

Automatizzazione dell'elasticità dei container basata sulle applicazioni

Per gli ingegneri di piattaforma e DevOps
che cercano di rendere operativa la velocità
di immissione sul mercato garantendo al
contempo le prestazioni delle applicazioni



Contenuto

03

Executive summary

07

Un approccio basato sulle app

03

Una promessa di velocità,
agilità, elasticità e scalabilità

08

Accelerare la trasformazione
digitale durante una pandemia

05

Piattaforma e infrastrutture

Executive summary

Il tuo vantaggio competitivo dipende da quanto rapidamente le idee si trasformano in transazioni di business e dalla loro efficacia per i tuoi clienti. Il fattore abilitante è la tecnologia.

I container offrono la velocità, l'agilità, l'elasticità e la scalabilità che stanno modificando in modo radicale il modo in cui costruiamo, distribuiamo ed eseguiamo queste applicazioni. Aprono inoltre le porte a un mondo in cui le applicazioni possono essere eseguite davvero ovunque: gli aggiornamenti e le nuove funzionalità vengono distribuiti diverse volte al giorno, e la domanda dinamica e fluttuante dei carichi di lavoro può essere gestita con la fornitura di un'infrastruttura elastica, ovunque e in qualsiasi momento. Kubernetes è una piattaforma che permette alle organizzazioni di essere agili ed elastiche, ma non gestisce i compromessi su come garantire le prestazioni pur rimanendo efficienti.

Nonostante la semplicità e l'agilità offerte dalla containerizzazione, la piattaforma di orchestrazione fornisce solo un modo per gestire il ciclo di vita di questi servizi, distribuendoli e mantenendoli nel modo da te descritto.

Le piattaforme di container non assicurano in modo nativo che i servizi soddisfino gli SLO e non possono gestire le risorse in modo dinamico

Le politiche basate sulle soglie non risolvono le prestazioni continue: questo approccio non ha mai funzionato e, considerata la velocità delle modifiche nelle piattaforme di container, la scalabilità automatica attivata da trigger senza correlazione può finire per causare problemi. Un'infrastruttura elastica è la chiave per la fornitura delle prestazioni, ma ha bisogno di un'analytics automatizzata che gestisca continuamente domanda, fornitura e vincoli per soddisfare gli obiettivi di livello di servizio (SLO).

Questo white paper affronta i concetti chiave da considerare per l'adozione di una piattaforma di container come strumento di gestione della tua attività, nonché come proteggere questo investimento con un'automazione che assicuri le prestazioni

minimizzando al contempo i costi e garantendo la conformità. Sottolinea perché è necessaria un'analisi basata su un approccio top-down per la gestione automatica di una piattaforma Kubernetes che esegua i tuoi servizi. Costruire per la scalabilità multicloud fin dalle prime fasi del tuo percorso dà alla tua organizzazione IT la "memoria muscolare" operativa che trasformerà radicalmente come (e quando) produrrà più innovazione.

Una promessa di velocità, agilità, elasticità e scalabilità

Kubernetes offre elasticità, ma non garantisce automaticamente il rispetto e la garanzia degli SLO delle applicazioni.

Il successo nell'adozione della containerizzazione dipende da quanto sarai capace di dare agli sviluppatori l'abilità di cui hanno bisogno, l'elasticità richiesta per adattarsi su larga scala alla domanda in continua fluttuazione e la garanzia che l'applicazione funzionerà alla velocità richiesta.

Adottare un approccio cloud-native e scomporre le applicazioni in set di servizi distinti porterà a uno sviluppo e a una distribuzione delle applicazioni più agili. I container offrono i pacchetti che rendono i tuoi servizi portatili e scalabili. Kubernetes fornisce un framework e punti di controllo per eseguire le applicazioni e i servizi digitali. Ma per fornire una piattaforma performante e di livello enterprise alla tua azienda, dovrai comunque aggiungere funzionalità che apportino l'elasticità così abilitata per soddisfare e garantire gli SLO delle applicazioni.

