



멀티/분산 클라우드, 차세대 클라우드를 향한 도전과 기회

- 클라우드바리스타 커뮤니티 제9차 컨퍼런스 -

AI 서비스 인프라를 위한 CB-Tumblebug의 여정

멀티 클라우드 인프라 통합 관리 기술

메인테이너@Cloud-Barista
손석호

시나몬 (Cinnamon) 한잔 어떠세요 ?

목 차

I

CB-Tumblebug 개요 (Quick Review)

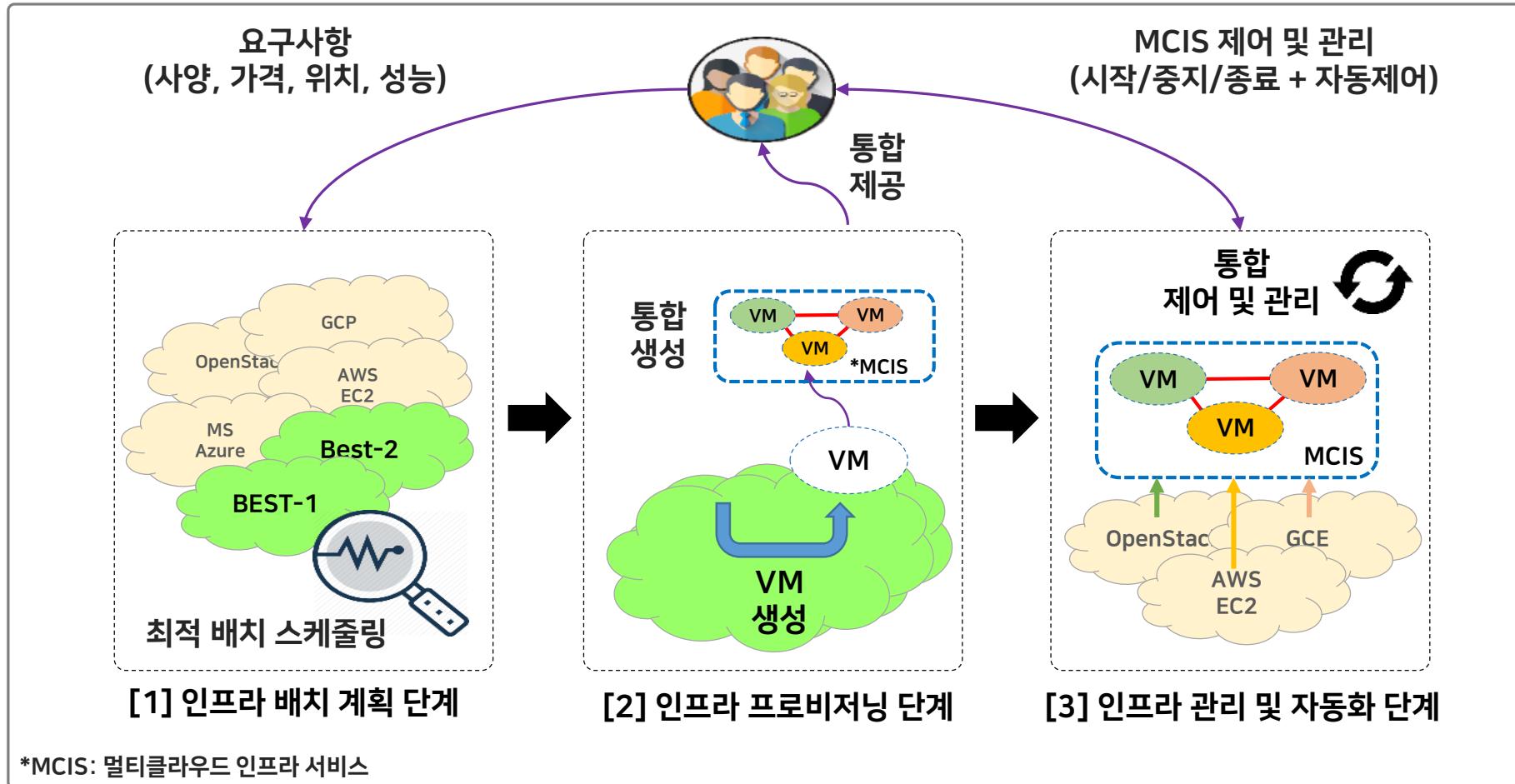
II

AI 서비스 인프라를 위한 CB-Tumblebug의 여정
(Self-hosted Multi-Cloud LLM Service)

CB-Tumblebug 개요



- 최적의 멀티 클라우드 인프라를 통합 배포, 제어, 관리하는 서비스 및 시스템



성능 벤치마킹 기반 최적

멀티클라우드 인프라

최적 클라우드 인프라
제공을 통한 효율성 증대

멀티클라우드 인프라
통합 운영 자동화

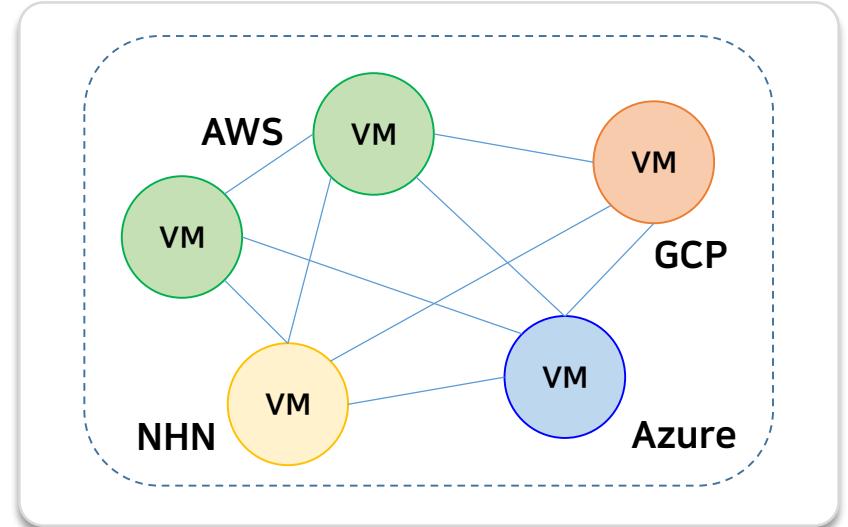
통합 제어, 정책 적용 등
관리 편의성 극대화

CB-Tumblebug 특징



- **멀티 클라우드 인프라 서비스 (MCIS: Multi-Cloud Infra Service)**

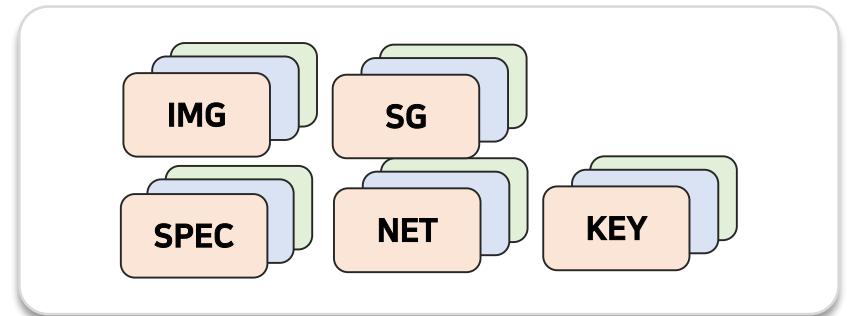
- 지역적으로 격리된 **다수의 클라우드 환경**에서
단일 목적(응용서비스, 애플리케이션 등)을 위해
하나 이상의 클라우드 인프라(가상머신 등)를
조합 및 상호 연계한 컴퓨팅 인프라 그룹
- 용도 : 멀티 클라우드 인프라의 통합 제어 및 관리



[MCIS 예시]

- **멀티 클라우드 인프라 리소스 (MCIR: Multi-Cloud Infra Resource)**

- **다수의 클라우드 환경에서 컴퓨팅 인프라 생성을 위해 관리**
하는 모든 리소스 (예: vNet, Image, Security Group, ...)
- 용도 : MCIS 구성 및 설정을 위한 리소스

