Hes·so

Haute Ecole Spécialisée
de Suisse occidentale

Fachhochschule Westschweiz

University of Applied Sciences and Arts
Western Switzerland

Traffic testbed and Carbon topography representation: **tools** to better measure, understand and analyse server lifecycle impacts

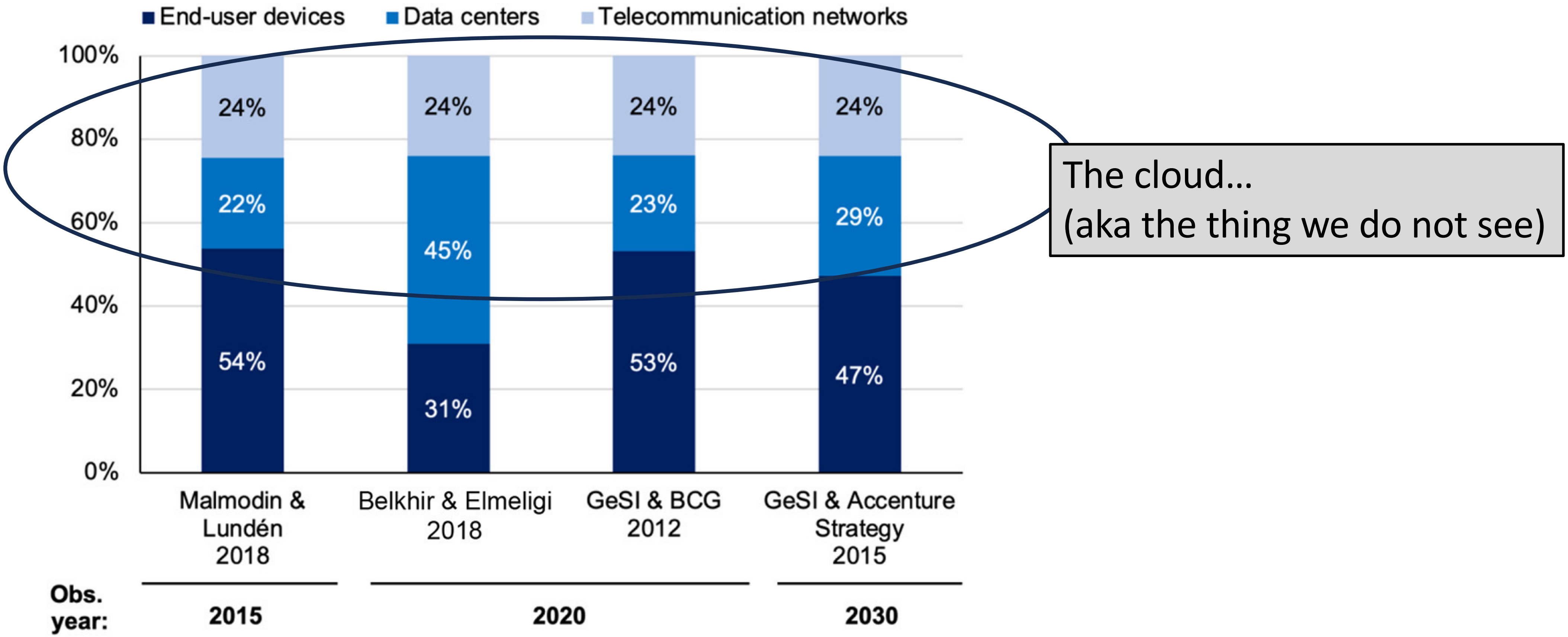
Haute école d'ingénierie et d'architecture Fribourg
Hochschule für Technik und Architektur Freiburg

esos
Électronique Soutenable Ouverte & Souveraine

Loïc Guibert, Olivier Weppe, **Sébastien Rumley**

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Foreword

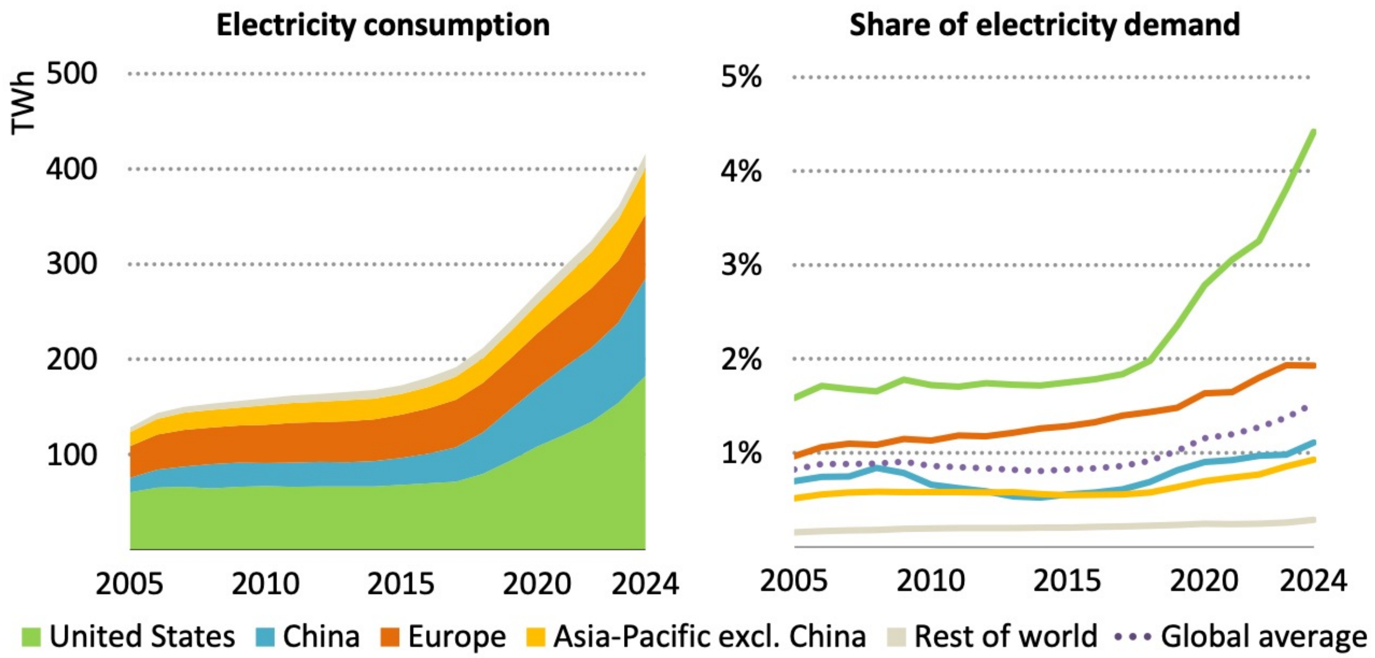


Share of greenhouse gaz emissions across sectors, according to different estimates [1]

[1] Jan C.T. Bieser, Ralph Hintemann, Lorenz M. Hilty, Severin Beucker, A review of assessments of the greenhouse gas footprint and abatement potential of information and communication technology, Environmental Impact Assessment Review, Volume 99, 2023,

Foreword

Figure 2.6 ▷ Electricity consumption of data centres by region, 2005-2024



The acceleration in data centre electricity consumption observed in 2017 was mainly driven by the United States and, to a lesser extent, by China

How-to reduce cloud impacts?

- Less of everything→ yes.. but this is sociology
  - I am a computer scientist (until now...)

There is also, and above all, a need for "primary" measurements, for "bottom-up" experiments.

- High-level (top-down) studies must be integrated with, and grounded in, these low-level experiments.

How many "infrastructure units" per request?

...questions.

