WHITE PAPER

European Data Platform for HPC and AI: A Path to Data Sovereignty

April 2026

An SRA 6 White Paper



ETP 4 HPC

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

Building the next generation of data-intensive HPC and AI infrastructures: a position paper

Co-leaders

Dr. Jean-Thomas Acquaviva (DDN)

Prof. Dr. -Ing André Brinkmann (Saarland University)

Table of contents

Executive Summary..... 3

Glossary of terms..... 4

1 Introduction..... 6

2 Reinventing Storage in Europe – Why Now?..... 8

3 AI storage workloads and their relation to HPC 10

4 KV Caching based on Ad-hoc File Systems 16

5 Landscape of the Data Economy..... 18

Data supply chain and data logistics 18

Data Market Places 18

Data catalogues.....19

Private data management.....19

Data management 20

6 Executing the Storage re-invention..... 21

7 Energy and Sustainability..... 22

Data reduction and data gravity.....22

Device efficiency and co-design with applications.....23

AI applied to AI infrastructure.....24

8 From Fragmentation to Consolidation 25

9 Conclusion and call to action..... 26

European Converged Data Platform Pilots: Innovation Program for a Pre-market Solution.....26

European Data Factory Transparency Initiative 26

European Sovereign Open Hardware: Research Program for future technological autonomy.....27

Contributing Authors 28

Executive Summary

This is a white paper released as part of the ETP4HPC's Strategic Research Agenda 6.

This White Paper argues for the need for a renewed focus on Storage and I/O needed in Europe, which has been largely ignored despite the critical importance of storage and data management for AI applications and Large Language Models. The paper presents the need for building a European Data Platform suitable for Exascale HPC and AI whilst keeping the needs of European technological sovereignty firmly in context. The paper presents technical directions and policy recommendations for the above.



Glossary of terms

Ad-Hoc File System - file systems that are ephemeral and spin off on demand, by a job, an application, or a project.

ADMIRE - grant agreement 956748, a EuroHPC project aiming at optimizing data movement and I/O using ad-hoc file system, malleability and intelligent I/O controller

(a)synchronous - Asynchronicity is the property of an I/O to be initiated and completed in the background without blocking computing. On the contrary, synchronous I/O blocks the CPU up to the completion of the I/O. Synchronous I/O is more efficient for data movement but at the cost of idle cycles for the computer. Asynchronous I/O optimises compute efficiency, but is harder to implement and reduces system predictability.

Attention Head - processing component within an AI model that focuses on one specific aspect of a text (such as grammar or context). Multiple heads work in parallel on the exact same input to build a multi-dimensional understanding of input data.

Attribute-Based Encryption - cryptographic approach where access to encrypted data is dependent on the user possessing specific attributes (e.g., role, organization), enabling fine-grained data sharing across distributed marketplaces.

Checkpointing - process consisting of dumping critical data structures from memory to persistent storage. By definition, a checkpoint can be read and is sufficient to allow the application to resume in a clearly defined state. In AI, checkpoint size depends on the size of the AI model. In HPC, the rule of thumb to estimate the checkpoint size is to consider half of the compute memory multiplied by the number of compute nodes involved in the parallel process.

Closed Source Model - e.g. Gemini, by opposition to open-source AI models such as Llama, Mistral, or Deepseek.

Computational Storage - recent evolution in the storage industry, under normalization by the SNAI, to allow storage devices or storage systems to perform computation on behalf of compute nodes. The goal is to reduce data movement by performing simple computation closer to the data source.

Context Window - maximum number of continuous tokens (words or sub-words) an AI model can ingest and "remember" at one time during a single prompt or conversation.

CXL, Compute eXpress Link - an open-standard of high-speed interconnect built on the PCIe standard that allows CPUs, GPUs, and memory devices to communicate with low latency while ensuring memory coherency.

Data logistics - operational orchestration of data movement across distributed infrastructures. Data logistics is complex due to the requirement to handle multiple environments with their own access and security policies, provide resilient connections, and deliver data at a high pace.

Dataset - structured collection of data serving as the fundamental unit of information for training, fine-tuning, or validating AI models.

DeepSeek Model - family of large language models known for specific architectural optimizations, such as using Multihead Latent Attention (MLA), which drastically reduces their KV cache memory footprint during inference.

DMTCP - Distributed MultiThread Checkpointing, is a tool to transparently checkpoint the state of multiple simultaneous applications, including multi-threaded and distributed applications. It operates directly on the user binary executable, without any Linux kernel modules or other kernel modifications.

Federated Learning - decentralized machine learning approach that allows models to be trained across multiple independent facilities holding local data, without the private data itself ever needing to be exchanged. Only the intermediate versions of the models are communicated across facilities.

Giga-Factory - the term used by the EU for large-scale supercomputer facilities with up to

100,000 AI-optimised chips specifically used for the development and training of next-generation AI models containing trillions of parameters.

GPFS - General Parallel File System (now known as IBM Storage Scale), is a high-performance clustered file system developed by IBM.

Grouped Query Attention - optimisation in transformer architectures that shares Key and Value representations across multiple Query heads, which helps reduce the memory bandwidth and KV cache size required during inference.

HBF, High Bandwidth Flash - emerging storage architecture designed to provide extremely high-throughput, low-latency access from GPUs directly to NAND flash, acting as a scalable extension of the memory hierarchy.

In-Situ analysis - practice of processing, filtering, or analysing data directly in the same location (typically the group of compute nodes involved in the computation) to minimize data movement over the network.

IO-SEA - grant agreement 955811, a EuroHPC project that focused on delivering a novel data management and storage platform for exascale computing, utilizing hierarchical storage and ephemeral ad-hoc file systems.

KV-Cache, Key-Value Cache - critical memory optimization technique used during LLM inference. It stores previously computed Key and Value vectors for past tokens so the system does not have to redundantly recompute them when generating the next token.

Llama-3 Models - family of open-source large language models developed by Meta, frequently used as benchmarks for performance and token context sizing.

LLM, Large Language Model - foundational AI model based on transformer architectures, trained on massive datasets to understand, process, and generate human-like text.

Lustre - open-source, distributed parallel file system designed originally for HPC. Lustre has been successful in AI, specifically for training systems.

Open-Source Model - AI model whose architecture and weights are made publicly available, allowing users to deploy and modify it independently on their own sovereign infrastructure (e.g., Llama, Mistral).

Malleability - ability of a computing workload to dynamically scale its resource usage (compute or storage) up or down at runtime, improving energy efficiency and overall system utilization.

Mistral Model - family of highly efficient open-source large language models developed by the European AI company Mistral AI.

Models - mathematical representations and parameter weights generated by machine learning algorithms after being trained on datasets, used to make predictions or generate content.

MLA, Multihead Latent Attention - specialized attention mechanism that compresses Key and Value vectors into a latent space, heavily reducing the KV cache size required during inference (notably used in DeepSeek models).

NAND flash - type of non-volatile storage technology that retains data without power, serving as the foundational building block for modern solid-state drives (SSDs) and high-performance storage systems.

RAG, Retrieval Augmented Generation - AI framework that improves the accuracy of an LLM by retrieving relevant facts from an external database and injecting them into the model's context window before it generates a response.

S3, Simple Storage Service - the de facto industry standard to store and manage objects, it has been designed by Amazon Web Service.

SCR, Scalable Checkpoint Restart - an HPC software library that enables fast, multi-level checkpoints across different storage tiers.

Token - the basic unit of data processed by an LLM, which can represent a whole word, part of a word, or a single character. Typically, 32 bits (the equivalent of 4 letters in text).

1 Introduction

This White Paper is based on our shared experience working in several EuroHPC innovation projects and advocates learning from these projects on building large-scale storage systems for exascale HPC and AI systems to enable an efficient data management lifecycle. The motivation for this white paper is that massive data volumes generated and managed by exascale systems for HPC and AI applications, along with their associated bandwidth and latency demands, are placing unprecedented pressure on storage infrastructures. When storage systems cannot deliver data to compute resources quickly enough, CPU and GPU utilization declines, immediately reducing overall system performance.

The ETP4HPC Strategic Research Agenda 6 already highlighted in 2024 that effective data management is critical for running AI workloads at full capacity and for fully leveraging data across the entire AI lifecycle:

> *"We recommend re-invigorated efforts in the area of data storage, data management and I/O innovations in light of AI developments in the upcoming work programmes, which seems to be completely lost after the H2020 programs. This will be a critical centre piece in the AI revolution."*

This is fully aligned with a declaration of Jensen Huang during NVIDIA GTC in March 2025:

> *"Remember, computing has three pillars. There's computing, you're looking at it. There's networking. As I mentioned earlier, Spectrum-X is going to be the world's enterprise and AI network. The third is storage. Storage has to be completely reinvented."*

In short, infrastructures such as HPC exascale computers, AI Factories, and future Giga factories will only be able to deploy their full potential if the data management aspects are considered thoroughly. The critical importance of storage is highlighted by industry leaders and academic experts. Yet, unlike initiatives in the United States and China, current European innovation programmes largely ignore the pivotal role of storage systems for exascale-size type systems for HPC and AI.

