Analysis of an Application on Red Storm

Courtenay T. Vaughan
Sue P. Goudy
Sandia National Laboratories
May 2005

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Red Storm

- Cray XT3
- 10368 processors connected in a 27 x 16 x 24 mesh
- Torus in z direction
- 2.0 GHz AMD Opteron processors

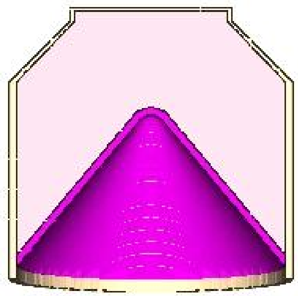
CTH

- Explicit, three-dimensional, multimaterial shock hydrodynamics code
- Uses several equations of state and material models
- Finite difference formulation on three-dimensional Cartesian mesh
- Has Automatic Mesh Refinement (AMR) capability
  - Not used for this study
  - Using flat mesh mode where each processor has an equal and consistent number of cells

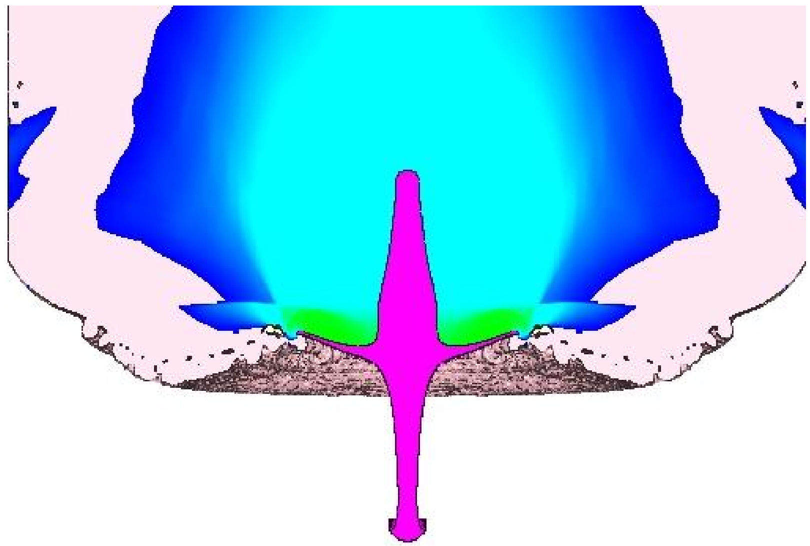
Shaped-Charge Problem

- Simulates the formation of a jet from a shaped-charge device
- Scaled problem with 90 x 216 x 90 cells per processor
  - Uses about 1 GB memory per processor
- Four materials including high explosive

Shaped-Charge Problem



0.0 ms



0.3 ms

Results

Number of processors	Time per Time Step	% Efficiency
1	11.83	100.0
2	14.23	83.1
4	14.86	79.6
8	17.17	68.9
16	17.49	67.6
32	18.70	63.2
64	18.86	62.7
128	19.73	59.9
256	19.86	59.6
512	21.95	53.9
1024	22.01	53.7
2048	22.16	53.4
4096	22.10	53.5
8192	24.69	47.9
10360	22.26	53.1

CTH

- Time stepping code
- Problem space is a rectilinear grid of cells
  - Update of variables in a cell may require values from the 26 neighboring cells
- Variables stored in three-dimensional arrays
  - Updated a k-plane at a time
  - May require operating on three k-planes at a time
- Values based on global operations over all of the cells are needed at times in each time step
  - Example: duration of next time step

Parallelization of CTH

- Processors arranged in a grid
- Each processor has a rectilinear grid of cells surrounded by a layer of ghost cells
  - Shares a face with neighboring processor
  - Data in ghost cells is updated by an exchange from real cells across the face several times a time step
  - Point to point communication
  - In each direction could communicate with 0, 1, or 2 neighbors
- Collective operations for global quantities

Basis of Model

- Computational complexity of $O(N^3)$ where N is the length of one edge of a processor's subdomain
- Communication complexity for the data exchanges is $O(N^2)$
- Communication complexity of collective operations is $O(\mathit{log}(P))$ where P is the number of processors

A Model of CTH

$$T = E(\kappa,\varphi)N^3 + C(\lambda + \tau k N^2) + S(\gamma\ log(P))$$

- T is the time per time step
- N is size of an edge of a processor's subdomain
- C and S are number of exchanges and collectives
- P is the number of processors
- *k* is the number of variables in an exchange
- λ and τ are latency and transfer cost
- γ is the cost of one stage of collective
- $E(\kappa,\varphi)$ is the calculation time per cell

Sandia National Laboratories

Parameters for model

- Obtained from Pallas benchmark
- Used PingPing benchmark for exchanges
  - λ = 8.3 μs
  - τ = 0.00102 μs/byte or 0.00816 μs/double precision
- Use AllReduce benchmark for collectives
  - γ = 10.5 μs/double precision

Application of Model

- Parameters for model depend on the problem
- For shaped-charge problem:
  - 4 materials
  - *k* = 40 (20 + 5 * # materials)
  - There are 58 places where exchanges may happen
    - C = 22 for 2 processors
    - C = 117 for 128 or more processors
    - One collective operation per (58 total)
  - There are 31 other collective operations
    - S = 89

Predictions of Model

- Average message size 600,000 double precision
  - Cost of message should be about 4.9 ms – large compared to latency of 8.3 μs
- Use time on one processor for computational time on multiple processors
  - Predict from11.94 seconds on 2 processors to 12.41 seconds on 10360 processors
- Model does not account for all of the time
  - Does not model time to assemble messages or the additional computation associated with ghost cells

Comparisons with Profiling

- Profiled code with CrayPat on several numbers of processors
  - Only able to profile MPI portion of code due to limitations of CrayPat
  - Ran simulations twice – once for a few time steps and once for more and subtracted times
- Volume of message traffic consistent with number and length of message predicted
- Time for exchanges about a factor of 2 larger than predicted

Sandia National Laboratories

More Comparisons

- Number of collective operations from profile consistent with model
- On 32 processors model predicts 4.7 ms for collectives while profile reports up to 4.8 seconds
  - Load imbalance
    - Expected with this problem
- This plus the difference for the exchange times accounts for 80% of the difference between model and actual time
- Similar on other numbers of processors

Sandia National Laboratories

Summary and Further Work

- Modeling has helped us to understand what the code is doing
- Plan to repeat with a better load balanced problem
- Plan to repeat with current version of code
- Plan to work at modeling the code running with AMR turned on