turbonomic
an IBM Company

Pourquoi l'AIOps ?
Performance des applications

Table des matières

Pourquoi la performance des applications ?

L'ère des applications modernes
À l'ère des applications modernes, où plus de deux tiers du PIB sont désormais numériques,[1] la performance des applications est la priorité absolue des CIO, car l'application *est* le business. La mise à disposition d'applications est la principale raison d'être de l'informatique. Le DSI n'a pas d'autre choix que d'assurer la performance des applications pour que son entreprise ne soit jamais freinée par l'informatique. En fait, le DSI est considéré comme quelqu'un qui échoue s'il n'assure pas la performance des applications ; ironiquement, il est acceptable, voire raisonnable, pour les DSI de dépasser les budgets prévus. En bref, le fait de ne pas assurer la performance des applications nuit à l'entreprise. Ce qui est difficile, c'est que les applications qui sont les premières et les plus longues à être congestionnées sont celles dont la demande augmente le plus. Ce sont souvent les applications les plus précieuses.

Le service d'hébergement des applications modernes
Les investissements dans le développement des applications dépassent largement le coût de l'hébergement des applications. Les entreprises qui investissent dans ce domaine sont conscientes et consentent à surdimensionner leur infrastructure et leurs environnements cloud pour réduire les risques liés aux performances des applications. Les infrastructures, tant dans les centres de données que sur le cloud, deviennent rapidement des produits jetables et, par conséquent, justifient moins bien le retour sur investissement. En outre, la mauvaise performance des applications crée une méfiance entre les propriétaires d'applications des lignes business (LOB) et les équipes ITOps et CloudOps qui fournissent le service d'hébergement des applications. Pensez aux LOB qui investissent des millions de dollars avec des centaines de personnes pour développer de nouvelles expériences utilisateur, mettant leur réputation en jeu, pour que les équipes d'exploitation livrent aux utilisateurs finaux l'expérience lente de la « roue de chargement ». Tout cet investissement dans l'amélioration de l'expérience de l'utilisateur final est perdu.

Le point de pression de plus en plus important pour la performance des applications
La saturation des applications se produit lorsque l'infrastructure - sur site ou sur le cloud - ne peut pas répondre à la demande de l'application et de ses utilisateurs finaux. La saturation des ressources de l'infrastructure est la cause la plus fréquente de la dégradation des performances des applications. En revanche, l'architecture du code d'application, surveillée à l'aide d'outils de gestion des performances des applications (APM), représente rarement moins de 10 % de la source de dégradation des performances des applications dans les environnements de production. De plus, étant donné que le développement d'applications se concentre sur la livraison d'un code de qualité en utilisant des processus améliorés, tels que l'intégration continue, la livraison continue (CICD), l'assurance qualité (QA), la pré-production et le stockage, la qualité du code ayant un impact sur les performances devient de plus en plus rare.

Un partage efficace des infrastructures requiert une hiérarchisation des priorités
Dans le datacenter sur site, les applications sont en souffrance lorsque la demande de ressources d'infrastructure partagées n'est pas hiérarchisée. Compte tenu de cette contrainte, les charges de travail des applications sont généralement surdimensionnées et mises en place sans compréhension suffisante des ressources disponibles et sans tenir compte du contexte. Lors du partage des ressources, toutes les applications puisent dans un pool de ressources commun, quelle que soit leur utilisation. L'allocation de ressources surdimensionnées et mal dimensionnées entraîne une congestion constante, ce qui se traduit par des violations des accords de niveau de service (SLA), un dépannage manuel inefficace, des ajustements perpétuels des ressources et, comme indiqué précédemment, des investissements non rentables dans le développement d'applications.

Les outils de surveillance actuels, réactifs et à ressource unique, ne comprennent pas la relation entre les applications et l'infrastructure et, par conséquent, reposent sur une interprétation et une intervention manuelles pour résoudre la congestion des ressources. Plus il faut de temps pour trouver une solution, plus la décision sur ce qu'il faut faire risque d'être obsolète.

