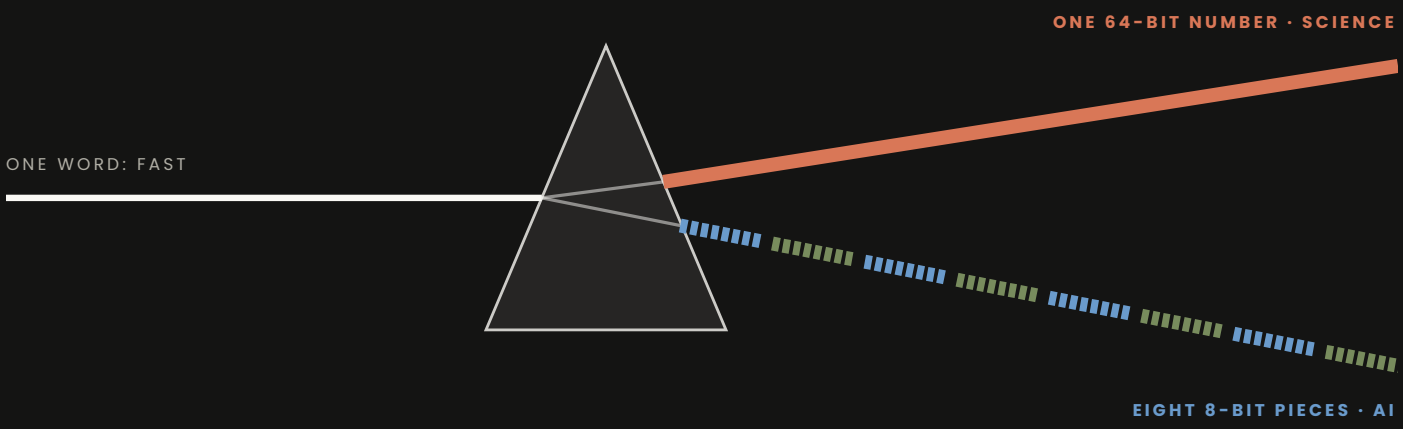
INFLECTION.

THE WEEKLY MAGAZINE OF INNOVATION

ISSUE 04 · FRIDAY, 20 NOVEMBER 2026 · DEEP DIVE — SUPERCOMPUTING



THE DEEP DIVE

The Scoreboard Went *Quiet*

The world's supercomputer ranking has just logged its fewest newcomers in 33 years, during the biggest computing build-out ever, and the machine at number one has no GPUs. Both facts have one cause. "Fast" has split into two kinds, and they no longer share an arithmetic.

- Why Nvidia's newer AI board lists 10 teraflops of double precision where the one before it listed 296
- Against the Grain: to get more 64-bit math, stop building 64-bit hardware
- Signals: a Starship caught from orbit · a drill kept cold in hot rock · silicon carbide on 300 mm wafers

01	**Two Kinds of Fast**	*The feature: why the list went quiet*	3–5
02	**Against the Grain**	*The Contrarian: eight bits are enough*	6
03	**Signals**	*Space · energy · semiconductors*	7
04	**By the Numbers**	*Seven figures worth remembering*	8
05	**The Long View**	*What the issue adds up to · sources*	9
06	**Next Issue**	*Back cover and the Lens question*	10

THE DISPATCH

A benchmark is a bet about what matters.

> *Ask what the number counts.*

A speedometer and an odometer both describe a car, and nobody confuses them. Computing has made do with one dial for 33 years: the TOP500, which ranks machines by how fast they solve a giant grid of equations in 64-bit arithmetic. This week the dial barely moved. Only 27 new machines joined the list, the fewest in its history, while the industry was adding hardware at a record pace.

The easy reading is that AI ate the supercomputer. The more useful one is that “fast” has become two things. AI is fast at cheap, coarse arithmetic, four and eight bits at a time. Science needs sixteen digits. The chips have drifted so far apart that the machine on top of the list has no GPUs at all, and one of Nvidia’s AI boards lists a thirtieth of the double precision of the board before it.

So this issue’s lens is to ask what each number counts. It explains the feature, it drives the contrarian column on page six, and it runs through our three Signals, where a rocket, a drill bit and a wafer each changed what was worth counting.

— THE EDITORS

Sources for every figure are listed on page 9. Vendor figures, company claims and our own arithmetic are labelled as such. Previously: Issue 03, base editing · Issue 02, building in orbit · Issue 01, photonic computing.

