Ponte Vecchio と Aurora は どうしてこうなったか？の話

牧野淳一郎
神戸大学/PFN

SC23 勉強会 2023/12/9

SC23 の最大のニュースは

間違いなく

ついに Aurora が Top500 にでてきたことと

その性能が衝撃的に悪かったことである

というわけで、どうしてそうなったかを推測し、我々はどうすればいいのかを考えたい

話の構成

- Ponte Vecchio の公称スペック
- Aurora の現状スコア
- 何が違うか？
- どうしてそうなったか？
- 同じ轍を踏まない方法は？

Ponte Vecchio の公称スペック

FP64 PEAK	52 TFLOPS
FP32 PEAK	52 TFLOPS
BF16 PEAK	832 TFLOPS
TDP	600W
メモリ	HBM2e x 8, 3.2TB/s

- 2023 年のマシンとしては HBM2e なのがちょっと。
- それでも、HPL で例えば実行効率 80%、色々あわせた平均電力が 1000W になったとしても 40GF/W、750W なら 53GF/W
- Green500 1 位にはならないかもしれないがそれなり。 450W 44TF という数字もある。
- Aurora は Ponte Vecchio 63744 枚。44TF 実行効率 70% で 1.96EF。1 枚 600W として 40MW。

Aurora の現状スコア

Linpack Performance	585.34 PFlop/s
Theoretical Peak	1,059.33 PFlop/s
Powe	24.687 MW

何が違うか？

- カード数
- カードあたり性能
- 実行効率
- 消費電力

カード数

- 「4,742,808 コア」と書いてある。1 ノードに SPR 2 個 (52 コア x2)、PVC 6 個 (128 コア x6) とすると 5439 ノード。
- SPR 無視して 32634 GPU
- 1PVC 44TF とすると 1.436 EF。そもそも Rpeak があってない

カードあたり性能

- Rpeak 1059.33PF から逆算するとPVC 1つ あたり 32.5 TF
- TDP 450W の 44TF の 75%になっている。
- 52TF の時にコアクロックが多分 1.6GHz なので、多分 900MHz でピーク性能が計算されている。44TF は 1.2GHz

実行効率

- これは書いてある通り 585.34/1059.33= 55.2%
- 高い効率とはいいがたいが、新しいアーキテクチャで初めてリストに載った時にこれくらいはまあ、、、
- (GRAPE-DR とかもっと全然低かったのであんまり効率のことを悪くいいたくない)

