Peta Ops, Associative Compute-In-Memory On-Board Processing: Coming to a Spacecraft in Your Orbit!

In-Orbit AI and Machine Learning will disrupt satellite services and applications: Earth-Observation operators are performing more on-board processing to extract the valuable insights in real-time for disaster management rather than downlink bandwidths of data for slow, cloud-based post processing on the ground. Smart, telecommunication transponders are autonomously re-configuring and optimising their frequency plans based on live traffic and link needs to maximise performance without intervention from the ground.

Space Debris Retrieval, De-Commissioning, In-Orbit Servicing, and Defence operators are using machine learning to perform space-domain awareness to identify and detect the trajectory and intent of other orbiting objects in real-time!

Current on-board digital processing is being realised using space-grade microprocessors, DSPs, FPGAs, and ASICs, as well as COTS GPUs, based on traditional von-Neumann/Harvard architectures, where instructions and data are fetched from external memory. Both these approaches are constrained by memory bandwidth limiting future needs to fuse and process terabytes of data from multiple sensors in real-time to enable the next generation of smart, *in-orbit* applications, *e.g.,* AI and Machine Learning.

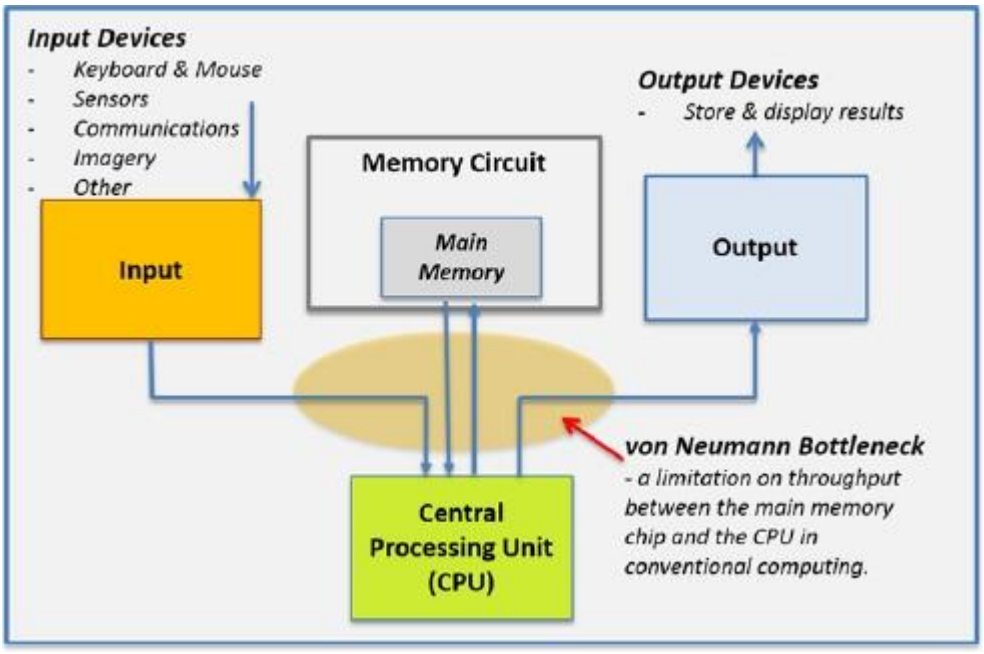


Figure 1: The von Neumann bottleneck constrains performance.

Associative Compute-In-Memory was first proposed as a concept in the 1960s but couldn't be realised because of the absence of semiconductor memory that was both high-density (storage capacity) and fast (speed). Recent advances in affordable SRAM memory with low access times, together with the increasing needs for data-centric applications, such as in-orbit machine learning, have revived interest in Associative Compute-In-Memory for certain types of applications.

Traditional computing architectures are handicapping the demand for increasing performance due to memory-bandwidth constraints and power consumption. Current on-board digital

processing has become very inefficient, with prohibitive energy and performance costs due to large-scale data movement between memories and processors, *e.g.,* a 700 W H100 GPU. Eliminating the need to retrieve data and instructions from off-chip storage will significantly increase performance and throughput (bandwidth), while simultaneously mitigating latency and reducing energy demands.

