UNIVERSITÉ SORBONNE PARIS NORD

Dimensionality Reduction Algorithms at the Edge

JRAF 2024 – Grenoble

Christophe Cérin

20 Novembre 2024

Objectifs

Constat : Déployer un service d'IA est de plus en plus populaire et facile à mettre en oeuvre.

Objectifs

Constat : Déployer un service d'IA est de plus en plus populaire et facile à mettre en oeuvre.

L'IA embarquée : une des solutions pour un numérique responsable ?

Objectifs

Constat : Déployer un service d'IA est de plus en plus populaire et facile à mettre en oeuvre.

L'IA embarquée : une des solutions pour un numérique responsable ?

Pour tenter de répondre à la question :

- Focus sur l'IA (apprentissage automatique) et IoT (IA embarquée) ;
- Mais pas l'IA des réseaux de neurones ;
- L'apprentissage automatique (machine learning) non supervisé ;
- Avec des applications dans le bâtiment intelligent ;
- Illustrations via la réduction de dimension et le clustering.

Réduire les impacts négatifs

Deux dimensions de réduction possibles :

- ***Efficience*** technologique qui rend les usages plus économes en ressources et moins polluant, sans les remettre en cause.
- ***Frugalité/sobriété :*** encadrement et réduction des usages (utilité), jusqu'à les remettre en cause…

