Towards Demystifying Intra-Function Parallelism in Serverless Computing

7th International Workshop on Serverless Computing (WoSC7)
In conjunction with, ACM/IFIP Middleware 2021

Michael Kiener, Mohak Chadha, Michael Gerndt

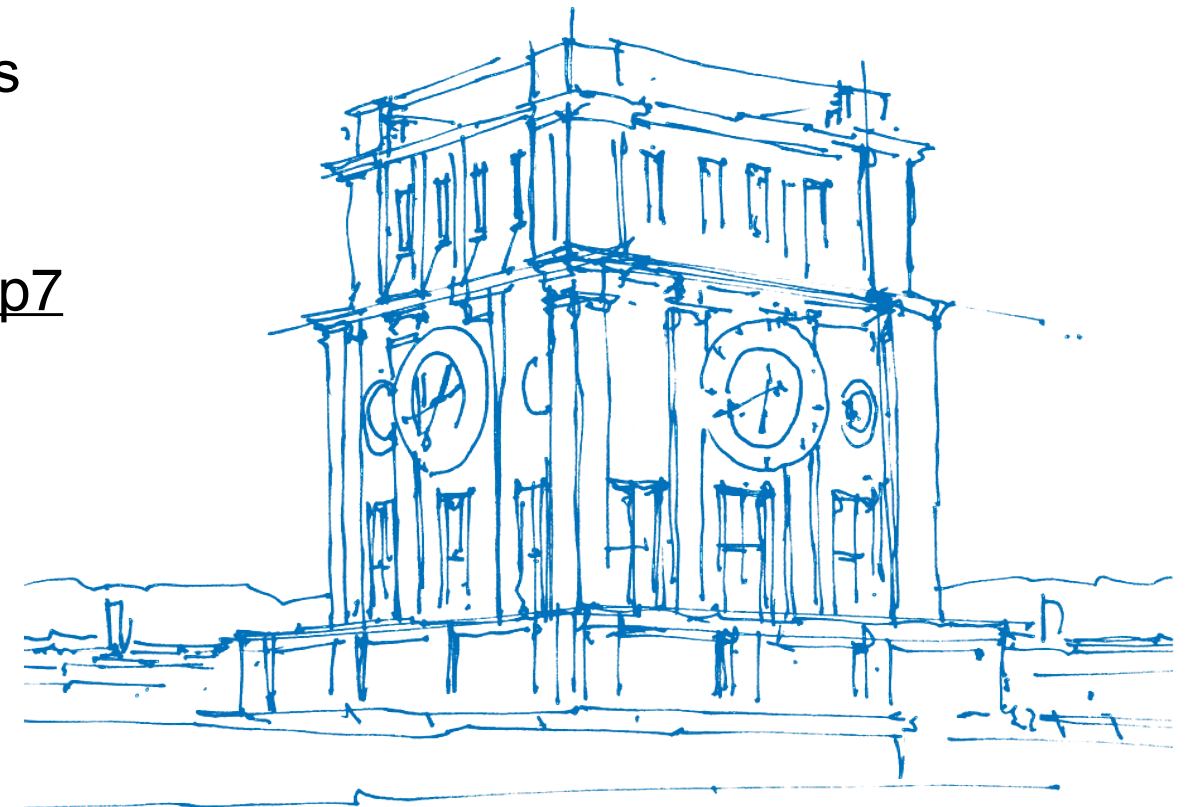michael.kiener@tum.de, mohak.chadha@tum.de, gerndt@in.tum.de

Technische University of Munich (TUM)

