北京大学高能效计算与应用中心学术报告

Invited Talk, Center for Energy-Efficient Computing and Applications

开放和开源环境下的大数据处理系统的软件研发

Prof. Xiaodong Zhang

Robert M. Critchfield Professor in Engineering

The Ohio State University

2014年12月19日 星期五 01:00pm

理科二号楼2736多功能厅

主办单位：高能效计算与应用中心

网络与信息系统研究所

报告摘要： 从数据处理的角度上看，大数据给计算机系统带来了几个新的挑战和需求。（1）已成熟的数据库系统，包括并行数据库系统，它们的架构不是为大数据而设计的，所以在可扩展性和容错性上都不能满足大数据的处理得要求。（2）大数据分析是全社会的各个领域里的重要任务，所以数据处理的基础设施（硬件和软件）必须是廉价和低成本的。现有数据处理的商业模式是不能满足这个应用需求的。（3）在大数据系统设计和实现以及数据分析中，需求很多新的软件工具。（4）数据处理的计算模式已由为高性能计算服务的Scale-Up模式转变为为高数据通量服务的Scale-Out模式。

我将介绍系统软件研发人员是如何为解决以上4个问题在开放和开源地环境下研发Apache Hive的。Apache Hive 是一个在全世界被广泛使用的大数据仓库。在过去的4年里，我们在用户需求的指导下，确认了在Hive原始版的几个重要缺陷，包括存贮结构，查询计划及其优化，还有数据处理的运行引擎。我会重点讲解学术研究为Hive在大数据分析中所起到的重要作用。我还会讨论一个IT领域里的可持续发展的趋势：为了保证其高质量和高可信度，通用的基础软件一定是开源的。

报告人简介： 张晓东是美国俄亥俄州立大学的Robert M. Critchfield讲席教授，并担任计算机科学与工程系主任。他的研究方向是计算机和分布式系统中的数据和存储管理。他主持研究的一些核心算法和系统设计已被广泛应用到主流的CPU芯片，以及主要操作系统,存贮系统，数据库系统和大型的分布式系统中，有效地优化或更新了计算机系统中的一些关键技术。

张晓东在北京工业大学获电气工程学士学位，在美国科罗拉多大学获计算机科学博士学位，并获得该校2011年度工程与应用科学的杰出校友奖。他还获得2010年中国计算机学会海外杰出贡献奖。他是国际计算机学会（ACM）Fellow，也是国际电气电子工程师学会(IEEE) Fellow。