

Cloud-Kosten senken: Warum Rabattsätze die Rechnung kaum bewegen

Von Thomas Faß · 1. Juli 2026 · aktualisiert 9. September 2026

Cloud-Kosten senkt man nicht durch weniger Nutzung, sondern durch Steuerung: passende Instanzgrößen (Right-Sizing), das Abschalten ungenutzter Ressourcen, den gezielten Einsatz von Commitment- und Spot-Rabatten sowie eine laufende Kostendisziplin (FinOps). Der Bedarf ist real: Laut Flexera schätzen Organisationen 2026, dass 29 % ihrer Cloud-Ausgaben verschwendet werden. Die größten Einsparungen entstehen dort, wo Architektur, Betrieb und Einkauf zusammenspielen – nicht durch einen einzelnen Trick.

Das Wichtigste in Kürze

- Der Rabattsatz entscheidet die Cloud-Rechnung fast nie. Bis zu 90 Prozent Nachlass auf eine Instanz, die kaum ausgelastet ist, ergeben eine günstige Instanz, die trotzdem Geld kostet. Die Rabattmodelle gibt es seit Jahren, 63 Prozent der Organisationen haben inzwischen ein FinOps-Team, knapp die Hälfte rechnet Stückkosten je Dienst – und bei 64 Prozent der cloudnutzenden Unternehmen in Deutschland sind die Betriebskosten 2025 trotzdem gestiegen. Für 2026 erwarten 54 Prozent weitere Steigerungen (Bitkom Cloud Report 2026, S. 16–17; Flexera 2026 State of the Cloud Report).
- Cloud-Kosten senkt man nicht durch weniger Nutzung, sondern durch Steuerung: Right-Sizing, Abschalten ungenutzter Ressourcen, gezielte Commitment- und Spot-Rabatte und laufende Kostendisziplin (FinOps).
- Der Bedarf ist real: Laut Flexera schätzen Organisationen 2026, dass 29 % ihrer Cloud-Ausgaben verschwendet werden.
- Die Säulen sind Transparenz, Governance und FinOps – ohne Sichtbarkeit keine Steuerung.
- Die größten Einsparungen entstehen dort, wo Architektur, Betrieb und Einkauf zusammenspielen – nicht durch einen einzelnen Trick.

Einleitung: Warum Cloud-Kosteneffizienz heute wichtiger denn je ist und warum ein Monitoring der AWS/Azure/GCP-Ausgaben unerlässlich ist.

Die Cloud hat sich als transformative Kraft für Unternehmen etabliert: Sie bringt Skalierbarkeit, Flexibilität und Innovationstempo, das On-Premises nicht erreicht. Doch mit diesen Vorteilen gehen auch neue Herausforderungen einher, insbesondere im Bereich der Kostenkontrolle. Ohne ein strategisches Management können Cloud-Ausgaben schnell unübersichtlich werden und die erwarteten Vorteile mindern. Hier setzt die Cloud-Kostenoptimierung (CCO) an, ein strategischer Prozess, der darauf abzielt, Cloud-Ausgaben zu reduzieren und gleichzeitig den maximalen Wert aus den genutzten Ressourcen zu ziehen. Das übergeordnete Ziel ist es, ein optimales Gleichgewicht zwischen Ressourcennutzung, Performance und Kosten zu finden.

Die Notwendigkeit einer effektiven Kostenkontrolle in der Cloud ist dringlicher denn je. Obwohl Cloud-Dienste eine hohe Agilität bieten, können sie, wenn sie nicht ordnungsgemäß verwaltet werden, zu unerwarteten Kosten führen. Der Flexera State of the Cloud Report beziffert den verschwendeten Anteil der Public-Cloud-Ausgaben zuletzt auf rund 29 %. Dies liegt oft an überprovisionierten Ressourcen, ungenutzten Instanzen oder einer ineffizienten Architektur. Eine konsequente CCO-Strategie hilft Unternehmen, diese Verschwendung zu identifizieren und zu eliminieren, wodurch die Cloud-Kosten effektiv kontrolliert und die Budgetierung sowie Prognose verbessert werden können.

Die Vorteile einer effektiven Cloud-Kostenoptimierung gehen weit über bloße Einsparungen hinaus. Unternehmen können nicht nur erhebliche, oft zweistellige prozentuale Kostensenkungen erzielen, sondern auch eine verbesserte Effizienz und Performance feststellen. CCO konzentriert sich darauf, die Leistungsfähigkeit von Anwendungen zu optimieren und die Ressourcennutzung zu verbessern, was zu schnelleren Anwendungen und einer besseren Benutzererfahrung führt.

Ein zentraler Aspekt der CCO ist die Maximierung des Return on Investment (ROI) von Cloud-Investitionen. Durch detaillierte Einblicke in die Ressourcennutzung und Kostenmuster können Unternehmen fundierte Entscheidungen treffen, Budgets effektiver zuweisen und sicherstellen, dass ihre Cloud-Ausgaben optimalen Wert liefern. Dies ermöglicht eine datengesteuerte Entscheidungsfindung, die über die reine Kostenreduzierung hinausgeht und die gesamte Geschäftsagilität verbessert. Schließlich hilft ein effektives Cloud-Management, unnötige Ausgaben zu vermeiden und effizientere, weniger komplexe Cloud-Umgebungen zu schaffen, wodurch das Phänomen des „Cloud Sprawl“ eingedämmt wird.

Die wahre Kunst der Cloud-Kostenoptimierung liegt im Erreichen eines optimalen Gleichgewichts zwischen Leistung und Kosten. Es geht nicht darum, blind Ausgaben zu kürzen, was die Performance beeinträchtigen und den Geschäftswert mindern könnte. Vielmehr ist das Ziel, die erforderlichen Leistungsziele mit den geringstmöglichen Kosten zu erreichen, wodurch der Wert der Cloud-Investitionen maximiert wird, ohne die betriebliche Exzellenz zu gefährden. Dieser anspruchsvolle Ansatz erfordert ein tiefes Verständnis der Workload-Anforderungen und eine kontinuierliche Anpassung der Ressourcen.

Ein oft übersehener Aspekt ist das Paradoxon der Cloud-Flexibilität. Die Leichtigkeit, mit der neue Ressourcen bereitgestellt und skaliert werden können, ist ein Hauptvorteil der Cloud. Doch genau diese Eigenschaft kann, wenn sie nicht sorgfältig verwaltet wird, zu einem Haupttreiber für unnötige Ausgaben werden. Das „Pay-as-you-go“-Modell kann sich schnell in ein „Pay-for-what-you-don't-use“ verwandeln, wenn Ressourcen überprovisioniert oder nach Gebrauch nicht wieder freigegeben werden. Dies unterstreicht die Notwendigkeit robuster Governance-Strukturen und automatisierter Prozesse, um die inhärenten Risiken dieser Flexibilität zu mindern.

