Use Tesla to provide first GPU VM Service in China

Feng Zhu

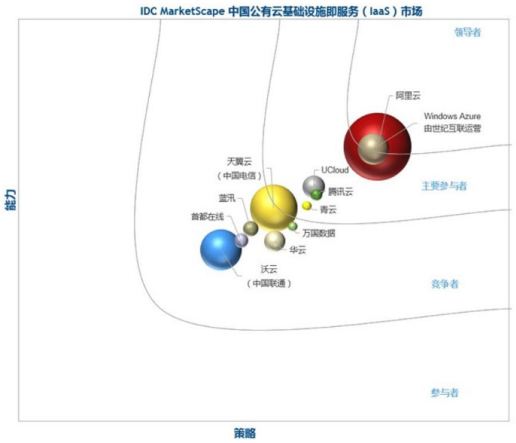
专注 • 服务 • 中立

Outline

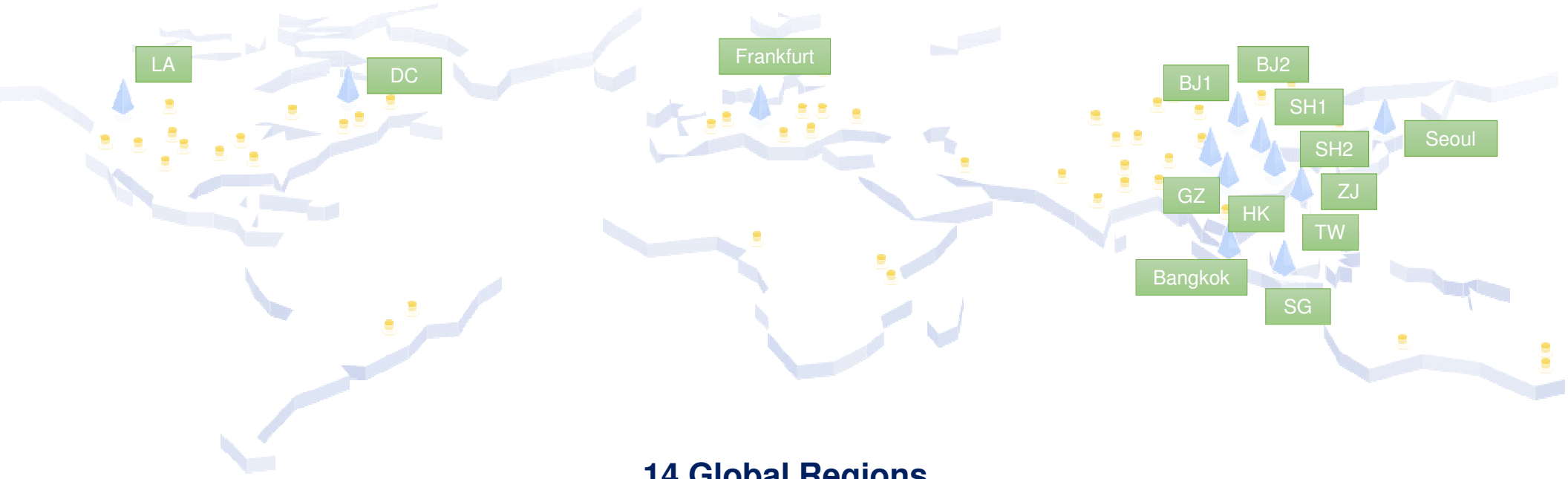
- UCloud Introduction
- K80 GPU VM
- P40 GPU VM
- UCloud GPU PaaS Service: UAI-Service
- UCloud GPU ecosystem

About UCloud

- **Top 3 IaaS Provider in China**
- **Found in 2012**
- **HQ in Shanghai**
- **Served 50,000+ Enterprise**

