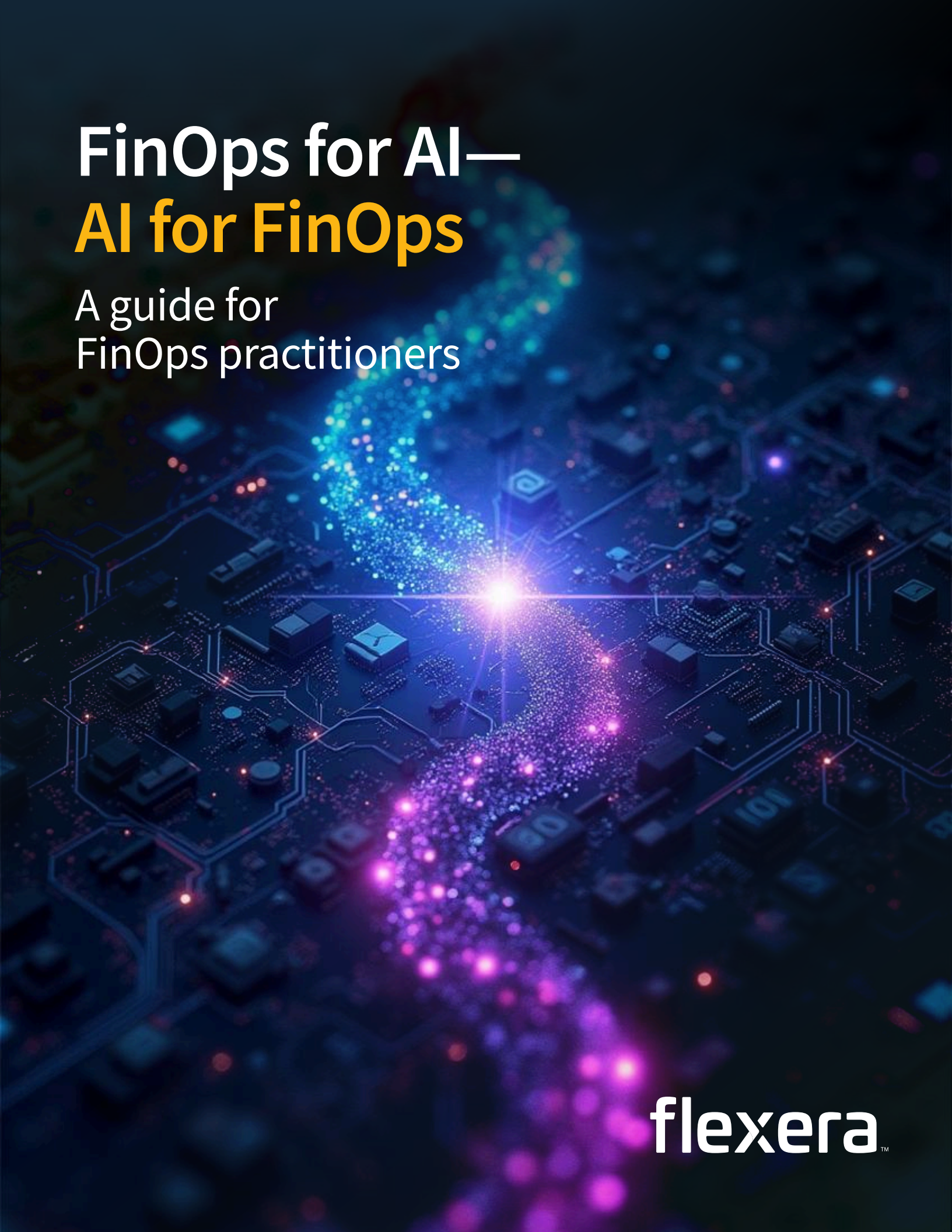
FinOps for AI—
AI for FinOps

A guide for
FinOps practitioners



flexera™

Table of contents

Preface

This e-book, "FinOps for AI—AI for FinOps" is a comprehensive resource for FinOps practitioners, CloudOps teams, CIOs, CTOs, decision-makers and other professionals who want to align innovation and efficiency as AI reshapes the cloud landscape.

This resource takes an in-depth look at:

- The relationship between FinOps and AI and how they impact each other
- The challenges AI brings to cloud cost management
- How your FinOps team can apply AI to their practice

Executive summary

In nearly every industry, AI is revolutionizing the way we work. For most of us, automation and AI—and, more recently, generative AI and agentic AI—are becoming integral to our daily workflows. They offer a powerful assist—accelerating processes, improving decision-making, boosting productivity and driving the creation of high-quality deliverables.

As much as AI is positively transforming business operations, it's also increasing costs. As organizations seek an edge over their competition, they're allowing employees broad access to AI resources and tools in pursuit of innovation and efficiency, so the rapid development of AI models and adoption of AI services have run relatively unchecked. From AI infrastructure to GenAI and agentic tools, cloud spend and software costs are rising fast, often without visibility or governance.

Amid shifting economic conditions, the rapid rise in AI-related costs is already forcing organizations to pause and re-evaluate how much they're getting in return for their AI investments. Forward-looking **businesses are asking their FinOps teams to manage these growing costs** and ensure that every AI dollar contributes to better business outcomes.

