ESSERC 2026 · KEYNOTE · 52ND IEEE EUROPEAN SOLID-STATE ELECTRONICS RESEARCH CONFERENCE

openchip

ENABLING THE DIGITAL WORLD

Design Space Exploration or Design Space Denial?

The case for rethinking chip-to-system co-design in the agentic era

Cesc Guim · CEO, Openchip

Palma · Thursday 10 September · Sala Magna · 09:40–10:20

openchip

Three things to get right.
Four fault lines in the way.

RESOURCES | INTELLIGENCE | PROGRESS

What humanity has to get right to build AI on human terms

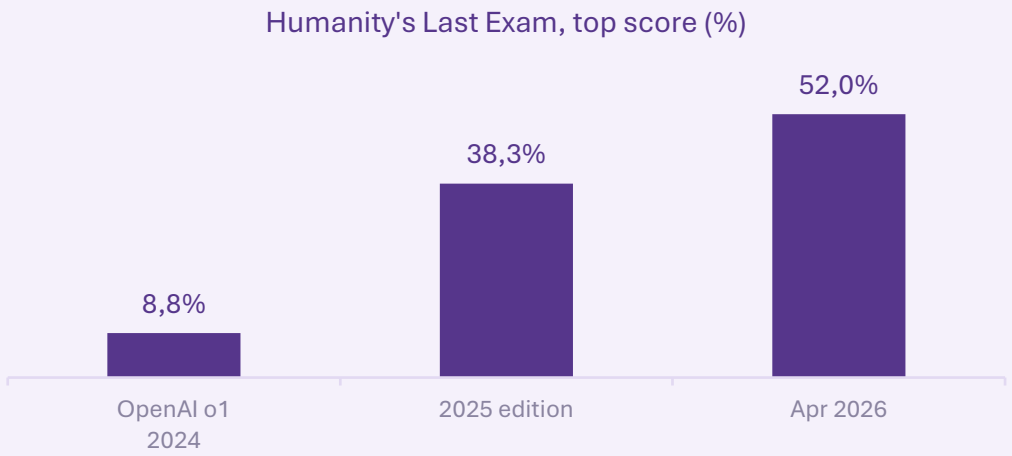
01	**Scale & demand**	Energy, compute, and the power wall
02	**Intelligence / watt**	General capability vs. specialization
03	**Trust**	Reliability, privacy, hallucination at scale
04	**Context**	AI's real currency — and its real cost

Part one — the facts

What changed in twelve months

The facts moved faster than most of our assumptions did.

The benchmarks moved fast. What they measure didn't keep up.



60→100%

SWE-bench Verified: real, unresolved bugs. Solved in a year.

30×

AI compute growth worldwide since 2021.

4 in 5

university students now use generative AI weekly.

The catch: every number above measures raw capability or raw scale. None measures whether an agent took the right real-world action, or whether the extra parameters were necessary to deliver it. Benchmarks still reward bigger — not more intelligent.

And the pace hasn't slowed — releases from the weeks before this talk:

Kimi K3
2.8T total, ~50B active · Aug 2026

DeepSeek-V4-Pro
1.6T total, 49B active · Aug 2026

Qwen3.8-Max
2.4T — largest open-weight yet

Sources: Stanford HAI AI Index 2026; Epoch AI; Local AI Zone, "Latest AI Developments: August 2026".

Usage is scaling faster than efficiency is improving.

This is becoming real usage

70%+

of organisations now use generative AI in at least one core business function.

$150B+

in a single hyperscaler's 2025 capex — cloud providers now build AI infrastructure like utilities build power plants.

#1 skill

“Agentic AI” is now the fastest-growing AI skill cluster in the global labour market, as “Chatbot” demand falls.

But models aren't efficient enough yet

565 TWh

global data-centre electricity demand in 2026 — up 26% in a single year.

84%/yr

growth rate of AI-optimised server power draw — will overtake all conventional servers by 2027.

~50%

of planned US data-centre builds are now delayed or cancelled. Texas moved fastest and absorbed a third of the tracked US buildout — yet even there, only 49 of 84 tracked facilities are actually operational today.

“More GPUs” stopped being a strategy the day it became a grid-interconnection queue.

Sources: Stanford HAI AI Index 2026; Lightcast contribution to AI Index 2026; Gartner (Jun 2026); IEA, “Key Questions on Energy and AI” (Apr 2026); LBNL; Silicon Report, “The AI data center power crunch” (Jul 2026).

Most of a “frontier” model never touches your query.

