INTERCONNECT
Data & Cloud Center

interconnect.nl

Whitepaper

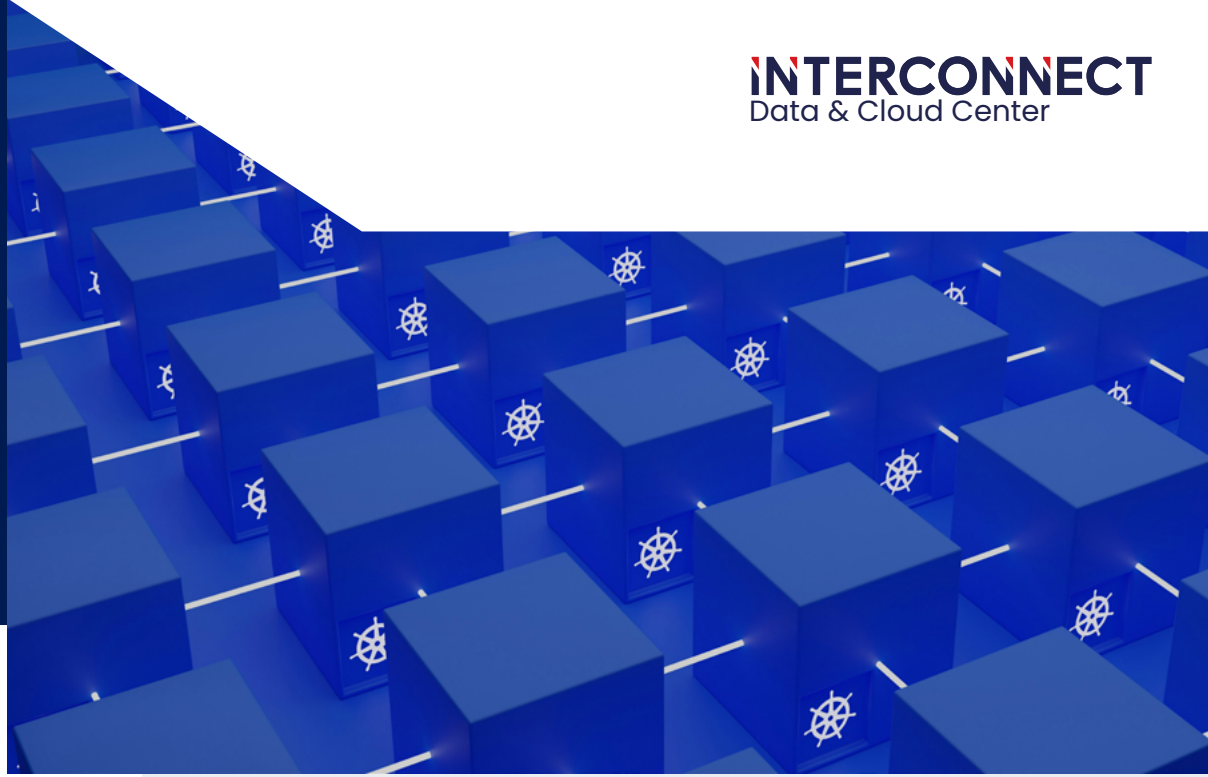
Managed Kubernetes voor slimme en schaalbare applicaties

T 073 – 88 000 00
M solutions@interconnect.nl

Het Sterrenbeeld 55
5215 MK 's-Hertogenbosch

Inleiding

MANAGED KUBERNETES VOOR SLIMME EN SCHAALBARE APPLICATIES



Digitale transformatie en flexibiliteit zijn vandaag de dag onmisbaar voor moderne organisaties. Tegelijkertijd groeit de behoefte aan krachtige platformen die applicaties efficiënt beheren, eenvoudig opschalen én voldoen aan strenge eisen op het gebied van datasoevereiniteit. Kubernetes heeft zich ontwikkeld tot de standaard voor containerorkestratie en biedt een robuuste infrastructuur voor schaalbaarheid, hoge beschikbaarheid en optimale resource-efficiëntie. Zeker nu controle over data en naleving van wet- en regelgeving steeds belangrijker worden, zoeken organisaties naar oplossingen die zowel technologische wendbaarheid als datacontrole garanderen.