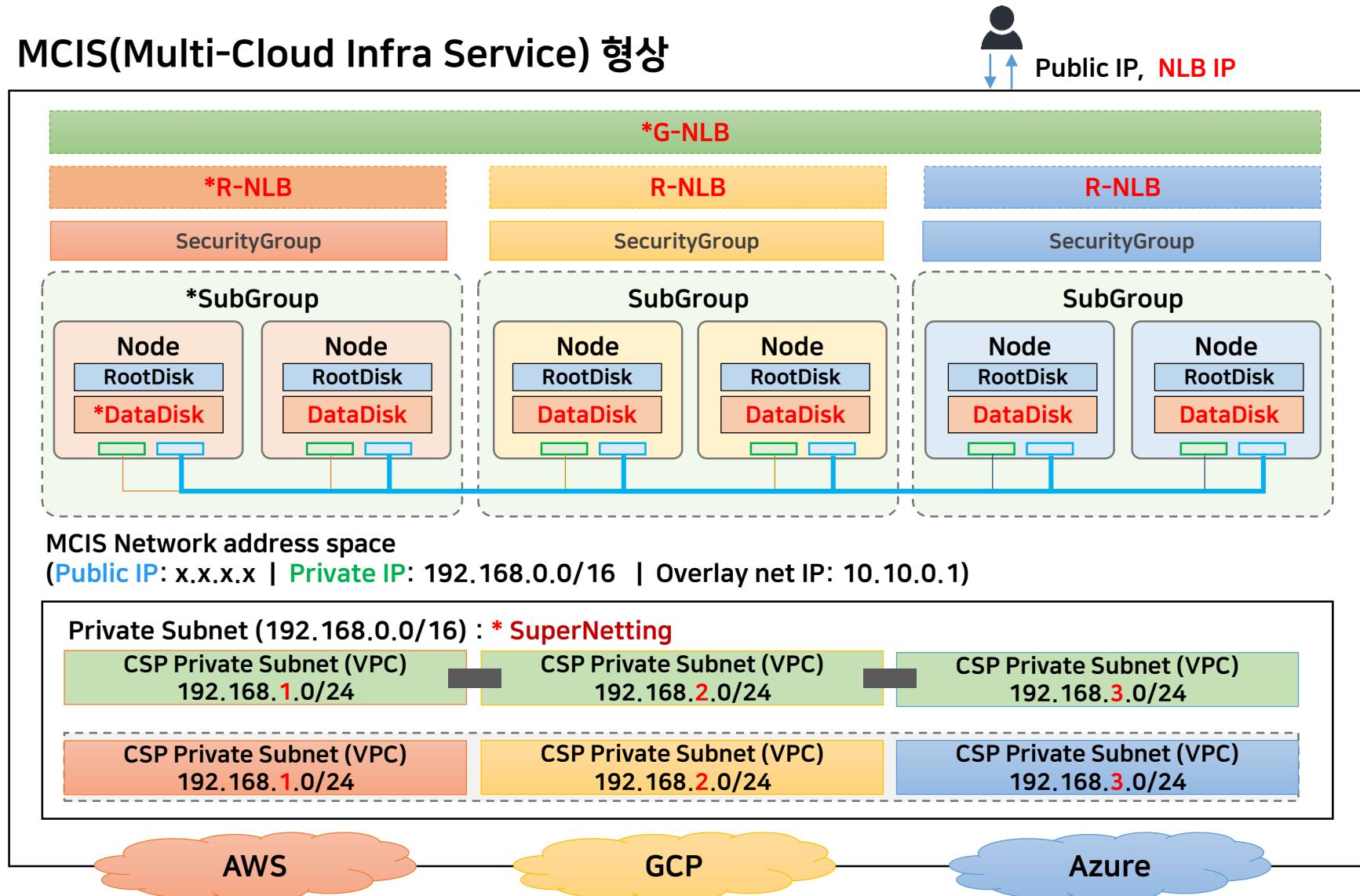


[MCIR 예시]

Cloud-Barista 멀티 클라우드 인프라 형상 - 2/2



MCIS(Multi-Cloud Infra Service) 형상



* NLB

네트워크 로드밸런서(L4)

- G-NLB: Global (SW 기반)
- R-NLB: Regional (CSP 서비스)

* SubGroup(Node Group)

동일 속성을 가진 Node 및 리소스그룹 (VM)

* DataDisk

Volume을 생성, 관리, 할당

* Site-to-Site VPN

CSP VPN 서비스를 통한 보안 터널링

* SuperNetting

더 많은 IP 주소 공간 활용, 라우팅 테이블 단순화, 터널링 등에 활용



(참고) CB-Tumblebug 기능 및 API



README Code of conduct Apache-2.0 license

CB-Tumblebug (Multi-Cloud Infra Service Management)

go report A+ build passing go 71.7% go.mod v1.21.6 repo size 34 MB go reference API Doc Swagger

release v0.8.0 release(dev) v0.8.11 license Apache-2.0 Slack SIG-TB

all contributors 41

A sub-system of Cloud-Barista Platform to Deploy and Manage Multi-Cloud Infrastructure.

- [CB-Tumblebug Overview \(Korean\)](#)
- [CB-Tumblebug Features \(Korean\)](#)
- [CB-Tumblebug Architecture \(Korean\)](#)
- [CB-Tumblebug Operation Sequence](#)

▶ [Note] CB-Tumblebug is currently under development

▶ [Note] Localization and Globalization of CB-Tumblebug

[[한국어](#), [English](#)]

목차

- [CB-Tumblebug 실행 및 개발 환경](#)
- [CB-Tumblebug 기여 방법](#)
- [CB-Tumblebug 실행 방법](#)
- [CB-Tumblebug 소스 빌드 및 실행 방법 상세](#)
- [CB-Tumblebug 기능 사용 방법](#)

CB-Tumblebug 실행 및 개발 환경

- Linux (추천: Ubuntu 22.04)
- Go (추천: v1.21.6)

[의존성 리스트 \(SBOM\)](#)

<https://github.com/cloud-barista/cb-tumblebug>

The screenshot shows a browser window displaying the API documentation for CB-Tumblebug. The URL is https://cloud-barista.github.io/api/?url=https://raw.githubusercontent.com/cloud-barista/cb-tumblebug/main/src/api/rest/doc/swagger.yaml. The page lists over 150 API endpoints categorized into two main sections: [Infra service] MCIS Provisioning management and [Namespace] Namespace management. The MCIS section includes endpoints for creating, deleting, listing, and managing MCIS objects and their subgroups. The Namespace section includes endpoints for creating, deleting, and listing namespaces.

API: 150+

- 라이브 API 상시 공개
- <https://cloud-barista.github.io/api/?url=https://raw.githubusercontent.com/cloud-barista/cb-tumblebug/main/src/api/rest/doc/swagger.yaml>

CB-Tumblebug 설치 및 활용 가이드

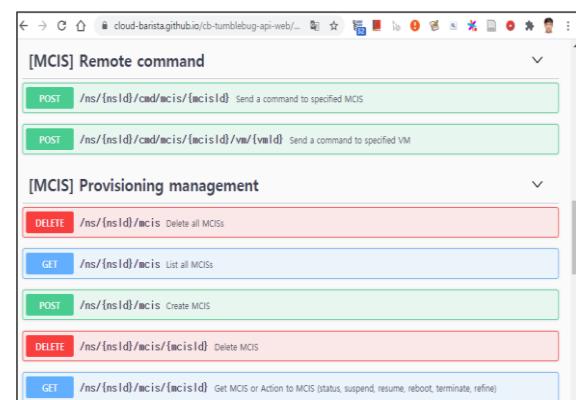


- 설치 및 실행 방법, API 규격 및 활용 방법 등 세부 내용 참고 : <https://github.com/cloud-barista/cb-tumblebug>
- REST API 및 활용 개요

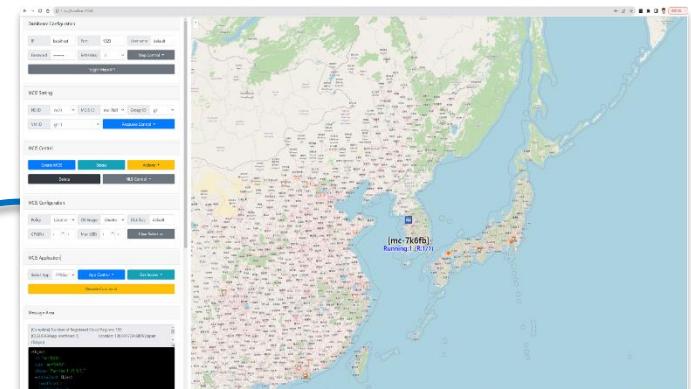
- CB-TB 서버 실행
- CB-TB 제어
 - REST API
 - CB-MapUI
 - Web Dashboards
 - 스크립트
- CB-TB로 생성한 MCIS 맛보기
 - 영상회의 서버 자동 배포
 - FPS, MMORPG 게임 자동 배포
 - Ansible 환경 자동 구성
 - Nginx 자동 배포
 - 클러스터 모니터링 도구 배포
 - ...



