Serverless Computing: An Investigation of Factors Influencing Microservice Performance

Wes Lloyd, Shruti Ramesh, Swetha Chinthalapati, Lan Ly, Shrideep Pallickara

April 20, 2018

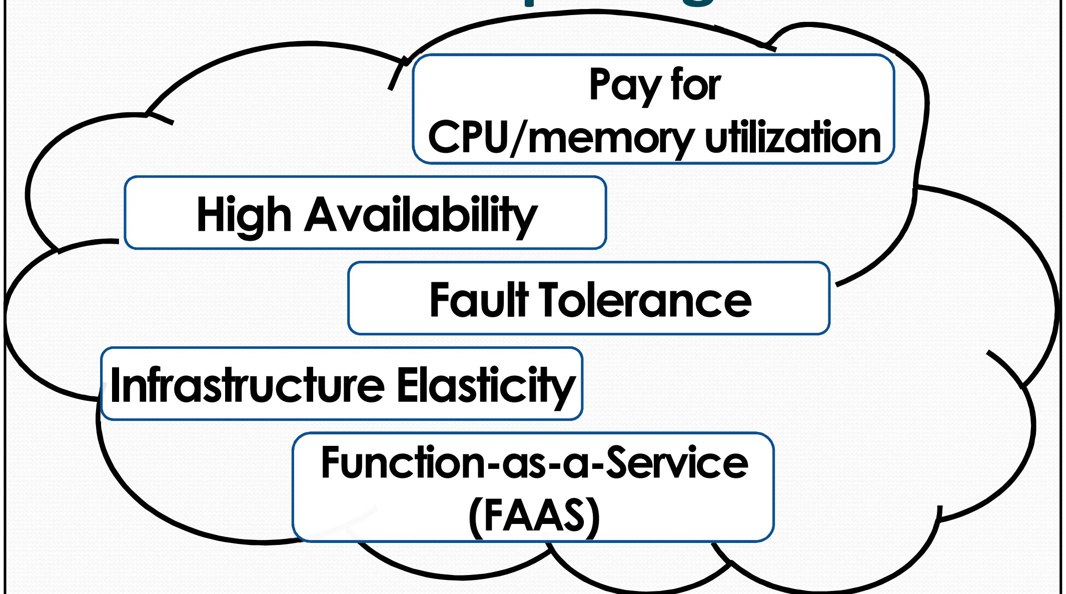
Institute of Technology, University of Washington, Tacoma, Washington USA

IC2E 2018: IEEE International Conference on Cloud Engineering

Outline

- Background
- Research Questions
- Experimental Workloads
- Experiments/Evaluation
- Conclusions

Serverless Computing



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 3

Serverless Computing

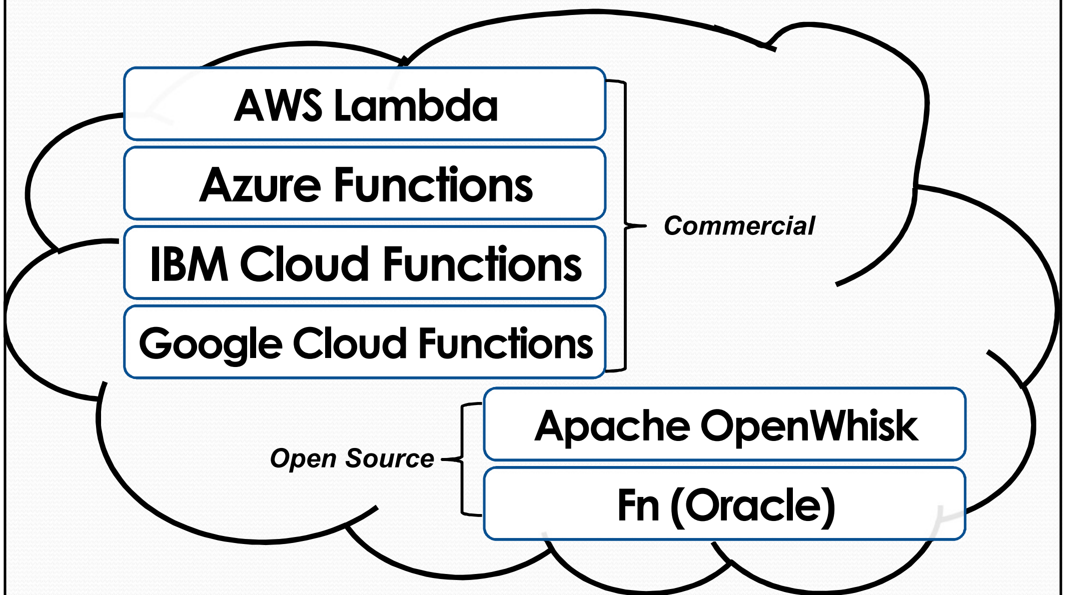
Why Serverless Computing?