Managed Kubernetes van Interconnect, in samenwerking met ACC ICT, voegt hier een waardevolle laag aan toe door complexe beheer- en onderhoudstaken te automatiseren. Dit stelt organisaties in staat zich volledig te richten op innovatie en ontwikkeling, in plaats van op het dagelijks beheer van de infrastructuur.

Omdat Interconnect al zijn datacenters in Nederland heeft en onder Europese wet- en regelgeving opereert, behouden klanten volledig zeggenschap over hun data. Dit biedt de zekerheid dat flexibiliteit en schaalbaarheid hand in hand gaan met gegevensbescherming en compliance.

Een belangrijk voorbeeld van automatisering is autoscaling: het dynamisch aanpassen van capaciteit op basis van de actuele belasting. Deze whitepaper legt de drie pijlers van autoscaling uit—Vertical Pod Autoscaler (VPA), Horizontal Pod Autoscaler (HPA) en de cluster-autoscaler—en laat zien hoe deze componenten samenwerken om resources efficiënt te beheren, kosten te optimaliseren en de schaalbaarheid van zowel stateless als stateful applicaties te maximaliseren.

Wilt u meer informatie of persoonlijk advies? Neem dan contact op met **solutions@interconnect.nl**

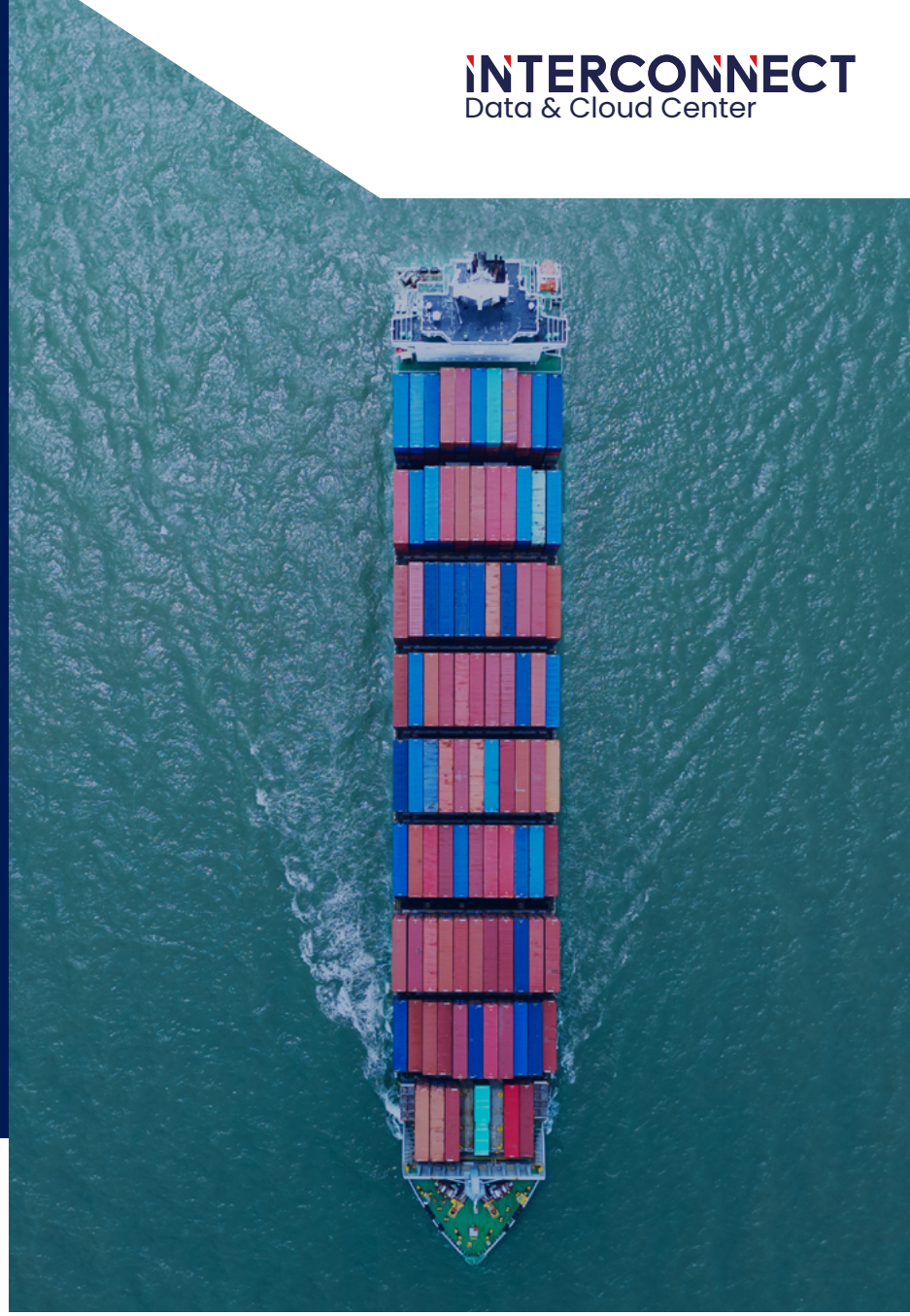
Resource Scaling met de Vertical Pod Autoscaler (VPA)