Le monde réél des architectures

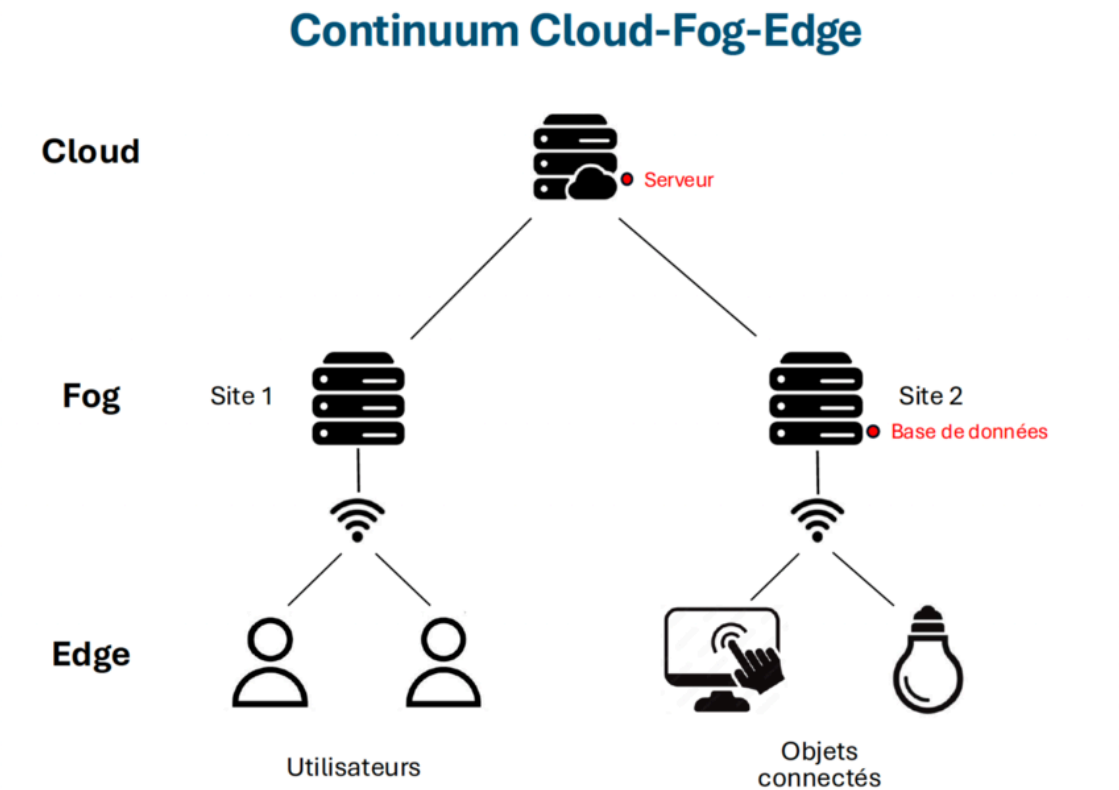


Figure 1: Edge = local computation ⇔ privacy

Tiers centre de données

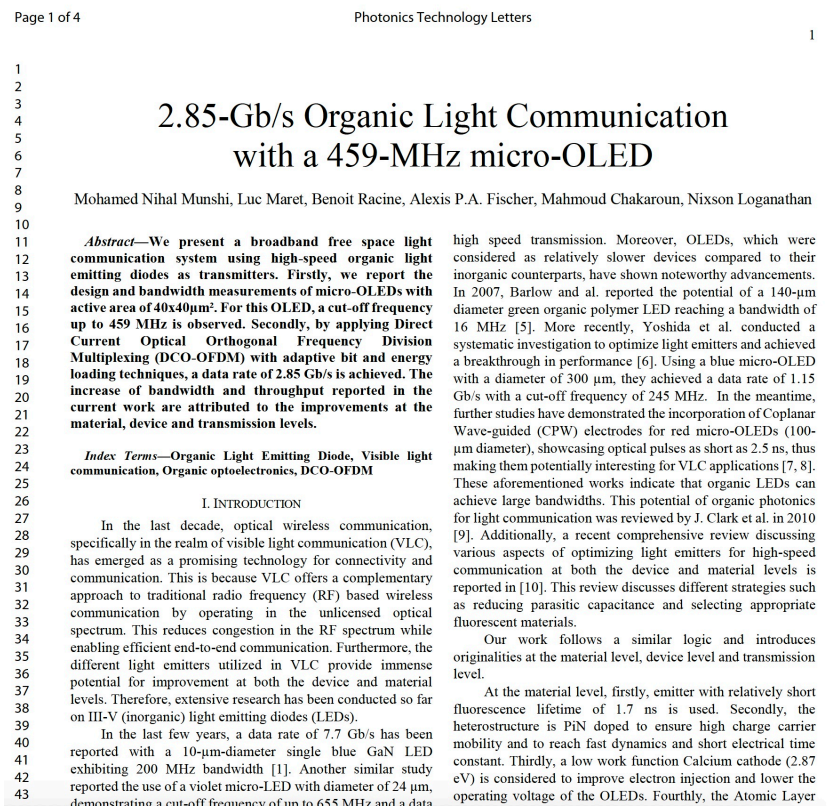
Échelle de performance (opérations flottantes / seconde)	Exemples de machine	Puissance (W et kW)	Puissance par mégaflops/s
Mégaflops/s = 10^6 flops/s	Intel 8086/8087, Motorola 68000 vers fin des années 1980	1,35 W à une température ambiante de 25 °C	1,35 W
Gigaflops/s = 10^9 flops/s	CRAY-2 (1985)	150-200 kW	200 W
Téraflops/s = 10^{12} flops/s	ASCI Red (1997)	850 kW	0,85 W
Pétaflops/s = 10^{15} flops/s	IBM Roadrunner (2008)	2 483 kW	0,002483 W
Exaflops/s = 10^{18} flops/s	Frontier (juin 2022)	22 703 kW	0,000022703 W

Fig. 2 – **Performances de calcul de systèmes HPC emblématiques et puissance électrique.** Même si la dernière colonne laisse à penser que les systèmes HPC, au fil du temps, sont de plus en plus efficients d'un point de vue de la puissance électrique par mégaflops, il faut rappeler que les 22 703 kW de puissance pour la machine Frontier représentent la consommation annuelle d'une ville française d'environ 90 000 habitants comme Versailles. La consommation énergétique en France est de 2 223 kWh par personne et par an. La valeur de 90 000 est obtenue en multipliant 22 703 kW par 24h par 365 jours et en divisant par 2 223 kWh. On suppose ici que la machine Frontier fonctionne 24h sur 24 et 365 jours par an ■

Nouveau classement du TOP500 de novembre 2024

The new El Capitan system at the Lawrence Livermore National Laboratory in California, U.S.A., has debuted as the most powerful system on the list with an HPL score of 1.742 EFlop/s.

