Ampere : la nouvelle arme de Nvidia pour s'imposer dans l'IA

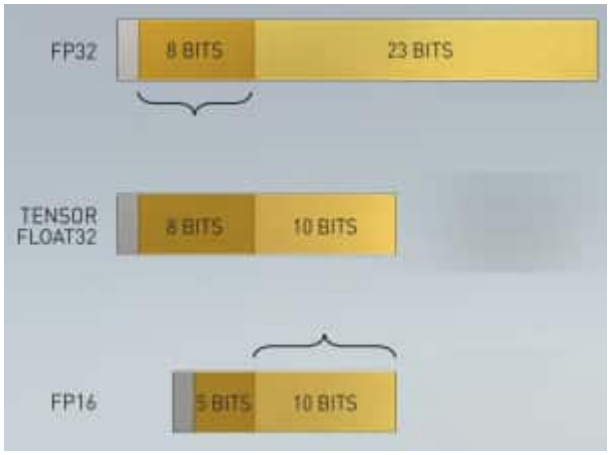
Le voile est levé sur Ampere.

Nvidia a officialisé, la semaine passée, cette architecture GPU qui succède à Volta, introduite en 2017.

L'accélérateur A100 en est le premier représentant.
Gravé en 7 nm, il embarque 54,2 milliards de transistors. En comparaison, le V100 (génération Volta) en est à 12 nm et 21,1 milliards de transistors.

Les principales avancées se font au niveau des cœurs Tensor, spécifiques aux réseaux de neurones.

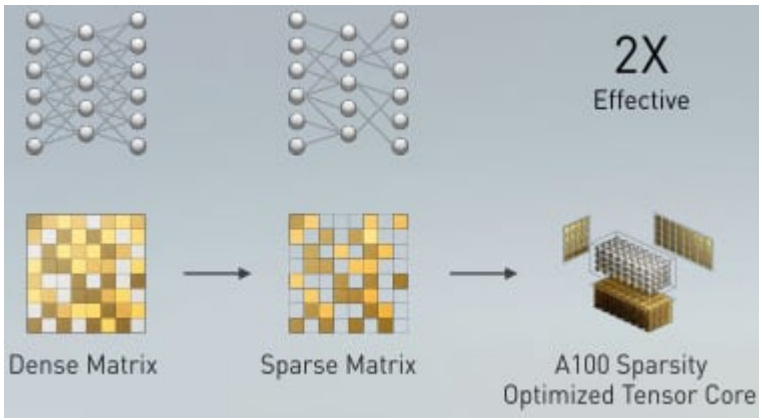
En premier lieu, on note un élargissement des niveaux de précision. Avec la prise en charge de bfloat16, de FP64… et de TF32.
Ce dernier conserve l'exposant 8 bits de FP32 (et donc sa portée), mais adopte la mantisse 10 bits de FP16, accélérant les performances.