Many features of distributed systems, that are challenging to deliver, are provided automatically

…they are built into the platform

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 4

Serverless Platforms



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 5

Serverless Computing

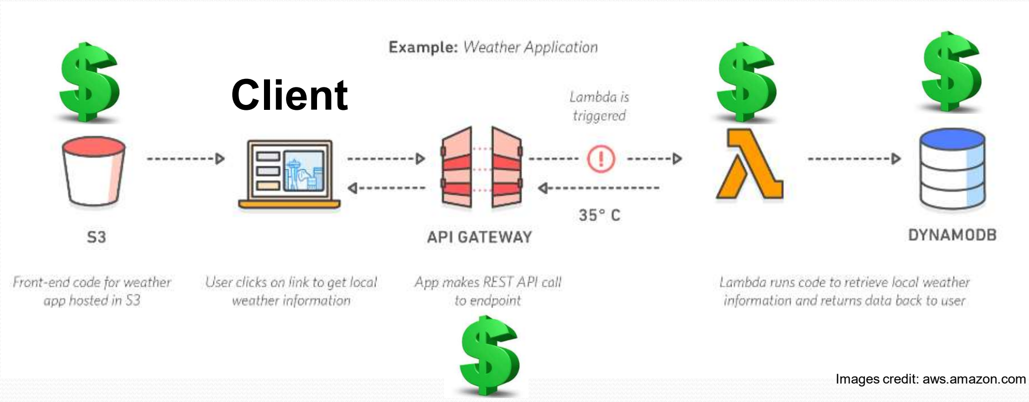
Research Challenges

Image from: https://mobisoftinfotech.com/resources/blog/serverless-computing-deploy-applications-without-fiddling-with-servers/

6

Vendor architectural lock-in

- Serverless software architecture requires external services/components



- Increased dependencies → increased hosting costs

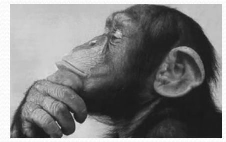
April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 7

Serverless Pricing Model

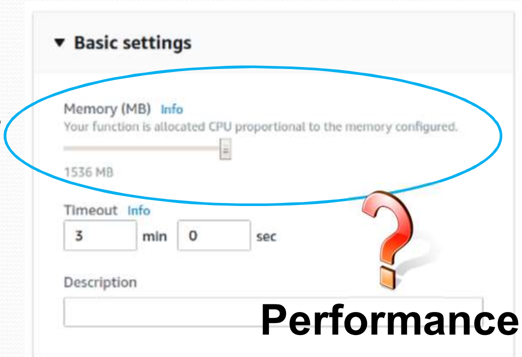
- **<u>EXAMPLE:</u>** AWS Lambda Pricing
- **<u>FREE TIER:</u>** first 1,000,000 function calls/month → FREE
  first 400 GB-sec/month → FREE
- Afterwards: *obfuscated pricing (AWS Lambda):*
  $0.0000002 per request
  $0.000000208 to rent 128MB / 100-ms

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 8

Serverless Computing
Memory reservation question…



- Lambda memory reserved for functions
- UI provides "slider bar" to set function's memory allocation
- CPU power coupled to slider bar: *"every **doubling** of memory, **doubles** CPU…"*
- **But how much memory does code require?**



Performance

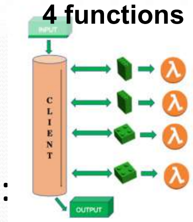
April 20, 2018 Serverless Computing: An Investigation of Factors Influencing Microservice Performance 9

Service Composition

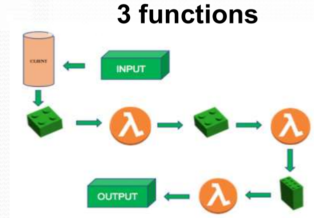
- **How should application code be composed for deployment to serverless computing platforms?**

Monolithic

Client flow control, 4 functions



Server flow control, 3 functions



- Recommended practice: Decompose code into many microservices
- Platform limits: code + libraries ~256MB
- **How does composition impact number of invocations, and memory utilization?**

Performance

April 20, 2018 Serverless Computing: An Investigation of Factors Influencing Microservice Performance 10

Freeze/Thaw Cycle

Performance

- Unused infrastructure is deprecated
  - ***But after how long?***
- Infrastructure: VMs, "containers"
- <u>**Provider-COLD / VM-COLD**</u>
  - "Container" images - built/transferred to VMs
- <u>**Container-COLD**</u>
  - Image cached on VM
- <u>**Container-WARM**</u>
  - "Container" running on VM

Image from: Denver7 – The Denver Channel News

Serverless Computing Research Challenges

- Vendor architectural lock-in
- Pricing obfuscation
- Memory reservation
- Service composition
- Infrastructure freeze/thaw cycle

Outline

- Background
- Research Questions
- Experimental Workloads
- Experiments/Evaluation
- Conclusions

Research Questions

RQ1: What are the performance implications of infrastructure ***elasticity*** for serverless computing?
(e.g. COLD vs. WARM performance)

RQ2: How does ***load balancing*** vary in serverless computing? How do computational requests impact load balancing, and ultimately performance?

Research Questions - 2

RQ3: What performance implications result from ***provisioning variation*** of container infrastructure?

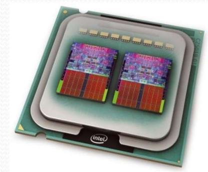
RQ4: What are the impacts on ***infrastructure retention*** based on microservice/function utilization?

RQ5: What performance implications result from microservice ***memory reservation*** size? How does memory reservation size impact container placement?