Chair of Computer Architecture and Parallel Systems

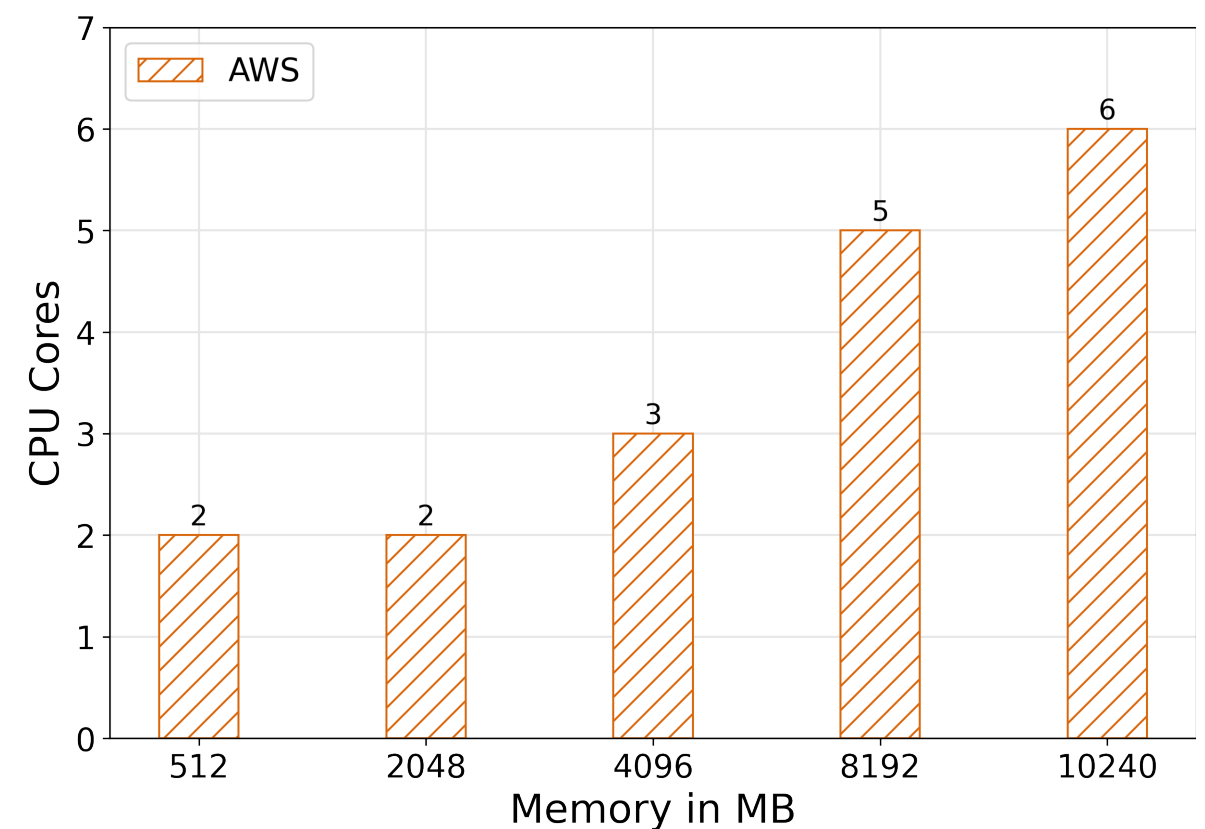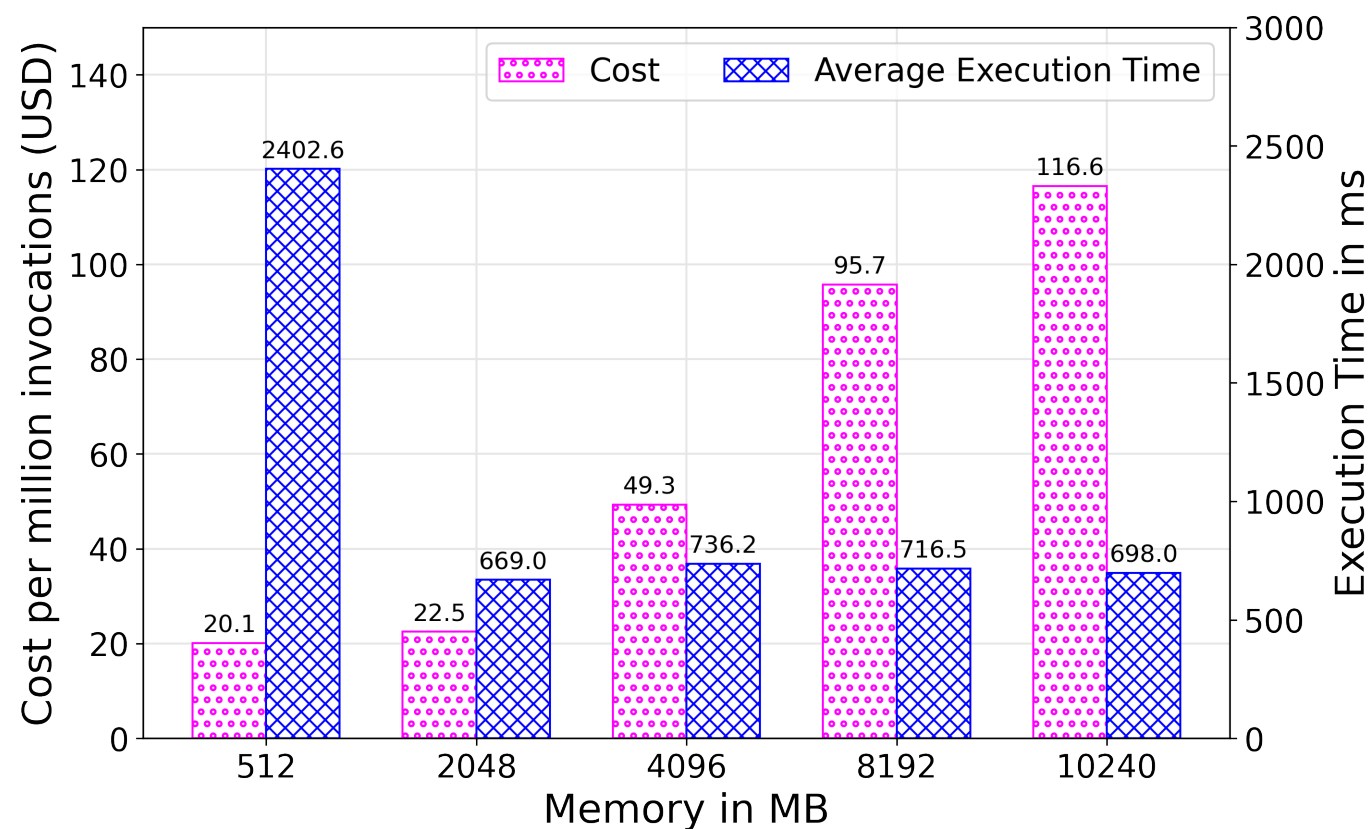Garching (near Munich), Germany
7th of November, 2021
https://www.serverlesscomputing.org/wosc7/papers/p7

