HOLLISTIC MONITORING AND DIAGNOSIS OF LARGE-SCALE APPLICATION PERFORMANCE DEGRADATION

Research Challenge
This project presents a methodology for characterizing congestion in large-scale high-speed 3D Torus networks. Production characterizations of congestion manifesting as hot spots can be difficult, since they require systemwide, coherent data on the state of the network. In addition, data must be collected at fidelities necessary to capture the relevant phenomena. Runtime evaluations that identify localized congestion and assess congestion duration can be used to trigger mitigation responses, such as resource scheduling, placement decisions, and dynamic application reconfiguration.

Methods & Codes
This project developed Monet, a generic framework for supporting congestion characterization and diagnosis in HPC systems. Monet aggregates and analyzes the network and I/O, resilience, and workload data. The network stall counters are used in the extraction of congestion regions. The identified congestion regions are then combined with other data sets (workload data, network failure data, and network performance data) to enable detection, diagnosis, and characterization of network congestion.

Why Blue Waters
Blue Waters is one of the few open-science capacity systems that provides a test bed for scaling computations to tens or hundreds of thousands of cores on CPUs and GPUs. It also enables the study of failures and degradations of applications in production petascale systems with its unique mix of XE6 and XK7 nodes. This allows understanding the performance–fault-tolerance continuum in HPC systems by enabling the investigation of application-level designs for mixed CPU and GPU node systems, and fault isolation in system components to mitigate failures at the application level.

Results & Impact
The project team has shown an example use case in which their analysis methodologies and framework, Monet, were used to detect and correctly diagnose the congestion for an execution of the Enzo application, which was reading from the file system at the time of detection. The project-developed tool generates evidence for system managers and users by producing timeseries data and statistical distributions of stall and traffic characteristics for the implicated application.