Tiers réseau (local) : OLED pour communiquer



Page 1 of 4 Photonics Technology Letters 1

2.85-Gb/s Organic Light Communication with a 459-MHz micro-OLED

Mohamed Nihal Munshi, Luc Maret, Benoit Racine, Alexis P.A. Fischer, Mahmoud Chakaroun, Nixson Loganathan

***Abstract*—We present a broadband free space light communication system using high-speed organic light emitting diodes as transmitters. Firstly, we report the design and bandwidth measurements of micro-OLEDs with active area of 40x40µm². For this OLED, a cut-off frequency up to 459 MHz is observed. Secondly, by applying Direct Current Optical Orthogonal Frequency Division Multiplexing (DCO-OFDM) with adaptive bit and energy loading techniques, a data rate of 2.85 Gb/s is achieved. The increase of bandwidth and throughput reported in the current work are attributed to the improvements at the material, device and transmission levels.**

***Index Terms*—Organic Light Emitting Diode, Visible light communication, Organic optoelectronics, DCO-OFDM**

I. Introduction

In the last decade, optical wireless communication, specifically in the realm of visible light communication (VLC), has emerged as a promising technology for connectivity and communication. This is because VLC offers a complementary approach to traditional radio frequency (RF) based wireless communication by operating in the unlicensed optical spectrum. This reduces congestion in the RF spectrum while enabling efficient end-to-end communication. Furthermore, the different light emitters utilized in VLC provide immense potential for improvement at both the device and material levels. Therefore, extensive research has been conducted so far on III-V (inorganic) light emitting diodes (LEDs).

In the last few years, a data rate of 7.7 Gb/s has been reported with a 10-µm-diameter single blue GaN LED exhibiting 200 MHz bandwidth [1]. Another similar study reported the use of a violet micro-LED with diameter of 24 µm, demonstrating a cut-off frequency of up to 655 MHz and a data high speed transmission. Moreover, OLEDs, which were considered as relatively slower devices compared to their inorganic counterparts, have shown noteworthy advancements. In 2007, Barlow and al. reported the potential of a 140-µm diameter green organic polymer LED reaching a bandwidth of 16 MHz [5]. More recently, Yoshida et al. conducted a systematic investigation to optimize light emitters and achieved a breakthrough in performance [6]. Using a blue micro-OLED with a diameter of 300 µm, they achieved a data rate of 1.15 Gb/s with a cut-off frequency of 245 MHz. In the meantime, further studies have demonstrated the incorporation of Coplanar Wave-guided (CPW) electrodes for red micro-OLEDs (100-µm diameter), showcasing optical pulses as short as 2.5 ns, thus making them potentially interesting for VLC applications [7, 8]. These aforementioned works indicate that organic LEDs can achieve large bandwidths. This potential of organic photonics for light communication was reviewed by J. Clark et al. in 2010 [9]. Additionally, a recent comprehensive review discussing various aspects of optimizing light emitters for high-speed communication at both the device and material levels is reported in [10]. This review discusses different strategies such as reducing parasitic capacitance and selecting appropriate fluorescent materials.

Our work follows a similar logic and introduces originalities at the material level, device level and transmission level.

