



I/O-Cool: Using I/Os for Temperature Regulation When Training on the Edge

YANIS MOHAMMEDI, LINA GANA, MERIEM BOUZOUAD, JALIL BOUKHOBZA
09/2026

Outline

1.CONTEXT

2.MOTIVATION

3.I/O COOL

4.EVALUATION

5.CONCLUSION

Outline

1.CONTEXT

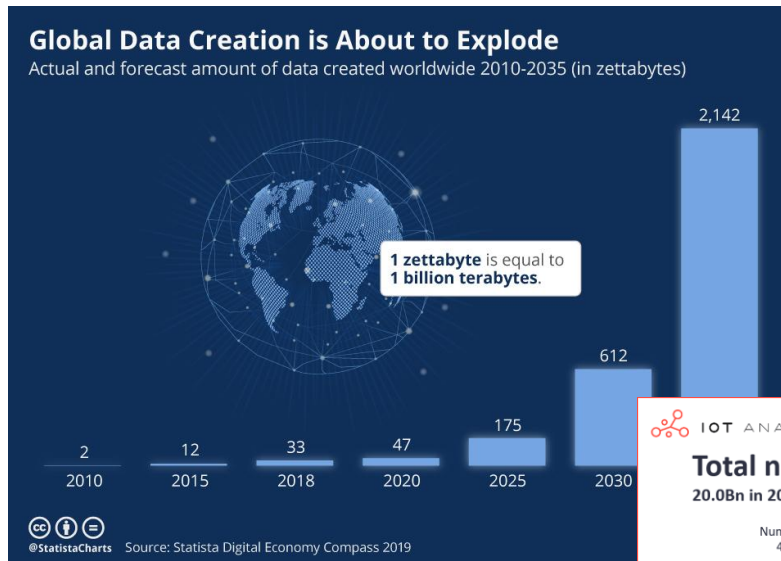
2.MOTIVATION

3.I/O COOL

4.EVALUATION

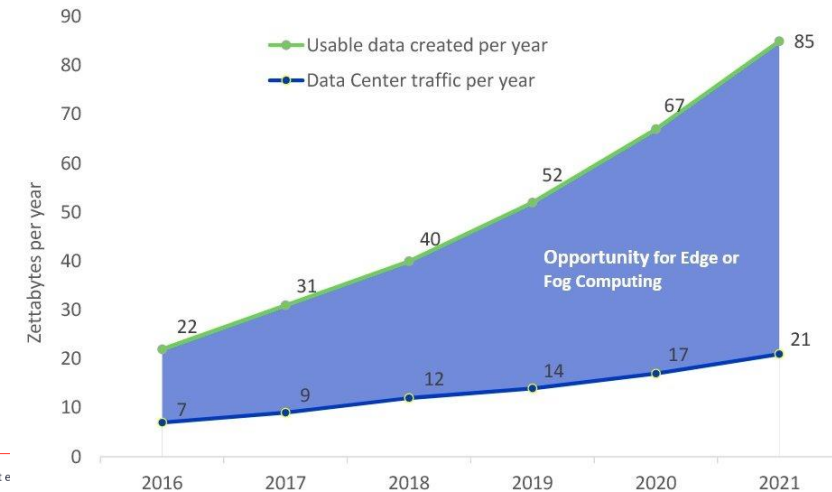
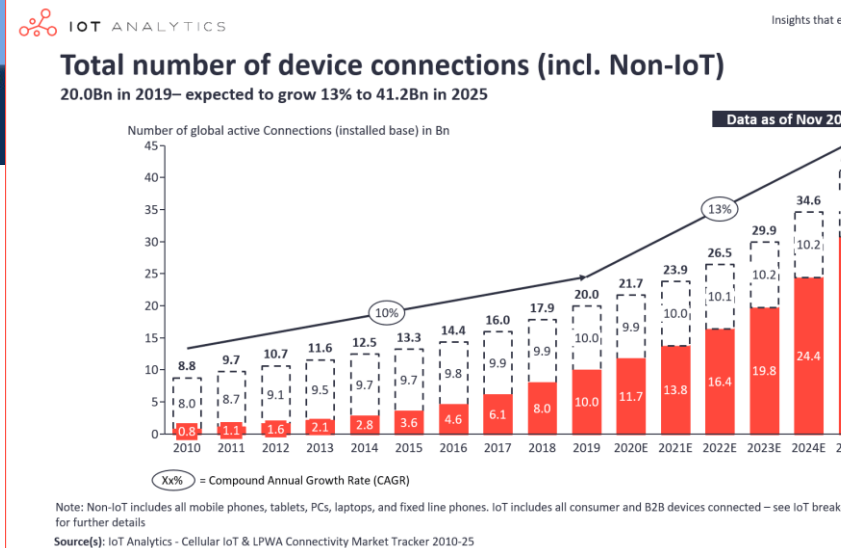
5.CONCLUSION