Outline

- Background
- Research Questions
- Experimental Workloads
- Experiments/Evaluation
- Conclusions

AWS Lambda Compute Bound Test Service



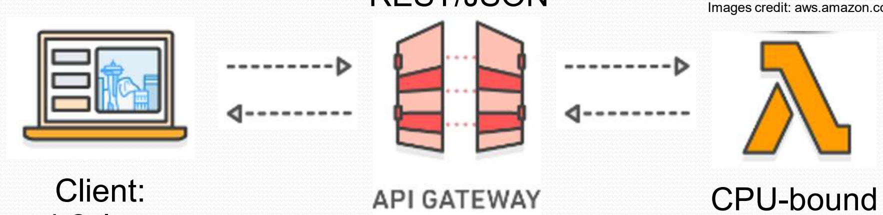
- Increasing stress levels 1 (none) → 9 (high) ***(non-linear)***
- Parameters:
  - Operand array size and number of calculation loops (0, 20, 100, 1,000, 10,000, 25,000, 100,000)
    → Operands stored in random array locations
    → Induces page faults when seeking random locations
  - Number of function calls per loop (0, 20, 1,000, 100,000, 300,000)

- Control CPU time of function as input parameter
- **Goal: observe impact of CPU time on infrastructure scaling, provisioning variation, retention, and service performance**

AWS Lambda Testing

REST/JSON

Images credit: aws.amazon.com



Client:
c4.2xlarge

BASH: GNU Parallel
Multi-thread client
"partest"

Up to 100 concurrent synchronous requests

Results of each thread traced individually

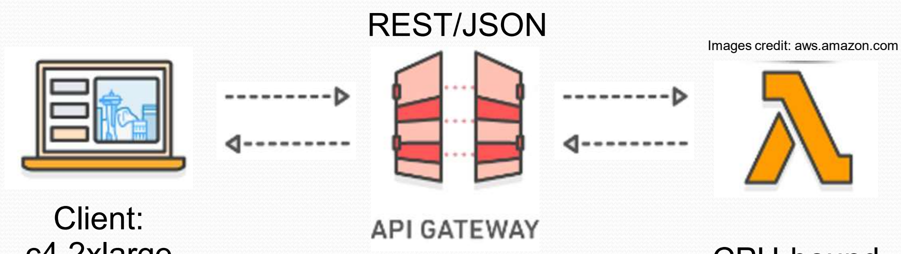
<u>Fixed-availability zone:</u>
EC2 client / Lambda server
us-east-1e

CPU-bound
Test Function

Max
service duration:
< 30 seconds

Memory:
128 to 1536MB

AWS Lambda Testing



Client:
c4.2xlarge

Automatic Metrics Collection:

New vs. Recycled Containers/VMs

# of requests per container/VM

Avg. performance per container/VM

Avg. performance workload

Standard deviation of
requests per container/VM

CPU-bound
Test Function

Container Identification
UUID → /tmp file

VM Identification
btime → /proc/stat

Linux CPU metrics

Azure Functions Testing

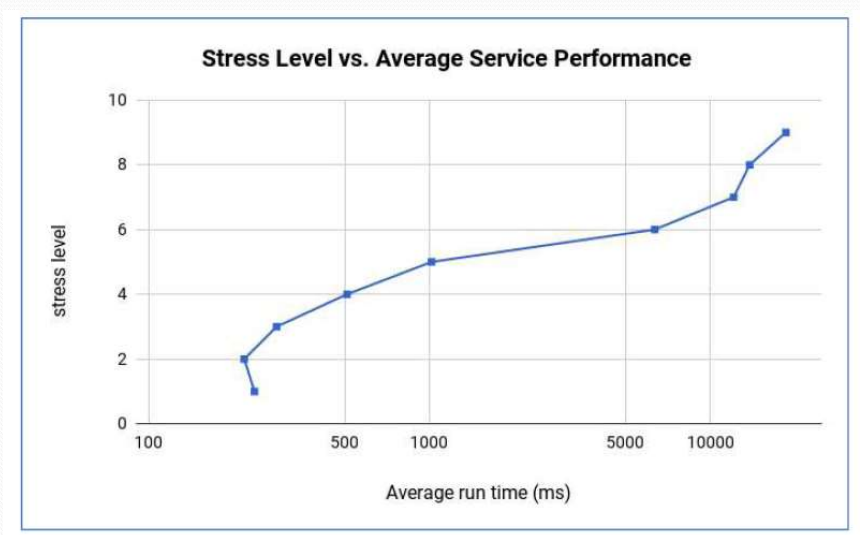
- Http-triggered function app, written in C#
- Logs to Azure Table storage (*similar to Dynamo DB*)
  - Unique app service instance IDs
  - Current worker process ID
- Consumption plan → auto-scaled infrastructure
  - vs. app service plan (*deployment to dedicated VMs*)
- Performance testing:
  Visual Studio Team System (VSTS)

Outline

- Background
- Research Questions
- Experimental Workloads
- Experiments/Evaluation
- Conclusions

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 21

CPU-Bound Lambda Test Service WARM Performance



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 22

RQ-1: Elasticity

What are the performance implications of infrastructure ***elasticity*** for serverless computing?

