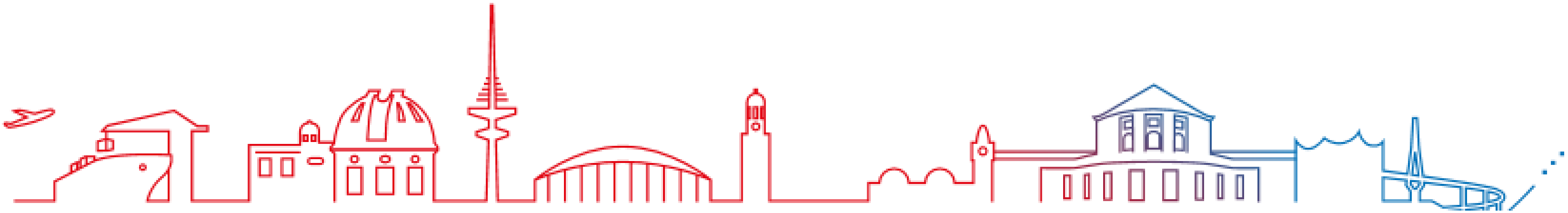
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Evaluating the Energy Efficiency of Cloud-Native Applications and Platforms

26.10.202
4

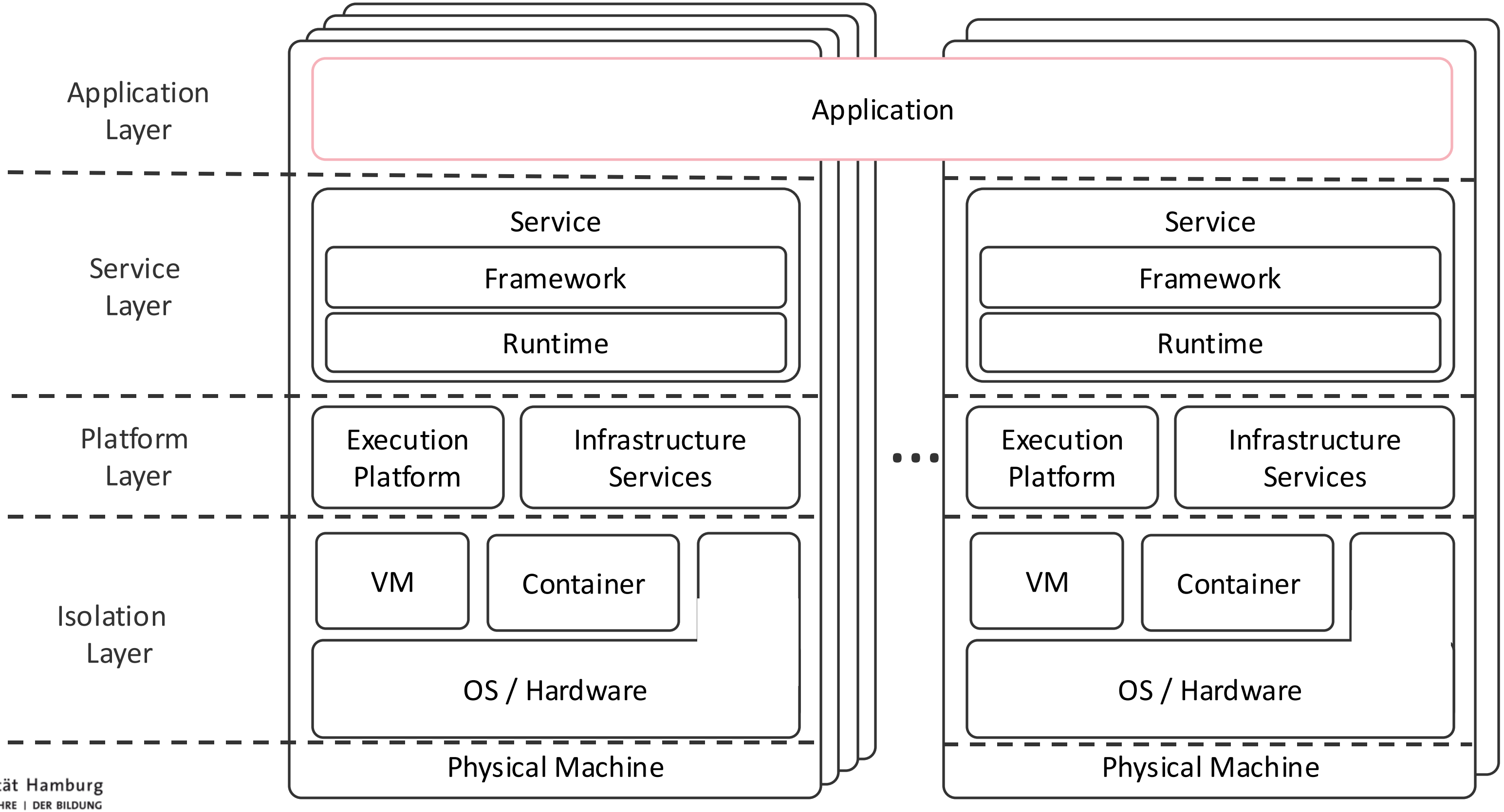
Sebastian Werner (Uni Hamburg & TU Berlin)