TUM Uhrenturm

Outline

Motivation

- Serverless services function like black-boxes
- Developer doesn't have full insights about underlying hardware & infrastructure
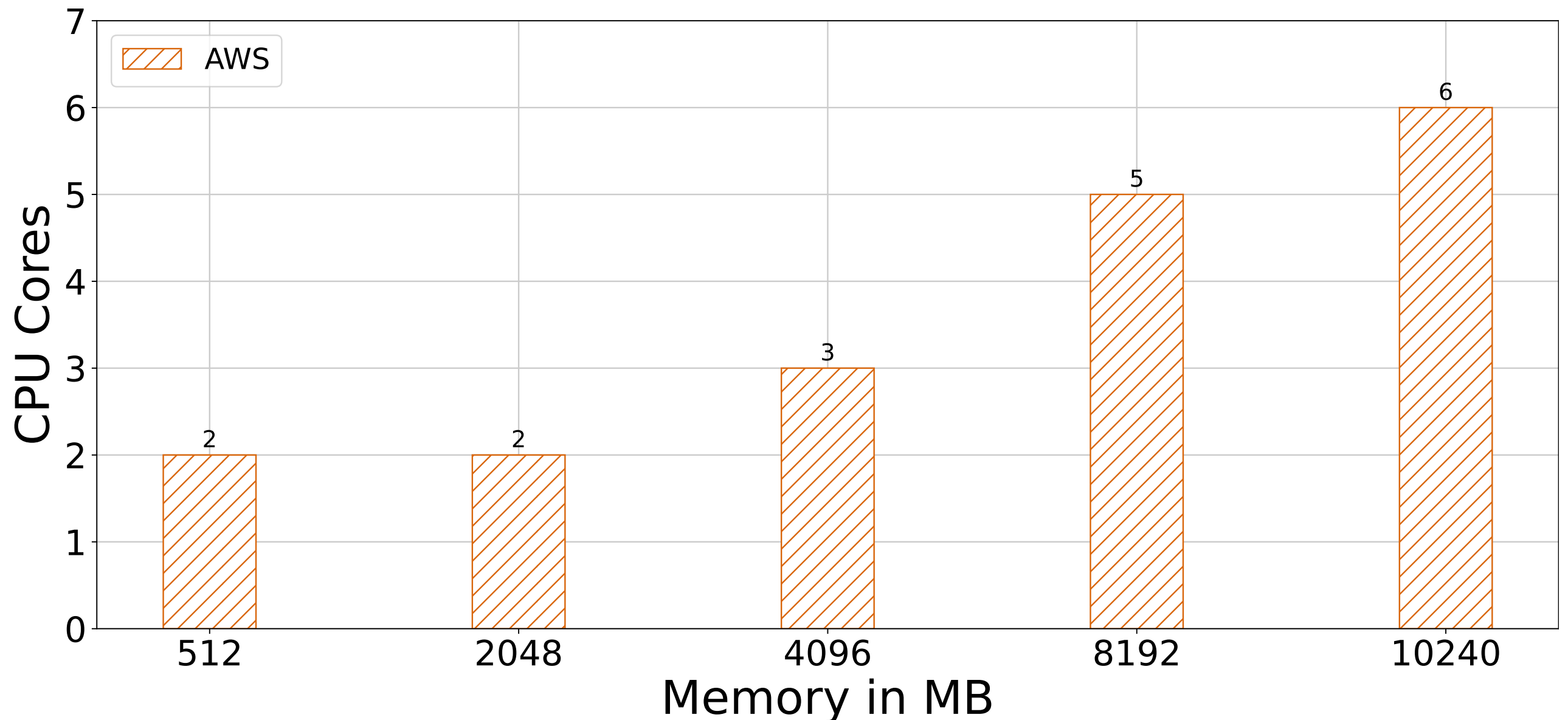- Underutilization of computing resources increases costs significantly