Context – data at the edge



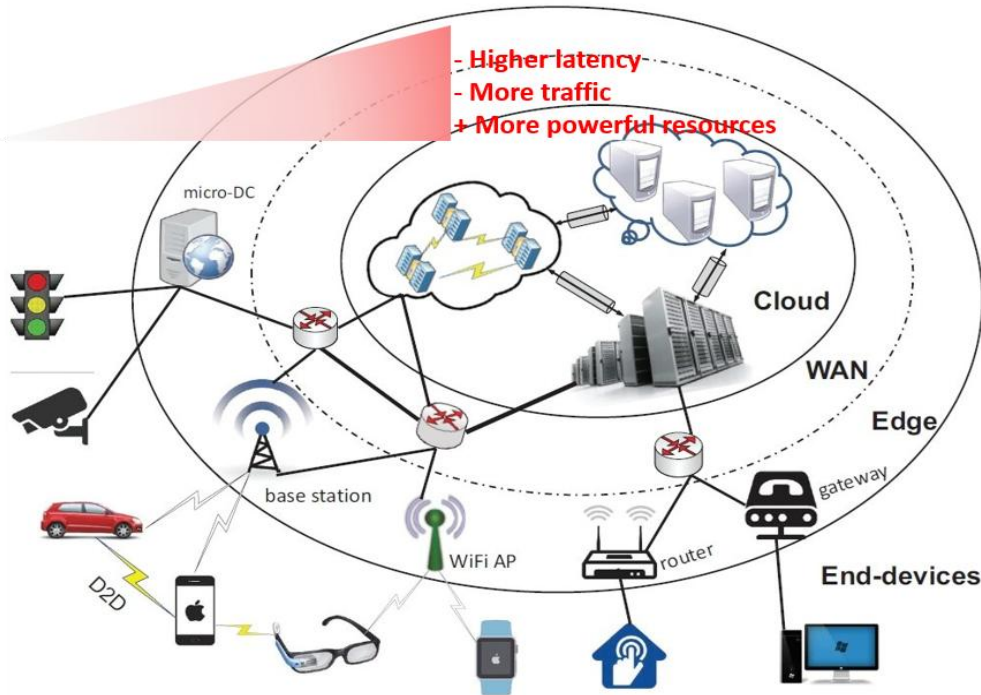
Data volume grows exponentially >500 Zettabytes by 2030!

IDC forecasts that 80 billion IoT devices will be connected by 2025



Data generated on devices > 4x data generated in data centers (2021 according to Cisco Global Cloud Index)

Context – Edge Computing



[1] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," in Proceedings of the IEEE, vol. 107, no. 8, pp. 1738-1762, Aug. 2019,.