In traditional data storage methods, information is typically organized in a sequential, structured format such as tables or files, accessed via specific queries. However, Associative Compute-In-Memory takes a different approach. Here, data is stored in a highly parallelized and distributed manner, facilitating simultaneous processing of multiple data elements. This innovative architecture enables the efficient handling of extensive data sets and complex queries, resulting in expedited results.

Associative Compute-In-Memory revolutionizes data access by abandoning rigid data structures. Instead of adhering to predetermined formats, data is retrieved based on its content or relationships, eliminating the need for addressing. This natural access method simplifies the exploration and analysis of intricate data sets, facilitating pattern recognition, predictive modelling, and insightful discoveries without the constraints of traditional storage systems.

By harnessing the inherent parallelism and swift data access capabilities, complex computations can be executed in real time, facilitating rapid insights and informed, data-driven decisions. Each processor bit is content-addressable, serving dual purposes of data storage and processing, thereby maximizing efficiency and performance.

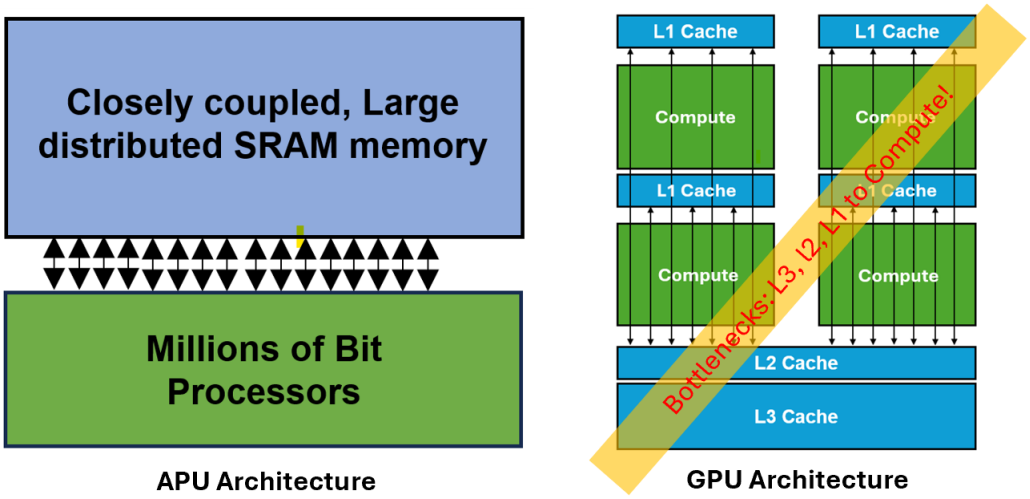


Figure 2: Associative Compute-In-Memory Architecture.

GSI Technology is in production with an Associative Processing Unit (APU) and is bringing to market a range of APU products for space applications. These are known as Gemini-I and Gemini-II. The latter can be configured for low-power, Edge-based processing delivering 6.6 TOPS of performance or high-performance computing offering a staggering 184 TOPS or 6.6 Peta Boolean operations per second! The following table compares the Gemini-II specifications:

	Gemini-II (TRL5)	Gemini-I (TRL9)
Operating Frequency[1]	1.4 GHz – 200 MHz	600 MHz – 200 MHz
Bit Processors	1M, (2-bit)	2M, (1-bit)
Associative (Compute) Memory	48Mb	48Mb
Boolean Ops/sec	5.87 Peta (4 per cyc)	1.26 Peta (1 per cyc)
Full Adder Ops/sec	1.47 Peta (1 per cyc)	0.314 Peta (1 per 4 cyc)
Compute Performance (8b ADD)	184 TOPS / 6.6[4] TOPS	38 TOPS
Compute Performance (4b MAC)	196 TOPS / 7.0[4] TOPS	44 TOPS
Local Memory (L1)	768Mb, distributed	96Mb, distributed
L1 <-> BP Data Bandwidth	367 Tb/sec	315 Tb/sec
I/O Interfaces	PCIe4 x16 (256 Gb/s) - Host Intfc PCIe4 x4 (64 Gb/s)[3] - Chip2Chip Intfc DDR4 x72 (204.8 Gb/s)[2] Ethernet 10G/1G, 32 GPIOs UART, I2C, SPI, JTAG 1149.6 eMMC/SD - boot code loading	UBus x64 (76.8 Gb/s)
Power (TDP)	130 W to 10 W[4]	80 W
Technology	16 nm	28 nm
Package (FCBGA)	37.5 mm x 37.5 mm (1296)	25 mm x 2 5mm (576)
Voltage	Core 0.8V	Core 0.95V