Methodology

- 3 Microbenchmarks
    - Atax
    - Go fast
    - MVT
- 2 Applications
    - Heat
    - Monte Carlo

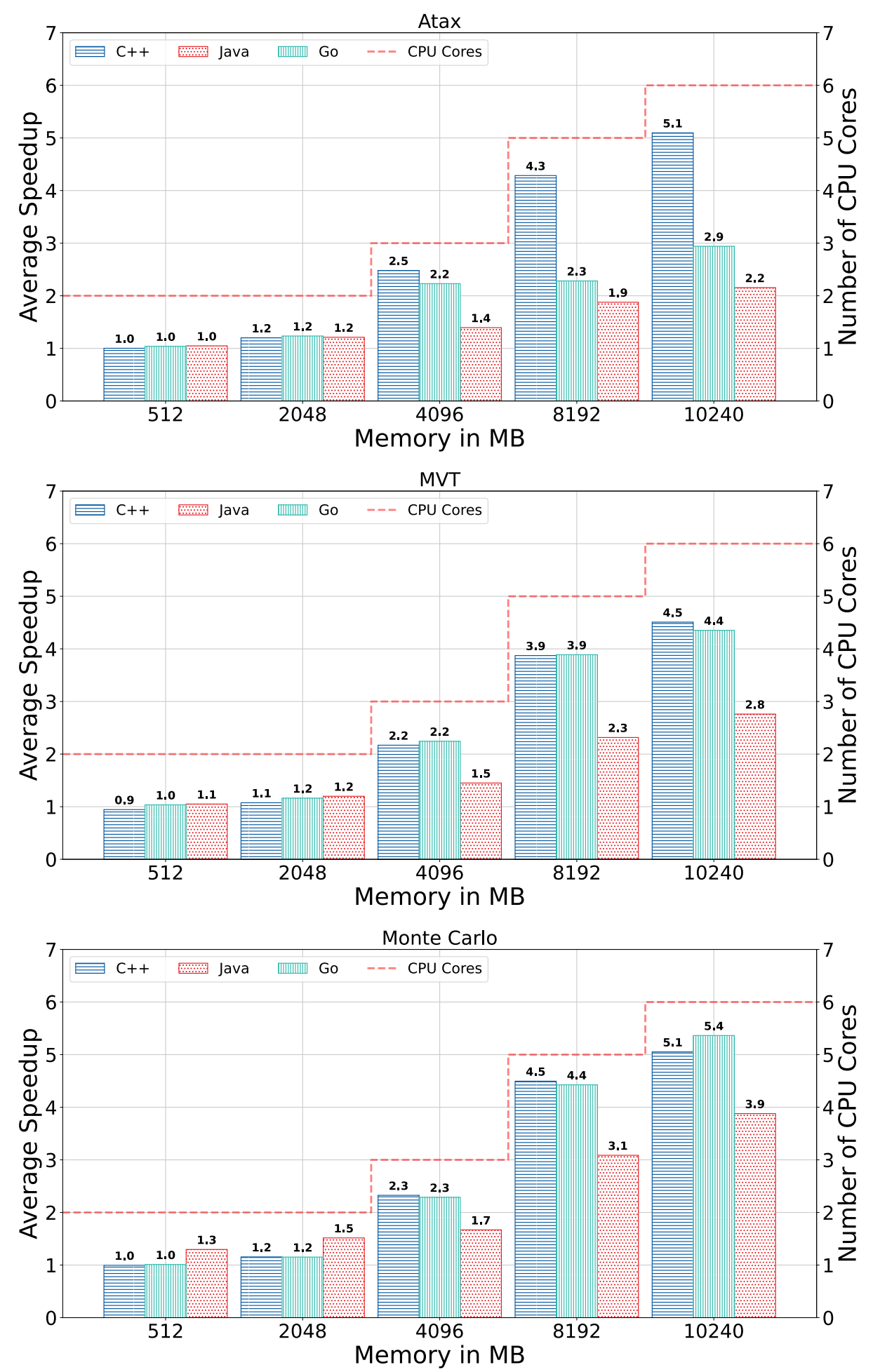