Portal Architecture Description

NVIDIA DGX-1

AI 研究主要利器

深度學習的最快途徑

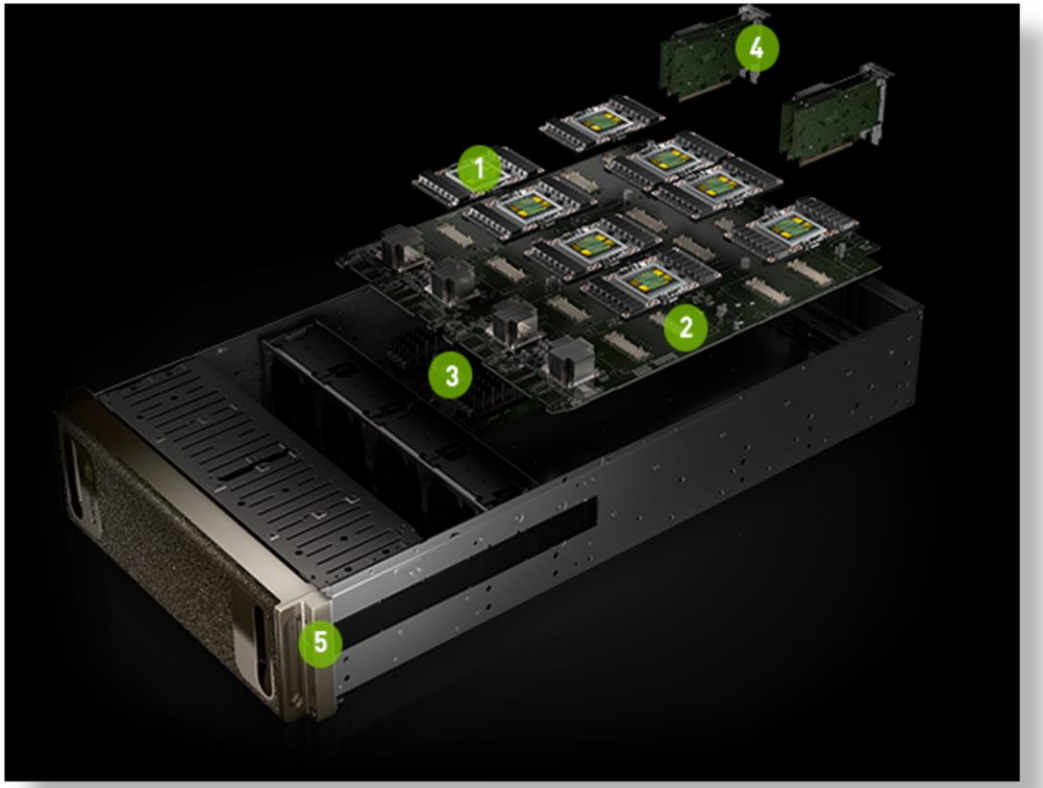
若您規劃在公司裡實作人工智慧，則須謹慎選擇並 整合複雜的應用軟體和硬體。NVIDIA® DGX-1™ 透過立即可用的解決方案加快您的計畫，讓您可以在數小 時內獲得見解，而非數週或數個月。

