



FINOPS & COÛTS

FinOps sur Azure optimiser coûts et réservations

Un guide opérationnel pour reprendre le contrôle de votre facture Azure : visibilité réelle des coûts, arbitrage entre réservations, savings plans et Spot, rightsizing et gouvernance durable.

11 pages · Matrice de décision engagements, modèle de tagging, grille d'audit et feuille de route incluses

72%

d'économie maximale annoncée par Microsoft pour une réservation Azure de 3 ans

65%

d'économie maximale pour un Azure Savings Plan for Compute sur 3 ans

90%

d'économie potentielle sur les VM Spot, pour des charges interruptibles

SOMMAIRE

Ce que vous allez trouver dans ce document

Ce document part du contexte économique du cloud Azure, détaille les mécanismes techniques d'optimisation et d'engagement, puis propose une feuille de route et une checklist directement applicables.

01	Synthèse — l'essentiel en une page	p.3
02	Pourquoi le FinOps Azure s'impose	p.4
03	Visibilité, tagging et allocation	p.5
04	Réservations, Savings Plans et Spot	p.6
05	Rightsizing, élasticité et PaaS	p.7
06	Gouvernance et culture FinOps	p.8
07	Grille d'auto-évaluation	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 ? CTO, DSI, responsables plateforme et FinOps, architectes cloud et équipes finance/achats IT qui gèrent une empreinte Azure significative (de quelques dizaines à plusieurs milliers de ressources) et cherchent à réduire durablement la facture sans freiner les équipes produit.

SYNTHÈSE

L'essentiel en une page

Reprendre le contrôle d'une facture Azure ne relève pas de la négociation commerciale, mais d'une discipline d'ingénierie : mesurer, arbitrer, engager, puis recommencer.

La dépense cloud est passée du statut de ligne budgétaire variable à celui de coût structurel majeur. Les enquêtes annuelles Flexera placent depuis plusieurs années l'optimisation des coûts cloud comme **priorité n°1** des organisations interrogées, lesquelles estiment gaspiller **environ 30 %** de leur dépense. Côté Azure, Microsoft affiche des remises pouvant atteindre **72 %** sur une réservation 3 ans face au pay-as-you-go, et jusqu'à **90 %** sur les machines virtuelles Spot : l'écart entre une empreinte pilotée et une empreinte subie se chiffre en millions d'euros sur trois ans pour une DSI de taille intermédiaire.

Cette tension s'aggrave pour deux raisons. D'abord, la grille tarifaire Azure est devenue combinatoire : réservations, savings plans compute, Spot, Azure Hybrid Benefit, tarifs Dev/Test et remises contractuelles MCA/CSP se recouvrent partiellement, et un engagement mal ciblé peut rester sous-utilisé pendant trente-six mois. Ensuite, la montée des charges data et IA — GPU, stockage chaud, trafic inter-régions — déplace la base de coûts vers des ressources chères, peu élastiques et mal couvertes par les réflexes FinOps historiques centrés sur la VM généraliste.

La bonne nouvelle est que ces leviers sont largement mécanisables. Une allocation de coûts fiable (tags, management groups, périmètres de facturation), un rightsizing piloté par les métriques réelles et une couverture d'engagements construite sur le socle stable — et non sur le pic — produisent des gains rapides sans ralentir les équipes produit. Ce document fournit la méthode de bout en bout : modèle de visibilité, règles d'arbitrage entre réservations, savings plans et Spot, seuils de décision chiffrés et rituels de gouvernance pour que les économies obtenues ne s'érodent pas au trimestre suivant.

72 %

remise maximale annoncée par Microsoft pour une réservation 3 ans face au pay-as-you-go

≈ 30 %

part de la dépense cloud jugée gaspillée par les répondants de l'enquête Flexera

5

leviers tarifaires Azure à arbitrer : RI, savings plan, Spot, Hybrid Benefit, Dev/Test

La question que traite ce document n'est donc pas « comment payer Azure moins cher ce trimestre », mais « comment construire une structure de coûts qui reste optimale alors que l'infrastructure, elle, change en permanence ». Les sections qui suivent proposent une méthode en huit étapes — du modèle de visibilité jusqu'à la gouvernance des engagements — assortie d'une checklist opérationnelle directement applicable à votre tenant.

