CUDA技術を利用したGPU（Graphics Processing Unit）コンピューティングの実際

前編

グラフィックス分野で磨かれた並列処理技術を汎用数値計算に応用

下馬場 朋禄，伊藤 智義

3Dグラフィックス処理の分野では，長年，並列処理技術が利用されてきた．こうしたグラフィックス分野で蓄積された並列処理技術の資産を汎用数値計算に応用しようというのが「GPU（Graphics Processing Unit）コンピューティング」と呼ばれる技術である．ここでは，NVIDIA社が提供しているCUDA（Compute Unified Device Architecture）環境を例に，GPUコンピューティングの概要と開発環境の構築手順を紹介する．（編集部）

3D（3次元）グラフィックスを扱うゲームや3D CAD（Computer Aided Design）などの処理を高速に行うための専用LSIとして，GPU（Graphics Processing Unit）があります[注1]．GPUの描画性能は，半導体技術の進歩とともに急速に向上しています．GPUは，グラフィックス・ボードの形でパソコンに組み込まれたり，PLAYSTATION 3やWiiといった家庭用ゲーム機に搭載されたりしています．最新のGPUのグラフィックス処理能力は，毎秒100億テクセル（テクセルとは，ポリゴンに貼り付けるテクスチャを構成する各ピクセル．3Dグラフィックスでは，ポリゴンと呼ばれる多角形にテクスチャと呼ばれる一種の壁紙を貼り付けたものを組み合わせて物体を構成する）を超えます．汎用的な計算を得意とする一般のマイクロプロセッサ（CPU）で同じ処理を実行した場合，GPUほどのグラフィックス処理能力を得ることはできません．

1. グラフィックス専用のGPUで汎用計算

この潤沢な演算能力に着目し，GPUで3Dグラフィックス処理以外の汎用的な計算を行わせる「GPGPU（General Purpose Computation on Graphics Processing Unit）」，あるいは「GPUコンピューティング」と呼ばれる手法がここ数年，注目を集めています．GPGPUは，画像処理はもちろんのこと，流体計算，電磁波シミュレーション，天文シミュレーション，たんぱく質などの挙動を解析する数値シミュレーション，バイオ・インフォマティクス，金融工学など，幅広い分野への適用が行われています．一般のマイクロプロセッサに比べて数倍～数百倍の計算速度を達成した，とする研究報告例もあります．

さて，本来，GPUは3Dグラフィックス処理専用のLSIでしたが，どのようにして汎用計算を行わせるのでしょう．本稿で解説を行うGPUは，後述する「統合シェーダ（Unified Shader）」と呼ばれる最新のアーキテクチャを採用しています．シェーダとは，GPU内に複数個搭載されている小型プロセッサ・コアのことを言います[注2]．

● 従来のGPU構成ではグラフィックスの知識が必要

ここでは便宜上，統合シェーダが登場する以前と以後のGPUを比較することで，その違いを見ていきます．

初めに，統合シェーダが登場する以前のGPUアーキテクチャの概略を**図1**に示します．このタイプのGPUは，基本的に3Dグラフィックス処理に特化したアーキテクチャを採用しています．その構造は，ポリゴンの頂点座標を計算する「頂点シェーダ」と，テクスチャのピクセル値を計算する「ピクセル・シェーダ」に分かれています．

頂点シェーダとピクセル・シェーダは，3Dグラフィックス処理の多様化に伴い，シェーダ言語と呼ばれる専用の言語（CgやHLSLなど）によってプログラミングできる構造になっています．これまでは，このシェーダ言語を利用して，汎用計算をGPU上へマッピングする例がありました．しかし，あくまでも3Dグラフィックス処理の高速化に力点を置いた構造になっていたため，汎用計算を目的として使用するには，いろいろな問題点がありました．

図1のように，入力データ（テクスチャに格納）はまず頂点シェーダで処理され，続いてピクセル・シェーダで処理されます．**図1**のGPUは頂点シェーダが8本，ピクセル・シェーダが24本あり，これらのシェーダが同時に並列処

注1：GPUはかつて，「3Dグラフィックス処理LSI」や「3Dグラフィックス・アクセラレータLSI」と呼ばれていた．

注2：コアとはLSI内部に組み込まれている大規模な回路ブロックのこと．CPUコアやDSPコア，PCIコア，メモリ・コア，アナログ・コアなどがある．これらとI/Oブロックなどを1チップ上に集積することで，LSIができ上がっている．

KEYWORD——CUDA，GPU，GPGPU，3Dグラフィックス，統合シェーダ，頂点シェーダ，ピクセル・シェーダ，SIMD，GDDR3，カーネル，グリッド

出典：PC Watch 後藤弘茂の Weekly 海外ニュース，NVIDIA からハイエンド GPU「G70」が登場（http://pc.watch.impress.co.jp/docs/2005/0622/kaigai193.htm）