...among other things):

- Understand how infrastructures relate to each individual demand and to flows of demand.
- Analyse infrastructure efficiency, and infrastructure impacts

"Ecological impact units" per "infrastructure unit"

Hes·so

Measuring the Impact of a Request Flow

SERVER

CONTROLLER

TRAFFIC: Testbed foR Assessing energy eFFiciency In throughput Computing

Loïc Guibert
HES-SO
Fribourg, Switzerland
loic.guibert@hefr.ch

Sébastien Reynaud
HES-SO
Fribourg, Switzerland
sebastienreynaud@proton.me

Olivier Weppe
INSA Rennes
Rennes, France
olivier.weppe@insa-rennes.fr

Kristjon Ciko
University of Oslo
Oslo, Norway
kristjoc@ifi.uio.no

Michael Welzl
University of Oslo
Oslo, Norway
michawe@ifi.uio.no

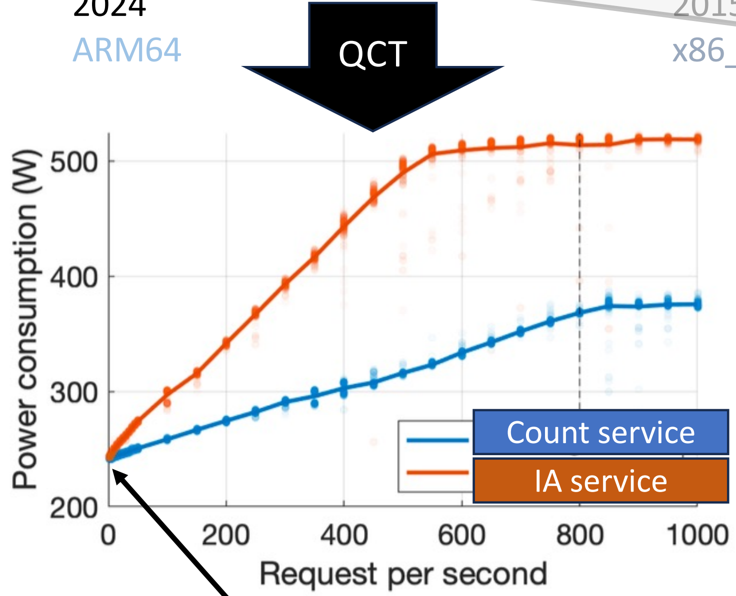
Sébastien Rumley
HES-SO
Fribourg, Switzerland
sebastien.rumley@hefr.ch

impact on energy consumption

1 request to a “BERT” (precursor to modern LLMs) = about 1-4 Joules

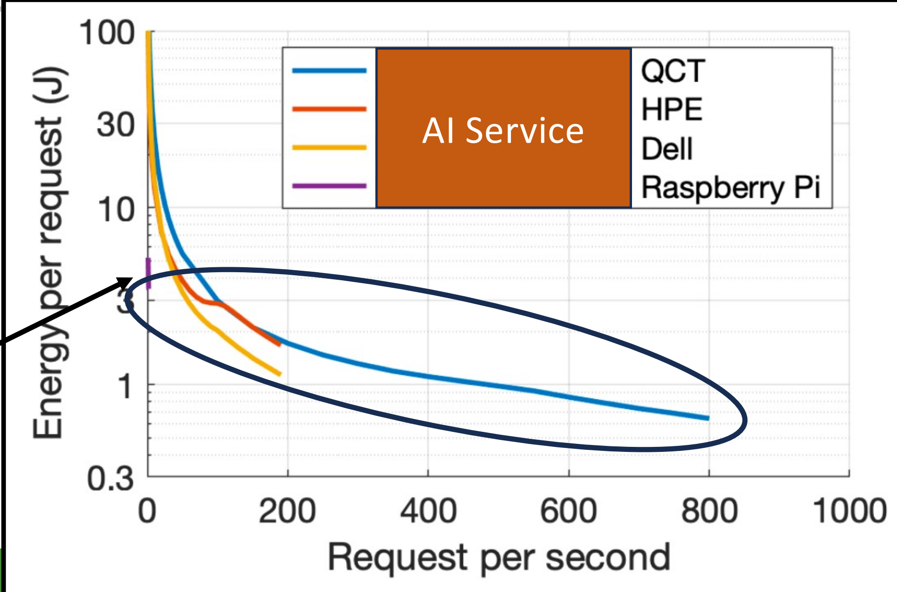
- **But machine must be sufficiently loaded!**
  - **At 200 req/s, the dell tower is more energy efficient**
  - **At 2 req/s, the Pi is the best deal…**

QuantaGrid S74G-2U (QCT) 2024 ARM64 | PowerEdge T430 (Dell) 2015 x86_64 | ProLiant DL360 Gen9 (HPE) 2015 x86_64 | Raspberry Pi 4B 2019 ARM64

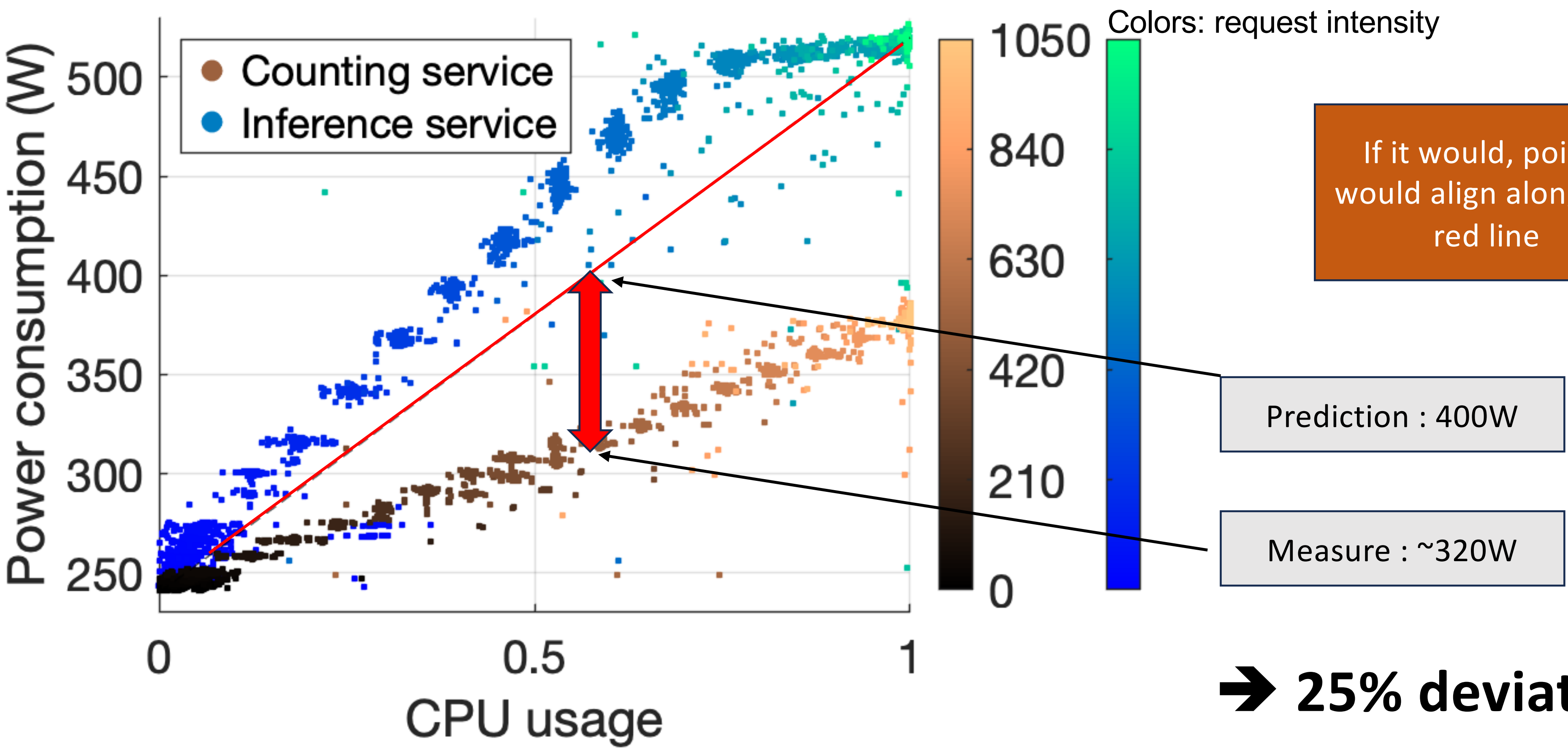


Idle power : 240W

Raspberry PI is less efficient at “its” full load, but efficient for low loads…



CPU usage is no good proxy for consumption!



