5.3 Plans for a GPU-accelerated MPAS-driven forecast system.

Hutchinson, Todd, Brett Wilt, James Cipriani, John Wong, *The Weather Company, an IBM Business*, Rich Loft, Bill Skamarock, Michael Duda, Dave Gill, and Raghu Kumar, *National Center for Atmospheric Research*

The Weather Company (TWC), an IBM Business, has been running numerical weather prediction (NWP) systems internally for the past 20 years. The systems have ranged from local domains driven by MM5 to continental-scale convective-allowing forecasts driven by WRF. Currently the next generation system is under development. It will be a global system driven by the Model for Prediction across Scales (MPAS) for the NWP and the Gridpoint Statistical Interpolation (GSI) software for the data assimilation. It will acquire the name "Deep Thunder", following on from the name given to NWP systems by IBM Research. GSI will be used to assimilate a variety of observations, including satellite (GOES-R, Himawari, POES, etc.), conventional (METARs, SYNOPs, cell-phone pressures, radiosondes), and radar. MPAS will integrate the GSI analyses forward in time to provide rapidly-updating global forecasts for the day ahead time period. Horizontal resolution will vary from convective-allowing over the populated areas of the world to coarser (~15km) over the rest of the world. The forecast output will be used throughout TWC product lines, including consumer (e.g., television, web and mobile forecasts), aviation (e.g., turbulence and convective hazards), energy demand, and insurance applications.

TWC, IBM Research, and NCAR have partnered to develop a GPU-accelerated software port of MPAS and to integrate this port into a future MPAS release. The Weather Company plans to run MPAS operationally on IBM Power9 systems that will contain 4-6 GPUs in each system. By utilizing GPUs, MPAS is expected to run significantly faster and thus, require less overall computing resources to achieve the same performance as compared to conventional CPU-only computing systems. Overall, the GPU-based systems will provide a significantly reduced datacenter footprint and electrical load, resulting in reduced ongoing operating costs.

As of April 2018, the dynamical MPAS core is running 2.7x faster on one nVidia Volta 100 GPU as compared to 2 Intel Broadwell CPUs. It is scaling reasonably well out to 16 GPUs. Further details on performance improvements for the dynamical core, as well as status on physics parameterization GPU ports will be provided at the Workshop.