7. APPLICATION MODERNIZATION

NEXT OnAir 2020
베스핀글로벌 세션브리핑

Optimize cost to performance on Google Kubernetes Engine

Google Kubernetes Engine에서 가격 대비 성능 최적화하기

Jerzy Foryciarz
Product Manager, Google Cloud

Joel Meyer
Chief Archiect, OpenX

Ivan Gusev
Principal Infrastrcture Architect, OpenX

#Kubernetes #Anthos #DevOps #Container Infrastructure

Week7. Application Modernization

Optimize cost to performance on Google Kubernetes Engine

Google Kubernetes Engine에서 가격 대비 성능 최적화하기

GKE(Google Kubernetes Engine)의 자동 확장 기능성

✔ **CPU, 메모리 등 리소스 사용량을 기준으로 애플리케이션 배포를 자동으로 확장하거나 축소합니다.**

✔ **Pod (파드): Pod는 Kubernetes에서 가장 작고 가장 기본적인 배포 가능한 객체입니다.**
파드는 클러스터에서 실행되는 프로세스의 단일 인스턴스를 나타냅니다.

✔ **중요한 2가지 요소: 워크로드 & 인프라**

- 워크로드
  - **HPA (Horizontal Pod Autoscaling)** : 사전에 설정해 둔 실제 리소스 사용량, 커스텀 측정항목, 외부 측정항목 기준에 따라 Pod를 자동으로 확장 및 축소시켜줍니다.
  - **VPA (Vertical Pod Autoscaling)** : 별도의 임계치값을 설정하지 않아도 CPU 및 메모리 요청과 한도의 값을 추천하거나 값을 자동으로 업데이트하여 Pod를 자동으로 확장 및 축소시켜줍니다.
- 인프라
  - **CA (Cluster Autoscaler)**: 워크로드의 요구에 따라 GKE 클러스터의 노드 풀 크기를 자동으로 조절시켜줍니다.
  - **NAP (Node Auto Provisioning)** : 사용자를 대신하여 노드 풀 집합을 자동으로 관리해 줍니다.

✔ **해당 4가지 확장성 기능 (HPA, VPA, CA, NAP)을 통해 GKE가 비용 및 성능을 최적화 합니다.**

Case (OpenX)

OpenX는 Ad exchange 기업이며 다양한 client에서 150-400 milliseconds 차이로 몇억개 ad request가 들어오는데 이러한 요청들을 처리하는 서비스 시스템 비용이 총 인프라 비용의 반이나 차지합니다.

✔ **고객 환경**

- 하루에 1,000억 이상의 요청사항
- 1천 5백만의 코드 라인 15000000
- 1만 5천개의 서버
- 5개 Region

✔ 클라우드 마이그레이션 전에는

RPM으로 포장하여 physical 하드웨어에 배포를 했지만 on-premise 인프라는 효율성이 제한되어 구글 클라우드 마이그레이션을 통해 효율성, 확장 및 안정성을 추구하였습니다.

✔ OpenX의 목표

- GKE 통해 traffic pattern기반으로 컴퓨팅 확장 및 최적화된 서비스 구현
- 파드(Pod)와 클러스터(Cluster)에 규모 최적화

방법 1. HPA(Horizontal Pod Autoscaling) 환경을 설정 및 커스텀

- HPA는 수립한 측정항목을 기준으로 Pod에서 실행되는 애플리케이션을 확장 처리하기 위한 기능입니다.
- CPU 사용률 또는 기타 커스텀 측정항목(예: 초당 요청 수)을 구성할 수 있습니다.
- OpenX는 코드 변동이 많아서 커스텀 측정보다 CPU 사용률 측정이 더 적합했습니다.
- 만약 워크로드가 IO 또는 네트워크 기반이면 거스텀 측정항목을 추천합니다.

방법 2. CA(Cluster Autoscaler) 환경 설정

- GKE의 CA는 워크로드 요구사항에 따라서 노드(Node) 자체를 동적으로 늘리거나 줄여주는 기능입니다.
- 수동으로 노드를 추가, 제거하거나 노드 풀을 과도하게 프로비저닝할 필요가 없습니다.
  대신 노드 풀의 최소 및 최대 크기를 지정하면 나머지는 자동으로 지정됩니다.
- CA를 사용했을때 OpenX의 워크로드에 더 적합한 VM을 매칭해줘 메모리와 CPU 사용률을 축소했습니다.
- 또한, CA에서 빈 노드를 빠르게 삭제할 수 있으며, 클러스터 축소 프로세스를 더 신속하게 수행 할 수 있습니다.

방법 3. Rightsizing Pod 요청 또는 VPA(Vertical Pod Autoscaling) 환경 설정

- VPA는 시간별로 Pod를 관찰하며 각 Pod에 필요한 최적 CPU 및 메모리 리소스를 점진적으로 찾아냅니다.
- 안정성과 비용 효율성을 위해서는 올바른 리소스 설정이 핵심입니다.
- 만약 리소스가 너무 크면 낭비로 인해 비용이 더 커집니다. 반대로, 리소스가 너무 작으면 애플리케이션이 제한될 수 있습니다.
- VPA는 리소스 사용량을 관찰하고, 관찰한 데이터를 통해 권장사항들을 제시하고, 그리고 해당 권장사항들을 자동으로 적용할 수 있습니다.
  예). Pod 사용률이 지속적으로 낮은 경우에는 VPA가 축소 작업을 합니다.

방법 4. CPU Manager 사용

- 애플리케이션들이 동일한 CPU 코어에서 실행되어야 할때 CPU Manager를 사용해야 합니다.
- CPU Manager는 GKE 리소스 요청 외에 CPU 할당에 대한 추가적인 개런티를 제공합니다.
- CPU Manager는 다른 파드들이 CPU 코어 사용을 방지하여 context switching을 최소화 합니다.

방법 5. BinPacking 최적화

- 모든 Pod를 일정한 사이즈로 맞추고 각각 노드가 해당 Pod 사이즈에 맞게 설정해야 합니다.
- BinPacking 최적화 할때 GKE의 DaemonSet와 Kernel 요소들을 고려해야 합니다.
- 또한, GKE의 sidecar과 init container의 요청들이 스케줄링 프로세스에 영향을 미치니 해당 요소들도 고려해야합니다.
- Pod 요청이 CPU의 일부만 차지하고 노드는 상대적으로 크게 설정하면 BinPacking이 최적화 됩니다.

베스핀글로벌 인사이트

구글 클라우드 Kubernetes는 고객사를 위해 비용 및 성능을 최적화하는데 다양한 솔루션을 제공합니다. 점점 더 커지는 Cloud IT 인프라 시대에 Kubernetes는 기업들이 애플리케이션 배포 및 실행과정의 효율성을 높이는 환경을 설계합니다. 베스핀글로벌은 GKE에 존재하는 다양한 기능들을 고객사의 cloud migration에 적용하여 GCP 비용을 최소화 시키면서 동시에 성능들을 최대화 시킬 수 있습니다.

베스핀글로벌은 Google Cloud를 가장 잘 아는 전문가이며, Google Cloud의 프리미어 파트너이자,
국내 최초 Google Cloud의 MSP(Managed Service Provider)입니다.
베스핀글로벌에서는 클라우드 문의나 Google Cloud 관련 무료 컨설팅을 진행하고 있습니다.
아래 문의로 편하게 연락주시기 바랍니다.

문 의 사 항 | 베스핀글로벌 구글사업부 sales.google@bespinglobal.com 070-7931-9600

참고 웹사이트 | https://cloud.withgoogle.com/next/sf/