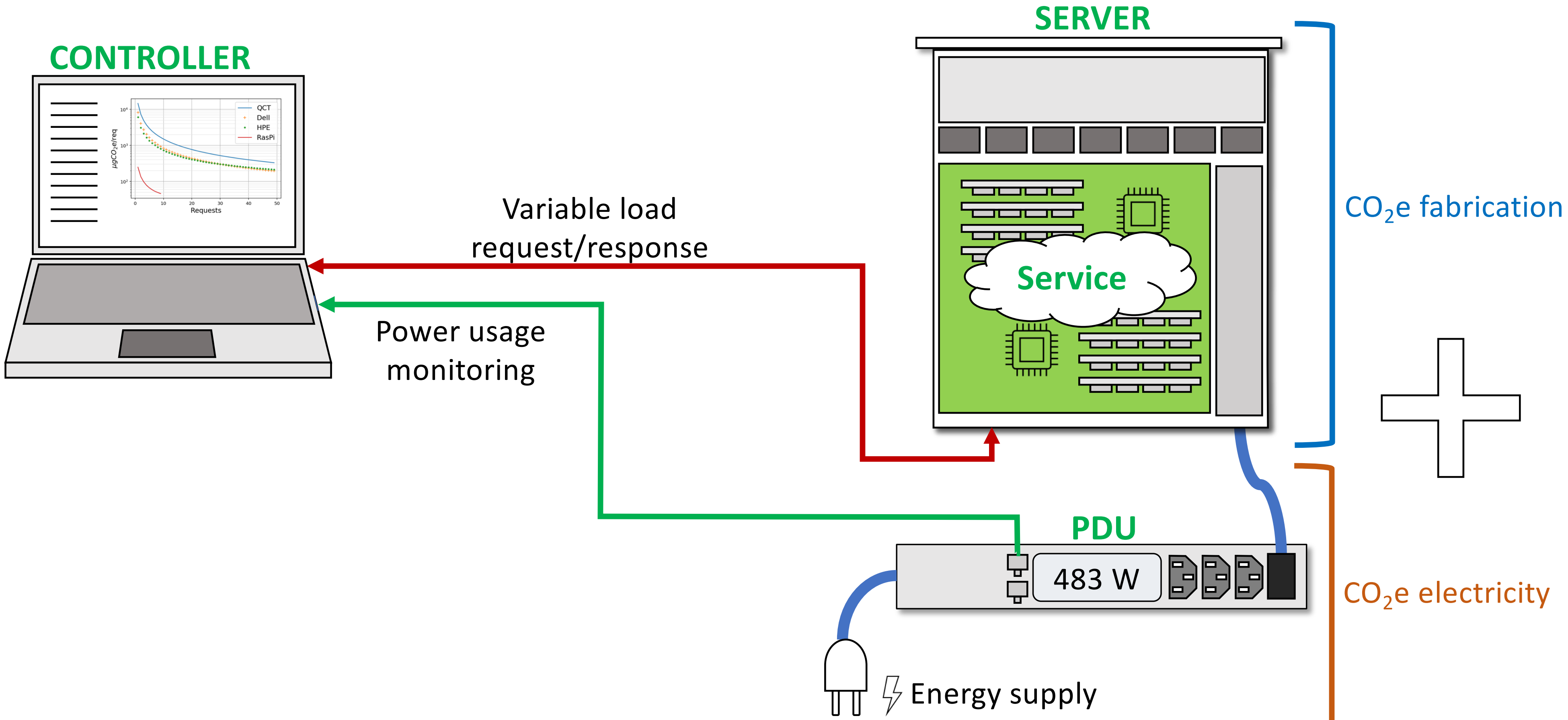
If it would, points would align along the red line

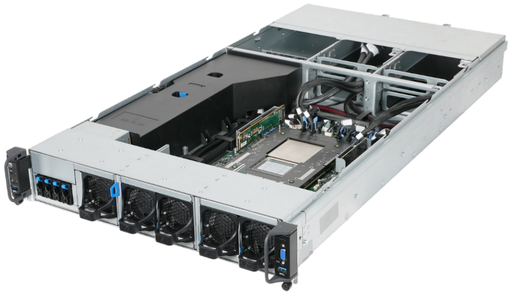
➔ **25% deviation!**

Contribution 1: TRAFFIC testbed proposition

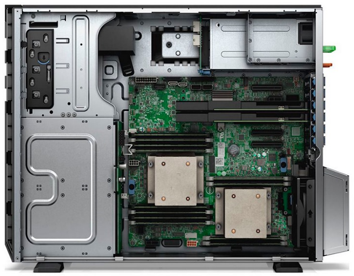
- Measure equipment when facing modulable load (i.e. traffic)
  - Actually, measure equipment ***when providing a service*** at a given load
  - What matters is not hardware efficiency, but “service-level” efficiency
- Relate impacts to service requests → functional units !

Next step: estimate CO_2e





QuantaGrid S74G-2U (QCT)
2024
ARM64
1023 $kgCO_2e$



PowerEdge T430 (Dell)
2015
x86_64
514 $kgCO_2e$

ProLiant DL360 Gen9 (HPE)
2015
x86_64
345 $kgCO_2e$

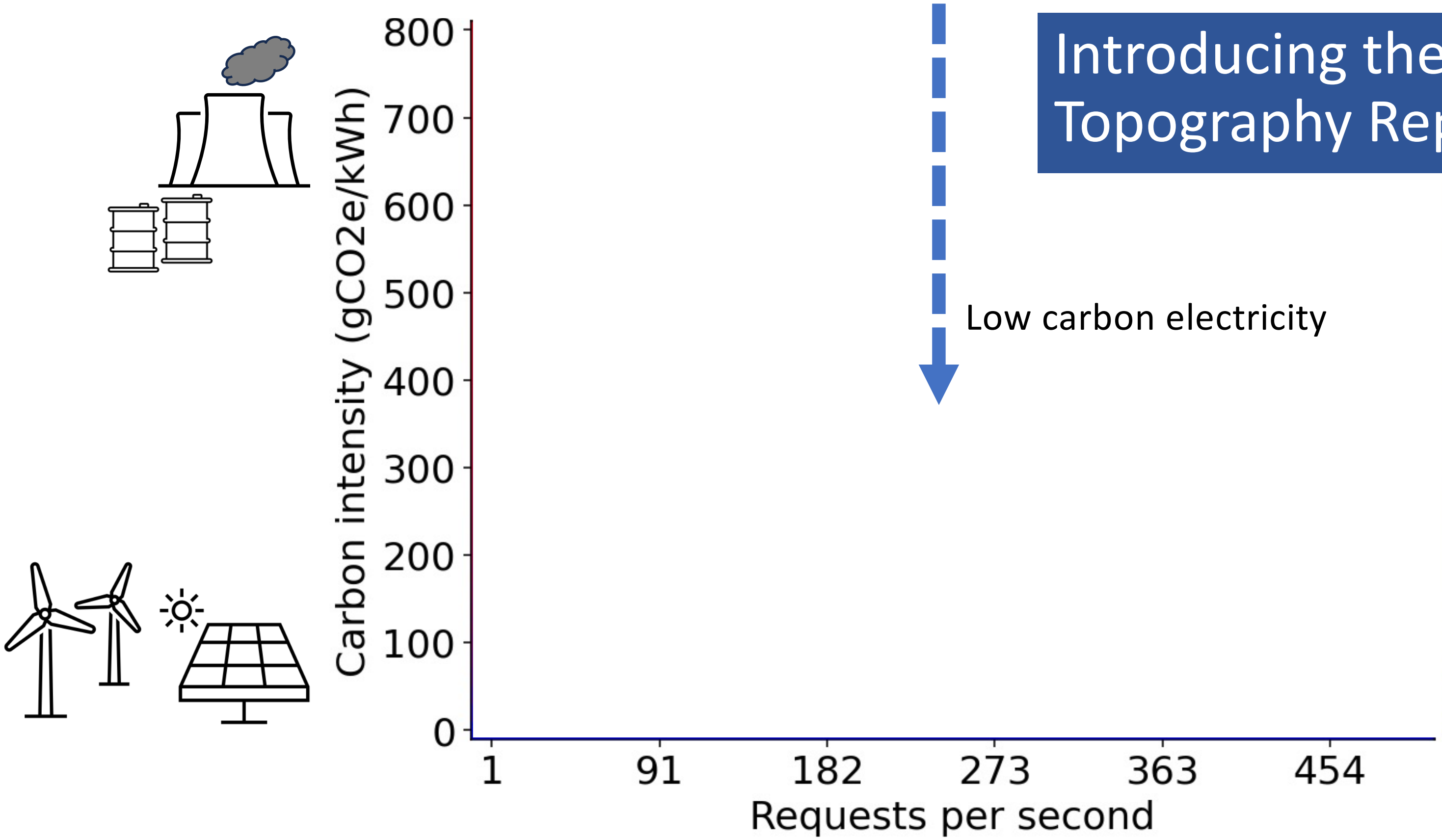
Raspberry Pi 4B
2019
ARM64
14 $kgCO_2e$

We then need to add electricity CO_2e
- For a given “service” load
- For the equipment lifetime