At the material level, firstly, emitter with relatively short fluorescence lifetime of 1.7 ns is used. Secondly, the heterostructure is PiN doped to ensure high charge carrier mobility and to reach fast dynamics and short electrical time constant. Thirdly, a low work function Calcium cathode (2.87 eV) is considered to improve electron injection and lower the operating voltage of the OLEDs. Fourthly, the Atomic Layer

Figure 3: Article CEA Grenoble - USPN

Transmissions optiques pour du LIFI à 2.8Gbit/s avec des semiconducteurs organiques ⇔ SC organique = moins d'énergie grise pour la fabrication que les SC traditionnels (pas besoin d'epitaxy) et faible consommation à hauteur du picoJoule par bit transmis au lieu du nJ/bit pour le wifi – Journée de lancement projet Européen HiSOPE : Jeudi 21 novembre –

Parenthèse enchantée 1 – A propos du modèle 3 tiers

Figure 4: https://github.com/christophe-cerin/Ecoindex-Revisited

Parenthèse enchantée – A propos d'Edge Computing

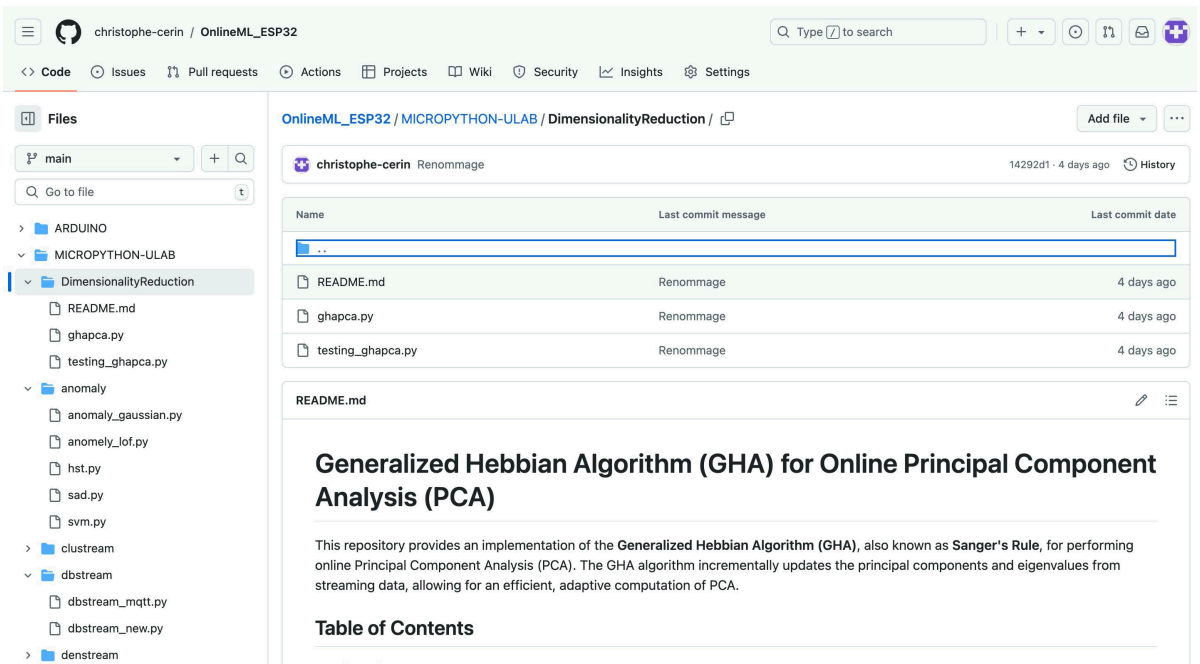


Figure 5: https://github.com/christophe-cerin/OnlineML_ESP32

– Désolé, mais ce projet n'est pas un projet de Deep Learning –

IA embarquée du projet Online ML on ESP32/ST32

Contexte & Objectifs

- Batiment (typiquement les capteurs de CO_2, température.. produisent une information toutes les 10 minutes) ;
- Les capteurs publient continuellement (MQTT ?) leurs données à un ESP32 qui implémente une intelligence
- Développer une librairie complète intégrant les briques Clustering, Anomaly Detection, Regression, Dimensionality Reduction, Time Series… (l'intelligence) et pour le cadre **Online** ;
- Pour ESP32/ST32 dans les éco-systèmes ARDUINO et MicroPython ;
- Avec une méthodologie d'étude permettant la reproductibilité (problématiques des datasets, de la mesure de la consommation, d'indicateurs qualitatifs et quantitatifs) ;