THE FEATURE

Two Kinds of Fast

For 33 years one test has ranked the world's fastest computers. This week it recorded its quietest update ever, with a machine on top that has no GPUs. The cause is not a slowdown. It is a fork in the arithmetic.

BY THE INFLECTION DESK · RESEARCH DRAWN FROM TOP500, NVIDIA, THE NEXT PLATFORM, THE REGISTER AND EPOCH AI · 9-MINUTE READ

Twice a year since 1993, the TOP500 has ranked the world's supercomputers by a single test: how quickly each can solve one enormous system of linear equations. On 16 November, at the SC26 conference in Chicago, the 68th edition arrived with its top ten unchanged and only 27 new machines among the 500, the fewest in the list's history. Combined performance crept up 4.4 percent, to 19.57 exaflops (an exaflop is a billion billion calculations a second). Yet the list's authors cite Epoch AI estimates that the AI computing capacity added each quarter has been growing about 2.5 times a year.

The problem: one dial, two engines

The test is called High-Performance Linpack, or HPL, and it has one stubborn rule. Every number in it is double precision, written FP64: 64 bits, about sixteen significant digits. Science needs them. A climate or combustion model multiplies and subtracts numbers billions of times over, and rounding errors compound like interest. Sixteen digits keep the error from swallowing the answer.

AI learned to live without them. A neural network is, loosely, a vast averaging machine, and averages forgive coarse inputs. So chip designers have spent a decade shrinking the numbers: 16 bits, then 8, then 4. A 4-bit number can take just sixteen values. Smaller numbers mean more multiplications per watt, and the circuits that do them, which Nvidia calls tensor cores, are where AI chips now put their muscle.

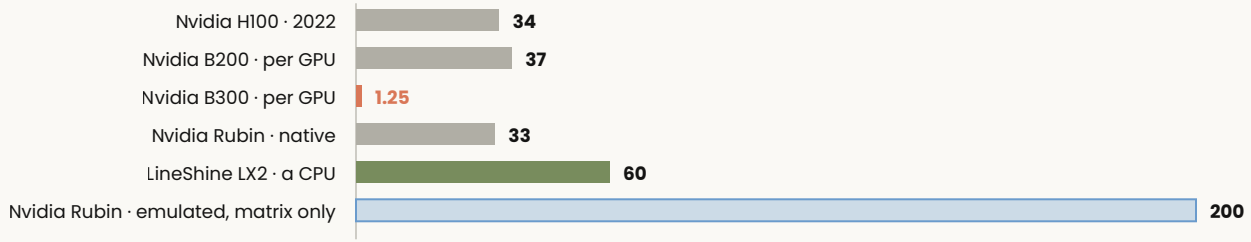
Where the hardware went

The list's own caretakers named the cause. "This suggests that growth has indeed shifted toward AI factories, and from capability systems to capacity systems," they wrote. A capability system is one machine built to run one enormous job. A capacity system runs many smaller jobs at once, which is what serving a chatbot looks like.

Epoch AI, which tracks AI data centres from satellite images, counts 96 sites holding the equivalent of about 15.8 million Nvidia H100 chips. It estimates that five companies (Amazon, Google, Meta, Microsoft and Oracle) held 71 percent of the world's AI computing power at the end of 2025. Most of that hardware has never been entered; the list is voluntary. And as the next page shows, some of the newest of it would score poorly.

What the week put on the table

DOUBLE PRECISION PER CHIP, AS LISTED · TERAFLOPS OF FP64, PEAK



Sources: Nvidia HGX specifications (board figures divided by eight GPUs); The Register (H100); TOP500 (LineShine's peak rating divided by its 45,360 processors, our arithmetic). The emulated figure is Nvidia's and applies to matrix multiplication only.

A champion with no GPUs

The machine at number one is LineShine, at the National Supercomputing Centre in Shenzhen. It entered in June at 2.198 exaflops, the first China-based system to lead since 2017, and nothing dislodged it this week. It is built from 13,789,440 cores in 304-core, Arm-based processors called LX2, and it has no accelerators. The TOP500 called it "the first system on the TOP500 to exceed two exaflops of sustained double-precision performance using CPUs only."

The design is widely read as an answer to US export controls on advanced GPUs, and it costs electricity: about 42 megawatts, the most in the top five. But divide its peak rating by its 45,360 processors and each lists roughly 60 teraflops of double precision. Now look at the chart.