Letztlich ist die Cloud-Kostenoptimierung ein Katalysator für verbesserte IT-Operationen und eine gesteigerte Geschäftsagilität. Der Prozess zwingt Unternehmen dazu, ihre Ressourcenbedürfnisse genauer zu verstehen und bessere Architektur-Entscheidungen zu treffen. Dies führt zu einer effizienteren Ressourcenzuweisung und dem Aufbau einer schlankeren, reaktionsschnelleren und leistungsfähigeren IT-Infrastruktur. Die Fähigkeit, Ressourcen dynamisch an die Geschäftsanforderungen anzupassen, ist ein direkter Ausfluss einer erfolgreichen CCO-Strategie und trägt maßgeblich zur Wettbewerbsfähigkeit in der digitalen Wirtschaft bei.

Die Säulen der Cloud-Kostenoptimierung: Transparenz, Governance und FinOps

Eine erfolgreiche Cloud-Kostenoptimierung basiert auf einem soliden Fundament aus Transparenz, einer klaren Governance und der Implementierung des FinOps-Frameworks. Diese drei Säulen sind untrennbar miteinander verbunden und bilden die Grundlage für nachhaltige Kosteneinsparungen und eine effiziente Ressourcennutzung.

Sichtbarkeit und Monitoring

Jede ernsthafte Anstrengung zur Cloud-Kostenoptimierung beginnt mit dem Erlangen umfassender Einblicke in die Ausgaben für jede einzelne Ressource. Ohne diese granularen Informationen ist es nahezu unmöglich, fundierte Entscheidungen darüber zu treffen, wo und wie Ausgaben reduziert werden können. Regelmäßige Überwachung und Berichterstattung über die Cloud-Nutzung und die damit verbundenen Kosten sind unerlässlich, um Trends und Anomalien frühzeitig zu erkennen. Cloud-Anbieter stellen hierfür eigene, leistungsstarke Tools zur Verfügung, die Echtzeit-Einblicke in die Ausgaben liefern. Dazu gehören der AWS Cost Explorer, Azure Cost Management + Billing und die GCP Cloud Billing Reports. Diese Tools bieten Dashboards und automatisierte Berichte, die eine kontinuierliche Übersicht über die Ausgaben liefern.

Ressourcen-Tagging und Kostenallokation

Eine der grundlegendsten und gleichzeitig mächtigsten Praktiken für die Kostentransparenz ist das Ressourcen-Tagging. Eine klare und konsistente Benennung und Kategorisierung von Ressourcen mittels Tags ist entscheidend, um deren Zweck zu verstehen und ihre Kosten effektiv zu verfolgen. Ohne ein robustes und konsistentes Tagging-System sind detaillierte Kostenanalysen, die Zuweisung von Kosten zu bestimmten Abteilungen oder Projekten und die Identifizierung ungenutzter Ressourcen erheblich erschwert oder unmöglich. Dies ist der Ausgangspunkt für jede effektive Kostenrückverrechnung und Verantwortlichkeit innerhalb einer Organisation.

Die Automatisierung des Tagging-Prozesses, beispielsweise durch den Einsatz von CloudFormation-Templates oder Lambda-Funktionen, spart nicht nur Zeit, sondern minimiert auch Fehler und gewährleistet eine durchgängige Konsistenz. Es ist von entscheidender Bedeutung, Kostenallokations-Tags in den Billing-Konsolen der Cloud-Anbieter zu aktivieren und regelmäßig zu pflegen. Dies ermöglicht detaillierte Einblicke in Ausgabenmuster und hilft, Bereiche mit Überausgaben präzise zu identifizieren. Tags ermöglichen es, Kosten spezifischen Abteilungen, Projekten oder Teams zuzuordnen, was nicht nur die Transparenz erhöht, sondern auch die Verantwortlichkeit fördert und letztlich zu einer besseren Kostenkontrolle führt.

FinOps-Framework

FinOps, kurz für Financial Operations, ist ein Cloud-Finanzmanagementansatz, der Unternehmen dabei unterstützt, Kosteneffizienz zu erreichen, während Leistung und Skalierbarkeit erhalten bleiben. Im Gegensatz zur traditionellen IT-Budgetierung, die oft isoliert erfolgt, integriert FinOps Finanz-, Engineering- und Business-Teams in eine gemeinsame Verantwortung für Cloud-Ausgaben. Diese Integration ist von entscheidender Bedeutung, da Cloud-Kosten dynamisch und nutzungsbasiert sind, was eine kontinuierliche Zusammenarbeit erfordert. Technische Teams (Ingenieure, [DevOps](#)) müssen die

Kostenfolgen ihrer Entscheidungen berücksichtigen, während die Finanzabteilung Echtzeit-Transparenz über technische Ausgaben erhält. Dies ist ein kultureller Wandel, der die Cloud-Kostenoptimierung zu einer gemeinsamen Aufgabe macht.

Die Kernprinzipien von FinOps umfassen:

- Teamzusammenarbeit: Cloud-Kostenmanagement ist eine gemeinsame Verantwortung über alle relevanten Teams hinweg.
- Geschäftswert-orientierte Entscheidungen: Cloud-Investitionen müssen durch einen messbaren Geschäftsnutzen gerechtfertigt sein, nicht nur durch reine Kostensenkung.
- Jeder übernimmt Verantwortung für die Cloud-Nutzung: Teams und Einzelpersonen sind für ihren Cloud-Verbrauch und ihre Optimierungen rechenschaftspflichtig.
- Zugängliche und zeitnahe Berichte: Echtzeit-Kostentransparenz ist unerlässlich — nur so lassen sich Entscheidungen proaktiv treffen.
- Ein zentralisiertes FinOps-Team treibt Best Practices voran: Eine dedizierte FinOps-Funktion gewährleistet Governance und Richtliniendurchsetzung, auch wenn alle Teams zur Kostenverwaltung beitragen.
- Nutzung des variablen Kostenmodells der Cloud: FinOps fördert den Einsatz von Autoscaling, Reserved Instances und Commitment-basierten Rabatten, um die dynamischen Cloud-Preise optimal zu nutzen.

