PIM and Various Computational Memory Solutions

MEMORY FOREST

2022. 04. 02

SK Hynix, Euicheol Lim

We Do Technology | SK hynix

CONTENTS

SK hynix

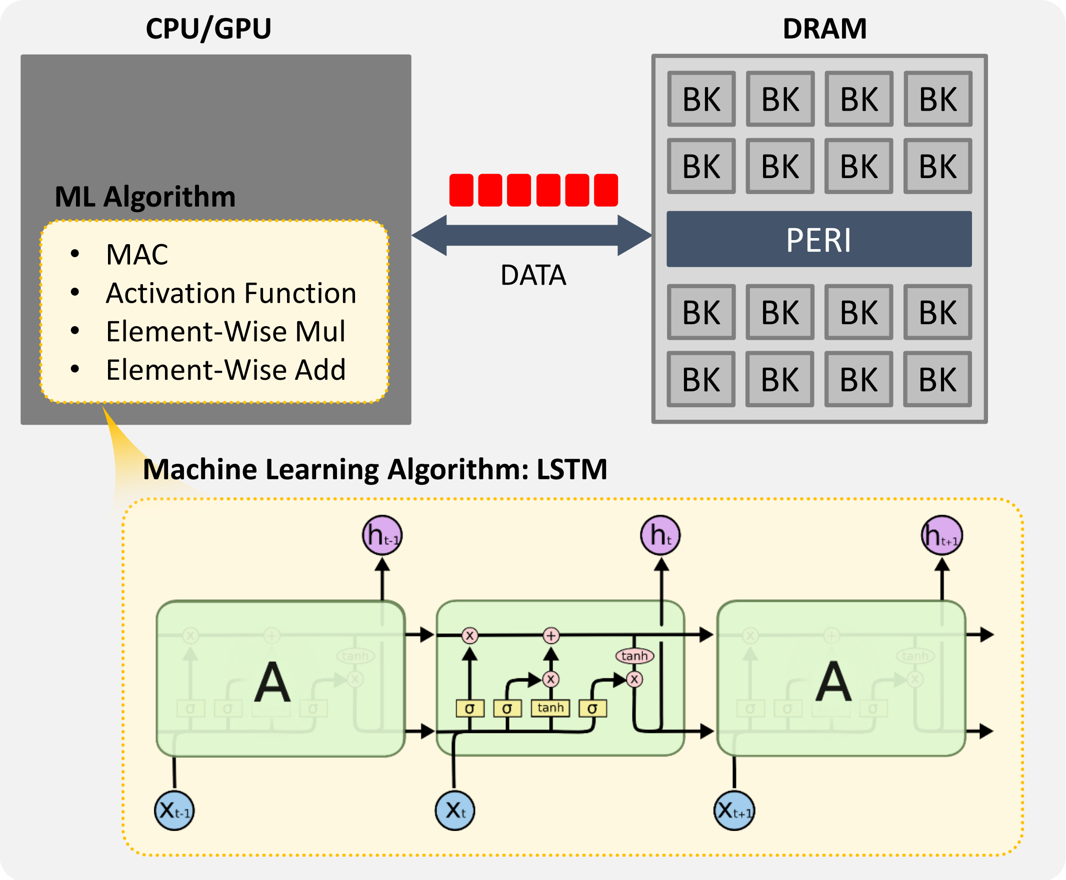
MEMORY FOREST

GDDR6-AiM Overview

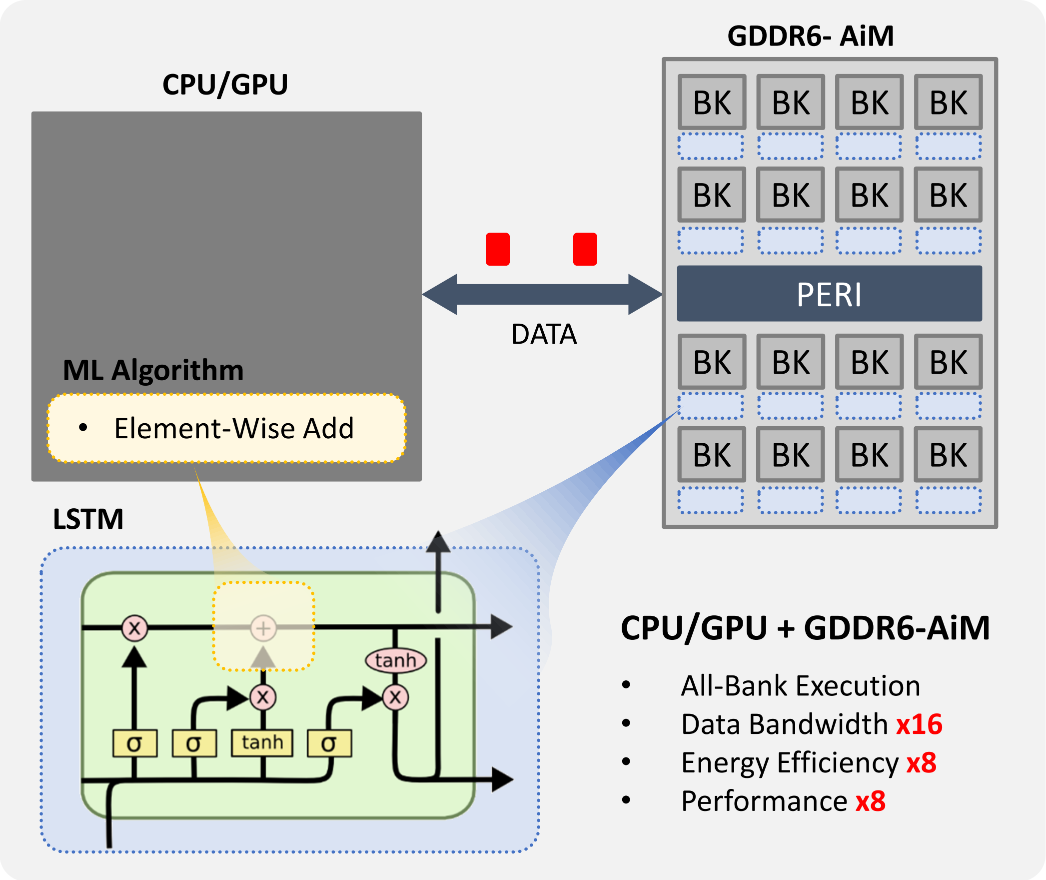
Definition

GDDR6-AiM is a **GDDR6**-based Accelerator-in-Memory (**AiM**) device targeted for memory-intense Machine Learning algorithm (**RNN, LSTM, MLP**) inference acceleration by offloading certain mathematical operations (**MAC, Activation Function, Element-Wise Multiplication**) from the host (**CPU, GPU, FPGA**).