(e.g. COLD vs. WARM performance)

23

RQ-1: AWS Lambda Latency Evaluation

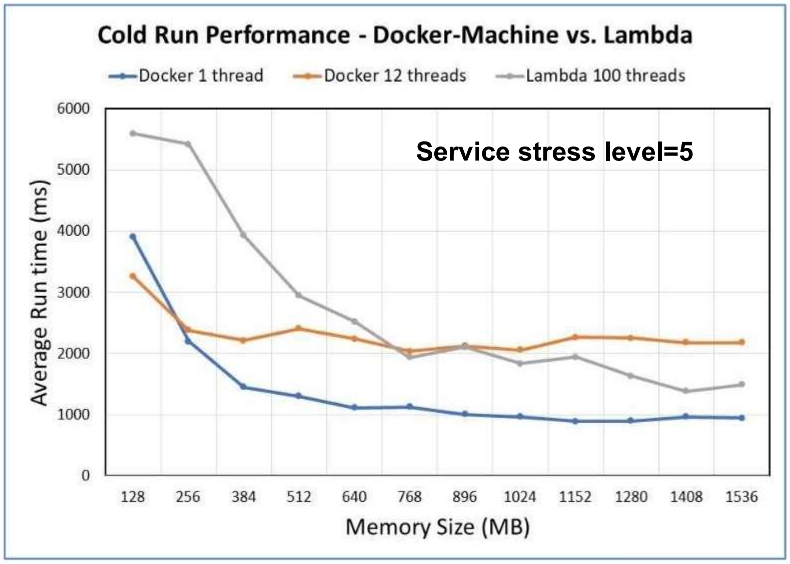
- **AWS Lambda Simulation**
- Harness c4.8xlarge 36 vCPU VM instance
  - Intel Xeon E5-2666v3 CPU – *same as Lambda*
- Lambda JAR file deployed Docker container(s)
  - **Set memory**: `docker run "-m <ram in MB>"`
  - **Set CPUs**: `docker run "—cpus <VCPUs>`
- Compare: 1 and 12 concurrent runs
  - Avg VM tenancy ~12.3 of all tests
- **How does Lambda scale CPU power?**

Literal Estimates:

Memory (MB)	Expected CPU%
128	16.6%
256	33.3%
384	50.0%
512	66.6%
640	83.3%
768	100.0%
896	116.7%
1024	133.3%
1152	150.0%
1280	166.60%
1408	183.30%
1536	200.00%

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 24

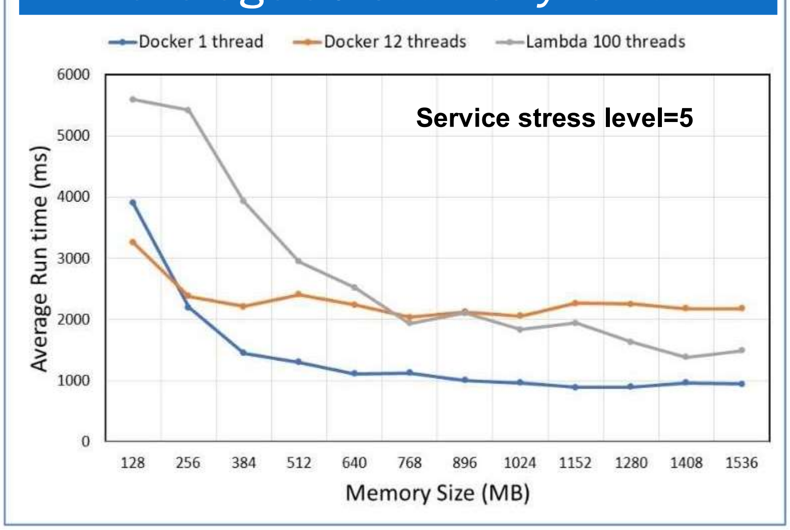
RQ-1: EC2/Docker vs. Lambda Performance Intel Xeon E5-2666 v3 - COLD



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 25

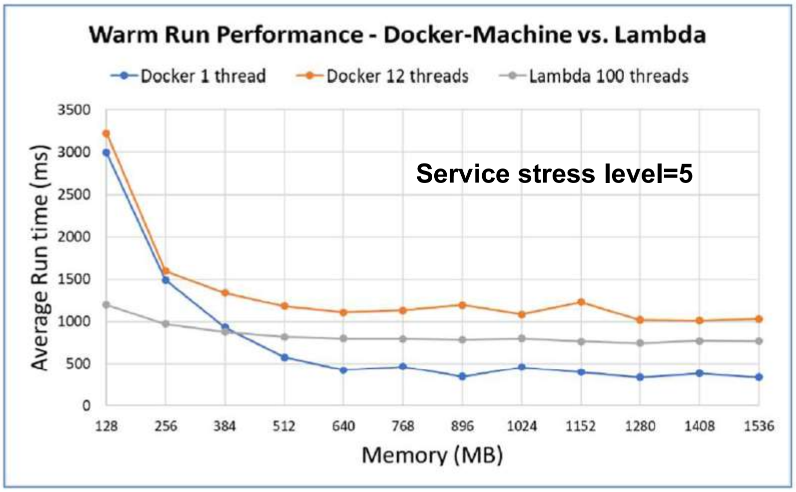
RQ-1 ... ance Intel ...