FinOps ist verbreitet — und die Verschwendung steigt gerade wieder. 63 Prozent der Organisationen stützen sich auf ein FinOps-Team, und 49 Prozent nutzen inzwischen Stückkostenrechnungen, um Cloud-Verbrauch an Geschäftsergebnisse zu koppeln; im Vorjahr waren es 40 Prozent. Gleichzeitig beziffert Flexera den verschwendeten Anteil der Cloud-Ausgaben für 2026 auf 29 Prozent — und nennt das den ersten Anstieg seit fünf Jahren. Im Vorjahresbericht lag der Wert bei 27 Prozent, nach einem Höchststand von 32 Prozent vier Jahre zuvor (Quelle: Flexera, 2025). Als Treiber nennt der Report die stark wachsenden KI-Workloads.

Das ist der eigentliche Befund. Die Disziplin nimmt zu, das Ergebnis verschlechtert sich. Beides zugleich ist nur erklärbar, wenn ein Kostenblock schneller wächst, als die Steuerung ihn einfangen kann.

Der neue Kostenblock heißt KI

Generative KI ist 2026 auf den dritten Platz der meistgenutzten Public-Cloud-Dienste gestiegen, 58 Prozent gegenüber 50 Prozent im Vorjahr. Diese Workloads verhalten sich anders als klassische Anwendungen: Der Verbrauch schwankt mit der Nutzung statt mit der Kapazitätsplanung, und die Nachfrage kommt aus Fachbereichen, die keine Cloud-Kostenstelle kennen.

Für die Steuerung heißt das zweierlei. Commitment-Rabatte greifen schlecht, weil man den Grundbedarf nicht kennt. Und die Zuordnung je Dienst wird von der Kür zur Pflicht — sonst erscheint der KI-Aufwand als anonymer Zuwachs in der Gesamtrechnung.

Budgetierung und Alerts

Das Festlegen von Budgets und das Konfigurieren von Warnmeldungen sind ein Kernaspekt des FinOps-Kostenmanagements, um unerwartete Ausgaben zu verhindern. Unternehmen sollten proaktiv Ausgabenlimits für Teams, Projekte und Workloads festlegen und diese in Echtzeit überwachen. Cloud-Anbieter-Tools wie AWS Budgets, Azure Cost Management und Google Cloud Billing bieten hierfür umfassende Funktionen.

Der Übergang von einem reaktiven zu einem proaktiven Kostenmanagement ist entscheidend. Anstatt auf den monatlichen Rechnungsschock zu warten, ermöglichen Echtzeit-Warnmeldungen bei Ausgabenanomalien oder unerwarteten Kostenspitzen eine schnelle Reaktion. Dies verhindert, dass Kosten außer Kontrolle geraten, bevor sie zu einem größeren Problem werden. Die Automatisierung von kostensparenden Maßnahmen, wie dem automatischen Abschalten von Ressourcen oder der Durchsetzung von Richtlinien, ergänzt diesen Ansatz und gewährleistet eine kontinuierliche Optimierung.

Übersicht über Cloud-Kostenmanagement-Tools:

- AWS Cost Explorer: Kostenvisualisierung, Budgetierung, Warnmeldungen, Prognosen, Rightsizing-Empfehlungen
- Azure Cost Management + Billing: Detaillierte Kostenanalyse, Budgetierung, Prognose, Warnmeldungen, Integration mit Azure Advisor
- GCP Cloud Billing Reports: Detaillierte Berichte, Dashboards (Billing Dashboard, Cost Table, Cost Breakdown), Budgetierung, Warnmeldungen

Fundamentale Strategien zur Multi-Cloud-Kostenoptimierung

Unabhängig vom gewählten Cloud-Anbieter gibt es eine Reihe fundamentaler Strategien, die in jeder Cloud-Umgebung angewendet werden können, um die Kosteneffizienz erheblich zu steigern. Diese anbieterunabhängigen Ansätze bilden das Rückgrat einer umfassenden Cloud-Kostenoptimierungsstrategie.

Rightsizing Ihrer Ressourcen

Rightsizing ist eine Schlüsselstrategie zur Optimierung von Cloud-Umgebungen und beinhaltet die präzise Anpassung von Instanztypen und -größen an die tatsächlichen Leistungs- und Kapazitätsanforderungen der Workloads zu den geringstmöglichen Kosten. Ein häufiger Fehler ist die Überdimensionierung von Instanzen, was zu erheblicher Ressourcenverschwendung führt. Umgekehrt können unterdimensionierte Instanzen zu Leistungseinbußen führen. Durch konsequentes Rightsizing können Unternehmen signifikante Einsparungen erzielen – bei überdimensionierten Ressourcen häufig einen erheblichen zweistelligen Prozentanteil der jeweiligen Kosten.

Der Prozess des Rightsizing umfasst typischerweise drei Schritte: Zunächst werden Nutzungs- und Leistungsmetriken von Volumes, virtuellen Maschinen und Instanzen analysiert. Anschließend wird überprüft, ob die aktuellen Ressourcen die vordefinierten Leistungsbenchmarks erfüllen oder ob ungenutzte Kapazitäten vorhanden sind. Basierend auf diesen Analyseergebnissen werden die Cloud-Ressourcen dann entsprechend hoch- oder herunterskaliert oder sogar beendet. Tools wie AWS Compute

Optimizer, Azure Advisor und Google Cloud Recommender sind hierbei unerlässlich, da sie unterausgelastete Ressourcen identifizieren und konkrete Empfehlungen für das Rightsizing geben können.

Eliminierung von Verschwendung (Cloud Sprawl)

Das Bezahlen für redundante, ungenutzte oder verwaiste Ressourcen ist eine der häufigsten und vermeidbarsten Ursachen für explodierende Cloud-Kosten. Dazu gehören Compute-Instanzen ohne aktive Benutzerinteraktionen, leere Datenbanken, nicht zugewiesene Speichervolumen, veraltete Snapshots oder ungenutzte Load Balancer. Branchenberichte zeigen, dass ein erheblicher Teil der gesamten Cloud-Ausgaben auf solche ungenutzten oder unterausgelasteten Ressourcen entfällt.

Regelmäßige Audits der Cloud-Konten sind unerlässlich, um solche unnötigen Elemente zu erkennen und zu entfernen. Es empfiehlt sich, einen internen Prozess zu etablieren, der festlegt, wann Cloud-Ressourcen stillgelegt oder Workloads heruntergefahren werden sollen, die nicht mehr benötigt werden, um unnötige Ausgaben proaktiv zu vermeiden. Die Verknüpfung von Rightsizing und Abfallbeseitigung ist dabei von großer Bedeutung. Während Rightsizing überprovisionierte, aber genutzte Ressourcen optimiert, zielt die Abfallbeseitigung auf vollständig ungenutzte oder verwaiste Ressourcen ab. Ein effektiver Ansatz beginnt damit, zuerst die vollständig ungenutzten Ressourcen zu identifizieren und zu eliminieren, bevor die verbleibenden aktiven Ressourcen präzise dimensioniert werden. Dies stellt sicher, dass keine Ressourcen optimiert werden, die eigentlich abgeschaltet oder gelöscht werden sollten.

