Exploring the Power Usage Effectiveness in Serverless Computing

Anderson Andrei Da Silva, Nathan Leblond,
Kellian Leveque, Romain Rouvoy

Inria, Univ. Lille, CNRS, UMR 9189 CRIStAL, France

GreenDays 2026 - Sophia

This presentation is based on: Assessing Power Usage Effectiveness in Serverless Computing Environments, presented at the 11th International Workshop on Serverless Computing (part of ACM/IFIP Middleware 2025) Nashville, TN, USA - December 15th, 2025

Table of Contents

Introduction

- As the computational power of machines, servers, and clusters has not ceased to grow, the energy consumed to use, cool, and maintain these computational resources has also grown, generating a huge ecological crisis in the domain.
  - The energy consumed by the data-centers themselves has already achieved 1% of the global energy consumed, and may continue to increase [1].
- Serverless computing has the potential of saving resources by reusing previous computational environments, and by releasing resources as soon as functions are executed, it is a promising paradigm to be used towards a greener computing [2, 3, 4].
  - However, the efforts made in this direction all rely on power usage measurements and energy consumption estimations.
  - To the best of our knowledge, no accurate metrics have been explored in the context of serverless computing.

Contributions

We propose a preliminary exploration of the *Power Usage Effectiveness* (PUE) of serverless. Our main contributions are:

- We quantify energy overheads in serverless executions;
- We propose a reproducible setup [5];
- We demonstrate how PowerAPI [6] provides fine-grained energy measurements in a serverless environment;
- We adapt the suite of benchmarks named FunctionBench [7];
- We propose a first approach on how to compute the named vPUE and cPUE of serverless functions;
- Finally, we detail our exploration of the different relations between the PUE of serverless functions and
  - their starting mode;
  - the size of serverless platforms;
  - the loading of serverless functions.

Preliminary Concepts

Power Usage Effectiveness:

- The *Power Usage Effectiveness* (PUE) can be defined as the ratio of the total energy consumed at the data-center level to the energy consumed by hosted IT equipment.
  - This metric is considered optimal when it is lowered to 1.
- More specifically, we consider extended versions of the PUE: the *virtual PUE* (vPUE) and the *combined PUE* (cPUE) of serverless applications [8, 9].
  - The vPUE computes the PUE of individual software layers, which can be stacked indefinitely and accounts for nested virtualization practices,
  - The cPUE reflects the end-to-end efficiency of a software platform.

Serverless Computing:

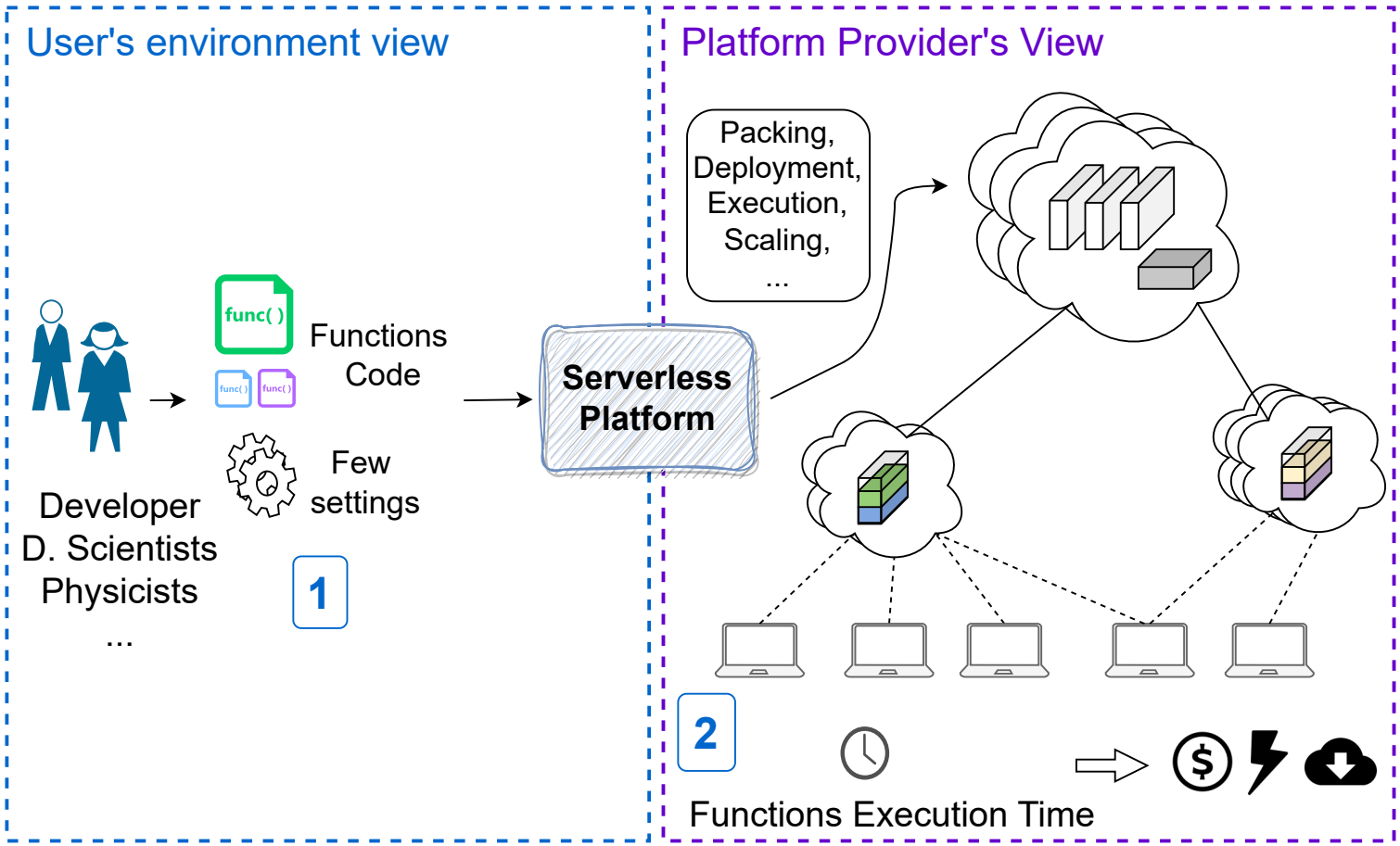
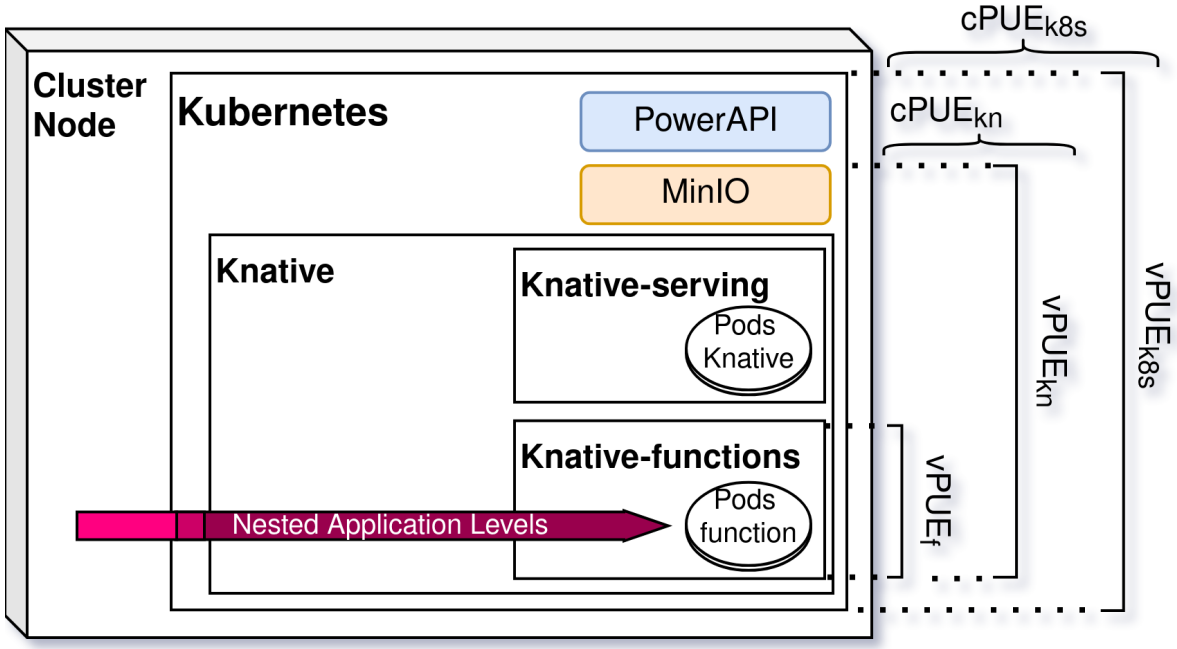


Figure: 1. Serverless computing illustration and definitions [10]

Table of Contents

Serverless vPUE and cPUE

As defined in [8], the vPUE and the cPUE are the ratios between the energy (in Joules in our case) of different components that make up our serverless layers and the energy consumed by our infrastructure. We compute them as illustrated in Figure 8, with the following the equations:



