



---

## Architecting for the Flash and Memory Supercycle

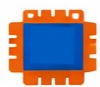
---

How to Reduce RAM Requirements, Extend Server Life, and Lower Infrastructure Costs When Flash and Memory Prices Keep Climbing

**A VergeIO White Paper**

March 2026

George Crump



**171%**  
DRAM Increase



**55-60%**  
Flash Price Jump



**Server**  
Delays

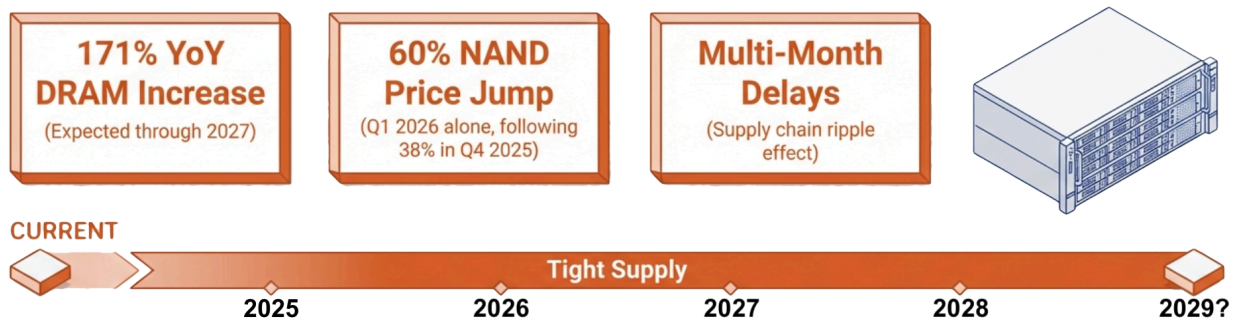


**VMware**  
Licensing Shift

# Opening

## The Impact

The data center industry is in the early stages of a pricing event that will reshape infrastructure purchasing decisions for the next two to three years. DRAM prices are expected to increase 171% year-over-year through 2027. NAND flash contract prices jumped 55–60% in Q1 2026 alone, following a 33–38% increase in Q4 2025. Enterprise SSD and server DRAM now represent the fastest-rising line items in data center budgets.



These price increases have already produced a second-order effect. Server orders that once shipped in weeks now face multi-month delays. Flash and memory shortages ripple through the entire supply chain, from component vendors to system integrators to end customers. Organizations that planned hardware refreshes for 2026 are discovering that the servers they budgeted for either cost significantly more than projected or are simply unavailable.

DDR4 production is winding down across all three major manufacturers. Samsung, SK Hynix, and Micron are executing a controlled end-of-life, shifting fabrication capacity toward DDR5, LPDDR5/x, and HBM. No manufacturer is adding new DDR4 capacity. DDR5 pricing carries steep premiums because AI infrastructure buyers will pay whatever the market demands for the fastest available memory. AMD is considering re-releasing older CPU generations that support DDR4 to give enterprise buyers a lower-cost option. This combination of factors is creating the **Flash and Memory Supercycle**.

## The Cause

AI factories are the primary driver. Companies building large language models and AI training infrastructure are purchasing flash and memory in volumes that distort the entire market.

OpenAI simultaneously purchased all available inventory from SK Hynix and Samsung, demonstrating how a single buyer can absorb supply that would otherwise serve thousands of enterprise customers. AI companies justify paying premium prices because their business models generate enough revenue per compute hour to absorb component cost increases that destroy traditional enterprise budgets.

Normal data center growth compounds the problem. Even without AI demand, enterprise data centers continue expanding to support digital transformation, hybrid cloud deployments, and growing data volumes. This baseline growth alone was enough to tighten supply. AI demand pushed an already-constrained market into a shortage.

The power problem adds a third pressure. New data center construction faces electrical capacity constraints in major markets. Organizations that cannot build new facilities must increase the density and efficiency of existing infrastructure. That means heavier workloads per server, which means more flash and more memory per node, increasing exposure to supercycle pricing. The organizations that find ways to run the same workloads on fewer servers gain a power and cooling advantage that compounds over the life of the equipment.

When will conditions improve? Industry analysts project tight supply and elevated pricing through 2027, with many forecasts extending into 2028. Some industry voices argue that structurally tight NAND will persist for up to a decade. The conservative baseline assumes at least two more years of pricing pressure. Planning around a shorter timeline carries significant risk.

## The Impact on IT Organizations

The supercycle creates three pressure points that IT organizations must address.