NVIDIA DGX-1
規格

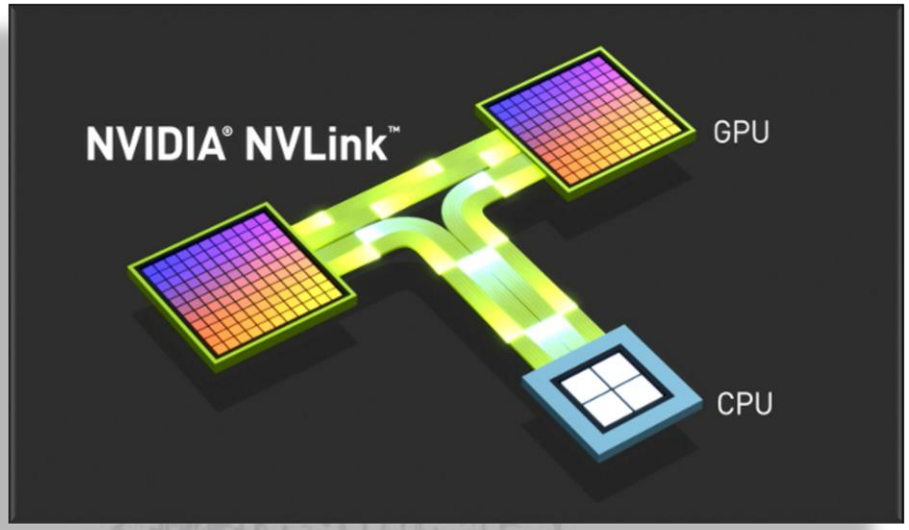
SYSTEM SPECIFICATIONS	
GPUs	8X Tesla V100
Performance (Mixed Precision)	1 petaFLOPS
GPU Memory	256 GB total system
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz
NVIDIA CUDA® Cores	40,960
NVIDIA Tensor Cores (on V100 based systems)	5,120
Power Requirements	3,500 W
System Memory	512 GB 2,133 MHz DDR4 RDIMM
Storage	4X 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Operating System	Canonical Ubuntu, Red Hat Enterprise Linux
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)
Packing Dimensions	1,180 D x 730 W x 284 H (mm)
Operating Temperature Range	5–35 °C

DGX-1的強大硬體組件

1. NVIDIA TESLA V100
第一個整合專為人工智慧量身打造之 Tensor
核心技術的 GPU 架構。
2. NEXT GENERATION NVIDIA NVLINK
每個 GPU 的高速互連速度每秒達 300 GB，
比目前的 PCIe Gen3 x16 互連速度快 10 倍。
3. 雙 Intel Xeon CPU
針對開機、儲存空間管理和深度學習架構協調。
4. QUAD EDR IB
兼具高頻寬與低延遲特性，通訊速度每秒共 800 GB。

