

広帯域メモリを用いた CGLA の LLM 向けスケーラビリティ評価

宗片 吉史[†] 安藤 拓翔[†] 中島 康彦[†]

[†] 奈良先端科学技術大学院大学 〒630-0192 奈良県生駒市高山町 8916-5

E-mail: †munakata.yoshifumi.nz5@naist.ac.jp

あらまし 本稿では, CGLA アクセラレータ IMAX 上での AI アプリケーション, 特に LLM 推論の実行を対象として, 3D 積層 DRAM に代表される広帯域メモリを備えた将来 IMAX における性能ボトルネックを予測するためのシミュレーションモデルを提示する. 評価の結果, 現行 IMAX においても, メモリ帯域の向上により 1.4 倍から 1.9 倍の性能向上が見込まれることを確認した. さらに, 本モデルに基づき, IMAX のスケーラビリティに応じたメモリ構成を検討するための見積り方法を示す. 加えて, 将来的に広帯域メモリを導入した際に顕在化すると考えられるホスト側性能のボトルネックへの対応策を検討し, IMAX の並列化およびカスケード接続を組み合わせた二次元的な拡張によるスケーラビリティ向上の方針を提示する.

キーワード CGLA, AI アクセラレータ, 3D 積層 DRAM, HBM

Scalability Evaluation of CGLA with High-Bandwidth Memory for LLM Inference

Yoshifumi MUNAKATA[†], Takuto ANDO[†], and Yasuhiko NAKASHIMA[†]

[†] Nara Institute of Science and Technology 8916-5 Takayama, Ikoma, Nara, 630-0192 Japan

E-mail: †munakata.yoshifumi.nz5@naist.ac.jp

Abstract This paper presents a simulation model for predicting performance bottlenecks in future IMAX architectures equipped with high-bandwidth memory, such as 3D-stacked DRAM, targeting the execution of AI applications—particularly LLM inference—on the CGLA accelerator IMAX. Based on this model, we further present an estimation method for studying memory configurations according to the scalability requirements of IMAX. In addition, we discuss approaches to addressing potential host-side performance bottlenecks that may emerge with the introduction of high-bandwidth memory, and present a scalability strategy based on two-dimensional expansion through the combination of IMAX parallelization and cascade interconnection.

Key words CGLA, AI Accelerator, 3D Stacked DRAM, HBM

1. はじめに

AI アプリケーションで用いられるモデルの大規模化に伴い, アクセラレータの性能は, 演算性能やメモリ容量だけでなく, メモリ帯域およびその利用効率によっても強く制約されるようになっていく.

我々が開発を進める CGLA (CPU-Grounded Linear Array) アーキテクチャである IMAX (In-Memory Accelerator eXtension) は, Manycore や GPGPU といったアーキテクチャと比較して, データの再利用性が高い. これは, PE と LMM の間でデータを直接受け渡せるためであり, 主記憶とアクセラレータ間のデータ転送に起因する制約は相対的に小さい. しかし, LLM 推論, 音声認識, 画像生成のように, 大容量の重みデータを頻繁に読み出

す AI アプリケーションでは, 依然として外部メモリ帯域が性能上のボトルネックとなる.

今後, モデルがさらに巨大化し, 電力やメモリ容量が限られるエッジ環境, あるいはデータセンターレベルでの CGLA 活用を検討するうえでは, スケーラブルな AI 実行基盤としての CGLA アーキテクチャにおいて, どこがボトルネックとなるのかを特定し, 現実的な改善手法を評価することが重要である. 特にメモリ階層の改善に着目すると, プロセス技術が微細化限界に近づく中で, HBM に代表される TSV を用いた 3D 積層 DRAM や, 将来的にはモノリシック 3D DRAM のような広帯域メモリの採用が有力な選択肢となる.

