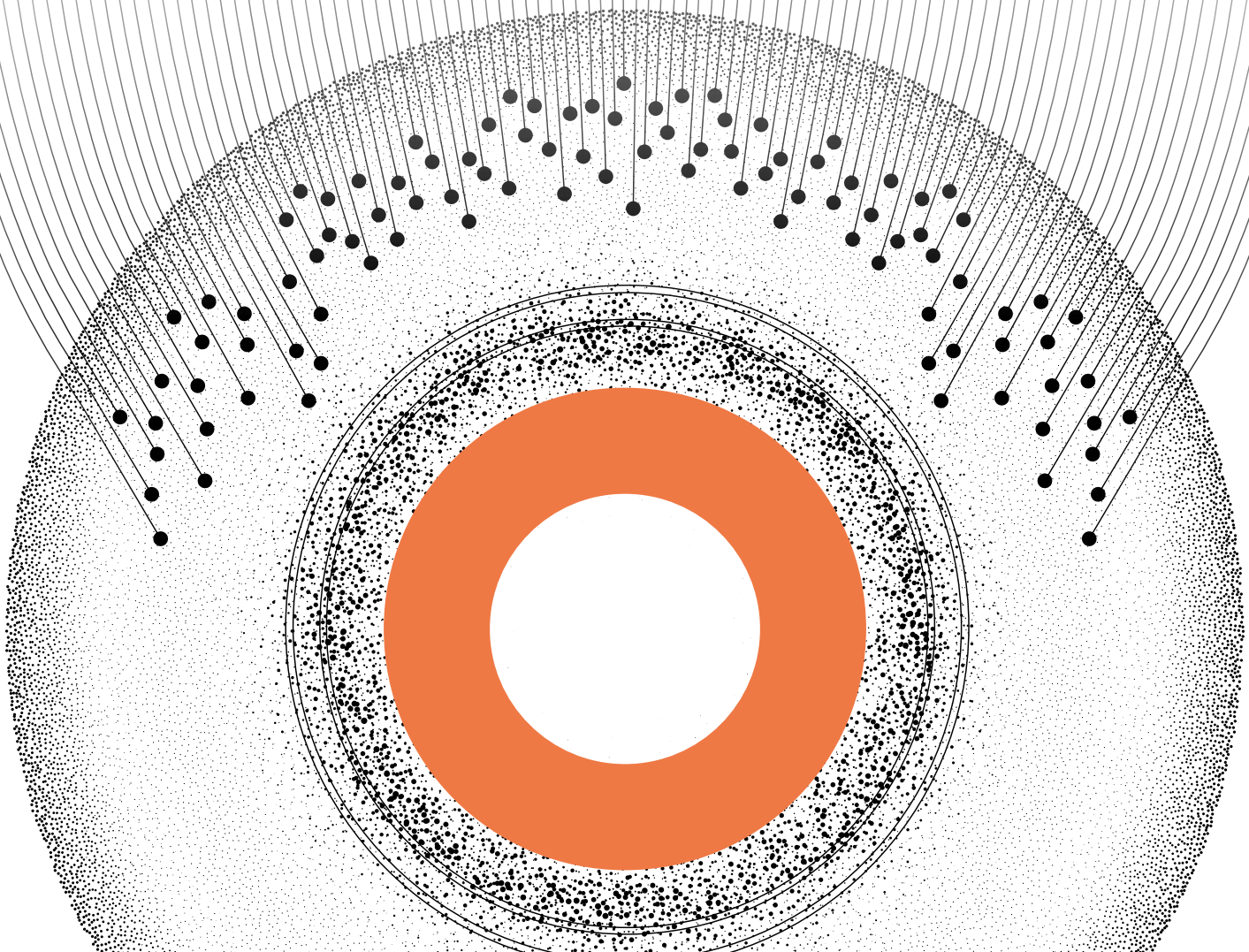
SCALITY | storage software

RELIABLE, SECURE, SUSTAINABLE

Redefining scale

Why next-gen AI and cloud data demand multidimensional scaling



Redefining scale

Why next-gen AI and cloud data demand multidimensional scaling

Before ChatGPT took the world by storm in late 2022, terms like "generative AI" and "LLM" were only heard in niche tech circles. But barely a few years later, AI is now dominating the mainstream business conversation, with the World Economic Forum estimating that 75% of companies will adopt AI by 2027. Organizations around the world are rushing to rethink their data storage strategies, finding that architectures that excelled in the pre-AI era are unable to meet the colossal data demands needed for AI model training, massive data lakes, and real-time inference.