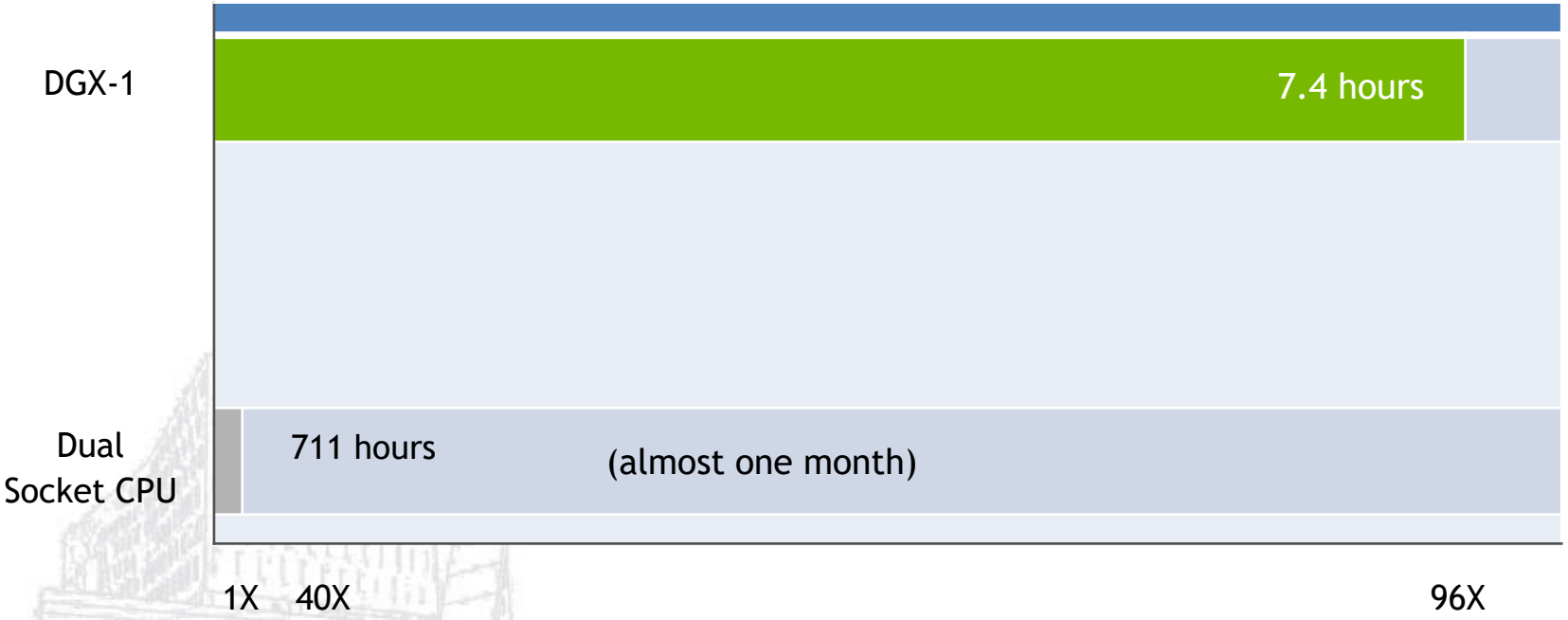


NVIDIA-NVLINK



- GPU 可以快速地處理大量數據。但是唯有當龐大數據可以被源源不絕傳送至 GPU，這項能力才可以徹底發揮，而 PCIe 互連技術往往無法跟上節奏。
- 為避免這樣的「交通壅塞」，我們針對 CPU 和 GPU 間，以及 GPU 間發明了更快速的互連技術，我們稱它為NVLink。
- 這是全球首個針對 GPU 的高速互連技術。
  NVIDIA NVLink 為下一世代的高效能運算(HPC)創造了 數據高速公路。比起 PCIe，這項技術能讓 GPU 和 CPU 彼此交換數據的速度提升5至12倍之多。

DGX-1: 比CPU運算快96倍



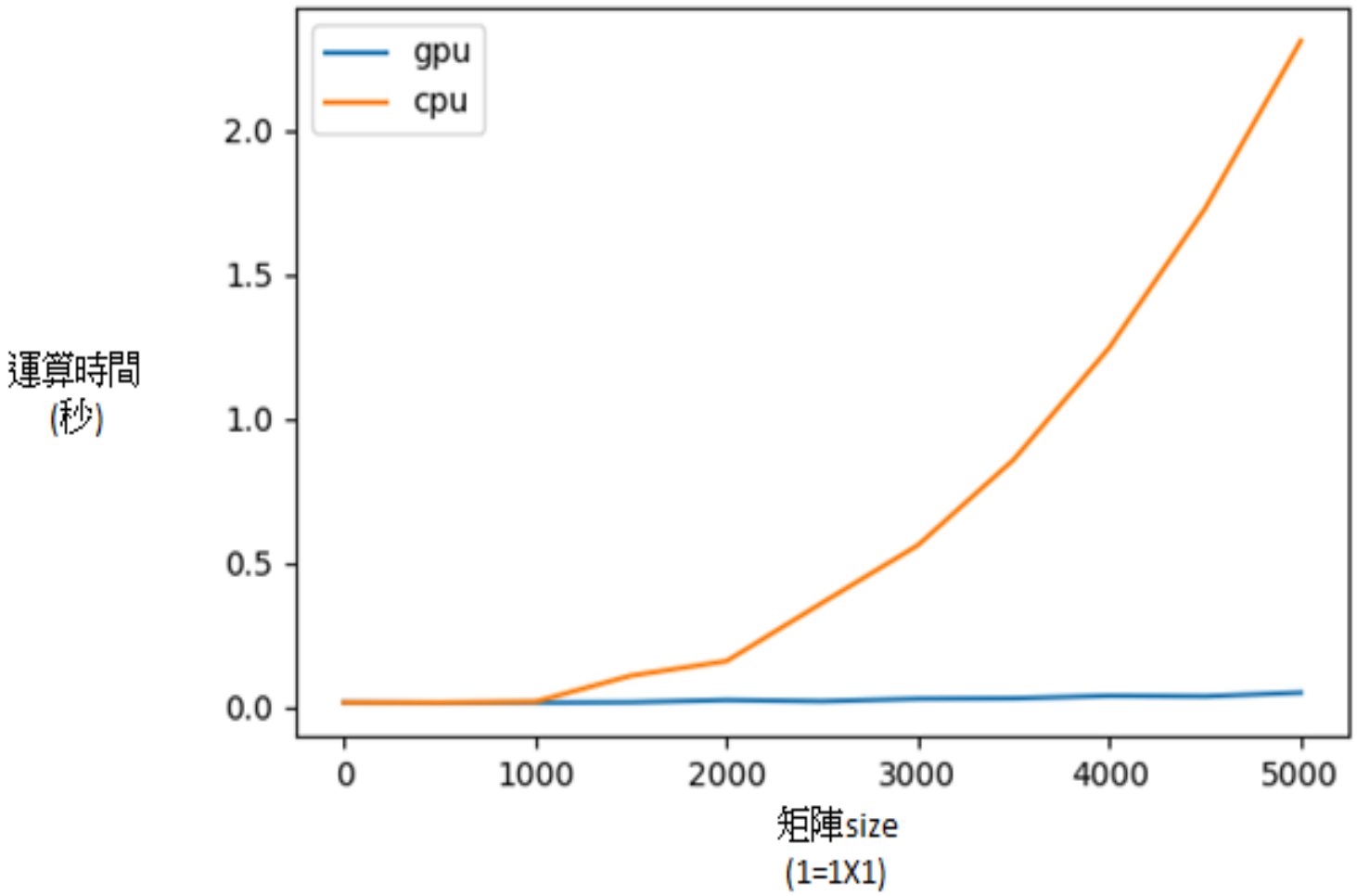
Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

