Delivering Faster Local AI to the ASUS GB300 Workstation

KV cache offloading with SupremeIO™ on the ASUS ExpertCenter Pro ET900N G3

NOVEMBER 2026

© 2026 Graid Technology. All rights reserved.

Table of Contents

EXECUTIVE SUMMARY 2
A CAPABLE AI PLATFORM, CLOSE TO THE PEOPLE USING IT 3
WHY RETAINING CONTEXT MATTERS 3
THE SOLUTION INSIDE THE WORKSTATION 4
HOW THE COMPARISON WAS RUN 5
RESULTS: LESS WAITING AND HIGHER COMPLETION THROUGHPUT 6
WHAT THIS MEANS FOR LOCAL AI DEPLOYMENT 7
CONCLUSION 7
TEST SCOPE 7
SOURCES 7

© 2026 Graid Technology. All rights reserved.

Executive Summary

The ASUS ExpertCenter Pro ET900N G3 brings large-model AI development and inference to the deskside. Powered by the NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip, it combines up to 20 PFLOPS of AI performance with 748 GB of coherent CPU-GPU memory. Built on NVIDIA DGX Station architecture, the tower gives teams a substantial local platform for working with demanding models and keeping their data under their control. ASUS platform overview

For interactive AI, the next question is how quickly that platform can respond as conversations grow and requests arrive together. Reusing the model's previously computed context can shorten the wait before a response begins. A local NVMe tier gives the inference software somewhere to retain that reusable key-value (KV) cache beyond accelerator memory.

This paper examines how ASUS and Graid Technology bring compute and storage together for local AI on the ET900N G3. One M.2 SSD serves as the boot device. Graid Technology's SupremeIO™ combines three additional M.2 NVMe SSDs into a GPU-accelerated RAID 5 volume that LMCache uses to retain and retrieve reusable context. Together, they put local flash capacity to work for faster AI responses.

In the DeepSeek-V4-Flash-0731 multi-turn workload at concurrency 32, the offload configuration reduced mean time to first token (TTFT) from **12.08 to 5.00 seconds** and increased aggregate completion throughput from **56.89 to 79.79 output tokens per second**. These are improvements of **58.6%** and **40.3%**, respectively. Both runs completed all 2,574 requests successfully.

© 2026 Graid Technology. All rights reserved.

A Capable AI Platform, Close To the People Using It

The ET900N G3 combines a Grace CPU and Blackwell Ultra GPU through NVIDIA NVLink-C2C. Its coherent memory architecture allows the processors to work with a large shared memory pool, while the NVIDIA AI software stack supports local development and deployment. ASUS packages these capabilities in a deskside tower with four internal M.2 NVMe positions. ASUS specifications

For an engineering team, research group, or enterprise AI lab, this creates a practical setting for working with large models locally. Compute, memory, and storage reside in one system. Teams can develop an application, serve it through a familiar inference engine, and evaluate it against their own data and request patterns.

This study focuses on a specific opportunity within that platform: using its internal NVMe storage to support context reuse. The GB300 continues to execute the model. SupremeIO™ and LMCache add a local cache tier that the serving stack can use when previously computed context is needed again.

Why Retaining Context Matters

An AI assistant may receive the same system instructions with every request. A document assistant may revisit the same source material across several questions. An agent may carry conversation history through a sequence of decisions. In each case, some of the context has already been processed.

The KV cache holds attention state created during that processing. When a compatible cached prefix is available, the serving stack can reuse it and avoid repeating the corresponding prefill computation. LMCache supports storing and retrieving this state across tiers that include CPU memory and local storage. LMCache technical report

For workloads with reusable context, retrieving cached state can shorten the prefill work needed before a response begins. A high-performance NVMe volume gives the serving stack a local storage tier for retaining and retrieving that state.

On the ET900N G3, the three-drive SupremeIO™ volume provides a high-performance NVMe tier for this purpose. It extends the cache hierarchy with local flash capacity alongside the platform’s coherent memory.

© 2026 Graid Technology. All rights reserved.