Energy Efficiency of Cloud-Native Applications

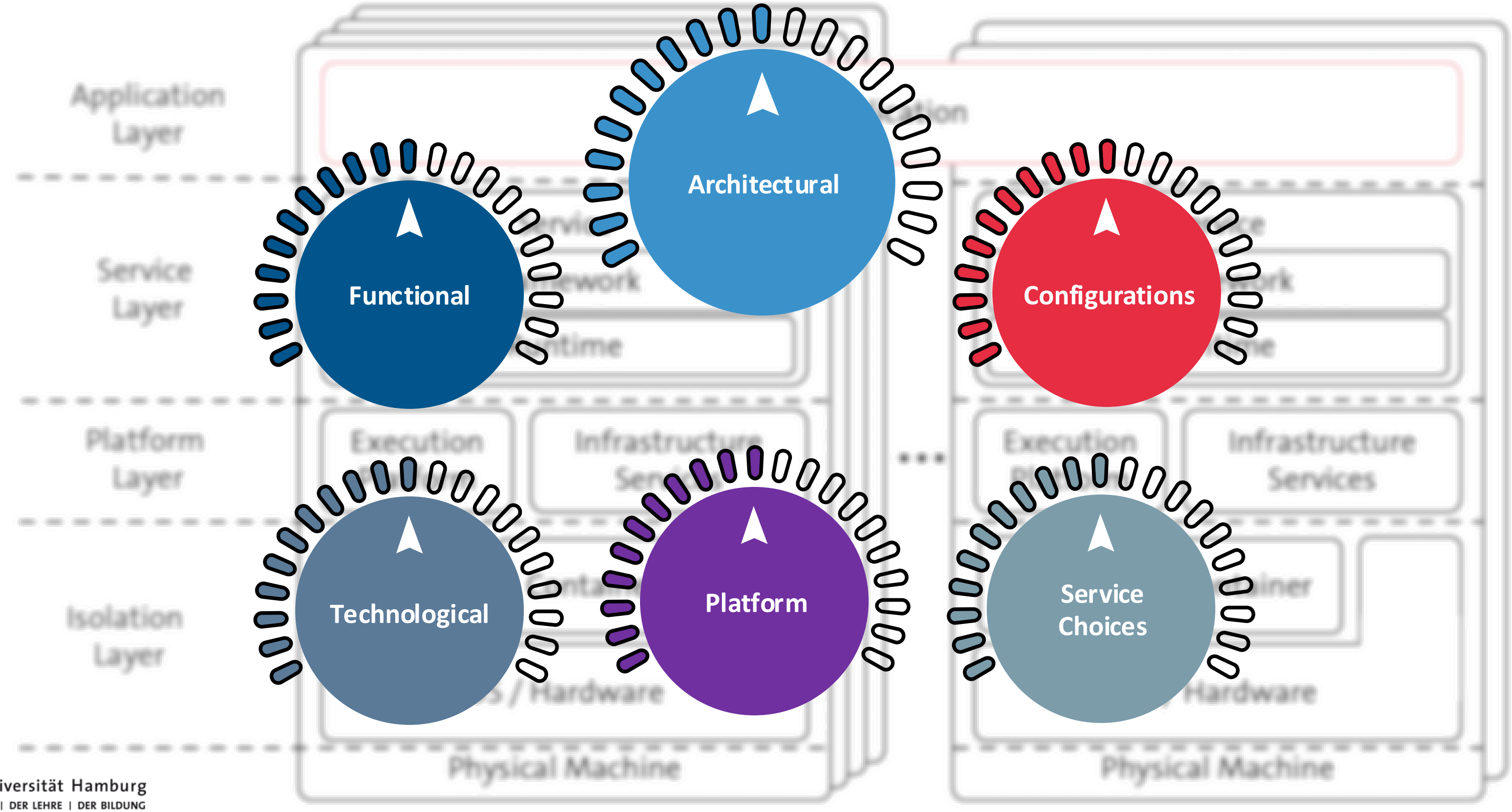
- 2.8–3.8% of total EU electricity use for data centers [1].
- Cloud-native applications consist of multiple layers of technologies, services, and platforms using these data centres.
- Identifying choices that impact energy efficiency is challenging.