Conventional System



In-Memory Accelerated System



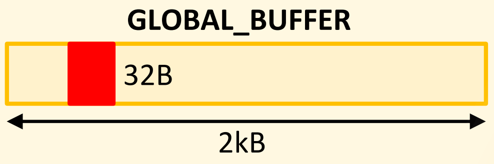
CPU/GPU + GDDR6-AiM

- All-Bank Execution
- Data Bandwidth **x16**
- Energy Efficiency **x8**
- Performance **x8**

GDDR6-AiM Overview

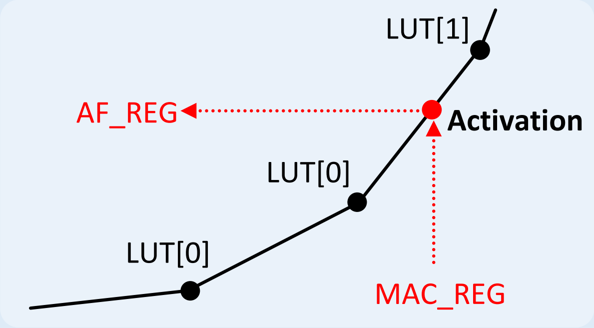
Global Buffer

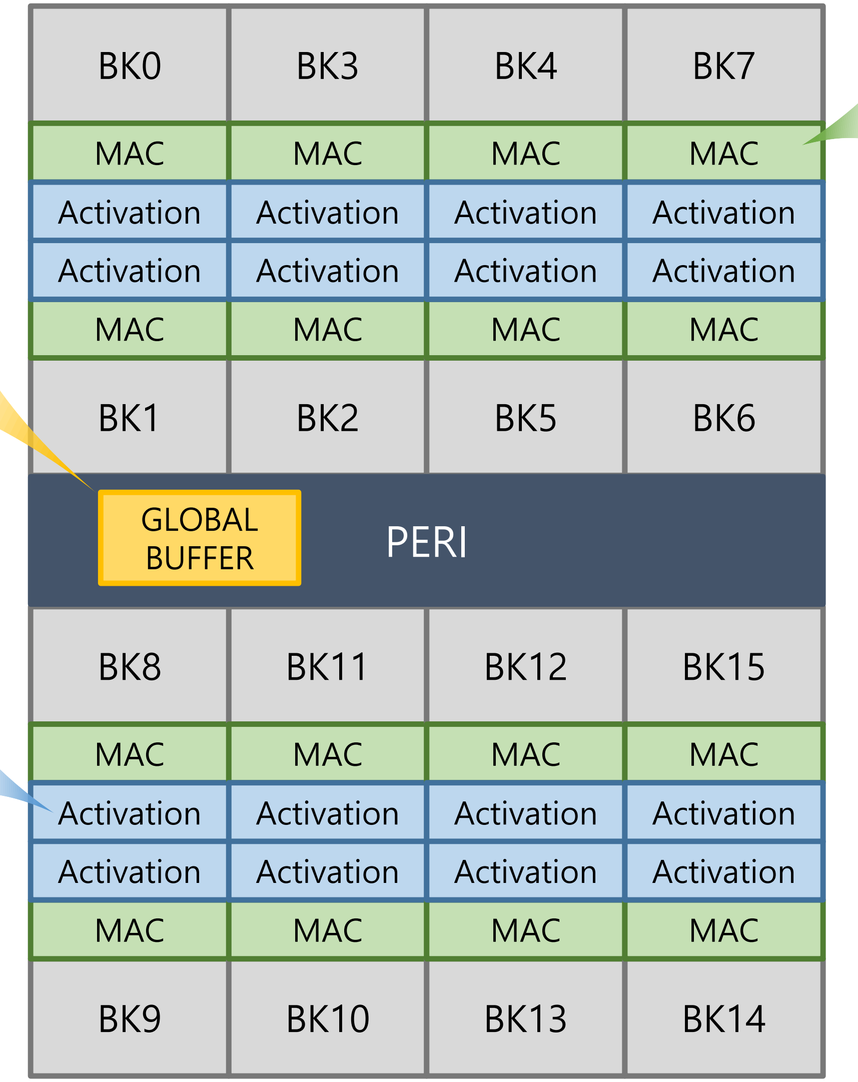
- Supplementary 2 kB SRAM buffer.
- Provides vector data for MAC.
- Supports 32B WRITE operations.



Activation Module

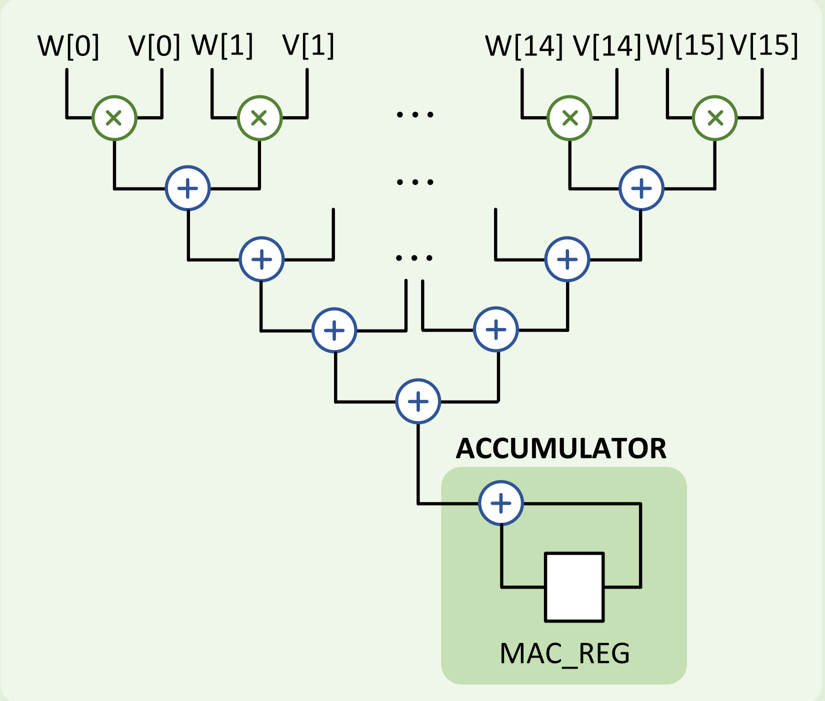
- Performs **Activation Function (AF)** computation by linearly Interpolating pre-stored AF template data using MAC calculation results.
- Activation results are stored in a dedicated **AF REG** set and can be later accessed by the user.