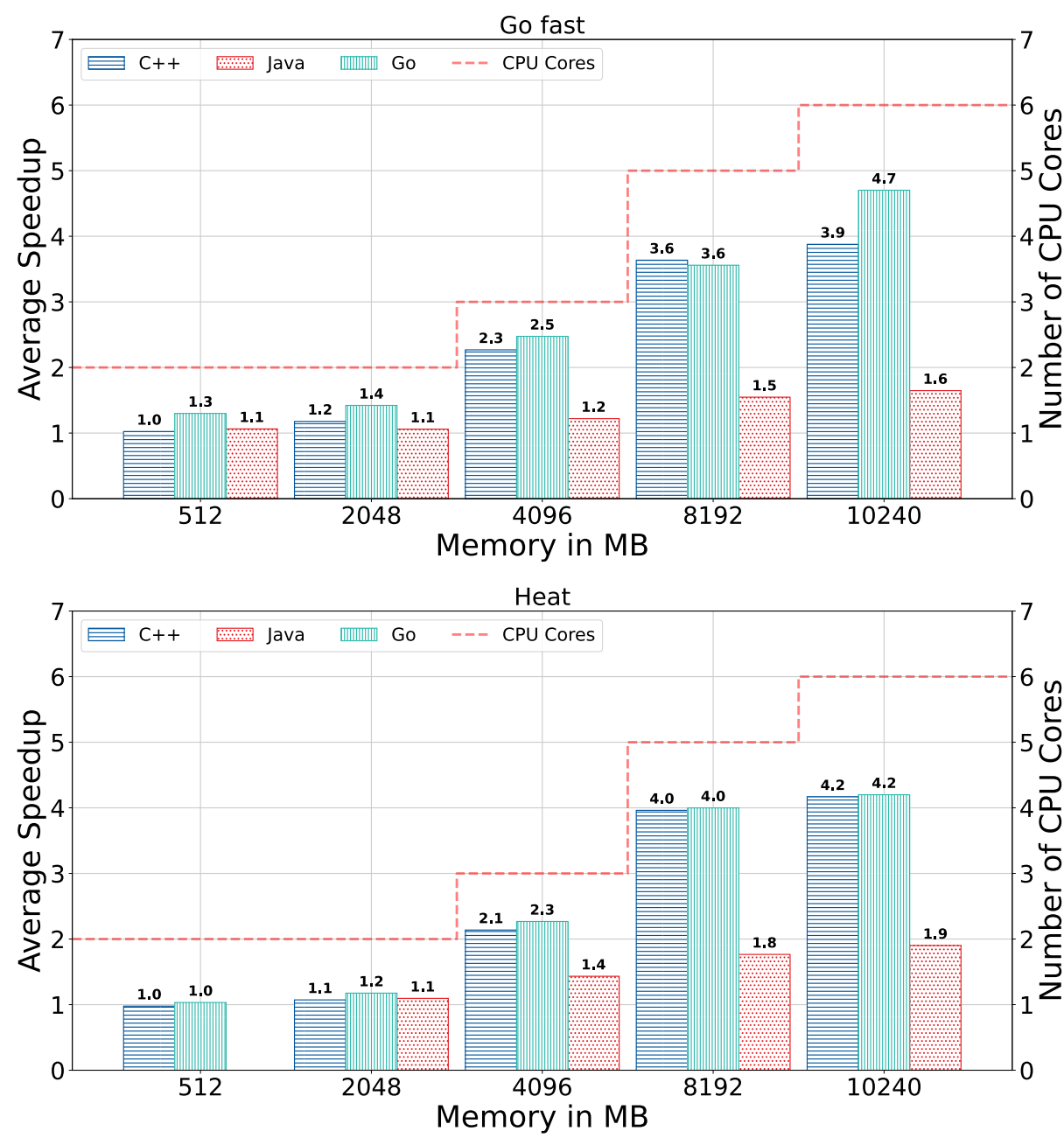| Language	Parallelization	Version	Compiler & Flags
C++	OpenMP	C++11	g++ -O3
Java	ExecutorService	Java 11	OpenJDK 11
Go	Goroutines	1.16	gc GOOS=linux

