Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service

Baolin Li, Siddharth Samsi,
Vijay Gadepally, Devesh Tiwari

Northeastern University

MIT LINCOLN LABORATORY

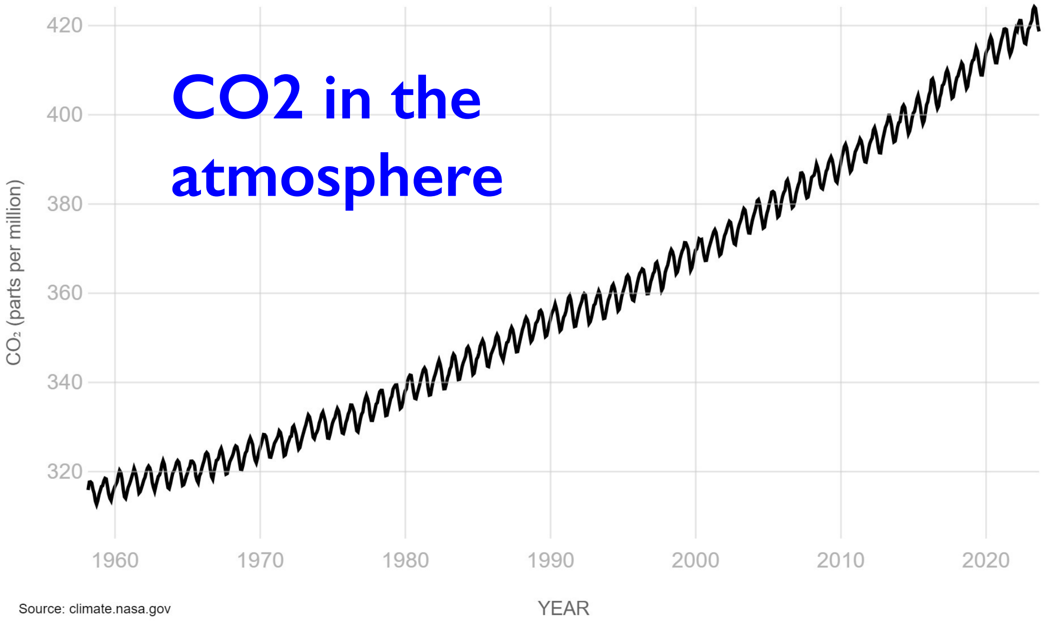
Reducing Carbon Emission Is of Critical Importance

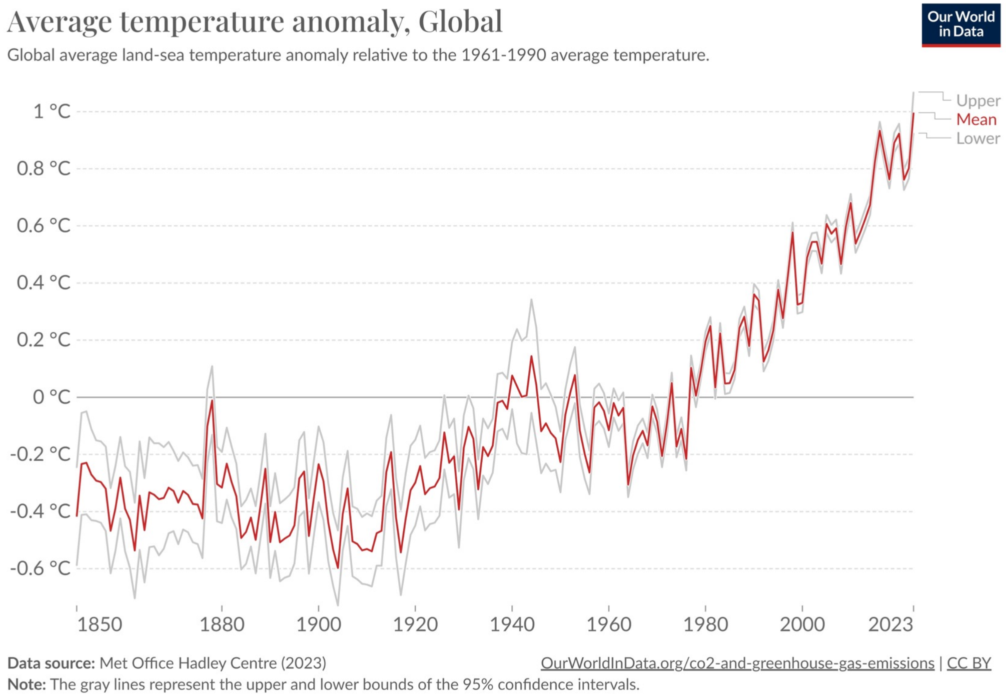
The Washington Post
Democracy Dies in Darkness

CLIMATE Environment Weather Climate Solutions Climate Lab Green Living Business of Climate

World is on brink of catastrophic warming, U.N. climate change report says

A dangerous climate threshold is near, but 'it does not mean we are doomed' if swift action is taken, scientists say





Machine Learning Inference Accounts for Significant Compute Cycles in Today's Datacenters

Inference represents 60% of their AI infrastructure emissions

David Patterson et. al., Computer'22

Expanded infrastructure capacity by 2.5x to meet ML inference demand

Carole-Jean Wu et al., MLSys'22

Inference is the big market, with an estimated 80 to 90% of cost of ML

Jensen Huang, GTC

The Gap between ML Inference and Sustainability

Totally Green: Evaluating and Designing Servers for Lifecycle Environmental Impact

Jichuan Chang Justin Meza Parthasarathy Ranganathan
Amip Shah Rocky Shih Cullen Bash
Hewlett Packard Laboratories, Palo Alto, USA
{jichuan.chang,justin.meza,partha.ranganathan,amip.shah,rocky.shih,cullen.bash}@hp.com

ASPLOS'12

Chasing Carbon: The Elusive Environmental Footprint of Computing

Udit Gupta[1,2], Young Geun Kim[3], Sylvia Lee[2], Jordan Tse[2],
Hsien-Hsin S. Lee[2], Gu-Yeon Wei[1], David Brooks[1], Carole-Jean Wu[2]

[1]Harvard University, [2]Facebook Inc., [3]Arizona State University

ugupta@g.harvard.edu carolejeanwu@fb.com

HPCA'21

Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters

Bilge Acun, acun@meta.com, Meta, USA
Benjamin Lee, leebcc@seas.upenn.edu, University of Pennsylvania, Meta, USA
Fiodar Kazhamiaka, fiodar@stanford.edu, Stanford University, USA
Kiwan Maeng, kwmaeng@meta.com, Meta, USA
Udit Gupta, uditg@meta.com, Harvard University, Meta, USA
Manoj Chakkaravarthy, mchakkar@meta.com, Meta, USA
David Brooks, dbrooks@eecs.harvard.edu, Harvard University, Meta, USA
Carole-Jean Wu, carolejeanwu@meta.com, Meta, USA

ASPLOS'22

SUSTAINABLE AI: ENVIRONMENTAL IMPLICATIONS, CHALLENGES AND OPPORTUNITIES

Carole-Jean Wu[1] Ramya Raghavendra[1] Udit Gupta[1 2] Bilge Acun[1] Newsha Ardalani[1] Kiwan Maeng[1] Gloria Chang[1] Fiona Aga Behram[1] James Huang[1] Charles Bai[1] Michael Gschwind[1] Anurag Gupta[1] Myle Ott[1] Anastasia Melnikov[1] Salvatore Candido[1] David Brooks[1 2] Geeta Chauhan[1] Benjamin Lee[1 3] Hsien-Hsin S. Lee[1] Bugra Akyildiz[1] Max Balandat[1] Joe Spisak[1] Ravi Jain[1] Mike Rabbat[1] Kim Hazelwood[1]

