Scalable IO for large-scale performance data

Participants: Jim Galarowicz, Todd Gamblin, Mark Krentel, John Mellor-Crummey, Felix Wolf

Problem: When performance measurement tools operate at very large processor configurations, they confront the problem of how to efficiently write process-local performance data in parallel to files. Experiences suggest that the traditional approach of having each process write a separate file does not scale. In this context, not only the aggregate data size but also the number of files seems to influence the I/O performance.

Approach: The sionlib library developed by Wolfgang Frings at the Jülich Supercomputing Centre solves this problem by mapping larger numbers of process-local files onto a single physical file. The file includes metadata to locate the data belonging to individual processes. The library has been designed in a platform-independent way, requiring only standard UNIX file system calls.

Discussion

- Having only one file might not provide optimal performance on Lustre. Both the number of files as well as the mapping of processes onto files should be configurable.
- How to use the library with hybrid applications: A collection of threads within an MPI rank coordinate themselves and write on behalf of the rank alone. Thread safety will be managed by the caller.
- Minor API changes/extensions.

Path forward

Interest in using sionlib expressed by HPCToolkit, OpenSpeedshop, and TAU. A summer student in Jülich will address the additional requirements during the next three months. Roll-out of a first release expected before SC08.