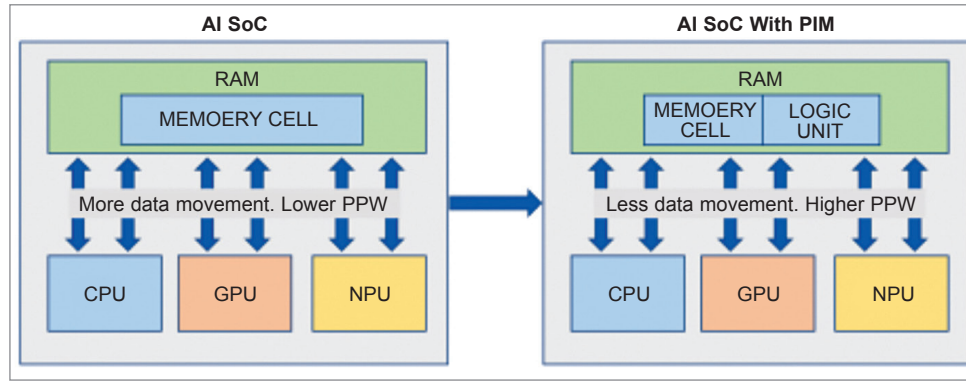
The Case Of PROCESSING-IN-MEMORY For AI SoCs

Processing-in-memory (PIM) has the potential to revolutionise AI SoCs, promising game-changing enhancements in speed, efficiency, and sustainability

The author, **CHETAN ARVIND PATIL**, is Senior Product Engineer at NXP USA Inc.



Emerging artificial intelligence (AI) applications will be more compute and memory-intensive than ever. The system-on-chip (SoC) architecture is crucial to manage such applications efficiently.

These SoCs rely on data pipelines and memory organisations to process the request. However, with every new AI application, the amount of data that needs to be processed is increasing 100 times, and the SoC design innovation cannot match it. Eventually, traditional SoCs will not be able to cater to the heavy demands of AI applications. Thus, several AI companies have started focusing on customised SoC solutions.

The emergence of AI SoC

This is where AI SoC also comes into the picture—Artificial intelligence system-on-a-chip. AI SoC uses either the aggregated (single silicon die) or disaggregated (multiple die, i.e., Chiplets) design and manufacturing methodology. What sets these apart from the traditional SoCs is that they are custom-designed to provide various types of processing elements (PEs)—CPU, GPU, NPU, etc, to ensure data is processed by maximising the number of tasks per given cycle.

Common challenges in AI SoCs

However, AI SoCs still suffer from the same issues as traditional SoCs. The data has to be moved continuously from lower-level memories (DRAM/SRAM) to upper-level memories (L3/L2/L1/registers). AI data-driven applications are in continuous contention to process as much data as possible, making AI SoCs spend clock cycles to move data from upper to lower-level memories and vice versa. All of this leads to slow processing and increases power consumption.

PIM solutions for AI SoCs

AI SoC should start adopting processing-in-memory (PIM) solutions to mitigate this bottleneck. PIM is a memory solution that combines logic functions and memory, allowing data processing at lower-level memories while also handling data with PEs with upper-level memory. PIM gets fabricated inside high-bandwidth memory (HBM), which not only brings the best of computing (within the memory) but also data transfer (to and from memory).

So far, Samsung and SK Hynix have developed a PIM memory solution. The data shared shows an 80-85% reduction in power consumption. It is a significant savings, given that AI applications are moving the computing industry towards TOPS—trillions of operations per second. Any solution that can increase the TOPS while improving performance per watt will be a game changer for the AI SoC domain.

The path forward for AI SoCs

Currently, there are no mass-produced AI SoCs with PIM-based features. With the promising solutions from Samsung and SK Hynix (with Micron also re-exploring PIM), there is a strong case for developing AI SoCs using this new memory architecture. It will not only speed up the processing request of AI applications, but when combined with more-than-Moore solutions like chiplets, it can revolutionise how server-grade data centres get designed, reducing the number of server racks—making data centres more energy efficient.

> ANY SOLUTION THAT CAN INCREASE THE TOPS WHILE IMPROVING PERFORMANCE PER WATT WILL BE A GAME CHANGER FOR THE AI SoC DOMAIN

AI SoCs with PIM will also require several system software-level changes. The applications must manage data processing with PEs and also with PIMs, all concurrently, without introducing a clock cycle penalty.

In summary, a memory-level solution that can speed up the data flow movement and simultaneously lower the number of clock cycles required to process trillions of data points will be a game changer for AI SoC. In this regard, PIM-powered AI SoC is definitely one such solution. EFY

"Never Put A Person Putting Faith In You Down"
—Sanjay Gupta, Spark Minda

"India Has Significant Demand, So Why Not Produce Indian Chips?"– Raja Manickam

₹100

NOVEMBER 2023

The New electronics FOR YOU

startups:

MINDGROVE TECHNOLOGIES

HARNESSING THE SHAKTI FOR MICROCONTROLLER CHIPS

RAPTEE

THE FIRST HIGH VOLTAGE ELECTRIC 2-WHEELER IN INDIA

RACENERGY

MOST ENERGY-DENSE ELECTRIC VEHICLE BATTERIES

An EFY GROUP Publication
Vol. 55 No. 11 • ISSN 0013-516X
Pages: 110 • UK £5; US $10

MUST READ

"High-Value Addition Can Be Achieved Only If The Percentage Of Local Value Addition Is Increased" — Jairaj Srinivas, Director General, CIMEI

Interesting Reference Designs Of EV Chargers

Trends And Test Challenges In Radar And Lidar Sensor Technologies And How To Solve Them

ChatGPT Terminal Made Using ESP32