[2] G. Boesch, **Edge Intelligence: Edge Computing and ML (2024 Guide)**, <https://viso.ai/edge-ai/edge-intelligence-deep-learning-with-edge-computing/>, 2023, accessed last on 21 sept 2024

[3] Thomas Bittman, Gartner analyst, [The Edge Will Eat the Cloud](#), 2017

❑ « ...paradigm that pushes computing tasks and services from the network core to the network edge » [1]

❑ « ...capturing, storing, processing, analyzing data near its source » [2]

- Move computation capabilities to the data (Vs data to the computing resources)

❑ Objectives:

- Reducing latency
- Preserving privacy + security
- Reducing network traffic

❑ Targets low resource devices:

- Compute/memory/storage/network
- energy constraint

Context – AI at the edge

❑ Edge intelligence:

- «providing reliable and real-time intelligent services on the network's edge» [1],
- « ... derives from the concept of combining edge computing and ML » [2]

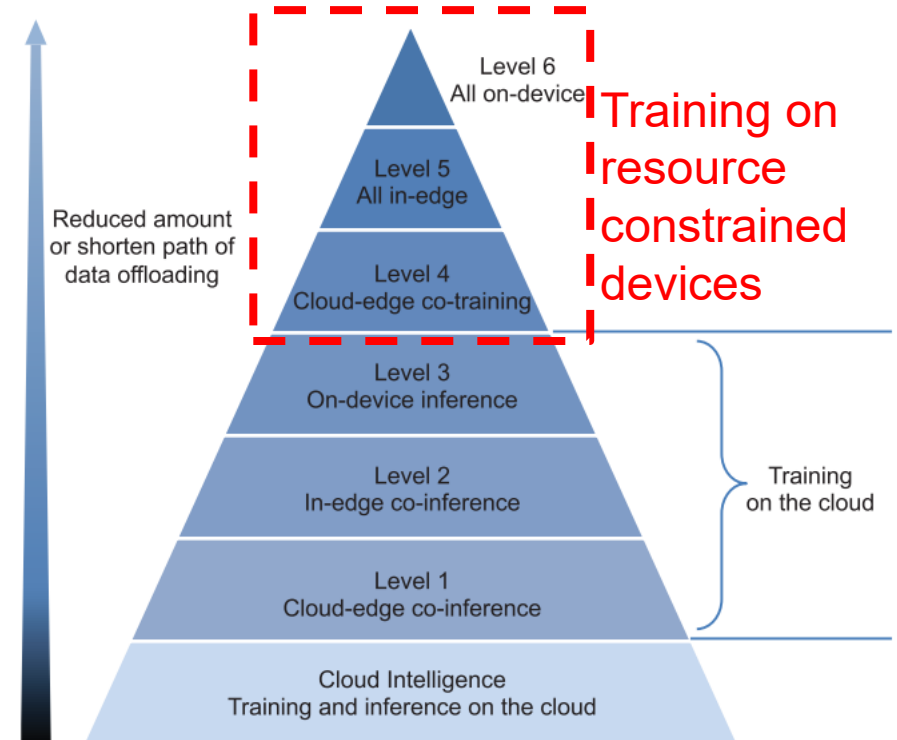
❑ Both learning and inference can be done at different levels

❑ Several SOTA work adapt ML algorithms to run on edge constrained devices (K-means, GMM, Random Forests)

- Reducing I/Os
- Alternating compute and I/O phases

[1] Mohammad Yahya Akhlaqi, Zurina Binti Mohd Hanapi, Task offloading paradigm in mobile edge computing-current issues, adopted approaches, and future directions, Journal of Network and Computer Applications, Volume 212, 2023, 103568, ISSN 1084-8045,
[2] Asif Mahmud Raivi, Sangman Moh, A comprehensive survey on data aggregation techniques in UAV-enabled Internet of things, Computer Science Review, Volume 50, 2023, 100599, ISSN 1574-0137,

Focus of this work



Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," in *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738-1762, Aug. 2019,