The Solution Inside the Workstation

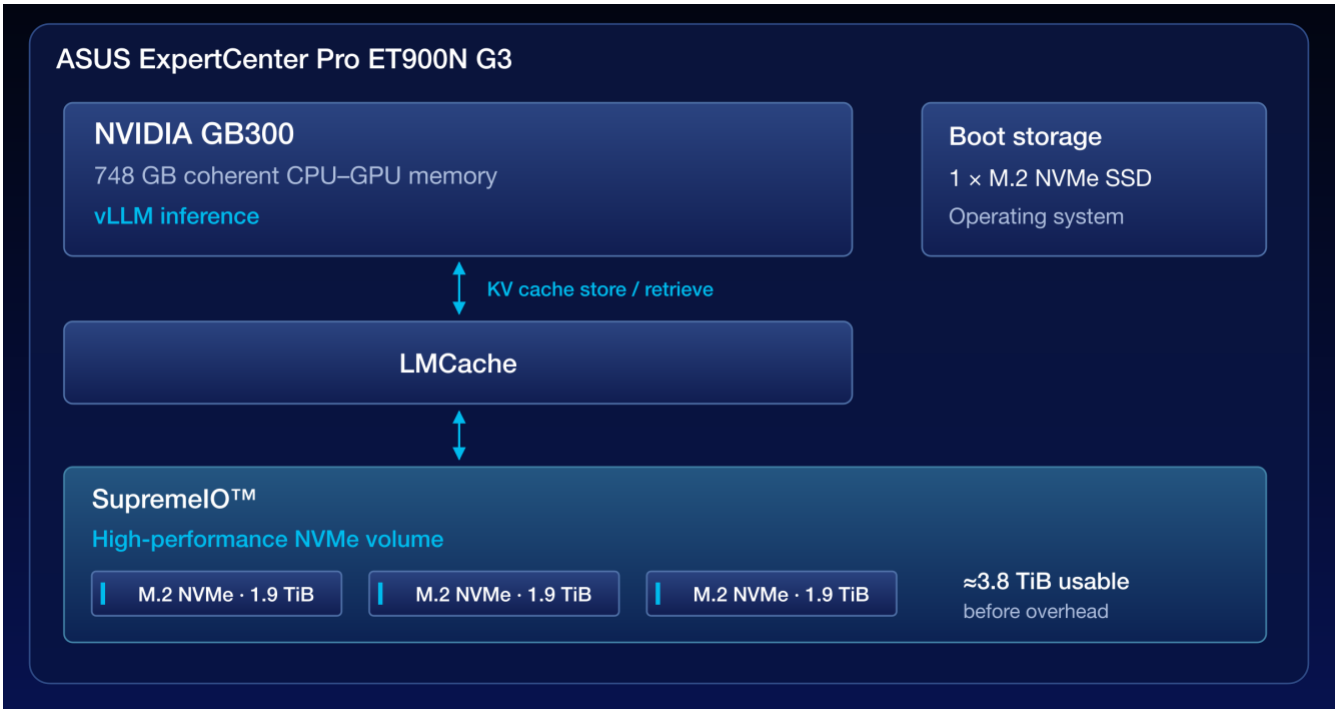


Figure 1. The ET900N G3 supplies the compute and memory foundation. LMCache manages reusable KV state on the local SupremeIO™ volume. Arrows represent logical cache movement. Source: supplied system information and test configuration.

SupremeIO™ is Graid Technology's GPU-accelerated, high-performance storage solution for AI workloads. It combines the workstation's three cache SSDs into one local NVMe volume for LMCache. LMCache manages reusable KV state, while SupremeIO™ provides the storage tier for retaining and retrieving it. In this configuration, vLLM and SupremeIO™ share the GB300 through NVIDIA Multi-Process Service (MPS).

Component	Tested configuration
Workstation	ASUS ExpertCenter Pro ET900N G3
Compute and memory	NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip; 748 GB coherent memory
Boot storage	1 × M.2 NVMe SSD
KV cache storage	3 × Samsung MZVLC2T0HBLD-00B07 M.2 NVMe SSDs; 1.9 TiB each
Cache volume	SupremeIO™ volume; RAID5; approximately 3.8 TiB before metadata and filesystem overhead
Serving stack	vLLM with LMCache in the offload case
Storage software	SupremeIO™ 2.0.1-1150.g5b994df2.bw103
Model	DeepSeek-V4-Flash-0731

The boot device remains separate from the cache volume. The architecture uses storage within the workstation and does not require an external storage appliance. This gives the serving stack approximately 3.8 TiB of local cache capacity before metadata and filesystem overhead.

© 2026 Graid Technology. All rights reserved.

How the Comparison Was Run

The test compared vLLM without KV cache offloading against vLLM with LMCache offloading to the SupremeIO™ volume. Both cases used the supplied EvalScope workload command.