However, **AI is also empowering FinOps practitioners** by improving the tools they use; AI increases the accuracy of forecasting models, drives more intelligent anomaly detection and even serves as an intelligent assistant that answers FinOps questions without requiring practitioners to write and run queries against a database.

What FinOps does for AI:

- Enables forecasting and allocation for cloud-based AI resource costs (e.g., GPUs, TPUs)
- Promotes alignment with intersecting disciplines such as SaaS management to keep unpredictable usage of AI applications (e.g., Copilot, ChatGPT) in check

What AI does for FinOps:

- Improves the accuracy of FinOps tooling by applying machine learning to activities such as forecasting, anomaly detection and rightsizing
- Democratizes cost information by enabling natural language interaction with FinOps data

Part one: FinOps for AI

For FinOps practitioners, AI represents both an opportunity and a challenge. On one hand, FinOps teams can leverage AI to help complete work faster and more easily. On the other hand, FinOps practitioners are tasked with helping their organizations understand and control what some are calling the "unmanageable" cost of this powerful technology.

FinOps for AI	AI for FinOps
What is FinOps for AI?	**What is AI for FinOps?**
You can think of FinOps for AI as just ... FinOps. But the resources and services you are optimizing are AI resources (e.g., GPUs) and services (e.g., Microsoft Copilot) instead of more traditional resources/services such as CPUs and database services.	AI for FinOps is using AI to help with the practice of FinOps. Examples included are asking an AI assistant in natural language how much an application costs to run or using AI-powered anomaly detection in your FinOps tooling.

Table 1: *What are FinOps for AI and AI for FinOps?*

The impact of AI on cloud costs

The surge in AI adoption is a primary catalyst for escalating cloud costs. The core challenge lies in the compute-intensive nature of AI, particularly the reliance on expensive graphics processing units (GPUs) or tensor processing units (TPUs) for model training and inference. Unlike traditional workloads, AI's unpredictable resource demands and the lack of mature cost management tools specific to AI make accurate forecasting and optimization difficult. FinOps practitioners are increasingly being tasked with managing these accelerating AI costs, highlighting the urgent need for specialized strategies.

GenAI and AI drive cloud expenses 30% higher—and 72% of IT and finance professionals say the spending has become unmanageable.*

*Source: Tangoe Report - State of Cloud: The Critical Role of Third-Party FinOps in Cloud Spending Control

The increased use of AI primarily drives up cloud spending in several ways:

Specialized hardware

AI workloads, especially for training large language models (LLMs) and complex machine learning models, require high-performance GPUs or TPUs. These instances are significantly more expensive than standard CPU instances.

Data ingestion, storage and egress

AI models are data hungry. The processes of collecting, cleaning, labeling and transforming massive datasets for AI training (ETL pipelines) incur substantial compute and storage costs. Furthermore, moving large volumes of data between cloud regions or out of the cloud (egress) can quickly escalate expenses.

Model training and operating

Training a single LLM like GPT-5 can consume thousands of compute hours, representing a massive upfront compute investment. Once trained, deploying and running AI models for real-time predictions or responses incurs ongoing costs, especially as user adoption scales.

Unpredictable workloads

AI workloads are often dynamic and bursty, making it challenging to leverage cost-saving mechanisms such as reserved instances (RIs) or savings plans (SPs), which are typically designed for predictable, long-term compute usage. This unpredictability leads to higher on-demand consumption.

Licensing and software

Beyond infrastructure, the use of specialized AI/ML platforms, tools, and third-party AI models can add significant licensing and subscription costs. “Shadow costs” can also emerge from unauthorized or unmonitored AI services spun up by various teams.

AI cost driver	Why it matters
Specialized hardware	GPUs and TPUs are significantly more expensive than CPUs
Data management	Hungry AI models require ingesting, storing and transferring large volumes of data
Model training and operations	Training and running AI models requires massive cloud resources, especially as usage scales
Unpredictable workloads	Dynamic workloads often require expensive on-demand resources that are hard to cover with commitments
Licensing and software	Beyond infrastructure, the use of AI tools often leads to “shadow” license and subscription costs

Table 2: How AI increases cloud costs

How FinOps can help with AI costs

Let's look at the pricing models associated with AI for some common business use cases, and how FinOps can help with each of them.

Subscription/license-based pricing

Generative and agentic AI services are transforming enterprise IT. *Flexera's 2025 State of the Cloud Report* found that 83% of organizations are currently using or experimenting with GenAI services—more than any other new PaaS service listed in the report's history.

