

Optimizing emissions for machine learning training

Sachini Ekanayake[^], Tapan Shah^{*}, Scott Evans^{*}

[^]: University of Albany,

^{*}: GE Vernova Research



GE VERNOVA

Our portfolio of energy businesses

AI and Emissions

Model	Total Training Energy Consumption (MWh)	Gross tCO2e model training
Evolved Transformer	7.5	3.2
T5	85.7	46.7
Meena	232	96.4
Gshard -600B	24.1	4.8
Switch Transformer	179	72.2
BERT	1.5	1.4
GPT-3	1287	552






**As AI models become ubiquitous, the numbers are going to scale.
Need for algorithmic, infrastructural and hardware innovations.**

Motivation

Best departing flights

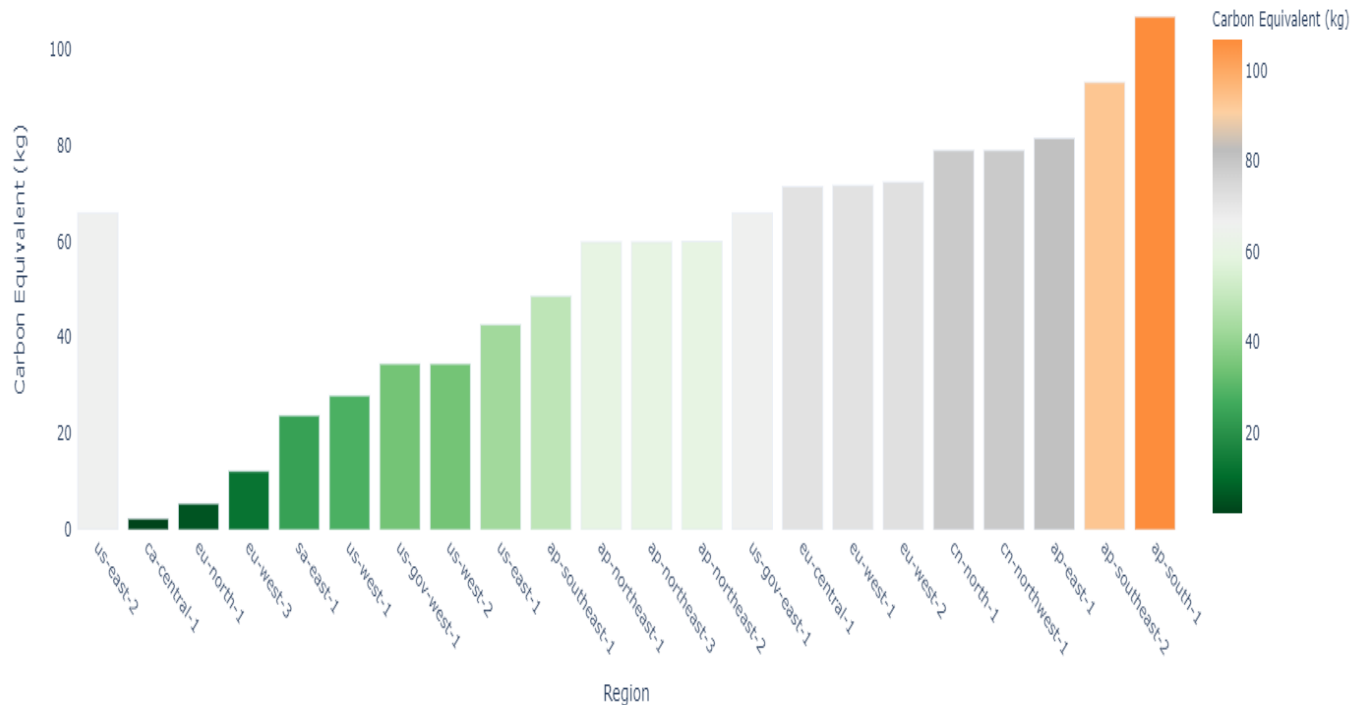
Ranked based on price and convenience ⓘ Prices include required taxes + fees for 1 adult. Optional charges and [bag fees](#) may apply.

