turbonomic
an IBM Company

LIVRE BLANC

Automatiser l'élasticité des conteneurs pilotés par les applications

Pour les ingénieurs de plateformes et DevOps qui souhaitent opérationnaliser la vitesse de commercialisation tout en assurant les performances applicatives

Résumé

Votre avantage concurrentiel dépend de la vitesse de transformation des idées en transactions commerciales et de leurs performances pour vos clients. La technologie en est le moteur.

Les conteneurs offrent la vitesse, l'agilité, l'élasticité et l'évolutivité qui modifient radicalement notre mode de construction, de déploiement et d'exécution des applications. Ils ouvrent la voie à un monde où les applications peuvent réellement s'exécuter n'importe où, où les mises à jour et les nouvelles capacités peuvent se déployer en production plusieurs fois par jour, et où la demande de charge de travail fluctuante dynamique peut se gérer grâce à une infrastructure élastique, n'importe où et n'importe quand. Kubernetes est une plateforme qui permet aux entreprises d'être agiles et élastiques, sans toutefois gérer les compromis sur la manière d'assurer les performances, tout en étant efficace.

Malgré la simplicité et l'agilité de la conteneurisation, la plateforme d'orchestration ne fournit qu'un moyen de gérer le cycle de vie de ces services : en déployant et en maintenant vos services de la façon dont vous les décrivez.

Les plateformes de conteneurs ne garantissent pas automatiquement que les services répondent aux objectifs de niveau de service (SLO) et ne sont pas non plus en mesure de gérer dynamiquement les ressources.

Les règles basées sur des seuils ne résolvent pas les problèmes de performances. Cette approche n'a jamais fonctionné, et pour la vitesse de changement des plateformes de conteneurs, la mise à l'échelle automatique déclenchée sans corrélation peut en réalité finir par entraîner des problèmes. Une infrastructure élastique est essentielle pour obtenir des performances, mais nécessite une analyse automatisée qui gère en permanence la demande, l'offre et les contraintes afin de répondre aux SLO souhaités.

Ce livre blanc examine les principaux concepts à envisager pour l'adoption des plateformes de conteneurs comme moyen de diriger une entreprise, et comment protéger cet investissement grâce à une automatisation qui assure des performances tout en minimisant les coûts et en étant en conformité. Il explique également pourquoi une analyse descendante est nécessaire à l'exécution de vos services par une plateforme Kubernetes qui s'autogère. Bâtir un environnement évolutif multicloud dès le départ permet à votre service informatique de développer une « mémoire musculaire » opérationnelle qui transformera radicalement la manière dont (et le moment où) vous apporterez davantage d'innovations envisagées par vos secteurs d'activité.

© 2021 Turbonomic, an IBM Company

Une promesse de vitesse, d'agilité, d'élasticité et d'évolutivité

Kubernetes permet l'élasticité, mais ne garantit pas automatiquement l'atteinte des SLO des applications.

Le succès de l'adoption de la conteneurisation dépend de la capacité des développeurs à bénéficier de l'agilité qu'ils recherchent, de l'élasticité qu'il leur faut pour s'adapter à grande échelle à l'évolution constante des demandes et pour garantir le fonctionnement de l'application à la vitesse requise.

L'adoption d'une approche cloud native et la décomposition de vos applications en ensembles de services distincts peuvent favoriser un développement et un déploiement d'applications plus agiles. Les conteneurs fournissent le conditionnement qui rend vos services portables et évolutifs. Kubernetes fournit un cadre et des points de contrôle pour exécuter vos applications et services numériques. Mais pour fournir une plateforme performante à l'échelle de votre entreprise, il vous faut encore ajouter des capacités pour exploiter l'élasticité offerte par la plateforme pour atteindre les SLO des applications.