Reproducible artifacts

Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

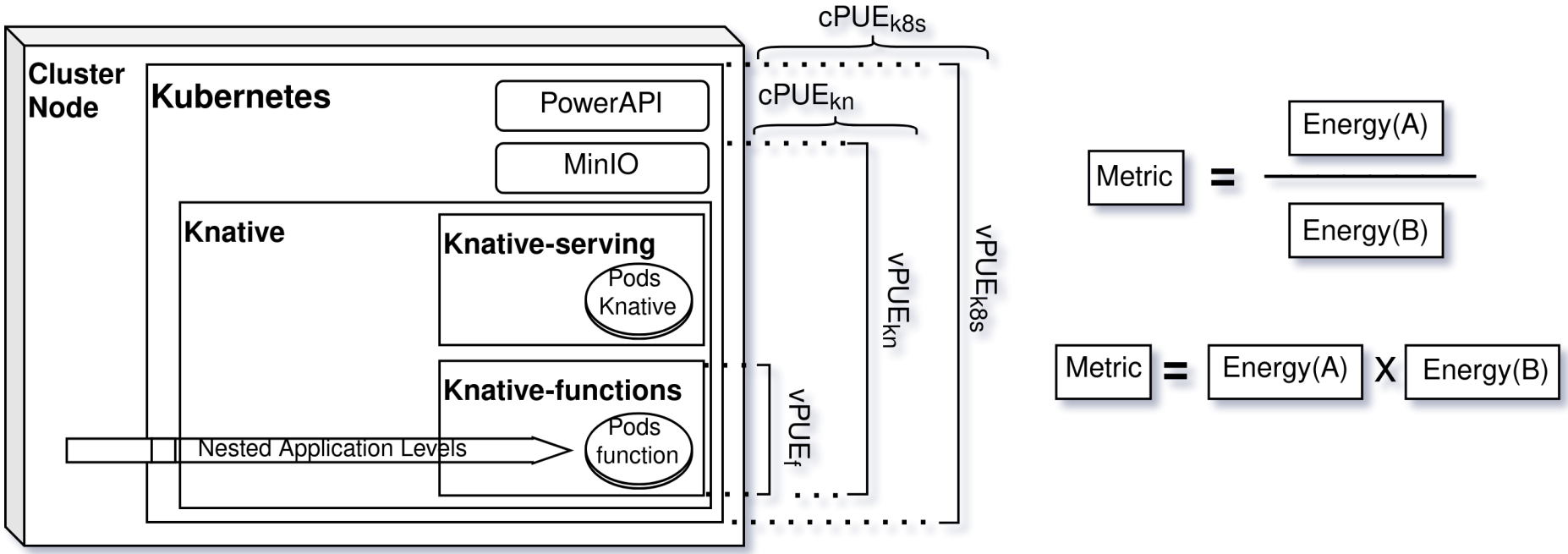


Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

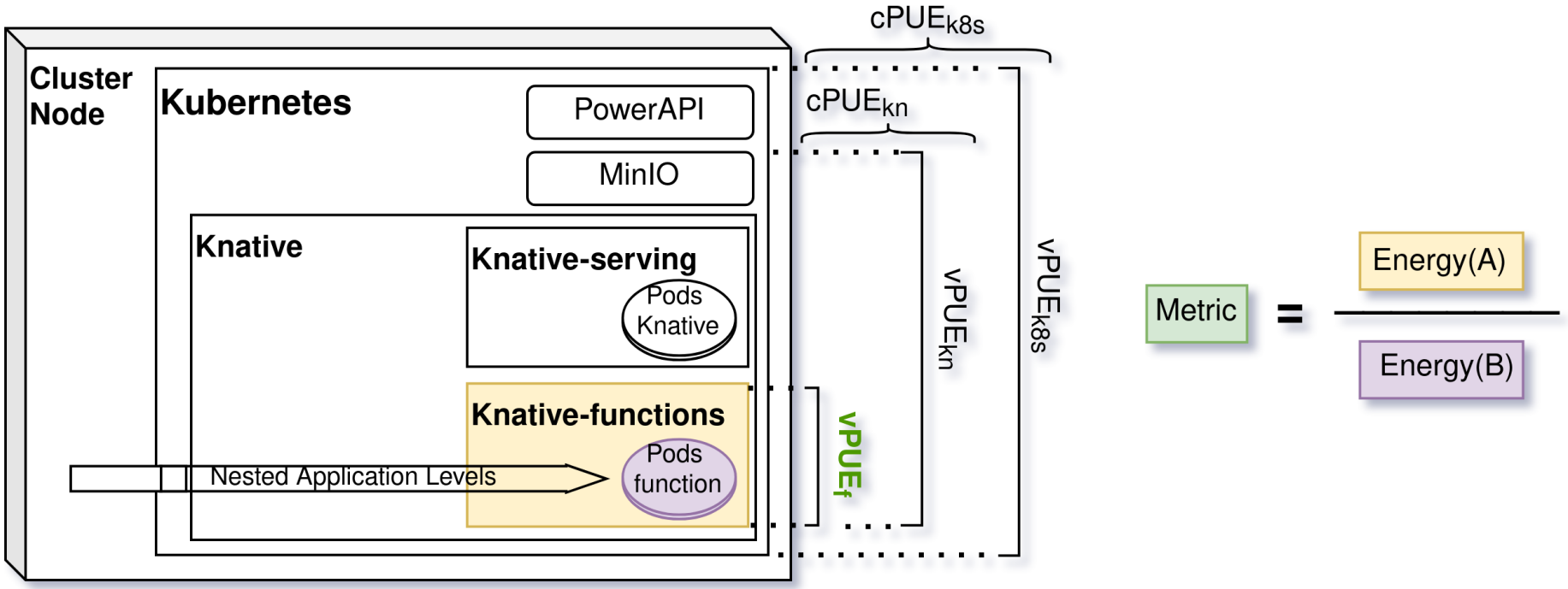


Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

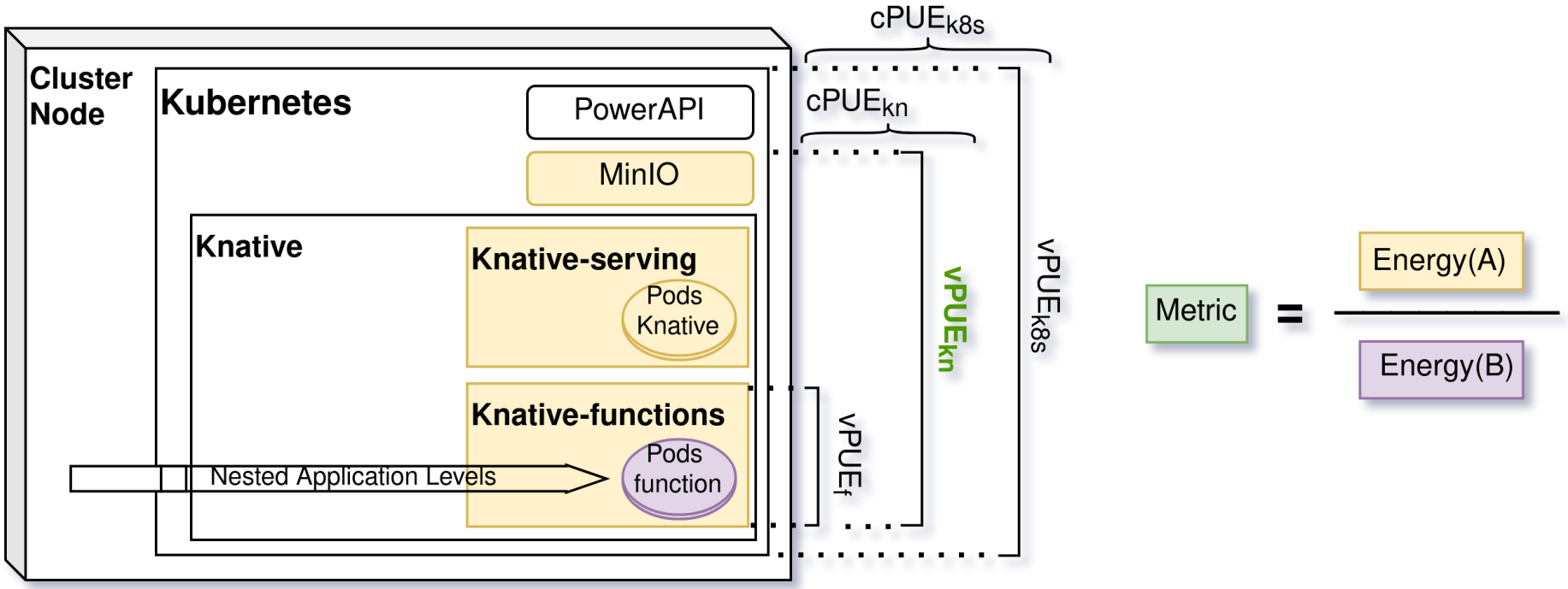


Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

