

Scaling Performance of CubePM and TVD MHD on Blue Gene

Hugh Merz: merz@cita.utoronto.ca

Canadian Institute for Theoretical Astrophysics

Overview

CubePM is a cubic decomposition 2-level particle mesh gravitational N-body solver that is used to simulate cosmological dark matter structure formation. CubePM is designed to maximize memory usage in order to simulate the largest number of bodies possible. It is designed to run in a weak-scaling mode where the problem size is increased in proportion to the number of processes used.

In the 2-level particle mesh method gravitational forces are split such that short-range forces are calculated locally and long-range forces are calculated globally. The global force calculations require a distributed memory 3D FFT, and while the FFTW group (www.fftw.org) provides an MPI version of their FFT library it is only decomposed in 1-dimension (slab decomposition). This decomposition limits the size of simulation that can be run using CubePM due to an increasing memory footprint as the number of processes is increased. Porting CubePM to Blue Gene utilizing the volumetric decomposition BG/L 3DFFT (M. ELEFThERIOU ET AL. IBM J. RES. & DEV. VOL. 49 NO. 2/3 MARCH/MAY 2005) allows us to avoid the increasing memory footprint and conduct much larger simulations, while maintaining efficient weak-scaling performance. In addition, the initial condition generator for CubePM and part of the post-processing pipeline were also ported to Blue Gene, allowing us to perform and verify the results of scientific simulations entirely on Blue Gene.

TVD MHD is a total variation diminishing magnetohydrodynamic solver that is used in a number of simulations at CITA, including a development version of CubePM. A set of weak and strong scaling tests on Blue Gene found weak-scaling to be near-linear and strong-scaling to perform well, although as in the case of CubePM it is rarely used in a strong-scaling scenario.

CubePM Scaling

The pipeline used in the scaling tests required generation of the initial particle distribution with a pre-processing code. This initial configuration is written to disk, which is then read in by CubePM and evolved, with the final configuration written back to disk. Both the initial and final states are verified by comparing their 3D gravitational power spectra to the desired initial spectrum and semi-analytically predicted final spectrum, respectively.

The parameters of interest used in the weak-scaling test were:

Particles / process	64 ³
global mesh size	[128 * (processes/dimension)] ³
physical volume	1 (Gpc/h) ³
initial redshift	140

Due to decomposition constraints in CubePM and the BG/L 3DFFT library the available process sizes for the simulation were limited to powers of two. With a maximum of 2 racks available at the Blue Gene COD Center this allowed for the following simulation sizes:

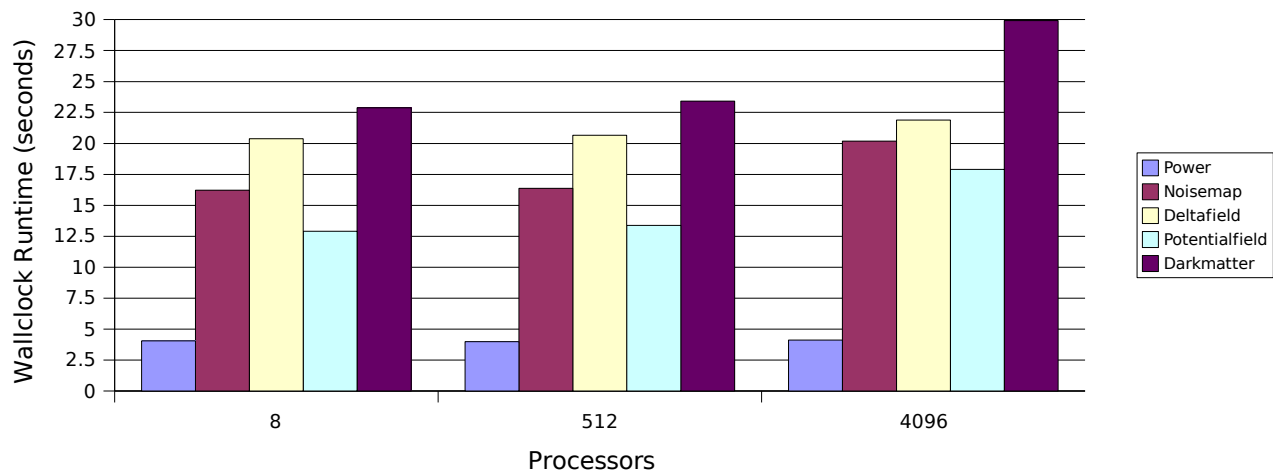
<i>Processes</i>	<i>Node Mode</i>	<i>Mesh Size</i>	<i>Particles</i>
8	Coprocessor	256	2,097,152
512	Coprocessor	1024	134,217,728
4096	Virtual	2048	1,073,741,824

I/O performance was found to severely degrade at larger process counts, as such the scaling results were focused on timing results from the communication routines in the codes rather than the total runtime and local, serial routines.

The initial condition generator performed as follows, with runtime in seconds:

<i>Processes</i>	<i>Subroutine (seconds)</i>				
	<i>Power</i>	<i>Noisemap</i>	<i>Deltafield</i>	<i>Potentialfield</i>	<i>Darkmatter</i>
8	4.05	16.23	20.39	12.9	22.88
512	3.99	16.38	20.66	13.38	23.42
4096	4.11	20.19	21.89	17.9	29.92

Initial Condition Generator Weak Scaling



Scaling from 8 to 512 processors is almost linear. Increasing to 4096 incurs a slight overhead, extending runtime by up to ~25%. This can partially be accounted for by differing work-loads on each process, especially in the darkmatter routine which experiences a non-uniform communication load that will be exacerbated by increasing the number of processes. Scaling the physical volume (held constant at 1Gpc) relative to the number of processes used would produce a more consistent workload and should be inspected in future tests.

