AI IS RESHAPING MODERN PC DESIGNS

NEXT-GENERATION THIN-AND-LIGHT LAPTOPS REQUIRE MORE MEMORY BANDWIDTH AND BETTER APPROACHES TO THERMAL MANAGEMENT

SUMMARY

As token volumes and cloud costs rapidly compound thanks to more agentic workloads, power users of AI are buying their own local inference hardware, and the PC industry is responding. This is helping to shift more AI inferencing to the edge. In particular, local inference on thin-and-light laptops is projected to become a core target for the industry, based on vendors' product roadmaps. The challenge is that these machines are coming up against hard physical limits — especially for thermal management — as they address the demands of local inference in an already space-constrained form factor.

Tokens per second is the performance metric that matters for local inference, and it mostly scales with memory bandwidth. Memory bandwidth, in turn, is a function of bus width, capacity, and speed. Most laptops today use 128-bit memory buses paired with LPDDR5, which caps bandwidth at roughly 150 GB/s. But that is not enough to run today's frontier-quality models at acceptable speeds.

This reality has brought the industry to an inflection point where memory and the associated functions to accommodate it must be improved. Chip designers are already taking steps to achieve this. Qualcomm has moved to 192-bit bus widths in its X2 Elite family. AMD and NVIDIA have 256-bit offerings today. Apple's M5 Max processor uses a 512-bit bus that operates at 614 GB/s, and is capable of running a high-end model like gpt-oss-120b entirely locally. All available evidence indicates that the next 18 months will bring a broader shift to 256-bit and wider architectures, paired with the introduction of LPDDR6, which delivers faster data rates, better power efficiency, and improved error correction. The confluence of LPDDR6 and improved bus widths capable of robustly supporting local inference is expected to become mainstream in 2027 and 2028.

The chip suppliers are doing their part. To meet the growing demand for local inference in notebooks, OEMs must now solve the spatial challenges that come with wider buses. A 256-bit interface requires about eight length-matched, shielded 32-bit channels with traces under 25mm, so memory must sit adjacent to the SoC — exactly where fans and heat pipes live today in most designs. Apple has solved this architecturally with on-package unified memory. Every other OEM must solve it in the chassis.

Copyright © 2026 Moor Insights & Strategy

Traditional fan-based cooling approaches will not readily work in the next generation of AI notebooks. Fans will either have to shrink (and get louder) or disappear entirely, giving way to other thermal-management solutions that allow for the necessary denser memory couplings surrounding the SoC. In this context, solid-state cooling is the most credible alternative on the table. With all this in mind, OEMs that treat thermal design as a fundamental design challenge rooted in geometry, rather than a late-stage design trade-off, set themselves up to lead in the emerging era of on-device inference.

The stakes for getting this right extend beyond consumer electronics. With GPU export controls creating real challenges, there is a rising global demand for local inference as an alternative to cloud computing. At this writing, the Mac Studio backlog through mid-2026 is the clearest signal of this. Premium laptops paired with distributed compute tools could become meaningful sovereign AI infrastructure, if the thermal and geometric problems get solved. Moor Insights & Strategy (MI&S) believes that the companies that resolve these constraints will define this burgeoning market.

Market Context and Bandwidth Requirements for AI

The computing needs of AI have shifted from training to inference, leading to a broader set of devices performing more AI compute and trying to generate tokens as quickly and efficiently as possible. As token volumes and associated costs continue to rise, many power users are buying their own inference devices to avoid the rapidly escalating costs of running cloud-based models. This remains true even given the ongoing global memory shortage. Illustrative is the demand for Apple's powerful Mac Studio device, which, as of late June 2026, was reported to be backordered until October and had similar delays as far back as April.[1] The broader market has also recognized the demand for powerful local inference. For example, NVIDIA offers the DGX Spark, which features a unified memory architecture, 273 GB/s of memory bandwidth, and the company's GB10 Grace Blackwell chip to maximize AI performance and efficiency. The company also recently introduced the RTX Spark, which leverages most of the same IP in a laptop form factor. Meanwhile, devices like HP's Z2 Mini G1a leverage AMD's Strix Halo chip architecture, which includes 256 GB/s of unified memory bandwidth for highly capable local AI performance.