But as impactful as it's already been, we must also acknowledge that AI is just the latest in a long line of technological shifts that have forced companies to reevaluate their storage paradigms. In the last 15 years alone, we've witnessed similar upheavals with the rise of cloud computing, the explosion of IoT data, the surge in ecommerce, the proliferation of streaming media, and more.

Each of these shifts has brought its own unique, and often conflicting, storage requirements — from massive capacity needs for media, to ultra-high transaction rates for IoT and real-time analytics — and often required organizations to build out costly new storage architectures to accommodate.

The tempo of these transformative developments also seems to be accelerating. The next disruptive shift could arrive at any moment, and just as most of us were blindsided by ChatGPT, we probably won't see it coming. This unpredictability makes it nearly impossible for organizations to confidently anticipate their future storage needs.

It's a great example of "what got you here won't get you there," except now you don't even know where "there" is — and worse, getting "there" may not be possible within your current storage framework.

Modern workloads are placing unprecedented demands on storage, stressing systems in new and novel dimensions beyond simple capacity.

Some examples include:

AI data lakes that require aggregating hundreds of petabytes of data from multiple external sources into a single repository that can be accessed across the data pipeline

ML applications that require supporting hundreds of millions of objects in a single S3 bucket

High-performance AI applications that demand either or both sides of the spectrum: sub-millisecond access to small objects for training purposes, and high-throughput access to large objects like image data for model learning

Cloud applications that require managing millions of S3 buckets, each mapped to a unique user

IoT applications that can send hundreds of thousands of transaction events per second, each needing to be stored as an individual small object in near real-time

Security monitoring applications that can generate — and require fast access to — petabytes of logging data

Media applications that require simultaneous access to thousands of terabyte-scale video objects

E-commerce applications that experience seasonal activity spikes requiring additional storage compute performance independent of capacity requirements

Cloud applications that require processing millions of user authentication requests per hour

The hidden cost of inflexible storage

In such a dynamic environment, flexible storage is like a life raft in a churning ocean of change.

Existing scale-out storage systems offered today typically excel at growing capacity, but often struggle to scale in areas like compute performance, metadata handling, or transaction capacity. Unfortunately, these are often the very areas where modern AI and cloud-centric workflows tax storage systems the hardest. Lack of scalability in these often-overlooked dimensions can lead to performance bottlenecks, operational headaches, and significant financial costs.

Most vendors of modern storage systems have to make design decisions that lead to tradeoffs. Should the system excel at capacity or performance (and which part of the performance spectrum)? These inflexibilities quickly become obstacles for achieving business goals with your data.

Operational and financial impacts

Performance bottlenecks: Systems may hit throughput or transaction ceilings long before capacity is maxed out, causing frustrating slowdowns.

Data migrations: When a storage system's scaling limits are reached, data must be moved to a costly new system, consuming valuable administrator time and making data temporarily unavailable to applications and users. Periodic forklift upgrades impose major disruptions.

Storage silos: Insufficiently flexible systems can necessitate multiple storage solutions for different workloads, adding complexity and inefficiency to data management.

Ongoing data management: Managing multiple silos requires continual balancing, moving, and copying data, adding hours to daily admin workload and causing frequent data unavailability.

Increased labor costs: Managing multiple storage solutions requires specialized admin skills and learning new systems, driving up costs.

Cost of downtime: Downtime due to data migrations, deployments, or system failures results in lost revenue and productivity.

High cost of proprietary hardware: Expensive, hardware-based all-flash storage solutions (e.g. Pure, VAST, IBM) add to both capital expenditures and operational complexity.

