization levels (-O5)
- NaN's and Inf's appear in results
  (but not always seen)
- all code seems to work at -O2
- need to verify program before production

Summary

- BG/L machines are already available to the LQCD community and likely more will be.
- Relatively simple to port and run applications.
- Well written C/C++ code can achieve good performance (1+ Tflops) with little extra effort.
- All benchmark results are just a first pass at running existing software and performance should improve as more optimizations are done.