



FINOPS & OPTIMISATION CLOUD

Réduire sa Facture AWS de 30 à 40 %

Le guide pratique pour combiner engagements tarifaires, rightsizing et automatisation — et transformer le FinOps en discipline continue plutôt qu'en audit ponctuel.

11 pages · Comparatif RI / Savings Plans / Spot, grille d'auto-évaluation et feuille de route incluses

29%

gaspillage cloud moyen constaté en 2026

72%

remise max. avec des Reserved Instances

90%

remise max. avec des instances Spot

SOMMAIRE

Ce que vous allez trouver dans ce document

Un parcours en six temps, de la compréhension des leviers de coût à la mise en œuvre d'une feuille de route FinOps opérationnelle.

01	Synthèse — l'essentiel en une page	p.3
02	Pourquoi le FinOps redevient un sujet de direction générale	p.4
03	Les leviers structurels : rightsizing, Graviton, stockage	p.5
04	Les engagements tarifaires : RI, Savings Plans, Spot	p.6
05	Automatiser le pilotage : scheduling, autoscaling, outils	p.7
06	Exemple illustratif : combiner les leviers	p.8
07	Grille d'auto-évaluation FinOps	p.9
08	Feuille de route en 5 étapes	p.10
09	Conclusion & à propos de Move2Cloud	p.11

À qui s'adresse ce document ? DSI, RSSI, responsables FinOps et dirigeants d'entreprises dont la facture cloud dépasse leurs prévisions, et qui cherchent une méthode reproductible plutôt qu'un audit ponctuel.

SYNTHÈSE

L'essentiel en une page

Après cinq années de baisse continue, le gaspillage cloud est reparti à la hausse en 2026 — et l'IA en est le principal moteur.

Selon le rapport Flexera 2026, le taux de gaspillage cloud moyen est remonté à **29 %** des dépenses, une première hausse après cinq ans de recul régulier. Pour une grande entreprise dépensant plus de 60 millions d'euros par an sur le cloud — le cas de 76 % des grandes organisations interrogées — cela représente environ **17 millions d'euros de dépenses évitables chaque année**. Le principal moteur de cette remontée : les instances GPU, qui coûtent 5 à 10 fois plus cher à l'heure que des instances CPU standards, et pour lesquelles les outils FinOps historiques n'ont pas encore pleinement adapté leurs contrôles.

Le signe encourageant : la discipline FinOps elle-même progresse vite. Selon le **State of FinOps 2026** de la FinOps Foundation, 98 % des organisations gèrent désormais activement leurs coûts d'IA, contre 63 % en 2025 et seulement 31 % en 2024. Mais la gouvernance peine à suivre : seules 28 % des organisations ont atteint une maturité d'optimisation automatisée.

La bonne nouvelle pour les équipes qui n'ont pas encore structuré leur démarche : les leviers classiques restent puissants. Combinés — engagements tarifaires (jusqu'à 72 % de remise en Reserved Instances, jusqu'à 90 % en Spot), rightsizing continu, nettoyage du stockage et automatisation du pilotage — ils permettent **typiquement de documenter une réduction de 30 à 40 % de la facture sur douze mois**, sans dégrader la performance ni la disponibilité.

29%

taux de gaspillage cloud moyen en 2026, en hausse pour la 1ère fois en 5 ans (Flexera)

17 M€

dépenses évitables par an pour une grande entreprise type au taux de gaspillage actuel

28%

seulement des organisations ont une gouvernance FinOps automatisée mature

Pour une organisation qui utilise AWS — startup en croissance, PME ou grande entreprise — la question n'est plus « faut-il faire du FinOps ? » mais « quels leviers actionner en premier, et comment en faire une pratique continue plutôt qu'un audit annuel ? ». Ce document propose une méthode pragmatique pour y répondre.

SECTION 02

Pourquoi le FinOps redevient un sujet de direction générale

Pendant cinq ans, le gaspillage cloud a régulièrement diminué à mesure que les entreprises structuraient leurs pratiques FinOps. Ce cycle s'est inversé en 2026, et trois évolutions expliquent pourquoi cette question remonte désormais jusqu'au comité de direction.