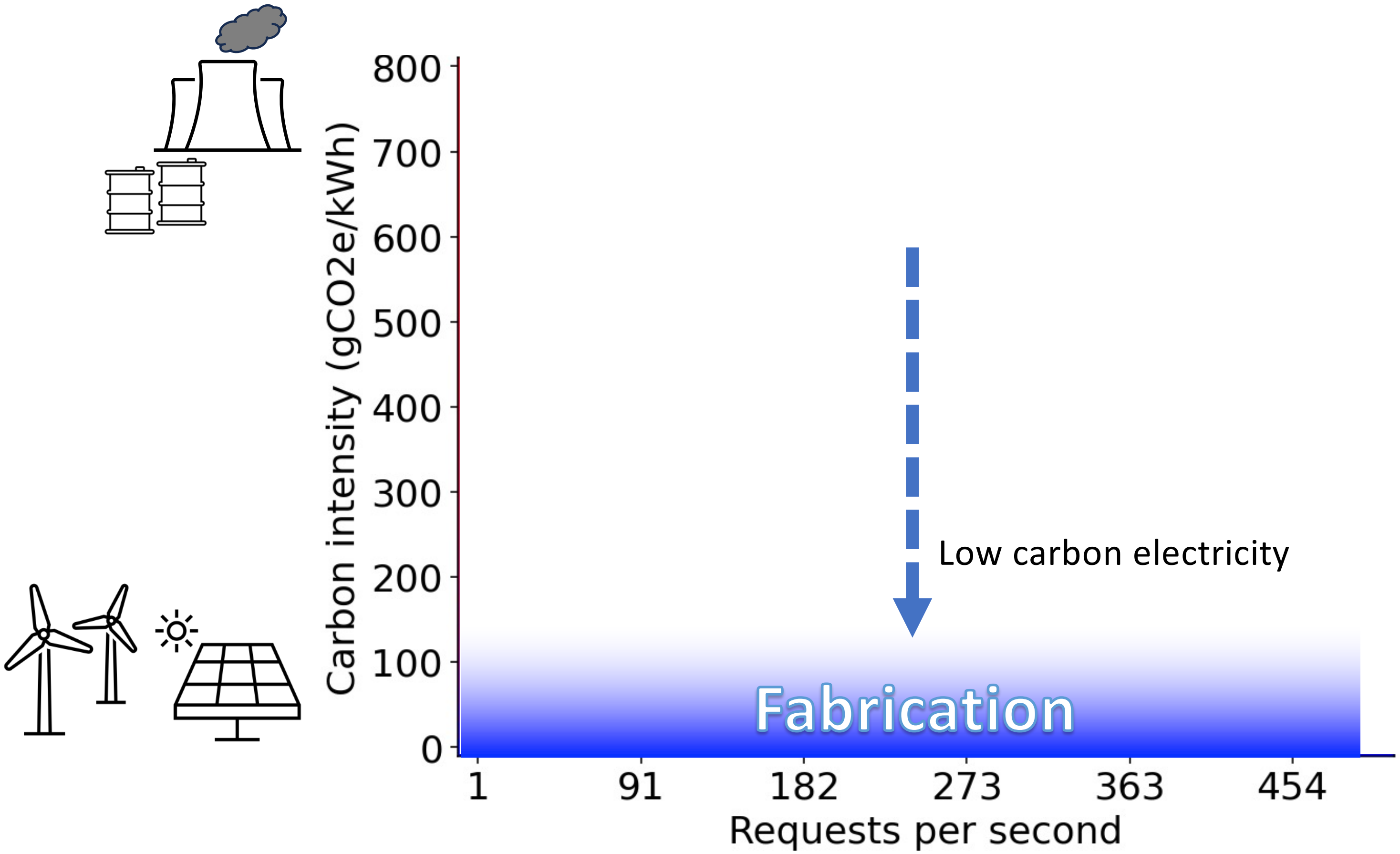
Example :
- QCT, 300 AI req/s = 400W
- 400W * 5 years = 17’520 kWh
- 17’520 kWh * 0.1 $kgCO_2e/kWh$ = 1’752 $kgCO_2e$

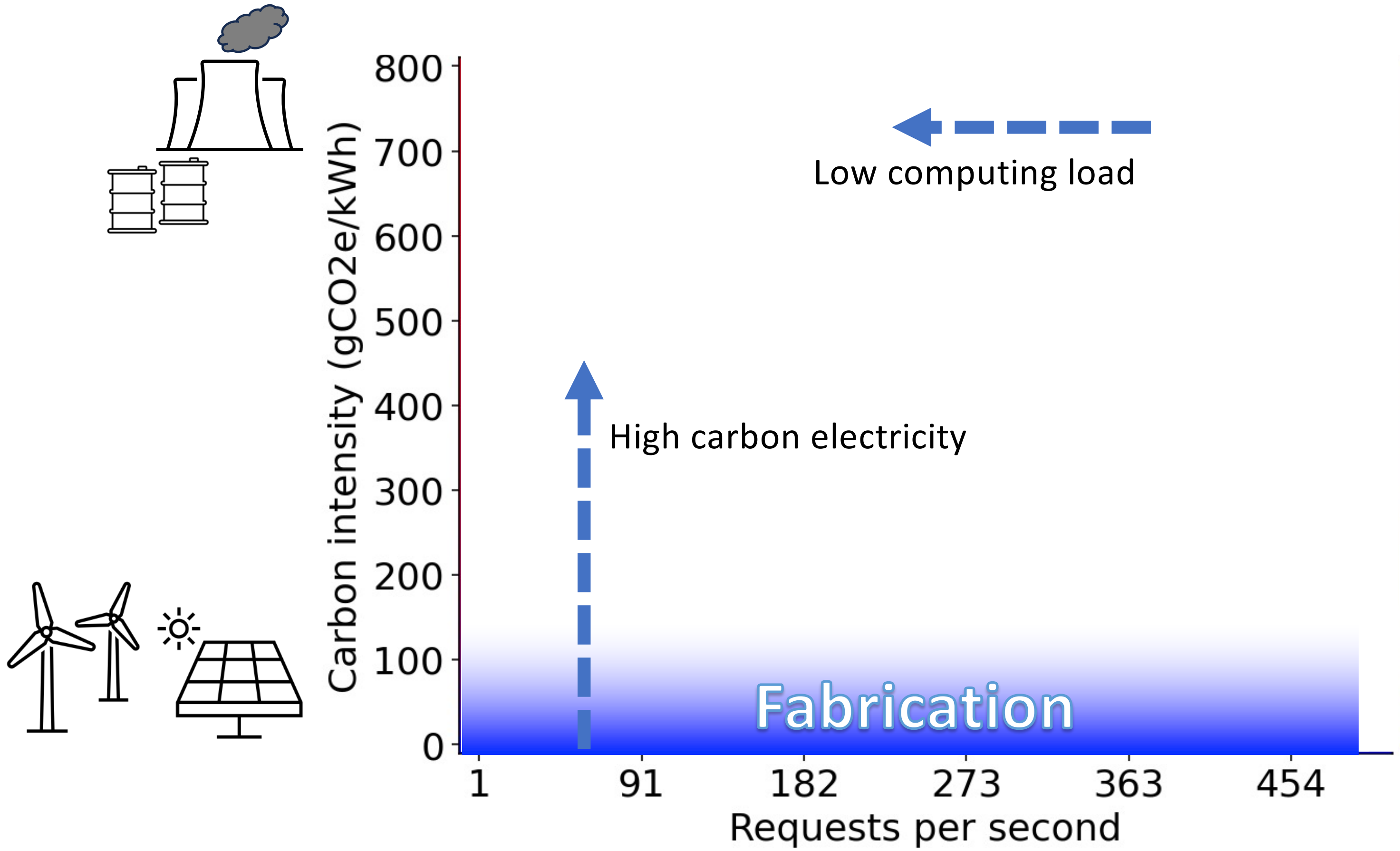
➔ **Total: 2’775** $kgCO_2e$
…in this particular case

What is dominating CO_2e ?