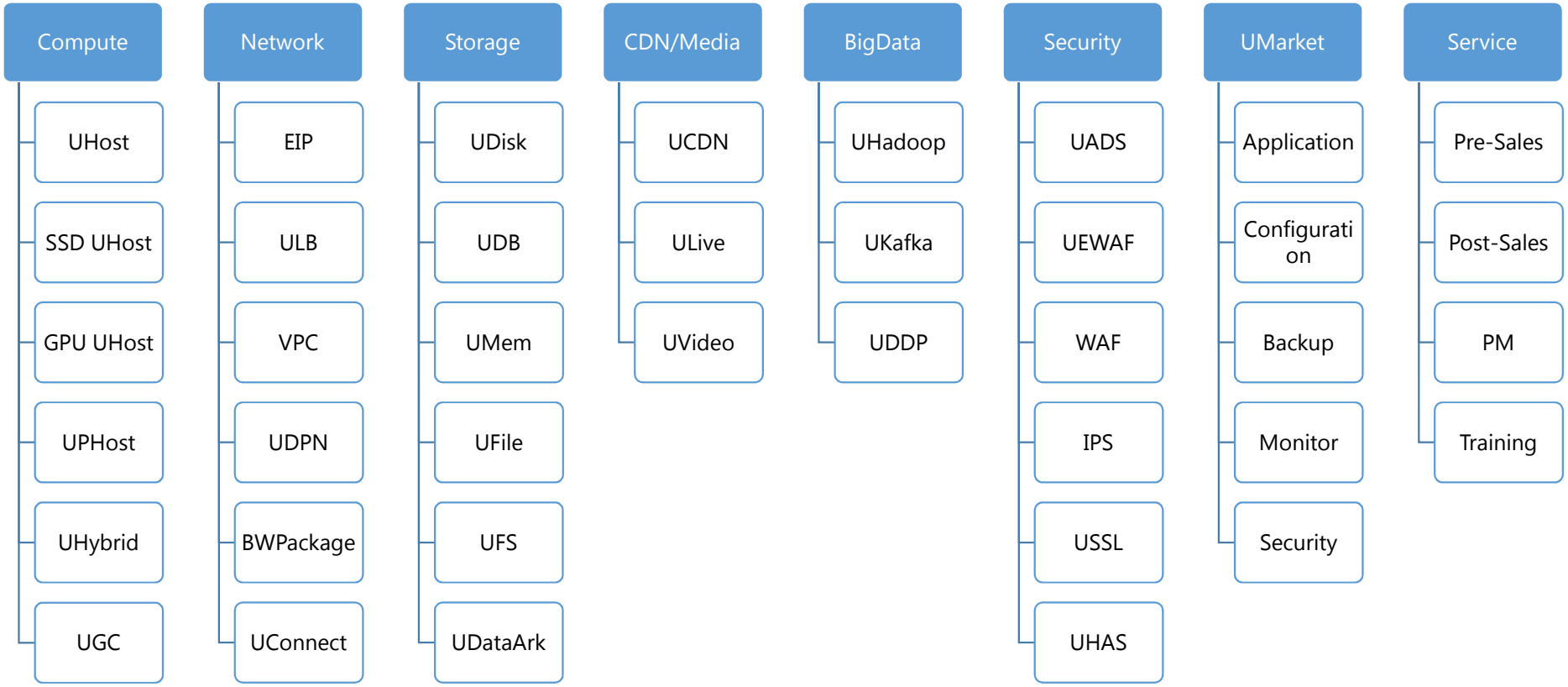


Data Centers



14 Global Regions

UCloud Product Line



GPU Timeline

- 2012 UCloud founded
- 2015.11 K80 GPU VM
- 2016.2 K80 GPU Physical Machine
- 2017.5 P40 GPU VM
- 2017.? P40 GPU Physical Machine

GPU Decision: Virtualization

PCI Pass-through	Grid
Single VM occupy GPU device	GPU device shared by multiple VMs
Performance guaranteed	No guarantee on performance
Limited by PCI device number （K80 = 2 * GPU）	no limit for PCI device
No license	License
√	X

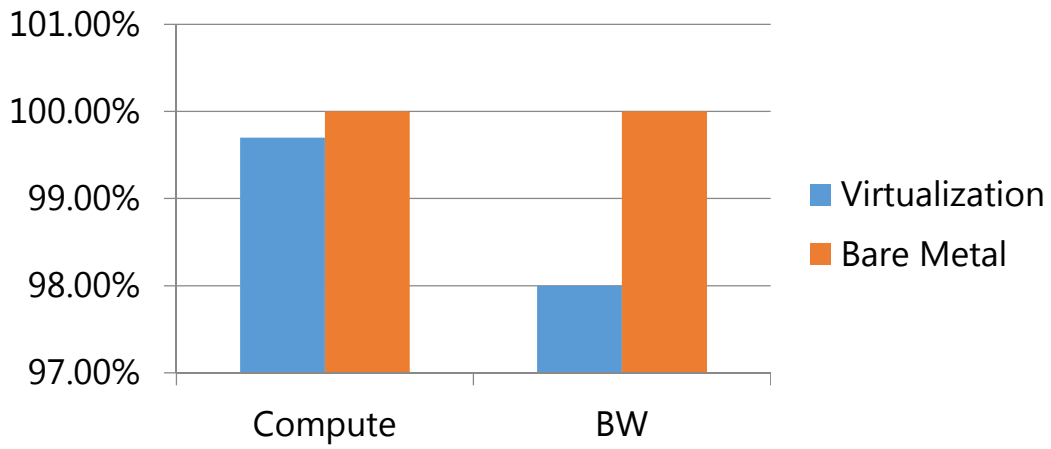
VM Advantage

- **Flexibility for VM configuration**
  - CPU、Memory、Disk size、GPU number are all flexible
  - SDN network flexible
- **Main OS all supported, Win/Linux**
  - CentOS 6.5/CentOS 7.0/Ubuntu 14.04/Ubuntu 12.04/Gentoo 2.2/Win 2008/Win 2012
- **Fast Deployment**
  - Based on self-defined image, can deploy 1000 VMs in 1 minute

VM Performance Degrade

- **Using Pass-through Technology, almost no degradation**

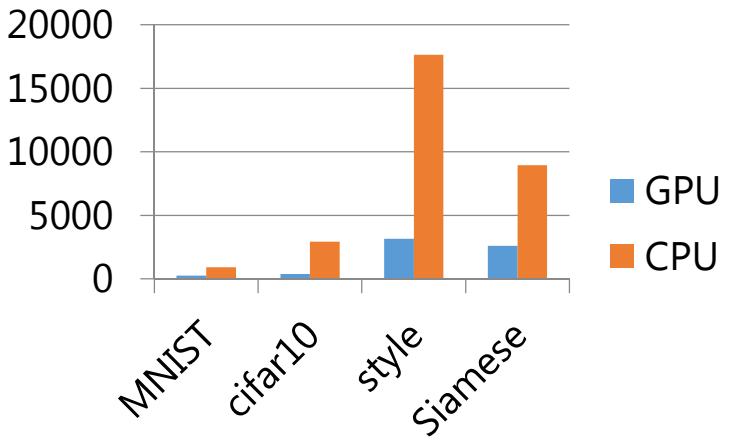
Degradation	Virtualization	Bare Metal
Computing	99.3%	100%
Bandwidth	98%	100%



UCloud GPU Virtualization – DL test

- **Caffe Performance （Ubuntu）**

Cases	iters	GPU(secs)	CPU(secs)	Speedup
LeNet on MNIST	10000	253.9	900.3	**3.5**
cifar10	5000	374.2	2931.0	**7.8**
Fine-tuning for style recognition	10000	3138.9	17634.8	**5.6**
Siamese Network Training	50000	2584.2	8937.1	**3.5**



UCloud GPU Virtualization – DL test(2)

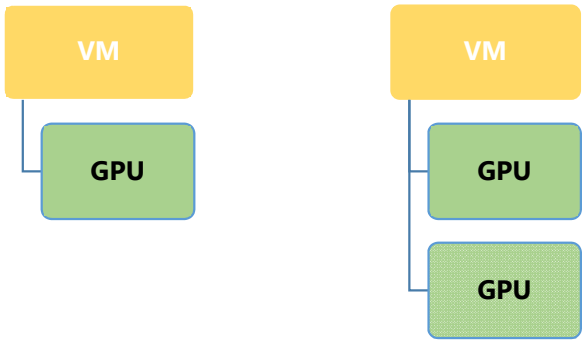
- **Theano/Keras (Ubuntu)**

Cases	iters	GPU(secs)	CPU(secs)	**Speedup**
imdb_lstm.py	20000	47	239	**5.1**
imdb_cnn.py	20000	11	563	**51.2**
imdb_cnn_lstm.py	20000	27	236	**8.7**
addition_rnn.py	45000	5	32	**6.4**
cifar10_cnn.py	50000	198	2670	**13.5**
babi_rnn.py	950	9	22	**2.4**
mnist_cnn.py	60000	23	1332	**57.9**
mnist_irnn.py	60000	270	451	**1.7**
mnist_mlp.py	60000	1	5	**5.0**