Network Performance of CPU & GPU

Compute Context

- TF Version: 1.14.0
- Platform: Linux-4.15.0-47-generic-x86_64-with-Ubuntu-18.04-bionic
- CPU: Intel Xeon E5-2698 V4 2.2 GHz
- CPU RAM: 504 GB
- GPU: Tesla V100-SXM2-32GB
- GPU RAM: 29.8 GB
- CUDA Version: 10.1
- CUDA Build: V10.1.243

Performance of Array Computing



Network Performance

Network	MobileNet-V2 [classification]		Inception-V3 [classification]		Inception-V4 [classification]		ResNet-V2-50 [classification]	
Batch size	50		20		10		10	
Size	224x224		346x346		346x346		346x346	
Model	Inference (ms)	training (ms)	Inference (ms)	training (ms)	Inference (ms)	training (ms)	Inference (ms)	training (ms)
CPU	1828 ± 148	9583 ± 1612	3234 ± 49	18448 ± 567	3671 ± 35	17194 ± 391	2825 ± 98	10810 ± 355
CPU+GPU	67.2 ± 116.1	143 ± 154	50.6 ± 4.0	176 ± 4	46.8 ± 2.7	195 ± 5	60.0 ± 1.6	99.0 ± 1.8

Network Performance

Network	VGG-16 [classification]		VGG-19 Super-Res [image-to-image mapping]		SRCNN 9-5-5 [image-to-image mapping]		LSTM-Sentiment [sentence sentiment analysis]	
Batch size	20	2	10		10		100	10
Size	224x224		256x256	224x224	512x512		1024x300	
Model	Inference (ms)	training (ms)	Inference (ms)	training (ms)	Inference (ms)	training (ms)	Inference (ms)	training (ms)
CPU	5288 ± 57	5277 ± 90	8216 ± 86	43212 ± 54	4791 ± 61	35565 ± 120	34748 ± 633	32963±111
CPU+GPU	53.9 ± 0.7	80.7 ± 0.6	62.3 ± 2.1	167 ± 1	54.4 ± 1.3	150 ± 6	586 ± 9	749 ± 103

AI Score of Device

	Device Inference Score:	Device Training Score	Device AI Score:
CPU	187	189	376
CPU+GPU	15925	16044	31969

1 - The final AI Score for this device was estimated based on its inference score