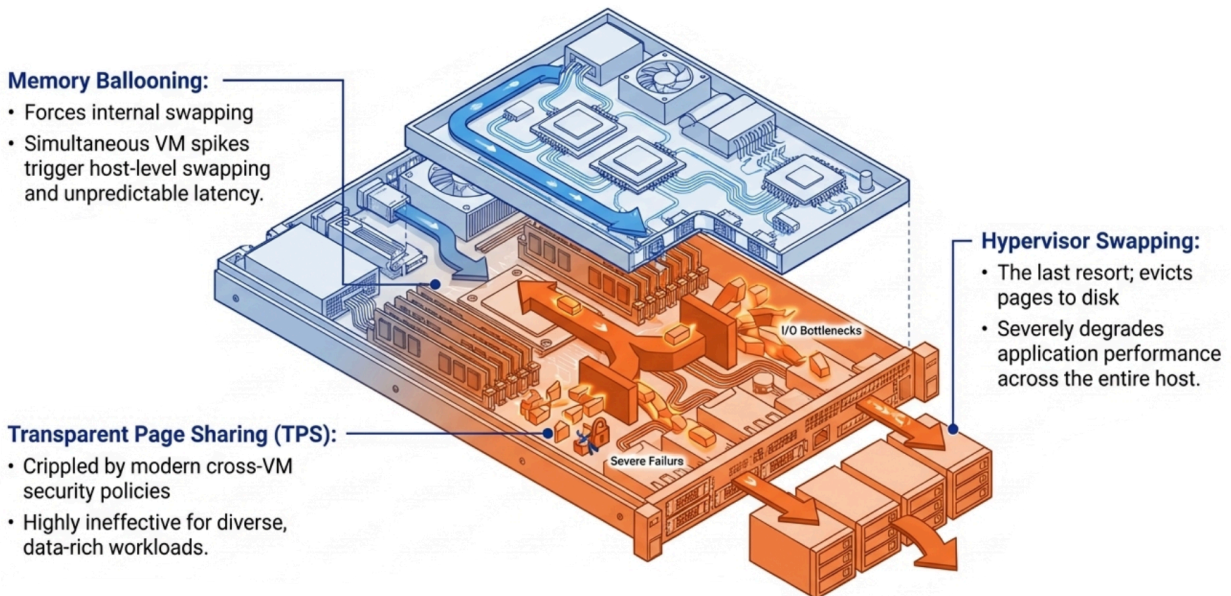
**Reduce server memory requirements.** Every gigabyte of DRAM installed in a server carries a higher cost than it did 18 months ago. Organizations that overprovisioned memory when prices were low now face renewal and expansion costs that break existing budgets. Reducing the amount of physical RAM required per workload directly lowers acquisition costs and extends the useful life of servers already in production.

**Reduce flash storage capacity.** Rising NAND prices tempt organizations to cut back on flash availability. Some teams consider abandoning dual-parity RAID or triple-mirror configurations to save on raw capacity. Trading data protection for cost savings is a dangerous path. The better approach reduces the total flash capacity required without weakening availability.

**Reduce server count.** Fewer physical servers means less flash and less memory to purchase. Consolidation driven by architectural efficiency, rather than by cramming more workloads onto oversubscribed hosts, addresses the supercycle at the infrastructure level. Fewer servers also reduce power consumption, cooling load, and rack space, addressing the coming power constraints that are limiting data center capacity across major markets.

## Workarounds: What the Industry Offers Today

Hypervisor vendors have developed several techniques to manage memory consumption across virtual machines. Each technique attempts to stretch available RAM further. None of them reduce the total physical memory the infrastructure requires. On the storage side, organizations are making dangerous compromises to control flash costs.



### Memory Ballooning

**What it is.** A balloon driver runs inside each guest operating system and cooperates with the hypervisor to reclaim unused RAM. When the hypervisor detects memory pressure, it tells the balloon driver to inflate by allocating pages inside the guest. The guest cannot use those pinned pages, and the hypervisor reassigns the underlying physical RAM to other VMs. When pressure drops, the balloon deflates and the guest regains access.

**What it does well.** Ballooning increases host RAM utilization by reclaiming memory from idle or overprovisioned VMs. It adjusts dynamically as workload demands change. It is less disruptive than hypervisor-level swapping because the guest OS decides which pages to give up, rather than the hypervisor blindly evicting pages from underneath the guest.

**Where it breaks down.** Ballooning works best when VMs have significant memory headroom. During a supercycle, organizations size VMs tighter to conserve expensive DRAM. Tighter sizing means less safe space for balloon inflation. When a VM actually needs its configured RAM, balloon inflation forces the guest to swap internally or drop its page cache. That triggers increased latency and disk I/O.

The worst failure mode occurs when multiple VMs spike simultaneously. All balloons try to deflate at the same time. The host runs out of physical RAM to hand back, triggering host-level swapping across every VM on the node. CPU and disk usage spike. Every workload on that host degrades regardless of its individual memory consumption. This cascading failure strikes unpredictably and resists prevention through monitoring alone.