CB-Tumblebug 서버: REST API/
Swagger API 지원



REST API Swagger dashboard



MapUI: 지도 기반 GUI 클라이언트

```
## Test setting for Regions of Cloud types
# Note: you can change order by replacing lines (automatically assign continuous numbers starting from 1)
# AWS (Total: 21 Regions / Recommend: 20 Regions)
NumRegion[$IndexAWS]=2

IY=0
AwsApSoutheast1=$((++IY)) # Location: Asia Pacific (Singapore)
AwsCaCentral1=$((++IY)) # Location: Canada (Central)
AwsUsWest1=$((++IY)) # Location: US West (N. California)
AwsUsEast1=$((++IY)) # Location: US East (N. Virginia)
AwsApNortheast1=$((++IY)) # Location: Asia Pacific (Tokyo)
AwsApSouth1=$((++IY)) # Location: Asia Pacific (Sydney)
AwsEuWest12=$((++IY)) # Location: Europe (Paris)
AwsEuWest1=$((++IY)) # Location: Europe (London)
AwsUsEast2=$((++IY)) # Location: US East (Ohio)
AwsUsWest2=$((++IY)) # Location: US West (Oregon)
AwsApNortheast3=$((++IY)) # Location: Asia Pacific (Seoul)
AwsEuCentral1=$((++IY)) # Location: Europe (Frankfurt)
AwsEuWest1=$((++IY)) # Location: Europe (Ireland)
AwsEuWest3=$((++IY)) # Location: Europe (Paris)
AwsEuNorth1=$((++IY)) # Location: Europe (Stockholm) - No t2.xxx Specs. t3 c5 m5 r5 ... are available
AwsSaEast1=$((++IY)) # Location: South America (São Paulo)
AwsApNortheast2=$((++IY)) # Location: Asia Pacific (Hong Kong) - Opt-In required
AwsMeSouth1=$((++IY)) # Location: Middle East (Bahrain) - Opt-In required
AwsAfSouth1=$((++IY)) # Location: Africa (Cape Town) - Opt-In required
AwsEuSouth1=$((++IY)) # Location: Europe (Milan) - Opt-In required
```

스크립트 기반 시험 자동화 도구

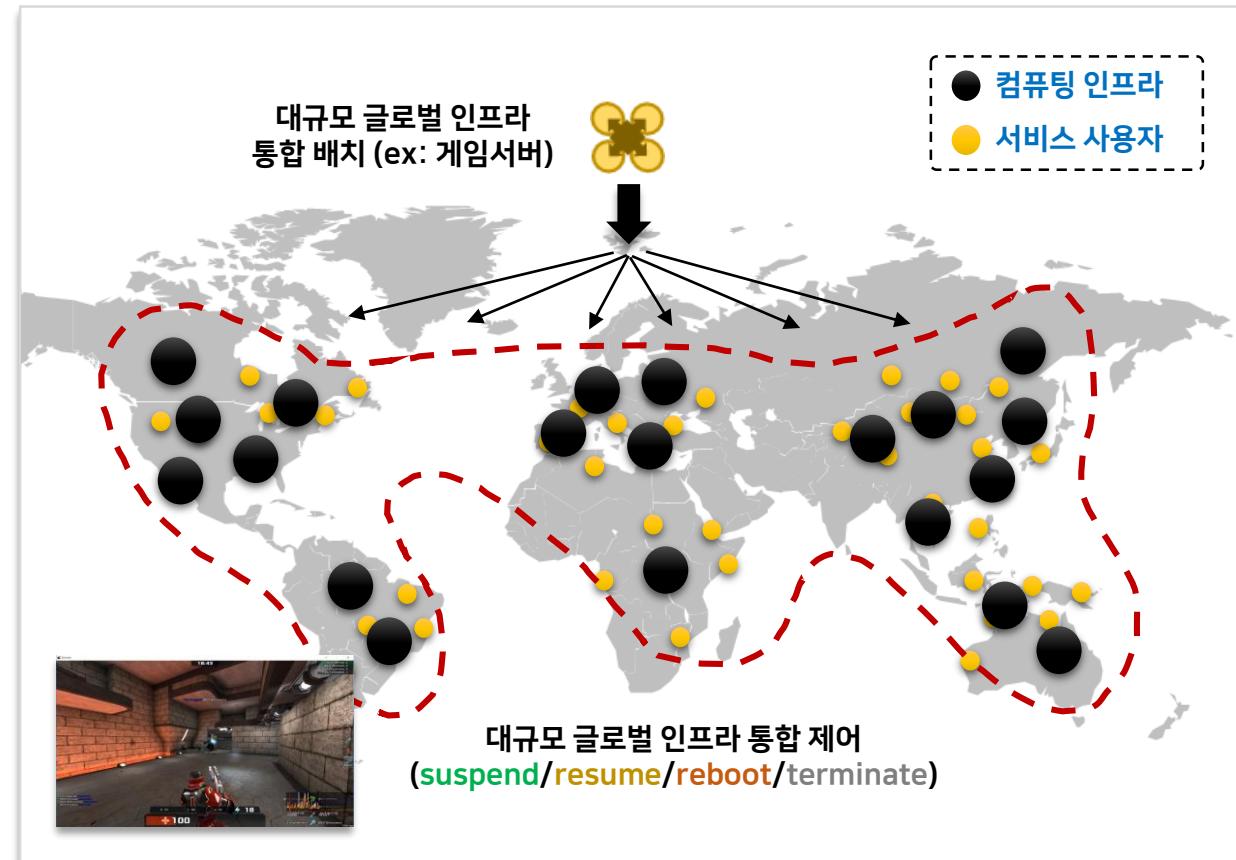
CB-Tumblebug 활용 사례



<최적배치 기반의 멀티 클라우드 기반 영상 회의 서비스>



< 글로벌 스케일 멀티 클라우드 FPS 게임 서비스>





CB-Tumblebug 활용 사례 (글로벌 3D FPS 게임 서비스)





클라우드바리스타 커뮤니티 제9차 컨퍼런스

AI 서비스 인프라를 위한 CB-Tumblebug의 여정

Multi-Cloud LLM Service by CB-TB

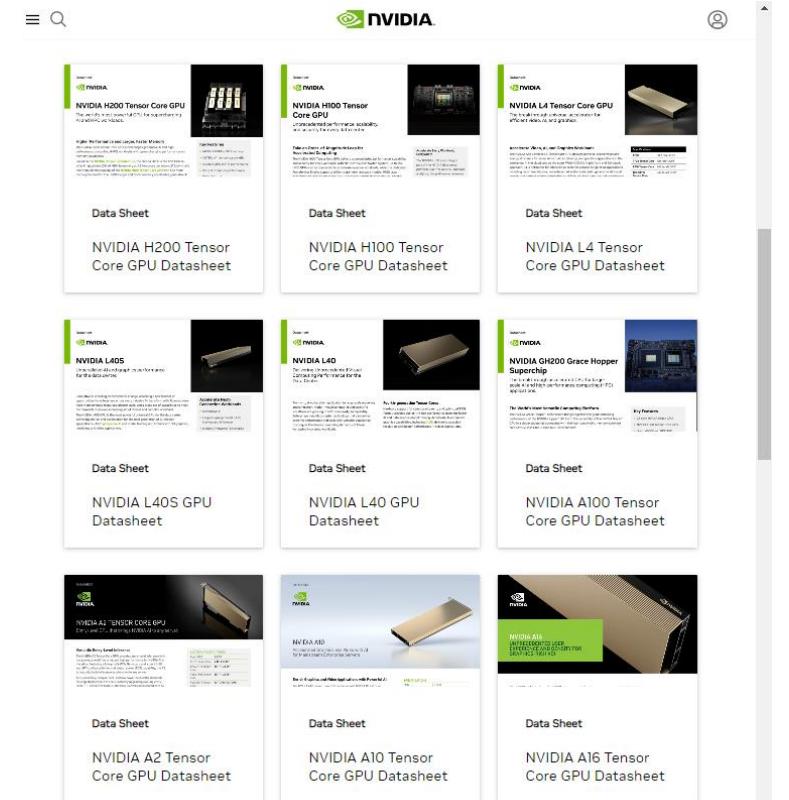
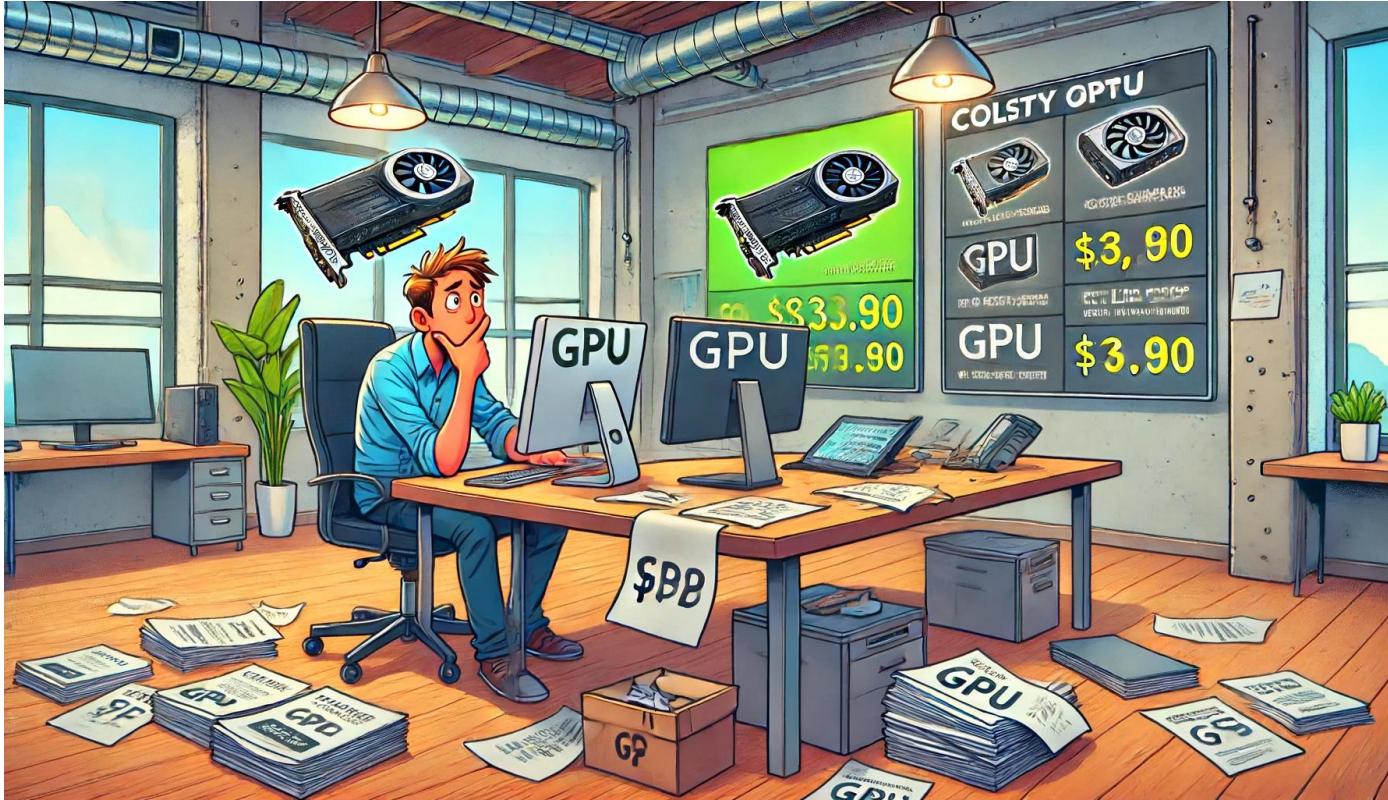
시나몬 (Cinnamon) 한잔 어떠세요 ?

