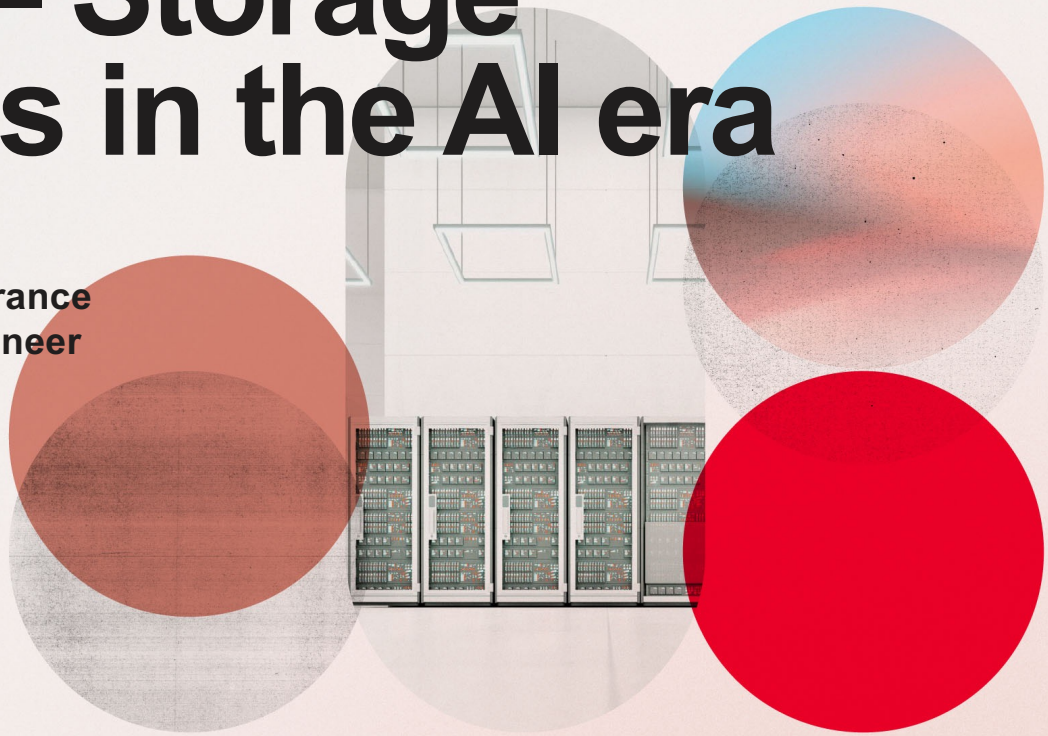
ddn

ddn.com

Teratec 2026 – Storage infrastructures in the AI era

Laurent Thiers – VP Sales, Head of DDN France
Emilien Tellier – Profesional Services Engineer





Training, a new traditional workload ?

Training and fine tuning

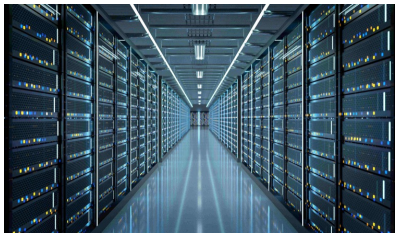
© Copyright 2026, DDN, Inc. All rights reserved.

Training and fine tuning - An HPC inheritance ?

AI training workloads are **similar to already well-known HPC workloads** but at a **new pace** and a **new always-increasing scale**.