Memory-intensive workloads expose a fundamental limitation. Large databases, in-memory caches, and analytics platforms actively use their full memory allocation. There is no unused RAM for the balloon to reclaim. Ballooning is either completely ineffective for these workloads or actively destructive.

## Transparent Page Sharing

**What it is.** Transparent Page Sharing (TPS) scans memory across VMs to identify identical pages. When it finds duplicates, it merges them into a single physical page with copy-on-write semantics. If twenty VMs run the same operating system, the shared OS binary pages consume physical RAM only once.

**What it does well.** TPS reduces the memory footprint of homogeneous environments where many VMs run identical operating systems and applications. It operates transparently without requiring changes inside the guest.

**Where it breaks down.** TPS only deduplicates identical pages. Application data, database contents, user-generated files, and runtime memory allocations produce unique pages that TPS cannot consolidate. In production environments with diverse workloads, TPS delivers modest savings.

Security concerns limit its deployment. VMware disabled large-page TPS by default in 2014 after researchers demonstrated that attackers exploited shared pages to leak data between VMs. Most production environments now run TPS in a restricted mode that limits deduplication to pages within a single VM, eliminating the cross-VM savings that made the technique attractive in the first place.

## Hypervisor-Level Swapping

**What it is.** When ballooning and compression cannot free enough memory, the hypervisor swaps guest pages to a .vswp file on disk. This is the last-resort reclamation mechanism. The hypervisor evicts guest memory pages to storage and retrieves them when the guest accesses them again.

**What it does well.** Swapping guarantees that the hypervisor reclaims memory regardless of guest cooperation. It prevents complete memory exhaustion on the host.

**Where it breaks down.** The performance penalty is severe. Even swapping to NVMe storage introduces latency orders of magnitude higher than RAM access. Swapping to spinning disk is catastrophic for application performance. Some newer configurations direct swap traffic to NVMe tiers to limit the damage, but this adds storage cost and complexity.

Hypervisor swapping also creates unpredictable performance for every VM on the affected host. The swapping VM suffers direct latency increases. Other VMs on the same host compete for the remaining memory and storage I/O bandwidth. The result is broad performance degradation that is difficult to diagnose because it manifests as application slowness rather than an obvious infrastructure alarm.

## Lowering Data Protection Levels

Rising flash prices push some organizations to reduce data protection to save on raw storage capacity. This is not a technical workaround. It is a calculated risk that trades resilience for cost savings.



### Eliminating RAID-DP:

- Downgrading to RAID 5 risks total array destruction if a second drive fails during multi-hour rebuild windows.



### Dropping Triple Mirroring:

- Double mirroring in distributed architectures exposes data to loss during routine maintenance windows.

**The Bottom Line: Supercycle economics tempt organizations to weaken protection, but the cost of extended downtime remains unchanged.**

**Eliminating RAID dual parity.** RAID-DP (RAID 6) protects against two simultaneous drive failures. Organizations under flash cost pressure may downgrade to single-parity RAID (RAID 5), which tolerates only one drive failure. With modern high-capacity SSDs, rebuild times after a single failure stretch to hours. A second failure during rebuild destroys the entire array. The probability of a second failure during a long rebuild window is not theoretical. It is a documented failure mode that RAID-DP exists to prevent.

**Dropping triple mirroring.** Triple-mirror configurations maintain three copies of every data block across separate drives or nodes. Some organizations drop to double mirroring to cut flash

consumption by a third. Double mirroring tolerates one failure. In a distributed storage architecture, a single-node failure during a maintenance window or firmware update can expose data to loss. Triple mirroring exists because production environments need to survive overlapping failure events.

**Abandoning mirroring entirely.** Under extreme cost pressure, some teams consider running workloads on single copies of data with no real-time redundancy, relying on periodic backups as their only protection layer. This eliminates the flash overhead of mirroring but creates a recovery gap measured in hours or days. Any failure between backup windows results in data loss. For production databases, transactional systems, and customer-facing applications, this approach is unacceptable.

Each of these decisions trades availability for flash capacity savings. The supercycle makes this trade-off tempting because the cost of additional SSDs has increased dramatically. But the cost of data loss or extended downtime has not decreased. The right response is not to weaken protection. It is to reduce the amount of flash capacity required by strong protection.

## The Core Problem with Current Techniques

Every workaround described above manages scarcity after it occurs or accepts increased risk. Ballooning shuffles RAM between VMs. TPS deduplicates identical OS pages. Swapping trade performance for capacity. Memory compression, another common technique, stores compressible pages in an in-RAM cache to delay swapping, but does not reduce total memory requirements. Reservations and limits redistribute available memory across priority tiers but do not create additional capacity. Lowering data protection levels saves flash at the direct expense of availability.