Une autre solution consiste à utiliser en permanence la demande dynamique des applications pour donner la priorité à l'allocation de l'ensemble des ressources de l'infrastructure partagée. Plus la courbe de demande d'une application est ascendante, plus sa priorité relative pour obtenir et préserver les ressources augmente. Au fur et à mesure que cette courbe de demande descend, sa priorité relative pour obtenir et conserver les ressources diminue. Cette redéfinition continue des priorités de l'ensemble des ressources partagées minimise la saturation, permet un partage fluide des ressources entre toutes les applications et garantit les performances des applications.

Le recours efficace au cloud public nécessite une connaissance de l'application
Dans le cloud public, les applications sont privées de ressources lorsque les instances - les ressources - restent insuffisantes pour répondre à la demande de l'application. Pour faire face à ce risque, les administrateurs de cloud surdimensionnent souvent les instances. Dans les deux cas, le provisionnement incorrect des instances résulte de la connaissance limitée qu'a l'administrateur du cloud des besoins en ressources de la demande d'application. En outre, les développeurs qui se concentrent sur la création de l'application ne s'intéressent généralement pas au profilage et à la prévision de ses besoins en ressources. Par conséquent, les ressources sont estimées au hasard et le fournisseur de cloud fait peser le risque, la charge et les conséquences de cette estimation sur le client.

En effet, le provisionnement d'une seule instance Amazon Elastic Compute Cloud (Amazon EC2) comporte des millions d'options de configuration pour déterminer le type de ressource, le matériel sous-jacent, la taille, les zones géographiques, les prix, les réservations, les plans d'économies, etc.

En outre, les changements dynamiques de la demande d'applications exigent une réévaluation continue de cette supposition utilisée pour sélectionner les instances du cloud. Enfin, la gestion de l'élasticité des instances d'applications du cloud public nécessite un calibrage continu en temps réel, en particulier avec l'adoption d'applications à courte durée de vie et basées sur des conteneurs. Par conséquent, la plupart des allocations d'instances dans le cloud public sont surdimensionnées et gérées manuellement et de manière réactive pour se prémunir contre les risques de performance et sans aucune compréhension de la demande de ressources des applications. De plus, les outils du fournisseur de cloud et de la plateforme de gestion du cloud basés sur la surveillance ne comprennent pas suffisamment la demande de l'application pour garantir les performances des applications dans le cloud. Ils sont limités à la visibilité des coûts, à la facturation historique et à l'allocation des coûts départementaux, ce qui n'a rien à voir avec la garantie des performances des applications exécutées sur un cloud public.

Le pourquoi de la gestion des ressources d'applications de Turbonomic pour IBM Cloud Paks

Turbonomic, une société IBM, fournit la base du service d'hébergement d'applications modernes en utilisant de manière unique une compréhension de la demande d'applications pour offrir des actions continues de ressourcement des applications et des analyses de performance afin d'assurer la performance des applications en temps réel, au fil du temps, 24h/24 7j/7 et 365j/an.

La gestion des ressources d'applications de Turbonomic pour IBM Cloud® Paks utilise un modèle de données commun afin que les clients puissent, en toute confiance, attribuer des ressources à leurs applications d'aujourd'hui, ainsi qu'aux applications modernes de demain, qu'elles soient exécutées sur place, dans des clouds publics ou à la périphérie.

La gestion des ressources applicatives (ARM) permet de faire des économies et d'améliorer les performances
L'ARM de Turbonomic, qui utilise des analyses basées sur l'intelligence artificielle, contribue à garantir les performances des applications en faisant correspondre en permanence les exigences de la demande applicative en temps réel avec les types et les tailles des ressources. Les actions de ressourcement comprennent le démarrage et l'arrêt, le placement initial et continu, la mise à l'échelle et le redimensionnement.

En revanche, des centaines d'outils - et même des feuilles de calcul - prétendent faire économiser de l'argent aux clients, mais leurs « recommandations » sont souvent basées sur de simples seuils d'alerte qui peuvent causer des problèmes de performance. Selon une étude récente, 39 % des professionnels des opérations financières (FinOps) ont indiqué que le principal défi à relever était d'inciter les ingénieurs à agir.[2] La raison ? le manque de confiance dans les outils existants utilisés pour générer ces « actions ».

