

Towards Sustainable and Accountable Hybrid Cloud Computing via Carbon Intensity Forecasting

M. Zanotto¹, G. Padovani¹, G. Iacca¹, G. Sipsos², S. Fiore¹

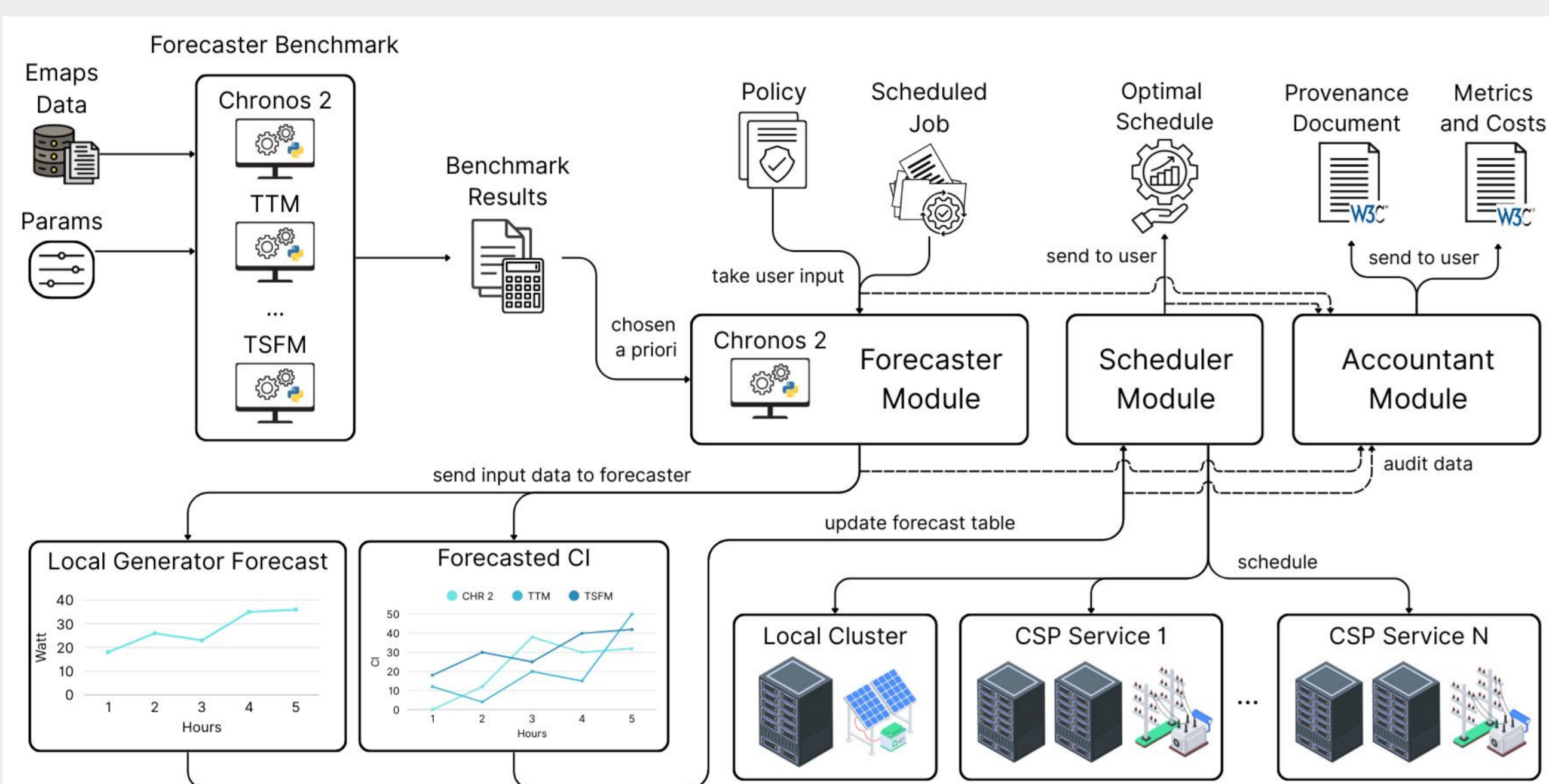
¹ University of Trento, ² EGI

Architecture overview

End-to-end architecture for sustainable and accountable management of cloud workloads in a hybrid cloud environment.

