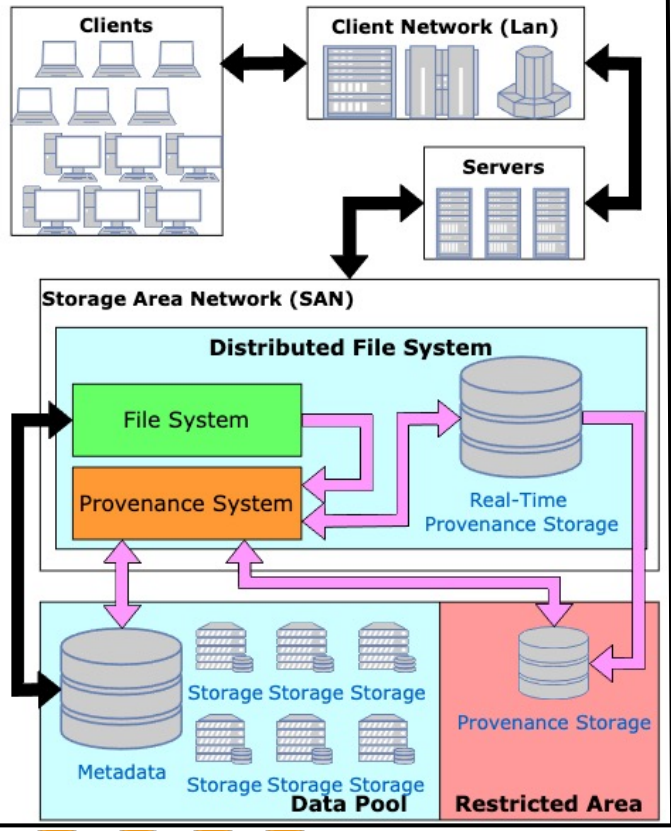
A Secure Provenance System Architecture for Distributed Storage Systems

PhD Project by **Hetong Jiang**[1], School of Information Technology and Electrical Engineering, Contact: hetong.jiang@uq.edu.au
Supervised by Ryan Ko[1], Guangdong Bai[1], David Abramson[1], Mohan Baruwal Chhetri[2]

Introduction

In this project, we would like to propose a secure provenance system architecture for Distributed Storage Systems. This approach addresses the difficulty of applying secure provenance in wide-area distributed file systems. This project is still at the beginning stage of design and implementation. Early-stage literature reviews have been done, which prove the project is noval and can be achieved.



Motivation

Several provenance systems exist, some of which can support secure provenance in a distributed environment. However, tracing and synchronising secure provenance multiple data centres across multiple physical locations is still missing. Most of the provenance systems for distributed storages are either for workflow provenance or not suitable for wide-area distributed architecture, which is not sufficient for security purposes.

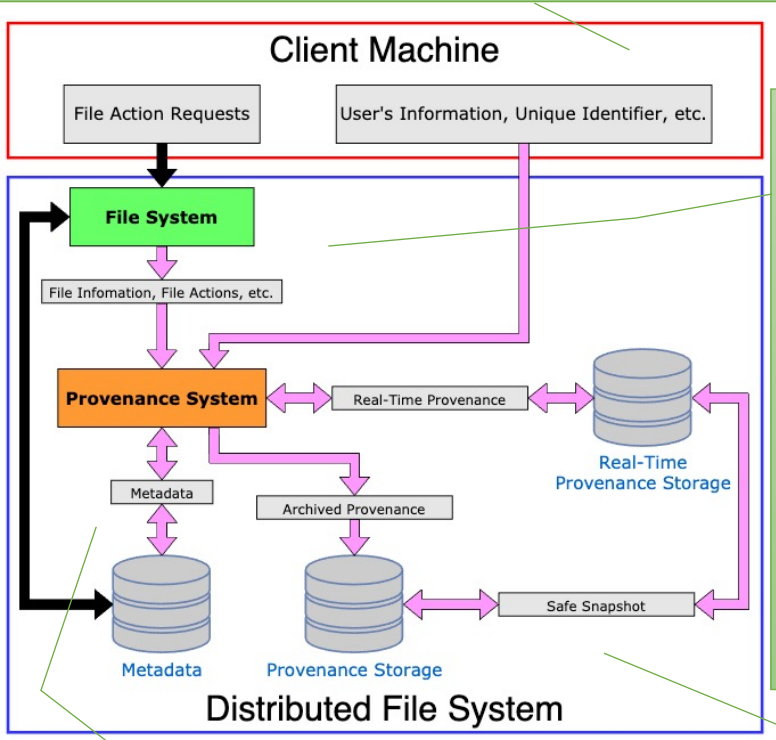
Design Goal

The proposed provenance system architecture allows the Storage Area Network(SAN) layer to collect complete and authentic provenance across wide-area distributed storage systems. It is designed as a part of the SAN, and all required information will be sent to and processed inside the provenance system. Depending on the setup, it allows the system to retain **secure provenance** and **versioned metadata**, even versioned files, if needed. Those provenances can be used to trace the complete lineage of a file's metadata, file usage and other security purposes such as forensic investigation and evidence preservation.

Outcome

The outcome of the project is a noval provenance system architecture for distributed storage systems that can collect complete and authentic provenance, stores the provenance with a specially designed storage that can ensure integrity and confidentiality, and provide provenance access that has high availability while being able to assuredly transfer provenance to the end consumers or next stop for further process.

The proposed provenance does not require installing any module in the user's machine. Instead, the whole provenance procedures rely on the SAN level provenance system. The provenance system will ask the higher level (Servers, LAN, Clients, etc.) for user's information. Each client will have a dedicated identifier, which will be stored in the provenance system along with the provenance records.



File action will go through the file system(SAN) as usual. In addition, the provenance system will record the file actions, optimise the records, combining with other information and store them into the provenance storage. The provenance system will record all actions through the file systems to ensure completeness and authenticity.

Metadata will work, as usual, in addition, the provenance system will record different versions of metadata, and metadata storage is able to request snapshots of metadata at a particular time point from the provenance system.

All provenance will be stored inside delicate provenance storage, it will take a part of the data pool of the SAN to take full advantage of the stability, with fast-accessing storage for real-time provenance accessing.

[1] The University of Queensland
[2] Data61, Docklands GS