Methodology: Data Healing Pipeline

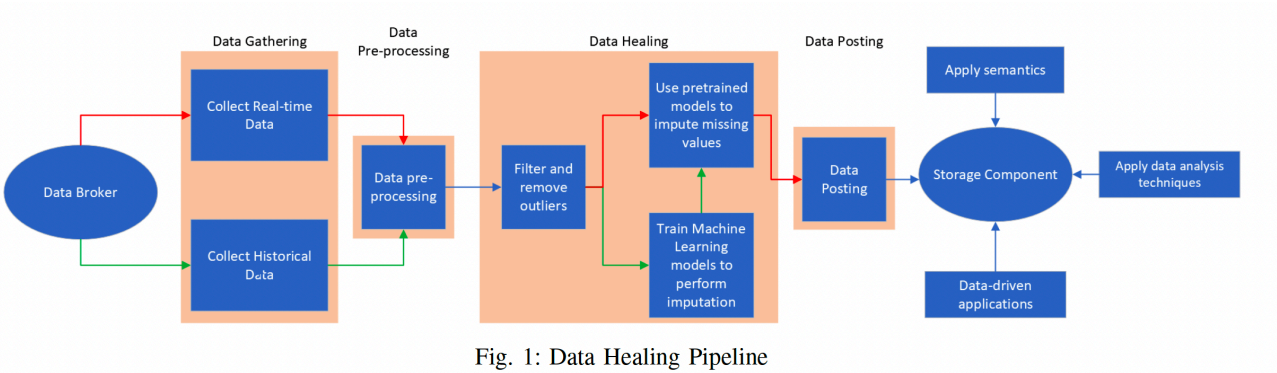


Fig. 1: Data Healing Pipeline

Figure 6: – SMART2B https://smart2bproject.eu/ project funded by the European Union's Horizon 2020 –

Key Performance Indicators (KPIs) examples: Total Number of Values Processed per Day, Global Number of Errors Filtered per Day, Outliers per Day, Duplicates per Day, Outliers per Device per Day.

Mainly, **qualitative** KPIs, and no need for a **powerful** computer and AI to produce such statistics on devices!

Methodology: Dimensionality Reduction (DR) case

- **Data Preparation:** a) convert categorical data to numerical data b) Prepare multiple datasets;
- **Dimensionality Reduction Algorithms:** choice of DR alg. (PCA, MDS, t-SNE, RP) for embedded systems;
- **Measurement Setup:** CodeCarbon library to compute the CO_2 emitted, and Time;
- **Execution and Data Collection**;
- **Performance Evaluation:** means of the energy and time consumed;
- **Evaluation of Preservation:** through Distance Preservation (compare pairwise distances between data points in the original and reduced spaces);
- **Analysis and Comparison:** trade-offs between energy efficiency, computation time, and algorithms' efficiency;
- **Generalization to Other Datasets:** Repeat steps 2 to 6 on different datasets to see if the results are consistent across various data types and sizes.

Détails sur l'algorithme en ligne de réduction de dimension

Generalized Hebbian Algorithm (GHA - https://en.wikipedia.org/wiki/Generalized_Hebbian_algorithm) for Online Principal Component Analysis (PCA)

- **Évite** la dépendance multicouche associée à l'algorithme de rétropropagation + **compromis** entre la vitesse d'apprentissage (paramètre γ) et la précision de la convergence.
- Principal Component Analysis (PCA) is a **dimensionality reduction** technique that finds the principal components of the data.
- i. ghapca_C: Updates the matrix Q (the eigenvectors) based on the current data point x, the projected data point y, and the learning rate γ.
- ii. ghapca: Performs the GHA update, calculating and updating the eigenvalues and eigenvectors of the data in an online fashion.