1. L'IA a changé l'échelle des coûts

Une instance équipée de GPU coûte 5 à 10 fois plus cher à l'heure qu'une instance CPU standard. Or la quasi-totalité des entreprises utilisent désormais l'IA générative sous une forme ou une autre, et 58 % des grandes organisations font tourner des charges d'IA générative en production — contre 50 % un an plus tôt. Les outils et processus FinOps conçus pour du calcul et du stockage classiques n'ont pas encore rattrapé cette nouvelle catégorie de dépense.

2. La visibilité reste le principal angle mort

Beaucoup d'équipes FinOps déclarent une visibilité limitée sur les moteurs réels de leurs coûts d'IA, ce qui rend difficile de mesurer le retour sur investissement ou d'identifier quelles équipes génèrent le plus de dépense. Sans cette visibilité, impossible de prioriser les leviers d'optimisation avec confiance.

3. Le fossé de gouvernance s'élargit

Seules 28 % des organisations ont atteint une gouvernance d'optimisation automatisée mature ; la majorité reste à un stade intermédiaire, sans contrôles spécifiques aux charges GPU. Résultat : le gaspillage progresse plus vite que la capacité des équipes à le détecter et le corriger manuellement.

« Le FinOps n'est plus un audit qu'on mène une fois par an — c'est une pratique continue, au même titre que la sécurité ou la fiabilité. »

Pour toute organisation dont la facture AWS progresse plus vite que l'usage réel, cela se traduit par une question simple à se poser dès aujourd'hui : **savez-vous précisément où va chaque euro dépensé, et qui, dans votre organisation, est responsable de le savoir ?**

SECTION 03

Les leviers structurels : rightsizing, Graviton, stockage

Avant de négocier le moindre engagement tarifaire, trois leviers structurels permettent souvent d'éliminer le gaspillage le plus évident.

LEVIER	CE QU'IL FAUT FAIRE CONCRÈTEMENT
Rightsizing continu	Surveiller en continu — pas lors d'un audit ponctuel — l'utilisation des instances. Une utilisation soutenue sous 40 % sur 14 jours est un signal fiable de sur-dimensionnement à corriger. Automatiser cette évaluation avec AWS Compute Optimizer plutôt que de la traiter manuellement.
Instances Graviton	Pour les charges Linux compatibles, les instances Graviton offrent un meilleur rapport prix/performance que leurs équivalents x86, souvent sans changement de code applicatif pour les langages de haut niveau.
Optimisation du stockage	Rétrograder les volumes EBS haute performance vers des types moins coûteux (gp3, st1, sc1) quand la performance ne le justifie pas, supprimer les volumes non attachés, et appliquer des politiques de cycle de vie sur les snapshots.

À retenir : ces trois leviers ne demandent aucun engagement contractuel ni changement d'architecture majeur — ils corrigent du gaspillage pur. C'est presque toujours le point de départ le plus rentable, avant même de négocier des remises.

Pourquoi le rightsizing ponctuel ne suffit plus

Un audit de rightsizing mené une fois par trimestre capture une photo instantanée qui se périmé en quelques semaines à mesure que les charges évoluent. Les organisations les plus matures traitent le rightsizing comme un processus continu, piloté par des seuils automatiques plutôt que par des revues manuelles — ce qui change fondamentalement le rythme auquel le gaspillage est détecté et corrigé.

SECTION 04

Les engagements tarifaires : RI, Savings Plans, Spot

Une fois le gaspillage évident éliminé, les engagements tarifaires constituent le levier le plus puissant — à condition de bien faire correspondre chaque option au bon type de charge.