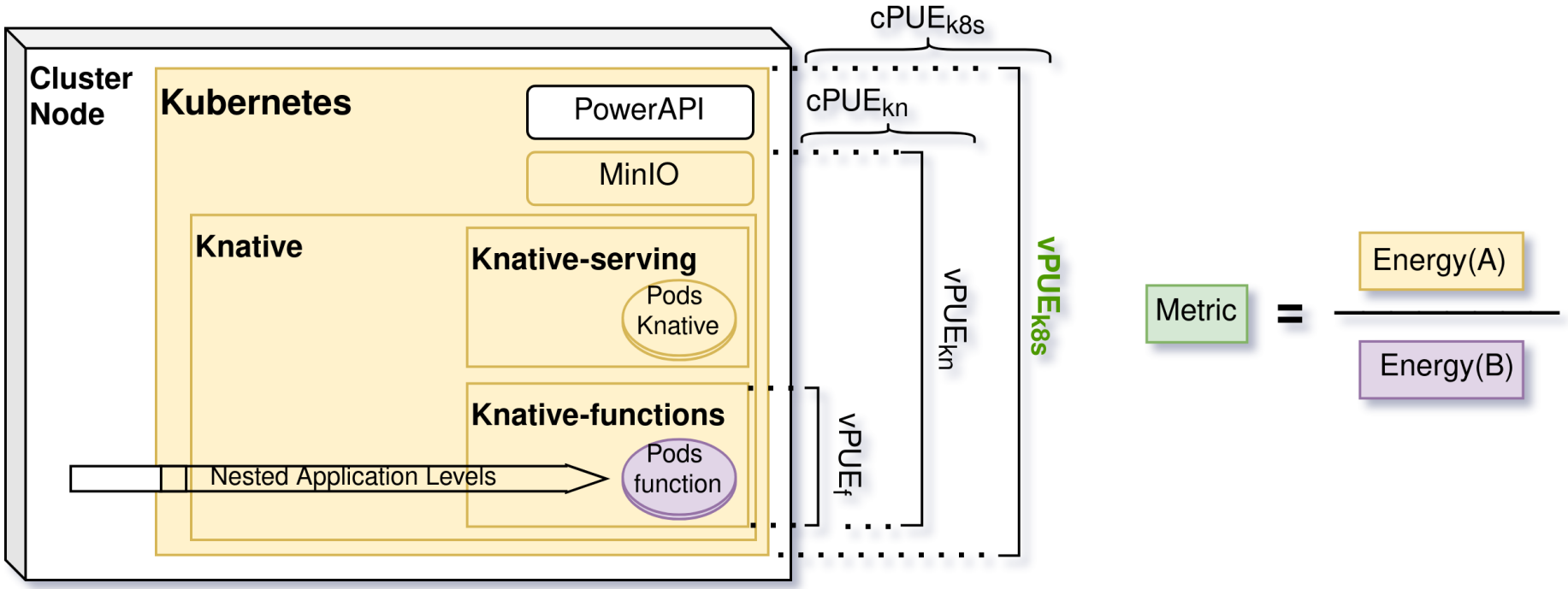


Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

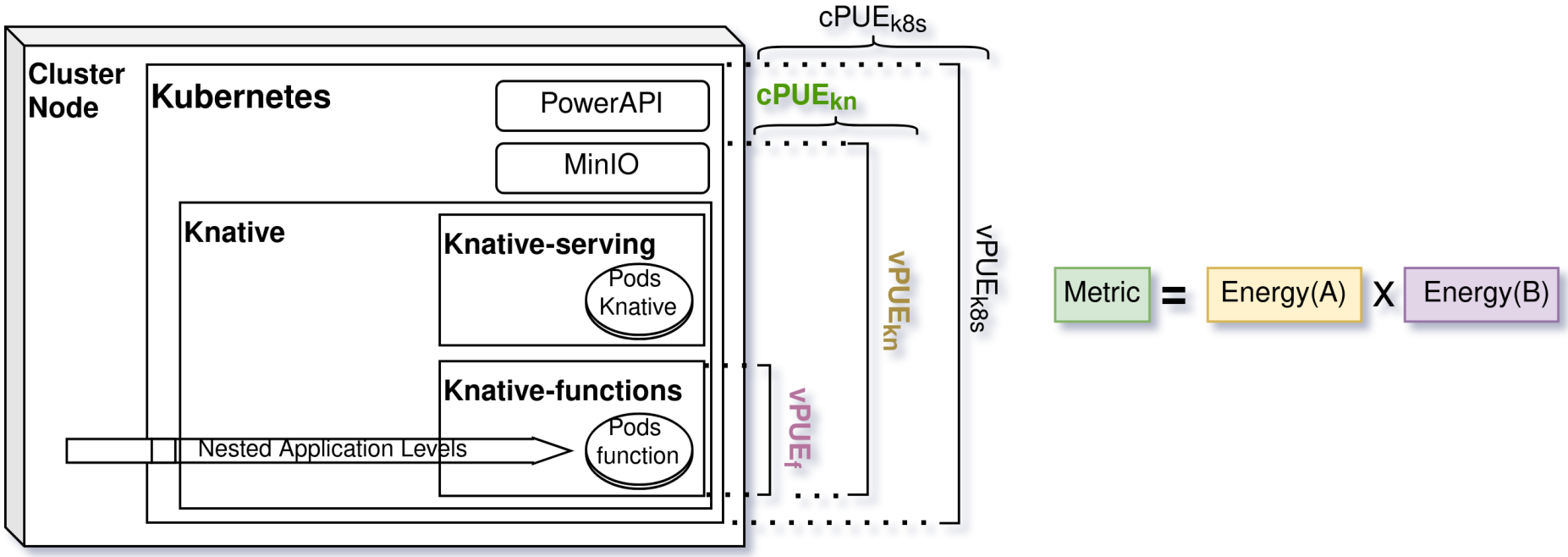


Figure: 2. Nested Application Levels and the PUE's computation.

Serverless vPUE and cPUE

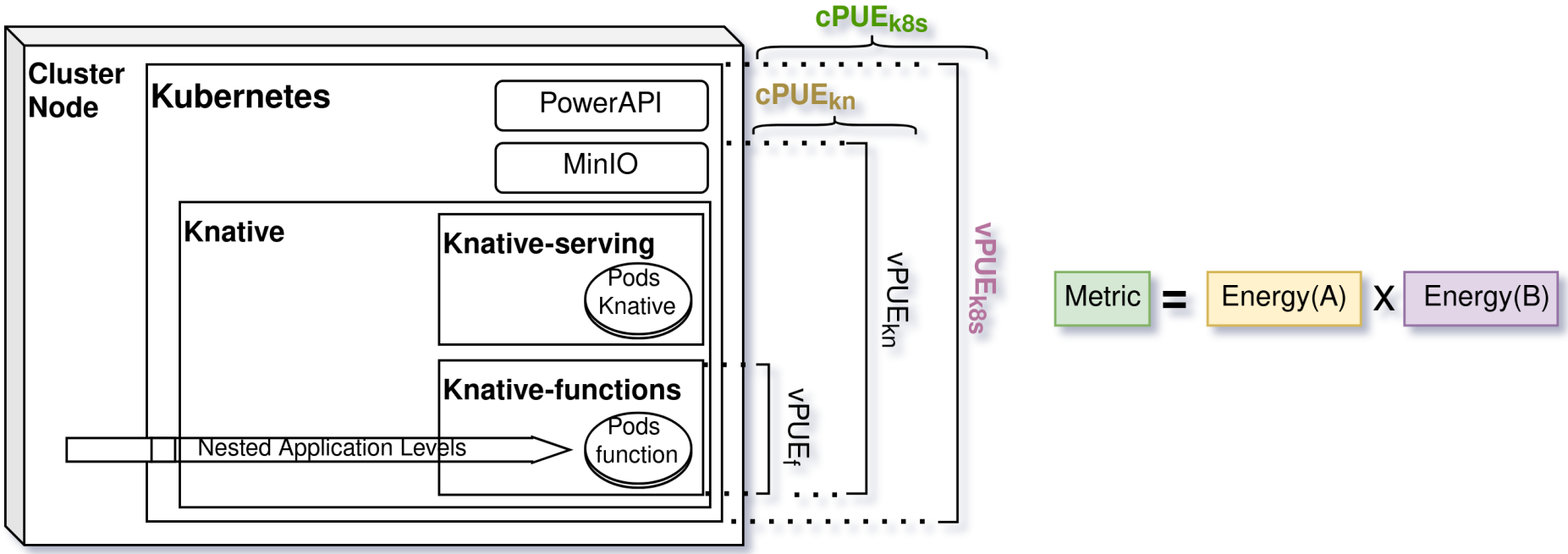


Figure: 2. Nested Application Levels and the PUE's computation.