In this white paper, we identify key challenges and propose actionable solutions through which Europe can build on prior EuroHPC results to become a leading player in next-generation HPC and AI infrastructure. Central to our vision are **large-scale storage systems**, engineered not only to serve as production backends for Europe's most demanding AI workloads but also to function as dynamic research platforms that drive innovation in storage for AI. These systems must be co-located with the most powerful publicly-funded HPC and AI infrastructures to ensure they are always sized for, and directly coupled to, the largest and most recent applications running in Europe. They must operate as true production environments, equipped with state-of-the-art data management, monitoring, and observability tools.

Continuous trace collection and lightweight observation of live workloads will enable rapid detection and response to any shifts in HPC and AI-driven storage demands, ensuring the infrastructure evolves in lockstep with the applications it serves. To support distributed AI workflows, these large-scale storage facilities will be interconnected in a next step across Europe, enabling federated learning while preserving data sovereignty.

Each facility will be operated by a consortium combining established industry players for infrastructure stability, European start-ups for innovation and agility, and academic partners for cutting-edge research, ensuring high TRL while maintaining dynamic adaptability. Together, this model transforms storage from a bottleneck into a strategic enabler. By deploying such large-scale, production-grade storage systems, which are co-located, interconnected, and research-integrated, Europe can significantly boost the efficiency of its AI and HPC ecosystems, reduce reliance on non-European

hardware and software, and create fertile ground for European AI start-ups to thrive. The focus is clear: without sovereign, scalable, and intelligent large-scale storage, Europe's ambitions in AI and exascale computing cannot be fully realized.

Following this introduction, we first establish why storage reinvention is urgent now, framing AI as a pivotal inflection point comparable to historical technological shifts like the introduction of cloud computing and mobile apps, and crucially, how evolving storage demands intersect with recent EuroHPC investments to create unique European opportunities. We then demonstrate how AI can leverage hard-won HPC storage lessons to avoid past pitfalls, and concrete use cases illustrate where AI's storage requirements strain current infrastructures, from distributed training to real-time inference. Finally, we present a scalable architecture emerging from EuroHPC projects: an open-source, purpose-built file system stack designed explicitly for AI's data velocity and sovereignty needs, turning legacy constraints into innovation catalysts. The paper concludes with a call to action, aligned with the first priority of European Commission's AI continent plan[1], to target investments that turn a vision into a sovereign, production-grade infrastructure.



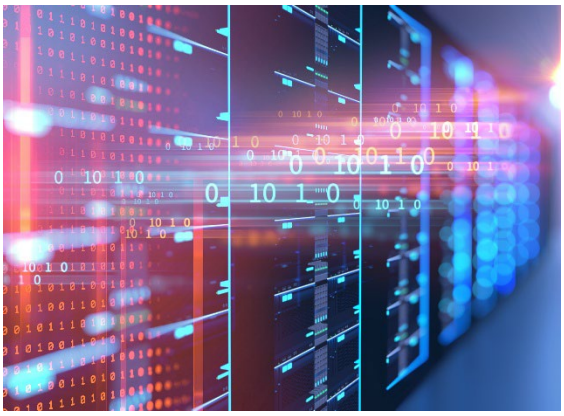
[1] https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan

2 Reinventing Storage in Europe – Why Now?

AI represents a fundamental discontinuity in the technical evolution of computing, comparable in scale and impact to the dot-com boom of the 1990s, the mobile-app revolution triggered by the iPhone in 2007, and the cloud-driven consolidation of the 2010s. Each of these inflection points not only reshaped user expectations and application architectures but also exposed the limitations of legacy infrastructure, forcing a rethinking of foundational technologies. Storage, in particular, has historically evolved in response to such disruptions: the rise of massively parallel systems in the 1990s gave birth to GPFS and Lustre; the cloud era demanded the creation of scalable, API-driven object storage like S3; and the adoption of SSDs compelled a re-engineering of hierarchical storage management to match new performance profiles. Today, AI's insatiable appetite for data, coupled with its unique access patterns, scale, and real-time demands, is exposing the inadequacies of storage systems designed for earlier paradigms.

This moment is not just an opportunity; it is a necessity.

Europe's AI and HPC infrastructure remains heavily dependent on non-EU storage technologies[2], a vulnerability amplified by NVIDIA's expanding dominance across GPUs, CPUs, networking, and AI software. With its GPU-Direct and DPU strategies, NVIDIA is extending its reach directly into the storage domain, while major OEMs such as VAST, DDN, and Weka are aligning their hardware and software ever more tightly with NVIDIA's ecosystem. This deepening integration raises the risk of vendor lock-in and exposes Europe to global supply chain fragility. In contrast, developing EU-owned storage technologies would reduce these geopolitical risks, enable genuine choice between NVIDIA, AMD, and emerging European chipmakers, and, critically, keep value creation, high-skill jobs, and innovation firmly anchored in Europe.

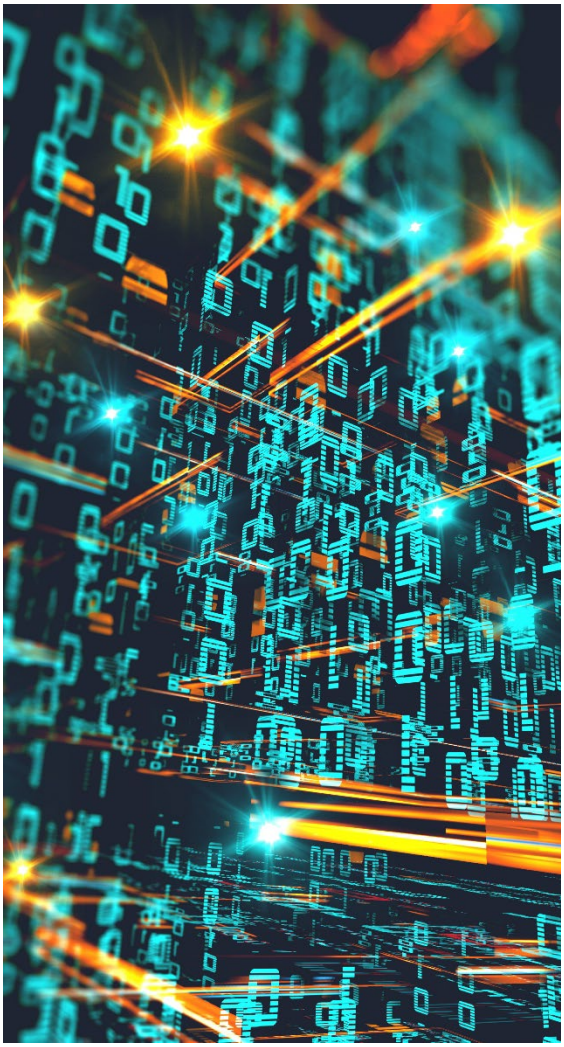


At the same time, the current hardware landscape offers concrete levers that Europe can exploit to build a sovereign, future-ready storage stack. AI frameworks typically require only relaxed storage semantics and benefit enormously from read-only caching, meaning that a lightweight, semantics-aware caching file system can already deliver most of the required performance. Modern AI nodes embed NVMe devices, providing a native foundation for distributed, high-performance data caching in front of external storage arrays. GPU-Direct allows accelerators to address NVMe media directly, bypassing the CPU and eliminating a traditional bottleneck in real-time, data-intensive workloads. Moreover, data processing units (DPUs), not only from NVIDIA, are becoming widespread; these programmable off-load engines can take over storage, networking, and security functions, freeing CPUs and GPUs for pure AI computation. Therefore, the strategic opportunity arises that by leveraging relaxed

[2] An analysis of sovereignty and AI infrastructure is available in the white paper "A European model for artificial intelligence" https://op.europa.eu/en/publication-detail/-/publication/e13d1844-e462-11ef-be2a-01aa75ed71a1

semantics, built-in NVMe, GPU-Direct, and DPU capabilities, Europe can design storage subsystems that are efficient, independent of external vendors, and ready for the exascale AI workloads of tomorrow.

The storage needs of AI are, in many respects, a direct extension of challenges already solved for HPC. Europe has accumulated deep experience in open-source storage software for high-performance environments: it has contributed to and deployed the Lustre parallel file system at scale, has built hierarchical storage solutions that incorporate ephemeral caching layers, has designed innovative key-value (KV) storage systems, and has pioneered ad-hoc file-systems tailored to specific scientific workloads. This know-how was consolidated in EuroHPC projects such as ADMIRE, IO-Sea, and EUPEX which delivered near production-grade, open-source storage stacks and demonstrated how to orchestrate exabyte-scale I/O under strict performance constraints. This knowledge base exists now and can be readily repurposed to construct large-scale AI storage systems, allowing Europe to turn proven HPC expertise into a decisive competitive advantage for AI. By coupling that expertise with the hardware shifts described above, Europe can create a storage ecosystem that is both sovereign and future-proof.



3 AI storage workloads and their relation to HPC

The field of HPC continues to scale rapidly and along a well-understood trajectory, growing in size, complexity, and capability with each generation. HPC workloads themselves are largely predictable and well-characterized, and the transition from monolithic applications to workflow- and service-oriented architectures is progressing steadily, without disruptive upheaval. Flash-based storage and tiered storage architectures have been mature and widely deployed for over a decade, enabling stable, high-performance I/O across most traditional HPC use cases. As a result, the most pressing and rapidly evolving storage challenges today do not stem from HPC alone but from the emergence of AI workloads, which impose novel, dynamic, and exponentially growing demands on data infrastructure that legacy systems were never designed to handle.