How can we evaluate the energy efficiency of continuously developing cloud-native applications and their platforms?

[1] Kamiya, G. and Bertoldi, P., Energy Consumption in Data Centres and Broadband Communication Networks in the EU, Publication of the European Union, Luxembourg, 2024, doi:10.2760/706491, JRC135926.



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

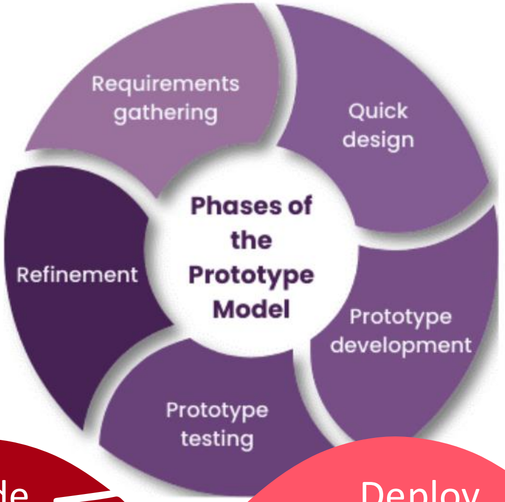


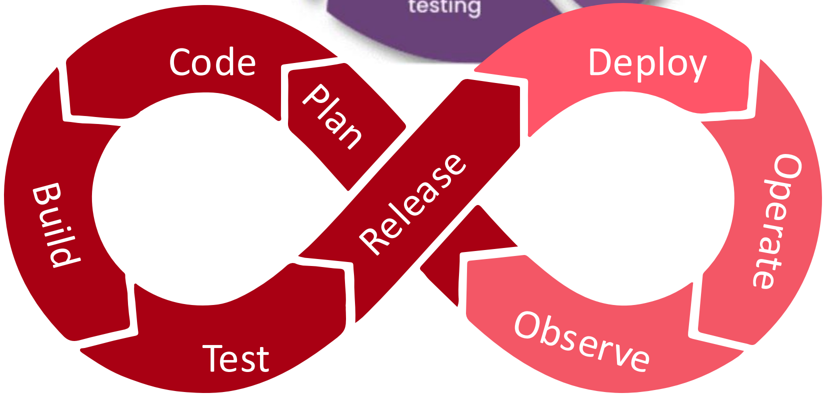
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Evaluating Energy Efficiency of Applications

What do we do in cases of evolving requirements and a large solution space?

- Prototyping
- Experimentation
- Software Quality Management





Clue

In support of this, we build **Clue**, a tool to continuously evaluate the energy-efficiency of changes in software development

- Git-oriented experiments
- CI/CD pipeline compatible
- Can be used to experiment/prototype

What do we evaluate (so far)?