[1] Configurable: 200 MHz to 1.4 GHz
[2] Using -3200 DDR DRAMs
[3] For chip-to-chip (board-to-board) interconnect
[4] One core configured at 200 MHz

Table 1: Specifications of Gemini-I and Gemini-II APUs

The evaluation module for the low-power Gemini-II APU is shown below:

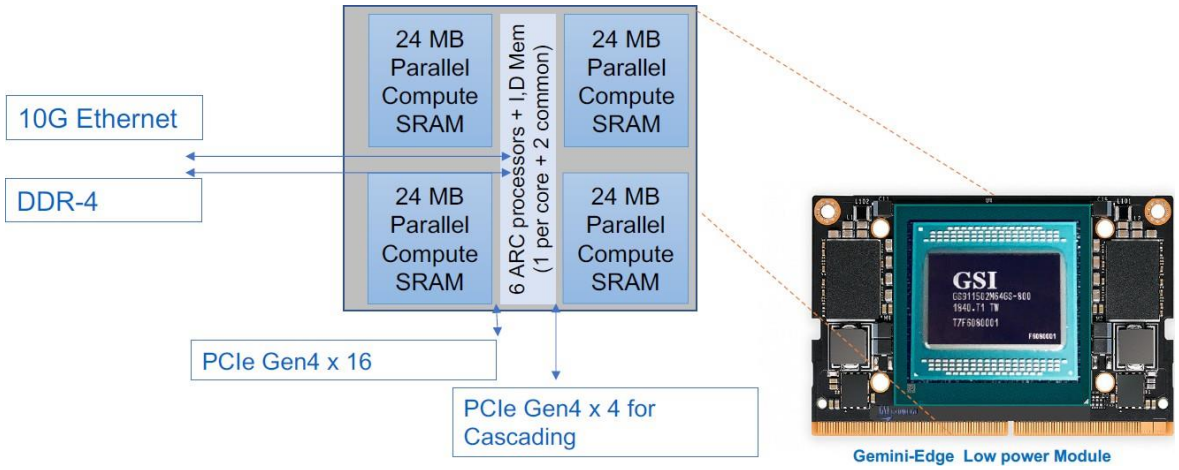
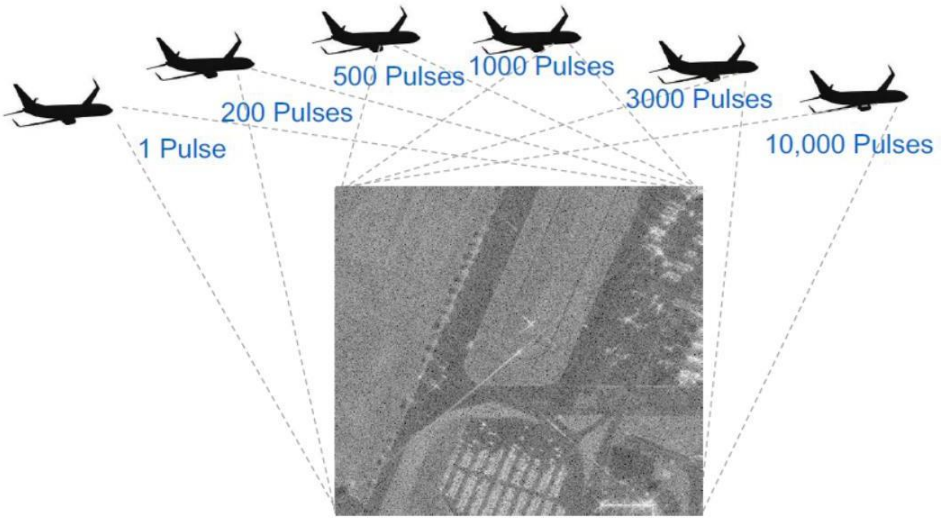


Figure 3: A Notional Gemini-II Low-Power APU Evaluation Module in Definition

The following examples demonstrate the superior computing performance of Associative Compute-In-Memory:

In SAR applications, GSI's APU can execute a back-projection algorithm, simultaneously processing all pixels in parallel for every incoming pulse stored within its memory. The APU architecture is segmented into four compute-in-memory tiles, each capable of storing and processing 384 x 256 = 98,304 pixels, resulting in a combined total of 393,216 pixels per chip.

