|  |  |
| --- | --- |
| **I/UCRC Executive Summary - Project Synopsis** | **Date**: |
|  **Center/Site**: Security and Software Engineering Research Center / Georgetown University |
|  **Tracking No**.: |  **Phone :** ( ) - |  **E-mail :** |
|  **Center/Site Director:** Eric Burger | **Type: (New or Continuing)** New |
|  **Project Leader:** Wenchao Zhou | **Proposed Budget**: $ |
|  **Project Description**: Operators of distributed systems often find themselves needing to answer forensic questions, to perform a variety of managerial tasks including fault detection, system debugging, accountability enforcement, and attack analysis. Provenance provides the fundamental functionality required to answer such forensic questions -- the capability to "explain" the existence (or change) of a certain distributed system state. However, the support of provenance often requires the primary system to maintain additional information which may hurt its performance. This project addresses the issues by considering recent advance in database architectures that leverages solid-state drives to enable efficient implementation of log-structured databases, in which provenance can be naturally supported (i.e., provenance can be regenerated from the logs) without impacting the performance of the primary system. This project will explore two research directions – 1) an efficient data structure that assists provenance operations. For instance, each log entry maintains a pointer to the last entry that updates the same data item, such that one can easily regenerate a specific system state without touching and replaying all log entries; 2) querying techniques that can allow provenance queries (such as root cause analysis) to be efficiently supported.  |
|  **Experimental plan**: This project plans to evaluate the performance of the proposed provenance systems against several applications including networking routing (a long-running application with simple execution logic), distributed hash table (a long-running application with complex logic), and cloud computing applications (batch applications). This project will evaluate the performance impact on the primary system, and the performance (e.g., querying latency) of the provenance system. |
|  **Related work elsewhere**: The project is related to two bodies of research literatures: 1) There has been extensive work on enabling provenance support for distributed systems, file systems, and scientific workflows. In these work, provenance is usually maintained in relational tables (often with fixed schemas) in addition to system state. 2) There has been a focus on leverage emerging hardware such as solid-state drives to improve the performance of database systems. For instance, the Hyder system proposed a new log-structured infrastructure that leverages the physical characteristics of SSDs to improve the performance of database systems. |
|  **How this project is different**: Existing work on provenance assumes traditional database systems that actively maintain the current state of the system in relational tables, where the support of provenance requires recording additional information (such as provenance updates or logs of non-deterministic events), and inevitably imposes additional computation and storage overhead. This project considers an orthogonal approach leverages the solid state drives to enable implementation of efficient log-based databases, in which provenance is naturally supported with little overhead. |
|  **Milestones for the current proposed year**: * Develop the log-based databases and its update and query interfaces
* Develop the provenance support based on the log-based databases
* Evaluate the performance of the provenance system based on example applications including network routing, distributed hash tables, and a cloud computing platform
 |
|  **Deliverables for the current proposed year**: * Papers that describe the design and implementation of the project
* Prototype implementation of the proposed provenance system
* Viability and performance evaluation reports
 |
|  **How the project may be transformative and/or benefit society**: This project could potentially provide an efficient platform that assists understanding and debugging the behavior of distributed systems.  |
|  **Research areas of expertise needed for project success:** Distributed systems, Databases, Networks |
|  **Potential Member Company Benefits:**  |
| **Progress to Date:**  |
| **Estimated Start Date**:  | **Estimated Knowledge Transfer Date**: |

The Executive Summary is used by corporate stakeholders in evaluating the value of their leveraged investment in the center and its projects. It also enables stakeholders to discuss and decide on the projects that provide value to their respective organizations. **Ideally, the tool is completed and shared in advance of IAB meetings to help enable rational decision making.**