

Verdancia Capital Ltd.

Taiwan Stratum Technology Fund

SPV I

Professional Overview

Investing In the Strategic Advantage of a Scarcity-Driven Supply Chain

September 10th, 2026

Strictly Private and Confidential



This Professional Overview is an excerpt from *Verdancia Capital SPV1 Investment Overview*. Access to the full version requires the execution of an NDA.



I. Executive Summary

- Verdancia's initial SPV investment **targets a foundational transition** in the global AI industry: as AI continues its rapid diffusion, demand for inference compute is now entering a phase of structural and sustained long-term expansion. *Inference is what powers real-world applications, it is where income is generated.*
- While training has benefited from years of investment in high-power GPUs, **purpose-built inference hardware is the next durable growth frontier.** It has been estimated that the size of the inference market will be USD 100 billion by 2035.¹
- There has long been a technical issue associated with the hardware dedicated to inference: **memory wall.** This is the bottleneck created when processors perform computations *far faster than data can be supplied from memory*, causing the hardware to sit idle while waiting for data, something that even the best hardware still can't fully address.²
- Verdancia's SPV investment is targeting a Taiwan-originated founder team that has developed a new line of inference processors built with a *novel technology: "wafer-on-wafer"*. This approaches stacks layers of memory and the SoC layer as a single unit, thus *ensuring exceptional inference performance in terms of available memory, accessible bandwidth, and compute throughput.* This technology has been lauded by Qualcomm, who is believed to be developing the same technology.³
- To highlight the importance of *semiconductor supply chain embeddedness*: Wafer-on-wafer is a craft that requires deep collaboration along the conception, engineering, and fabrication processes. By being part of the supply chain fabric, the SPV company gains access to proven manufacturing capabilities and feedback loops, **including scarce foundry and component support to ensure mass production.**
- The proposed size of the investment is USD 50 million, representing 20% of the target company, implying a post-money valuation of USD 250 million. The investment period is projected to be six years, in a 4 + 2 configuration. The expected return profile includes MOIC of 4-5x and IRR of 35-40%.

¹ <https://www.globemarketresearch.com/reports/ai-inference-chip-market>

² <https://wccfttech.com/micron-says-ai-memory-wall-worsening-compute-outruns-hbm-bandwidth-by-3x-every-two-years/>

³ <https://www.jonpeddie.com/news/qualcomm-pushes-compute-next-to-memory/>

- The proposed SPV aligns directly with Verdancia's investment thesis: (i) *innovation rooted in the depth of Taiwan's tech ecosystem, and (ii) strong embeddedness within Taiwan's semiconductor fabrication*. Given the scale of the coming inference requirement, this SPV *offers a compelling opportunity to secure an early, meaningful position in a critical market* — particularly for investors engaged in AI infrastructure.
 - It is reminded that interested parties will invest in a **single-asset fund** structured by Verdancia Capital. As such, the investor is subject to an annual management fee and profit-sharing program with Verdancia upon exit.
-

II. SPV Investment – The Technology, Products, & Capital Call Milestones

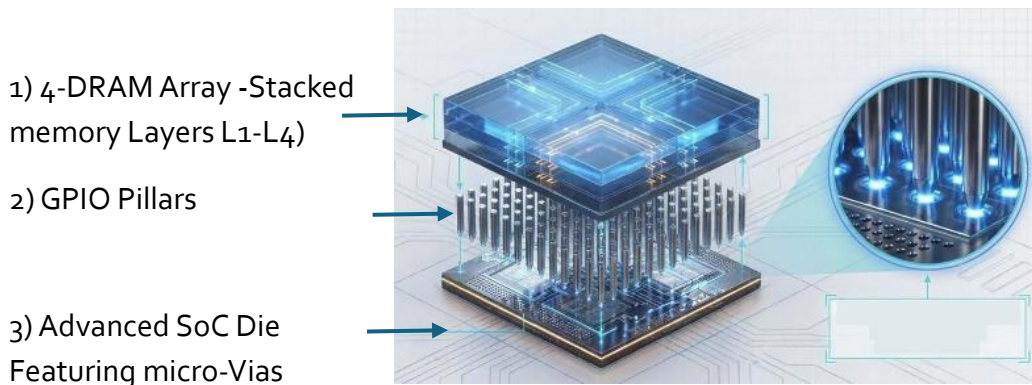
- Since the first LLMs emerged in 2024–25, AI compute has largely centered on training and GPUs. But as AI diffuses rapidly into everyday life, *inference is gaining prominence because it is where AI applications interface with the physical world - inference is where income occurs.*
- When general-purpose GPUs perform inference, they run into the “memory wall”: data cannot move through the system fast enough to keep the GPU fully utilized.
- The SPV company employs a **differentiated “wafer-on-wafer” architecture** that results in materially improved *inference efficiency, lower costs, and significantly reduced power consumption.*
- The Company has developed i) fully integrated rack-level inference solutions built around the new chips and ii) the software interface and compiler tools, enabling immediate compatibility with prevailing industry AI workflows.
- The key advantage of Verdancia’s SPV project company is its **deep supply-chain embeddedness**. This is reflected in i) its **supply-chain level collaboration with leading foundry and memory-wafer partners**, and ii) in **supply chain commitment to provide scarce production capacity and the memory chips required for mass manufacturing.**
- The funding will support production-readiness and mass-production (MP) of the first chip product. It will also support manufacturing of the Company’s indigenous rack-level servers and continuing pre-production development of more powerful chips.

