B-010

Ada Lovelace アーキテクチャのマルチ GPU クラスタシステムによる 4K 解像度位相型実時間電子ホログラフィ
Real-time 4K-resolution phase-type electroholography using multi-GPU cluster system based on Ada Lovelace architecture

土居 明可† Haruka Doi　三谷 永久† Towa Mitani　和田 翔夢† Shoumu Wada　髙田 直樹‡ Naoki Takada

1. はじめに

電子ホログラフィは，「究極の三次元テレビ」になると考えられている．しかし，計算機合成ホログラム(CGH:Computer-Generated Hologram)の計算は膨大であり，実用化を妨げている．CGH には振幅型と位相型がある．位相型 CGH は，計算量が振幅型 CGH の 2 倍になるが，再生像は振幅型と比べて明るくなる．さらに，位相型 CGH では共役像が再生されず，再生像を大きく表示できる．

本研究では，三次元物体を点群で表現したポイントクラウドモデルを用いる．さらに，鮮明な再生像を得るため，4K 解像度による位相型 CGH を用いる．その CGH 計算は使用するデータ量に比べ演算量が多く並列化に向いている．また，計算された CGH は画像として出力されることから，Graphics Processing Unit(GPU)に適している．GPU を用いた電子ホログラフィの研究が盛んに行われている[1-4].

近年，GPU は浮動小数点演算性能の向上に加え，深層学習の計算に適したアーキテクチャへと進化を遂げている．NVIDIA 社 Ampere アーキテクチャ以降の GPU では，CUDA コア数に対する Special Function Unit (SFU)の数が減り，三角関数の計算がボトルネックとなった．そこで，著者らは三角関数の計算を低減させた CGH 計算アルゴリズムを提案し，性能を十分引き出すことに成功した[5]．しかし，CGH の計算時間は，物体点数と CGH の解像度に比例する．4K 解像度位相型 CGH による三次元動画再生をリアルタイムで実現するには，さらなる高速化が望まれる．

2022 年に NVIDIA 社より，Ada Lovelace アーキテクチャを採用した GPU が発売され，更なる計算高速化が可能となった．しかし，消費電力と重量が増し，複数の GPU を搭載したマルチ GPU 環境を構築するのは容易ではない．本研究では，CGH 計算用に Ada Lovelace アーキテクチャの GPU(NVIDIA GeForce RTX 4090)を 12 枚搭載したマルチ GPU クラスタシステムを開発した．最終的に，Ampere アーキテクチャの GPU(NVIDIA GeForce RTX 3080)に対して理論性能比に近い約 2.8 倍の CGH 計算高速化を実現した．

2. 計算機合成ホログラム

2.1 従来の CGH 計算

物体を構成する点数をN_p，物体点を点光源とする．フレネル近似を用いて，位相型ホログラム面上の点$(x_\alpha, y_\alpha, 0)$における位相$\phi(x_\alpha, y_\alpha, 0)$は，次式となる．

$$\phi(x_\alpha, y_\alpha, 0) = \tan^{-1}\frac{\sum_{j=1}^{N} A_j \sin\left[\frac{\pi}{\lambda z_j}\{(x_\alpha - x_j)^2 + (y_\alpha - y_j)^2\}\right]}{\sum_{j=1}^{N} A_j \cos\left[\frac{\pi}{\lambda z_j}\{(x_\alpha - x_j)^2 + (y_\alpha - y_j)^2\}\right]} \quad (1)$$

ここで，三次元物体上の点jの座標を(x_j, y_j, z_j)とした．A_jはj番目の物体点の光の強度を示す．λは三次元情報の記録に使用される参照光の波長である．

2.2 三角関数の計算を低減させた計算手法[5]

Ampere アーキテクチャ以降の GPU に適応させるため，式(1)に加法定理を用いて次のようにする．

$$\phi(x_\alpha, y_\alpha, 0) = \tan^{-1}\frac{\sum_{j=1}^{N} A_j (\sin X \cos Y + \cos X \sin Y)}{\sum_{j=1}^{N} A_j (\cos X \cos Y - \sin X \sin Y)} \quad (2)$$

ここで，$X = \frac{\pi}{\lambda z_j}(x_\alpha - x_j)^2$，$Y = \frac{\pi}{\lambda z_j}(y_\alpha - y_j)^2$.

