Key Takeaways

- As enterprises scale, cloud optimization can significantly reduce infrastructure costs over time by enhancing resource efficiency and reducing cloud waste. The following are areas to examine for optimization opportunities: storage, computing, networking and management.

- With cloud storage, it is important to set a lifecycle policy for governing how data should be stored. Auto-tiering solutions help automate this governance and ensure data is being stored efficiently and cost-effectively.

- Right-sizing computing resources to match workloads is a widely practiced method for optimizing cloud costs, and can be achieved through a combination of autoscaling, reserved instances, and spot instances. Emerging use cases like configuration tuning and machine learning (ML) model tuning maximize computing efficiency for further cloud cost savings.

- For network efficiency, Networking-as-a-Service (NaaS) or Virtual Wide Area Networks (VWAN) create dynamic, programmable network infrastructure; however, Software-Defined WAN (SD-WAN) solutions still hold their benefit of simplifying operational management and creating a real-time path for network traffic. To optimize network costs, egress charges should be monitoring using Network Performance Monitoring solutions.

- Many cloud management platforms (CMPs) incorporate cost optimization tools into their offerings for total control and visibility over the cloud environment. Current products can be divided into cost-focused and architecture-focused platforms, with a shift in the market towards architecture-focused offerings.

- Due to the cloud's inherent flexibility and rapid feature deployment, cloud optimization is an ongoing process. It requires the enterprise to critically analyze where and how it leverages the cloud, being mindful that workload performance directly impacts cloud costs. Along with tools, successful optimization requires modernized policies and procedures.

What is it

Cloud optimization is the process of eliminating cloud resource waste by selecting, provisioning, and right-sizing the resources spent on specific cloud features. Cloud optimization strategies include determining the most efficient way to allocate cloud resources among different use cases. The goal is usually to reduce waste while increasing performance in the cloud.

Why It Matters

Cloud infrastructure drives efficiencies by allowing enterprises to leverage infrastructure and platform resources needed at scale; however, pressures on cloud cost margins can start to outweigh the benefits as enterprises continue to grow. Increasing awareness of long-term cloud cost implications can provoke companies to consider alternatives, such as repatriation of workloads from cloud to on-premise. Shifting workloads from cloud will take significant effort and diverge from the on-demand benefits cloud services provide. Cloud optimization is an approach that can effectively reduce infrastructure costs over time while maintaining performance, flexibility and scalability.

Approaches

This report breaks cloud optimization into the following sections: storage, computing, networking and management.

Storage

As storage optimization is not a "one-size fits all" approach, it is important for enterprises to evaluate the required performance and retention of the stored data and develop storage policies that will meet their needs.

- Performance: Consider where the data will be accessed from and find a storage location that will provide a fast, cost effect pathway. Storage in multi-regional locations create higher availability, but could also potentially increase egress charges.
- Retention: Set a lifecycle policy for each type of stored data based on length of storage (short-term or long-term), frequency of retention and criticality.

Tiered storage leverages an enterprise's storage policy and places the data in the appropriate tiers based on business value. This ensures hot tier data (frequently accessed) is available for quick retrieval, whereas cold data (less accessed) is stored in a more cost-effective manner. **Auto-tiering solutions** automatically manage the tiered storage by determining which datasets are most accessed by the applications and can place frequently utilized data in faster storage retrieval tiers and less used data in lower-cost storage tiers. As data usage changes, the tiers will automatically adjust to meet the business needs.

Computing

The key to optimizing cloud computing is right-sizing, a process of matching computational resources to workload performance and capacity requirements at the lowest possible cost. This is achieved either by changing the CPU capacity of an instance or modifying the number of instances running a workload. Right-sizing means upgrading, downgrading, or terminating assets based on actual usage. While downgrading and terminating assets reduces computing costs, upgrading assets ensures the enterprise is ready to meet surges in demand, minimizing downtime and latency. Right-sizing solutions leverage a combination of autoscaling, reserved instances and spot instances to optimize cloud computing spend.

Autoscaling – Dynamically right-sizes instances based on workload. Types of autoscaling are:
- Reactive (default type): Scales in response to real-time changes in traffic; effective solution, except when facing rapid changes in demand
- Scheduled: Anticipates changes in traffic at scheduled times in addition to scaling reactively; best used when workloads regularly change at certain times of day
- Predictive: Uses predictive analytics to anticipate when to scale; best for detecting large, imminent spikes in advance, coping with regional outages, and dealing with variable traffic patterns throughout the day

Reserved instances – Buyer commits to paying a lower price for a set amount of computing resources over long periods of time; best for steady, predictable workloads

Spot instances – Cheaper than on-demand instances but may be terminated at any time by cloud provider; best for data analysis or batch workloads that don't require continuous availability

A specialized group of solutions are taking the idea of right-sizing one step further by dynamically adjusting the configuration settings of instances to match the unique resource needs of different workloads. **Configuration tuning** improves upon the capabilities of autoscaling solutions because it addresses factors other than CPU power that effect runtime efficiency of workloads.

As machine learning (ML) continues to be deployed across all industries, **ML model tuning** is an emerging avenue for cloud computing optimization. Most firms opt to run ML models on GPUs in the cloud, since it's costly to manage GPUs in-house. However, ML in the cloud becomes a cost issue if models are slow and inefficient. Solutions like OctoML solve this issue by automating the tuning of ML models to run efficiently on target hardware (CPUs, GPUs, accelerators, etc.).

Networking

Network optimization plays an important role in managing the increasingly large amount of data and network traffic flowing through enterprise infrastructure. The ultimate goal is to balance costs, performance, availability, and security.

In order to optimize the network as needed, it is important to implement a programmable infrastructure that creates a dynamic, flexible network. **Networking-as-a-Service (NaaS) and Virtual Wide Area Networks (WAN)** are viable alternatives to VPN and SD WAN, which offer on-demand provisioning and management, simplified operations and API integration to network automation. Certain solutions, such as Alkira and Teridion, actively monitor network performance and can add more capacity as utilization rises or use optimal paths as conditions change between points in the network. With the consumption-based pricing model, customers can scale with demand while optimizing network costs and performance. These WANS also use providers' resources to connect the enterprise's cloud services and locations, alleviating the need on the enterprise to incur high speed internet costs for network connection. While rapid evolution in **Software-defined WAN (SD WAN)** has slowed, it still holds a market for enterprises looking to interconnect the network and streamline operational management. Initially the perceived benefit was the real-time, contextual path selection of network traffic; however, the real benefit became the simpler operational management, which is normally complex to achieve. Leveraging SD WAN extends the benefits such as traffic prioritization, security, and automatic provisioning and deployment to cloud access.

The shift to hybrid and multi-cloud infrastructure can increase network traffic flow, putting a strain on the network costs. To obtain network traffic cost optimization, egress charges should be monitored, since they can account for a significant portion of overall network costs. Egress charges are often priced in gigabytes for data transferred out of the cloud and can be difficult to spot, as they are scattered across cloud bills. **Network performance monitoring tools** can help achieve this visibility by allowing users to map and filter by egress charges.

Management
In many cases, cost optimization tools are bundled with features in visibility, automation and orchestration, governance, and security as full-service cloud management platforms (CMPs). When choosing a CMP, the enterprise should look for platforms that can dive deep into their subscriptions and provide intelligent recommendations for reducing app footprint, optimizing storage usage and its cost, and getting rid of unused instances. They must also consider the make-up of their cloud environment, seeking out platforms that support multi-cloud or hybrid cloud configurations. A comprehensive CMP will provide relevant tools for the development, operations, and financial stakeholders in an organization.

Today's third-party cloud management solutions fall into two general categories: cost-focused and architecture-focused. While **cost-focused platforms** take a reactive approach to cutting down cloud fees, **architecture-focused platforms** seek to proactively reduce costs through well-designed infrastructure. In recent years, there has been a "shift left" in the CMP market towards these architecture-focused platforms, as solutions that focus solely on cutting costs face significant competition from native cost management tools.

Trace3 Innovation's Point of View
Among the major cloud providers, there is no shortage of native tools built to address cost optimization. These tools tend to be designed for general purpose, single cloud use, leaving several niches in the market for third party solutions to fill. Emerging solutions in this space provide continuous optimization and automated tuning to build upon native optimization capabilities. Development of tools that can span hybrid or multi-cloud environments is another area where third party solutions have a competitive advantage. As the market continues to evolve, traditional on-premise technologies are also evolving to support hybrid use cases. The enterprise's toolset should align with their cloud strategy and may require a rationalization exercise to determine product fit to business needs, the skill sets of teams, and optimization goals.

This report is focused on the optimization of the following cloud disciplines: storage, computing, networking and management. While the tools and techniques highlighted in this piece will lead to savings on the monthly cloud bill, no tool is an effective substitute for robust architecture. To achieve true optimization, the enterprise must take a continuous position that focuses on use cases, workloads and architectures that are best suited for cloud. Cloud-native architectures such as microservices and serverless provide for the best optimization strategies, along with resiliency and portability. Adopting a clear cloud governance structure will help align processes and procedures to the optimization strategy.

Cloud providers are continuously lowering the barrier to entry, providing new features and giving the enterprise endless options in the marketplace. While extremely convenient, the costs can add up quickly. With a proactive and continual focus on cloud resources, every organization can truly adopt a cloud operating model that is optimized for both performance and cost.

Solutions
All vendors provided are examples and are not meant to be an exhaustive list. Emerging technologies are subject to significant changes in market share and relative capability.

Storage
Automatic Tiering

Our obsession for optimization fuels every decision we make, from the source of our data to our automated approach. Zesty Disk™ automatically adjusts storage volumes in real-time based on the application needs, achieving optimal disk utilization, and a dramatic reduction in EBS spend.

DataCore Software is a provider of operating system and infrastructure software for the emerging Storage Area Networking (SAN) marketplace. The company distributes its products through Systems Integrators, technical Value Added Resellers, and Original Equipment Manufacturers worldwide

Cloudian is a provider in scale-out object storage. The company's flagship product, Cloudian HyperStore, enables service providers and enterprises to build hybrid cloud storage solutions.

GigaSpaces Technologies provides software middleware for deployment, management and scaling of mission-critical applications on cloud environments through two main product lines, XAP In-Memory Computing and Cloudify. Hundreds of Tier-1 organizations worldwide are leveraging GigaSpaces' technology to enhance IT efficiency and performance, from top financial firms, e-commerce companies, online gaming providers, healthcare organizations and telecom carriers.

Netapp's ONTAP automates storage tiering feature lets you cut costs for data that isn't immediately in use. Instead of keeping snapshots, disaster recovery environments, and infrequently-used active data on performant block storage, data tiering allows users to store such data in a capacity tier on inexpensive block storage on Amazon S3 or Azure Blob.

Computing
Right-sizing

CAST AI brings resource and cost optimization to Kubernetes using an AI-driven optimization engine that applies cluster changes based on real-time workload conditions. It selects the most cost-effective instance types and bin-packs pods for maximum utilization. The platform uses spot instances that offer 70-90% discounts and are automatically selected for any stateless workloads. Compatible across AWS, GCP, and Azure.

Zesty Commitment Manager automatically adjusts EC2 commitments in real-time based on the environment's capacity requirements, achieving optimal commitment utilization and dramatic reduction in EC2 spend. Zesty automatically reacts in real-time by purchasing Reserved Instances and/or Savings Plans upon increase in compute usage and by selling Reserved Instances and/or Savings Plans upon decrease in compute usage.

Kubecost allows you to see allocated spend across all native Kubernetes concepts, so you can provide your teams with transparent, accurate cost data reconciled with your actual cloud bill. Join your cluster costs with spend from AWS, GCP, and Azure for complete cost visibility. Kubecost automatically generates insights you can use to save 30-50% or more on your cloud spend, without exposing your private information.

Harness Cloud Cost Management is a dedicated product focusing on cloud cost visibility. Answering the question "how much is my application costing" for engineers is difficult, especially with the multi-tenant or multi-app nature of modern cloud native infrastructure. Harness Cloud Cost Management unlocks cloud cost visibility to solve this problem.

Replex's complete insights into the efficiencies of tech, business units, services and functions provide real-time cost analyses and reporting, customized for each organization. Replex helps you make proactive and informed rightsizing decisions based on near real-time data and frees up IT budget through reallocation, so you can focus on innovation and growth.

CloudZero isn't just a better way to slice and dice your AWS bill. By enriching your cost with services metadata, telemetry, and more, the CloudZero platform lets you see your costs from any angle. Engineering can self-serve and explore the cost of their architecture and apps, while finance can measure the ROI of your technical investments – all without any tagging.

YotaScale ingests data from various sources such as AWS CloudWatch, Prometheus and Chronosphere, and uses ML to aggregate this data into spending forecasts, rightsizing recommendations and anomaly detection. YotaScale also provides automated tagging management to ensure that costs are accurately attributed at the team and app level.

Virtana Optimize's powerful analysis and recommendation tools enable you to confidently cost-optimize your cloud resources. Automatically optimize Amazon EC2 and Azure VM instances with our rightsizing recommendations. Tune sizing based on your organization's risk tolerance with "What-if" analysis that includes CPU, memory, I/O, and ingress and egress charges. Find the ideal resource settings before purchasing discount commitments.

Densify is the only technology that leverages patented, predictive machine learning-powered analytics to perform advanced modeling of workload patterns, and provide precise optimization directives —not naive suggestions. Machine learning precisely and automatically right-sizes and right-types your public cloud IaaS, ensuring app performance at the lowest possible cost per instance.

Spot's machine learning and automation scale to exactly meet application needs using the most efficient mix of instances and pricing models, eliminating overprovisioning and waste. Our software works with leading cloud platforms, services, and tools so that you can simplify and automate your cloud infrastructure wherever your workloads and applications run and however you run them.

Cloudability is the simple way to track, control and share the costs for all of your IaaS, PaaS and SaaS. Track EC2s, S3s and Reserved Instances with easy-to-understand reports that keep you tied to your current spend as well as your predicted spend for the whole month. Customized budgets alerts let you know that your costs are going up before they spike, and IAM support means you give Cloudability read-only access to your billing and usage data.

Nutanix Beam unifies private and public cloud costs into a single console along with automated detection of cost anomalies. Beam provides cost optimization recommendations by identifying unused or underused cloud resources. Beam also optimizes purchase plans based on your consumption needs and helps you implement automation policies.

Configuration Tuning

With Opsani, it is possible to tweak and perfect your runtime parameters that are too complex to touch. Opsani tunes your runtime infrastructure precisely to the workload and your business performance goals. By using precisely what you need, you get the performance that you need and quickly see your costs slashed.

Granulate is a real-time continuous optimization software that increases throughput and reduces latency, server dependency and costs with no code changes required. Granulate's agent automatically learns your application's specific resource usage patterns and data flow. The intelligent agent tailors OS-level decisions regarding CPU, locks, caches and memory accesses to improve an infrastructure's application specific performance.

StormForge analyzes, optimizes, and refines your cloud-native application configurations so you can release with confidence and meet your goals. Use Machine Learning to analyze how your application will perform in real-world scenarios and automatically implement recommendations for optimization. Gain actionable insights to release your cloud-native applications without the risk of downtime, poor performance or inflated cloud costs.

Concertio provides AI-powered performance optimization tools that boost performance by tailoring the many system settings (in processors, firmware, OSs and applications) to the running workloads. By continuously monitoring and dynamically optimizing system settings, Concertio Optimizer tools effectively transform general-purpose systems into high-performant, tailor-made systems.

Akamas exploits patented reinforcement learning techniques to automatically optimize hundreds of interdependent service configurations, delivering full-stack optimization at business speed. Akamas unique AI solves otherwise intractable real-world optimization problems, requiring the exploration of millions of system configurations. The Akamas reinforcement learning engine intelligently explores the configuration space and converges to the optimal configuration within hours.

Pepperdata Capacity Optimizer automatically optimizes your cluster resources, recapturing wasted capacity so you can run more applications and get the most out of your infrastructure investment. On a typical cluster, Capacity Optimizer uses machine learning (ML) to analyze the resource usage of each node in real time. The result: CPU, memory, and I/O resources are automatically optimized to increase utilization.

Networking

NaaS/VWAN

The Mode Group is a developer of autonomously controlled networks. Mode now provides Mode Core, a software-defined core network (SD-CORE) that gives enterprises affordable private network reliability and quality of service, which can be used in combination with SD-WAN and last-mile internet. SD-WAN providers and enterprises use Mode Core side-by-side with MPLS and internet as an affordable, elastic, reliable private network with SLA guarantees for UC, SaaS, remote access, and site-to-site.

Teridion delivers Teridion for Enterprise, a public cloud-based WAN service that delivers carrier-grade, SLA-backed performance powered by a cloud-native approach to routing that draw on deep learning, which brings hierarchical and centralized routing to enterprise networking to improve WAN, application and SaaS performance.

Alkira is a multi-cloud network platform that aggregates public, private, and remote cloud networks.

Cato Networks provides organizations with a software-defined and cloud-based secure enterprise network. Cato delivers a single networking and security platform that securely connects all enterprise locations, people and data. The Cato Cloud reduces MPLS connectivity costs, eliminates branch appliances, provides direct, secure internet access everywhere, and integrates mobile users and cloud infrastructures to the enterprise network.

Our Network-as-a-Service (NaaS) platform makes it simple for you to interconnect yourself or your customers globally with an on-demand network orchestration capability.
With Infiny by Epsilon, you can spin up connections to leading cloud providers, internet exchanges, data centre operators and procure global inbound numbers.

SD WAN

GTT Communications (NYSE:GTT) is a cloud networking provider to multinational clients that operates a global Tier 1 IP network with an interconnected Ethernet service platform. The company provides highly reliable, scalable, and secure cloud networking services, as well as Internet, managed services, and voice services.

Versa Networks solutions enable service providers and large enterprises to transform the WAN and branch networks to achieve unprecedented business advantages. Versa's VNF software provides unmatched agility, cost savings, and flexibility vs. traditional network hardware.

Palo Alto Networks (NYSE: PANW) is a global cybersecurity company. It seeks to help address security challenges using developments in artificial intelligence, analytics, automation, and orchestration. By delivering an integrated platform and empowering a growing ecosystem of partners, the company seeks to protect organizations across clouds, networks, and mobile devices.

Cisco provides networking solutions. The company is involved in designing, manufacturing, and selling Internet protocol (IP)-based networking products to the communication and information technology (IT) industry. Its major products include switches, routers, network access, IP telephony, optical networking, security, storage area networking, home networking, and wireless technology. The company provides these products and services for transporting data, voice, and video traffic across the internet, intranets, and extranets.

Network Performance Monitoring

AppNeta offers IT Performance Management with a cloud-delivered service for integrated, end-to-end visibility across networks and applications. AppNeta delivers a SaaS portfolio of end-user experience monitoring services with cross application performance visibility and network performance insight.

Zenoss works with organizations to ensure their IT services and applications are always on. Specializing in software-defined IT operations, Zenoss uniquely collects all types of machine data, building real-time IT service models that train machine learning algorithms to predict and eliminate outages in hybrid IT environments, dramatically reducing downtime and IT spend.

LiveAction provides software designed to simplify network management, including voice, video and application troubleshooting in real time. Initially developed to aid the U.S. Department of Defense in the operation of its networks, the solution features a visual display, real-time big data analytics for decision-making, and integration with routers and switches for network control.

Kentik is a cloud-based network visibility and analytics solution that provides a panoramic view of any network. Kentik processes tens of billions of data records per day, equipping service providers, Web enterprises and network operations teams with actionable, usable insights that enable making quick, cost-effective decisions.

Datadog unifies the data from servers, databases, applications, tools and services to present a unified view of the infrastructure. These capabilities are provided on a SaaS-based monitoring and data analytics platform that enables multiple teams working collaboratively on the infrastructure to avoid downtime, resolve performance problems and ensure that development and deployment cycles finish on time.

Management
Architecture-Focused Platforms

Env0 provides automated, collaborative workflow management for cloud deployments on Terraform, Terragrunt and custom flows. Env0 enables teams to jointly govern cloud deployments with self-service capabilities and provides visibility into GitOps workflows. Implement any policy you desire with the freedom of Open Policy Agent guardrails. Env0 also helps you prevent drifts with centralized remote-runs processes.

The Morpheus orchestration platform is infrastructure agnostic and provides full control over bare metal, VM and container-based systems to all stakeholders. Their platform provides complete application lifecycle management, including provisioning, backup, recovery, logging, monitoring, reporting, and access control. With Morpheus, IT organizations speed up application delivery, simplify workload migration, eliminate cloud and vendor lock, and provide greater operational efficiency

Pulumi is a modern infrastructure as code platform that allows you to use familiar programming languages and tools to build, deploy, and manage cloud infrastructure. Pulumi enables teams to deliver infrastructure and applications together and faster, increasing agility, reducing risks, and speeding up innovation. Pulumi's secure cloud backend (the Pulumi Service) provides built-in state management and encrypts configuration secrets in transit and at rest.

Scalr is a remote operations backend that harnesses the power of Terraform to aid organizations in infrastructure management. Scalr extends the provisioning capabilities of Terraform to the entire application lifecycle, providing a developer-focused workflow in Git. Scalr also provides management for roles and teams, a policy engine, cost estimation and policy dry runs.

Terraform by HashiCorp is an infrastructure as code (IaC) tool that allows you to build, change, and version infrastructure safely and efficiently. This includes low-level components such as compute instances, storage, and networking, as well as high-level components such as DNS entries, SaaS features, etc. Terraform allows for provisioning across multicloud and hybrid cloud environments and is one of the most widely used infrastructure automation tools available.

Ansible by Red Hat is a tool that automates cloud provisioning, configuration management, application deployment, intra-service orchestration, and many other IT needs. It uses no agents and no additional custom security infrastructure, so it's easy to deploy - and most importantly, it uses readable YAML language to describe automation jobs. Ansible's library of cloud support modules make it easy to provision instances, networks, and complete cloud infrastructure wherever you need.

Cost-Focused Platforms

Cloudify is built on opensource architecture with a focus on flexibility and scale. Model your services once, and then automate entire lifecycles from provisioning to configuration. Using Cloudify, developers and IT departments can finally integrate with any external service and natively with all major cloud infrastructures. Our technology simplifies orchestration and ensures all services are managed from one entry point.

CloudSphere provides governance across infrastructure optimization, standards compliance and security posture in your multicloud. We are the only cloud governance platform that monitors real time contextual data points in the cloud, providing guardrails with actionable governance alerting and reporting. Operate with confidence as auto-tagging supplements continuously update resource tags in the dynamic cloud environment.

Flexera One provides a rich set of capabilities across cloud and on-premise resources including: visibility, provisioning, orchestration and automation, governance, and cost optimization. Flexera One offers a unique approach to cloud cost optimization that enables cloud governance teams to work collaboratively with business units and cloud resource owners to optimize spend.

Centilytics gives you complete and granular visibility into your ever-changing cost and usage of the cloud. Knowing how to control your cloud costs becomes a breeze when the right data comes together in a single dashboard. Centilytics builds a meaningful picture of all your cloud expenses and breaks down the complex cloud billing into actionable insights.

CloudCheckr CMx is next-generation cloud management featuring a new experience for delivering total visibility into infrastructures and organizational management—by individuals, teams, and roles. With the ability to unify disparate data sources across an enterprise, businesses achieve immediate and actionable insights on how to run and scale public cloud infrastructure.

Logic Monitor allows you to visualize your entire ecosystem – on-prem, cloud, and microservices – within one platform. Get a breakdown of where you're spending the most and least amount of money to optimize costs moving forward. See 'at-a-glance' where you can cut costs and optimize spend with detailed ROI analysis. Automated alerts flag spending thresholds and reserved instance expirations.

With Snow Commander (formerly Embotics), you can immediately start providing governance and control over the consumption of your data center and cloud assets. Snow Commander simplifies cost management, providing you with visibility across cloud platforms, applications and business units. This new perspective enables you to reduce costs and maximize performance by allocating the correct resources for each workload across your organization.

RackWare provides an intelligent highly automated Hybrid Cloud Management Platform that extends across physical and virtual environments. Supporting a suite of services including Disaster Recovery and Backup, Hybrid Cloud Management, and Cloud Migration and Replication, computing resources—physical, virtual, and cloud machines—can be easily and automatically scaled up or down as demand fluctuates.