Compressible Tasks in Green Data Centers as Grid-Forming Support Assets

Based on work by

Anna Vandi, Ramon Aparicio-Pardo, Guillaume Urvoy-Keller.

I3S, Université Côte d'Azur/CNRS

Anna Vandi, Ramon Aparicio-Pardo, Guillaume Urvoy-Keller. Compressible Tasks in Green Data Centers as Grid-Forming Support Assets. IEEE International Conference on Communications (ICC 2026), IEEE, May 2026, Glasgow, United Kingdom. ⟨hal-05326830v2⟩

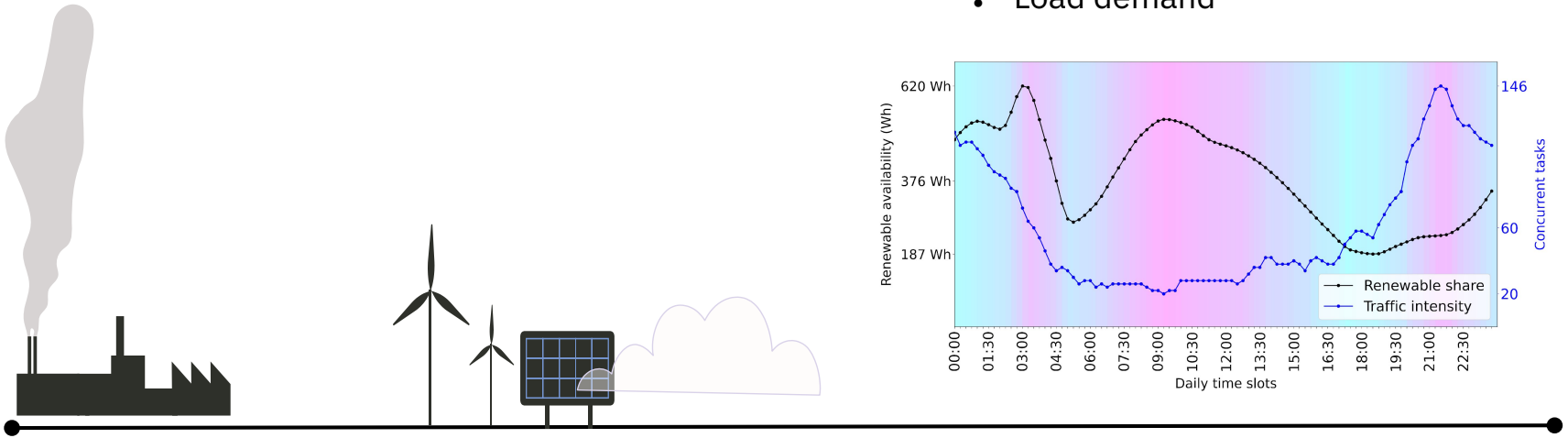
Green Data-Centers equipped with microgrids

Demand in data centers **increases**

Green Data-Centers Equipped with **Microgrids**

Uncertainty/Mismatch challenge

- Renewable Power Generation [1]
- Load demand



[1] Y. Lian, Y. Li, Y. Zhao, C. Yu, T. Zhao, and L. Wu, “Robust multi-objective optimization for islanded data center microgrid operations”, Applied Energy, 2023

Renewable energy integration challenges grid stability

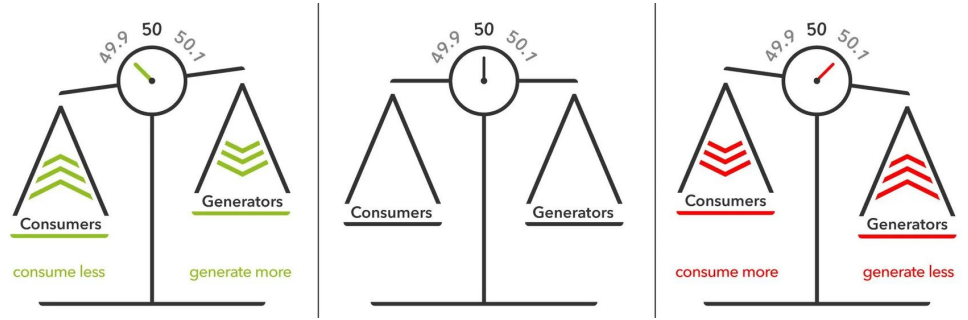
Intermittency of Generation :

supply/demand mismatch

Reduced Inertia:

Conventional turbines provide natural inertia, stabilizing frequency [2].

- **Frequency (50 Hz) Deviations**



Supply > Demand

Supply < Demand

[2] F. Ahmed et al., “Dynamic grid stability in low carbon power systems with minimum inertia,” Renewable Energy, 2023. https://www.sciencedirect.com/science/article/pii/S0960148123003774
[3] ICS Investigation Expert Panel, “Grid Incident in Spain and Portugal on 28 April 2025: Factual Report,” European Network of Transmission System Operators for Electricity (ENTSO-E), Tech. Rep., Oct. 3, 2025.
source img: https://www.next-kraftwerke.com/knowledge/afrr

Strategies to stabilize Renewable-Rich Grids

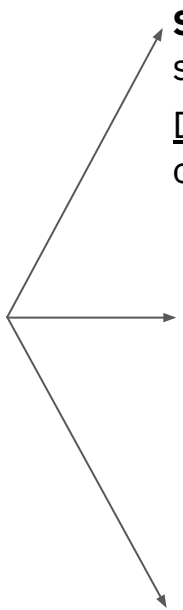
- Classical ***hardware*-based** solutions[2]
  - Energy Storage Systems (ESS)
  - Advanced Inverter Technologies as Grid-Forming Inverters (GFMI)

- Novel paradigm: ***Demand-Side Flexibility***
  - Flexible Workloads as **Grid-Forming** Assets

[2] F. Ahmed et al., “Dynamic grid stability in low carbon power systems with minimum inertia,” Renewable Energy, 2023. https://www.sciencedirect.com/science/article/pii/S0960148123003774
[3] ICS Investigation Expert Panel, “Grid Incident in Spain and Portugal on 28 April 2025: Factual Report,” European Network of Transmission System Operators for Electricity (ENTSO-E), Tech. Rep., Oct. 3, 2025.