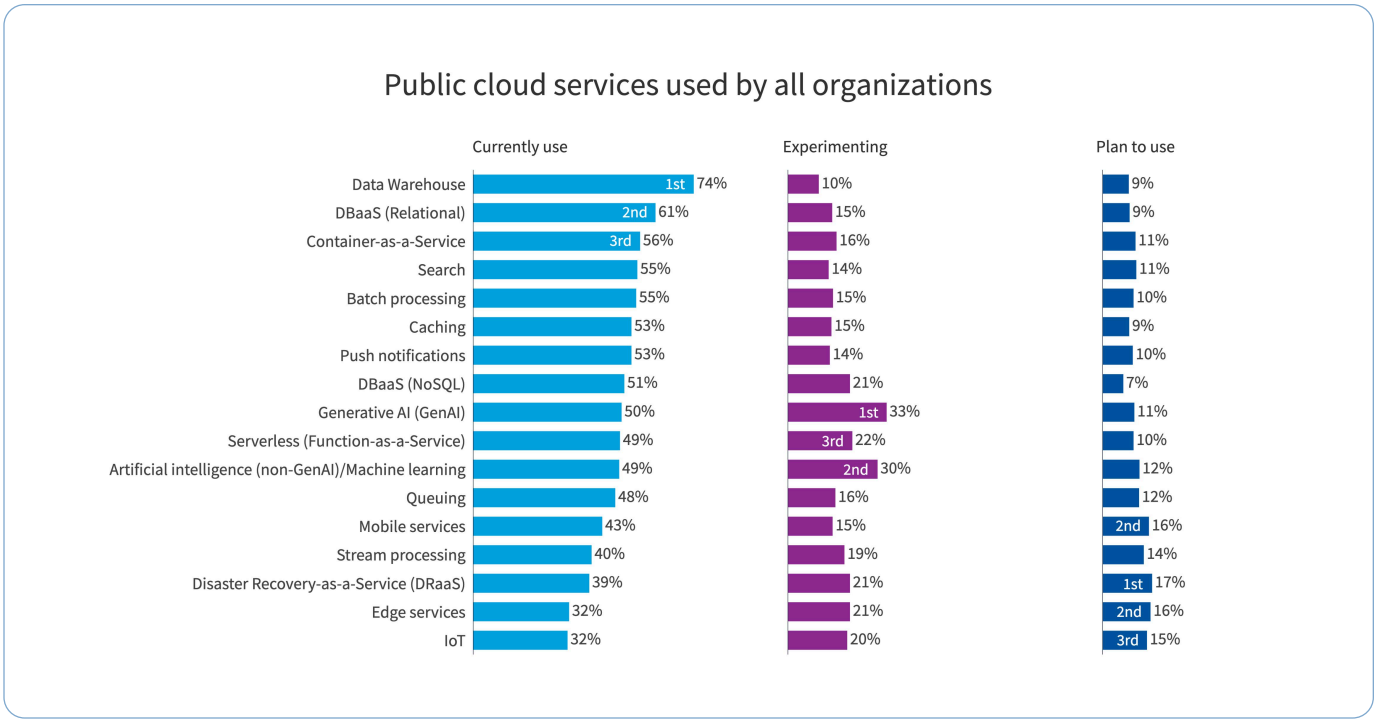


Figure 1: *83% of organizations are currently using or experimenting with GenAI services**

AI assistants such as ChatGPT and Gemini often come with a free version, but users can unlock richer functionality with paid versions, usually via subscription. AI assistants are built on LLMs that specialize in text-related tasks. Typically, users interact with AI assistants by asking questions or making requests in plain language using voice or chat. These questions or requests are processed using natural language processing (NLP) to generate text or image-based responses.

AI agents are typically categorized by their relative autonomy to AI assistants. They are designed for specific tasks (e.g., meeting summarization) and don't require an NLP prompt to begin their work. However, the distinction between AI assistants and AI agents is not always clear, as AI assistants can act as the interface for underlying AI agents, as with Microsoft Copilot.

AI assistants and agents can be billed as standalone subscriptions (e.g., ChatGPT), **or as add-ons or feature upgrades to an existing software license** (e.g., Microsoft 365 Copilot requires a Microsoft 365 license).

***Source:** Flexera 2025 State of the Cloud Report

How FinOps can help with subscription/license-based AI costs

AI applications and services are widely used throughout organizations, but they are not always sanctioned. Because these tools are typically billed via subscriptions or licenses, FinOps teams have not traditionally been called upon to get visibility into this shadow IT usage, or to optimize the cost of software licenses and SaaS spend. Instead, ITAM professionals have traditionally been responsible for software asset management.

However, the scope of FinOps is changing. The 2025 State of FinOps Report found that FinOps teams are being asked to manage additional IT costs beyond IaaS (Infrastructure as a service, a.k.a. "cloud") costs; specifically, SaaS and licensing costs. As the walls between ITAM and FinOps come down, professionals in both disciplines are coming together to solve what is inherently a data challenge across the organization.