Outline

1.CONTEXT

2.MOTIVATION

3.I/O COOL

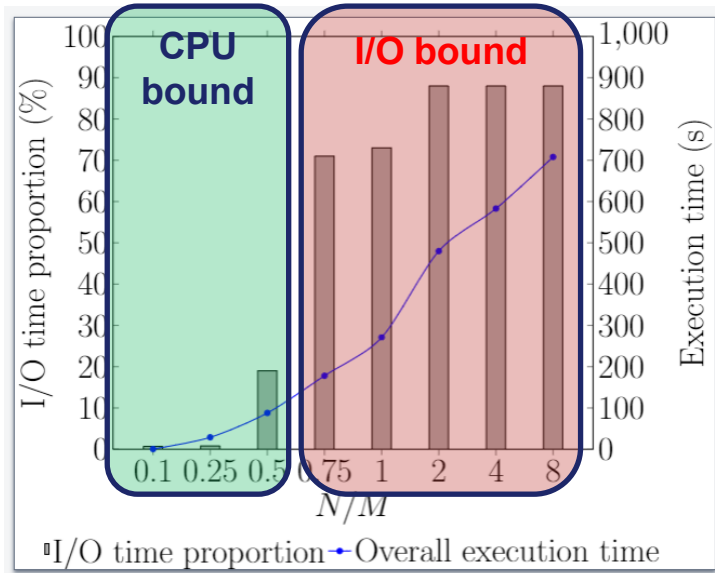
4.EVALUATION

5.CONCLUSION

Motivation – Training at the edge (you can't help yourself from swapping !)

□ AI training on memory-limited devices

- Spanning the dataset several times
- I/O swapping issues



Algorithm 1 General solution pattern

```
1: Subdivide the dataset into equal chunks that fit in the memory
2: for each chunk  $i$  of the dataset do
3:   Read chunk data                                ▶ I/O phase
4:   while stop criteria is not reached do           ▶ Compute phase
5:     Perform model related processing.
6:     Update the sub-model
7:   end while
8:   Store submodel  $m_i$ 
9: end for
10: Agregate all models  $m_i$ 
```

J. Boukhobza, et. al. 2025. Design and Optimization for AI/ML Acceleration on Resource-constrained and Edge Systems. In Proceedings of the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES '25).

Motivation – how about temperature

❑ **Edge devices are rarely equipped with cooling systems → passive cooling/heat sinks**

❑ **Training/compute phases provoke a temperature rise**

- (uncontrollable) CPU throttling (training time increase)
- sudden power off
- faster wear out of hw

❑ **SOTA solutions mainly work on**

- Frequency scaling (DVFS) → reduce frequency to cool down
- Load balancing → balance thermal load accross cores (assumes the presence of cold cores)



