IaaS : la bataille des GPU embraye sur Ampere

Passage en phase commerciale acté pour les VM A2. Après quasiment un an d'expérimentation, elles entrent au catalogue public de Google Cloud.

La gamme comprend cinq types d'instances. Toutes reposent sur des cœurs de processeur Xeon Cascade Lake. Et surtout, sur des GPU A100. Jusqu'à 16 en l'occurrence, avec la possibilité de leur assortir un maximum de 3 To de SSD local.

Modèle GPU	Machine type (Type de machine)	GPU	Mémoire du GPU	Processeurs virtuels disponibles	Mémoire disponible
NVIDIA® A100	a2-highgpu-1g	1 GPU	40 Go GDDR6	12 processeurs virtuels	85 Go
	a2-highgpu-2g	2 GPU	80 Go GDDR6	24 processeurs virtuels	170 Go
	a2-highgpu-4g	4 GPU	160 Go GDDR6	48 processeurs virtuels	340 Go
	a2-highgpu-8g	8 GPU	320 Go GDDR6	96 processeurs virtuels	680 Go
	a2-megagpu-16g	16 GPU	640 Go GDDR6	96 processeurs virtuels	1360 Go

Les VM A2 sont pour le moment disponibles dans trois régions Google Cloud. Dont une en Europe (Pays-Bas). Le coût des GPU s'ajoute à celui des instances et des autres ressources (disques et images, mise en réseau). Ceux associés à des ressources préemptives sont facturés sur le même modèle. Ceux qui ne le sont pas peuvent bénéficier de remises passé une certaine durée d'utilisation mensuelle.

Les prix ci-dessous valent pour les ressources GPU, RAM et CPU à la demande dans la région Pays-Bas (europe-west-4).

Modèle	GPU	Mémoire des GPU	Prix des GPU (USD)	Prix des GPU préemptifs (USD)	Prix de l'engagement sur un an (USD)	Prix de l'engagement sur trois ans (USD)
NVIDIA® A100	1 GPU	40 Go HBM2	3,100 $ par GPU	0,93 $ par GPU	1,953 $ par GPU	1,395 $ par GPU
	2 GPU	80 Go HBM2				
	4 GPU	160 Go HBM2				
	8 GPU	320 Go HBM2				
	16 GPU	64 Go HBM2				

On trouve aussi des instances en A100 sur Azure, mais en préversion (depuis novembre 2020). Plus précisément un modèle : la ND96asr, dans la série ND A100 v4. Elle embarque 8 GPU.
Les tarifs qui suivent valent pour la région Europe de l'Ouest. Les premiers sont pour des VM Linux ; les seconds, pour des VM Windows (non pris en charge sur les A2 de Google Cloud).

Instance	Cœur	RAM	Stockage temporaire	GPU	À l'utilisation	1 an réservé (% d'économies)	3 an réservé (% d'économies)	Spot (% d'économies) *
ND96asr A100 v4	96	900 Gio	6 500 Gio	8x A100 (NVlink)	29,8166 €/heure	21,0504 €/heure (~29% savings)	12,6720 €/heure (~58% savings)	11,9267 €/heure (~60% savings)

Instance	Cœur	RAM	Stockage temporaire	GPU	À l'utilisation	1 an réservé (% d'économies)	3 an réservé (% d'économies)	Spot (% d'économies) *
ND96asr A100 v4	96	900 Gio	6 500 Gio	8x A100 (NVlink)	33,5406 €/heure	24,7744 €/heure (~26% savings)	16,3960 €/heure (~51% savings)	13,4163 €/heure (~60% savings)

AWS aussi a lancé une instance sur base A100 en novembre dernier, mais pas en préversion : la p4d. Comme les A2, elle repose sur des Xeon Cascade Lake (fréquence de base : 2,2 GHz ; Turbo : 2,9 GHz en monocœur et 3,7 en multicœur).

Instance	GPU	Processeurs virtuels	Mém(Gio)	Bande passante réseau	GPUDirect RDMA	Pair à pair GPU	Stockage	Bande passante EBS
p4d.24xlarge	8	96	1 152	ENA et EFA 400 Gbit/s	Oui	NVSwitch 600 Go/s	8 SSD NVMe 1 To	19 Gbit/s

	vCPU	ECU	Mémoire (Gio)	Stockage des instances (Go)	Utilisation de Linux/UNIX
p4d.24xlarge	96	345	1 152 Gio	8 disques SSD 1000	35,39655 USD par heure

Illustration principale © NVIDIA