Strategies to stabilize Renewable-Rich Grids

- Novel paradigm: ***Demand-Side Flexibility***
  - Flexible Workloads as **Grid-Forming** Assets



Shift in space: tasks across different servers/data centers

Drawbacks: latency, resource constraints, overhead [4],[6]

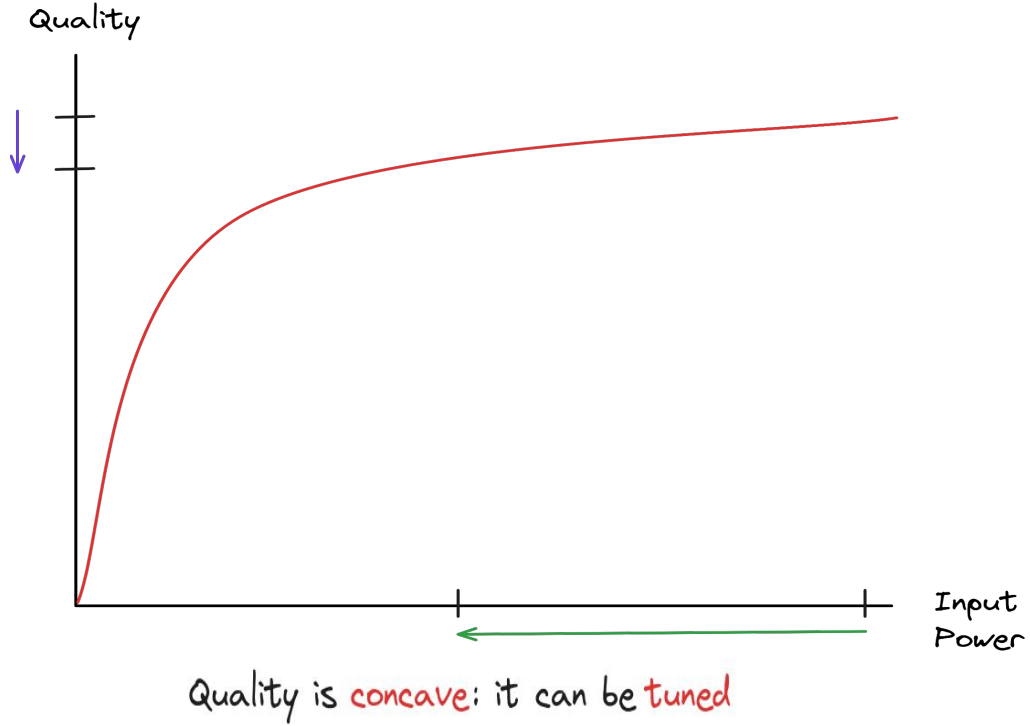
Shift in time: delaying or pausing/resuming workloads

Drawbacks: not always possible, more servers always ON [5]

PROPOSAL: "Shift" in size "compressible" tasks

[4] W. E. Gnibga et al., "Latency, energy and carbon aware collaborative resource allocation with consolidation and qos degradation strategies in edge computing," in 2023 IEEE 29th ICPADS.
[5] B. Acun et al., "Carbon explorer: A holistic framework for designing carbon aware datacenters," 2023.
[6] W. E. Gnibga et al., "Energy-related impact of redefining self-consumption for distributed edge datacenters," in 2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC).

“Compressible” tasks



Video Transcoding Tasks

Transcoding output quality can be tuned [7]

Image Classification Tasks

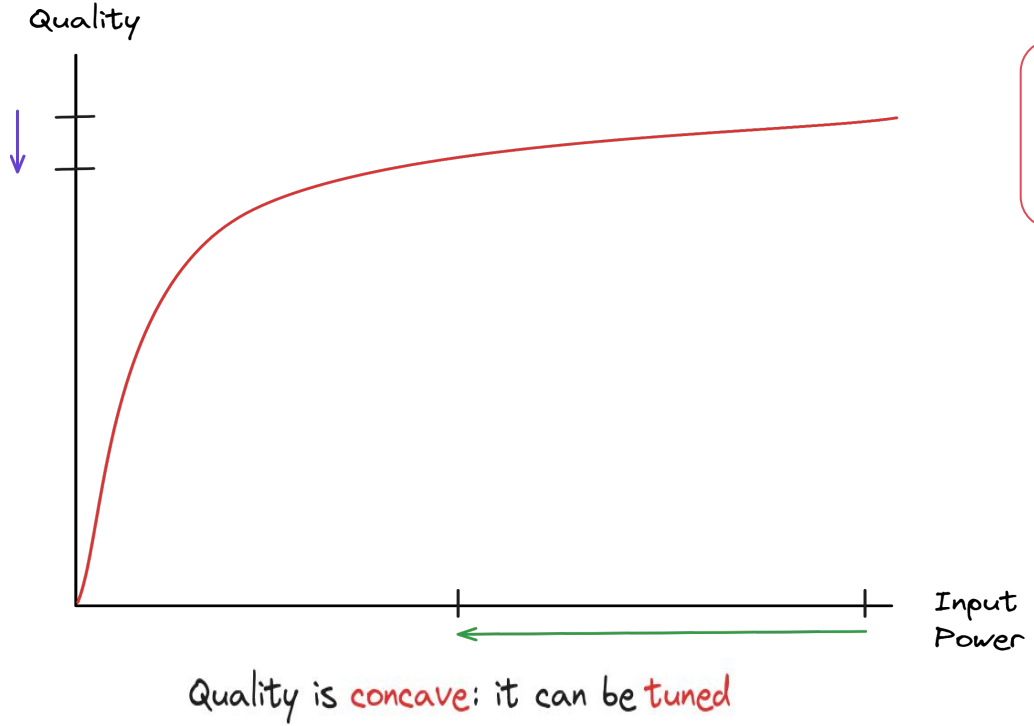
Inference optimization has impact on accuracy [8]

