Barcelona Supercomputing Center
Centro Nacional de Supercomputación

//Resilient

Towards Resilient EU HPC Systems: A Blueprint

This document aims to spearhead a Europe-wide discussion on HPC system resilience and to help the European HPC community define best practices for resilience. It analyses a wide range of state-of-the-art resilience mechanisms and recommend the most effective approaches to employ in large-scale HPC systems. These guidelines will be useful in the allocation of available resources, as well as guiding researchers and research funding towards the enhancement of resilience approaches with the highest priority and utility. Although it is focused on the needs of next generation HPC systems in Europe, the principles and evaluations are applicable globally.

This document is the first output of the ongoing European HPC resilience initiative and it covers individual nodes in HPC systems, encompassing CPU, memory, intra-node interconnect and emerging FPGA-based hardware accelerators. With community support and feedback on this initial document, we will update the analysis and expand the scope to include other types of accelerators, as well as networks and storage.

The need for resilience features is analysed based on three guiding principles:

1. The resilience features implemented in HPC systems should assure that the failure rate of the system is below an acceptable threshold, representative of the technology, system size and target application.
2. Given the high cost incurred by uncorrected error propagation, hardware errors should be detected and corrected frequently and at low overhead, which is likely only possible in hardware.
3. Overheating is one of the main causes of unreliable device behaviour. Production HPC systems should prevent overheating while balancing power/energy and performance.

Based on these principles, the main outcome of this document is that the following features should be given priority during the design, implementation and operation of any large-scale HPC system:

- ECC in main memory
- Memory demand and patrol scrub
- Memory address parity protection
- Error detection in CPU caches and registers
- Error detection in the intra-node interconnect
- Packet retry in the intra-node interconnect
- Reporting corrected errors to the BIOS or OS (system software requirement)
- Memory thermal throttling
- Dynamic voltage and frequency scaling for CPUs, FPGAs and ASICs
- Over-temperature shutdown mechanism for FPGAs
- ECC in FPGA on-chip data memories as well as in configuration memories

The remaining state-of-the-art resilience features surveyed in this document should only be developed and implemented after a more detailed and specific cost–benefit analysis.