Passons au cas du clustering (1/2)

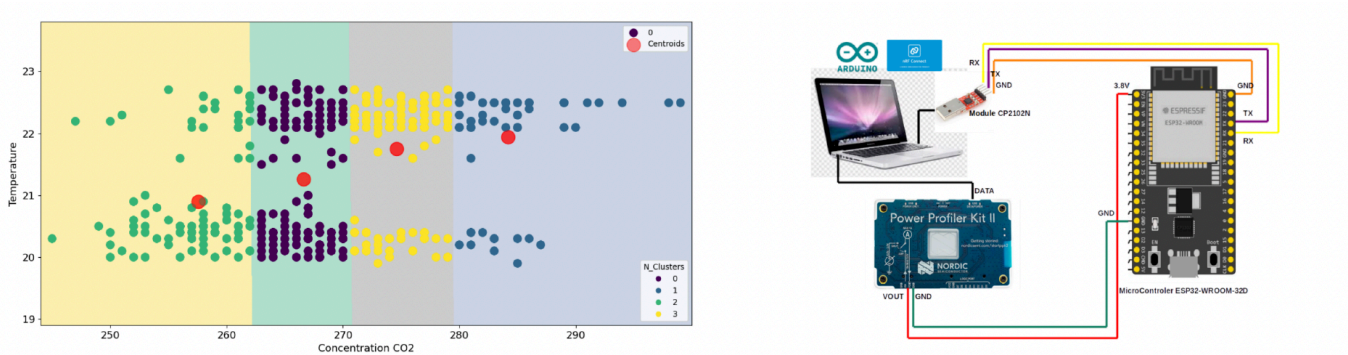


TABLE I
ELECTRICITY CONSUMPTION OF THE K-MEANS FOR A WINDOW SIZE OF w DATA AND k CLUSTERS

Sampling = 10 000 samples/s	Iteration #	Current Intensity average (mA)	Current Intensity max (mA)	Time second(s)	Charge Coulomb (C)
w=512, k=4, itmax=1, d=1024	1	101,85	442,22	27,65	2,82
	2	100,72	437,89	24,73	2,49
w=128, k=4, itmax=8, d=1024	1	99,39	434,98	49,08	4,88
	2	100,91	439,83	49,19	4,96
	3	98,03	444,97	50,67	4,97
w=128, k=2, itmax=8, d=1024	1	102,62	440,93	52,11	5,35
	2	100,68	444,91	51,09	5,14
	3	99,46	434,81	53,72	5,34

Figure 7: – Distance based vs. Density based –

Passons au cas du clustering (2/2)