AI workloads can be divided into two broad strands: training for model building and inference for model execution.

Training: Early AI training workloads were read-dominated: model checkpoints were small and infrequent, placing minimal write pressure on persistent storage. Starting around 2021, publications from US national labs depicted a more balanced read/write footprint for ML workloads running on huge supercomputers[3]. The surge of Large Language Models (LLMs) dramatically increased this trend, and now there is a critical need for protecting GPU training with checkpointing, especially as training scales to thousands of GPUs across distributed systems, where even a single node failure can derail days or weeks of computation.

Checkpointing is primarily a risk management and operational strategy during the training of an LLM. The probability of a hardware failure or infrastructure interruption when coordinating hundreds or thousands of GPUs over weeks or even months is very high. The mitigation strategy, well known in HPC, is to periodically save a complete snapshot of the model's weights and optimiser state to persistent storage. This checkpointing acts as an insurance, allowing to resume training from the most recent checkpoint after a crash, rather than suffering the catastrophic loss and restarting from scratch. Frequent checkpointing ensures progress is preserved and enables efficient recovery, making it indispensable for large-scale, long-running AI training jobs.

In addition, there is an intrinsic need for checkpointing in the specific case of training. These snapshots provide crucial opportunities to evaluate the model's progress mid-training, mitigate overfitting by allowing a reversion to an earlier, more stable state, and serve as branching points for future fine-tuning experiments.

Checkpointing in AI is essentially identical to the one that classical HPC has been solving for years[4]. Large-scale scientific simulations have long relied on coordinated checkpoint/restart to safeguard multi-day runs on thousands of nodes, using parallel I/O to a shared, POSIX-compliant file system, asynchronously staging data in burst-buffer or node-local NVMe, and then flushing it to permanent storage in a coordinated fashion.

When a failure occurs, however, the naïve approach of letting every rank reload its checkpoint simultaneously creates a sudden traffic spike that can saturate network bandwidth, exceed storage-throughput limits, and overwhelm connection resources. To tame these spikes, in hierarchical distribution schemes, a small set of leader nodes first fetches the checkpoint from the parallel file system and then fans out the data to the remaining ranks, dramatically reducing contention on the shared fabric. In addition, multi-level checkpointing

[3] A. Paul, A. Karimi, F. Wang: Characterizing Machine Learning I/O Workloads on Leadership Scale HPC Systems. 29th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Houston, TX, USA, November 3-5, 2021, 1-8.

[4] V. Sistla, Y. Niu, H. Basudan, and S. Devabakthini: AWS Storage Blog: Architecting scalable checkpoint storage for large-scale ML training on AWS, 2025.

spreads checkpoint data across storage tiers. High-frequency checkpoints reside on node-local NVMe or burst-buffer, while full, coarse-grained checkpoints are written less often to the global file system. This tiered strategy preserves resiliency while keeping I/O traffic within the limits of the underlying hardware.

Modern AI training pipelines adopt the same pattern: each GPU cluster writes a local copy of the checkpoint to fast NVMe, then a collective "stripe-out" operation transfers the data to a parallel file system. In practice, the dominant storage backends for AI training on European supercomputers are HPC-grade file systems such as Lustre, IBM Spectrum Scale (GPFS), and BeeGFS. Their ability to stripe data across many storage nodes supplies the aggregate bandwidth required for huge checkpoint streams.

Critically, this convergence means that HPC storage architectures can form the backbone of European AI training. Open-source parallel file systems, already deployed at scale in EuroHPC facilities, provide the ideal foundation for a continent-wide AI infrastructure. Their POSIX compliance, petabyte-scale resilience, and vendor-agnostic design eliminate dependency on proprietary US or Chinese storage stacks while delivering the aggregate bandwidth required for exascale checkpoint streams. Unlike closed Original Equipment Manufacturer (OEM) solutions tightly coupled to NVIDIA ecosystems, these open technologies enable Europe to scale independently across heterogeneous hardware (AMD, RISC-V, EU chipmakers), retain full data sovereignty through transparent, auditable code, and accelerate innovation via community-driven development (e.g., future EuroHPC projects similar to ADMIRE or IO-SEA projects).

Consequently, **leveraging mature HPC storage for AI training, paired with checkpoint-management libraries like SCR[5] and DMTCP[6], allows Europe to bypass reinventing foundational I/O layers.** This strategic reuse of open-source infrastructure not only slashes time-to-deployment but also redirects R&D focus toward AI-specific optimizations (adaptive checkpointing, model parallelism), while ensuring storage sovereignty remains anchored in European hands.

Inference, the second major workload in AI, remains highly dynamic and is currently the fastest-evolving area of the stack. Early systems significantly underestimated inference-time data requirements, which are now increasing rapidly as models rely on longer context windows and multi-step reasoning. A context window defines the number of interdependent tokens a model must process to generate an output.

The state of an ongoing interaction with a chatbot is captured by the model parameters together with the key-value (KV) cache, which represents the attention state of the conversation. The key (K) matrices encode how tokens attend to one another, while the value (V) matrices store the information that is weighted and combined to produce outputs. Each processed token appends one key vector and one value vector per attention head and per transformer layer to the KV cache, causing KV memory usage to grow linearly with sequence length. The size of each vector depends on the model's hidden dimensions and numerical precision. Figures 1.a and 1.b illustrate the positioning of a KV cache in the inference process.

[5] https://github.com/llnl/scr

[6] https://github.com/dmtcp/dmtcp

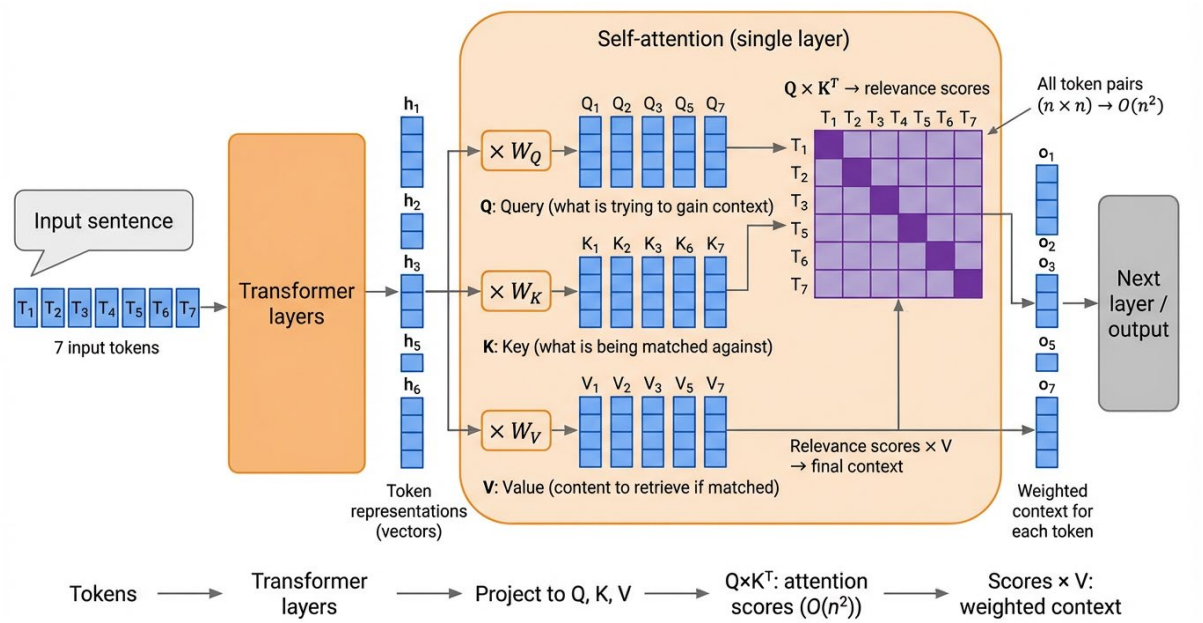


Figure 1.a: The initial sentence of 7 tokens is processed through the transformation layers of the model to generate vector values for every token. Inside the attention mechanism, these vectors are projected into three categories: the Query (what is currently trying to gain context), the Key (what is being matched against), and the Value (the content information we will retrieve if there is a match). A QxK multiplication is performed first to determine relevance scores, and these scores are then multiplied by V to create the final weighted context for every token. Complexity is in O(n²).

