# Workshop on Resiliency in High Performance Computing (Resilience) in Clusters, Clouds, and Grids

Stephen L. Scott[1,2] and Chokchai (Box) Leangsuksun[3]

[1] Stonecipher/Boeing Distinguished Professor of Computing,
Tennessee Tech University, USA
[2] Oak Ridge National Laboratory, USA
[3] SWEPCO Endowed Associate Professor of Computer Science,
Louisiana Tech University, USA

Clusters, Clouds, and Grids are three different computational paradigms with the intent or potential to support High Performance Computing (HPC). Currently, they consist of hardware, management, and usage models particular to different computational regimes, e.g., high performance systems designed to support tightly coupled scientific simulation codes and commercial cloud systems designed to support software as a service (SAS). However, in order to support HPC, all must at least utilize large numbers of resources and hence effective HPC in any of these paradigms must address the issue of resiliency at large-scale.

Recent trends in HPC systems have clearly indicated that future increases in performance, in excess of those resulting from improvements in single- processor performance, will be achieved through corresponding increases in system scale, i.e., using a significantly larger component count. As the raw computational performance of these HPC systems increases from today's tera- and peta-scale to next-generation multi peta-scale capability and beyond, their number of computational, networking, and storage components will grow from the ten-to-one-hundred thousand compute nodes of today's systems to several hundreds of thousands of compute nodes and more in the foreseeable future. This substantial growth in system scale, and the resulting component count, poses a challenge for HPC system and application software with respect to fault tolerance and resilience.

Furthermore, recent experiences on extreme-scale HPC systems with non-recoverable soft errors, i.e., bit flips in memory, cache, registers, and logic added another major source of concern. The probability of such errors not only grows with system size, but also with increasing architectural vulnerability caused by employing accelerators, such as FPGAs and GPUs, and by shrinking nanometer technology. Reactive fault tolerance technologies, such as checkpoint/restart, are unable to handle high failure rates due to associated overheads, while proactive resiliency technologies, such as migration, simply fail as random soft errors can't be predicted. Moreover, soft errors may even remain undetected resulting in silent data corruption.

The goal of this workshop is to bring together experts in the area of fault tolerance and resiliency for HPC to present the latest achievements and to discuss the challenges ahead.