These are post-consumption memory management techniques and risk-based storage compromises. They react to scarcity after workloads have already consumed physical RAM, or they accept weaker data protection because the cost of doing it right has increased. During a supercycle, reacting to scarcity is not enough. Organizations need an architecture that reduces the amount of flash and physical memory the infrastructure requires in the first place, without compromising protection.

## The Architectural Comparison

The alternative to managing scarcity is to build infrastructure that, by design, consumes less flash and memory. This requires changes at the hypervisor and platform level that go beyond the workaround techniques traditional vendors offer.

Other virtualization platforms claim data reduction and efficiency features. Nutanix, VMware vSAN, and every HCI vendor advertise deduplication and compression. The difference is where and how these techniques operate. Most competitors run post-process deduplication or narrow inline deduplication within a single storage pool. Their memory management still relies on the same ballooning, TPS, and swapping techniques described above. VergeOS combines global inline storage deduplication with a global deduplicated RAM cache, reducing consumption at both the memory and storage layers simultaneously. No other platform delivers this combination.

### Efficiency: Less Overhead, Less Memory Per VM

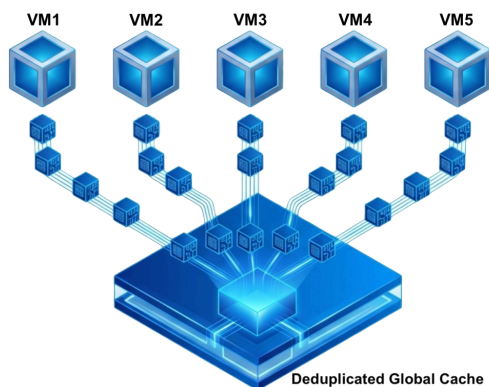
Traditional virtualization stacks carry significant memory overhead. VMware environments running vSphere, vSAN, vCenter, and NSX dedicate substantial RAM to management processes before a single VM even starts. Each layer of the stack reserves memory for its own operations. The cumulative overhead consumes tens of gigabytes per host that are unavailable to production workloads.

VergeOS takes a different approach. The entire platform, including hypervisor, storage, networking, and data protection, runs at only 2–3% memory overhead. VMware environments running vSphere, vSAN, vCenter, and NSX consume a significantly higher percentage of installed RAM before a single VM even starts. Every percentage point of overhead saved is RAM returned to production workloads.

The efficiency difference is measurable in production. Topgolf, a global sports entertainment company operating more than 100 venues, reduced each venue's infrastructure from six-node VxRail clusters to three-node VergeOS clusters while supporting the same workloads. That is a 50% reduction in physical servers per site, driven by the platform's lower overhead and more efficient resource consumption. Across 100+ venues, the consolidation eliminated hundreds of servers from the company's global footprint. Alinsco Insurance reported similar results. Servers originally constrained under VMware gained capacity headroom under VergeOS on the same hardware, with no additional RAM purchased.

## VergeOS Techniques

### Global Deduplicated Hypervisor-Level Cache



Traditional architectures give each VM its own storage cache. If ten VMs on the same host access the same data blocks, each VM maintains its own cached copy. That means ten copies of identical data sitting in RAM across ten separate VM caches.

VergeOS implements a global deduplicated cache at the hypervisor level. The cache operates below the VM layer and deduplicates cached blocks across all workloads on the node. When ten VMs access the same data, the cache stores one copy. This design produces three benefits during a supercycle.