K80 Physical Machine

Hardware	Specification
GPU	Tesla K80
CPU	Intel E5 2630*2
Memory	192G
Disk	2T SSD
Networking	10G NIC*4

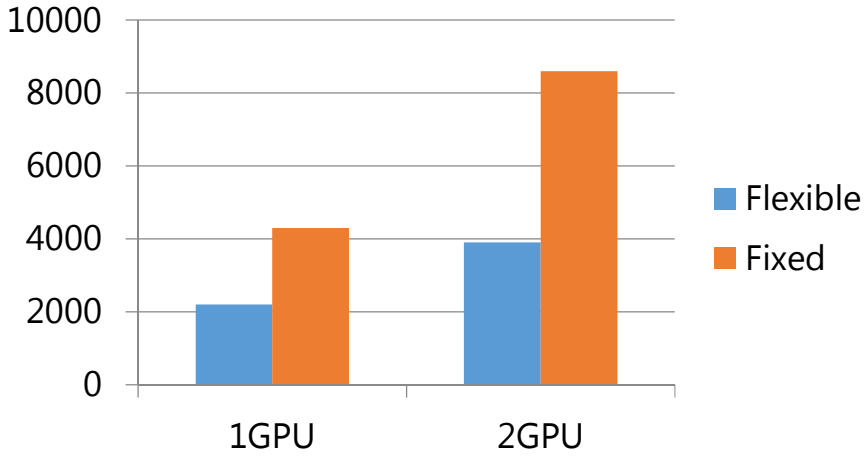
VM Configuration - K80



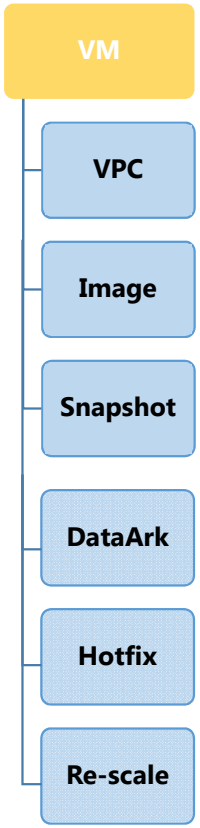
CPU	4C	8C	16C	
Memory	8G	16G	32G	64G
Disk	100G - 1T			

Flexible VM Save Cost

Configuration	Fixed	Flexible
CPU	16C	4C
Memory	96G	8G
Disk	1T	100G
GPU	1	1
Price	4,300 CNY/month	2,200 CNY/month
USD Price	$615/m	$315/m



GPU VM Features



Networking

Self-defined images, deploy 1000 VMs in 1 min

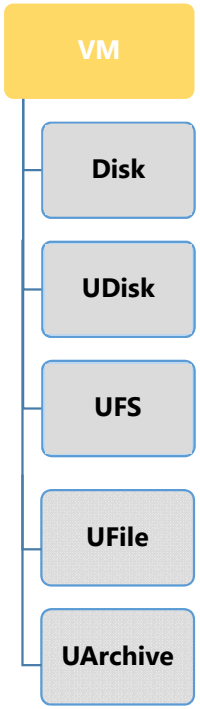
Data backup

24 continuous data protection, call rollback to any second

Kernel patch without system shutdown

Resize CPU/Memory/Disk anytime

Storage Solution



Local SSD disk

NAS, no limit on device numbers

NFS file system

Object storage

Low cost cloud archive

Create GPU VM

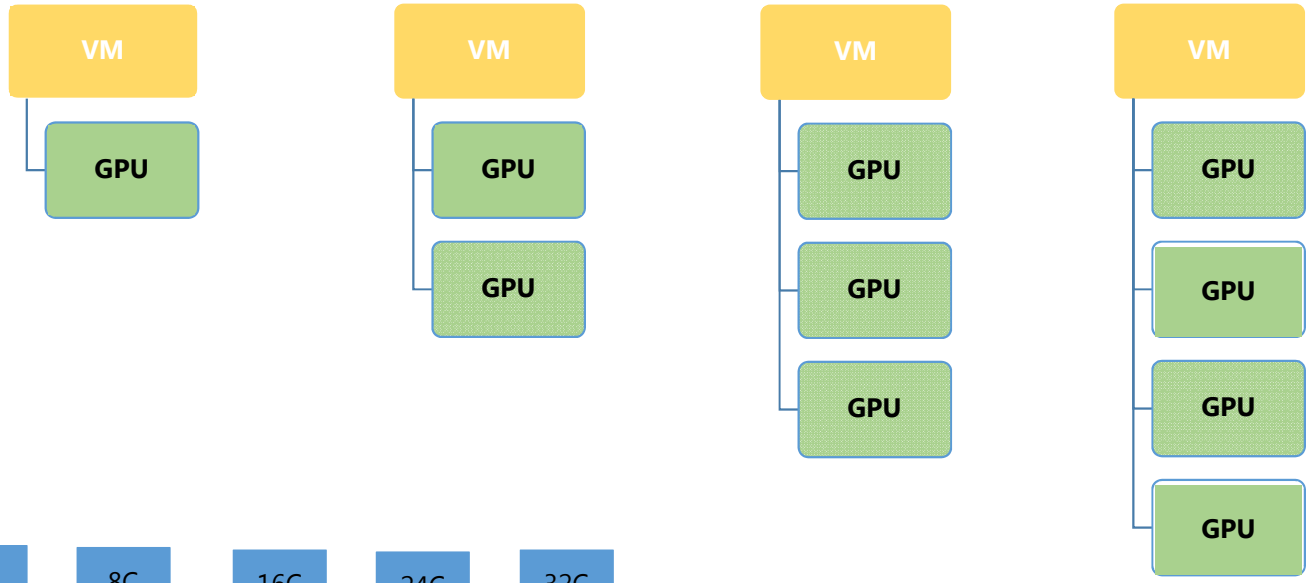
Create GPU VM

Create GPU VM

P40 Physical Machine

Hardware	Specification
GPU	Tesla P40*4
CPU	Intel E5 2650*2
Memory	256G
Disk	3T SSD
Networking	10G NIC*4