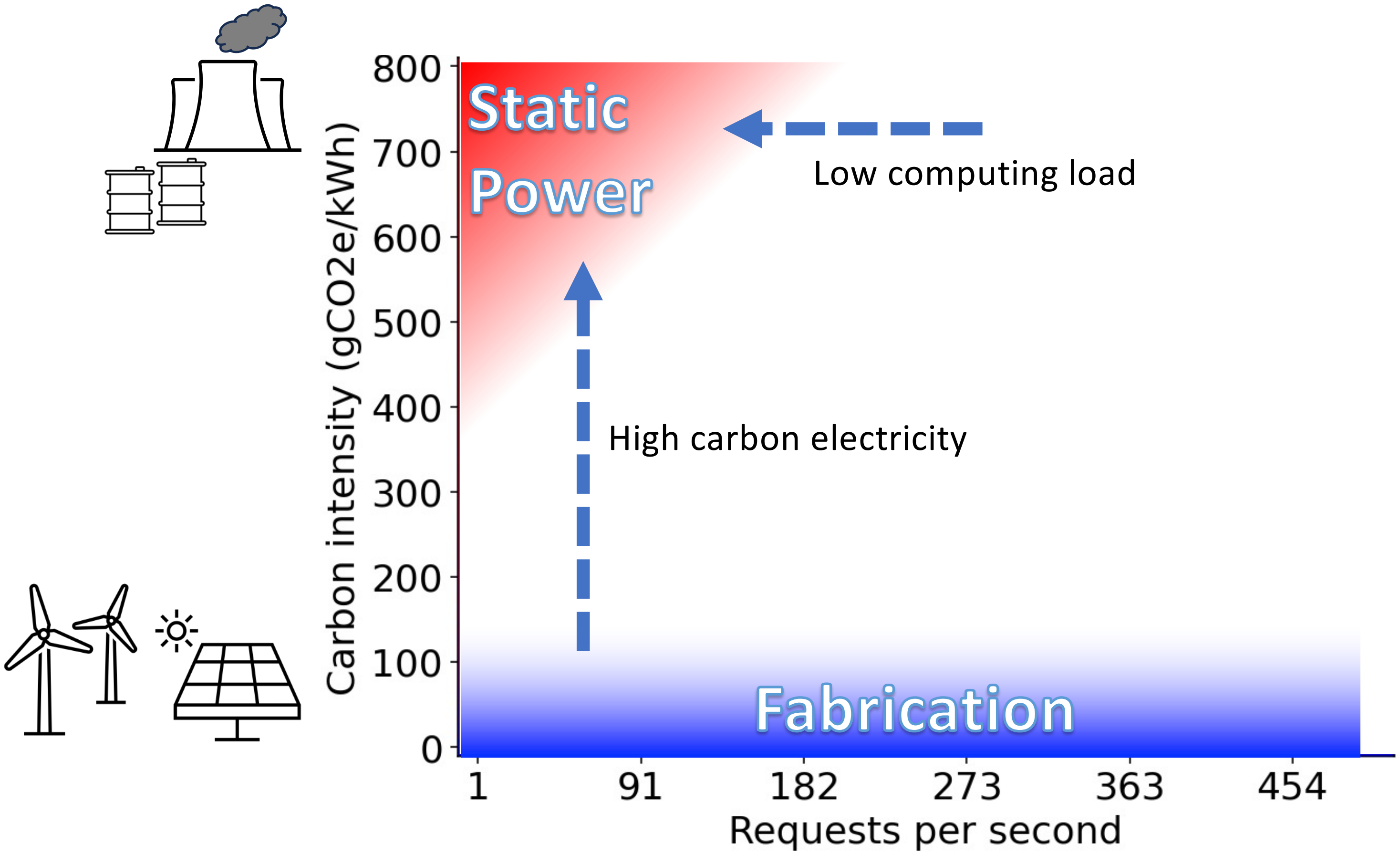


Introducing the Carbon Topography Representation…

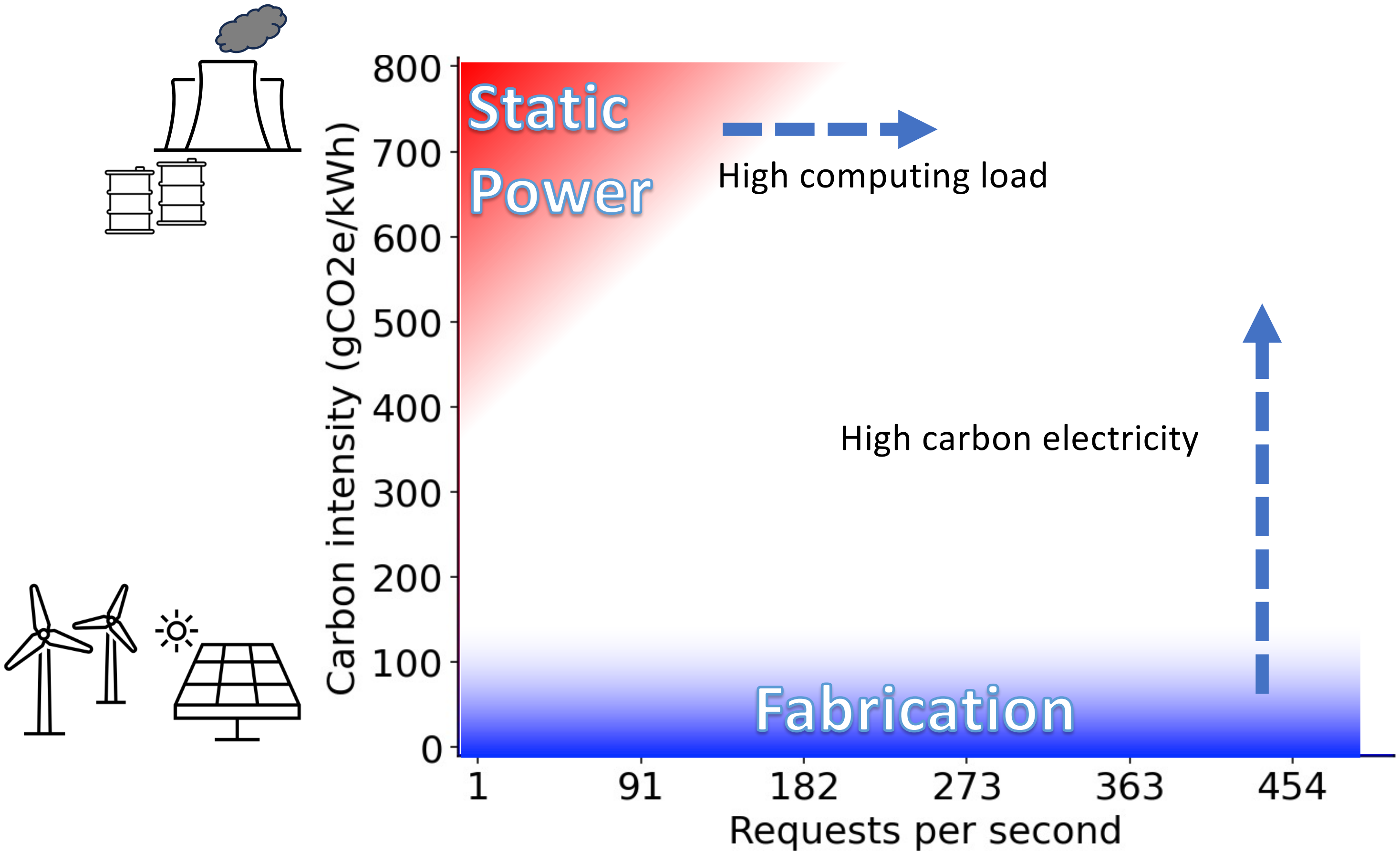

800
700
600
500
400
300
200
100
0
Carbon intensity (gCO2e/kWh)
1
91
182
273
363
454
Requests per second
Low carbon electricity
Fabrication