Sort by: ↑↓

 jetBlue	3:43 PM – 12:28 AM⁺¹ JetBlue · American	5 hr 45 min SFO–JFK	Nonstop	333 kg CO ₂ +14% emissions ⓘ	 \$238 round trip	▼
	1:25 PM – 9:50 PM Delta	5 hr 25 min SFO–JFK	Nonstop	367 kg CO ₂ +26% emissions ⓘ	\$238 round trip	▼
	4:05 PM – 12:38 AM⁺¹ Delta	5 hr 33 min SFO–JFK	Nonstop	367 kg CO ₂ +26% emissions ⓘ	\$238 round trip	▼
	7:00 AM – 3:30 PM Alaska	5 hr 30 min SFO–JFK	Nonstop	222 kg CO ₂ -24% emissions ⓘ	\$328 round trip	▼

Similarly, while training AI models, cognizance of appropriate compute resourcing (location and time is essential)

Dependence on location



API and data provided by [CodeCarbon](https://mlco2.github.io) — CodeCarbon 0.0.1 documentation (mlco2.github.io)

- Same deep learning models trained on differently located compute infrastructure.
- At certain locations, the carbon emission can be ~30 times lower compared to average.

**Optimal Location of model training can reduce carbon by
>1000%**

Dependence on time

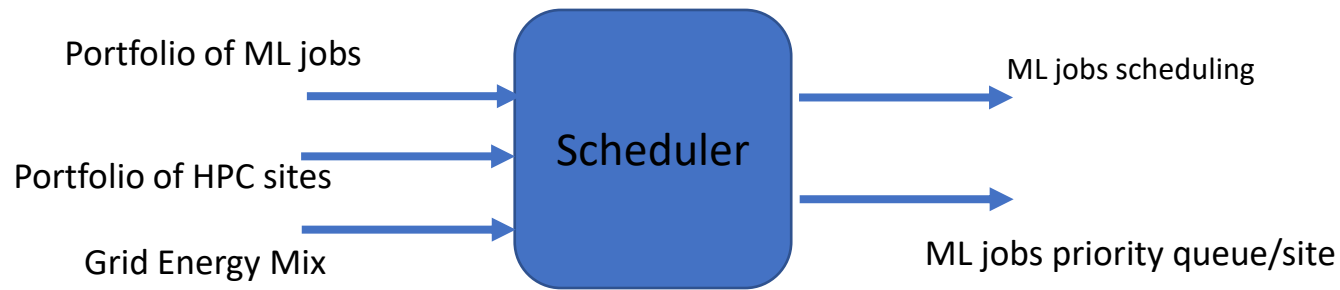


API and data provided by [CodeCarbon](#) — [CodeCarbon 0.0.1 documentation \(mlco2.github.io\)](#)

- 1) Same deep learning models trained in Ohio at different timestamps.
- 2) At certain timestamps, the carbon emission can be half the average.

Optimal Time of Model Training can half the carbon emissions

Problem Formulation



Assumptions:

- Communication costs are not considered in the optimization.
- Spatio-temporal energy mix forecast are available.
- Computation time and energy are approximations derived as a function of ML model (SVM/NN/RF), #features and hardware.