ABSTRACT
This paper explores the environmental impact of the super-linear growth trends for AI from a holistic perspective, spanning *Data*, *Algorithms*, and *System Hardware*. We characterize the carbon footprint of AI computing by examining the model development cycle across industry-scale machine learning use cases and, at the same time, considering the life cycle of system hardware. Taking a step further, we capture the operational and manufacturing carbon footprint of AI computing and present an end-to-end analysis for *what* and *how* hardware-software design

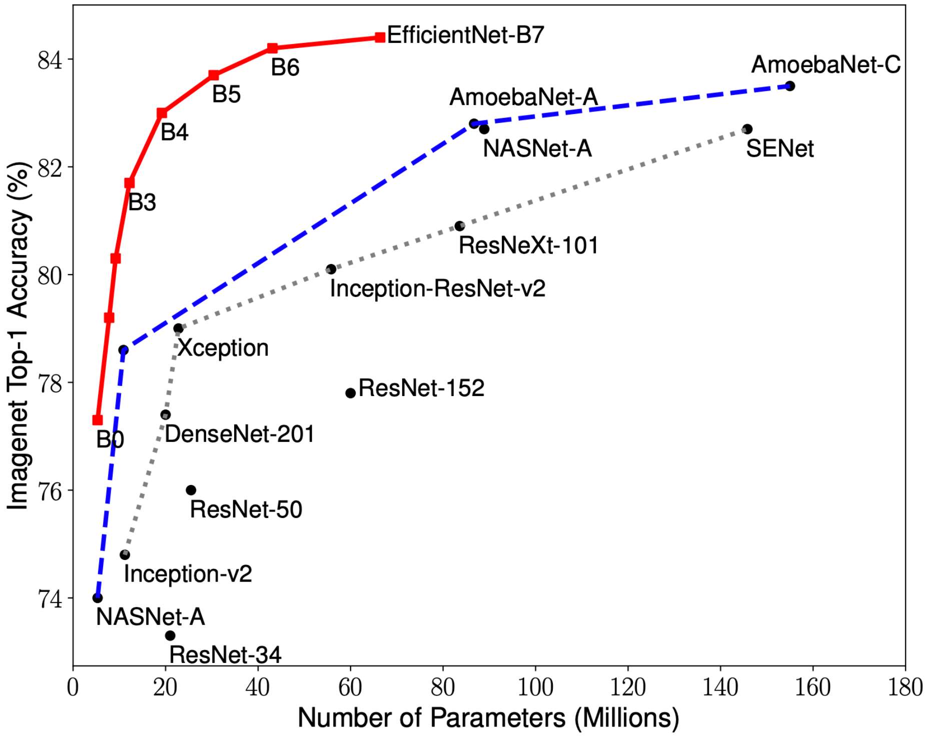
MLSys'22

No carbon-aware ML inference solution yet!

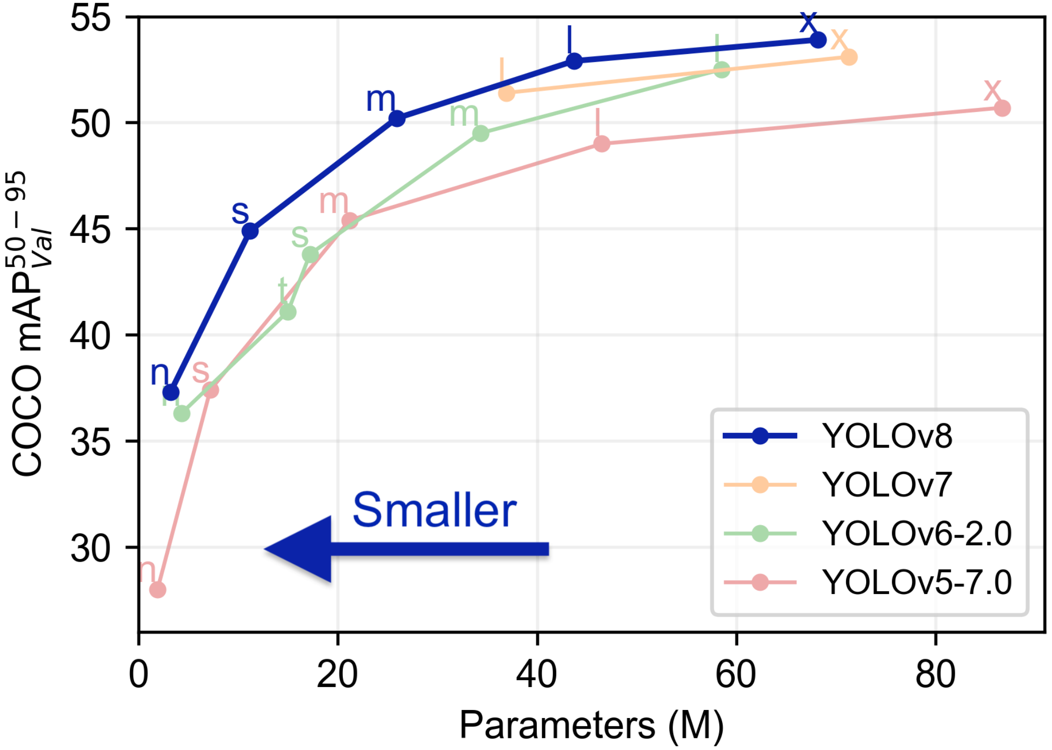
Opportunity I: Mixed Quality Models

The same model architecture can have a family of model variants with different number of parameters and sizes, yielding different accuracy levels.

Image Classification



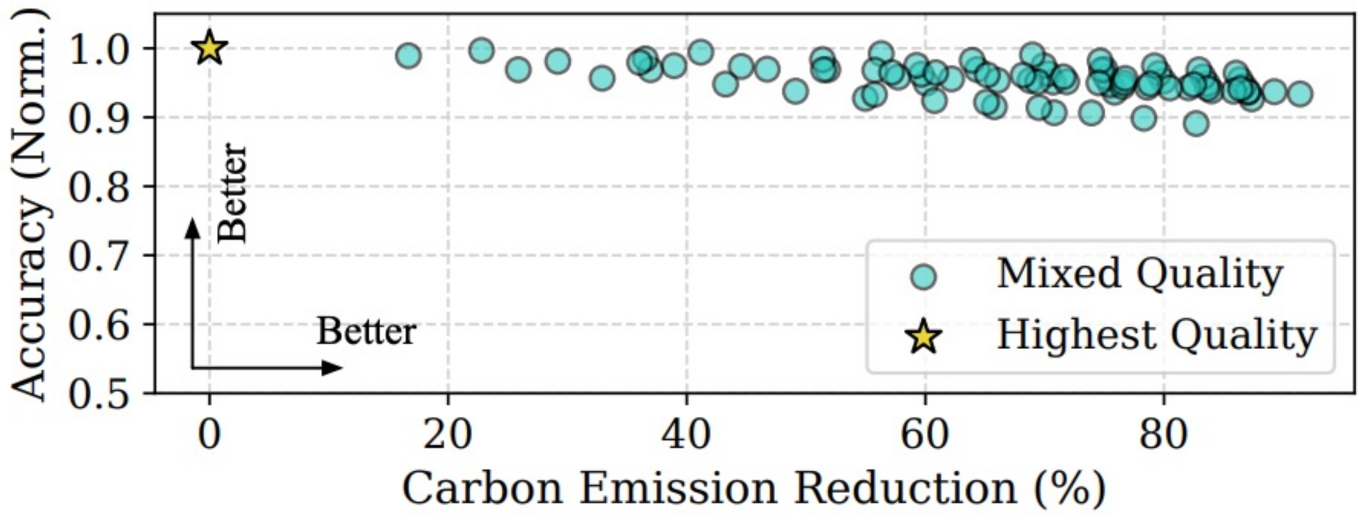
Object Detection



Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019.