Resource Scaling met de Vertical Pod Autoscaler (VPA) is een methode waarbij Kubernetes automatisch meer of minder CPU en geheugen toewijst aan een pod op basis van het daadwerkelijke gebruik.

In een standaard Kubernetesconfiguratie krijgt elke pod een vaste hoeveelheid CPU en geheugen toegewezen, bijvoorbeeld 1 CPU en 2GB RAM. Zonder autoscaling zou een pod dus altijd vastzitten aan die resources, ongeacht of hij die capaciteit wel of niet nodig heeft.

De Vertical Pod Autoscaler (VPA) biedt hier een oplossing door een pod dynamisch meer resources toe te wijzen als dat nodig is, tot het maximum dat een node kan leveren. Dit is met name handig voor stateful applicaties, zoals relationele databases (bijvoorbeeld MySQL en PostgreSQL) of legacy monolithische applicaties. Deze applicaties kunnen lastig horizontaal geschaald worden omdat ze intern een state bijhouden.

Bij verticaal schalen verandert de pod zelf, wat het proces relatief simpel maakt. Maar er is wel kans op downtime als de node tijdelijk niet genoeg resources heeft of als Kubernetes de pod naar een andere node moet verplaatsen.



Horizontal scaling met de Horizontal Pod Autoscaler (HPA)

Horizontaal schalen met de Horizontal Pod Autoscaler (HPA) houdt in dat Kubernetes automatisch extra pods toevoegt of verwijdert om de prestaties van een stateless applicatie te optimaliseren. Bij toenemende vraag worden extra pods ingezet om meer resources te gebruiken, terwijl bij afnemende vraag het aantal pods wordt verminderd om resources efficiënter in te zetten.

De Horizontal Pod Autoscaler (HPA) is ideaal voor cloud-native applicaties die ontworpen zijn om stateless te zijn. Stateless applicaties houden geen interne data bij, of regelen de opslag van deze data extern. Hierdoor kunnen er eenvoudig nieuwe pods opstarten wanneer extra capaciteit nodig is, zonder dat dit effect heeft op de gebruikers.

