ΔΙΑΛΕΞΗ

" Optimized Large Scale Systems"

Christos Kozanitis
Research Collaborator
FORTH-ICS

Περίληψη – Abstract

In the light of the end of the Moore's Law and Dennard scaling eras, which affect the evolution of CPUs and RAM capacity, existing Big Data solutions do not scale sustainably. Today's distributed analytics frameworks optimize mainly for coding simplicity and make liberal use of cluster resources, as they were initially designed under the assumption that the main bottlenecks in distributed computing come from the network and storage. On the other hand, not only do dataset sizes grow exponentially, data processing algorithms also become too sophisticated to run efficiently in distributed infrastructures. Moreover, deployments at scale require the use of expensive infrastructures to run for days.

My work copes with all those challenges as follows: 1) We enable Apache Spark to become more memory efficient for iterative analytics workloads by trading RAM for Disk at no significant CPU overhead. 2) In the case of sophisticated data processing algorithms, we show how a more careful partitioning improves the scalability of large scale collaborative filtering. 3) And in the case of infrastructure costs, we enable the Kubernetes scheduler to automatically allocate resources to applications.

Christos Kozanitis is a research collaborator at FORTH-ICS. He received his M.S. and Ph.D in Computer Science and Engineering from the University of California, San Diego in 2009 and 2013 respectively. Parts of his phd work influenced products from companies such as Cisco and Illumina. He also held a two-year postdoctoral appointment at the AMP Lab of the University of California, Berkeley, where he used and adapted state of the art big data technologies, such as Apache Spark SQL, Apache Parquet and Apache Avro to process large amounts of DNA sequencing data. He leads, on behalf of FORTH, the Greek Project Sentitour (Erevno-Dimiourgo-Kainotomo) and he also participates with a leading role in the H2020 funded project Evolve. His current research interests involve the improvement in software, storage and hardware level of modern distributed frameworks in order to speed up the processing of big data workloads.

Δευτέρα 26/10/2020 – 15:00-16:00