NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Privacy-preserving Computing with Mix-sensitivity Data on Hybrid Clouds**

Speaker: Mr Zhang Chunwang

Date/Time: 11 September 2014, Thursday, 04:00 PM to 05:30 PM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Chang Ee Chien, Associate Professor, School of Computing

Abstract:

hybrid cloud could seamlessly integrate a private and a public cloud, offering increased scalability and cost-effectiveness. Nevertheless, if data are handled on the hybrid cloud without security measures, sensitive information could be leaked to the untrusted public cloud. In this thesis, we are interested in privacy-preserving computing with mixed-sensitivity data on hybrid clouds. We look at both general MapReduce computation as well as an application on video surveillance streams.

The MapReduce framework is designed for only one (logical) cloud and may leak sensitive information if used in a hybrid cloud with sensitive data. In view of this, we propose extending MapReduce by augmenting each key-value pair with a sensitivity tag. The tags enable fine-grained dataflow control during execution to prevent information leakage. More importantly, the tagging provides increased flexibility by allowing sophisticated security polices and facilitating complex MapReduce computation with chained jobs. To address potential performance issues introduced by the security constraint, we exploit useful properties of the MapReduce functions and present three scheduling modes which can rearrange the computation for increased efficiency while maintaining MapReduce correctness. A generic security framework is also provided for analyzing what kind of information a scheduler can leak through MapReduce computation on hybrid clouds. Experiments on Amazon EC2 show that our prototype on Hadoop is able to preserve data-privacy while effectively outsourcing computation and reducing inter-cloud bandwidth usage.

We next consider processing of mixed-sensitivity video surveillance streams. The challenge falls on how to schedule multiple tasks over a large number of video streams onto the two clouds. We first present a stream processing model that is specifically designed for the hybrid cloud setting. Based on this model, we formalize the scheduling challenge as an optimization problem that minimizes the monetary cost to be incurred on the public cloud,

with a set of resource, security and Quality-of-Service (QoS) constraints. Our proposed scheduler exploits useful properties of the hybrid clouds for more efficient solutions, and thus allows for larger instances. Both the simulations and proof-of-concept system evaluation on Amazon EC2 demonstrate its effectiveness and efficiency.

We conclude that the hybrid cloud is a practical and effective solution for privacy-preserving large-scale data processing. With the well-designed scheduling mechanisms, the overheads incurred by the security constraint can also be significantly reduced.