296 becomes 10

Nvidia publishes specifications for its eight-GPU server boards. The HGX B200 lists 296 teraflops of FP64. The newer HGX B300, built for AI, lists 10, while its dense 4-bit rating rose by half. Both are on sale. Per chip, that is about 37 teraflops against 1.25. On paper, one LineShine processor carries the native double precision of 48 of those GPUs.

> *"It's still FP64. It's not mixed precision. It's just done and constructed in a different way from the hardware perspective."*
>
> DAN ERNST · NVIDIA, TO THE REGISTER

Sixty-four bits from eight

Ernst was describing the cleverest idea in this story. Nvidia's maths library can now compute a double-precision matrix product without double-precision hardware. The method, the Ozaki scheme, dates from 2012: split each 64-bit number into small whole-number pieces, multiply the pieces exactly on the low-precision tensor cores, then stitch the answer back together. Nvidia lists 200 teraflops for matrix multiplication done this way on its newest chip, Rubin. That is six times Rubin's native 33, itself one fewer than the H100 of four years ago.

And a chip for the rest

On the day the list came out, Nvidia also announced Rubin HPC, a separate GPU for simulation with more than 50,000 CUDA cores, aimed at the sparse and stencil codes that emulation does not help. It is due in 2028, and Nvidia has published no double-precision figure for it. AMD projects up to 288 teraflops of native FP64 for its MI430X, due in 2027.

FROM LAB TO MARKET

What it means, and what it does not

Proven: the list's 27 newcomers are a record low, a CPU-only machine leads it, and on vendors' own spec sheets the native double precision of AI chips has stalled or fallen. **Not yet shown:** that the quiet is permanent. Timothy Prickett Morgan of The Next Platform argues that "the stall is more about roadmap timing", with buyers waiting on the next generation of chips and networking, and he expects the shortfall to be made up over the next few lists.

For anyone buying compute, "how fast?" now needs a second question: fast at which arithmetic? AI capacity is contracted by the megawatt. On 16 November TensorWave said it will supply OpenAI with 457 megawatts over six years. Scientific capability is still scored in 64-bit flops. A megawatt of one is no longer a known quantity of the other.

Three routes to double precision are on offer. Wide-vector CPUs, LineShine's route, run existing code unchanged and pay in power. Emulation on AI chips is cheap where it fits: in February the Texas Advanced Computing Center said it expected two to three times native speed on its Horizon system, and would submit Linpack results both ways. Science accelerators are due in 2027 and 2028.

For the CTO. Sort your workloads by arithmetic before you buy or rent: which precision, and matrix-shaped or vector-shaped? An "H100-equivalent" measures AI-fast only. Two boards from one chip family can differ thirtyfold in double precision while matching each other at 8 bits. That reading of Nvidia's table is ours, not Nvidia's.

Timeline: Vera CPU racks, early 2027 · AMD MI430X, 2027 · next TOP500, mid-2027 · Rubin HPC, 2028.

FIELD NOTES · HOW TO DO 64-BIT MATH ON 8-BIT HARDWARE

1. **Split.** Turn each double-precision matrix into several matrices of small whole numbers: its remainders after dividing by a set of small numbers that share no factors.
2. **Multiply.** Run each small-number product on the fast low-precision tensor cores. The pieces are small enough that nothing is rounded. Use more pieces and you keep more digits.
3. **Recombine.** The Chinese Remainder Theorem rebuilds the one answer that fits every remainder, and the result is turned back into floating point.

Try it. A number leaves 2 when divided by 7, 3 when divided by 11 and 5 when divided by 13. Only one number below 1,001 fits: 135. Three small facts pin down one larger one. The Ozaki scheme does the same with whole matrices.

GLOSSARY

FP64. Double precision: numbers stored in 64 bits, about sixteen significant digits.

HPL. High-Performance Linpack, the dense equation-solving test that ranks the TOP500.

Capability vs capacity. One machine for one huge job, versus many machines for many jobs.

Ozaki scheme. A way to compute high-precision matrix products from many low-precision ones.

THE CONTRARIAN

Eight Bits Are *Enough*

Everyone wants their 64-bit hardware back. The stranger bet is that nobody needs it.