To put this into perspective, for a 100-million-pixel image, the computation time ranged from 2.6 to 5.5 seconds based on the number of APUs utilized.



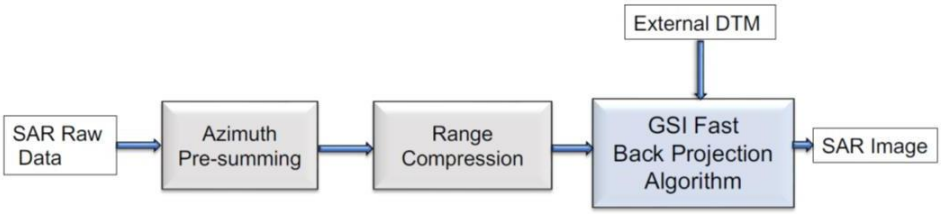


Figure 4: APU Implementation of SAR Back-Projection Algorithm.

In LIDAR applications, an SSIM algorithm has been deployed on an APU alongside a COTS GPU to gauge the similarity between two images. The APU can store the reference image within its internal memory while computing a covariance metric at each point. For instance, it can compare 400 x 400 pixels with a sliding window of 100 x 100 without necessitating additional I/O operations.

The APU showcases a performance improvement of 25 times over a COTS GPU, and nearly 10 times when factoring in both performance and power consumption.

Figure 5: APU Implementation of Image Identification and Classification.

The Gemini-I APU is ideal for change detection applications: the 16 Leda-S server can achieve near real-time image identification and change in a mobile server - something that is not possible with a CPU or a GPU!

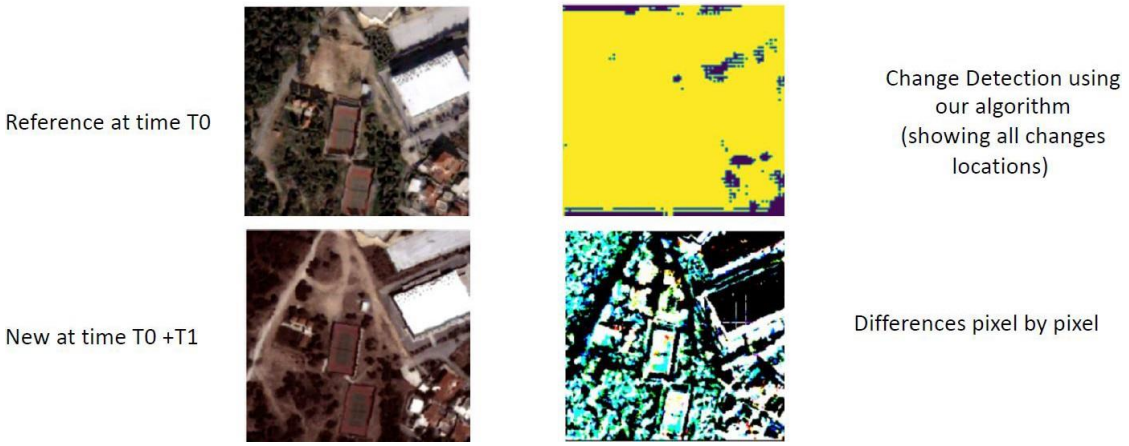


Figure 6: APU Implementation of Change and Trajectory Algorithms.

For cryptography applications, the Gemini-I APU delivers 9 times better performance than a COTS GPU.