Les nouvelles exigences pour assurer la performance des applications modernes
La plateforme AIOps de Turbonomic utilise un modèle de données commun complet qui ingère, normalise et gère toutes les ressources partagées dont dépendent les performances des applications. Plus important encore, elle crée la topologie de la relation de la chaîne d'approvisionnement - « la couture » - entre chaque dépendance de ressourcement, de l'application à l'infrastructure.

En revanche, les outils manuels et fragmentés ne peuvent pas garantir les performances, car ils surveillent les ressources de manière isolée et n'ont qu'une perspective limitée, voire inexistante, des applications. La performance des applications exige de comprendre et de gérer toutes les dépendances des ressources dans les quantités, l'ordre et le calendrier exacts.

Conclusion

À l'ère des applications modernes, l'application est le business. Les applications devenant plus complexes, avec plus de dépendances, et les environnements plus divers et distribués, le risque pour les performances des applications et l'expérience des utilisateurs augmente de façon exponentielle. La seule façon de relever ces défis est de mettre en œuvre une approche axée sur les applications qui évalue en permanence la demande des applications et l'offre de ressources disponibles et qui génère des recommandations exploitables auxquelles les responsables des opérations et des applications peuvent se fier. Au fil du temps, à mesure que la confiance s'installe, l'étape suivante consiste à automatiser ces actions. Il en résultera des applications très performantes, une réduction massive des dépenses informatiques et la possibilité de stimuler l'innovation commerciale. C'est la magie d'AIOps.

Plus de deux tiers du PIB mondial est désormais numérique.[1]

- La mise à disposition d'applications est la principale raison d'être de l'informatique
- Les CIO doivent s'assurer que l'évolution de l'entreprise n'est jamais freinée par l'informatique.
- Le développement d'une application coûte 3 fois plus cher que son hébergement.
- Les entreprises surdimensionnent leurs ressources afin d'atténuer le risque de dégradation des performances des applications.
- La saturation des ressources de l'infrastructure est la source la plus fréquente de la dégradation des performances des applications.
- Les applications souffrent lorsque la demande de ressources d'infrastructure partagées n'est pas hiérarchisée.
- Les outils de surveillance actuels, axés sur les ressources et ne comprenant pas la demande des applications, ne peuvent que réagir aux situations négatives.
- La redéfinition continue des priorités de l'ensemble des ressources partagées, en fonction de la demande des applications, est le seul moyen de garantir les performances des applications.

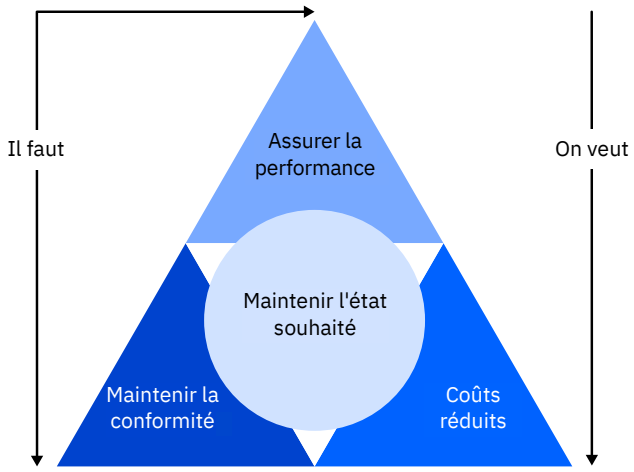

Figure 1. Les applications sont dans l'état souhaité où les performances sont assurées tout en maintenant la conformité au moindre coût.

- Les administrateurs du cloud ont une connaissance limitée des besoins en ressources d'une application.
- Les développeurs d'applications se concentrent sur la logique commerciale et reportent les décisions de ressourcement sur le personnel informatique.
- La sélection des bonnes ressources de cloud public est complexe, avec des millions de choix de configuration.
- La plupart des sélections d'instances de cloud sont surdimensionnées afin de minimiser le risque de dégradation des performances.
- Le dimensionnement des instances du cloud public nécessite de connaître la demande des applications.

