Extended abstract proposal: Who is driving storage research? Questioning the priorities behind SSD research

Abstract

The rapid evolution of data storage technologies, particularly the transition from hard disk drives (HDDs) to solid-state drives (SSDs), has significantly impacted computing ecosystems. SSDs are now essential components in large-scale machine learning systems and cloud computing infrastructures.

However, much of this technological progress has been influenced by factors that are often overlooked, including commercial interests, military applications, and biases in how storage systems are optimized. These influences contribute to the underutilization of modern storage technologies, despite their potential to drive more efficient and decentralized computing systems.

This talk will explore the broader dimensions of software storage-defined systems (SDS) and argue that modern SSD capabilities are often underused due to narrow industry focus. The technological potential of storage systems is frequently overlooked, with consumers typically receiving only the byproducts of research and industry innovation.

1. Undone science in storage systems

Despite the widespread adoption of SSDs in both consumer and enterprise markets, their full potential remains largely untapped. For example, a modern SSD can achieve up to 3 million IOPS (I/O operations per second), but existing software stacks often reduce this to only around 5,000 IOPS when advanced features such as snapshots or network clustering are used. This underutilization is partly due to a persistent focus on HDD-based storage models in many sectors, especially in enterprise environments, where SSDs are often overshadowed by large, existing HDD clusters. The gap is driven by technical limitations in software design and a lack of research into adapting SDS architectures to fully exploit SSD capabilities. Challenges in maintaining and evolving complex open-source projects like ZFS also hinder innovation in this area.

For instance, the Vitastor project [1], a fork of the Ceph storage system [2] developed by a single individual over several years, rethinks the assumptions of Ceph to optimize performance for SSD-only clusters. Benchmarks [3] show that Vitastor provides multiple-fold performance improvements over Ceph's.

2. The Deep Learning mania (so-called "AI") and data infrastructure

As deep learning systems are considered to be the new gold mine, the role of data storage in commodity infrastructure has been increasingly sidelined. The immense data requirements of deep learning systems typically prioritize keeping GPUs fed with data, often at the expense of broader storage needs. This narrow focus has resulted in the neglect of other critical storage system requirements, especially in non-cloud environments.

For example, the FAST conference on storage systems [4] now dedicates two entire sections to machine learning systems [5], emphasizing the growing importance of storage infrastructure in deep learning systems. Despite this, storage solutions still largely focus on cloud computing and specialized, often untested, hardware features.

3. Military-industrial influence and storage design

The military-industrial complex has long been a key player in shaping modern storage technologies. Military organizations require large-scale data storage for operations such as monitoring internet traffic or storing surveillance footage. These needs influence the design of specialized storage systems, often prioritizing long-term reliability and security over consumer-facing features.

For example, the porting of ZFS to Linux was funded through a contract with the U.S. Department of Energy and developed at the Lawrence Livermore National Laboratory. This project helped bring advanced storage technologies to a broader audience: the Linux kernel users, reflecting how institutional support can drive important technical developments that influence both military and civilian sectors, analogous to nuclear research consequences.

4. Pursuing a paradigm shift

Current research in storage systems often reflects the limitations of existing paradigms, where funding biases and corporate interests dictate the research agenda. A shift toward recognizing the potential of software storage-defined systems could open new opportunities for community-driven technological solutions. By focusing on low-cost, high-reliability storage systems, it would be possible to reduce dependence on centralized services and empower smaller, decentralized networks. However, this shift is constrained by a lack of research into storage technologies specifically designed for decentralized or consumer-driven systems.

For example, the Garage project [6] from the Deuxfleurs non-profit (association loi 1901) [7], co-funded by public research [8] and the European Union via NLnet [9], provides an open-source implementation of the Amazon S3 API [10] designed for commodity hardware. This project aims to democratize object storage [11] by offering an alternative to commercial solutions. By focusing on providing S3-compatible storage for small-scale deployments, Garage has supported grassroots movements seeking to self-host their own internet infrastructure. Unlike other implementations such as MinIO [12], which later monetized [13] their offerings [14], Garage has maintained its commitment to a community-based approach.

Conclusion

This talk will explore how modern SSDs are underused, particularly in the context of deep learning systems, military applications, and societal influences. It will argue for a broader research agenda in software storage-defined systems to help counterbalance the economic advantages held by large corporations and promote grassroots technological innovation.

By broadening the scope of research, we can address critical infrastructure needs, reduce the dominance of centralized systems, and foster a more equitable and sustainable technological future. For example, projects like the Solid project [15] aim to ensure that data remains decentralized and under user control, providing a model for future storage systems that prioritize individual sovereignty and community empowerment.

Bibliography

1. Vitastor project. Available at: https://vitastor.io/en
2. Ceph project. Available at: https://ceph.io/en/
3. Vitastor benchmarks. Available at: https://vitastor.io/en/docs/performance/comparison1.html
4. FAST Conference on storage systems. Available at: https://www.usenix.org/conference/fast25
5. FAST Conference Technical Sessions on Machine Learning systems. Available at: https://www.usenix.org/conf%C3%A9rence/fast25/technical-sessions
6. Garage project. Available at: https://garagehq.deuxfleurs.fr/
7. Deuxfleurs non-profit organization. Available at: https://deuxfleurs.fr/
8. NLnet European Union funding for Garage project. Available at: https://nlnet.nl/project/Garage/
9. NLnet European Union funding for Garage admin UI. Available at: https://nlnet.nl/project/Garage-AdminUI/
10. Amazon S3 API documentation. Available at: https://docs.aws.amazon.com/AmazonS3/latest/API/Type_API_Reference.html
11. Object storage on Wikipedia. Available at: https://en.wikipedia.org/wiki/Object_storage

12. MinIO GitHub repository. Available at: https://github.com/minio/minio
13. MinIO Monetization discussion. Available at: https://github.com/minio/minio/issues/21584
14. How to self host your own S3 in 2025. Available at: https://jamesoclaire.com/2025/05/27/how-to-self-host-your-own-s3-in-2025/
15. Solid project for decentralized data. Available at: https://solidproject.org/