Performance Optimization of a Petascale-enabled Finite Volume Solver

P. Hadjidoukas

Chair for Computational Science
ETH Zurich

with Diego Rossinelli, Babak Hejazialhosseini
Fabian Wermelinger, Jonas Sukys
Petros Koumoutsakos

Exascale Applications and Software Conference, EASC 2015
Flow Simulations complement theory and experiments

- Here: Cloud Cavitation Collapse

**CHALLENGE**

The gap

between hardware capabilities and achievable performance by flow solvers
HPC and CFD: The Gap

Chombo, Flash, Raptor, Uintah,.... < 7 % of the available performance

Achievements:

- 55% of the peak performance - 13 Trillion cells, Unprecedented time to solution
- Simulations on 1.6M cores of Sequoia IBM BG/Q supercomputer
- ACM Gordon Bell Prize 2013 (for peak performance)
THIS TALK

• A finite volume, two phase flow solver at 14.4 PFLOP/s

• How did we get 11 PFLOP/s (55% of peak)?

• How did we improve the performance to 14.4 PFlops (72% of peak)

Performance update for the 2013 Gordon Bell finalist:

14.4 PFLOP/s Simulations of Cloud Cavitation Collapse

D. Rossinelli, B. Hejazialhosseini, P. Hadjidoukas, C. Bekas, A. Curioni, A. Bertsch, S. Futral, S. Schmidt, N. Adams, P. Koumoutsakos

October 8, 2013

We report a 31% improvement in the sustained performance of our simulations. We now measure the highest sustained peak performance at 14.43 (previously 10.99) PFLOP/s. The present performance corresponds to 72% (previously 55%) of the nominal peak of Sequoia, the IBM BG/Q system at the Lawrence Livermore National Laboratory.

Methodology
All runs were performed using the 96-rack Sequoia BG/Q system at Lawrence Livermore National Laboratory, and use the IBM HPC Toolkit for BG/Q to measure performance figures, as in our original submission. We also employ the identical weak scaling problem sizes as in the previous simulations.
CAVITATION and DESTRUCTION

AROUND for Performance

HARNESS for Drug Delivery

- Detrimental to the lifetime of high pressure injection engines and ship propellers
- Instrumental to kidney lithotripsy and ultrasonic drug delivery
STATE OF THE ART (2013)

- EXPERIMENTS:
  - Formulation of cloud interaction parameter, cloud radius versus collapse time (Brennen and co-workers.)
  - Averaged quantities, damage assessments (Lohse, Keller, Bose and others)

- THEORY/ MODELING
  - 1D - Rayleigh-Plesset equation (1949)
  - Single bubble, ODE, perfectly-spherical collapse, incompressible flow, singular behavior

- SIMULATIONS
  - Single Bubble (Colonius, Caltech), Multiple bubbles with models
  - 3D shock Bubble (ETHZ, SC’12)

- STATE OF THE ART: 120 bubbles, under-resolving and coarse-graining (Adams, TUM)
FINITE VOLUME SIMULATIONS
Roofline and the 7 Dwarfs

ALGORITHMS & DATA STRUCTURES
- Operational Intensity (FLOP/Byte ratio)
- FLOP/Instruction density

Compressible Flow Solvers

Arithmetic Intensity

- Sparse matrix (SpMV)
- BLASS1,2
- Structured grids (Stencils, PDEs)
- Spectral methods (FFTs)
- Dense matrix (BLAS3)
- N-body (Particle Methods)
SETTING THE STATE OF THE ART

**PFLOPS (% Peak)**

14.4 PFLOPS (72%)
0.1 - 3% (TUM)
1.3 - 6.4% (2 racks - WENO) (Stanford)

**TIME TO SOLUTION** (no I/O)

\[ T_w = \Delta wt \cdot \frac{N_c}{N_p} \] (Stanford paper)

\[ T_w = 1.8 \]

\[ T_w = 29.7 \text{ (TUM)} \]

\[ T_w = 16.3 - 39.0 \text{ (Stanford)} \]

**SIZE (Comp. Elements)**

1.3 E13 - 15K bubbles
1.2 E08 - 0.15K bubbles
0.4 E13 - Turbulence

**I/O Compression**

10-100X

(*) Stanford: SC’13, Bermejo-Moreno et al, Compressible Homogeneous Turbulence
THE SOFTWARE: CUBISM-MPCF (C++)