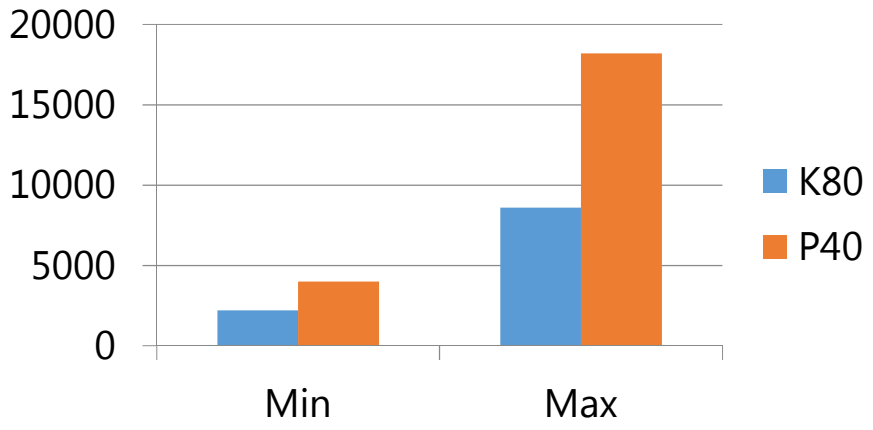
VM Configuration - P40



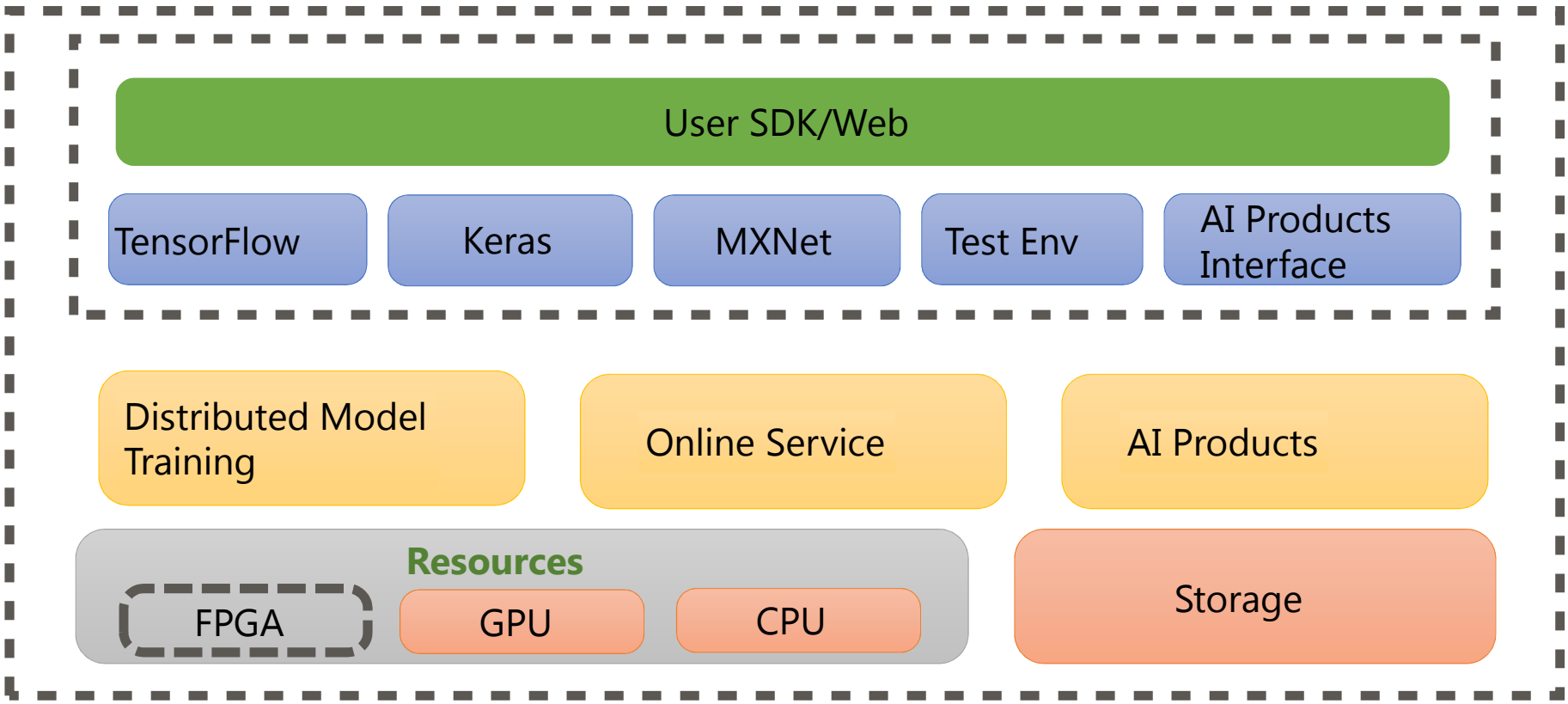
CPU	4C	8C	16C	24C	32C	
Memory	8G	16G	32G	64G	96G	128G
Disk	100G - 1T					