Design of Experiments

Parameters	Batch 1	Batch 2
Workload Size	24 (4 functions × 2 inputs × 3 repetitions)	200 (2 functions × 100 repetitions)
Platform Size	1 control plane + 1 worker	1 control plane + 1, 2, 3 workers
Execution Mode	Cold, Warm and Hot	Hot
Concurrency Target	70%	50, 70, 100%
Total of executions	72	3,600

Table: 1. Design of Experiments. Total of 3672 experiments.

We evaluated 5 functions with different inputs. In total, we cover 10 combinations of functions and inputs, from FunctionBench [7, 11, 12]:

- Linpack (35000 and 5000 nxn matrices);
- Chameleon (4000 and 8000 nxn matrices);
- Matrix Multiplication (25000 and 40000 nxn matrices);
- Model Training (50 and 100 Mb datasets);
- Video Processing (50 and 100 Mb video sizes).

Table of Contents

Performance over the functions execution mode - vPUE

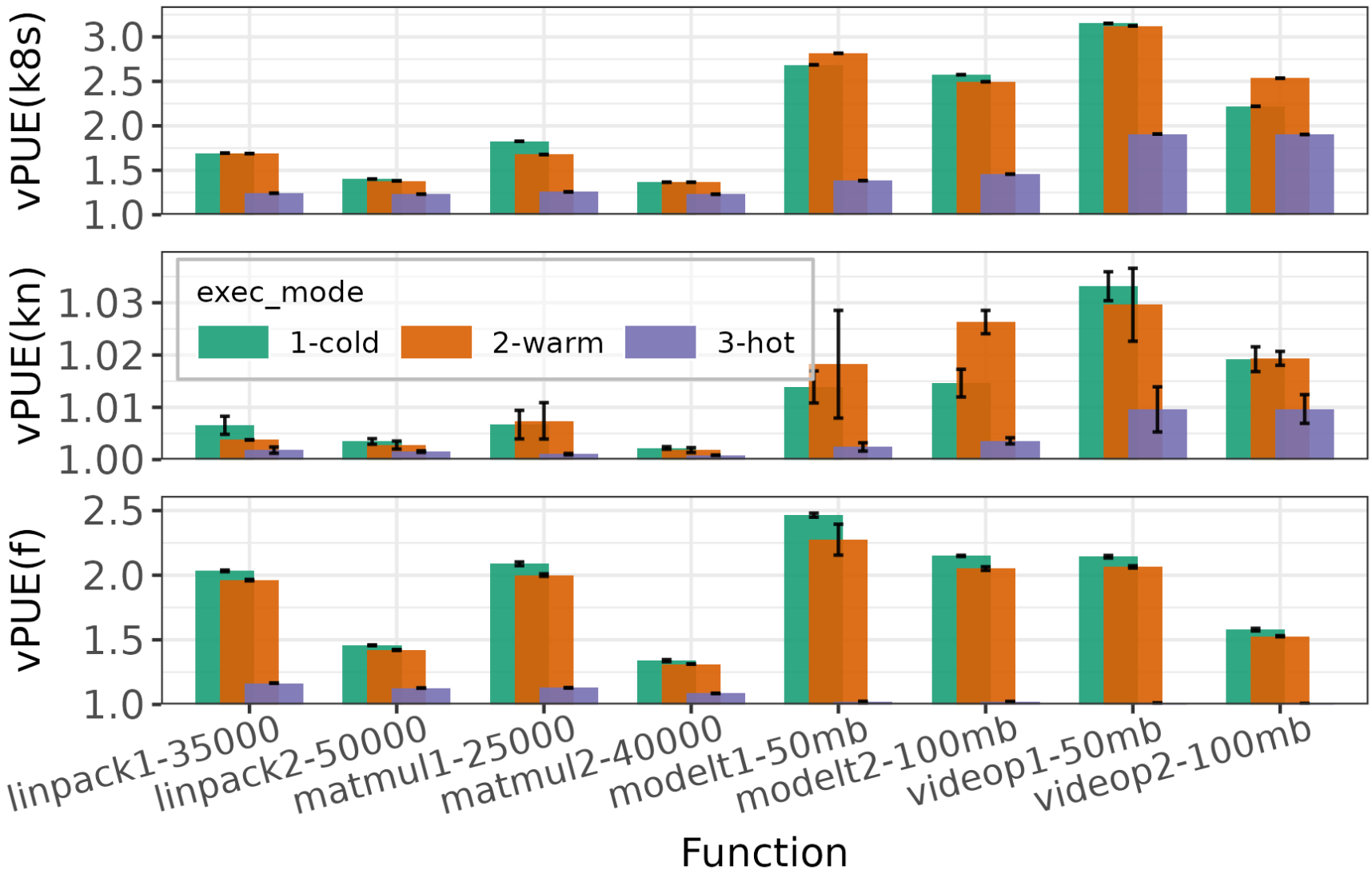


Figure: 4. vPUE for different parameters & functions.

Overall Serverless cPUE

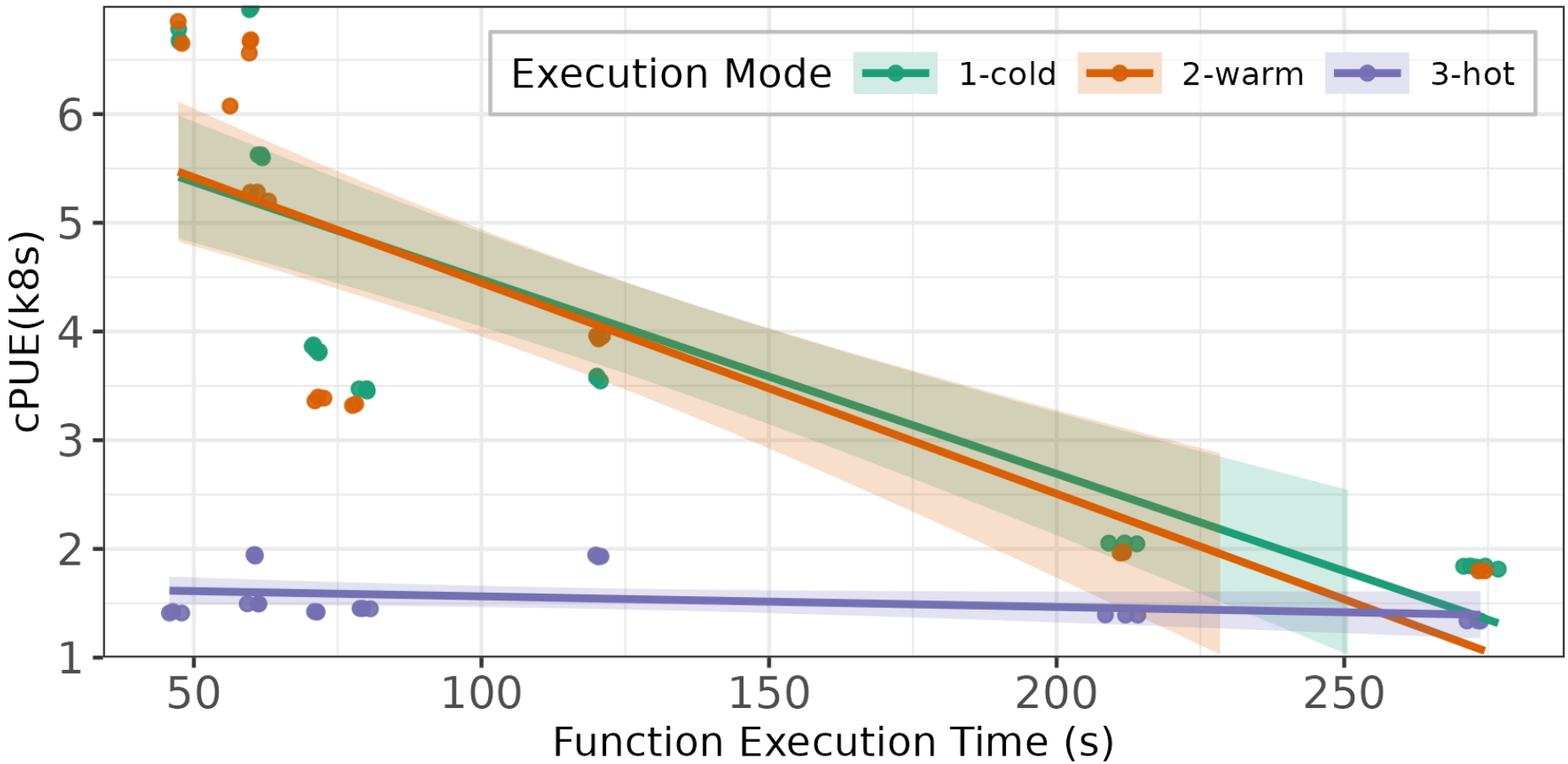


Figure: 6. Linear regression of function execution time over the cPUE.