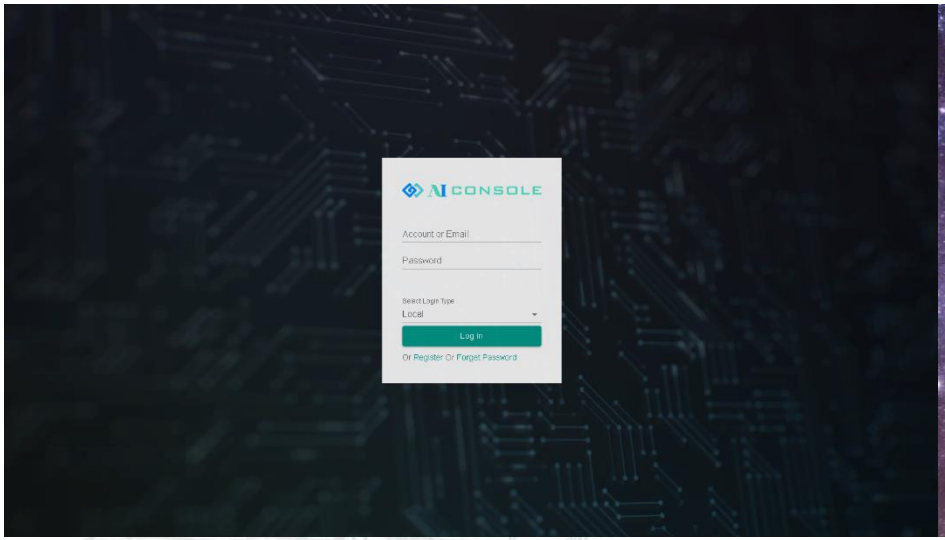
2 - The final AI Score for this device was estimated based on its training score

3 - This device might be using unofficial / prototype hardware or drivers

4 - These are the results of an early prototype. The results of the commercial device might be different