LLM - Text Generation Tasks

Inference optimization has impact on accuracy [9]

[7] Ramon Aparicio-Pardo, Karine Pires, Alberto Blanc, and Gwendal Simon. 2015. Transcoding live adaptive video streams at a massive scale in the cloud. (MMSys '15)
[8] H. Cai, C. Gan et al., “Once-for-all: Train one network and specialize it for efficient deployment,” in International Conference on Learning Representations, 2020.
[9] Xia, Heming, et al. "Tokenskip: Controllable chain-of-thought compression in llms." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

“Compressible” tasks



Video Transcoding Tasks

Transcoding output quality can be tuned [7]

Image Classification Tasks

Inference optimization has impact on accuracy [8]

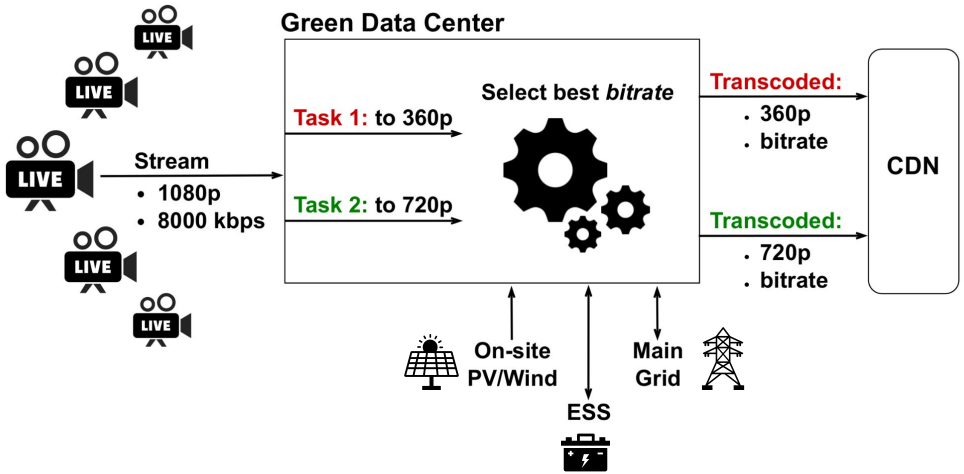
LLM - Text Generation Tasks

Inference optimization has impact on accuracy [9]

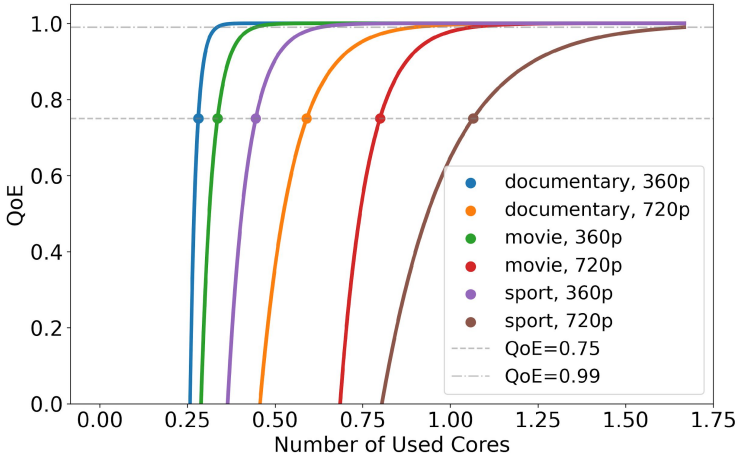
[7] Ramon Aparicio-Pardo, Karine Pires, Alberto Blanc, and Gwendal Simon. 2015. Transcoding live adaptive video streams at a massive scale in the cloud. (MMSys '15)
[8] H. Cai, C. Gan et al., “Once-for-all: Train one network and specialize it for efficient deployment,” in International Conference on Learning Representations, 2020.
[9] Xia, Heming, et al. "Tokenskip: Controllable chain-of-thought compression in llms." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

A case study: live video transcoding

The System.



QoE for content type and resolution.



Anna Vandi, Ramon Aparicio-Pardo, Guillaume Urvoy-Keller. Green Data Centers as Grid-Forming Support Assets. IEEE International Conference on Communications (ICC 2026), IEEE, May 2026, Glasgow, UK

Renewable Energy Aware Task Allocation (REATA)

Oracle: Convex Formulation

$$\max_{\{\mathbf{x}\}} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} U_j\left(f_{tj} = F \cdot x_{tj}\right)$$

s.t.

$$\sum_{j\in\mathcal{J}_i} P \cdot x_{tj} \leq \max(R_t, P_{ti}^{\min}),$$

$$x_{tj} \geq X_j^{\min},$$

$$x_{tj} \in \mathbb{R}_{\geq 0},$$

$$S_{ti} = \max\left(0, P_{ti}^{min} - R_t\right)$$

$x_{tj} \in \mathbb{R}_{\geq 0}$	CPU usage by task j at slot t.
$U_j(x_j) \in \mathbb{R}_{\geq 0}$	Utility (QoE) of task j at CPU usage x_j.
$U^{min} \in \mathbb{R}_{\geq 0}$	Minimal utility (QoE) level.
$X_j^{min} \in \mathbb{R}_{\geq 0}$	Value for x_j at minimal QoE level.
$P \in \mathbb{R}_{\geq 0}$	Maximum power consumption (in *Watts*)
$R_t \in \mathbb{R}_{\geq 0}$	Renewable power (*Watts*) at slot t.
$S_{ti} \in \mathbb{R}_{\geq 0}$	Power *imported* at slot-interval pair (t, i).
$F \in \mathbb{R}_{\geq 0}$	CPU core frequency (in *GHz*).