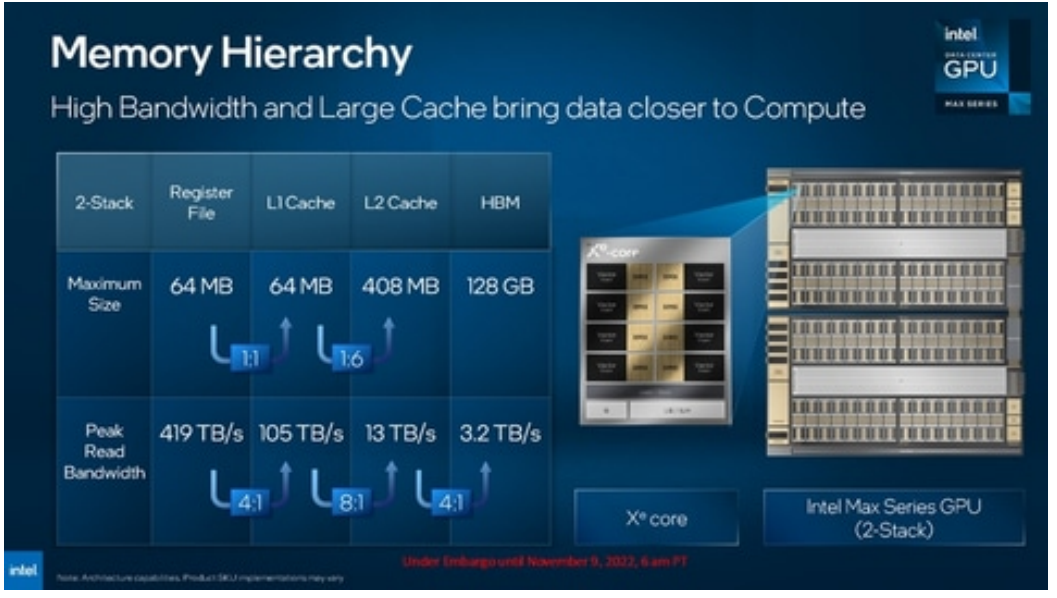
消費電力

- 全体で 24.7MW。1 ノード 4.54KW で、 SPR2 個 が 1KW もっていったとして PVC 1 つ 600W。
- SPR2 個で 1.5KW くらいとしても 500W。
- カタログスペックでは 1.6GHz である程度の効率で動いて TDP 600W と書いてあったが、現実は 60% の 900MHz のクロック、効率 55%、すなわち名目ピークの 1/3 の性能で消費電力が 600W くらいまでいっているということ。
- 電力性能がカタログスペックの 1/2 くらい、言い換えると、もしも 1.6GHz で動かしたらカタログスペックの 2 倍以上の電気を食うということ (900MHz なら多分電圧も落としてるので 3 倍くらいかも)

これが最大の問題: 電力消費が目標の 2-3 倍。

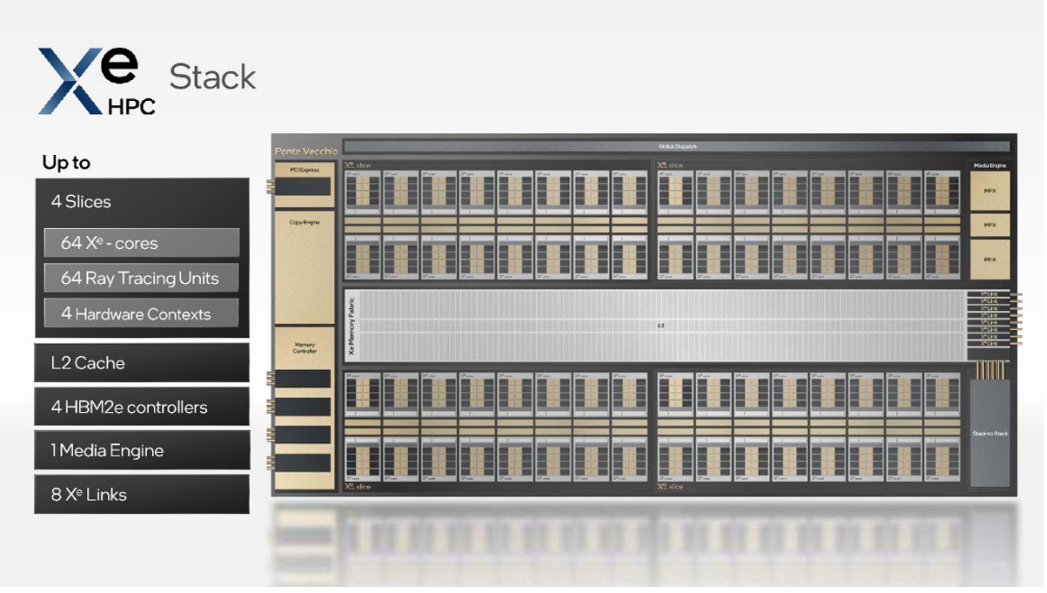
どうしてそうなったか？

牧野がだいぶ前から使い回してたスライド



B/F の値
レジスタ: 8? (普通 16 いる。行列乗算器だと 8 はありえる)
L1: 2
L2: 0.25
メモリ 0.06
A64fx に比べても数字が小さい
それでも電力消費が大きい原因になっている

PVC のレイアウト(?)



- これが「スライス」。4「スライス」で1「チップ」
- 中央の灰色が L2 キャッシュ。これ実は3次元実装で別ダイ

PVC のレイアウト(?)