Automatisches Skalieren (Autoscaling)

Autoscaling ist der automatisierte Prozess, der die Menge an Rechenressourcen in einer Cloud-Umgebung basierend auf der aktuellen Nachfrage anpasst. Diese dynamische Anpassungsfähigkeit ist von entscheidender Bedeutung, um sowohl optimale Leistung als auch Kosteneffizienz zu gewährleisten. Autoscaling eliminiert die Ineffizienzen, die durch Überprovisionierung (das Bezahlen für ungenutzte Ressourcen) oder Unterprovisionierung (was zu Leistungsengpässen führen kann) entstehen.

Es gibt verschiedene Typen des Autoscaling:

- **Horizontales Skalieren (Scaling Out/In):** Diese häufiger genutzte Methode beinhaltet das Hinzufügen oder Entfernen von Instanzen (z. B. VMs oder Server) innerhalb eines Workloads. Der Traffic wird dabei oft durch einen Load Balancer gleichmäßig verteilt.
- **Vertikales Skalieren (Scaling Up/Down):** Hierbei wird die Rechenleistung einer einzelnen Instanz angepasst, indem Ressourcen wie Speicher oder CPU-Kapazität hinzugefügt oder entfernt werden.
- **Prädiktives Autoscaling:** Nutzt historische Daten und Machine-Learning-Algorithmen, um zukünftige Anforderungen vorherzusagen und Ressourcen proaktiv anzupassen.
- **Reaktives Autoscaling:** Reagiert in Echtzeit auf Änderungen der Nachfrage durch Überwachung der tatsächlichen Nutzung.

Autoscaling stellt sicher, dass die Ausgaben direkt mit der tatsächlichen Nutzung korrelieren, was zu vorhersehbareren und besser verwaltbaren Kosten führt. Es kann als ein dynamischer Rightsizing-Mechanismus verstanden werden. Anstatt manueller, periodischer Anpassungen bietet Autoscaling eine Echtzeit- und adaptive Lösung für die schwankende Natur von Cloud-Workloads. Es adressiert direkt die

„Spitzen und Täler“ der Nachfrage und verhindert sowohl Über- als auch Unterprovisionierung kontinuierlich. Dies positioniert Autoscaling als die ideale, automatisierte Weiterentwicklung des Rightsizing für dynamische Workloads, die seine Überlegenheit gegenüber manuellem Rightsizing in solchen Szenarien unterstreicht.

Speicheroptimierung

Die Auswahl der richtigen Speicheroptionen und -klassen ist ein kritischer Faktor für die Cloud-Kostenoptimierung. Cloud-Speicheroptionen lassen sich im Allgemeinen in drei Kategorien einteilen: Objektspeicher, Dateispeicher und Blockspeicher. Für jeden Datentyp ist es entscheidend, die passende Speicherklasse basierend auf der Zugriffshäufigkeit und den Leistungsanforderungen zu wählen. Beispielsweise sollten selten genutzte Backups in einem Archiv-Speicher-Tier wie Amazon S3 Glacier gespeichert werden, da dies wesentlich günstiger ist als performantere Tiers, die für häufigen Zugriff konzipiert sind.

Die Speicheroptimierung ist jedoch kein einmaliger Vorgang, sondern ein kontinuierliches Lebenszyklusmanagement. Datenzugriffsmuster ändern sich im Laufe der Zeit; häufig genutzte Daten können zu selten genutzten Archivdaten werden. Daher ist die Implementierung von Lebenszyklusregeln, wie z. B. Amazon S3 Lifecycle Rules, unerlässlich, um Daten automatisch basierend auf ihren Nutzungsmustern in kostengünstigere Speicherklassen zu verschieben. Ohne solche automatisierten Richtlinien verbleiben Daten lange nach ihrer aktiven Nutzung in teuren Tiers, was zu erheblicher Speicherverschwendung führt. Regelmäßige Überprüfungen der Speichernutzung, einschließlich des Löschens veralteter Daten und ungenutzter Buckets, sind ebenfalls notwendig, um unnötige Kosten zu senken.

Strategische Preismodelle und Rabatte über Cloud-Anbieter hinweg

Die Cloud-Anbieter AWS, Azure und GCP bieten eine Vielzahl von Preismodellen und Rabatten an, die bei strategischer Nutzung erhebliche Einsparungen bringen. Ein tiefes Verständnis dieser Optionen und ihrer Anwendungsfälle ist entscheidend, um die Cloud-Ausgaben im Griff zu behalten.

Was ein „bis zu“-Rabatt wirklich sagt

Die Prozentsätze in diesem Kapitel stammen aus der Preisdokumentation der Anbieter und sind korrekt. Sie beschreiben aber den günstigsten möglichen Fall: die längste Bindung, die passende Maschinenfamilie, die richtige Region, durchgehende Nutzung. Wer eine dieser Bedingungen nicht erfüllt, landet unterhalb – oft weit unterhalb.

Drei Größen entscheiden vorher. Erstens die Auslastung: Ein Rabatt wirkt auf den Preis, nicht auf ungenutzte Kapazität. Zweitens die Bindungsdauer: Ein dreijähriges Commitment auf eine Architektur, die in achtzehn Monaten anders aussieht, ist kein Sparmodell, sondern ein zweiter Lock-in – welche Anwendung überhaupt wie lange in ihrer heutigen Form bleibt, entscheidet sich bereits bei der [Cloud-Migration](#). Drittens die Zuordenbarkeit: Wer nicht weiß, welcher Dienst welche Kosten verursacht, kann weder abschalten noch verhandeln.

Ein Rabatt auf eine Ressource, die niemand braucht, ist ein Rabatt auf null Nutzen.

Commitment-basierte Rabatte

Das Konzept der Commitment-basierten Rabatte beruht auf der Verpflichtung zu einer bestimmten Nutzungsmenge über einen festen Zeitraum, typischerweise ein oder drei Jahre. Im Gegenzug erhalten Unternehmen erhebliche Preisnachlässe im Vergleich zu den flexibleren On-Demand-Preisen. Diese Modelle sind ideal für Workloads mit vorhersehbaren und stabilen Nutzungsmustern.