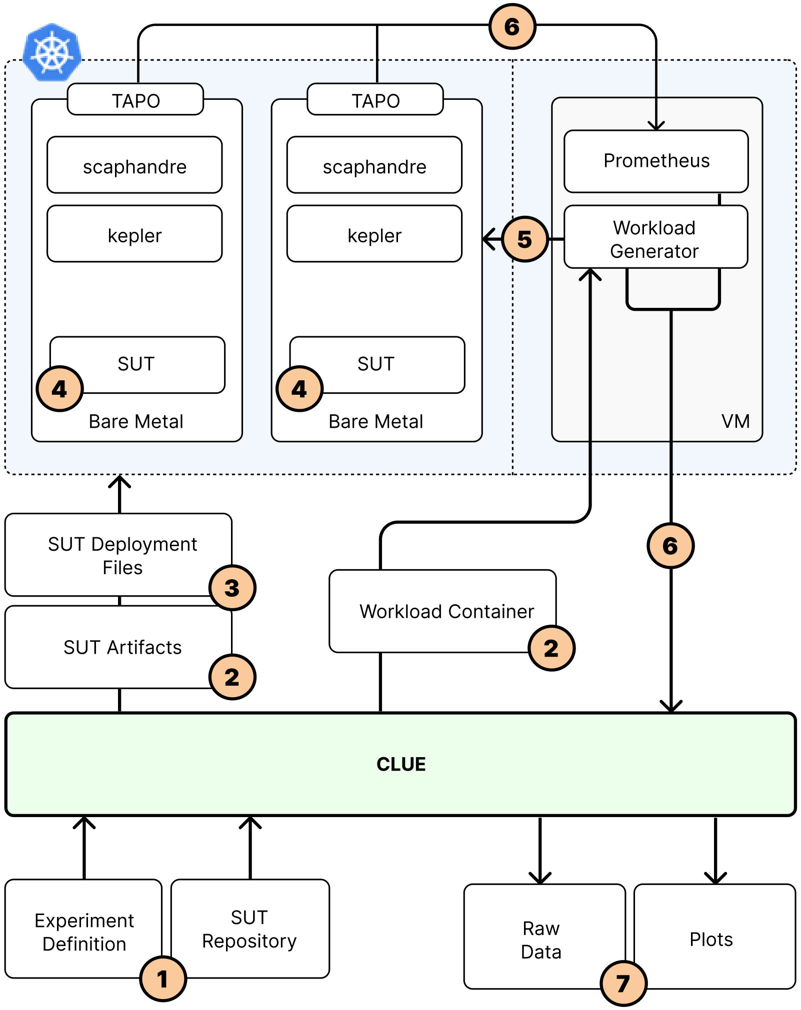
- System Qualities
- Energy Consumption*
- Resource Utilization
- Carbon Intensity

Name	
	Sustainability
Request Consumption (WR)	Ws
Runtime Overhead (RO)	[0..1]
Resource Utilization (RU)	[0..1]
Resource Efficiency (RE)	Ws
Auxiliary Costs (AC)	Ws
Software Carbon Intensity (SCI)	g_{CO2e}
	Quality
Total Costs (TC)	$
Failure Rate (FR)	[0..1]
	Performance
Requests (Rqs)	$\mathbb{R}$
Latency (Lat)	s

* We use existing energy meters, e.g., Kepler, that use estimation and sampling.