Opportunity I: Mixed Quality Models

Using mixture of model variants saves carbon without significantly impacting accuracy



Applications and corresponding model variants

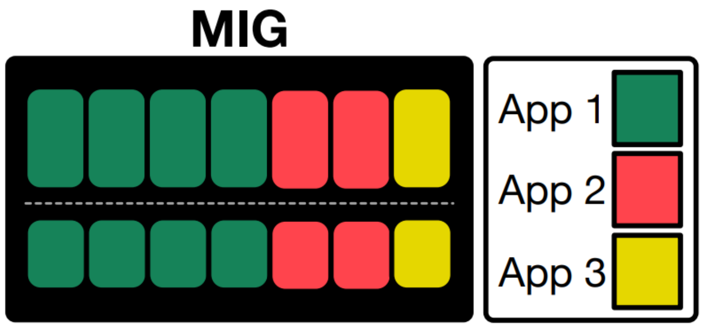
Application	Dataset	Architecture	Variants
Object Detection	MS COCO [50] (Microsoft)	YOLOv5 [51] (Ultralytics)	YOLOv5l, YOLOv5x, YOLOv5x6
Language Modeling	SQuADv2 [52] (Stanford)	ALBERT [21] (Google)	V2-base, V2-large, V2-xlarge, V2-xxlarge
Image Classification	ImageNet [53] (Princeton/Stanford)	EfficientNet [22] (Google)	B1, B3, B5, B7

Opportunity II: GPU Partitioning

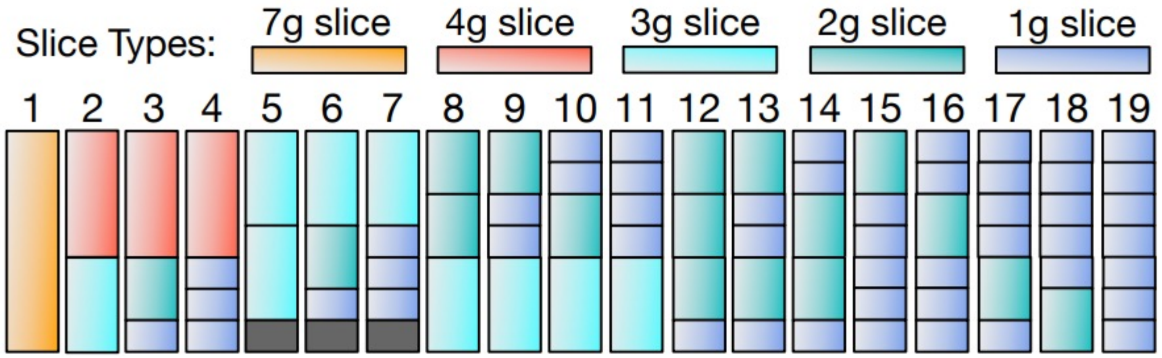
When GPU is underutilized, it can be partitioned into multiple individual GPU slices

Multi-Instance GPU (MIG)

Slice	Compute	Memory	Cache	Max Count
7g.40gb	7 GPC	40 GB	Full	1
4g.20gb	4 GPC	20 GB	4/8	1
3g.20gb	3 GPC	20 GB	4/8	2
2g.10gb	2 GPC	10 GB	2/8	3
1g.5gb	1 GPC	5 GB	1/8	7



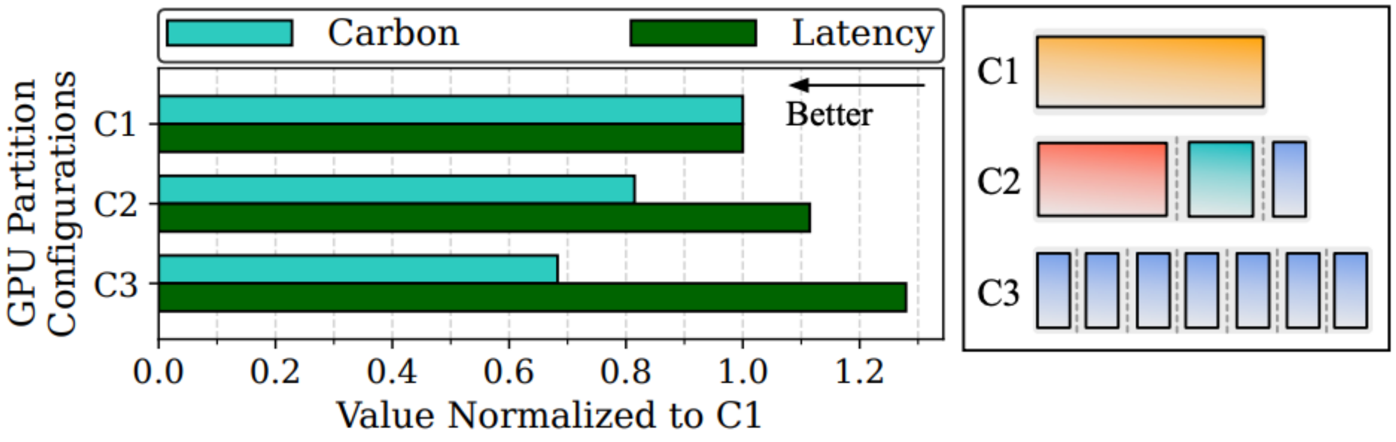
19 different ways to partition a GPU



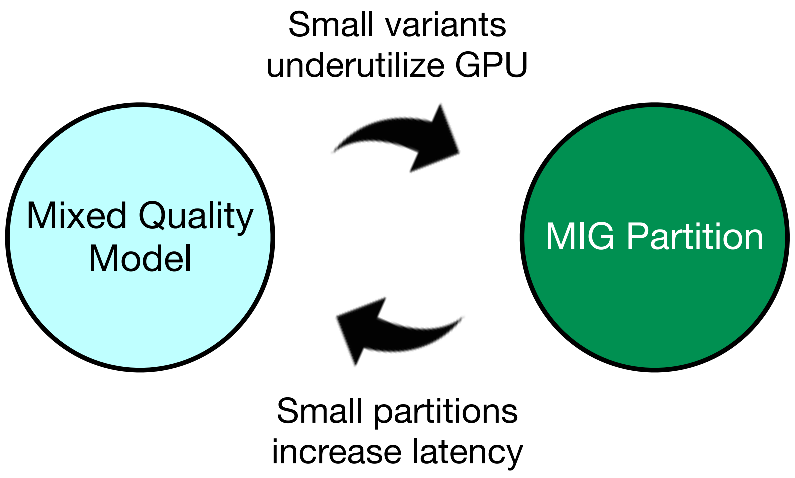
Li, Baolin et. al. "MISO: exploiting multi-instance GPU capability on multi-tenant GPU clusters." In Proceedings of the 13th Symposium on Cloud Computing, pp. 173-189. 2022.

