

Korea × Europe AI Semiconductor Alliance

From the AI Memory Wall to Post-von-Neumann Silicon

A public 5-minute pitch for consortium building

Prof. Tae-Wan Kim
Seoul National University
Founder & CEO, LLM Core AI, Inc.
taewan.kim@llmcore.ai | taewan@snu.ac.kr

KERC · Horizon Europe Cluster 4 · Brussels · 15–16 September 2026



The bottleneck has moved: from FLOPs to memory

Frontier AI is no longer only a compute problem. It is a memory-movement, bandwidth, and energy problem.

Long context

KV cache grows

Memory traffic

Energy wall

Symptom

Consequence

Korea–Europe opportunity

KV cache explosion

Bandwidth + latency dominate long-context inference

Compression-first architecture

Data movement energy

Moving bits can cost more than computing them

PIM/CIM + memory-centric silicon

AI chip dependency

GPU stacks concentrate value and control

Trusted sustainable AI compute alliance

A third strategic pillar for AI semiconductors

In a world shaped by U.S.–China technology rivalry, Korea and Europe can build a trusted, sustainable and impact-oriented alternative.

U.S.
Platform & GPU dominance

- Hyperscale AI clouds
- Mature software stacks
- Capital concentration

China
Sovereignty push

- Full-stack national investment
- Local ecosystem scaling
- Strategic independence drive

Korea × Europe
The alliance path

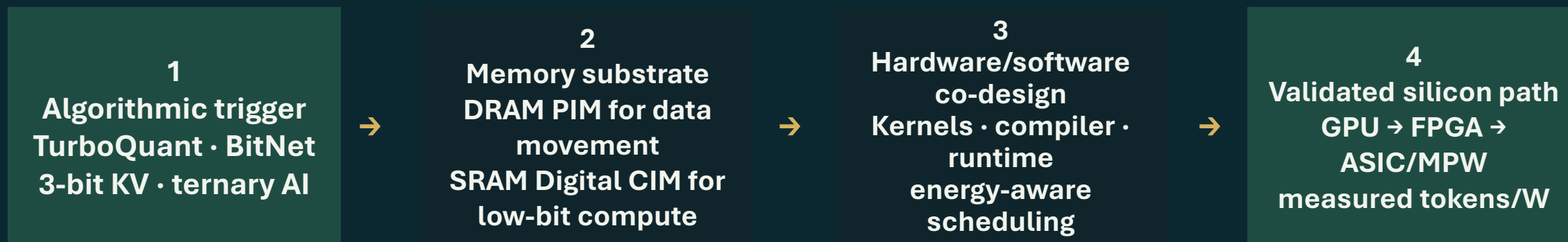
- Memory + semiconductor speed
- Deep science + trusted AI
- Industrial validation + open ecosystems

**The vision is not to imitate. It is to lead where the next bottleneck is emerging:
memory-centric, post-von-Neumann AI.**

Europe's research depth × Korea's memory and manufacturing speed = a credible third pillar for sustainable AI
silicon

Post-von-Neumann AI: start with the algorithm, end in silicon

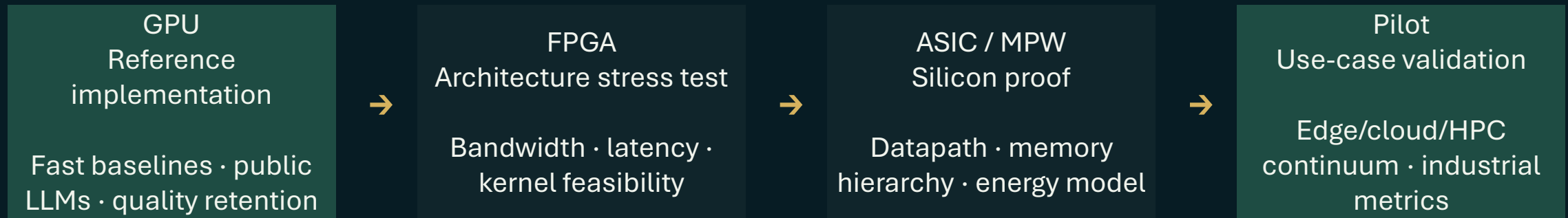
Algorithmic compression becomes most powerful when the memory system and chip architecture are co-designed for it.