Assumed tenancy of ~12 service requests per container for Lambda: *average across many tests*

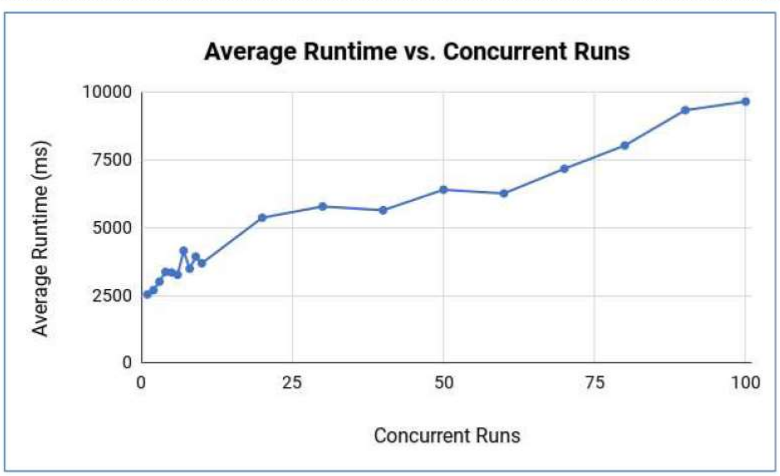


April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 26

RQ-1: EC2/Docker vs. Lambda Performance Intel Xeon E5-2666 v3 - WARM



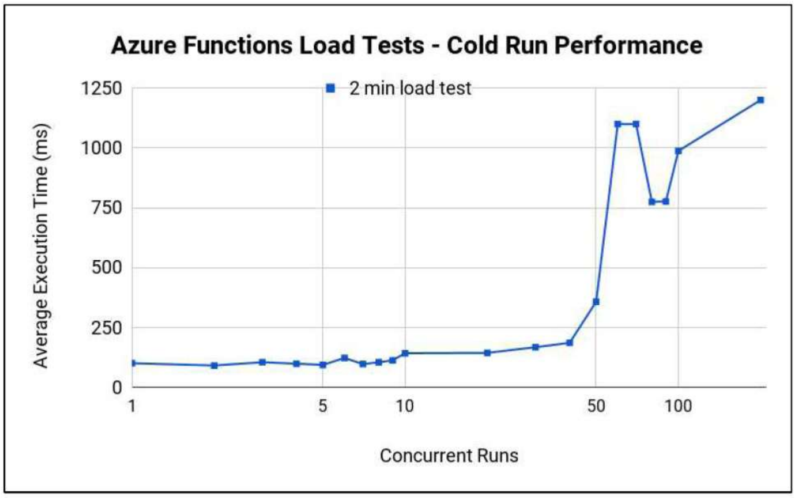
RQ-1: AWS Lambda SCALE UP Performance



Service stress level=4

RQ-1: Azure functions COLD Performance

includes "container" initialization



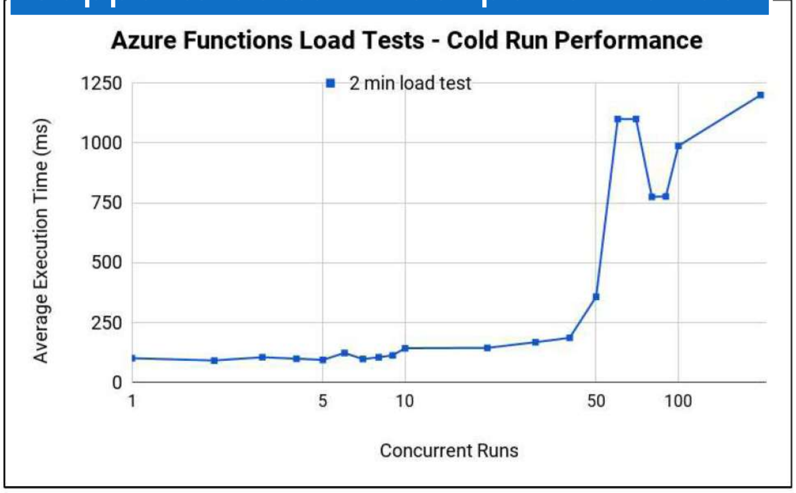
Up to 4 VMs automatically created

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 29

RQ-1 ... ce

incluc

VMs are allocated as opposed to individual container instances. Supports better initial performance.



Up to 4 VMs automatically created

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 30

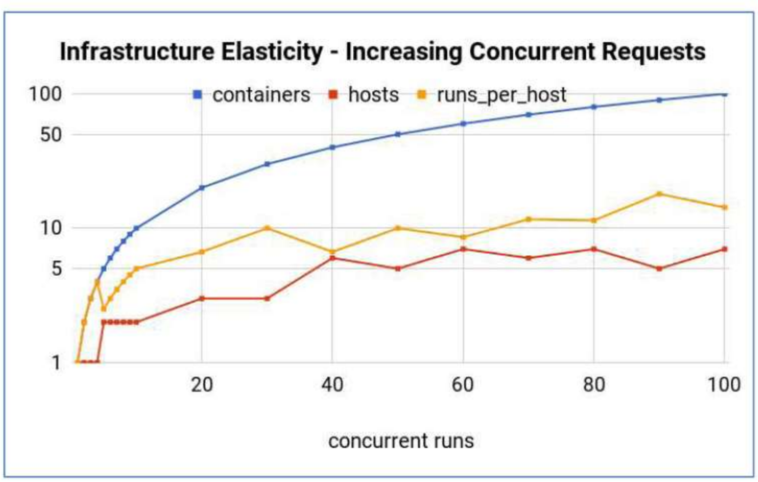
RQ-2: Load Balancing

How does ***load balancing*** vary in serverless computing?