Opportunity II: GPU Partitioning

More efficient usage of GPU by partitioning also saves carbon per request

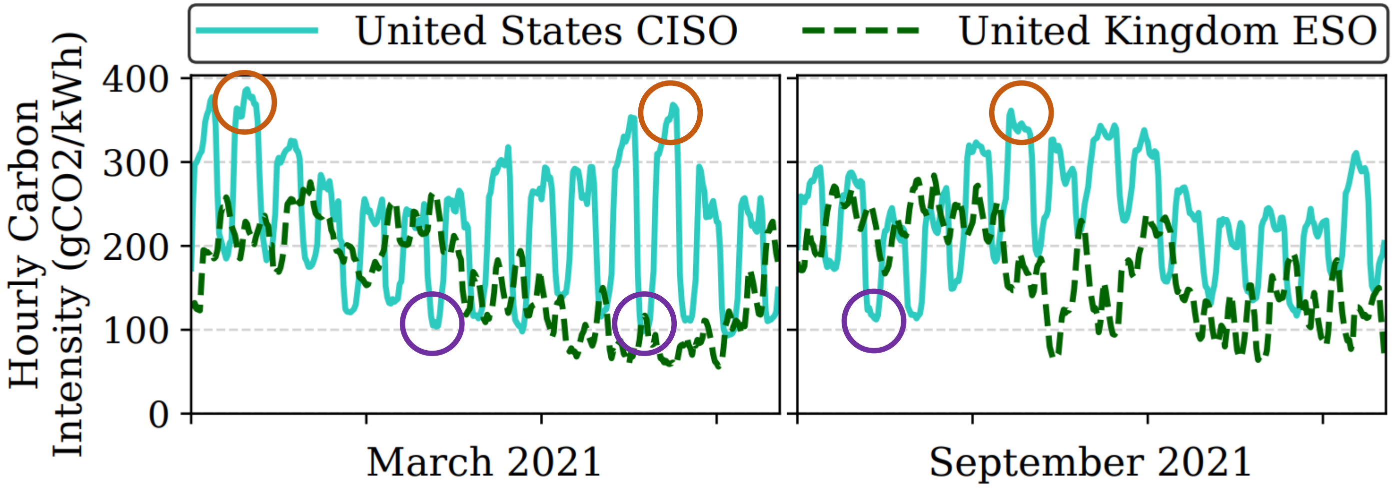


Model variants and MIG-based GPU partitioning complements each other



Opportunity III: Carbon Intensity Variation

Configuring model variants and GPU partition allows us to reduce carbon emission, but this needs to exploited carefully in conjunction with the carbon intensity of the energy source



High gain from saving energy

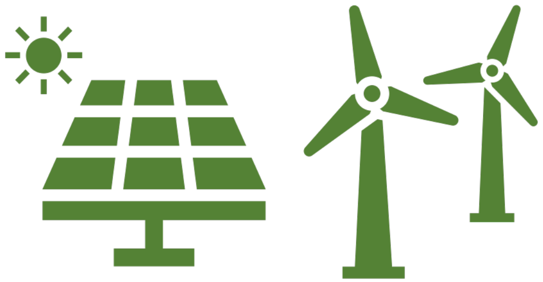
Low gain from saving energy

$$C_{\text{op}} = I_{\text{sys}} \cdot E_{\text{op}}$$

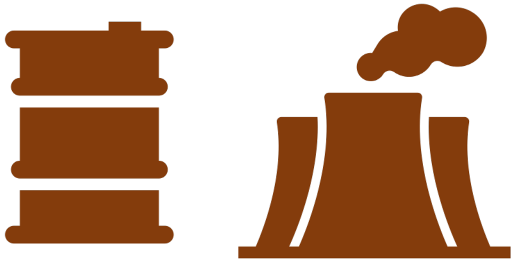
Operational carbon emission = carbon intensity x energy

Carbon-Aware Machine Learning Inference

How much effort we put into saving energy should depend on current carbon intensity



Low carbon intensity:
aim for quality!



High carbon intensity:
aim for reducing carbon footprint!

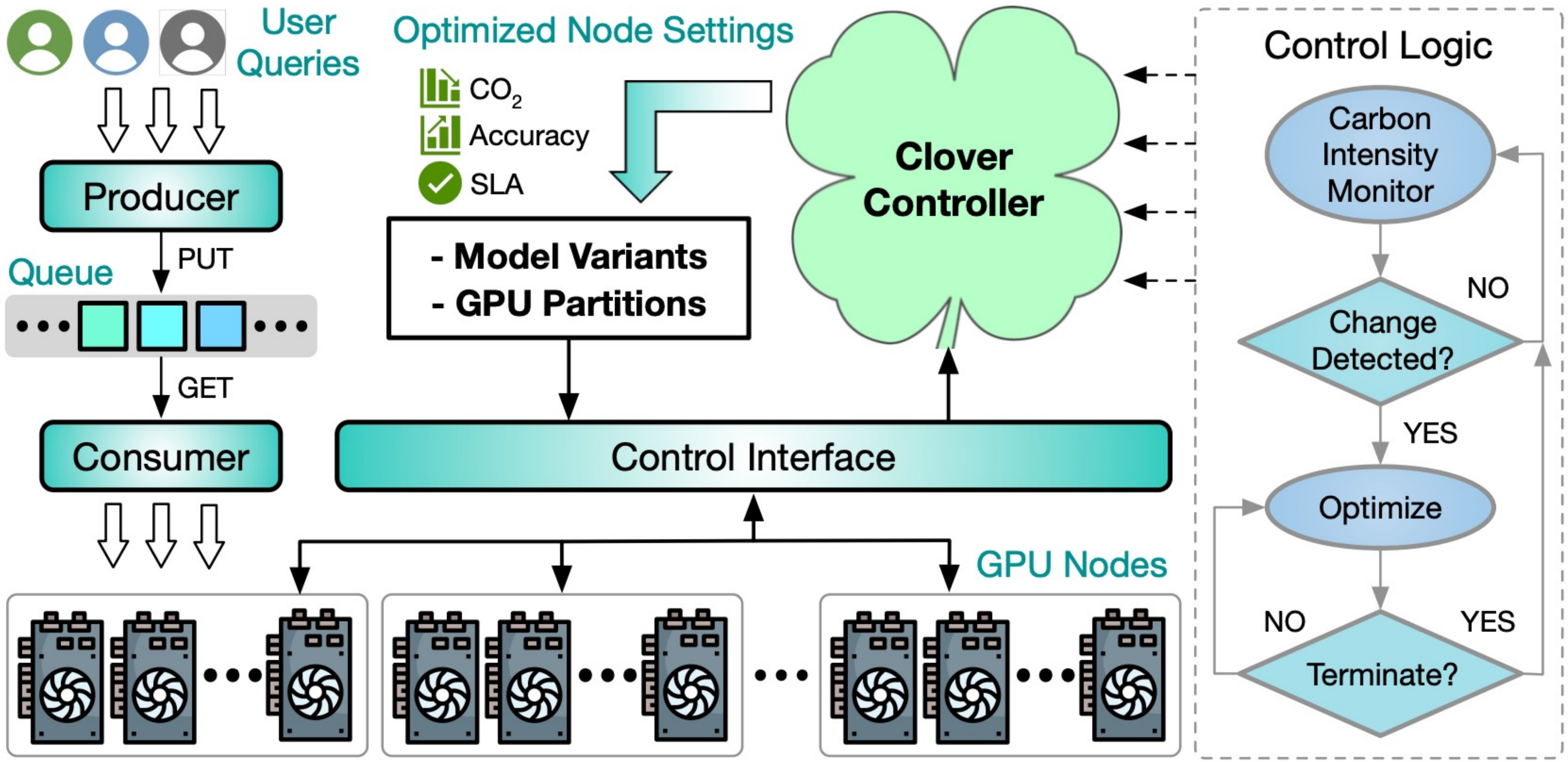
How to build a carbon-aware system for ML inferences?