~9–15%

of parameters actually activate per token in today's leading MoE models — a 200B-class model may fire as few as 30B.

≤40%

is the figure most often cited across recent papers and podcasts for how much of a foundational LLM activates for a specific use case — consistent with the harder MoE numbers alongside it.

10–30×

cost savings NVIDIA Research reports from routing agentic workloads to small, specialised models instead of a frontier LLM by default.

The paper making this case formally: ***“Small Language Models are the Future of Agentic AI”*** (NVIDIA Research, arXiv 2506.02153). Its argument isn't just that SLMs are cheaper — it's that most agent work is a small, repetitive loop of specialised sub-tasks, which is exactly what a smaller, specialised model is suited for. A separate 2026 study (Cao et al.) proposing a Performance-Efficiency Ratio metric found 0.5–3B models achieve superior efficiency across every NLP task they tested.

Sources: futureagi.com SLM/LLM 2026 comparison (Gemma 4, Qwen3.5 activation ratios); Belcak et al., “Small Language Models are the Future of Agentic AI”, NVIDIA Research, arXiv 2506.02153; Cao et al. (2026), Performance-Efficiency Ratio.

openchip

Four fault lines nobody has fully closed.

01

RESOURCES

Scale & demand

02

PROGRESS

Intelligence / watt

03

TRUST

Reliability & privacy

04

VALUE

Context — the real currency

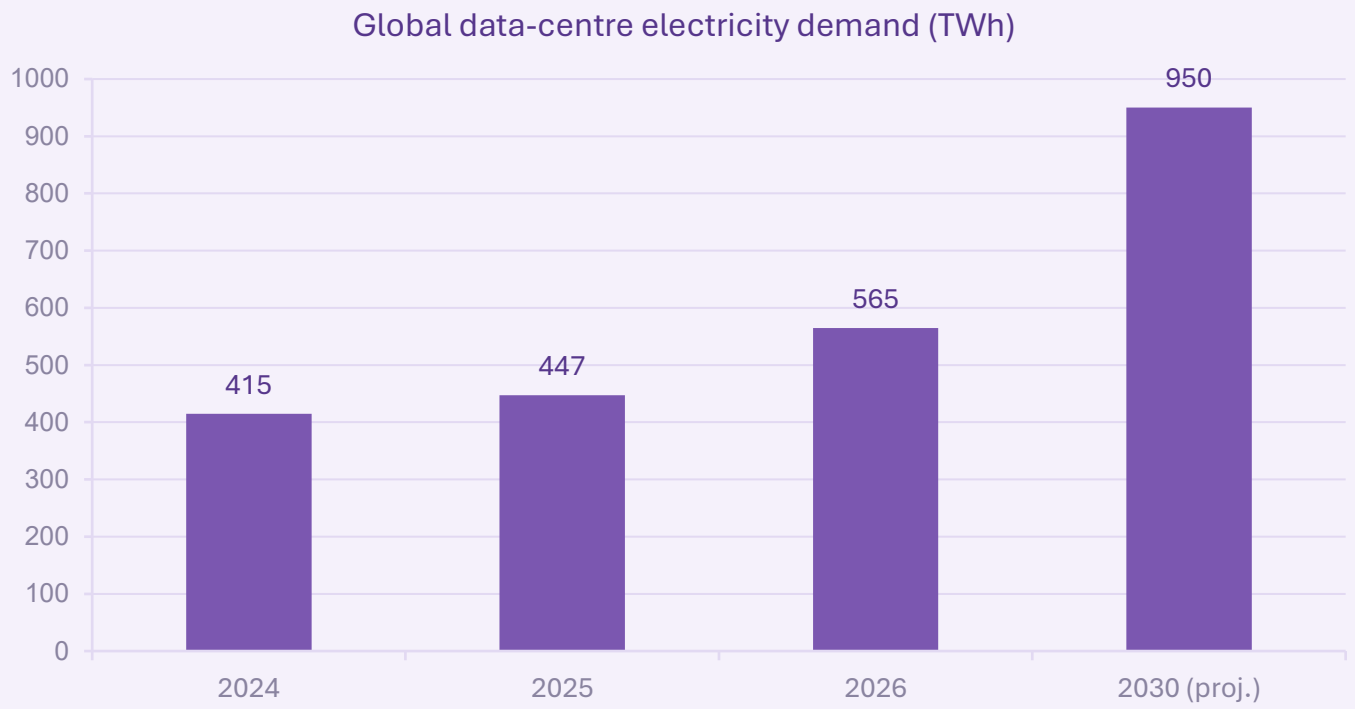
Four topics. Each one: the honest problem, then what we're actually building at the chip and system level to answer it.