Déploiement accéléré et CI/CD - boucle de rétroaction de production

Pour accélérer la mise sur le marché, il est essentiel d'adopter la bonne approche CI/CD (intégration continue/déploiement continu) basée sur l'automatisation. Dans le rapport State of DevOps 2018 de DORA, les personnes interrogées ont cité des améliorations significatives résultant de la mise en œuvre de l'approche CI/CD :

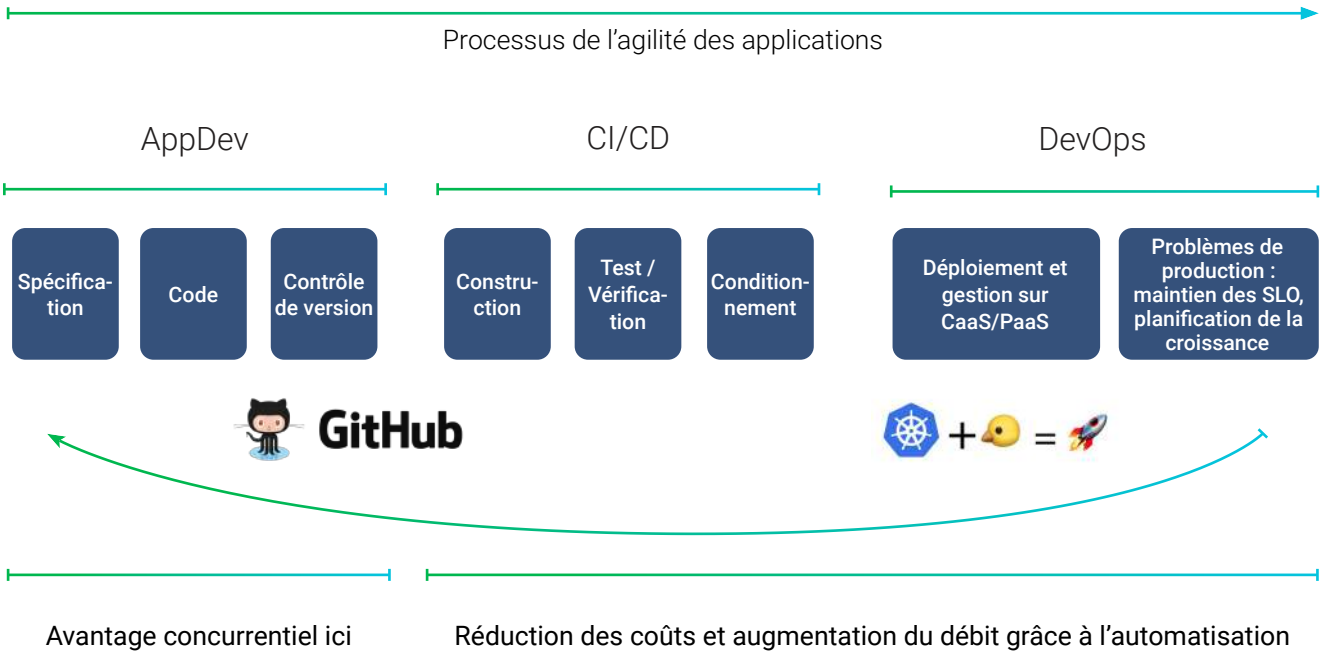
Fréquence de déploiement	Hebdomadaire-mensuel	→	Horaire-quotidien
Variation du délai de mise en œuvre	1 à 6 mois	→	1 à 7 jours
Variation du taux d'échec	46 à 60 %	→	0 à 15 %

La vitesse s'accompagne du besoin de disposer d'un moyen de gérer l'évolution constante de la production ainsi que d'une boucle de rétroaction sur les performances de vos services et les prévisions concernant les besoins de l'infrastructure. L'objectif est de disposer d'un moyen de définir vos SLO et de faire en sorte que la plateforme vous explique comment configurer vos conteneurs et votre infrastructure afin de réduire le risque de problèmes de performances.