Technology lever	Best hardware direction	Why it matters
TurboQuant / KV-cache compression	DRAM-side PIM and memory-traffic reduction	Long-context AI becomes practical when memory traffic shrinks
BitNet / ternary inference	SRAM-based Digital CIM and low-bit datapaths	Compute becomes add/subtract/accumulate friendly
Runtime + metrics	GPU/FPGA/ASIC validation loop	Speed: measure, iterate, tape out, validate

The Game of Speed: verify before the world catches up

AI algorithms change faster than chips. The only winning method is a fast validation loop from public benchmarks to silicon proof.



Success metrics

tokens / watt

latency

bandwidth

memory footprint

quality retention

reproducibility

The project is not just another AI model. It is an AI-semiconductor execution engine.

What SNU + LLM Core AI bring to the consortium

One team bridging AI algorithms, hardware architecture, and semiconductor validation.

Efficient AI algorithms

TurboQuant · 3-bit KV cache · BitNet/ternary inference · long-context LLMs

Hardware/software co-design

HW-aware kernels · runtime mapping · memory hierarchy · heterogeneous computing

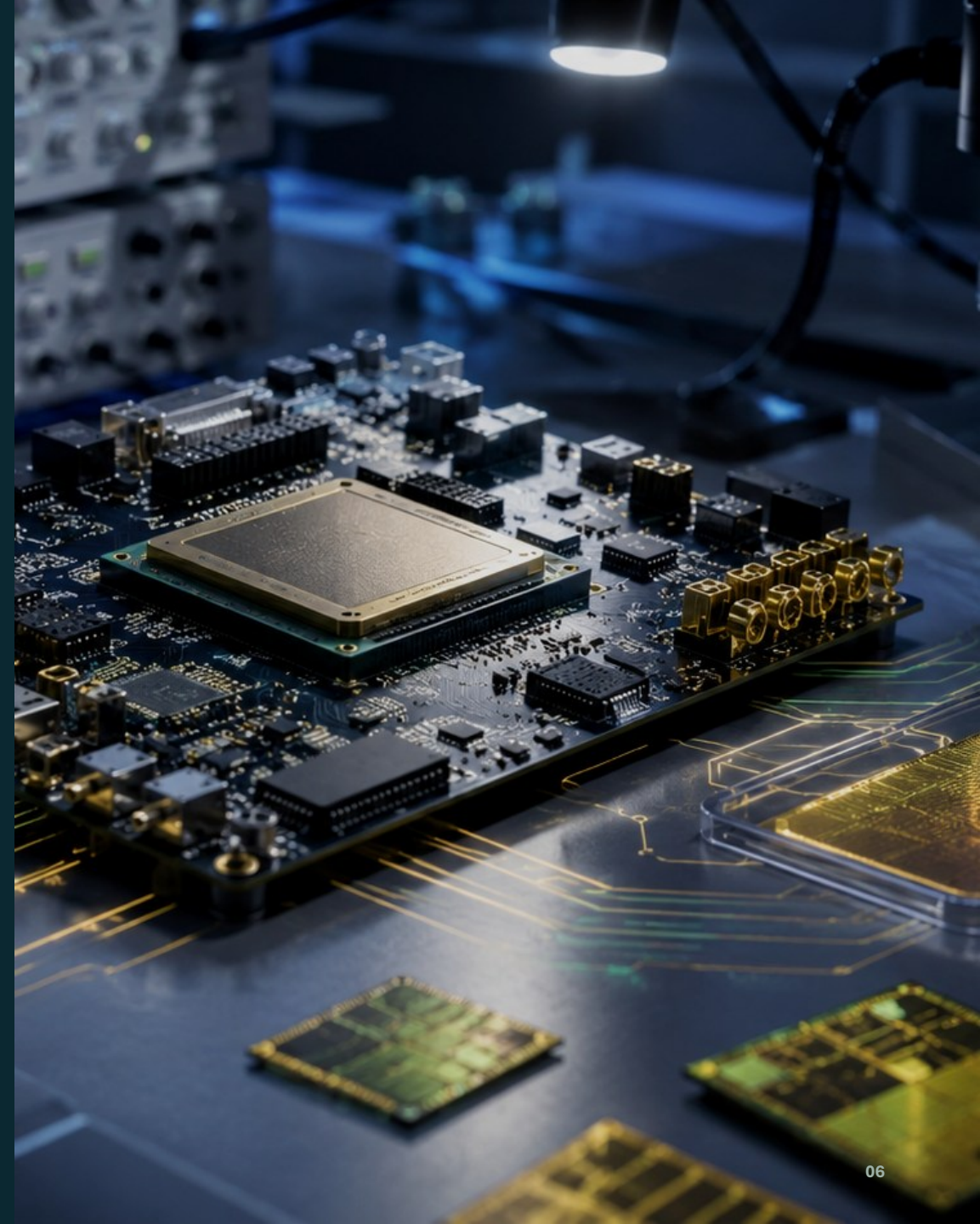
Prototype validation path

GPU baselines · FPGA validation · ASIC/MPW-oriented architecture · energy metrics

Korea bridge

Memory semiconductor ecosystem · rapid industrial validation · Korea–Europe R&D links

Prof. Tae-Wan Kim · Seoul National University · Founder & CEO, LLM Core AI, Inc.



Best Horizon Europe fit: compute continuum + sustainable AI

Primary target: HORIZON-CL4-2027-04-DATA-03

New approaches for decentralized, federated and sustainable AI data processing · RIA

Call requirement	Our fit	Potential consortium partner