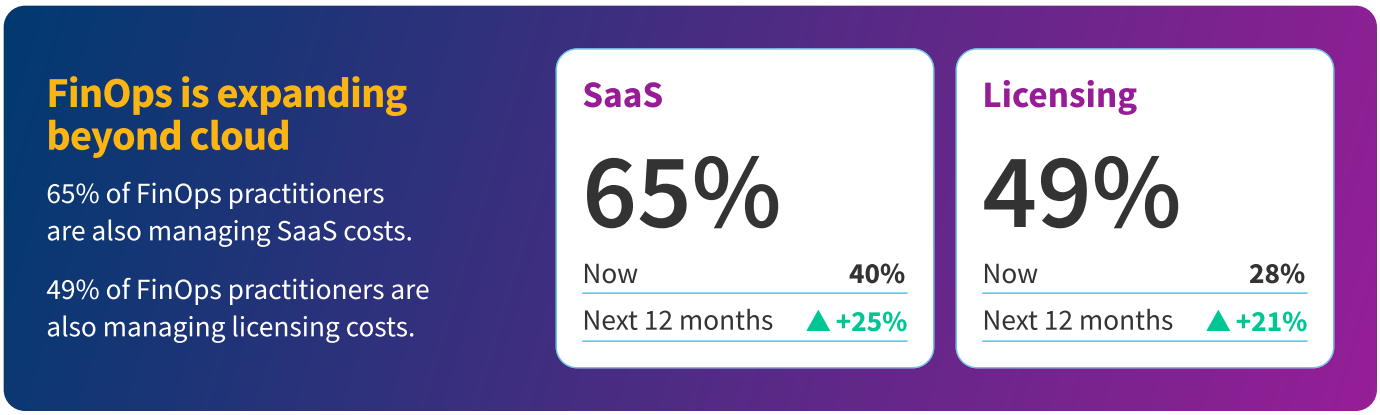


*Figure 2: The expanding scope of FinOps**

Consumption-based pricing

Many AI services today are shifting from simple subscription/license pricing to more complex pricing based on metered usage (consumption)—such as number of API calls, amount of time using the service, or GB of data stored—much like pricing for cloud infrastructure and services.

Emerging pricing models are increasingly tied to business outcomes, such as Zendesk's number of automated resolutions. In this model, if the AI agent handles 100 conversations and successfully resolves 80 of them without human intervention, you'll be charged for those 80 automated resolutions.

With AI resources and services, usage-based and outcome-based prices are often further obscured by metering based on tokens or credits. Salesforce recently abandoned its Agentforce pricing model of $2 per conversation in lieu of Flex Credits, which can be purchased in increments of 100,000 ($500) and are drawn down at different rates for different agentic activities.

These usage-based and outcome-based pricing models use proxies for cost, which make it harder to predict spending. Providers tout the fairness of only paying for what you use, and the ability to control spending by reducing usage if necessary. **However, seeing how much of a service is being used and by whom requires granular, real-time visibility. This is where the practice of FinOps has much to offer.**

***Source:** FinOps Foundation 2025 State of FinOps survey

How FinOps can help with consumption-based AI costs

Consumption-based pricing was first introduced with public cloud infrastructure and continues to be applied to AI resources and services from both cloud providers and third parties. FinOps practitioners play a crucial role in helping their organization gain visibility into consumption-based resources/services and providing recommendations for optimizing the cost of that usage.

In fact, the practice of FinOps was developed specifically to manage costs incurred:

- In a decentralized manner by employees throughout the organization
- With little visibility for finance teams
- As ongoing operational expenses (Opex)

Before cloud and FinOps, IT costs were typically incurred:

- In a centralized manner by procurement processionals
- With high visibility for finance teams
- As upfront capital expenditures (Capex)

Figure 3: *FinOps brings visibility to cloud billing data*

Tokens and credits add a layer of pricing complexity for FinOps practitioners, but in the end, it's still about tracking usage to optimize cost. FinOps teams can apply the following well-established capabilities to resources that are used to build AI models, workloads or applications, and to AI services that employees use throughout the organization:

Ingest and normalize billing data

The cost of cloud-based AI resources and services is included in cloud billing data files just like any other cloud resource or service. However, invoices for AI services that are sold via subscription or by license will need to be ingested and normalized with cloud billing data. A comprehensive FinOps tool (such as Flexera Cloud Cost Optimization) can ingest and normalize all cost data, whether billed by consumption or via subscription or license.

Implement a tagging strategy

Just like with more traditional cloud and SaaS costs, tagging AI resources and services enables accurate cost **allocation** and custom **reporting**. Tagging is supported natively for many cloud and AI services, but a FinOps tool should offer enhanced tagging and labeling capabilities that go beyond native support.

Forecast costs and establish budgets

FinOps teams should incorporate the cost of AI resources and services into their cost projections to help with the development of budgets for different teams, projects or applications.

Optimize rates

Many AI resources are eligible for commitment discounts, and FinOps teams should manage them just like commitment discounts for traditional cloud resources.

Optimize usage

As with all things cloud, usage = cost. FinOps teams should meet regularly with engineering teams to review and optimize the usage of compute, storage, data transfer and other resources/services that engineering uses for AI workloads.

Manage cost anomalies

FinOps teams should incorporate the cost of AI resources and services into their cost projections to help with the development of budgets for different teams, projects or applications.

FinOps Capability icons by FinOps Foundation

Engaging new personas

FinOps teams are used to working with a variety of personas throughout the organization. AI introduces a host of new stakeholders into the mix.

When it comes to managing the cost of subscription-based or license-based software and SaaS, **end-user employees** throughout the organization are signing up for services and incurring costs. FinOps teams should liaise with **SaaS management teams** who have the tools and skills to uncover shadow IT, reduce vendor sprawl and negotiate company-wide discounts.

If your organization is developing AI models or building AI-powered applications or services, FinOps teams may need to engage with **data scientists, data engineers** and **developers** who develop models, prepare data and build AI into applications.

Taking an incremental and iterative approach

FinOps and engineering teams may feel overwhelmed by new pricing models (e.g., tokens) and new terminology (e.g., GPUs), but just like with all FinOps activities, optimizing AI spend should be approached incrementally and iteratively.