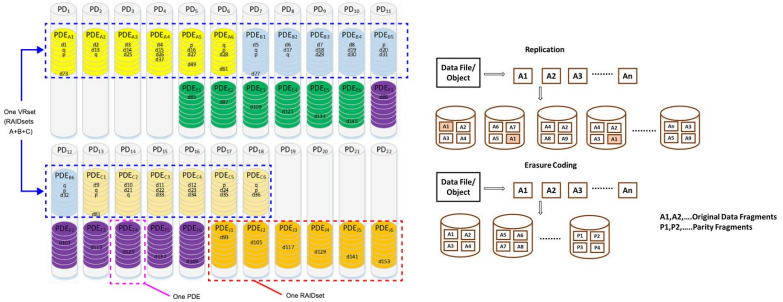
Data center footprint

- Low latency network footprint (ports & transceivers counts)
- Power efficiency, GB / s / Watt
- Rack units



Data protection

- Layout efficiency
- Redundancy levels
- Erasure coding & dynamic protection



Performance

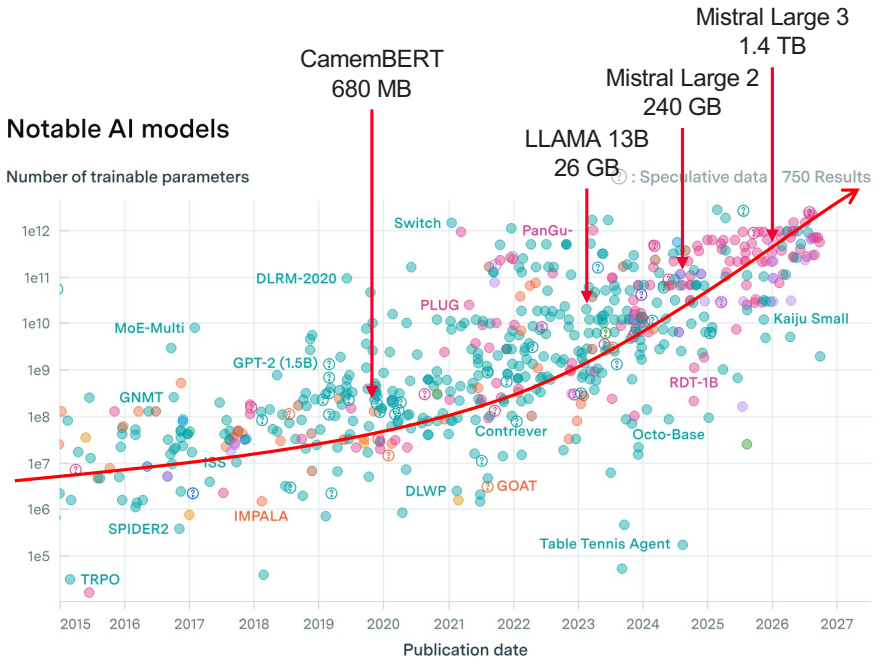
- Sequential
- Latency
- GB / s / Rack Unit