SECTION 02

Pourquoi le FinOps Azure s'impose

Le FinOps Azure n'est pas une mode de gouvernance : c'est la réponse à trois dynamiques simultanées — une dépense devenue structurelle, une tarification de plus en plus combinatoire et des décisions de coût prises par les équipes d'ingénierie. Voici pourquoi le sujet ne peut plus attendre le prochain exercice budgétaire.

1. Le cloud est devenu un poste de coût structurel

La dépense cloud n'est plus un budget projet mais un poste de coût structurel, scruté au même titre que la masse salariale. Dans les enquêtes annuelles Flexera, l'optimisation des coûts cloud reste la priorité n°1 des répondants, qui estiment gaspiller près d'un tiers de leur dépense.

2. Une grille tarifaire devenue multi-dimensionnelle

Azure superpose désormais au moins cinq leviers tarifaires : réservations 1 ou 3 ans, savings plans compute, Spot, Azure Hybrid Benefit et tarifs Dev/Test. Avec des remises annoncées jusqu'à 72 % (RI 3 ans) et 90 % (Spot), l'arbitrage n'a plus rien d'intuitif.

3. Le coût se décide dans le code, pas dans le tableau

L'essentiel de la facture se décide en amont, dans le code et les templates IaC : famille de SKU, niveau de redondance, tier de stockage, trafic inter-régions. Sans boucle de retour vers les équipes produit, le FinOps reste un exercice comptable a posteriori.

« Sur Azure, la remise ne se négocie pas une fois par an : elle se conçoit, se mesure et se pilote chaque mois. »

Les sections suivantes détaillent la mécanique : visibilité réelle des coûts, arbitrage entre réservations, savings plans et Spot, rightsizing, puis gouvernance durable.

SECTION 03

Visibilité, tagging et allocation

Sur une facture Azure non gouvernée, il est courant que 20 à 40 % de la dépense ne soit rattachable à aucune application identifiée : tant que ce chiffre n'est pas proche de zéro, toute optimisation reste une conversation d'opinion.

PRATIQUE	CE QU'ELLE APPORTE
Hiérarchie de management groups	Structurer les abonnements par entité, environnement et criticité donne un premier axe d'allocation natif, indépendant de la qualité du tagging, et le point d'application naturel des politiques.
Azure Policy en mode Modify + remédiation	Plutôt que d'auditer, on applique : héritage automatique des tags du resource group vers les ressources, et refus de création pour les tags obligatoires manquants, avec tâches de remédiation sur l'existant.
Export quotidien des données de coût	Un export planifié des coûts amortis vers un compte de stockage, puis leur exploitation dans Power BI ou un entrepôt, lève les limites de l'interface Cost Analysis et permet de croiser coûts, tags et métriques métier.

À retenir : Figez une taxonomie courte — 5 à 6 tags maximum (owner, application, environnement, centre de coût, criticité) — rendez-la obligatoire par Azure Policy au niveau du resource group, et basez tous vos reportings sur le coût amorti, jamais sur le coût facturé brut.

Le cas des coûts partagés : AKS, réseau et observabilité

Les coûts partagés sont le point de blocage de 90 % des démarches d'allocation : un cluster AKS mutualisé, une passerelle ExpressRoute, un espace de travail Log Analytics ou un Application Gateway portent des coûts réels qu'aucun tag de ressource ne permet de ventiler. Deux stratégies coexistent. La première, pragmatique, consiste à isoler physiquement : un resource group — voire un abonnement — par produit, quitte à perdre un peu de mutualisation. La seconde consiste à répartir au prorata d'un usage mesuré : sur AKS, les métriques de requests/limits par namespace (via OpenCost ou les solutions de cost allocation Kubernetes) donnent une clé de répartition défendable ; sur Log Analytics, la table `Usage` fournit le volume ingéré par type de données et par source. L'essentiel est de documenter la clé, de l'accepter collectivement et de ne plus la rediscuter chaque mois : une clé imparfaite mais stable vaut mieux qu'une clé exacte mais contestée.