**TASKS**

- Minimize the memory traffic
- Maximize FLOP/Byte and FLOP/instructions
- Maximize IL, DL, TL and Cluster Level Parallelism
- Exploit BG/Q features
- Efficient wavelet-based compression
HIGH THROUGHPUT SIMULATIONS OF COMPRESSIBLE TWO-PHASE FLOWS

Babak Hejazialhosseini - Diego Rossinelli - Christian Conti - Petros Koumoutsakos
- SC’12

CODE: CUBISM@CSE LAB
C++, SSE/AVX intrinsics
Parallel patterns
Roofline performance analysis

NUMERICS
Finite Volume Method:
WENO5/HLL
Low storage RK3

SIMULATIONS
Shock-bubble interaction at Mach 3
47K cores - 30% of peak
250B elements
Core/Node Performance: The Roofline of BG/Q

Operational Intensity: FLOP count over off-chip memory transfer

BG/Q node ridge point: (7.3 FLOP/Byte, 204.8 GFLOP/s)

Perf = min(PB × OI, PP)

Kernels
- RHS
- DT
- UPDATE
- upper bound

GFLOP/s

1000

100

10

1

1/10

1

10

100

Operational Intensity (FLOP/Byte)
**CORE Layer:**  \( O_{I(RHS)} \): from 1.4 to 21

1: Block-based memory layout
- Increases *spatial locality*

2: IL and DL Parallelism
- 1 thread exclusively processes 1 block
- SoA-> explicit vectorization of *all* kernels
- exploit *common subexpressions in the RHS* *(SC’13)*
**CORE Layer (cont.)**

3: Increase **temporal locality**
- Buffers for active data-slices (e.g. in WENO, HLLE)
- **Fusion** of the RHS substages (**SC’13**)

4: exploit **BG/Q features**
- **QPX instructions** (expose as many FMAs as possible)
  - `vec_madd(a, b, c) = a*b+c`
OpenMP parallelization - 64 threads
Depth-first thread placement
Reduced load imbalance by:
  - Dynamic loop scheduling
  - Work per block amortizes OpenMP overheads
Non-blocking P2P communication for halo blocks
- 6 messages to neighbor ranks, size: 3-30MB
- Communication Time ~ Time for processing 1 block
## How Did We Reach 14.4 PFLOP/s?

From 55% (submission) to 72% (update)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Techniques</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>More efficient C/T overlap</td>
<td>+7%</td>
</tr>
<tr>
<td></td>
<td>Faster packing/unpacking of data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Better load balancing</td>
<td></td>
</tr>
<tr>
<td>Core/Node</td>
<td>Tuned data prefetching</td>
<td>+7%</td>
</tr>
<tr>
<td></td>
<td>Faster ghost reconstruction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Optimized code for B.C.</td>
<td></td>
</tr>
<tr>
<td>Core</td>
<td>Improved level of accuracy (better reciprocals)</td>
<td>+3%</td>
</tr>
</tbody>
</table>
**Issue**: time of `MPI_Waitall()` was not negligible!
- Observed when the number of compute nodes increases

**Solution**: asynchronous progress communication at the PAMID layer of the BGQ MPI implementation.
- One dedicated hardware thread assigned to MPI asynchronous progress

**How it works:**
1. The main thread issues the necessary `MPI_Irecv/Isend` calls
2. **OpenMP parallel region** with 63 threads: 1 thread - 1 inner block
3. After this parallel region, the main thread calls `MPI_Waitall()`.
4. **OpenMP parallel region** with 64 threads: rest of the inner blocks + halo blocks processed using a dynamically scheduled **OpenMP for loop**.
IMPROVED MEMORY MANAGEMENT

- Linear stream prefetching of data with depth = 1
- Deactivation of loop unrolling around WENO kernel
  - avoid register spilling
- Faster packing/unpacking of ghost data
  - optimized built-in __bcopy() function
Initial C/T overlap scheme:
- 1st stage: 2744 inner blocks to 64 threads
- 2nd stage: 1352 halo blocks to 64 threads

Updated C/T overlap scheme:
- 1st stage: 63 blocks to 63 threads
- 2nd stage: 4033 blocks to 64 threads

Boundary conditions: faster with __bcopy()
MORE IMPROVEMENTS

- Increased numerical accuracy
  - Reciprocal with Newton-Raphson scheme and two passes
    - Initial submission: single pass
    - vec_swdiv (QPX): uses two passes

- Additional minor fine tuning options:
  - Compilation of the core layer with -O3 instead of -O5
  - Decrease of stack size of OpenMP threads from 1MB to 512KB
  - Compilation with the non-debug version of the IBM XLC compiler
## From NODE to SEQUOIA