Workload parameter	Setting
Dataset and seed	EvalScope random_multi_turn; seed 42
Concurrency	32
Traces and requests	256 traces; 2,574 evaluated requests
Turns	8-12 configured turns per trace
Prefix and prompt segments	4,096-token prefix; 256-512 tokens per generated prompt segment
Output limit	Up to 256 tokens per response
Streaming and warm-up	Streaming enabled; 32 warm-up requests

The multi-turn workload averaged approximately 25,390 input tokens and 36-37 output tokens per request. It exercises repeated context with short responses, a useful pattern for evaluating context reuse.

TTFT measures the wait before a response starts. Completion throughput measures aggregate output tokens produced per second across the run. The headline throughput result is the serving system's aggregate output rate.

The comparison measures the end-to-end benefit of enabling LMCache offloading to the SupremeIO™ volume on the ASUS GB300 workstation.

© 2026 Graid Technology. All rights reserved.

Results: Less Waiting & Higher Completion Throughput



Figure 2. Mean TTFT was 58.6% lower and aggregate completion throughput was 40.3% higher with LMCache offloading to the SupremeIO™ volume. Each bar represents one run at concurrency 32. Source: EvalScope benchmark reports for the tested configurations.

The average wait for the first token fell by about seven seconds. At the same time, the serving system produced more output tokens per second across the workload. For applications built around repeated context, this combination points to a useful improvement in both response start time and overall serving throughput.

Supporting measure	No KV cache offload	SupremeIO™ + LMCache offload	Observed change
Request throughput	1.56 req/s	2.20 req/s	41.0% higher
Mean request latency	19.375 s	14.297 s	26.2% lower
Mean first-turn TTFT	11.424 s	4.217 s	63.1% lower
Mean subsequent-turn TTFT	12.150 s	5.084 s	58.2% lower

The improvement in subsequent-turn TTFT is particularly relevant to conversational applications: the observed benefit extended beyond the first response. Mean first-turn TTFT improved by 63.1%, while mean subsequent-turn TTFT improved by 58.2%.

© 2026 Graid Technology. All rights reserved.

What This Means for Local AI Deployment

For teams already using the ET900N G3, SupremeIO™ turns internal NVMe capacity into a high-performance storage tier for AI context reuse alongside GB300 compute. The measured outcome is straightforward: responses began sooner on average, and the system completed more output work per second under the tested load.

The approach is most relevant to applications with reusable context, such as assistants with shared instructions, document analysis with repeated questions, and agents that revisit conversation history. Teams developing these applications can evaluate the same local cache approach against their own models and request patterns.

The solution makes use of the workstation's internal M.2 storage, with a dedicated boot device and a separate high-performance cache volume. This keeps the deployment within one system and gives teams a clear path to evaluate KV cache offloading on their existing ASUS GB300 platform.

Conclusion

ASUS and Graid Technology address complementary requirements for local AI. The ASUS ExpertCenter Pro ET900N G3 provides a substantial foundation for local AI: GB300 compute, a large coherent memory pool, and internal NVMe storage in one deskside system.

Adding LMCache offloading to a three-M.2 SupremeIO™ high-performance volume improved the supplied multi-turn test in two practical ways: **58.6% lower mean TTFT** and **40.3% higher aggregate completion throughput**. The result shows how high-performance local storage can complement the ASUS platform's compute and memory, helping applications that repeatedly work with shared context respond sooner and serve more output.

To learn how the ASUS ExpertCenter Pro ET900N G3 can support your local AI workloads, contact ASUS for a pre-sales consultation.

For more information about Graid Technology and storage solutions, visit the Graid Technology website.

Test Scope

Results reflect one run per configuration of DeepSeek-V4-Flash-0731 on the ASUS ExpertCenter Pro ET900N G3 at concurrency 32. The reported improvements apply to the named mean-TTFT and aggregate-throughput metrics for the combined LMCache and SupremeIO™ offload configuration. Results may vary with model, software settings, context reuse, and workload. These figures are not per-request latency guarantees.

Sources

ASUS ExpertCenter Pro ET900N G3 overview and specifications

ASUS ET900N G3 launch announcement, June 15, 2026

LMCache technical report

© 2026 Graid Technology. All rights reserved.

ABOUT GRAID TECHNOLOGY

Graid Technology keeps AI infrastructure moving. We remove the bottlenecks that slow training, inference, and every workload in between, ensuring resilient, scalable performance across any deployment. Headquartered in Silicon Valley with global operations and R&D in Taiwan, Graid Technology keeps the infrastructure layer moving for the AI era. Visit graidtech.com.

© 2026 Graid Technology. All rights reserved.