- **Qui décide comment les ressources devraient être allouées aux services ? Comment cette décision est-elle prise : test de résistance, analyse comparative par rapport à des SLO établis, etc. ?**
- **Comment mesurer les performances ? Votre pipeline CI/CD présente-t-il une boucle de rétroaction pour assurer la configuration adéquate des conteneurs et des pods ?**
- **Comment garantir une capacité toujours suffisante pour les nouveaux déploiements ?**

© 2021 Turbonomic, an IBM Company

Options	Limites	Réponse de Turbonomic
Analyse manuelle des données d'utilisation des conteneurs/pods pour déterminer les spécifications des ressources	• Configuration de la collecte des données • Main-d'œuvre pour l'analyse	• Analyse descendante pilotée par les applications déterminant le dimensionnement de vos conteneurs • Boucle de rétroaction dans le pipeline CI/CD • Opportunités de réduction des demandes lorsqu'elles ne sont pas requises
Analyse manuelle des données des ressources à partir de tous les points de la pile pour déterminer la capacité de production	• Main-d'œuvre pour la collecte des données en provenance de plusieurs sources • Main-d'œuvre pour l'analyse	Analyse basée sur l'utilisation pour identifier les besoins en ressources dans l'ensemble de la pile



Plateforme et infrastructure

Pourquoi il vous faut une gestion de la pile complète pilotée par les applications

Quel que soit votre choix de la technologie d'orchestration de conteneurs (PCF ou OpenShift/Kubernetes) et/ou de l'infrastructure sous-jacente (cloud privé, cloud public, cloud hybride, multicloud et/ou sans système d'exploitation), les défis opérationnels de votre PaaS restent les mêmes :

- **Comment déterminer si la capacité est suffisante pour répondre à la demande actuelle et évolutive ?**
- **Comment décider quand exécuter davantage de nœuds d'application ?**
- **Comment décider quand suspendre l'activité ?**
- **Comment gérer les pics de demande ?**
- **Comment tirer parti des ressources du cloud public pour le cloud bursting ?**
- **Comment garantir haute disponibilité et résilience dans l'ensemble de la pile ?**
- **Comment appliquer les contraintes de l'entreprise ?**

L'élasticité offerte par les plateformes de conteneurs permet de fournir la somme des demandes moyennes de vos applications au lieu de la somme de leurs pics de demande. Pour en tirer parti et réaliser les gains potentiels, il convient de disposer d'une plateforme de contrôle piloté par les applications et descendant qui assure que les applications bénéficient des ressources dont elles ont besoin au moment où elles en ont besoin pour fonctionner. Autrement dit, une plateforme qui adapte en permanence ces ressources aux fluctuations de la demande.

© 2021 Turbonomic, an IBM Company

Options	Limites	Réponse de Turbonomic
Exploitation de prestataires de services fournissant des groupes de mise à l'échelle automatique (ASG, ensembles de disponibilité, etc.)	• Règles basées sur des seuils • Mise à l'échelle impossible d'un nœud spécifique : tous les nœuds doivent être identiques (mêmes contraintes, labels de nœuds, etc.)	• SLO pilotés par les applications et descendants • Adaptation continue des ressources de l'infrastructure pour répondre à la demande d'applications • Adaptation continue, verticale/horizontale, des conteneurs, des pods et des nœuds appropriés • Placement continu des pods dans les nœuds appropriés
Analyse des données des ressources à partir de tous les points de la pile pour déterminer la capacité de production	• Main-d'œuvre pour la collecte des données en provenance de plusieurs sources • Main-d'œuvre pour l'analyse	• Analyse basée sur l'utilisation pour identifier les besoins en ressources dans l'ensemble de la pile • Adaptation continue, verticale/horizontale, des conteneurs, des pods et des nœuds appropriés • Déclenchement continu d'actions pour éviter les goulots d'étranglement