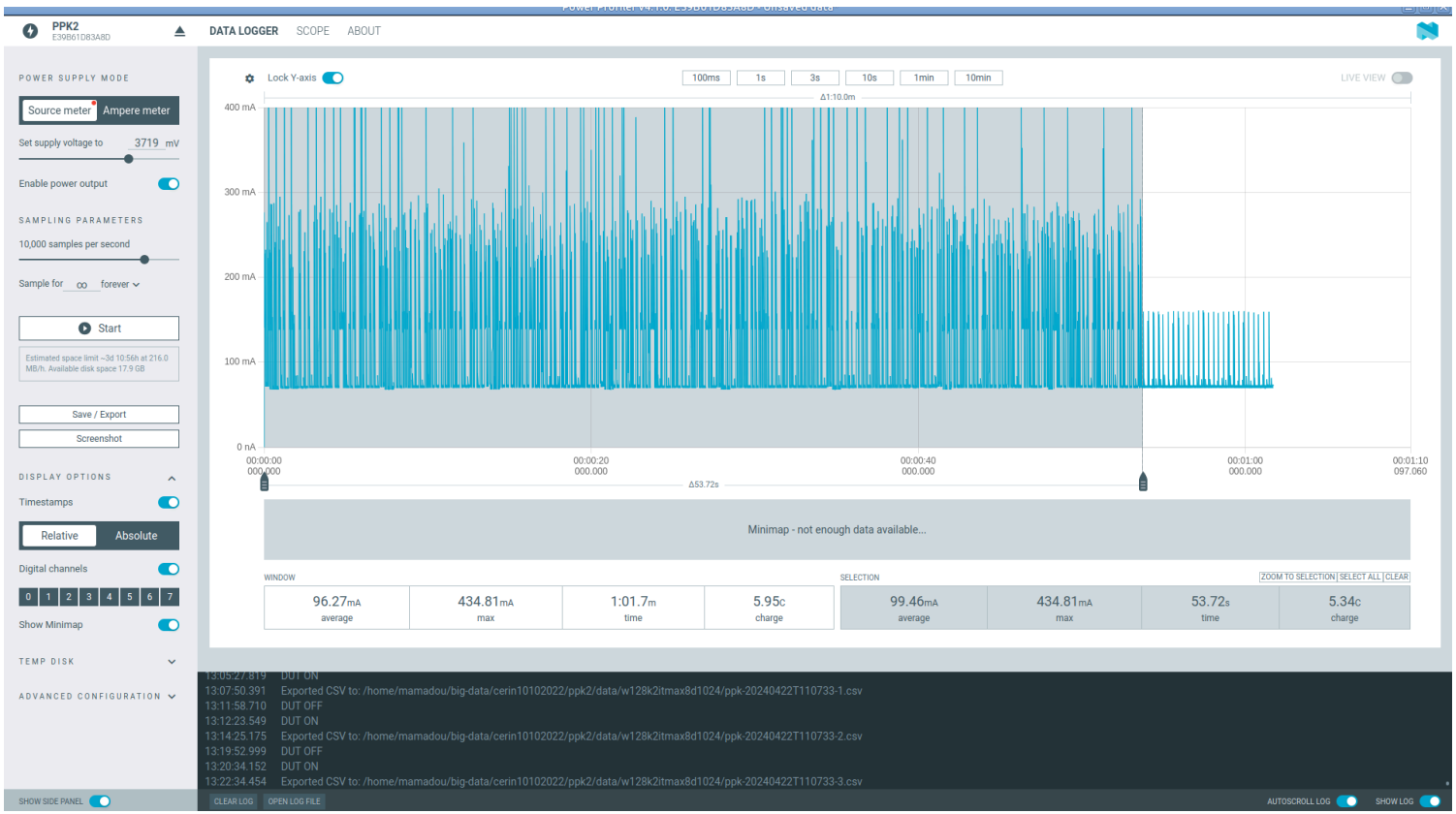


Figure 8: – Power measures and learning –

Conclusion (1/2)

La crise environnementale est avérée et ses conséquences sont sans précédent.

La croissance économique est la base de nos sociétés (occidentales) et le Numérique (IA) est un des principaux moteurs de la croissance :

- L'IA peut jouer un rôle positif partiel dans la crise, mais…
- … l'IA est un grand accélérateur dans un contexte où il nous faut réduire.
- Réduire quoi, et quand avec de l'IA embarquée ?

  Voir "Is TinyML Sustainable?, Communications of the ACM, 2023, doi 10.1145/3608473" et "Footprint of a microcontroller", https://www.st.com/content/st_com/en/about/sustainability/sustainable-technology.html

Conclusion (2/2)

Quelle place pour le edge computing ?

Un paradigme séduisant :

Conclusion (2/2)

Quelle place pour le edge computing ?

Un paradigme séduisant :

- Effectuer les traitements au plus proche des données et avec une meilleure efficacité énergétique.
- Il est clair que certains modèles ne peuvent être entrainés complètement sur de l'edge…
- mais l'IA n'est pas que les réseaux neuronaux !
- Comment éviter/minimiser les effets indirects et rebonds ?

Conclusion (2/2)

Quelle place pour le edge computing ?

Un paradigme séduisant :

- Effectuer les traitements au plus proche des données et avec une meilleure efficacité énergétique.
- Il est clair que certains modèles ne peuvent être entrainés complètement sur de l'edge…
- mais l'IA n'est pas que les réseaux neuronaux !
- Comment éviter/minimiser les effets indirects et rebonds ?

…en s'interrogeant sur la matérialité d'un service numérique et sur les usages.

I. Merci pour votre attention

Questions ?