Automatizando a elasticidade de contêineres orientada por aplicações

Para engenheiros de plataforma e DevOps que buscam operacionalizar a velocidade de lançamento no mercado e garantir o desempenho das aplicações

Índice

Resumo executivo

Sua vantagem competitiva depende da rapidez com que ideias se transformam em transações de negócios e do desempenho delas para seus clientes. A tecnologia é a facilitadora.

Os contêineres oferecem a velocidade, agilidade, elasticidade e a escalabilidade que está mudando fundamentalmente a maneira como desenvolvemos, implementamos e executamos essas aplicações. Eles anunciam um mundo em que as aplicações realmente podem ser executadas em qualquer lugar. As atualizações e novos recursos podem ser colocados em produção diversas vezes por dia e a demanda dinâmica e flutuante das cargas de trabalho pode ser gerenciada com o fornecimento de infraestrutura elástica, em qualquer lugar e a qualquer hora. O Kubernetes é uma plataforma que permite que as empresas sejam ágeis e elásticas, mas ela não gerencia compensações sobre como garantir desempenho e manter a eficiência ao mesmo tempo.

Apesar de toda a simplicidade e agilidade que a containerização oferece, a plataforma de orquestração fornece apenas uma maneira de gerenciar o ciclo de vida desses serviços: implementando e mantendo seus serviços da maneira determinada por você.

As plataformas de contêiner não asseguram desde o início que os serviços atendem aos objetivos de nível de serviço (SLO) e não podem gerenciar recursos dinamicamente
As políticas baseadas em limites não resolvem a questão do desempenho contínuo. Essa abordagem nunca funcionou e, devido à velocidade das mudanças nas plataformas de contêiner, o ajuste automático de escala acionado sem correspondência pode acabar causando problemas. A infraestrutura elástica é importante para aumentar o desempenho, mas precisa de uma análise de dados automatizada que gerencie continuamente a demanda, o fornecimento e as restrições para atender aos SLOs (objetivos de nível de serviço) desejados.

Este white paper aborda os principais conceitos a serem considerados ao adotar uma plataforma de contêiner, como a maneira de administrar seus negócios e como proteger esse investimento com a automação, que garante desempenho enquanto reduz custos e mantém a conformidade.

Este documento descreve por que você precisa de uma análise de dados orientada de cima para baixo para que uma plataforma Kubernetes autogerenciada execute seus serviços. A construção de uma escala multinuvem no início da sua jornada proporciona à sua organização de TI a “memória muscular” operacional que transformará substancialmente como (e quando) você entrega mais inovação.

Um compromisso com a velocidade, agilidade, elasticidade e escalabilidade

O Kubernetes oferece elasticidade, porém, isso não garante automaticamente que você atenderá e cumprirá os SLOs das aplicações.

O sucesso na adoção da conteinerização depende de como você oferece aos desenvolvedores a agilidade de que precisam, a elasticidade necessária para se adaptar em escala às demandas continuamente flutuantes e a garantia de que as aplicações vão desempenhar na velocidade esperada.

Adotar uma abordagem nativa em nuvem e separar suas aplicações em conjuntos distintos de serviços pode impulsionar o desenvolvimento e a implementação mais ágeis das aplicações. Contêineres oferecem o empacotamento que torna os seus serviços portáteis e escaláveis. O Kubernetes oferece uma estrutura e pontos de controle para executar suas aplicações e serviços digitais. No entanto, para fornecer uma plataforma de escala corporativa de bom desempenho para o seu negócio, ainda é necessário incluir recursos para liberar a elasticidade permitida pela plataforma para atender e garantir os SLOs da aplicação.

Implemente mais rápido com CICD e feedback da produção
A metodologia certa de integração contínua/implementação contínua (CICD), baseada em automação, é essencial para alcançar o prazo de lançamento no mercado mais rápido. No relatório Google Cloud State of DevOps 20211, os entrevistados citaram melhorias significativas devido à implementação da metodologia CICD:

Frequência de implementação	Semanal – mensal	Por hora – diária
Mudar tempo de avanço	Mais de seis meses	Menos de uma hora
Mudar índice de falhas	16% a 30%	0% a 15%

Com a velocidade surge a necessidade de ter um modo de gerenciar as mudanças constantes na produção e ter um loop de feedback quanto ao desempenho de seus serviços e quanto à previsão do que é necessário para a infraestrutura. O objetivo é ter uma maneira de definir seus SLOs e receber feedbacks da plataforma sobre como configurar seus contêineres e a infraestrutura para reduzir o risco de problemas de desempenho.