Distribuisce più velocemente con CI/CD e feedback di produzione

La corretta metodologia CI/CD (implementazione e integrazione continua), basata sull'automazione, è la chiave per un time-to-market più rapido. Nel report Google Cloud State of DevOps 2021¹, gli intervistati hanno nominato gli importanti miglioramenti dovuti all'implementazione di CI/CD:

Frequenza di implementazione	Settimanale–mensile	Oraria–quotidiana
Modifica del tempo di esecuzione	Più di sei mesi	Meno di un'ora
Modifica del tasso di malfunzionamento	16%–30%	0%–15%

La velocità porta con sé anche l'esigenza di riuscire a gestire le costanti modifiche nella produzione e di avere un loop di feedback riguardo alle prestazioni dei servizi, oltre alla necessità di prevedere cosa è richiesto all'infrastruttura. L'obiettivo è capire come definire i tuoi SLO e ottenere dalla piattaforma un feedback su come configurare i container e l'infrastruttura per ridurre il rischio di problemi nelle prestazioni.

- Chi decide come allocare le risorse ai servizi? Come lo si decide? Stress test e benchmarking rispetto a SLO stabiliti e altro.
- Come si misurano le prestazioni? Esiste un loop di feedback nella pipeline CICD che garantisca che container e pod siano configurati adeguatamente?
- Come si fa a garantire che ci sia sempre una capacità sufficiente per le nuove implementazioni?

Trova le tue risposte IBM® Turbonomic

Opzioni	Limitazioni	Risposte di IBM® Turbonomic
Analizza manualmente i dati di utilizzo di container e pod per determinare le specifiche delle risorse.	<ul style="list-style-type: none"> – Impostazione della raccolta dati – Manodopera per l'analisi 	<ul style="list-style-type: none"> – Analisi top-down basata sulle applicazioni che determina come dimensionare i container – Feedback su CICD – Opportunità di ridurre le richieste quando non sono necessarie
Analizza manualmente i dati delle risorse da tutti i punti dello stack per determinare la capacità di produzione.	<ul style="list-style-type: none"> – Manodopera per raccogliere i dati da più origini – Manodopera per l'analisi 	Analisi basata sull'utilizzo per identificare la necessità di risorse in tutto lo stack

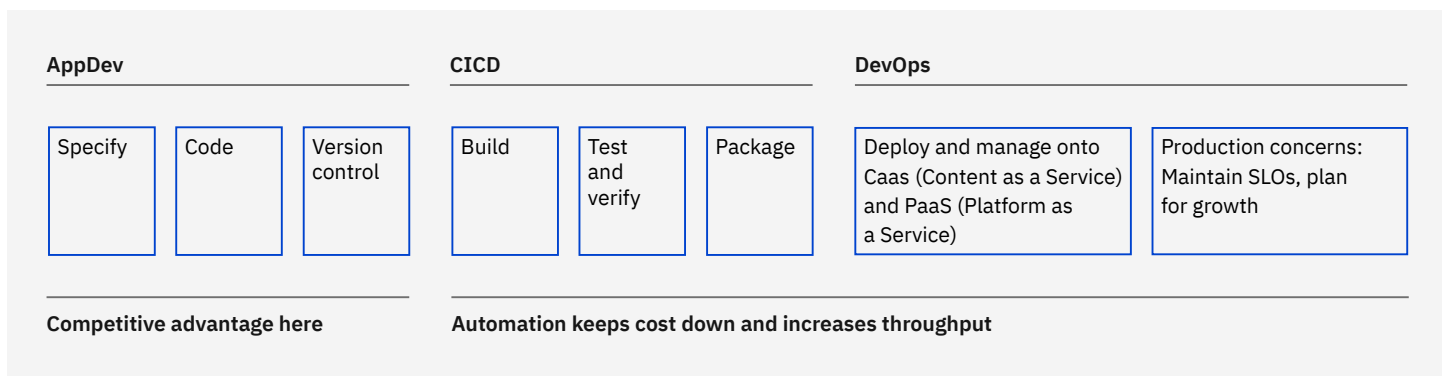


Figura 1. Processo per l'agilità delle applicazioni.

Piattaforma e infrastrutture

Perché hai bisogno di una gestione completa dello stack basata sulle app

Qualunque sia la piattaforma contenitore che sceglierai o la sua struttura sottostante (cloud privato, cloud pubblico, cloud ibrido, multicloud o hardware), le sfide operative del tuo PaaS (Platform-as-a-Service) saranno le stesse:

- Come determinare se c'è abbastanza capacità per soddisfare le richieste attuali e quelle in crescita?
- Come decidere quando attivare altri nodi di applicazione?
- Come decidere quando sospenderli?