어느 날 갑자기 찾아온 *LLM



*LLM (Large Language Model): 방대한 양의 텍스트 데이터로 훈련된 AI 모델 (프롬프트!)

담당자의 고민..



인프라는 어떻게 구축하지? 어떤 GPU를 어떻게 구입하지? GPU 드라이버 설정은 어떻게 하지? ..
LLM은 뭐지? 어떻게 서비스로 배포하지? 비용은 어떻게 하지? 서비스 확장성은 어떻게 제공하지? ..



어떤 GPU를 어떻게 구입하지?

danawa

검색어를 입력해주세요.

NVIDIA H100 HBM3 94GB NVL
H100 / 스트림 프로세서: 16896개 / PCIe5.0 / HBM3 / 지원기능: 멀티VGA 지원 / 사용전력: 400W
등록월 2024.05. | 상품의견 4건 | 관심
주요부품 > 그래픽카드(VGA)

NVIDIA H100 HBM2e 80GB PCIe
H100 / PCIe x16 / HBM(2세대) ECC / 지원기능: 멀티VGA 지원 / 사용전력: 350W
등록월 2024.04. | 상품의견 1건 | 관심
주요부품 > 그래픽카드(VGA)

NVIDIA L40S D6 48GB
L40S / 스트림 프로세서: 18176개 / PCIe4.0x16 / GDDR6 ECC / 출력단자: DP1.4 / 사용전력: 350W / 전원 포트: 16핀(12VHPWR) x1 / 가로(길이): 266.7mm
등록월 2024.04. | 관심
주요부품 > 그래픽카드(VGA)

NVIDIA A40 D6 48GB
A40 / 스트림 프로세서: 10752개 / PCIe4.0x16 / GDDR6 ECC / 출력단자: DP1.4 / 지원기능: 멀티VGA 지원 / 사용전력: 300W / 가로(길이): 266.7mm
등록월 2024.04. | 관심
주요부품 > 그래픽카드(VGA)

NVIDIA A30 HBM2 24GB
A30 / PCIe4.0x16 / HBM2 / 출력단자: DP1.4 / 지원기능: 멀티VGA 지원 / 사용전력: 165W / 가로(길이): 266.7mm
등록월 2024.04. | 관심
주요부품 > 그래픽카드(VGA)

어디서 뭘 사야 하나...

*주문 전 재고 문의 바랍니다.

엔비디아 NVIDIA H100 94GB NVL

구입하기도 어려운 상황?!

챗GPT가 등장한 2022년 말부터 현재까지 AI 칩 부족 현상은 이어지고 있다. 구글, 아마존, 메타 등 빅테크 기업들이 저마다 AI 서비스를 제공하기 위해 엔비디아의 GPU 서버를 대량으로 사들이고 있어서다. 메타는 올해 말까지 H100 35만개를 확보할 예정이며, 마이크로소프트(MS)는 엔비디아 AI 칩 재고를 180만개까지 늘린다는 계획이다.

업계 관계자는 “엔비디아가 AI 칩을 생산하는 데 걸리는 시간(리드타임)이 수개월로 늘어나 공급이 수요를 따라가지 못한 지는 이미 오래”라며 “1년 내내 기다려도 제품을 받지 못하는 기업들도 부지기수”라고 말했다. 오픈AI는 AI 칩 공급 부족 문제를 해결하기 위해 자체 AI 칩을 개발하는 방안을 검토 중이다. 또 MS와 1000억달러(약 134조6000억원)를 들여 AI 슈퍼컴퓨터를 포함한 데이터센터를 구축할 계획을 세우고 있다.

출처: 일본 머스크 “엔비디아 AI 칩 10만개 확보하겠다”… 8조 투자 받고 AI 개발 속도 (ChosunBiz, 2024.05.28.)



LLM은 무엇이고, 어떻게 서비스화 하지?

Google

LLM

전체 이미지 뉴스 동영상 쇼핑

LLM은 주어진 프롬프트에 대처 위해 방대한 양의 텍스트 데이터를 모델들은 인간 언어를 이해해서 뛰어납니다. 이를 통해 다양한 도구로 사용될 수 있습니다.

Bureau Works

대형 언어 모델(Large Language Model)이란 무

Amazon Web Services

대규모 언어 모델(LLM)이란 무

Cloudflare

대규모 언어 모델(LLM)이란 무

datamaker

LLM vs LMM : 미래의 언어 모델은 LLM (Large Language Model), 방대한 양의 생성할 수 있는 능력을 갖춘 모델입니다. 예

동영상

인공지능(AI)

YouTube · ETE

2024. 1. 11.

Google

LLM 추론

전체 이미지 쇼핑 동영상 뉴스

Databricks

LLM 추론 성능 엔지니어링: 모

AI타임스

LLM 추론 속도 300배까지 향상

aws

대규모 언어 모델(LLM)이란 무

Cloudflare

대규모 언어 모델(LLM)이란 무

datamaker

LLM vs LMM : 미래의 언어 모델은 LLM (Large Language Model), 방대한 양의 생성할 수 있는 능력을 갖춘 모델입니다. 예

동영상

인공지능(AI)

YouTube · ETE

2024. 1. 11.

Google

LLM 시스템 요구사항

전체 이미지 뉴스 동영상 쇼핑

브런치스토리

[번역] 거대언어모델(LLM) 가이드

요즘IT

LLM 서비스를 위협으로부터 지키는

daewoo kim - Medium

LLM(Large Language Model)을 학습하는

GitHub Pages

초거대 언어 모델의 수업시대

Sionic AI

LLM의 추론 능력 높이기 : Self

NVIDIA Developer

LLM 기술 마스터하기: 인퍼런스

ITWorld Korea

“앱 구축하는 것보다 어렵다” LLM

ITWorld Korea

“앱 구축하는 것보다 어렵다” LLM

Google

llm 헬프 만들기

전체 이미지 동영상 뉴스 쇼핑

Kanaries Docs

LLM 모델로 Streamlit 챗봇 만들

velog

LangChain을 사용하여 챗봇 만

ITWorld Korea

단 몇 분 만에 뚝딱... '나만의 AI

Elastic

LLM 선택: 2024년 오픈 소스 LLM

jiniai.biz

매칭 시스템의 정확도를 어떻게 향

ITWorld Korea

“앱 구축하는 것보다 어렵다” LLM

YouTube · 인공지

YouTube · OpsNote

구글 챗마(gemini)

Google

llm langchain

전체 이미지 동영상 뉴스 쇼핑

LangChain

LLMs | LangChain

ITWorld Korea

“LLM 개발을 더 간편하게” 랭체

DataCamp

How to Build LLM Applications

IBM

LangChain이란 무엇인가요?

vLLM

READMD

Apache-2.0 license

Code

sywangyi and ywang96

[Bugfix]if the content is started with ":"(resp...