Fault line 01 — resources

Scale & the power wall

“More GPUs” is not a strategy. It's a symptom.

Demand is outrunning the grid, not the fabs.



Ireland

21% of national electricity (2022) → 32% forecast by 2026

Virginia, US

some grid zones already exceed 39% AI-driven load

Hyperscalers

~10 GW of nuclear / SMR capacity committed since 2023 to skip the interconnection queue

Sources: IEA, "Key Questions on Energy and AI" (Apr 2026); Gartner (Jun 2026); LBNL / US DOE.

openchip

One coupled system — and the metric that flipped underneath it.

Monolithic die → **Chiplets (UCIe 2.0)** → **Packaged (CoWoS + HBM4)** → **Optical I/O**

HBM4 + optics, coupled to compute

0.8V 64GB HBM4 stacks, CoWoS Gen 6, and on-package optical I/O move together — packaging capacity is the bottleneck now.

Cooling is a floorplan constraint

Direct-to-silicon microfluidic cooling and 25µm EMIB-T bumps are load-bearing decisions made alongside compute.

$20B. NVIDIA absorbed Groq's core chip tech in 2026 — proof that real architectural diversity (Cerebras, Tenstorrent, hyperscaler silicon) still doesn't buy durable independence without open interfaces.

Vertical power delivery

Sub-1V core supplies fed from 48V+ rails through micro-ohm interconnect — see yesterday's power-management keynote for the full story.

Reliability and RAS, not just power

Die-to-die interfaces need their own DFT strategy — silent data corruption from test escapes is now a named, published risk, not a rare edge case.

Data movement, co-designed

Moving the right bytes to the right place needs the memory controller, interconnect and scheduler designed together as one hardware–software problem — placement decided in software, enforced in silicon, not a driver-level afterthought.

Sources: PatSnap chiplet/HBM4 outlook; Sacra, Tenstorrent analysis; Silicon Report, "Groq, Cerebras, and Tenstorrent" (Jul 2026); Mitra et al., "Silent data corruption by 10× test escapes".

Model and system co-design have to move together — on an open floor.

Co-design only works if the model side is open, too

Chip and system design can adapt to real workloads only if the models (LLM, SLM, and everything specialised in between) are open enough to co-design against — open weights and open foundations, not a black box on the other side of an API.

That takes academia and industry actually side by side

Not academia publishing papers industry ignores, and not industry building silicon academia never gets to test against real architectures on. Genuine co-design needs both in the same room, on the same hardware, early.

The honest starting point: academia operates at a different scale than industry

A university lab's compute budget looks nothing like a hyperscaler's, and that gap is widening, not closing. But scale is exactly what shared infrastructure and open consortia are built to offset — the right response isn't closing the gap alone, it's building the bridges below, early and often.

Real answers are forming, not just the problem: national compute resources (the US NAIRR initiative), cloud research-credit programmes tied to open publication, and consortia — the Trillion Parameter Consortium, the Open Compute Project — built to bring industry, academia and national labs onto shared infrastructure. None of it is sufficient yet; all of it has to scale together with the systems themselves.

Sources: arXiv 2509.25397 (EleutherAI quote); arXiv 2603.05225, "AI+HW 2035" (§7); NAIRR Pilot (US); Trillion Parameter Consortium; Open Compute Project, OCP Barcelona 2026.

Fault line 02 — progress

General intelligence vs. intelligence / watt

We didn't just find a new scaling law. We found an axis we'd been ignoring: watts.

We bet the industry on one scaling law.

77%

of real-world AI queries are practical, everyday tasks that never needed a frontier-scale model in the first place.

1,300×

year-over-year growth in token volume — even the vendors selling frontier compute are now pricing a spectrum, not one model.

Even NVIDIA now maps its own inference business as a deliberate ladder — free, mid, premium — rather than a single frontier product.

The largest context windows sit at the top of that ladder, priced highest — not treated as the only tier that matters.

If the biggest GPU vendor on Earth is hedging on "one model to rule them all," maybe we should too.

Sources: intelligence-per-watt.ai research programme (Stanford / Together AI, 2025–26); NVIDIA Developer Blog, GTC 2026.

There's no Intelligence-per-Watt metric yet. That's the gap.

The closest attempt — Saad-Falcon et al., Stanford / Together AI — measures generic benchmark accuracy per watt. Real progress; still not the metric this room needs.