Pipeline for management of cloud workloads in three core modules:

- Forecaster:** time series ML forecaster
- Scheduler:** linear programming optimization model
- Accountant:** provenance-based reporting



Evaluation

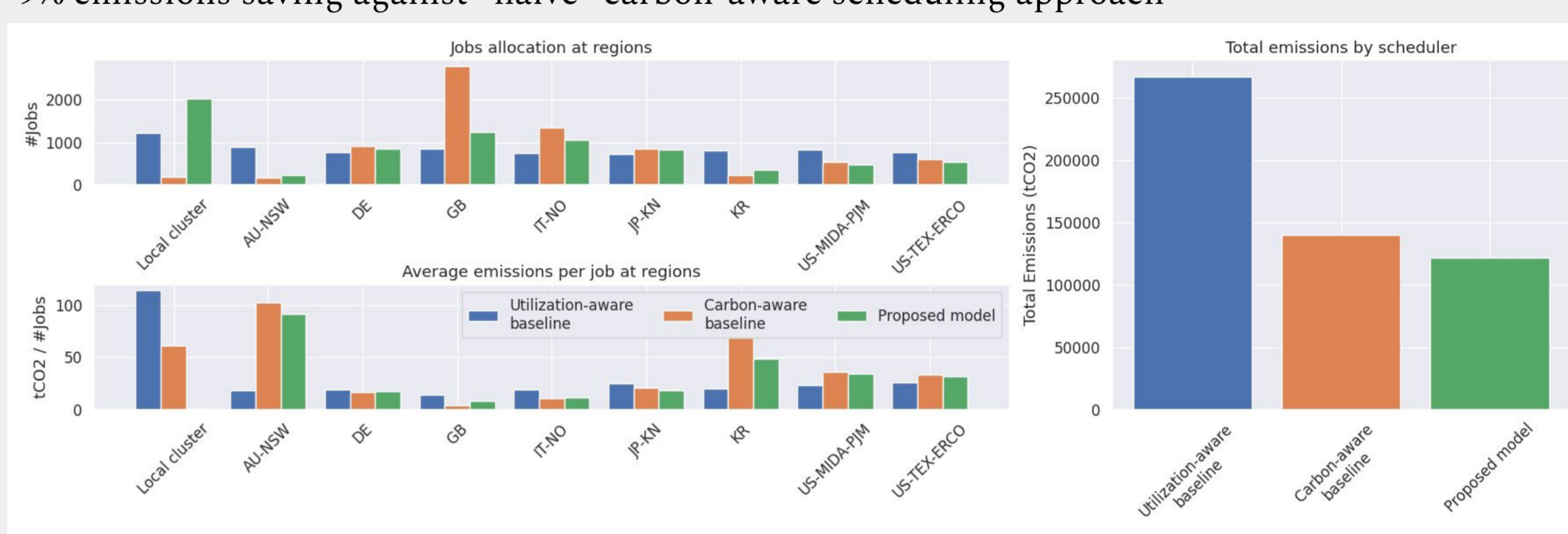
Evaluation of proposed architecture through hybrid cloud scheduling simulation

- Local cluster with 32 GPUs + Solar panels
- Public cloud services with 64 GPUs located across evaluation regions