ONLINE-REATA Algorithm

Algorithm 1: ONLINE-REATA: abstract task allocation

Data: $\mathcal{J}_{i-1}$, $\mathcal{H}_{i-1}$, $\mathcal{L}_{i-1}$, $\hat{N}_i$, S_{i-1}, R
k: arriving task at interval i (slot t).
Result: Optimal allocation x_k for task k

1 **for** *each arriving task* k **do**
2 **if** $|\mathcal{J}_{i-1}| + 1 \geq \hat{N}_i$ **then**
3 $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
4 **else**
5 Generate artificial tasks $\mathcal{A}_{il}$ based on $\hat{N}_i$;
6 $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\} \cup \mathcal{A}_{il}$;
7 Compute λ using (8) with $(\hat{\mathcal{H}}_i, \mathcal{L}_{i-1}, R, S_{i-1})$;
8 **if** $\lambda > U_k'(X_k^{min})$ **then**
9 $x_k \leftarrow X_k^{min}$;
10 $\mathcal{L}_i \leftarrow \mathcal{L}_{i-1} \cup \{k\}$;
11 **else**
12 $x_k \leftarrow -\frac{1}{\theta_k F} \ln\left(\frac{\lambda}{\tau_k \theta_k F}\right)$;
13 $\mathcal{H}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
14 $\mathcal{J}_i \leftarrow \mathcal{H}_i \cup \mathcal{L}_i$;

[10] S. Boyd et al., Convex Optimization. Cambridge University Press, 2004.

Renewable Energy Aware Task Allocation (REATA)

Oracle: Convex Formulation

$$\max_{\{\mathbf{x}\}} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} U_j\left(f_{tj} = F \cdot x_{tj}\right)$$

Max Sum Utilities

s.t.

$$\sum_{j\in\mathcal{J}_i} P \cdot x_{tj} \leq \max(R_t, P_{ti}^{\min}),$$

Power Budget

$$x_{tj} \geq X_j^{\min},$$

$$x_{tj} \in \mathbb{R}_{\geq 0},$$

Minimum Allocation (QoE)

$$S_{ti} = \max\left(0, P_{ti}^{min} - R_t\right)$$

x_{tj} decision variable

$x_{ti} \in \mathbb{R}_{>0}$	CPU usage by task j at slot t.
$U_j(x_j) \in \mathbb{R}_{\geq 0}$	Utility (QoE) of task j at CPU usage x_j.
$U^{min} \in \mathbb{R}_{\geq 0}$	Minimal utility (QoE) level.
$X_j^{min} \in \mathbb{R}_{\geq 0}$	Value for x_j at minimal QoE level.
$P \in \mathbb{R}_{\geq 0}$	Maximum power consumption (in *Watts*)
$R_t \in \mathbb{R}_{\geq 0}$	Renewable power (*Watts*) at slot t.
$S_{ti} \in \mathbb{R}_{\geq 0}$	Power *imported* at slot-interval pair (t, i).
$F \in \mathbb{R}_{\geq 0}$	CPU core frequency (in *GHz*).

ONLINE-REATA Algorithm

Algorithm 1: ONLINE-REATA: abstract task allocation

Data: $\mathcal{J}_{i-1}, \mathcal{H}_{i-1}, \mathcal{L}_{i-1}, \hat{N}_i, S_{i-1}, R$
k: arriving task at interval i (slot t).
Result: Optimal allocation x_k for task k

1 **for** *each arriving task* k **do**
2 **if** $|\mathcal{J}_{i-1}| + 1 \geq \hat{N}_i$ **then**
3 $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
4 **else**
5 Generate artificial tasks $\mathcal{A}_{il}$ based on $\hat{N}_i$;
6 $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\} \cup \mathcal{A}_{il}$;
7 Compute λ using (8) with $(\hat{\mathcal{H}}_i, \mathcal{L}_{i-1}, R, S_{i-1})$;
8 **if** $\lambda > U_k'(X_k^{min})$ **then**
9 $x_k \leftarrow X_k^{min}$;
10 $\mathcal{L}_i \leftarrow \mathcal{L}_{i-1} \cup \{k\}$;
11 **else**
12 $x_k \leftarrow -\frac{1}{\theta_k F} \ln\left(\frac{\lambda}{\tau_k \theta_k F}\right)$;
13 $\mathcal{H}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
14 $\mathcal{J}_i \leftarrow \mathcal{H}_i \cup \mathcal{L}_i$;

[10] S. Boyd et al., Convex Optimization. Cambridge University Press, 2004.

Renewable Energy Aware Task Allocation (REATA)

Oracle: Convex Formulation

$$\max_{\{\mathbf{x}\}} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} U_j\left(f_{tj} = F \cdot x_{tj}\right)$$

s.t.

$$\sum_{j\in\mathcal{J}_i} P \cdot x_{tj} \leq \max(R_t, P_{ti}^{\min}),$$

$$x_{tj} \geq X_j^{\min},$$

$$x_{tj} \in \mathbb{R}_{\geq 0},$$

$$S_{ti} = \max\left(0, P_{ti}^{min} - R_t\right)$$

$x_{tj} \in \mathbb{R}_{\geq 0}$	CPU usage by task j at slot t.
$U_j(x_j) \in \mathbb{R}_{\geq 0}$	Utility (QoE) of task j at CPU usage x_j.
$U^{min} \in \mathbb{R}_{\geq 0}$	Minimal utility (QoE) level.
$X_j^{min} \in \mathbb{R}_{\geq 0}$	Value for x_j at minimal QoE level.
$P \in \mathbb{R}_{\geq 0}$	Maximum power consumption (in *Watts*)
$R_t \in \mathbb{R}_{\geq 0}$	Renewable power (*Watts*) at slot t.
$S_{ti} \in \mathbb{R}_{\geq 0}$	Power *imported* at slot-interval pair (t, i).
$F \in \mathbb{R}_{\geq 0}$	CPU core frequency (in *GHz*).