- Come si gestisce il picco della domanda?
- Come utilizzare le risorse del cloud pubblico per il bursting?
- Come assicurare alta disponibilità (HA) e resilienza su tutto lo stack?
- Come applicare i vincoli di business?

L'elasticità fornita dalle piattaforme contenitore offre la possibilità di effettuare il provisioning per la somma della domanda media delle tue applicazioni invece che per la somma del picco. Per sfruttare questa abilità, offrire una piattaforma scalabile verso l'alto o verso il basso secondo le fluttuazioni della domanda richiede un software che prenda decisioni di resourcing continue, per garantire che le applicazioni abbiano le capacità di calcolo, archiviazione e rete di cui hanno bisogno, nel momento in cui ne hanno bisogno.

Trova le tue risposte IBM® Turbonomic

Opzioni	Limitazioni	Risposte di IBM® Turbonomic
Eseguito su service provider che forniscono gruppi di scalabilità automatica, come gruppi affiliati di servizi (ASG), set di disponibilità e altro.	<ul style="list-style-type: none">– Politiche basate sulle soglie– Impossibile scalare uno specifico nodo: tutti i nodi devono avere gli stessi vincoli, le stesse etichette eccetera	<ul style="list-style-type: none">– SLO basati su applicazioni top-down– Regola continuamente le risorse dell'infrastruttura per far fronte alla domanda di applicazioni– Scala continuamente in alto, in basso, in verticale e orizzontale i contenitori, pod e nodi adeguati– Posiziona continuamente i pod nei nodi adeguati
Analizza i dati risorse da tutti i punti dello stack per determinare la capacità di produzione.	<ul style="list-style-type: none">– Manodopera per raccogliere i dati da più origini– Manodopera per l'analisi	<ul style="list-style-type: none">– Analisi basata sull'utilizzo per identificare la necessità di risorse in tutto lo stack– Scala continuamente in alto, in basso, in verticale e orizzontale i contenitori, pod e nodi adeguati– Attiva continuamente le azioni per prevenire i colli di bottiglia

Operare per gli SLO su larga scala

Lo scopo della piattaforma contenitore è eseguire le applicazioni al livello di servizio desiderato per la tua attività. Devi garantire sempre le prestazioni, anche di fronte all'aumento del numero di applicazioni. Solitamente, i clienti impiegano più di 12 mesi per le prime 1-3 applicazioni. Per quelle successive, grazie al vantaggio delle competenze e delle best practice acquisite, possono impiegare altri 6-12 mesi. Quando i settori di attività capiscono cosa è possibile, l'aumento del numero di singoli servizi da gestire supera le capacità umane. Anche se hai costruito servizi senza stato sfruttando la natura effimera dei contenitori, qual è la tua tolleranza verso il calo delle prestazioni dell'esperienza dei tuoi utenti finali? Cosa puoi fare per gestire non solo la domanda, ma il tasso di modifica crescente? La risposta è l'automazione, ottenuta tramite azioni basate sull'analisi dei compromessi riguardanti il numero di istanze dei servizi richieste per garantire gli SLO, la configurazione delle dimensioni e del posizionamento del carico di lavoro e la messa a disposizione delle risorse conformi dall'infrastruttura.

Le soglie non risolvono il problema

Una piattaforma contenitore garantirà che tu abbia a disposizione un numero minimo di servizi: se uno va in crash, tenterà di riattivarlo. Ma se vuoi assicurarti di fornire un'esperienza dell'utente di buon livello, il sistema dovrà rispondere prima che si verifichino cali delle prestazioni o crash. Puoi impostare la scalabilità automatica nativa orizzontale per soddisfare la domanda, ma dovrai decidere quali sono le metriche che esprimono al meglio le risorse richieste, configurare soglie e limiti superiori e inferiori, testare ed estrapolare se funzioneranno sotto la domanda di produzione e poi ripetere la procedura per tutti i servizi implementati. E ora immagina di avere più di 100 servizi per una sola applicazione. Tutte queste

politiche non saranno correlate fra loro. Come garantire che l'aggiunta di altri pod di un servizio non crei congestione in un'altra area? Stai clonando un pod mal configurato che ha bisogno prima di scalare verticalmente? Come gestire il sovraccarico dei nodi, individuare gli elementi adiacenti rumorosi e identificare le risorse allocate inutilizzate che potrebbero essere liberate per soddisfare questa domanda?