OPTION	REMISE MAX.	FONCTIONNEMENT	À RÉSERVER À
Reserved Instances	72%	Engagement 1 ou 3 ans sur un type d'instance et une région ; paiement upfront total, partiel ou nul. Les RI convertibles permettent de changer de type d'instance.	Charges stables et prévisibles sur le long terme
Savings Plans	jusqu'à 72%	Engagement en euros/heure, applicable automatiquement à travers familles d'instances, régions et services éligibles — plus flexible que les RI classiques.	Charges stables mais dont la composition technique peut évoluer
Instances Spot	90%	Capacité inutilisée d'AWS, vendue à prix variable selon la demande ; AWS peut la reprendre avec un court préavis.	Charges tolérantes aux pannes : batch, CI/CD, rendu, clusters Kubernetes sans état

Piège fréquent : déployer des instances Spot sur des charges avec état sans mécanisme de reprise (statefulsets non protégés, files d'attente sans persistance) transforme une économie de 90 % en incident de production. Le Spot n'est rentable que combiné à une architecture réellement tolérante aux pannes.

Comment arbitrer entre les trois

Une approche pragmatique : couvrir le socle stable et prévisible (bases de données, services cœur de métier) avec des Savings Plans pour garder la flexibilité ; réserver les Reserved Instances classiques aux charges dont le type d'instance ne changera pas sur la durée de l'engagement ; et pousser vers le Spot tout ce qui est réellement interruptible — traitement par lots, environnements de build, workers stateless.

SECTION 05

Automatiser le pilotage : scheduling, autoscaling, outils

Les leviers structurels et les engagements tarifaires ne tiennent dans la durée que s'ils sont pilotés par des automatismes, pas par des revues manuelles.

Éteindre ce qui ne sert à rien la nuit

Arrêter les environnements de non-production (développement, recette, démo) en dehors des heures ouvrées divise quasiment par deux leur coût : stopper une instance 12 heures par jour réduit son coût de près de 50 %. Des outils comme AWS Instance Scheduler automatisent cette extinction sans intervention manuelle.

Faire évoluer l'autoscaling au-delà du CPU

De nombreuses politiques d'autoscaling ne surveillent que le CPU ou la mémoire, ce qui sous-dimensionne ou sur-dimensionne mal la capacité réelle nécessaire. Des métriques plus fidèles à l'usage réel — nombre de requêtes, latence, profondeur de file d'attente — produisent un dimensionnement plus précis et donc plus économe.

OUTIL AWS	RÔLE DANS LE PILOTAGE FINOPS
CloudWatch	Collecte des métriques d'utilisation qui alimentent les décisions de rightsizing et d'autoscaling
Cost Explorer	Analyse historique des dépenses par service, tag, compte — base de tout reporting FinOps
Compute Optimizer	Recommandations automatiques de rightsizing basées sur l'utilisation réelle observée
Cost Anomaly Detection	Alerte automatique dès qu'une dérive de coût sort des schémas habituels, sans attendre la facture mensuelle

À retenir : un outil de détection d'anomalie qui alerte une équipe en quelques heures vaut mieux qu'un rapport mensuel qui révèle la dérive trente jours trop tard.

SECTION 06

Exemple illustratif : combiner les leviers

Pour rendre ces leviers concrets, voici un exemple pédagogique — une entreprise fictive dépensant 100 000 € par mois sur AWS, sans démarche FinOps structurée.

LEVIER APPLIQUÉ	CONTRIBUTION ESTIMÉE
Savings Plans sur les charges stables identifiées	– 22 %
Instances Spot sur les charges batch et CI/CD	– 6 %
Rightsizing continu des instances sous-utilisées	– 5 %
Nettoyage du stockage (volumes orphelins, snapshots)	– 3 %
Extinction automatique des environnements non-prod	– 4 %

≈ 35%

réduction cumulée illustrative sur 100 000 €/mois, soit environ 35 000 €/mois

Important : cet exemple est illustratif et pédagogique — les contributions de chaque levier ne s'additionnent jamais mécaniquement dans un environnement réel, et les résultats varient fortement selon le profil de charge. Il illustre un ordre de grandeur atteignable en combinant plusieurs leviers, pas un engagement de résultat.

Ce qui rend ce scénario réaliste : aucun des cinq leviers ne dépend d'un investissement en infrastructure ou d'une refonte applicative. Ce sont des ajustements opérationnels et contractuels, activables en quelques semaines pour les trois premiers, et en continu pour les deux derniers.

SECTION 07

Grille d'auto-évaluation FinOps

Avant de lancer un chantier d'optimisation, commencez par situer votre maturité actuelle sur ces six questions.