Poursuite des SLO à grande échelle

L'objectif d'une plateforme de conteneurs est d'exécuter vos applications au niveau de service souhaité pour votre entreprise. Il faut assurer en permanence les performances à mesure que le nombre d'applications augmente. En général, nous constatons chez nos clients que cela prend plus de 12 mois pour les 3 premières applications. Pour les applications ultérieures, avec l'avantage des compétences acquises et des meilleures pratiques, cela peut prendre 6 à 12 mois de plus. Lorsque les secteurs d'activité découvrent les possibilités, l'ampleur du nombre de services individuels à gérer dépasse les capacités de gestion de l'homme. Peut-être avez-vous construit des services sans état et peut-être considérez-vous les conteneurs comme du bétail et non comme des animaux domestiques, mais quelle est votre tolérance à la dégradation des performances dans l'expérience de votre utilisateur final ? Que pouvez-vous faire pour gérer non seulement la demande, mais aussi l'accélération du changement ? La réponse réside dans l'automatisation, à travers des actions fondées sur une analyse des compromis sur le nombre d'instances de votre service nécessaires à l'atteinte de vos SLO, la configuration de votre charge de travail (taille et placement), et la mise à disposition de ressources conformes depuis l'infrastructure.

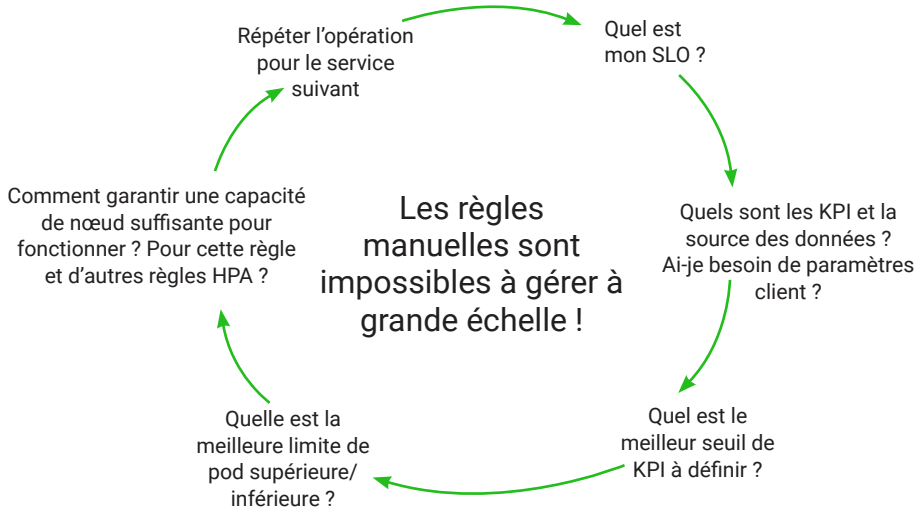
© 2021 Turbonomic, an IBM Company

Les seuils ne résolvent pas le problème

Une plateforme de conteneurs vous garantira un nombre minimal de services disponibles ; si l'un tombe en panne, elle tentera de le rétablir. Mais si vous voulez garantir une bonne expérience utilisateur, il faut que le système réponde avant que la dégradation des performances et qu'une panne se produisent. Vous pouvez toujours envisager de définir une mise à l'échelle automatique horizontale native pour répondre à la demande, mais il vous faut décider quel ou quels paramètres expriment le mieux les ressources nécessaires, configurer des seuils et des limites supérieure/inférieure, tester et extrapoler si cela fonctionnera sous la demande de production, puis répéter l'opération pour chaque service déployé. Imaginez que vous disposiez de plus de 100 services pour une seule application. Chacune de ces règles n'a aucune corrélation entre elles. Comment veillez-vous à ce que l'ajout de pods supplémentaires d'un service n'introduise pas de congestion dans un autre domaine ? Clonez-vous un pod qui a mal été configuré et qui a d'abord besoin d'une mise à l'échelle verticale ? Comment gérez-vous la congestion des nœuds, répondez-vous aux voisins bruyants et identifiez-vous les ressources allouées inutilisées qui pourraient être débloquées pour répondre à cette demande ?

En outre, la configuration de vos conteneurs, pods et règles de mise à l'échelle automatique HPA ou cluster n'est pas un exercice qui se fait en une fois. Les meilleurs efforts doivent être sans cesse surveillés et redéfinis s'ils ne le sont pas. Que pourraient faire vos équipes du temps gagné si elles n'avaient pas à définir et à redéfinir manuellement ces seuils ?