Inoltre, configurare contenitori, pod e programmi di scalatura automatici dei pod orizzontali (HPA) o politiche per i programmi di scalatura automatici dei cluster non è un lavoro che si fa una volta e basta. Gli sforzi di ipotesi devono sempre essere monitorati e, se non lo sono, vanno ridefiniti. Cosa potrebbero fare i tuoi team con il tempo risparmiato non dovendo impostare e reimpostare manualmente queste soglie?

La corretta impostazione di queste configurazioni ha implicazioni dirette per un rollout di successo della tua strategia di trasformazione digitale. Basta qualche implementazione sbagliata per rallentare notevolmente l'adozione delle piattaforme e dei sistemi che stai costruendo. E impiegare troppo tempo o manodopera nella configurazione manuale di questi punti di controllo può ostacolare in maniera significativa la capacità della tua organizzazione di diventare platform-first. Puoi concederti il lusso di questo ritardo? Quello che serve è un sistema di controllo che gestisca i compromessi di tutte le risorse e definisca le richieste e i limiti di scalabilità verticale del container, il numero di pod richiesto e le decisioni di posizionamento per ridistribuire i pod e gestire le risorse del cluster utilizzando un unico motore di analitica.

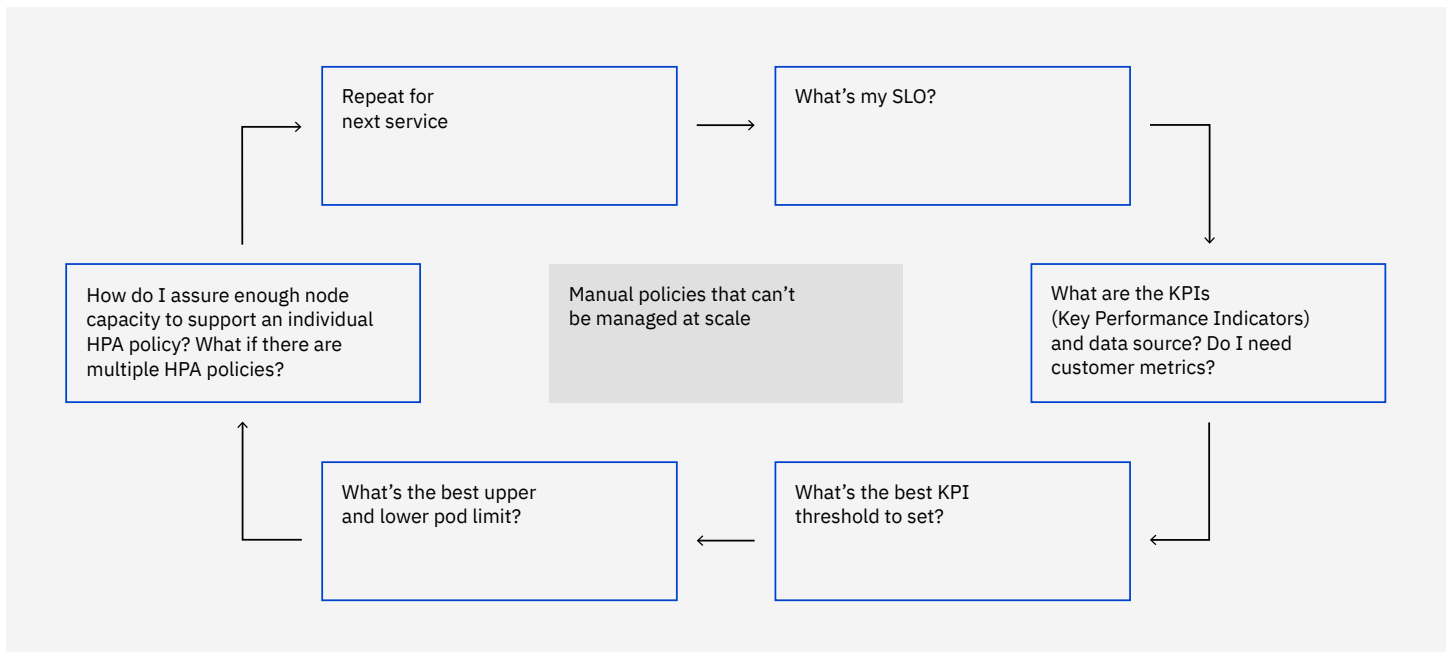


Figura 2. Le politiche manuali non possono essere gestite su larga scala

Trova le tue risposte IBM® Turbonomic

Opzioni	Limitazioni	Risposte di IBM® Turbonomic
Politiche basate sulle soglie HPA per l'attivazione dello scaling in e out dei pod	<ul style="list-style-type: none">– Configura per servizio– In base alla media di tutti i pod per il servizio– KPI e soglie definiti manualmente, con limiti superiori e inferiori dei pod	<ul style="list-style-type: none">– SLO basati su applicazioni top-down– Utilizza i dati sui tempi di risposta per ottenere la scalabilità orizzontale dei servizi e soddisfare gli SLO– Scala continuamente in alto e in basso, in verticale e orizzontale i contenitori, pod e nodi adeguati– Posiziona continuamente i pod nei nodi adeguati– Regola continuamente le risorse dell'infrastruttura per far fronte alla domanda di applicazioni
Politica di scalatura automatico verticale del pod basata sulle soglie per scalare verticalmente i container	<ul style="list-style-type: none">– Deve essere definito per tutti i servizi– Progetto beta: utilizza a tuo rischio e pericolo– Non accede alla capacità di agire del nodo	
Lascia crashare i pod per ridistribuirli su un nodo migliore	Scarsa esperienza dell'utente per le transazioni del pod pronto ad andare in crash	