Clover Objectives and Key Ideas

Objectives

Give the current carbon intensity, adjust the mixture of model variants and MIG partition to optimize the combined objective of accuracy & carbon

Clover System Overview



Optimizing the dual objective of accuracy and carbon

$$\Delta Accuracy = \frac{A(\boldsymbol{x}^p, \boldsymbol{x}^v) - A_{base}}{A_{base}} \times 100\%$$

Accuracy

GPU Partition

Model variant

Highest accuracy possible

$$\Delta \text{Carbon} = \frac{C_{\text{base}} - E(\boldsymbol{x}^p, \boldsymbol{x}^v) \cdot ci}{C_{base}} \times 100\%$$

Energy per request

Carbon intensity

Base carbon

Combined objective function using a coefficient

$$\boxed{f(\boldsymbol{x}^p, \boldsymbol{x}^v) = \lambda \cdot \Delta Carbon + (1 - \lambda) \cdot \Delta Accuracy}$$

Optimization

$$\max_{\boldsymbol{x}^p, \boldsymbol{x}^v} f(\boldsymbol{x}^p, \boldsymbol{x}^v)$$
$$\text{s.t. } L(\boldsymbol{x}^p, \boldsymbol{x}^v) \leq L_{tail}$$

Carbon-Aware Formulation

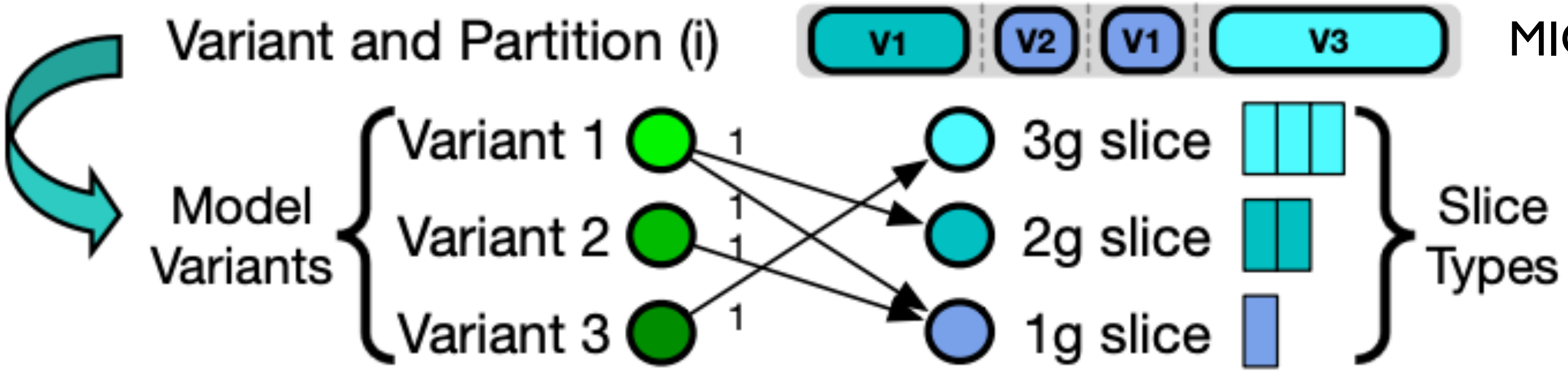
Why does this optimization problem formulation make Clover carbon-aware?

$\lambda = 0.1$, $C_{base}=1000$	Configs	ΔAccuracy	$E(x^p,x^v)\cdot ci$	ΔCarbon	**Objective**	Preference
ci = 500	Config. A (E = 0.4)	-4.0	200	80	**4.4**	✓
ci = 500	Config. B (E = 1.2)	-2.0	600	40	**3.2**	✗
ci = 100	Config. A (E = 0.4)	-4.0	40	96	**6.0**	✗
ci = 100	Config. B (E = 1.2)	-2.0	120	88	**7.0**	✓

Optimality between two configurations depends on the carbon intensity

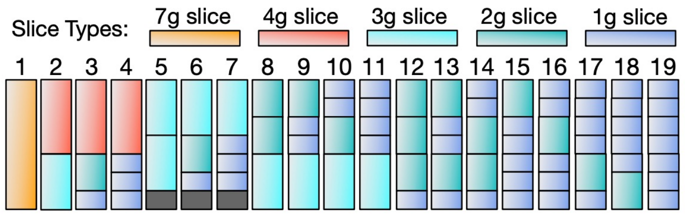
How to optimize the Clover objective?

Model the configurations as a bipartite graph and apply neighbor search based on graph similarity



MIG config = 9

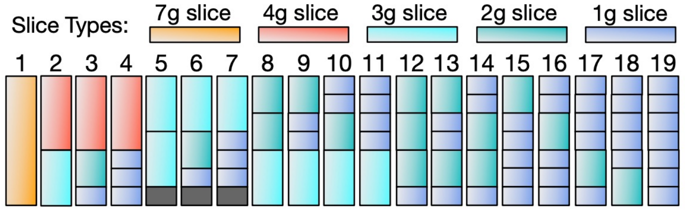
Edge Weight: number of instances hosted on slice type



Why model the configurations as graphs?

Removal of configurations that yield the same objective function values

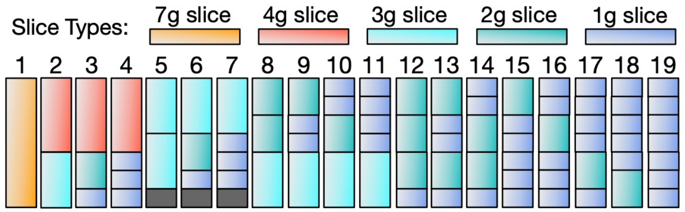
- MIG provides performance isolation – only the slice type matters
- Which GPU the variant is hosted or the order of variants in a GPU changes the x^p, x^v representation, but they would eventually result in the same graph representation



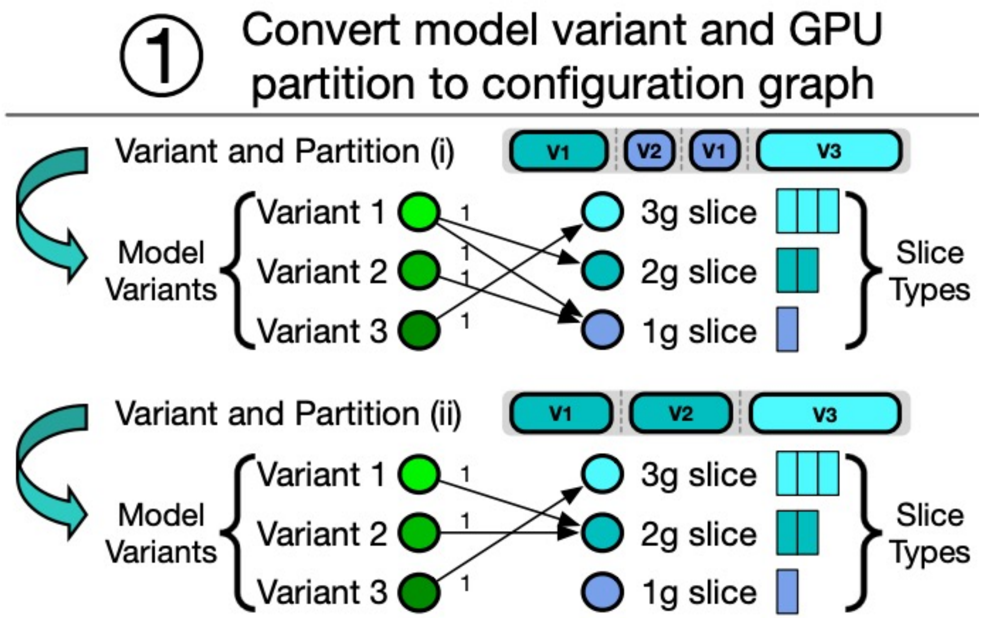
Why model the configurations as graphs?