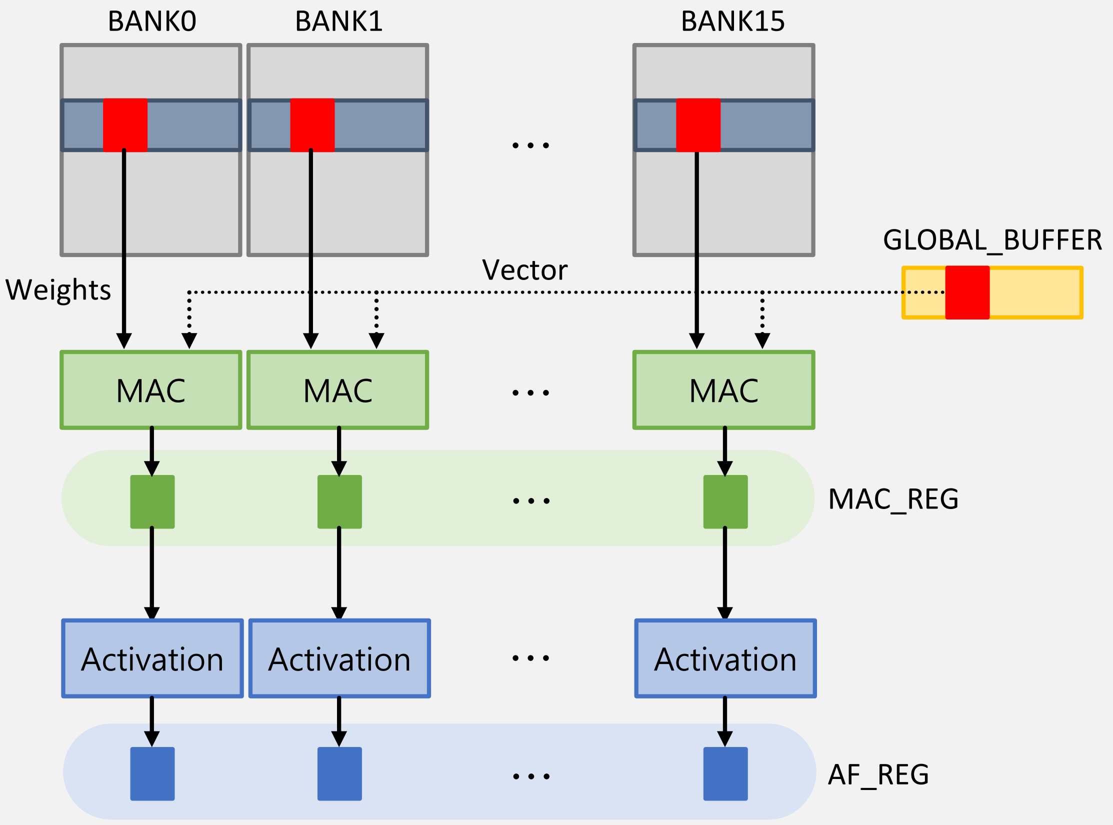
Multiply-And-Accumulate (MAC)

- Performs MAC operation on **sixteen** bfloat16 weight and vector elements (corresponds to a single DRAM column access, i.e. 32 Bytes).
- Computation results are stored in a dedicated **MAC_REG** set and can be later accessed by the user.



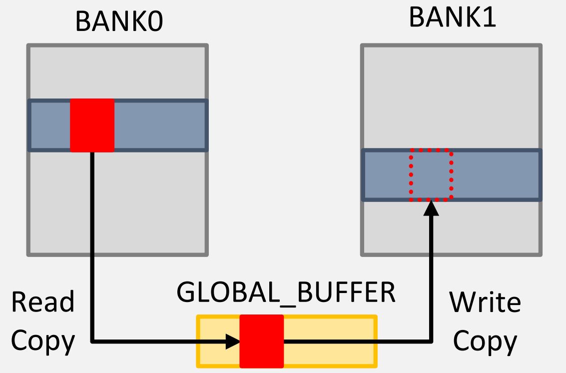
GDDR6-AiM Operations

All-Bank Operations (MAC, Activation Function)



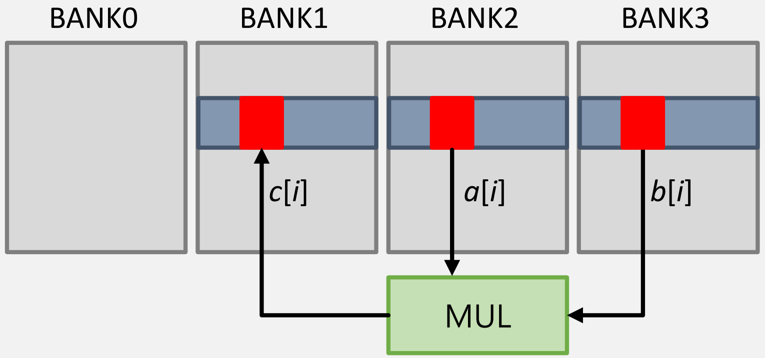
- **MAC** and **Activation Function** operations can be performed in all banks in parallel.
- **Weight** data is sourced from **Banks**; **Vector** data is sourced from the **Global Buffer**.
- **MAC** results are stored in latches collectively referred to as **MAC_REG**.
- **Activation Function** are stored in latches collectively referred to as **AF_REG**.

In-Channel COPY



- **Global Buffer** acts as FIFO register.
- **Read Copy** fills the FIFO, **Write Copy** transfers FIFO contents to a bank.

Element-Wise Multiplication



- Underlying expression: $c[i] = a[i] \cdot b[i]$
- One operation per **Bank Group** can be performed in parallel.

CONTENTS

MEMORY FOR EST

AI Service Business Case - Grab Driver

- Adopt data/analytics/AI platforms due to growing needs for real-time/large-scale/AI-based analysis.
- [1] AI Modeling enables prediction, [2] DB/Analytics stores and analyzes key biz data

Driver's App

DRIVER

Personalized real-time profit calculation & prediction report

Every 15-20 minutes

Personalized weekly/ monthly statistics & outlook report

Front-end: Service Application SW

Service Application

Back-end: Platform Application SW

Database : transaction/real-time analysis

Stores transaction/expenses information

Live Report (real-time)
Activity status/prediction for each driver

[2] Analytics : advanced analysis

Weekly, Monthly Report (Batch)
By driver/region/time of day…
Business analysis, e.g. profit-cost…

3) Data Lake Platform: storage of raw data

Data Repository (w/ Catalog)
- Operation/customer/external data
- Including image/video data