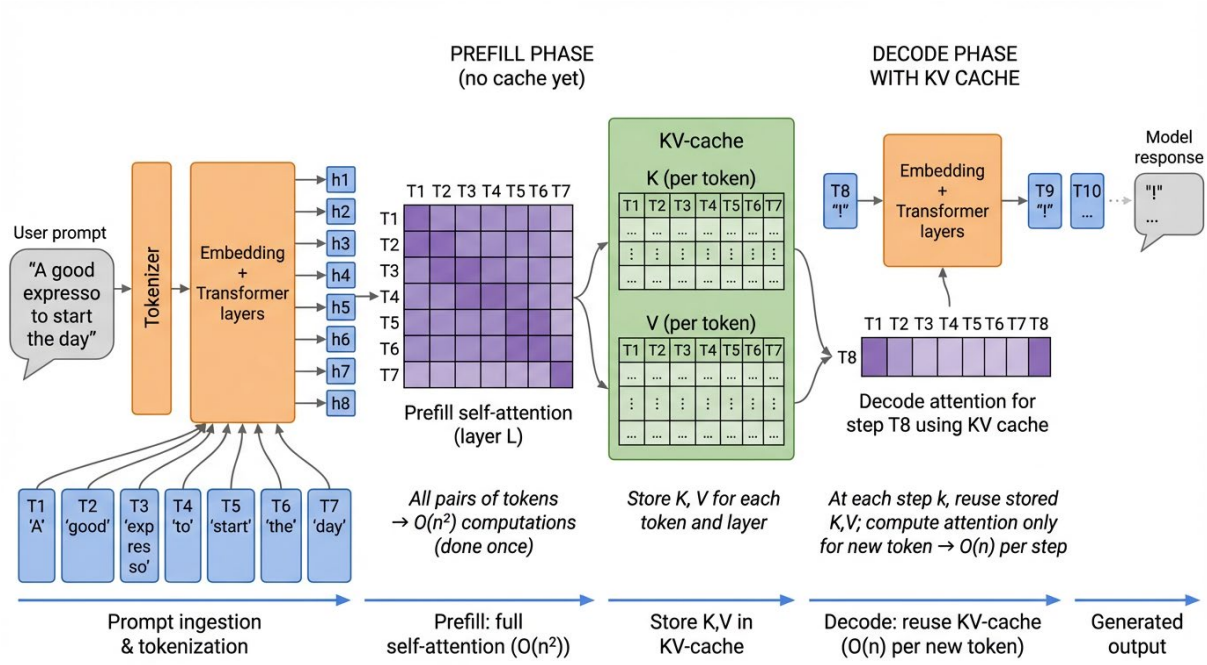


Figure 1.b: The LLM inference pipeline with KV cache operates in two distinct phases to optimise performance. In the initial "prefill" phase, the entire prompt is processed in parallel, and the computed Key (K) and Value (V) vectors for every token are stored in the GPU memory cache. Subsequently, in the "decoding" phase, the model generates tokens one by one; for each new step, it only computes the Query (Q), Key, and Value for the single newest token, retrieving the entire history of past K and V vectors from the cache to perform attention, thereby avoiding redundant re-computation of the preceding context. Complexity is now in O(n). Notice that K and V are constant within the context; Q depends on each token and cannot be cached.

The size of the KV cache therefore heavily depends on model parameters[7]. Assuming Float16 parameters and the open-source model Llama-3-70B, the size per token is 305 KByte. The European Mistral Model 7B is about 122 KByte per token, and the widespread public model DeepSeek v3 has an even larger footprint of 1.63 MByte per token[8]. Using Multi-head Latent Attention (MLA), DeepSeek is able to reduce this to 69 KByte per token.

As a result, the memory required to maintain the conversation state grows linearly with the number of tokens in the context window, and KV cache storage often dominates inference memory requirements for long contexts. Retrieval-Augmented Generation (RAG) further accelerates this trend by injecting a corpus of tokenized documents into the reasoning process, while chain-of-thought reasoning compounds the effect by generating long intermediate reasoning traces that must be retained in the KV cache, increasing inference-time memory pressure even when those tokens are not part of the final output.

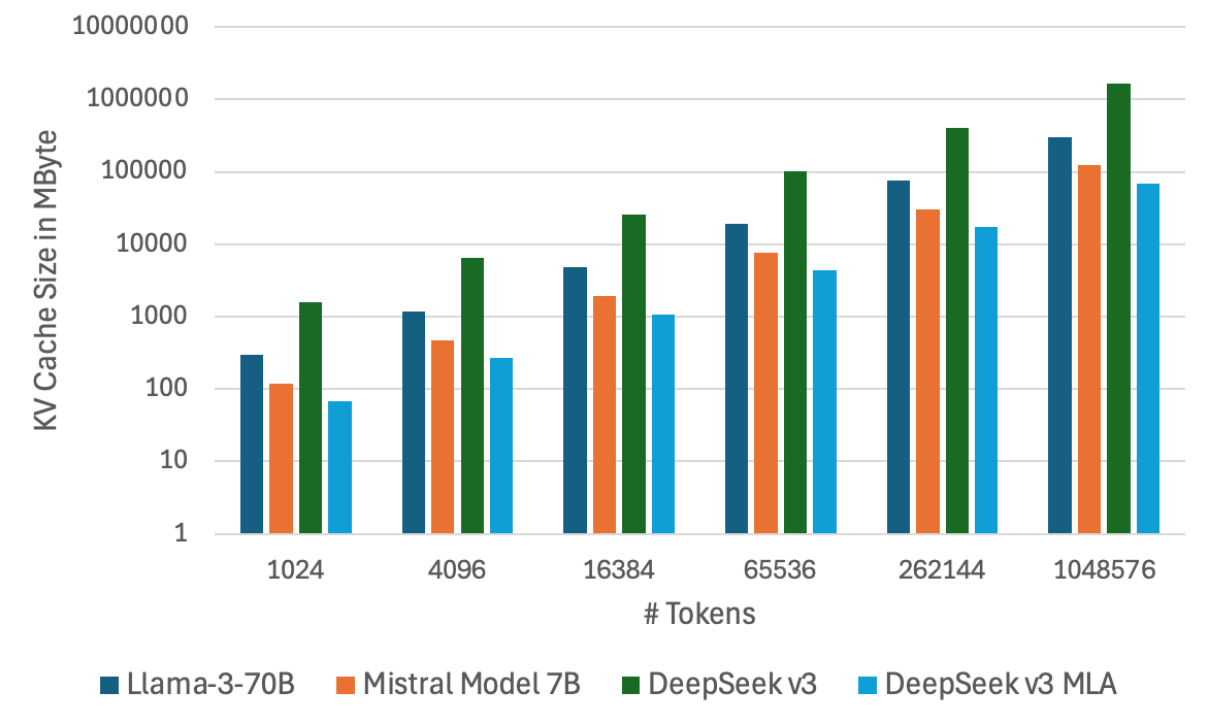


Figure 2: KV Cache size depending on number of tokens and model type

The user might interrupt and continue the interaction with an LLM at any time. When restarting a conversation, the underlying system has to choose between massive recomputation (quadratic in the number of tokens) or keeping the complexity linear at the expense of storing and loading vast quantities of intermediate results (KV caching). While a 'large' context window was previously considered to be around 4K tokens, modern inference engines must now handle windows in the range of millions of tokens. For instance, running inference on a Llama-3-70B parameter model with a 1-

[7] P. Lienhart: LLM Inference Series: KV Caching, a deeper look, https://medium.com/@plienhar, 2025

[8] All kv sizes per token have been calculated based on https://lmcache.ai/kv_cache_calculator.html.

Actual runtime memory usage may be lower if architectural or compression techniques are applied.

million-token context generates approximately 300 GB of intermediate data (see Figure 2).

These intermediate results have either to be recomputed or stored in GPU HBM memory or moved on to DRAM and persistent storage. Recomputing is too slow and too expensive, storing in GPU HBM memory is too expensive, and only the combination of DRAM and persistent storage allows it to meet both the performance and the economical requirements.

Motivation Example: Mooncake KV Cache

The "Mooncake" architecture[9], deployed by Moonshot AI, illustrates a trade-off between storage and computation. Mooncake is currently a large AI cluster of 24,000 GPUs, it is used mostly to train and serve the Kimi language model. Kimi and Mooncake have been extremely successful, leading to massive workloads exceeding 100 billion tokens daily. Mooncake addresses a critical scaling challenge: the explosion of intermediate data in the KV cache. A specific feature of the Kimi model is the large context windows used to deliver the most accurate answer to the end-user. As a downside, to prevent costly recomputation, Mooncake stores the full state of many user contexts in memory. However, the HBM of the GPU and the local memory cannot accommodate these token volumes, and recomputing it for every turn remains prohibitively expensive.

Therefore, Mooncake overcomes this by disaggregating the KV cache into a distributed pool of DRAM and SSDs, leveraging cheaper storage bandwidth to substitute for costly GPU compute cycles. This approach both reduces memory pressure and avoids recomputation of intermediate attention states, significantly improving quality of service. Such efficiency is crucial for inference workflows with human-in-the-loop, where responses must be served within ~100ms to feel instantaneous.

The development of the Mooncake architecture and its novel KV cache stem from limitations observed in production while serving real customers. As a side note, even with this KV cache, the system currently operates only at 2–3% of its theoretical peak performance. This suggests that AI infrastructure is mirroring the historical evolution of HPC, past initial deployment, it is now poised for a major optimization phase.

KV caching underscores the importance of storage in modern AI systems. Traditional GPU memory cannot keep up with massive intermediate data from long context windows. **Without new approaches, systems face a trade-off between expensive recomputation or efficient persistence of attention states.** Architectures like Mooncake, which disaggregate the KV cache across DRAM and SSDs, show that integrating tiered and networked storage can offload GPUs, reduce recomputation, and maintain low-latency inference. KV caching demonstrates that scalable storage is now a critical component of the AI stack, not just raw compute. **However, even KV caching is closely related to HPC, as Mooncake has been inspired by experiences from HPC ad-hoc file systems.**

[9] R. Qin, Z. Li, W. He, J. Cui, F. Ren, M. Zhang, Y. Wu, W. Zheng, X. Xu: "Mooncake: Trading more storage for less computation – a KVCache-centric architecture for serving LLM chatbot." In *23rd USENIX Conference on File and Storage Technologies (FAST)*, pp. 155-170. 2025.

