Keynote Talk

Data Centric Computing for Internet Scale Enterprises

Yuqing Gao
IBM T. J. Watson Research Center
Yorktown Heights, NY

ABSTRACT

In the era of exploding internet usage, social and mobile, enterprises are facing both the challenges and business opportunities that are introduced by Big Data, which has the characteristics of high volume, high velocity, and high variety. Big Data and the emergence of Internet-facing workloads will blur the separation between traditional transactional and analytics workloads. To extract business value and make actionable insight from the unprecedented volume of the data with the agility required from the business, it requires transformational innovations from many fronts. For example, in data management layer, how unstructured data is stored and retrieved efficiently, how data-intensive analytic computation can be done on commercial systems effectively, how the distributed cache should be designed to make use of the latest network protocols so the network-connected memory data can be accessed remotely and seamlessly. Moreover, the trend also motivates many architectural and technological advancement, such as moving from a transaction-centric to a data-centric architecture that supports extreme low and predicable latency, massive scale-out, high concurrency, and real-time situational awareness and analytics, and that requires orders of magnitude improvement over existing systems across each of these characteristics. At the same time, new applications in the Mobile and social space leverage new open source software stacks written in multiple programming languages, e.g., Java, JavaScript, Ruby, PHP, where the developer chooses the best tool for the job. How a polyglot runtime platform can be built that serves as a best practice platform for the programmers' community and in the meantime, optimized for enterprises with elastic, lightweight, resilient, agile runtime for business computing. Last, but not least, how the benchmarks should be enriched to measure the new runtimes, new data-centric systems and architectures.

In this talk, I will talk about some of the research activities at IBM Research that addresses these challenges. We examined several enterprise-grade java workloads running on commercial multicore systems for massive parallelization, identified lock contentions and worked towards a streamlined methodology for lock-contention analysis of Java workloads. I will use this to describe the excitement around node.js framework. I will also talk about our design of data centric computing systems, particularly, in the area of data access latency, data ingestion, and massive scale-out distributed caching in the exemplary context of an eCommerce application. I will describe the architecture of a global secondary index to greatly improve data access latency of Hadoop Database (HBase), an open-source key-value distributed datastore. I will describe an innovative distributed caching system that exploits low latency interconnects to utilize hash maps of data keys on each server for local lookup while data resides and are accessed across clustered systems. The distributed cache can achieves 100 to 1000-fold performance gain over many caching methods. Last, I will talk about our early activities in developing technologies for an elastic, scalable, resilient polyglot runtime system. I will conclude with my views on the challenges for benchmarking community for next decade.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Performance

Keywords

Big data, data access, workload characterization

ICPE'13, April 21–24, 2013, Prague, Czech Republic.
ACM 978-1-4503-1636-1/13/04.