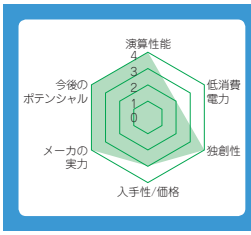
第3章 NVIDIA/グーグル/インテル/IBM/マイクロソフト/中国

大手ITメーカのAIチップ

中森 章



1：王者NVIDIAの貫禄！ AIサーバ向けGPU

Volta (NVIDIA)

● NVIDIAのAIに対する基本思想

現在のAIブームの立役者はNVIDIAであると言っても過言ではありません．NVIDIAはGPUを使ってテンソル処理が高速に実現できることを実証しました．それだけに，競合他社がテンソル処理（のみ）に特化したチップを開発する中で，グラフィックス処理も可能なGPUでAI（ディープ・ラーニング）分野の覇権を握ろうとしています．

NVIDIAの戦略は明確です．クラウド（サーバ）からエッジまでをサポートする「GPUによる全方位展開」です．具体的には，GPUのスケーラビリティで，同社のPascalアーキテクチャを用いて640個のCUDA（Compute Unified Device Architecture）コアを1単位として，そのコア数を変えることで各分野に適応させる考えです．このコアの集まりをGPC（Graphics Processor Cluster）と呼びます．このGPCをGP102，GP104，GP106にそれぞれ，6基，4基，2基搭載しています．さらにGP102をマルチコア化し，少ないコア数でエッジでのAI処理やADAS（Advanced Driving Assistant System）対応，膨大なコア数でスーパ・コンピュータ級のサーバ処理に対応させます．

● よりAI用に振った最新Voltaアーキテクチャの特徴

しかし，2017年に発表されたVoltaアーキテクチャのGPUからは様子が異なります．CUDAコア1本で押してきたGPUにテンソル・コアを搭載しました．これはFP16と呼ばれる半精度（16ビット）の浮動小数点を要素とする4×4行列（A，B，Cとする）において，

$A \times B + C$

という演算を1秒間に約120兆回行うことができる（120TOPS）コアです（GV100の場合）(2)．明らかに，ますますAIに特化するという意思の表れだと思われます．また，Voltaでは，1つのGPC内のCUDAコア数も（10基から14基と）1.4倍となっており，単精度浮動小数点（FP32）も強化されています．

図1に，Volta（GV100）のブロック図と，VoltaのSM（Streaming Multiprocessor）のブロック図を示します．GV100は80基のSM，640基のテンソル・コア，5120基のCUDAコアを備えています．力技でAI処理を行うというNVIDIAらしい構成です．

● Voltaはサーバ側用

NVIDIAのGV100は，AI処理用GPUといってもアクセラレータです．それを制御するCPUの存在なしでは自律動作できません．CPUとのインタフェースはNVLinkなので，CPUとしてはIBMのPOWER9を想定していると思われます．これは明らかにサーバ側を狙った製品です．