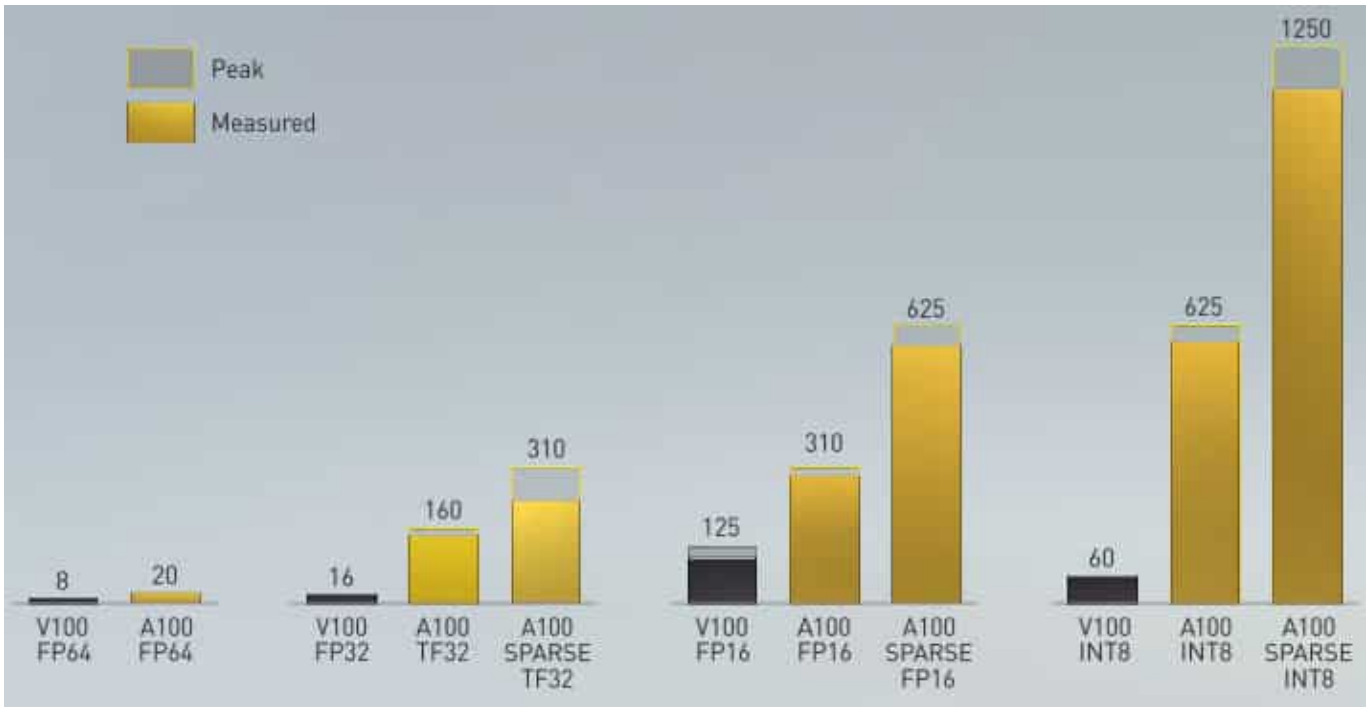


Pour les procédures d'inférence, l'A100 prend en charge les formats INT8, INT4 et INT1. Plus besoin, donc, de s'appuyer sur l'offre Turing, qui avait tendance à compléter Volta sur ce volet.

Avec parcimonie

Entre Volta et Ampere, le débit FMA des cœurs Tensor quadruple.
L'A100 est par ailleurs capable d'exploiter la propriété de parcimonie des réseaux de neurones, avec à la clé un doublement des performances.