[1] AI Modeling Platform

Generates AI-driven prediction model

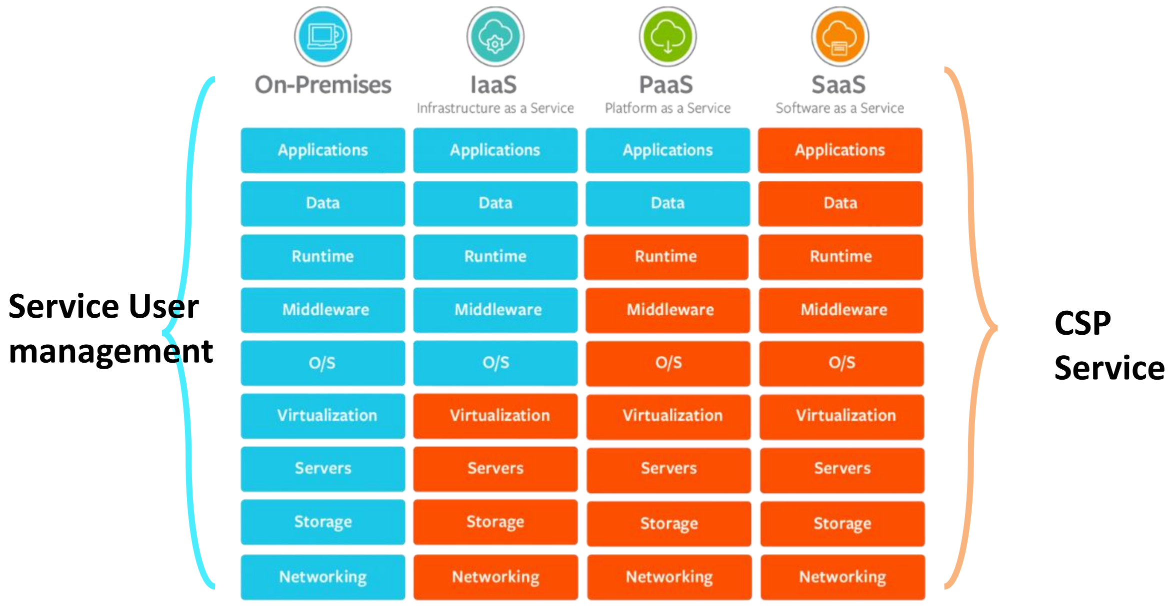
Develops AI prediction model using all data, past and present

On-premise vs. Cloud (IaaS, PaaS, SaaS)

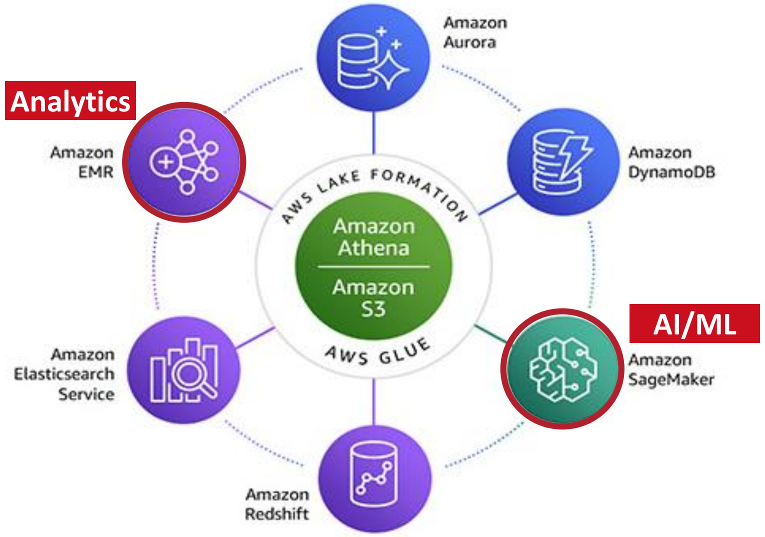
- On-premise enterprise → Center of gravity shifted to Cloud (IaaS, PaaS, SaaS) from 2010
- The cloud service is enhancing user convenience and strengthening AI/ML and Analytics PaaS

IaaS, PaaS & SaaS

- **IaaS:** HW resource provision → platform built by Service User
- **PaaS:** + Platform provision → Application development by Service User
- **SaaS:** + Application provision → Service built by Service User



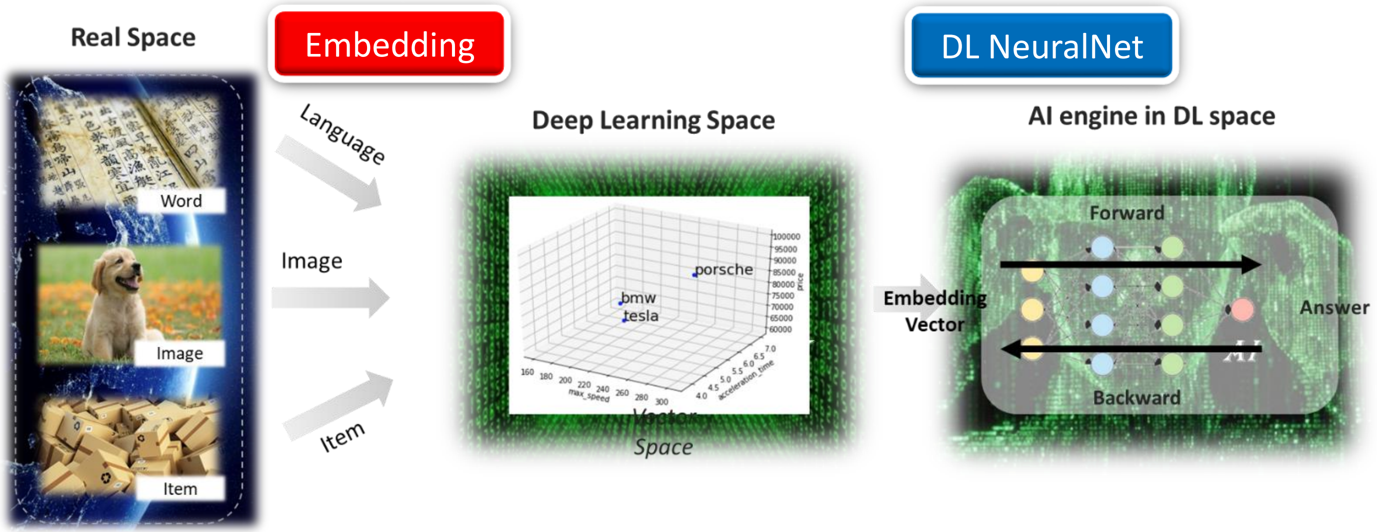
AI/ML, Analytics PaaS based on Data lake (AWS)



- **[CSP]** Opportunities for service/infrastructure optimization (cost reduction)
- **[Service User]** IT Cost/Business Reduction
- **CSPs understand domain specific customer requirements and workloads**

AI Model → Embedding + DL Neural Network

- **Embedding**
  - To translate Real Space Data into Deep Learning Space Data
  - Memory intensive function
- **DL NeuralNet (Transformer/MLP…)**
  - To do Neural Network in Deep Learning Space
  - Memory intensive + Computing intensive function



- **Embedding intensive AI Service: Recommendation**
- **DL NN intensive AI Service: NLP, Vision, …**

Data pipeline for AI service

- Real-time data analytics and AI systems are built as pipelines for data processing using AI.
- In the future, architectural convergence between AI-Analytics is expected.