1 — Intelligent action taken

classified per segment — healthcare, finance, logistics...

→

2 — Outcome measured

did it actually achieve the goal, end to end?

↓

3 — Resources counted

energy & carbon — at the system and datacentre level

←

4 — Feedback signal

adjusts the model, the routing, the next action

(loops to 1)

Some future model may eventually score its own intelligence-per-watt, per action, per segment. We're not there — today this is a design target, not a dashboard.

Two things this needs that don't exist as products yet: classification per segment — healthcare, finance and logistics define a 'good' intelligent action completely differently — and agentic-level monitoring wired into a live feedback loop, not a leaderboard number measured once and left alone.

Source: Saad-Falcon et al., "Intelligence per Watt: Measuring Intelligence Efficiency of Local AI," Stanford / Together AI, 2025–26 (arXiv 2511.07885).

Design for the spectrum, not the peak.

Sparse, routed compute

Mixture-of-experts activates a fraction of parameters per token — hardware that does expert-parallel routing well turns “bigger model” into “cheaper token,” not just “slower token.”

Precision as a first-class knob

FP4 / FP8 inference paths (e.g. NVFP4-class formats) trade precision the model doesn't need for watts it does.

RISC-V as the orchestration layer

Inside modern accelerators, RISC-V increasingly isn't the matmul engine — it's the open, customizable control plane that schedules tensor units without per-chip licensing tax.

And, honestly — more still needed

Sparsity, precision and open orchestration make today's architecture more efficient. None of the three, alone or combined, is guaranteed to be enough for the next real leap in capability. That's still an open research question, not a solved one.

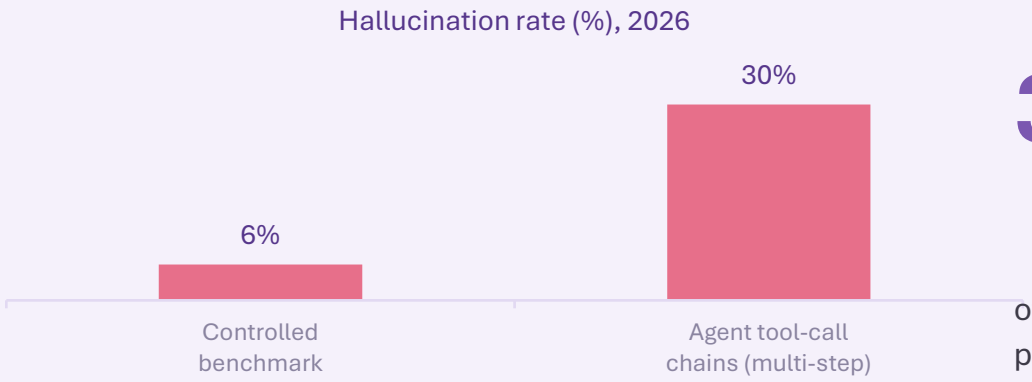
We said it plainly a year ago: the shift is from monolithic models to distributed, agentic systems — it's not about scaling bigger, it's about scaling smarter.

Fault line 03 — trust

Reliability, privacy & the cost of being wrong

We spent two years making the answer more accurate. We only just started asking whether the agent took the right action.

“Solved” on the benchmark. Not in the workflow.



34%

of enterprises report a customer-facing incident from an AI hallucination in the past year; regulated-industry fixes average $50K+ each.

5–6% on narrow benchmarks — real progress — but 20–40% of tool-call chains still involve a hallucinated step.

The regulatory floor just moved, too. 2 August 2026: EU AI Act transparency duties (Art. 50) and general-purpose-model obligations took effect on schedule. The toughest tier — full high-risk system obligations — was pushed to December 2027 by the Digital Omnibus, adopted this June. The rules are catching up to yesterday's AI while agentic systems are already three moves ahead.

Sources: arXiv 2605.17062 (2026 frontier hallucination re-evaluation); Deepchecks LLM Evaluation Benchmarks 2026; Arthur AI; EU Digital Omnibus on AI (Jun–Jul 2026).

Trust has to run silicon to system — not bolted on at either end.

Silicon: open, vendor-neutral roots of trust

Caliptra (Microsoft, Google, AMD, and the Open Compute Project) and OpenTitan (Google, now shipping in production hardware) are open-source silicon root-of-trust specs built for exactly this — CPUs, GPUs, DPUs and TPUs attested the same way, regardless of who made the chip.