- 物理的にこういう構造かどうかは？
- とはいえ L2 は少なくともスライス内で物理共有

L2キャッシュと消費電力

- チップ上でデータを横に移動させると電気を食う。概ね 1cm で 1pJ/bit (1V で)
- これは配線の太さにあまりよらない。基本的にキャパシタンスからくるから。
- 傾向としては細くなると遅延小さくするために必要なバッファやパイプラインレジスタが増えるので、段々電力増える。
- 13TB で配線 1cm あるとそれだけで大体 100W。バッファやパイプラインレジスタも同じくらい食うと L2 キャッシュだけで 200W。HBM2e メモリも 8 個もあれば 160W くらいにはなる。
- つまり、TDP の半分くらいを L2 と DRAM だけで食いつぶしている。

同じ轍を踏まない方法は？(1)-他の GPU

- H100 は？
  - 単純に、L2 のバンド幅を抑えている
  - H100 L2 バンド幅は white paper に書いてない。5.5 TB/s くらいの模様
- AMD MI250X も多分 6.4TB/s
- shared L2 だとこの辺にしとかないと厳しい
- とはいえこれだと B/F=0.1 とか。意味があるキャッシュとして使えるバンド幅ではない。

同じ轍を踏まない方法は？(1)-原理的には？

- 問題は「バンド幅の高い」「物理的に共有された=長い配線のある L2」
- 従って、原理的にはバンド幅を下げるか配線短くするか
- H100 とか: バンド幅下げる (実際 A100 からあんまり増えてない)
- 他の方向=物理共有をやめる
- これは GPU アーキテクチャのままでは無理
- 大規模なローカルメモリをもつアーキテクチャの例: PEZY-SC, Sunway 26010, GRAPE-DR, MN-Core

以下ちょっとだけ歴史的発展の話を。

スカラー計算機→ベクトル計算機

- スカラー計算機の進歩: 増えるトランジスタを「1 つの演算器を速くする」に
- 主記憶は磁気コアで、半導体よりはるかに遅いのが想定
- CDC 7600 (S. Cray の設計) あたりが最後
- S. Cray はこの後、4 プロセッサ並列の CDC 8600 を開発していたが、完成前に CDC やめて Cray Research を設立。ベクトルの Cray-1 を開発する。

スカラー計算機では

- IC の開発で使えるゲート数が 1 演算器を超えて大きくなった
- 半導体メモリの開発でメモリが非常に高速になった

ことを生かせなかった。
ベクトル: 半導体メモリは有効利用できた。

ベクトル計算機→マイクロプロセッサ超並列

- ベクトル計算機の進歩: 1 プロセッサのパイプライン数や主記憶を共有するプロセッサの数を増やす
- メモリもプロセッサも多数の IC から構成されて、沢山内部配線があるので、その間の配線も沢山あっても大丈夫、が前提
- Cray-1 くらいのプロセッサが 1 チップにはいると話が破綻。
- 1 プロセッサチップと少数のメモリチップで 1 ノードを構成し、その間はネットワークでつなぐ構成が有利になる。
- 初期の代表: Cray T3D。

マイクロプロセッサ超並列→GPU

- 1チップの中に沢山プロセッサがはいるようになる。
- これは実はシステムレベルでのベクトル並列プロセッサの限界と同じ。
- マルチ(メニー)コアプロセッサでは、キャッシュコヒーレンシを維持した階層キャッシュとオフチップの共有メモリ
- キャッシュコヒーレンシの維持のための電力が支配的に
- GPU では(OS 動かしたりしないので)キャッシュコヒーレンシを緩和したり、捨てたりできる。

GPU→???