本稿では現行の CGLA 型アーキテクチャである IMAX3 上での LLM 推論処理を詳細に観測し, 3D 積層 DRAM のような広

帯域メモリを採用した将来構成において、どの程度の性能向上が見込めるかを評価する方法を提示する。さらに、データの再利用性が高いという IMAX の特徴を最大限に活かすための設計方針について議論する。本稿の主な貢献は次の 2 点である。

- IMAX 上での LLM 推論実行を対象に、フェーズ分解に基づくトレース駆動の性能見積り枠組みを示し、広帯域メモリ導入時の性能改善を定量化した。

- 広帯域化後に顕在化するホスト側制御およびデータ供給のボトルネックを整理し、ホスト増強, DMA 構成最適化, 広帯域メモリ導入, カスケード接続を含む段階的なスケールアップ方針を示した。

先行研究では, IMAX 上での LLM 実行における実測分析やボトルネックの観測が中心であり, 現行構成における挙動の理解に主眼が置かれていた。これに対して本稿では, それらの観測結果を出発点として, 3D 積層 DRAM のような広帯域メモリを導入した将来構成における性能見積りへと議論を拡張し, メモリ帯域の向上がどこまで有効であり, どの段階でホスト制御や DMA 構成が新たな律速要因となるかを設計指針として整理する。

2. IMAX における LLM 推論のボトルネック分析

2.1 スケーラブルな AI 実行基盤としての CGLA

AI をきっかけとするエネルギー需要の急増に伴いスケラブルな省電力計算基盤の確立が急務となっている。従来型半導体のまま電力効率を追求する 1 つの方法がデータフローアーキテクチャであり, その中でも汎用性とコンパイル負荷の軽減を両立する選択肢が CGLA である。特に CGLA の汎用性の高さ, データ再利用性の高さは AI アプリケーションとの相性が見込めることから, AI OSS アプリケーションを CGLA 上に実装する先行研究が進められている。[1]~[4]

2.2 LLM 推論実行のボトルネック

先行研究で指摘されている通り, IMAX 上での AI アプリケーションの実行, とりわけ LLM 推論のアクセラレーションにおいては, デコードフェーズにおける LOAD 操作が実行時間の大部分を占める。

今後, 推論モデルの規模がさらに拡大するにつれて, 外部メモリ帯域およびデータ転送方式が性能上の律速要因となる可能性が高い。この課題に対する打開策として, HBM や 3D 積層 DRAM のような広帯域メモリの採用が考えられる。しかし, その効果を定量的に見積もるためには, 単純な帯域モデルだけでなく, 実行時のフェーズ構造を考慮した分析が必要である。

図 1 には, Qwen3 0.6B Q3_K.S を IMAX3 上で実行した際に, 入力トークン数を 8 で固定し, 各出力トークン数条件について 10 回測定した平均総実行時間の比較を示す。出力トークン数が増加するほど, すなわち処理が重くなるほど, IMAX を用いない ARM 単体実行に対する IMAX の優位性が低下する傾向が見られる。

出力するトークン数が増えるほど DMA におけるオーバーヘッドの蓄積が大きくなっている可能性が示唆される。

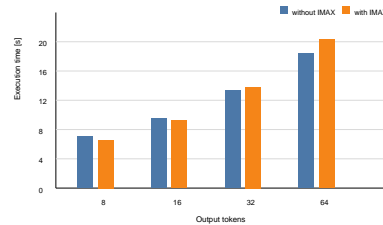


図 1 出力トークン数に対する実行時間の比較

2.3 フェーズ分解と帯域依存性

IMAX 上での処理は, ホスト制御, DMA によるデータ転送, ローカルメモリへの配置, 演算実行, および結果回収など複数のフェーズから構成される。便宜上ここでは表 1 に記載の 7 つのフェーズに分解する。

表 1 性能見積りで用いるフェーズと具体的な処理内容の対応

フェーズ	行う処理
LOAD	入力を主記憶 (DRAM) から IMAX のローカルメモリへとロードする処理
DRAIN	IMAX のローカルメモリ上の計算結果を主記憶 (DRAM) へ書き戻す処理
EXEC	IMAX 上のバースト演算
CONF	PIO インタフェースを介して IMAX の命令を転送する処理
REGV	IMAX の内部レジスタの初期化
RANGE	アクセスするデータのアドレス範囲を決定する処理
ARM	ARM プロセッサ上で行われるオフロード外の処理

