UNCLASSIFIED

Information Science and Technology Seminar Speaker Series

Gagan Agrawal
Ohio State University

Programmability, Performance, and Dealing with Memory Constraints in In-Situ Analytics

Wednesday, July 29, 2015

3:00 - 4:00 PM

TA-3, Bldg. 1690, Room 102 (CNLS Conference Room)

Abstract: Neither the memory capacity, memory access speeds, nor disk bandwidths are increasing at the same rate as the computing power in current and upcoming parallel machines. This has led to considerable recent research on in-situ data analytics. However, many open questions remain on how to perform such analytics, especially in view of memory constraints in systems. This talk will describe two different directions of research in this area.

Developing an efficient in-situ implementation involves many challenges, including parallelization, data movement or sharing, and resource allocation. Based on the premise that MapReduce can be an appropriate API for specifying scientific analytics applications, we present a novel MapReduce-like framework that supports efficient in-situ scientific analytics, and address several challenges that arise in applying the MapReduce idea for in-situ processing. Specifically, our implementation can load simulated data directly from distributed memory, and it uses a modified API that helps meet the strict memory constraints of in-situ analytics. The framework is designed so that analytics can be launched from the parallel code region of a simulation program. We have developed both time sharing and space sharing modes for maximizing the performance in different scenarios, with the former even avoiding any copying of data from simulation to the analytics program.

Second, we propose an approach that utilizes bitmap index (bitmaps) as the summary structure, performs further data reduction (such as time-steps selection) using just them, and subsequently, stores only the selected bitmaps for post-analysis. We construct compressed bitmaps on the fly, show that many kinds of in-situ analyses can be supported by bitmaps without requiring the original data (and thus reducing memory requirements for in-situ analysis), and instead of writing the original simulation output, we only write the selected bitmaps to the disks (reducing the I/O requirements). We also demonstrate that we are able to use bitmaps for key offline analysis steps.

Biography: Gagan Agrawal is a Professor of Computer Science and Engineering at Ohio State University. He received his B. Tech degree from IIT Kanpur and MS and PhD degrees from University of Maryland, College Park. His research interests include parallel programming systems and tools, and various aspects of `big data', including data management and data mining. He has published more than 250 papers in these areas and supervised 25 PhDs to completion.

For more information contact the technical host Curt Canada, cvc@lanl.gov, 665-7453 or James Ahrens, ahrens@lanl.gov, 667-5797.

Hosted by the Information Science and Technology Institute (ISTI)

Los Alamos NATIONAL LABORATORY EST. 1943

ISTI INFORMATION SCIENCE & TECHNOLOGY INSTITUTE