The unfashionable read

The natural response to a board that lists 10 teraflops of double precision is to demand the old circuits back, and the market is obliging with chips aimed at science for 2027 and 2028. The contrarian position comes from Satoshi Matsuoka, director of the RIKEN centre that operates Japan's Fugaku supercomputer. His paper is titled "FP8 is All You Need", and its subtitle promises to debunk hardware FP64 as "the HPC Holy Grail". His model puts the ceiling for emulated double precision near 473 teraflops on Rubin, and large matrix products at about half that, against 33 in native hardware.

It looks wrong until you price it. AI demand has made low-precision multiplication extraordinarily cheap. Nvidia lists 17.5 petaflops of dense 8-bit math on one Rubin chip beside 33 teraflops of FP64, so the chip does about 530 small operations in the time of one native 64-bit one. Paying a large bill in small coins is absurd unless the coins are nearly free. Emulation wins whenever a 64-bit result costs fewer than 530 coins, and Nvidia's figure for big matrices implies a price of about 88.

What follows (our inference)

Precision stops being a property of the chip and becomes a setting on the job. If that holds, the asset is the conversion layer, and above all the proof that its answers are right. Here is a proposal to test, not a finding: pair emulation with a receipt. A 1977 result, Freivalds' check, tests a matrix product with a few cheap matrix-times-vector multiplications instead of redoing it. Small coins buy the answer, and a few large ones pay for the audit.

WHY THE CONSENSUS DISAGREES

01 **It reaches a tenth of the field.** A paper posted on 9 November, whose authors include AMD fellow Nicholas Malaya, calls emulation "an optimization for the subset of workloads whose dominant computation is sufficiently large, dense, and GEMM-shaped", and puts that subset at "less than 10% of the HPC ecosystem". AMD sells native FP64, so weigh the source. The claim is checkable.

02 **It is not the same arithmetic.** The Register reports that Nvidia's emulation is not fully IEEE-compliant: it does not handle signed zeros, NaNs or infinities, though Nvidia says it has added checks for them. Malaya says it uses about twice the memory and breaks down on poorly conditioned problems.

03 **The big numbers are vendor or model numbers.** The 200 teraflops is Nvidia's specification and the 473 comes from Matsuoka's model. TechTimes notes that no independent accuracy test on Rubin has been published.

04 **The vendors are hedging.** If eight bits were enough, Rubin HPC and the MI430X would not need to exist. And most science code waits on memory, not arithmetic: on the list's sparse-code benchmark, LineShine reaches about 1 percent of its headline score.

SPACE

22 h

A rocket's upper stage went to orbit and came home to its tower

SpaceX's Starship lifted off from Starbase, Texas, on 13 November. After about 22 hours in space the upper stage plunged back through the atmosphere, relit three of its six engines, flipped upright and was caught by the mechanical arms of its launch tower. CBS News called it the first time a Starship had been recovered from orbit. If the catch can be repeated, the unit of spaceflight shifts from launches to turnarounds. The other half of the plan failed: the Super Heavy booster was diverted offshore and lost, which Elon Musk put down to a "slight underperformance of engines". Refuelling in orbit, which Moon missions need, has yet to be shown.

SOURCE · CBS News, William Harwood, 14 Nov 2026 · cbsnews.com/news/spacex-catch-starship-rocket/

ENERGY

341°C

To drill hotter rock, a geothermal firm made the drill colder

The electronics that steer a drill bit are the weak point in very hot rock. Mazama Energy says it has field-tested a fix at Newberry, Oregon: pumping supercritical carbon dioxide, CO_2 pushed past the point where liquid and gas are distinct, down the well as the drilling fluid. In an August demonstration reported on 16 November, its steering tools stayed at about 91°C (195°F) while cutting rock at 341°C (645°F), and drilling ran 20 percent faster. The company calls it the first field use of the fluid in a deep geothermal well. The figures are its own, and the rock was still short of the 374°C "superhot" threshold it is aiming for.

SOURCE · Mazama Energy via GlobeNewswire, 16 Nov 2026 · globenewswire.com/news-release/2026/11/16/3403585/0/en/mazama-energy-achieves-drilling-breakthrough-that-advances-the-economics-of-superhot-geothermal.html

SEMICONDUCTORS

2.3×

Silicon carbide jumps a wafer size, on machines built for silicon

Silicon carbide makes the tough power switches inside electric cars, solar inverters and, increasingly, the power supplies of AI data centres, but it has been stuck on small wafers. On 9 November Infineon said it had processed 300-millimetre silicon carbide wafers into working power devices at its fab in Villach, Austria, which it calls a first for a full-scale fab. A 300 mm wafer yields about 2.3 times as many chips as a 200 mm one. The quieter detail is the method: the wafers ran on existing silicon manufacturing equipment. There is no production date. Infineon says the ramp will be set with customers, and its 200 mm roadmap continues.