Data Pipeline

Data Collection
- Data increase and large capacity due to IoT, 5G, etc.
- High Network Bandwidth

Up to EBs

Data pre-processing (Apache Spark)
- Streaming and Batch
- Growing demand for real-time data processing

Feature Engineering* → **Offline Pre-Processing****

Up to PBs — Up to 100TBs

Data Use
- Unstructured data analysis and larger ML model size
- Requires High performance computing & large memory

NLP Model training

Memory Intensive
- Embedding
- Weight Update
- Nearest Neighbor

Compute Intensive
- Transformer
- MLP, CNN
- Matrix Multiplication
- Regression, SVMs

Visualization
- Natural Language query

* **Feature Engineering**: Filtering, Imputing, Filling, Discretization, Bucketing, Split, Combine, ..
** **Preprocessing**: Normalizing, Encoding, Shuffle, ..

Data Store

Requirements	Massive capacity	Capacity, Bandwidth	Performance, Bandwidth	Latency, Bandwidth

System①: Real time Data Analytics

Data Analytics

System②: Massive AI

Embedding

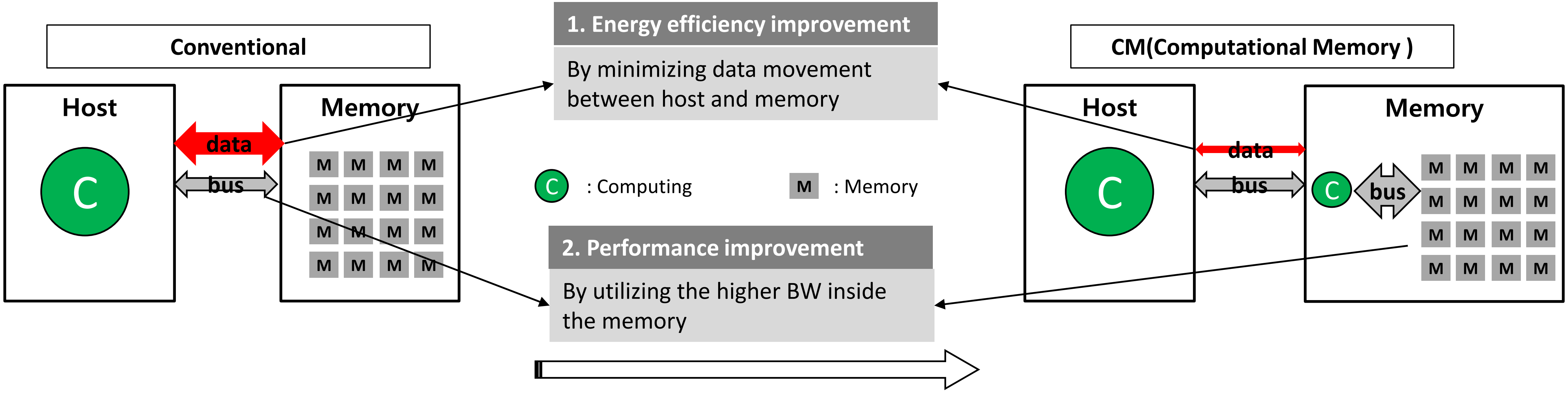
DL NeuralNet

CONTENTS

MEMORY FOREST

Back to the Basic – Computational Memory

- **Computational Memory Concept**
  - By performing some host operations on the memory side, energy efficiency and performance are improved



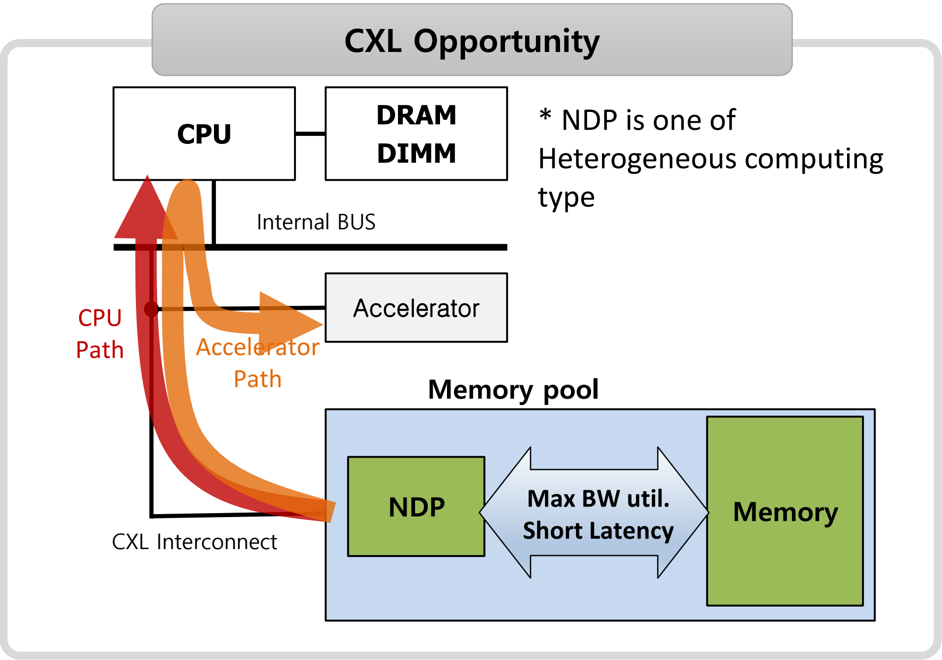
Offloading conditions

1. ***Low OI (operational intensity)***
2. ***Massive parallelism***
3. ***Data reduction***

Die level computational memory
→ PIM, GDDR6-AiM

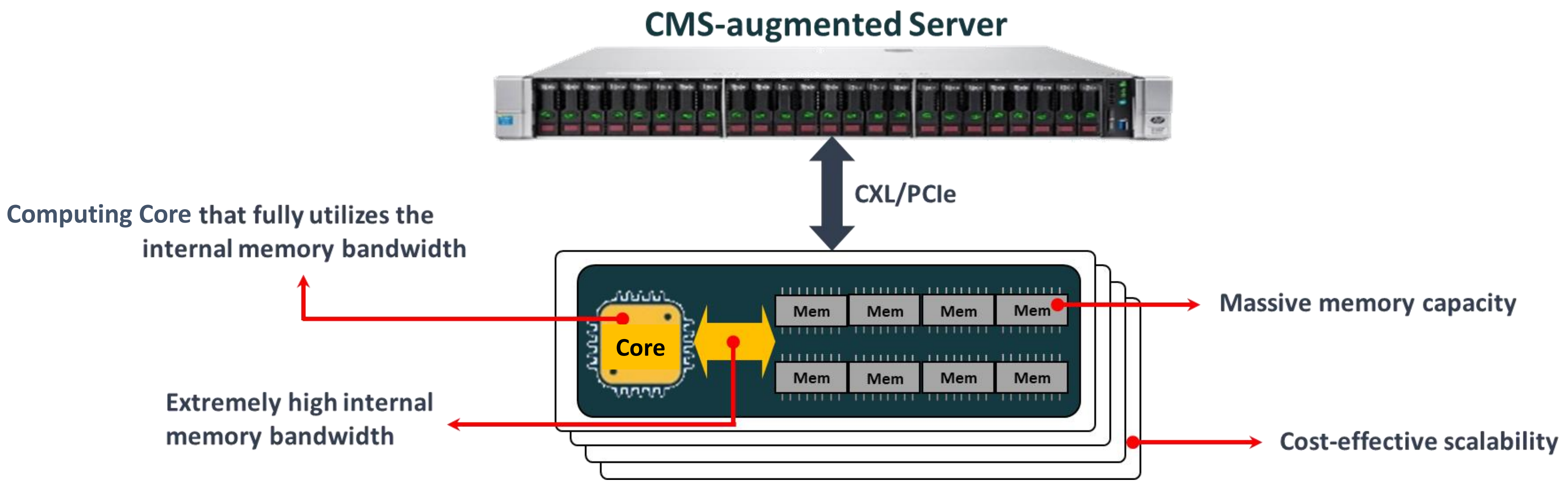
Check point : CXL - Heterogeneous Interconnect

