

The constraint is no longer silicon. It is **electricity** — and the speed of building to consume it.

Eight companies presented across the AI infrastructure stack at COMPUTEX 2026 in Taipei: chips, power delivery, racks, cooling, storage and networking. They did not coordinate their messages. They arrived at the same one.

FIVE STATEMENTS THAT CAPTURE WHAT WAS SAID

PAGE ONE

01 · Power density has moved fifty-fold in five years

A typical AI rack drew 10–20 kW three years ago. NVIDIA's NVL72 draws 150–200 kW today. The next platform generation targets roughly one megawatt in a single rack. Texas Instruments framed it historically: 5 kW per rack twenty-five years ago — enough for eight Taipei households — against one megawatt today, enough for 1,600.

02 · The 48-volt data centre is finished

At a megawatt per rack, a 48 V backplane would need to carry over 20,000 amps. The industry has moved to 800 V DC — or ± 400 V — out of physical necessity, not preference. Flex is already discussing 1,500 V DC. Every speaker on power independently confirmed the transition.

03 · Grid growth cannot meet AI growth

Endura put global grid generation growth at roughly 3.6% to 2030, against data centre demand growing about six times faster and AI demand alone twice as fast as the grid. Efficiency is the only near-term lever that has not been pulled.

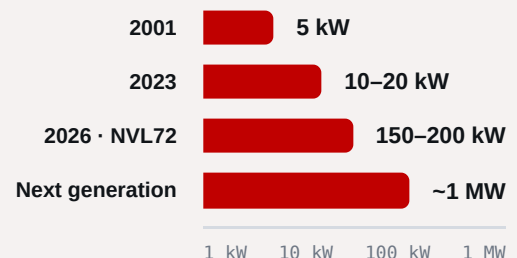
04 · Execution, not technology, is the stated bottleneck

Flex was explicit: “*The real constraint isn't the technology or the innovation, it's execution at scale.*” Supermicro quantified the mismatch — silicon refreshes in under twelve months, while a data centre takes twelve to twenty-four months to build, and two to three years at frontier scale.

05 · The unit of account is now tokens per watt

Both NVIDIA and Endura made the shift explicit. A data centre is no longer a cost centre running applications; it is a factory converting megawatts into tokens, measured on tokens per watt and tokens per dollar. This is the single most useful reframing for anyone selling power, cooling or land into this market.

RACK POWER DENSITY, 2001 TO NEXT GENERATION



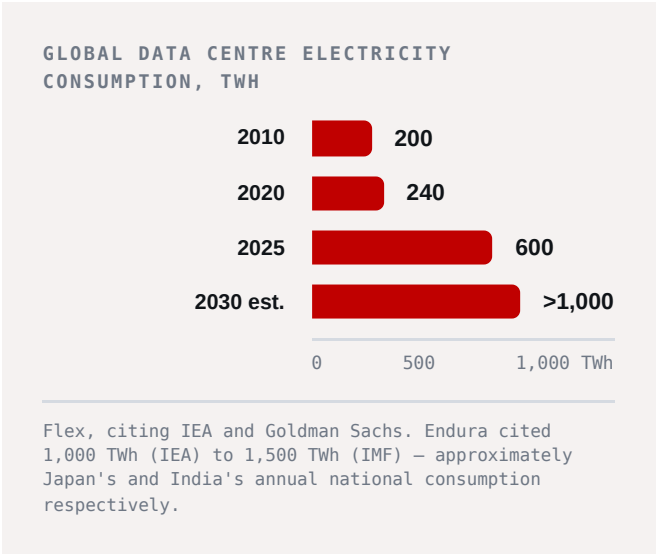
Logarithmic scale — each division is a ten-fold increase. Sources: Texas Instruments and Flex, COMPUTEX 2026.