- コヒーレンシ諦めても、階層キャッシュで物理的に遠くにデータ移動させること自体が性能の限界になってきている
- と書いた以上、階層キャッシュを諦めて、なるべく物理的に遠くにデータ移動させないのが対応。
- PEZY-SC(キャッシュもあるけど)、Sunway SC26010、MN-Core等では演算器の物理的に近くに大容量メモリ配置

ではローカルメモリにすればそれでいいか？

- すでに問題になっているのが DRAM
- DDR5 メモリは 20pJ/bit くらい。電圧 1.3V とか。。
- LPDDR5 と GDDR6x: 10 pJ/bit くらい。配線が DDR5 のモジュールより短くて電圧も低い。
- HBMx: 3-4 pJ/bit。概ね 25mm くらいまで配線短くなっている。

3pJ/bit は 64 ビットワード移動に 0.2nJ なので、HBM でも 5GW 移動すると 1J 消費する。 B/F=4 の計算機作ると 10GF/W しかでない。 0.1 にしても DRAM メモリだけで 400GF/W までしかいかないのでキャッシュの分とかいれると 100GF/W あたり。

つまり

PVC の轍を踏まないためには、「データのチップ上・基板上の横方向の移動」を限りなく小さくする必要がある。

シリコンフォトニクスでとかいう話があるが、今のところ光変換のコストのほうが大きい

必要な方向

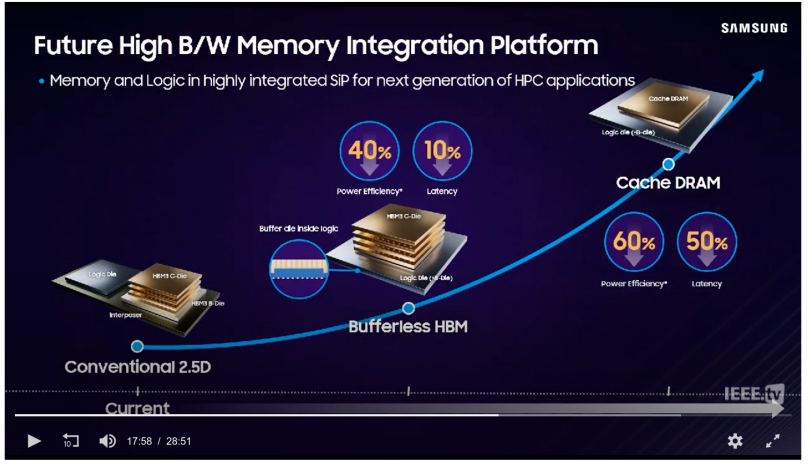
- 基本的には、 DRAM をプロセッサダイの上 (下かも) につむ。 HBM でやってることと変わりはない。
- 但し、 HBM では DRAM ダイの中心に集中しているのは配線を、もうちょっと細かい単位でだす必要がある (横方向の移動を減らす)
- プロセッサアーキテクチャ自体も、演算器の物理的に近くに分散したメモリがある形に移行する必要がある。
- 90 年代の SIMD 超並列計算機 (CM-2、MasPar とか) に類似。
- ポスト富岳 FS の神戸大学チームはこういう方向を検討している
- A 社とか B 社とか C 社とかが 3DDRAM 使う計画ありとの情報が色々。

Samsung の公開情報

IEEE.tv Tune in to where technology lives. IEEE

Home | Channels | Live | Events | Education | My Videos | Premium

Home > Heterogeneous Integration Platform for Next Generation Computing ('Beyond Moore')



Heterogeneous Integration Platform for Next Generation Computing ('Beyond Moore')

★★★★☆ 160 views

Download Share

Published on February 24, 2023

HBM に対して 40-60% の電力削減

Cache DRAM で 1pJ/bit 程度

(現状の HBMx は 2.5-3pJ/bit 程度)

まとめ

- PVC のアーキテクチャは、物理的に現実的な限界を超えて L2 バンド幅を高くしていて、結果的に電力性能目標を達成できなくなったようにみえる
- これは共有キャッシュをもつ GPU アーキテクチャ自体の限界。H100 とかは意味があるかどうかわからないところまで L2 バンド幅を下げている
- この問題の「解決」には、共有キャッシュから分散メモリへの移行が必要
- さらに、DRAM も 3D 実装にして分散メモリにしないといけない
- 神戸大学・PFN のポスト富岳 FS では最初からこの方向を検討
- 各社やってる気配あり。26-27 年にはでてくるんじゃないか。