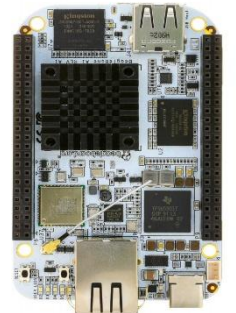
Raspberry Pi



Google Coral dev



Arduino Nano 33 BLE
Sense



BeagleBone AI



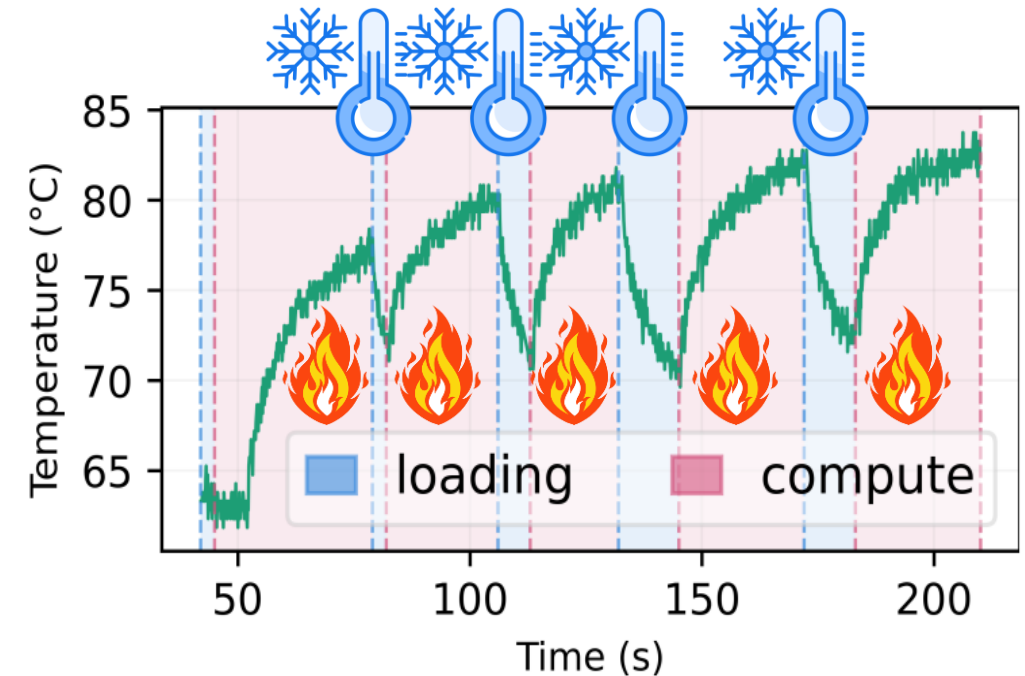
NVIDIA Jetson



BeagleBone Black

Motivation – I/Os, a natural cooling system

- ❑ Alternating compute (heat up) and I/O phase loading (cool down)
- ❑ CPU activity drops sharply during I/O phases
- ❑ Temperature follows a logarithmic increase in compute phases
- ❑ I/Os act like a natural cooling windows



Temperature change during the training of GMM on Rasp4b

How to use I/Os as a temperature regulator between compute phases?

Outline

1.CONTEXT

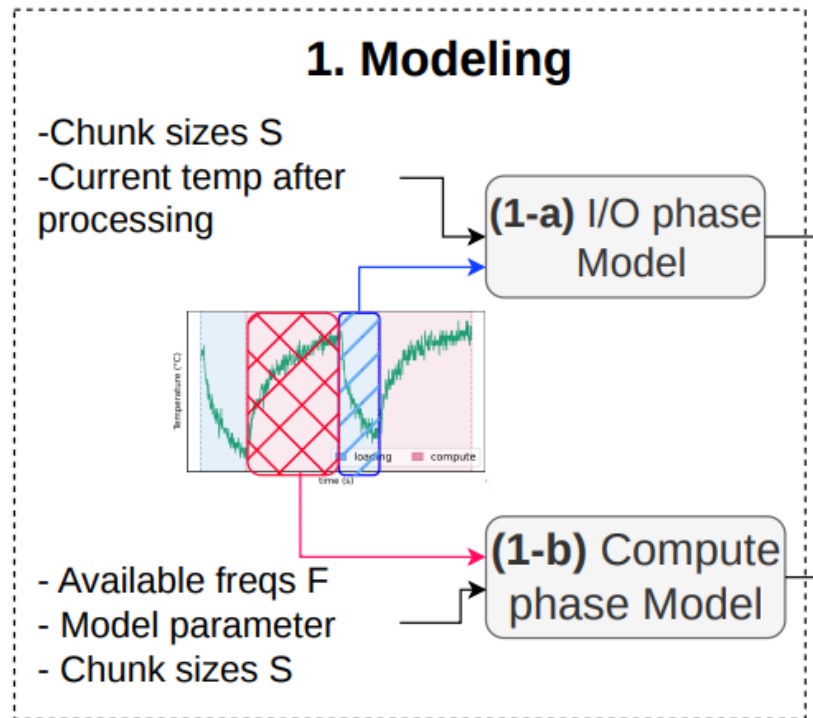
2.MOTIVATION

3.**I/O COOL**

4.EVALUATION

5.CONCLUSION

I/O-Cool: Using I/Os for Temperature Regulation When Training on the Edge



I/O-Cool – Modeling phase

□ I/O phase temperature model

- Logarithmic decay: Lightweight online logarithmic regression model.