System: agentic infrastructure needs it more than chat ever did

When agents from different owners share the same infrastructure, hardware attestation — not policy documents — is what lets them cooperate without extending blind trust up and down the stack.

The gap in between: attesting the model, not just the data

Confidential computing today mostly protects data in use. What's still missing is AI-trusted confidential computing — proving which model actually ran, unmodified, on verified silicon, and carrying that proof all the way up to the system consuming it. That's the kind of problem active IP work is targeting right now, including inside our own portfolio.

And the single biggest lever on reliability isn't a bigger model — it's better-grounded context. Which brings us to fault line four.

Sources: Caliptra (chipsalliance/OCP); OpenTitan (Google / lowRISC); "When Agents Handle Secrets" survey (May 2026); "Open Problems in Technical AI Governance" (2024–26); Openchip in-flight IP, hardware AI-model attestation.

Fault line 04 — value

Context is AI's real currency

It's the single biggest lever on quality we have — and our hardware still treats it as an afterthought.

openchip

Context grew 5,000× in five years. The cost grew with it.

70–80%

of total decode latency at a 64K-token context is spent on attention alone — not the rest of the model.

87%

drop in hallucination rate with well-structured retrieval versus none — context is the cheapest reliability fix we have.

More context an agent accumulates means more that has to be watched, not less:

Observability has to scale with context

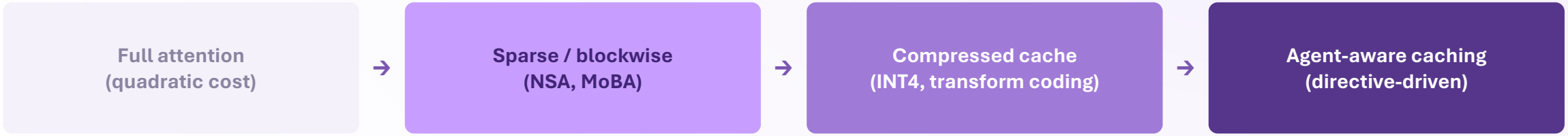
The more an agent retrieves and remembers, the more essential it is to monitor, in real time, what actually got retrieved and why — not just whether the answer looked right.

Explainability and attestation, not just logs

Being able to explain a decision and cryptographically attest to what informed it — after the fact isn't enough for agentic stacks acting autonomously at scale.

Sources: Yuan et al., "Native Sparse Attention", DeepSeek-AI (arXiv 2502.11089); Atlan, RAG grounding analysis (2026); public model release notes.

Attention stopped being “free” to make it fast again.



Sparse attention, natively trained

DeepSeek's Native Sparse Attention and Mixture-of-Block Attention (2025) reach full-attention quality while cutting 64K-token latency sharply — already shipping in production models like DeepSeek-V3.2.

KV-cache compression, not just eviction

INT8 KV-cache is already the shipping default in production inference stacks; INT4 is now available and cuts cache memory to roughly a third of FP16. ICLR 2026 work borrows video-codec-style transform coding to compress it further still.

Caching designed for agents, not chat turns

New systems (2026) treat the KV cache as a position-independent, directive-driven resource built for long-running tool-using agents — not single back-and-forth conversations.

Sources: arXiv 2502.11089 (NSA), 2502.13189 (MoBA); OpenVINO 2026.2 release notes; ICLR 2026 “KV Cache Transform Coding” (NVIDIA & U. Warsaw); arXiv 2606.01065 (Leyline).

The next fight isn't per-chip. It's per data centre.

Memory pooling at rack scale

CXL 4.0 (Nov 2025) doubles bandwidth to 128 GT/s and pools 100+ TB of cache-coherent memory across racks — memory finally scales independently of compute.

A shared, reusable context tier

NVIDIA's new context-memory storage tier and Marvell's rack-scale CXL and optical shared-memory platforms (announced in early August 2026) let multiple servers reuse one warm KV-cache pool instead of every node paying to rebuild it.

The new scoreboard: tokens per dollar

Not tokens per second. As context windows and agent fleets grow, the economics of AI infrastructure are being rewritten around memory efficiency, not raw FLOPs.

Sources: CXL Consortium, CXL 4.0 (Nov 2025) & FMS 2026 recap; Marvell, AI Memory Infrastructure Portfolio (4 Aug 2026); arXiv 2606.12556 (ITME).

Part three

Where this goes in four to five years

Not a bigger model. A better-designed system around it.

Four shifts, one design philosophy.

FLOPS	One model, monolithic	Context per server	Tools you call
↓	↓	↓	↓
Intelligence / watt	**Specialists, orchestrated**	**Context pooled & shared, data-centre-wide**	**Agent hooks the system is built around**

This isn't hypothetical — it just happened at DAC 2026.

Synopsys, Cadence and Siemens all shipped fully autonomous EDA agents this year — a verification agent claiming up to 50× faster time-to-validated RTL, autonomous debug closure cutting cycle time ~40%, and the first EDA workflow to reach “Level 5” autonomy. Chip design is now one of the clearest places agentic AI is already changing how the work gets done — not someday, this summer.

Sources: Synopsys, Cadence, Siemens & NVIDIA, 2026 DAC Chips to Systems Conference (Jul 2026); Synopsys–AMD–Microsoft announcement (27 Jul 2026).

openchip

Intelligence per watt — scored the way a brain would score it.

Parameters, bandwidth and peak FLOPS describe how big a system is, not what it's worth. The brain is the best working proof intelligence and power draw don't have to move together: it runs on ~20 watts because its memory is tiered, not because it's small.

~20 W runs the entire human brain, all at once.

Working / active
GPU HBM — fast, expensive, small

Short-term / session
CXL-tiered memory — warm, shared

Long-term / consolidated
NVM / flash-backed — dense, cheap, slow

Not brain-like, brain-inspired: tiered, event-driven, only the relevant path lights up.

Intelligence / Watt

Domain know-how delivered per joule, for the task in front of the system — not accuracy on a generic benchmark.

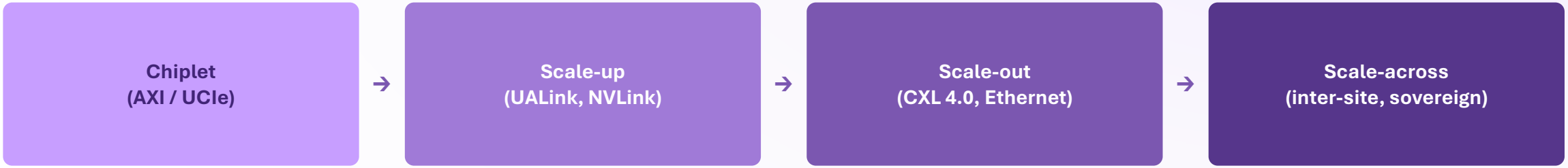
Memory / Watt

Tiered, brain-like memory — not one flat pool paying HBM-class energy for everything it touches.

Getting either metric right — and how far each can still be pushed — is an open research question. That's exactly the kind of problem this room is built for.

Sources: Nature Nanotechnology, neuromorphic device networks (Mar 2026); "Beyond Accuracy", arXiv 2511.14136; METR, arXiv 2503.14499.

One coherency story, from the chiplet to the data centre.



The stack is being built to connect on purpose: UALink 2.0's new Chiplet Spec (Apr 2026) is explicitly built compatible with UCIe 3.0, so the scale-up fabric and the on-package interconnect share one coherency story instead of two.

Scale-up is becoming a memory fabric, not just a link

UALink 1.0 already reaches 1,024 accelerators at 200 GT/s per lane; CXL 4.0 doubles memory bandwidth to 128 GT/s with multi-rack pooling.

Scale-out has to hold shape under load

Ultra Ethernet targets predictable performance across huge accelerator populations — a different problem than raw bandwidth.

Scale-across is the newest, least-settled layer

Extending coherent context beyond one building's walls — sovereignty and cross-company deployments will live or die here.

Sources: Hot Interconnects 2026 (Scale Up/Out/Across framing); UALink Consortium, UALink 2.0 (Apr 2026); CXL Consortium, CXL 4.0 (Nov 2025).

Agentic AI is changing chip floorplans, not just software stacks.

Intel's Hot Chips 2026 answer: three purpose-built dies, one agentic system.

Diamond Rapids	Crescent Island	Wildcat Lake
high-performance orchestration	efficient inference GPU	intelligent edge / client SoC

Tied together with UCIe chiplet interconnect and Foveros Direct 3D packaging — orchestration, inference and edge as one designed system, not three chips that happen to share a rack.

Agent context caching as a hardware primitive

The same KV-cache-sharing shift covered under fault line 04 — memory pooling, context-memory storage tiers — is really an orchestration hook: hardware that lets the scheduler move context, not just tokens.

Discovery and scaling need accelerated paths

If agents are meant to discover and recruit other agents at scale, that lookup and hand-off can't be a slow, general-purpose software path forever.

Source: Intel Newsroom, "Intel Outlines Architectures for Agentic AI at Hot Chips 2026" (Aug 2026).

Proprietary won't scale an ecosystem — open standards already exist at every level.

ISA / core

RISC-V

Silicon root of trust

OpenTitan · Caliptra

Rack / datacentre hardware

OCP (Open Compute Project)

Security certification

ISO/IEC 15408 (Common Criteria)

115+ companies co-govern UALink alone. Coupling accelerator choice to one proprietary fabric compounds lock-in across multi-hundred-million-dollar decisions. No European or open-ecosystem player can out-proprietary the incumbent — the only way in is winning the open layer, at every level above.

Are we training at any cost, for a sliver of use?

MoE models already activate ~11% of parameters per token; 77% of real queries never needed frontier scale. We keep training bigger flagships anyway — a rational bet, or a sustainability question we're not asking?

Sources: UALink Consortium (2026); DeepSeekMoE, arXiv 2502.11089; Saad-Falcon et al., arXiv 2511.07885; Caliptra (CHIPS Alliance/OCP); OpenTitan (Google/lowRISC); ISO/IEC 15408.

openchip

Two targets worth being unreasonable about.

1,000×

more intelligence per watt than today — not 2×, not an incremental roadmap bump. Judged against the domain-specific metric from earlier in this section, not a generic benchmark.

1,000×

less space and energy to encode the same real, usable know-how — compressing what a system knows, not just how fast it computes.

Neither target is reachable by scaling today's architecture harder. Both are reachable by designing the system around what intelligence actually costs.

openchip

One last sketch — no products, just the brief

From computing to cognition

If we started today, this is what we'd ask our engineers to build.

openchip

Stop provisioning for the biggest model. Provision for the one running now.

Addressable space:
trillions of parameters

→

Automatic, real-time
mutation

→

Active, right now:
tens of billions

Hardware that mutates to the use case, not the other way round

The model space can hold trillions of parameters. What's actually active at any moment should be tens of billions — decided automatically, in real time, by the task in front of it.

Automatic data-type adaptation

Precision that adjusts itself to what a task actually needs, not a format frozen at design time.

Interfaces built around intelligence/watt, not fixed allocation

Resource allocation as a live decision driven by the metric from earlier in this talk — not a static policy set once and left alone.

openchip

We don't need better caches. We need something that remembers.

Memories, not cache lines

Not tagging and caching data — associating it the way memory actually works: bound to when it happened and what it followed from. Temporality and causality as first-class properties, not metadata bolted on after.

New hardware schemes for storing and connecting context

Emulate the associative structure directly in how information is stored and linked, not just in the software layer sitting on top of flat storage.

Privacy, privacy, privacy

Our memories don't get scanned. We share what we choose, when we choose, shaped for who's asking. That selective, audience-aware disclosure needs to be architecture — not a policy written after the fact.

Infinite capacity, zero-cycle access

Index, manage and reach the relevant memory without paying a latency penalty for how much you've accumulated.

Fewer monoliths. More minds, well connected.

Today

One model. Trillions of parameters. One place it has to live.

What we need

Many smaller minds, each hundreds of billions of parameters, activated and connected in real time — only where the work actually is.

Neurons grouped by model, models grouped like brains

Different nodes and servers hosting different specialised models — interconnected as efficiently as the chiplet-to-scale-across coherency stack described earlier in this talk allows.

A scale-out mechanism for sharing know-how, not just weights

Today, humans do this with councils of scientists convening to compare notes. Systems need the equivalent — a real mechanism for scaling and connecting compute, memory and storage together, not each in isolation.

The part that has to hold all three of the above together.

Infrastructure that connects

Everything above only works if compute, memory and the models sitting on different machines can reach each other as if they were one system — the coherency stack from earlier in this talk, made real.

Infrastructure that's honestly sustainable

Not offset-sustainable. Efficient and sustainable energy as a design input from day one, not a claim added at launch.

Infrastructure that's composable, not fixed

Scales up, down and sideways without a rebuild — the same instinct behind memory pooling and open interconnects, applied to the whole system, not just one layer of it.

None of this is a product pitch. It's the brief.

Three questions to sit with.

RESOURCES

We keep optimising the model. When did we last optimise joules per correct answer?

TALENT

We're fighting over who trains the smartest model. Almost no one is fighting over who builds the chips underneath it.

PROGRESS

We keep asking how to reach general intelligence. The sharper question: how much intelligence does this problem need — and at what cost?

Which of our current design assumptions are actually helping us — and which are just walls of the box we've stopped noticing we're in?

openchip

Call to action.

Drive and evaluate through intelligence and value added

Judge every design and every deployment against know-how delivered per watt, per dollar, per domain — not parameter count, not peak TFLOPS, not leaderboard position.

Drive designs at scale

Chase the standards-based path — chiplet to scale-up to scale-out to scale-across — not a proprietary shortcut that wins one company's benchmark and nobody else's ecosystem.

Choose the metric before the next model, not after

Intelligence/Watt, Memory/Watt and trusted Context aren't retrospective scorecards. Decide what we're optimising for before the next architecture gets frozen, not once the results are in.

Bring academia and industry into the same room, earlier

This room is exactly where that has to start — open architectures, shared infrastructure, and research questions industry can't answer alone. Co-design doesn't work as a handoff after the fact.

This is the agenda for the next five years — not a prediction of them.

Thank you.

Let's hear where you disagree.

Cesc Guim

CEO, Openchip

linkedin.com/in/francesc-guim

openchip.com

References

— Stanford HAI, The 2026 AI Index Report

— Lightcast, Stanford AI Index 2026 labour-market chapter

— IEA, “Key Questions on Energy and AI” (Apr 2026)

— Gartner, data-centre power forecast (Jun 2026)

— LBNL / US DOE, data-centre electricity estimates (2024)

— Saad-Falcon et al., “Intelligence per Watt”, arXiv 2511.07885 (2025)

— NVIDIA Developer Blog, “Scaling Token Factory Revenue” (2026)

— Yuan et al. (DeepSeek-AI), “Native Sparse Attention”, arXiv 2502.11089 (2025)

— Lu et al., “MoBA: Mixture of Block Attention”, arXiv 2502.13189 (2025)

— OpenVINO 2026.2 release notes (INT4 KV-cache compression)

— Staniszewski & Łańcucki, “KV Cache Transform Coding”, ICLR 2026

— CXL Consortium, CXL 4.0 specification (Nov 2025) & FMS 2026 recap

— Marvell, AI Memory Infrastructure Portfolio announcement (Aug 2026)

— PatSnap / TechInsights, chiplet & advanced-packaging outlooks (2026)

— Intel Foundry & Marvell, ECTC 2026 disclosures

— “When Agents Handle Secrets”, confidential-computing survey (May 2026)

— EU Digital Omnibus on AI, provisional agreement (Jun–Jul 2026)

— arXiv 2605.17062, 2026 frontier hallucination re-evaluation

— Deloitte, Global Semiconductor Talent Shortage (2022)

— Semiconductor Industry Association, 2030 workforce projections (2023)

— Synopsys, Cadence, Siemens & NVIDIA, DAC 2026 (Jul 2026)

— Openchip, BER10 launch announcement (30 Jul 2026)

— HPCwire, Openchip–NEC RISC-V VPU coverage (Nov 2025)

— EE Times Europe, Openchip / Cesc Guim interview (2026)

— “Beyond Accuracy” (CLEAR framework), arXiv 2511.14136 (2025)

— METR, “Measuring AI Ability to Complete Long Software Tasks”, arXiv 2503.14499 (2025)

— Nature Nanotechnology, protonic nickelate neuromorphic devices (Mar 2026)

— “Deliberative Curation”, multi-agent knowledge bases, arXiv 2606.00007 (2026)

— “Open Challenges in Multi-Agent Security”, arXiv 2505.02077 (2025)

— UALink Consortium, UALink 2.0 white paper (Apr 2026)

— Hot Interconnects 2026, Scale Up / Out / Across program

— Intel Newsroom, “Agentic AI Architectures at Hot Chips 2026”

BER10: what taking the design space seriously looks like.

BER10

Openchip's first silicon — launched 30 July 2026

- 64-bit RISC-V, sub-2nm Gate-All-Around process
- Boots a real Linux, runs real software — not a lab demo
- Co-developed vector processing units with NEC for the Aurora HPC platform — the same scalar/vector interface tension this abstract opened with
- 120+ engineers, 15 nationalities, built for European digital sovereignty

Sovereignty as a design constraint

Not an afterthought bolted on — security, scalability and sustainability were architectural requirements from line one.

Built on the open standard, not around it

RISC-V means the control plane is customizable without a per-chip licensing tax — exactly the orchestration layer fault line 02 described.

Commercial production: 2028

This is a multi-year bet on the systems-level view this whole talk has argued for — made before it was fashionable to make it.

Source: Openchip, BER10 launch announcement (30 Jul 2026); HPCwire.