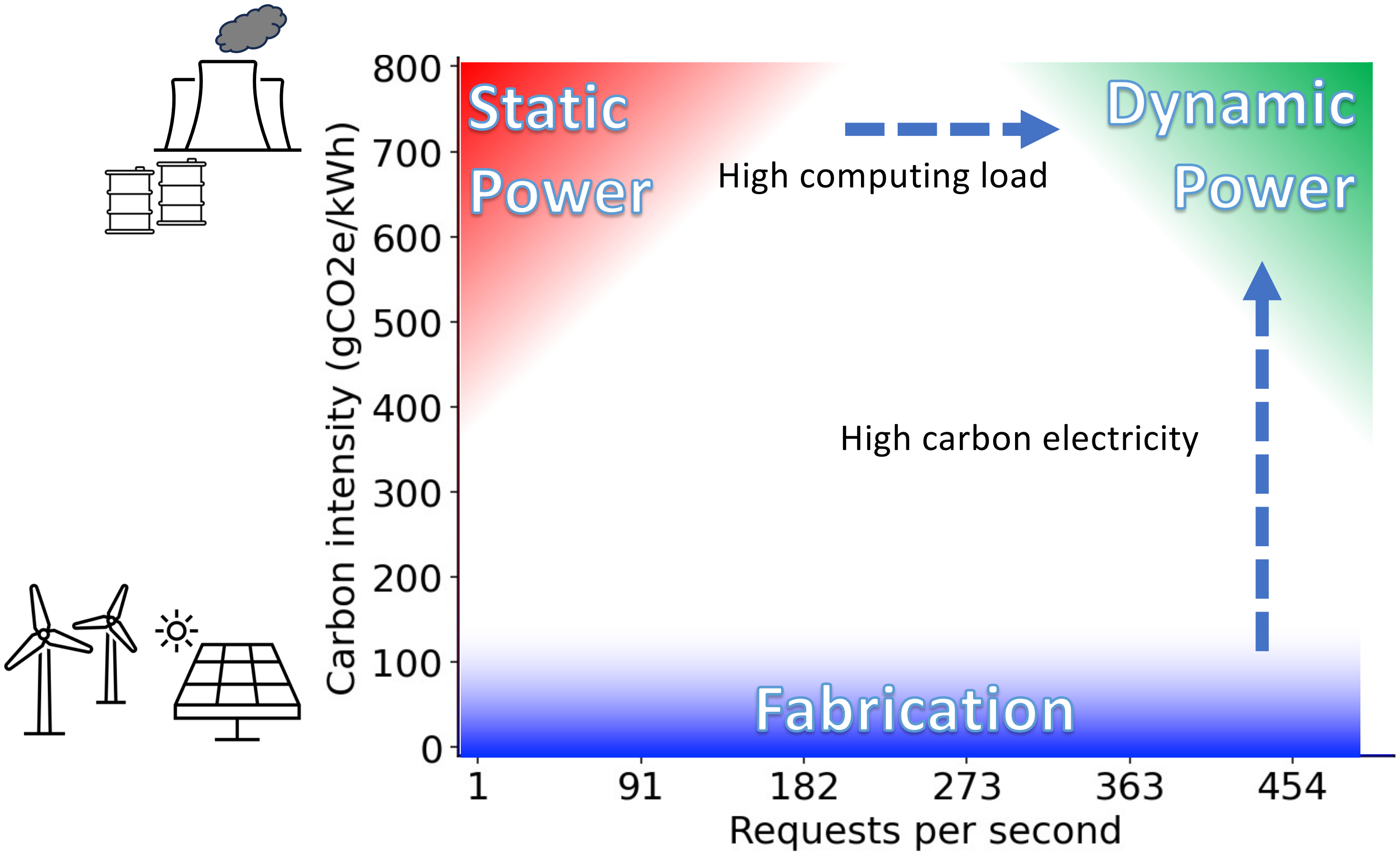


Low computing load
High carbon electricity
Fabrication
Carbon intensity (gCO2e/kWh)
800
700
600
500
400
300
200
100
0
1
91
182
273
363
454
Requests per second



Static Power
Low computing load
High carbon electricity
Fabrication
Carbon intensity (gCO2e/kWh)
800
700
600
500
400
300
200
100
0
1
91
182
273
363
454
Requests per second


800
700
600
500
400
300
200
100
0
Carbon intensity (gCO2e/kWh)
Static Power
High computing load
High carbon electricity
Fabrication
1
91
182
273
363
454
Requests per second

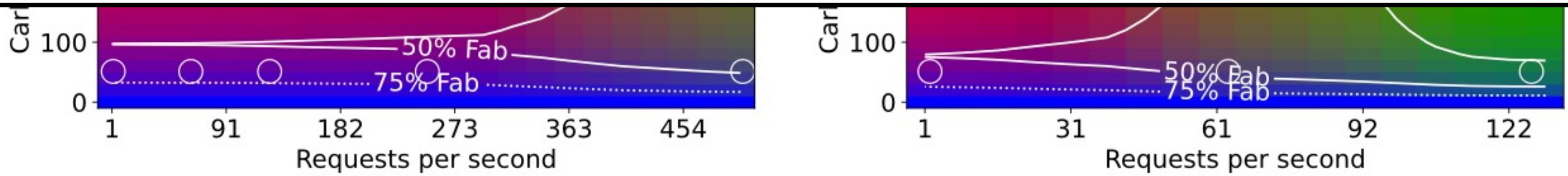

Static Power
High computing load
Dynamic Power
High carbon electricity
Fabrication
Carbon intensity (gCO2e/kWh)
800
700
600
500
400
300
200
100
0
1
91
182
273
363
454
Requests per second

What is dominating CO_2e ? It depends…

HotCarbon 2025

Carbon Topography Representation: Improving Impacts of Data Center Lifecycle

OLIVIER WEPPE, Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, France
DAVID BEKRI, HES-SO, Switzerland
THIBAUT MARTY, Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, France
LOÏC GUIBERT, HES-SO, Switzerland
LOUISE AUBET, Resilio, Switzerland
JEAN-CHRISTOPHE PRÉVOTET, Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, France
MAXIME PELCAT, Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, France
SÉBASTIEN RUMLEY, HES-SO, Switzerland

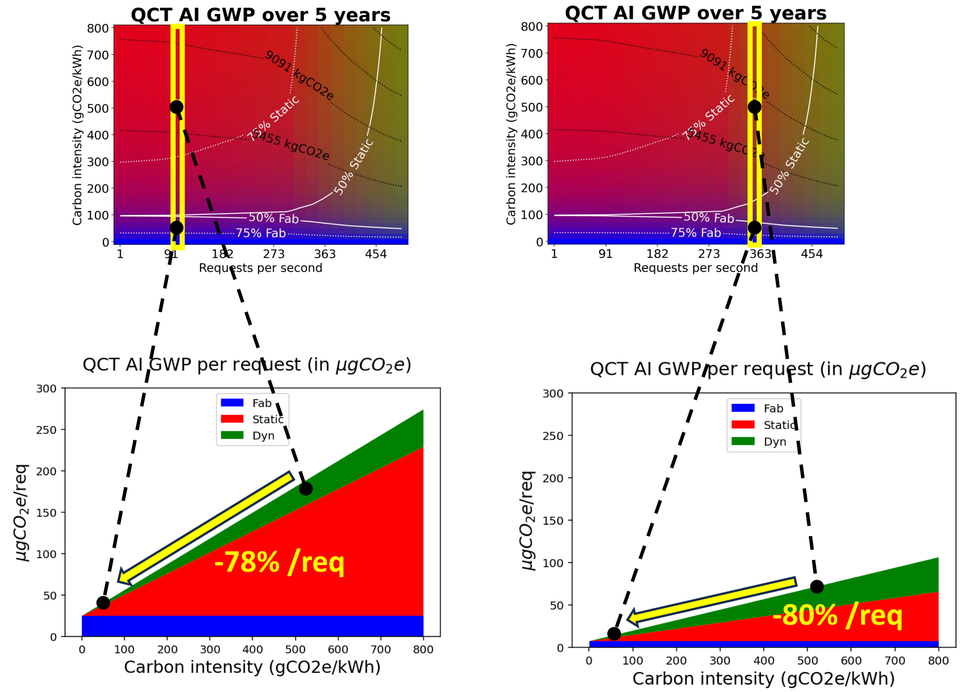


In low-carbon electricity countries (<100gCO_2e/kWh), fabrication appears to dominate impacts