Can scale to arbitrary system size without adding vertices/edges to the graph

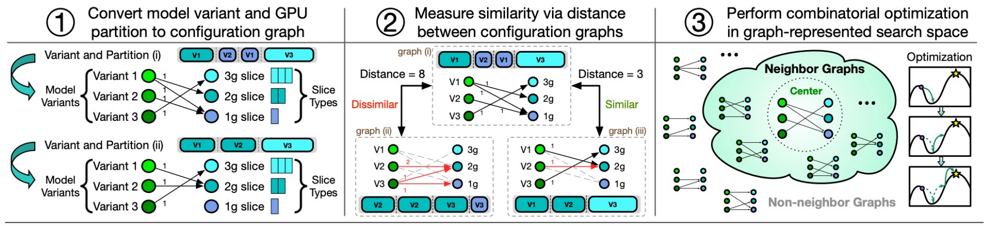
- The graph size only depends on number of model variant and GPU slice types
- The graph configurations are additive – when adding more GPUs to the system, we simply add the edge weights of the new GPUs to current graph. But in x^p, x^v representation, we need to increase the dimensionality.



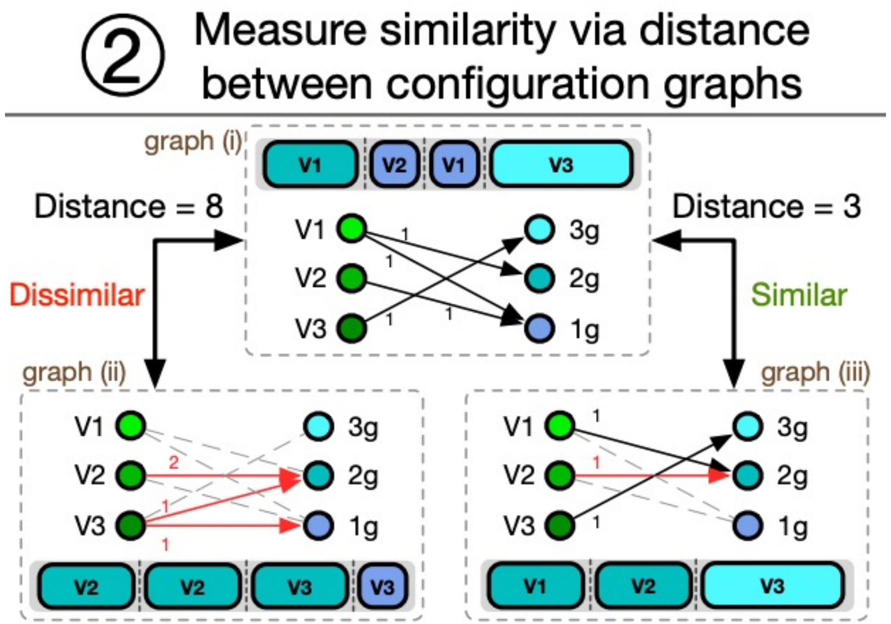
Clover Optimization Workflow I

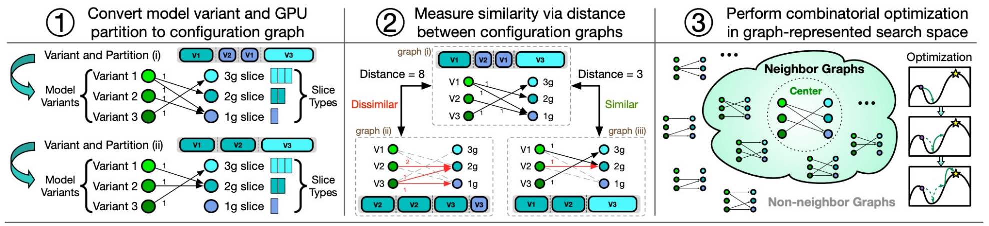


Create one graph representation for services on all GPUs in the system



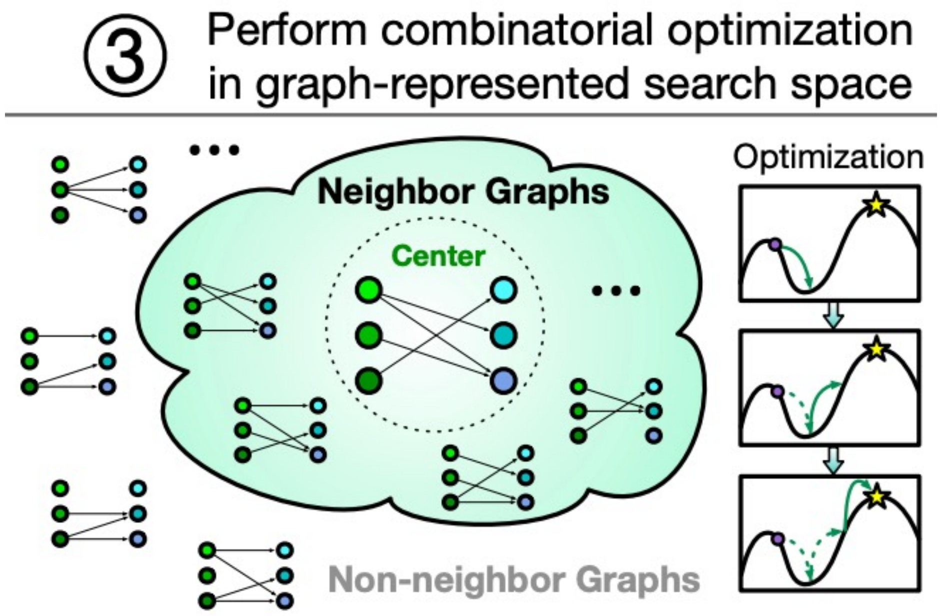
Clover Optimization Workflow II





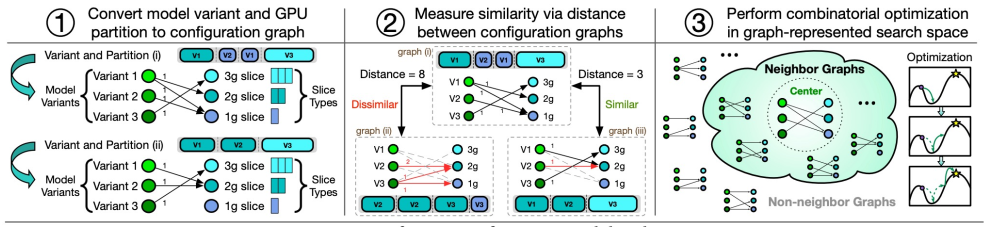
Similarity between two graph representations are measured by graph editing distance (GED)

Clover Optimization Workflow III



Apply neighborhood search algorithm to optimize in graph space.

Clover uses Simulated Annealing.



