The University of Chicago Computer Science Department

PRESENTS:

"Scaling database systems to high-performance computers"

Spyros Blanas

Assistant Professor, The Ohio State University

Abstract:
We are witnessing the increasing use of warehouse-scale computers to analyze massive datasets quickly. This poses two challenges for database systems. The first challenge is interoperability with established analytics libraries and tools. Massive datasets often consist of images (arrays) in file formats like FITS and HDF5. To analyze such datasets users turn to domain-specific libraries and deep learning frameworks, and thus write code that directly manipulates files. We will first present ArrayBridge, an open-source I/O library that allows SciDB, TensorFlow and HDF5-based programs to co-exist in a pipeline without converting between file formats. With ArrayBridge, users benefit from the optimizations of a database system without sacrificing the ability to directly manipulate data through the existing HDF5 API when they want to.

The second challenge is scalability, as warehouse-scale computers expose communication bottlenecks in foundational data processing operations. This talk will focus on data shuffling and parallel aggregation. We will first present an RDMA-aware data shuffling algorithm that transmits data up to 4X faster than MPI. This is achieved by switching to a connectionless, datagram-based network transport layer that scales better but requires flow control in software. We will then present a parallel aggregation algorithm for high-cardinality aggregation that carefully schedules data transmissions to avoid unscaleable all-to-all communication. The algorithm leverages similarity to transmit less data over congested network links. We will conclude by highlighting additional challenges that need to be overcome to scale database systems to massive computers.

Bio:
Spyros Blanas is an assistant professor in the Department of Computer Science and Engineering at The Ohio State University. His research interest is high performance database systems, and his current goal is to build a database system for high-end computing facilities. He has received the IEEE TCDE Rising Star award and a Google Research Faculty award. He completed his Ph.D. at the University of Wisconsin–Madison where part of his Ph.D. dissertation was commercialized in Microsoft SQL Server as the Hekaton in-memory transaction processing engine.

Monday, October 15, 2018
3:00 pm
JCL 390
Host: Aaron Elmore

People in need of assistance should call 773-702-3508 in advance.
For additional information on future CS talks please visit: http://www.cs.uchicago.edu/events