How do computational requests impact load balancing, and ultimately performance?

31

RQ-2: COLD Lambda Infrastructure for Scaling

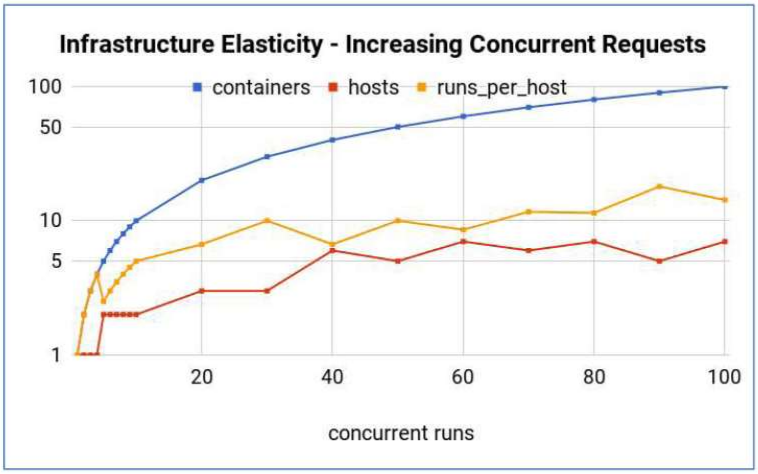


Service stress level=4

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 32

RQ… Infr…

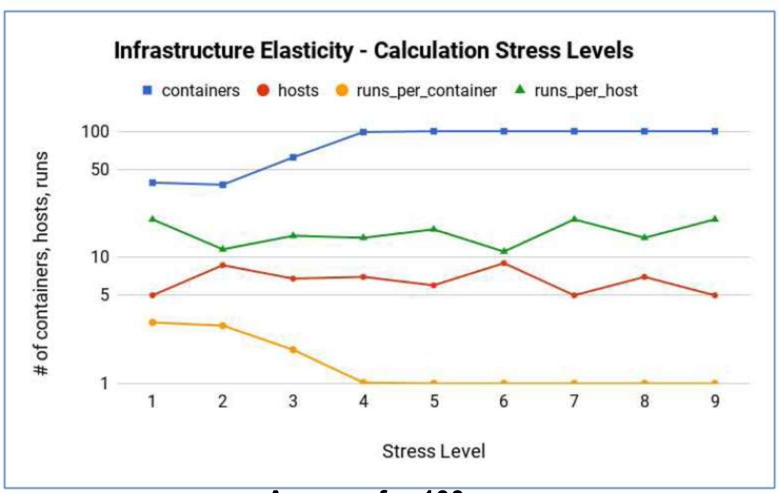
COLD service requests receive separate container instances to amortize startup overhead



Service stress level=4

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 33

RQ-2: WARM Lambda Infrastructure for Scaling

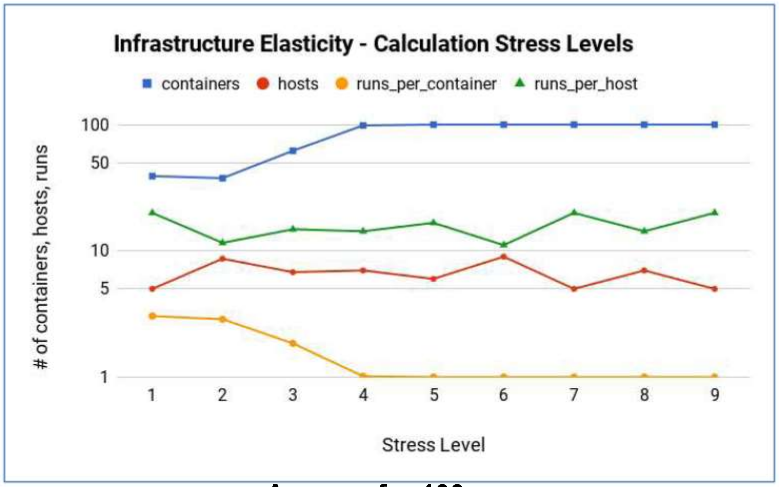


Average for 100 runs

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 34

RQ... Infr...