Analysis of CubePM weak-scaling performance has to take into account load balancing as the simulation evolves. The initial particle load per process is almost homogeneous, but in late stages it will depend on the size of the initial density fluctuations that were used to generate the initial conditions. In the 4096 process run it was found that some processes had up to 4x their initial number of particles by the end of the simulation. This is an algorithmic limitation of CubePM and needs to be clearly separated from the performance of the underlying numerical methods. By comparing measurements at the start of the simulation this imbalance can largely be ignored.

CubePM wallclock runtime per timestep, averaged over the first 10 timesteps:

<i>Processes</i>	<i>Wallclock Runtime / Timestep (seconds)</i>
8	20.2
512	20.4
4096	20.5

This suggests that the code scales extremely well in the load balanced state. Focusing only on routines where significant communication occurs, again averaged over the first 10 timesteps:

<i>Processes</i>	<i>Wallclock Runtime / Timestep (seconds)</i>	
	<i>Particle Pass</i>	<i>Coarse Mesh Force</i>
8	1.11	0.27
512	1.11	0.28
4096	1.25	0.35

The particle pass routine communication pattern consists of single message exchanges of varying size with face-centered neighbors in the process space, while the coarse mesh force routine contains single message exchanges of fixed size with face-centered neighbors as well as the global 3D FFT. In both cases the code scales almost perfectly from 8 to 512 processes, and encounters ~25% overhead scaling to 4096.

As a rough gauge, the total runtime for the dark matter evolution over 97 timesteps was only 12% longer for the 4096 process case with respect to the 8 process case:

<i>Processes</i>	<i>Wallclock Runtime (seconds)</i>
8	1947
4096	2209

TVD MHD Scaling

TVD MHD scaling was measured by advecting a circularly polarized Alven wave through a uniform magnetic field for 5 timesteps and then taking the average wallclock runtime per timestep. All of the scaling tests were done using virtual node mode. Weak scaling was tested using a mesh with 120^3 zones / process, while strong scaling was measured using a 240^3 zone mesh for all process counts.

Weak scaling results:

<i>Processes</i>	<i>Wallclock Runtime / Timestep (seconds)</i>
1	33.9
8	33.7
512	33.0
1728	35.0

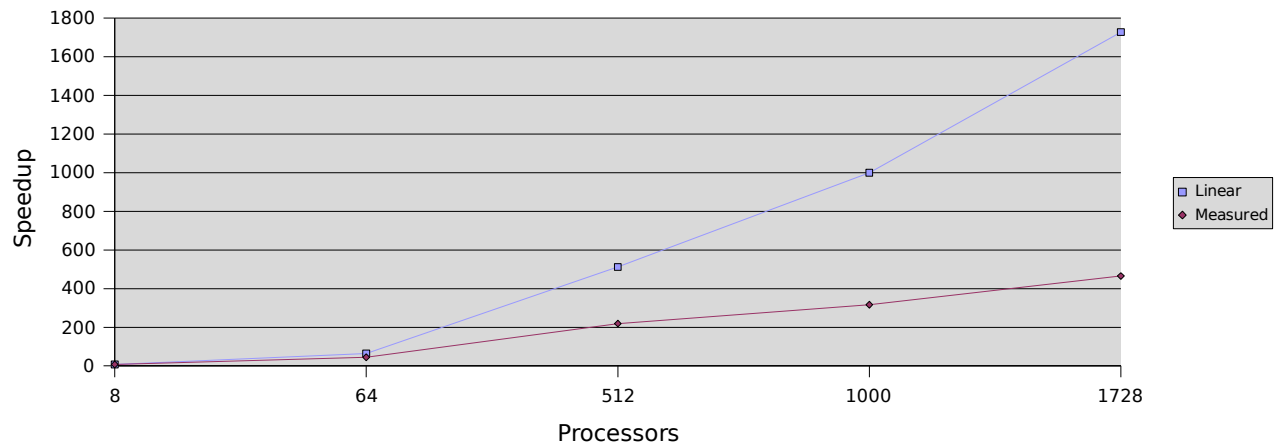
These results indicate that the TVD MHD algorithm will scale almost linearly in a weak scaling scenario up to thousands of processors.

Due to insufficient memory the strong scaling test commenced at 8 processes, assuming the speedup is roughly 8X that of a 1 process case. This may artificially inflate the speed up results and as such this data should only be qualitatively considered.

Strong scaling results:

<i>Processes</i>	<i>Wallclock Runtime / Timestep (seconds)</i>	<i>Speed Up</i>
8	33.72	8
64	6.01	44.9
512	1.23	219.3
1000	0.852	316.6
1728	0.58	465.1

TVD MHD Strong Scaling Speedup



There is a notable performance improvement at every process size, although communication overhead clearly eclipses the amount of computation per process for all process sizes greater than 64.

Conclusions

Cubepm has been successfully ported and tested on Blue Gene using the BG/L 3D FFT library. Weak scaling tests show that it performs very well up to 4096 processors, with only a marginal ~10% increase in runtime. The 4096 processor simulation modeled the evolution of 1 billion dark matter particles over 100 timesteps in less than 40 minutes.