

SciDAC Software Overview

Robert Mawhinney

SciDAC Software Committee Meeting

FNAL

November 7-8, 2008

1. Brief description of CPS functionality
2. Describe recent additions
3. Some BG/Q generalities

SciDAC Personnel

- Chulwoo Jung
- Enno Scholz (through 9/31/08)
- Oliver Witzel (starting 10/1/08)
- Stratos (Efstratios Efstathiadis)

Others

- Bob Mawhinney (Columbia)
- Peter Boyle (Edinburgh)
- Petreczky, Schmidt, Soelder (BNL thermo)
- Blum, Dawson, Izubuchi, Y. Aoki, Li, (RBC)

CPS

- LQCD code incorporating many current algorithmic features and important physics measurements
 - * Supported fermion actions: DWF, ASQTAD, P4, Clover, Wilson, relativistic heavy quarks,
 - * Evolution algorithms: RHMC, R, phi
 - * Many observables from spectroscopy to matrix elements
- Heavily used for ensemble generation and measurements
 - * $T = 0$ DWF ensembles ($32^3 \times 64 \times 16$) (ANL and QCDOC)
 - * $T = 0$ and $T \neq 0$ P4 staggered ensembles (QCDOC, NYBlue)
 - * $T = 0$ and $T \neq 0$ P4 staggered ensembles (LLNL)
 - * $N = 1$ SUSY ensembles (NYBlue)

Recent CPS Additions I

- Addition of QIO functionality (primarily Scholz)
 - * Extensively used for storage of DWF quark propagators
 - * Not currently used for gauge field storage, still NERSC format
 - * Need concurrency control in QIO to avoid overloading file system
 - * Managing/storing partfiles tedious (workflow tools?)
 - * Partfiles useful for parallel I/O bandwidth, but conversion to single-file generally needed since generally reloaded on different machine
- I/O flexibility likely becoming more important on larger machines to make sure this is not weak link
 - * Need maximum flexibility in choosing writing nodes and file types
- Considerable effort expended to check files between RBC and LHPC

Recent CPS Additions II

- **Multidimensional parallel transport (Jung)**
 - * Needed to make use of high-dimensional networks
 - * Functionality used in gauge actions, fat links, ...
 - * Low software latency/overhead for QMP important
 - * Will performance suffer with many tens of threads on SMP node?
- **Optimization for BG/L and BG/P (primarily Jung)**
 - * Using QCDOC assembly kernels for Dirac operator
 - * Improved kernels for BG/P developed by Boyle - to be included
 - * Mapping problems to partitions challenging, particularly thermo
 - * Communications via QMP and SPI

Recent CPS Additions III

- Job replication (Jung)
 - * Fixed, large machine partitions may not match physics jobs
 - * Introduce a fictitious sixth dimension into CPS software
 - * Use 6d coordinate to run different jobs within partition
 - * Parameter files, I/O decoupled
 - * In production at LLNL on BG/L
- Improving QIO file flattening (Stratos)
 - * QCDOC parallel disk arrays - direct ethernet to flattening server
 - * Flattening server is 6 TByte RAID disk with 16 GBytes of memory
 - * Server flattens propagators at production rate of 2, 4k QCDOC's

CPS and Hardware Debugging

- CPS developed for QCDSF and used to debug hardware
 - * Test reproducibility of calculation without slowdown
 - * In CG, keep checksums of vectors every iteration
- Extended these feature in port of CPS to QCDOC
 - * Reporting of failing node improved
 - * Communication checksums added
- Proving vital for work on BG/L and BG/P
 - * Chulwoo found bad chips in two partitions at ANL after acceptance
 - * Caused considerable consternation at IBM
 - * Chulwoo worked with IBM engineers to located bad chips
- Important feature to improve. SciDAC software should not degrade this functionality, and hopefully improve it.

BG/Q

- Columbia/Edinburgh working with IBM on this machine
- 200 TFlops/rack
- All machines of this scale will require massive threading
- Cores will run hardware threads
- Dirac operator will be aggressively threaded assembly
- Challenge: how to get decent efficiency out of non D-slash parts?
 - * Default - one small local volume per thread using QMP over MPI
 - * Will MPI be well implemented so latency is low?
 - * SMP on every node - need threaded code and low thread overhead
 - * Structure of memory systems, internode latency and communications bandwidth important for optimal choice
 - * We know this for BG/Q, via NDA
- Can we do something more generic?