While leading consumer small-form-factor (SFF) desktops use large memory allocations, wider buses, and capable chipsets to deliver powerful local inference, much of the industry still pairs 128-bit memory buses with LPDDR5 memory, limiting inference

[1] Rajesh Pandey, "Mac mini and Mac Studio face long shipping delays," Cult of Mac, April 3, 2026.

Copyright © 2026 Moor Insights & Strategy

performance. This limitation arises because memory bandwidth is the main indicator for how many tokens per second an AI device can generate. Until recently, many consumer devices were capped at around 150 GB/s in peak memory bandwidth. As discussed above, however, we are already seeing silicon vendors including Qualcomm, AMD, and NVIDIA move to buses of 192 bits, 256 bits, and beyond. Apple's new M5 Max represents a sea change by pairing 512-bit memory buses with 128GB of RAM to deliver bandwidth speeds as high as 614 GB/s. This performance enables frontier-quality open-source models such as gpt-oss-120b to run entirely locally with cloud-like performance, demonstrating the emerging possibilities for local inference distributed across consumer-friendly form factors. In the bigger picture, all of these flagship configurations signal a shift beyond 128-bit buses and herald increases in memory bandwidth thanks to wider buses, higher capacities, and faster speeds.

Also primed to accelerate local inference enablement is the upcoming transition towards LPDDR6 memory, whose improvements over prior generations compound the advantages of wider buses. LPDDR6 not only delivers faster data rates (up to 14.4 Gbps) but also wider bit widths, enhanced power management, and improved reliability and error correction.[2] LPDDR6 is set to debut in 2026, with memory manufacturers including Samsung, SK Hynix, and Micron already sampling to customers. Widespread adoption is expected in 2027 and 2028. The new technology's promise of local inference is clearly top-of-mind, with Samsung even focusing on LPDDR6's optimization for on-device AI capabilities at CES 2026 earlier this year.[3]

Given the clear market demand for capable local inference machines, we expect 2027–2028 to be the timeframe when many vendors move beyond 128-bit controllers into 256-bit (and beyond) memory controllers paired with LPDDR6. While the memory crunch may affect the initial availability of LPDDR6, we believe that there will be a continuing appetite for the transition to faster memory with higher densities. Indeed, this transition could alleviate some of the capacity issues caused by the datacenter AI boom as more inference shifts to local edge devices. Still, the global memory shortage could persist and impede the expected growth of on-device AI.

[2] Scott Knowlton, "LPDDR6 vs. LPDDR5 and LPDDR5X: What's the Difference?," Synopsys, February 10, 2026.
[3] See Samsung, "LPDDR6: World's First Next-Gen LPDDR Optimized for High-Performing On-Device AI," December 3, 2025.

Copyright © 2026 Moor Insights & Strategy

Physical Constraints: Geometry, Power, and Thermal Limits

The discussion to this point has largely focused on SFF units because there are currently limited options — not much beyond Apple's M5 Pro and Max lines — that offer highly capable local inference in a traditional laptop form factor. For a laptop to be considered thin-and-light, it needs to be 18mm or less in thickness and usually below 2kg in overall weight. These constraints present challenges for the PCB layout, power delivery, and thermals, as there is only so much physical space available to dissipate all the heat generated. In the thin-and-light category, OEMs are continuously engaged in a balancing act between device performance and thermal limitations, all while still accounting for acoustic considerations. Years of work has improved the performance and capabilities of thin-and-light laptops, but the stark physical limitations remain. And even though some laptops have been made slightly larger to accommodate user requirements, space inside the chassis is eternally at a premium.

With the increasing demand for on-device inference, the space problem compounds further: Larger and more capable SoCs must also be paired with additional memory. OEMs that want to deliver performant on-device inference must directly address these thermal and spatial limitations to give users the local inference they are clamoring for in the traditional thin-and-light form factor.

Delivering 256-bit or higher memory buses requires roughly eight 32-bit memory channels with length-matched and shielded traces. This requires shorter memory traces (under 25mm) with more precise signals to maintain signal integrity. That, in turn, means locating the memory chips much closer to the memory controller on the SoC, likely in a very tight configuration. This kind of array is not optional, but absolutely necessary to support capable on-device inference. While emergent technologies such as LPCAMM might also solve some of the challenges created by wider buses and higher bandwidth by simplifying configuration and interfaces, even LPCAMM creates geometric challenges within existing thin-and-light internal layouts.

