CMP SCI 635: Modern Computer Architecture

Profiling a warehouse-scale computer

Name: Kunjal Panchal Date: 5th Dec, 2019 Student ID: 32126469

Paper: 23 – Data Centers [Kanev15]

Strengths:

1. WSC applications are well-known for their long tails and flat execution profiles, which are best addressed with scalable automatic optimization over many code locations.

2. The instruction footprint of WSC workloads in particular is often over 100 times larger than the size of a L1 instruction cache (i-cache) and can easily overwhelm it. This paper shows that it is expanding at rates of over 20% per year. This results in instruction cache miss rates which are orders of magnitude higher than the worst cases in desktop-class applications, commonly represented by SPEC CPU benchmarks.

Weaknesses:

1. The paper mentions the metric to be used for the case when only profiling data is available, but which metric should be used when evaluating schedulers for warehouse-scale (cloud) clusters, which have machines of different types and sizes, heterogeneous workloads with dependencies and constraints on task placement, and long-running services that consume a large fraction of the total resources?

2. We do not have to think about the cloud as it doing the same old datacenter workflow in a new UI in someone else's datacenter. We CAN do that, but that's not where the cloud excels. It is not about how to maximize server uptime and think about application uptime. We should be trying to consume platform services first and doing everything we reasonably can to avoid deploying a server we can log into. It's not about hosting VMs better, it's about hosting applications better.

Questions/Assertions:

1. Google built a model for offline what-if explorations based on collected far-memory traces, that can model one week of an entire WSCs far memory behaviour in less than an hour. The best parameter configuration found by this process is periodically deployed to the WSC with a carefully monitored phased rollout. The big advantage of the ML based approach is that it can continuously adapt to changes in the workload and WSC configuration without needing constant manual tuning.

2. Although an edge-case, the only thing the cloud is bad for is when we need really low latency access to a large dataset. An app that was fully bottlenecked by read i/o per second (to a dataset too large to fit in memory). A $500 box with a cheap local SSD would be orders of magnitude faster than the biggest EC2 instance. AWS has been great for everything else.

3. Cache partitioning in scaled cloud application may avoid the bottleneck but should we consider the issue of fairness in case of cloud or warehouse?

4. WSC servers typically execute thousands of different applications, so the kernels that matter most across the fleet (the "datacenter tax") may not be significant for a single workload, and are easy to overlook in application-by-application investigations.

5. Data centre scale production environments are concerned with not only the average case but also tail performance. Making the study of the impact of running at-scale inferences for recommendation systems in production-environments, an important topic.