- We formulate a constrained optimization problem
- The objective is to minimize the carbon emissions.
- Input: ML Task characterized by model used, quality expectation, dataset properties, expected computation time and energy requirement, energy mixes at the HPC site
- Auxiliary input: Priority of the job, start time margins
- The decision variables are the location and time window

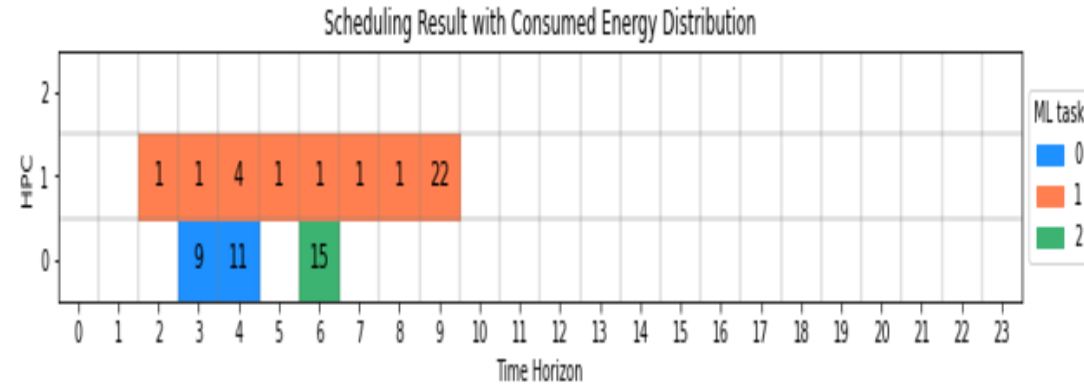
Empirical Results on Simulate Data

Example Setting

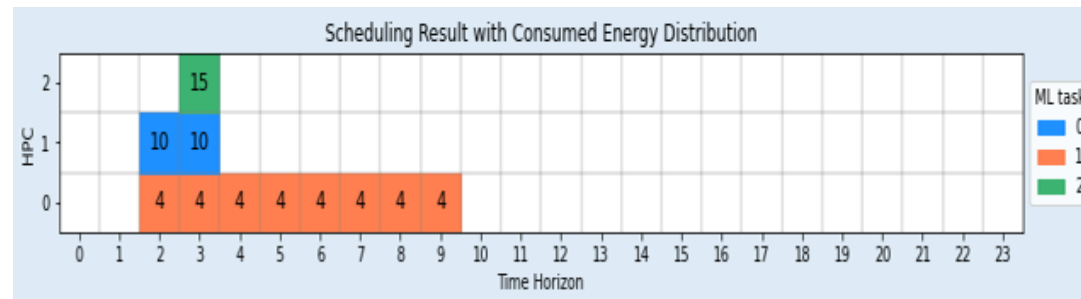
ML Task j	0	1	2
Expected Energy Requirement	20	32	15
Arrival Time	2	2	3
Computation time	2	8	1
Start time Margin	2	4	6

Key Observations

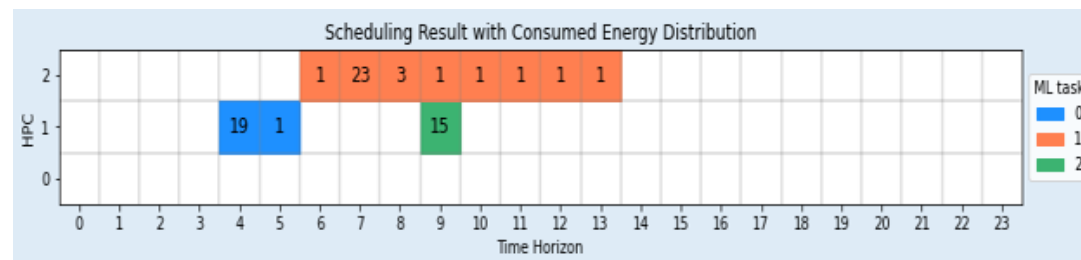
- For optimal scheduling, the training starts much later, using margin for minimizing emissions
- Across multiple setting, 10 and 4 times less emission as compared to random and greedy scheduling, respectively.



Optimal Scheduling



Greedy Scheduling



Random Scheduling

Next Steps

- Add uncertainty into expected computation and energy required as well as energy mix forecasts.
- Conduct simulations for a large real world set of ML tasks.
- Incorporate communication costs into the optimization.
- Interface the scheduler as a wrapper to a corporate/cloud HPC scheduling service.