SECTION 04

Réservations, Savings Plans et Spot

Sur un parc de calcul stable, le seul fait de ne rien engager coûte structurellement 30 à 60 % de plus que le prix accessible : le tarif à la demande est, sur Azure, le tarif le plus cher du catalogue.

MÉCANISME	REMISE INDICATIVE	FONCTIONNEMENT	À RÉSERVER À
Reserved Instances	Jusqu'à ~72 % sur 3 ans	Engagement sur une famille de VM dans une région donnée, 1 ou 3 ans ; flexibilité d'instance au sein de la série, échange possible sous conditions.	Socle stable et architecturalement figé : bases de données, VM applicatives de production, charges legacy peu susceptibles de changer de famille.
Savings Plan for compute	Jusqu'à ~65 % sur 3 ans	Engagement sur un montant horaire en dollars, appliqué automatiquement aux ressources de calcul éligibles, toutes familles et régions confondues (VM, App Service, Container Instances, Functions Premium).	Parcs de calcul en mouvement : modernisation en cours, multi-régions, mix IaaS/PaaS, arbitrages de familles fréquents.
Machines virtuelles Spot	Jusqu'à ~90 %	Capacité excédentaire d'Azure, récupérable par la plateforme avec un préavis de 30 secondes, avec politique d'éviction par prix ou par capacité.	Batch, CI/CD, rendu, entraînement de modèles, pools de nœuds AKS interruptibles — tout ce qui est stateless, reprenable et sans SLA de disponibilité.

Piège fréquent : Acheter des réservations avant d'avoir fait le rightsizing revient à verrouiller son gaspillage pour trois ans. Rightsizez d'abord, stabilisez deux à trois mois, engagez ensuite — et n'engagez jamais sur la base d'un pic saisonnier ou d'un environnement de pré-production temporaire.

Dimensionner l'engagement : raisonner en couverture, pas en volume

La bonne question n'est pas « combien de réservations acheter » mais « quel pourcentage de ma consommation horaire est réellement incompressible ». Tracez la courbe horaire de votre consommation de calcul sur 90 jours : la base plate du bas — celle que vous consommez 24h/24, 365 jours par an — est le périmètre engageable. L'usage courant consiste à couvrir 70 à 85 % de cette base, en laissant délibérément une marge non engagée pour absorber les migrations, les décommissionnements et les changements d'architecture. Portez ensuite les engagements à la portée « partagé » (shared scope) pour qu'ils s'appliquent automatiquement à l'ensemble des abonnements de l'accord de facturation, et suivez deux KPI mensuels distincts : le taux de couverture (part de la consommation bénéficiant d'une remise) et le taux d'utilisation des engagements (part de l'engagement effectivement consommée). Un taux d'utilisation inférieur à 98 % signale un engagement mal dimensionné ou une ressource éteinte sans que l'engagement ait été réaffecté.

SECTION 05

Rightsizing, élasticité et PaaS

Azure Advisor ne remonte par défaut aucune recommandation fondée sur la mémoire : sans agent de collecte configuré, le rightsizing s'appuie sur le seul CPU et passe à côté de l'essentiel.

Rightsizing piloté par les métriques, pas par l'intuition

Raisonnez au P95 sur 30 jours, pas à la moyenne : une VM dont le P95 CPU plafonne à 15 % et la mémoire à 40 % est surdimensionnée de deux crans. Activez la collecte mémoire via Azure Monitor Agent avant toute campagne.

Changer de famille avant de changer de taille

Les générations récentes (Dv5/Ev5, ou Dpsv5/Epsv5 sur Arm Ampere) offrent un meilleur rapport performance/prix que les Dv3/Ev3 encore massivement déployées. Une migration de famille est souvent plus rentable qu'un cran de taille en moins.

Élasticité : la ressource la moins chère est celle qui est éteinte

Un environnement de dev/test arrêté la nuit et le week-end ne tourne que 40 à 50 heures sur 168. Automatisez l'extinction (Start/Stop VMs, Automation, tags d'horaires) et activez l'autoscale AKS avec un node pool Spot pour les charges interruptibles.

LEVIER	EFFET ATTENDU
Extinction hors heures ouvrées (dev/test)	Jusqu'à ~70 % d'heures de calcul en moins sur les environnements non productifs, sans aucun impact sur la production.
Lifecycle management du stockage Blob	Bascule automatique Hot → Cool → Cold → Archive selon l'âge : le coût au Go est divisé par un facteur significatif sur les données froides.