このうち, LLM 推論においては各トークン生成のために大容量の重みを主記憶から繰り返し読み出す必要があることから, 外部メモリ帯域の影響を最も強く受けるのは LOAD である。

DRAIN および REFILL もデータ転送を伴うため帯域に依存する処理ではあるものの, 計算結果や再配置対象のデータ量は重み読出しに比べて小さく, 呼び出し頻度も限定的である。実測でも LOAD が総実行時間の 54.91% を占める一方で, DRAIN は 1.76%, REFILL は 1.07% にとどまっており, end-to-end (E2E) 評価における時間寄与は限定的である。

本稿での LLM 推論を念頭に置いた外部メモリ帯域の向上に伴うスケラビリティ評価においては, LOAD フェーズの処理高速化によって E2E 実行時間がどの程度短縮されるかを中心に分析することが合理的であるため, LOAD の高速化がどの程度 E2E 実行時間の短縮に寄与するかを定量的に評価することを目的とする。このため本稿では, 広帯域化の一次近似として LOAD のみを帯域依存項として扱う。

ここでの REFILL は, LMM 上に保持しきれないデータを再度ロードする処理であり, 実質的には LOAD の再実行に相当する。

2.4 広帯域化による性能見積りモデル

広帯域化の効果を見積もるため, 本稿では観測された実行時間のうち LOAD が全体の 54.91% を占めるとみなし, この成分のみが帯域拡大に伴って短縮されると仮定する。すなわち, 全実

表2 Q3.K系 LLM 実行のフェーズ別時間内訳

フェーズ	実行時間 [ms]	割合 [%]
LOAD	80389.1	54.91
DRAIN	2576.8	1.76
EXEC	28193.0	19.26
CONF	18.7	0.01
REGV	18955.3	12.95
RANGE	14691.6	10.04
REFILL(再 LOAD)	1566.0	1.07
TOTAL	146390.4	100.00

行時間に占める並列化可能部分を $p = 0.5491$, 帯域非依存の逐次部分を $1 - p = 0.4509$ と置き, 広帯域化後の性能を評価する.

基準帯域を B_0 , 評価対象の帯域を B とし, 倍率を

$$k = \frac{B}{B_0} \quad (1)$$

と定義すると, 正規化された実行時間は

$$\frac{T(B)}{T(B_0)} = (1 - p) + \frac{p}{k} \quad (2)$$

与えられる. したがって, 広帯域化による高速化率 $S(k)$ は

$$S(k) = \frac{T(B_0)}{T(B)} = \frac{1}{(1 - p) + \frac{p}{k}} = \frac{1}{0.4509 + \frac{0.5491}{k}} \quad (3)$$

となる.

例えば, 帯域を 2 倍, 4 倍, 8 倍に拡大した場合の高速化率は, それぞれ $S(2) = 1.38$, $S(4) = 1.70$, $S(8) = 1.92$ となる. また, 帯域を無限大まで拡張できた場合でも, 高速化上限は

$$S_{\max} = \lim_{k \rightarrow \infty} S(k) = \frac{1}{1 - p} = \frac{1}{0.4509} \approx 2.22 \quad (4)$$

にとどまる.

このモデルは, LOAD を帯域依存項, それ以外を固定項として分離し, 広帯域化だけで達成できる改善幅を明示するための一次近似である. したがって, もし 2.22 倍を超える高速化を狙うのであれば, その他のフェーズの高速化が不可欠である.

2.5 広帯域メモリ導入時の帯域見積もり

現行の IMAX3 は主記憶として OS 空間 8GB, DMA Buffer 空間 4GB, バス幅 192bit, 3,900 Mbps の LPDDR4 メモリを搭載しており, IMAX 本体と AXI4 バスによって接続されている. 以下にこのメモリが LPDDR5, LPDDR6, HBM3 といった広帯域メモリに置き換わった場合, 帯域がどの程度拡大するかを示す. LPDDR4-6 までは 256bit, HBM3 が 1024bit のバス幅を持つと仮定する. LOAD の速度が帯域に一時比例で短縮されると仮定すると, 2 倍の高速化であれば LPDDR6 の下限, 4 倍の高速化であれば LPDDR6 の上限, 8 倍の高速化であれば HBM3 の下限が必要になることが予想される. 一方で, HBM3 のような TSV を用いた 3D 積層 DRAM による帯域向上の効果を最大限に引き出すためには, 単に外部メモリの帯域を高めるだけでなく, その帯域を IMAX レーン群へ効率的に分配するための DMA 構成やソフトウェアスタックの最適化も必要になること