L'importance de ne pas se tromper avec ces configurations a une incidence directe sur le succès du déploiement de votre stratégie de transformation numérique. Une poignée de mauvais déploiements peut ralentir considérablement l'adoption des plateformes et des systèmes en cours de développement. Et trop de temps et de main-d'œuvre consacrés à la configuration manuelle de ces points de contrôle peuvent considérablement limiter la capacité de votre entreprise à privilégier une approche axée sur les plateformes. Votre entreprise peut-elle se permettre ce retard ? Il vous faut un système de contrôle qui soit capable de gérer les compromis dans toutes les ressources, de définir les limites et les demandes de mise à l'échelle verticale des conteneurs, le nombre de pods nécessaires et les décisions de placement à prendre pour la redistribution des pods, ainsi que de gérer les ressources du cluster à l'aide d'un moteur d'analyse unique.



© 2021 Turbonomic, an IBM Company

Options	Limites	Réponse de Turbonomic
HPA (Horizontal Pod Autoscaler) : règle basée sur un seuil pour exécuter la mise à l'échelle évolutive et horizontale des pods	• Configuration par service • En fonction de la moyenne de tous les pods du service • Définition manuelle des KPI et des seuils, ainsi que des limites de pods supérieure/inférieure	• SLO pilotés par les applications et descendants • Exploitation des données de temps de réponse pour favoriser la mise à l'échelle horizontale des services afin de répondre aux SLO • Adaptation continue, verticale/horizontale, des conteneurs, des pods et des nœuds appropriés • Placement continu des pods dans les nœuds appropriés • Adaptation continue des ressources de l'infrastructure pour répondre à la demande d'applications
VPA (Vertical Pod Autoscaler) : règle basée sur un seuil pour exécuter la mise à l'échelle verticale des conteneurs	• Définition obligatoire pour chaque service • Projet bêta (utilisation à vos risques et périls) • Pas d'accès à la capacité du nœud pour prendre des mesures	
Laisser les pods tomber en panne pour qu'ils se redéploient sur un meilleur nœud	• Mauvaise expérience utilisateur dans les transactions sur un pod qui s'apprête à tomber en panne	
Prometheus : solutions d'observabilité, collecte et consolidation des données	• Analyse des données indisponible • Actions indisponibles	

© 2021 Turbonomic, an IBM Company

Une approche pilotée par les applications

Les SLO des applications doivent favoriser l'infrastructure

La conteneurisation des applications essentielles est un investissement qui présente de nombreux avantages. Mais pour tirer pleinement parti de ces avantages de vitesse, d'élasticité et de portabilité, il vous faut des logiciels qui vous permettent de prendre les bonnes décisions en matière de ressources au bon moment, 24 heures sur 24, 7 jours sur 7, tous les jours de l'année. Sinon, la complexité vous ralentira.

Seul Turbonomic relie vos applications essentielles à la plateforme Kubernetes et à l'infrastructure sous-jacente, où que s'exécutent vos applications. En fonction de la demande d'applications en temps réel et de la prise en compte des contraintes et des interdépendances à chaque couche de la pile (de la logique à la physique), le logiciel détermine les actions appropriées au bon moment pour garantir que les applications bénéficient toujours exactement de ce dont elles ont besoin pour fonctionner. Pour s'exécuter en temps réel, selon une planification ou dans le cadre de votre pipeline DevOps.

Dimensionnement intelligent : comment dimensionner les conteneurs ?

- Automatisation associée au déploiement. Exécution et poursuite du redimensionnement dans le cadre du pipeline (par exemple, YAML, Jenkins, etc.).
- Automatisation en temps réel. Exécution dynamique via Kubernetes.

Placement continu : quand déplacer les pods ? Vers quels nœuds ?

- Exécution dynamique en temps réel via Kubernetes. Uniquement pour les services sans état (sans interruption).

Mise à l'échelle dynamique : quand exécuter la mise à l'échelle horizontale (ou verticale) du cluster ? De combien ?