A Novel Technology & Superior Performance

When general-purpose GPUs are tasked with LLM inference, they must move enormous volumes of data through memory token-by-token. Despite offering massive compute throughput — often 989 to 2,250 TFLOPS across NVIDIA’s current generation — their standard memory bandwidth of 4.8–8.0 TB/s is insufficient to feed the compute units at the required rate. The result is a classic memory-wall stall: the GPU’s ALUs sit idle, consuming power, while waiting for weights and KV-cache data to trickle through a bandwidth bottleneck that was never designed for real-time, high-intensity inference workloads.

The industry is not oblivious to these constraints but is limited by both supply-chain (high costs and availability concerns) and engineering realities. Achieving superior inference performance *requires an architecture that balances available memory, memory bandwidth, and physical proximity to the SoC*. Designers can add substantial memory through HBM, yet its bandwidth is ultimately constrained by standardized interfaces to the processor. On the other hand, SRAM offers far greater bandwidth, but its memory capacity is inherently limited by die-area constraints.

Verdancia's SPV target company (the "Company" or "Project") employs a "wafer-on-wafer" architecture. The design stacks four layers of DRAM arrays, interconnected through specialty "pillars" that also *provide high-bandwidth vertical links* rather than relying on standard interconnects. These *memory layers are then bonded directly onto a custom SoC die, forming a unified compute-and-memory structure*. The configuration is shown below in **Illustration I**.



Compared with industry peers, the wafer-on-wafer approach ***delivers both abundant memory and maximized bandwidth simultaneously***.

Relative to the competitors, INFPRO₁, the first within the family of chips, possesses the capability for up to *500 TFLOPS of compute throughput, 64 GB of memory, and 48TB/s of memory bandwidth, which compare significantly favorably to what the competitors can currently offer*. Please also refer to the *Competitive Landscape & Project Valuation* section below for more market dynamics.

To facilitate large-scale adoption of the INFPRO family, the Company has developed fully integrated rack-level inference hardware built around the INFPRO chips, complete with proprietary software interface and compiler tools. Together, the software and the hardware represent a turnkey system that can be readily applied within any existing

data-center or model-serving environment. Datacenters and enterprises can deploy INFPRO-based racks without modifying their existing infrastructure.

"Wafer-on-Wafer" - Advantages of Supply Chain Embeddedness

Wafer-on-wafer is a craft that requires deep collaboration along the conception, engineering, and fabrication processes, something that outside competitors cannot readily access. This approach ensures that engineering feasibility, performance targets, and cost structures are validated early and continuously. The IRPVF chips are a result of this collaboration amongst the Target Company, Taiwan's main memory maker, and a key foundry.

In the case of INFPRO chips, Memory Maker must pre-fabricate the DRAM wafers to precise specifications so they can align and interconnect through nearly one million specialty GPIO pillars, which also serve as high-bandwidth vertical links. Once the memory module is completed, Key Foundry will need to fabricate the SoC and bond it to the stacked memory wafers, a process that likewise requires exact alignment and specification compliance. In other words, "wafer-on-wafer" is custom fabrication taken to a deeper level. Without being situated within the supply chain with longstanding relationships, it is more difficult to achieve the same, intended product specifications.