As an additional **lesson learned** from the recent evolution of AI workloads, our strategy to address the complexity of emerging, unstable workloads is to acknowledge that the industry is traversing a transient stage. Therefore, in addition to designing new storage systems based on the current requirements of AI workflows, we believe it is critical to install probes on production systems. Monitoring via advanced telemetry the evolution of the production grade AI will enable it to detect weak signals and trends originating directly from the most versatile component of the equation: the AI software itself. Ultimately, as illustrated by the Kimi motivation example, effective anticipation is based on early detection combined with robust knowledge of operational constraints and research state-of-the-art.



4 KV Caching based on Ad-hoc File Systems

Modern AI hardware motivates moving KV cache functionality closer to storage. Contemporary AI nodes integrate multiple high-capacity NVMe SSDs, large memory resources, and extremely high-bandwidth network connectivity, offering tens of terabytes of potential cache capacity per node. However, node-local SSDs alone are insufficient as a KV cache solution because they cannot easily be shared across nodes and are unsuitable for persistent or coordinated storage across distributed workloads. This limitation highlights the need for a software layer that can aggregate local storage resources while preserving accessibility and performance at cluster scale.

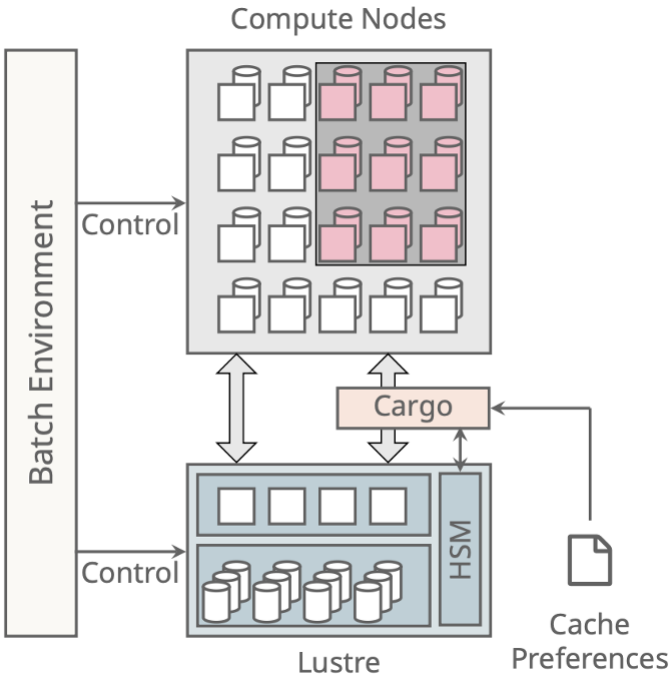


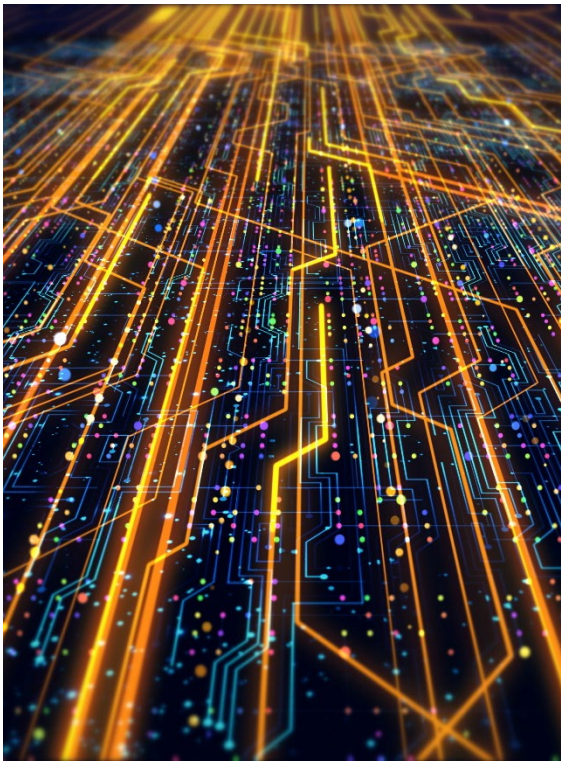
Figure 3: Combining storage in compute nodes (marked in red) to build ephemeral ad-hoc file systems.

Ad-hoc file systems provide exactly this capability and are a promising foundation for implementing key-value caching in modern AI systems (see Figure 3). They originate in supercomputing and are instantiated dynamically within an application context and combine the storage capacity of multiple compute nodes into a unified namespace accessible across participating nodes[10]. Ad-hoc file systems can be optimised for the relaxed storage semantics required by frameworks, which strongly benefit from fast, read-only caching. At the same time, hardware trends already support such an approach. AI nodes increasingly integrate high-performance NVMe storage, GPU-Direct enables direct interaction between GPUs and storage to reduce CPU bottlenecks, and data processing units (DPUs) enabling storage, networking, and security functions to be offloaded from the host processor. Together, these developments create an architectural environment in which storage can act as an efficient extension

[10] N. O. Almaaitah, J. García Blas, G. Sanchez-Gallegos, J. Carretero, M.-A. Vef, A. Brinkmann: A comparative study of ad-hoc file systems for extreme scale computing. Future Gener. Comput. Syst. 170: 107815 (2025).

of memory, enabling scalable and high-throughput caching mechanisms suited for AI workloads.

These new AI requirements on storage build an opportunity to leverage Europe's long-standing expertise in HPC storage systems, which closely resemble the requirements of AI data infrastructure. European research and deployment efforts around open-source storage software, such as parallel file systems, hierarchical storage with ephemeral caching layers, and ad-hoc file systems, have already addressed challenges similar to those faced by large-scale AI systems. Knowledge consolidated in the major EuroHPC initiatives IO-SEA and Admire has demonstrated how distributed storage can efficiently support data-intensive computation, meaning that much of the conceptual and technical groundwork required for AI KV caching already exists and can be transferred with relatively low effort.



5 Landscape of the Data Economy

The data market is a domain connected to storage and data infrastructure. The disruption introduced by AI is happening within an already constituted data ecosystem. It is worth mentioning certain challenges and evolutions in the data landscape to understand the context in which this storage reinvention is unfolding. Furthermore, the fact that several large-scale European initiatives already support the data market underscores the unsustainable situation of storage infrastructure that is currently left ignored by European funding agencies.

Data supply chain and data logistics

Europe produces a considerable amount of high-quality scientific data thanks to its variety of world-leading scientific instruments. These big science instruments cover a large number of scientific domains, for instance, light sources with ESRF, telescopes with Astron, Earth Observation with the Copernicus program, or medical imaging with Neurospin. At the industrial level Europe hosts a large and diverse ecosystem with instrumented factories, telcos, and hospitals. The UK, which is back in the fold of European HPC R&D, also leads cutting-edge data-intensive scientific experiments, such as the Square Kilometre Array.

The AI data supply chain defines the end-to-end lifecycle, from acquisition, cleaning, outlier detection, transformation, and delivery. This chain is operationalized through data logistics, ensuring an efficient flow from diverse sources to the final consumption agents: the deployed machine learning models for training or embeddings generation for inference.

Data logistics focuses on the operational aspects of managing data movement, ensuring that data is available where and when it is needed. This involves orchestrating data pipelines, optimizing data storage solutions, and implementing robust data governance practices to maintain data quality, security, and compliance. Effective data supply chain and logistics practices are essential for enabling AI models to learn from high-quality data, adapt to new information, and deliver accurate and reliable insights.

While Europe possesses large-scale data suppliers and a portfolio of sovereign data logistics software and infrastructure, these assets are not sufficient. Implementing an operational interconnection at scale serving AI workloads requires a significant step forward. The establishment of cooperative AI factories with data and models that seamlessly transfer and exchange data in a peer-to-peer network still lacks the correct characterization of data gravity and workflow optimization. All the pieces of the infrastructure puzzle are gradually being put in place, but some critical software links are missing or not yet emerging as a viable production-grade system. CERN with FTS, or Nodeum with its Data mover for AI Factory are examples of successful data logistics software. These efforts could be leveraged and extended to build the first version of a European AI data logistics backbone.

A successful data logistics backbone aims beyond being federative. It aims to be inclusive: to leverage the pre-existence of the vast EU public data hubs and act jointly with emerging national initiatives such as the Research Data Management (RDM) from the German NFDI. Therefore, interoperability and compliance with standards are critical.

Data Market Places

The aforementioned datahubs are part of an emerging continental data marketplace that is tightly intertwined with data logistics. As an illustration, in scientific/medical and industrial practices, data are generated in a distributed way and are periodically updated. The distributed ecosystem of platforms through which this data flows from producers to consumers operate within the compute continuum rather than a federation of HPC platforms. Consequently, AI workloads (training, fine-tuning, inference) and their related data preparation phases typically

span platforms of very different natures: from HPC systems to the cloud and out to edge devices.