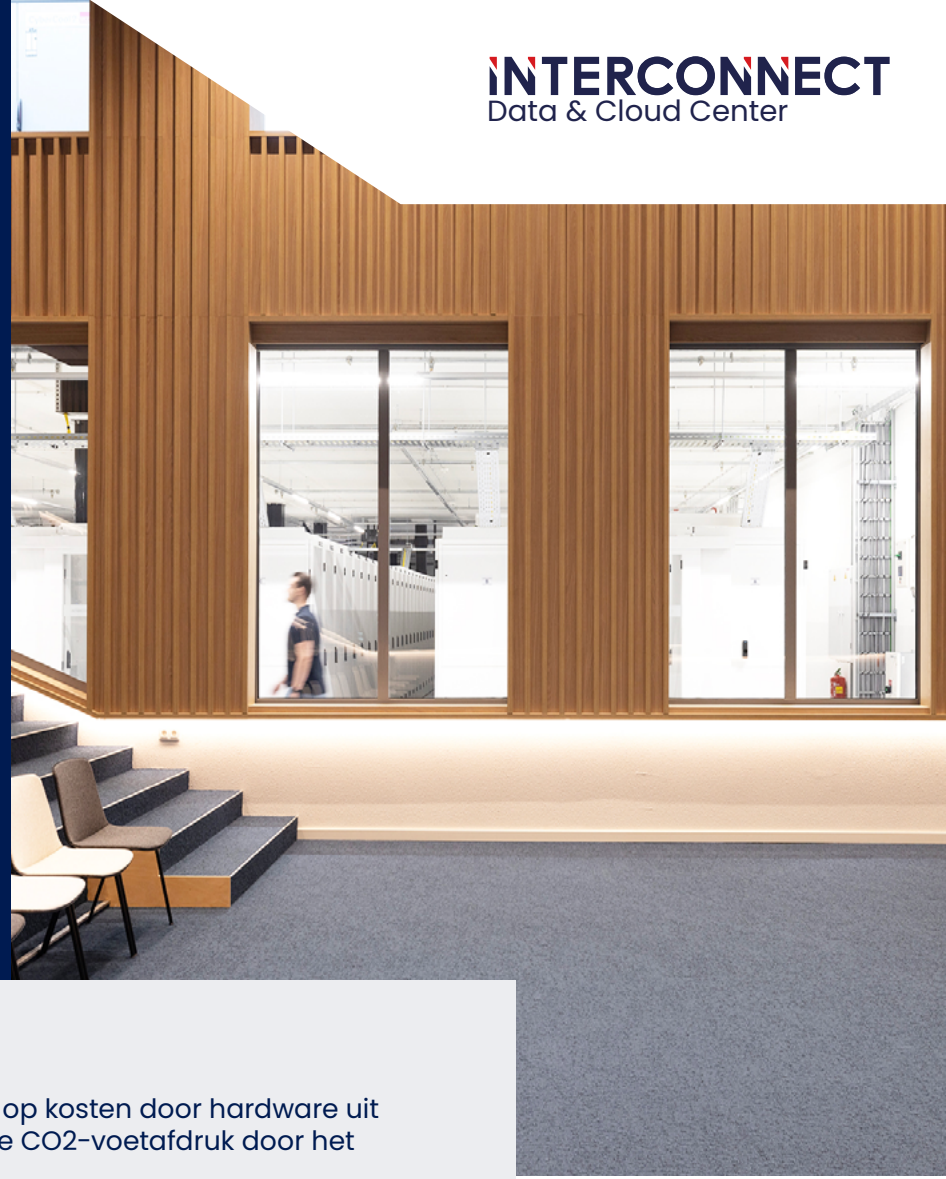
Deze horizontale schaalmethode kijkt niet naar de resources van individuele nodes, maar naar het cluster als geheel. Het doel is om resources te zien als één grote pool waar de pods uit kunnen putten. Hierdoor kunnen deze onafhankelijk van elkaar werken, terwijl het systeem blijft functioneren; zelfs als een van de pods uitvalt.

Cluster-autoscaler: dynamische node scaling

Wanneer er onvoldoende hardware beschikbaar is in het cluster om nieuwe pods op te starten, komt de cluster-autoscaler in actie. De cluster-autoscaler scant voortdurend of alle aangevraagde pods een plek kunnen vinden in het cluster. Zo niet, dan voegt deze automatisch nieuwe nodes toe totdat er genoeg capaciteit beschikbaar is.