SECTION 06

Gouvernance et culture FinOps

Le FinOps Framework structure la démarche en trois phases itératives — Inform, Optimize, Operate — et le piège le plus fréquent consiste à vouloir optimiser avant d'avoir informé.

Le rituel mensuel et la chaîne de responsabilité

Une revue mensuelle de 45 minutes, réunissant les responsables produit, la plateforme et le contrôle de gestion, suffit si elle s'appuie sur un jeu de KPI stable : coût amorti par produit, coût unitaire (par transaction, par client actif, par environnement), taux de couverture et taux d'utilisation des engagements, et volume de gaspillage identifié non traité. Le showback — montrer à chaque équipe ce qu'elle consomme — précède toujours le chargeback : refacturer avant d'avoir construit la confiance dans les chiffres garantit des mois de contestation sur la méthode d'allocation plutôt que des actions d'optimisation. Chaque écart significatif doit se traduire par un ticket nommé dans un backlog, avec un porteur et une échéance, sinon la revue devient un exercice de constat.

Garde-fous automatisés et coût dans la pull request

La gouvernance durable repose sur des garde-fous automatisés plutôt que sur la vigilance humaine. Trois mécanismes couvrent l'essentiel : des budgets Azure assortis d'alertes à 80 % et 100 % du prévisionnel, routées vers l'équipe propriétaire et non vers une boîte générique ; des Azure Policy restreignant les SKU, régions et tailles de VM autorisées par environnement, qui empêchent structurellement le déploiement d'une GPU en recette ; et la détection d'anomalies de Cost Management, qui signale les ruptures de tendance avant la fin du mois. À cela s'ajoute le décalage à gauche : estimer le coût d'un changement d'infrastructure directement dans la pull request, via un outil d'estimation sur le code Terraform, transforme l'arbitrage financier en décision d'ingénierie prise au bon moment — avant le déploiement, pas trente jours après la facture.

À retenir : Nommez un cost owner par produit, pas par équipe d'infrastructure : tant que le coût n'apparaît pas dans la revue de backlog du produit qui le génère, il restera le problème de quelqu'un d'autre.

Une démarche FinOps ne produit pas un gain unique mais une trajectoire : les premiers mois dégagent les économies évidentes — ressources orphelines, environnements oubliés, surdimensionnements grossiers —, les suivants installent les engagements, et c'est seulement à partir du troisième ou quatrième cycle que le coût unitaire devient un indicateur piloté au même titre que la latence ou le taux d'erreur. Ce qui distingue les organisations qui tiennent la trajectoire de celles qui retombent au bout de six mois tient rarement à l'outillage : c'est la régularité du rituel et la clarté de la responsabilité. Mesurer, arbitrer, engager, puis recommencer.

SECTION 07

Grille d'auto-évaluation

Passez ces six questions en revue avec vos équipes plateforme et finance avant d'engager le moindre budget sur des réservations ou des savings plans.

- ☐ **Disposez-vous d'une taxonomie de tags formalisée (5 à 6 clés maximum) et rendue obligatoire par Azure Policy au niveau des resource groups ?** Sans tags fiables et imposés à la création, aucune allocation de coût par application ou par équipe n'est exploitable.
- ☐ **Vos reportings de coûts s'appuient-ils sur le coût amorti plutôt que sur le coût facturé brut ?** Le coût facturé concentre les achats de réservations sur un seul mois et fausse toute comparaison de tendance.
- ☐ **Avez-vous configuré la collecte des métriques mémoire pour alimenter les recommandations de rightsizing ?** Sans agent de collecte, Azure Advisor ne raisonne que sur le CPU et laisse passer la majorité des surdimensionnements réels.
- ☐ **Avez-vous stabilisé vos dimensionnements pendant deux à trois mois avant tout achat de réservation ou de savings plan ?** Engager sur un parc non optimisé revient à verrouiller contractuellement son gaspillage pour un à trois ans.
- ☐ **Suivez-vous mensuellement votre taux de couverture et votre taux d'utilisation des engagements existants ?** Une réservation sous-utilisée coûte plus cher que le pay-as-you-go qu'elle était censée remplacer.
- ☐ **Existe-t-il un rituel récurrent (mensuel ou par sprint) où les coûts sont arbitrés avec un responsable nommé par application ?** La régularité du rituel et la clarté de la responsabilité distinguent les démarches qui tiennent dans la durée de celles qui s'essoufflent.

