# Adaptive GPU Power Capping: Balancing Energy Efficiency, Thermal Control and Performance

Tanish Desai*, Jainam Shah*, Gargi Alavani, Snehanshu Saha, Santonu Sarkar
Department of Computer Science & Information Systems, BITS Pilani – Goa Campus, APPCAIR

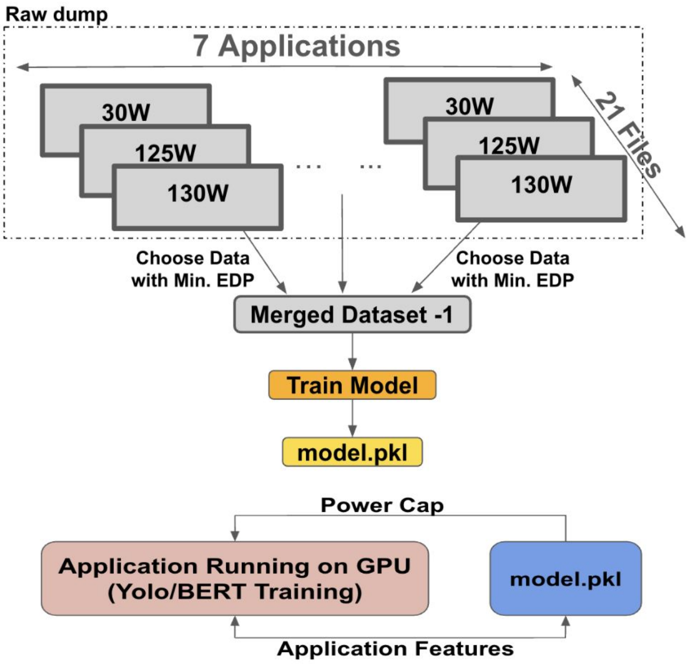
## Introduction

We present an ML-driven, real-time GPU power-capping strategy—leveraging utilization, memory use, temperature and frequency—to adaptively set optimal caps. This yields up to **12.9%** energy savings, **11.4%** lower temperatures, and only a **2.7%** performance hit.

## Methodology

**Training Pipeline**

- **Data Collection:**
  - Performance data collected from three GPU kernels (DenseNet, CUDA matrix multiplication, CNN image processing).
  - Programs created by running individual and combined kernels on an NVIDIA RTX 4000 Ada GPU.
  - Metrics recorded: power, temperature, energy, GPU utilization, memory utilization, and frequency.
  - Dataset constructed by selecting the power cap minimizing Energy Delay Product (EDP).
- **Model Selection and Training:**
  - Models evaluated: Linear Regression, Random Forest, Decision Tree, XGBoost, CatBoost.
  - k-fold cross-validation used to prevent overfitting.




## Model Performance

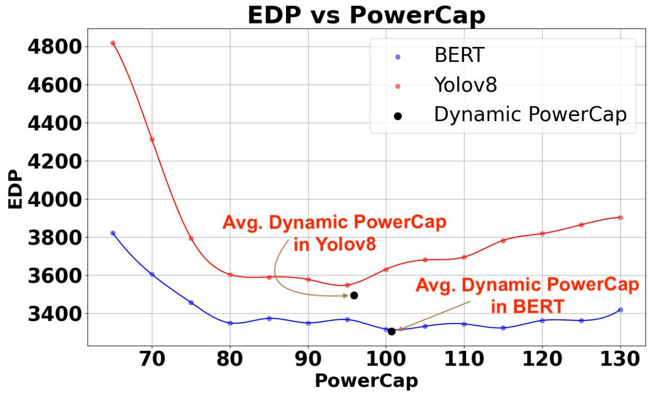
- **Table 1: Model Comparison**
  Shows MSE, MAE, and R² for each model.
  - CatBoost achieved minimum MSE and highest R² score.

**Table 1: Minimisation Metric: EDP**

| Model | MSE | MAE | $R^2$ |
|---|---|---|---|
| Linear Regression | 18,979,841.72 | 3048.73 | 0.8869 |
| Random Forest Regressor | 4,728,518.83 | 628.92 | 0.9719 |
| Decision Tree Regressor | 10,701,882.51 | 627.56 | 0.9369 |
| XGBoost Regressor | 7,901,473.70 | 686.46 | 0.9521 |
| CatBoost Regressor | 4,018,389.22 | 834.17 | 0.9761 |

## Results

- **Benchmark Applications:**
  - The benchmark applications we used included training a YOLOv8 model and fine-tuning a BERT model.
- **Dynamic vs Static Power Capping:**
  - Dynamic model converges rapidly to optimal power cap during execution.
  - Achieved significantly higher energy savings and temperature reductions than any static power cap when tested for YOLO, and delivered comparable energy savings for BERT.
  - Minor performance loss observed




## Application Metrics

- YOLOv8: 12.87% energy gain, 11.38% temp reduction, 2.69% performance loss.
- BERT: 6.45% energy gain, 10.56% temp reduction, 3.26% performance loss.

**Table 2: Performance, Energy, and Power Metrics for Applications**

| Application | Performance Loss | Energy Gain | Temp. Gain | Avg. Dynamic Power Cap | Best Static Power Cap (EDP) |
|---|---|---|---|---|---|
| Yolov8 | 2.69% | 12.87% | 11.38% | 95.875 | 95 |
| BERT | 3.26% | 6.45% | 10.56% | 100.669 | 100 |

## Conclusion

- Dynamic power capping using machine learning significantly improves energy efficiency and thermal control with minimal performance loss.
- Enables smarter, greener supercomputing practices.
- Future work: Extend to multi-GPU systems and new architectures for broader applicability.

## References

[1] D. Zhao, S. Samsi, J. McDonald, B. Li, D. Bestor, M. Jones, D. Tiwari, and V. Gadepally, "Sustainable Supercomputing for AI: GPU Power Capping at HPC Scale," in Proc. 2023 ACM Symposium on Cloud Computing (SoCC '23), pp. 588–596, ACM,2023. DOI: 10.1145/3620678.3624793.
[2] J. McDonald, B. Li, N. Frey, D. Tiwari, V. Gadepally, and S. Samsi, "Great Power,Great Responsibility: Recommendations for Reducing Energy for Training Language Models," in Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1962–1970, ACL, 2022. DOI: 10.18653/v1/2022.findings-naacl.151.
[3] G. A. Prabhu, T. Desai, S. Potdar, N. Gogari, S. Saha, and S. Sarkar, "Estimating Power Consumption of GPU Application Using Machine Learning Tool," in Proc.2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI),pp. 734–739, 2024.
[4] J.-R. Yu, C.-H. Chen, T.-W. Huang, J.-J. Lu, C.-R. Chung, T.-W. Lin, M.-H. Wu, Y.-J. Tseng, and H.-Y. Wang, "Energy Efficiency of Inference Algorithms for Clinical Laboratory Data Sets: Green Artificial Intelligence Study," J. Med. Internet Res., vol.24, no. 1, e28036, 2022. DOI: 10.2196/28036.
[5] P. Stanley-Marbell, "How Device Properties Influence Energy-Delay Metrics and the Energy-Efficiency of Parallel Computations," in Proc. Workshop on Power-Aware Computing and Systems (HotPower '15), pp. 31–35, ACM, 2015. DOI: 10.1145/2818613.2818744.