Table of Contents

Conclusion

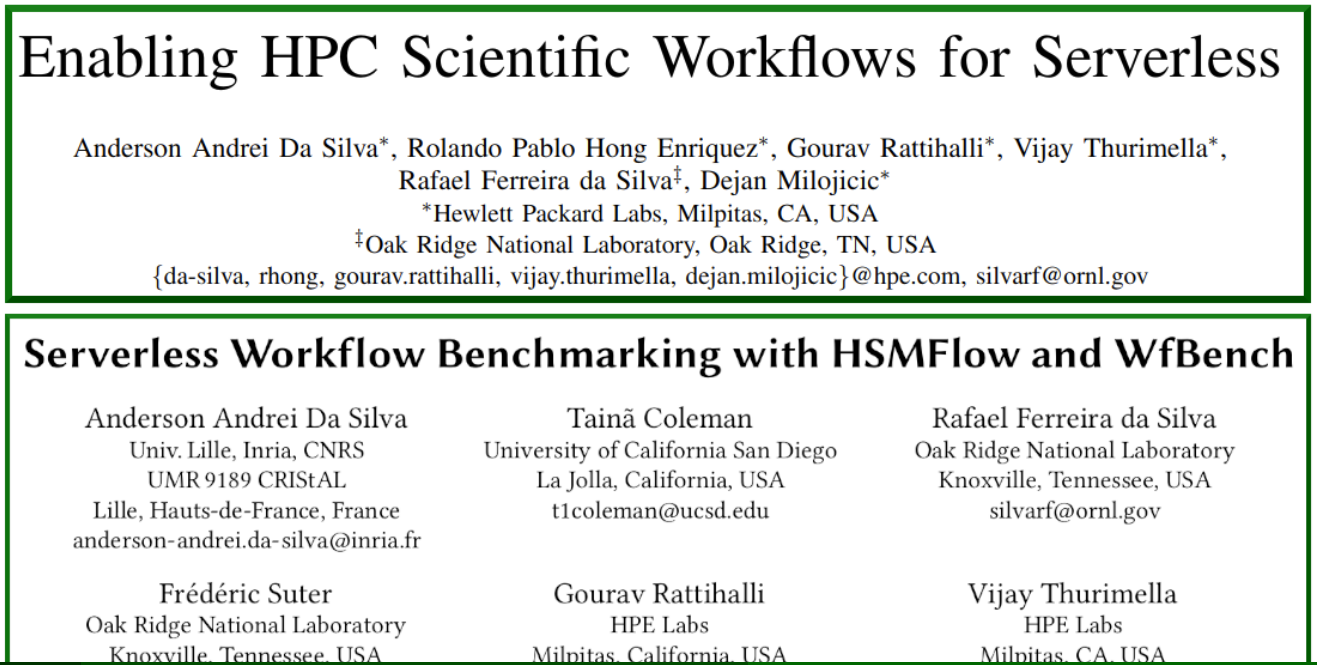
Our findings acknowledge that

- The starting mode of functions in a serverless platform is critical.
- Hot start-ups are really beneficial in terms of the overall PUE.
- Even as serverless was designed for very fast executed functions (from milliseconds to a few minutes), they do not have good performances in terms of PUE.
  - In fact, as long as we use the deployed functions for a longer duration, their PUE is improving, achieving close to optimal performance.
  - On the contrary, our results showed an increase of by 2.5× of $vPUE_f$ for cold and warm-started functions, in comparison to hot ones; arriving at by 7× worst performance in a global view within $cPUE_{k8s}$.

Scientific Perspectives

We plan to expand our experimentation setup:

- To include a wider range of serverless functions, also including workflow executions;
- To add more PUE metrics to improve our exploration campaign, such as the *partial PUE* (pPUE) to better model the PUE of external but included services, such as *MinIO*.
- To investigate the cPUE of scientific workflows executed on serverless platforms:



Enabling HPC Scientific Workflows for Serverless

Anderson Andrei Da Silva*, Rolando Pablo Hong Enriquez*, Gourav Rattihalli*, Vijay Thurimella*,
Rafael Ferreira da Silva‡, Dejan Milojicic*
*Hewlett Packard Labs, Milpitas, CA, USA
‡Oak Ridge National Laboratory, Oak Ridge, TN, USA
{da-silva, rhong, gourav.rattihalli, vijay.thurimella, dejan.milojicic}@hpe.com, silvarf@ornl.gov

Serverless Workflow Benchmarking with HSMFlow and WfBench

Anderson Andrei Da Silva
Univ. Lille, Inria, CNRS
UMR 9189 CRIStAL
Lille, Hauts-de-France, France
anderson-andrei.da-silva@inria.fr

Tainã Coleman
University of California San Diego
La Jolla, California, USA
t1coleman@ucsd.edu

Rafael Ferreira da Silva
Oak Ridge National Laboratory
Knoxville, Tennessee, USA
silvarf@ornl.gov

Frédéric Suter
Oak Ridge National Laboratory
Knoxville, Tennessee, USA

Gourav Rattihalli
HPE Labs
Milpitas, California, USA

Vijay Thurimella
HPE Labs
Milpitas, CA, USA

Acknowledgments

This work is done in the context of the Inria – Qarnot PULSE project: `https://www.inria.fr/en/pulse`, `https://defi-pulse.github.io/`. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

References

[1] "Data centres and data transmission networks, iea, paris, tech." https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks, 2023, accessed: October-2024.

[2] A. A. Da Silva, Y. Georgiou, M. Mercier, G. Mounié, and D. Trystram, "Foa-energy: A multi-objective energy-aware scheduling policy for serverless-based edge-cloud continuum," in *Proc. of the 40th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '25. ACM, 2025. [Online]. Available: https://doi.org/10.1145/3672608.3707941

[3] S. Werner, M. Kähler, and A. Hakamian, "Code once, run green: Automated green code translation in serverless computing," 09 2025.

[4] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," 2008.

[5] "Assessing power usage effectiveness in serverless computing environments - support repository," https://gitlab.com/andersonandrei/serverless-pue-workshop/, 2025, accessed: November-2025.

[6] G. Fieni, D. R. Acero, P. Rust, and R. Rouvoy, "PowerAPI: A Python framework for building software-defined power meters," *Journal of Open Source Software*, vol. 9, no. 98, p. 6670, Jun. 2024. [Online]. Available: https://hal.science/hal-04601379

[7] J. Kim and K. Lee, "FunctionBench: A Suite of Workloads for Serverless Cloud Function Service," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*. IEEE, Jul. 2019.

[8] G. Fieni, R. Rouvoy, and L. Seinturier, "*x*pue: Extending power usage effectiveness metrics for cloud infrastructures," *IEEE Transactions on Sustainable Computing*, 2025. [Online]. Available: http://dx.doi.org/10.1109/TSUSC.2025.3549687

[9] ——, "SmartWatts: Self-Calibrating Software-Defined Power Meter for Containers," in *CCGRID 2020 - 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing*, May 2020. [Online]. Available: https://inria.hal.science/hal-02470128

[10] A. A. Da Silva, "Apprentissage sur l'impact des politiques d'ordonnancement multi-objectifs pour les fonctions serverless dans le edge-cloud continuum," Ph.D. dissertation, 2023, thèse de doctorat dirigée par Trystram, DenisMounié, Grégory et Georgiou, Yiannis Informatique Université Grenoble Alpes 2023. [Online]. Available: http://www.theses.fr/2023GRALM078

[11] "Functionbench repository," https://github.com/ddps-lab/serverless-faas-workbench, 2025, accessed: October-2025.

[12] "Functionbench adated to knative," https://github.com/andersonandrei/serverless-faas-workbench/, 2025, accessed: October-2025.

Exploring the Power Usage Effectiveness in Serverless Computing

Anderson Andrei Da Silva, Nathan Leblond,
Kellian Leveque, Romain Rouvoy

Inria, Univ. Lille, CNRS, UMR 9189 CRIStAL, France

GreenDays 2026 - Sophia
This presentation is based on: Assessing Power Usage Effectiveness in Serverless Computing Environments, presented at the 11th International Workshop on Serverless Computing (part of ACM/IFIP Middleware 2025)
Nashville, TN, USA - December 15th, 2025

Thank you very much for your attention!