Optimization strategies: #1

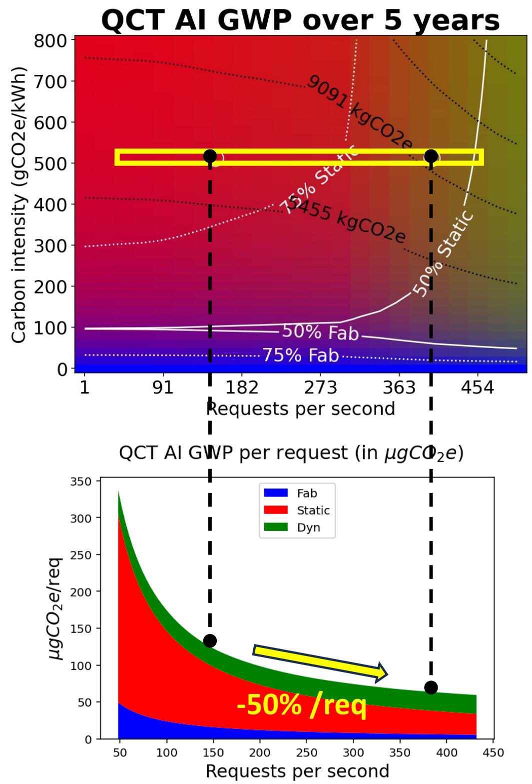
- High Carbon electricity: <u>use low carbon energy</u>
  Ex: $500\ gCO_2e/kWh \rightarrow 50\ gCO_2e/kWh$

  - At 20% load:
    $-121\ \mu gCO_2e$ /req (**-78%**)
  - At 70% load:
    $-56\ \mu gCO_2e$ /req (**-80%**)



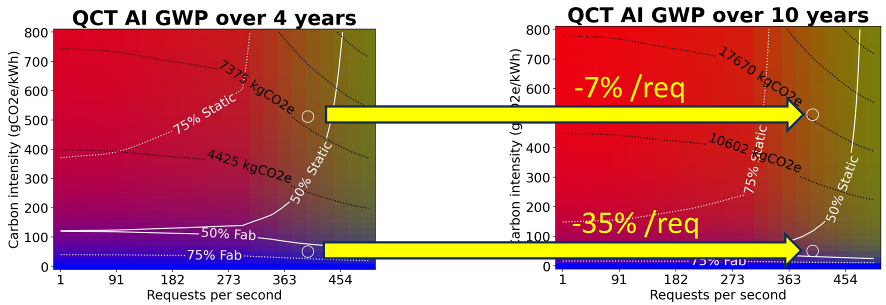
Optimization strategies: #2

- <u>Maximize server load</u>
  Ex : 30% load to 80% load:
  $-62\ \mu gCO_2e$ /req. (**-50%**)



Optimization strategies: #3

3. Low carbon & high load: <u>increase server lifetime</u>
   Ex: from 4 years lifetime to 10 years lifetime
   - At 80% load with $50gCO_2e/kWh$:
     $-5.5\ \mu gCO_2e$ /req. (**-35%**)



Hes·so

Conclusions

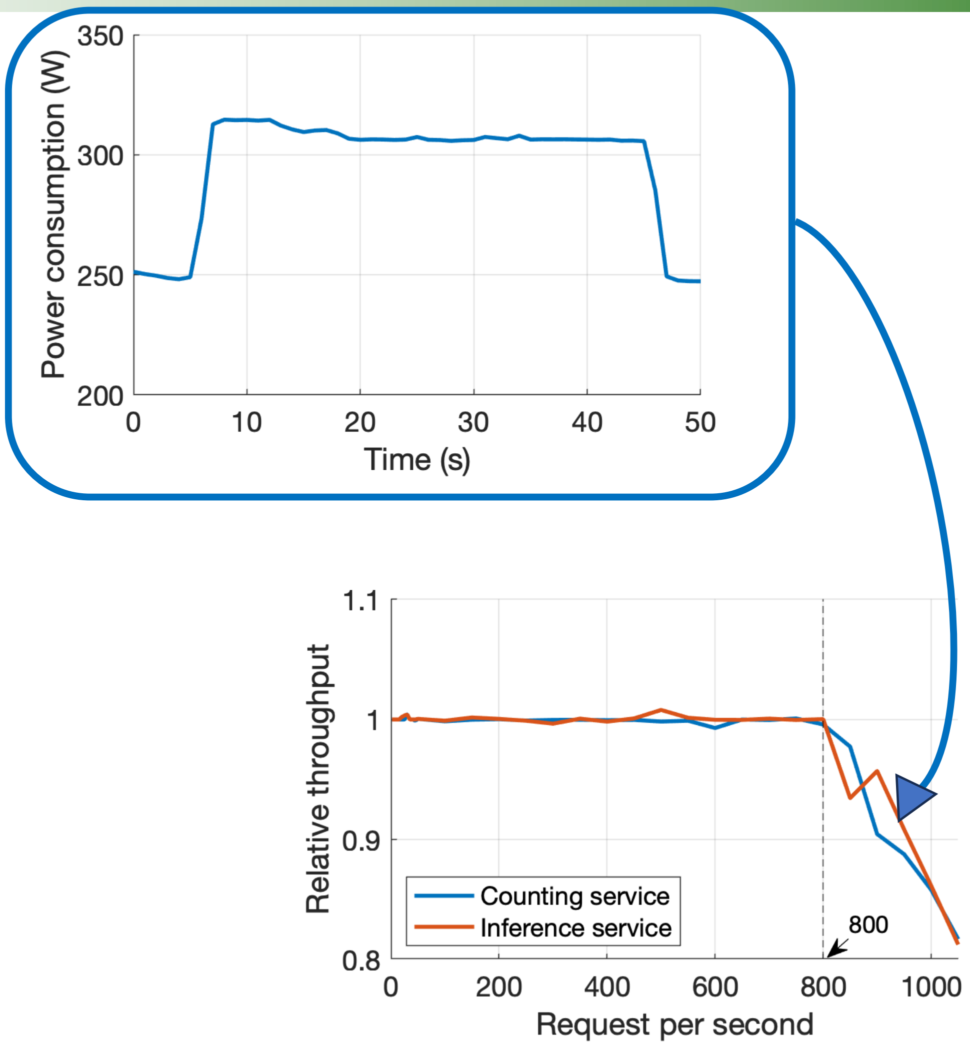
- We need more measurements! More testbeds!
  - The “TRAFFIC” approach is one approach, there are others
    - Introduction of functional units (service requests)
- Careful with simplistic models!
  - What is CPU usage exactly? What is server utilization?
- Newer machines not necessarily better
  - Especially if they remained unused → rebound effect
- Impacts vary a lot!
  - Geography, usage → very different conclusions