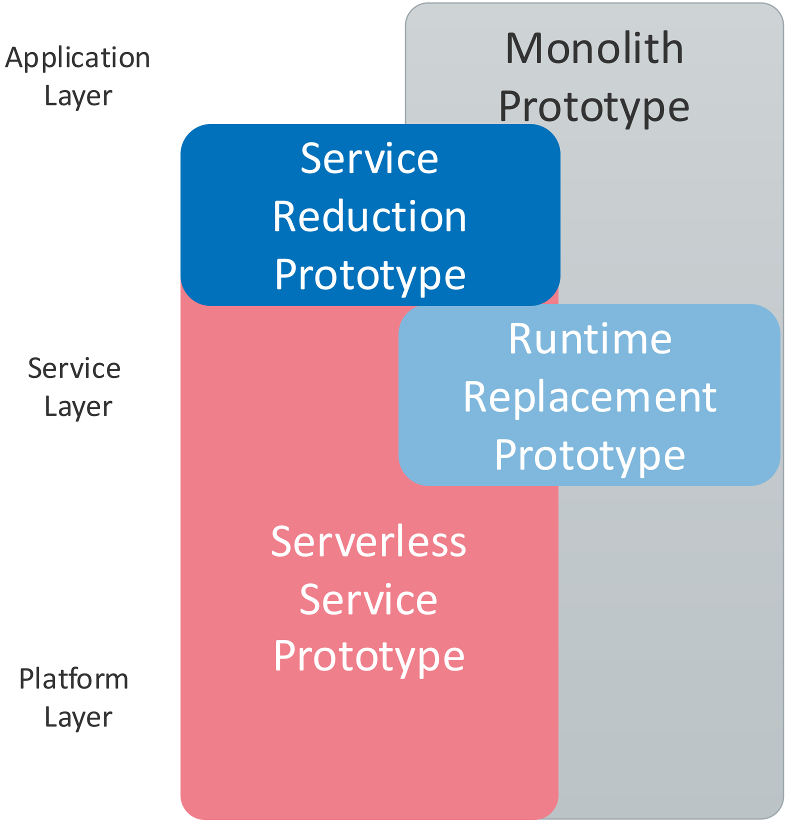
How does clue work?

- Need to define an experiment
- Requires IaC-capable deployments to a staging/dev environment
- Relies on Prometheus for collecting measurements



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Seeming Clue in action.



1. Tee Store[2]
2. On-Prem bare-metal Kubernetes cluster, with socket meters (for inner validation)
3. 4x Workload profiles

   [Fixed, Backoff, Stress, Shape]

[2] J. von Kistowski et. al. , "TeaStore: A Micro-Service Reference Application for Benchmarking, Modeling and Resource Management Research," *2018 MASCOTS*, doi: 10.1109/MASCOTS.2018.00030.

System Quality

Branch	Latency p95 [s]	Failure Rate [%]	Costs		
			Projected Total [$]	Projected Consumed [$]	Per Request [¢/1000]
Baseline	0.17 - 16.37	3.5 - 11.51	0.58 - 0.84	0.27 - 0.41	24.01 - 0.26
Runtime Replacement	0.10 - 12.42	2.3 - 0.03	0.58 - 0.82	0.27 - 0.40	23.11 - 0.10
Monolith Architecture	0.04 - **42.78**	0.89 - **41.80**	0.16 - 0.26	0.08 - 0.11	10.10 - 0.77
Service Reduction	0.20 - 8.36	1.9 - 1.78	0.69 - 0.86	0.28 - 0.41	24.98 - 0.10
Serverless	1.76 - 15.38	5.1 - 9.31	**4.08 - 4.60**	0.67 - 0.94	**63.49** - 0.53

Comparing Pulsing and Stress Workloads

System Quality

Serverless has a high (idle/startup) cost, that also impacts performance and reliability.

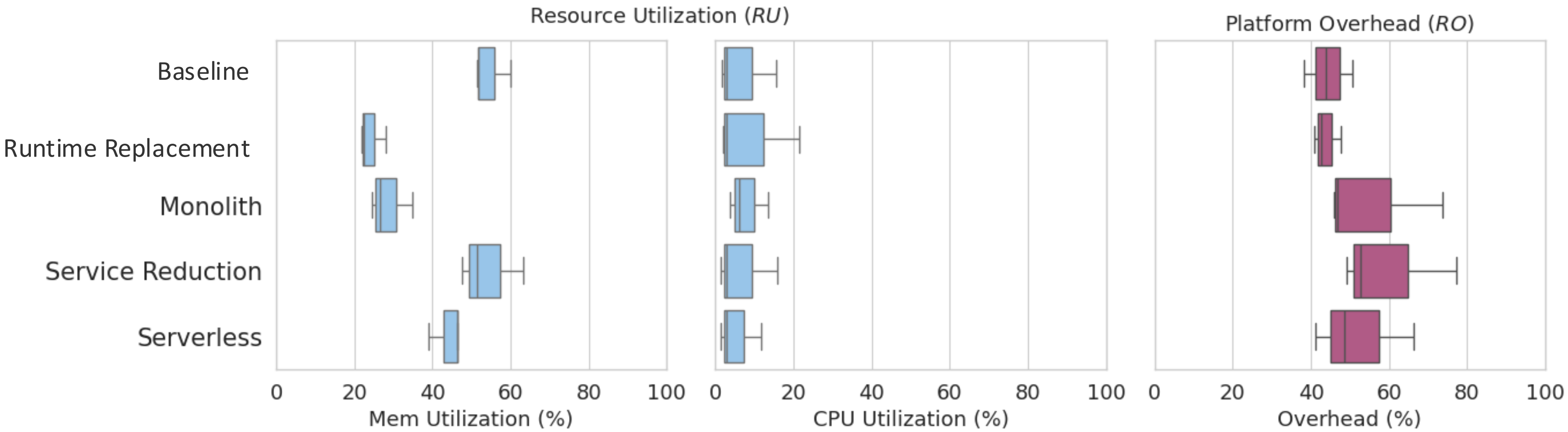
The **Monolith** can’t scale enough for the stress workload but has very good idle cost and performance.

