HeROcarmin

Heterogeneous Resources Orchestration for carbon minimization

March 25, 2025

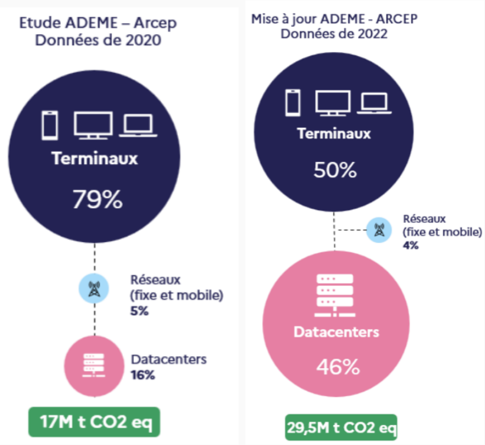
Maëlys Eymond [1]
Vincent Lannurien [1]
Laurent Lemarchand [2]
Jalil Boukhobza [1]

[1] Lab-STICC, CNRS UMR 6285, ENSTA, Institut Polytechnique de Paris
[2] Lab-STICC, CNRS UMR 6285, Université de Bretagne Occidentale

Context

Context – Datacenters and carbon emissions

- Digital technologies: **4.4% of global carbon footprint** in France [1]
- **Datacenters: 2nd source of pollution** among digital technologies
  - As high as 46% of total digital services footprint [2]
- Usage of foreign datacenters:
  - More carbon-intensive electricity grid = worse carbon footprint

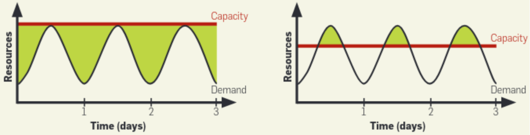


[1] [9] Etienne Lees Perasso, Caroline Vateau, Firmin Domon, et al. *Evaluation de l'impact environnemental du numérique en France et analyse prospective.* Tech. rep. ADEME, 2022

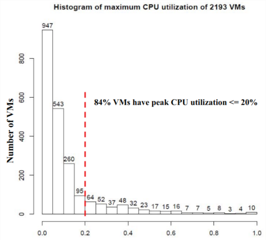
[2] [2] Thomas Brilland et al. *Evaluation de l'impact environnemental du numérique en France.* Tech. rep. ADEME, 2025

Context – Datacenters and resource usage

- **Datacenter rightsizing problem** [3]:
  - Should be able to handle peak load (QoS requirements): *over provisioning*
  - Should minimize idle resources (energy efficiency): *under utilization*
- **Datacenter solution**: *over committing*
  - Risk: QoS violations



- Cloud resource usage:
  - Generally considered **<15%** [4] on avg.
  - 10 to 30% of servers in US datacenters are **idle** [5]
  - Idle hosts consume around **60% of their peak power** [6]



[3] [1] Michael Armbrust et al. "A View of Cloud Computing". *Commun. ACM* (2010)

[4] [14] Arunchandar Vasan et al. "Worth Their Watts? - An Empirical Study of Datacenter Servers". HPCA '16. 2010

[5] [11] Arman Shehabi et al. *United States Data Center Energy Usage Report*. Tech. rep. 2016

[6] [8] Massoud Pedram. "Energy-Efficient Datacenters". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2012)

Context – Serverless computing

- Cloud service models: **renting** (a VM, a container, …)
- **Serverless** service model [7]:
  - Allocation unit: **time** instead of **resource**
  - Fine granularity: **user request**
  - Responsibility: shifting from **cloud user** to **cloud provider**
- Opportunity for service providers:
  - Optimize the platform for QoS metrics
    - With dynamic resource management
  - **Minimize energy consumption**

[7] [10] Johann Schleier-Smith et al. "What Serverless Computing is and Should Become: The next Phase of Cloud Computing". *Commun. ACM* (2021)

Context – Problem statement

Research Question 1

How to estimate the carbon fooprint for a serverless platform?

Research Question 2

How to optimize the underlying infrastructure for carbon footprint and Quality of Service?

Context – Problem statement

Research Question 1

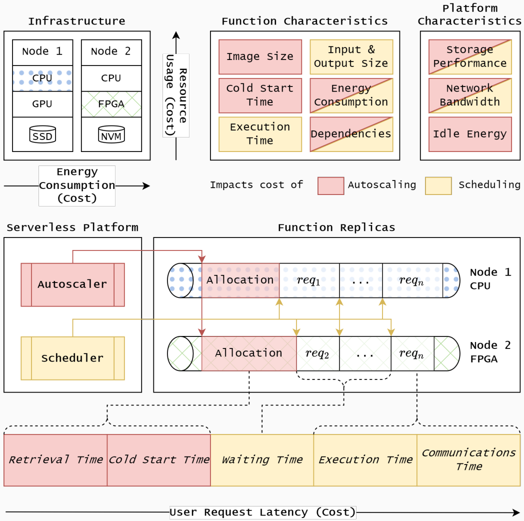
How to estimate the carbon fooprint for a serverless platform?