GSI's Gemini-I Software offering is as follows:

- Linux Native libraries and tools:
  - Synthetic Aperture Radar
    - We scale SAR BP algorithm across multiple boards

        - Includes Python Jupyter notebook examples
    - DNA Computational Sequencing
        - Accelerates the sequence candidate and filtering stage
    - Fast Vector Search
        - Support both KNN/exact-match and approximate search for billion scale vector databases
        - Includes Python demo applications showing visual search and multi-modal semantic search use cases
    - Cryptographic Library
        - Accelerates brute-force search on SHA-based hash spaces
        - Password cracking utility
- APU Cloud products:
    - Web-service REST-based APIs
    - We can provide:
        - A managed service in GSI's cloud in Santa Clara
        - Installation and management tools to support your own private APU cloud
- Application plugins:
    - BIOVIA Pipeline Pilot plugin for molecule search
- Compiler
    - Custom micro-code development using Python
    - GNU Compiler tool chain for C/C++ compatible programming
    - Emulation tools
    - Self-guided low-programming tutorial

Gemini-II software adds additional AI capability:

- Pytorch 2.0 support (requires SOL license)
- Model support
    - convolutional neural networks
        - MobileNet
        - Resnet
    - vision transformer
        - DEIT

In terms of radiation hardness, Gemini-I has completed preliminary testing with a SEL immunity of 60MeV/cm^2/mg and a TID immunity of 150krad (Si). Gemini-II will be tested for both with higher levels of immunity expected.

The Gemini-I APU is packaged in a 576-pin, 25mm x 25mm PGA while Gemini-II will be offered in a 1296-pin, 37.5 mm x 37.5 mm PGA, as shown below. The part can be ordered in commercial grade now and there are plans to offer a military temperature version.

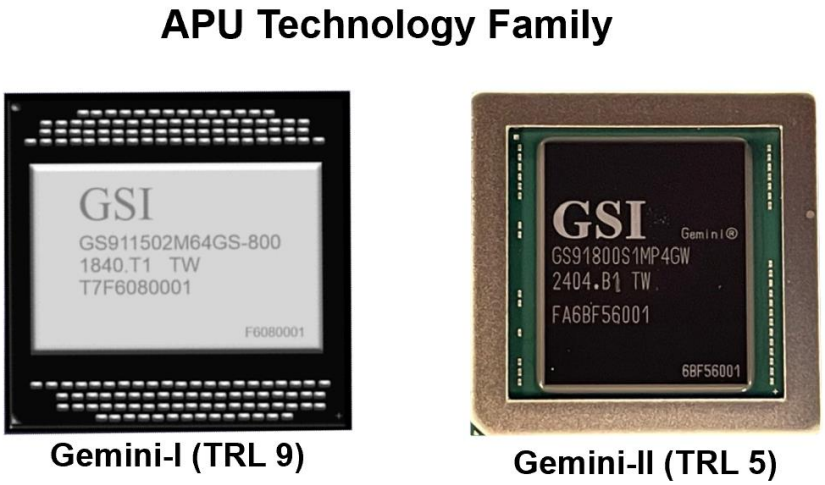


Figure 7: Gemini I and II packaged devices.

Associative Compute-In-Memory is poised to revolutionize on-board processing in space with a road map to offer an extraordinary 420 TOPS at 12 W (Trillion Operations Per Second) of computing performance. This capability paves the way for the next generation of intelligent, autonomous applications *in-orbit*! The potential of Associative Compute-In-Memory promises to redefine the landscape of space exploration and satellite operations, introducing unprecedented efficiency and innovation.

Production boards of the Gemini-I APUs are available now to prototype and de-risk Associative Compute-In-Memory for your applications. Gemini-II APUs will be available late 2024 to evaluate and embed in your platforms, and further information can be viewed at: https://gsitechnology.com/compute/. Spacechips will be developing on-board processing products baselining both configurations of Gemini-II parts.

GSI also allows you to assess their APU hardware remotely to de-risk and implement your application via their server.

Dr. Rajan Bedi is the CEO and founder of Spacechips, which designs and builds a range of advanced, re-configurable, L to K-band, ultra-high throughput SDRs, transponders and on-board processors, AI-enabled, Edge-based OBCs and Mass-Memory Units for telecommunication, Earth-Observation, navigation, 5G, internet, SIGINT and M2M/IoT satellites. Spacechips also offers Space-Electronics Design-Consultancy, Avionics Testing, Technical-Marketing, Business-Intelligence and Training Services. (www.spacechips.co.uk).