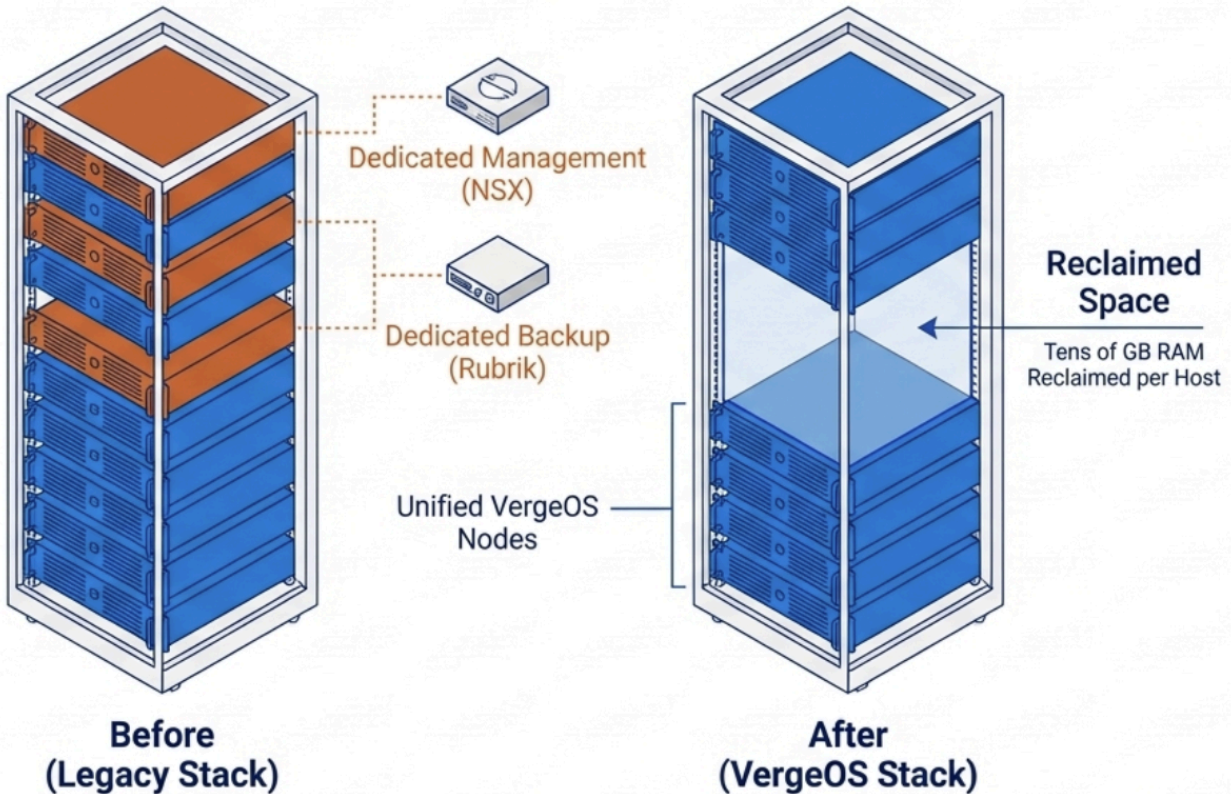
First, the deduplicated cache eliminates redundant cached data, directly reducing RAM consumption. In environments with common operating system images, shared application binaries, or overlapping datasets, the savings compound across every VM on the node.

Second, eliminating per-VM caching removes the memory overhead that traditional architectures assign to each VM's individual cache. That RAM returns to the available pool for production workloads or additional VM density.

Third, a global deduplicated cache improves storage read performance. Deduplication means more unique blocks fit in the same physical RAM, so cache hit rates increase. Higher hit rates reduce reads from flash storage. Fewer flash reads lower I/O latency and reduce wear on SSDs, extending drive life during a period when replacements cost significantly more.

### Integrated Infrastructure Services

Traditional enterprise virtualization requires separate products for networking and data protection. VMware environments typically run NSX for software-defined networking and additional products for backup and disaster recovery. Each product consumes memory on every host where it operates.



NSX alone reserves significant RAM per host for distributed firewall, routing, and overlay network processing. Third-party backup agents, replication engines, and snapshot managers add further memory overhead. The total memory that infrastructure services consume, not VMs running production workloads, represents a substantial percentage of installed RAM on each host.

Here are some example memory consumption prior to the first VM being created:

- **vSAN OSA:** 4–10+ GB per host for storage alone, depending on disk group config
- **vSAN ESA:** ~32 GB on a 64 GB system (51% overhead), with supported minimums of 128–512 GB RAM per host
- **Nutanix CVM:** 24–32 GB reserved per node

(Neither includes ESXi overhead, vCenter, NSX, or backup agents)

- **VergeOS:** 2–3% for the entire stack — on a 512 GB host that's roughly 10–15 GB total

VergeOS integrates networking and data protection into the platform itself. Software-defined networking runs as part of the operating system rather than as a separate product with its own memory footprint. The networking stack consumes less memory than NSX because it does not

require a separate management plane, distributed agents, or overlay processing engines running in dedicated VMs.