實機運行

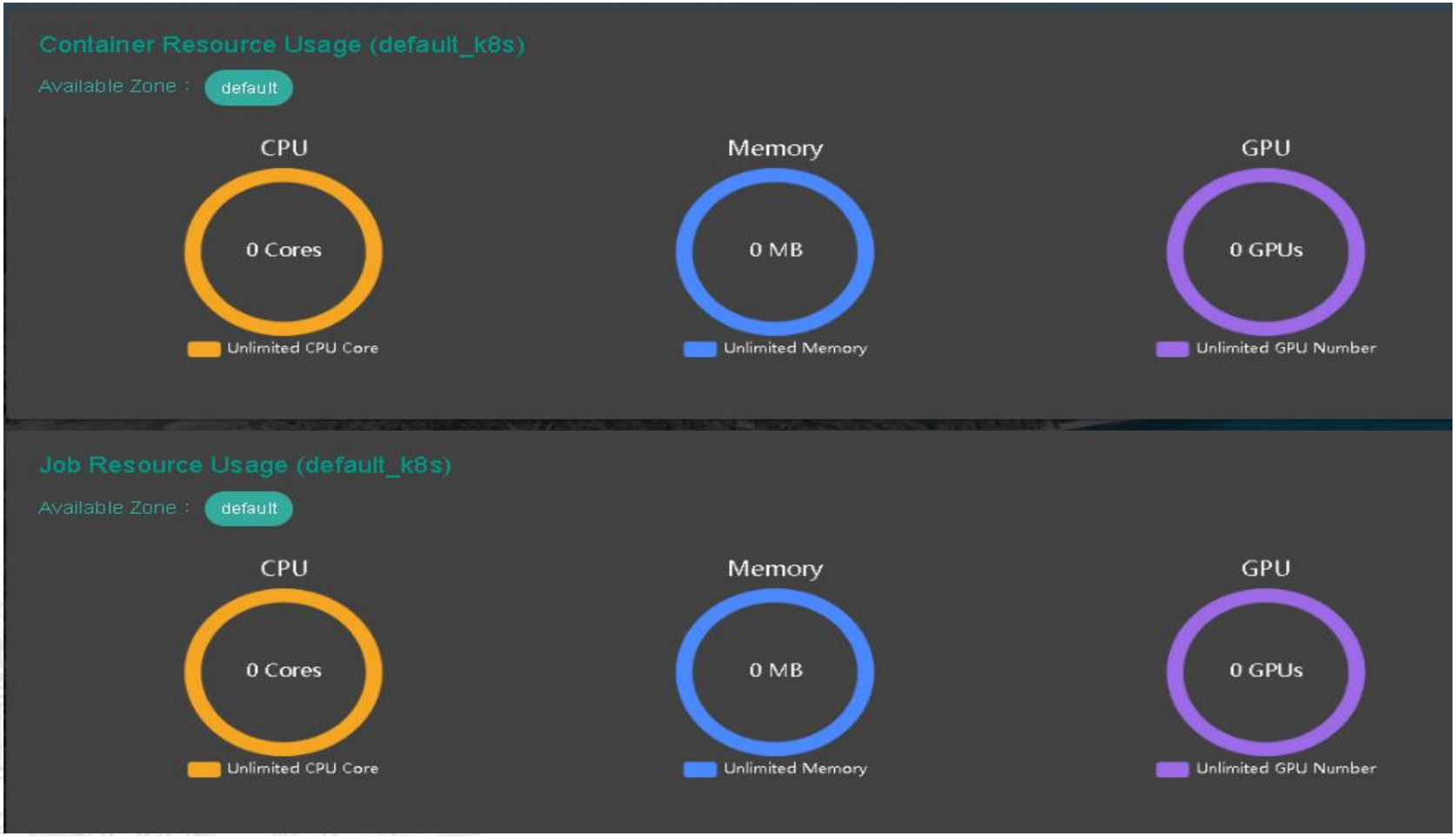
登入畫面


AI CONSOLE
Account or Email
Password
Select Login Type
Local
Log in
Or Register Or Forget Password