AWS:

- **Reserved Instances (RIs):** Bieten bis zu 72% Rabatt auf On-Demand-Preise. Sie erfordern eine Verpflichtung zu einem spezifischen Instanztyp, Betriebssystem und einer AWS-Region. Sie sind ideal für stabile Workloads, bei denen Instanztyp und Region über den Verpflichtungszeitraum unverändert bleiben. RIs sind für Dienste wie EC2, RDS, Redshift, ElastiCache, OpenSearch Service und DynamoDB verfügbar.
- **Savings Plans:** Bieten ebenfalls bis zu 72% Rabatt auf Compute-Nutzung. Sie sind flexibler als RIs, da sie keine Auswahl von Größe, Betriebssystem oder Tenancy erfordern und sich automatisch auf berechnete Compute-Dienste wie EC2, Fargate und Lambda anwenden.

Azure:

- **Azure Reservations (Reserved VM Instances):** Ermöglichen Einsparungen von bis zu 72% im Vergleich zu Pay-as-you-go-Preisen bei einer Verpflichtung über ein oder drei Jahre. Sie sind ideal für stabile Anforderungen und bieten eine hohe Budgetvorhersehbarkeit. Verfügbar für VMs, SQL Database, Cosmos DB und weitere Dienste.

GCP:

- **Committed Use Discounts (CUDs):** Bieten laut Google-Doku bis zu 55 % Rabatt auf On-Demand-Preise (bei Speicher-optimierten Maschinenserien bis zu 70 %) bei einer Verpflichtung über ein oder drei Jahre zur Nutzung einer Mindestmenge an Ressourcen (z. B. vCPUs oder Speicher). Sie sind ideal für feste Workloads und helfen, variable Ausgaben in Fixkosten umzuwandeln.

Ein entscheidender Grundsatz bei der Nutzung dieser Rabatte ist das Prinzip „Erst Rightsizing, dann Reservieren“. Eine Verpflichtung zu einem 1- oder 3-Jahres-Plan für überdimensionierte Instanzen birgt das Risiko, in ineffiziente Ausgaben zu geraten, da ein Großteil des Rabatts durch die unnötige Kapazität zunichtegemacht wird. Daher sollte die Optimierung der Ressourcennutzung (Rightsizing, Abfallbeseitigung) immer vor der Verpflichtung zu langfristigen Rabatten erfolgen, um die maximalen Einsparungen zu erzielen. Dieser sequentielle Ansatz ist eine bewährte FinOps-Praxis.

Ein Wechsel ist teuer, aber nicht unmöglich. Von den deutschen Unternehmen nennen 41 Prozent den zu hohen finanziellen Aufwand als Grund gegen einen Anbieterwechsel, 45 Prozent den personellen Aufwand und ebenfalls 45 Prozent die Komplexität ihrer Anwendungen; der technische Aufwand liegt mit 32 Prozent deutlich dahinter. Angeführt wird die Liste von Lock-in-Effekten mit 59 Prozent (Quelle: Bitkom Cloud Report 2026). Die Werte sind Mehrfachnennungen und ergeben zusammen nicht 100 Prozent — sie beschreiben eine Rangfolge, keine Aufteilung.

Gleichzeitig hat rund jedes dritte Unternehmen den Wechsel bereits vollzogen: 26 Prozent einmal, 8 Prozent mehrfach, weitere 20 Prozent planen ihn (Quelle: Bitkom Cloud Report 2026). Vor jedem mehrjährigen Commitment steht deshalb dieselbe Frage: Was kostet es, in zwei Jahren anders zu entscheiden? Wer das nicht beziffern kann, kauft den Rabatt mit einer unbekannten Option – dieselben Bitkom-Werte tragen auch die Abwägung in der [Multi-Cloud-Strategie](#).

Spot-Instanzen / Preemptible VMs

Diese Preismodelle nutzen überschüssige oder ungenutzte Kapazitäten der Cloud-Anbieter zu erheblich reduzierten Kosten, oft bis zu 90% Rabatt im Vergleich zu On-Demand-Preisen. Der wesentliche Unterschied und das damit verbundene Risiko ist, dass diese Instanzen jederzeit mit kurzer Vorankündigung (z. B. 2 Minuten bei AWS) beendet werden können, wenn der Anbieter die Ressourcen zurückbenötigt.

Daher sind sie ideal für fehlertolerante, nicht-kritische Workloads, die Unterbrechungen tolerieren können, wie z. B. Batch-Verarbeitung, Entwicklungs-/Testumgebungen, Rendering oder Datenanalyse.

- AWS Spot Instances: Bieten bis zu 90% Rabatt ohne langfristige Verpflichtung.
- Azure Spot Virtual Machines (Spot VMs): Bieten ebenfalls bis zu 90% Rabatt und können jederzeit beendet werden.
- GCP Spot VMs (Nachfolger der Preemptible VMs): Sind laut Google-Doku bis zu 91 % günstiger als reguläre VMs und können jederzeit gestoppt werden.

Weitere Sparoptionen

Neben den genannten Modellen bieten die Cloud-Anbieter weitere spezifische Sparmöglichkeiten:

- Azure Hybrid Benefit: Ermöglicht die Nutzung bestehender On-Premises Windows Server- und SQL Server-Lizenzen (sowie Linux-Lizenzen) in Azure – laut Microsoft mit Einsparungen von bis zu 85 % bei SQL Server bzw. bis zu 80 % bei Windows Server, da nur für die Compute-Rate gezahlt werden muss.
- GCP Sustained Use Discounts (SUDs): Werden von Google Cloud automatisch angewendet, wenn bestimmte Ressourcen (z. B. Compute Engine VMs) für einen erheblichen Teil des Abrechnungsmonats genutzt werden. Der Rabatt steigt mit längerer Nutzungsdauer – laut Google-Doku bis zu 30 % netto bei durchgehender Monatsnutzung – ohne vorherige Verpflichtung.

Die strategische Nutzung dieser Preismodelle erfordert ein ausgeklügeltes Verständnis der Workload-Eigenschaften wie Vorhersehbarkeit, Fehlertoleranz und Kritikalität. Es gibt kein einzelnes Allheilmittel; die kostengünstigste Strategie beinhaltet oft ein strategisches Mischen verschiedener Modelle: On-Demand für unvorhersehbare Spitzen, Commitment-basierte Modelle für stabile Grundlasten und Spot-/Preemptible-Instanzen für unterbrechbare Aufgaben. Dies ermöglicht das beste Kosten-Leistungs-Verhältnis.