Le soluzioni di osservabilità di Prometheus raccolgono e consolidano i dati	<ul style="list-style-type: none">– Non fornisce l'analisi dei dati– Non fornisce azioni	

Un approccio basato sulle app

Gli SLO delle applicazioni devono guidare l'infrastruttura

La containerizzazione delle applicazioni mission critical è un investimento che presenta diversi vantaggi ma, per poter cogliere il beneficio di velocità, elasticità e portabilità, hai bisogno di un software in grado di prendere le giuste decisioni di resourcing al momento giusto, 24 ore su 24, 7 giorni su 7, 365 giorni all'anno. Altrimenti, la complessità ti rallenterà.

IBM® Turbonomic Application Resource Management unisce le tue applicazioni mission-critical alla piattaforma Kubernetes e all'infrastruttura sottostante ovunque vengano eseguite le tue applicazioni. Il software, basato sulla domanda delle applicazioni in tempo reale e in grado di tenere in conto dei vincoli e delle interdipendenze a tutti i livelli dello stack (da quello logico a quello fisico) determina l'azione giusta da intraprendere al momento giusto per garantire che le applicazioni ottengano sempre esattamente ciò di cui hanno bisogno per funzionare. Eseguito in tempo reale, pianificato o come parte della tua pipeline DevOps.

Dimensionamento intelligente: come si dimensionano i container?

- Automatizza con la distribuzione: esecuzione e persistenza si ridimensionano come parte della pipeline, ad esempio YAML, Jenkins e altri.
- Automatizza in tempo reale: esegui in modo dinamico tramite Kubernetes.

Posizionamento continuo: dove devi spostare i pod?

Su quali nodi?

- Esegui in modo dinamico e in tempo reale tramite Kubernetes. Solo per servizi senza stato non distruttivi.

Scalabilità dinamica: quando hai bisogno di scalare all'esterno o all'interno del cluster? Di quanto?

- Esegui dinamicamente la scalabilità del cluster in tempo reale attraverso l'infrastruttura come codice o l'API Kubernetes Cluster.

Scalabilità basata sugli SLO: quando devi scalare all'esterno o all'interno dei pod per soddisfare gli SLO dei tempi di risposta delle applicazioni? Di quanto?

Prerequisiti per la scalabilità basata sugli SLO:

- Le applicazioni sono progettate per microservizi senza stato orizzontali.
- Hanno una definizione e un'origine dei dati degli SLO che Kubernetes non fornisce.

Cosa significa questo tipo di automazione intelligente per te, per i tuoi team e per il tuo business? Di seguito sono elencati i benefici esclusivi offerti da IBM® Turbonomic, che tu esegua Kubernetes on-premise, sul cloud, su server bare metal o in qualsiasi loro combinazione.

