Resilience Design Patterns: A Structured Modeling Approach of Resilience in Computing Systems

Christian Engelmann, Oak Ridge National Laboratory
engelmannc@ornl.gov

Resilience to faults, errors, and failures in extreme-scale high-performance computing (HPC) systems is a critical challenge. Resilience design patterns (Figure 1) offer a new, structured hardware/software design approach for improving resilience by identifying and evaluating repeatedly occurring resilience problems and coordinating corresponding solutions. Initial work identified and formalized these patterns and developed a proof-of-concept prototype to demonstrate portable resilience [1,2]. This recent work created performance, reliability, and availability models for each of the identified 15 structural resilience design patterns and a modeling tool that allows (1) exploring the performance, reliability, and availability of each pattern, and (2) investigating the trade-offs between patterns and pattern combinations [3,4].

The model for each of the 15 structural design patterns consists of a flowchart and state diagram, identifying its dynamic error/failure-free behavior and when handling errors/failures. It also includes mathematical models for performance (error/failure-free execution time and under error/failure conditions), reliability (probability of not experiencing an error/failure) and availability (portion of time a system provides correct service). The reliability and availability models rely on exponential error/failure distribution to make a modeling approach possible. Other distributions, such as Weibull, would require a simulation approach. The modeling tool relies on parametrized descriptions of patterns to calculate and plot performance, reliability and availability. For example, Figure 2 shows the results for a 2-level checkpoint/restart (CR) solution, with fine-grain CR at the compute node or accelerator level and coarse-grain CR at the job level. Complex horizontal and vertical pattern combinations can be modeled to understand system behavior.

This work concludes a 6-year research and development effort in understanding the fault, error and failure characteristics of extreme-scale HPC systems and in the design pattern approach for improving resilience. It created new concepts, methods and proof-of-concept prototypes for understanding the resilience problem and for designing resilient HPC systems. The major lesson learned was that extreme-scale HPC systems often display unexpected fault/error/failure modes. Therefore, resilience needs to become an integral part of the HPC hardware/software ecosystem through codesign, such that the burden for providing resilience is on the system by design and not on the operator or user as an afterthought. The resilience design pattern approach offers this capability by identifying, classifying, quantifying and coordinating the detection, containment and mitigation properties of individual resilience solutions and their vertical and horizontal compositions within an extreme-scale HPC system, avoiding coverage gaps and overprotection.

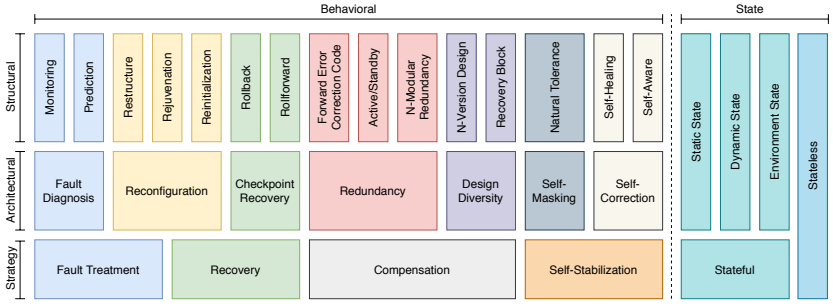
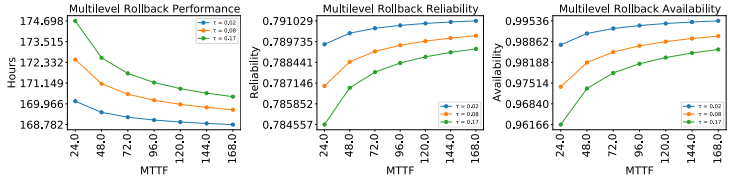


Figure 1: Classification of resilience design patterns



(a) Performance (b) Reliability (c) Availability

Figure 2: Multi-level Rollback performance, reliability, and availability

[1] S. Hukerikar and C. Engelmann. Resilience Design Patterns: **A Structured Approach to Resilience at Extreme Scale**. *Journal of Supercomputing Frontiers and Innovations (JSFI)*, volume 4, number 3, pages 4-42, 2017.

[2] R. Ashraf, S. Hukerikar, and C. Engelmann. **Pattern-based Modeling of Multiresilience Solutions for High-Performance Computing.** In *Proceedings of the 9th ACM/SPEC International Conference on Performance Engineering (ICPE) 2018.*

[3] M. Kumar and C. Engelmann. **Models for Resilience Design Patterns.** In Proceedings of the *10th Workshop on Fault Tolerance for HPC at eXtreme Scale (FTXS) 2020.*

[4] M. Kumar and C. Engelmann. RDPM: **An Extensible Tool for Resilience Design Patterns Modeling.** In Proceedings of the *14th Workshop on Resiliency in High Performance Computing (Resilience) in Clusters, Clouds, and Grids 2021.*

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, program managers Robinson Pino and Lucy Nowell.