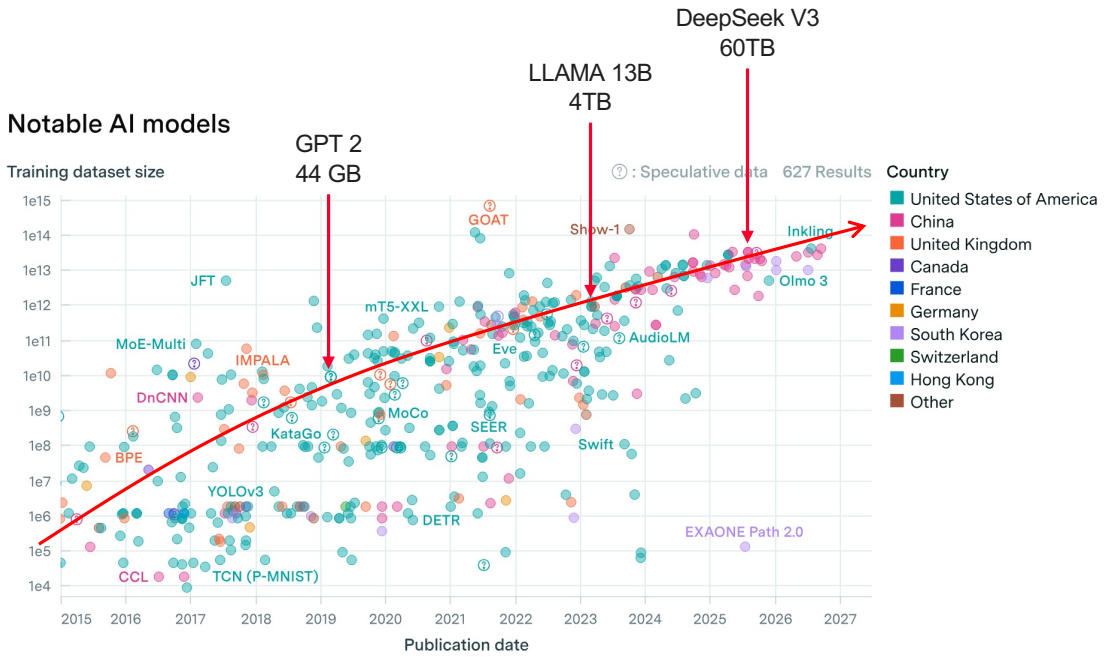


ddn

© Copyright 2026, DDN, Inc. All rights reserved.

Training and fine tuning - An HPC inheritance ?





The storage needs to **follow the increasingly demanding needs** of these already familiar workloads.

© Copyright 2026, DDN, Inc. All rights reserved.