Thank you

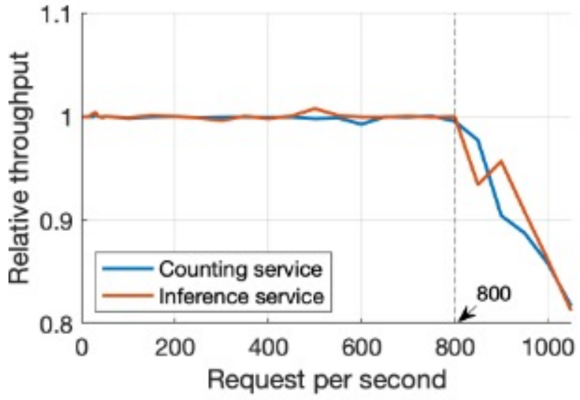
Extra Slides

One experiment

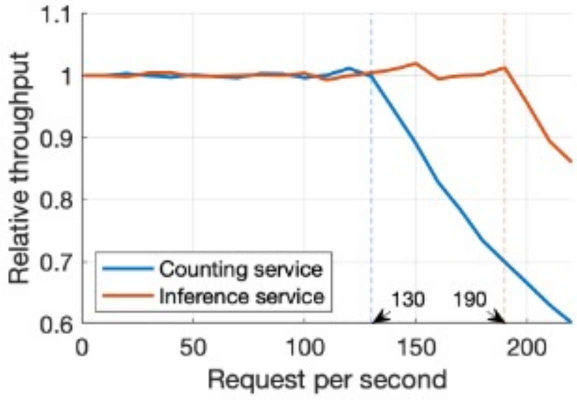
- Inputs
  - … one hardware
  - … one service
  - … one parameter to the service
    - For example, count until 280 millions
  - … one load parameter
    - For example, 300 requests per second
- Outputs
  - (relative) throughput
  - Average power
  - CPU utilization
  - CPU frequency
  - …



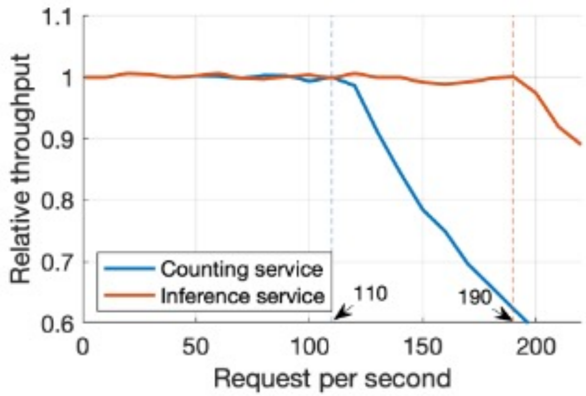
Throughput



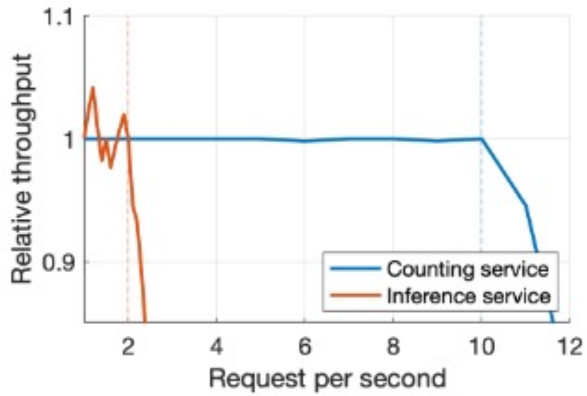
(a) QCT



(b) HPE



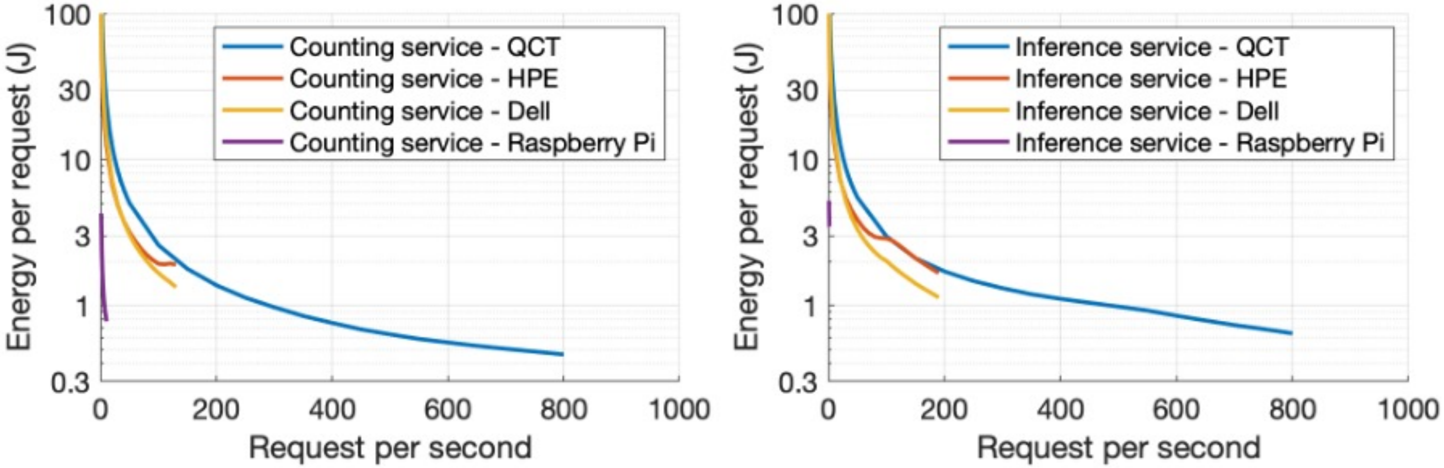
(c) Dell



(d) Raspberry Pi

	Counting	Inference
QCT	800	800
HPE	130	**190**
Dell	110	**190**
RasPi	**10**	2

Energy usage comparisons

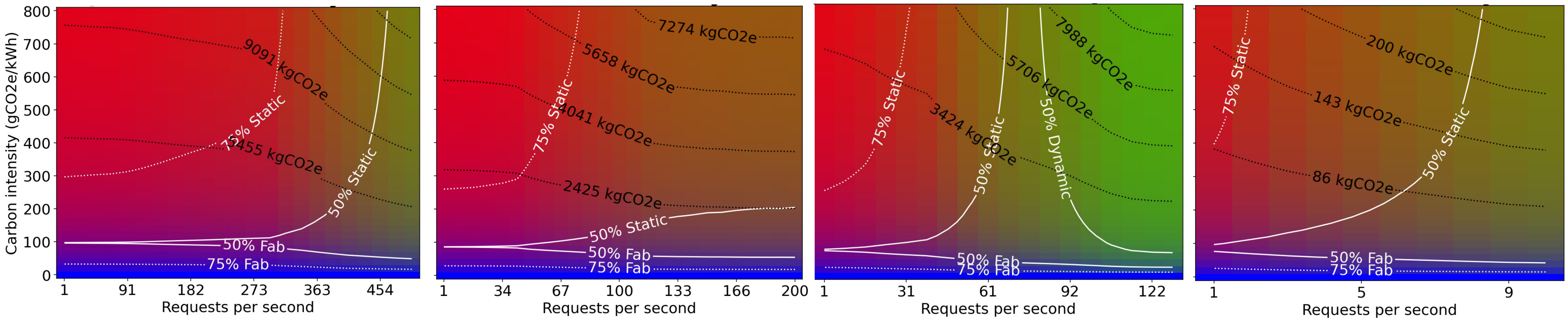


(a) Counting service **(b) Inference service**

@ 2 req/s	Counting	Inference
QCT	115 J	116 J
HPE	100 J	101 j
Dell	103 J	103 J
RasPi	**5 J**	**7 J**

Carbon Topography Representation

QuantaGrid S74G-2U (QCT) | PowerEdge T430 (Dell) | ProLiant DL360 Gen9 (HPE) | Raspberry Pi 4B



[4] Olivier Weppe, David Bekri, Thibaut Marty, Loïc Guibert, Louise Aubet, Jean-Christophe Prévotet, Maxime Pelcat, and Sébastien Rumley. 2025. Carbon Topography Representation: Improving Impacts of Data Center Lifecycle. SIGENERGY Energy Inform. Rev. 5, 2 (July 2025), 41–47.

Normalisation