Colloquium

Dynamic Performance Profiling of Data Caches

Ymir Vigfusson
Reykjavik University

Abstract: Scalable data replication protocols and layers, such as streaming, multicast and caching, enable large data-driven distributed systems to be practical. As a concrete example, large-scale in-memory object caches like memcached are now widely used to accelerate popular web sites and to reduce burden on backend databases. Yet operators still have limited visibility into how these caches should be set up to optimally accommodate the workloads they see. How much would the cache performance improve from additional cache space, or by adding more cache servers to the pool? Since resources come at a cost, to what extent would request latencies deteriorate if cache memory were repurposed for a different service?

In this talk, I'll focus on some of the latest research questions pertaining to scalable data replication and large-scale distributed caches. In particular, I'll home in on the challenge of providing online monitoring of the cost and benefits of memory space in a large-scale cache, enabling cache operators to answer the questions above without requiring extraneous trace collection and manual offline tuning. I will introduce general and efficient algorithms for dynamically estimating hit rate curves – histograms of cache hit rate as a function of memory size – which can be plugged into cache replacement policies such as LRU.

Extensive simulations on cache benchmarks indicate that these methods provide accurate estimates of hit rate at different cache sizes. Experiments on an implementation of these methods in memcached showed that hit rate curves were dynamically estimated at over 98

Monday, February 24, 2014, 4:00 pm
Mathematics and Science Center: W303

Mathematics and Computer Science
Emory University