Die Flexibilität der Cloud-Modelle birgt auch versteckte Kosten. On-Demand-Preise bieten maximale Flexibilität, aber zu den höchsten Kosten. Commitment-basierte Preise bieten erhebliche Rabatte im Austausch für reduzierte Flexibilität und eine Verpflichtung. Spot-/Preemptible-VMs bieten noch tiefere

Rabatte, aber mit dem Risiko einer Unterbrechung. Dies verdeutlicht einen klaren Kompromiss: Höhere Flexibilität bedeutet im Allgemeinen höhere Kosten, während höhere Vorhersehbarkeit und die Toleranz gegenüber Unterbrechungen größere Einsparungen eröffnen. Unternehmen zahlen einen Aufpreis für Agilität und mangelnde Weitsicht. Daher ist das Verständnis der Workload-Vorhersehbarkeit für die Wahl des richtigen Preismodells von größter Bedeutung. Investitionen in Prognosen und Workload-Analysen führen direkt zu größeren Kosteneinsparungen.

Anbieterspezifische Tools und Empfehlungen

Jeder große Cloud-Anbieter – AWS, Azure und GCP – bietet eine Reihe von nativen Tools und spezifischen Empfehlungen, die Unternehmen bei der Kostenoptimierung innerhalb ihrer jeweiligen Cloud-Umgebungen unterstützen. Diese Tools sind oft die erste und grundlegendste Verteidigungslinie im Kampf gegen unnötige Cloud-Ausgaben.

AWS-spezifische Tools und Empfehlungen

AWS stellt eine umfassende Palette an Diensten zur Kostenverwaltung und -optimierung bereit:

- **AWS Cost Explorer:** Dieses Tool ermöglicht es Benutzern, ihre AWS-Kosten und -Nutzung über die Zeit zu visualisieren, zu verstehen und zu verwalten. Es bietet auch Rightsizing-Empfehlungen und hilft dabei, die Einsparungen aus Savings Plans zu verfolgen.
- **AWS Compute Optimizer:** Dieses Tool liefert Empfehlungen für optimale EC2-Instanztypen, Amazon EBS-Volumes und AWS Lambda-Funktionen basierend auf der tatsächlichen Nutzung. Es kann auch Empfehlungen für die Migration zu Graviton-Prozessoren geben, die ein besseres Preis-Leistungs-Verhältnis bieten.
- **AWS Trusted Advisor:** Bewertet die AWS-Umgebung kontinuierlich anhand von Best Practices in verschiedenen Kategorien, einschließlich der Kostenoptimierung. Es identifiziert ungenutzte Ressourcen, wie z. B. im Leerlauf befindliche EC2-Instanzen oder nicht zugeordnete Elastic IP-Adressen, und schlägt konkrete Maßnahmen zur Behebung vor.
- **Amazon S3 Intelligent-Tiering:** Dieser Speicherdienst automatisiert die Datenverschiebung zwischen verschiedenen S3-Speicherklassen basierend auf den Zugriffsmustern, wodurch Speicherkosten ohne manuelle Eingriffe optimiert werden.
- **Graviton-Prozessoren:** AWS's eigene ARM-basierte Prozessoren bieten für viele Workloads ein signifikant besseres Preis-Leistungs-Verhältnis und sind eine wichtige Option zur Kostenoptimierung, insbesondere für rechenintensive Aufgaben.

Azure-spezifische Tools und Empfehlungen

Für Azure-Umgebungen stehen folgende Tools und Strategien zur Verfügung:

- **Azure Cost Management + Billing:** Eine umfassende Suite nativer Tools, die zur Überwachung, Analyse, Budgetierung und Prognose von Azure-Ausgaben dient. Sie bietet detaillierte Kostenanalysen und integriert sich nahtlos mit Azure Advisor für personalisierte Empfehlungen.

- **Azure Advisor:** Bietet personalisierte Empfehlungen zur Kostenoptimierung, indem unterausgelastete Ressourcen identifiziert, Rightsizing-Möglichkeiten vorgeschlagen und langfristige Rabattmodelle empfohlen werden.
- **Azure Hybrid Benefit:** Ermöglicht erhebliche Kosteneinsparungen (laut Microsoft bis zu 85 % bei SQL Server, bis zu 80 % bei Windows Server) durch die Nutzung bestehender On-Premises-Lizenzen in Azure-VMs und anderen Diensten. Dies reduziert die Notwendigkeit, neue Lizenzen zu erwerben.
- **Auto-Shutdown für VMs:** Die Implementierung automatischer Abschaltzeiten für nicht-produktionsbezogene VMs ist eine einfache, aber effektive Methode, um Kosten für ungenutzte Kapazitäten außerhalb der Betriebszeiten zu vermeiden.

GCP-spezifische Tools und Empfehlungen

Google Cloud Platform bietet ebenfalls eine Reihe von Funktionen zur Kostenoptimierung:

- **Cloud Billing Reports:** Die Google Cloud-Plattform stellt detaillierte Berichte und Dashboards (wie das Billing Dashboard, die Cost Table und den Cost Breakdown) zur Verfügung, um Ausgaben zu überwachen, Trends zu erkennen und Kostenspitzen zu identifizieren. Dies ermöglicht eine granulare Kontrolle über die Ausgaben.
- **Committed Use Discounts (CUDs) & Sustained Use Discounts (SUDs):** Diese wurden bereits in Abschnitt 4 detailliert, sind aber zentrale GCP-spezifische Preisstrategien, die hier als Empfehlung für die Nutzung aufgeführt werden. CUDs bieten Rabatte für feste Verpflichtungen, während SUDs automatisch für längere Nutzungsdauern gewährt werden.
- **Custom Machine Types:** Diese Funktion ermöglicht die Erstellung von VMs mit benutzerdefinierten vCPU- und Speicher-Konfigurationen. Dies erlaubt es, die Workload-Anforderungen präzise zu erfüllen und Überprovisionierung zu vermeiden, was zu Kosteneinsparungen führt.
- **Spot VMs (früher Preemptible VMs):** GCPs Äquivalent zu Spot-Instanzen, ideal für fehlertolerante, kurzlebige Workloads zu stark reduzierten Preisen, weil sie überschüssige Kapazität nutzen.
- **Storage Class Optimization:** Die Auswahl der richtigen Speicherklassen (Standard, Nearline, Coldline, Archive) und die Implementierung von Lebenszyklusrichtlinien zur automatischen Datenverschiebung in kostengünstigere Tiers sind entscheidend für die Optimierung der Speicherkosten.