### % of peak performance

(4K blocks per node - 8Gb per node - 512^3 per node)

<table>
<thead>
<tr>
<th>KERNEL</th>
<th>Node</th>
<th>Sequoia</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHS</td>
<td>72.3%</td>
<td>71.8%</td>
<td>efficient C/T overlap</td>
</tr>
<tr>
<td>DT</td>
<td>19.9%</td>
<td>13.2%</td>
<td>global reduction (MPI_Allreduce)</td>
</tr>
<tr>
<td>UPDATE</td>
<td>2.3%</td>
<td>2.3%</td>
<td>local operations</td>
</tr>
<tr>
<td>ALL</td>
<td>65.5%</td>
<td>64.8%</td>
<td></td>
</tr>
</tbody>
</table>

- **RHS:** 14.4 PFLOP/s, 72% of peak
- **OVERALL:** 12.1 PFLOP/s, 65% of peak
### PERFORMANCE ON SEQUOIA

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>RHS</th>
<th>TtS (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>% of peak performance</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIAL</td>
<td>50.4%</td>
<td>54.6%</td>
<td>18.3</td>
</tr>
<tr>
<td>UPDATED (1-PASS)</td>
<td>61.1%</td>
<td>68.5%</td>
<td>15.2</td>
</tr>
<tr>
<td>UPDATED (2-PASS)</td>
<td>64.8%</td>
<td>71.8%</td>
<td>17.0</td>
</tr>
<tr>
<td><strong>PFLOP/s</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIAL</td>
<td>10.14</td>
<td>10.99</td>
<td>18.3</td>
</tr>
<tr>
<td>UPDATED (1-PASS)</td>
<td>+1.16</td>
<td>+2.80</td>
<td>-3.1</td>
</tr>
<tr>
<td>UPDATED (2-PASS)</td>
<td>+2.96</td>
<td>+3.44</td>
<td>-1.3</td>
</tr>
</tbody>
</table>
SOME INTERESTING FACTS

- Source code: github.com/cselab/CUBISM-MPCF

- QPXEMU module: QPX to SSE translation
  - Limited access to BG/Q (2 days/week @ IBM Zurich)
  - Not direct access to JUQUEEN and SEQUOIA

- Not access to a BGQ platform for >1 year after SC13

- Some issues for the production runs afterwards
  - MPI Collective I/O on large number of nodes
  - Processor Overheating
On Tue, Mar 10, 2015 at 11:33 AM, SC Support Team <sc@fz-juelich.de> wrote:

Dear JUQUEEN user,
yes indeed, all our overtemperature events in March came from your application:

03.03.15 14:34:06 R02-M1-N06 F I 2322965 HWERR01 0004014D ::
This board was powered off due to overtemperature. : NodeTm2Reg=0xC0000000

05.03.15 15:32:29 R33-M1-N01 F I 2323670 HWERR01 0004014D ::
This board was powered off due to overtemperature. : NodeTm2Reg=0xC0000000

10.03.15 08:42:44 R02-M1-N06 F I 2323878 HWERR01 0004014D ::
This board was powered off due to overtemperature. : NodeTm2Reg=0xC0000000

2015-03-03 12:08:48 2015-03-03 14:35:03 146 pra0913 juqueen1c1.223921.0 2052448 LL15030312064874 R02-M1 8192 2954 - abnormal termination b

2015-03-05 15:04:53 2015-03-05 15:33:26 28 pra0913 juqueen1c1.225066.0 2056255 LL15030515030489 R33-M1 8192 3374 - END_JOB control action

2015-03-10 08:12:22 2015-03-10 08:43:42 31 pra0913 juqueen1c1.226134.0 2062798 LL15031008095035 R02-M1 8192 2878 - abnormal termination b

Nevertheless this is a hardware problem, where your program seems to put some stress on the node(board)s. We have identified the nodes in question and worked on them, including screwing down the cooling units, etc. and we are monitoring the temperatures more closely now.

So please continue to resubmit the application and hopefully it will not run into that problem again. Sorry for the inconveniences.
50K bubbles, $\beta = 119$, $t = 0 \ldots 2.5$
FERMI (CINECA), 2 racks, 24h++
10K++ time steps
OUTLOOK

- Lossy and Lossless compression of 3D simulation data
- Performance optimization of CUBISM-MPCF on NVIDIA GPUs
- Uncertainty Quantification Studies
THANKS TO:
THANK YOU!