Results - Cores to vCPU Mapping
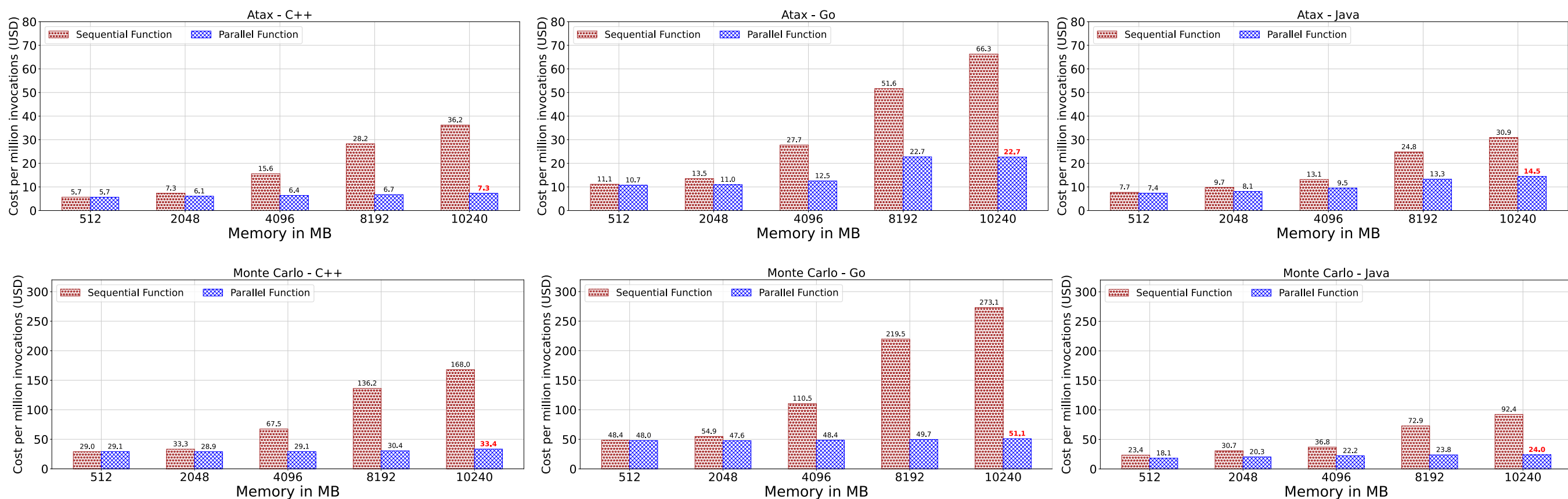
CPU Cores Allocation