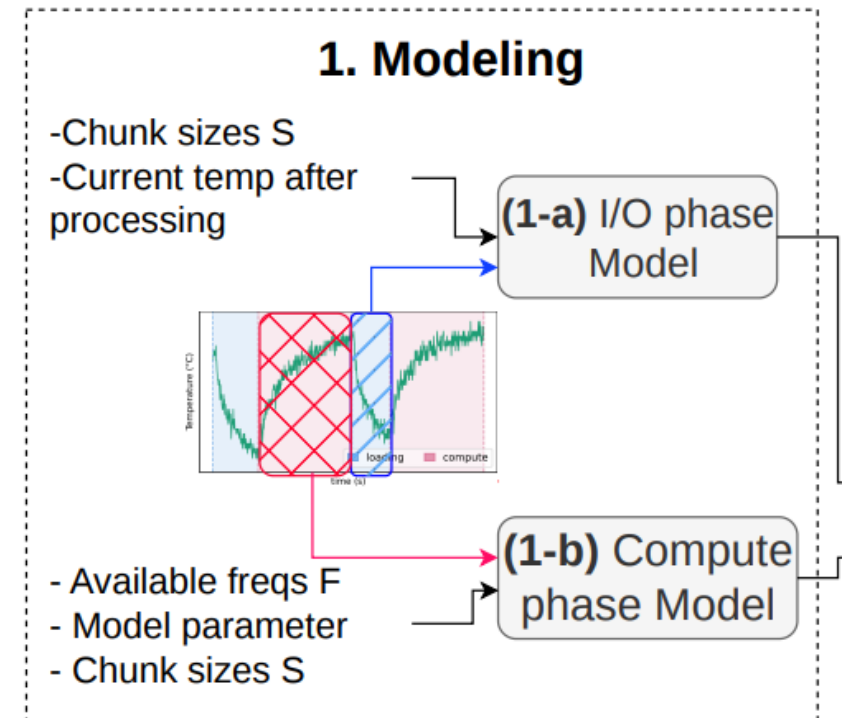
$$T_{I/O}^l = T_{curr} - \alpha \cdot \log(t^l + 1)$$

□ Compute phase temperature model

- multiple linear regression mode

$$T_{proc}^l = \beta_0 + \beta_1 \cdot s + \beta_2 \cdot F^l + \beta_3 \cdot T_{I/O}^l + \beta_4 \cdot k$$