Die nativen Tools der Cloud-Anbieter dienen als erste Verteidigungslinie und bilden die Grundlage für fortgeschrittene Optimierungsstrategien. Sie bieten die notwendige Transparenz und die Möglichkeit, Richtlinien durchzusetzen, bevor komplexere Maßnahmen ergriffen werden. Diese Tools sind entscheidend, um die Cloud-Ausgaben zu visualisieren, zu analysieren und erste Optimierungsschritte einzuleiten. Sie ermöglichen es Organisationen, schnell Bereiche mit Ineffizienz zu identifizieren und die Ressourcennutzung zu verbessern.

Fazit und Empfehlungen

Die Cloud-Kostenoptimierung ist in der heutigen dynamischen Cloud-Landschaft keine Option mehr, sondern eine geschäftliche Notwendigkeit. Die Analyse zeigt, dass Unternehmen ohne proaktive Strategien

und Tools einem erheblichen Risiko unerwarteter und unnötiger Ausgaben ausgesetzt sind. Das Kernziel der CCO geht dabei über die reine Kostensenkung hinaus; es ist das Erreichen eines optimalen Gleichgewichts zwischen Leistung, Effizienz und Kosten, um den maximalen Geschäftswert aus Cloud-Investitionen zu ziehen.

Die Flexibilität und On-Demand-Natur der Cloud, obwohl vorteilhaft, birgt das Paradoxon, dass sie ohne sorgfältige Verwaltung leicht zu Verschwendung führen kann. Dies erfordert einen strategischen, disziplinübergreifenden Ansatz, der technische und finanzielle Disziplinen miteinander verbindet.

Zentrale Empfehlungen für eine effektive Cloud-Kostenoptimierung:

- **Etablieren Sie eine FinOps-Kultur:** Verankern Sie die gemeinsame Verantwortung für Cloud-Ausgaben im gesamten Unternehmen. Fördern Sie die Zusammenarbeit zwischen Finanz-, Engineering- und Business-Teams, um Ausgabenentscheidungen auf den Geschäftswert auszurichten. Ein dediziertes FinOps-Team kann Best Practices vorantreiben und die Governance sicherstellen.
- **Sorgen Sie für umfassende Transparenz durch Tagging und Monitoring:** Implementieren Sie eine strikte und konsistente Tagging-Strategie für alle Cloud-Ressourcen. Dies ist die Grundlage für eine genaue Kostenallokation und die Identifizierung von Verschwendung. Nutzen Sie die nativen Kostenmanagement-Tools der Cloud-Anbieter (AWS Cost Explorer, Azure Cost Management, GCP Cloud Billing Reports) für Echtzeit-Monitoring und detaillierte Berichte.
- **Implementieren Sie proaktive Kostenkontrollen:** Gehen Sie von einem reaktiven zu einem proaktiven Ansatz über. Legen Sie Budgets fest und konfigurieren Sie Echtzeit-Warnmeldungen für Ausgabenanomalien. Automatisieren Sie das Abschalten von nicht-produktionsbezogenen Ressourcen außerhalb der Betriebszeiten und implementieren Sie Richtlinien zur Ressourcenbereinigung.
- **Priorisieren Sie Rightsizing und Abfallbeseitigung:** Beginnen Sie mit der Eliminierung von ungenutzten oder verwaisten Ressourcen (Cloud Sprawl), bevor Sie aktive Ressourcen anpassen. Nutzen Sie Tools wie AWS Compute Optimizer, Azure Advisor und Google Cloud Recommender, um die Größe Ihrer Instanzen und Speichervolumen präzise an die tatsächlichen Workload-Anforderungen anzupassen.
- **Setzen Sie Autoscaling strategisch ein:** Für dynamische Workloads ist Autoscaling der effektivste Mechanismus für kontinuierliches Rightsizing. Implementieren Sie horizontale und vertikale Skalierungsstrategien, einschließlich prädiktiver Ansätze, um Ressourcen dynamisch an die Nachfrage anzupassen und Über- oder Unterprovisionierung zu vermeiden.
- **Nutzen Sie Preismodelle intelligent und diversifiziert:** Verstehen Sie die Eigenschaften Ihrer Workloads (Vorhersehbarkeit, Fehlertoleranz, Kritikalität) und wählen Sie die passenden Preismodelle. Kombinieren Sie Commitment-basierte Rabatte (Reserved Instances, Savings Plans, Committed Use Discounts) für stabile Grundlasten mit Spot-/Preemptible-Instanzen für fehlertolerante, unterbrechbare Workloads. Berücksichtigen Sie auch Hybrid-Vorteile wie den Azure Hybrid Benefit. Achten Sie darauf, dass Rightsizing vor der Verpflichtung zu langfristigen Rabatten erfolgt, um die Einsparungen zu maximieren.

- Optimieren Sie die Speichernutzung kontinuierlich: Wählen Sie die richtige Speicherklasse basierend auf Zugriffshäufigkeit und Leistung. Implementieren Sie automatisierte Lebenszyklusrichtlinien, um Daten in kostengünstigere Tiers zu verschieben, sobald sich ihre Nutzungsmuster ändern.

Die Cloud-Kostenoptimierung ist ein fortlaufender Prozess, keine einmalige Aufgabe. Durch die konsequente Anwendung dieser Strategien können Unternehmen nicht nur ihre Ausgaben erheblich senken, sondern auch ihre betriebliche Effizienz steigern, die Leistung verbessern und die Agilität ihrer Cloud-Infrastruktur nachhaltig sichern.

Bleibt die unbequeme Bilanz. Alles, was dieser Artikel beschreibt, ist verfügbar: Right-Sizing-Werkzeuge, Commitment-Modelle, Spot-Kapazität, Stückkostenrechnung, FinOps-Teams. Nichts davon ist neu, nichts davon ist geheim. Und trotzdem stiegen die Betriebskosten bei 64 Prozent der Unternehmen (Quelle: Bitkom Cloud Report 2026), und die Verschwendung nimmt erstmals seit fünf Jahren wieder zu.

Der Engpass liegt selten im Werkzeug. Er liegt darin, dass jemand die Verantwortung für eine Ressource trägt, ihre Kosten sieht und die Befugnis hat, sie abzuschalten. Das ist eine Organisationsfrage, und sie wird von keinem Rabattsatz beantwortet — sie gehört in den [Cloud- und FinOps-Betrieb](#) hinein, nicht daneben.