1. Savez-vous précisément quelle part de votre facture est en on-demand vs engagée (RI/Savings Plans) ?

Sans cette visibilité de base, impossible de prioriser où négocier un engagement.

2. Le rightsizing est-il un processus continu, ou un audit ponctuel ?

Une revue trimestrielle capture une photo qui se périmé en quelques semaines.

3. Vos environnements non-production sont-ils arrêtés automatiquement en dehors des heures ouvrées ?

Si la réponse est non, c'est souvent le gain le plus rapide à obtenir.

4. Utilisez-vous des instances Spot sur vos charges réellement tolérantes aux pannes ?

Batch, CI/CD, workers sans état : ce sont les meilleurs candidats à 90 % de remise.

5. Connaissez-vous le coût de vos charges GPU/IA séparément du reste de votre facture ?

Sans cette séparation, impossible de mesurer le retour sur investissement de vos projets d'IA.

6. Une alerte automatique prévient-elle une équipe en cas de dérive de coût anormale ?

Découvrir une dérive sur la facture mensuelle signifie un mois de gaspillage non détecté.

Comment lire vos réponses : trois « non » ou plus indiquent généralement un potentiel d'économie proche des 30 à 40 % évoqués en synthèse. C'est le point de départ naturel d'un audit FinOps structuré.

SECTION 08

Feuille de route en 5 étapes

Une démarche progressive, pensée pour montrer des résultats dès les premières semaines.

- 1 Mesurer**
Mettre en place un système de tags cohérent (idéalement aligné sur le standard FOCUS) pour attribuer chaque euro dépensé à une équipe, un projet ou un workload.
- 2 Éliminer le gaspillage évident**
Supprimer les volumes non attachés, arrêter les ressources orphelines, corriger le sur-dimensionnement le plus flagrant identifié par Compute Optimizer.
- 3 Engager**
Couvrir les charges stables avec des Savings Plans, et migrer les charges tolérantes aux pannes vers le Spot.
- 4 Automatiser**
Déployer le scheduling automatique des environnements non-prod et faire évoluer l'autoscaling vers des métriques métier plutôt que le seul CPU.
- 5 Gouverner**
Mettre en place la détection d'anomalies, un showback ou chargeback par équipe, et une revue mensuelle courte plutôt qu'un audit annuel lourd.

CONCLUSION

Le FinOps, un réflexe plutôt qu'un projet

Les organisations qui traitent le FinOps comme une pratique continue — au même titre que la sécurité — documentent des économies durables de 30 à 40 %, sans compromis sur la performance ni la disponibilité.

Avec la montée des charges d'IA et de GPU, les enjeux ne vont faire que croître. Commencer par la visibilité et les leviers structurels, avant même de négocier le moindre engagement, reste la base la plus solide quelle que soit la maturité de départ.

Move2Cloud	est un cabinet de conseil spécialisé en infrastructure, cloud et transformation IT, partenaire AWS Advanced. Nous accompagnons les organisations dans l'audit, l'optimisation et la sécurisation de leur architecture cloud.
Contact	Move2Cloud — Paris · move2cloud.fr
Prochaine étape	Un audit FinOps initial permet de répondre en quelques semaines aux six questions de la page 9, avec une feuille de route priorisée et chiffrée.

Sources

ProsperOps, « Amazon EC2 Cost Optimization Guide for 2026 » — Flexera, rapport 2026 sur l'état du cloud (synthèse via algeriatech.news, « Flexera 2026: Cloud Waste Returns — AI GPU Is Why ») — FinOps Foundation, « State of FinOps 2026 Report » (synthèse via usu.com, « 6 Takeaways from the State of FinOps Report 2026 ») — documentation AWS (Savings Plans, Reserved Instances, Compute Optimizer, Cost Anomaly Detection).

Document rédigé à titre d'exemple et d'illustration par Move2Cloud à des fins pédagogiques. Les pourcentages d'économies présentés en section 06 sont illustratifs ; les résultats réels varient selon le profil de charge de chaque organisation. Ce document ne constitue pas un engagement contractuel de résultat.