Omgekeerd, wanneer er overcapaciteit is, kan de cluster-autoscaler overbodige nodes verwijderen. Dit bespaart kosten en zorgt voor een efficiënter gebruik van compute-resources.

Ook kan het prioriteren van bepaalde pods nuttig zijn; door resources van minder belangrijke pods tijdelijk vrij te maken, kan het schalen nog sneller verlopen voor de pods die belangrijker zijn voor de dienstverlening.



Milieu- en kostenvoordelen

Met de cluster-autoscaler kan een organisatie significant besparen op kosten door hardware uit te schakelen wanneer deze niet nodig is. Bovendien reduceert het de CO2-voetafdruk door het energieverbruik te verlagen.

Dankzij Kubernetes' mogelijkheden voor het combineren van autoscaling-opties—zoals het afstemmen van Pod Disruption Budgets en Pod Priority—is het mogelijk om stabiele, automatisch schaalbare applicaties te draaien zonder de performance op cruciale momenten te belemmeren.

De andere kant van de medaille

Autoscaling is een krachtig hulpmiddel om IT-resources efficiënt te beheren, maar zonder goede configuratie brengt het ook risico's met zich mee. We kijken eerst naar de problemen die ontstaan zonder autoscaling, waarna vervolgens de valkuilen van verkeerd ingestelde autoscaling worden toegelicht.

Overprovisioning: te veel betalen voor ongebruikte resources

Zonder autoscaling wordt vaak een vaste hoeveelheid capaciteit gereserveerd, ongeacht de daadwerkelijke behoefte. Dit leidt tot verspilling, vooral bij wisselende belasting. Denk aan ontwikkelomgevingen die in het weekend onnodig blijven draaien, terwijl er niemand aan het werk is. Het gevolg? U betaalt voor resources die u niet gebruikt. Vanuit zakelijk perspectief is dit simpelweg zonde van het budget.

Underprovisioning: slechte prestaties tijdens piekbelasting

Wanneer er niet genoeg capaciteit beschikbaar is, ontstaan er problemen bij piekbelasting. Applicaties worden traag of crashen zelfs, wat directe gevolgen heeft voor gebruikerservaringen en omzet. Prestaties zijn een terugkerend aandachtspunt in IT, en de angst om te veel te betalen zorgt vaak voor een te voorzichtig resourcemanagement. Autoscaling biedt een oplossing door automatisch extra capaciteit beschikbaar te stellen wanneer dat nodig is.

Handmatig schalen: te traag en arbeidsintensie

Handmatig schalen kan een alternatief lijken, maar dit is verre van ideaal. Het vereist uitgebreide monitoring en een team dat 24/7 beschikbaar is om in te grijpen. Dit proces is te traag om plotselinge pieken, zoals op Black Friday, op te vangen. De kans op fouten en vertragingen is groot, wat leidt tot onnodige risico's.

De risico's van verkeerd geconfigureerde autoscaling

'Thrashing': constant op- en afschalen

Slecht ingestelde thresholds kunnen ervoor zorgen dat het systeem continu opschaalt en weer afschaalt. Dit 'thrashen' veroorzaakt overbodige overhead, zoals het herverdelen van verkeer en het opnieuw opstarten van pods. Het gevolg is dat het systeem meer bezig is met zichzelf organiseren dan met het verwerken van gebruikersverkeer.

Scaling lag: te traag reageren op veranderingen

Als metrics of thresholds niet goed zijn ingesteld, reageert het systeem te laat. Een veelvoorkomend voorbeeld is het gebruik van CPU-belasting als enige metric. Tegen de tijd dat de CPU piekt en opschaling start, zijn de nieuwe resources vaak te laat beschikbaar. Gebruikers ervaren dan al performanceproblemen.