Ihr nächster Schritt

Wenn Ihre Cloud-Rechnung steigt und niemand genau sagen kann, welcher Dienst dahintersteckt, ist das kein Preisproblem. Wir bringen Kostenzuordnung, Auslastung und Commitment-Struktur in eine Form, in der Entscheidungen möglich sind — mit deutschsprachigen Senior-Teams und Steuerung in Deutschland.

Häufige Fragen

Warum ist die Cloud oft teurer als gedacht?

Weil einfaches Verschieben in die Cloud ohne Optimierung die Rechnung nur verlagert — meist nach oben. Überdimensionierte Instanzen, vergessene Ressourcen und fehlende Kostentransparenz summieren sich. Laut Flexera 2026 gilt fast ein Drittel der Cloud-Ausgaben als verschwendet.

Was bringen Commitment- und Spot-Rabatte konkret?

Wer planbare Grundlast auf ein bis drei Jahre festlegt, erhält deutliche Nachlässe: AWS nennt für Savings Plans bis zu 72 %, Google für Committed Use Discounts bis zu 55 % (bei bestimmten Maschinentypen mehr). Für unterbrechbare Lasten bietet AWS Spot bis zu 90 % Rabatt gegenüber On-Demand. Die Kunst ist, die Rabattform zur Lastform passend zu wählen.

Ist Kostenoptimierung ein einmaliges Projekt?

Nein. Cloud-Kosten sind ein laufendes Thema (FinOps): kontinuierliches Monitoring, Right-Sizing, Tagging und Verantwortlichkeit statt einmaliger Aufräumaktion. Ohne feste Disziplin wächst die Rechnung wieder — meist unbemerkt.

Was ist FinOps?

FinOps ist die betriebliche Disziplin, Cloud-Kosten kontinuierlich zu steuern — mit gemeinsamer Verantwortung von Technik, Betrieb und Einkauf. Statt Kosten einmal zu senken, macht FinOps sie dauerhaft transparent und beeinflussbar.

Was bedeutet Right-Sizing?

Right-Sizing bedeutet, Ressourcen auf die tatsächlich benötigte Größe anzupassen, statt sie überdimensioniert laufen zu lassen. Viele Cloud-Rechnungen sind zu hoch, weil Instanzen größer gebucht sind, als die reale Last verlangt.

Was ist Cloud Sprawl und wie vermeidet man ihn?

Cloud Sprawl ist das unkontrollierte Wuchern ungenutzter oder vergessener Ressourcen — laufende Instanzen, verwaiste Volumes, Test-Umgebungen. Vermeiden lässt es sich durch Transparenz, Tagging und konsequentes Abschalten, was nicht gebraucht wird.

Wie hilft Autoscaling beim Sparen?

Autoscaling passt die Kapazität automatisch an die Last an: bei wenig Nachfrage laufen weniger Ressourcen, bei Spitzen mehr. So zahlt man für Bedarf statt für Vorhaltung — ein zentraler Hebel gegen Verschwendung.

Thematische Vertiefungen

[Multi-Cloud-Strategie](#)

[DevOps in der Cloud: Beschleunigung Ihrer Softwareentwicklung](#)

[Cloud, DevOps & FinOps \(Expertise\)](#)

Quellen

- Flexera: 2026 State of the Cloud Report, Pressemitteilung vom 18. März 2026. Befragung von 753 Cloud-Entscheidern, global. Belegt: 63 % FinOps-Team; 49 % Stückkostenrechnung, im Vorjahr 40 %; verschwendeter Anteil der Cloud-Ausgaben 29 % mit dem ersten Anstieg seit fünf Jahren; generative KI 58 %, im Vorjahr 50 %. Flexera veröffentlicht kein Erhebungsdesign über die Stichprobengröße hinaus. — <https://www.flexera.com/about-us/press-center/flexera-finds-cloud-value-is-rising-while-ai-waste-grows>
- Flexera: 2025 State of the Cloud Report. Belegt die Lesart der Kennzahl als Anteil: „estimated wasted cloud spend on IaaS and PaaS has decreased, reaching a low of 27 % this year“, nach einem Höchststand von 32 % vier Jahre zuvor. — <https://resources.flexera.com/web/pdf/Flexera-State-of-the-Cloud-Report-2025.pdf>
- Bitkom e. V.: Cloud Report 2026 — Status Quo & Trends in Wirtschaft und Politik, 17. Juni 2026. CATI-Befragung von 603 Unternehmen in Deutschland ab 20 Beschäftigten, KW 14 bis KW 20 2026. Belegt: 64 % gestiegene Betriebskosten 2025 (19 % stark, 45 % leicht) und 14 % / 40 % Steigerungserwartung 2026 (S. 16, Basis n=517); Wechselbarrieren 59 / 45 / 45 / 41 / 32 % und Wechselquoten 26 / 8 / 20 % (S. 20). — <https://www.bitkom.org/sites/main/files/2026-06/17062026-bitkom-studienbericht-cloud-report-2026.pdf>
- Amazon EC2 Spot Instances (bis zu 90 % Rabatt gegenüber On-Demand) — <https://aws.amazon.com/ec2/spot/>

- **AWS: EC2-Kaufoptionen — Savings Plans (bis zu 72 %)** — <https://docs.aws.amazon.com/decision-guides/latest/decision-guides/ec2-purchasing-options-aws-how-to-choose.html>
- **Google Cloud: Committed Use Discounts für Compute Engine (bis zu 55 % / 70 %)** — <https://docs.cloud.google.com/compute/docs/instances/committed-use-discounts-overview>
- **AWS: Amazon EC2 Reserved Instances (bis zu 72 % gegenüber On-Demand)** — <https://aws.amazon.com/ec2/pricing/reserved-instances/>
- **Microsoft Azure: Reserved Virtual Machine Instances (bis zu 72 % gegenüber Pay-as-you-go)** — <https://azure.microsoft.com/en-us/pricing/reserved-vm-instances/>
- **Microsoft Azure: Spot Virtual Machines (bis zu 90 % Rabatt)** — <https://azure.microsoft.com/en-us/products/virtual-machines/spot/>
- **Microsoft Azure: Azure Hybrid Benefit (bis zu 85 % SQL Server / 80 % Windows Server)** — <https://azure.microsoft.com/en-us/pricing/hybrid-benefit/>
- **Google Cloud: Spot VMs (bis zu 91 % Rabatt)** — <https://docs.cloud.google.com/compute/docs/instances/spot>
- **Google Cloud: Sustained Use Discounts (bis zu 30 % netto, automatisch)** — <https://docs.cloud.google.com/compute/docs/sustained-use-discounts>