Then there is the issue of foundry and component scarcity, which industry observers cite as a key bottleneck in the development of the inference market.⁴ Due to longstanding relationships, *both Memory Maker and Key Foundry have committed to providing DRAM wafers and the needed 12nm foundry capacity, respectively, in support of full production* once the Project reaches the production stage. These commitments represent critical assurances for the Company and materially strengthen the Project's viability.

Funding Milestones & Timeline

The Investment shall be funded in three tranches, the first shall be provided upon Closing, with the latter two subject to the achievement of development and pre-production execution milestones. Please note that the timeline for **the capital calls assume a November 15th, 2026, Closing**.

⁴ <https://www.businesstimes.com.sg/international/global/whats-stopping-asias-chipmakers-taking-nvidia-answer-lies-bottleneck>

Capital-call timeline is:

- 1) Tranche I: Initial funding of USD 15 million will be provided upon Closing.
- 2) Tranche II: Capital call for the second tranche of funding, in the amount of USD 20 million, is scheduled between **February 1st and March 31st, 2027**.
- 3) Tranche III: For USD 15 million, capital call is scheduled between **June 1st and July 31st, 2027**.

Details pertaining to milestones will be provided in the expanded Investment Overview.

III. Competitive Landscape for Inference Chips

- There have been other efforts to develop inference- and decode-specific compute, and several teams have made meaningful progress. Most of these solutions remain in various stages of testing or early production. The transition toward massive, dedicated inference and decode compute is therefore still in its early phases.
- Relative to competing solutions, the INFPRO1 demonstrates superior performance and hardware metrics across compute throughput, memory capacity, and available bandwidth.
- Apart from a few cases where Korean foundries have been used for chip fabrication, most competing products rely on TSMC and its associated supply chain. In contrast, the INFPRO product line benefits from deep, material collaboration with key supply-chain partners, giving it a structural advantage in both development and production.
- The committed foundry capacity from Key Foundry and the dedicated memory-wafer support from Memory Maker position the INFPRO chips to scale rapidly.
- A key consideration in Verdancia's assessment of the SPV Company's valuation is superior performance. Although the INFPRO products will require another year before entering mass production, their advantages in dedicated performance enhancement and supply-chain support should compound meaningfully once deployed in service.

Competitive Landscape

The emergence of purpose-built inference chips followed naturally after the development of GPUs optimized for training. Yet the computational intensity of training has driven hyperscalers to reserve vast amounts of advanced packaging, HBM supply, and foundry allocation. GPUs have effectively monopolized critical supply-chain resources, leaving inference-specific architectures in earlier stages production and deployment. This dynamic underscores why dedicated inference and decode compute remains a second wave, and still early in its commercial adoption.

This second wave—centered on purpose-built inference and decode compute—*may prove more durable than the first. Training is a prerequisite for AI, but **it is inference that generates recurring revenue, drives unit economics, and determines real-world adoption.***

As a result, we expect a growing number of new entrants in the near term, each aiming to capture a share of the rapidly expanding inference market.

Yet, the market faces inherent bottlenecks that could slow, even derail, progress by some of these new entrants: supply-chain and component constraints. Memory components shall be in short supply, some speculate for years, and foundry capacity for advanced node fabrication is an even more scarce resource.

The following tables provide a list of competitors engaged in inference chip development. The first table compares performance metrics while the second table outlines status, characteristics, and valuation. The list of names is not exhaustive and leaves out some ongoing projects. One such company is Japan's ***Preferred Networks ("PFN")***, whose once-touted MN-Core L1000 is awaiting additional subsidies to proceed. ***Qualcomm*** is understood to be developing an inference chip by utilizing the Taiwan supply chain. The list also does not include efforts coming out of China, where a parallel tech ecosystem is now taking place.