- **CXL interconnect that connects Heterogeneous Computing Elements will become the center of Server System**
  - CXL is an interconnect to support high speed connection between Host Processor and Accelerators/Memory Device.
  - CXL supports 3 protocols based on PCIe Gen5.0
    - 1) CXL.io, 2) CXL.cache, 3) CXL.mem
- **Opportunity: Value added Memory solution available**
  - Unlike conventional DIMMs, CXL-connected memory protocol enables hand-shaking communication, enabling additional functions on memory (ex, DRAM cache, Data processing engine…)
- **With the advent of Memory-intensive Killer Application (AI) and Memory Semantic Interconnect (CXL), research and deploy of CXL memory-based Computational Memory is expected to accelerate.**

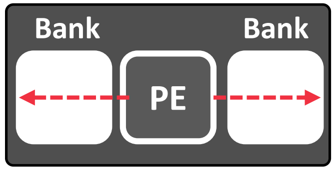


Card level CM - CMS (CXL base Computational Memory Solution)

- **Higher Performance by fully utilizing Memory BW + energy saving by data reduction + low cost high capacity**
  - Computing core can efficiently handle data-intensive workloads by fully utilizing memory bandwidth in card
  - Data reduction in the cards can significantly improve energy consumption by data movement
  - Cost-effective scalability makes the system to easily scale-up and out without having to pay for expensive servers just to increase the number of memory channels

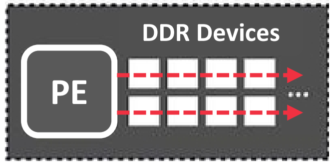


Various CM solution (Characteristics & Challenges)



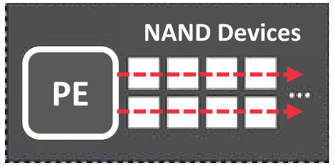
Die level CM - PIM

- Bank level parallelism
- Applied within one memory device die
- Small memory capacity per processing node
- Need to define Host interface and standardization



Card level CM - CMS

- Channel level parallelism
- Applied across multiple memory devices
- Larger memory capacity per processing node than PIM
- CXL interface available



Storage level CM - CSD

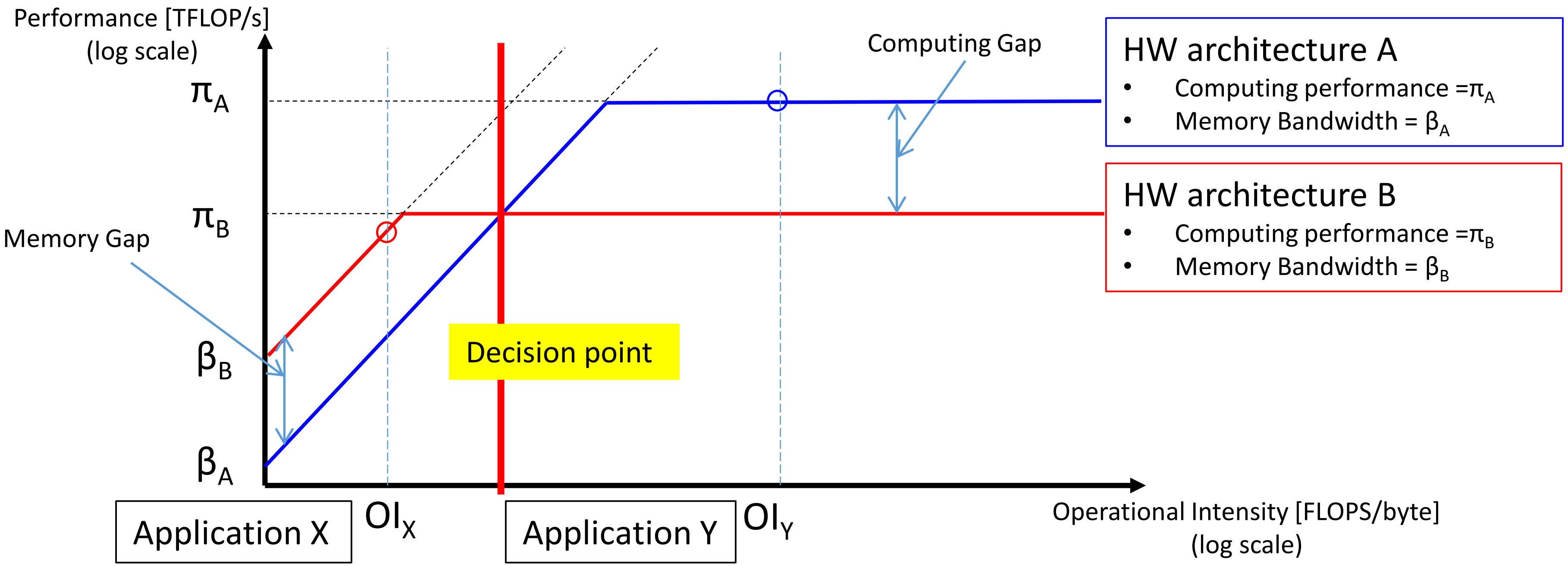
- NAND flash level parallelism
- Applied across multiple NAND flash devices
- Larger memory capacity per processing node than CMS
- Block interface → KV interface