Beyond capacity:
Why true storage scalability requires multidimensional thinking

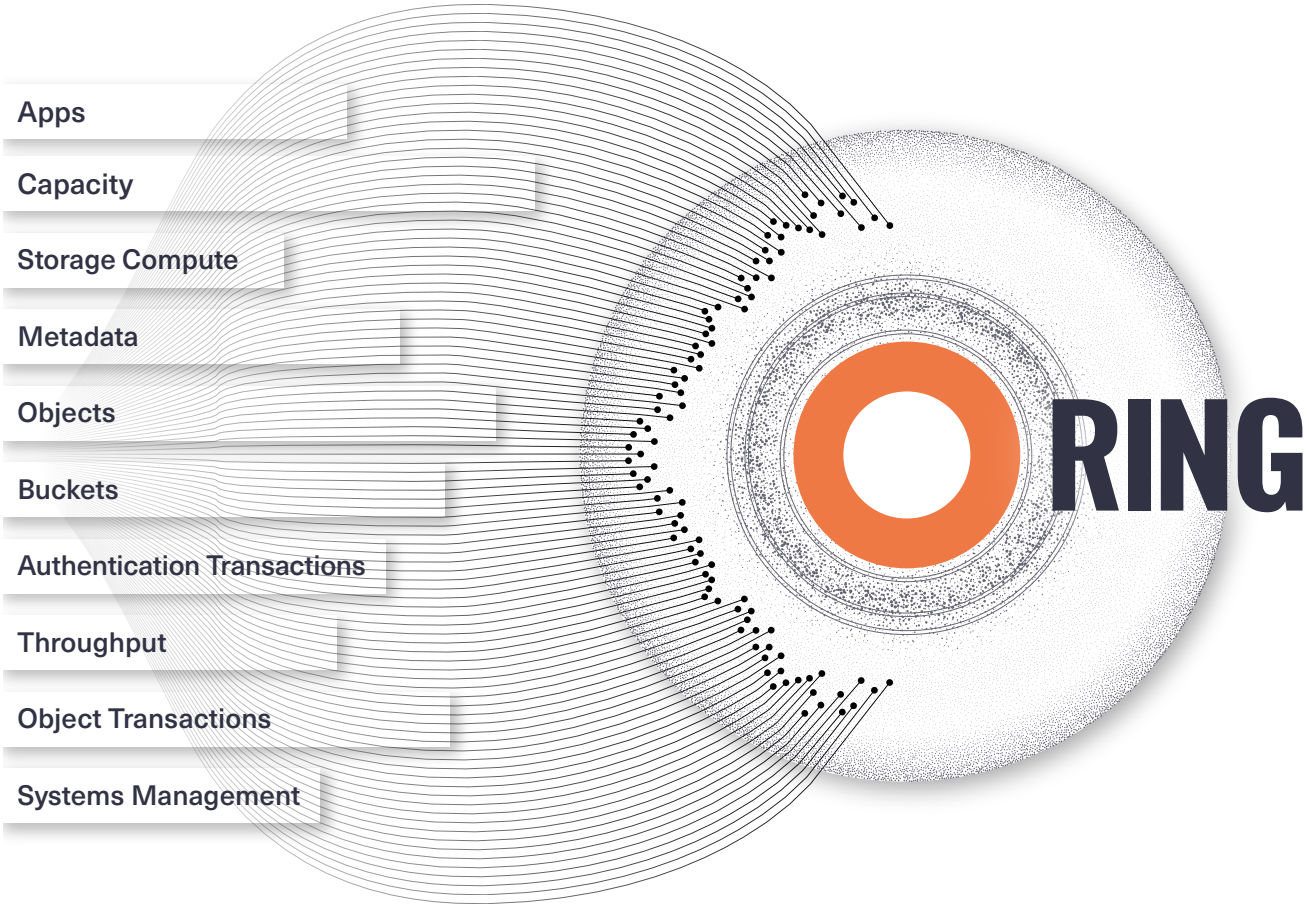
Simply put, scaling isn't just about storage capacity anymore.

To keep pace with emerging workloads and unknown future demands while avoiding the pitfalls of inflexible storage, a modern storage infrastructure must be able to seamlessly scale in any aspect that might be conceivably required.

Scality RING is that storage. Built on gold-standard S3 object storage architecture, it was designed from the ground up to be intrinsically scalable across ten different dimensions, allowing it to effortlessly adapt to any workload you throw at it — even ones you haven't thought up yet.

These multidimensional scaling capabilities make RING the most flexible storage solution on the market. RING XP, a new extreme-performance configuration of RING, provides additional flexibility by enabling microsecond-level access latencies and high transaction rates for small object workloads. This level of performance is previously unheard of for object storage systems.

THE TEN DIMENSIONS OF SCALE



Scaling apps

Solves: Need to address a range of requirements from single workloads to many

Today's storage infrastructures are tasked with supporting an ever-expanding number of applications, each with their own unique performance and capacity requirements. Nearly all of today's large enterprise, government, and service provider IT environments require data access from multiple applications. Private and public clouds are multi-tenant by definition, with multiple apps and use cases all sharing a common storage infrastructure. Business-critical AI and analytics data lakes require apps for everything from data ingestion and preprocessing to visualization and reporting.

Whereas traditional storage solutions struggle to support such large numbers of concurrent workloads, RING is built to handle it all. Whether it's one application or thousands, RING's architecture scales effortlessly and on-demand, delivering the power and flexibility needed to support the full spectrum of modern applications — all on a single system.

Scaling capacity

Solves: Need to store and protect increasing data volumes to grow with your business

Built on a unique distributed architecture inspired by hyper-scale cloud designs, RING offers near-limitless scalability, expanding to virtually unbounded numbers of storage nodes and millions of disk drives with ease. With modern storage server designs now reaching capacities of multiple petabytes per server, this means RING can effortlessly scale into the exabyte range.

For maximum flexibility, RING allows you to incrementally expand by adding disks, servers, racks, or even entire data centers — all while staying 100% online and available, with zero service disruption. A single RING can even stretch across multiple data centers in a synchronous, geo-stretched deployment, further increasing storage capacity and providing strong protection against cascading failures or multiple simultaneous hardware outages.

Scaling storage compute

Solves: Need to increase resources to deal with the demands of higher data volumes

Storage demands aren't always about adding more space. In many scenarios, workloads require increased compute performance without the need for additional storage capacity. For example, E-commerce platforms often face seasonal spikes in user activity, requiring more compute power to handle traffic without needing more disk space. Unfortunately, many traditional storage systems can't scale performance resources without the addition of new storage servers — an inefficient and costly limitation.

RING breaks this mold with its disaggregated architecture, which allows for independent scaling of compute and capacity. By decoupling services onto dedicated servers and resources, RING enables businesses to scale exactly what they need — whether it's increasing S3 services to handle more API requests or boosting metadata services for higher ops/second — without the added expense of unnecessary capacity.

Scaling metadata

Solves: Need to efficiently manage/access growing metadata volumes and accommodate data augmentation for AI applications

The backbone of every storage system is its metadata database — the system's master catalog of all stored data. With storage requirements now frequently reaching hundreds of petabytes and beyond, the ability to efficiently manage and quickly access huge volumes of metadata has never been more critical.

Most storage systems rely on fixed-size databases for metadata storage. These rigid designs can't be easily expanded as data demands grow, creating performance bottlenecks in critical metadata services that can require costly, time-consuming migrations to resolve. Other systems store metadata alongside the data itself, restricting the ability to run complex queries and introducing performance bottlenecks.

RING solves this problem by storing metadata in a scale-out distributed cluster on flash storage. By distributing metadata across multiple servers that form a "consistency cluster," metadata storage and processing power can be easily grown as needed by simply adding more disks, servers, or even additional clusters. This flexible approach ensures metadata services always keep pace as system requirements grow, delivering consistent metadata performance at any scale without disruption.

Scaling S3 objects

Solves: Need to simplify applications that manage large numbers of objects

Once an unthinkable feat for all but the largest cloud storage services, widespread adoption of the S3 API has made storing millions of objects in a single bucket not just feasible, but essential for modern workloads. The ability to manage fewer, larger buckets rather than balancing storage across multiple smaller buckets is a boon for simplicity, but it requires a storage system that can truly scale without compromising performance.

Many systems claim to support buckets with millions of objects, but in practice suffer from performance degradation in key operations like lookup performance (GET Object requests) and new object insert operations (PUT Object requests) as bucket sizes grow.

RING rises to this challenge with an architecture designed to scale effortlessly to billions of objects per bucket with zero hit to system performance. Whether it's listing, inserting, or retrieving objects, RING ensures that even as object counts continue their upward climb, critical system operations remain fast and efficient.

Scaling S3 buckets

Solves: Need to separate data by workload in large multi-tenant environments

In cloud infrastructures, storage is often managed by provisioning new S3 buckets for each tenant, user, or application. This approach enables granular, user-level control over bucket-specific policies like security and lifecycle management. But this flexibility comes at a cost. For common use cases like backup-as-a-service or storage-as-a-service, the number of buckets can quickly scale into the millions — and as the number of buckets grows, so too does the strain on the storage system's performance.

Traditional storage systems often impose hard limits on the number of buckets or suffer from significant performance degradation as bucket counts climb. With RING, there are no such constraints — its distributed architecture and sophisticated, all-flash metadata service are purpose-built to support scaling to millions of buckets and beyond, ensuring low-latency, high-performance bucket management regardless of how large the system grows.

Scaling S3 authentications per second

Solves: Need to ensure data privacy through strong security protocols

In both private and public cloud environments, user authentication and security policy requests can place a massive strain on storage systems. In private clouds, even a modest load of 1,000 users simultaneously uploading hundreds of documents can generate hundreds of thousands of S3 authentication requests per second. In public cloud scenarios, this demand can multiply exponentially to millions of requests per second, driven by hundreds of thousands — or even millions — of users. Each authentication request also involves evaluating user and bucket-level access policies to maintain data privacy, further amplifying the computational burden on the storage system.

RING's architecture uniquely addresses this challenge through its dedicated Vault service, which manages authentication and access control requests. Unlike fixed-capacity solutions, RING's Vault service is highly scalable and can be deployed on a disaggregated cluster of dedicated servers to handle authentication independently from other storage operations. This allows for seamless scaling of processing power dedicated to user authentication, ensuring that RING can handle even the most massive private and public cloud environments without sacrificing performance.

Scaling throughput

Solves: Need to quickly process large object data (e.g. media applications)

Modern workloads such as video streaming, high-resolution medical imaging, and big data analytics place incredible demands on a storage system's throughput — the ability to move large files and volumes of data quickly and efficiently. Organizations working with such huge files can quickly find themselves reaching the limits of their storage system's throughput, and lacking the flexibility to easily scale in this critical dimension as demands on the system grow.

In comparison, RING handles even the largest of objects with ease. It achieves scale-out throughput by increasing system resources for S3 API services and, when needed, by disaggregating these services onto dedicated servers. Additionally, RING's S3 Connectors can be scaled to any number as required, and their stateless nature allows load-balancing via standard HTTP/IP techniques. Together, these features allow RING to deliver exceptionally high throughput levels, reaching dozens of gigabytes per second and beyond.

Scaling objects per second

Solves: Need to quickly process small object data (e.g. AI training and fine-tuning)

Conversely, workloads that require managing vast quantities of small objects — such as IoT sensor data, log files, or microservices transactions — demand a storage system optimized for high transaction rates (objects per second). These workloads generate immense numbers of small data points that need to be processed and retrieved with minimal latency, something that traditional storage systems often struggle to accommodate at scale.

RING's architecture is built for these high-transaction environments, with optimizations focused on its S3 metadata service. By using fast flash storage for metadata and indexes, which are cached in memory, RING ensures near-instantaneous access to small objects, enabling extremely high transaction rates. In addition, RING further enhances performance by enabling configuration of data durability policies (erasure coding and replication) based on object size, optimizing overhead to match data requirements. This ensures that organizations can manage millions of small objects efficiently without compromising on speed or performance.

Meet RING XP

The world's most flexible object storage is now the fastest

For many storage systems, the most demanding workloads are those requiring ultra low-latency, high transaction rate access to millions (or even billions) of small data objects. RING XP was designed to tackle these extreme performance requirements with ease.

Through a streamlined, lightweight architecture that efficiently utilizes NVMe flash, RING XP delivers microsecond-level access latencies, making it the perfect solution for workloads that access small (KB or less) objects, such as AI model training, fine tuning, and inferencing, real-time analytics, and other latency-sensitive, transaction-heavy environments.

Because RING XP is implemented as a tier within a broader RING deployment, organizations finally have a cost-effective option for managing an entire AI data pipeline with a single unified storage environment. Instead of managing separate systems for low-latency, high-performance access during AI training and inference, and then moving data to capacity-optimized archival storage, RING's unified architecture streamlines data movement and management. This reduces complexity, lowers costs, and ensures seamless, consistent access to data at every stage of the AI workflow.

Scaling systems management

Solves: Need to efficiently manage large-scale systems or deployments with fewer people

As organizations scale, so do the demands on their infrastructure's ability to generate, capture, and manage critical operational data. System health and performance must be monitored through key performance indicators (KPIs), and infrastructure services must create ultra-granular logs of user activity, admin activity, and system events. In large-scale cloud deployments, this can quickly add up to staggering volumes of data — from gigabytes to terabytes every week, potentially reaching petabytes of system logs and metrics over time.

While these massive logs can easily swamp the capabilities of lesser storage systems, RING was designed with the logging and metrics needs of enterprise and cloud-scale deployments in mind — a capability it's repeatedly proven in some of the world's most demanding production deployments.

Multidimensional scaling future-proofs your storage

The rapid acceleration of AI, cloud computing, and other emerging technologies has made one thing clear — the demands placed on storage systems today are more complex than ever before, and traditional storage solutions are increasingly being stretched beyond their limits. Succeeding in this environment demands storage designed for the workloads of the future, not the past.

Scality RING rises to meet this challenge. Whether you need to expand capacity, increase throughput, manage millions of S3 objects, or optimize for high-transaction workloads, RING provides the flexibility and power you need. And with the addition of RING XP, organizations can push their performance even further, achieving microsecond-level access latencies for even the most transaction-intensive workloads.

RING's multidimensional scalability ensures that your storage infrastructure isn't just prepared for today's challenges — it's ready for the unknown demands of tomorrow. In an unpredictable landscape where the next big technological shift is always around the corner, Scality RING provides the adaptability, performance, and resilience needed to future-proof your data strategy.

SCALITY | **storage software**
RELIABLE, SECURE, SUSTAINABLE

Visit **www.scality.com** San Francisco • Paris • Washington, D.C • Tokyo • London