Comment lire vos réponses : Trois « non » ou plus indiquent qu'il faut consolider la visibilité et le rightsizing avant toute stratégie d'engagement.

SECTION 08

Feuille de route en 5 étapes

Cette trajectoire se déroule sans interrompre la production : chaque étape produit un gain mesurable et crée les conditions de la suivante, plutôt que d'imposer une transformation en bloc.

1

Établir la visibilité

Déployez la taxonomie de tags par Azure Policy, reconstruisez les reportings sur le coût amorti et établissez une répartition claire entre coûts directement attribuables et coûts partagés. Objectif : qu'aucune ligne significative de facture ne reste orpheline.

2

Éliminer le gaspillage évident

Supprimez les disques non attachés, les adresses IP publiques réservées sans usage, les snapshots anciens et les environnements de test oubliés. Ces gains immédiats financent la suite de la démarche et crédibilisent l'approche auprès des équipes.

3

Rightsizer les ressources

Activez la collecte des métriques mémoire, croisez-les avec le CPU et les IOPS, puis redimensionnez progressivement en commençant par les environnements hors production. Introduisez l'arrêt planifié et l'autoscaling là où la charge le permet.

4

Engager par paliers

Après deux à trois mois de consommation stabilisée, couvrez le socle incompressible par des réservations, la part variable par des savings plans, et basculez les charges tolérantes à l'interruption sur Spot. Engagez par paliers plutôt qu'en une fois.

5

Ancrer la gouvernance

Instaurez un rituel mensuel d'arbitrage avec un responsable nommé par application, des alertes de budget par environnement et un suivi du coût unitaire rapporté à l'usage métier. Réévaluez couverture et utilisation à chaque cycle.

CONCLUSION

Le coût unitaire, une métrique d'ingénierie comme une autre

Les organisations qui traitent le coût cloud comme une métrique d'ingénierie — mesurée, attribuée, arbitrée en continu — obtiennent bien plus qu'une baisse ponctuelle de facture : elles gagnent la capacité de corréler une décision d'architecture à son effet économique, et donc d'arbitrer en connaissance de cause entre performance, résilience et dépense.

La difficulté du FinOps sur Azure n'est pas technique : les outils de visibilité, les leviers d'engagement et les mécanismes d'élasticité existent et sont documentés. Elle est organisationnelle. Tant que la facture reste un sujet de fin de mois traité par la direction financière sans les équipes qui provisionnent les ressources, les mêmes dérives réapparaîtront. Remettre la décision de dépense au plus près de celle de conception, avec des données fiables et un rituel court, est l'investissement qui paie le plus durablement.

Move2Cloud	est un cabinet de conseil spécialisé en infrastructure, cloud et transformation IT. Nous accompagnons les organisations dans l'audit, l'optimisation financière et la gouvernance de leur environnement cloud.
Contact	Move2Cloud — Paris · move2cloud.fr
Prochaine étape	Un audit FinOps de quelques jours — revue de la taxonomie de tags, analyse du coût amorti sur douze mois, inventaire des ressources orphelines et diagnostic de couverture des engagements — suffit généralement à transformer les « non » de la checklist ci-dessus en un plan d'action chiffré.

Sources

FinOps Foundation — FinOps Framework et State of FinOps (édition annuelle) — Microsoft — Azure Well-Architected Framework, pilier Cost Optimization — Microsoft Learn — Documentation Microsoft Cost Management, réservations et savings plans Azure — Microsoft Learn — Cloud Adoption Framework for Azure, gouvernance et conventions de nommage et de balisage — Flexera — State of the Cloud Report (édition annuelle)

Document rédigé à titre d'exemple et d'illustration par Move2Cloud à des fins pédagogiques. Les mécanismes de tarification Azure, les familles de machines et les conditions d'engagement évoluent régulièrement : vérifiez les conditions en vigueur avant tout achat.