Experiences with Lattice QCD on the Blue Genes at Juelich

10. October 2013 | Stefan Krieg
Motivation: Physics results from LQCD on BG

Budapest-Marseille-Wuppertal Coll., Science 322, 1224

Science: Top-10 Scientific Breakthrough 2008
Motivation: Physics results from LQCD on BG

![Diagram showing data points for \( \Delta M_N \), \( \Delta M_\Sigma \), and \( \Delta M_\Xi \) with error bars.](image)

arXiv:1306.2287
The new machine

Specs:
- 28 Racks
- 7x2x2Racks
  = 28x8x8x8x2 nodes
  = 28,672 nodes
  = 458,752 cores
  = 1,835,008 HW tds.
- 5.9 Pflop
- Top500 #7
- Public: PRACE, GCS, NIC
- local use (JARA)
- Production: Jan. 2013
  (8->16->24->28)
Mitglied der Helmholtz-Gemeinschaft

S. Krieg, 32e Forum ORAP
10. Oktober 2013

BGQ Chip architecture

- 16+1 core SMP
  Each core 4 way hardware threaded
- Transactional memory and thread level speculation
- Quad float point unit on each core
  204.8 GF peak node
- Frequency target of (1.6) GHz
- 563 GB/s bisection bandwidth to shared L2
  (BGL at LLNL has 700 GB/s system bisection)
- 32 MB shared L2 cache
- 42.6 GB/s DDR3 bandwidth
  (2 channels each with chip kill protection)
- 10 intrarack interprocessor links each at 2.0GB/s
- 1 I/O link at 2.0 GB/s
- 4-8 GB memory/node
- ~30 Watts chip power

2 GB/s I/O link (to I/O subsystem)

10*2GB/s Intrarack (5-D torus)

** chip I/O shares function with PCI_Express
# Blue Gene/Q vs. Blue Gene/P

<table>
<thead>
<tr>
<th>Blue Gene/Q</th>
<th>Blue Gene/P</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Midplane = 105 Tflop/s</td>
<td>• 16 MP = 111 Tflop/s</td>
</tr>
<tr>
<td>• 32768 HW threads</td>
<td>• 32768 HW threads</td>
</tr>
<tr>
<td>• = 3.2 Gflop/s per thread</td>
<td>• = 3.4 Gflop/s per thread</td>
</tr>
<tr>
<td>• 5d interconnect:</td>
<td>• 3d interconnect:</td>
</tr>
<tr>
<td>• 512 = 2x4x4x4x4</td>
<td>• 8192 = 16x16x32</td>
</tr>
</tbody>
</table>

→ USE master threads if possible!
## Comparison of relevant hardware specs

<table>
<thead>
<tr>
<th></th>
<th>Blue Gene/P</th>
<th>Blue Gene/Q</th>
<th>Q vs. P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>300.000 threads</strong></td>
<td></td>
<td><strong>6.291.456 threads</strong></td>
<td>21</td>
</tr>
<tr>
<td><strong>1 Pflop/s</strong></td>
<td></td>
<td><strong>20 Pflop/s</strong></td>
<td>20</td>
</tr>
<tr>
<td>Cache: <strong>2.0 MB/core</strong></td>
<td></td>
<td>Cache: <strong>2.0 MB/core</strong></td>
<td>1</td>
</tr>
<tr>
<td>Cache: <strong>109 GB/s BW</strong></td>
<td></td>
<td>Cache: <strong>563 GB/s BW</strong></td>
<td>6</td>
</tr>
<tr>
<td>Mem: <strong>0.5 GB/thread</strong></td>
<td></td>
<td>Mem: <strong>0.5 GB/thread</strong></td>
<td>1</td>
</tr>
<tr>
<td>Mem: <strong>13.6 GB/s BW</strong></td>
<td></td>
<td>Mem: <strong>42.6 GB/s BW</strong></td>
<td>3</td>
</tr>
<tr>
<td>FPU: <strong>4 Flop/c/core</strong></td>
<td></td>
<td>FPU: <strong>8 Flop/c/core</strong></td>
<td>2</td>
</tr>
<tr>
<td>FPU: <strong>13.6 Gflop/s/node</strong></td>
<td></td>
<td>FPU: <strong>204.8 Gflop/s/node</strong></td>
<td>15</td>
</tr>
</tbody>
</table>
16+1 core SMP
Each core 4 way hardware threaded
Transactional memory and thread level speculation
Quad float point unit on each core
204.8 GF peak node
Frequency target of (1.6 ) GHz
563 GB/s bisection bandwidth to shared L2
(BGL at LLNL has 700 GB/s system bisection)
32 MB shared L2 cache
42.6 GB/s DDR3 bandwidth
(2 channels each with chip kill protection)
10 intrarack interprocessor links each at 2.0GB/s
1 I/O link at 2.0 GB/s
4-8 GB memory/node
~30 Watts chip power