Underprovisioning: slechte prestaties tijdens piekbelasting

Overmatig agressieve scaling rules kunnen leiden tot onnodig hoge kosten. Bijvoorbeeld, een kleine verkeerspiek kan een disproportionele scaling-reactie uitlokken, waarbij veel meer resources worden geprovisioneerd dan nodig. Of een fout in de baseline-capaciteit kan ervoor zorgen dat het systeem nooit naar een lager resourcegebruik schaalt, zelfs niet tijdens rustige uren.

Onjuiste plaatsing van schaling en pod

Kubernetes biedt veel configuratiemogelijkheden, zoals pod priorities en topology spread constraints. Onjuist gebruik hiervan kan leiden tot problemen zoals ongelijkmatige verdeling van workloads, nodes die te snel vol raken, of kritieke applicaties die concurreren met minder belangrijke processen. Het goed configureren van autoscaling vereist zowel technische kennis als inzicht in de architectuur van uw applicaties.

Managed Kubernetes samengevat

Zonder een goed geconfigureerde autoscaling riskeert u inefficiëntie, instabiliteit en onverwachte kosten. Managed Kubernetes biedt hiervoor een oplossing met krachtige tools zoals Vertical Pod Autoscaler, Horizontal Pod Autoscaler en de cluster-autoscaler. Deze maken het mogelijk om resources optimaal af te stemmen op applicatiebehoeften, wat resulteert in betere prestaties, kostenbesparing en een duurzame inzet van compute-resources.

Bovendien blijven organisaties met Managed Kubernetes van Interconnect en ACC-ICT in volledige controle over hun data. Dit is een belangrijk voordeel in het kader van datasoevereiniteit. Door infrastructuur binnen Nederland te hosten en te voldoen aan Europese regelgeving, kunnen bedrijven schalen zonder concessies te doen aan compliance of controle. Hiermee creëren organisaties een flexibele, efficiënte én soevereine basis voor groei in een dynamische cloudomgeving.

MANAGED KUBERNETES

INTERCONNECT
Data & Cloud Center

Uw developers

Bouwen, beheren en implementeren van containerized applicaties

Uw operations engineer, direct ondersteund door ons

Applicatie specifieke beheertaken, zoals het netwerkbeleid

Geavanceerdere Kubernetes optimalisaties, zoals monitoring en alerting

Basis Kubernetes beheer en onderhoud

Volledig beheerd voor u

Operating systeem

Fysieke en virtuele infrastructuur

OVER INTERCONNECT

Interconnect is een honderd procent Nederlands Data- en Cloud Center dat uw IT-infrastructuren toekomstbestendig maakt. Dit doen wij sinds 1995 door IT-professionals te ondersteunen met het opzetten van betrouwbare IT-oplossingen en leveren we een veelzijdig spectrum aan datacenterdiensten.

Wij beheren twee uitstekend beveiligde en kwalitatief hoogwaardige datacenters (Tier 3) in Noord-Brabant. De datacenters, gevestigd in Eindhoven en 's-Hertogenbosch, bieden volledig Nederlandse datasoevereiniteit, leveren een uptimegarantie van 99,9% en hebben samen een vloeroppervlak van bijna 8.000 vierkante meter.

De twee vestigingen vormen samen het grootste datacenter van Zuid-Nederland. Met Managed Kubernetes van Interconnect nemen wij het beheer van de onderliggende infrastructuren voor onze rekening, waardoor ontwikkelaars zich volledig kunnen focussen op het deployen van software die waarde toevoegt aan uw business.

NEEM HIER CONTACT OP MET ONZE SPECIALISTEN

Heeft u vragen over Managed Kubernetes van Interconnect?

Vrijblijvend advies nodig?
De beste oplossing voor uw organisatie hangt af van verschillende factoren. Kunt u ondersteuning gebruiken bij het aanscherpen van uw IT-strategie? Onze experts staan voor u klaar. Ze voorzien u graag van een gepersonaliseerd en deskundig advies. Bel 073-88 000 00, stuur een e-mail naar **solutions@interconnect.nl** of vul het contactformulier op onze website in.