Approaching AI spend incrementally: For each capability, teams should review the FinOps Maturity Model and determine what level of maturity makes sense for their organization. For example, see the maturity assessment for cost allocation:

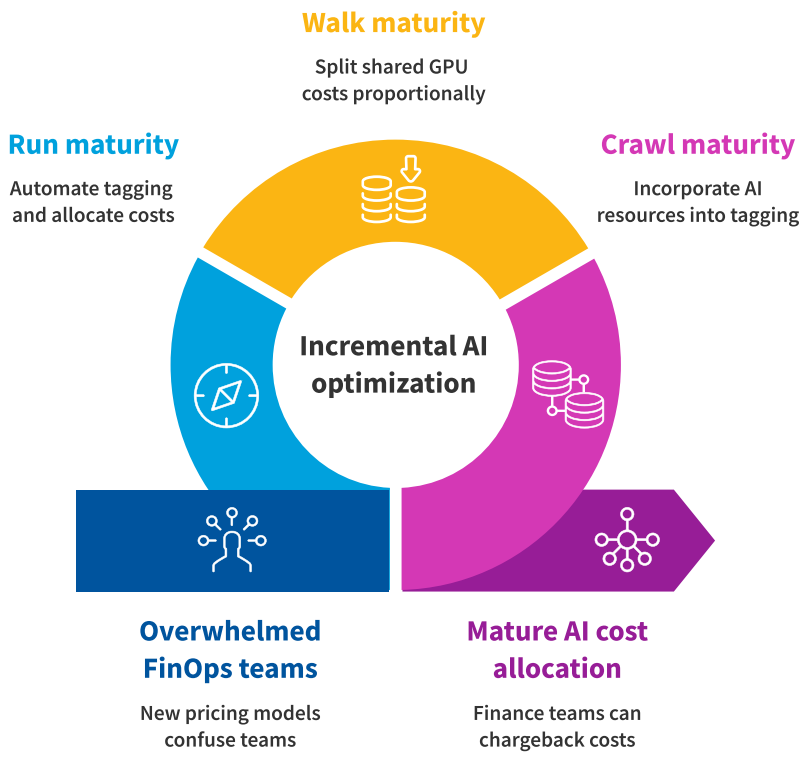


Figure 4: Optimizing AI costs incrementally

- At a **Crawl** level of maturity, allocating AI costs might amount to incorporating AI resources into your tagging strategy, ensuring they are labeled by application, engineering team or business unit
- At a **Walk** level of maturity, AI cost allocation might mean you can split shared GPU costs proportionally by usage, rather than simply splitting the cost evenly across teams
- At a **Run** level of maturity, perhaps your team automatically tags AI resources at provisioning and allocates 100% of AI costs by usage; finance teams can charge back accordingly

Improving AI spend iteratively

Work iteratively using the FinOps phases as a guide for providing critical capabilities to everyone in your organization.

1 Inform

Provide visibility to key stakeholders where they need it. Leadership and finance teams will likely want rollup reports delivered weekly or monthly. Engineering teams will need more frequent visibility at a granular resource level. Understand what each stakeholder requires so you can meet their unique needs.

2 Optimize

Collaborate with SaaS management teams to uncover shadow AI usage and optimize license and subscription costs. For cloud-based AI costs, meet regularly with DevOps teams to find ways to eliminate waste and reduce costs over time.

3 Operate

Coordinate with procurement and with engineers to establish policies for AI usage, and leverage automation to enforce compliance with those policies. Then, complete the cycle over again, making tweaks as your organization's needs and goals change.

Figure 5: FinOps Phases by FinOps Foundation

The business impact of AI cost optimization

Providing granular visibility into AI costs, reducing wasteful AI spending, optimizing AI resource usage and connecting AI costs to business outcomes are critical for maintaining tolerance for AI spending. The desire for agility and rapid innovation can quickly dissipate if soaring costs or changing economic conditions force leadership to shift their strategy from growth to conservation. FinOps practitioners play a valuable role in helping leadership understand the return on AI investments. This understanding often results in:

- **Experimentation and innovation**—more time for the development and adoption of AI projects or applications, both internally and by customers
- **Enhanced productivity**—freedom to use AI services to do work faster, enhance decision-making and generate deliverables
- **Improved unit economics**—lower cost to deliver AI-powered services improves gross margins and increases profitability
- **Competitive advantage**—new and differentiated offerings at competitive price points position your company ahead of the pack

Part two: AI for FinOps

We've gone through the challenges that AI introduces for FinOps teams, but the power of AI can be leveraged to improve the practice of FinOps itself.

Cloud providers and FinOps vendors are incorporating AI into their tooling to democratize cost data and help practitioners practice FinOps faster and more easily.

> "While other FinOps tools offer simple capabilities without deep intelligence, Flexera goes further by leveraging AI to forecast spend, detect anomalies and rightsize infrastructure more intelligently. It works across your multi-cloud environment too—making it a powerful and more complete package."
>
> **Jay Litkey**
> Senior Vice President of Cloud and FinOps at Flexera

Flexera incorporates AI into its entire product suite to help organizations manage cost more intelligently. Flexera's FinOps portfolio offers these AI-powered capabilities:

Forecasting and budgeting

Cloud Cost Optimization in Flexera One uses machine learning to analyze up to 24 months of historical spending to predict future needs.