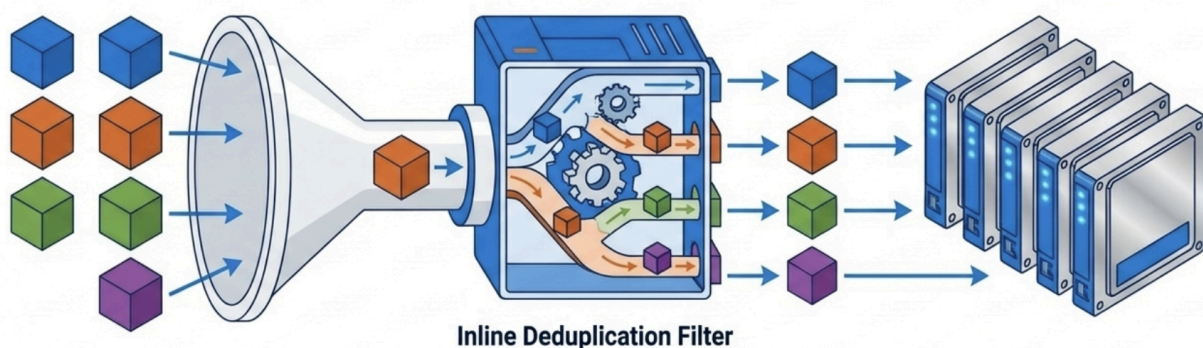
Data protection follows the same integrated model. Snapshot, replication, and disaster recovery capabilities operate within the platform rather than through external products. The protection engine shares the platform's existing memory management rather than maintaining its own separate allocation, consuming a fraction of the memory that third-party solutions require.

Topgolf validated this integration at scale. After migrating to VergeOS, the company eliminated its siloed backup infrastructure. The platform's built-in data protection, including ioClone, ioGuardian, and ioReplicate, replaced the separate backup product entirely. That removed an additional layer of software licensing, memory overhead, and operational complexity from every venue and data center node. Across 100+ locations, the savings in licensing alone contributed to millions in annualized cost reduction.

This integration matters during a supercycle because every gigabyte consumed by infrastructure services is a gigabyte unavailable for production workloads. Reducing infrastructure overhead directly increases the number of VMs each host supports, which reduces the total number of servers an organization needs to purchase at supercycle prices.

### Bringing Down the Cost of Flash

The memory-side techniques reduce RAM requirements. The flash-side challenge is different. Organizations need to store the same amount of data with less raw flash capacity, without compromising performance or availability. VergeOS addresses this through three capabilities.



**Safely use commodity flash with high performance.** Traditional storage architectures require enterprise-grade SSDs to deliver consistent performance and endurance. Enterprise SSDs carry significant price premiums over commodity NVMe drives. VergeOS delivers high storage performance on commodity flash hardware. The platform's storage engine manages wear leveling, garbage collection, and I/O scheduling at the software layer, reducing dependence on the drive's internal controller for performance consistency. Organizations can deploy lower-cost NVMe drives without sacrificing the throughput and latency their workloads require.

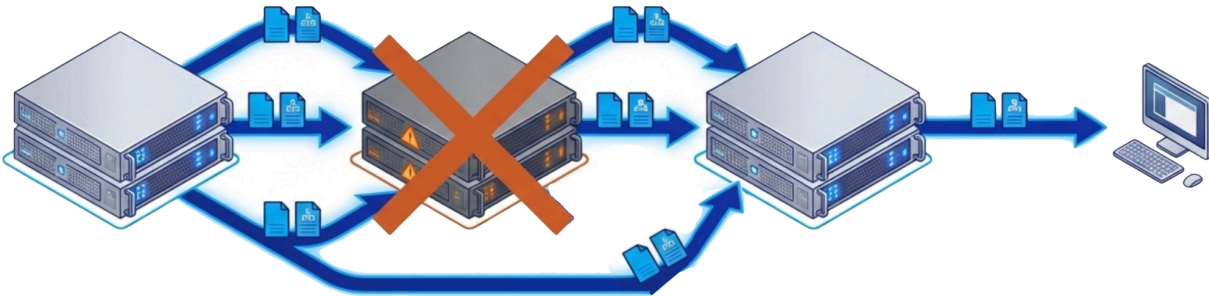
During a supercycle, the price gap between enterprise and commodity flash widens. The ability to use commodity drives safely translates directly into lower storage acquisition costs. The endurance concern that drives enterprise SSD selection is addressed at two levels. Global inline deduplication reduces the total volume of data written to each drive, directly extending drive life by lowering write amplification. ioGuardian's synchronous replication protects against drive failures gracefully, meaning a commodity drive that wears out faster than an enterprise drive is replaced without data loss or downtime. The combination of reduced writes and fault-tolerant replication makes commodity flash a safe production choice.

**Global inline deduplication.** VergeOS performs inline deduplication at the storage layer across all workloads on the cluster. Unlike memory-level TPS, which only deduplicates identical memory pages within narrow constraints, storage-level deduplication operates on every data block written to flash. Environments running multiple instances of the same operating system, application binaries, container images, or overlapping datasets see significant capacity reduction. The deduplication runs inline, meaning the system identifies and eliminates duplicate blocks before they consume flash capacity. This directly reduces the number of SSDs an organization needs to purchase. In environments with moderate to high data similarity, deduplication ratios of 2:1 or higher are common, cutting effective flash costs proportionally.

Topgolf's deployment illustrates the storage efficiency at scale. Each venue reduced storage requirements from 20 TB under VxRail to 5 TB per node under VergeOS, a direct result of global deduplication and the platform's integrated storage architecture, eliminating the overhead of separate caching and backup tiers.

Global deduplication also plays a critical role in offsetting the capacity cost of mirrored writes. When ioGuardian writes two or three synchronous copies of data for protection, deduplication reduces the effective capacity consumed by those copies. The combination of replication-based protection and global deduplication delivers strong data availability without the raw-flash overhead that traditional mirroring requires.

**ioGuardian RF3+ (N-X) protection at N+1 cost.** This is where VergeOS addresses the data-protection trade-off that supercycle economics impose on other platforms. Traditional triple-mirror protection requires 3x the raw flash capacity. RAID-DP requires less overhead than mirroring but introduces write penalties from parity calculations and suffers degraded read performance during drive failures. Both approaches are expensive when flash prices are elevated, and both carry performance trade-offs under failure conditions.



ioGuardian takes a fundamentally different approach. Instead of erasure coding or parity-based RAID, ioGuardian uses synchronous two-way or three-way replication. Every write is replicated to two or three nodes simultaneously. This design delivers three advantages that parity-based and erasure-coded systems cannot match.

First, write performance. Synchronous replication involves no parity calculations, no Reed-Solomon encoding, and no write amplification due to erasure-coding overhead. A write is a write, replicated in parallel across nodes. The system delivers full write-throughput without the CPU overhead and latency that parity-based systems introduce with every I/O operation.

Second, performance during failure. When a node or drive fails in a parity-based system, the surviving nodes must reconstruct missing data from parity blocks before serving reads. That reconstruction competes with production I/O, degrading performance for every workload on the cluster until the rebuild completes. ioGuardian's replication model eliminates the need for reconstruction entirely. When a node fails, surviving copies serve reads directly at full speed. Production workloads see no performance degradation during the failure window.

Third, rapid rebuild. Reconstructing data from parity is computationally expensive and time-consuming, especially with high-capacity drives. ioGuardian rebuilds by copying data from surviving replicas, a simpler and faster operation. Shorter rebuild windows mean less time spent in a degraded state and a lower probability of an additional failure during recovery.

The cost equation closes the argument. RF3+ protection with three-way synchronous replication nominally requires 3x raw capacity. Global inline deduplication reduces that overhead by eliminating duplicate blocks across all three copies. The effective flash cost of RF3+ protection under VergeOS approaches N+1 rather than the full 3x that traditional mirroring demands. Organizations get protection that exceeds RAID-DP and matches triple mirroring, with write performance and failure behavior that surpass both, at a flash capacity cost that fits within supercycle budgets.

## Use the Servers You Already Own

The techniques described above reduce the amount of flash and memory each server requires. The next step uses existing servers, reducing the overall number of servers the environment requires.

VergeOS installs on standard x86 servers from any manufacturer. The platform abstracts the underlying hardware, meaning it does not depend on a specific vendor, model, or configuration to operate.

Organizations migrating from VMware install VergeOS on the same physical servers they already run. There is no hardware forklift. The servers, memory, and flash storage already purchased and deployed remain in production. VergeOS runs standard VM formats and does not require proprietary hardware, so the decision to adopt it does not create a new lock-in dependency.

Alinsco Insurance, a Texas-based insurer, demonstrated this directly. The company ran VMware and vSAN on a five-node VxRail cluster in Fort Worth. After migrating to VergeOS, those same VxRail servers and their internal SSDs continued to remain in production. There was no hardware replacement. The servers that Dell originally sold as part of a VMware-specific appliance continued to run production workloads on a different platform. During a supercycle, the ability to keep existing hardware in service and avoid purchasing replacement servers at inflated prices represents immediate capital savings.

The efficiency gains from global deduplicated caching, low platform overhead, and integrated infrastructure services translate directly into higher consolidation ratios. An organization running 10 hosts under VMware typically consolidates to 7 or 8 hosts under VergeOS while maintaining the same workload count. Topgolf saw even greater consolidation, reducing from six nodes to three per venue. That consolidation frees servers from production duty.

Those freed servers create immediate options. One server becomes a dedicated ioGuardian node, delivering RF3+ data protection without requiring a new hardware purchase. The remaining servers become parts donors. Their DRAM and NVMe drives are pulled and redistributed across the active production nodes, increasing memory and flash capacity where workloads need it most.