- Quem decide sobre como os recursos devem ser alocados aos serviços? Como eles tomam essa decisão? Testes de estresse e referenciais comparativos com relação aos SLOs definidos e assim por diante.
- Como medir o desempenho? Há um loop de feedback em seu pipeline de CICD para garantir que os contêineres e os pods estejam configurados corretamente?
- Como você garante que sempre há recursos suficientes para novas implementações?

Encontre suas respostas do IBM Turbonomic

Opções	Limitações	Respostas do IBM Turbonomic
Analisar manualmente os dados de utilização do contêiner e do pod para determinar as especificações dos recursos.	– Configuração de coleta de dados – Mão de obra para análise de dados	– Análise de dados top-down orientada por aplicações que determina como dimensionar seus contêineres – Feedback na metodologia CICD – Oportunidades para reduzir solicitações quando não forem necessárias
Analisar manualmente os dados de recursos de todos os pontos da pilha para determinar a capacidade da produção.	– Mão de obra para coletar dados de diversas fontes – Mão de obra para análise de dados	Análise de dados baseada em utilização para identificar as necessidades de recursos em toda a full-stack

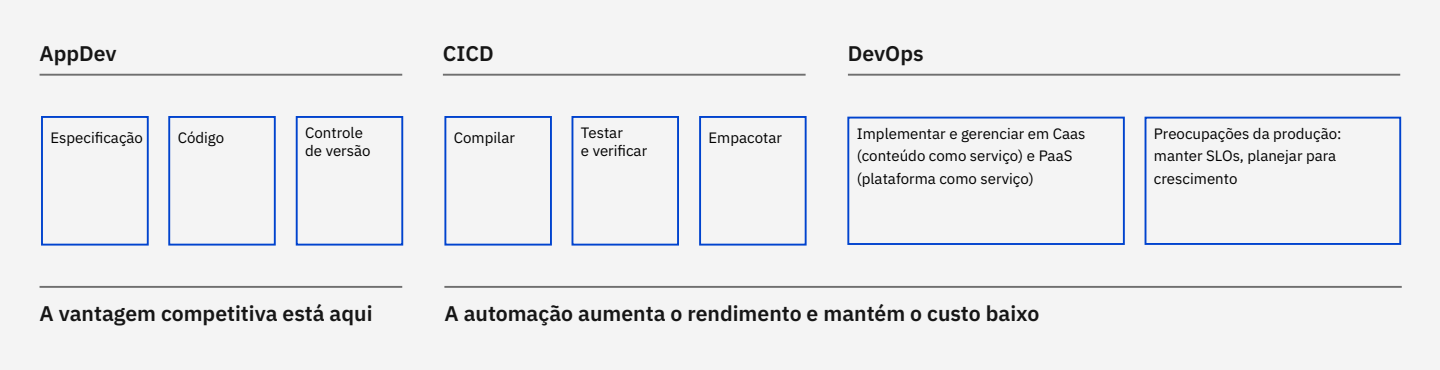


Figura 1. Processo para agilidade de aplicações.

Plataforma e infraestrutura

Por que você precisa de gerenciamento full-stack orientado por aplicações?
Independentemente da plataforma de contêiner, ou infraestrutura subjacente, que você escolher (nuvem privada, nuvem pública, nuvem híbrida, multinuvem ou até mesmo bare metal), os desafios operacionais de sua plataforma como um serviço (PaaS) serão os mesmos:

- Como determinar se há capacidade suficiente para acomodar a demanda atual e em escala?
- Como decidir o momento para ativar mais nós de aplicações?
- Como decidir o momento para suspender?
- Como lidar com o pico de demanda?
- Como utilizar os recursos de nuvem pública para bursting?
- Como garantir alta disponibilidade (HA) e resiliência em toda a pilha?
- Como impor as restrições de negócios?

A elasticidade ativada por plataformas de contêiner oferece a oportunidade de fornecer para o total das demandas médias de sua aplicação em vez de o total das demandas de pico de suas aplicações. Para aproveitar essa capacidade, a entrega de uma plataforma que aumenta e diminui a capacidade continuamente conforme as flutuações de demanda, requer um software que tome decisões de fornecimento constantemente para assegurar que as aplicações tenham a computação, o armazenamento e a rede necessários sempre que precisarem deles.