Contact info: anderson-andrei.da-silva@inria.fr,
silva.andersonandrei@gmail.com,
https://www.linkedin.com/in/andersonandrei/

Serverless executions & Challenges

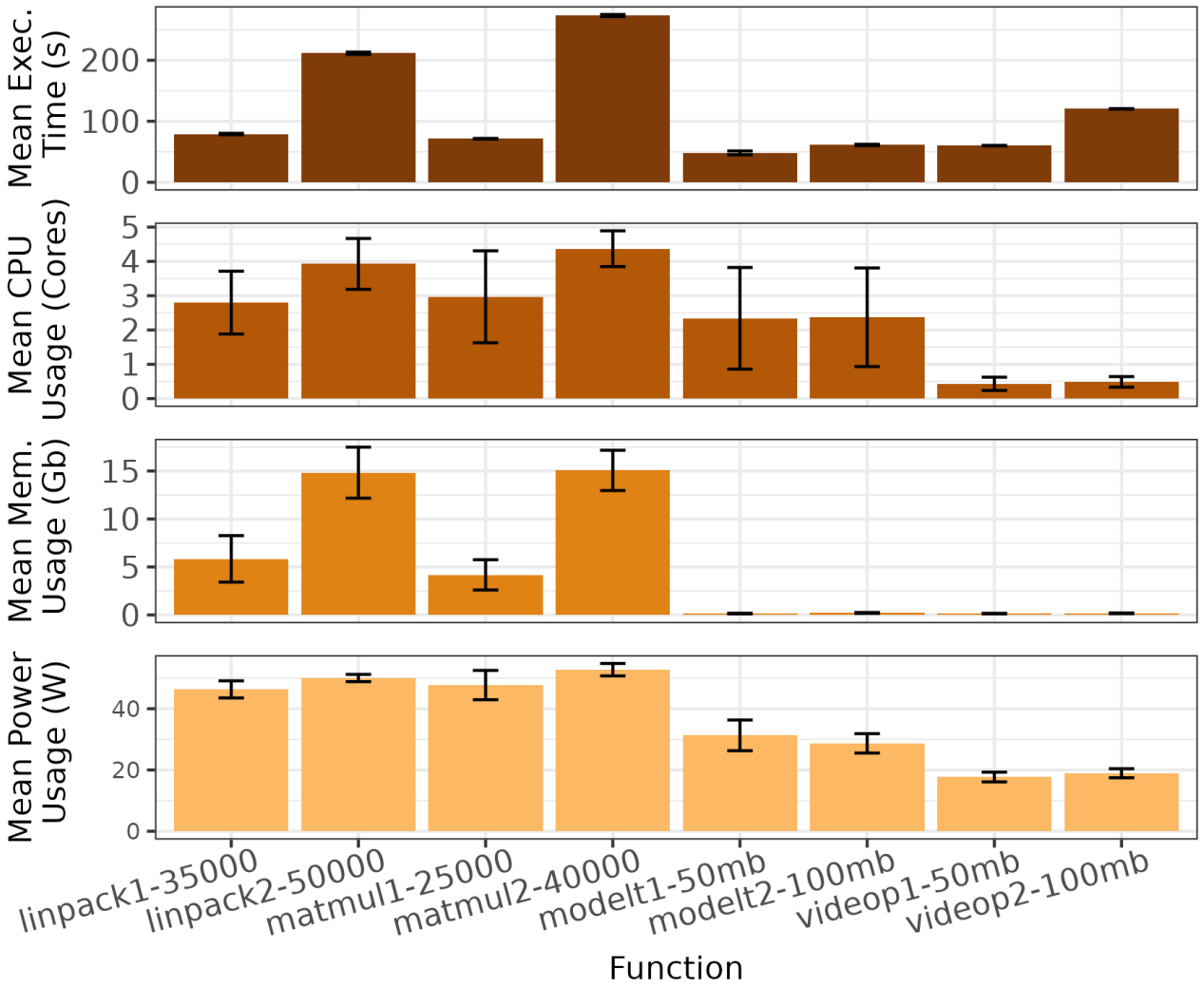


Figure: 3. Execution time and resource usage of the serverless functions

Performance over the functions execution mode - cPUE

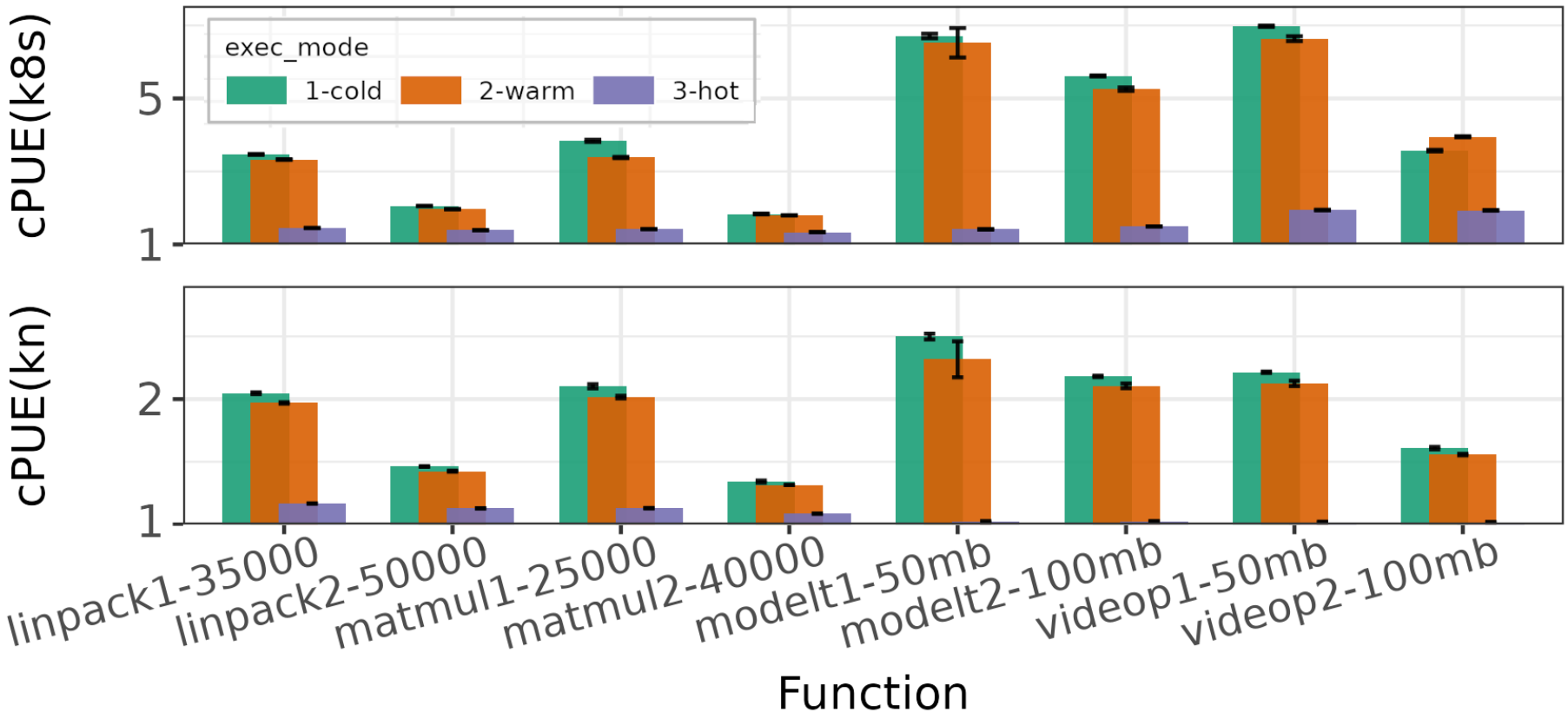


Figure: 5. cPUE for different parameters & functions.