Table 1:

Company (country)	Chip Name/Architecture	Peak Compute Throughput	Memory Bandwidth	On-chip Memory Capacity
INFPRO (Taiwan)	INFPRO1 / 2 / 4 (3D- DRAM ASIC)	1,000 – 4,000 TFLOPS	48.0 – 192.0 TB/s	64 – 256 GB 3D-DRAM
Etched(U.S.)	Soho(Trans-former)	~2,000 TFLOPS	~5.0 TB/s	144 GB HBM3e
FuriosaAI (Kor)	RNGD (SRAM + LPDDR5X)	512 TOPS	~1.5 TB/s	512MB SRAM + 48GB LPDDR
Rebellions (Kor)	Rebel (LLM NPU / HBM3)	~1,000 TFLOPS	~2.0 – 3.2 TB/s	32 – 64 GB HBM3
Tenstorrent (U.S.)	Wormhole (RISC-V + Tensix)	~1,000 TOPS (INT8)	~1.6 TB/s (GDDR6)	32 – 64 GB GDDR6
d-Matrix (U.S.)	Jayhawk (In-compute memory DIMC)	~1,000 TFLOPS (MXFP6)	High on-chip bandwidth	~100s MB SRAM + external
DeepX (Kor)	DX-M1 / V3 (低功耗 NPU)	~200 – 500 TOPS	~512 GB/s	16 – 32 GB LPDDR5
Hailo (Israel)	Hailo-10 / 15 (數據流 NPU)	~40 – 100 TOPS	Low on-chip bandwidth	8 – 16 GB LPDDR4/5

The INFPRO line of chips clearly delivers superior inference compute, abundant on-package memory, and the highest bandwidths among emerging entrants. Based on simulation results and performance verification, the INFPRO₁ executes text and voice inference *workloads at roughly one-tenth the power consumption while achieving six- to eight-fold efficiency gains*. This “best-of-both-worlds” performance—combining ample memory with exceptional bandwidth—is a *direct outcome of its wafer-on-wafer engineering approach*.

Table 2:

Company (country)	Est. Valuation	Characteristics	Stage of Development
INFPRO (Taiwan)	USD 200- 250 mio	48–192 TB/s ultra-high memory bandwidth, delivering second-level low-batch decoding with large on-card memory, best TCO.”	Entering Final Tape-out and Mass Production.
Etched	~ USD 10 Billion	TSMC N4P 4nm, SK Hylix/Micron 144GB HBM3e, 5.0TB/s bandwidth only	Returned-Wafers, early- stage Rack-Scale verification, yet to enter mass-production(MP)
FuriosaAI (Kor)	~USD 1.0- 1.5 billion	Pairing 512MB SRAM with LPDDR; 1.5 TB/s bandwidth only.	2-Gen RNGD chips now in engineering sample (ES) stage; in application testing stage with LG AI Research (Exaone) & local Naver Cloud.
Rebellions (Kor)	~USD 2.3 billion	Funded by Korea Telecom (KT) & Samsung; ample local resource; employs HBM but did not address memory wall.	1st-gen ATOM chips deployed in KT data centers; next-gen Rebel chip in ES stage.
Tenstorrent (U.S.)	~USD 3.2 billion	Led by Jim Keller; employs open-source RISC-V structure; GDDR6 still constrained by limited bandwidth.	Wormhole chips entered production; Blackhole chips have been produced and being tested; performance being verified.

d-Matrix (U.S.)	~USD 2 billion (Private)	In-compute memory design with better efficiency; relies on aggressive quantization (INT8/MXFP), large models still require multi-card chaining	In-compute memory Chiplet (DIMC) delivered to Early Access; data-center testing with Microsoft.
DeepX (Kor)	~USD 1 billion (Private)	Designed for low-power vision and lightweight inference workloads; compute and bandwidth insufficient for server-grade LLM deployment.	
Hailo (Israel)	~USD 1.2 billion (Private)	Edge/auto-specific AI chips; not for large scale applications.	

The key takeaway from Table 2 is the development stage of the various entrants. Most have completed tape-out but remain in the ES (engineering-sample) or early-testing phase, with only the ATOM chips confirmed to be in active deployment. On this basis, the INFPRO1 is approximately 1 to 1.5 years behind. However, the INFPRO program retains two decisive advantages: performance and supply-chain stability. The INFPRO architecture is clearly driven toward enhanced performance, with superior memory allocation and a more advanced design philosophy. Equally important, the project is deeply embedded within the Taiwan supply chain, ensuring that company growth is not constrained by supply-side limitations and positioning the INFPRO line to scale reliably once in production.

Valuation estimates in Table 2 are telling. Companies with the capability to develop and deliver products in this new arena of inference specialization are being rewarded with a premium for staking out an early place. The INFPRO chips have the potential to out-compete vs. their peers.