.buildkite

.github

benchmarks

cmake

csrc

docs

examples

rocm_patch

tests

vllm

[View all files](#)



Easy, fast, and cheap LLM serving for everyone

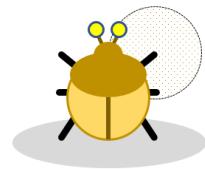
[Documentation](#) | [Blog](#) | [Paper](#) | [Discord](#)

Ray Summit CPF is Open (June 4th to June 20th)!

There will be a track for vLLM at the Ray Summit (09/30-10/02, SF) this year! If you have cool projects related to vLLM or LLM inference, we would love to see your proposals. This will be a great chance for everyone in the community to get together and learn. Please submit your proposal [here](#)



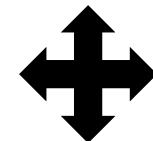
CB-Tumblebug + GPU x (Ollama + Open WebUI)



CB-Tumblebug

Deploy and Manage
Multi-Cloud Computing Infra

<https://github.com/cloud-barista/cb-tumblebug>



Ollama



Open WebUI



혹시 같은 고민을 가지고 계셨다면, 이 자리에 잘 오셨습니다. :)

그럼, 열차 출발합니다.
(소요시간: 13분)

**SELF HOSTED MULTI-CLOUD LLM
BY CB-TUMBLEBUG
(&OLLAMA&WEBUI)**

Behind the scenes (part 1)

그 동안 어떤 일이 있었을까요?



Back to the 2019

사실은 처음부터..GPU..

CLOUD BARISTA

최적 멀티 클라우드 인프라를 찾아서

클라우드 서비스 성능.. 알고 보면 많이 달라요

손석호
멀티 클라우드
인프라 통합 관리
프레임워크 리더

멀티 클라우드에서는 자원들의 성능 및 특성이 매우 다양하므로, 최적의 멀티 클라우드 인프라 서비스 제공 필요

Q&A : contact-to-cloud-barista@googlegroups.com

제2차 오픈 컨퍼런스 : CB-Tumblebug-멀티 클라우드 인프라 통합 운영 관리(Multi-Cloud Infrastructure Service Management)

VM 성능 비교 (AWS vs GCP)

Category	AWS	GCP
1 core	1.2x	1.4x
2 cores	1.0x	0.8x
8 cores	1.0x	1.1x

VM 가격 비교 (AWS vs GCP)

Category	AWS	GCP
2-core on-demand price	\$0.08	\$0.10
Price normalized to match relative machine's performance	\$0.08	\$0.12

GPU 성능 및 비용 비교 (AWS vs GCP)

Category	AWS	GCP
Time to Train (Minutes)	~10	~20
Cost per Hour (\$ USD)	~\$0.10	~\$0.05

<https://towardsdatascience.com/maximize-your-gpu-dollars-a9133f4e546a> by Jeff Wiles



그때 더 팔어야 했는데.. (\$: 그때 더 샀어야 했는데..)

2,400% ↑

최적 배치 기능 개선: GPU 등록 및 검색 지원

```
AWS,all,c4.large,0.114,78.3,,,,,,default,default,,,
AWS,all,c4.4xlarge,0.907,80.74,,,,,,default,default,,,
AWS,all,c4.2xlarge,0.454,92.55,,,,,,default,default,,,
AWS,all,g5.xlarge,1.006,34.18,,,,,,default,default,gpu,NVIDIA A10G,1,24,
AWS,all,g5.2xlarge,1.212,67.18,,,,,,default,default,gpu,NVIDIA A10G,1,24,
AWS,all,g5.4xlarge,1.624,78.18,,,,,,default,default,gpu,NVIDIA A10G,1,24,
AWS,all,g5.8xlarge,2.448,89.18,,,,,,default,default,gpu,NVIDIA A10G,1,24,
AWS,all,g5.12xlarge,5.672,99.18,,,,,,default,default,gpu,NVIDIA A10G,1,24,
AWS,all,p4d.24xlarge,32.77,99.88,,,,,,default,default,gpu,NVIDIA A100,1,320,
azure,all,Standard_B1s,0.0132,33.85,,,,,,default,default,,,
azure,all,Standard_B1ms,0.0264,33.85,,,,,,default,default,,,
```

github.com/cloud-barista/cb-tumblebug/blob/main/assets/cloudspec.csv

```
AWS,ap-northeast-1,ami-039ed92b1a75d78cc,Debian 10,,
AWS,ap-northeast-1,ami-006b8cbbcb9386e7,Windows Server 2012 R2,,
AWS,ap-northeast-1,ami-0bded5720a0796acb,Ubuntu 20.04,Deep Learning OSS Nvidia Driver
AMI GPU TensorFlow 2.15 (Ubuntu 20.04) 20240423,G4dn G5 G6 Gr6 P4 P4de P5
AWS,ap-south-1,ami-0123b531fc646552f,Ubuntu 18.04,,
AWS,ap-south-1,ami-068257025f72f470d,Ubuntu 22.04,,
```

github.com/cloud-barista/cb-tumblebug/blob/main/assets/cloudimage.csv

- GPU 입력 가능하게 정보 확장
- Spec, Image가 순차적으로 등록 및 테스트 진행되고 있음
(아직 제한적, 3 CSP++)

Recommended Spec and CSP region

Recommended Spec	g6.xlarge
Estimated Price(USD/H)	\$ 0.805 (at least)
Selected Image OS Type	Ubuntu22.04
-----	-----
Provider	AWS
Region	us-east-2
ConnectionConfig	aws-us-east-2
-----	-----
vCPU	4
Mem(GiB)	16
RootDiskType	default
RootDiskSize(GB)	200
-----	-----
AcceleratorType	GPU
AcceleratorModel	NVIDIA L4
AcceleratorCount	1
AcceleratorMemoryGB	24

Show All Recommendations ▾	
Search:	
10	entries per page
Spec	CSP
Region	CPU
Mem	Cost
GPU	Model
Model	Mem

1	g6.xlarge	AWS	us-east-2	4	16	\$ 0.805	1	NVIDIA L4
2	g6.xlarge	AWS	us-east-2	4	16	\$ 0.805	1	NVIDIA L4
3	g6.xlarge	AWS	us-east-2	4	16	\$ 0.805	1	NVIDIA L4
4	g6.2xlarge	AWS	us-east-2	8	32	\$ 0.978	1	NVIDIA L4
5	g6.2xlarge	AWS	us-east-2	8	32	\$ 0.978	1	NVIDIA L4

최적 배치 기능 개선: GPU 등록 및 검색 지원

commonSpec ID 획득 방식 개선

/mcisRecommendVm 필터링 옵션 확대

- 기존

- 산술 조건: `cpu, memory, cost`
- 문자열 조건: `provider, region, specname`
- 확장시 소스에서 조건 개별 추가

- 개선

- 산술 조건 & 문자열 조건:
Spec Object Struct의 모든 항목 활용 가능
- 확장시 개별 조건 추가 필요 없음
(내부적: type 동적 판별을 위한
`reflect`)

- [/mcisRecommendVm API](#)를 사용하면, 사용자가 요청한 사양 및 조건에 맞는 Spec 리스트를 획득할 수 있음
- 필터링 및 우선순위가 적용 가능하며, 다양한 파라미터의 조합이 가능하므로 별도의 옵션을 알고 사용할 필요.

필터링 옵션

ex: "metric": "providerName"

산술 일치 조건 "`<=`", "`>=`", "`==`" (ex: "operator": "`<=`", "operand": "4")

- `vCPU json: "vCPU"`
- `MemoryGiB json: "memoryGiB"`
- `CostPerHour json: "costPerHour"`
- `AcceleratorCount json: "acceleratorCount"`
- `AcceleratorMemoryGB json: "acceleratorMemoryGB"`

문자열 포함 조건 (ex: "operand": "gcp")

- `ProviderName json: "providerName"`
- `RegionName json: "regionName"`
- `CspSpecName json: "cspSpecName"`
- `AcceleratorModel json: "acceleratorModel"`
- `AcceleratorType json: "acceleratorType"`
- `Description json: "description"`

무한 Testing & Scripting (NVIDIA+Cuda, Ollama, Open WebUI)

잘 될 때까지 잘 되게 하라 (아마도 아직 부족..)

cb-tumblebug / scripts / usecases / llm / 	
 seokho-son	Fix echo for Update deployOpenWebUI.sh
Name	Last commit message
 ..	
 README.md	Update llm readme
 <u>deployOllama.sh</u>	Remove disabled echo function in deployOllama.sh
 <u>deployOpenWebUI.sh</u>	Fix echo for Update deployOpenWebUI.sh
 <u>installCudaDriver.sh</u>	Update installCudaDriver.sh
 llmServer.py	Fix model agrs bug in the llm script
 startServer.sh	Update startServer.sh
 statusServer.sh	Ehance llm usecase code
 stopServer.sh	Update llm usecase with easy setting

[github.com/cloud-barista/
cb-tumblebug/tree/main/scripts/usecases/llm](https://github.com/cloud-barista/cb-tumblebug/tree/main/scripts/usecases/llm)

Model library

Ollama supports a list of models available on ollama.com/library

Here are some example models that can be downloaded:

Model	Parameters	Size	Download
Llama 3	8B	4.7GB	ollama run llama3
Llama 3	70B	40GB	ollama run llama3:70b
Phi 3 Mini	3.8B	2.3GB	ollama run phi3
Phi 3 Medium	14B	7.9GB	ollama run phi3:medium
Gemma	2B	1.4GB	ollama run gemma:2b
Gemma	7B	4.8GB	ollama run gemma:7b
Mistral	7B	4.1GB	ollama run mistral
Moondream 2	1.4B	829MB	ollama run moondream
Neural Chat	7B	4.1GB	ollama run neural-chat
Starling	7B	4.1GB	ollama run starling-lm
Code Llama	7B	3.8GB	ollama run codellama
Llama 2 Uncensored	7B	3.8GB	ollama run llama2-uncensored
LLaVA	7B	4.5GB	ollama run llava
Solar	10.7B	6.1GB	ollama run solar

Note: You should have at least 8 GB of RAM available to run the 7B models, 16 GB to run the 13B models, and 32 GB to run the 33B models.