This feature uses a non-linear model to learn seasonal and other cyclical demands, improving the accuracy and reliability of spend forecasts

Users can convert forecasts to budgets and use policies to alert users to unexpected budget variance

Users can calculate forecasts and budgets by custom business grouping (such as team, application or cost center) for decentralized cost planning

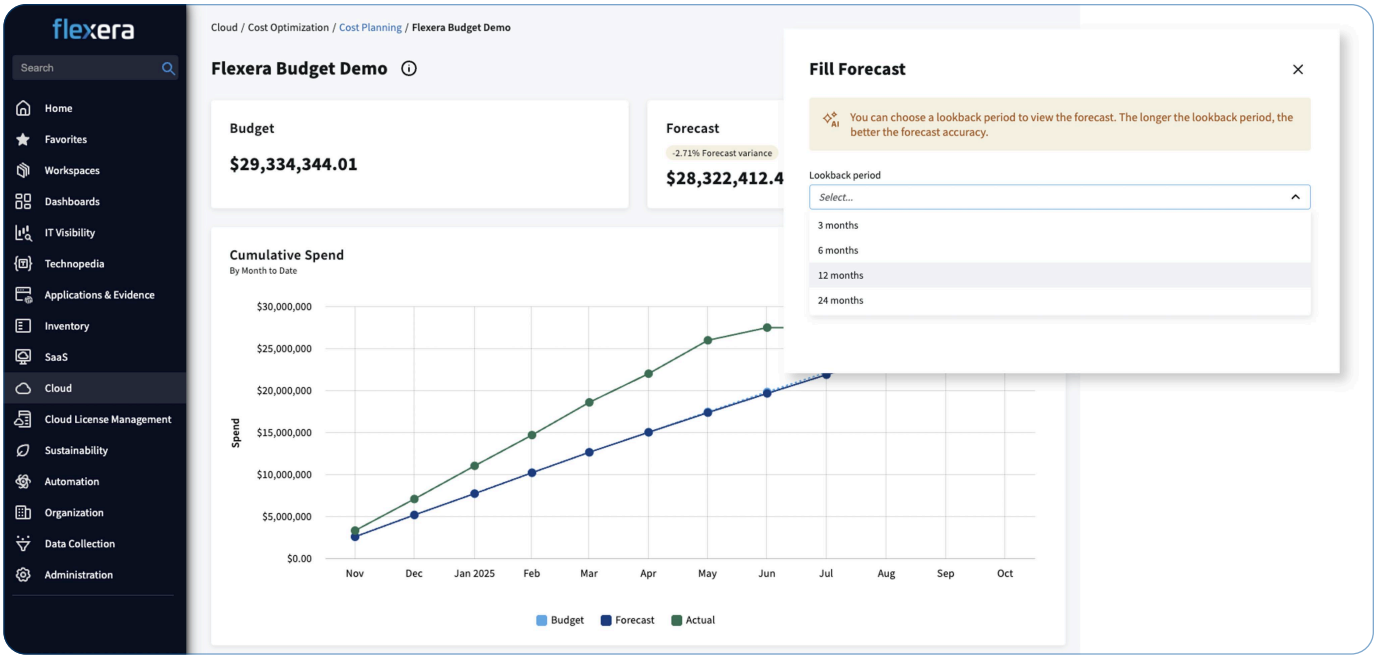


Figure 6: Flexera One Cloud Cost Optimization leverages AI for spend forecasting

Anomaly management

Cloud Cost Optimization in Flexera One has released enhanced detection and root cause analysis for cost anomalies.

Machine learning identifies cost anomalies within a specified threshold, minimizing false positives and negatives

Comprehensive root cause analysis will be provided for each detected anomaly, including potential factors such as resource utilization, pricing changes or billing errors

Users can filter anomalies by various criteria such as resource type, cost impact and time period to facilitate fast remediation