When OEMs have faced spatial design challenges in the past, they have responded with design trade-offs such as reducing the battery size, increasing the z-height, or simply accepting more frequent performance throttling. Unfortunately, each of these traditional trade-offs is highly undesirable for delivering user-friendly local inference. So, unless they want to ignore market demand for on-device inference — which would come with harsh competitive downsides — OEMs must yet again innovate to update their designs for thin-and-light notebooks that meet user requirements.

Copyright © 2026 Moor Insights & Strategy

CLIENT SILICON ARCHITECTURE ASSESSMENT

Different silicon architectures may enable PC OEMs to take different approaches to the spatial and thermal challenges presented by local inference in thin-and-light laptops.

Apple has made itself a leader in PC silicon with its M-Series, which introduced the unified memory architecture. The M-Series leverages Apple's strengths in mobile, custom Arm-based processors and scales them up to a higher-performance tier for Mac PCs. Since the introduction of the A4 in 2010, Apple's CPU performance has been a leadership story, but its neural engine (NPU) has mostly been right-sized for Apple's own workloads with a 16-core design. However, with the M5, Apple has introduced GPU Neural Accelerators that scale with the GPU configuration, meaning that the M5 Max now also features a powerful 40-core GPU. Apple claims[4] that this enables 4x faster AI inference than its previous M4 Max. Apple's unified memory architecture helps it address the spatial constraints discussed above, but only at the expense of flexibility and upgradability. We have already discussed the sea change represented by the M5 Max bus; given Apple's MLX software and memory architectures, we believe Apple is ready to make the transition to LPDDR6 in due course. What is clear: Thanks to its unified-memory approach, Apple is already delivering frontier-class local inference on Arm-based platforms that run comfortably within a laptop or SFF desktop.

Qualcomm's latest Snapdragon X2 Elite family of chips pairs the company's top Oryon CPU cores with its fastest-ever NPU, which runs at 85 TOPS (an upgrade from 45 TOPS). The TOPS focus represents a continuation of Qualcomm's strategy of relying on its NPU for AI tasks, which the company argues could deliver some of the best power-efficient AI compute in the industry. Raising the NPU to 85 TOPS also indicates that Qualcomm and its partners expect to fill that NPU with many concurrent workloads. The GPU is capable of AI compute as well, albeit not to the same degree as the NPU. Qualcomm's X2 family offers bus widths of up to 192 bits with peak bandwidths of 228 GB/s. This wider-than-standard bus, along with the highly performant chip, suggests that Qualcomm may be positioning itself to compete in the local inference space. This is further supported by reports that Qualcomm has been among the first to sample Samsung's LPDDR6X memories.[5]

[4] Apple, "Apple introduces MacBook Pro with all-new M5 Pro and M5 Max, delivering breakthrough pro performance and next-level on-device AI," March 3, 2026.
[5] Asif Iqbal Shaik, "Qualcomm chips could use Samsung's LPDDR6X memory next year," SamMobile, February 13, 2026.

Copyright © 2026 Moor Insights & Strategy

AMD's Ryzen Halo has been at the center of the company's client AI strategy, with early support for tools like OpenClaw on Radeon GPUs. The Ryzen Halo (Ryzen AI Max+) family represents the highest-performance chips available with both CPU and GPU on the same SoC. AMD claims that Qwen 3.5 35B can run on it at 45 tokens per second with a single agent.[6] The 256-bit bus widths and up to 256 GB/s bandwidth of the Ryzen AI Max+ further enable the cited local performance.[7] We expect AMD to continue producing the Halo line of products as the maximum AI configuration for its client devices, whether SFF desktops or high-powered laptops. AMD also recently announced Gorgon Halo with support for up to 192GB of local memory, further increasing bandwidth and capacity for AI tasks. The company is rumored to be developing a new line of Halo chips that would support 384-bit LPDDR6 configurations, delivering theoretical memory bandwidth of up to 691 GB/s.

In our own testing, we found the performance delta between the HP Z2 Mini G1a and the ZBook Ultra using the same chip was very narrow. This affirms that high-performance AI inference is possible in a mobile form factor, so long as there is enough cooling, which today requires a lot of fans and a thicker design — as in the Zbook Ultra.

