Streaming Data-Parallel Algorithms Enable Cosmology Data Analysis for Large Halos

Christopher Sewell

Report
Streaming Data-Parallel Algorithms Enable Cosmology Data Analysis for Large Halos

**Objectives**

**Milestone**
- Implement application-specific visualization and/or analysis operators needed for in-situ use by LCF science codes
- Use VTK-m to take advantage of multi-core and many-core technologies

**Target Application**
- The Hardware/Hybrid Accelerated Cosmology Code (HACC) simulates the distribution of dark matter in the universe over time
- An important and time-consuming analysis function within this code is finding halos (high density regions) and the centers of those halos

**Impact**

**Recap of previously highlighted work**
- Data-parallel algorithms for halo and center finding implemented using VTK-m allowed the code to take advantage of parallelism on accelerators such as GPUs, and for the code to be portable across architectures
- On Titan, this enabled MBP centers to be found on the GPU ~50x faster than using the pre-existing algorithms on the CPU (with one rank per node)

**New challenge**
- At late time steps, particles become more concentrated in some nodes, resulting in great load imbalance for center finding
- This makes inefficient use of resources on large Titan runs, with most of the 16k nodes waiting idly for hours as a few nodes process large halos

**Accomplishments**

**Streaming Solution**
- During the initial full-system analysis task, all halos are found, but, instead of computing centers for all halos, particles in halos larger than a specified size are just output to disk using HACC’s GenericIO library
- These large halos can then be streamed into memory from file to find centers one at a time as single-node jobs
- Since no nodes wait idly, and since the jobs can be run even on another machine (such as Moonlight at Los Alamos), resource allocations are not wasted

**Science Impact**
- This streaming solution allowed halo analysis to be completed on the late time steps of a very large $8192^3$ particle data set across 16,384 nodes on Titan for which analysis using the existing CPU algorithms was not feasible
- This is the first time that the c-M relation has been measured from a single simulation volume over such an extended mass range (see graph at left)

**Publications**

---

Concentration-mass relation over the full mass range covered by the Q Continuum simulation at redshift $z = 0$ (points with error bars) and the predictions from various groups. The yellow shaded region shows the intrinsic scatter. All predictions and the simulation results are well within that scatter.