S6.04  Science Data Discovery in Extremely Large Data Environments

Lead Center: GSFC

Participating Center(s): JPL, LaRC

This subtopic focuses on supporting science data discovery for extremely large data environments through developing innovative cloud and large cluster based science data discover applications, application development tools, and performance monitoring tools. Specific areas for which proposals are being sought:

- **Science discovery applications**: Applications for science data discovery, data mining, data search, and data subsetting that scale to extremely large data sets in cloud or large cluster computing environments.
- **Application development tools**: Integrated ecosystem of tools for developing applications for high performance processing environments, including cloud computing, high performance cluster, and GPU processing environments, that support software development for science data discovery applications, including support for compilation, debugging, and parallelization.
- **Performance monitoring tools**: Integrated tools to collect, analyze, store, and present performance data for cloud computing and large scale cluster environments, including tools to collect data throughput of system hardware and software components such as node and network interconnects (GbE, 10 GbE, and Infiniband), storage area networks, and disk subsystems, and tools to capture data on a user configured basis, that allow extensibility for new metrics, and verification of the configuration and health of a system.

Research should be conducted to demonstrate technical feasibility during Phase I and show a path toward a Phase II hardware/software demonstration, and when possible, deliver a demonstration unit for functional and environmental testing at the completion of the Phase II contract.