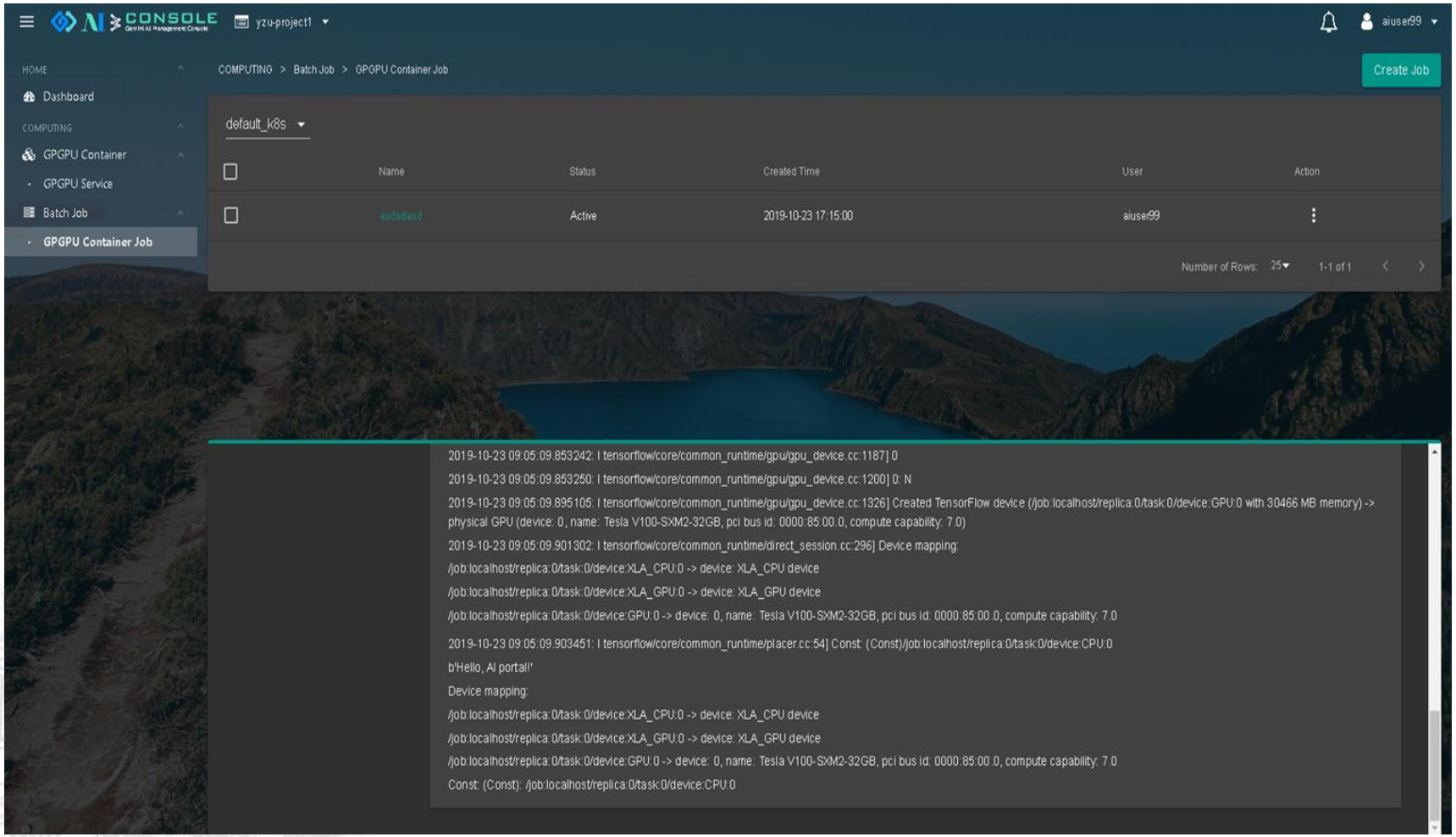
密碼修改


自助密码服务
邮件
元智大學
Yuan Ze University
邮件发送密码重置链接
输入您的用户名重置您的密码。收到邮件后，点击链接完成重置密码。
用户名
用户名
提交

資源監測


Container Resource Usage (default_k8s)
Available Zone : default
CPU
0 Cores
Unlimited CPU Core
Memory
0 MB
Unlimited Memory
GPU
0 GPUs
Unlimited GPU Number
Job Resource Usage (default_k8s)
Available Zone : default
CPU
0 Cores
Unlimited CPU Core
Memory
0 MB
Unlimited Memory
GPU
0 GPUs
Unlimited GPU Number

實機運行


CONSOLE
yzu-project1
aiuser99
HOME
Dashboard
COMPUTING
GPGPU Container
GPGPU Service
Batch Job
GPGPU Container Job
COMPUTING > Batch Job > GPGPU Container Job
Create Job
default_k8s
Name
Status
Created Time
User
Action
asdsdasd
Active
2019-10-23 17:15:00
aiuser99
Number of Rows: 25
1-1 of 1
2019-10-23 09:05:09.853242: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1187] 0
2019-10-23 09:05:09.853250: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1200] 0: N
2019-10-23 09:05:09.895105: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1326] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 with 30466 MB memory) -> physical GPU (device: 0, name: Tesla V100-SXM2-32GB, pci bus id: 0000:85:00.0, compute capability: 7.0)
2019-10-23 09:05:09.901302: I tensorflow/core/common_runtime/direct_session.cc:296] Device mapping:
/job:localhost/replica:0/task:0/device:XLA_CPU:0 -> device: XLA_CPU device
/job:localhost/replica:0/task:0/device:XLA_GPU:0 -> device: XLA_GPU device
/job:localhost/replica:0/task:0/device:GPU:0 -> device: 0, name: Tesla V100-SXM2-32GB, pci bus id: 0000:85:00.0, compute capability: 7.0
2019-10-23 09:05:09.903451: I tensorflow/core/common_runtime/placer.cc:54] Const: (Const)/job:localhost/replica:0/task:0/device:CPU:0
b'Hello, AI portal!'
Device mapping:
/job:localhost/replica:0/task:0/device:XLA_CPU:0 -> device: XLA_CPU device
/job:localhost/replica:0/task:0/device:XLA_GPU:0 -> device: XLA_GPU device
/job:localhost/replica:0/task:0/device:GPU:0 -> device: 0, name: Tesla V100-SXM2-32GB, pci bus id: 0000:85:00.0, compute capability: 7.0
Const: (Const): /job:localhost/replica:0/task:0/device:CPU:0

THANKS FOR WATCHING