Runtime Replacement outperforms in all categories.

Branch				Consumed [$]	equest [¢/1000]
Baseli					4.01 - 0.26
Runti					3.11 - 0.10
Monolith Architecture	0.04 - **42.78**	0.89 - **41.80**	0.16 - 0.26	0.08 - 0.11	10.10 - 0.77
Service Redu					10
Serverless	1.76 - 15.38	5.1 - 9.31	**4.08 - 4.60**	0.67 - 0.94	**63.49** - 0.53

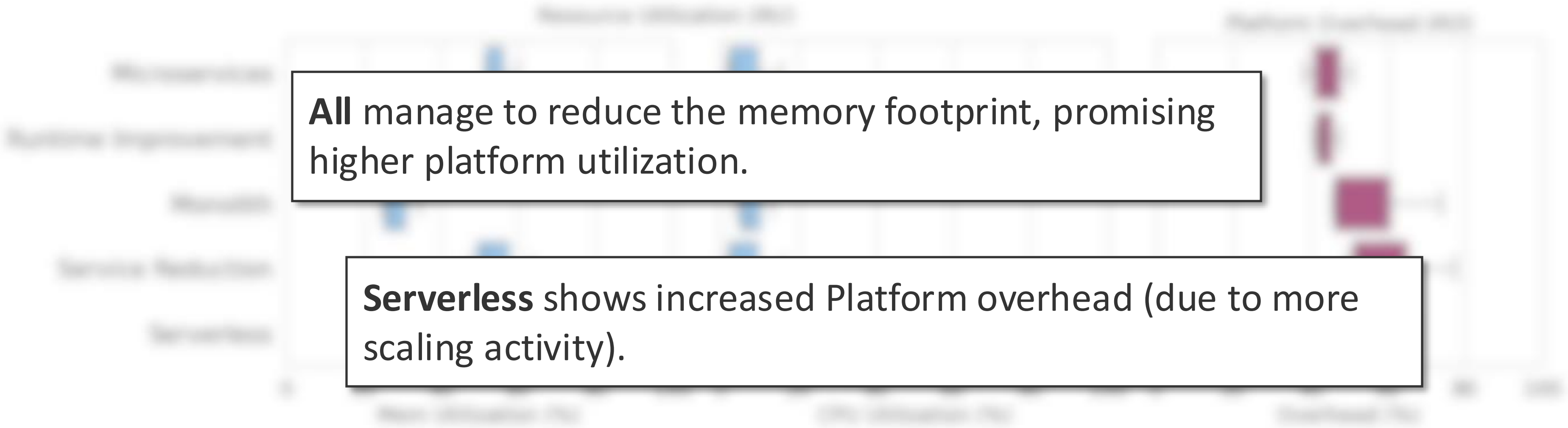
Comparing Pulsing and Stress Workloads

Resource Utilization



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

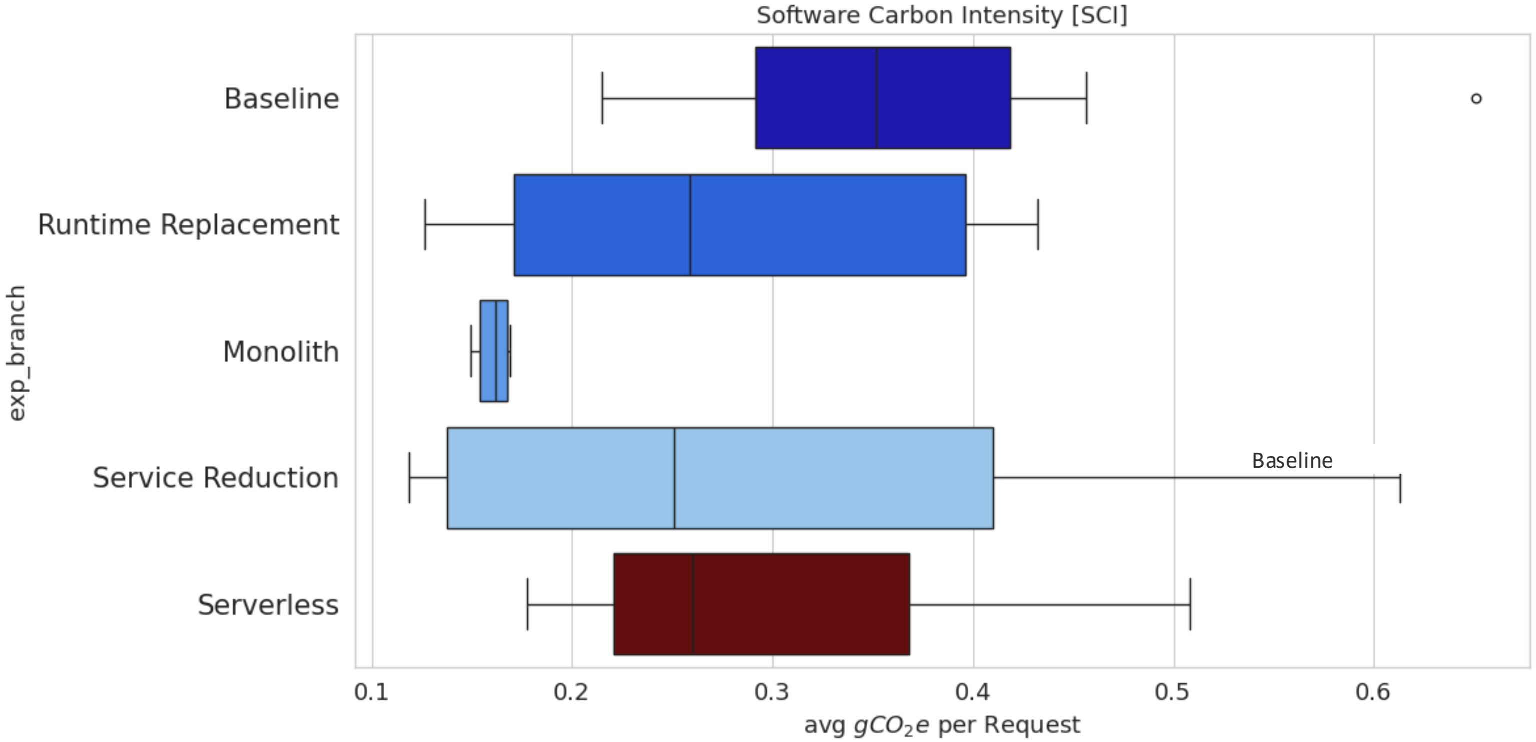
Resource Utilization



All manage to reduce the memory footprint, promising higher platform utilization.