表3 IMAX AXI バス帯域幅の拡大

メモリ	転送速度	拡大幅
LPDDR4	3,900 Mbps	-
LPDDR5	4,800-6,400 Mbps	1.23x-1.41x
LPDDR6	11,000-14,000 Mbps	2.82x-3.59x
HBM3	7,000-12,000 Mbps	9.57x-13.68x

が予想される.

3. 広帯域化後のスケールアップ方針

3.1 ホスト側のスケールアップ

現行のエッジ向け IMAX3 ではホストプロセッサとして 2 コアの ARM Cortex-A72 プロセッサを採用しているが, レーン数の増加に対してホスト側の処理能力が十分に追従できていない. 図 2 に IMAX のレーン数を増やし並列化を進めた際に, 3 レーン以上で ARM プロセッサの処理時間の増加が観測され, ホスト側制御が E2E 実行時間のボトルネックとなっている様子を示している. LLM 推論の実行においては, 単純な演算発行だけでなく, DMA 設定, 転送先バッファ管理, 命令配送, アドレス範囲設定など, 細粒度の制御が高頻度に発生するため, ホスト側の処理能力が追従できていない状態で並列化を進めるとメモリ帯域の拡大による高速化が頭打ちになる.

今後は ARM プロセッサを 8 コアまで増やすことで, ARM 処理の増加を軽減しながら 8 レーンまでの効果的な IMAX 並列化を目指していく.

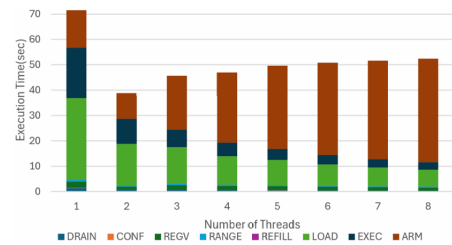


図2 スレッド数の増加に伴う ARM 処理の増加

3.2 2次元的なスケールアップ

前節までは現行のエッジ向け IMAX3 を前提にスケーラビリティの評価を行っていたが, さらなるスケールアップのためには帯域の拡大幅が 3D 積層 DRAM による大幅な帯域向上を前提としたアーキテクチャの再構成の検討も行う必要がある.

IMAX の特徴の 1 つに, メモリ機能を内部に保有していることからカスケード接続が容易な点があげられる. IMAX2 においては 64 ユニットの IMAX を 4 段カスケード接続する構成を採用した. 一方で IMAX3 においては 8 レーン並列実行を念頭に置いた構成がされており, 広帯域化に伴い LOAD フェーズの高速化が頭打ちになった場合は, 各 8 レーンにつき 4 段の IMAX をカスケード接続することでユニット数単位では IMAX3 の 32 倍までのスケールアップを見据えている. 今後さらにデータセンター向けのノード, ラックサーバーの構成を考えるうえでも並列化とカスケード接続による 2 次元方向で

のスケーリングの可能性を検討していく。

3.3 ソフトウェアスタックによる改善余地

単純なホストプロセッサの増強のみでなく、制御処理の削減や遅延の隠蔽、LMM 容量に応じたチャンクサイズの設定といったソフトウェアスタックの最適化による改善余地も依然として大きい。

以上を踏まえると、将来の IMAX 設計における優先順位は、まずホスト側の処理能力増強によってレーン並列化時の制御ボトルネックを緩和し、次に DMA 構成およびソフトウェアスタックを最適化して拡大した帯域を各レーンへ効率的に供給できるようにし、そのうえで LPDDR6 や HBM3 のような広帯域メモリを導入して LOAD フェーズの短縮効果を最大化する、という順で整理できる。さらにその先のスケールアップとして、並列化とカスケード接続を組み合わせた二次元方向の拡張を検討することが有効である。

