IDC

White Paper, sponsored by:
IBM

AI-Driven Capabilities in Enterprise Storage Systems Bolster System Optimization, Automation, and Cyber-Resilience

Written by:
Carol Sliwa
Research Director, Storage Systems, Core Infrastructure

Published:
July 2026

Table of contents

IDC opinion **3**

Situation overview **3**

Considering IBM FlashSystem **4**

Challenges/opportunities **6**

Conclusion **7**

Definitions **7**

About the IDC analyst **8**

IDC opinion

The rapid growth of AI and analytics workloads and the increasing frequency and sophistication of ransomware and other cyberattacks are two of the most consequential trends confronting IT organizations today. As these workloads multiply and threats intensify, enterprise storage administrators are under pressure to optimize system performance and ensure data protection and recovery, while also meeting business demands, often on accelerated timelines. With data volumes scaling to petabytes and hybrid cloud and multicloud approaches now the norm, the need for more intelligent, automated storage infrastructure management has become critical as organizations grapple with mounting IT complexity and, in some cases, a lack of internal storage expertise.

One of the most promising developments in the long-standing quest to simplify, if not automate, the administration of storage systems is the introduction of AI-driven capabilities. Many storage arrays leverage AI/ML to monitor storage system health, detect anomalies that could indicate performance or security issues, and trigger alerts to help IT departments orchestrate a response. Now, many are taking the next step by adding agentic AI capabilities to execute commands and automate manual storage tasks.

IBM's 2026 FlashSystem release, with its new fifth-generation FlashCore Module (FCM), introduced AI agents that can take recommended actions, such as provisioning capacity, migrating application data to the optimal storage array, and enforcing data protection policies. FlashSystem also adds natural language processing, allowing administrators to ask plain-language questions via a chat interface as well as refine and confirm the AI engine's suggestions before applying them. IBM views the new FlashSystem.ai capabilities as an always-on storage co-administrator that can handle mundane tasks, help facilitate high utilization, proactively address performance issues, and strengthen security.

For enterprise IT leaders, the core question becomes: At what point does AI-driven storage management deliver sufficient accuracy, transparency, and control to serve as a trusted partner in day-to-day operations rather than an emerging capability that remains a work in progress? As the storage industry explores various design choices, IBM's approach combines hardware-accelerated analysis in its custom FCM with natural language interaction and human-in-the-loop oversight to give administrators visibility into the AI recommendations and the ability to accept, modify, or reject them. Pairing agentic automation with administrator review is a credible path forward. Still, vendors must continue to back their claims with examples of positive real-world outcomes to give customers greater confidence in the AI-driven technology.

Situation overview

AI-driven infrastructure optimization and automation are now high priorities for enterprise IT organizations. IDC's December 2025 *Future Enterprise Resiliency & Spending (FERS) Survey* Waves 8 and 9 finds that the technology spending areas most immune to budget reduction are AI initiatives, infrastructure and IT operation optimization, and security, risk, and compliance. Looking at where spending will grow in 2026, the projected increases will focus primarily on

building customized AI agents to automate business processes, enhancing cyber-recovery and resilience, and modernizing infrastructure in corporate datacenters. The critical software priority within that modernization push is investing in infrastructure automation and AI-enabled observability solutions.

Those broad IT priorities flow directly into data storage infrastructure. IDC's Quarterly Enterprise Infrastructure Tracker: Workloads 2025H2 Forecast (May 2026) shows the top workloads driving OEM storage capacity shipments in 2030 will be structured databases and data management at 16.4 exabytes (EB), followed by AI life cycle at 15.0EB. As always, organizations must keep mission-critical databases and key business applications running at peak performance without interruption. At the same time, they now also need to modernize their storage infrastructure to address the data throughput and capacity demands of new AI and analytics workloads.