Furthermore, the intellectual property and license of use attached to this data are of complex varieties that can hardly be grouped into a few categories but should be managed as a point-to-point relation between producers and consumers. Most data of interest in AI applications are entirely private. They can be used for a specific purpose, for a limited period of time: some data collections can only be virtually grouped, but cannot be moved to a central location.

The data market model is still bound to the paradigms that inhibit the usage of shared infrastructure for any private or industrially valuable datasets, and associated policy for their sharing. To illustrate the problems this causes, all the typical techniques in the realm of AI and HPC exploit additional (cached) data copies to boost performance. Those copies are under the control of the consumer (not of the producer) and they cannot be easily destroyed (e.g., because the usage time is over) or controlled (e.g., limiting the number of times they can be used). Over time, organizations began to recognize that merely storing data was insufficient. Historical and operational data held latent value – insights that could inform business strategy, enhance operational efficiency, or fuel innovation. To support this transition, two pressing needs emerge: the need for data catalogues, and the need for a new model for private data management.

Data catalogues

To organize the market, data catalogues are provided to ease navigation and data identification and retrieval out of these data hubs. The scientific community knows how to develop and how to maintain data catalogues, for instance RUCIO from CERN and now SKA, or the data portfolio from Destination Earth. An ecosystem of scientific data produced in a highly distributed way by different laboratories that cannot be concentrated in one place, will require shared and machine actionable metadata catalogues in accordance with the principles of open science (F.A.I.R[11]) that are designed to monitor and support workflows across the data supply chain (see previous section).

Private data management

Data privacy can be implemented at the infrastructure level, using multi-tenancy to segregate virtually shared infrastructures. At the AI framework level, it is possible to apply distributed AI techniques. For instance, federated learning (FL) makes it possible to virtually pool datasets in large collections while maintaining the data private, thus under the full control of the data owner. Also, FL proved to be an effective federation protocol across supercomputing (cross-facility federated learning[12]). If properly managed, FL has the potential to trigger an entirely new data market, where data is "rented" and not just copied, making it possible to effectively control data ownership and license & track data movement via its data catalogue (which could also become a data marketplace). At the level of the data itself, other technologies can be used, such as *Attribute-Based Encryption*. These techniques are complementary and can serve the distributed data management (ownership, veracity, etc.) value chain and marketplace.

It should be noted that data marketplaces are already supported by an array of programs, including a flagship initiative named SIMPL[13] with a budget of 150M€ dedicated to the development of European data marketplaces.

[11] Coming from the scientific community, the Findable, Accessible, Interoperable and Reusable principles are the cornerstone of scientific data management. FAIR principles aim to improve the quality, usability, and impact of data. They are natural guidelines for AI data readiness.

[12] https://www.sciencedirect.com/science/article/pii/S1877050924016909

[13] https://digital-strategy.ec.europa.eu/en/policies/simpl

Data management

The last three years have dramatically disrupted data management. The original vision focused on data preservation (moving older data to less expensive archival media) has been replaced with a paradigm of living data. Data is continuously collected, analysed, interpreted, and leveraged to create real-time value. Data is no longer considered a liability but an asset.

Proper data management practices encompass a large breadth of challenges, from seamless collection, storage, and retrieval of vast amounts of data. These challenges apply to training and validating AI models. By implementing robust data management strategies, AI Factories can enhance data quality, reduce redundancy, and ensure data integrity, thereby improving the accuracy and reliability of AI models.

Optimised data management leads to significant cost savings by reducing the need for excessive computational power and associated storage. This is particularly important in cloud-based environments where costs are tied to use. Data management impacts a large array of fields where Europe has the highest ambitions, such as healthcare, finance, and autonomous vehicles, where real-time data processing and decision-making are crucially applied. The Flagship Destination Earth program has emphasized the critical aspect of data management at scale. As illustrated by the Giga-factory program[14], AI systems are racing toward ever-increasing scale. Overlooking efficient data management will hamper performance and scale, return on investment, and overall competitiveness.

In addition to moving data at a fast pace, to fully de-risk infrastructures and be ready for future AI workloads, data solutions must efficiently manage datasets. A dataset is the fundamental quantum of information for any training process. This means metadata needs to be updated, browsed and shared across cooperative data hubs and consumers through a registration mechanism. This metadata sharing and registration capability is essential for multi-cloud operations. Therefore, any future-proof architecture must include an integrated, federable, and efficiently scalable metadata catalogue.

Our vision is that future AI systems should integrate warehouse capabilities within the data platform enhanced with an internal synchronous metadata catalogue and ultra-efficient listing capabilities. Two performance criteria emerge for dataset management solution: time to first byte and metadata parsing.

Data management and data logistics are at the core of the EuroHPC federated platform project[15]. These aspects are thus already handled by major European initiatives and can be let out of the scope of this agile white paper, which is advocating for new initiatives. We have discussed the importance of these topics for the sake of consistency, but in the remainder of this paper, we will focus on the more storage-specific aspects of data.

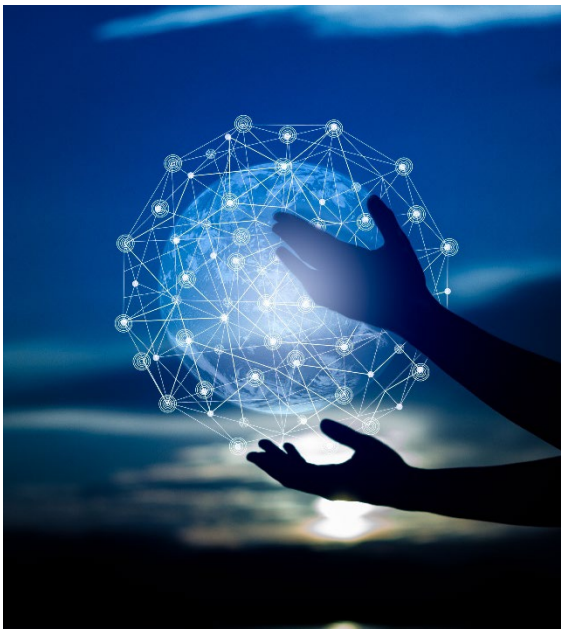


[14] https://commission.europa.eu/topics/competitiveness/ai-continent_en

[15] https://my-eurohpc.eu/

6 Executing the Storage re-invention

Reinventing storage is a highly ambitious goal that requires flawless execution. To prevent fragmentation and ensure future storage developments remain relevant at scale, the challenges listed below must be addressed with a clear focus on driving production business outcomes. Inspired by Moonshot AI's Mooncake architecture for their Kimi system, we can establish a virtuous cycle between production environments and R&D facilities. Telemetry and observability have a major role to play here. Because the AI software stack is evolving at a breakneck pace, the ability to detect real-world production trends is essential to prevent misaligned research. Finally, we conclude with a proposed path toward increased hardware sovereignty by targeting an emerging tier within the storage hierarchy.



7 Energy and Sustainability

To be market-viable, any European storage must encompass sustainability. We can broadly consider sustainability and efficiency at two different levels: (a) Data reduction and data movement/gravity and (b) storage device efficiency and co-design.

Data reduction and data gravity

Multiple techniques are available to improve sustainability and notably to optimise the energy footprint associated with data, including three main and broad approaches: (a) compression and deduplication, (b) moving processing closer to data, and (c) workload consolidation and malleability.

First, reducing data is an effective way to minimise energy. Compression and deduplication are often coalesced as data reduction techniques. They are actually two reduction techniques for the at-rest capacity. But the relationship is not additive, compression tends to prevent deduplication, and only client-side compression minimizes network traffic. At the opposite end, sending uncompressed data over the network and performing compression on the server side can be stacked on top of a data deduplication stage, offering cumulative benefits from a capacity gain. Exploring ad-hoc file system usage for AI is also interesting. They can leverage existing deployments by providing balanced I/O distribution and linear scalability across arbitrary numbers of client nodes. AIStore from NVIDIA, a lightweight, built-from-scratch storage stack tailored for AI applications, is very close to the ad hoc file systems, an approach where Europe has some important scientific and technical assets.

Second, a more ambitious approach, exemplified by the UK-based Excalidata project, considers *in situ* data processing directly at the level of storage servers. Sending the code where data is located drastically optimises the energy footprint and significantly reduces network contention. In the case of data-intensive AI workloads, filtering, compression, and feature extraction could occur directly within storage devices. Computation can be added at the server level, storage pool level, or storage device level, providing a range of possibilities to offload bespoke functions from HPC/AI workflows. Effective deployment requires innovation in the software stack and a deep understanding of the energy footprint across different storage tiers, enabling intelligent offloading of computations to maximize efficiency. Activities in standards bodies such as SNIA on computational storage can also be leveraged by European actors to accelerate adoption and interoperability. By integrating computational storage into modular HPC and AI systems, HPC/AI data centres achieve higher throughput, lower energy consumption, and scalable performance, making it essential for large-scale, hybrid, and data-local AI deployments in research, industry, and sovereign supercomputing environments. Early projects in Europe such as SAGE/Sage2[16] leading to IO-SEA have already provided some interesting ideas in this direction, which can be further exploited. Pioneered in HPC, these efforts reflect the exact trade-off between computation and data movement that is currently defining AI inference. A prime example is the KV cache[17] architecture discussed in Section 4.

