IEEE International Symposium on Workload Characterization – 2022

Title: Performance Engineering in the Public Cloud

Author: Padma Apparao, Intel Corporation

Abstract:

The cloud has indeed been a blessing for many application developers and innovators. Ability to consume compute easily, from anywhere, elastically scale it, plus have state-of-art hardware and software technologies to choose from has democratized technology evolution and cloud growth. Enterprises are accelerating their adoption of cloud, with the various public clouds, AWS, Azure, GCP, Ali, Tencent all propositioning service offerings such as IaaS, PaaS and SaaS. The workloads that are being migrated to the cloud, and the workloads that are born in the cloud have a choice of cloud, their compute, and service offerings. While there are challenges and solutions all the way from deployment of the workloads to the cloud, to the functionality and delivering to the expected performance, in this tutorial we will address only the challenges one faces when the performance in the cloud does not match the expected performance. We will discuss and help the audience understand how one can evaluate the performance in the cloud, and how the performance data can give insights into the underlying hardware architecture.

Workload characterization in the cloud is not significantly different from that on bare-metal in an on-prem datacenter. The techniques to evaluate workload sensitivity to compute, memory, IO and Storage is the same and the way a workload interacts with the underlying CPU, adjacent platform technologies of memory, storage, network has not changed. These fundamentals will continue to remain important and must be characterized for the new consumption model of the cloud compute. How workloads leverage the underlying CPU instructions like vector operations, SIMD, AESNI etc. sensitivity of the workloads to memory latencies, bandwidth is also needed but these are harder to characterize as the CSP vendors abstract the HW information. There are additional challenges like determining the right cloud instance to land a workload on, and some not so obvious issues that the cloud presents like the ‘noisy neighbor’ interference problem.

To address these challenges and the needs of workload performance characterization for the cloud we need a process, methodology and tools. In this talk, we will address the best practices for deployment of workloads, the processes of benchmarking to ensure consistency and repeatability of metrics and finally the tools needed to debug and analyze the workload to gain insights into the usage of the underlying hardware and software architectures.

The process of deploying the workload and ensuring repeatability and consistency is handled by tools such as Terraform, Ansible and Intel DC-Perf-Kit Benchmark Harness. From a methodology perspective we first choose proxy workloads such as HammerDB-OLTP to represent TPC-C, TPCDS to represent BigData workloads, secondly, we analyze these workloads with respect to core, cache, memory, storage, and network sensitivity to remove any application level bottlenecks. Performance debug tools come in many flavors such as OS based tools turbostat, vmstat, perfmon and Intel v-Tune. Additionally, architectural

tools such as eMon, perf, pcm that give information about the HW events, although limitedly available on CSP and their instances.

In this tutorial we will discuss some examples and results of NUMA node sensitivity analysis, removing CPU usage limitations, scaling up vs. scaling out of workload, benefits of the Intel QAT engine for crypto acceleration, benefits of Intel DL-Boost technology for deep-learning inferencing and other novel accelerated technologies help improve workload/use case performance.

What matters a lot for customers is to apply the best practices accrued by the process of benchmarking into their cloud solutions. While static collection of telemetry data is useful, there is always a need to collect runtime data via Grafana and Prometheus which are observability tools. We will show some data from these tools and educate the audience on how the data can help identify performance bottlenecks in the workloads when deployed in the public cloud.