WHITEPAPER

Magnition

The Origins of CHOPT

Foundational Research Behind AI-Powered System Design at Magnition

2025

Introduction

For leaders responsible for building and delivering cloud, storage, or CDN platforms, performance and efficiency are directly tied to business results. Whether you're developing next-generation storage arrays, distributed file systems, or global CDN services, your enterprise customers expect high quality, consistent reliability, and a tangible ROI.

This whitepaper illustrates why single-tier caching models can't meet the demands of today's storage and CDN platforms, and how the Magnition platform empowers technology leaders to deliver measurable value, secure competitive advantage, and maintain full transparency for their customers.

A Legacy Problem that Needs a Modern Solution

Historically, platform teams have relied on single-tier memory hierarchy caching models and design tools to make architecture and investment decisions. An important example includes the widely taught Belady's MIN, a theoretical "best case" for cache efficiency, assuming all memory is the same and every data access should be cached. Similarly, single-tier LRU models are often used to help with critical sizing and architectural design choices. In the past, tools like these were considered sufficient to compare new designs, guide hardware spend, and justify premium SKUs.

However, those assumptions no longer hold. Modern storage and CDN platforms now span several layers—DRAM, NVM, SSD, and CXL-attached memory or distributed caches. Each tier has unique costs, speeds, and trade-offs. Simple cache hit rates can hide expensive bottlenecks, inefficient use of high-value memory, or missed opportunities for cost savings.

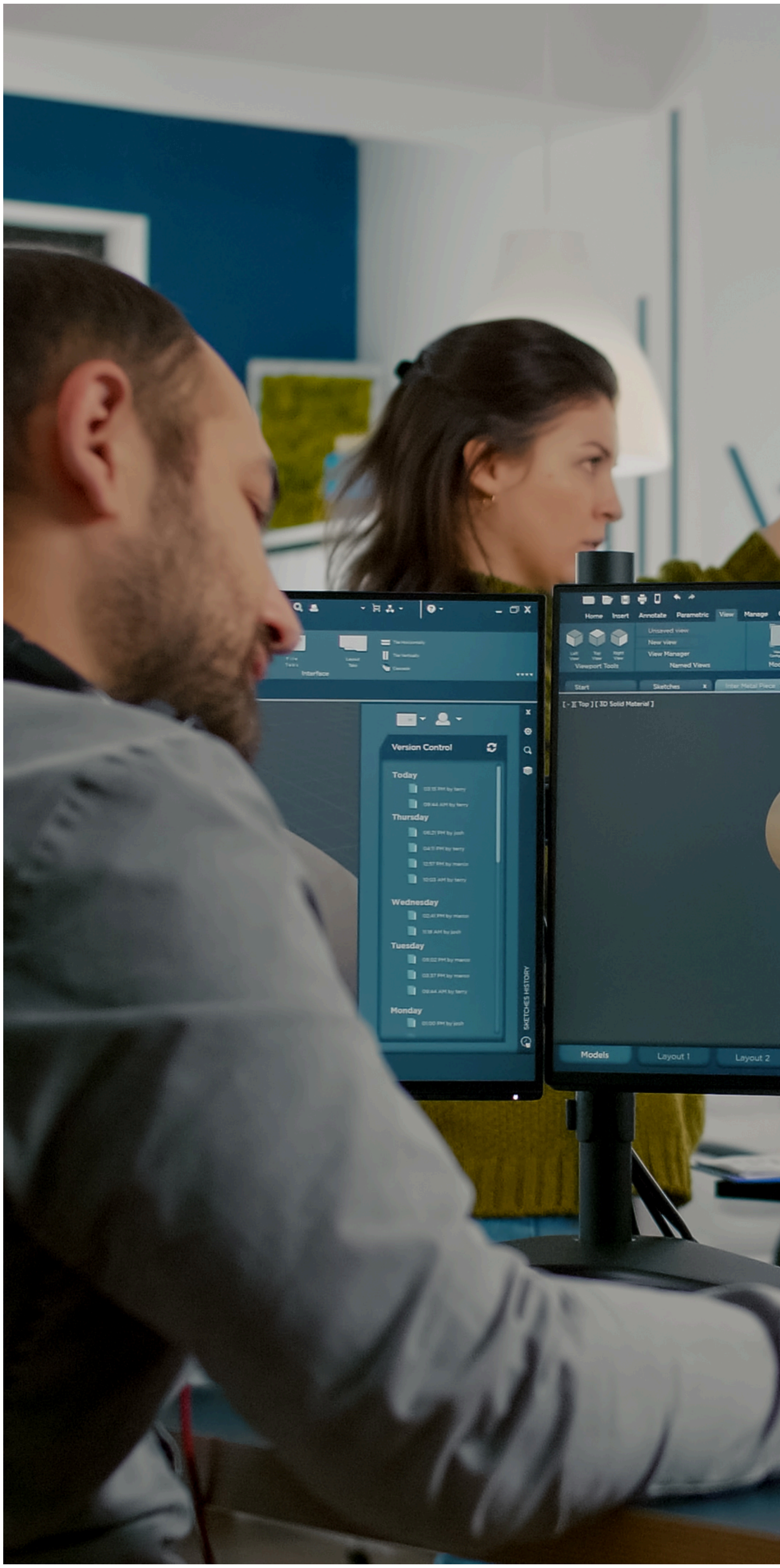
Single-tier models overlook critical choices like when to bypass cache, how to handle asymmetric read/write penalties, and how to optimize end-to-end performance across multiple tiers and geographies. Simply put, you can no longer use single-tier models for modeling the ground reality of your multi-tier memory and storage systems.

Balancing Performance, Cost, and Business Risk

For SVPs and GMs in storage and CDN businesses, this isn’t just a technical problem—it’s a business risk. Relying on outdated benchmarks can lead to over-provisioned products, squeezed margins, or platforms that can’t deliver on their performance claims. It can slow down innovation, create uncertainty in roadmap decisions, increase time-to-market, and make it hard to demonstrate differentiated value to customers who demand transparency and measurable ROI.

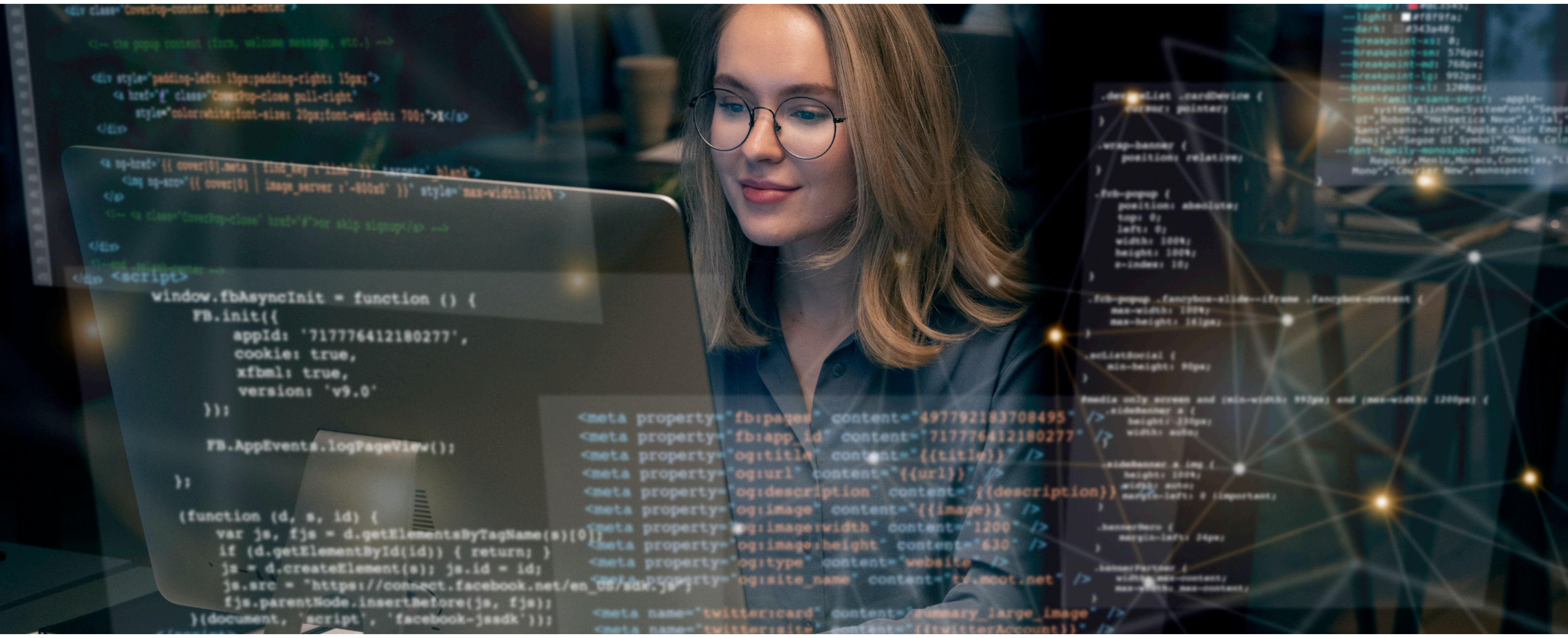
Magnition addresses this new reality with a research-driven approach designed for complex, multi-tier platforms that has received 3 of the highest awards in computing research. For example, our Choice-aware Offline Placement Theory (CHOPT) pushes past the performance ceilings of existing placement methods determining exactly what to cache, when to bypass, and applying real-world cost metrics across every memory and storage tier.

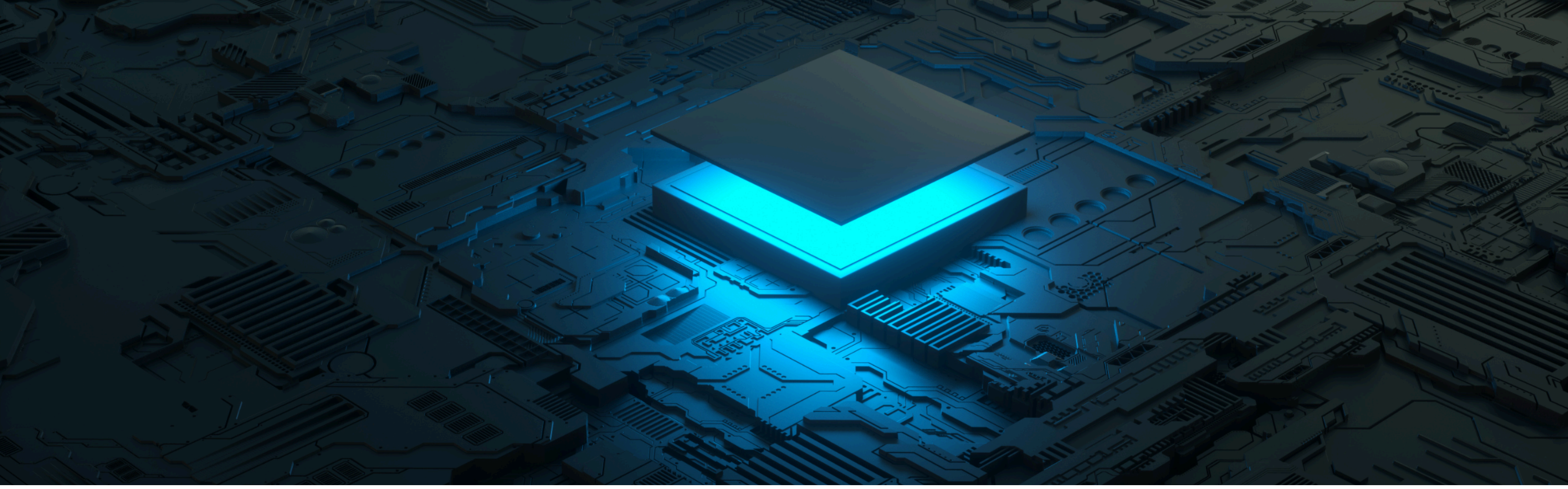
Similarly, our Test-of-Time award-winning Spatially Hashed Approximate Reuse Distance Sampling (SHARDS) technology enables fast, accurate simulation on real, production-sized workloads using the actual data patterns that matter to your customers. Combined, Magnition’s technology makes complex storage, cloud, and CDN systems design and modeling achievable for the first time in history.



With Magnition System Designer, platform business leaders and architects can:

- Quantify true system headroom and efficiency across any platform or deployment
- Model and justify investments in new memory or caching technologies with hard numbers, not assumptions
- Diagnose and eliminate waste, improving platform margins and reducing unnecessary spend
- Predict and de-risk the impact of new product features, tiering policies, or SLA requirements before rollout
- Offer data-driven performance and TCO guarantees to enterprise and carrier customers, differentiating in a crowded market





Why Single-Tier Memory Hierarchy Models Break Down

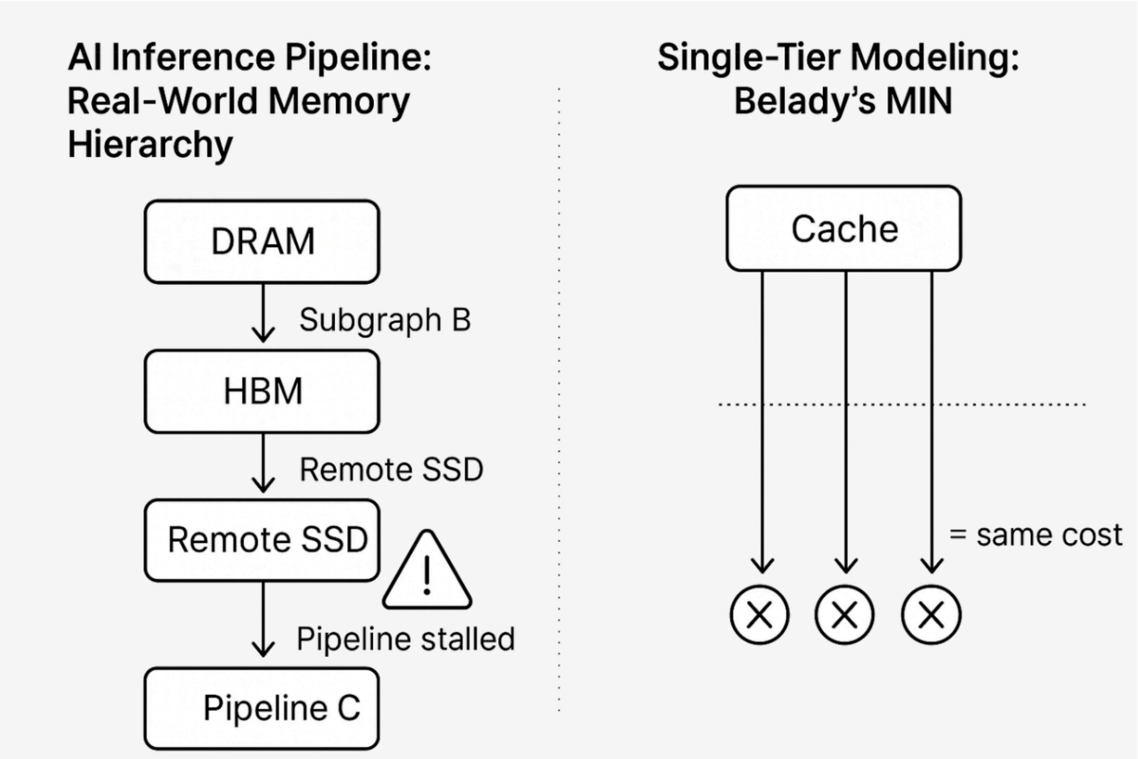
To deliver reliable performance and predictable ROI, storage and CDN platform leaders need new tools that reflect the true complexity of today’s architectures. However, the industry’s most familiar benchmarks, rooted in single-tier caching models, no longer map to the multi-layered systems your teams design and your customers depend on.

Belady’s MIN (or OPT), often cited as the gold standard for “optimal” caching, was built for an era where memory meant a single layer, access costs were uniform, and every request was best served from the cache. In theory, Belady’s approach is clean and even elegant. In practice, it assumes a world that no longer exists.

Here’s where single-tier models break down against the realities of modern storage and delivery architectures:

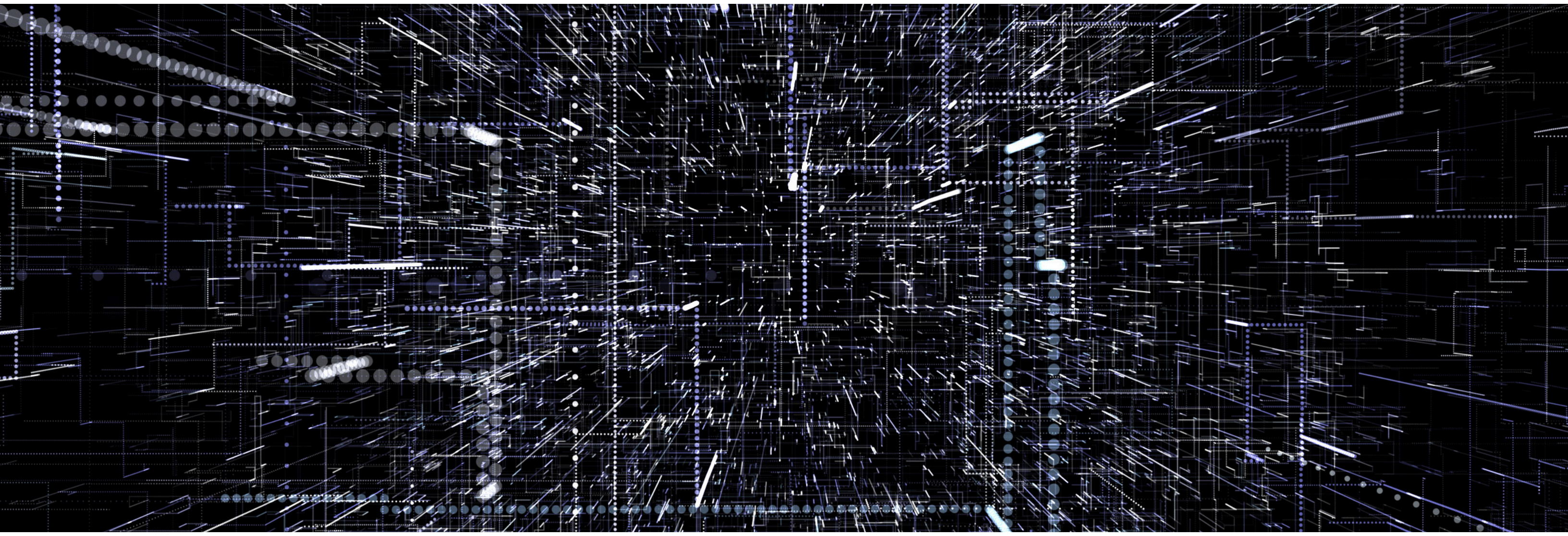
Belady’s Assumption	Real-World Risks
Uniform access latency	DRAM, NVM, SSD, and CXL tiers show 10x–1000x latency differences
Mandatory caching	Modern systems often bypass the cache entirely for certain workloads
Uniform miss penalty	Cost varies based on request size, data locality, read vs. write, and downstream effects
Flat hierarchy	Hierarchies now span 3–5 tiers with heterogeneous performance envelopes

In real-world AI inference pipelines, for example, a single model can stretch across DRAM, high-bandwidth memory (HBM), and remote SSD. If just one subgraph’s data lands in a cold, slow tier, it can stall the entire pipeline, even if other layers are fast. But single-tier models like Belady’s MIN treat all misses as equal, masking these bottlenecks and making it difficult for platform architects to diagnose or avoid systemic risks.


AI Inference Pipeline: Real-World Memory Hierarchy
DRAM
Subgraph B
HBM
Remote SSD
Remote SSD
Pipeline stalled
Pipeline C
Single-Tier Modeling: Belady's MIN
Cache
= same cost

Furthermore, the hit rate itself is increasingly misleading. In traces we've analyzed, two systems with identical hit rates can show >2x difference in end-to-end latency, due to tier-specific penalties and bypass logic.

Simply put, single-tier models and metrics don't explain or predict behavior in modern architectures. They misguide optimization and make system upgrades harder to evaluate objectively.



CHOPT: Redefining Optimal Data Placement

CHOPT (Choice-aware OPT) redefines what “optimal” means in real systems. It is the first general-purpose model to incorporate:

- **Cache bypass:** CHOPT models the possibility that the optimal action is to skip cache entirely
- **Multi-tier hierarchy support:** It works across N-level memory structures (e.g., DRAM→CXL→SSD)
- **Performance asymmetry:** It treats read and write penalties independently
- **Non-uniform costs:** Each memory tier has configurable latency, bandwidth, and capacity constraints

The Core Insight: Min-Cost Max-Flow

CHOPT reframes data placement as a network optimization problem. The workload trace is transformed into a temporal graph. Nodes represent request-time slots, and edges encode allowable transitions (e.g., cache retain, evict, bypass) and their associated costs (e.g., latency, memory usage).

The optimization goal becomes a min-cost max-flow computation over this graph. This transforms data placement from a greedy local decision into a globally optimal one, under full knowledge of future accesses.

CHOPT’s output is an upper bound on system performance for a given memory configuration, storage configuration, and workload. It does not predict what will happen; it tells you what could happen if placement were ideal.

Why This Matters

This offline-optimal baseline is invaluable. With CHOPT:

- Engineers can evaluate new memory configurations (e.g., adding CXL) without needing a production deployment
- You can quantify potential headroom, e.g., “how much better could this workload run if we had perfect placement?”
- Benchmarking is finally possible for new caching and tiering strategies, a previously intractable problem

Without the capabilities of CHOPT, system designers would have to rely on manual tuning, trial-and-error, and ad hoc performance metrics. Using CHOPT provides a mathematical foundation for architectural evaluation and workload optimization.

SHARDS: Scaling Offline Analysis to Production-Scale Workloads

Offline analysis with full future knowledge is computationally intensive. Computing CHOPT for a 1B+ request trace with 5 memory tiers would require terabytes of memory and thousands of CPU hours.

This is where SHARDS (Spatially Hashed Approximate Replay for Data Systems) comes in.

What SHARDS Does

SHARDS applies probabilistic spatial sampling to decompose large traces into representative shards. It exploits spatial locality and request distribution to extract workload slices that preserve key access patterns.

Rather than simulate the entire trace, CHOPT operates on the reduced, representative dataset. SHARDS then uses statistical reconstruction to extrapolate global metrics from shard-level outcomes.

Key properties:

- 10–100x reduction in compute time and memory usage
- <2% typical error margin across tested workloads
- Supports billion-request traces with commodity resources

How SHARDS and CHOPT Work Together

Stage	**Action**
Trace ingestion	***SHARDS partitions the workload and selects a statistically representative subset***
Simulation	***CHOPT computes the optimal placement on that subset***
Scale Up	***SHARDS scales up results from the statistical subset to generate system-level metrics***
Output	***Engineers receive accurate upper bounds on latency, hit ratio, and origin load***

This combination enables true offline-optimal modeling for real-world, production-sized workloads, something no other system design tool can claim.

CHOPT + SHARDS in Action: Benchmarking the Modern Memory Hierarchy

What does CHOPT reveal when applied to real systems? We evaluated CHOPT and Belady’s MIN across a broad collection of real-world workload traces from content delivery, block storage, and AI workloads. These workloads span multi-tier memory stacks, including DRAM, SSD, and CXL-backed storage.

Benchmark Setup

- **Workloads:** CDN, object store, and transformer inference (multi-GB models with real-time serving latency requirements)
- **Memory stacks:** DRAM→CXL→SSD; DRAM→NVM→HDD; DRAM→HBM→remote SSD
- **Traces:** Production and benchmark traces from real-world systems, including CDN, virtual machine block storage, and program memory workloads, evaluated using spatial sampling for efficient simulation.
- **Simulated using:** SHARDS-sampled CHOPT computation with under 2% MAPE (mean absolute percentage error)

Observations and Results

Using CHOPT reveals critical performance opportunities and inefficiencies that single-tier models like Belady’s MIN simply cannot detect.

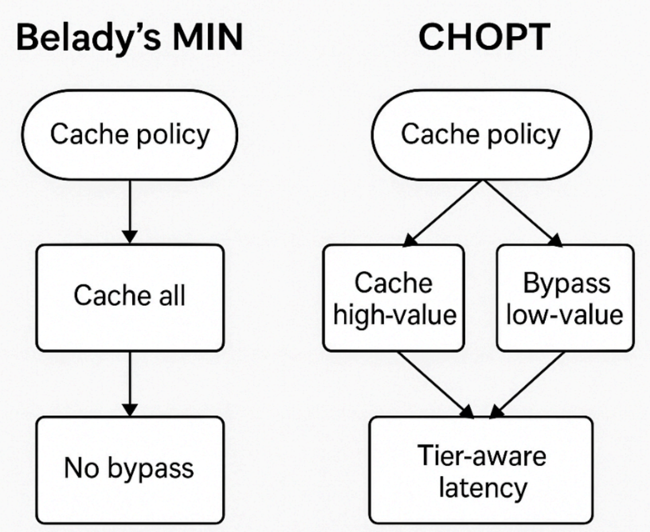
Simulation Results: CHOPT vs. Belady’s MIN — Latency Gains Across Real Systems

- **8.2%** faster on DRAM–NVM workloads (PARSEC benchmark)
- **44.8%** lower latency on VM block I/O
- **25.4%** lower latency on CDN cache

CHOPT consistently identified optimal cache bypass strategies that Belady MIN could not

- The performance gap was largest in AI serving workloads, where memory pressure was highest and latency asymmetries (read vs. write) were most pronounced
- In many cases, CHOPT suggested fewer objects in the cache yielded lower latency, simply because smarter placement avoided polluting high-speed tiers with low-value data
- The implication is clear: using Belady-based assumptions leads to over-provisioned caches, suboptimal write amplification, and missed opportunities for efficiency.

CHOPT identifies "what to cache" and "what not to." This distinction is fundamental to designing high-performance, cost-efficient systems.



From Research to Product: CHOPT and SHARDS in Magnition System Designer

CHOPT and SHARDS are not just academic breakthroughs, they are operationalized at the core of Magnition System Designer, turning advanced theory into a competitive advantage for modern storage and content delivery platforms.

With Magnition, platform teams and architects can:

Capability	Results
Model real-world workloads	Simulate at scale using proprietary or customer traces
Evaluate multi-tier memory configurations	Compare DRAM, CXL, SSD, NVM with true latency/cost impact
Simulate placement and bypass policies	Visualize object flow and identify optimal bypass points
Benchmark system headroom	Quantify the gap between the current state and the theoretical best
Guide hardware and architecture decisions	Run "what if" scenarios before making capital investments

Magnition uniquely answers the questions that matter for platform and business leaders:

- Is this caching layer helping or hurting my platform's margins and SLAs?
- What's the upper bound on performance and efficiency if I add new memory tiers?
- Can I justify tuning or policy changes with evidence, not guesswork?

No other solution delivers these answers, grounded in a mathematically sound, offline-optimal reference that matches the complexity of today's storage and CDN architectures.

Setting a New Standard for Platform Innovation

Memory and data placement are now critical levers for platform value and business impact. Today, memory and storage are fragmented across multiple, asymmetric tiers. In these environments, guessing is not an option. Single-tier memory hierarchy metrics and legacy benchmarks simply cannot keep up.

Magnition System Designer marks a turning point. By embedding CHOPT and SHARDS, Magnition delivers a “true north” for platform design, a clear, actionable upper bound that accounts for bypass, tier-specific costs, and real-world workload patterns. Platform teams move from trial-and-error to evidence-based architecture, from overspending to targeted investment, and from generic tuning to transparent, differentiated results for enterprise and global customers.

For leaders building the next generation of distributed storage, CDN, or AI platforms, Magnition System Designer delivers:

- Higher confidence in investment decisions and product launches
- Clear differentiation with transparent, customer-facing value
- Faster innovation and reduced engineering risk

You don’t have to settle for intuition or outdated models. With Magnition, you can model the real world, simulate what’s possible, and build systems that truly deliver at scale, for your customers, and for your bottom line.

Book a Demo Discussion Today!

Magnition

About Magnition

Magnition System Designer delivers AI-driven tools that help enterprises design, build, and optimize large-scale distributed systems with real-world accuracy and efficiency.

From concept to deployment, our platform ensures that your systems are efficient, reliable, and scalable. This helps you innovate faster, reduce costs, and improve time-to-market.

Find out more at https://magnition.io