ONLINE-REATA Algorithm

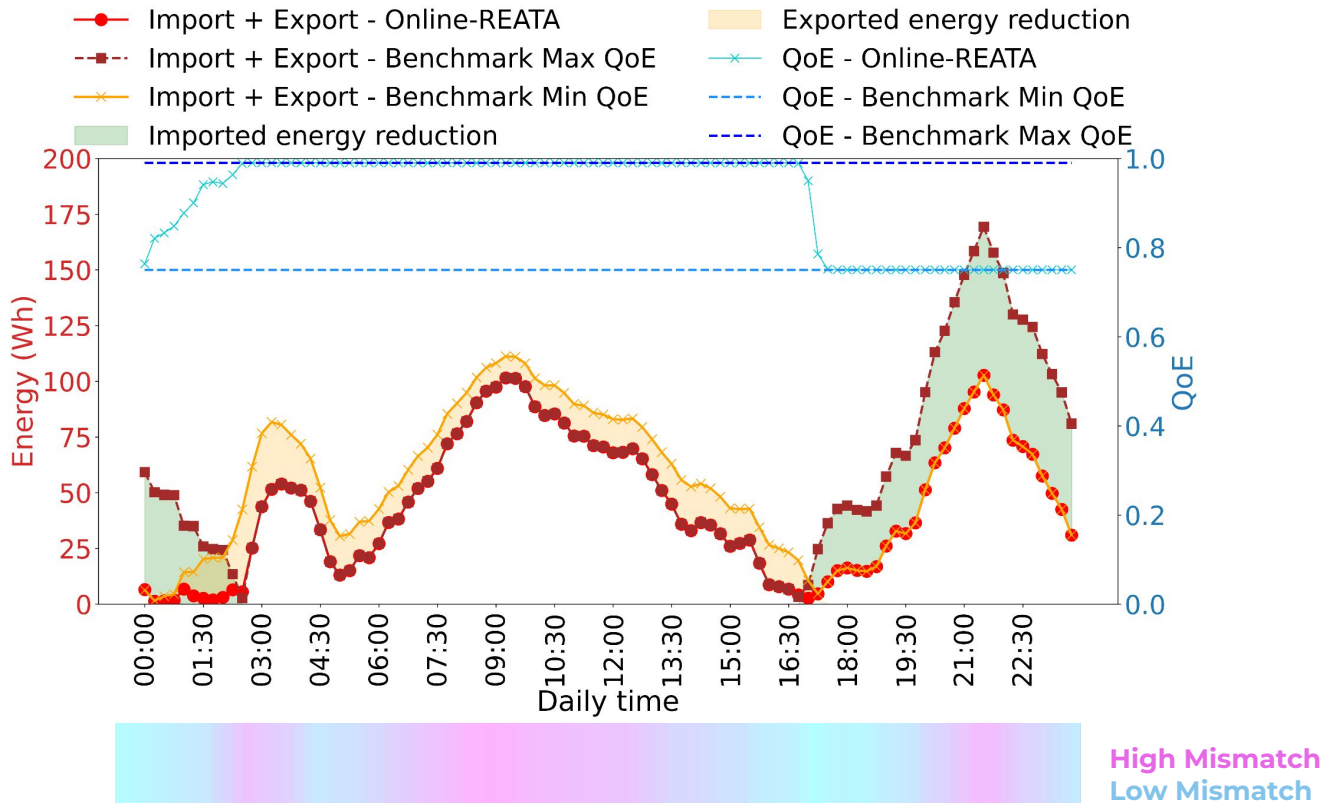
Algorithm 1: ONLINE-REATA: abstract task allocation

Data: $\mathcal{J}_{i-1}$, $\mathcal{H}_{i-1}$, $\mathcal{L}_{i-1}$, $\hat{N}_l$, S_{i-1}, R
k: arriving task at interval i (slot t).
Result: Optimal allocation x_k for task k

for *each arriving task* k **do**
  if $|\mathcal{J}_{i-1}| + 1 \geq \hat{N}_l$ **then**
    $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
  else
    Generate artificial tasks $\mathcal{A}_{il}$ based on $\hat{N}_l$;
    $\hat{\mathcal{H}}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\} \cup \mathcal{A}_{il}$;
  Compute λ using (8) with $(\hat{\mathcal{H}}_i, \mathcal{L}_{i-1}, R, S_{i-1})$;
  if $\lambda > U_k'(X_k^{min})$ **then**
    $x_k \leftarrow X_k^{min}$;
    $\mathcal{L}_i \leftarrow \mathcal{L}_{i-1} \cup \{k\}$;
  else
    $x_k \leftarrow -\frac{1}{\theta_k F} \ln\left(\frac{\lambda}{\tau_k \theta_k F}\right)$;
    $\mathcal{H}_i \leftarrow \mathcal{H}_{i-1} \cup \{k\}$;
  $\mathcal{J}_i \leftarrow \mathcal{H}_i \cup \mathcal{L}_i$;

