Abstract:

Serverless computing in the cloud, or functions as a service (FaaS), poses new and unique systems design challenges. While the automated resource management offered by serverless computing simplifies many aspects of developing and operating applications for users, several novel features of serverless complicates its automated resource management for cloud providers. One outcome of design decisions in the automated resource management of serverless platforms is execution performance variability.

In this paper, we conduct one of the first detailed in situ measurement studies of performance variability in AWS Lambda, a leading FaaS platform. We utilize novel measurement techniques to characterize the performance isolation and stability of multiple resource types utilized by serverless functions. We find that these variations exist at many levels: between physical servers within regions, across time as diurnal patterns, and across different regions. We also find that these performance variances are stable on short timescales, and they can be used to successfully model and predict the near-future performance of functions.

We then design and evaluate an end-to-end system that takes advantage of this resource variability to exploit the FaaS consumption based pricing model, in which functions are charged based on their fine-grain execution time rather than actual low-level resource consumption. By using both light-weight resource probing and function execution times to identify attractive servers in serverless platforms, customers of FaaS services can cause their functions to execute on better performing servers and realize a cost savings of up to 10% in the same AWS region.

Textbook:

Operating Systems: Three Easy Pieces
by Remzi H Arpaci-Dusseau (Author), Andrea C Arpaci-Dusseau (Author)

https://www.amazon.com/Operating-Systems-Three-Easy-Pieces/dp/198508659X/ref=sr_1_1?crid=DDDW6Q0WLPE0&keywords=operating+systems+three+easy+pieces&qid=1582920460&sprefix=operating+systems+%2Caps%2C147&sr=8-1

Papers:

1)

Title: Dominant Resource Fairness: Fair Allocation of Multiple Resource Types
Link: Dominant Resource Fairness: Fair Allocation of Multiple Resource Types

Brief Description:

This work proposes a new resource allocation strategy titled “Dominant Resource Fairness”, which is a generalization of max-min fairness to multiple resources.

2)

Title: Peeking Behind the Curtains of Serverless Platforms
Link: Peeking Behind the Curtains of Serverless Platforms
Brief Description:

This work conducted a measurement survey of 3 popular cloud functions platforms, and reverse engineered the underlying software architecture of those platforms. Both cold start and resource allocation information is measured.

3)

Title: Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds
Link: Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds
Brief Description:

This work shows that resource packing strategies can unintentionally introduce new ‘placement gaming’ vulnerabilities, which can allow attackers to co-locate VMs with target VMs, allowing for side channel attacks to be conducted.

4)

Title: My VM is Lighter (and Safer) than your Container
Link: My VM is Lighter (and Safer) than your Container | Proceedings of the 26th Symposium on Operating Systems Principles
Brief Description:

This paper explores the security (isolation) - performance tradeoff between virtual machines and linux containers. The authors show that when VMs are stripped down and optimized heavily, VMs have comparable performance to containers (and even surpass them in several cases).

5)

Title: Architectural Implications of Function-as-a-Service Computing
Link: https://parallel.princeton.edu/papers/micro19-shahrad.pdf
Brief Description:

This work is a measurement study on how microarchitectural components can impact the workloads of cloud function providers. The paper shows that branch predictors and the last level CPU cache can both have significant performance isolation problems with co-tenant workloads in AWS Lambda.

6)

Title: Xen and the Art of Virtualization
Link: Xen and the Art of Virtualization
Brief Description:

This paper describes the design and implementation of the Xen hypervisor, as well as the concept of paravirtualization, which is introduced in this paper.

7)

Title: More for your money: exploiting performance heterogeneity in public clouds
Link: https://dl.acm.org/doi/10.1145/2391229.2391249
Brief Description:

This work describes how placement gaming can be utilized to exploit a combination of poor performance isolation and hardware heterogeneity in a public cloud. The authors show that it is possible to exploit VM scheduling algorithms to get better performance for the same cost on Amazon EC2.

8)

Title: Serverless Computing: One Step Forward, Two Steps Back
Link: Serverless Computing: One Step Forward, Two Steps Back
Brief Description:

This paper outlines the benefits and challenges of the new compute model introduced by serverless computing. Three key issues are pinpointed (function lifetime, I/O bottlenecks, and communication via slow storage), and the authors explain both the causes of these issues as well as potential impacted applications.

9)

Title: SAND: Towards High-Performance Serverless Computing
Link: SAND: Towards High-Performance Serverless Computing
Brief Description:

This paper introduces a new sandboxing mechanism for multi-function serverless based applications. The key insight that is leveraged is that functions that are invoked by the same user of a cloud platform that are ‘chained’ together within the context of a larger application require less sandboxing than functions that do not belong to the same application. In addition to providing two tiers of sandboxing, the system also provides an efficient IPC-like communication mechanism for functions contained within the same application to communicate quickly with each other.