As business stakeholders increasingly demand rapid capacity provisioning, IT departments are often stretched thin and may lack specialized storage expertise. Growing data volumes spanning on-premises arrays and public cloud environments compound the management complexity for those teams. Adding to the load, ransomware has become a board-level concern, intensifying the need for stronger security and faster, more automated detection and recovery. Regulatory frameworks such as the European Union's General Data Protection Regulation and AI Act, along with industry-specific compliance mandates, impose additional obligations.

Against this backdrop, AI-driven capabilities and automation represent the next wave of innovation, aiming to ease the burden on storage administrators by alleviating labor-intensive management tasks, strengthening security postures, accelerating recovery from cyberattacks, and optimizing systems to deliver consistent performance at scale.

Considering IBM FlashSystem

FlashSystem has been IBM's leading product in the external storage array market since 2020. IBM launched the block-based FlashSystem in 2013, leveraging technology from its acquisition of Texas Memory Systems, and completed a major redesign in 2014, changing it from a speed-focused device into a full-featured enterprise storage system. Storage technologies that IBM has added over the years include data compression and deduplication, thin provisioning, dynamic tiering, multisite and active-active replication, immutable "safeguarded copy" snapshots, and latency-lowering nonvolatile memory express (NVMe) technology.

Some of FlashSystem's most important capabilities are enabled by its custom-engineered FlashCore Module. IBM is one of only two major storage vendors to build its own NAND flash modules rather than procuring off-the-shelf solid-state drives (SSDs) from third parties. In line with industry trends, IBM increased the density of its FCMs over time, starting at single-level cell (SLC) flash storing one bit per cell to multilevel cell (MLC), triple-level cell (TLC), and the current quad-level cell (QLC), at four bits per cell, to help customers lower costs and datacenter footprint. The main trade-off as flash density increases is a decline in endurance and write speed, rendering QLC best suited to read-intensive applications in many storage arrays. However, IBM's design goal was to make its QLC-based FCMs suitable for any enterprise

workload, so the company developed firmware to enable the flash modules to use higher-performance SLC pages for hot data and higher-capacity QLC pages for colder data.

IBM sees another key point of distinction in its implementation of computational storage, via the FlashCore Module's onboard field-programmable gate array (FPGA) and embedded processor core. IBM's FCM handles compression, encryption, and, in the latest generation, deduplication directly on the drive, offloading them from the FlashSystem controller to free up array resources for other tasks. IBM maintains that shifting those processor-intensive functions to the drive, coupled with the use of QLC flash even in its highest-performance storage systems, allows FlashSystem to use less powerful controller processors and less DRAM, helping reduce system costs, lower power consumption, and shrink the storage impact on datacenter floorspace.

With the latest series of FCM developments, IBM has turned its focus to AI technology. The fourth-generation FCM, released in 2024, adds AI-driven ransomware threat detection and cyberattack prevention. IBM says the FCM's FPGA collects information on every input/output (I/O) operation, not merely a sample, to enable the system to recognize unusual activity that could signal the start of an attack in real time. The flash module aggregates more than 40 types of data statistics. In tandem, the FlashSystem array's built-in AI inference engine continuously evaluates them, comparing behavior patterns against IBM Research–developed ML models trained on real-world ransomware, and varying the sampling window to balance sensitivity and nuisance alerts. Because the heaviest computation runs in the FCM's FPGA and embedded cores, IBM says it can analyze every I/O without the performance penalty that would occur if the FlashSystem controller did the same work. According to IBM, the AI-driven FCM can detect the start of a ransomware encryption process within a minute. Alerts surface either through IBM's cloud-based Storage Insights software, which can apply additional filtering, or directly on the array for organizations that cannot connect to the cloud. IBM claims telemetry gathered across production systems over roughly two years has helped reduce the false-positive rate to less than 1%. FlashSystem can automate the creation of immutable "safeguarded copy" snapshots for continuous data protection. IBM offers a cyber-recovery guarantee to restore immutable snapshots from safeguarded copies within 60 seconds, mitigating the risk of lost revenue, regulatory exposure, and recovery costs associated with a cyberattack.