Research Question 2

How to optimize the underlying infrastructure for carbon footprint and Quality of Service?

Background

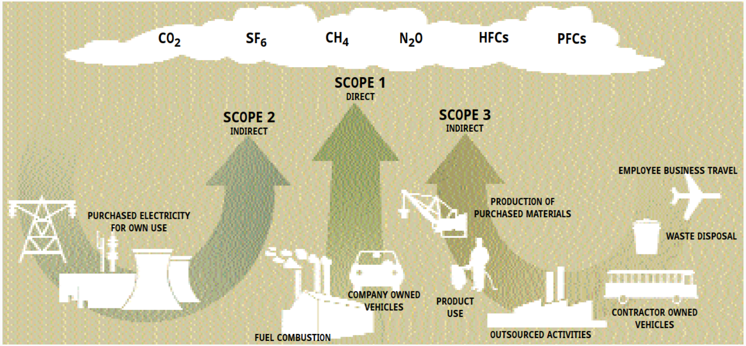
- Idle resources impact global platform cost [8]
  - Consolidating workloads decreases dynamic energy consumption
  - But servers are still powered on and draw static power



[8] [6] Vincent Lannurien. *HeROsim: An Allocation and Placement Simulator for Evaluating Serverless Orchestration Policies.* 2024

Background – Carbon footprint modeling

- Scoped footprint [9]: direct, indirect
- Life Cycle Analysis [10]: evaluate impact thoughout the product's lifecycle (production, distribution, use, end-of-life)
- Metric: CO_2eq [11]



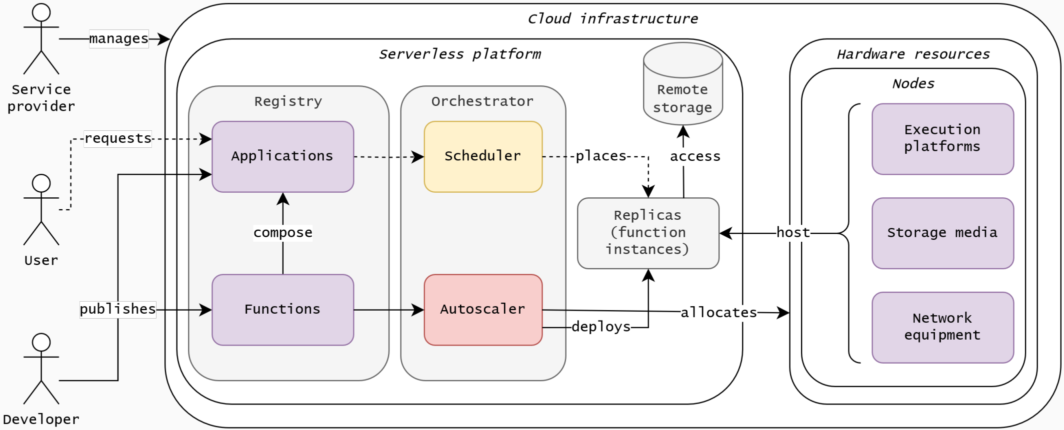
[9] [12] . *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard*. revised ed. 2004

[10] [4] Mehmet Ali Ilgin and Surendra M. Gupta. "Environmentally conscious manufacturing and product recovery (ECMPRO): A review of the state of the art". *Journal of Environmental Management* (2010). ISSN: 0301-4797

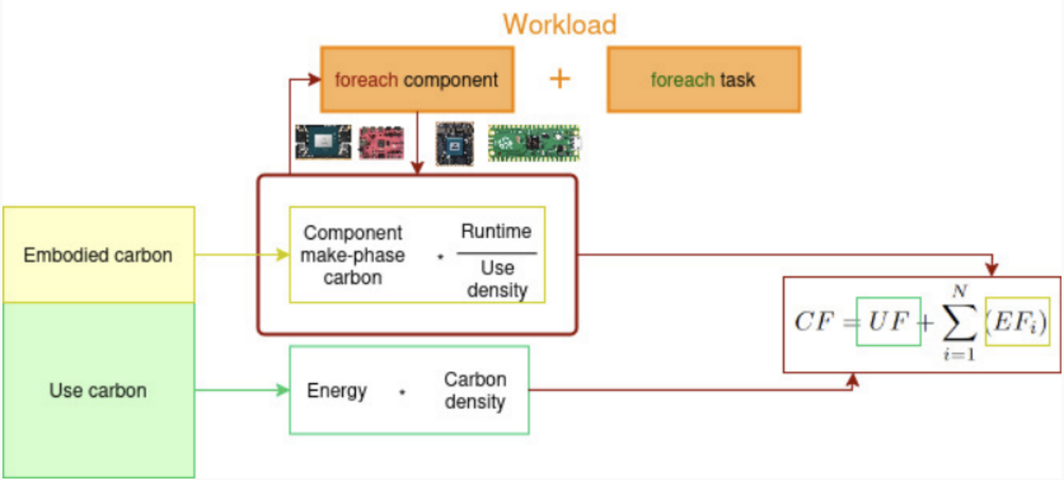
[11] [13] OAR US EPA. *Greenhouse Gas Equivalencies Calculator*. https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator. Data and Tools. 2015

Contribution

HeROcarmin – Contribution: Considered infrastructure

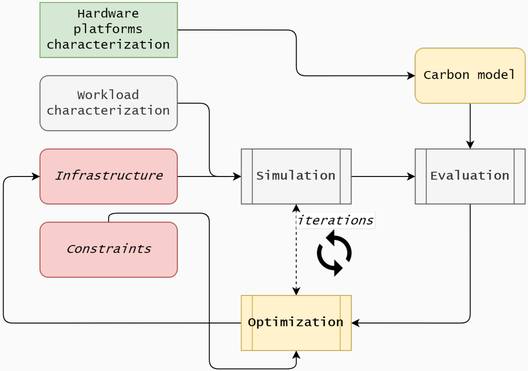


HeROcarmin – Contribution: Serverless carbon footprint

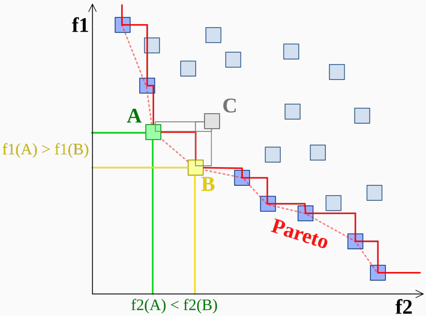


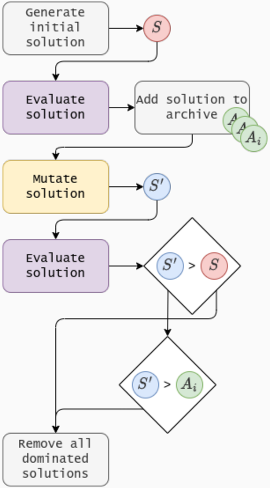
HeROcarmin – Contribution: Optimization loop

1. Characterization: Workloads and execution platforms
2. HeROcarmin loop:
   2.1 *Optimization module* – Generate an infrastructure
   2.2 *Simulation module* – Run a scenario through various orchestration policies
      - Input data: Infrastructure description (node count w/ selection of hardware platforms)
      - Output data: QoS metrics and carbon footprint
   2.3 *Optimization module* – Evaluate infrastructre quality



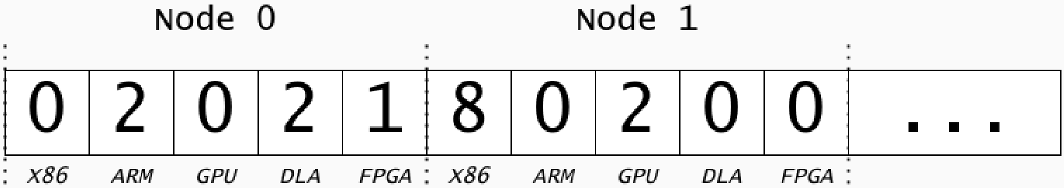
Metaheuristic method: PAES [12] (*Pareto Archived Evolution Strategy*)





[12] [5] J. Knowles and D. Corne. "The Pareto Archived Evolution Strategy: A New Baseline Algorithm for Pareto Multiobjective Optimisation". 1999

[13] Johann Dréo, "Optimum de Pareto", 2006. `https://commons.wikimedia.org/wiki/File:Front_pareto.svg`



- Mutations :
  - **Add** a node (made of random platforms)
  - **Remove** a (random) node
  - **Increase** (random) resource type count on a (random) node
  - **Decrease** (random) resource type count on a (random) node

HeROcarmin – Contribution: Some results (WIP)

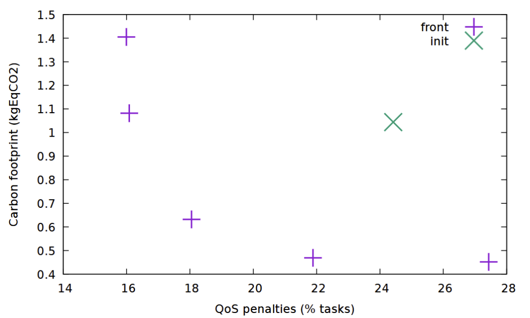


Figure 1: 10-minute scenario

- How does that compare? [14]
    - Average French person: 10 TEqCO2/year
    - Round-trip Paris - New York: 2 TEqCO2
    - Smartphone: 86 kgEqCO2
    - 1 kg rice: 3.36 kgEqCO2
    - Wooden chair: 2 kgEqCO2
    - **4.6 km by car: 1 kgEqCO2**

