Energy Efficient Any-Time I/O Adaptive K-means

Hafsa Kara Achira, Camélia Slimani, Mouloud Koudil, Jalil Boukhobza

Energy Efficient **Any-Time I/O Adaptive K-means**

Hafsa KARA ACHIRA[1,2], Camélia SLIMANI[1],
Mouloud KOUDIL[2,3], Jalil BOUKHOBZA[1],

[1] ENSTA Bretagne, Lab-STICC, CNRS, UMR 6285, F-29200 Brest, France
[2] Ecole nationale Supérieure d'Informatique (ESI-Alger)
[3] Laboratoire des Méthodes de Conception des Systèmes (LMCS), Alger, Algérie

1- Context: Deadline-constrained ML limits in Edge devices

Edge Intelligence (EI) : paradigm that brings intelligent applications closer to data collection devices.

Edge Intelligence constraints :

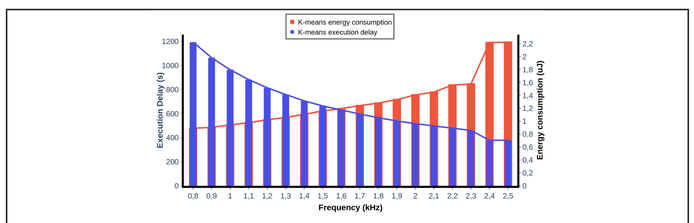
- **Memory:** The current growth of data volume to process has surpassed the scaling capabilities of DRAM memory [1].
- **Time:** Intelligent applications are required to deliver timely responses as they process large amounts of data.
- **Energy:** ML models training are resource-hungry which drains rapidly the power budget.

Problem statement: How to design a memory-constrained deadline-aware K-means algorithm for embedded devices while reducing energy consumed.

K-means algorithm :

- K-means is widely used for clustering in EI.
- K-means execution time is affected by dataset size, memory workspace size, centers initialization, and the required number of iterations to converge.
- Predicting K-means delays for real-time is crucial for applications such as anomaly detection [2], or preprocessing data before AI models as CNN [3].

DVFS: Power management technique that adjusts online the voltage and frequency of a processor.



(1) Frequency Impact on delay and energy

2- Related Work

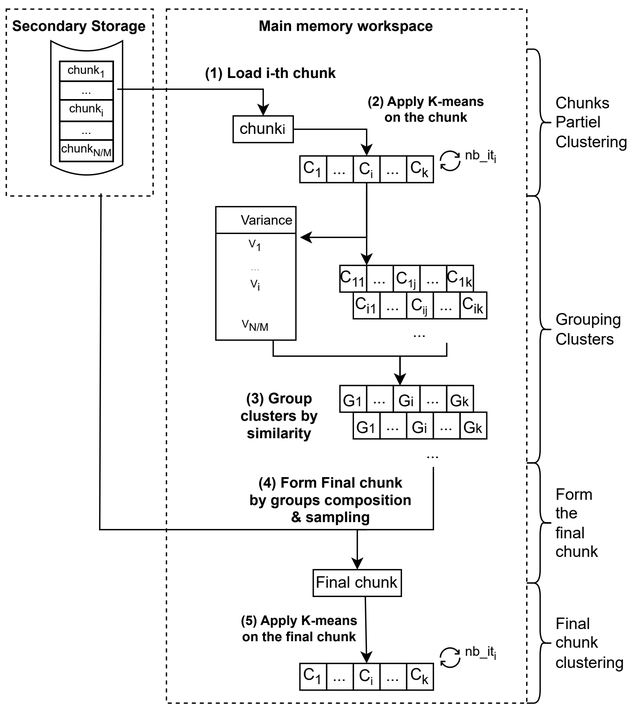
- [4] presents a stereo depth estimation with CNN on GPU that performs depth estimation in stages, allowing the model to provide ongoing estimates when queried.
- [5] proposes a fast inference framework that learns to speed up inference at run-time by combining a flexible sampling technique with deterministic message-passing to reduce computations in general regressors as random forests.
- [6] The authors propose adapting the minimum number of observations of a data stream before calculating the best attribute for node splitting to achieve precision within confidence intervals.

None of the previous studies adapt K-means to satisfy a given deadline on embedded devices.

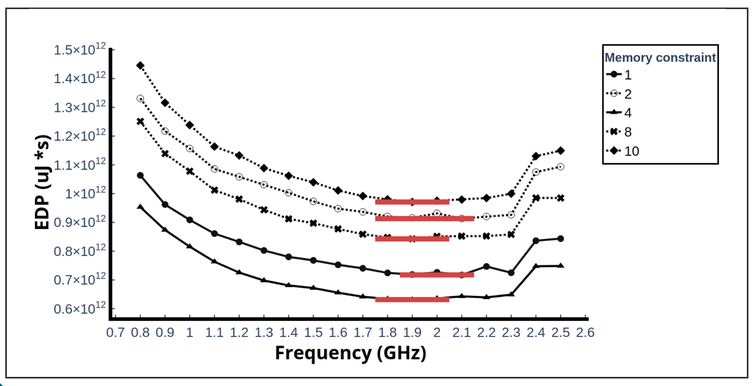
3- $K-$MLIO [7] Energy Analysis

- **K-MLIO (for K-Means with Low I/O Overhead)** is a state-of-the-art I/O optimization for the K-Means algorithm that employs a divide-and-conquer approach to mitigate I/O overhead.