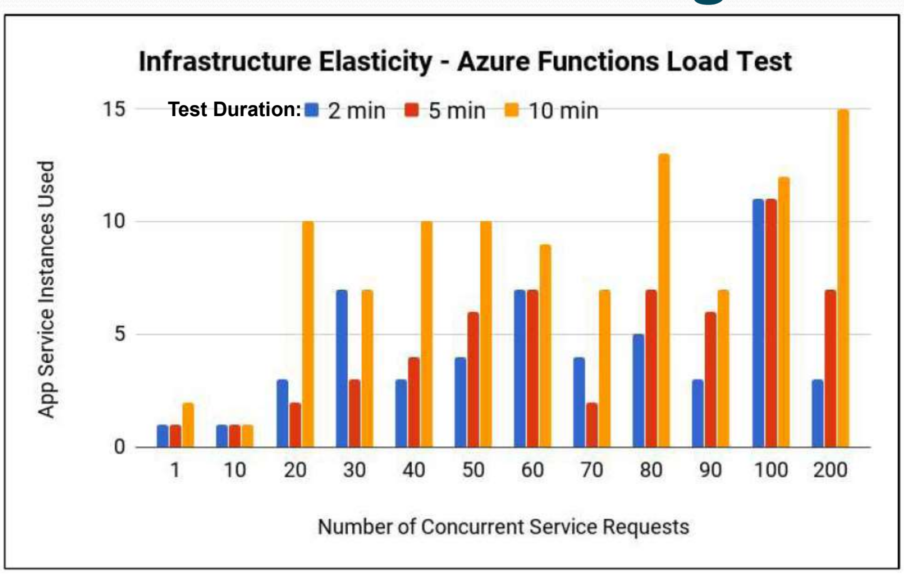
WARM service requests share container instances unless CPU requirements are increased



Average for 100 runs

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 35

RQ-2: COLD Azure Functions Infrastructure for Scaling



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 36

RQ-3: Provisioning Variation

What performance implications result from ***provisioning variation*** of container infrastructure?

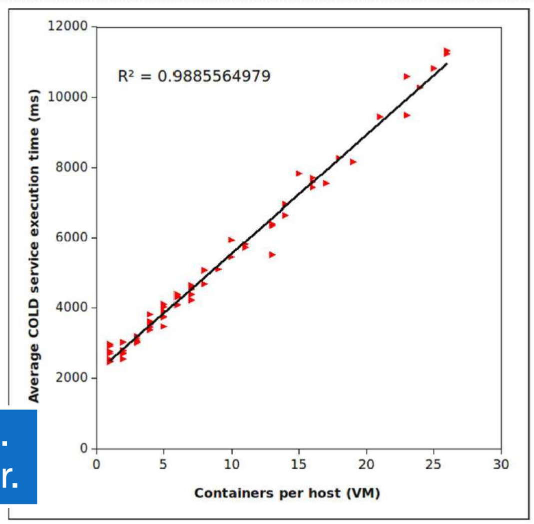
37

RQ-3: Cold Lambda service performance vs. Container Placement

When more containers were placed on the same VMs for COLD service requests, Lambda Performance suffered up to 5x !

The impact of tenancy vs. performance is quite clear.

Service stress level=4



April 20, 2018 Serverless Computing: An Investigation of Factors Influencing Microservice Performance 38

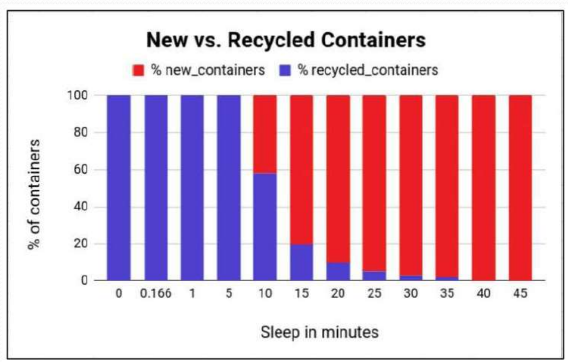
RQ-4: Infrastructure Retention

What are the impacts on ***infrastructure retention*** based on microservice/function utilization?

39

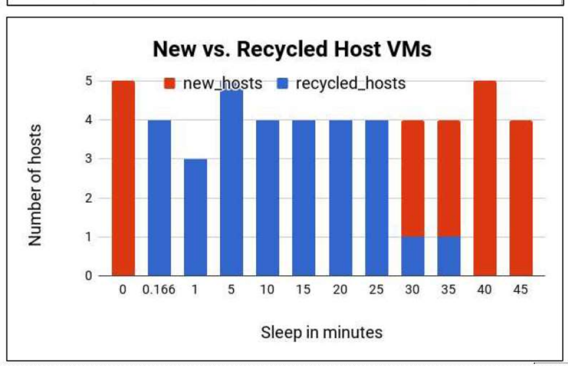
RQ-4:

Lambda Container Recycling



Service stress level=4

Lambda Virtual Machine Recycling



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* **40**

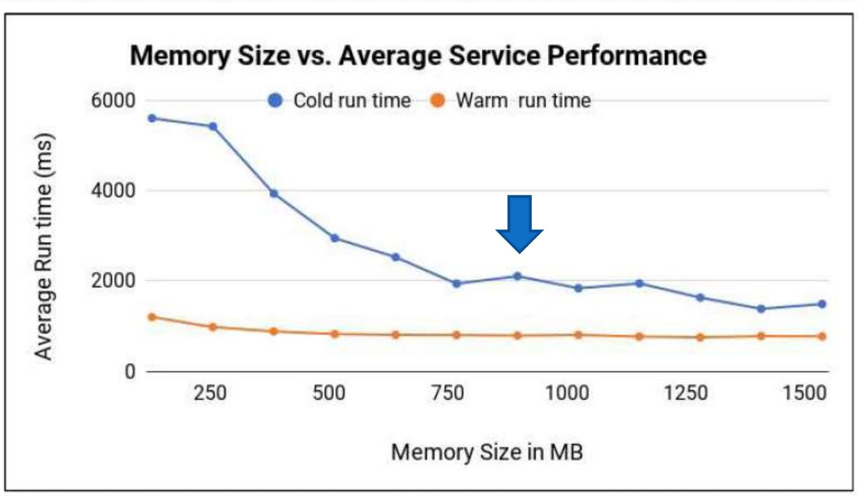
RQ-5: Memory Reservation

What performance implications result from microservice ***memory reservation*** size?