Third, energy efficiency and sustainability are directly related to achieving high utilization of the underlying compute and storage infrastructure. Idle infrastructures tend to consume energy and contribute to the increasing pressure for building ever larger systems, especially for AI-related processing. Similar trends were observed in the Cloud era, when data centres were required to serve massive, planet-scale workloads. A main approach to achieve

[16] https://cordis.europa.eu/project/id/671500

[17] It should be noted that YanRong, a parallel file system in China, has recently announced KV cache acceleration support for inference. To our knowledge, no European or open-source solutions provide such a feature.

high(er) infrastructure utilization is to consolidate multiple workloads on each computing and storage infrastructure. This requires embracing malleability and elasticity by the workloads and the infrastructure so that each workload uses at any point in time the resources it requires, leaving the rest of the system resources for other applications and workloads. Consolidation and malleability are fundamental techniques that have the potential to improve the efficiency of computing and storage in the era of AI by a significant factor. For instance, training and inference are seldom run on the full system; instead, most workloads use only a fraction of the infrastructure. Therefore, an energy-conscious and sustainable approach to AI needs to cater for different requirements, ranging from efficient single-thread/single-node performance for data pre-/post-processing to GPU-heavy processing to data-intensive phases and memory interference, creating a complex and constantly evolving optimization landscape.

Device efficiency and co-design with applications

Storage devices themselves have a significant role in energy efficiency and sustainability.

First, NAND-Flash storage devices that dominate in modern storage systems perform internally complex data management functions, such as garbage collection, logical-to-physical block mapping, error detection, and correction. These functions contribute significantly to energy consumption, especially as data rates increase and device capabilities are required to keep up. Existing efforts[18, 19] strive to improve these internal functions and have the potential to lead to more efficient and more sustainable devices in the era of massive AI processing.

Second, new requirements, such as KV caching for inference workloads, create new opportunities. A significant difference between modern AI workloads and the cloud workloads of the previous 20 years is that they exhibit more limited runtime behaviour. This allows us to consider co-designing the storage devices in a manner that optimises the behaviour of these workloads. Such approaches have proven challenging in the past, e.g., with the use of computational devices and KV SSDs[20]. However, they are now gaining renewed attention as promising directions. Analysing, understanding, and offloading to storage devices functions, such as KV caching, have the potential for order-of-magnitude improvements in performance and efficiency.

Third, storage devices historically have been designed and optimised for host-level workloads. However, today, GPUs are emerging as the main engines of computation. Therefore, co-design efforts that reconsider the interfaces of storage devices so that GPUs become first-class citizens have a significant role. A promising track of research to streamline further the data plane is to implement the file system client directly within the GPU architecture. Both NVIDIA and AMD are now supporting direct memory access to transfer data from NVMe storage or network fabrics directly into GPU memory, bypassing the host CPU and system RAM entirely. This zero-copy approach drastically reduces latency and improves bandwidth by comparison with the legacy methods associated with bounce buffers, ensuring that the PCIe bus is utilized exclusively for effective data movement rather than redundant copying. However, a file system client is more than a data plane; the control plane with metadata actions and data positioning is critical. If GPUs are powerful parallel machines, they remain limited in their capabilities to handle branches and

[18] Matias Bjørling, Abutalib Aghayev, Hans Holmberg, Aravind Ramesh, Damien Le Moal, Gregory R. Ganger, George Amvrosiadis: ZNS: Avoiding the Block Interface Tax for Flash-based SSDs. USENIX ATC 2021: 689-703.

[19] Michael Allison, Arun George, Javier Gonzalez, Dan Helmick, Vikash Kumar, Roshan R. Nair, and Vivek Shah. 2025. Towards Efficient Flash Caches with Emerging NVMe Flexible Data Placement SSDs. In Proceedings of the Twentieth European Conference on Computer Systems (EuroSys '25). Association for Computing Machinery, New York, NY, USA, 1142–1160. https://doi.org/10.1145/3689031.3696091.

[20] Duffy, C., Shim, J., Kim, S. H., & Kim, J. S. (2023). Dotori: A key-value ssd based kv store. *Proceedings of the VLDB Endowment*, *16*(6), 1560-1572. https://www.vldb.org/pvldb/vol16/p1560-duffy.pdf

complex control flow codes. The field of research dedicated to file systems for GPUs addresses exactly these limitations. In the long term, the results of this research could be implemented on top of existing European file systems.

Fourth, there are efforts to examine how Flash-based devices can be co-designed to become an integral part of the memory hierarchy, especially in AI-based systems. Efforts related to HBF (High Bandwidth Flash) aim to provide high-throughput access from GPUs to NAND Flash, overcoming limitations in DRAM scaling capacity[21]. Such approaches create a vast configuration space, covering device-level design issues, caching mechanisms and policies, and workload adaptation and optimization with significant potential to improve efficiency, sustainability, and cost of computing infrastructures for AI.

Today, flash-based storage devices dominate high-performance storage systems and are almost exclusively of non-European origin. This creates significant risks when managing and serving data. On the other hand, Europe has built extensive expertise in various layers of modern storage systems, including server-level architectures, interconnects, and high-end controllers for offloading acceleration purposes. Flash storage devices can be seen as performing three main functions: (a) host-device interfacing, (b) device-flash interfacing, and (c) block management in the file-translation layer (FTL). Europe already has significant expertise in (a)[22] and is well positioned to develop expertise in (c) as well[23]. This can be leveraged for designing European-based storage devices and their complex, device-level software stack as part of a future-proof architecture, contributing to data integrity and I/O performance, and technological sovereignty.

AI applied to AI infrastructure

Our modern data-intensive factories are instrumented in the very same way as the traditional industrial factories. Production lines are continuously probed and optimised; similarly, digital AI factories are instrumented with sophisticated telemetry. By continuously learning from real-time workloads behaviour, AI can dynamically rebalance data across tiers, fine-tune caching policies, and even anticipate component degradation before it impacts performance. This agentic way to manage data placement opens the field to efficient tiered architectures harnessing different storage technologies within a unique system. This can considerably improve the resilience of the European supply chain when key elements suffer from supply, price, or other uncertainty or volatility, such as flash storage device shortages for the next two years. Being able to adapt storage operation (using AI-based optimization of system behaviour) to alternatives, such as hard drives that exhibit longer-term stability, can shield infrastructure from such uncertainty.

Therefore, exploring long-term use of AI-based optimization in AI infrastructures, and in particular in the storage system itself, creates a virtuous cycle where telemetry fuels AI models, and those models in turn drive progressively smarter, more efficient, and more resilient storage architectures.

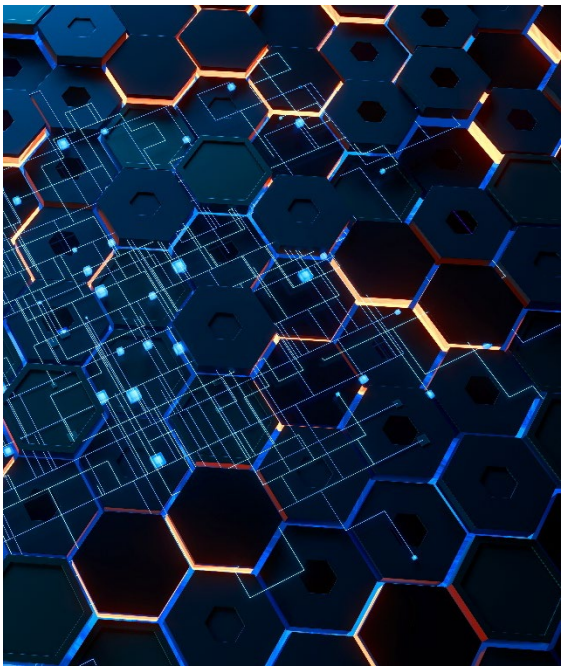
[21] Ma, Xiaoyu and David Patterson. “Challenges and Research Directions for Large Language Model Inference Hardware.” ArXiv abs/2601.05047 (2026): https://doi.org/10.48550/arXiv.2601.05047.

[22] Swissbit is an European SMEs developing advanced integration and interfacing for NAND chips.

[23] The concept of Zone Named SSD is coming from Western Digital researchers based in Denmark.

8 From Fragmentation to Consolidation

Europe is rich in innovative designs and initial Proofs of Concept, but most of them remain at limited TRLs. The continental landscape suffers from significant fragmentation, preventing these isolated successes from evolving into viable, production-grade systems.



Addressing this challenge requires moving beyond scattered pilots and consolidating efforts under a cohesive pre-market program designed to accelerate industrial readiness. This consolidation is the prerequisite for a more mature operational strategy that can be deployed along the idea of bridging the divide between research and production by implementing a Transparency Act to effectively monitor existing sites; and second, securing the foundation of this infrastructure by enhancing the resilience of the storage supply chain through sovereignty actions.

