SE&T Colloquium Series-Fall 2016

Speaker	Dr. Il-Hyung Cho Department of Computer Science and Information Systems
Title	***Big Data Analysis using Apache Hadoop on Amazon Clouds***
Abstract	There have been various approaches to improve computing performance over the years. A simple approach was to make a processor faster with faster clock speed and more transistor counts in the CPU. This approach had worked well and the processor performance had doubled every two years, as predicted by the Moore's law, until early 2000s. Since then, the top clock speed has remained the same at under 4 GHz due to the physical limitation, and people started looking at multi-core technology for higher performance. However, increasing the number of CPU cores did not scale well due to the lack of support from application software and underlying OS, as well as the limitation of transistor counts in a single chip. Parallel and distributed processing technology has been around for quite a while for high performance computing since late 70's, and recently cluster, grid, and cloud based computing became popular. With the advent of Big Data (tera bytes or even peta bytes of data) analysis, it is practically impossible to process such big data on a single machine simply because the data is too big. Super computers are good for complex number crunching applications (e.g., weather forecast, simulation, games, …), but not suitable for such large data processing. Also, super computers are off limits to normal business where big data needs to be analyzed. The current Cloud technology allows thousands or tens of thousands of virtual machines configured to work together to solve big data problem. For example, even if the data size is 1 petabyte (10^{15} byte), once distributed to 10,000 nodes Cloud, each node (machine) only needs to deal with 500 gigabyte of data, which is quite large to process, but yet manageable for each individual machine. Apache™ Hadoop® is an open source software project that can be used to efficiently process large datasets. Hadoop allows combining commodity hardware together to analyze massive data sets in parallel. Amazon Cloud supports Hadoop, and this presentation shows how to deploy a Hadoop cluster on Amazon Clouds with real life example.
Date	Tuesday, September 13
Time	4:10-5:00pm
Place	Pioneer 240
	Refreshments will be served at 4:00pm.