Experimental Methodology

Setup

- 5 nodes
- 2 AMD EPYC 7542 CPUs each node
- 2 NVIDIA A100 GPUs each node

Metrics

- Carbon Emission
- Accuracy
- SLA

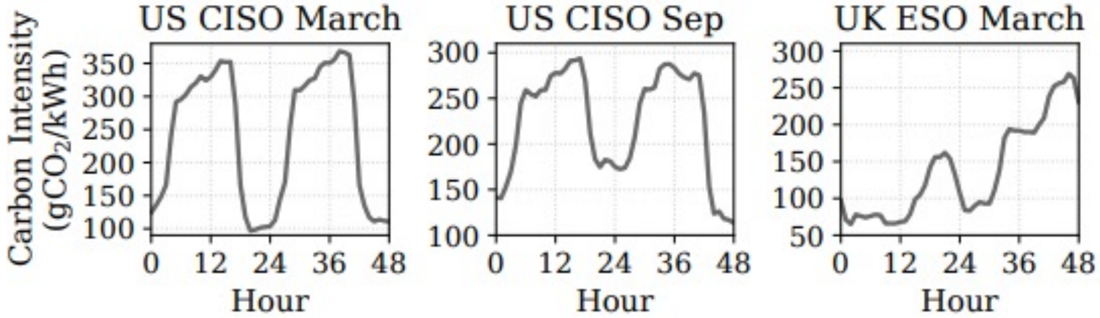
Workloads

- YOLO (Object Detection)
- ALBERT (Language Modeling)
- EfficientNet (Image classification)

Schemes

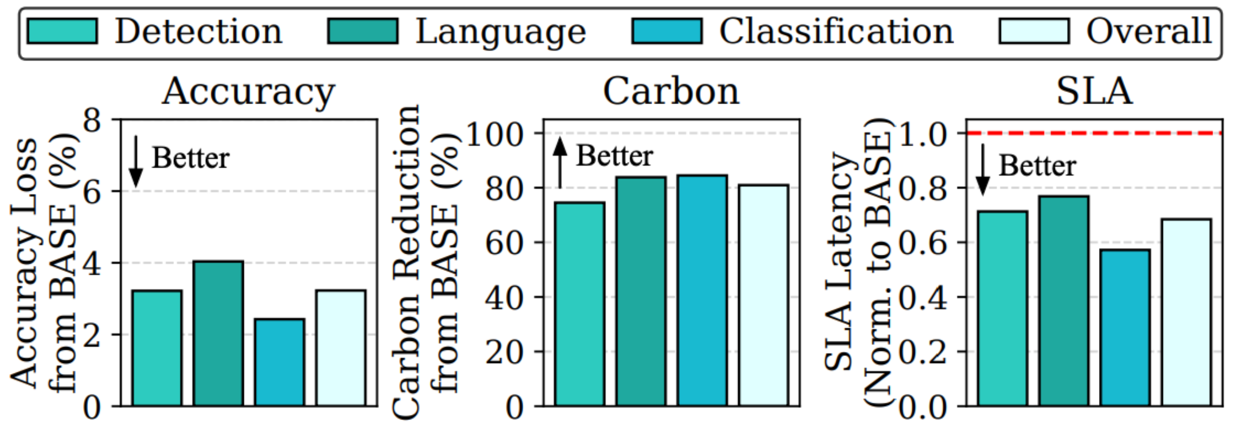
- BASE: highest-quality model, exclusive GPU
- CO2OPT
- Blover: basic Clover w/o graph
- ORACLE

Real carbon intensity trace

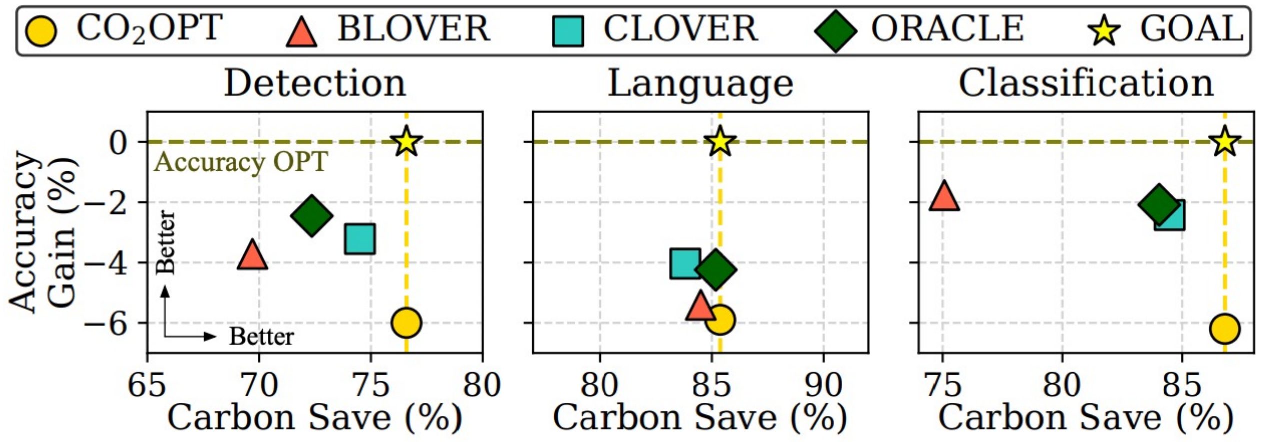


Clover significantly reduces carbon emission with negligible accuracy degradation