SOURCE · Infineon Technologies press release, 9 Nov 2026 · infineon.com/press-release/2026/infpr202611-028

27

New systems on the November 2026 TOP500: the fewest in the list's 33-year history.

2.198 EF

LineShine's Linpack score in exaflops, from 13.8 million CPU cores and no GPUs.

296 to 10

Teraflops of double precision listed for Nvidia's eight-GPU HGX B200 board, then for the newer HGX B300.

71%

Share of the world's AI compute held by five hyperscalers at the end of 2025, on Epoch AI's estimate.

6×

Nvidia's emulated double-precision matrix speed on Rubin against its native figure: 200 teraflops to 33. Vendor numbers.

1%

LineShine's score on HPCG, the benchmark that mimics sparse science code, as a share of its Linpack score.

15.8 million

H100-equivalents of AI compute at the 96 data-centre sites Epoch AI tracks, drawing 15.1 gigawatts. Most new AI hardware never reaches the TOP500.

The surprise hides in the unit

Four stories this week, one pattern. The TOP500 counts 64-bit flops, and the industry began building for 4-bit ones, so the count went quiet while the building went on. Starship has been scored in launches; a ship caught by its own tower changes the unit to turnarounds. Mazama's rock was already hot; the number that mattered was the temperature of the electronics behind the bit. Infineon's news was less the material than the diameter of the wafer and the age of the machines it ran on.

The warning label is the same for all four. Emulated double precision rests on vendor benchmarks and one researcher's model, and a paper this month puts its reach below a tenth of scientific computing. Starship lost its booster. Mazama's figures are its own, and its rock was not yet superhot. Infineon has no production date. Each is a first, and firsts are where the error bars are widest.

What to watch: whether the list's newcomers rebound in 2027, as The Next Platform expects; whether a Linpack run done with emulated arithmetic is submitted, and how the list treats it; and the double-precision figure Nvidia has yet to publish for Rubin HPC.