Intel's Core Ultra 3 Series, codenamed Panther Lake, brings the company back into this competition, with a 50 TOPS NPU paired with an Xe3 GPU running up to 12 GPU cores. The GPU can reach up to 120 TOPS of AI performance, which is impressive, but the memory bandwidth lags the competition: Limited by a 128-bit bus and DDR5 memory, it has a maximum of 153.6 GB/s of memory bandwidth. Intel also hasn't really positioned Panther Lake as an AI powerhouse, and its maximum memory configuration is limited to 96GB, which constrains the bandwidth it can deliver and the size of the AI models it can handle for local inference.

NVIDIA has just one client device shipping today, which currently runs only on Linux; its Windows variant RTX Spark is slated to ship this fall. The DGX and RTX Spark are based on the GB10, which NVIDIA partnered with MediaTek to create. This platform is a dual-die single-chip design with eight memory modules — up to 128GB of memory — tightly placed next to the GB10 chip. NVIDIA claims 1 petaflop of FP4 AI compute, but in our own testing of DGX Spark, we found that it does a fairly good job of running 70B models and has no issues locally running 120B models such as gpt-oss-120b. It uses a 256-bit bus with LPDDR5x unified memory at speeds up to 9400 MT/s, delivering

[6] Usman Pirzada, "Run OpenClaw Locally On AMD Ryzen™ AI Max+ Processors and Radeon™ GPUs," AMD, March 13, 2026.
[7] AMD, "AMD Ryzen™ AI MAX+ 395 Processor: Breakthrough AI Performance in Thin and Light," March 17, 2025.

Copyright © 2026 Moor Insights & Strategy

roughly 301 GB/s of raw bandwidth. DGX Spark-based systems of varying performance levels are expected to appear in thin-and-light laptops in the near future for gaming and AI applications. NVIDIA is also expected to iterate on this platform down the road, possibly even with Intel's help, and we can envision it presenting a competitive platform in the 2027 or 2028 timeframe, though exact specifications are still unclear.

Some architectures are still limited by their 128-bit buses and will need to be rearchitected around 256-bit (or greater) buses to enable high-performance local AI. OEMs that want to address market needs and capitalize on forthcoming advances in client silicon will need to carefully account for the physical memory constraints that these updated architectures introduce as they refine their laptop designs.

OEM Design Impact and Thermal Design Responses

Beyond the work being done by chip makers, many leading PC OEMs are trying to balance the requirements of on-device AI inference with what's technically possible today in a thin-and-light laptop. Given the unique spatial demands for the memory required to enable high-performance local AI, there is a high probability that they will find that existing fan-based thermal solutions may not be sustainable for future laptops.

Based on the OEMs' traditional design logic, they must decide how much fan noise is acceptable while allotting the available SoC-adjacent memory placement space (<25mm) that a 14- or 16-inch laptop can afford. Apple has addressed this for its M5 Max MacBook Pros in part by using smaller, louder fans, but primarily by freeing SoC-adjacent space by placing memory directly on the package. OEMs that do not use unified memory must address thermal and spatial challenges by other means. When facing these challenges in the past, many OEMs compromised performance for thinness or compromised acoustics for thermal dissipation. Again, this is the traditional design logic at work.

However, given the unique spatial demands for memory required to enable high-quality local AI performance, it is highly likely that existing thermal designs will not be adequate for the future of inference-first computing. More specifically, cooling architectures built around heat pipes and fans may not be sufficient to meet the technical requirements for AI inference because areas within the chassis previously used for fans are likely to be crowded by memory, routing, and other necessary features. To adequately cool thin-and-light laptops running inference locally, fans would either need to disappear entirely, shrink significantly (and thus get louder), or be relocated in a way that fundamentally changes the thin-and-light design into something quite different.

Copyright © 2026 Moor Insights & Strategy

Emergent active cooling solutions may offer OEMs relief from their spatial and acoustic dilemma as the silicon advances to support truly performant local AI. Solid-state cooling offers compelling alternatives, including Frore's AirJet technology and Ventiva's ionic-cooling platform. Both of these help address the geometric constraints in the footprint around the SoC by offering smaller, more flexible cooling approaches than traditional fans. Ventiva's technology provides the added benefit of silent operation, while Frore offers ultra-quiet operation below the noise levels of traditional fans.