これにより，あらかじめx, y方向それぞれで三角関数の値を保持しておく．加法定理により，各画素の値を四則演算のみで計算することができる．この時，加法定理による演算は四則演算のみであり，三角関数の演算は行われない．その結果，式(2)で三角関数の計算を低減できる．

3. マルチ GPU クラスタシステム

図 1 にマルチ GPU クラスタシステムを示す．CGH 表示ノードは 1 枚の GPU を搭載した 1 台の PC で構成する．6 枚の GPU を搭載した PC を 2 台用いて CGH 計算ノードを構成する．図 1 のシステムに図 2 のパイプラインアルゴリズムを実装する．CGH 計算ノード上の各 GPU に三次元動画の各フレームが割り当てられ CGH 計算を行う．計算された CGH はネットワークを介して CGH 表示ノードへ転送される．CGH 表示ノードでは，計算ノードから転送された

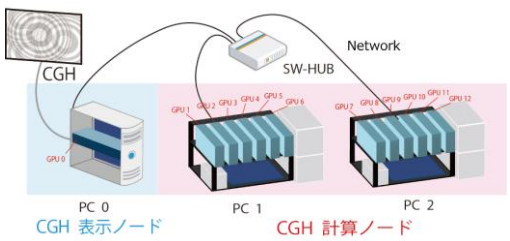

図 1　マルチ GPU クラスタシステム

† 高知大学大学院総合人間自然科学研究科
‡ 高知大学教育研究部自然科学系理工学部部門

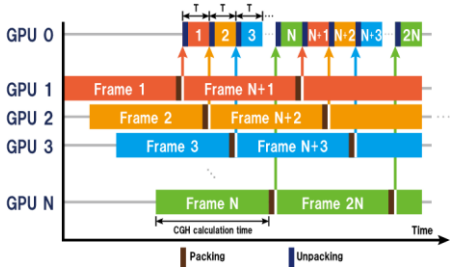

図 2　パイプライン方式

CGH 画像を順番に空間光変調器(SLM)に表示する．本研究で使用する Ada Lovelace アーキテクチャの GPU は，消費電力が増加している．それに伴い発熱量も増加しているため，冷却装置も大きくなった．本研究では，消費電力，筐体，重さも格段に大きくなった GPU に適するように，CGH 計算ノードの設計を見直した．システムで使用する GPU は，コストパフォーマンスを考慮し，RTX 4090 を使用した．まず，2 枚の GPU を搭載した PC を構築し，CGH 計算負荷をかけて GPU をフル稼働させ，消費電力と GPU の発熱量をモニタリングした．その結果，計算ノード 1 台あたり 1200W の電源 3 台で GPU へ電力を供給することで GPU を 6 枚まで搭載できることが確認された．従来のフレームでは GPU の筐体と重さに対応できないため，フレームを自作した．汎用的な PC では複数の GPU を搭載することは難しい．CGH 計算は計算量が多いがデータ量が少ないため，高い転送速度を要求しない．そこで，PCI Express (PCIe)の 1 レーンのみを利用することで，多くの GPU を単一のマザーボードに搭載できる．

一方，4K 解像度(3840×2048)の位相型 CGH のデータ量は約 63Mbit である．転送速度が 1Gbps のネットワークの場合，1 枚の位相型 CGH の転送時間は 63ms となる．三次元動画のリアルタイム再生を実現するには，1 秒間に 30 枚の位相型 CGH を表示する必要がある．そこで，10 Gigabit Ethernet(GbE)を使用し，ノード間通信によって生じるボトルネックを解消する．そのため，7 本の PCIe スロットと 10GbE を搭載したマザーボード(ASUS WS X299 SAGE/10G)を用いた．