2 GB/s I/O link (to I/O subsystem)
10*2GB/s Intrarack (5-D torus)
** chip I/O shares function with PCI_Express
BGQ Chip architecture

- 16+1 core SMP
  Each core 4 way hardware threaded
- Transactional memory and thread level speculation
- Quad float point unit on each core
  204.8 GF peak node
- Frequency target of (1.6) GHz
- 563 GB/s bisection bandwidth to shared L2
  (BGL at LLNL has 700 GB/s system bisection)
- 32 MB shared L2 cache
- 42.6 GB/s DDR3 bandwidth
  (2 channels each with chip kill protection)
- 10 intrarack interprocessor links each at 2.0 GB/s
- 1 I/O link at 2.0 GB/s
- 4-8 GB memory/node
- ~30 Watts chip power

2 GB/s I/O link (to I/O subsystem)

10*2GB/s Intrarack (5-D torus)

** chip I/O shares function with PCI_Express
Messaging unit (mu)

- Direct access to L2
- Direct access to network HW
- Fully user programmable
- Performs PtP and collective communications
- Runs independent of cores:
  - Sends data
  - Receives data and stores to memory subsystem
- Shared resource for all 17 cores
Comparison of communication specs

<table>
<thead>
<tr>
<th>Blue Gene/P</th>
<th>Blue Gene/Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Persistent communication:</strong></td>
<td><strong>Persistent communication</strong></td>
</tr>
<tr>
<td>- Inject descriptors once</td>
<td>- Same setup as for BGP</td>
</tr>
<tr>
<td>- Reuse as often as possible</td>
<td>- However: no address associated with reception counter</td>
</tr>
<tr>
<td>- Communication startup slow (as slow as non-persistent generic communication)</td>
<td>- HW:</td>
</tr>
<tr>
<td>- Communication restart fast: manipulate (memory) Fifo head &amp; rec. and inj. counters</td>
<td>- 512 Fifos (= 32 fifos/core)</td>
</tr>
<tr>
<td>- HW:</td>
<td>- 512 BAT (base address table) entries (= 32 entries/core)</td>
</tr>
<tr>
<td>- 4 Fifo groups, 32 Fifos/group</td>
<td>- Number of counters “unlimited” → BAT entry required</td>
</tr>
<tr>
<td>- 4 Cntr. groups, 64 couters/group</td>
<td>- BAT entry required also for direct put operation.</td>
</tr>
</tbody>
</table>
LQCD code

- Vanilla (C/MPI), Cuda, BG/Q (extensive use of macros)
- Threaded (pthreads, master/workers)
- Parallelization strategy for BG/Q:
  - 16/4 (AxBxCxDxEx16)(x4)
  - Use shmem window to communicate between processes
  - Synchronize threads/processes with A2 barrier (shmem)
  - Node layout: wrap 2 dimensions into a 4d torus
- Implementation strategy:
  - Use permutes
  - Always stay in 2\textsuperscript{nd} level cache
Going from P to Q

- In many cases there is a 1 to 1 mapping from BG/P to BG/Q
- Obvious examples:
  - vec_madd = __fpmadd
  - vec_sub = __fpsub
- Less obvious ones:
  - vec_xmadd = __fxcpmadd
  - vec_xxmadd = __fxcxma
  - vec_xxcpnmmadd = __fxcxnsma
  - vec_xxnpmadd = __fxcxnpma
- Loads can isolate lower half of register → emulate BG/P (latency!)
  - vec_ld2, vec_st2 = __lfpd, __stfspd
Treading