P40 Price

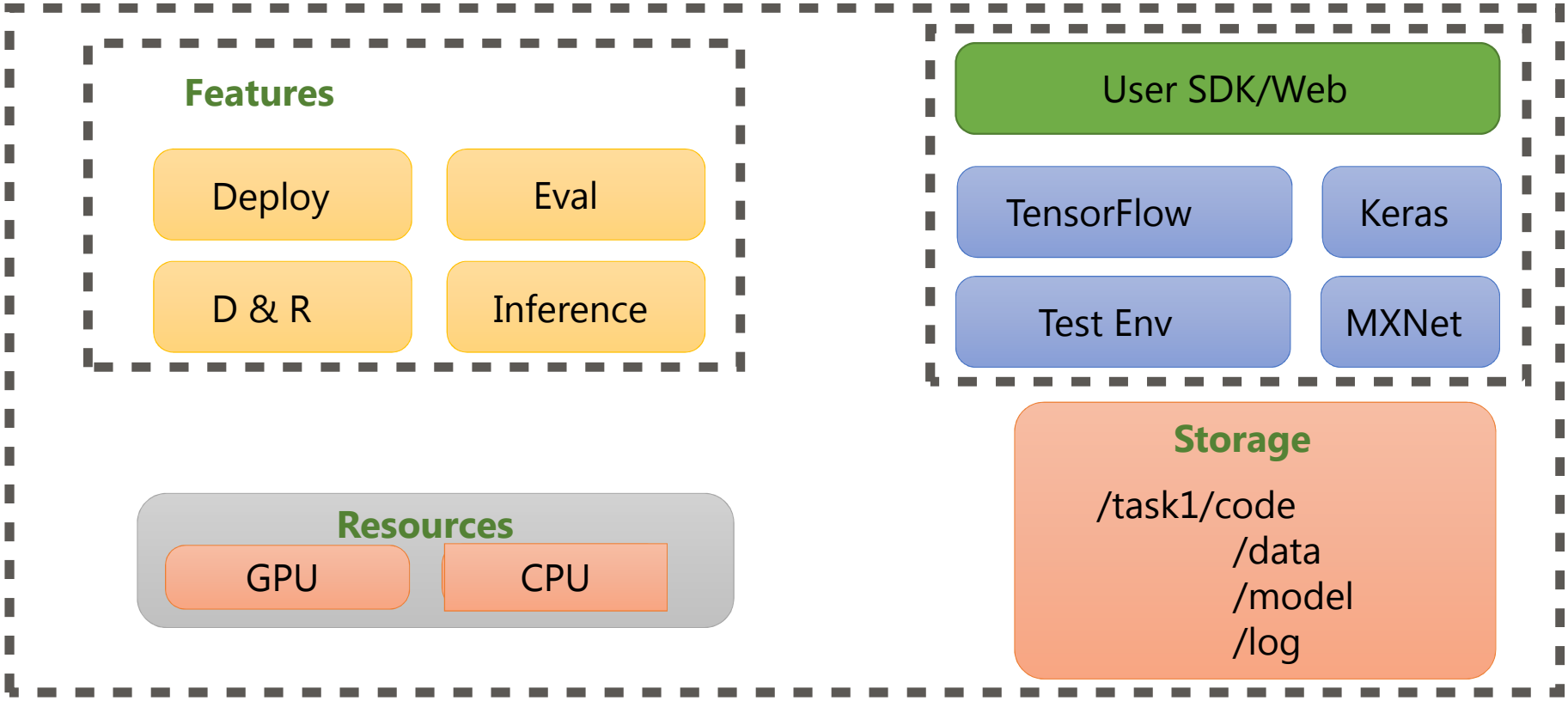
Configuration	Spec 1	Spec 2
CPU	4C	32C
Memory	8G	128G
Disk	100G	1T
GPU	1	4
Price	4,000 CNY/month	18,200 CNY/month
USD Price	$570/m	$2,600/m