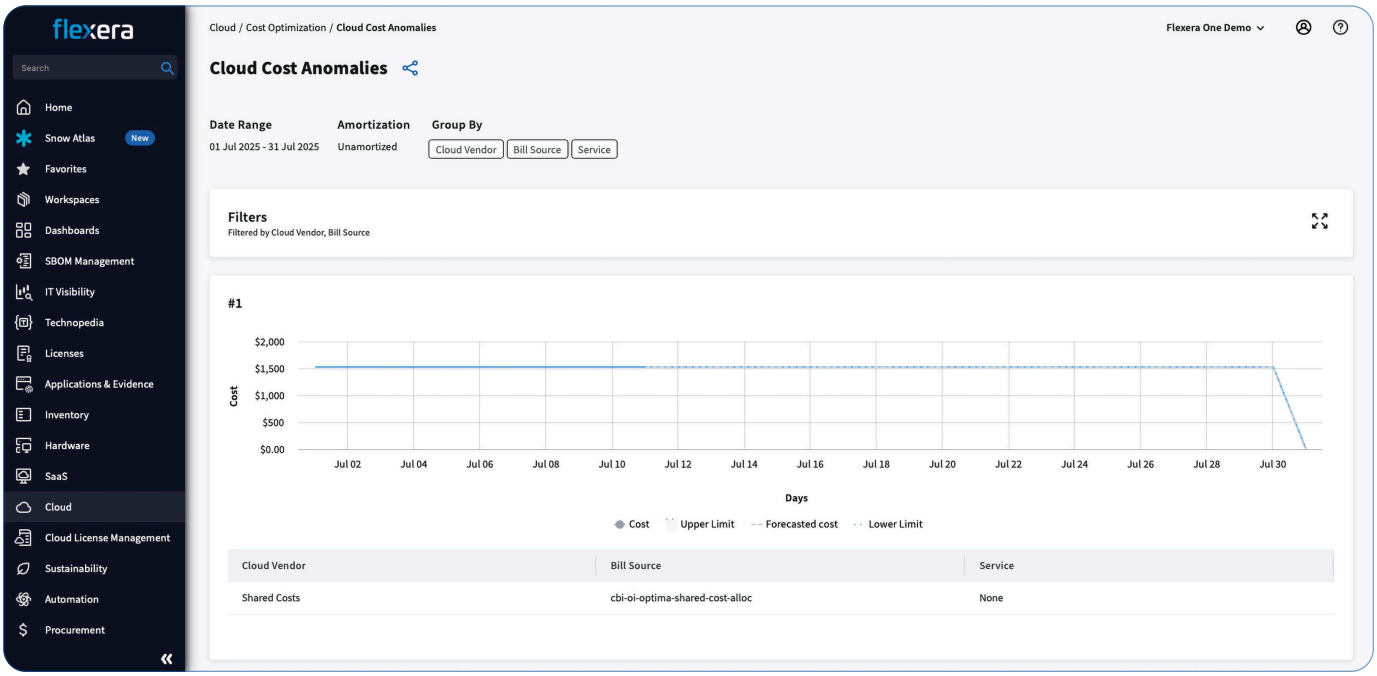


Figure 7: AI-powered cost anomaly detection in Flexera One Cloud Cost Optimization

Predictive VM autoscaling

Flexera's virtual machine optimization solution, Spot Elastigroup, offers AI-driven scaling of virtual machines (VMs) to deliver right-on-time availability in burst scenarios.

- Machine learning algorithms predict the future load of your application up to 2 days in advance and proactively scale the number of instances to accommodate predicted peak traffic, right when this peak is predicted
- Users can choose between *Predict and scale* and *Predict only*

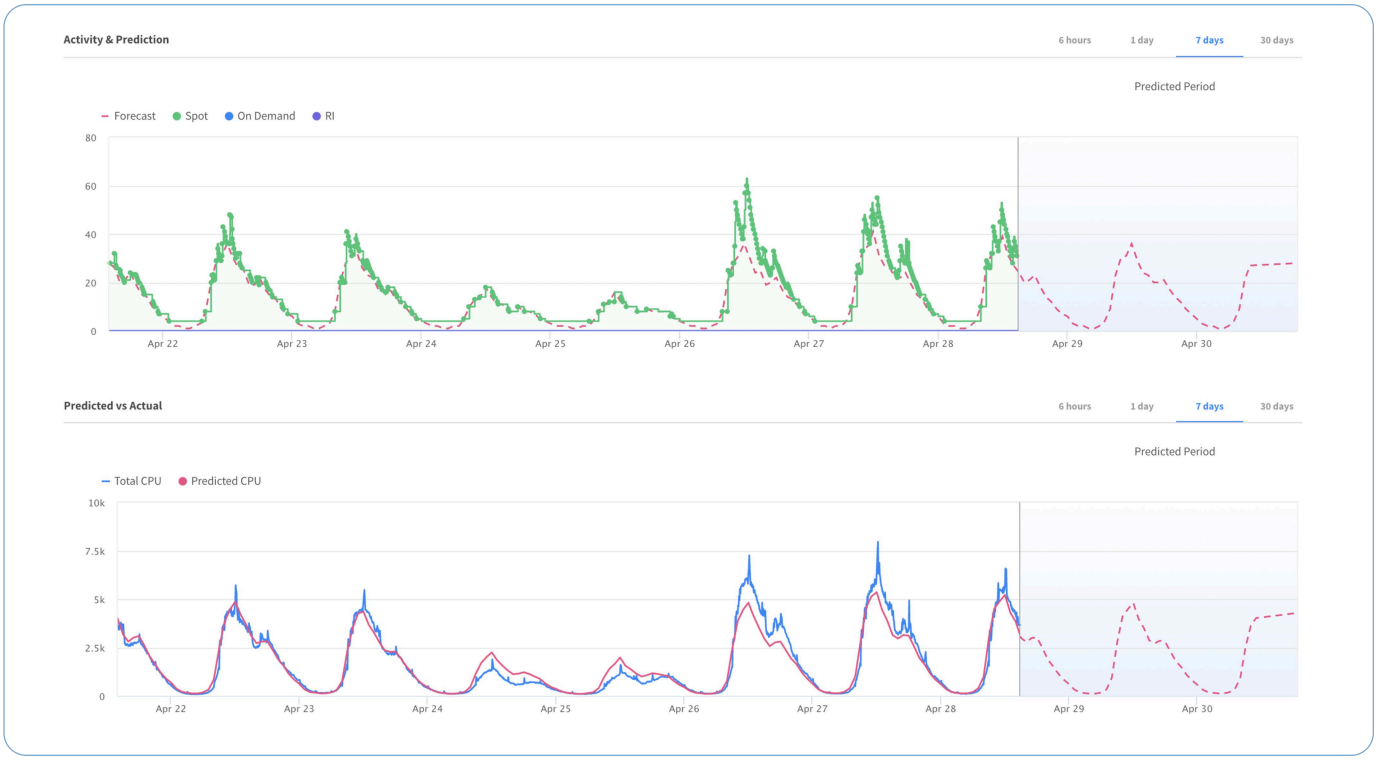


Figure 8: *Flexera One Virtual Machine Optimization uses AI to predict capacity needs and auto-scale VMs accordingly*

Kubernetes rightsizing

Spot Ocean, Flexera's container optimization solution, automates Kubernetes (K8) infrastructure management for Amazon Elastic Kubernetes Service (EKS) and Azure Kubernetes Service (AKS), dynamically scaling K8 clusters in response to real-time demand. This targeted, intelligent automation reduces costs significantly and removes manual oversight, allowing engineers to focus on their primary responsibilities.

