LEGION

The complexity of the exascale systems that will be delivered, from processors with many cores to accelerators and heterogeneous memory, makes it challenging for scientists to achieve high performance from their simulations. Legion provides a data-centric programming system that allows scientists to describe the properties of their program data and dependencies, along with a runtime that extracts tasks and executes them using knowledge of the exascale systems to improve performance, thus shielding scientists from this complexity.

Increasing hardware specialization, power, and cost constraints will result in exascale systems with billion-way concurrency, a growing gap between memory and network latency and floating-point performance, heterogeneity in both processing and memory capabilities, and more dynamic performance characteristics due to power capping and highly tapered network topologies. Achieving sustained performance on these systems will require significant advances in latency hiding, minimizing data movement, and the ability to extract additional levels of parallelism from applications.

The Legion parallel programming system is a data-centric system for writing portable high-performance programs targeted at distributed, heterogeneous architectures designed to address these challenges. Legion presents abstractions which allow programmers to describe the properties of their program data, such as independence and locality. By making the Legion programming system aware of the structure of program data, it can automate many of the tedious tasks programmers currently face, including correctly extracting task- and data-level parallelism and moving data around complex memory hierarchies. A novel mapping interface provides explicit programmer-controlled placement of data in the memory hierarchy and assignment of tasks to processors in a way that is orthogonal to correctness, thereby enabling easy porting and tuning of Legion applications to new architectures to achieve performance.

The Legion team is focusing on developing new and modified features and integrating them into their programming system to address application requirements unique to the ECP, including better support for complex data structures, scalable data partitioning mechanisms, more versatile decomposition into different forms of parallelism, and more flexible and performant mechanisms to map computations and data to hardware.

Progress to date

- The Legion team provided regular releases of the software that reflect bug fixes, new features, performance improvements, and target system support. The features released are dependent upon testing, evaluation, and input from application teams.
- The team demonstrated significant performance improvements on real-world applications, up to a 7× performance increase over the baseline MPI version of a combustion simulation (S3D) and up to a 2.5× performance increase over the MPI+OpenACC version, and the ability to conduct experiments previously out of reach.
- The team obtained up to a 3× performance improvement in the training time for machine learning models.

PI: Pat McCormick, Los Alamos National Laboratory

Collaborators: Los Alamos National Laboratory, Stanford University, SLAC National Accelerator Laboratory, Argonne National Laboratory, NVIDIA