The new 2026 FlashSystem builds on its AI-driven foundation with the launch of agentic automation. IBM's stated objective is to create a storage system that can self-learn, self-improve, analyze threats, proactively flag performance and security risks, and take action based on models trained on billions of data points collected from telemetry and years of real-world operational data. Delivered as a layer of data services called FlashSystem.ai, the software is designed to help administrators manage, monitor, diagnose, and remediate issues across the data path and execute routine decisions that previously required manual intervention. The AI functionality runs primarily on the on-premises FlashSystem arrays, with the system also connecting to cloud-based IBM Storage Insights software.

For enterprise users, the practical effect can be a shift from reactive troubleshooting to proactive, policy-driven operation. FlashSystem.ai agents can execute labor-intensive tasks such as storage provisioning, data migration, non-disruptive workload placement on an optimal IBM or third-party array, SLA enforcement, and adherence to security, data protection, and compliance requirements. IBM designed the agents to adapt to application behavior and reduce

audit and compliance documentation time by auto-assembling records of configurations, protection policies, and operational history. Administrators can view the reasoning behind each AI-driven recommendation, interact with the system through a natural-language chat interface, and accept, modify, or reject automated suggestions. The human-in-the-loop safeguard helps reduce the risk of incorrect resource provisioning, errant data migration, or suboptimal workload placement. IBM claims these capabilities could significantly reduce routine storage management tasks compared with performing them manually in a drive's graphical user interface.

IBM illustrated the experience at the 2026 FlashSystem launch, where an administrator used a single chat interface to complete three different jobs:

- The first provisioned capacity for an application team's ticket, including the disaster recovery policy and the choice of array, then updated and closed the ticket with a full record of the commands and reasoning behind the recommendation.
- The second resolved an SLA problem by non-disruptively migrating a workload from an old array to a new one within a FlashSystem grid, with disaster recovery protection policies moving with it.
- The third closed a cyber-resilience gap by applying safeguarded snapshot policies to unprotected volume groups. The system automatically summarized and logged each action, providing the administrator with an audit trail of the AI-driven activity.

In addition to AI functionality, the fifth-generation Flash Core Module improves storage efficiency over prior models by supporting a maximum raw capacity of 105.6TB per drive and adding data deduplication alongside existing compression capabilities. IBM says the on-board FPGA's function-offload capabilities allow FCM drives to use 85% to 95% of their potential capacity without performance degradation, versus roughly 70% if deduplication and compression ran on host controllers. Higher usable capacity means organizations could address storage requirements with fewer drives and reduce capital expenses, rack space, and energy costs. The FCM supports quantum-safe cryptography algorithms to guard against emerging threats. Three new 2026 FlashSystem models address varying deployment needs:

- **FlashSystem 5600.** Up to 633TB raw and 2.4PB "effective" capacity (after data reduction and RAID overhead) and 2.6M I/O per second (IOPS) in a single rack unit, targeting smaller application needs and space-constrained edge and remote office locations.
- **FlashSystem 7600.** Up to 1.6PB raw and 7.2PB effective capacity and 4.3M IOPS in a 2U system, aimed at large, virtualized environments and analytics workloads.
- **FlashSystem 9600.** Up to 3.3PB raw and 11.8PB effective capacity and 6.3M IOPS in a 2U system, consolidating space over the prior 9500 model (1.5PB in 4U), for mission-critical workloads in regulated industries such as government, financial, healthcare, large and AI-driven applications, and consolidated use cases, with energy efficiency of 0.8 watts per usable terabyte at maximum capacity.

Challenges/opportunities

AI-driven storage management can offer significant benefits, but the tools and capabilities can also produce false positives and errant results that could disrupt storage systems and business

operations. Tuning models and algorithms to minimize false alarms without creating blind spots to genuine threats remains a challenge across the industry, as normal system behavior can mimic suspicious activity. Storage vendors such as IBM have worked to improve cyberthreat detection, claiming accuracy rates exceeding 99%.