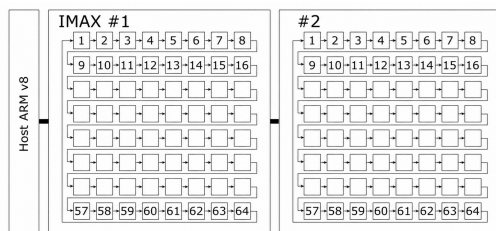


図3 2チップ構成 IMAX のカスケード接続

4. 今後の課題

本稿で示した見積り枠組みを将来的な IMAX の実設計へ接続していくために主として以下の三点を検討する必要がある。

(1) 3D 積層 DRAM に代表される広帯域メモリを導入した際のインターコネクト技術の検討である。現行 IMAX で採用されているインターコネクトでは、将来想定されるメモリ帯域を十分に活用できず、転送系が新たな律速要因となる可能性がある。そのため、高速リンクの方式やトポロジーを含めた再設計が必要となる。

(2) DMA バッファと IMAX 間のデータ転送量、転送時間、およびバッファ使用量を、より詳細に計測・分析するための基盤整備である。本稿の見積りは実機トレースに基づく一次近似であるが、モデル精度をさらに高めるためには、転送挙動をより細粒度に把握できる観測手段が必要である。

(3) 向上した演算性能とメモリ帯域を最大限に活用するためのスケジューリング方策の検討である。特に、将来構成では、DMA 転送、命令配送、lane 利用の重なり方が全体性能を大きく左右するため、ハードウェア構成と整合したランタイム制御の設計が重要となる。

5. おわりに

本稿では、現行 IMAX 上での LLM 推論実行をフェーズ単位で精査し、その結果に基づいて 3D 積層 DRAM に代表される広帯域メモリを備えた将来的 IMAX 構成における性能向上の見積り方法を示した。特に、実行時間を帯域依存フェーズと

帯域非依存フェーズに分離することで、広帯域メモリの導入がどこまで有効であり、どの段階で固定オーバーヘッドが支配的になるかを、同一の枠組みで議論できることを示した。

得られた結果は、広帯域メモリの導入が LOAD 支配の緩和に有効である一方で、それだけではエンドツーエンド性能は比例的には向上せず、DMA 構成、ホスト側制御、ランタイム設計を含めたシステム全体の最適化が不可欠であることを示している。したがって、将来 CGLA の設計においては、メモリ帯域の拡張を単独で評価するのではなく、データ供給機構と制御系を含めた統合的な設計空間として捉える必要がある。

今後は、実測に基づくモデル精度の向上に加え、広帯域メモリに適したインターコネクト、転送観測基盤、およびランタイムスケジューリングの設計を含めて検討を進める予定である。

謝 辞

本研究の一部は JST 戦略的創造研究推進事業先端的カーボンニュートラル技術開発 (ALCA-Next) JPMJAN23F4 の支援を受けたものである。また、本研究の一部は科学研究費補助金 (基盤研究 (A) 課題番号 22H00515) による。加えて、本研究は東京大学 VDEC 活動を通して、日本シノプシス合同会社の協力で行われたものである。

文 献

- [1] T. Akabe, V. T. D. Le, and Y. Nakashima, "IMAX: A Power-Efficient Multilevel Pipelined CGLA and Applications," *IEEE Access*, vol. 13, pp. 31899–31911, 2025.
- [2] T. Ando, Y. Eto, and Y. Nakashima, "A Detailed Analysis of LLM Execution on IMAX3 and Initial Evaluation of IMAX4 Prototype for Server Environment," *Proc. SASIMI*, 2025.
- [3] T. Ando, Y. Eto, A. Takeuchi, and Y. Nakashima, "Efficient Kernel Mapping and Comprehensive System Evaluation of LLM Acceleration on a CGLA," *IEEE Access*, vol. 13, pp. 199631–199646, 2025.
- [4] T. Ando, A. Takeuchi, Y. Eto, Y. Munakata, and Y. Nakashima, "Q-Snap: Quantization-Aware Dynamic Chunking for LLM Execution on a CGLA," *Proc. ICISN*, 2026.