- Offline training once per platform



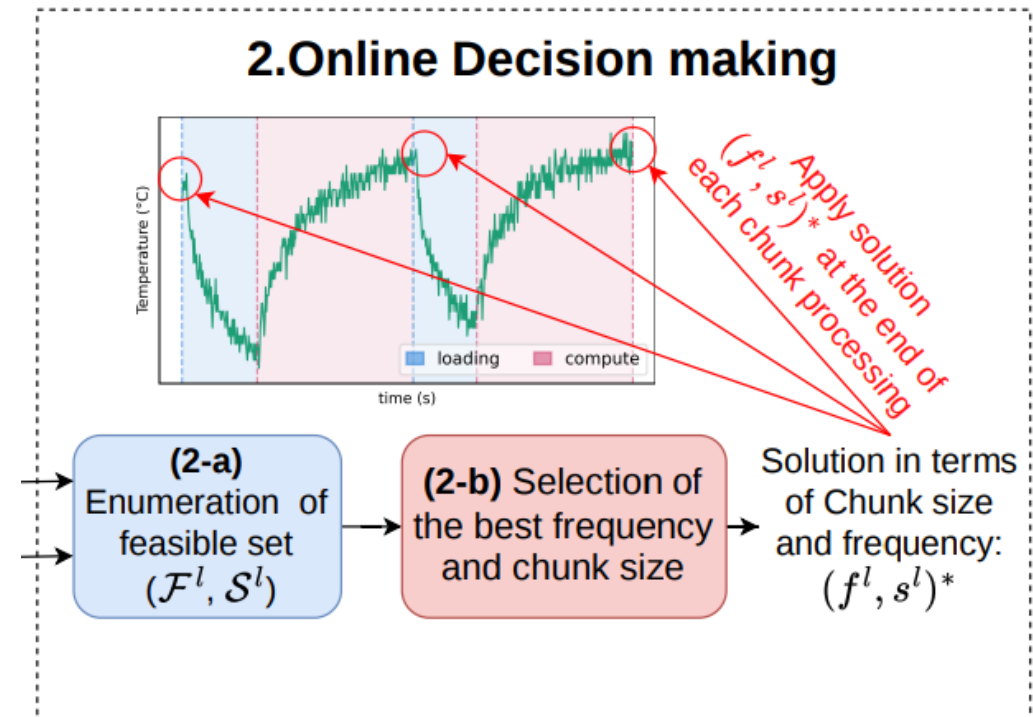
I/O-Cool – Online decision making

❑ Choose the right I/O size and CPU frequency for a given chunk

- → at the end of a chunk processing

❑ Enumerate all candidate pairs (f, s) that satisfy temperature limit

- → use of previous compute and I/O models
- $1\text{MB} < s < \text{memory workspace available}$
- Estimate the cooling for a given size (I/O model) then the heat for processing that same size (Compute model)
- Draw a Pareto front with 2 objectives to maximize
 - » CPU frequency (priority)
 - » Chunk size



Outline

1.CONTEXT

2.MOTIVATION

3.I/O COOL

4.EVALUATION

5.CONCLUSION

Evaluation – Experiment 1

I/O-Cool vs SOTA

❑ Goal : compare I/O-Cool with SOTA work

- zTT[1], CamBoost[2], and original[3] version
- ML models: GMM and K-means

❑ Temperature limit : 80°C (given in the specs)

❑ Evaluation metrics:

- Temperature Limit satisfaction rate: T_{sat} , avg and max temperature
- Execution time

❑ Dataset : 200 MB with 10 clusters, dim=30, and $sepVal = -0.6$, $[N/M] = 10$



- [1]S, Kim, K. Bin, S. Ha, K. Lee, and S. Chong. 2021. zTT: learning-based DVFS with zero thermal throttling for mobile devices. In Proceedings of the 19th MobiSys '21. ACM, 41–53.
- [2]S. H Choi, J. Kong, and S. W. Chung. 2024. ComBoost: An Instruction Complexity Aware DTM Technique for Edge Devices. In Proceedings of the 29th ACM/IEEE ISPLED. ACM, 1–6
- [3] M. Bouzouad, Y. Benhamadi, C. Slimani, and J. Boukhobza. 2024. Pigmmalion: a partial incremental gaussian mixture model with a low i/o design. In Proceedings of the 39th ACM/SIGAPP SAC. 428–435

Evaluation – Experiment 1

I/O-Cool vs SOTA

❑ **I/O-Cool ensures the best thermal satisfaction rate**

- 45% to 60% better than SOTA

❑ **Execution time: better or competitive with zTT, and outperforms ComBoost**

K-means training

Scenario	Time (s)	T _{Sat} %	Temp _{Avg.} (°C)	Temp _{Max} (°C)
I/O-Cool	266.45 ± 8.5	100%	75.50 ± 1.5	79.00 ± 0.7
Original	238.00 ± 6.2	12.00%	81.30 ± 0.2	85.05 ± 0.2
zTT	278.00 ± 30.0	55.00%	78.10 ± 2.1	82.50 ± 0.3
ComBoost	295.00 ± 25.0	65.00%	77.00 ± 1.2	80.50 ± 0.5

GMM training

Scenario	Time (s)	T _{Sat} %	Temp _{Avg.} (°C)	Temp _{Max} (°C)
IO-Cool	316.83 ± 11.60	90.58%	77.38 ± 1.25	82.17 ± 1.63
original	280.43 ± 3.53	19.05%	81.75 ± 0.20	85.09 ± 0.21
zTT	299.94 ± 5.70	30.80%	80.42 ± 0.23	84.48 ± 0.24
Comboost	358.85 ± 25.21	45.56%	79.15 ± 0.80	82.53 ± 0.24