Workload analysis for actual use case

SW framework support such as Compiler, API, Library, device driver

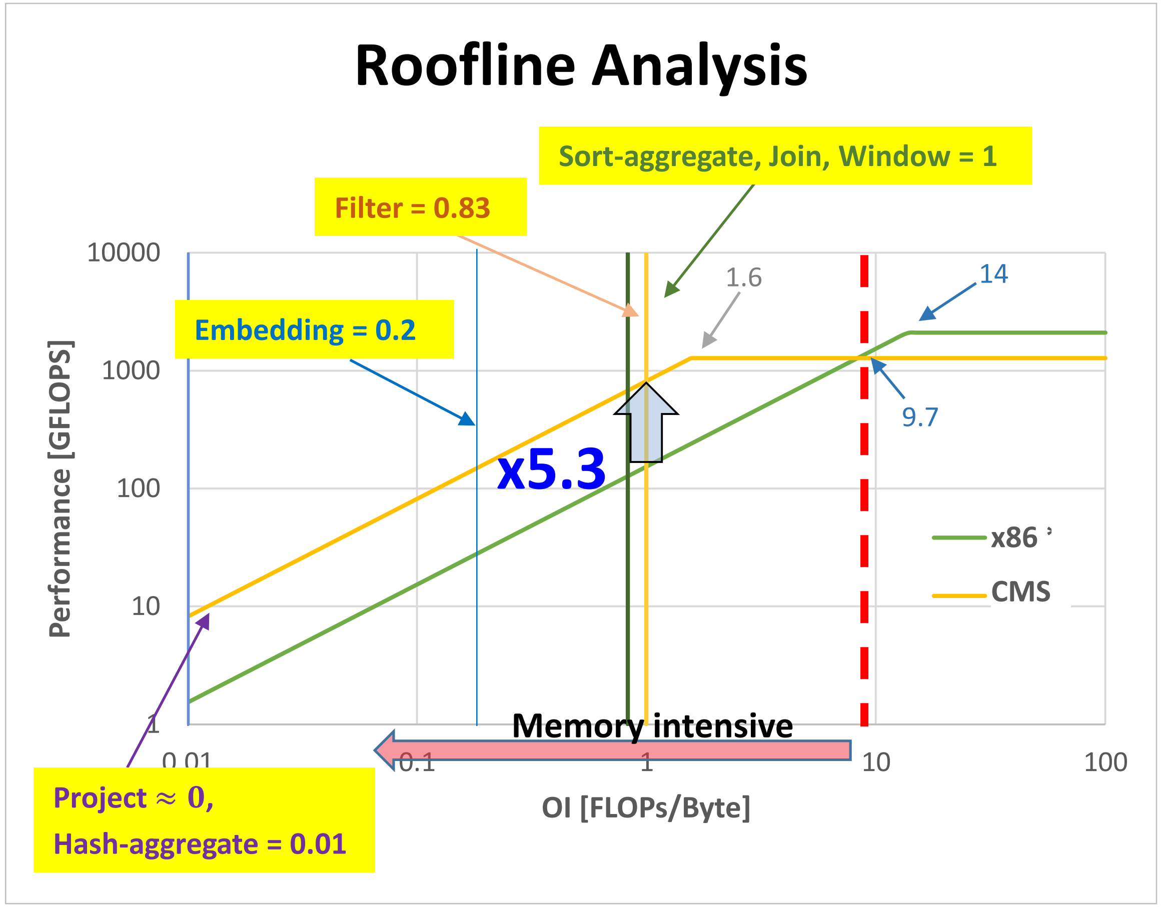
Roofline Analysis

- **Methodology that can analyze HW architecture suitable for a specific processing algorithm**
  - W (Work) : # of operations performed by a given application
  - Q (Memory Traffic) : # of bytes of memory transfers incurred during execution of application
  - OI (Operational Intensity) = W/Q : # of operations per byte of memory traffic.
  - P (Attainable performance) = min (π, β x OI) : In given HW, π is max processing performance, β is the max bandwidth



Workload Analysis – Data Analytics, Embedding

- Representative Data Analytics functions have overall low operational intensity characteristics → Memory bound
- The embedding operation is also a memory bound operation with very low operational intensity.
- So, if these operations are operated in the Computational Memory, performance and power gains can be obtained.



Workload Analysis – DL Neural network

- **Matrix multiplication, which has been the main target of PIM, is losing its memory-intensive characteristics as the batch increases and algorithm evolves.**
- **There is still an opportunity for offloading of memory-intensive functions regardless of batch size increase such as layer normalization and any kind of function for data itself.**

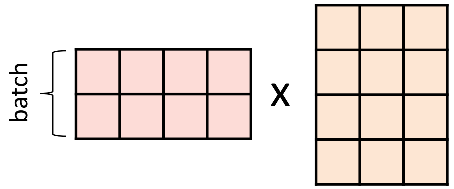
Batch vs. OI

- **GEMV (related w/ weight)**

  small batch → memory intensive, large batch → computing intensive
- **Normalization, Optimizer (related w/ data)**

  Even in large batch → memory intensive

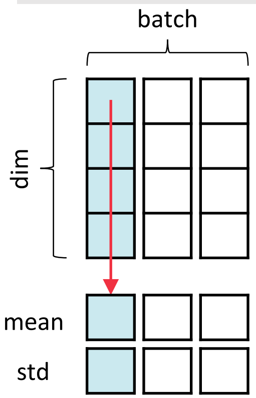
GEMV



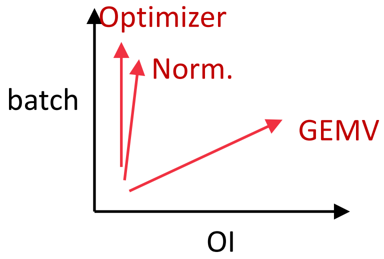
Vectors

Matrix (Reused)

Layer Normalization

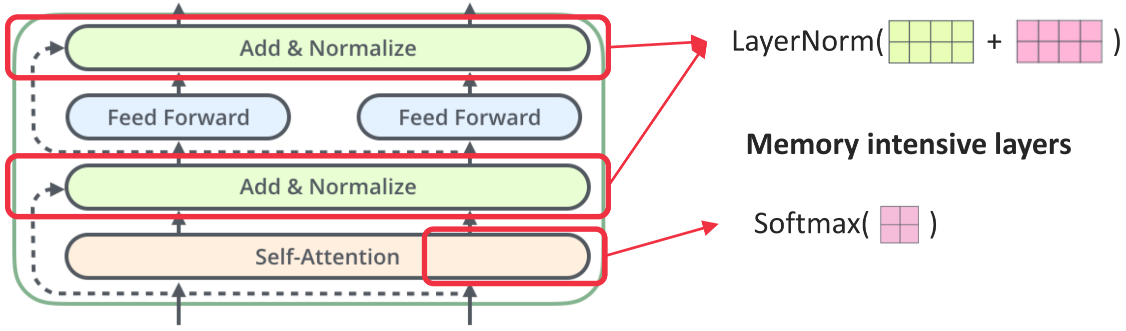


Batch vs. OI



Low OI Layers in Transformer

- **In Transformer, Memory intensive operation takes significant portion of the total execution time**
- **There is also memory intensive operation in Attention layer (softmax, biases, dropout)**

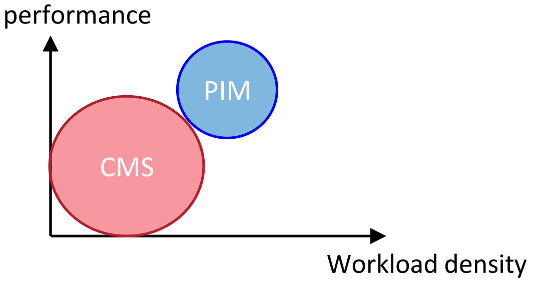


Single Transformer Block *Source: https://arxiv.org/abs/2007.00072

Workload Analysis – Workload density

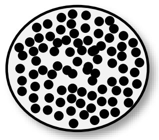
Workload Density – How dense is the data to be processed in the memory