Saves carbon emission by 80% while operating under SLA latency

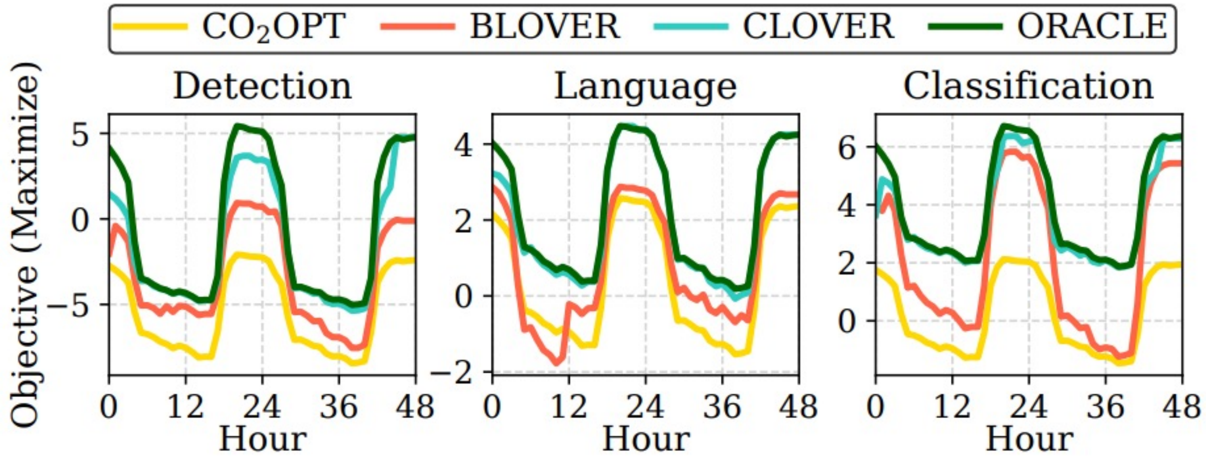


Clover outperforms competing schemes and is always closest to ORACLE

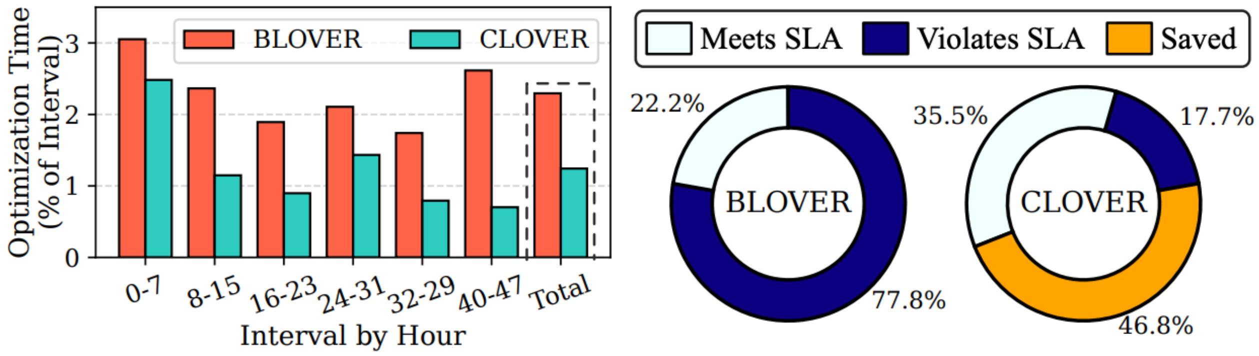


Clover's effectiveness comes from its superior optimization process

Clover gets closer and closer to ORACLE over time

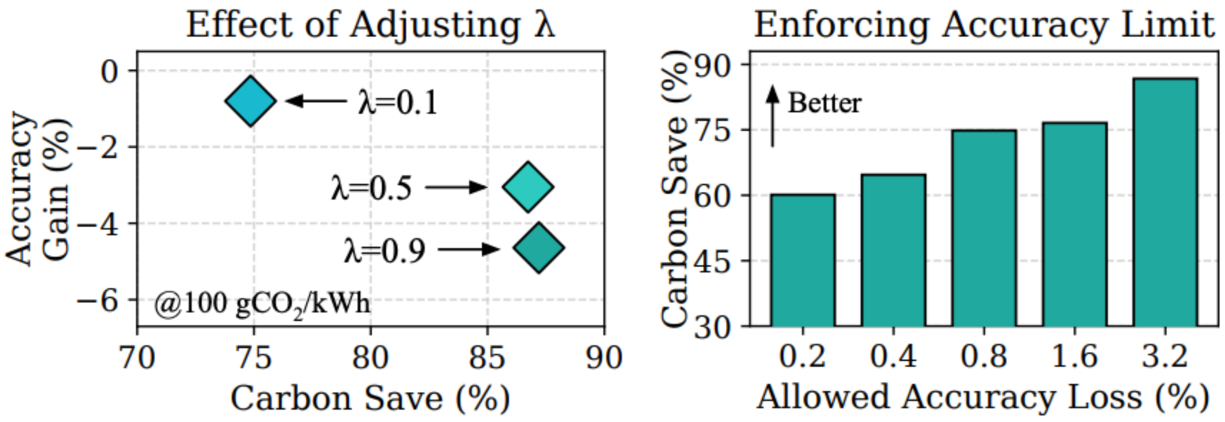


Clover has much lower optimization overhead compared to Blover

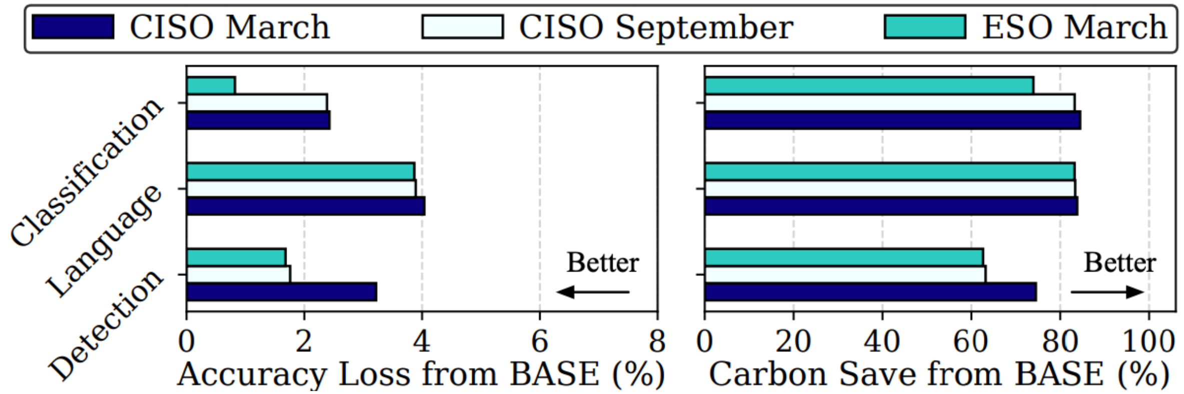


Clover is adaptive and robust

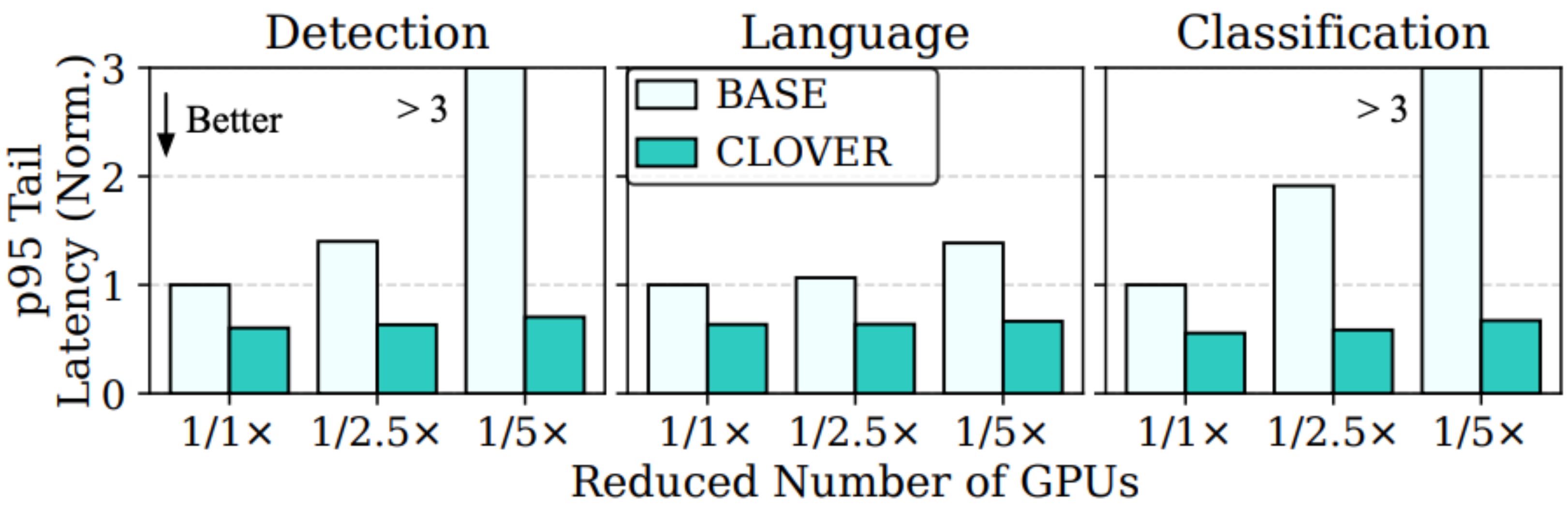
User can control the trade-off between accuracy and carbon, and even enforcing accuracy limit



Clover is effective across geographical regions and seasons with varying carbon intensity



Clover reduces the number of GPUs needed to meet service target (embodied carbon savings)



Clover's co-location and mixed-quality serving enable reductions in number of GPUs

This is essentially reducing the carbon emission needed to produce these devices (embodied carbon)

Clover Summary of Key Contributions

Clover is the first carbon-aware machine learning inference system.

Clover actively configures the model variant mixture and GPU partition to adapt to the varying carbon intensity levels.

Clover uses a novel graph-space optimization method to significant reducing carbon emission while maintaining high service quality.

Contact

Baolin Li

li.baol@northeastern.edu

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001, and United States Air Force Research Laboratory Cooperative Agreement Number FA8750-19-2- 1000. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering, or the United States Air Force. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein