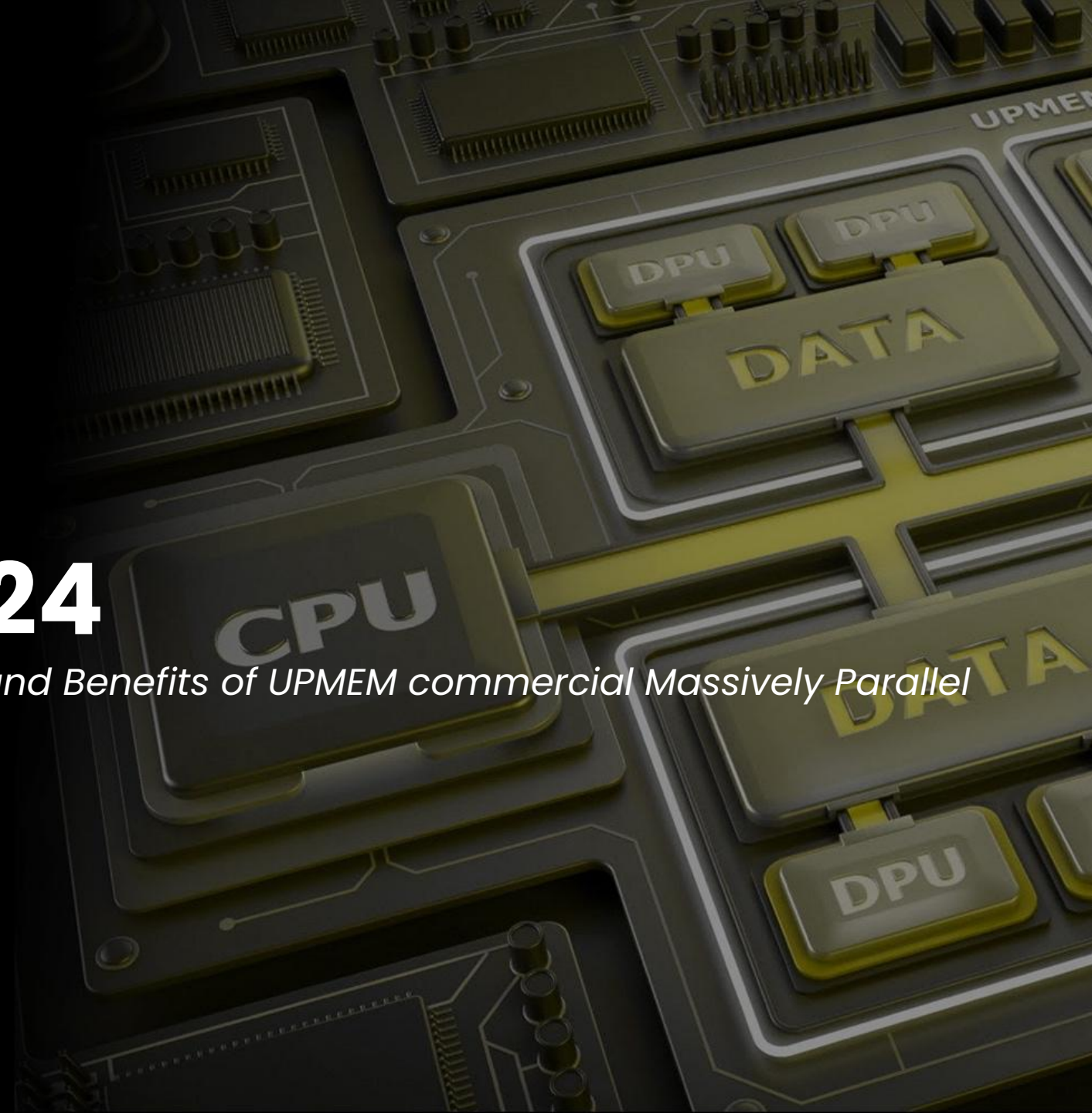


ABUMPIMP 2024

The 2nd Symposium on Applications and Benefits of UPMEM commercial Massively Parallel Processing-In-Memory Platform

August 26, 2024

EURO-PAR CONFERENCE 2024

Overcome data and energy bottleneck thanks to PIM

Founded: 2015

Employee Count: ~20

Headquarters: Grenoble, France

Total Patents: 11

Gilles Hamou

CEO / Co-Founder

Track Record:

Co-owner @ Oscaro.com
Scaled Oscaro.com to $100M revenue from inception
Founded & scaled Plantes-et-Jardins.com
Senior Manager @ RSM
Case Leader @ BCG

Education:

MBA INSEAD
Eng. Centrale Paris

Fabrice Devaux

CTO / Co-Founder

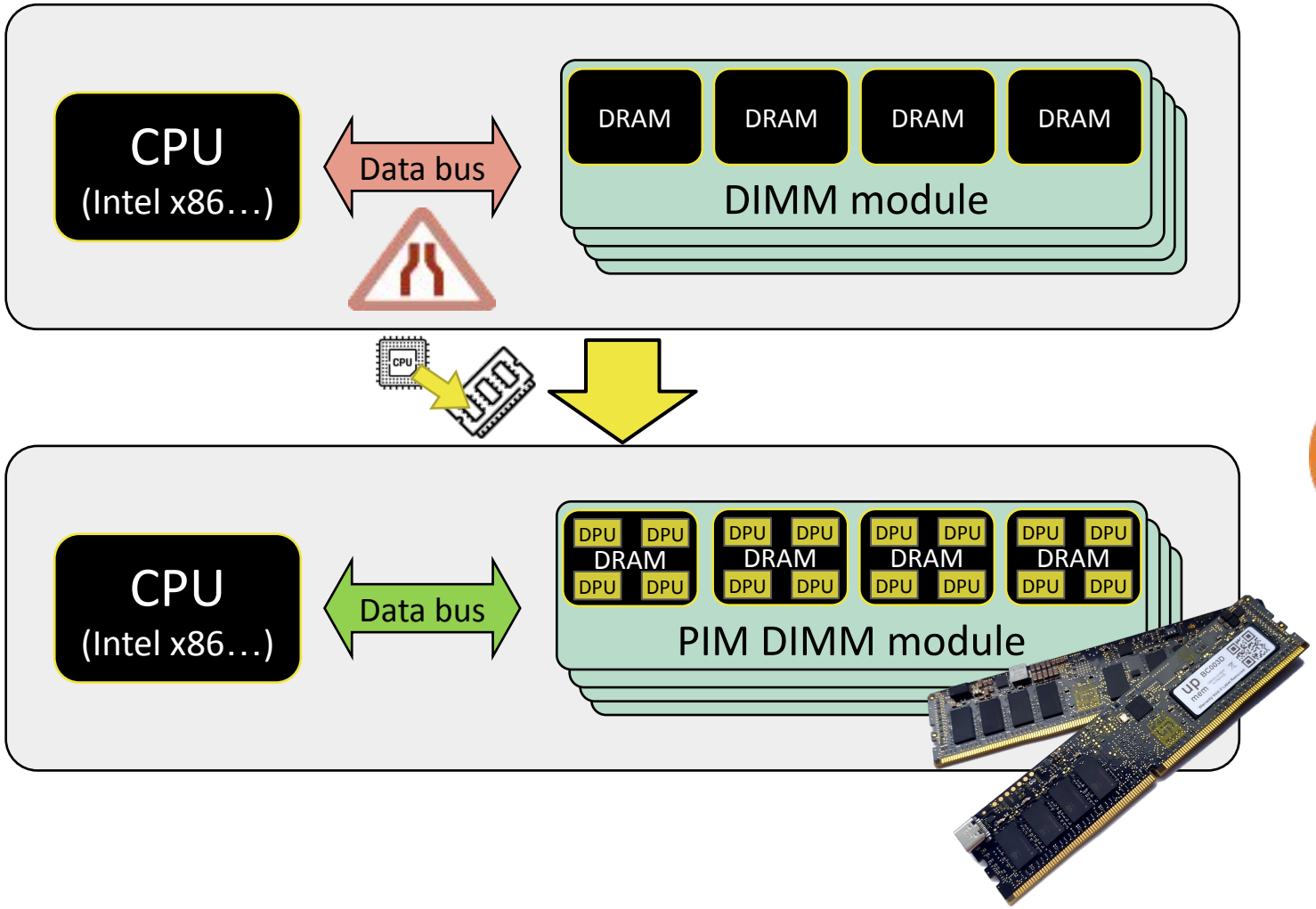
Track Record:

Senior Staff SWE @ VMWare
Co-owner, CTO @ Trango Virtual Processors, sold to VMware
CPU Architect @ STMicroeletronics

Education:

DEA, Microelectronics, Pierre and Marie Curie University

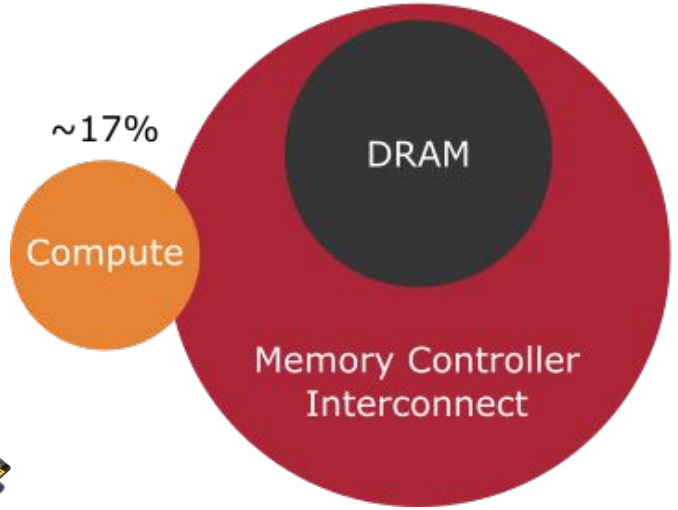
Overcome Limitations of Traditional Compute-Centric Architectures for Data-Intensive Workloads Thanks to Processing in Memory



Total system (server with compute node) energy consumption

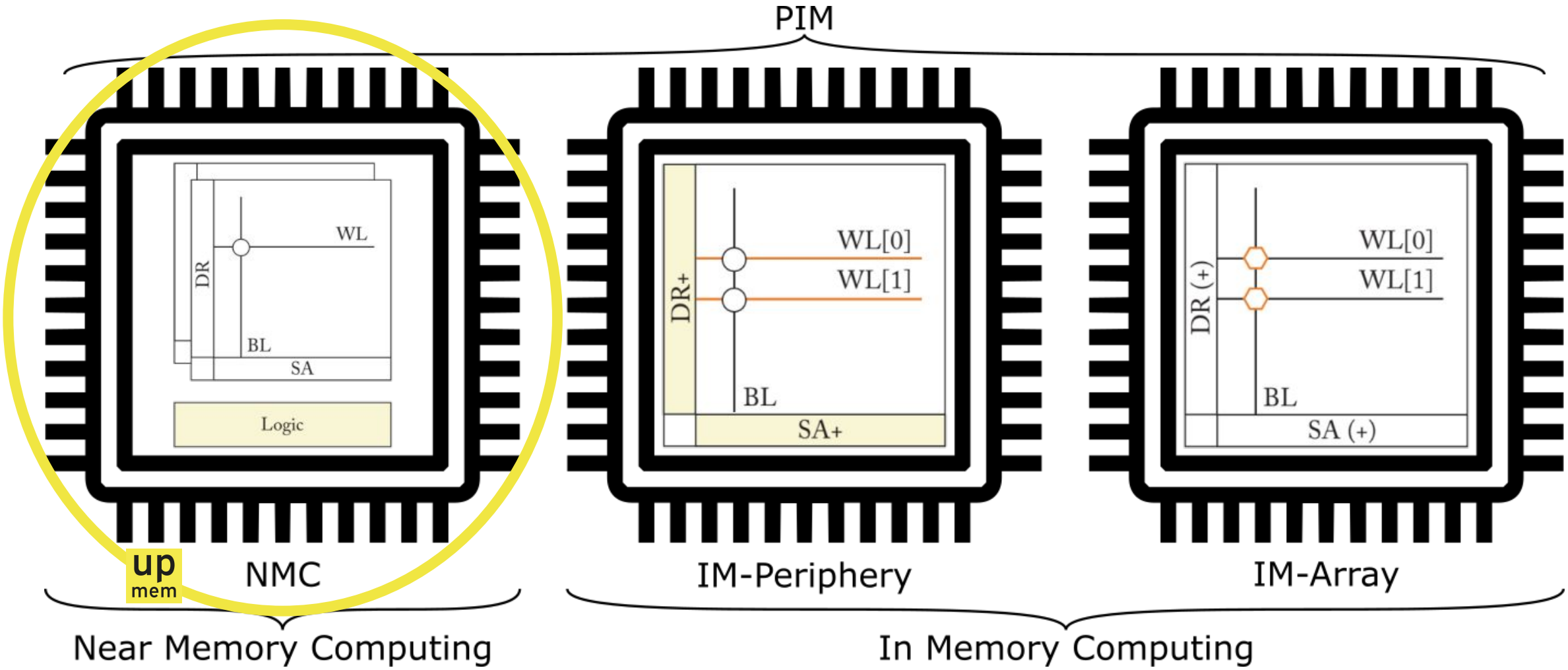
~63%

in memory and data transfers



Source: SK hynix CEO Seok-Hee Lee's keynote at the GSA Memory+ 2021 conference, confirming Lawrence Berkeley Lab results.

Taxonomy of processing in memory (PIM)



Source: Das, Reetuparna., Wang, Xiaowei., Fujiki, Daichi., Subramaniyan, Arun. In-/near-Memory Computing. United States: Morgan & Claypool Publishers, 2021.

Proven capacity to benefit a wide range of applications

Genomics

Databases

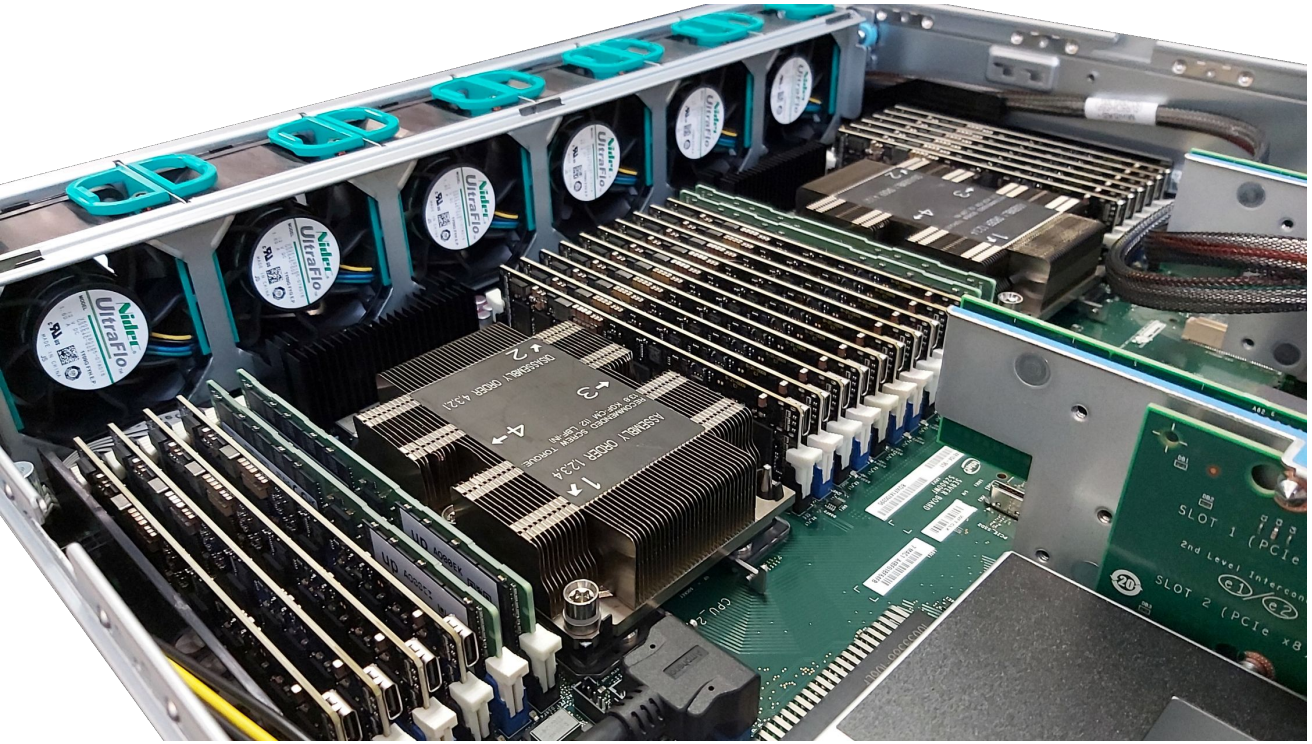
Data analytics

Image Processing

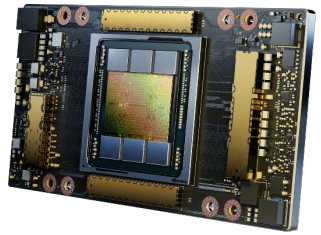
AI

Cryptography

Compression



VS

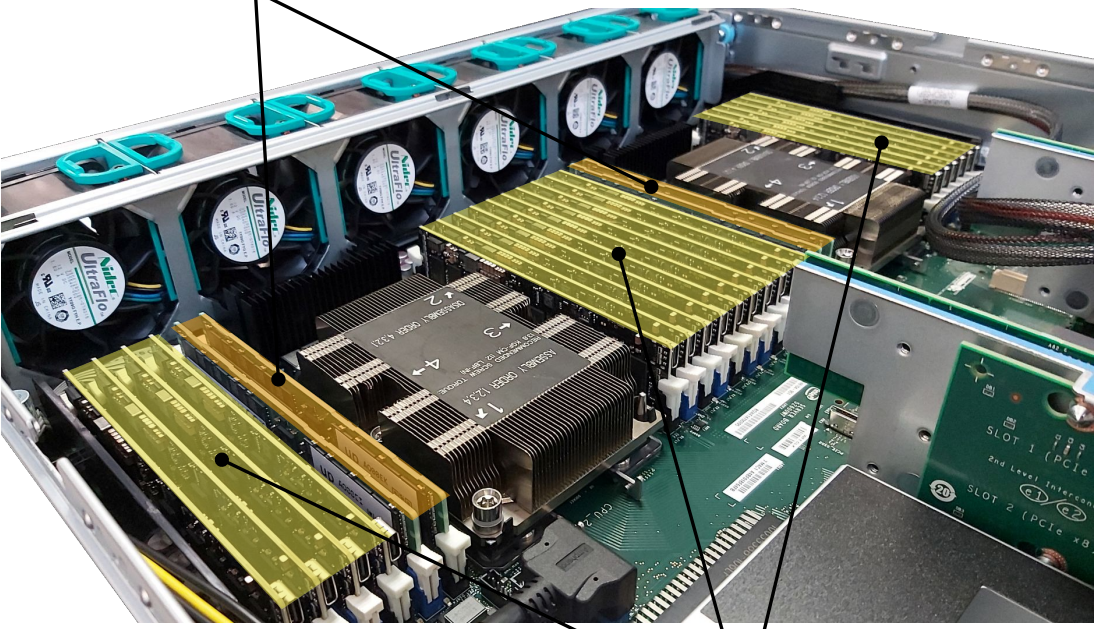


up mem

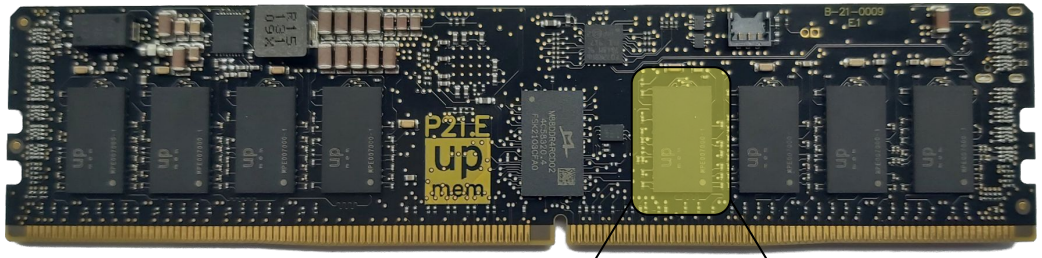
Technology Overview

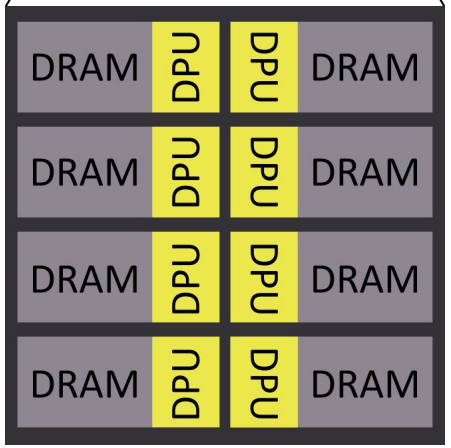
A standard application server populated with PIM DIMMs

4 Standard DIMMS



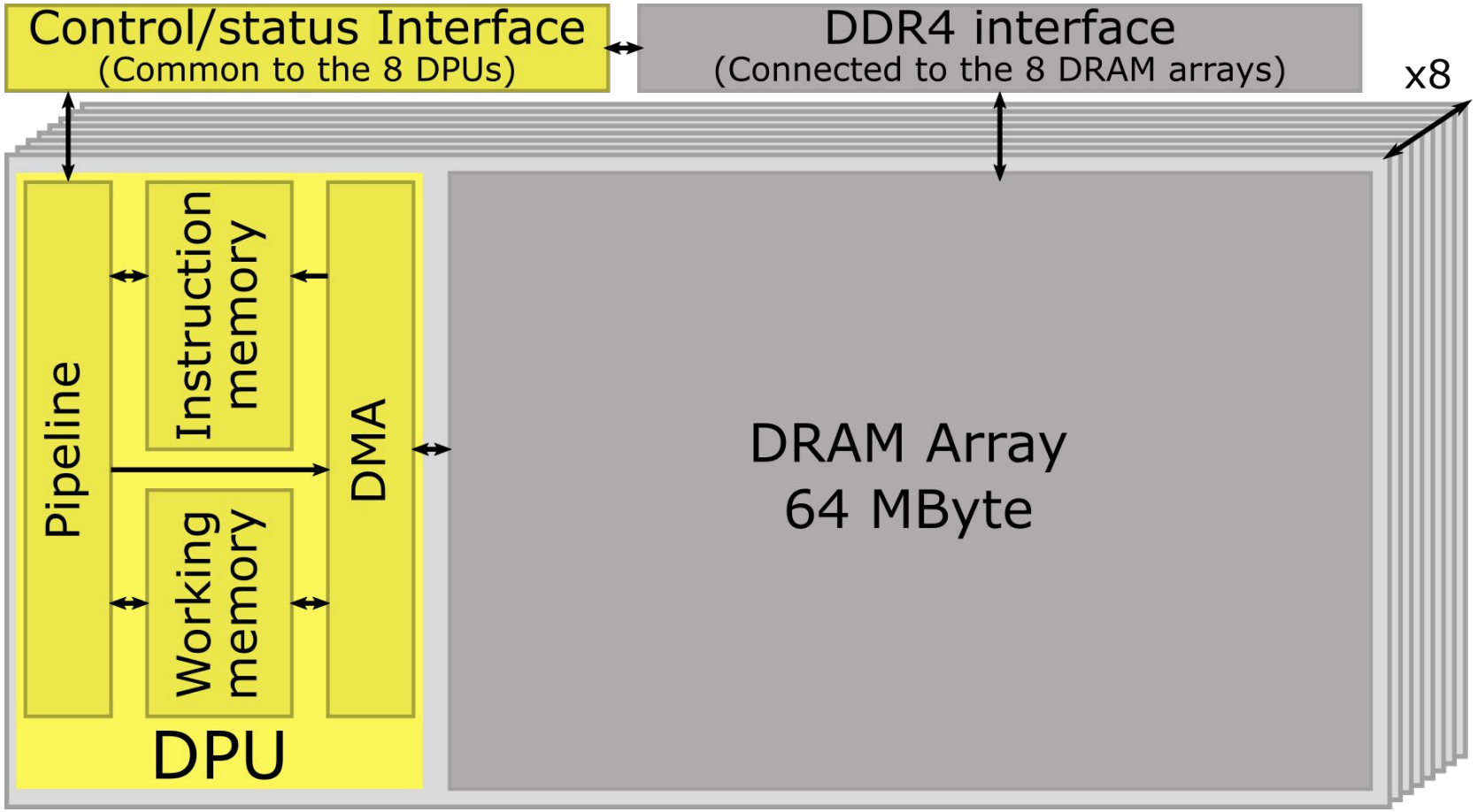
20 PIM DIMMS





2560 DPUs for 160GB of PIM DRAM

A DPU is a simple modern general-purpose processor



- Shared access with the host CPU to a DRAM bank
- Instruction and data caches replaced by instruction RAM and a Working RAM
- Independent and asynchronous
- 16 independent threads per DPU
- No direct communication channel among DPUs

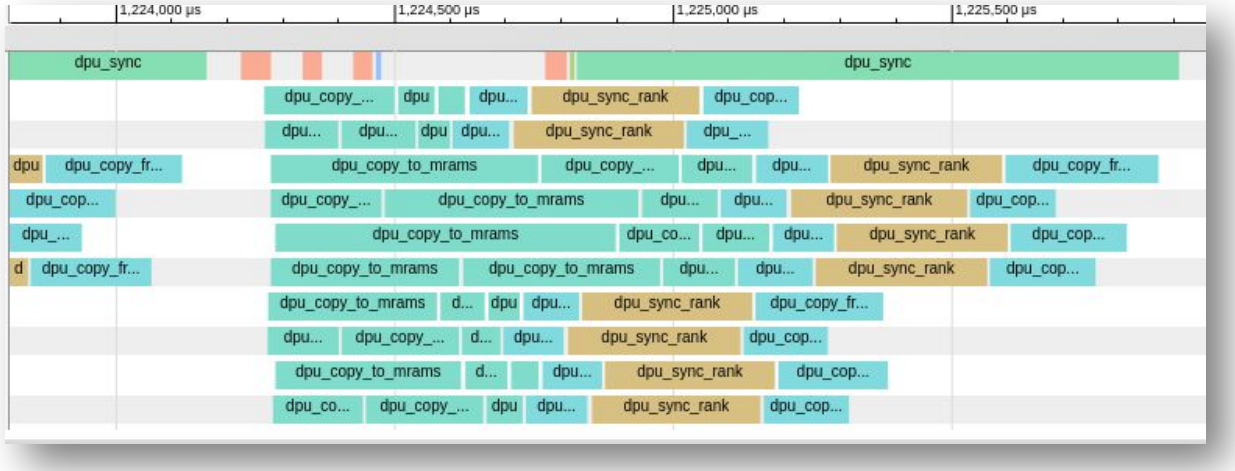
up mem

A set of tools for smooth application porting

x86 program written in C, C++ or python with C functions to call routines on the DPUs

UPMEM SDK contains:

- A Full-featured runtime library for the DPU
- Management and communication libraries to encapsulate easily all the Host to DPU operations
- An LLVM based C-compiler using LLVM 12.0
- A LLDB based debugger
- Programming tools: profilers, simulator…
- Server BIOS binaries
- Linux driver for x86 servers Validated on Redhat, Ubuntu and Debian.



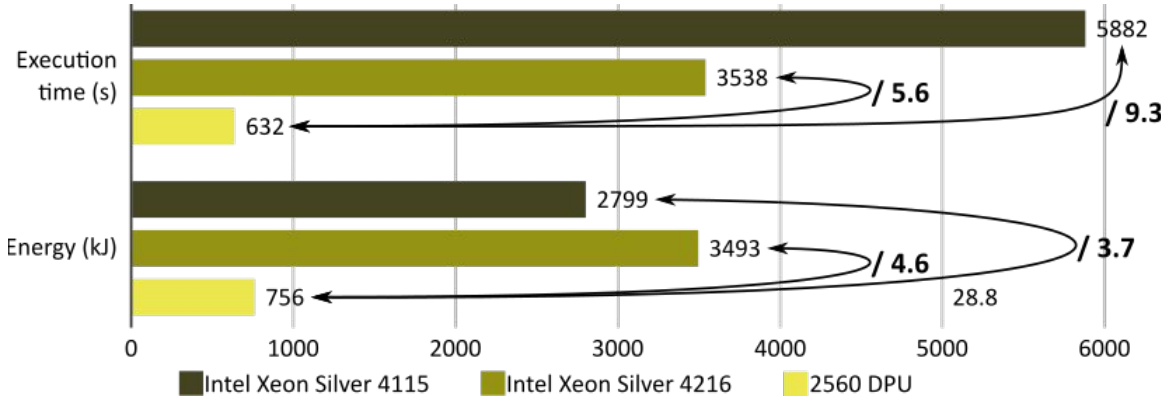
PIM Applications

Genomics

FASTQ file (read of 120 nucleotides):

@1
GGTTCACTGCAACCTCCGCCTCCCAGGTTCAAGCAATTATCCTGC
+
CCC0CGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGGGD
@2
AGTGAATTCATTCAGAAAGATGAATTTCCTGTGTAATCTCATTATG
+
CC3ACEGGGGGGGGGGGCGGG1GCGGGGGGGGGGGGGG>G

Long read alignment : adaptive KSW2

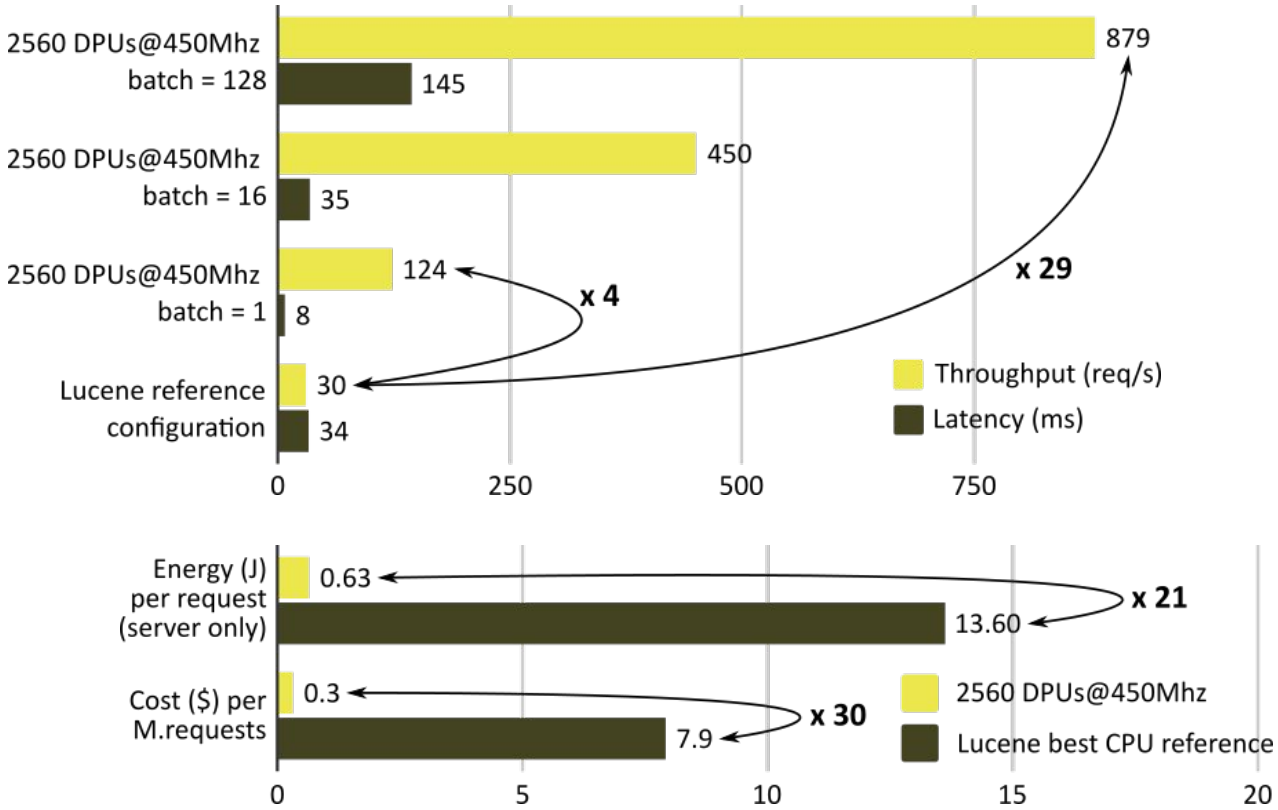


		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

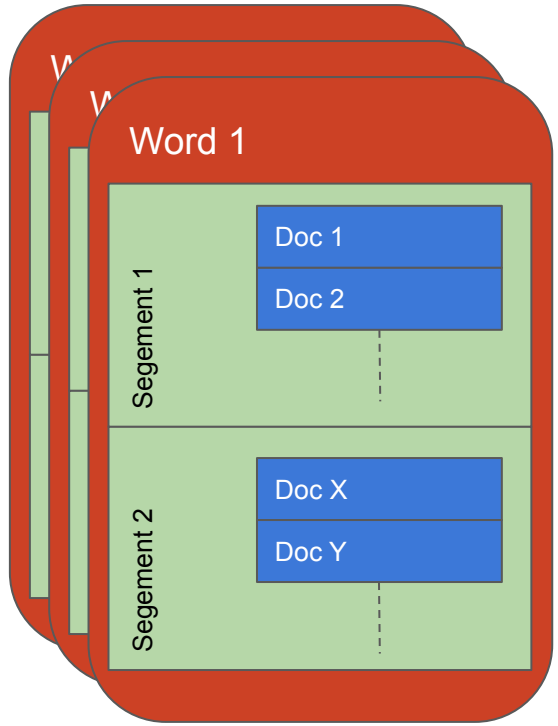
https://github.com/upmem/usecase_dpu_alignment

Analytics : Index Search

- An index search engine identifies items in a database from keywords specified by the user (web pages, text documents, e-commerce product…)
- **UPIS**: Engine for exact phrase match



Inverted index

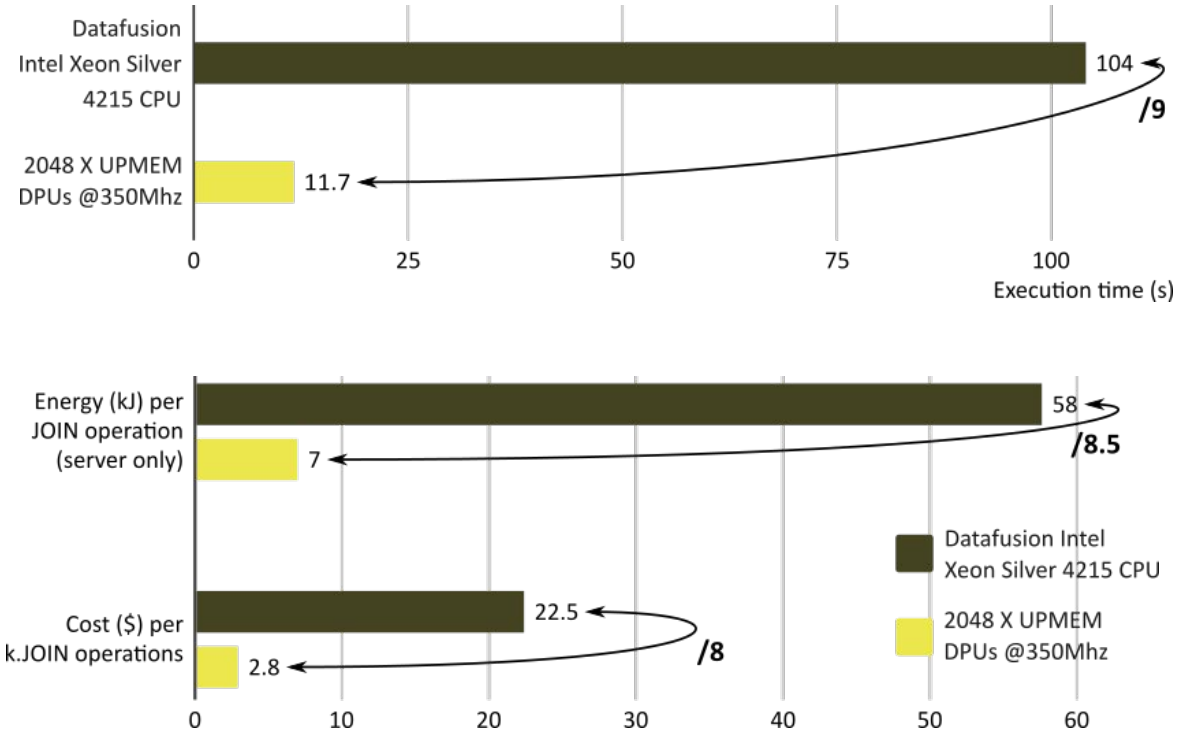


- **PIM Lucene**: extension of Apache Lucene

https://github.com/upmem/usecase_UPIS
https://github.com/upmem/pim-lucene

Analytics : Hash Join

- **Parallel hash-based join** on DPU
- 4G rows per table (32GB of random data)



Left Table

Date	CountryID	Units
1/1/2020	1	40
1/2/2020	1	25
1/3/2020	3	30
1/4/2020	2	35

Right Table

ID	Country
1	USA
2	Canada
3	Panama
4	Spain

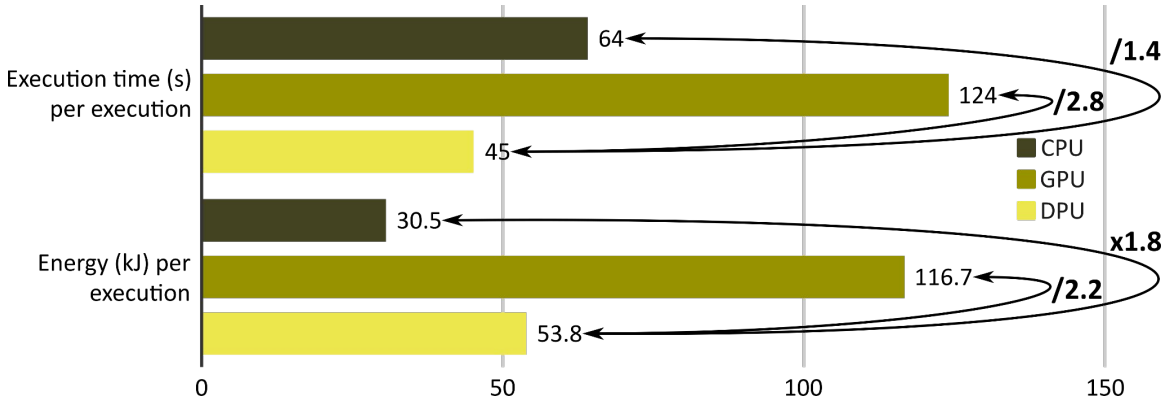
Merged Table

Date	CountryID	Units	Country
1/1/2020	1	40	USA
1/2/2020	1	25	USA
1/4/2020	2	35	Canada
1/3/2020	3	30	Panama
null	*null*	*null*	Spain

https://github.com/upmem/dpu_olap

ML : K-means Clustering

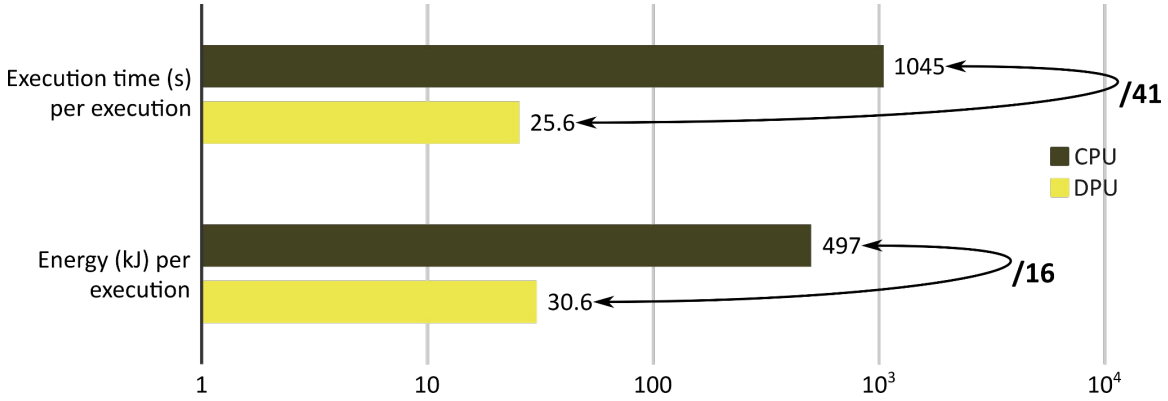
- **K-means** : partition the dataset into K distinct non-overlapping subgroups (clusters)
- ¼ days Criteo dataset

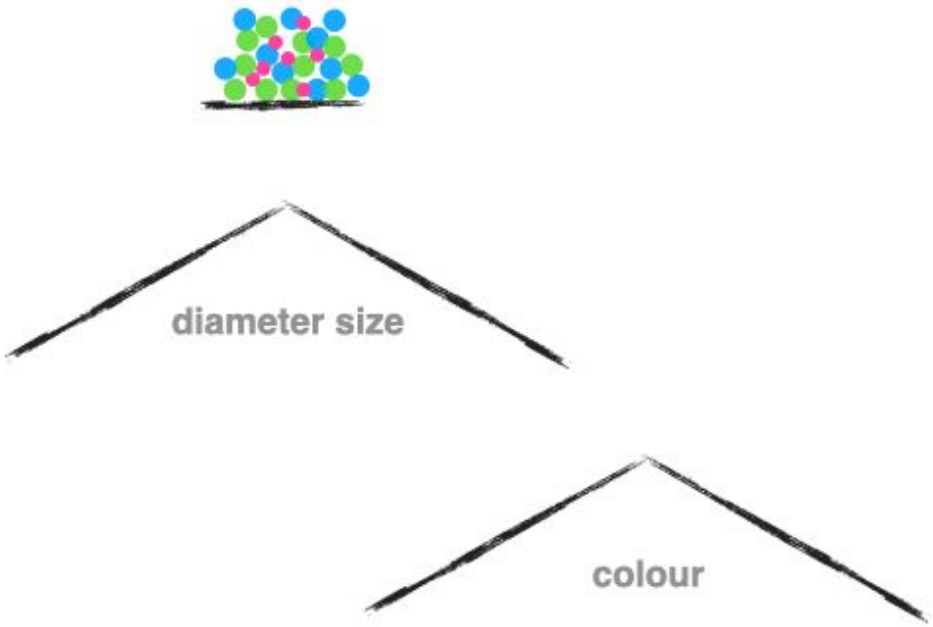


animation: shabal.in
https://github.com/upmem/dpu_kmeans

ML : Decision Trees

- **CART** training implemented on DPU : builds a binary-search tree which represents a partitioning of the feature space
- 2 days Criteo dataset

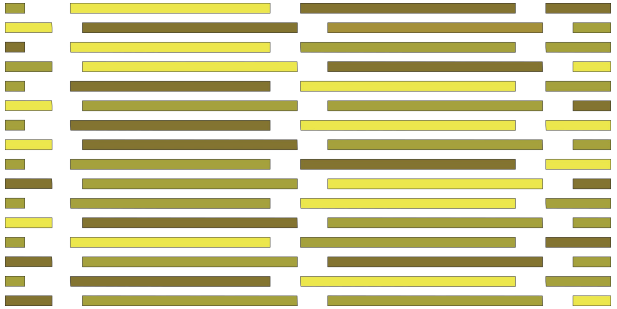




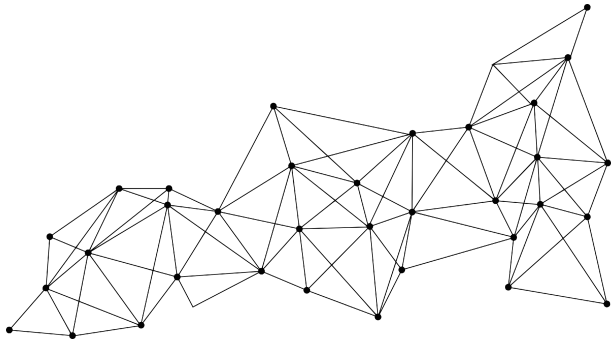
- Next step: XGBoost on PIM, throughput implementation

animation: ml2gifs.substack.com
https://github.com/upmem/scikit-dpu

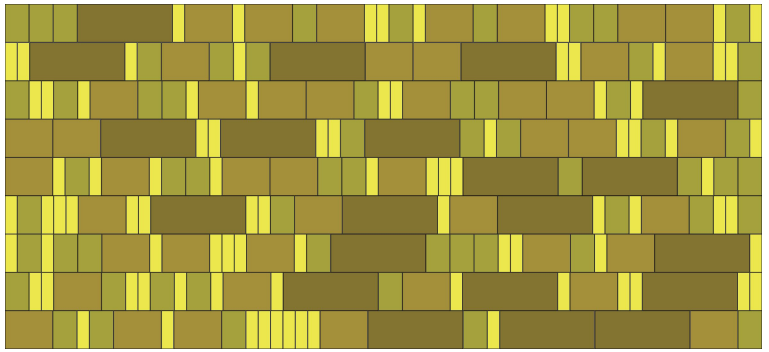
Algo patterns when PIM deliver great acceleration



Highly parallel operations with fine granularity partitioning



Irregular data access patterns



Algorithms with data of different types and sizes, difficult to vectorise

Research Projects

Collaborative projects

EIC Accelerator

ENERGIA/PIM

EIC Pathfinder

BioPIM

horizon europe

SustAInML

STRATUM 3D
3D Decision Support Tool for Brain Tumor Surgery

EUROPEAN UNION
EUROPEAN DEFENCE FUND

ΤΙΡΕΣΥΑΣ

ARCHYTAS

FRANCE 2030

ODYSSAI

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

GenoPIM
Processing-in-Memory for Genomics

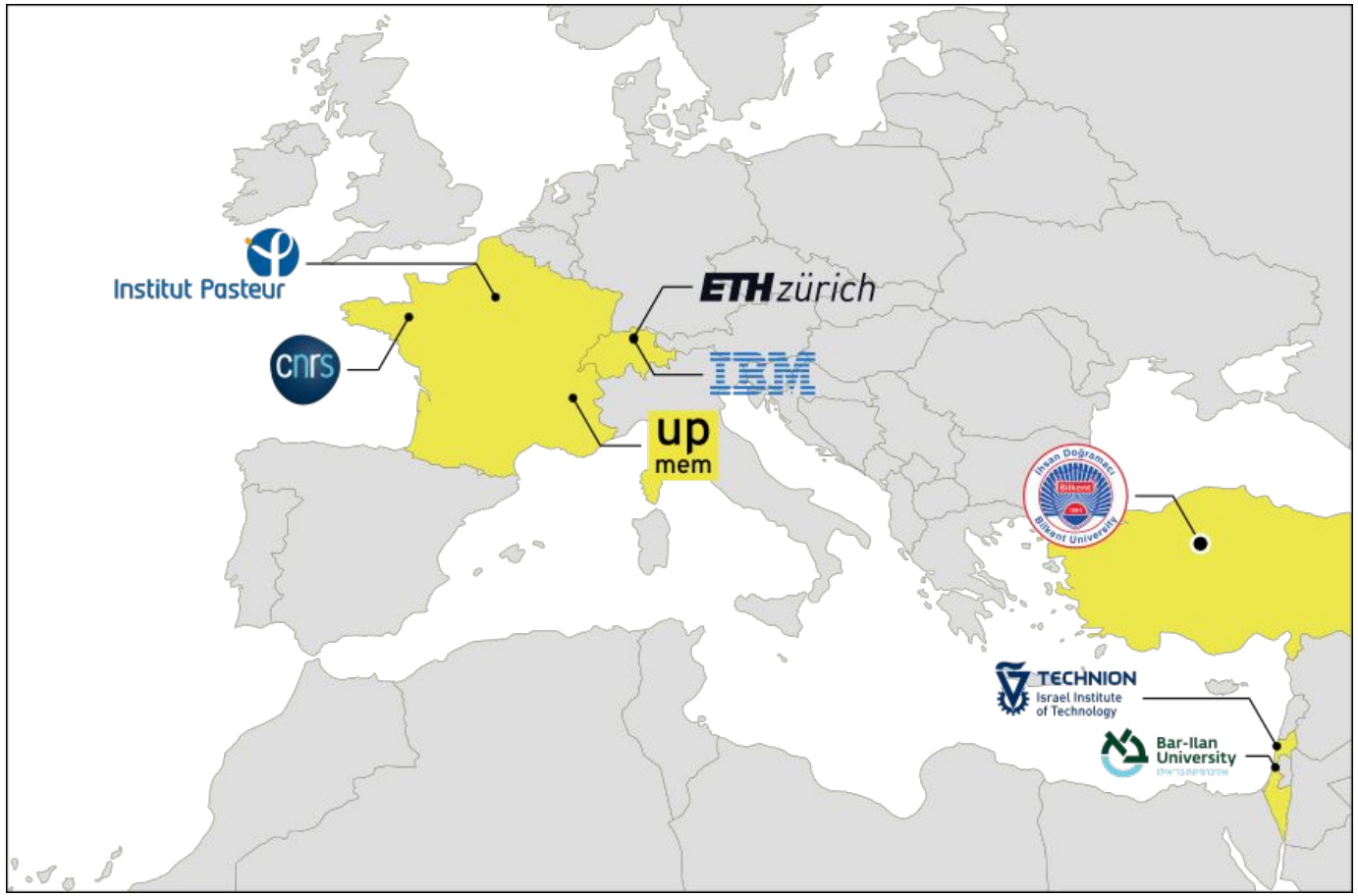
upmem

BioPIM

EIC Pathfinder

Co-designing algorithms and data structures commonly used in bioinformatics together with several types of PIM architectures to obtain the highest benefit in cost, energy, and time savings.

Coordinator	Bilkent Univ.
Start	05/22
Duration	4 years
Type of action	EIC Pathfinder
Total Budget	3 M€
Target TRL	3-4



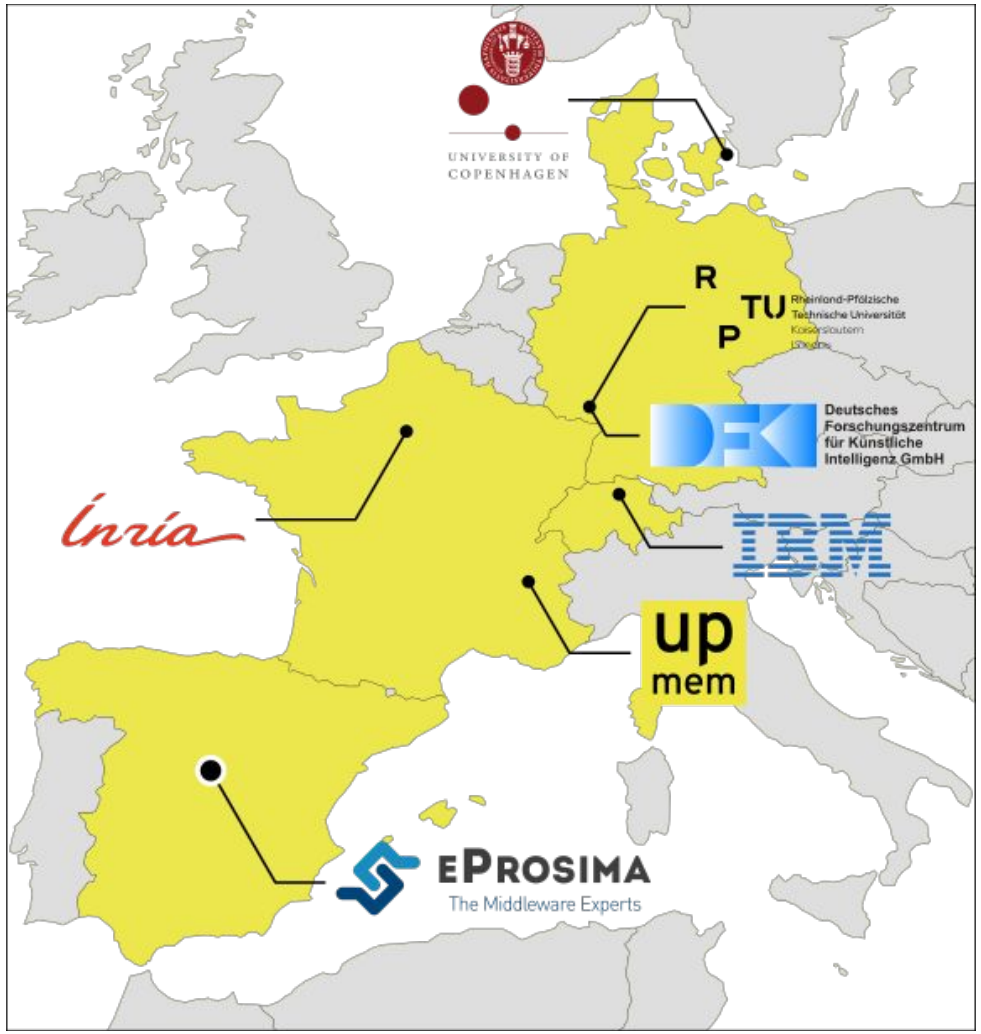
up mem

SustAInML

horizon europe

Sustainable, interactive ML framework development for Green AI that will comprehensively prioritize and advocate energy efficiency across the entire life cycle of an application and avoid AI-waste.

Coordinator	eProsima
Start	10/22
Duration	3 years
Type of action	HE-RIA
Total Budget	4.3 M€
Target TRL	4-5



up mem

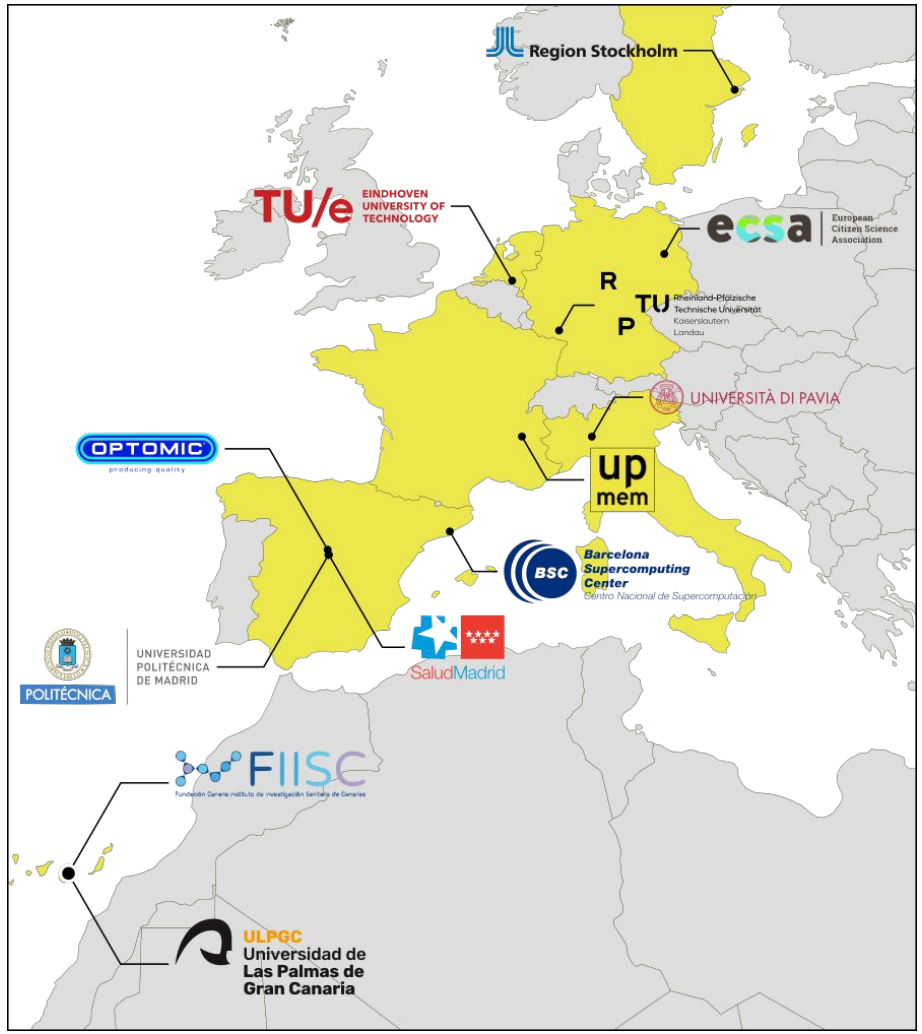
STRATUM 3D

3D Decision Support Tool for Brain Tumor Surgery

horizon europe

3D decision support tool for brain surgery guidance and diagnostics based on multimodal data processing through AI algorithms that will be integrated as an energy-efficient Point-of-Care computing tool.

Coordinator	ULPGC
Start	01/12/2023
Duration	5 years
Type of action	HE-IA
Total Budget	10.7 M€
Target TRL	>7



up mem

ARCHYTAS

EUROPEAN UNION
EUROPEAN DEFENCE FUND

The ARCHYTAS project explores advanced AI accelerators for defense, using novel technologies like optoelectronics, processing in memory, and neuromorphic devices. It integrates these with CMOS systems in a multi-chip setup and develops new programming models for improved performance and productivity in parallel systems.

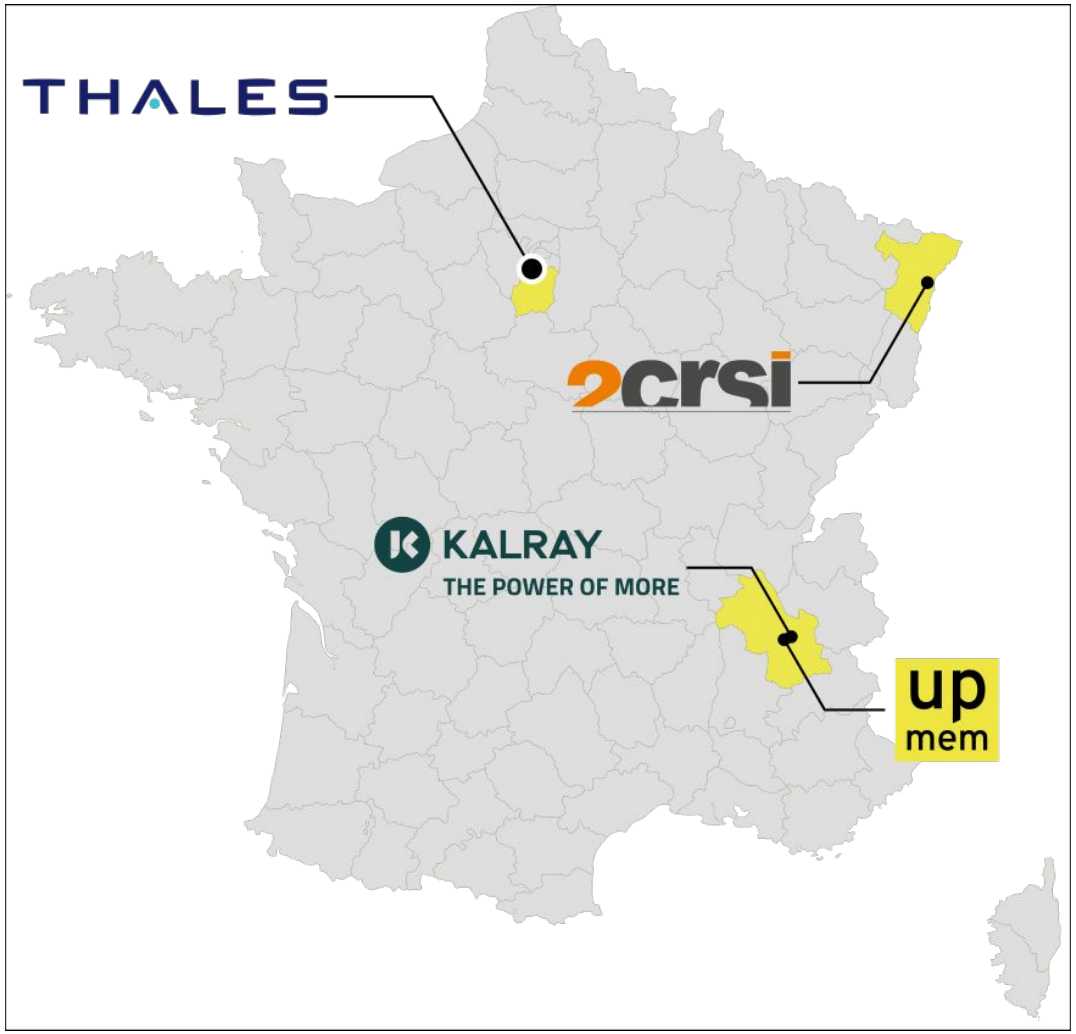
Coordinator	Iveco DV
Start	01/25 (TBC)
Duration	3 years
Type of action	EDF-RA
Total Budget	20 M€
Target TRL	<4

up mem

ODYSSAI

The ODYSSAI project aims to design a modular edge computing system for embedded AI applications in constrained environments. This open and secure system will use open source software and hardware technologies to achieve significant energy efficiency by implementing specialised computing accelerators for AI in optimised hardware platforms. The system will also simplify programmability and enable seamless deployment in the cloud.

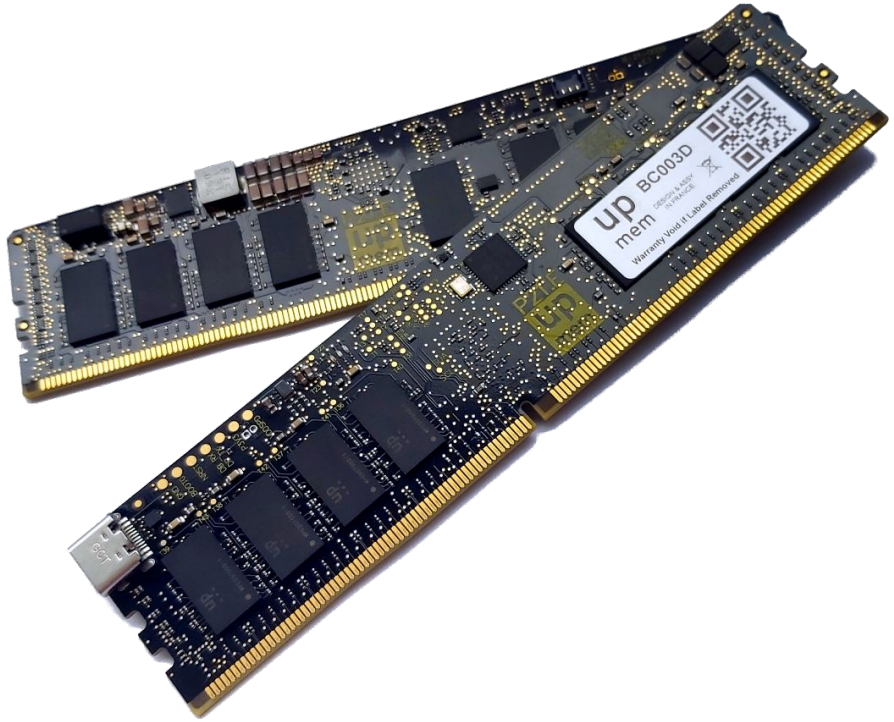
Coordinator	Thales
Start	12/23
Duration	3 years
Type of action	FRANCE 2030
Total Budget	7,87 M€
Target TRL	6-8



Technology Roadmap

Gen 1B PIM DRAM Modules

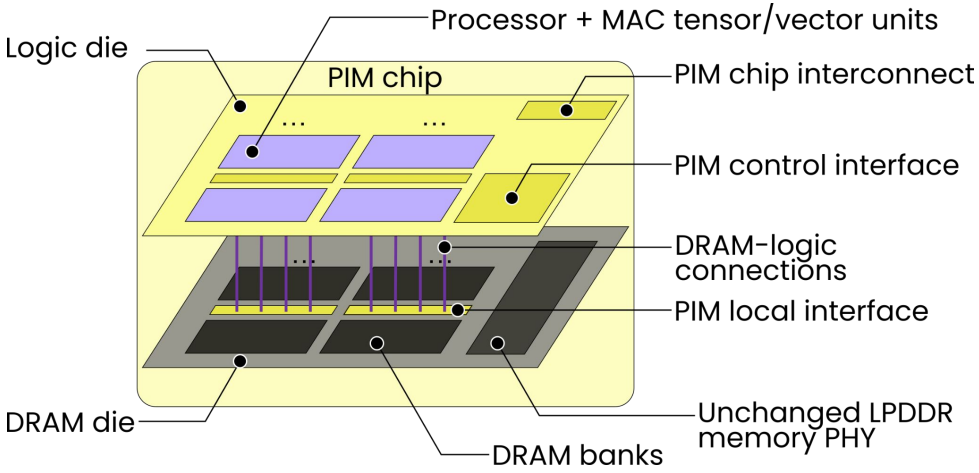
- Released last May
- Frequency increased to 400 MHz or up to 40% lower power consumption at same frequency (350 MHz)
- Host access to WRAM while the DPU owns the bank
- New HW monitoring features
- DPU switch off capability → Idle consumption ↘ by 90%



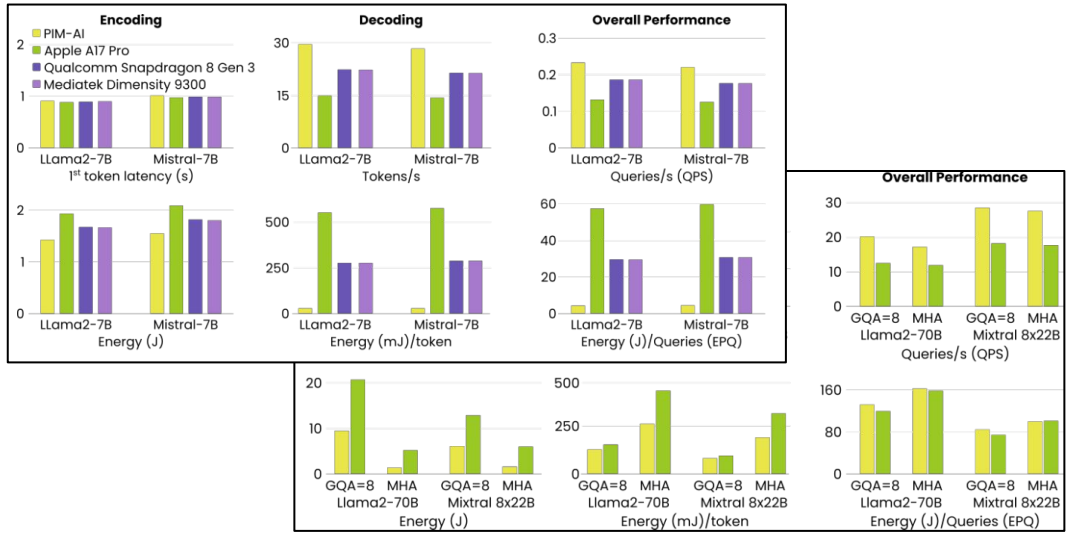
up
mem

Next-Generation Modules: PIM-AI

- Stacked Die Configuration:
  - Combining DRAM and logic dies in a single chip
  - 4 Linux-capable RISC-V processors
  - Tensor and vector units
- High Bandwidth and Low Energy Consumption:
  - 102.4 GB/s
  - Read/write energy consumption of just 0.95 pJ/bit
- Flexible Operation Modes:
  - Standard memory mode for conventional tasks
  - Accelerated PIM mode for performance-intensive AI operations
- Simulators
  - Pytorch simulator to be open sourced by the end of the year
  - QEMU and/or Gem5 simulator to be developed



Mobile



Cloud

Servers

From Skylake SP to Ice Lake SP

- Work in progress
- We expect to be able to ship these new servers by the end of the year

Exploration of Arm manycore processors

- Focus on Altra Ampere
- Part of the OdyssAI project

Investigation of other platforms

- Part of the STRATUM project
- Platforms not yet defined

Cloud infrastructure

In Numbers

- 10 servers
- 70 teams (+ 15)
- ~ 300 active users (+ ~100)
- ~ 40 000 hours booked (+ ~10 000)

Evolutions

- Gen 1B progressively deployed
- Service storage capacity (local disk, sftp for dataset pre-loading…)

Event Overview

Today's agenda

TIME	TITLE	SPEAKER(S)
09:00 – 09:15	Session welcome and aims	Yann FALEVOZ (UPMEM)
09:15 – 10:00	Keynote: UPMEM PIM platform for Data-Intensive Applications	Sylvan BROCARD (UPMEM)
10:00 – 10:30	Coffee break + Posters	—
10:30 – 11:00	Keynote: Next Generation UPMEM PIM DRAM for AI Applications	Cristobal ORTEGA (UPMEM)
11:00 – 11:22	Research paper: uPIMulator: A Flexible and Scalable Simulation Framework for General-Purpose Processing-In-Memory (PIM) Architectures	Bongjoon HYUN (KAIST)
11:23 – 11:45	Invited talk: Processing in Memory Virtualization	Dufy TEGUIA (UGA / Orange Innovation) / Jiaxuan CHEN (McGill University)
11:46 – 12:07	Research paper: SimplePIM: A Software Framework for Productive and Efficient Processing-in-Memory	Geraldo F. OLIVEIRA (ETHZ)
12:08 – 12:30	Research paper: High-level programming abstractions and compilation for near and in-memory computing.	Jeronimo CASTRILLON (TU Dresden)
12:30 – 13:30	Lunch Break + Posters	—
13:30 – 13:52	Research Paper: PID-Comm: A Fast and Flexible Collective Communication Framework for Commodity Processing-in-DIMMs	Si Ung NOH (Seoul National University)
13:53 – 14:15	Keynote: PIM Lucene	Sylvan BROCARD (UPMEM)
14:16 – 14:37	Research Paper: PIM-tree: A Skew-resistant Index for Processing-in-Memory	Hongbo KANG (Tsinghua University)
14:38 – 15:00	Research Paper: Enhancing Personalized Recommender Systems with PIM-Rec: Leveraging Processing-In-Memory Technology for Efficient AI	Niloofar ZARIF (University of British Columbia)
15:00 – 15:30	Coffee break + Posters	—
15:30 – 15:52	Research Paper: BIMSA: Accelerating Long Sequence Alignment Using Processing-In-Memory	Alejandro ALONSO-MARIN (BSC)
15:53 – 16:15	Research Paper: Compression of genomic data	Dominique LAVENIER (Univ. Rennes, CNRS-IRISA & Inria)
16:16 – 16:37	Research Paper: In-memory acceleration for HE with UPMEM PIM	Mpoki MWAISELA (University of Neuchâtel)
16:38 – 16:45	Closing	UPMEM

Useful links

- Website
- Resource page
- Github
- SDK

Thank you

Yann FALEVOZ, In charge of of Lab Relationship Management and Tech Marketing
yfalevoz@upmem.com