LA-UR-24-28158

Author(s): Robin E. Preble

Mentors: Conor Robinson, Graham Van Heule

Title: Cluster Care: Reducing Downtime with Automated Node Failure Recovery

Abstract:

Maintaining the functionality of thousands of nodes in large clusters is a labor-intensive task for system administrators. Nodes often fail for common and well documented reasons with established remediation procedures. Manually intervening to fix each of these nodes is time consuming and can necessitate urgent responses, requiring employees to come in during off hours. This presentation gives an overview of Cluster Care, an automated solution to streamline the remediation process for node failures within clusters. This tool collects data on the current state of each node in the cluster and automatically executes configurable remediation procedures based on this information. Information about node states and actions performed is logged for tracking in Splunk. The system's modular design allows for extensive customization, enabling admins to configure monitoring tools, action commands, and mappings from node states to specific remediations, making it easy to adopt the tool for use in a variety of cluster environments. With Cluster Care's robust checks and automated remediations, users can expect to see improved availability for production workflows.