Serverless executions & Challenges

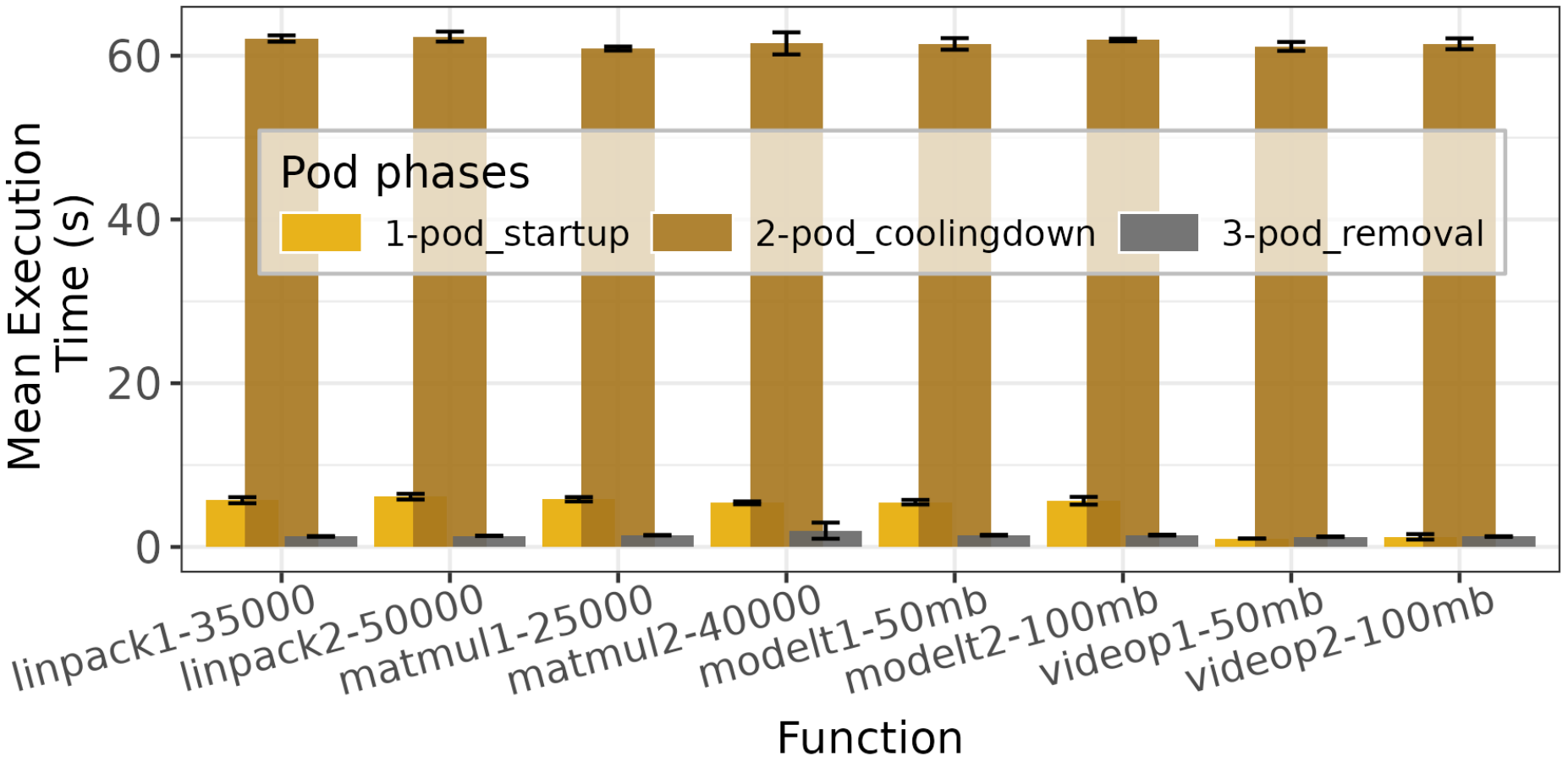


Figure: 7. Mean duration of different phases of pods execution.

Impacts on number of nodes

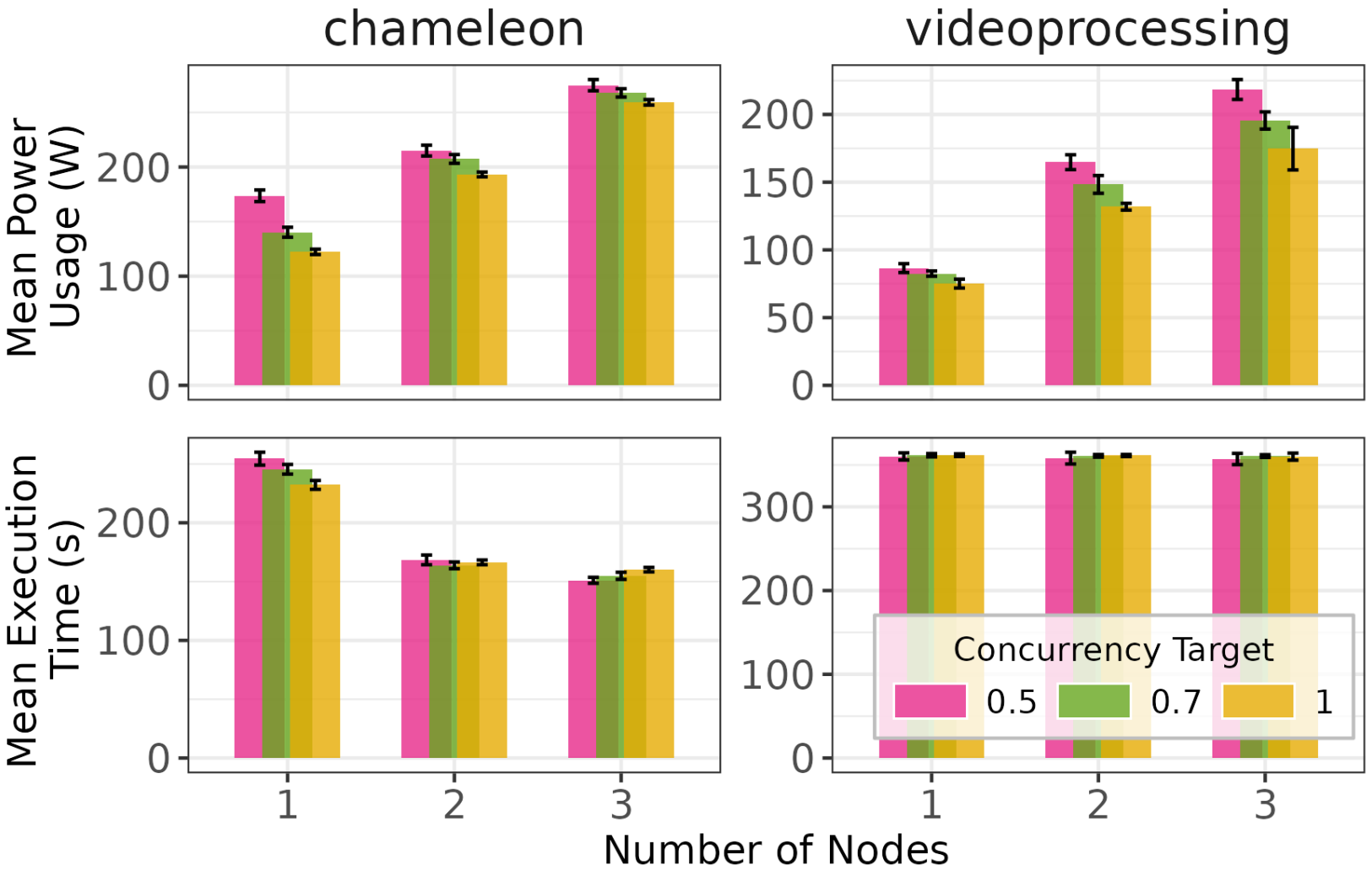


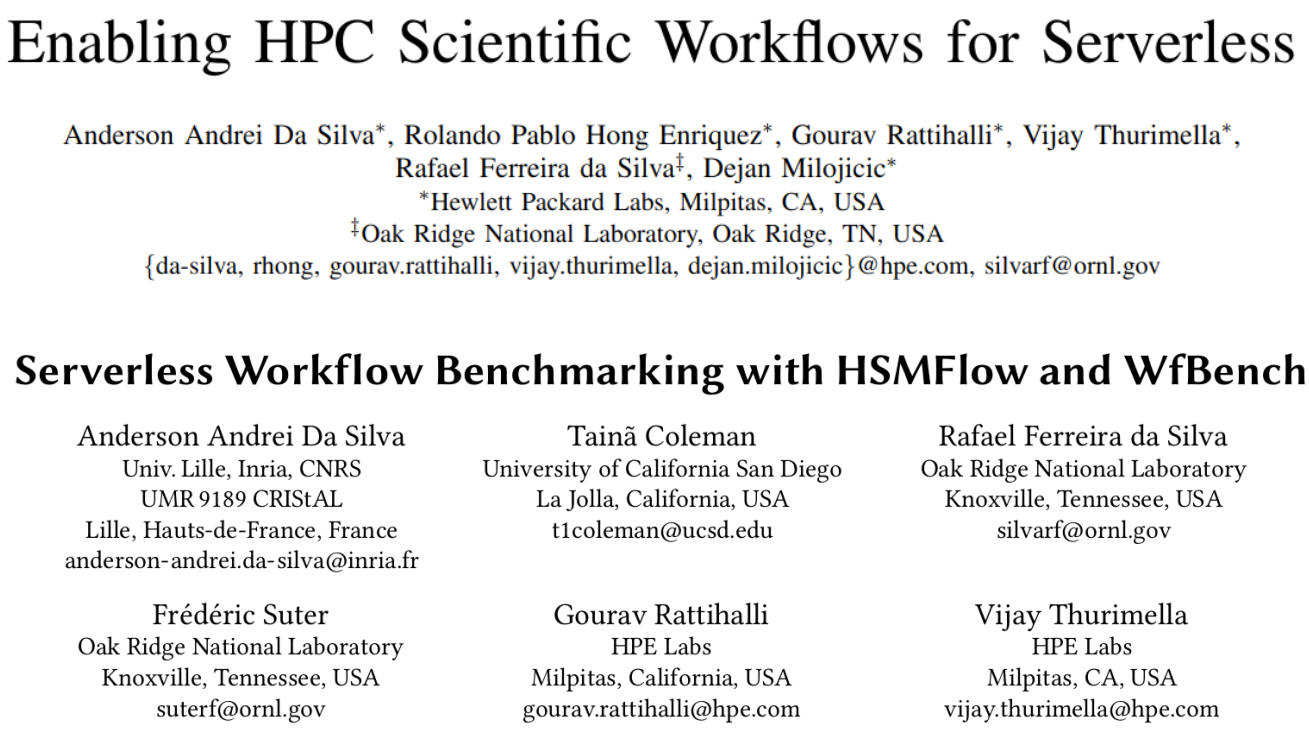
Figure: 8. The effects of concurrency auto-scaling & number of nodes.

Enabling HPC Scientific Workflows for Serverless

Motivation:

1. **Scientific applications have become more complex**, combining different domains and tasks;
2. Their **deployment and management** have become more challenging;
3. **Resource provisioning** has also become more challenging.