How does memory reservation size impact container placement?

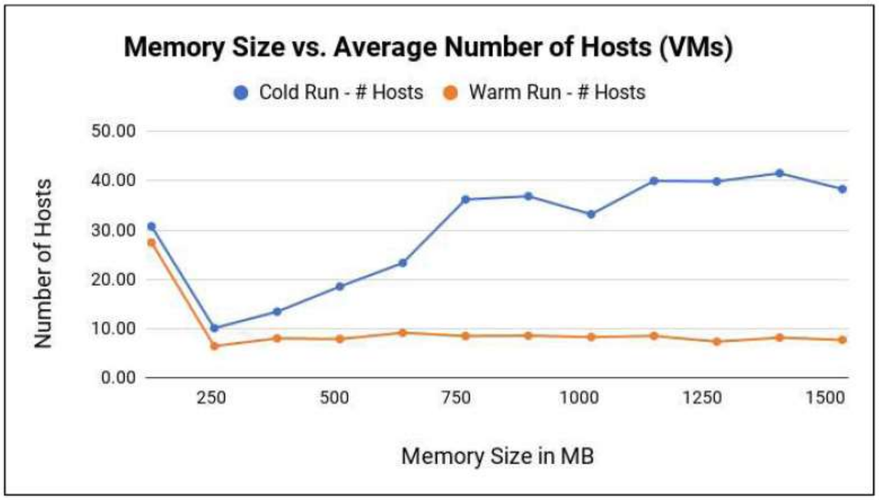
RQ-5: Slider Bar Test: Memory vs. CPU power

Service stress level=4



RQ-5: Slider Bar Test II: Infrastructure vs. Memory Reservation

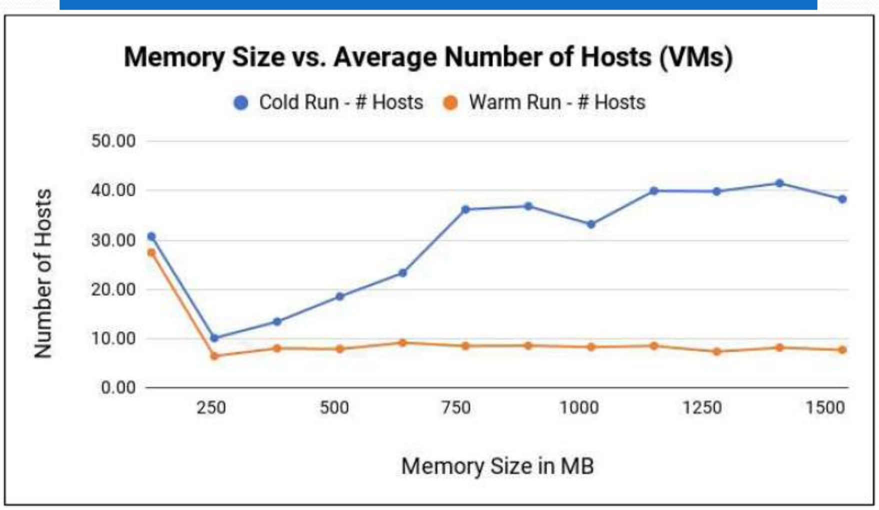
Service stress level=4



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 43

RQ-5
Infra … on

Increasing the memory reservation size results in more hosting infrastructure



April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 44

Outline

- Background
- Research Questions
- Experimental Workloads
- Experiments/Evaluation
- Conclusions

Conclusions

- **RQ-1 Elasticity**: Extra infrastructure is provisioned to compensate for initialization overhead of "container" startup
  - VM COLD: up to ~20x slower than WARM
  - Container COLD: ~5x slower than WARM
- **RQ-2 Load Balancing**: Better when COLD. WARM runs only use all original infrastructure when CPU-bound execution time is similar to container initialization execution time
  - Must increase stress level to harness available infrastructure

Conclusions - 2

- **RQ-3 Provisioning Variation**: Bad placement can lead to ~4.6x degradation in COLD service performance
- **RQ-4 Infrastructure Retention**:
  3 distinct performance states:
  VM COLD, ***Container COLD***, ***WARM***
  - Containers begin to disappear after 10 minutes
  - VM hosts deprecated after ~40 minutes
- **RQ-5 Memory Reservation**:
- For non memory-bound service, performance improves up to ~512-640MB

April 20, 2018 *Serverless Computing: An Investigation of Factors Influencing Microservice Performance* 47

Questions