N	M	K
Data-set size	Chunk size	Clusters number

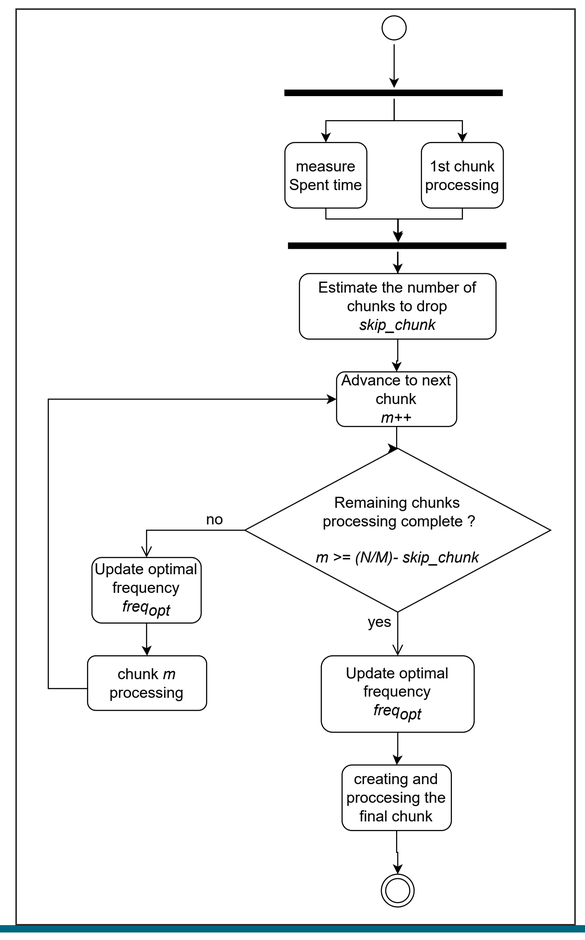


- K-MLIO [7] stabilizes the I/O proportion of the total execution time to less than 4% regardless of the memory constraint ($\frac{N}{M}$).
- The results of EDP analysis indicate that one can leverage frequency to achieve a favorable energy-delay trade-off.



4- Energy efficient dead-line aware $K-$MLIO

- **Principle :** We run the K-means per chunk (as in K-MLIO) on the data subset that makes it respond at the given deadline at the cost of a small clustering error through online data analysis.
- After measuring the processing time of the 1st chunk, we speculate on the number of chunks to drop *skip_chunk* to respect the deadline.
- Then, DVFS is applied before processing each chunk to update the optimal frequency $freq_{opt}$ required to execute the Worst-Case Execution Time $WCET(freq_{opt})$ of the remaining chunk.
- In case a slack time emerges when a chunk converges more quickly, the frequency $freq_{opt}$ is decreased while respecting the deadline.

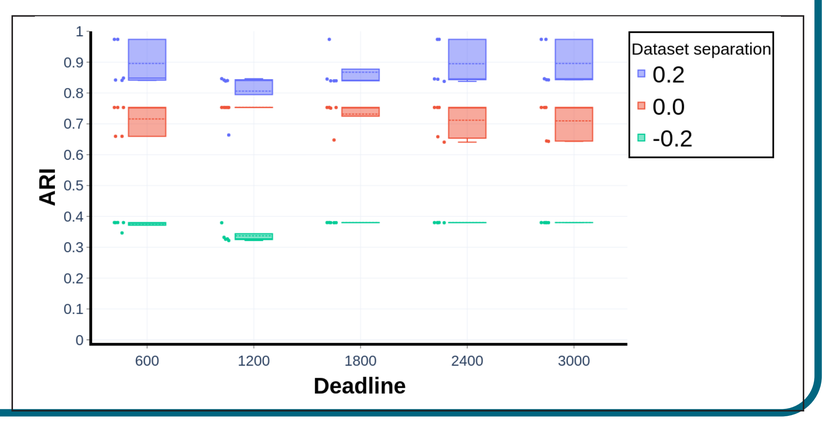


5- Evaluation

Objective : demonstrate that our solution respects the deadline while maintaining good precision (and reducing energy consumption).

Experiment Setup

Memory Work Space	300 Mo
$\lfloor N/M \rfloor$	10
Cluster Separations	-0.2, 0.0, 0.2
Time constraints	from 600s to 3000s



6- Conclusion and Future Work

- We proposed a memory-constrained deadline-aware K-means algorithm.
- The precision of the proposed solution is comparable with the original K-MLIO version **while the deadlines were always met**.

For future work:

- Evaluate energy consumption reduction on embedded devices.
- Evaluate on real data-sets and on embedded devices.

References

[1] Jeongdong Choe. Memory technology 2021: Trends challenges. In *2021 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pages 111–115, 2021.

[2] Y. Kitagawa and al. Anomaly prediction based on k-means clustering for memory-constrained embedded devices. In *2017 16th IEEE ICMLA*, pages 26–33, Dec 2017.

[3] Bing He, FengXiang Qiao, Weijun Chen, and Ying Wen. Fully convolution neural network combined with K-means clustering algorithm for image segmentation. In Xudong Jiang and Jenq-Neng Hwang, editors, *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, volume 10806, page 108062O. International Society for Optics and Photonics, SPIE, 2018.

[4] Yan Wang, Zihang Lai, Gao Huang, Brian Wang, Laurens van der Maaten, and Kilian Weinberger. Anytime stereo image depth estimation on mobile devices, 10 2018.

[5] S.M. Eslami, D. Tarlow, P. Kohli, and John Winn. Just-in-time learning for fast and flexible inference. *Advances in Neural Information Processing Systems*, 1:154–162, 01 2014.

[6] Eva Garcia-Martin, Albert Bifet, and Niklas Lavesson. Energy modeling of hoeffding tree ensembles. *Intelligent Data Analysis*, 25:81–104, 01 2021.

[7] Camelia Slimani, Stéphane Rubini, and Jalil Boukhobza. K -mlio: Enabling k -means for large data-sets and memory constrained embedded systems. In *IEEE MASCOTS*, pages 262–268, 10 2019.