To meet the emerging requirements of this PC segment, both vendors will need to demonstrate cooling capability for thin-and-light laptop designs with TDPs ranging from 20W to 28W or higher. Ventiva also needs to ensure that designs have lower airflow impedance to account for its device's more modest pressure head, while Frore needs to find ways to overcome its limited flow rate when compared to other cooling solutions. There are indicators of success: Ventiva demonstrated its cooling solution at CES 2026 on an AMD platform running Microsoft's Phi-4 14B model at 15 tokens per second — a much larger model than AMD itself demoed using the same silicon that week.

Advanced vapor chamber designs such as ultra-thin loop heat pipes and two-phase vapor chambers offer other alternatives for dissipating heat with minimal z-height compromises, but these haven't traditionally scaled well beyond mobile applications or into devices using dozens of watts of power. While there are high-voltage vapor chamber designs capable of dissipating up to 2kW, those solutions cannot fit inside of a laptop or SFF desktop. Many of the ultra-thin vapor chamber designs also rely on novel materials such as graphene or advanced thermal interface materials (TIMs) such as liquid metal to transfer heat from the dies. Advanced materials like these may introduce additional cost or maintenance challenges down the road.

Many of these passive thermal advancements are niche solutions that don't necessarily address all the needs of future AI-enabled PCs. The vendors behind these solutions need to prove that they are scalable in manufacturing and have long-term viability for OEMs to consider them. For example, graphene has long been promised as a means for solving cooling problems, and it has been implemented as a novel material in many cooling solutions — yet virtually none of these graphene advancements have reached scale or broad adoption. Still, there is hope that some of these solutions will achieve manufacturing or cost breakthroughs, which would surely be welcomed by PC OEMs.

Copyright © 2026 Moor Insights & Strategy

Sovereign AI and Edge Infrastructure Considerations

Given the current limitation of high-end GPU exports from the United States, there is an increasing interest in local AI inference computing across the world, especially in China. As mentioned earlier, this has driven demand for devices such as Apple's Mac Studio. With the advent of better SoCs such as the M5 Max paired with appropriate memory configurations, we could see the premium AI-capable laptop market reach beyond the consumer electronics sphere. These devices could become critical agentic compute infrastructure and, when paired with other distributed local computing devices, enable considerable cost-effective AI computing. This could extend even to enterprise and sovereign settings — where on-device inference conveys added benefits for privacy as it enables the use of AI on sensitive data that need never traverse the cloud.

One of the biggest challenges in making this a reality is the spatial bottleneck posed by today's fan-based thermal solutions. Solve this challenge, and enterprises and sovereign buyers, especially those working under export restrictions, could drive more demand for a new class of machine that sidesteps cloud inference entirely.

Analyst Conclusions and Recommendations

Cooling is now a major gating factor for local inference in laptops. Denser memory architectures are encroaching on the chassis real estate that fans and heat pipes have historically occupied. Apple has shown that solving this at the architecture level, through unified memory on package, can unlock frontier-class local inference in a thin-and-light laptop form factor. Lacking an on-package unified memory path, however, other OEMs need a different answer. The unforgiving geometry of the thin-and-light form factor suggests that the answer likely cannot include fans, making performant and efficient on-device inference increasingly difficult to achieve with traditional design approaches; this should drive OEMs to consider alternatives to the traditional cooling model. The 2027–2028 window, when LPDDR6 and 256-bit-plus buses appear primed to go mainstream, looms as the deadline to find one.

As such, OEMs should elevate thermal design to a first-order focus. In parallel, silicon vendors should co-engineer reference designs with advanced thermal partners such as Ventiva to shorten time-to-market. The companies that lead the next era of edge AI in form factors geared for consumers, enterprises, and sovereign buyers will likely be those that resolve the thermal and geometric challenges posed by on-device AI.

Copyright © 2026 Moor Insights & Strategy

Important Information About This Paper

Contributor

Anshel Sag, Vice President and Principal Analyst, AR/VR/XR, 5G Mobility, PCs, Smartphones, Graphics

Publisher

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

This document, including any supporting materials, is owned by Moor Insights & Strategy. This publication may not be reproduced, distributed, or shared in any form without Moor Insights & Strategy's prior written permission.

Disclosures

Ventiva commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

© 2026 Moor Insights & Strategy. Company and product names are used for informational purposes only and may be trademarks of their respective owners.