→ **If the workload is sparse compared to the memory capacity per PE, data reduction per PE is reduced and frequent data movement between PEs is induced.**



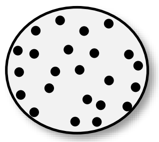
The higher workload density, the more appropriate for PIM

High workload density

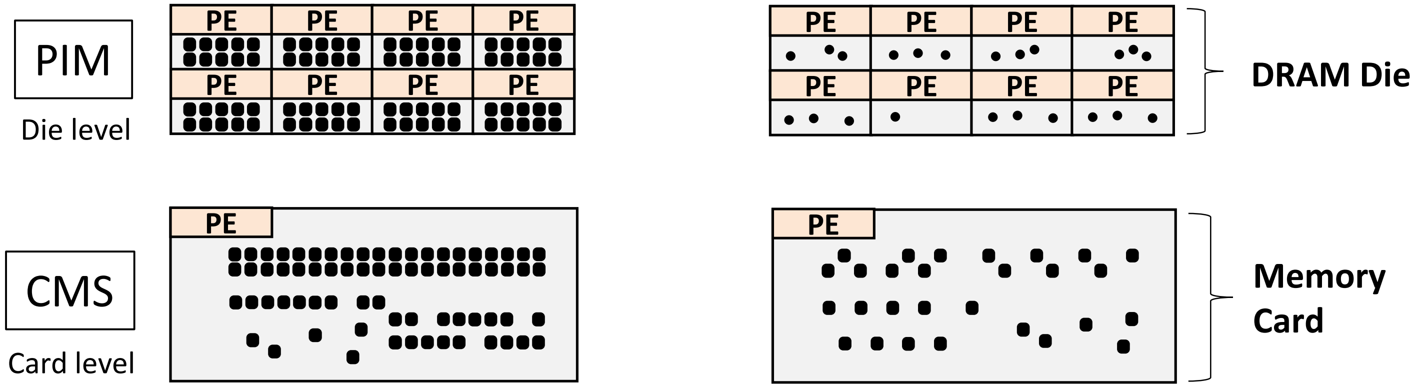


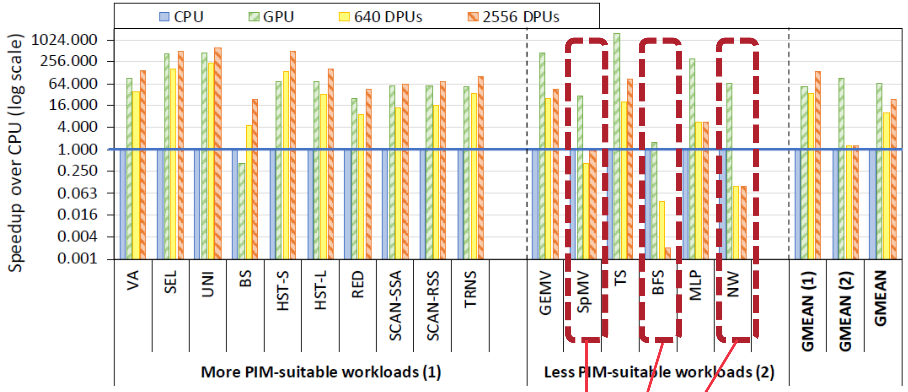
Deep learning NN
(MLP, CNN, Transformer…)
Data Analytics

Low workload density



Embedding (DLRM)
Graphs
Big Data/Data Analytics





Gómez-Luna, Juan, et al. "Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture."(2021)

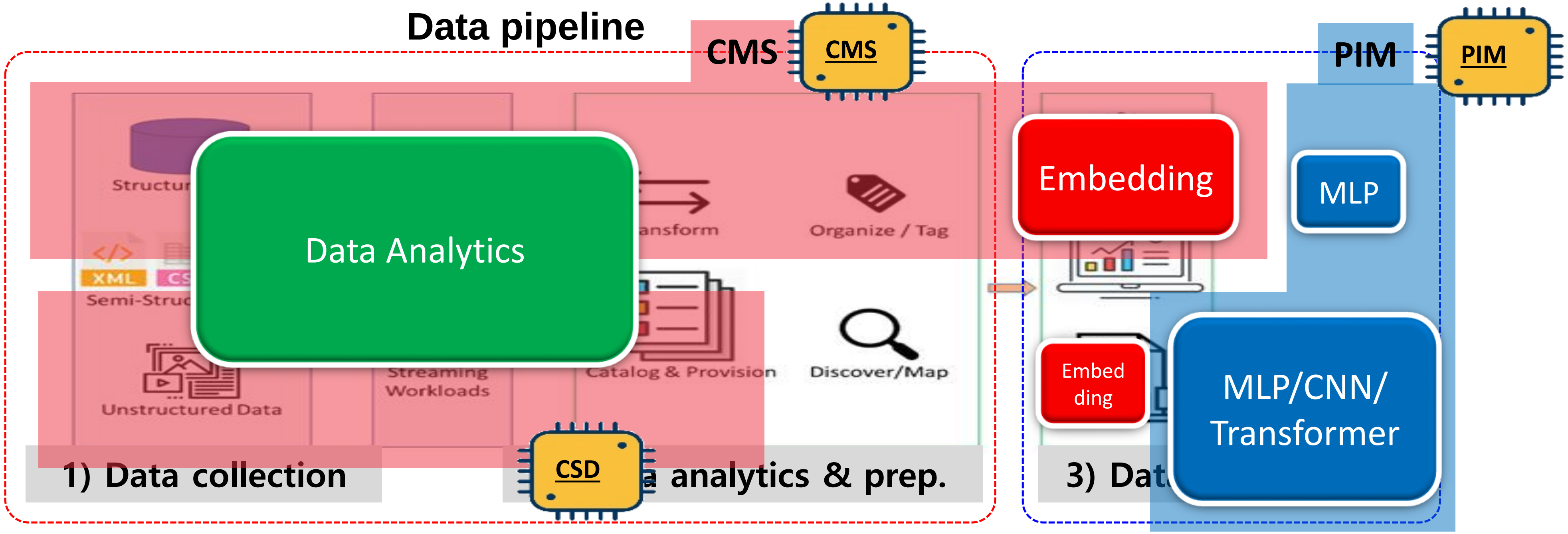
Sparse Data, large data movement among processing node.

CONTENTS

MEMORY FOREST

Summary

- **Value addition memory base solution can be deployed in whole data pipeline of AI/Data Analytics system**
  - Die level CM : PIM
  - Card level CM : CMS
  - Storage level CM : CSD



End of Document