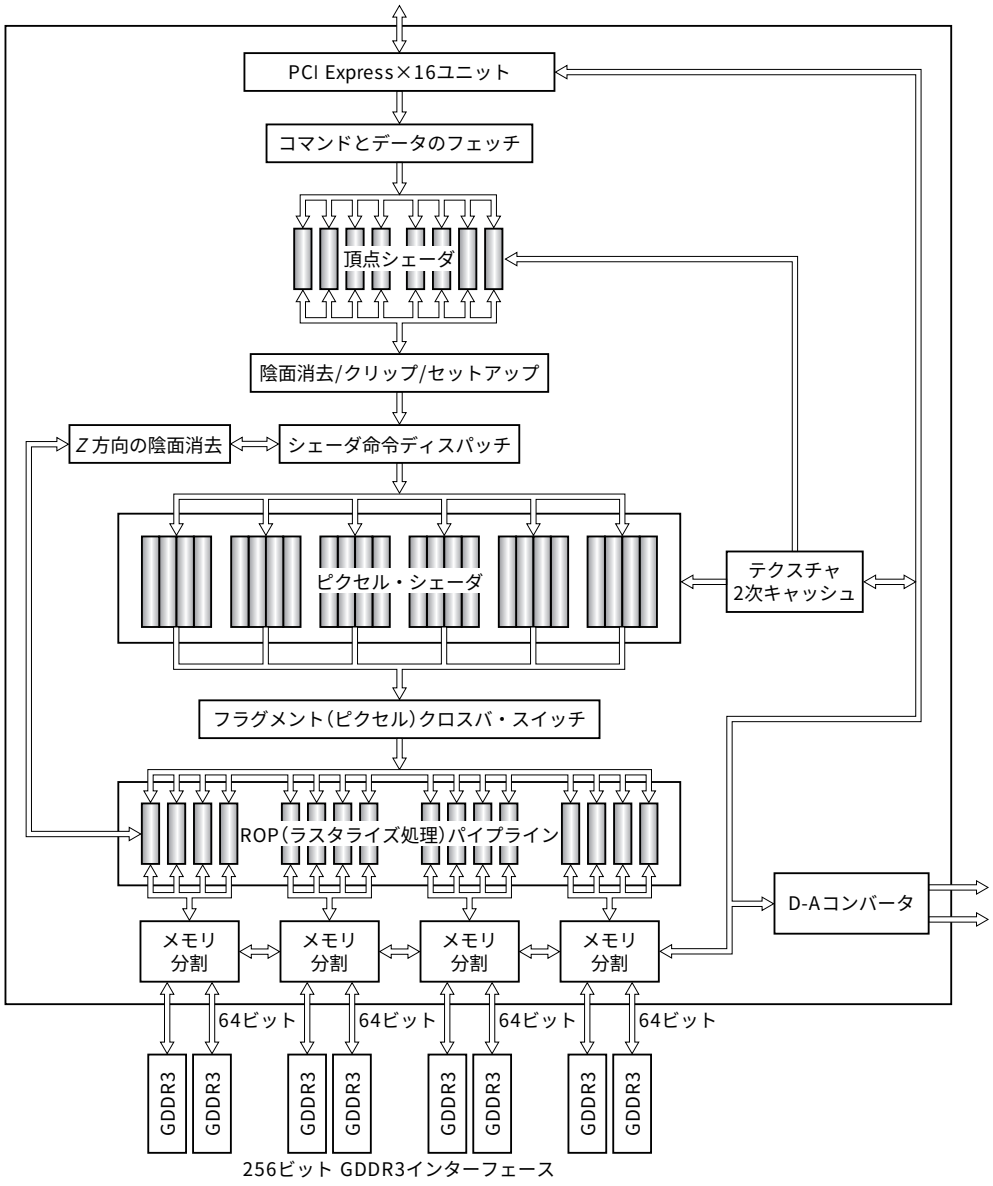

図1
統合シェーダ以前の GPU の構成

GeForce 7800 GTX（G70）のブロック図を示す．3D グラフィックス処理に特化したアーキテクチャを採用している．ポリゴンの頂点座標を計算する「頂点シェーダ」と，テクスチャのピクセル値を計算する「ピクセル・シェーダ」が処理の中心になる．

理を行うことで，計算処理を高速化しています．各シェーダでは，目的の計算を行わせるようにシェーダ言語でプログラミングを行います．このような構造で汎用計算を行わせるには，まずテクスチャに入力データをセットし，シェーダ言語でテクスチャに含まれたデータを加工していきます．計算結果はフレーム・メモリに格納されます．

このような工程を経るため，3D グラフィックスの視点や視線方向などの設定の仕方によっては，入力データを格納したテクスチャが変形する現象が起こる場合があります．つまり，自分が意図したデータとは異なるデータへと変形されてしまい，正しい計算結果が得られないわけです．これを避けるため，テクスチャが変形しないように頂点シェーダの設定を正しく行う必要があります．また，テクスチャなどの 3D グラフィックス処理や，それを操作する 3D グラフィックス・ライブラリ（OpenGL や DirectX）の知識が不可欠となります．このようなことが，GPU 上で汎用計算を行わせる場合に問題となります．

● NVIDIA 社が CUDA と呼ばれる環境を用意

そこで，このような制約に束縛されず，GPU 上で自由に汎用計算を行わせるための「CUDA（クーダ）（Compute Unified Device Architecture）」と呼ばれる新しいハードウェア・アーキテクチャ，およびソフトウェア環境が NVIDIA 社に