“Cruise control” per le tue app: i tuoi team impostano SLO di tempi di risposta. Il software basato su AI assicura che la piattaforma e l'infrastruttura sottostante forniscano sempre le risorse richieste per rispettare questi SLO ovunque le app siano eseguite.

Minimizza il lavoro manuale: sviluppatori, DevOps e site reliability engineers (SRE) non hanno bisogno di impostare soglie, vincoli o politiche di scalatura automatica. Il software prende le giuste decisioni relative alle risorse al posto tuo, fornendo azioni effettivamente automatizzabili.

Non spendere troppo in capacità: non c'è bisogno di affidarsi agli sviluppatori per le decisioni di resourcing. Spesso ricorrono all'overprovisioning tanto per stare sul sicuro, vero? Il nostro software, invece, determina esattamente quali servizi di risorse sono necessari, il tutto in base alla domanda delle applicazioni.

Accelera DevOps con sicurezza: aumenta con sicurezza la frequenza e la scala delle distribuzioni. La nostra analisi si integra con i tuoi flussi di lavoro DevOps, assicurando che i nuovi servizi e quelli già esistenti funzionino sempre.

Pianifica la crescita con più facilità: simula l'onboarding di nuovi servizi con il nostro software. Determina esattamente quanti altri nodi ti servono per supportare la nuova crescita.

Clienti in primo piano

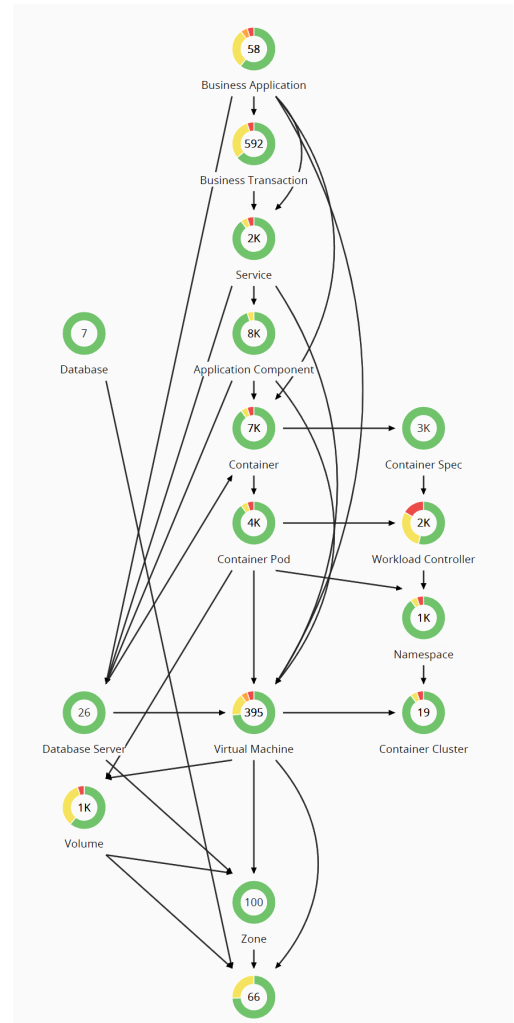
Accelerare la trasformazione digitale durante una pandemia

Il resourcing dinamico di IBM® Turbonomic all'interno della piattaforma Kubernetes e l'infrastruttura sottostante hanno mantenuto bassi i tempi di risposta.

Questo cliente è una delle più grandi compagnie di assicurazioni del Sud America, e conta oltre 6 milioni di clienti. Il suo approccio standard di settore alla gestione del resourcing di ambienti esistenti e di nuova generazione stava rallentando la trasformazione digitale e la risposta dell'azienda alla pandemia.

L'automazione di IBM® Turbonomic ha mantenuto i tempi di risposta bassi anche durante il picco della domanda nel periodo delle feste

Questo cliente ha un'app aziendale che si integra con una delle compagnie aeree low-cost più grandi della regione. Tramite quest'app è possibile prenotare l'assicurazione di viaggio, e il picco che vediamo nella Figura 3 è relativo alle vacanze di Pasqua. Anche se la domanda sull'app aumentava, il resourcing dinamico di IBM® Turbonomic all'interno della piattaforma Kubernetes e dell'infrastruttura sottostante ha permesso di mantenere bassi i tempi di risposta.