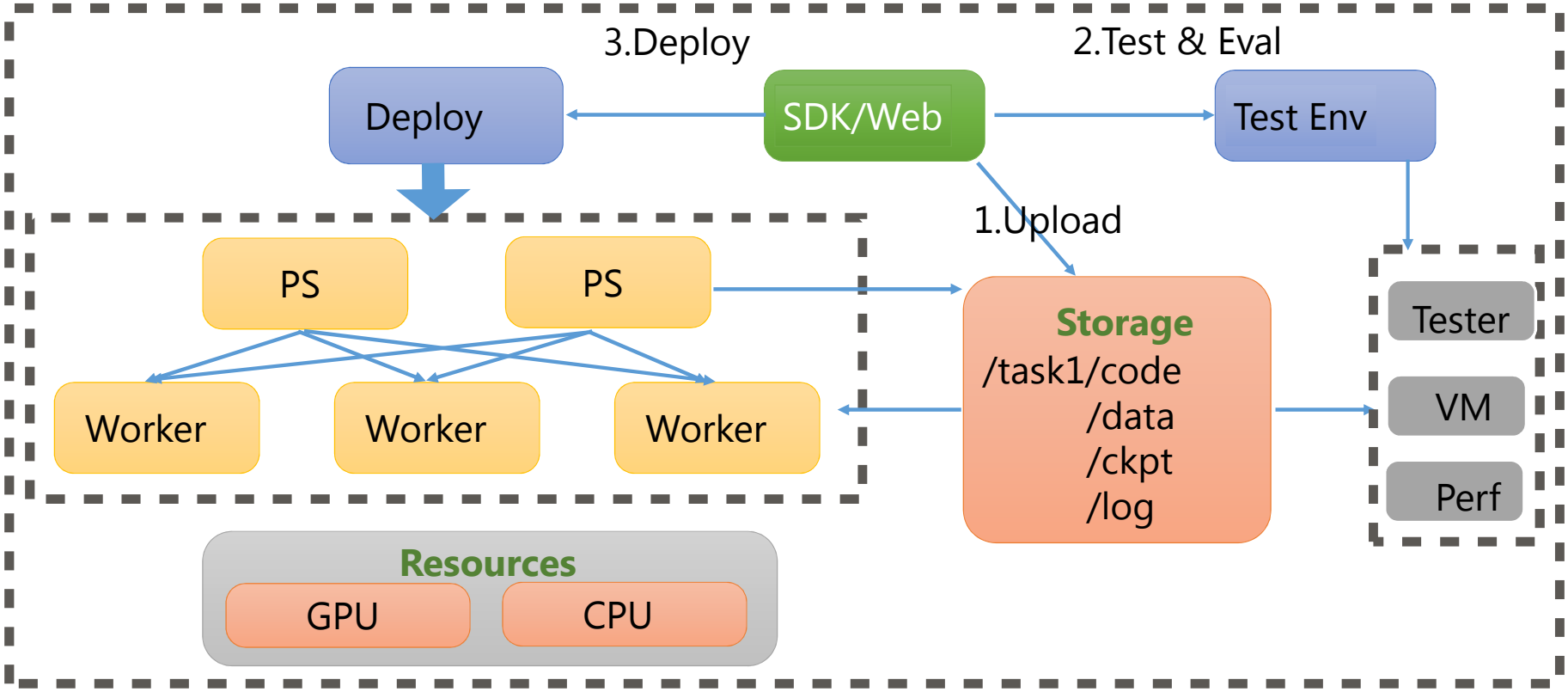
UAI-Service Overview



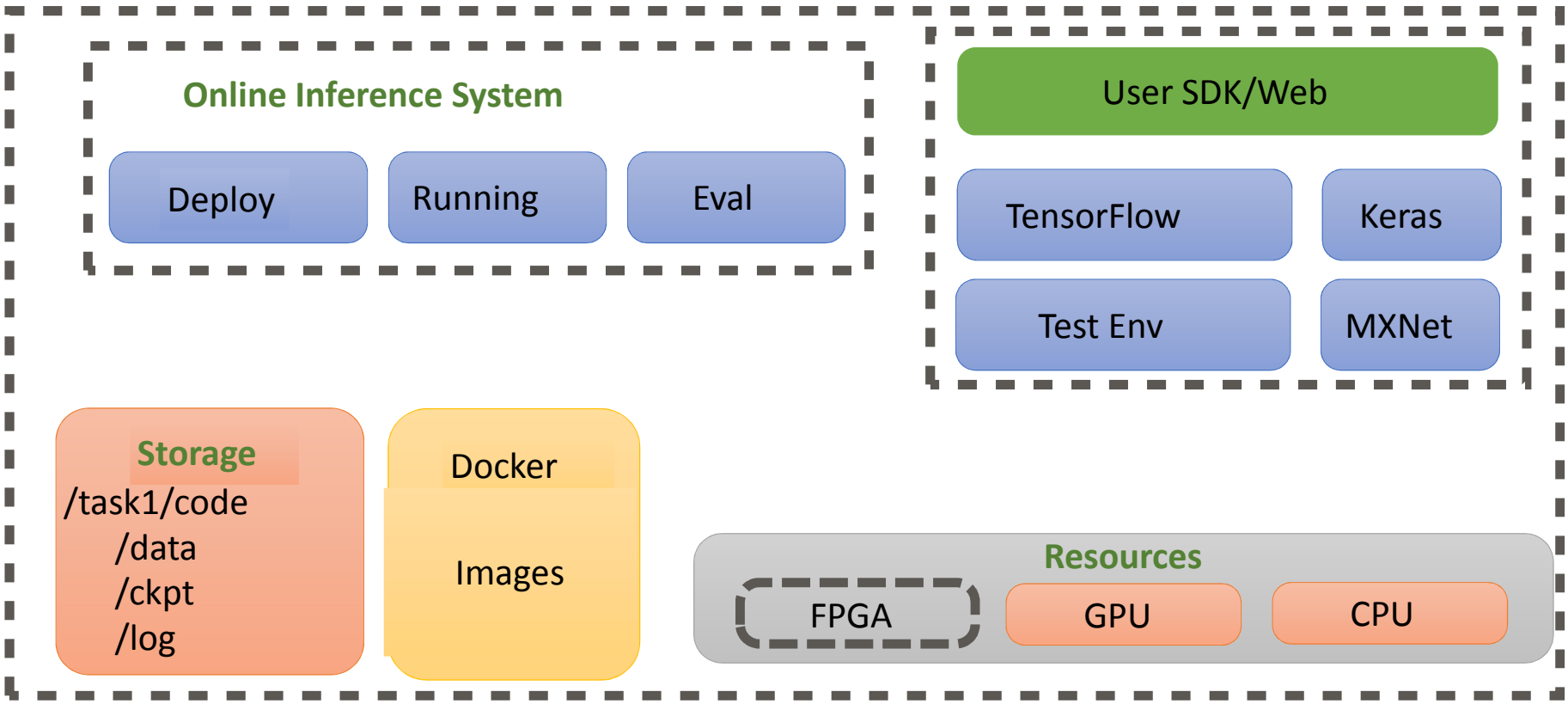
Distributed Training Layout



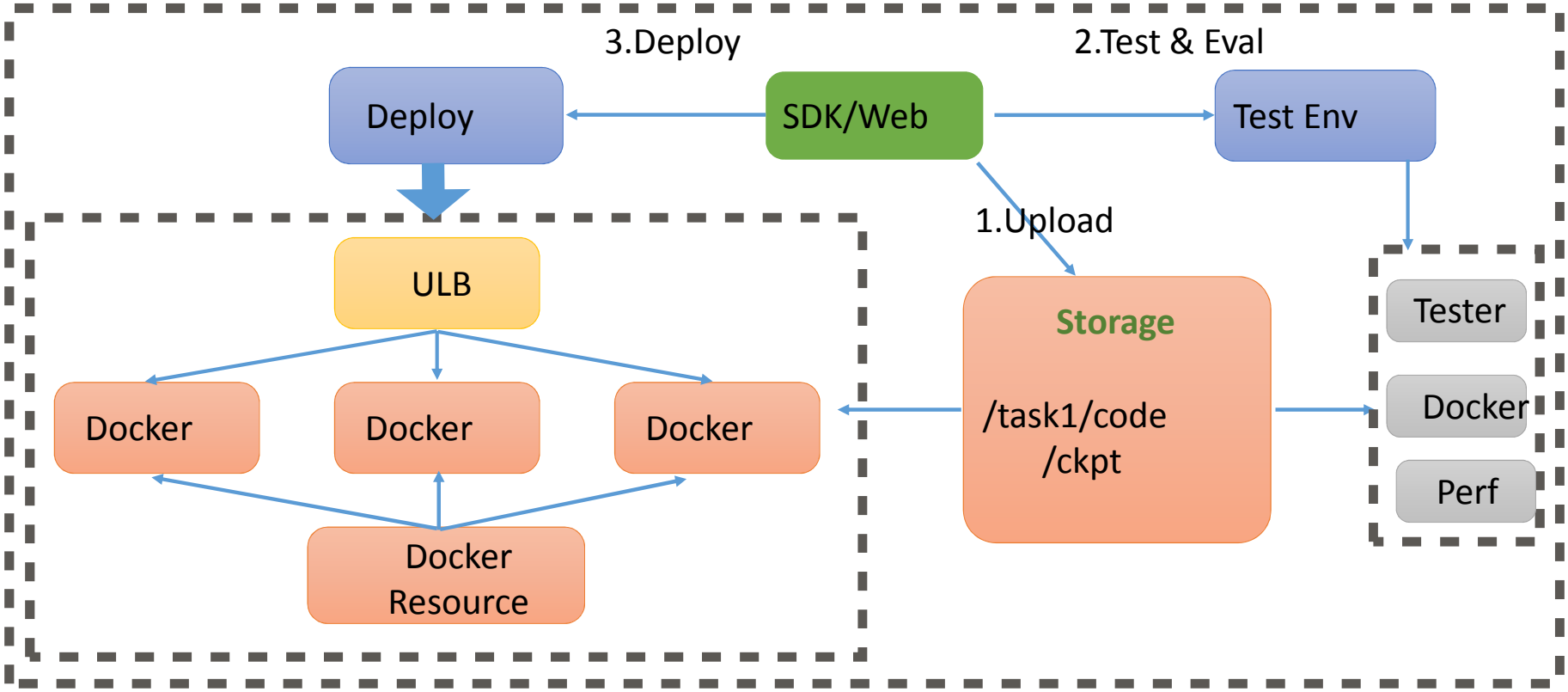
Distributed Training Process



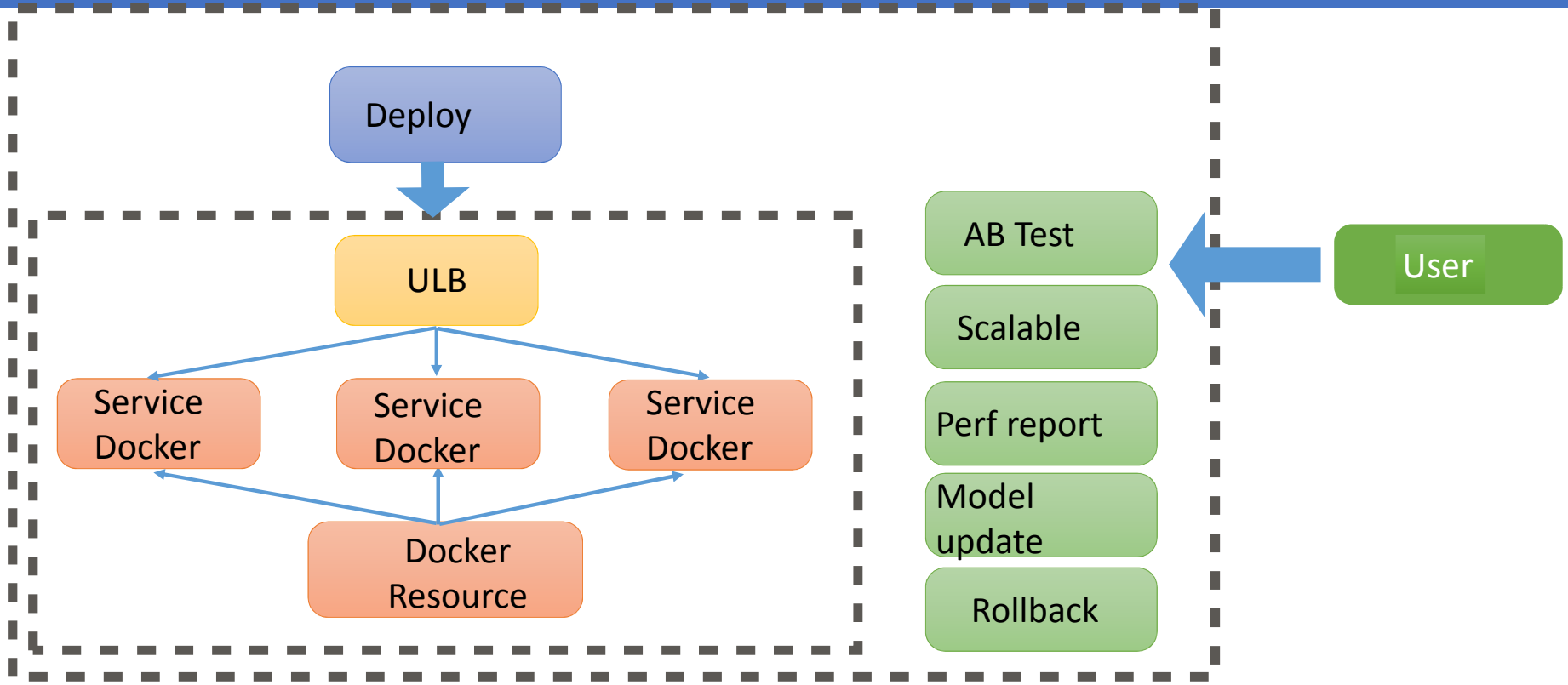
Online Inference Layout



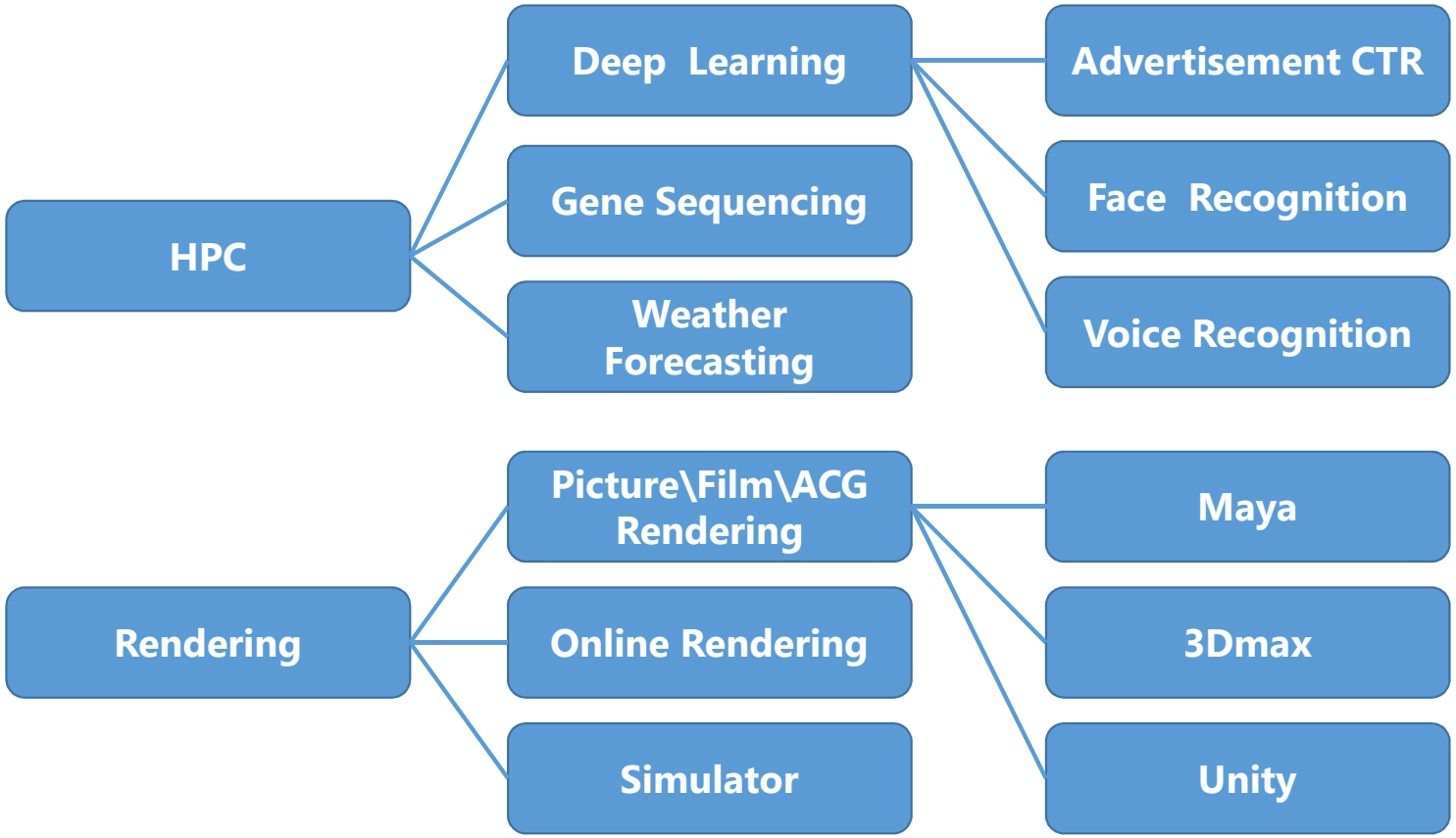