SOURCES & FURTHER READING

1. **TOP500**, November 2026 list — https://top500.org/lists/top500/2026/11/
2. **TOP500**, "LineShine Debuts at No. 1 as the TOP500 Enters a New Global Exascale Era," 23 Jun 2026 — https://top500.org/news/lineshine-debuts-no-1-top500-enters-new-global-exascale-era/
3. **Tom's Hardware** (A. Shilov), "China retains fastest supercomputer in the world crown as AI data centers soak up new hardware," 16 Nov 2026 — https://www.tomshardware.com/tech-industry/supercomputers/china-retains-fastest-supercomputer-in-the-world-crown-as-ai-data-centers-soak-up-new-hardware-top500-records-fewest-new-systems-in-33-years
4. **Network World** (D. Dubie), "China's LineShine tops US systems in supercomputer ranking," 16 Nov 2026 — https://www.networkworld.com/article/4249297/chinas-lineshine-remains-worlds-fastest-supercomputer.html
5. **The Next Platform** (T. P. Morgan), "Supercomputer Installs Take A Dive In The Second Half Of 2026," 16 Nov 2026 — https://www.nextplatform.com/hpc/2026/11/16/supercomputer-installs-take-a-dive-in-the-second-half-of-2026/5310221
6. **The Next Platform**, "A Deep Dive On China's 'LineShine' All-CPU, Exaflops-Class Supercomputer," 25 Jun 2026 — https://www.nextplatform.com/hpc/2026/06/25/a-deep-dive-on-chinas-lineshine-all-cpu-exaflops-class-supercomputer/5262439
7. **The Next Platform**, "TACC Explores Mixed Precision And FP64 Emulation For HPC With Horizon," 3 Feb 2026 — https://www.nextplatform.com/hpc/2026/02/03/tacc-explores-mixed-precision-and-fp64-emulation-for-hpc-with-horizon/4092147
8. **Nvidia**, HGX platform specifications, read 17 Nov 2026 — https://www.nvidia.com/en-us/data-center/hgx/
9. **The Register** (T. Mann), "Nvidia leans on emulation to squeeze more HPC oomph from AI chips in race against AMD," 18 Jan 2026 — https://www.theregister.com/2026/01/18/nvidia_fp64_emulation/
10. **StorageReview** (B. Beeler), "NVIDIA SC26: Rubin HPC GPU With 50,000 CUDA Cores Due in 2028," 16 Nov 2026 — https://www.storagereview.com/news/nvidia-sc26-rubin-hpc-gpu-vera-cpu-hpc-rack-nemotron-for-science-quantum
11. **TechTimes**, "Nvidia Rubin HPC Debuts With 50,000+ Cores," 16 Nov 2026 (AMD MI430X figures) — https://www.techtimes.com/articles/330867/20261116/nvidia-rubin-hpc-debuts-50000-cores-top-supercomputers-hit-3-percent-peak-sparse-code.htm
12. **TechTimes**, "NVIDIA Says FP64 Emulation Lifts Rubin Matrix Math to 175 TFLOPS; Vector Code Sees No Gain," 16 Nov 2026 — https://www.techtimes.com/articles/330792/20261116/nvidia-says-fp64-emulation-lifts-rubin-matrix-math-175-tflops-vector-code-sees-no-gain.htm
13. **S. Matsuoka**, "FP8 is All You Need (Part 1): Debunking Hardware FP64 as the HPC Holy Grail," arXiv:2606.06510 — https://arxiv.org/abs/2606.06510
14. **K. Swirydowicz, N. Malaya et al.**, "FP64 Performance and the Limits of Ozaki Emulation on HPC Systems," arXiv:2611.13332, 9 Nov 2026 — https://arxiv.org/abs/2611.13332
15. **K. Ozaki, Y. Uchino, T. Imamura**, "Ozaki Scheme II," arXiv:2504.08009 — https://arxiv.org/abs/2504.08009
16. **D. Mukunoki**, "DGEMM without FP64 Arithmetic," arXiv:2508.00441 (the scheme's 2012 origin) — https://arxiv.org/abs/2508.00441
17. **Epoch AI**, "Five hyperscalers now own over two-thirds of global AI compute," 14 Apr 2026 — https://epoch.ai/data-insights/hyperscalers-control-most-compute
18. **Epoch AI**, "Data on AI data centers," updated 6 Nov 2026 — https://epoch.ai/data/ai-data-centers
19. **TensorWave**, "TensorWave to Provide OpenAI with 457 Megawatts of AMD-Powered AI Infrastructure," 16 Nov 2026 — https://tensorwave.com/blog/tensorwave-to-provide-openai-with-457-megawatts-of-amd-powered-ai-infrastructure
20. **RIKEN**, Center for Computational Science (director; Fugaku) — https://www.riken.jp/en/research/labs/r-ccs/index.html
21. **R. Freivalds**, "Probabilistic Machines Can Use Less Running Time," IFIP Congress 1977; summary at https://en.wikipedia.org/wiki/Freivalds%27_algorithm
22. **CBS News** (W. Harwood), "SpaceX pulls off dramatic midair catch of Starship rocket after fiery plunge from orbit," 14 Nov 2026 — https://www.cbsnews.com/news/spacex-catch-starship-rocket/
23. **Mazama Energy** via GlobeNewswire, "Mazama Energy Achieves Drilling Breakthrough That Advances the Economics of Superhot Geothermal," 16 Nov 2026 — https://globenewswire.com/news-release/2026/11/16/3403585/0/en/mazama-energy-achieves-drilling-breakthrough-that-advances-the-economics-of-superhot-geothermal.html
24. **Infineon Technologies**, "Infineon marks innovation milestone with world's first 300-millimeter power silicon carbide (SiC) wafer," 9 Nov 2026 — https://www.infineon.com/press-release/2026/infpr202611-028

INFLECTION.

THE WEEKLY MAGAZINE OF INNOVATION

THE LENS · A QUESTION FOR YOUR WEEK

Which number on your dashboard quietly assumes that two different kinds of fast are the same thing?

NEXT ISSUE · FRIDAY, 27 NOVEMBER 2026

A fresh deep dive from a different domain, three new Signals, and a contrarian column that argues with the consensus.

ONE 64-BIT NUMBER, OR EIGHT 8-BIT PIECES. THIS ISSUE WAS ABOUT THE DIFFERENCE.

RESEARCHED, WRITTEN & DESIGNED WITH CLAUDE. TYPESET IN POPPINS & LORA ON THE ANTHROPIC PALETTE.
ISSUE 04 · 20 NOVEMBER 2026