THE NUMBERS, AS PRESENTED ON STAGE

PAGE TWO

Demand and capital

MEASURE	FIGURE
Global data centre electricity consumption	200 → 240 → 600 → >1,000 TWh 2010 · 2020 · 2025 · 2030
AI portion of that consumption	90 → >400 TWh by 2030
Annual AI data centre capital spend	~US\$1tn / year by 2029
IT infrastructure spend in play	~US\$1.5tn
Power implied by announced projects	~190 GW
Cumulative AI chip power deployed	2 → 13 GW in two years
Token price range	US\$0.40 – 25 / million



Power delivery

MEASURE	FIGURE
Processor current draw	500 A → 2,000 A → >10,000 A
Processors per rack	~40 (2022) → >500 (2026)
Power per processor package	>2.5 kW now, >10 kW by 2029
Backplane voltage	48 V → 800 V DC / ±400 V
Gain from vertical power delivery	~50 kW saved per 1 MW rack
Distribution loss on the server board	15–30% before the transistors
Gain from in-package regulation	~85% less loss, +10–25% tokens/W
Power smoothing gain	~40% more GPUs per GW

Cooling and build

50–100 kW Practical ceiling for air cooling per rack. Liquid is mandatory beyond it.	+50% Rise in OCP coolant flow guidance in two years — 1.0 to 1.5–2.0 L/min/kW.
122 days xAI Colossus, 100,000+ GPUs, against 2–3 years for a conventional build.	PUE 1.1 Samsung Heavy Industries' 50 MW floating data centre, built in under a year.

A caution on all of the above. These are figures as stated on stage by speakers presenting their own commercial positions, citing third-party forecasts without supplying the underlying documents. Treat them as accurately transcribed claims, not validated data.

SIX THEMES THAT CUT ACROSS ALL EIGHT SESSIONS

PAGE THREE

01 · Energy is the first design decision

NVIDIA structured its session around “extreme co-design” across five layers — energy, chips, infrastructure, models, applications — and put energy at the bottom deliberately.

The consequence is a change in load *shape*, not just size. AI training runs tens of thousands of GPUs in lockstep, producing a square wave at gigawatt scale. NVIDIA's answer is a facility-level energy management layer that stores during troughs and discharges into peaks. Claimed return: **40% more GPUs per contracted gigawatt** — which is to say interconnection, not chip supply, was the limit.

Read across: this is a utility-side product, not an IT product — a dispatchable behind-the-meter asset whose customer is an operator negotiating a connection.

04 · Liquid cooling is no longer optional

Air tops out at 50–100 kW per rack. Beyond that, liquid is mandatory — and NVIDIA explained why from the other direction: copper interconnect only reaches a few metres at current signalling rates, so GPUs must be packed densely enough to be joined by copper. **The cooling problem is a by-product of the interconnect solution.**

Requirements are moving faster than certification: OCP coolant flow guidance rose 50% in two years, so recently designed systems are already undersized.

Not raised by any speaker: waste heat. At a megawatt per rack, rejected heat is a district-scale resource and a siting variable.

02 · The 800 V transition, with standards unready

Four of eight speakers described the same migration. The reason is not efficiency ambition but Ohm's law: 800 V cuts current by an order of magnitude, reduces copper mass, and moves AC–DC conversion out of the IT rack entirely.

Flex named the real problem: **UL and IEC are not harmonised on 800 V DC**, and companies are shipping against a certification target that has not stabilised. PUE is a second example — calculated differently under EN, ASHRAE and each hyperscaler's internal spec.

Read across: a time-limited opening. The specification is being written in 2026–27. Participation after it settles is compliance, not influence.

05 · Storage became a performance bottleneck

Solidigm's example: a 16-word prompt the user sees as 25 tokens is processed as roughly **42,000 tokens** once instructions, retrieval, tools and history are assembled — generating about 13 GB of KV cache per request. Across 150 MLPerf submissions, the correlation between storage bandwidth and GPU utilisation was effectively perfect.

Western Digital reframed it as economics: at 200 exabytes, all-flash versus all-disk differs by about **US\$96 billion** — roughly three million GPUs. The stable answer is about 20% flash, 80% hard disk.

Read across: nine racks of drives replaced by one rack of SSD at 50 PB is a power-budget argument. Freed watts go to GPUs.

03 · Power delivery moved inside the package

TI, MPS and Endura presented the same roadmap from three vantage points: lateral delivery, then vertical beneath the processor, then integrated voltage regulators inside the ASIC package.

MPS revised its own forecast upward mid-stream — having predicted 5 kW packages by 2029, mid-2026 devices already exceed 2.5 kW, so the new figure is **over 10 kW per package**. “Imagine you pack six kitchen ovens into one tiny package, and you still want to take the heat out of it.”

Read across: a co-development thesis, not a component-sales thesis. Endura: “you cannot design the chip first and figure out the power later.”

06 · Speed of deployment is the competitive axis

Silicon refreshes in under a year; a frontier data centre takes two to three. A project begun against one hardware generation can be obsolete at commissioning. Supermicro's answer is modularity — pre-validated building blocks so deployment becomes assembly rather than engineering.

The sharpest illustration came from a guest on Supermicro's stage. Samsung Heavy Industries showed a **50 MW floating data centre** — LNG storage and solid oxide fuel cells, seawater cooling, PUE 1.1, zero water usage, built in under a year.

The decisive advantage was regulatory: on land, construction starts only after permitting. For a vessel, permitting runs in parallel.

WHERE THE OPENINGS ARE — AND HOW FAST THEY CLOSE

PAGE FOUR

None of the eight speakers addressed Japan or Southeast Asia directly. That is the point worth noticing: an entire conference programme described unmet demand that maps closely onto Japanese, Korean and Taiwanese industrial capability — and **the only Asian industrial company on any of these stages was Korean.**

THE OPENING	WHO IT FITS	HOW FAST IT CLOSES
Flexible load and behind-the-meter storage — flexibility as a connection asset, priced at 40% more GPUs per contracted gigawatt	Utilities and retail arms, BESS integrators, grid-forming inverter makers, trading houses	Immediate. Queues and tariffs are negotiated now, ahead of equipment standardisation.
The 800 V DC standards window — UL and IEC unharmonised, specification still being written	Power semiconductor makers, switchgear and protection, connector and busbar specialists, industry associations	2026–27. After convergence, the spec reflects whoever was in the room.
Components for in-package power — a named, unmet shopping list	Passive and magnetics manufacturers, advanced substrate and packaging houses, materials suppliers	Near term. When the 2029 package architecture fixes, supplier positions fix with it.
Thermal systems at gigawatt scale — plus waste heat, which no speaker raised	Pump, heat exchanger and chiller makers, fluid chemistry, heavy engineering, industrial parks for heat offtake	Near term for equipment. Waste heat is a three-to-five-year siting argument.
On-site generation and the floating data centre — shipbuilding, LNG, SOFC and seawater cooling recombined into a product	Shipbuilders and heavy industry, LNG value chain, fuel cell developers, trading houses, port and prefectural authorities	Urgent. A Korean competitor has a product and a stage. Months, not years.

MPS closed its session by asking the audience for four things: high-frequency magnetic materials, low-ESL and low-ESR capacitors, low-CTE PCB substrates, and embedded active and passive packaging — within a common-footprint ecosystem. Those are Japanese and Taiwanese supplier categories almost by definition. The question is whether the response comes as a merchant supplier of parts, or as a co-development partner inside the design cycle. The second is more defensible, and it is the one being solicited.

Where Asia Focus fits

We were in the room in Taipei, and we read the AI power build-out where it happens first — at the US demand frontier, in the technology roadmap, and inside Japan's power industry. If any of the openings above are yours, the useful conversation is about who to talk to and how quickly. Stand D1.

Startup & Tech Scouting · Strategic Sales & Support · GX/DX Business Transformation · Global Innovation Tours



Full 20-page summary – scan to request

SOURCES AND CONFIDENCE

Built from Otter.ai transcripts of eight COMPUTEX 2026 infrastructure sessions recorded on site, Taipei, 2–5 June 2026: NVIDIA (Kevin Deierling), Flex (Paul Diemer), Supermicro (Vik Malyala), Texas Instruments (Jeffrey Morroni), Monolithic Power Systems (Jinghai Zhou), Endura Technologies (Davood Yazdani), Solidigm (Avi Shetty), Western Digital (Ahmed Shihab). Figures are as stated on stage and are not independently verified; speakers cited IEA, Goldman Sachs, IMF and IDC forecasts without supplying the underlying documents. Speaker names and titles corrected against public sources; the NVIDIA attribution is inferred rather than confirmed.