Encontre suas respostas do IBM Turbonomic

Opções	Limitações	Respostas do IBM Turbonomic
Executar em provedores de serviço que fornecem grupos de ajuste de escala automático, como grupos de serviço afiliados (ASGs), conjuntos de disponibilidade e assim por diante.	– Políticas baseadas em limites – Não é possível escalar um nó específico: todos os nós devem ter as mesmas restrições, rótulos de nó e assim por diante	– SLOs top-down orientados a aplicações – Continuamente ajusta os recursos da infraestrutura para atender à demanda das aplicações – Continuamente aumenta e diminui a capacidade, verticalmente e horizontalmente, dos contêineres, pods e nós certos – Continuamente posiciona os pods nos nós adequados
Analisar os dados de recursos de todos os pontos na pilha para determinar a capacidade de produção.	– Mão de obra para coletar dados de diversas fontes – Mão de obra para análise de dados	– Análise de dados baseada em utilização para identificar as necessidades de recursos em toda a full-stack – Continuamente aumenta e diminui a capacidade, verticalmente e horizontalmente, dos contêineres, pods e nós certos – Continuamente aciona ações para evitar gargalos

Operando para SLOs em escala

O propósito da plataforma de contêiner é executar suas aplicações no nível desejado de serviço aos seus negócios. É preciso garantir o desempenho continuamente conforme o aumento do número de aplicações. Normalmente, vemos os clientes levarem mais de 12 meses para até as três primeiras aplicações. Para as aplicações seguintes, com o benefício das qualificações adquiridas e melhores práticas, podem ser necessários entre seis a doze meses adicionais. Quando as linhas de negócios entendem e aprendem o que é possível realizar, a escala do número de serviços individuais a serem gerenciados deixa de ser possível gerenciar manualmente. Mesmo que você tenha criado serviços stateless, aproveitando a natureza efêmera dos contêineres, qual é a sua tolerância para redução do desempenho da experiência do seu usuário final? O que você pode fazer para gerenciar não apenas a demanda, mas a crescente taxa de mudança? A resposta está na automação, por meio de ações que são baseadas em uma análise de compensações de quantas instâncias de serviço são necessárias para garantir o SLO, a configuração do tamanho e posição de sua carga de trabalho e a disponibilização de recursos em conformidade por meio da infraestrutura.

Limites não resolvem os problemas

Uma plataforma de contêineres garantirá que você tenha um número mínimo de serviços disponíveis. Se um deles for interrompido, ele tentará ativá-lo novamente. Mas se você deseja garantir uma boa experiência do usuário, é recomendável que o sistema responda antes que haja redução do desempenho e ocorra uma interrupção. É possível definir o ajuste de escala horizontal automático nativo para atender à demanda, mas é necessário decidir quais métricas expressam melhor os recursos requeridos, configurar os limites superiores e inferiores, testar e extrapolar se funcionará sob a demanda de produção e, em seguida, repetir isso para cada serviço implementado. Imagine se você tivesse mais de 100 serviços para uma única aplicação? Essas políticas não têm relação entre si. Como garantir que a inclusão de mais pods de um serviço não causará congestionamento em outra área? Você está clonando um pod que foi mal configurado e precisa de um ajuste de escala vertical primeiro? Como você lida com o congestionamento de nós, considerando-se os “vizinhos barulhentos”, e identifica recursos alocados não utilizados que podem ser liberados para atender a essa demanda?

Além disso, a configuração de contêineres, pods e escaladores automáticos de pod horizontal (HPA) ou políticas de ajuste automático de escala de cluster não é um exercício que se faz uma única vez. Os melhores esforços devem ser monitorados continuamente e redefinidos se falharem. O que suas equipes poderiam fazer com o tempo que seria economizado se elas não precisassem definir e redefinir manualmente esses limites?

A importância de acertar essas configurações tem implicação direta no lançamento bem-sucedida de sua estratégia de transformação digital. Algumas implementações inválidas podem atrasar significativamente a adoção de plataformas e sistemas em desenvolvimento. E gastar tempo e mão de obra em excesso configurando manualmente esses pontos de controle pode ser um grande obstáculo à capacidade da sua empresa de se tornar centrada em plataforma. Os seus negócios podem arcar com esse atraso? É necessário ter um sistema de controle que possa gerenciar compensações de todos os recursos e definir os limites e solicitações de escala vertical do contêiner, o número de pods requeridos e as decisões de posicionamento para redistribuir os pods e gerenciar os recursos do cluster usando um único mecanismo de análise de dados.