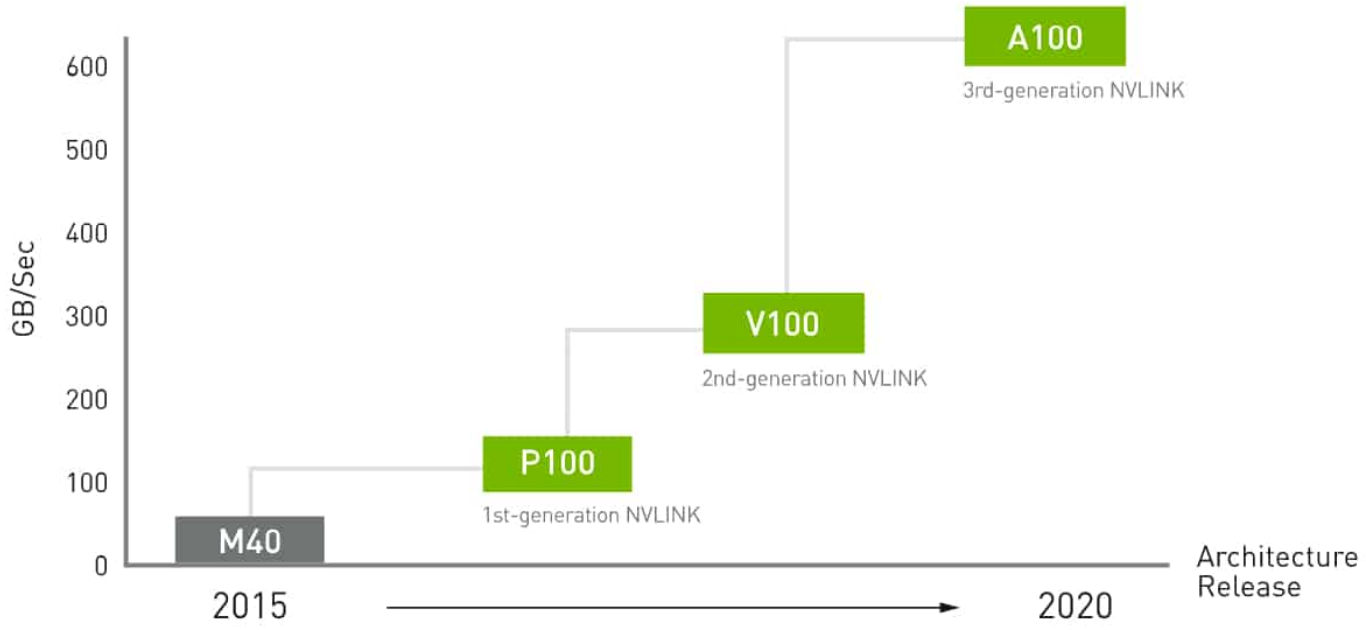




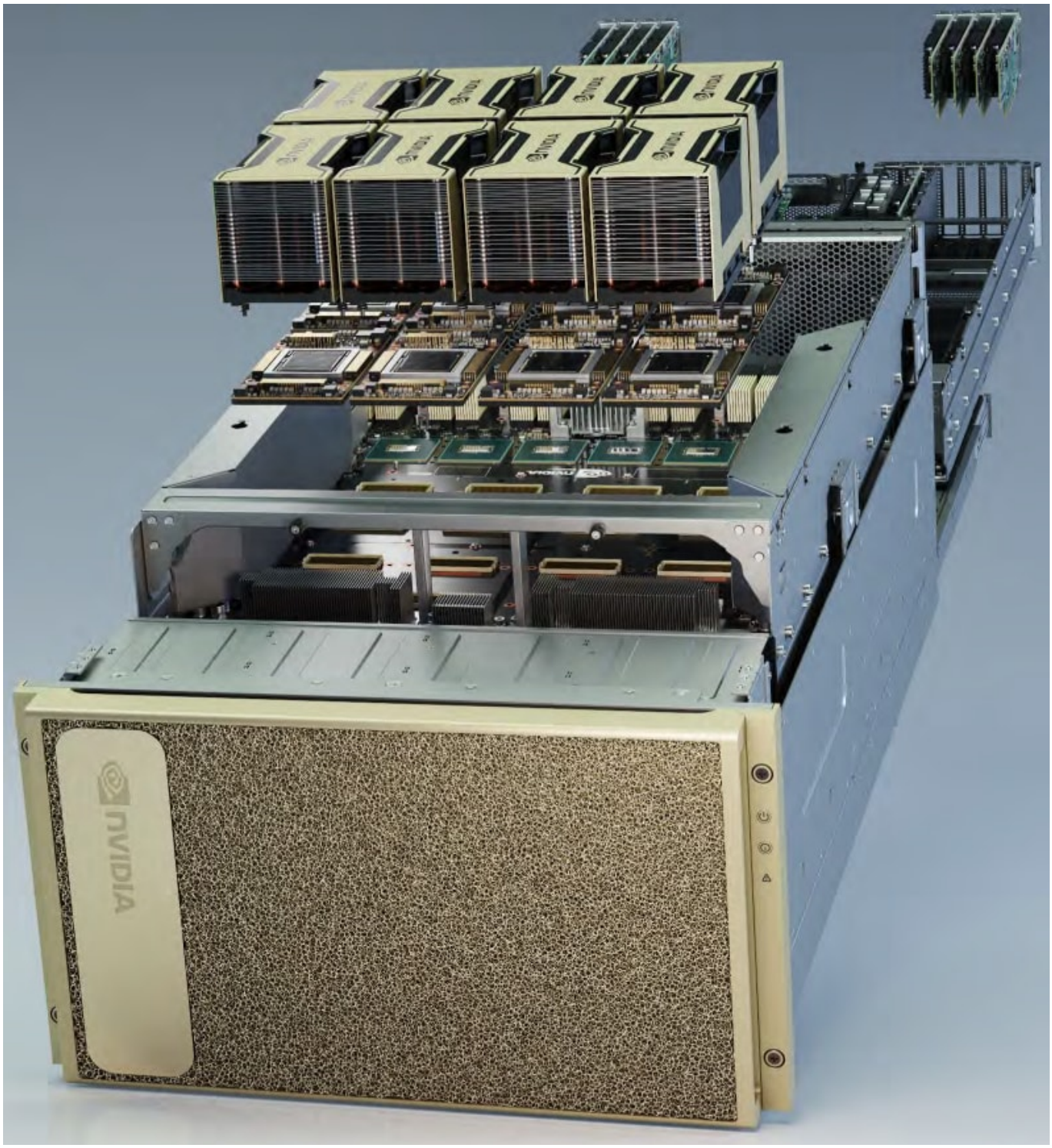
Autre avancée par rapport à Volta : la fonction de virtualisation des GPU permet désormais de dédier à chacun sa RAM et son cache L2.

MIG Instance	SMs Per Instance	Memory Per Instance	# Instances Per GPU	Target Workload
MIG 1g.5gb	14	5 GB	7	Jupyter Notebooks For Development, Model Tuning, Inference, Light HPC
MIG 2g.10gb	28	10 GB	3	Inference, Light HPC
MIG 3g.20gb	42	20 GB	2	Light Training, Inference, HPC
MIG 4g.20gb	56	20 GB	1	Light Training, Inference, HPC
MIG 7g.40gb	98	40 GB	1	Training, HPC

En parallèle, la technologie NVLink voit sa bande passante doublée, à 600 Go/s. Elle permet toujours de mettre en cluster jusqu'à 16 GPU.



L'A100 présente une enveloppe thermique de 400 W, contre 300 à 350 W pour le V100.
On le trouve au format SXM (mezzanine), notamment au sein du serveur DGX A100. Celui-ci comprend 8 accélérateurs A100, 15 To de stockage, 1 To de RAM, deux CPU AMD Rome 7742 (à 64 cœurs chacun) et des contrôleurs Mellanox. Prix annoncé : 199 000 $.



On trouve aussi l'A100 dans la 2e génération du système DGX SuperPOD, qui atteint les 700 Pflops sur 1 120 GPU.

Illustrations © Nvidia