Commitment discount management

Flexera One Cloud Commitment Management (formerly Spot Eco) uses machine learning to make intelligent commitment purchasing decisions based on real usage patterns. AI-powered commitment "blending" tailors a unique portfolio that is aligned with the types and quantities of cloud resources used.

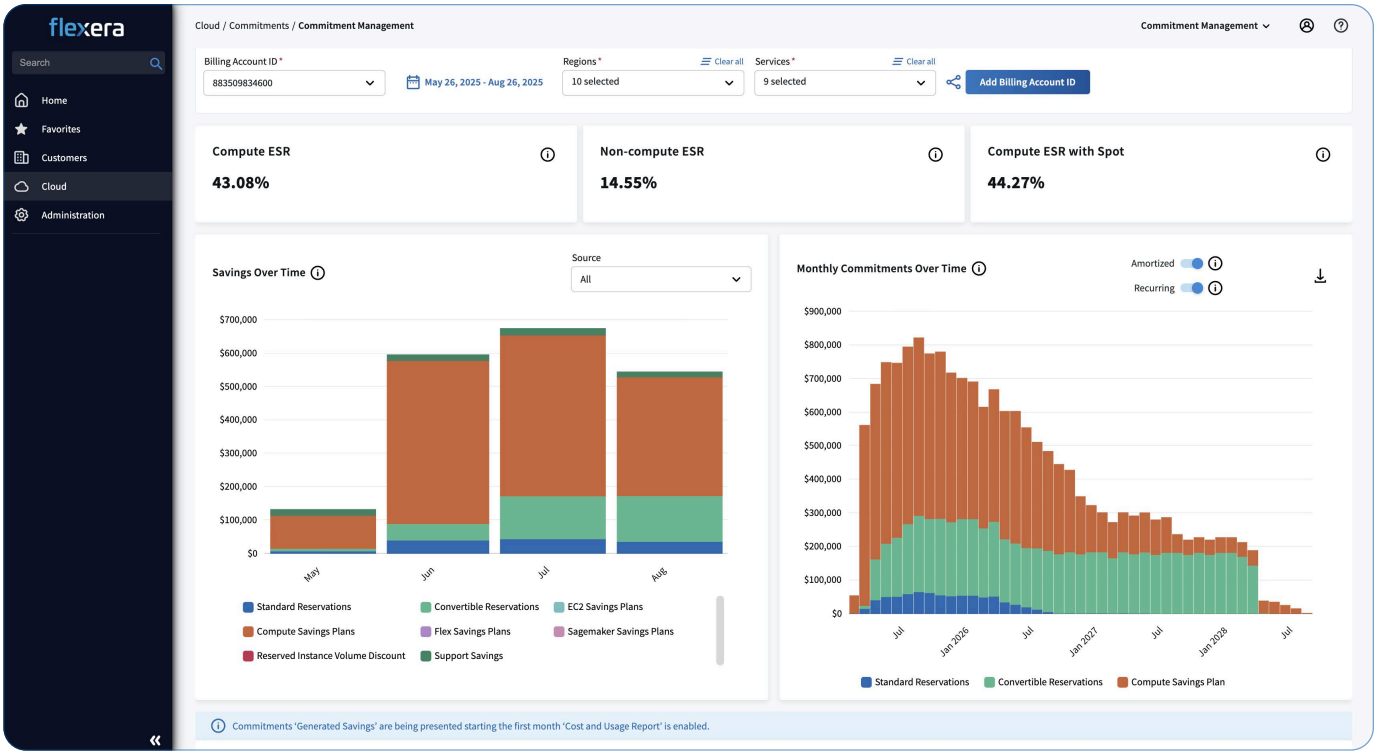


Figure 9: *Flexera One Cloud Commitment Management intelligently blends commitment discounts for maximal resource coverage*

Tagging

On the roadmap for later in 2025, Flexera plans to add **AI-based tagging** in Flexera One to help maintain tagging adherence as new resources are added to your environment.

GenAI for policy creation

Also on the roadmap for this year is the introduction of **generative AI for policy creation**. This update will make it easier for users to create new governance policies without writing code or using the policy interface.

Looking into the future of AI and FinOps

There's no doubt that AI is transforming how we all work, and FinOps practitioners will benefit from AI being incorporated into FinOps tooling. Practitioners are also in a unique position to help businesses get more value from AI investments, by guiding the organization toward responsible AI spending that drives greater efficiency and profit.

As AI and FinOps continue to influence each other, the integration of advanced AI tools will become smoother, making cost management more precise and proactive. FinOps teams will need to stay flexible and keep their skills up to date to make the most of AI. By building a culture that values collaboration and data-driven decisions, organizations can ensure that AI investments aren't just justified but also embraced and supported by everyone. This alignment will be crucial for maintaining financial health and driving growth in the AI-driven future.