Online Inference Process



Online Inference API/SDK

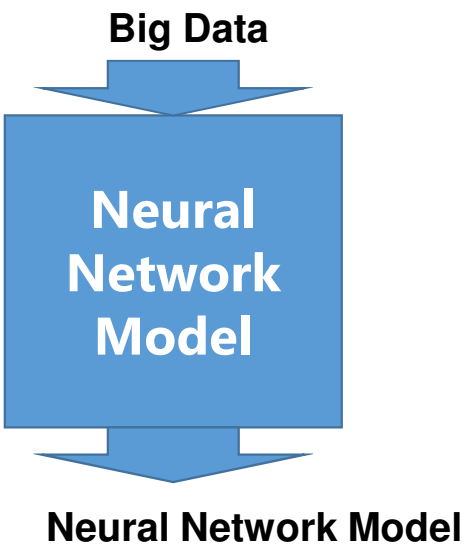


GPU Scenario



GPU Scenario

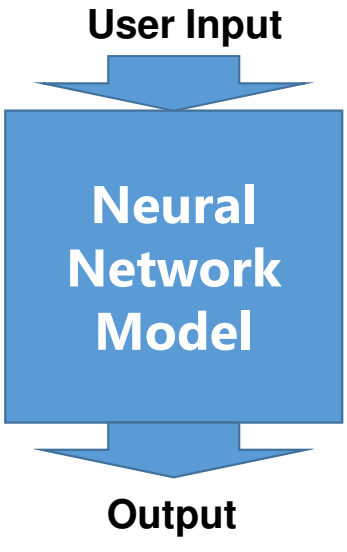
Training



Compute-Intensive（GPU）

- Advertisement CTR
- Face Recognition
- Voice Recognition

Online Service



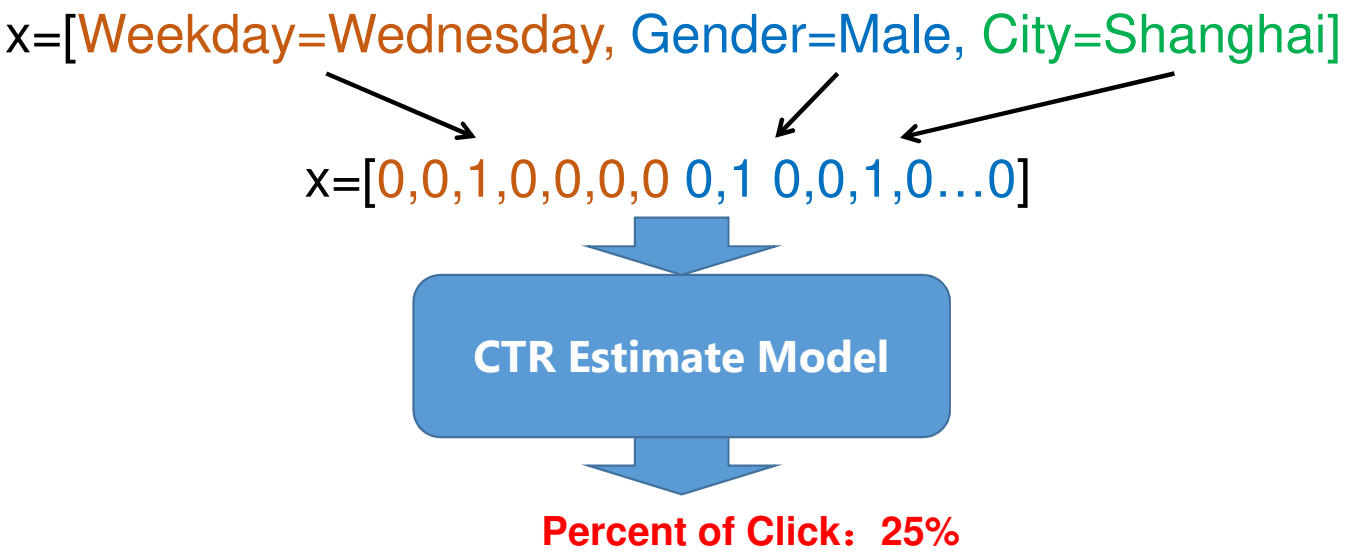
Compute-Sensitive（GPU）

GPU Scenario: Example

- **CTR click-through-rate estimation**
  - Precise Advertise: according to personal historical data
  - Ad bidding & ranking：CTR click rate * Bid Price
  - Video、UGC recommendation：watching ratio

GPU Scenario: Example

- **CTR click-through-rate estimation**



Thank You

www.ucloud.cn

UCLOUD 专业云计算服务商