Response Time

69 Business Applications (@tw0jb_10sjqc)

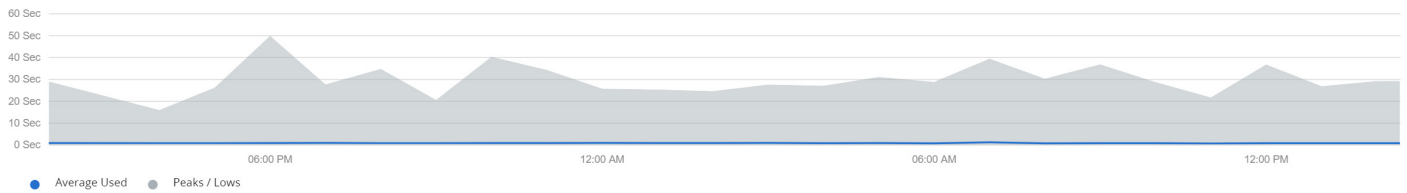


Figura 3. Una vista completa dello stack della singola applicazione aziendale e dei suoi tempi di risposta, con i tempi di risposta dell'automazione mantenuti bassi anche durante i picchi di domanda

57 applicazioni mission-critical

- Esempio, il GPS in auto: denuncia il furto del veicolo, chiede un preventivo per nuove polizze eccetera
- ~3000 pod (composti da ~7000 container)
- Collegato a Dynatrace

Automatizzato

- Ridimensionamento dei container (staging)
- Posizionamento continuo (tutto)

~70%

di riduzione dei ticket

Informazioni su IBM® Turbonomic, una società IBM

IBM® Turbonomic Application Resource Management fornisce un software di gestione delle risorse delle applicazioni (ARM) che i clienti possono utilizzare per garantire le prestazioni delle applicazioni e la governance, effettuando il resourcing in modo dinamico in ambienti ibridi e multcloud. IBM® Turbonomic network performance management (NPM) fornisce soluzioni moderne di monitoraggio e analisi che garantiscono prestazioni di rete continue su larga scala in rete multivendor per aziende, fornitori e provider di servizi gestiti.

Per saperne di più sull'automazione intelligente di IBM® Turbonomic, visita ibm.com/it-it/cloud/turbonomic o parla con un [rappresentante IBM](#).

© Copyright IBM Corporation 2022

IBM Italia S.p.A.
Circonvallazione Idroscalo
20054 Segrate (Milano)
Italia
IBM Corporation
New Orchard Road
Armonk, NY 10504

Prodotto negli Stati Uniti d'America
Marzo 2022

IBM e il logo IBM sono marchi o marchi registrati di International Business Machines Corporation, negli Stati Uniti e/o in altri paesi. Altri nomi di prodotti e servizi potrebbero essere marchi registrati di IBM o di altre aziende. Un elenco aggiornato dei marchi registrati IBM è disponibile su ibm.com/it-it/trademark.

IBM® Turbonomic è un marchio registrato di Turbonomic Inc., una IBM Company.

Le informazioni contenute nel presente documento sono aggiornate alla data della prima pubblicazione e possono essere modificate da IBM senza preavviso. Non tutte le offerte sono disponibili in ogni Paese in cui opera IBM.

Gli esempi citati relativi a clienti sono presentati unicamente a scopo illustrativo. Gli attuali risultati in termini di prestazione possono variare a seconda delle specifiche configurazioni e delle condizioni operative. La valutazione e la verifica del funzionamento di qualsiasi altro prodotto o programma con prodotti e programmi IBM sono responsabilità dell'utente. LE INFORMAZIONI RIPORTATE NEL PRESENTE DOCUMENTO SONO DA CONSIDERARSI "NELLO STATO IN CUI SI TROVANO", SENZA GARANZIE, ESPLICITE O IMPLICITE, IVI INCLUSE GARANZIE DI COMMERCIALIZZABILITÀ, DI IDONEITÀ A UN PARTICOLARE SCOPO E GARANZIE O CONDIZIONI DI NON VIOLAZIONE. I prodotti IBM sono coperti da garanzia in accordo con termini e condizioni dei contratti sulla base dei quali vengono forniti.

¹ State of DevOps 2021, Google Cloud, 2021