Evaluation – Experiment 2

I/O-Cool tradeoffs

❑ **Goal : Adaptability of I/O-Cool to temperature limits**

❑ **Temperature limits: 80, 83, 85°C**

❑ **Evaluation metrics:**

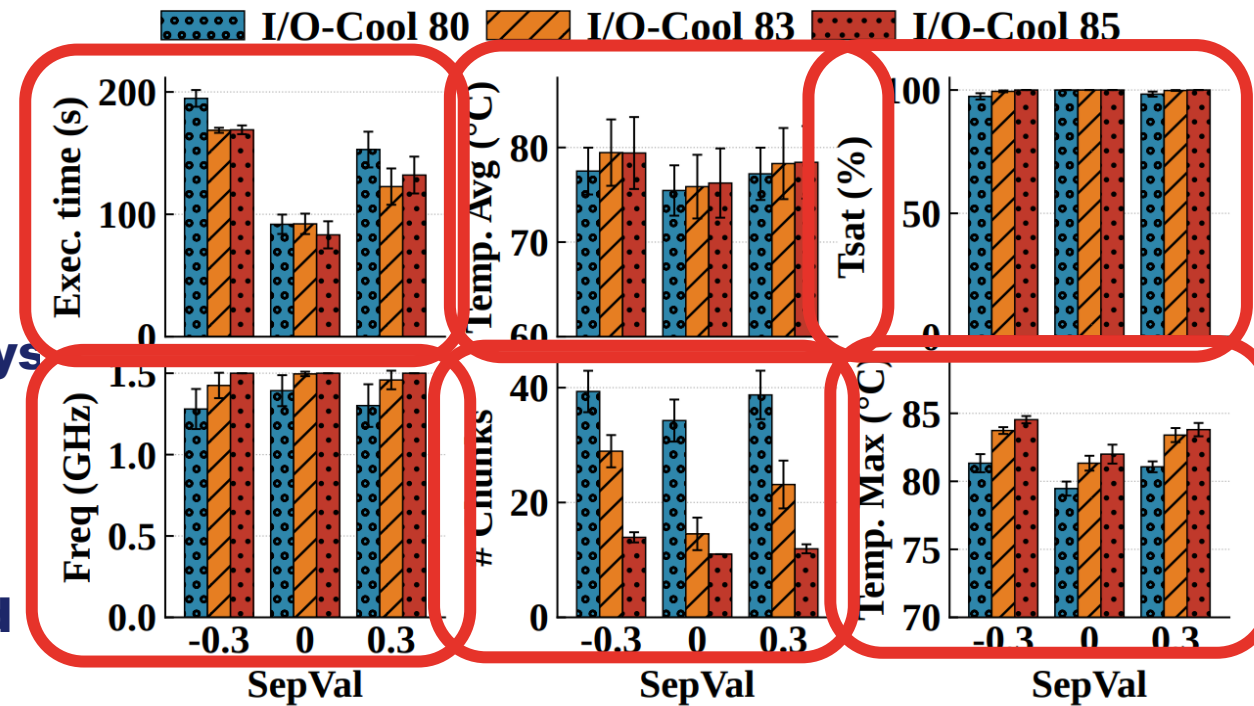
- Temperature Limit satisfaction rate: T_{sat} , avg and max temperature
- Execution time

❑ **Dataset : 100 MB synthetic dataset with $K=10$ clusters, $dim=20$, under a memory constraint $[N/M] = 10$ with $sepVal=\{-0.3,0,0.3\}$**

Evaluation – Experiment 2

I/O-Cool tradeoffs

- ❑ Frequency and chunking are both useful for controlling temperature raise, dual lever
- ❑ The satisfaction rate is always high regardless of the dataset/temperature limit
- ❑ Higher thermal limits require fewer frequency reductions and adjust chunking as primary mechanism.
- ❑ IO-Cool does not affect clustering quality



ARI for GMM
(the higher the better)

SepVal	I/O-Cool80	I/O-Cool83	I/O-Cool85	Original
-0.3	0.572	0.571	0.572	0.569
0	0.898	0.896	0.897	0.898
0.3	0.998	0.996	0.995	0.993

Outline

1.CONTEXT

2.MOTIVATION

3.I/O COOL

4.EVALUATION

5.CONCLUSION

Conclusion and perspectives

❑ I/O-COOL → I/Os as an additionnal lever for temperature regulation for I/O-then-compute patterns

- Evaluated on 2 ML training algorithms: K-Means and GMM and 2 platforms (only one reported in the paper)
- 90-100% temperature limit satisfaction

❑ Perspectives:

- Generalize to
 - » other ML models
 - » other data-intensive applications
 - » other boards
- Better model initialization
- Online (light) I/O and compute temperature model learning