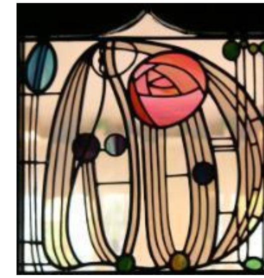


Carbon-Conscious Scalable Data Analysis w/ the *Ichnos* Carbon Footprint Estimator

FONDA Workshop – 25 Nov 2025

Dr Lauritz Thamsen

University of Glasgow

https://lauritzthamsen.org

General Motivation

- **Computing's carbon footprint** is rising rapidly
- **Big data analytics** are routinely identified as **one driver** of computing's rising emissions
- There is often **limited insight** into the footprint of specific applications

Carbon-Conscious Computing Lab at the University of Glasgow

Computer systems research on

- **Performance profiling & prediction**
- **Adaptive resource management**
- **Carbon-aware computing**
- **Carbon footprint estimation**

for data-intensive systems on distributed compute infrastructure

Website: https://lauritzthamsen.org/lab/

Lauritz Thamsen

Kathleen West

James Nurdin

Youssef Moawad

Max MacDonald

Tobias Fröhlich

Acknowledgements

- This talk presents **joint work** with:
  - Kathleen West, Youssef Moawad,
    Magnus Reed, Yehia Elkhatib – UofG
  - Vasilis Bountris, Philipp Thamm, Ulf Leser – HU Berlin
  - Giulio Attenni – Sapienza Rome

- Initial results were presented at the **1st International Workshop on Low Carbon Computing** (LOCO 2024)

- The work was supported by **EPSRC** (UKRI154)

Starting Point: Linear Power Models

- Many carbon footprint assessment methodologies (e.g. CCF and GA) **estimate** energy consumption based on resource utilization using **linear power models**

- Relatedly, much of the **research on energy-efficient and carbon-aware scheduling** – including ours – relies on such methodologies

Alternatives: Monitoring & Estimating

- Workflow and analytics systems do not automatically track energy and emissions, leaving two options:

1. *Monitor* energy consumption (and then translate to emissions)

1. Record energy usage using external or use built-in power meters (e.g. RAPL)
2. Use carbon intensity data to estimate emissions

- Requires setup **before** application execution
- May not have access to power meters on **shared resources**

2. *Estimate* energy consumption (and then translate to emissions)

1. Record resource utilization metrics (e.g. CPU usage)
2. Use methodologies (e.g. GA or CCF) to estimate energy consumption using resource utilization metrics and carbon intensity data

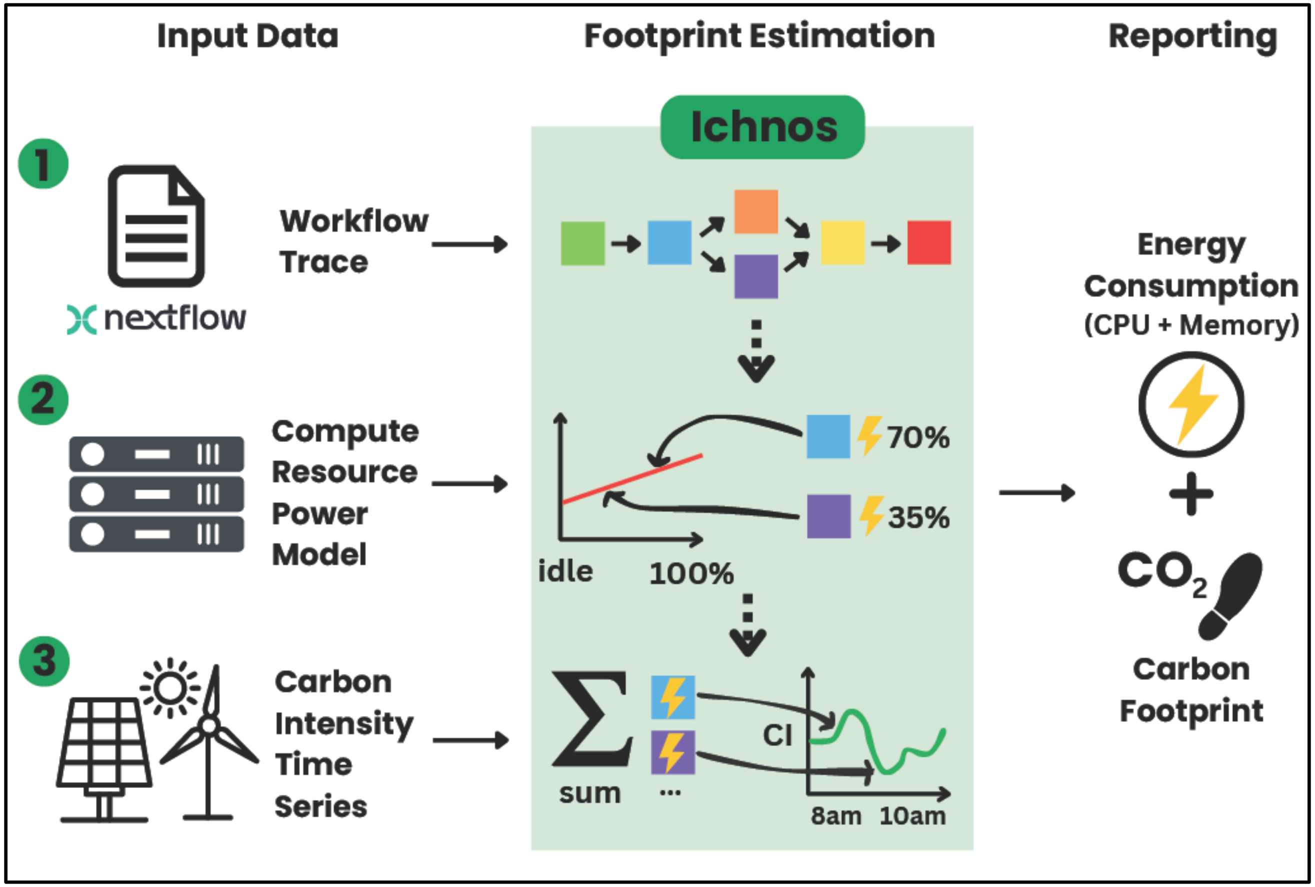
- Can often be done **afterwards** and **without full access**

Our Own Estimator and Experiments

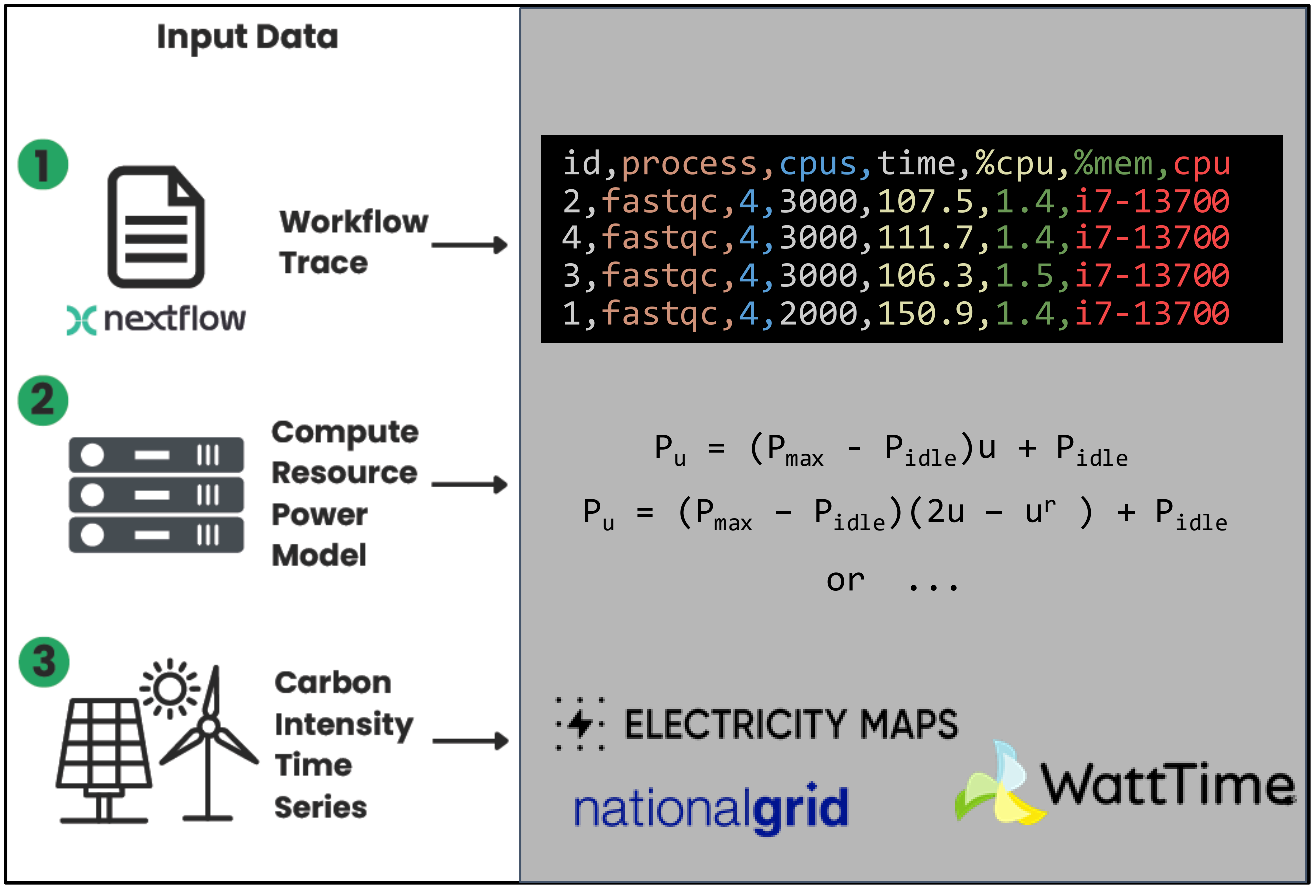
- Objective 1: **Build an estimator tool – Ichnos –** that works for Nextflow (and possibly other systems)
  - Enabling post-hoc estimation, using resource utilization data (e.g. from existing traces)
  - Estimating CPU and memory energy consumption as well as operational emissions at task-level

- Objective 2: **Understand how accurate resource utilization-based estimates** are for our workloads
  - How accurate are these estimates for scalable data analysis on compute clusters?
  - Can estimates be improved without requiring access to low-level hardware counters or measurement devices?

Ichnos: Design



Ichnos: Input Data

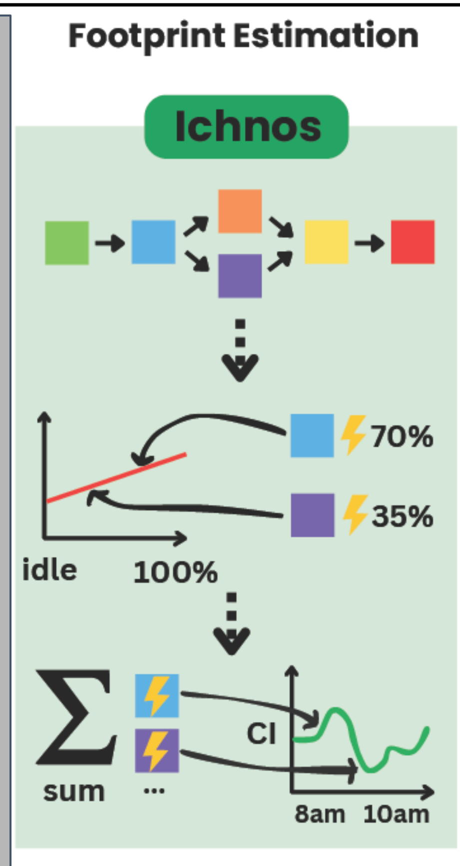


Ichnos: Footprint Estimation

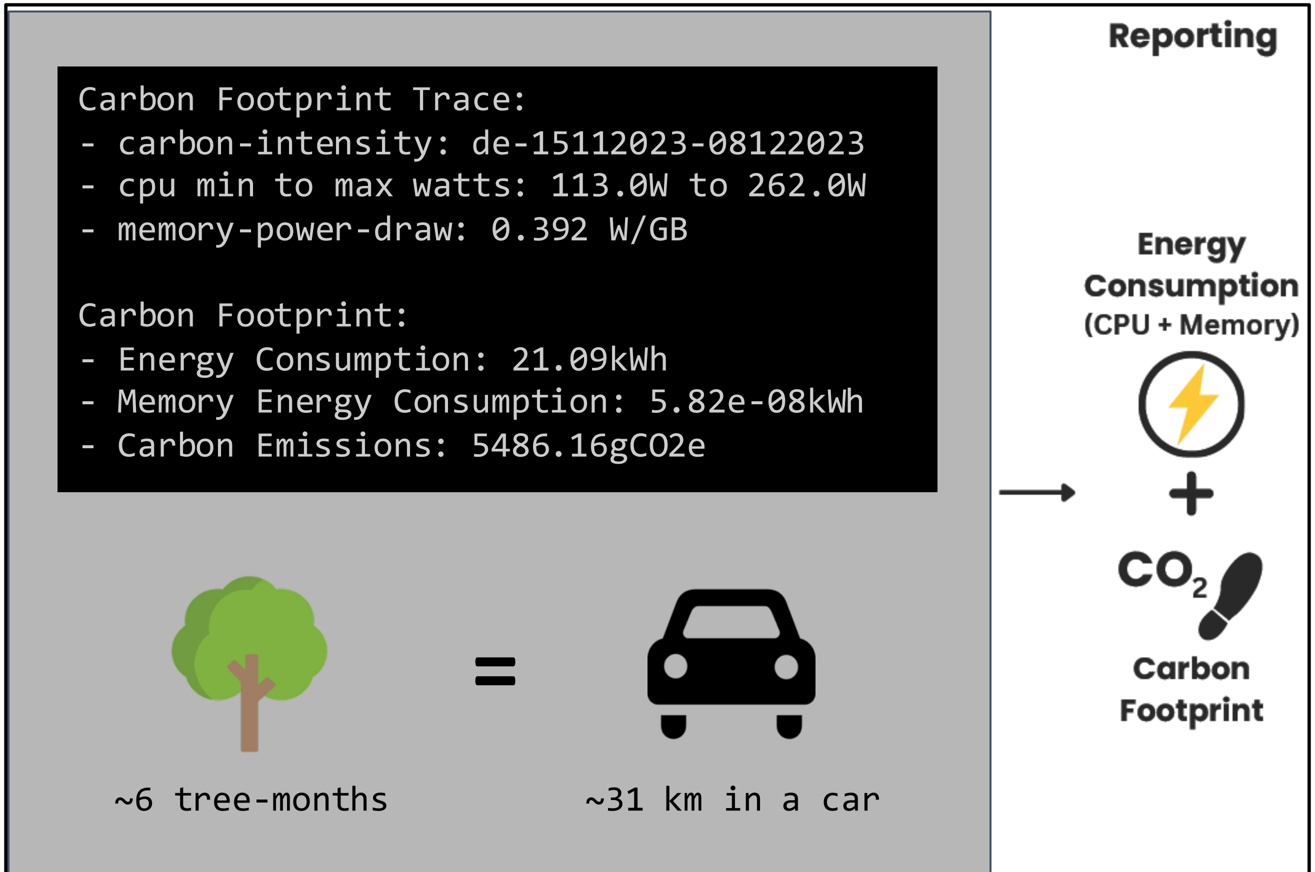
extract resource usage data

energy consumption estimation

translate consumption into carbon emissions

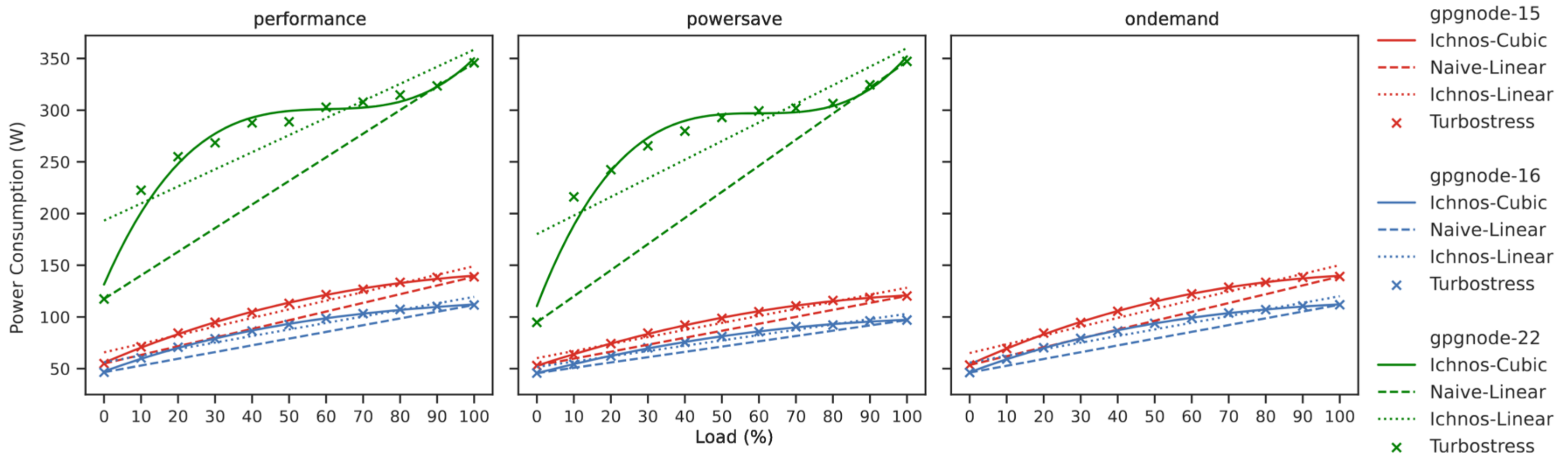


Ichnos: Outputs



Experiments: Power Model Fitting

- **RAPL measurements** taken at **different CPU loads** on cluster nodes, using different Intel governors



GPG nodes 01-20: 2 * **Intel Xeon E5-2640 2GHz**, 64Gb RAM, 2 HDDs

GPG nodes 21-22: 2 * **Intel Xeon Gold 6426Y 2.5GHz**, 128 Gb RAM, 1 SSD + 2 HDDs

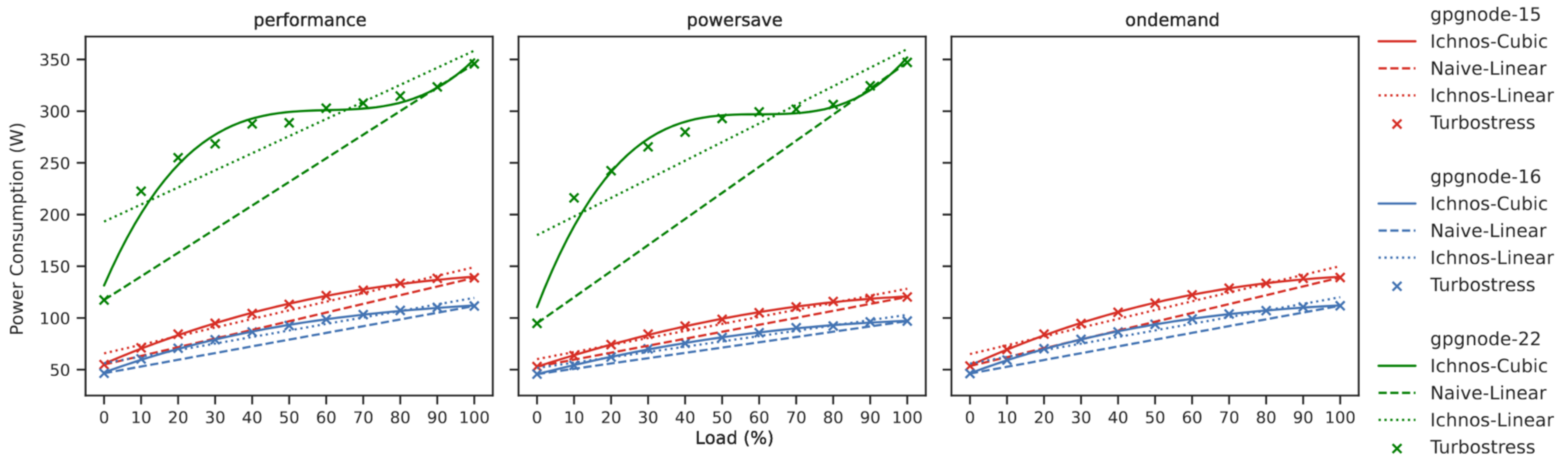
Experiments: Estimates Using Models

- **Estimating nf-core Ampliseq's** energy consumption using the different power models on our cluster nodes

Node	Governor	Perf (kWh)	Ichnos−Cubic (kWh)	Error (%)	Ichnos−Linear (kWh)	Error (%)	Naive−Linear (kWh)	Error (%)	GA (kWh)	Error (%)
gpgnode-13	ondemand	0.161	0.135	16.1	0.144	10.3	0.121	24.7	0.28	82.9
gpgnode-14	performance	0.161	0.138	14.2	0.146	9.1	0.124	22.8	0.026	83.7
"	powersave	0.159	0.143	9.8	0.150	5.6	0.136	14.4	0.029	81.4
gpgnode-15	performance	0.168	0.147	12.4	0.155	7.4	0.134	19.9	0.027	83.6
"	powersave	0.178	0.157	11.7	0.165	7.3	0.148	16.7	0.031	82.4
gpgnode-16	ondemand	0.139	0.124	10.8	0.131	5.4	0.113	18.8	0.026	81.4
gpgnode-22	performance	0.165	0.131	20.7	0.159	3.9	0.101	38.7	0.003	98.0
"	powersave	0.163	0.031	81.0	0.150	8.0	0.085	47.9	0.003	98.0

The Impact of Non-Linearity

Again, 11 RAPL data points fitted for the GPG nodes:

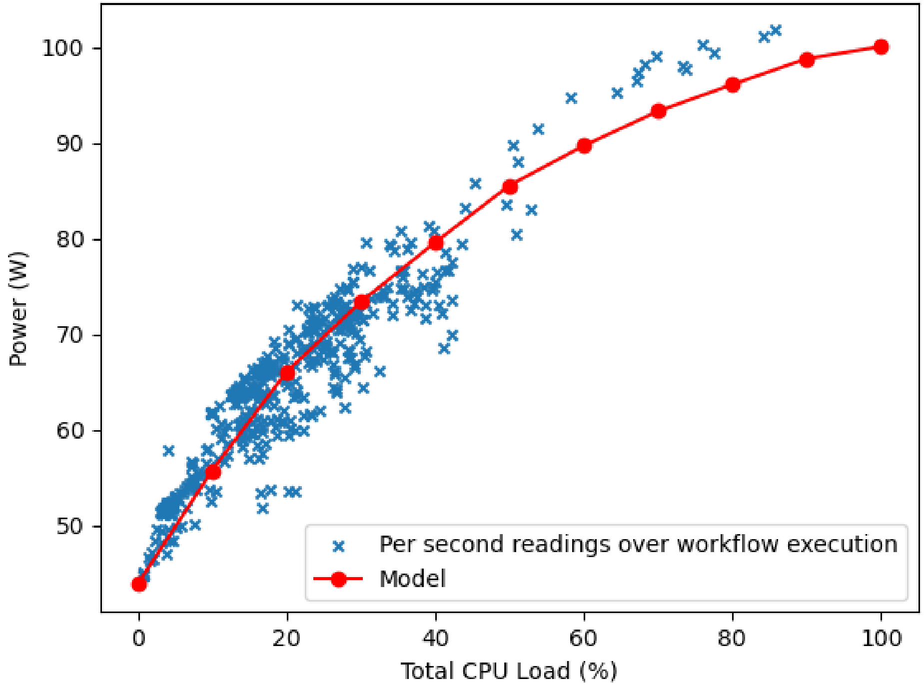


→ Issue 1: Using non-linear models with **coarse-grained utilization averages**

→ Issue 2: Non-linear power draw depends on **overall load on shared resources**

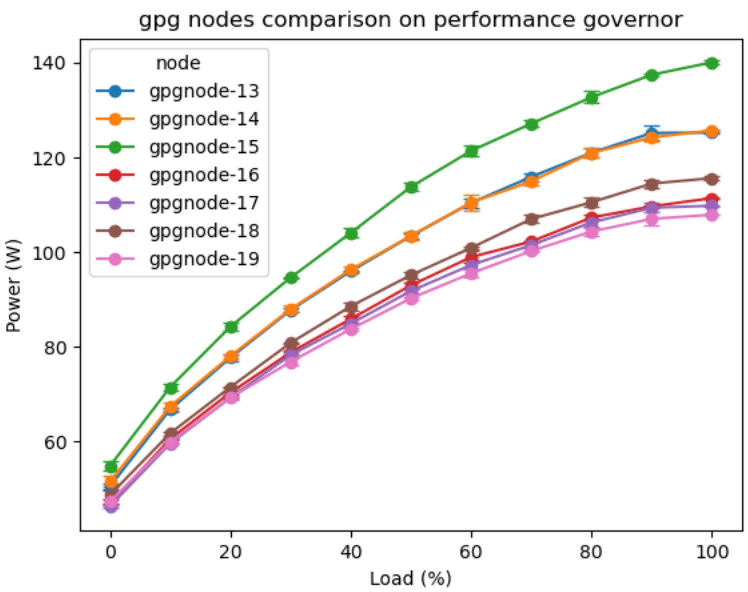
Model Accuracy Is Inherently Limited

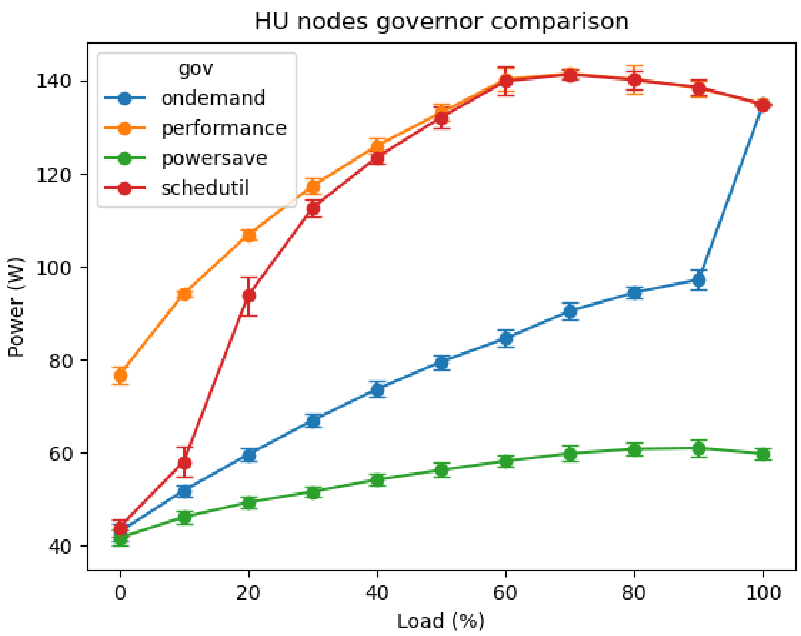
- CPU power draw depends on more than utilization, so utilization-based estimation accuracy is inherently limited



Why Still Fit Power Models?

- Power draw of even homogeneous nodes and of different processor settings can vary drastically





More Results: Distributed Execution

- Various workflows executed over multiple nodes on both the GPG and HU clusters:

cluster	workflow	run	ichnos (kWh)	rapl (kWh)	error (%)
hu	rnaseq	1	2.27	2.11	7.13
hu	rnaseq	2	2.27	2.44	6.77
hu	chipseq	1	3.66	3.85	5.1
hu	chipseq	2	3.68	3.87	4.9
hu	sarek	1	3.08	3.05	0.92
hu	sarek	2	3.1	3.09	0.38
hu	rangeland	1	0.45	0.53	15.78
hu	rangeland	2	0.38	0.38	1.58
gu	rnaseq	1	1.59	1.78	10.76

Outlook

- Support for **additional systems** via a general trace format (such as for **Spark**)
- Support for **embodied emissions** estimates (via the Boavizta API)
- Support for **additional impacts** (such as water and land use)

Conclusion

- Linear power models allow estimates **with ≤ 20% error**
- CPU%-based estimates are **inherently limited**, but it is still important to fit **node-specific models**
- Ichnos, as a practical tool, is **ongoing work**

Contact

Lauritz.Thamsen, Kathleen.West,
& Youssef.Moawad @ glasgow.ac.uk

https://lauritzthamsen.org/lab/

https://casperproject.gitlab.io/

LOCO'24 paper:
https://arxiv.org/abs/2411.12456

Ichnos code:
https://github.com/GlasgowC3lab/ichnos