There will be a U-REaSON seminar event on this Friday, April 13th, from 12:00pm to 1:00pm in room 205 of the Civil Engineering Building. Speakers for this week are Chao Chen (12:00pm - 12:30 pm) and Jialin Liu (12:30 pm -1:00 pm).

Speaker: Chao Chen

Topic: Dynamic Active Storage for High Performance I/O

Abstract-- Many High-End Computing applications in critical areas of science and technology are becoming more and more data intensive. These applications transfer large amounts of data from storage nodes to compute nodes for processing, which is costly and bandwidth consuming and often dominates the applications run time. Active storage provides a promising solution for these applications by moving appropriate processing tasks from processing nodes to storage nodes. The prior research has achieved considerable progress and developed several active storage models. However, the existing studies have neglected that it is not optimal to offload every operation to storage nodes, especially when data dependence exists among operations.
Such data dependence often exists in applications, such as the Geographic Information System and medical image processing. In such situation, it often wastes bandwidth for the conventional active storage architectures, because it is needed to transfer dependent data sets among storage servers. In order to resolve this problem, this paper presents a novel Dynamic Active Storage
(DAS) system that analyzes the bandwidth requirement of an operation and then determines whether it is suitable for offloading to storage servers.
Furthermore, based on the analysis of the bandwidth requirement, this paper proposes a data layout strategy for the DAS system to reduce the bandwidth requirement. Experimental tests have been conducted, and the results have confirmed that this new active storage architecture outperforms existing architectures. It significantly reduces the data movement caused by data dependence and improves applications performance over existing strategies.
It has a potential for high performance I/O in high-end computing.

Speaker: Jialin Liu

Topic : FSD: Fast Query with Integrated Statistical Information in Scientific Datasets

Abstract—Scientific datasets, such as HDF5 and PNetCDF, have been used in many scientific applications in data intensive fields. These file formats and programming interfaces provide efficient access to large volume of data storage. Researchers have been interested in integrating modern database techniques and parallel file I/O into the management of scientific datasets, in which I/O performance and query efficiency are both important criteria. In this research, we analyze how the subsetting partition can affect the access and analysis efficiency of datasets. Based on subsetting extraction, we present a new idea of adding statistical information into the datasets. The statistical information illustrates the data distribution features, and the parallel access code can utilize these metadata to perform fast query. Though the added metadata may increase the original data size, we evaluate this trade-off issue though experiments. This research is the first study that utilizes statistical information with different ways of subsetting to perform fast query. It is currently evaluated with the PNetCDF library, but can also be implemented in other scientific data management libraries. The idea we present in this research can lead to a new datasets design and can have an impact.