9 Conclusion and call to action

From individual citizens to large scientific instruments, European societies are generating enormous amounts of data. However, there has been no serious European initiative to establish the storage industry required to store and provide this data. The emergence of AI and its substantial IT requirements now present a unique opportunity to update outdated IT infrastructure practices and invest in developing new data storage technologies that ensure data sovereignty. Europe already has a rich ecosystem of competitive, research-level approaches to data storage for HPC. These approaches must be transformed into industry-ready solutions that can compete with massive initiatives developed abroad. Therefore, we recommend fostering deeper collaboration among European technological leaders and experts in HPC and AI-supported by a well-defined Research and Innovation program in this area.

European Converged Data Platform Pilots: Innovation Program for a Pre-market Solution

In line with the successful programs for developing European processor and network technologies, the European Commission should now prioritize Converged Data Platform Pilots. The corresponding innovation program must be substantial enough to create an adaptive storage platform that is flexible enough to efficiently support various use cases, including large-scale AI training and inference, HPC, and cloud computing. Decoupling the data platform from a specific hardware architecture and embracing the diversity of the storage ecosystem are necessary conditions for a future-proof data system.

In the absence of a trillion-dollar industrial leader in Europe and given the fragmented industrial landscape, open-source solutions are the continent's most viable strategic pathway to building its own data platform. Open-source has consistently proven to be a competitive and robust alternative to proprietary systems, offering the flexibility and resilience needed to establish European data autonomy. Therefore, Europe can rely on its world-class open-source assets. Components such as high-performance Lustre file system distributions, ad-hoc file systems, and key-value stores must be standardized and qualified for a single, at-scale, high-TRL storage pilot. Such pilots will pave the way for a European sovereign data platform. Incubation within AI Factories and other EuroHPC infrastructures will demonstrate the initiative's industrial viability.

European Data Factory Transparency Initiative

The deployment of European AI Factories over the last few years represents a major infrastructure endeavour. This effort should be extended to fully capitalize on this unique network of systems for the design of future sovereign data platforms. We propose a European Data Factory Transparency Initiative. This initiative will generalize the instrumentation of these infrastructures, offering an unparalleled opportunity to collect detailed workload information running across multiple sites with diverse user communities. Gaining this world-unique insight will enable the detection of weak signals and emerging trends, thus providing a crucial competitive advantage.

- **Necessity of Profiling: I/O tracing and profiling** tools are mandatory for understanding AI storage workloads. They reveal the fine-grained, temporal patterns (such as short bursts, random accesses, or sequential streaming) that specifically characterize AI training and inference.
- **Bottleneck Identification:** Correlating the low-level I/O view with higher-level system profiling makes it possible to pinpoint exactly where storage latency or bandwidth limits become system-wide bottlenecks in complex machine learning workflows.

Leveraging this instrumentation and applying advanced AI modelling techniques, the

Transparency Initiative will directly drive innovation in the emerging AI-centric storage field. At the operational level, we propose to build upon results delivered by previous EuroHPC projects, such as ADMIRE and IO-SEA, and to publish the collected metrics and analysis following an open-data approach. The successful implementation of this data transparency initiative will deliver a world-class catalogue of operational data and will become a major research inspiration for future system designers.

In addition to the previous recommendations focusing on innovation and exploitation, the following recommendation will be about research. We consider that a specific research call is mandatory to further push the hardware autonomy of the European storage stack.



European Sovereign Open Hardware: Research Program for future technological autonomy

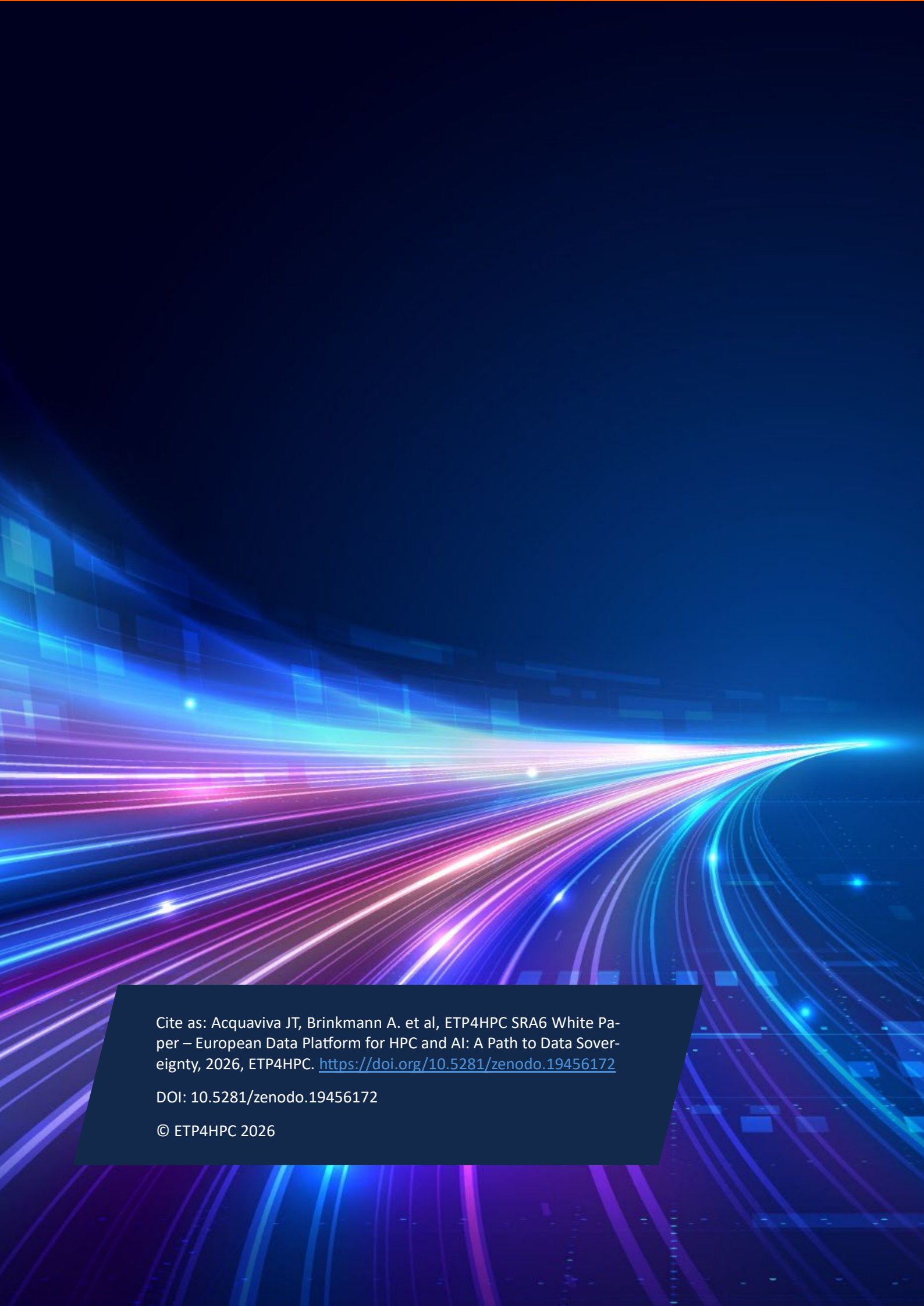
Non-European manufacturers currently dominate Europe's storage market, especially in the production of key components such as NAND flash and advanced controllers. However, Europe possesses significant strengths in embedded software and cybersecurity, which are crucial for building intelligent, efficient, and secure storage systems. By leveraging this expertise and taking advantage of emerging standards such as CXL, which open a unique window of opportunity for rethinking memory and storage architectures, Europe can focus on system integration and software-defined storage that maximize performance, reliability, and interoperability. We therefore propose to establish a research track centred on embedded software to build data storage systems focusing on the Open Source Hardware (OSH) paradigm and emerging standards such as CXL (Compute eXpress Link) amongst others. This research track would enable Europe to take a first step to create high-value storage solutions that combine limited external hardware inputs with strong domestic innovation in software and system design, fostering strategic autonomy and competitiveness in the data infrastructure ecosystem.

We consider an EU High Bandwidth Flash (EU-HBF) as the best opportunity at the moment in the market. HBF is an emerging AI technology, and Europe's significant planned investments in AI infrastructure position European HBF as a valuable innovation opportunity worth exploring. Such an initiative could be articulated with the Chip-JU[24] to deliver fully, from hardware to firmware, sovereign storage devices.

[24] https://www.chips-ju.europa.eu/CallDetails/?id=455a3eb8-bcce-f011-bbd3-7c1e5276d834

Contributing Authors

AMD	Jose Noudohouenou Georgios Markomanolis
Carlos III Universidad	Jesús Carretero Javier Garcia Blas
CINI / ICSC	Marco Aldinucci
TU Darmstadt	Felix Wolf
DDN	Jean-Thomas Acquaviva
MPCDF	Erwin Laure
ETP4HPC	Sai Narasimhamurthy
FORTH	Angelos Bilas
CEA	Philippe Deniel
CERN	Martin Barisits
University of Reading	Bryan Lawrence
ECMWF	Simon Smart
Saarland University	André Brinkmann
Nodeum	Valéry Guilleaume



Cite as: Acquaviva JT, Brinkmann A. et al, ETP4HPC SRA6 White Paper – European Data Platform for HPC and AI: A Path to Data Sovereignty, 2026, ETP4HPC. https://doi.org/10.5281/zenodo.19456172

DOI: 10.5281/zenodo.19456172

© ETP4HPC 2026