Distributed edge / cloud / HPC compute continuum	Runtime and hardware mapping for heterogeneous AI inference	AI/HPC centre · cloud/edge platform
Model compression and diverse hardware architectures	TurboQuant, BitNet, FPGA/ASIC validation, PIM/CIM direction	Semiconductor RTO · FPGA/ASIC partner
Energy and sustainability measurement	tokens/W, bandwidth, latency, memory footprint, quality retention	Sustainability / data-centre / measurement partner
Real-world use cases in at least two domains	Public benchmarks plus industrial validation scenarios	Manufacturing · mobility · healthcare · edge AI users

Indicative framing from the Work Programme
RIA · around €17.5M EU contribution per project · TRL 3 → TRL 6–7 · total topic budget €35M

Source: European Commission, Horizon Europe Work Programme 2026–2027, Cluster 4, “Digital, Industry and Space” (downloaded 2026).

Let us build the Korea–Europe AI silicon alliance

Wanted: complementary European partners who can turn a technology thesis into a winning Horizon Europe proposal.

Partner role	What we want to build together
Coordinator / proposal leader	Competitive proposal structure, governance, exploitation, EU impact logic
AI/HPC / cloud-edge infrastructure	Compute-continuum testbed for long-context AI and distributed inference
Semiconductor / FPGA / ASIC RTO	PIM/CIM-oriented architecture, FPGA validation, silicon proof path
Industrial validation partner	Manufacturing, mobility, edge AI, data-centre or healthcare use cases
Sustainability / energy expert	Transparent measurement of tokens/W, carbon, energy, bandwidth and latency

Next 90 days

1-page concept

WP map

Partner gaps

Benchmark PoC

Proposal sprint

From the memory wall to trusted, sustainable, post-von-Neumann AI silicon.