Seventh Workshop on Data Mining in Earth System Science (DMESS)

The Roosevelt New Orleans, New Orleans, Louisiana, USA

November 18, 2017

Scalable Algorithms for Clustering Large Geospatiotemporal Data Sets on Manycore Architectures

Vamsi Sripathi[1]

[1]Intel Corporation, Hillsboro, Oregon, USA

The increasing availability of high-resolution geospatiotemporal data sets from sources such as observatory networks, remote sensing platforms, and computational Earth system models has opened new possibilities for knowledge discovery using data sets fused from disparate sources. Traditional algorithms and computing platforms are impractical for the analysis and synthesis of data sets of this size; however, new algorithmic approaches that can effectively utilize the complex memory hierarchies and the extremely high levels of available parallelism in state-of-the-art high-performance computing platforms can enable such analysis. We describe a hybrid parallelism (MPI-OpenMP) based implementation of accelerated k-means clustering and some optimizations to boost computational intensity and utilization of wide SIMD lanes on state-of-the art multi- and manycore processors, including the second-generation Intel Xeon Phi ("Knights Landing") processor based on the Intel Many Integrated Core (MIC) architecture, which includes several new features, including an on-package high-bandwidth memory.