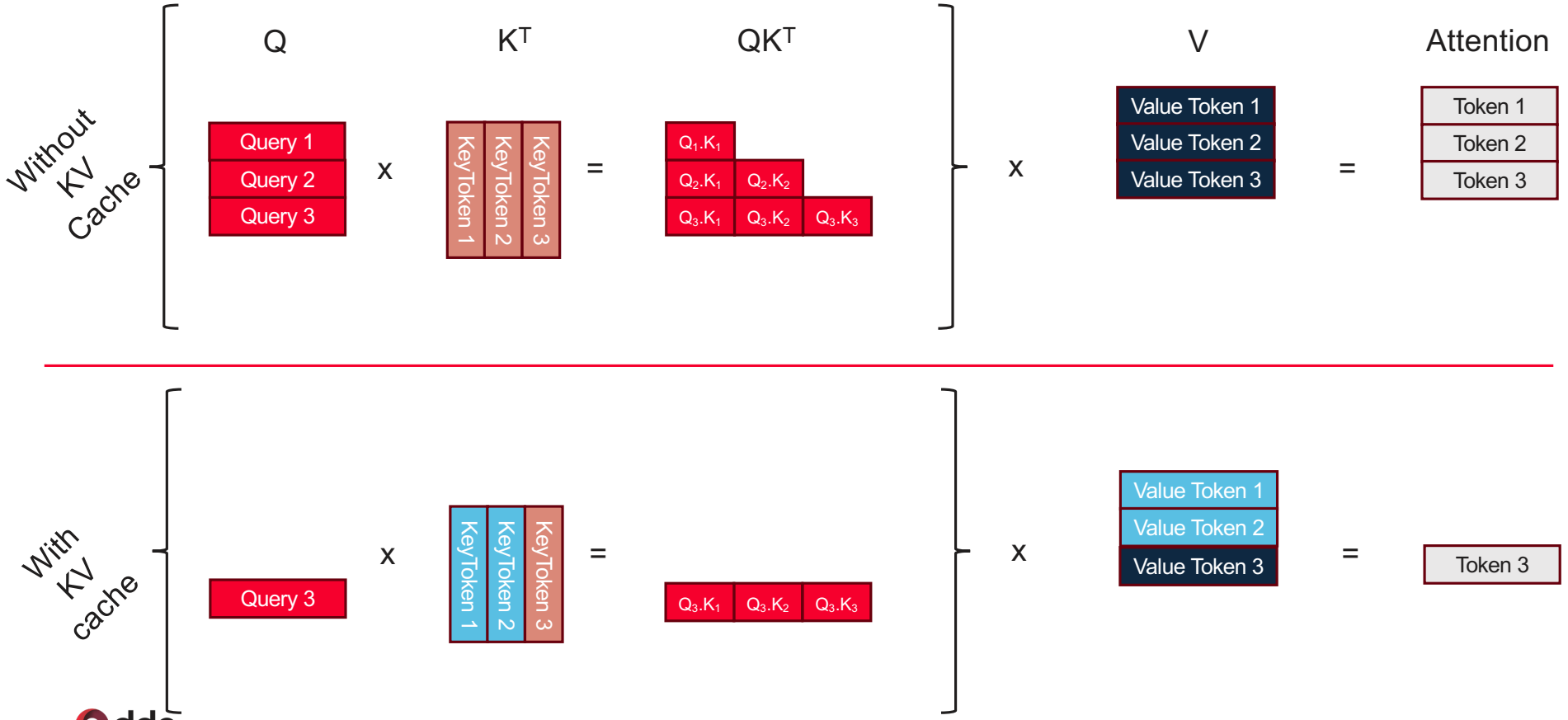


Storage as a vector of AI agentic inference performance

Inference and Key-Value cache

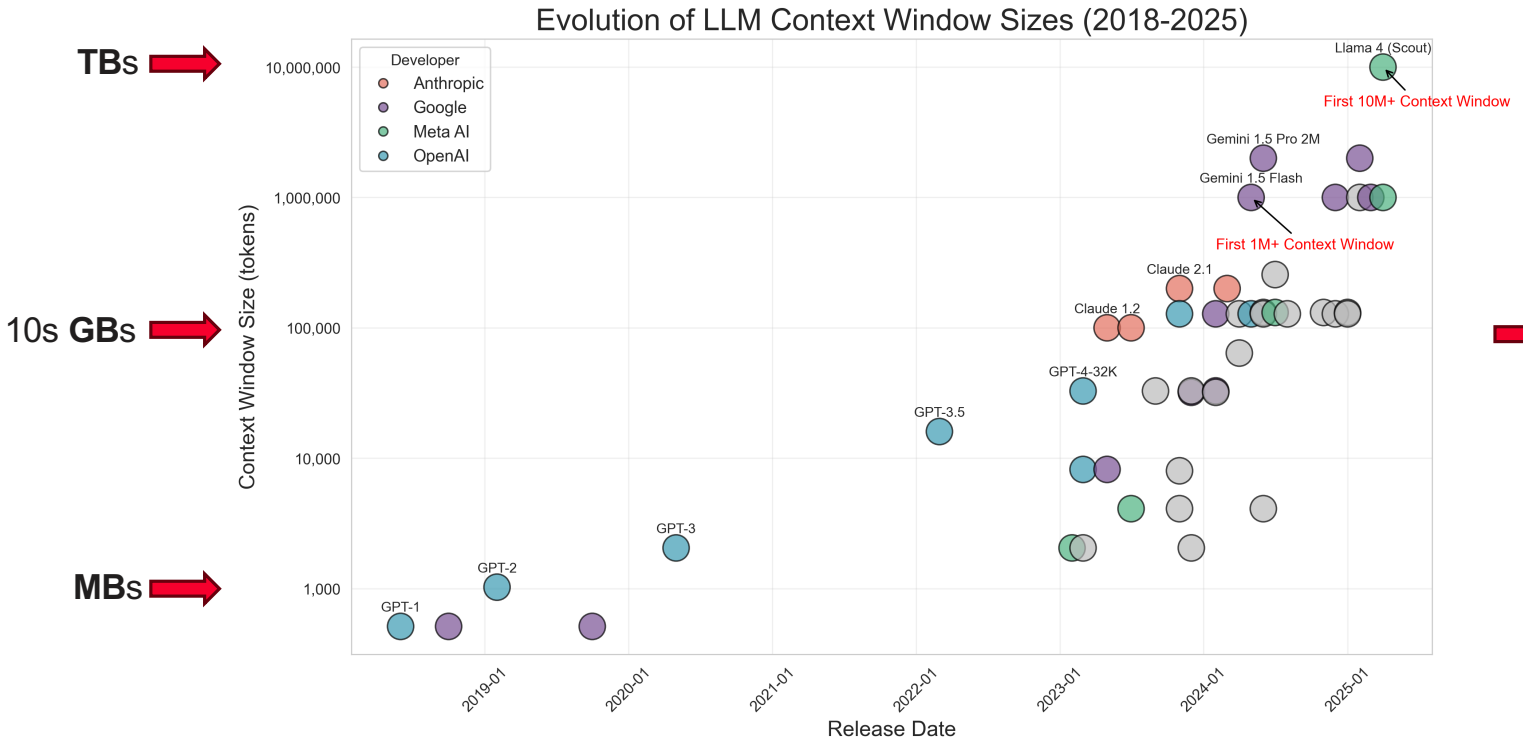
© Copyright 2026, DDN, Inc. All rights reserved.