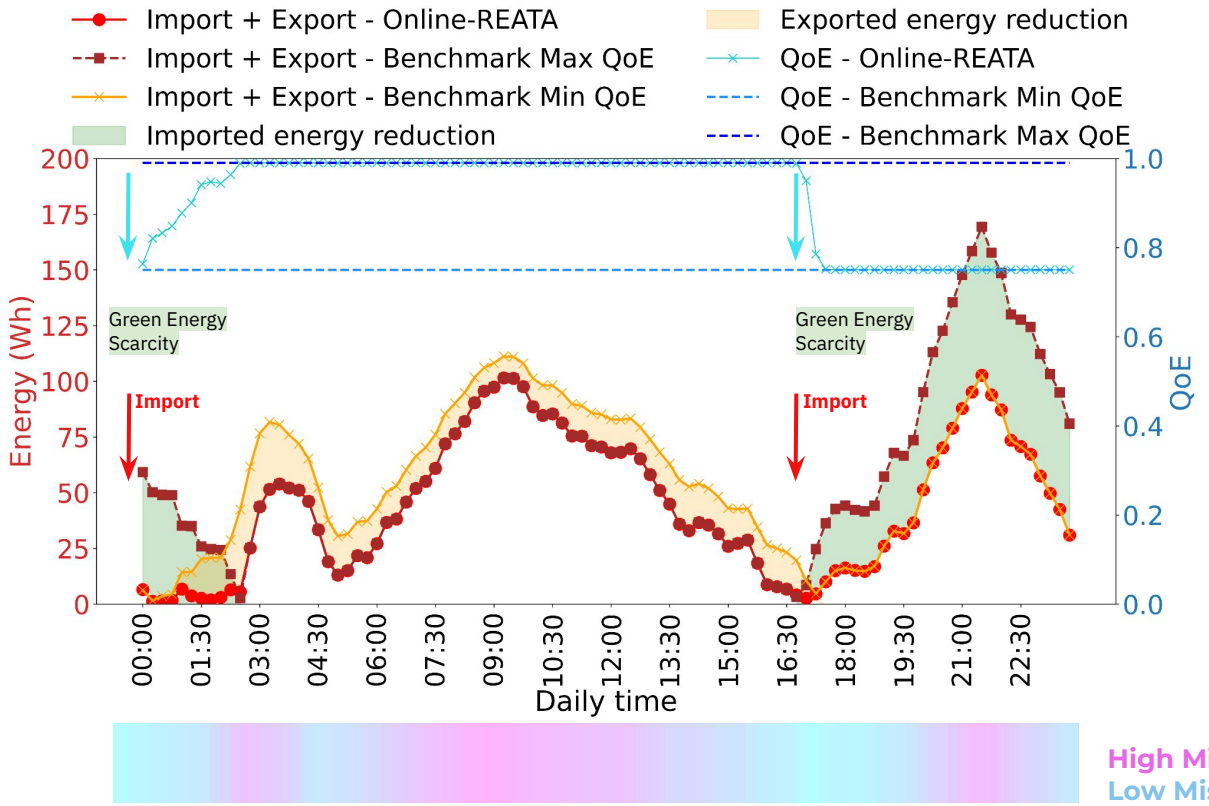
[10] S. Boyd et al., Convex Optimization. Cambridge University Press, 2004.

Some results:

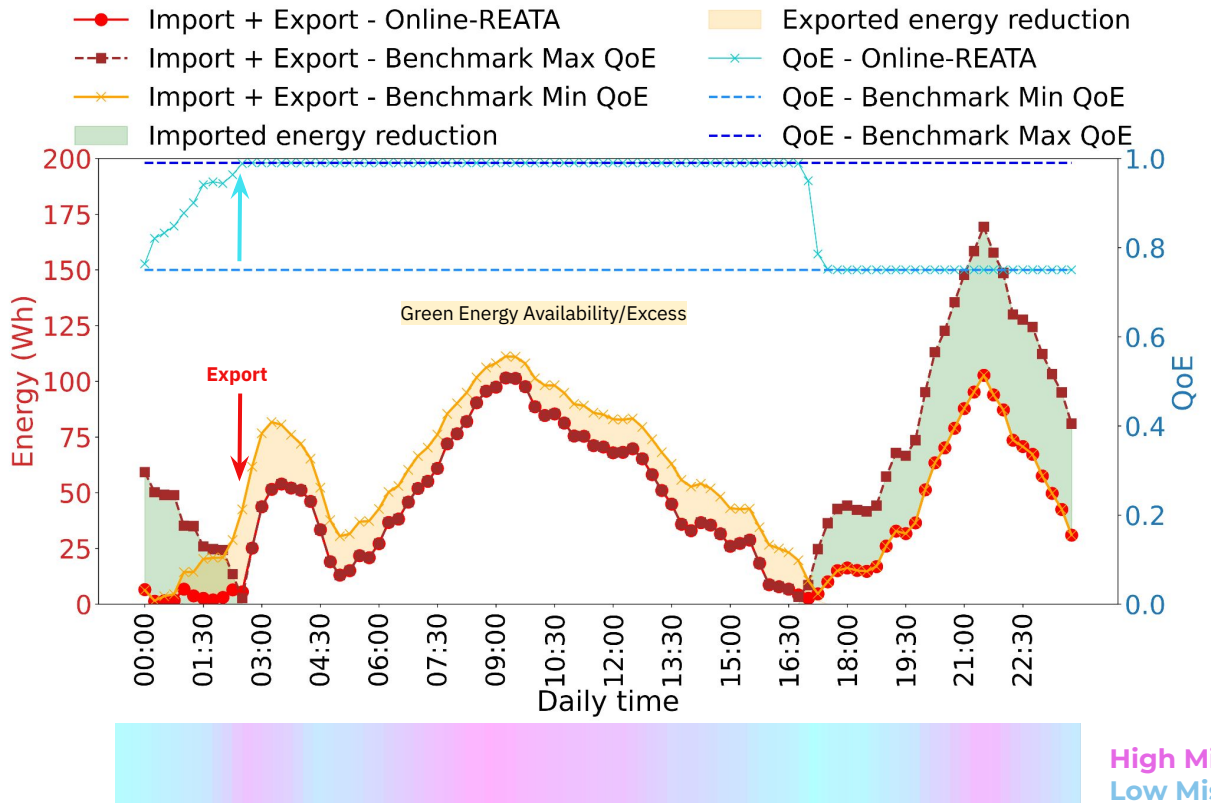


Some results:

Compression



Some results:



Decompression

Some results

Self-Sufficiency Rate (SSR) [11]

- share of load covered by renewables

$$SSR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2load + ESS2load + Grid2load}$$

- 📌 High SSR ⇒ less import from the main grid

Self-Consumption Rate (SCR) [11]

- share of renewable energy used over total available

$$SCR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2grid + PV2load + PV2ESS}$$

- 📌 High SCR ⇒ less export to the main grid

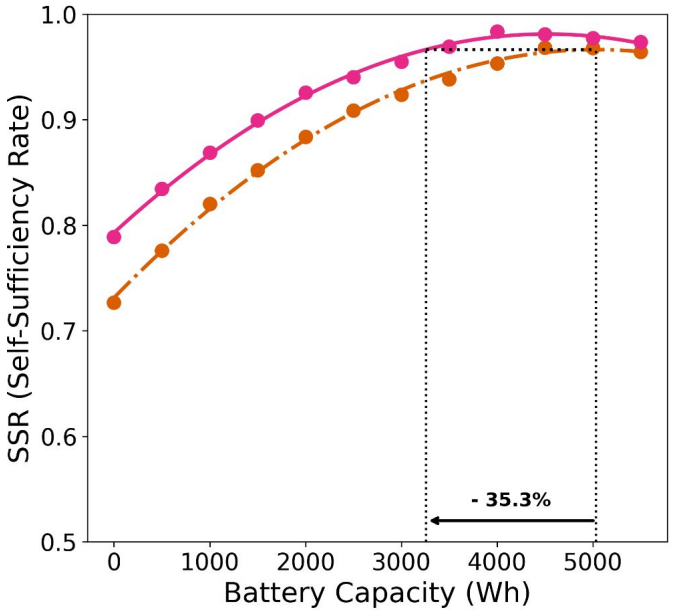
[11] D. Kim et al., “Improved metrics for evaluating self-consumption and self-sufficiency rates in ess-integrated renewable energy systems,” Renewable Energy, vol. 247, p. 123059, 2025.

Some results

Self-Sufficiency Rate (SSR) [11]

- share of load covered by renewables

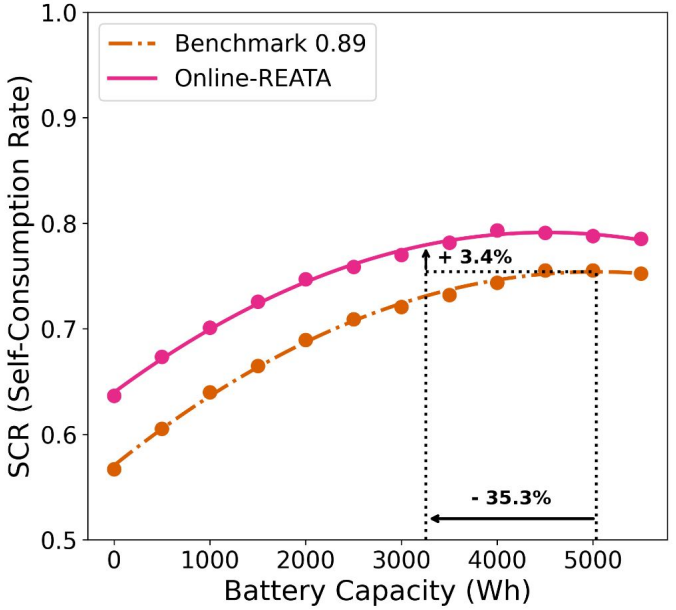
$$SSR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2load + ESS2load + Grid2load}$$



Self-Consumption Rate (SCR) [11]

- share of renewable energy used over total available

$$SCR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2grid + PV2load + PV2ESS}$$



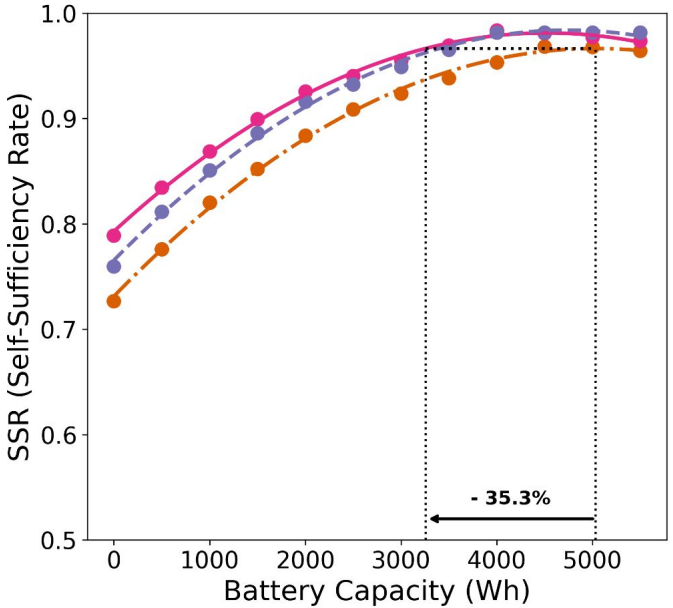
[11] D. Kim et al., “Improved metrics for evaluating self-consumption and self-sufficiency rates in ess-integrated renewable energy systems,” Renewable Energy, vol. 247, p. 123059, 2025.