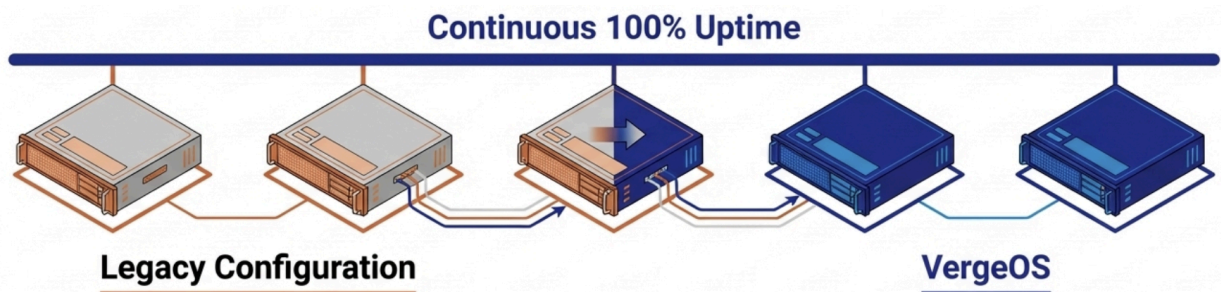
The redistribution does not require uniform hardware. VergeOS supports mixed node types and mixed node roles within the same cluster. One node running 256 GB of RAM and another running 192 GB operate together without restriction. A node with four NVMe drives and a node with six NVMe drives contribute to the same storage pool. The platform's storage engine normalizes performance across nodes with different drive counts, preventing the inconsistent behavior that heterogeneous clusters produce on platforms that expect identical hardware configurations. This flexibility eliminates the requirement that every node in a cluster match the same specification, a constraint that forces organizations on other platforms to purchase identical hardware even when existing inventory holds usable components.

## Migration Without Disruption

Architecture only matters if organizations can reasonably adopt it. The strongest technical argument fails if the migration itself is expensive, risky, or requires extended downtime.

VergeOS supports node-by-node migration from VMware. An organization evacuates workloads from one host, installs VergeOS on that host, migrates VMs onto the new platform, and repeats the process across the remaining hosts. Production workloads continue running during the migration. There is no maintenance window that takes the entire environment offline.

Alinsco Insurance completed this process on a production VxRail cluster that runs the company's mission-critical insurance application. The system cannot tolerate downtime. Alinsco migrated node by node during business hours with zero downtime. Critical web servers were moved at night out of an abundance of caution, but even those migrations produced no service interruption. The entire five-node VxRail environment transitioned to VergeOS without a single maintenance window.



Topgolf is executing the same approach across more than 100 venues nationally, converting VxRail clusters to VergeOS site by site. The company reports zero failures since migration and 50% faster site boot times compared to the VMware environment it replaced. Live upgrades now happen during business hours, and processes that previously required 30 to 45 minutes of manual effort per site now run automatically through the platform's API.

This migration model matters during a supercycle because it eliminates the capital expense of standing up a parallel environment. The organization does not need to purchase new servers to build a VergeOS cluster alongside the existing VMware environment. It converts the servers it already owns, one at a time, while production continues.

## Conclusion

The Flash and Memory Supercycle is not a short-term pricing fluctuation. It is a structural market condition driven by AI demand, constrained fabrication capacity, the end-of-life of DDR4, and data center power limitations. Organizations that wait for prices to normalize will wait for years.

Traditional memory management techniques react to scarcity after it occurs. They shuffle RAM between VMs, deduplicate a narrow slice of identical OS pages, or swap guest memory to disk at severe performance cost. On the storage side, organizations are weakening data protection to save on flash capacity, trading resilience for short-term savings. These are not strategies. They are concessions.

The architectural approach is different. VergeOS reduces memory requirements through a global deduplicated cache that eliminates redundant data across VMs, with only a 2–3% platform overhead, and integrated infrastructure services that consume less RAM than the multi-product stacks they replace. On the flash side, commodity drive support lowers acquisition costs, global inline deduplication reduces capacity requirements and drive wear, and ioGuardian delivers synchronous multi-copy protection with full write performance, zero degradation during failure, and rapid rebuild at a fraction of the flash overhead traditional approaches demand.

The production evidence supports these claims. Topgolf reduced server count by 50% per venue across 100+ locations, eliminated backup software licensing, and saved millions annually. Alinsco Insurance migrated a mission-critical environment from VxRail during business hours, with zero downtime, and extended the life of hardware already in production. Both organizations run on the servers they already own.

During a supercycle, architecture is not optional. It is the primary lever IT organizations have to control costs without sacrificing performance or availability. The organizations that treat this market condition as a catalyst for architectural improvement will emerge with more efficient, more resilient, and less expensive infrastructure than those that rely on workarounds and compromises.