Inference and Key-Value cache – Quick Introduction



© Copyright 2026, DDN, Inc. All rights reserved.

Inference and Key-Value cache – The context size issue

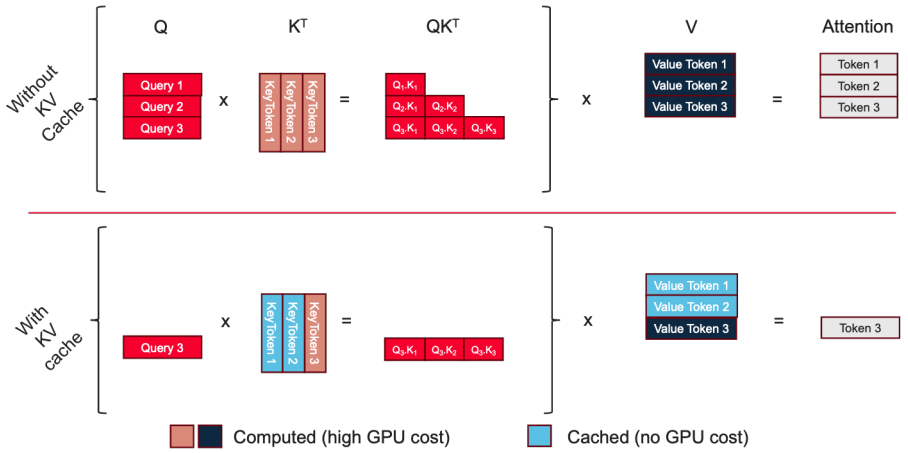


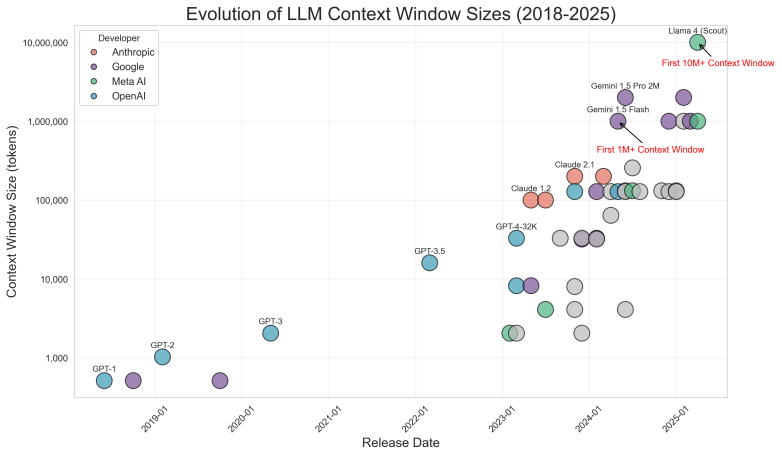
Context's memory consumption increases

Source: https://www.meibel.ai/post/understanding-the-impact-of-increasing-llm-context-windows

© Copyright 2026, DDN, Inc. All rights reserved.