Usage of real-world ML workloads execution traces

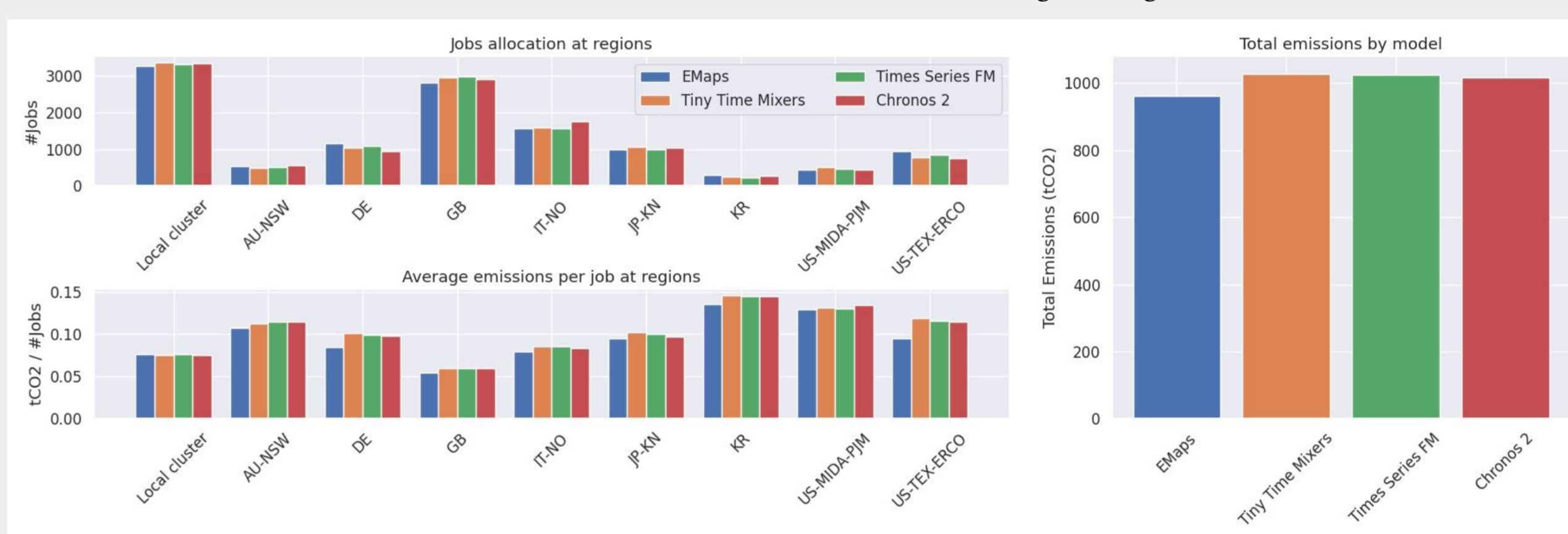
Baseline evaluation:

27% emissions saving against utilization-aware, carbon agnostic baseline
9% emissions saving against "naive" carbon-aware scheduling approach



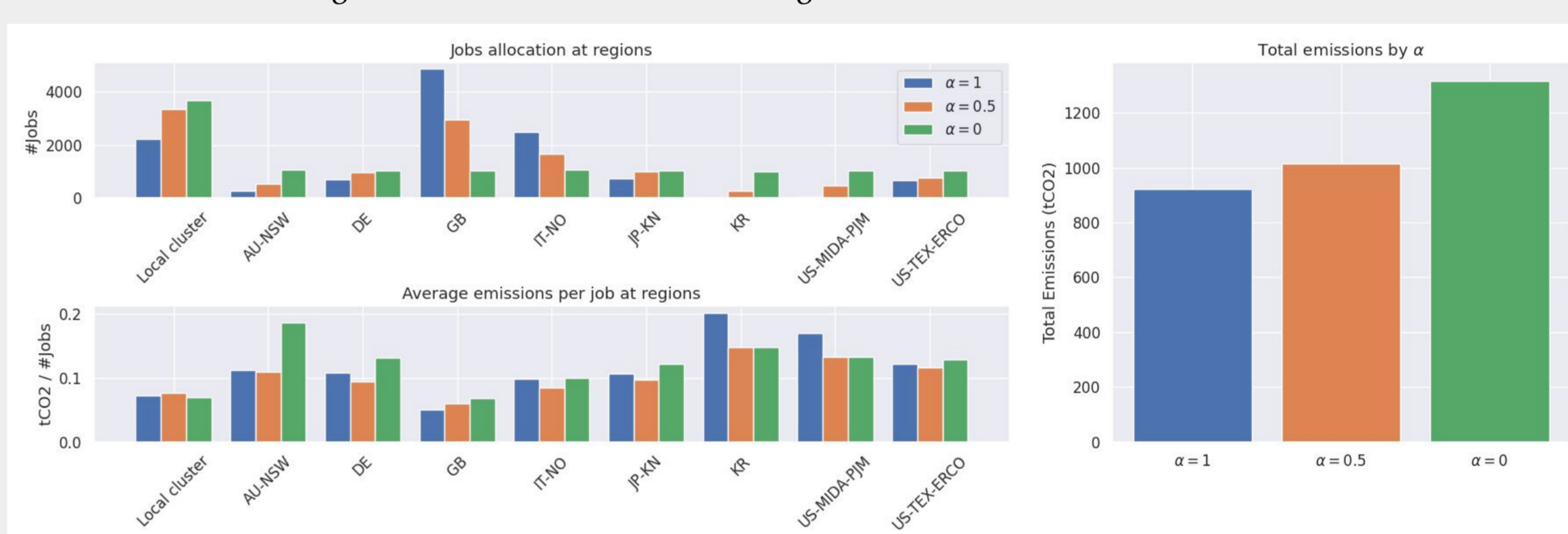
Forecasters evaluation:

Best model forecasts (Chronos2) cause 6% emissions overhead against ground truth data



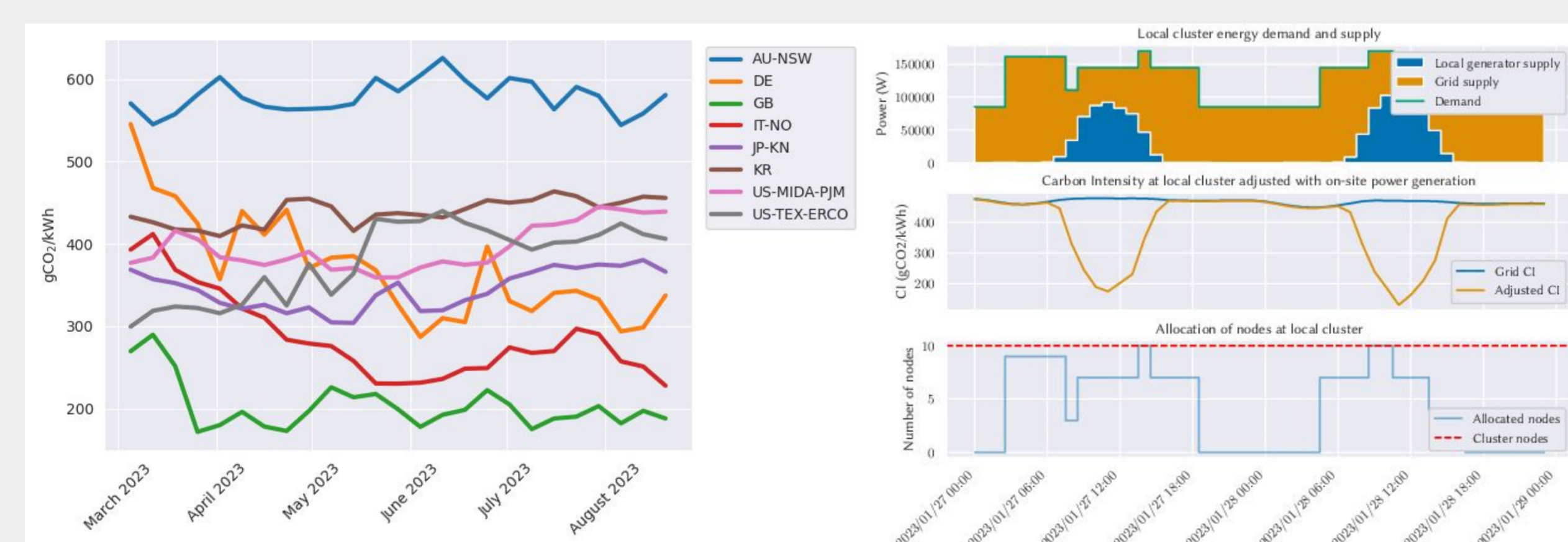
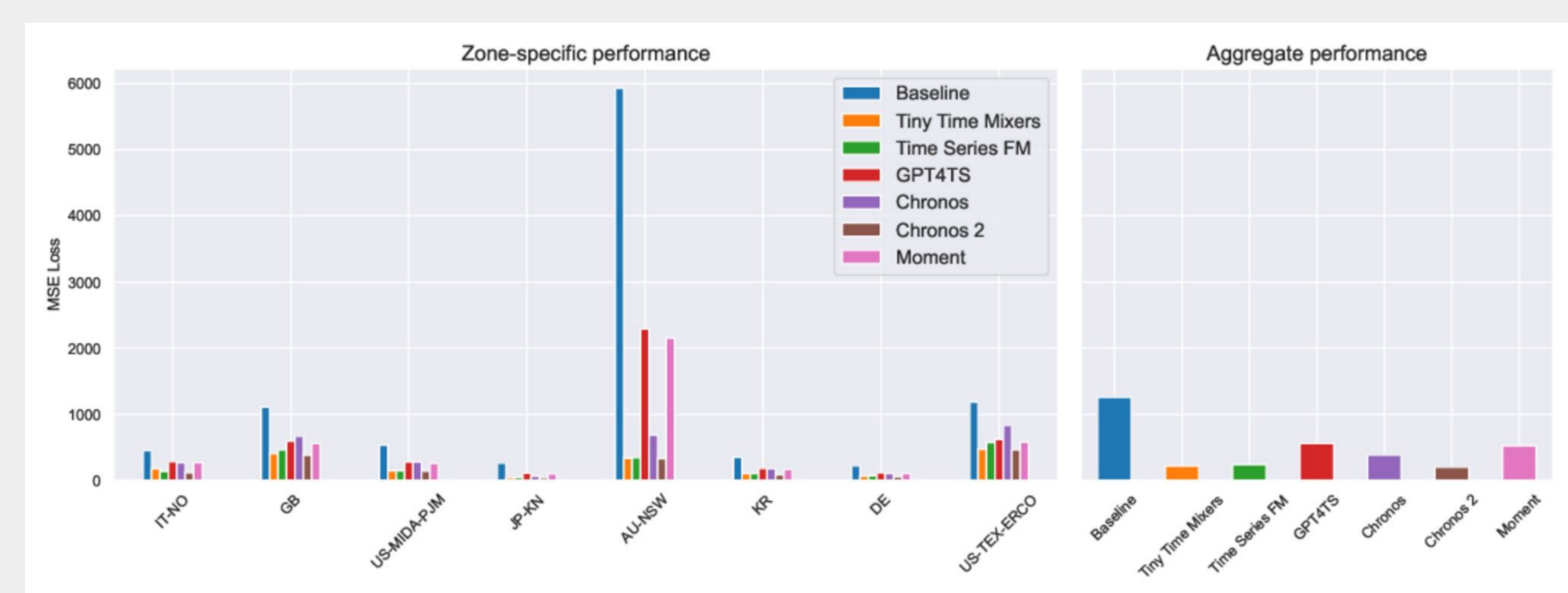
Trade-off evaluation:

Balanced scheduling maximizes local cluster usage with 10% emissions overhead



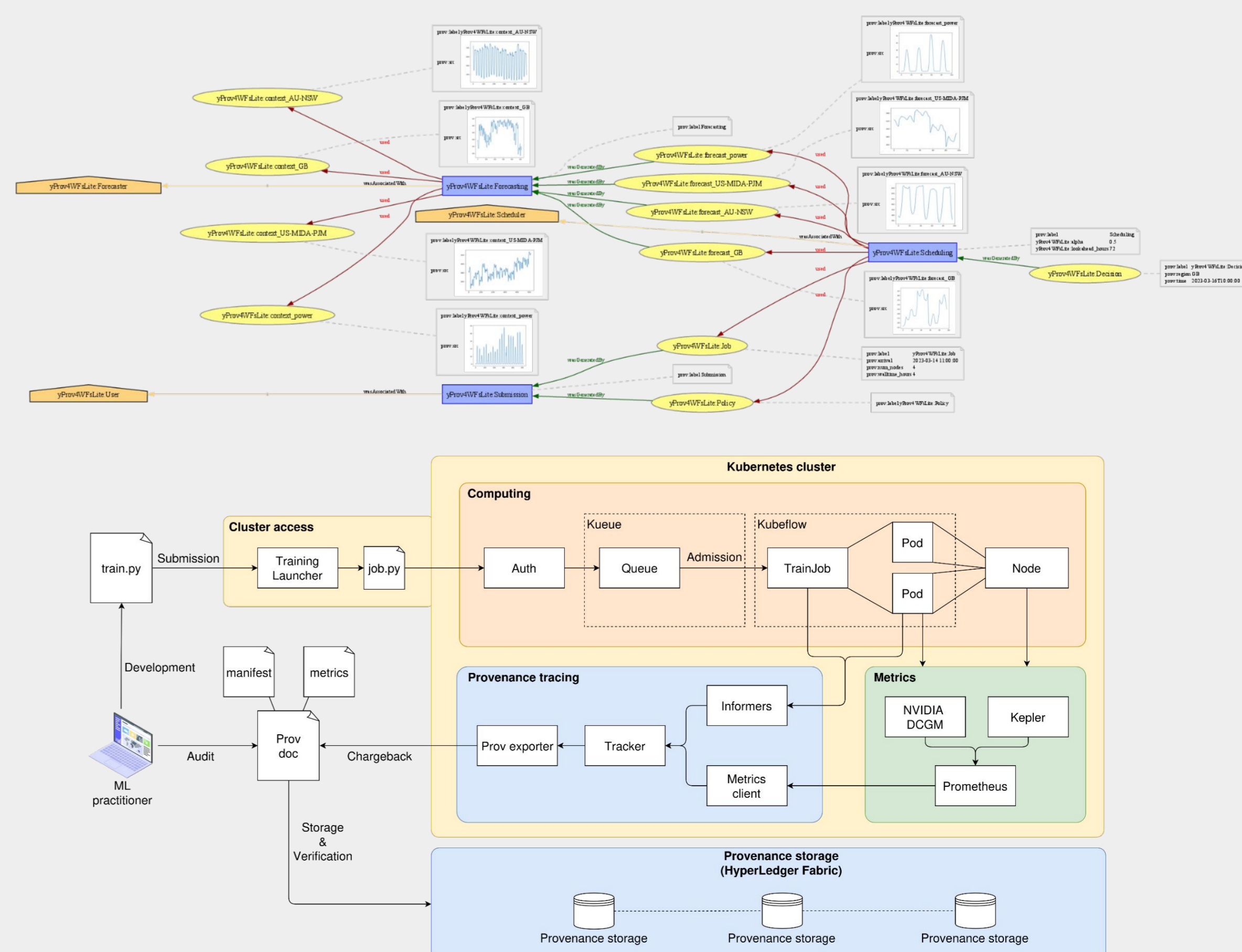
Carbon Intensity & power forecasting

- Low carbon approach: zero-shot forecasting avoiding further training or finetuning
- Benchmark of state-of-the-art time series forecasters
 - Chronos2 results as best performing model
- Local carbon intensity forecasts adjusted considering on-site renewable power generation
 - Weighted contributions in the electricity supply
- Evaluation over 2023 data in 8 regions with different CI profiles
 - AU-NSW, DE, GB, IT-NO, JP, KR, US-TEX-ERCO, and US-MIDA-PJM



Provenance accounting

- Provenance-based accounting module to enable transparent accountability assessments of the automated decisions of the system
- Production of a provenance document in W3C-PROV standard format to describe management of cloud resources
- Description of information used in the management (CI forecasts, resources manifests), algorithmic decisions, execution metrics
- Support for transparent and verifiable chargeback and reporting system
- Creation of historical record of fine-grained information on management of resources to support optimizations through agentic AI approach



Carbon-aware scheduling

Linear Programming (LP) model for optimally allocating workloads such that:

- Emissions are minimized:** space and time shifting scheduling approach to delaying jobs or geographically offloading to periods and regions with minimal carbon intensity
- Local cluster usage is maximized:** costs optimization by amortizing local hardware resources purchase and reduction in public cloud services expenses

α parameter to control trade-off between objectives

Constraints on solution feasibility and respect of usage quota limits

Optimization solved in Python PuLP

$$x_{ij} = \begin{cases} 1, & \text{if workload is allocated at cluster } j \text{ at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

$$E(t, j) = x_{ij} \cdot \sum_{\tau=t}^{t+D} CI_j(\tau)$$

$$L(t, j) = x_{ij} \cdot \begin{cases} N, & \text{if } j = l, \\ 0, & \text{otherwise.} \end{cases}$$

$$t^*, j^* = \operatorname{argmin}_{t, j} \{ \alpha E(t, j) - (1 - \alpha) L(t, j) \}$$

such that:

$$\sum_{j=1}^J x_{ij} = 1 \quad t + D \leq T$$

$$\forall \tau \in [t, t + D], N + N_i(\tau) \leq N^{max}, Q^{max}$$