Some results

Self-Sufficiency Rate (SSR) [11]

- share of load covered by renewables

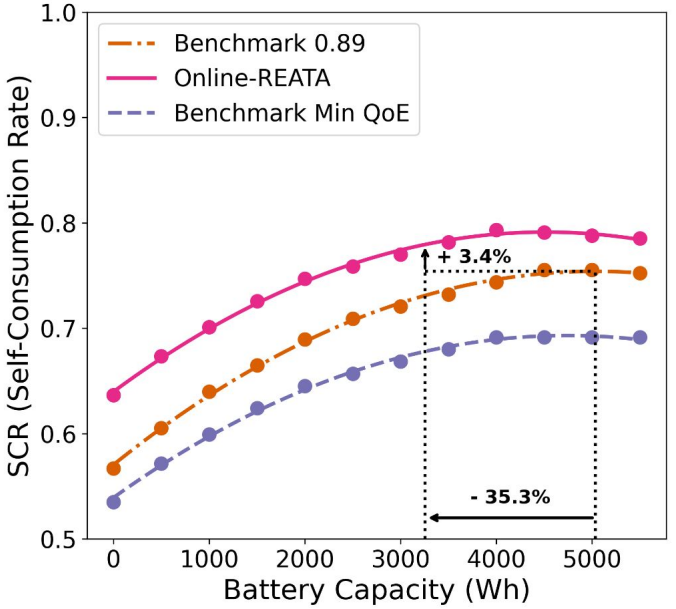
$$SSR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2load + ESS2load + Grid2load}$$



Self-Consumption Rate (SCR) [11]

- share of renewable energy used over total available

$$SCR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2grid + PV2load + PV2ESS}$$



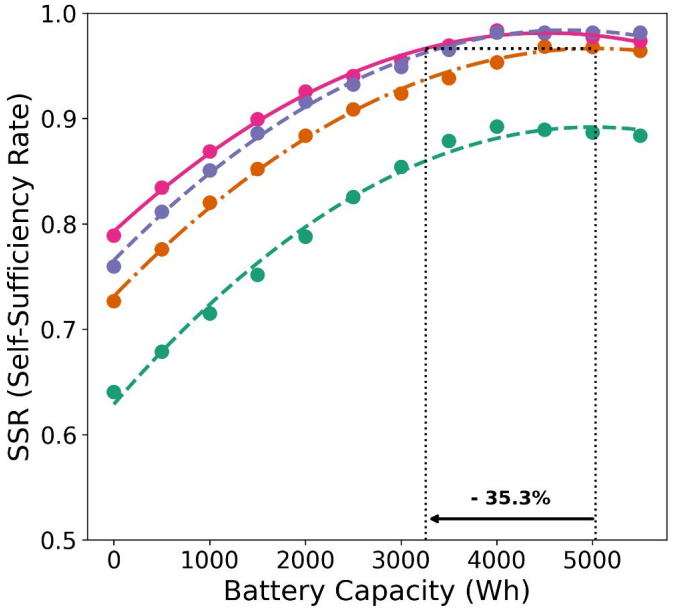
[11] D. Kim et al., “Improved metrics for evaluating self-consumption and self-sufficiency rates in ess-integrated renewable energy systems,” Renewable Energy, vol. 247, p. 123059, 2025.

Some results

Self-Sufficiency Rate (SSR) [11]

- share of load covered by renewables

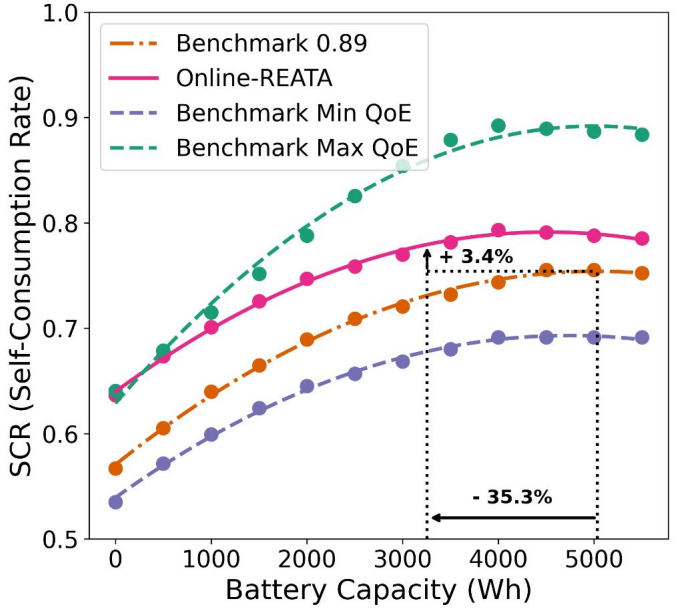
$$SSR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2load + ESS2load + Grid2load}$$



Self-Consumption Rate (SCR) [11]

- share of renewable energy used over total available

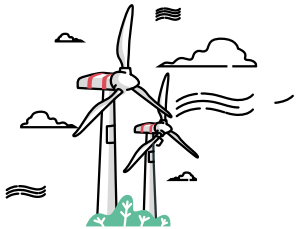
$$SCR_{PV+ESS} = \frac{PV2load + ESS_{PV}2load}{PV2grid + PV2load + PV2ESS}$$



[11] D. Kim et al., “Improved metrics for evaluating self-consumption and self-sufficiency rates in ess-integrated renewable energy systems,” Renewable Energy, vol. 247, p. 123059, 2025.

Future Directions:

- Extension to other **compressible tasks** (RLMs and CoT pruning)
- Model selection (Object detection, text generation, code generation,...)
- Combination with **time and space shifts** (batches as GFM asset, servers consolidation)
- **Distributed** environments integration
- Social Incentives



Thank You!

vandi@i3s.unice.fr