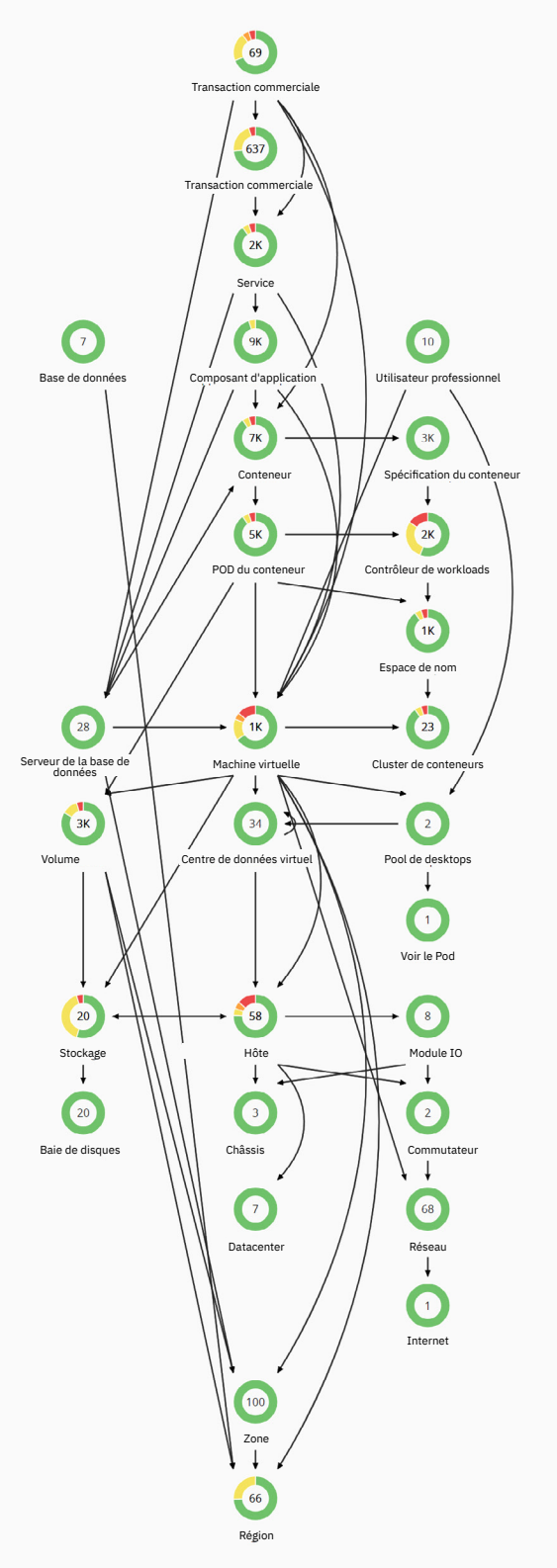

Figure 2. Turbonomic crée automatiquement la topologie des relations de la chaîne d'approvisionnement.

A propos de Turbonomic, une société IBM

Turbonomic, une société IBM, fournit un logiciel de gestion des ressources d'application (ARM) utilisé par les clients pour assurer la performance et la gouvernance des applications en leur attribuant dynamiquement des ressources dans des environnements hybrides et multiclouds. La gestion de la performance du réseau (NPM) de Turbonomic fournit des solutions modernes de surveillance et d'analyse pour aider à assurer une performance continue du réseau à l'échelle des réseaux multifournisseurs pour les entreprises, les transporteurs et les fournisseurs de services gérés.

Pour plus d'informations, rendez-vous sur ibm.com/cloud/turbonomic.

Compagnie IBM France
17 avenue de l'Europe
92275 Bois-Colombes Cedex

Produit aux États-Unis d'Amérique
Novembre 2021

[1] IDC Reveals 2021 Worldwide Digital Transformation Predictions, IDC, 29 octobre 2020.
[2] State of FinOps Report 2021, The FinOps Foundation, 2021.

turbonomic
an IBM Company