- Exécution dynamique en temps réel de la mise à l'échelle du cluster via l'infrastructure en tant que code (IaC) ou l'API du cluster Kubernetes.

Mise à l'échelle définie par les SLO : quand exécuter la mise à l'échelle horizontale (ou verticale) des pods pour répondre aux SLO en matière de temps de réponse des applications ? De combien ?

Conditions préalables à la mise à l'échelle définie par les SLO :

- Les applications sont conçues pour des microservices sans état horizontaux.
- Elles présentent une définition et une source de données en matière de SLO (non fournies par K8s).

© 2021 Turbonomic, an IBM Company

Qu'est-ce que cela signifie pour vous, vos équipes et votre entreprise ? Vous trouverez ci-dessous un aperçu des avantages uniques offerts par Turbonomic, que vous exécutiez Kubernetes sur site, dans le cloud, sans système d'exploitation, ou en adoptant plusieurs de ces approches.

« Régulez la vitesse » de vos applications : vos équipes définissent les SLO en matière de temps de réponse, et un logiciel basé sur l'intelligence artificielle assure que la plateforme et l'infrastructure sous-jacente fournissent en permanence les ressources dont elles ont besoin pour répondre à ces SLO, où que s'exécutent les applications.

Minimisez la main-d'œuvre manuelle : développeurs, ingénieurs DevOps et ingénieurs en fiabilité de site (SRE) n'ont pas besoin de définir des seuils, des contraintes ou des règles de mise à l'échelle automatique. Le logiciel prend les bonnes décisions en matière de ressources à votre place, en vous proposant des actions réellement automatisables.

Ne dépensez pas inutilement en matière de ressources : inutile de vous en remettre aux développeurs pour prendre des décisions en matière de ressources (puisqu'ils ont souvent tendance à surestimer les besoins par simple mesure de précaution, n'est-ce pas ?). Notre logiciel détermine exactement les ressources nécessaires pour chaque service, le tout en fonction de la demande d'applications.

Accélérez le DevOps en toute confiance : augmentez en toute sécurité la fréquence et l'évolutivité des déploiements. Nos analyses s'intègrent à vos flux de travail DevOps, ce qui garantit le fonctionnement continu des services existants et nouvellement déployés.

Planifiez rapidement et facilement votre croissance : simulez l'intégration de nouveaux services avec notre logiciel. Déterminez exactement le nombre de nœuds supplémentaires qu'il vous faut pour générer une nouvelle croissance.

En savoir plus

Rendez-vous sur turbonomic.com/kubernetes

© 2021 Turbonomic, an IBM Company

Lumière sur un client

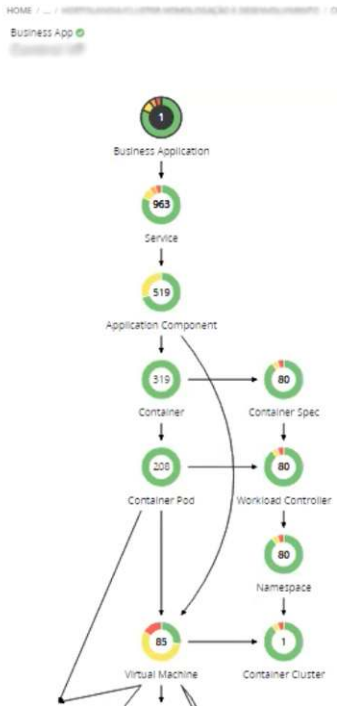
Accélérer la transformation numérique durant une pandémie

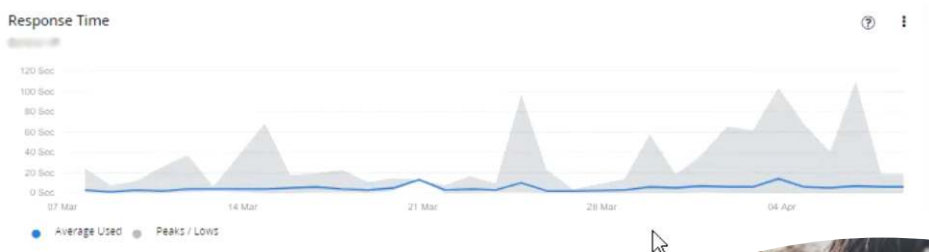
Les ressources dynamiques de Turbonomic au sein de la plateforme Kubernetes et de l'infrastructure sous-jacente ont permis de maintenir un niveau faible de temps de réponse.