원격 커맨드 개선: 다중 커맨드 / 빌트인 함수 지원

Put multiple commands to forward

[Commands]

Command 1:	client_ip=\$(echo \$SSH_CLIENT awk '{print \$1}'); my_ip=\$(hostname -l)	Clear
Command 2:	echo SSH client (bastion) IP is: \$client_ip	Clear
Command 3:	echo IPs of my network: \$my_ip	Clear
Command 4:		Clear
Command 5:		Clear
Command 6:		Clear
Command 7:		Clear
Command 8:		Clear
Command 9:		Clear

[Select target]

- MCIS: aws-ap-northeast-1-aws-application-migration-service-replication-server-i-05014598057263010
- SUBGROUP: aws-ap-northeast-1-aws-application-migration-service-replication-server-i-05014598057263010
- VM: aws-ap-northeast-1-aws-application-migration-service-replication-server-i-05014598057263010-1

Execute **Cancel**

- 빌트인 키워드: **\$\$Func()**
 - target 리소스 지정 방식
 - this: 커맨드가 실행될 스스로를 명시
 - mcis.vm 패턴: 리소스를 “.” 으로 구분하여 명시
 - prefix, postfix, separator
 - 앞뒤에 스트링 추가, 리스트를 구분자포함 스트링으로 변경
- 빌트인 함수
 - **GetPublicIP**: 커맨드가 실행되는 자신(또는 명시한 VM)의 IP 스트링으로 교체
 - **\$\$Func(GetPublicIP(target=this, prefix=http://, postfix=:3000))**
→ http://3.15.18.101:3000
 - **GetPublicIPs**: 커맨드가 실행되는 자신(또는 명시한 MCIS) 의 IP 스트링으로 교체
 - **\$\$Func(GetPublicIP(target=this, separator=+, postfix=:3000))**
→ 3.15.18.101:3000 + 5.25.153.31:3000 + 7.215.23.33:3000
- **AssignTask**: MCIS 내의 각 VM에 지정된 Task 리스트 내의 Task를 할당
 - **\$\$Func(AssignTask(task='llama3, solar, mistral, phi3, gemma, mixtral, llava, yi, falcon2, llama2'))**
→ VM1: llama3 , VM2: solar, VM3: mistral, ...

시스템 개선: safe startup

- 개요

- CB-TB 초기화시, CB-SP에 자동 초기화 작업 진행 개선
- 연결 및 활성화된 CB-SP가 없으면 CB-TB 오류 발생

- 개선 사항

- CB-TB 서버 실행시, CB-SP Health check (API: `readyz`)
 - 3초 단위로 60회 시도
 - 180초 내에 확인되지 않으면, 서버 종료

```

| tencent | sa-saopaulo | Sao Paulo | (-23.55:-46.65) | sa-saopaulo
+-----+
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 1/60
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 2/60
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 3/60
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 4/60
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 5/60
12:38AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 6/60
12:39AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 7/60
12:39AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 8/60
12:39AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 9/60
12:39AM INF main.go:189 > CB-Spider at http://localhost:1024/spider is not ready. Attempt 10/60
12:39AM INF main.go:186 > CB-Spider is now ready. Initializing CB-Tumblebug...
12:39AM INF main.go:219 > /home/son/.cloud-barista/credentials.yaml
+-----+
| CREDENTIALHOLDER | CLOUD SERVICE PROVIDER | CREDENTIAL KEY | CREDENTIAL VALUE
+-----+
| admin           | alibaba            | clientid       | *****
| admin           | alibaba            | clientsecret   | *****
| admin           | aws                | clientid       | .....
| admin           | .....              | .....          | .....
| admin           | nhncloud           | username        | *****
| admin           | tencent             | clientid       | *****
| admin           | tencent             | clientsecret   | *****
+-----+
12:39AM DBG core/common/namespace.go:199 > [Get namespace] ns01
12:39AM DBG core/common/namespace.go:201 > /ns/ns01
12:39AM INF main.go:304 > [Initiate Multi-Cloud Orchestration]
12:39AM INF api/rest/server/server.go:86 > REST API Server is starting

```

CB-TB
READY

Multi-cloud infrastructure management framework

<https://github.com/cloud-barista/cb-tumblebug>



시스템 개선: 시스템 초기화, 고도화 (init.py)

```
Checking server health...
Tumblebug Server is healthy.

Registering credentials and Loading common Specs and Images takes time
Do you want to proceed ? (y/n) : y

Registering all valid credentials for all cloud regions...
- ibm: {'credentialName': 'ibm', 'credentialHolder': 'admin', 'providerName': 'ibm', 'keyValueInfoList': [{'key': 'ApiKey', 'value': '*****'}]}
- alibaba: {'credentialName': 'alibaba', 'credentialHolder': 'admin', 'providerName': 'alibaba', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}]}
- aws: {'credentialName': 'aws', 'credentialHolder': 'admin', 'providerName': 'aws', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}]}
- gcp: {'credentialName': 'gcp', 'credentialHolder': 'admin', 'providerName': 'gcp', 'keyValueInfoList': [{'key': 'client_id', 'value': '*****'}, {'key': 'ClientEmail', 'value': '*****'}, {'key': 'private_key_id', 'value': '*****'}, {'key': 'PrivateKey', 'value': '*****'}, {'key': 'ProjectID', 'value': '*****'}]}
- azure: {'credentialName': 'azure', 'credentialHolder': 'admin', 'providerName': 'azure', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}, {'key': 'TenantId', 'value': '*****'}, {'key': 'SubscriptionId', 'value': '*****'}]}
- ktcloud: {'credentialName': 'ktcloud', 'credentialHolder': 'admin', 'providerName': 'ktcloud', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}]}
- openstack: Incomplete credential data, Skip
- ncp: {'credentialName': 'ncp', 'credentialHolder': 'admin', 'providerName': 'ncp', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}]}
- ktcloudvpc: {'credentialName': 'ktcloudvpc', 'credentialHolder': 'admin', 'providerName': 'ktcloudvpc', 'keyValueInfoList': [{'key': 'IdentityEndpoint', 'value': '*****'}, {'key': 'Username', 'value': '*****'}, {'key': 'Password', 'value': '*****'}, {'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}, {'key': 'DomainName', 'value': '*****'}, {'key': 'ProjectID', 'value': '*****'}]}
- ncpvpc: {'credentialName': 'ncpvpc', 'credentialHolder': 'admin', 'providerName': 'ncpvpc', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {'key': 'ClientSecret', 'value': '*****'}]}
- nhncloud: {'credentialName': 'nhncloud', 'credentialHolder': 'admin', 'providerName': 'nhncloud', 'keyValueInfoList': [{'key': 'IdentityEndpoint', 'value': '*****'}, {'key': 'Username', 'value': '*****'}, {'key': 'Password', 'value': '*****'}, {'key': 'DomainName', 'value': '*****'}, {"key": "TenantId", "value": "*****"}]}
- tencent: {'credentialName': 'tencent', 'credentialHolder': 'admin', 'providerName': 'tencent', 'keyValueInfoList': [{'key': 'ClientId', 'value': '*****'}, {"key": "ClientSecret", "value": "*****"}]}

Loading common Specs and Images...
Progress: 100% | 240/240 [03:39<00:00, 1.09s/s]

Loading completed (3.65 minutes)
Registered Common specs
- Successful: 3363, Failed: 1354
Registered Common images
- Successful: 467, Failed: 128

Cleaning up...
Environment cleanup complete.
son@son:~/go/src/github.com/cloud-barista/cb-tumblebug$
```

```
more cb-tumblebug
bash cb-tumblebug
npm cb-mapui
bash cb-tumblebug
```

지원 대상 확대: 국내 CSP 프로비저닝 지원

**CB-TB 국내 CSP 지원 시연
[MCIS 통합 제어]**

Behind the scenes (part 2)

아직 잘 안되는 포인트

H100 생성, 실패 : 사용자 자원 퀘터 부족 (AWS)

Are you sure you want to create this MCIS?

h100 (1 node(s))

Usage Period	Estimated Cost
Hourly	\$98.3200
Daily	\$2359.6800
Monthly	\$73150.0800

(Do not rely on this estimated cost. It is just an estimation using spec price.)