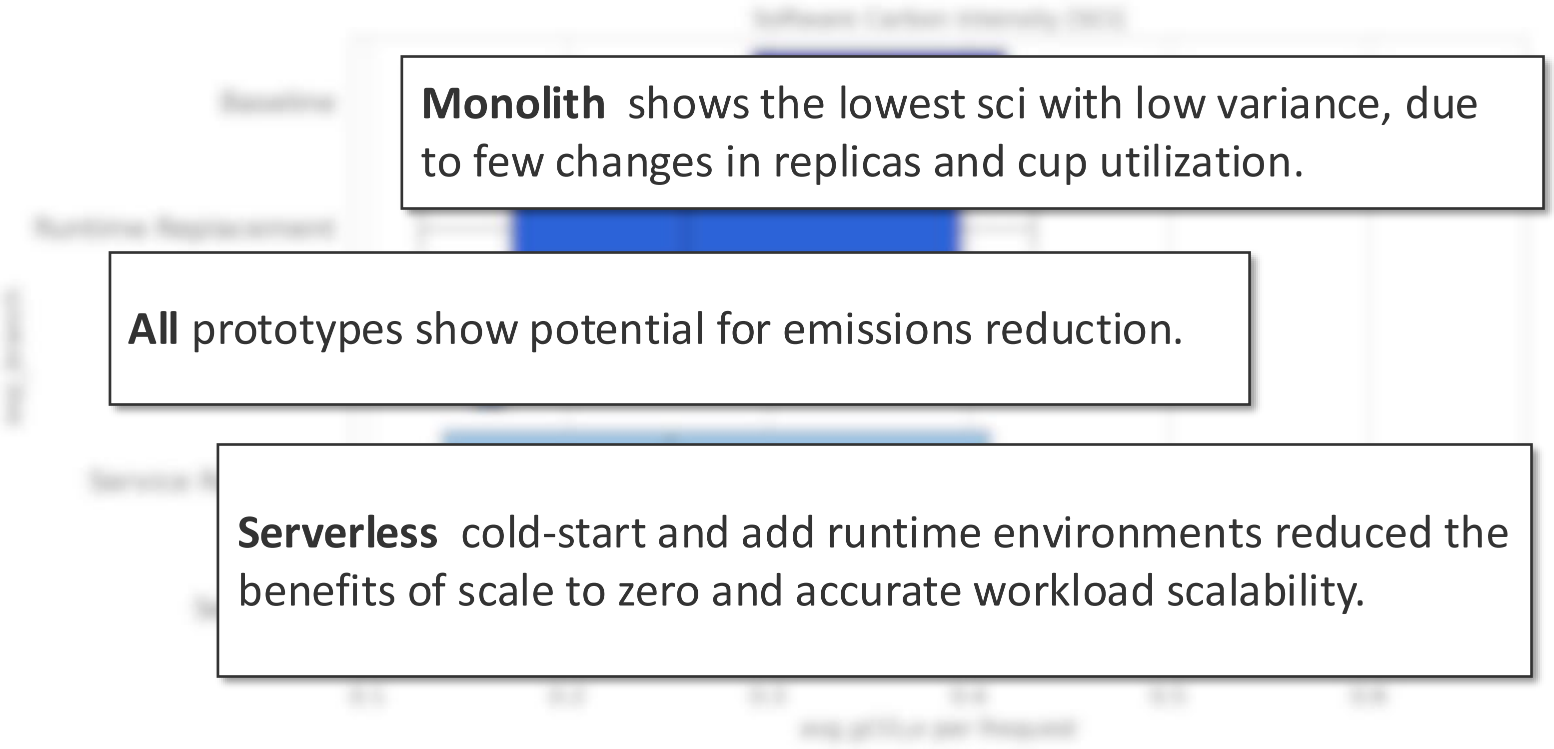
Serverless shows increased Platform overhead (due to more scaling activity).

Energy Consumption



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Energy Consumption



Monolith shows the lowest sci with low variance, due to few changes in replicas and cup utilization.

All prototypes show potential for emissions reduction.

Serverless cold-start and add runtime environments reduced the benefits of scale to zero and accurate workload scalability.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Con-Clue-sion

(sorry I had to)

- Serverless is surprisingly not always saving energy despite scale to zero
- Continues Prototyping can lead to application specific improvements
- Cloud energy meters already sensitive enough to evaluate changes across the full cloud-native stack

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Contact

sebastian.werner@uni-hamburg.de

sw@ise.tu-berlin.de

in/bastiwerner/

Clue

ISE-TU-Berlin/Clue