Proposed solutions:



Enabling HPC Scientific Workflows for Serverless

Anderson Andrei Da Silva*, Rolando Pablo Hong Enriquez*, Gourav Rattihalli*, Vijay Thurimella*,
Rafael Ferreira da Silva‡, Dejan Milojicic*
*Hewlett Packard Labs, Milpitas, CA, USA
‡Oak Ridge National Laboratory, Oak Ridge, TN, USA
{da-silva, rhong, gourav.rattihalli, vijay.thurimella, dejan.milojicic}@hpe.com, silvarf@ornl.gov

Serverless Workflow Benchmarking with HSMFlow and WfBench

Anderson Andrei Da Silva
Univ. Lille, Inria, CNRS
UMR 9189 CRIStAL
Lille, Hauts-de-France, France
anderson-andrei.da-silva@inria.fr

Tainã Coleman
University of California San Diego
La Jolla, California, USA
t1coleman@ucsd.edu

Rafael Ferreira da Silva
Oak Ridge National Laboratory
Knoxville, Tennessee, USA
silvarf@ornl.gov

Frédéric Suter
Oak Ridge National Laboratory
Knoxville, Tennessee, USA
suterf@ornl.gov

Gourav Rattihalli
HPE Labs
Milpitas, California, USA
gourav.rattihalli@hpe.com

Vijay Thurimella
HPE Labs
Milpitas, CA, USA
vijay.thurimella@hpe.com

A Framework for HPC Scientific Workflows on Serverless

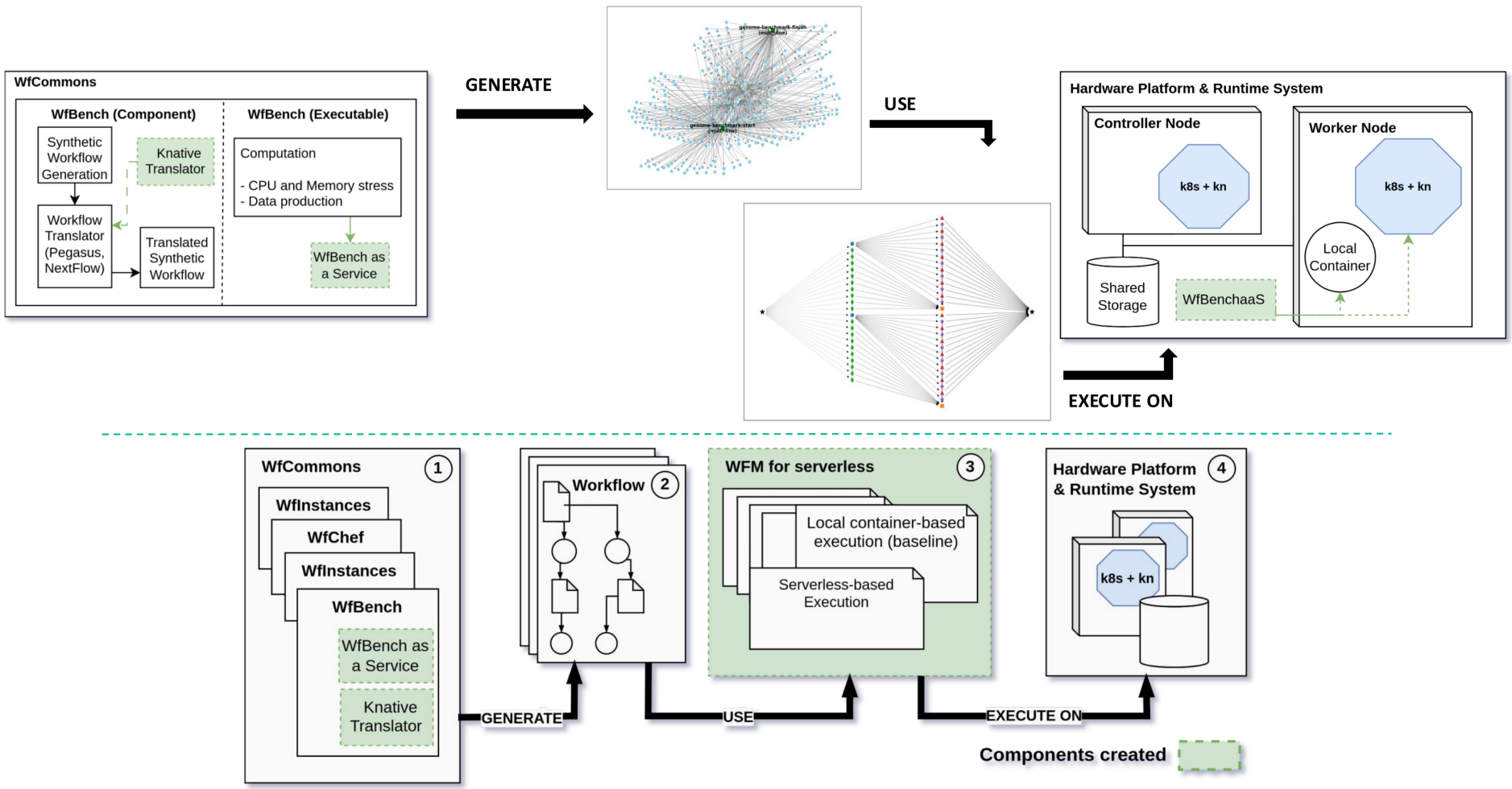


Figure: Architecture of our framework.

Serverless Computing vs Local Containers

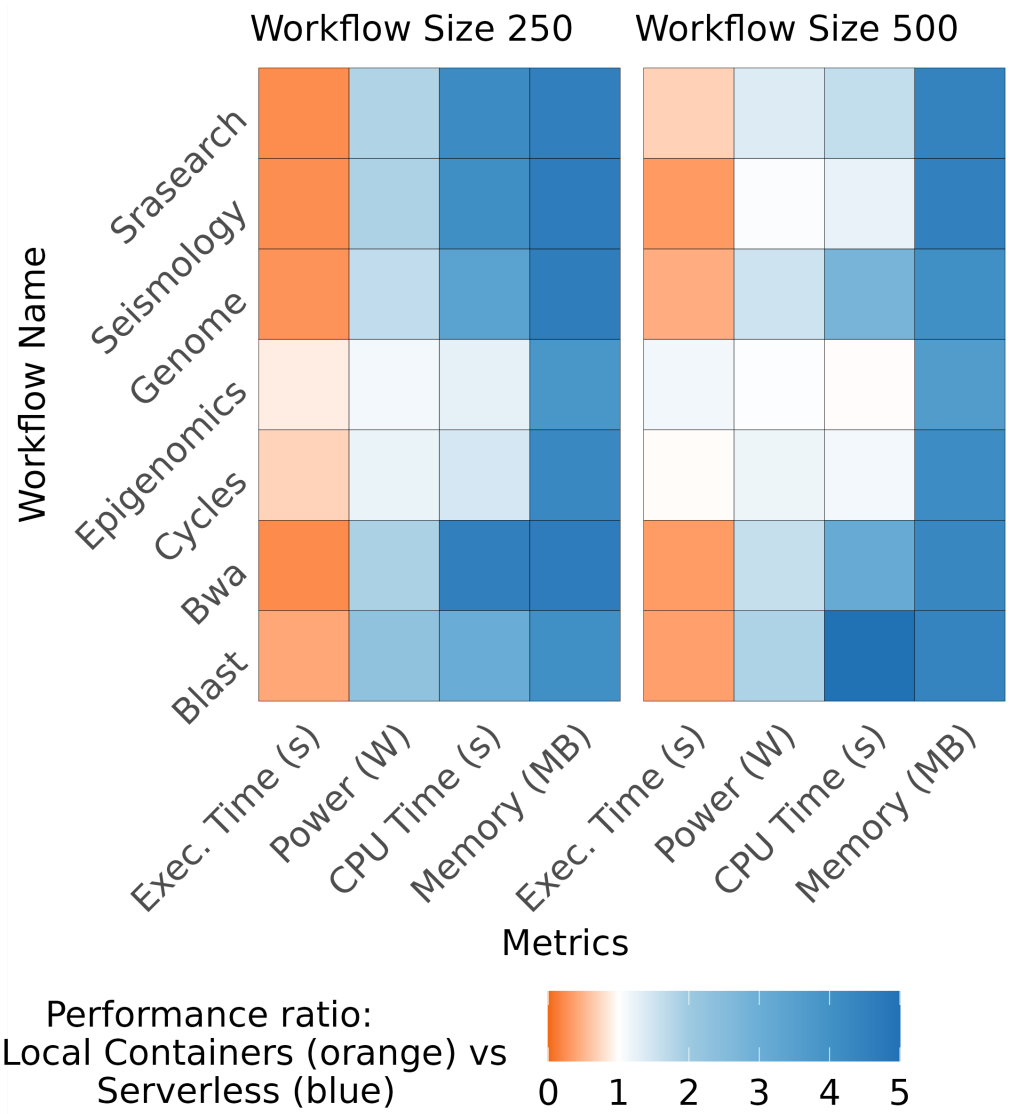


Figure: Comparison between best setups for Serverless and Local Containers.