[#1] SubGroup Name	g1 (1 node(s))
Estimated Price(USD/1H)	\$98.3200 (\$98.32 * 1)
Spec	aws+us-east-2+p5.48xlarge
vCPU	192
Mem(GiB)	2048
Accelerator	GPU (NVIDIA H100)
RootDisk(GB)	default (type: default)
Selected Image OS Type	Ubuntu22.04

Hold VM provisioning of the MCIS
 Deploy CB-Dragonfly monitoring agent

Confirm **Cancel**



Status: Failed

VcpuLimitExceeded: You have requested more vCPU capacity than your current vCPU limit of 64 allows for the instance bucket that the specified instance type belongs to.

Please visit <http://aws.amazon.com/contact-us/ec2-request> to request an adjustment to this limit.

사용자 자원 쿼터 부족 (AWS)

Recommended Spec and CSP region

Recommended Spec	p5.48xlarge
Estimated Price(USD/1H)	\$ 98.32 (at least)
Selected Image OS Type	Ubuntu22.04

Provider	AWS
Region	us-east-2
ConnectionConfig	aws-us-east-2

vCPU	192
Mem(GiB)	2048
RootDiskType	default
RootDiskSize(GB)	default

AcceleratorType	GPU
AcceleratorModel	NVIDIA H100
AcceleratorCount	8
AcceleratorMemoryGB	640

Show All Recommendations ▾

Enter the number of VMs for scaling (1 ~ 10)

Confirm Cancel

aws Services Search [Alt+S] VPC EC2 Service Quotas > AWS services > Amazon Elastic Compute Cloud (Amazon EC2) > Running On-Demand P instances Request increase at account level

Service Quotas

Dashboard AWS services Quota request history Organization Quota request template

Running On-Demand P instances

Details

Description Maximum number of vCPUs assigned to the Running On-Demand P instances.

Quota code L-417A185B

Quota ARN arn:aws:servicequotas:us-east-2:635484366616:ec2/L-417A185B

Utilization	Applied account-level quota value	AWS default quota value	Adjustability
0	64	0	Account level

Recent quota increase requests

View open quota increase requests or requests that were recently closed.

Filter by service or quota

Service	Quota name	Status	Requested quota value	Request date	Last updated date
Amazon Elastic Compute Cloud (Amazon EC2)	Running On-Demand P instances	Requested	400	2024년 6월 17일	2024년 6월 17일

쿼터 증가 요청 필요

사용자 자원 큐터 부족 (GCP, Azure, ... 마찬가지)

The screenshot displays three separate cloud service interfaces:

- Azure:** Shows a list of VM sizes (e.g., NV12s_v3, NV16as_v4) with "Request quota" buttons. A yellow box highlights the "Azure" logo.
- GCP:** Shows the "Quotas & System Limits for project 'etri-jhseo-test'". It includes sections for "Set up quota & system limit alerts", current usage (0/90%), and a quota table. A yellow box highlights the "GCP" logo.
- New Quota Request:** A modal window titled "New Quota Request" with tabs for "Successful" (0), "Partial" (0), and "Failed" (1). It contains a message: "We were unable to adjust your quota. Submit a support ticket so that a support engineer can assist you in adjusting your quota for your Standard NVADSA10v5 Family vCPUs in North Europe for Microsoft Azure."

방도가..

심지어.. CSP 자원 부족 (H/A100)

Status: Failed

Operation errors: Quota 'GPUS_PER_GPU_FAMILY' exceeded. Limit: 0.0 in region us-west4

LLM 배포 관련 스크립트의 제약! (급하게 만드느라..)

[cb-tumblebug / scripts / usecases / llm / installCudaDriver.sh](#)

seokho-son Update installCudaDriver.sh 56b32cd · 2 days ago History

Code Blame 75 lines (62 loc) · 2.77 KB Raw ⌂ ⌃ ⌄ ⌅ ⌆ ⌇

```
1 #!/bin/bash
2
3 # This script installs NVIDIA drivers and CUDA on Ubuntu 22.04.
4 # It downloads necessary files from NVIDIA's official website,
5 # installs the drivers and CUDA, sets the required environment variables,
6 # and reboots the system to apply changes.
7 # https://developer.nvidia.com/cuda-downloads?target_os=Linux&target_arch=x86_64&Distribution=Ubuntu&Version=22.04
8
9 # Check for NVIDIA GPU
10 echo "Checking for NVIDIA GPU..."
11 GPU_INFO=$(sudo lspci | grep -i nvidia)
12 if [ -z "$GPU_INFO" ]; then
13     echo "No NVIDIA GPU detected or an error occurred. Exiting..."
14     sudo lspci
15     exit 1
16 else
17     echo "NVIDIA GPU detected:"
18     echo "$GPU_INFO"
19     echo "Check root disk size"
20     df -h / | awk '$NF=="/" {print "Total: "$2, "Available: "$4}'
```

- 현재! 깡통 Ubuntu 22.04만 지원합니다!
- Sustainable 하지 않습니다!
- CSP 이미지를 사용하면 좋겠지만, 찾는 것도 일입니다.

Select Target Platform
Click on the green buttons that describe your target platform. Only supported platforms will be shown. By downloading and using the software, you agree to fully comply with the terms and conditions of the [CUDA EULA](#).

Operating System	Linux	Windows						
Architecture	x86_64	arm64-sbsa	aarch64-jetson					
Distribution	Amazon-Linux	Debian	Fedor	KylinOS	OpenSUSE	RHEL	Rocky	SLES
Version	Ubuntu	WSL-Ubuntu						
Installer Type	20.04	22.04						
deb (local)	deb (network)	runfile (local)						

[github.com/cloud-barista/cb-tumblebug/
blob/main/scripts/usecases/llm/installCudaDriver.sh](https://github.com/cloud-barista/cb-tumblebug/blob/main/scripts/usecases/llm/installCudaDriver.sh)

developer.nvidia.com/cuda-downloads

오류. 기타 등등 (이제는 조심해야!)

파산 방지

정확하지 않은 추정 가격이라도
GUI에 급히 추가한 이유가 있습니다.

(시스템 로그에도 좀 더 신경 썼습니다. ++ZeroLog)

Are you sure you want to create this MCIS?

H100-클러스터-만들면-파산 (70 node(s))

매시간: 9백만
매월: 66억 7천만

Usage Period	Estimated Cost
Hourly	\$6492.0000
Daily	\$155808.0000
Monthly	\$4830048.0000

(Do not rely on this estimated cost. It is just an estimation using spec price.)

[#1] SubGroup Name	g1 (20 node(s))
Estimated Price(USD/1H)	\$1966.4000 (\$98.32 * 20)
Spec	aws+us-west-2+p5.48xlarge
vCPU	192
Mem(GiB)	2048
Accelerator	GPU (NVIDIA H100)
RootDisk(GB)	default (type: default)
Selected Image OS Type	Ubuntu22.04

[#2] SubGroup Name	g2 (20 node(s))
Estimated Price(USD/1H)	\$1771.2000 (\$88.56 * 20)
Spec	gcp+us-east4+a3-highgpu-8g
vCPU	208
Mem(GiB)	1872
Accelerator	GPU (NVIDIA H100)
RootDisk(GB)	default (type: default)
Selected Image OS Type	Ubuntu22.04

아직은 단순한 인프라 구성. 최적화는..언제쯤..

NVIDIA Data Center Products	
GPU	Compute Capability
NVIDIA H100	9.0
NVIDIA L4	8.9
NVIDIA L40	8.9
NVIDIA A100	8.0
NVIDIA A40	8.6
NVIDIA A30	8.0
NVIDIA A10	8.6
NVIDIA A16	8.6
NVIDIA A2	8.6
NVIDIA T4	7.5
NVIDIA V100	7.0
Tesla P100	6.0
Tesla P40	6.1
Tesla P4	6.1
Tesla M60	5.2
Tesla M40	5.2
Tesla K80	3.7
Tesla K40	3.5
Tesla K20	3.5
Tesla K10	3.0