Comptant plus de 6 millions de clients, ce client est l'une des plus grandes compagnies d'assurance en Amérique du Sud. Son approche standard de la gestion des ressources des environnements « existants » et « nouvelle génération » ralentissait sa transformation numérique et sa réponse à la pandémie.

L'automatisation de Turbonomic a permis de maintenir un niveau faible de temps de réponse au pic de la demande durant les vacances de Pâques.

Ce client dispose d'une application commerciale qui s'intègre à l'une des plus grandes compagnies aériennes low-cost de la région. La réservation de l'assurance voyage se fait à partir de cette application ; le pic que l'on observe sur le graphique se rapporte à la période des vacances de Pâques (sur plusieurs jours). Alors que la demande sur l'application augmentait, les ressources dynamiques de Turbonomic au sein de la plateforme Kubernetes et de l'infrastructure sous-jacente ont permis de maintenir un niveau faible de temps de réponse.



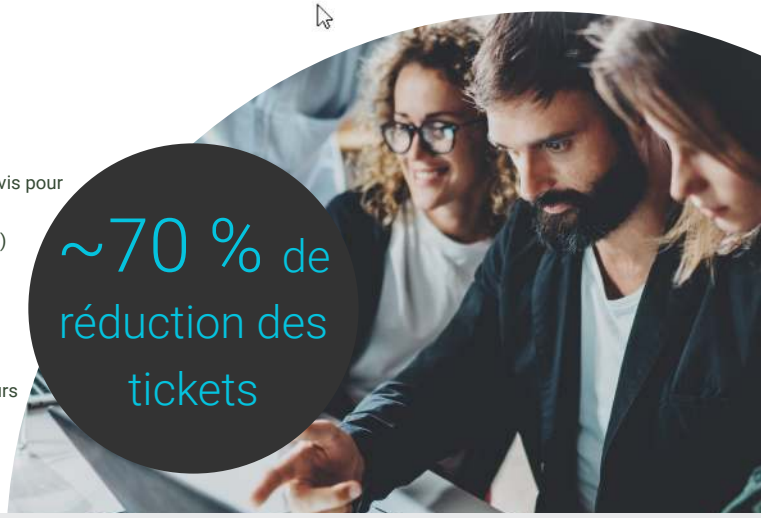


57 applications essentielles

- Par exemple, GPS dans la voiture : signalement du vol du véhicule, devis pour de nouvelles règles, etc.
- (~7 000 conteneurs / ~3 000 pods)
- Liaison à Dynatrace

Automatisation

- Redimensionnement des conteneurs (échelonnement)
- Placement continu (tous)



À propos de Turbonomic

Turbonomic permet aux entreprises et aux organismes et entités publiques de bénéficier d'une agilité et d'une élasticité dans leurs environnements sur site et multicloud afin de pouvoir arrêter de se focaliser sur la gestion réactive des opérations informatiques pour concentrer leurs budgets et leurs ressources humaines précieuses sur l'innovation de l'entreprise.

La plateforme de Turbonomic est conçue pour automatiser la gestion des ressources applicatives, du physique au virtuel, en passant par le cloud, les conteneurs et l'Internet des objets, dans tous les environnements du client. Nous pensons que lorsqu'un logiciel fait ce qu'il fait le mieux, l'homme peut faire ce qu'il fait le mieux.

Pour demander une démonstration afin de découvrir ce que Turbonomic peut faire pour votre environnement hybride ou multicloud, cliquez ici.

© 2021 Turbonomic, an IBM Company