Results - Performance
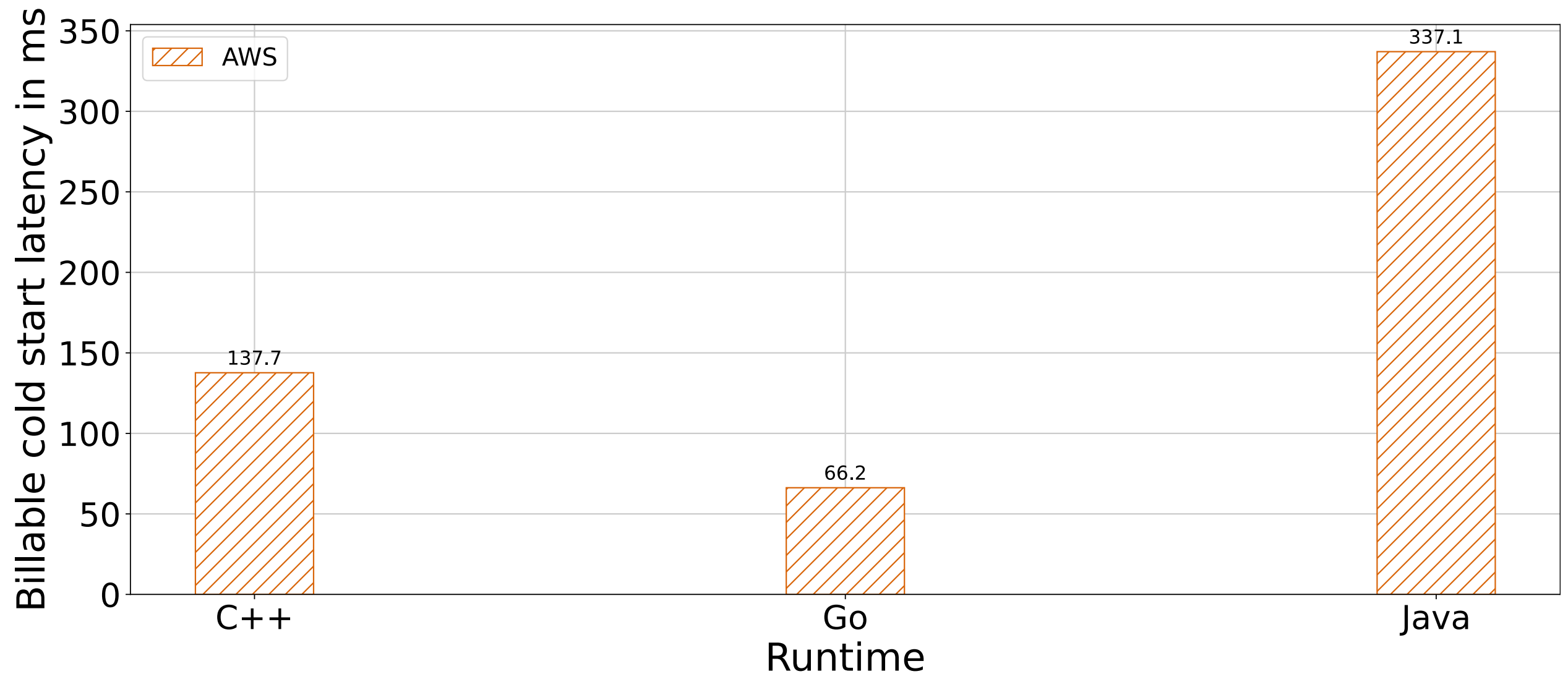
Parallel Speedups for AWS Lambda

Results - Costs

Cost Comparison for AWS Lambda

Language	Benchmark	Configuration	Maximum Cost Savings
C++	Monte Carlo	10240MB	80,1 %
Java	Monte Carlo	10240MB	74,1 %
Go	Monte Carlo	10240MB	81,3 %

Cold starts

Impact of billable time difference of cold starts

Conclusion

- Allocated CPUs not always equal to vCPUs

- Parallelization efficiency heavily depends on service and configuration

- Maximum cost savings achieved by parallelizing
    - AWS: 81%

Thank you for your attention

Questions?