[14] [7] Tristan Nitot. *Quels sont les rapports entre numérique et changement climatique ?* 2022

Perspectives

Limits – Short term

- What has been done: ground work
  - DONE Carbon model implementation in simulation
  - DONE Optimization loop using PAES
- What must be done:
  - FIXME Compare outcomes across orchestration policies
  - FIXME Configurable constraints to be enforced in solutions at runtime
    - Carbon budget for the services provider
    - Required average QoS level across user requests
  - TODO Improve the carbon footprint model
    - Network communications
    - Storage footprint (over time)

Limits – Future work

- Geographic dimension:
  - Pollution through high electricity use (approx. 4% US electricity in 2023)
  - Sometimes sourced from fossil fuels; sometimes ”clean”
  - But datacenters are also backed by diesel generators (outages)
  - Chemical waste from hardware manufacturing

- Cost for society: not everyone pays the same toll
  - Health impacts in the US forecast to approx. 20 bn USD/year by 2030 [15]
  - Heavily skewed towards low-income populations due to datacenters location (West Virginia, Ohio)
  - Digital products contribute to reaching planetary boundaries: metals, minerals, water...

[15] [3] Yuelin Han et al. *The Unpaid Toll: Quantifying the Public Health Impact of AI*. 2024

Thank you.

References

[1] Michael Armbrust et al. “A View of Cloud Computing”. *Commun. ACM* 53.4 (Apr. 2010), pp. 50–58. DOI: `10.1145/1721654.1721672`.

[2] Thomas Brilland et al. *Evaluation de l’impact environnemental du numérique en France*. Tech. rep. ADEME, 2025.

[3] Yuelin Han et al. *The Unpaid Toll: Quantifying the Public Health Impact of AI*. Dec. 2024. DOI: `10.48550/arXiv.2412.06288`. arXiv: `2412.06288 [cs]`. (Visited on 02/24/2025).

[4] Mehmet Ali Ilgin and Surendra M. Gupta. “Environmentally conscious manufacturing and product recovery (ECMPRO): A review of the state of the art”. *Journal of Environmental Management* 91.3 (2010), pp. 563–591. ISSN: 0301-4797. DOI: `https://doi.org/10.1016/j.jenvman.2009.09.037`. URL: `https://www.sciencedirect.com/science/article/pii/S0301479709003417`.

[5] J. Knowles and D. Corne. “The Pareto Archived Evolution Strategy: A New Baseline Algorithm for Pareto Multiobjective Optimisation”. *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*. Washington, DC, USA: IEEE, 1999, pp. 98–105. ISBN: 978-0-7803-5536-1. DOI: `10.1109/CEC.1999.781913`. (Visited on 03/23/2025).

[6] Vincent Lannurien. *HeROsim: An Allocation and Placement Simulator for Evaluating Serverless Orchestration Policies*. Feb. 2024.

[7] Tristan Nitot. *Quels sont les rapports entre numérique et changement climatique ?* Lyon, 2022.

[8] Massoud Pedram. “Energy-Efficient Datacenters”. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31.10 (Oct. 2012), pp. 1465–1484. DOI: `10.1109/TCAD.2012.2212898`.

[9] Etienne Lees Perasso, Caroline Vateau, Firmin Domon, et al. *Evaluation de l’impact environnemental du numérique en France et analyse prospective*. Tech. rep. ADEME, 2022.

[10] Johann Schleier-Smith et al. "What Serverless Computing is and Should Become: The next Phase of Cloud Computing". *Commun. ACM* 64.5 (Apr. 2021), pp. 76–84. DOI: 10.1145/3406011.

[11] Arman Shehabi et al. *United States Data Center Energy Usage Report*. Tech. rep. LBNL–1005775, 1372902. June 2016, LBNL–1005775, 1372902. DOI: 10.2172/1372902.

[12] Bryan Smith, World Business Council for Sustainable Development, and World Resources Institute, eds. *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard*. revised ed. o.O: World Resources Institute and World Business Council for Sustainable Development, 2004. ISBN: 978-1-56973-568-8.

[13] OAR US EPA. *Greenhouse Gas Equivalencies Calculator*. https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator. Data and Tools. Aug. 2015. (Visited on 03/25/2025).

[14] Arunchandar Vasan et al. "Worth Their Watts? - An Empirical Study of Datacenter Servers". *HPCA - 16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture*. HPCA '16. Bangalore: IEEE, Jan. 2010, pp. 1–10. ISBN: 978-1-4244-5658-1. DOI: 10.1109/HPCA.2010.5463056.