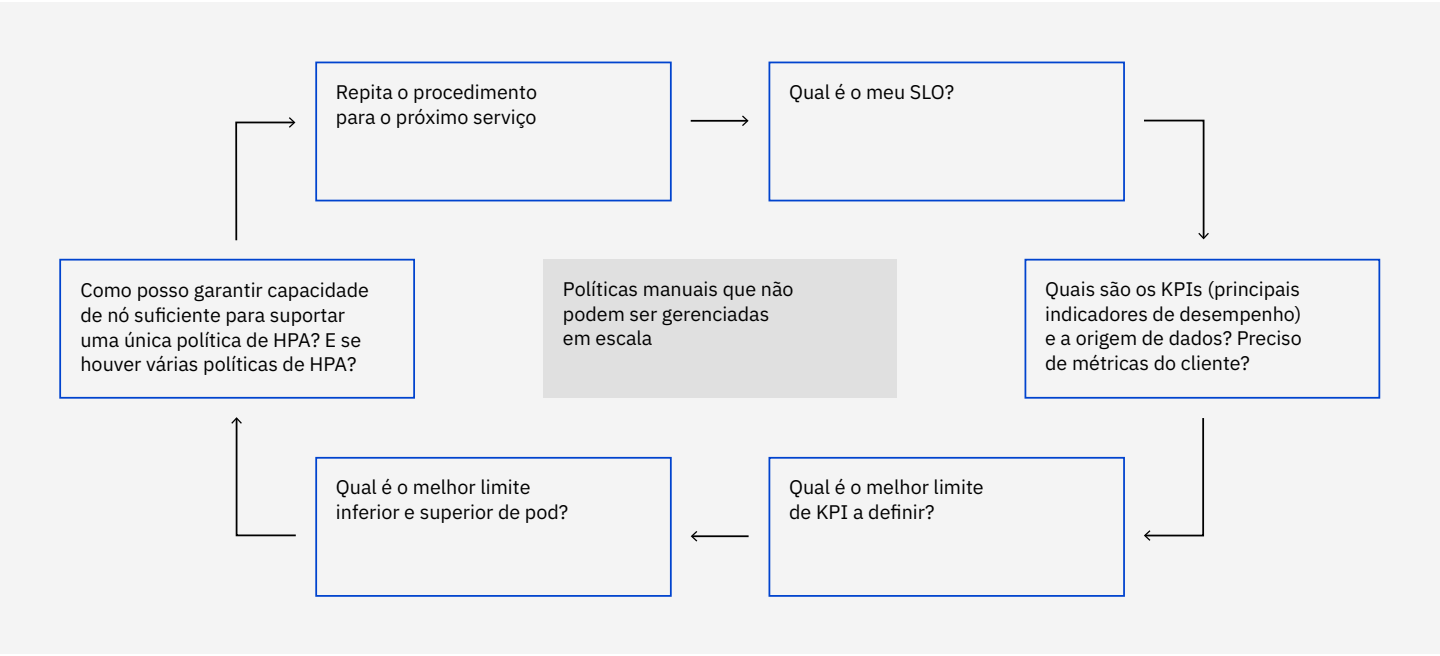


Figura 2. As políticas manuais que não podem ser gerenciadas em escala

Encontre suas respostas do IBM Turbonomic

Opções	Limitações	Respostas do IBM Turbonomic
Política baseada em limite HPA usada para acionar o aumento e a redução de capacidade de pods	– Configura por serviço – Com base na média de todos pods para o serviço – KPIs e limites definidos manualmente, além de limites superiores e inferiores de pod	– SLOs top-down orientados a aplicações – Usa dados de tempo de resposta para promover o ajuste de escala horizontal de serviços para atender os SLOs – Continuamente aumenta e diminui a capacidade, verticalmente e horizontalmente, dos contêineres, pods e nós certos – Continuamente posiciona os pods nos nós adequados – Continuamente ajusta os recursos da infraestrutura para atender à demanda das aplicações
Política baseada em limite de escalador automático de pod (VPA) vertical para ajustar verticalmente a escala dos contêineres	– Deve ser definido para todos os serviços – Projeto beta: usar sob sua própria responsabilidade e risco – Não acessar a capacidade do nó para tomar ações	
Permite que os pods sejam interrompidos para reimplementá-los em um nó mais apropriado	Experiência do usuário ruim para transações no pod que estão preparadas para serem interrompidas	
As soluções de observabilidade do Prometheus coletam e consolidam os dados	– Não fornecer análise de dados – Não fornecer ações	

Uma abordagem orientada por aplicações

Os SLOs de aplicações devem conduzir a infraestrutura
A conteinerização de aplicações de missão crítica é um investimento com muitos benefícios. Mas para aproveitar todos benefícios de velocidade, elasticidade e portabilidade, é necessário um software para tomar as decisões corretas sobre a alocação de recursos no momento certo, 24 horas por dia, 7 dias por semana, 365 dias por ano. Caso contrário, a complexidade irá atrasá-lo.

O IBM Turbonomic Application Resource Management conecta suas aplicações de missão crítica à plataforma Kubernetes e à infraestrutura subjacente, essencialmente onde as suas aplicações são executadas. Com base na demanda de aplicações em tempo real e levado em conta as restrições e interdependências em cada camada da solução, da lógica à física, o software determina as ações certas no momento certo para ajudar a garantir que as aplicações sempre obtenham exatamente o que precisam para executar. Execute em tempo real, planejado ou como parte do pipeline de DevOps.

Dimensionamento inteligente: como você deve dimensionar os contêineres?
- Automatize com a implementação, execute e persista no redimensionamento como parte do pipeline, por exemplo, YAML, Jenkins e assim por diante.
- Automatize em tempo real, execute dinamicamente por meio do Kubernetes.

Posicionamento contínuo: quando é necessário mover os pods? Para quais nós?
- Execute dinamicamente e em tempo real por meio do Kubernetes . Apenas para serviços stateless sem interrupção.

Ajuste de escala dinâmico: quando é necessário aumentar ou retornar a capacidade do cluster? Em quanto?
- Execute dinamicamente o ajuste de escala do cluster em tempo real por meio da infraestrutura como código ou da API do cluster Kubernetes.

Ajuste de escala baseado em SLO: quando é necessário aumentar ou retornar a escala de pods para atender aos SLOs de tempo de resposta das aplicações? Em quanto?

Pré-requisitos para o ajuste de escala orientado por SLO:
- As aplicações são projetadas para microsserviços stateless horizontais.
- Elas têm uma definição e origem de dados de SLO que o Kubernetes não fornece.

O que esse tipo de automação inteligente representa para você, suas equipes e seus negócios? A seguir estão os benefícios exclusivos que o IBM Turbonomic oferece, independentemente de você executar o Kubernetes localmente, na nuvem em servidores bare metal ou em qualquer combinação.

“Piloto automático” para suas aplicações: suas equipes definem SLOs de tempo de resposta. O software impulsionado por IA ajuda a garantir que a plataforma e a infraestrutura subjacente sempre forneçam os recursos necessários para atender a esses SLOs, independentemente de onde as aplicações são executadas.

Reduza o trabalho manual: desenvolvedores, DevOps e engenheiros de confiabilidade de sites (SREs) não precisam definir limites, restrições ou políticas de ajuste de escala automático. O software toma as decisões certas sobre recursos para você, oferecendo ações que podem realmente ser automatizadas.

Não gaste demais com capacidade: não é necessário depender de desenvolvedores para tomar decisões sobre recursos. Eles geralmente sobreprovisionam apenas por segurança, certo? Nosso software determina exatamente os serviços de recursos que são necessários, tudo com base na demanda das aplicações.

Acelere o DevOps com confiança: aumente a frequência e a escala das implementações com segurança. Nossa análise de dados pode ser integrada aos fluxos de trabalho do DevOps, ajudando a garantir o desempenho contínuo de serviços recém-implementados e já existentes.

Planeje o crescimento com mais facilidade: simule a integração de novos serviços com nosso software. Determine exatamente quantos nós adicionais serão necessários para poder expandir os seus negócios.

Destaques de cliente

Acelerando a transformação digital durante a pandemia

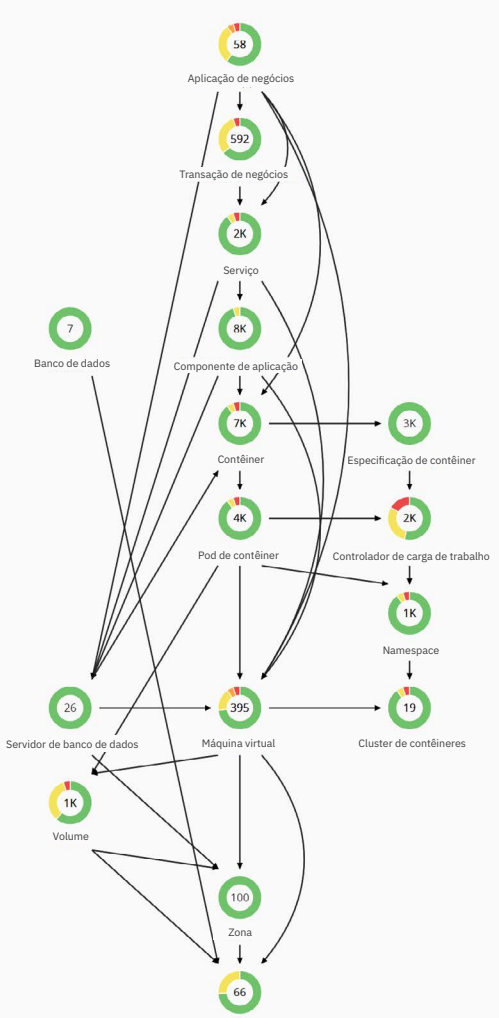
Os recursos dinâmicos do IBM Turbonomic dentro da plataforma Kubernetes e a infraestrutura subjacente mantiveram o tempo de resposta baixo.

Este cliente é uma das maiores companhias de seguros na América do Sul, com uma carteira com mais de 6 milhões de clientes. Sua abordagem padrão do setor para gerenciar os recursos de ambientes existentes e de nova geração estava desacelerando a transformação digital e a resposta da empresa à pandemia.

A automação do IBM Turbonomic manteve o tempo de resposta baixo durante o pico de demanda em feriados

Este cliente tem uma aplicação corporativa que se integra com uma das maiores companhias aéreas de baixo custo em operação na região.

O seguro de viagem é reservado por meio dessa aplicação, então o pico visto na Figura 3 está relacionado ao feriado prolongado de Páscoa. Embora a demanda da aplicação tenha aumentado, a alocação de recursos dinâmica do IBM Turbonomic na plataforma Kubernetes e a infraestrutura subjacente conseguiram manter o tempo de resposta baixo.



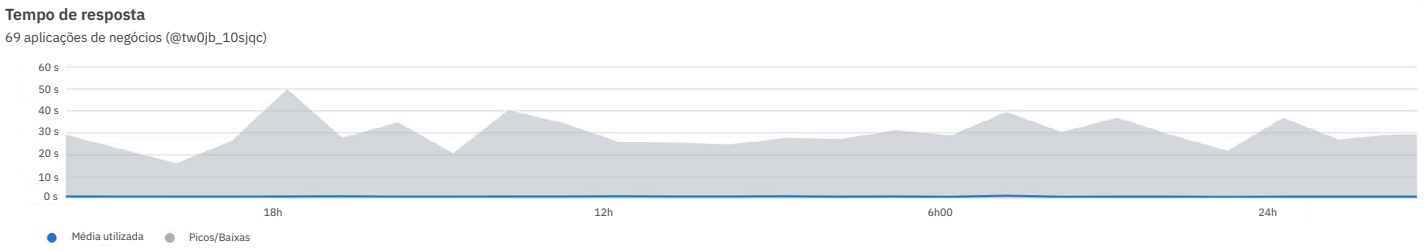


Figura 3. A visualização full-stack da aplicação corporativa individual e seu tempo de resposta com automação mantido baixo, mesmo durante o pico de demanda

57 aplicações de missão crítica
- Por exemplo, o dispositivo de GPS no carro para reportar roubo de veículos, cotações para novas políticas e assim por diante
- Aproximadamente 3.000 pods (composto por cerca de 7.000 contêineres)
- Integrado ao Dynatrace

Automatizado
- Redimensionamento de contêiner (preparação)
- Posicionamento contínuo (todos)

Redução de chamados em aproximadamente 70%

Sobre a IBM Turbonomic, uma empresa IBM

O IBM Turbonomic Application Resource Management fornece o software Application Resource Management (ARM), que é utilizado por clientes para ajudar a garantir o desempenho e o controle das aplicações ao fornecer recursos dinamicamente às aplicações em ambientes híbridos e multinuvem. O gerenciamento de desempenho de rede (NPM) do IBM Turbonomic fornece soluções modernas de monitoramento e análise de dados para ajudar a garantir o contínuo desempenho da rede em escala entre as rede de diversos fornecedores para empresas, operadoras e provedores de serviços gerenciados.

Para saber mais sobre a automação inteligente do IBM Turbonomic, visite ibm.com/cloud/turbonomic ou fale com um representante IBM.

IBM Brasil
Rua Tutóia, 1157
CEP 04007-900 São Paulo - SP

Produzido nos Estados Unidos da América
Março de 2022

[1] State of DevOps 2021, Google Cloud, 2021