Table 21: Technical Specifications per Compute Capability

	Compute Capability													
Technical Specifications	5.0	5.2	5.3	6.0	6.1	6.2	7.0	7.2	7.5	8.0	8.6	8.7	8.9	9.0
Maximum number of resident grids per device (Concurrent Kernel Execution)	32		16	128	32	16	128	16		128				
Maximum dimensionality of grid of thread blocks									3					
Maximum x -dimension of a grid of thread blocks [thread blocks]									$2^{31}-1$					
Maximum y- or z-dimension of a grid of thread blocks									65535					
Maximum dimensionality of thread block									3					
Maximum x- or y-dimensionality of a block									1024					
Maximum z-dimension of a block									64					
Maximum number of threads per block									1024					
Warp size									32					
Maximum number of resident blocks per SM							32		16	32	16	24	32	
Maximum number of resident warps per SM							64		32	64	48		64	
Maximum number of resident threads per SM							2048		1024	2048	1536		2048	
Number of 32-bit registers per SM									64 K					
Maximum number of 32-bit registers per thread block	64 K	32 K	64 K	32 K						64 K				
Maximum number of 32-bit registers per thread									255					
Maximum amount of shared memory per SM	64 KB	96 KB	64 KB	96 KB	64 KB	96 KB	64 KB	164 KB	100 KB	164 KB	100 KB	228 KB		
Maximum amount of shared memory per thread block ³²				48 KB									²²⁷ KB	
Number of shared memory banks														
Maximum amount of local memory per thread														
Constant memory size									64 KB					
Cache working set per SM for constant memory		8 KB	4 KB											
Cache working set per SM for texture memory		Between 12 KB and 48 KB		Between 24 KB and 48 KB					64 KB	KB ~ 192 KB	28 KB	28 KB	28 KB	
Maximum width for a 1D texture object using a CUDA	65536								131072	KB	128 KB	128 KB	256 KB	

Compute Capability
다양한 결정 요소 (NVIDIA)

.. 연구 필요..

cloud-barista / cb-tumblebug

Issues 35 Pull requests 3 Discussions Actions Projects 2 Wiki Security Insights Settings

Filters Q isissue isopen Labels 29 Milestones New issue

35 Open 295 Closed

Failure when provisioning the first Data Disk after creating MCIS
No CB-TB objects' values in structs related to TbCluster* (like TbClusterInfo, TbClusterNodeGroupInfo, ...) bug
Enhance vmDynamic API request body enhancement
cb-spider recognition failure error in docker-compose environment enhancement wip
System default sshkey resource can be removed by rollback from another mcis provisioning bug
Typographical Errors in provisioning.go enhancement good first issue
Extend options to specify Id2d target and step for MCIS provisioning enhancement
K8s cluster provisioning process based on CB-MapUI
Support configuration for seamless K8s cluster provisioning feature request
Umbrella: Support seamless K8s cluster provisioning feature request
[NCP, NCPVPC, NHN CLOUD, KT CLOUD, KT VPC] request for region, zone list and Spec/image test set information question
Azure k8s cluster provisioning failure bug
Remote access/command for a cluster feature request
Watch Swagger API dashboard performance issue enhancement
[KT Cloud VPC] To use NLB, needs to support the subnet management features with a fixed name. enhancement
[NCP Classic] When manages S/G, need to apply the name user generated in the console as it is. enhancement
OpenTofu TF PoC integration feature request
Support Built-in functions in parameters of a request body feature request
Is it valid that scripts that creating vNet/securityGroup/sshKey/... for all csp regions? question
Check and Incorporate OpenAPI 3.1 from new swaggo release feature request
Prevent multiple requests for ScaleOut and ScaleIn feature request
Register all existing CSP resources for each cloudConnection to CB-TB objs feature request
Data types of fields in core/mcis enhancement
Need a tool to retrieve price of Cloud VM Specs feature request help wanted
Enhance codes for mcis control by applying switchCase enhancement

사실은.. 아직 해야 할 게 너무 많아요..

(힘들어 죽겠어요..)

살려주세요.

Please..



github.com/cloud-barista/cb-tumblebug/issues

So, we need you ! (not just a new feature)

The screenshot shows the GitHub repository page for 'cloud-barista/cb-tumblebug'. The repository has 1,100 commits and 15 forks. It features a sidebar with links to 'Code', 'Issues', 'Pull requests', 'Discussions', 'Actions', 'Projects', 'Wiki', 'Security', 'Insights', and 'Settings'. The 'About' section describes it as a 'Cloud-Barista Multi-Cloud Infra Service (MCIS) Management Framework'. It lists various contributors and their commits. The 'Contributors' section shows 10 contributors with their GitHub profiles and icons. A chart at the bottom shows the distribution of languages: Go (66.2%), Shell (33.6%), and Other (0.2%).

Cloud-Barista Multi-Cloud Infra Service (MCIS) Management Framework

Contributors: 10

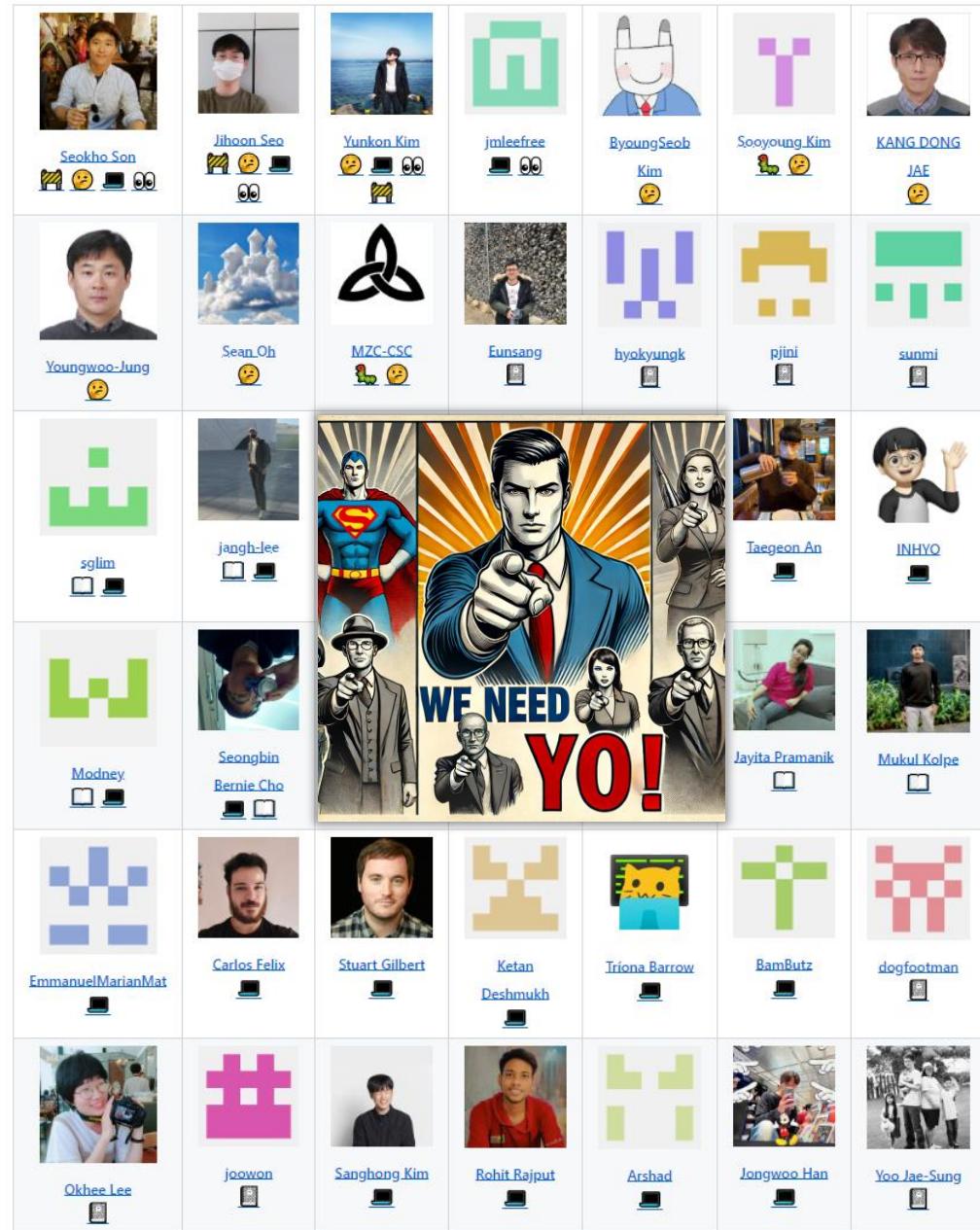
Languages: Go 66.2%, Shell 33.6%, Other 0.2%

CB-Tumblebug GitHub

<https://github.com/cloud-barista/cb-tumblebug>

Contributors

Thanks goes to these wonderful people (emoji key):





Thank you!



@seokho-son

@seokho-son

@seokho son (CNCF / K8s Slack)

Senior Researcher & Special Fellow, ETRI

Ambassador, CLOUD NATIVE COMPUTING FOUNDATION

Maintainer, Cloud-Barista CB-Tumblebug ☕

Maintainer, M-CMP Platform ☕

Maintainer, CNCF Cloud Native Glossary ☕

Lead, Kubernetes SIG-Docs Subproject 🛡

Lead, Kubernetes SIG-Docs Korean I10n Team 🛡

멀티 클라우드에 진심인 사람들의 이야기

멀티/분산 클라우드, 차세대 클라우드를 향한 도전과 기회

Cloud-Barista Community the 9th Conference

감사합니다.

shsonkorea@etri.re.kr

손석호