Inference and Key-Value cache – Two phenomenon, one issue





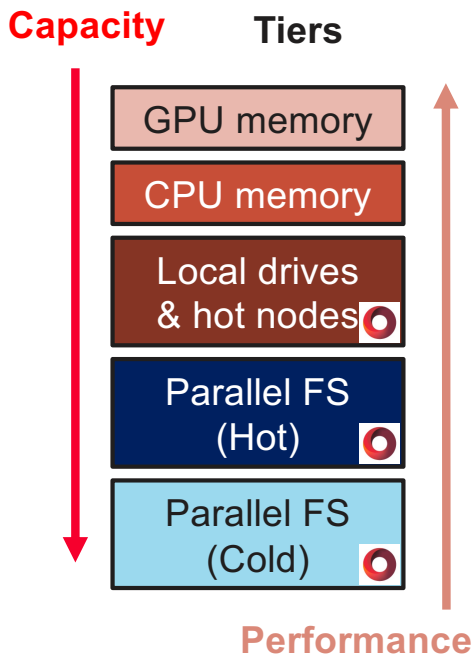
KV cache is allocated in GPUs HBM

Increasing context size doesn't fit in memory anymore

Issue :
We're running out of memory.

© Copyright 2026, DDN, Inc. All rights reserved.

Inference and Key-Value cache – The solution, a hierarchized memory

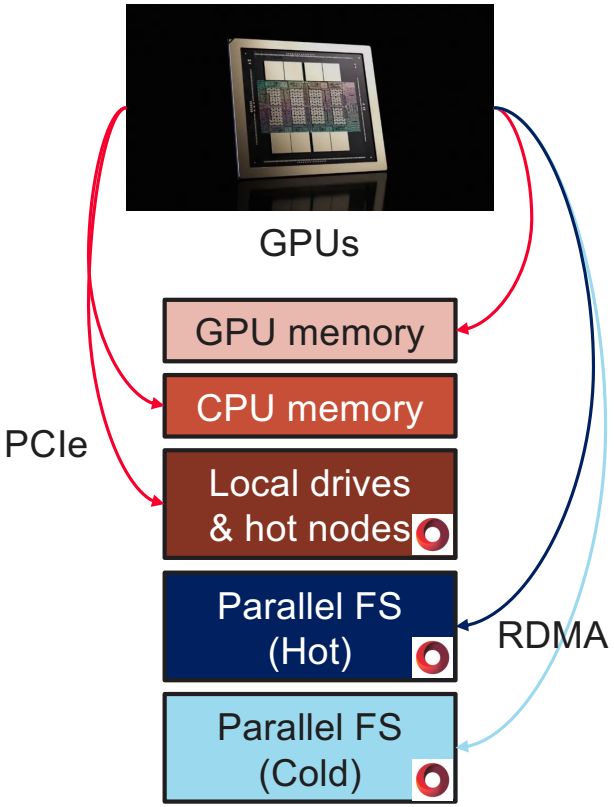


Usage	Locality
On KV generation	**GPU-local** cache
Automatic **spillover**	**Node-local** cache
Hot KVs cache	
Hot KVs cache	**Shared** cache for **all cluster's GPUs**
Cold KVs cache (ended sessions, disconnected users)	

The **storage solutions need to be performant enough** to enable a **speedup compared to KV-cache recomputation**.

© Copyright 2026, DDN, Inc. All rights reserved.

Inference and Key-Value cache – An already-proven solution



Tested on a GB200 platform, NVIDIA software stack :
Dynamo, vLLM, LMCache & NIXL (POSIX)

On a sovereign cloud, using a sovereign foundation model (Mistral Large)

Time to first token divided up to 15x.

Workloads duration reduced up to **60%**, as many **more requests served in the same timeframe**.

Reduces up to 62% the energy footprint of the workload (including storage's footprint).

ddn

© Copyright 2026, DDN, Inc. All rights reserved.



Merci!

Emilien Tellier – etellier@ddn.com

ddn.com

© Copyright 2026, DDN, Inc. All rights reserved.