4. 結果

Ampere アーキテクチャと同様に，Ada Lovelace アーキテクチャは，浮動小数点演算を行う CUDA コアに対して三角関数を高速に計算する SFU の数が少ない．そのため，三角関数の計算を低減させた計算手法が有効となる．本システムに，本手法を実装する．図 3 に構築したシステムを示す．CGH 表示ノードの GPU に NVIDIA GeForce GTX 1080Ti を 1 枚，CGH 計算ノードに RTX 4090 を 12 枚使用した．

3840×2048 の 4K 解像度を持つ，位相型 CGH の表示時間間隔の結果を図 4 に示す．図 4 の赤色の実線より，約 25 万点までの物体点数において，30fps で実時間再生が可能であることが分かった．また，各 CGH 転送データは約 63Mbit であるため，CGH 表示ノードへの転送時間は約 7ms である．よって，図 3 より，物体点数が 51200 点以下の場合，各 CGH 計算ノードから CGH 表示ノードへの CGH 転送時間がボトルネックとなり，表示時間間隔が一定になっている．

図 4 の青色の実線は，表示ノード，計算ノードともに Ampere アーキテクチャの RTX 3080 を 13 枚使用したシステムの計測結果である．Ada Lovelace アーキテクチャは Ampere アーキテクチャのシステムより約 2.8 倍高速化されている．ここで，RTX 4090 と RTX 3080 の理論性能は，それぞれ 85.03TFLOPS と 30.29TFLOPS である．理論性能において，RTX 4090 は約 2.8 倍高速化されている．よって，本システムは理論性能の比に等しい高速化を実現している．

図 3　12 枚の GPU を搭載した CGH 計算ノード

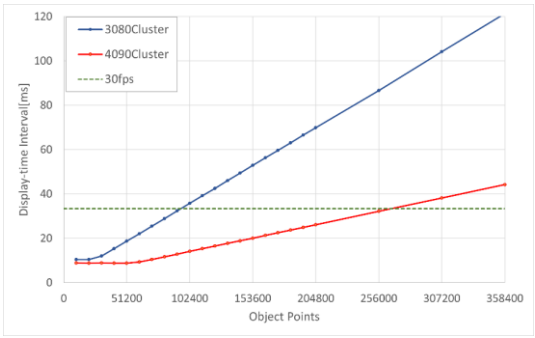

図 4　CGH 表示時間間隔

5. まとめ

Ada Lovelace アーキテクチャの GPU を搭載したマルチ GPU クラスタシステムを開発した．

GPU の消費電力と重量が増したため，CGH 計算ノードにおいてフレームを自作し，汎用的な部品を組み合わせることで省スペース化と低コスト化を図った．最終的に，4K 解像度位相型 CGH における，約 25 万点からなる三次元物体を 30fps で実時間再生することに成功した．

謝辞

本論文の一部は，日本学術復興会の科研費・基盤研究(C)(課題番号 21K11996, 24K15048)および競輪の補助によって行われた．

参考文献

[1] T.Sugie et al,“High-performance parallel computing for next-generation holographic imaging”, Nature Electron. 1, 254–259 (2018).
[2] N.Takada et al,“Fast high-resolution computer-generated hologram computation using multiple graphics processing unit cluster system”, Applied Optics 51, pp. 7303-7307 (2012).
[3] H.Sannomiya et al,“Real-time spatiotemporal division multiplexing electroholography for 1,200,000 object points using multiple-graphics processing unit cluster”, Chinese Optics Letters, Vol.18, Issue 7, pp.070901 (2020).
[4] T. Takada. “Real-Time Electroholography Based on Multi-GPU Cluster,” Hardware Acceleration of Computational Holography, ed. by T. Shimobabaand T. Ito. Springer Nature Singapore, pp. 227-240 (2023).
[5] T. Mitani et al,“Fast Calculation of Amplitude-modulated Computer Generated Hologram with Multiple Ampere-GPU Cluster System”, Proc. of the International Display Workshop vol. 26, pp 498-499 (2021).