- BG/P code was not threaded $\rightarrow$ 1 core = 1 process
- BG/Q, by default we use 1+3 threads/core
- Started with openMP, but found large latencies
- Moved to pthreads using a master, worker setup
  $\rightarrow$ pthreads/worker setup requires rewriting of loops
- pthread barrier too slow, use A2 atomics
- Avoid wasting cycles:
  - Spinning (waiting) workers take cycles from master running serial code
  $\rightarrow$ Use Wakeup-Unit: threads sleep and wake up on demand
Node-wide barriers – BG/Q L2 atomics

```c
__INLINE__ void L2_Barrier(L2_Barrier_t *b, int numthreads)
{
    uint64_t target = b->start + numthreads;
    uint64_t current = L2_AtomicLoadIncrement(&b->count) + 1;

    if (current == target) {
        b->start = current;  // advance to next round
    } else {
        while (b->start < current);  // wait for advance to next round
    }
}
```
Low-level communication, BGP/BGQ

- SPI based
- Code interface requires flexible communication using tags/communicators
- Standard comms are made persistent and are freed in times of need
- Restarts check if the tag/communicator has been force freed, otherwise performs a simple restart
- **Q:** SMP (1/64) or multiple process modes (eg. 16/4)
- Global comms proceed via SMP window (shm_open/mmap)
- **Q:** PtP comms contain no global ops (i.e. no global sycs)
Performance

Number of BG/P racks

- $64^3 \times 144$ Clover
- 37.5% peak
- 1.3 GF/c/s

Number of BG/Q racks

- $128^3 \times 144$ Clover
- 37.5% peak
- 4.8 GF/c/s
Performance

Number of BG/P racks

Performance / TFlop/s

1  24  48  72

4K  98K  196K  295K

64³x144 Clover
37.5% peak
1.3 GF/c/s

Number of BG/Q racks

Performance / TFlop/s

6  12  24

393K  786K  1573K

128³x144 Clover
37.5% peak
4.8 GF/c/s
Performance

Number of BG/P racks

Number of BG/Q racks

Performance/Peak

Kernel memory footprint/MB

# Parallel tasks

perf single
perf double
mem single
mem double

10. Oktober 2013 S. Krieg, 32e Forum ORAP
Solver performance (production)

- Multishift CG: 3.2 Gflop/core/s (3x6x6x6 loc. lat.) (CG more efficient, BiCGstab similar)
- Multilevel method (A. Frommer, K. Kahl, S.K., B.Leder, and M.Rottmann, 1303.1377, 1307.6101)
  - Smoother: 3.4 Gflop/core/s (12x6x6x3 loc. lat., 48³x96)
  - Restriction/interpolation: 3.6/6.6 Gflop/s
  - Coarse grid operator: 1.2/2.2 Gflop/s (2x2x1x1 loc. lat.)
  - Total 1.9 Gflop/s
  - Setup: 2.4 Gflop/s
Conclusions

• Efficiency on BG/Q appears to much more peaked when compared to BG/P
• Peak value remains unchanged
• Solver performance noticeably lower than for dslash
• Continuing tendency towards more complex solvers (e.g. MG)
  → Tuning becomes more difficult
  → So far, performance results of complex solvers do not match those of ordinary ones
  → Full vectorization of code potentially necessary
One more thing...

Number of BG/Q racks

Performance/TFlop/s

128³x168
Clover

37.5% peak
4.8 GF/c/s

# Parallel tasks

459K
918K
1835K
One more thing…

Number of BG/Q racks

<table>
<thead>
<tr>
<th>Numbers</th>
<th>Performance/TFlop/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>128³x168 Clover</td>
</tr>
<tr>
<td>14</td>
<td>37.5% peak 4.8 GF/c/s</td>
</tr>
<tr>
<td>28</td>
<td>perfect weak scaling…</td>
</tr>
</tbody>
</table>

# Parallel tasks

- 459K
- 918K
- 1835K