Agentic automation introduces the additional risk of incorrect or suboptimal actions if an AI agent mischaracterizes workload requirements and provisions inadequate capacity, migrates data to a poorly suited array, or applies a policy incorrectly. IBM mitigates the risk by allowing customers to review FlashSystem.ai's reasoning and recommendations and approve, modify, or reject suggested changes before an agent executes an action. Giving organizations the option to adjust the level of automation enables a gradual transition to autonomous operation. IBM and other storage vendors can instill greater confidence in AI-driven capabilities by providing independently validated test results and real-world proof points to back their claims.

Conclusion

AI-driven capabilities are becoming increasingly critical for reducing complexity, optimizing performance, and strengthening security and cyber-resiliency in modern enterprise storage environments. Staffing-challenged IT organizations face demands for quick capacity provisioning amid growing volumes of data spanning on-premises and public cloud environments. The rise of data-intensive AI and analytics workloads, alongside existing mission-critical databases and business applications, further intensifies the need for storage systems that can deliver high performance, availability, and scalability while minimizing labor-intensive administration.

IBM FlashSystem exemplifies the trend toward AI-driven storage with new agentic AI capabilities designed to co-manage storage systems with administrator input. FlashSystem.ai agents can execute actions — such as provisioning capacity, migrating data, and enforcing data protection policies — while natural language processing enables administrators to query the system via a chat interface. IBM FlashCore Modules add AI-driven ransomware detection that IBM says can identify the start of an attack in less than a minute, with a false-positive rate below 1%.

Although AI-driven storage management can deliver significant benefits, challenges remain around minimizing false positives and ensuring reliable automation. With its newly added agentic AI capabilities, IBM FlashSystem enables customers to approve, modify, or reject recommended actions. Given the flexibility and potential upside of new AI-driven capabilities, IBM FlashSystem offers a viable option for consideration to enterprises seeking to simplify their data storage infrastructure.

Definitions

A **field-programmable gate array (FPGA)** is an integrated circuit that can be programmed or reprogrammed to suit different purposes after the manufacturer releases the chip, without

physically changing the hardware. An FPGA offers greater flexibility than a traditional fixed-function chip, such as an application-specific integrated circuit (ASIC), which has logic hardwired during manufacturing and cannot be altered after leaving the factory.

NAND flash is a type of nonvolatile memory that does not require power to retain data. Flash types available today are:

- Single-level cell (SLC), storing 1 bit per cell
- Multilevel cell (MLC), storing 2 bits per cell
- Triple-level cell (TLC), storing 3 bits per cell
- Quad-level cell (QLC), storing 4 bits per cell

As flash density increases, performance, endurance, and manufacturing costs decline. TLC is currently the dominant flash type in enterprise storage systems, but QLC use is rising.

About the IDC analyst



Carol Sliwa
Research Director, Storage Systems, Core Infrastructure

Carol Sliwa is a Research Director in IDC's enterprise infrastructure global research domain and part of the core infrastructure subdomain. Carol is the lead analyst for IDC's global enterprise storage infrastructure coverage. Her core research spans block, file, and object storage, including scale-out parallel and distributed file systems for AI and performance-intensive computing environments and trends in flash media. Approaching 20 years covering the enterprise data storage industry, Carol provides qualitative and quantitative insights and advice to technology vendors, IT buyers, and the investment community. Her experience in technology journalism also adds a distinctive perspective on how the IT industry has developed and adapted technologies, platforms, and systems over time to address the evolving needs of IT organizations worldwide.

More about Carol Sliwa

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.
One Beacon Street, Suite 33100, Boston, MA 02108, USA
T +1 508 872 8200

blogs.idc.com | LinkedIn @IDC | www.idc.com

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community make fact-based technology decisions and achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. CCPA