RISC-V Days Tokyo 2021 Autumn
2021年11月17日

esperanto.ai

カスタム**機械学習（ML）**拡張を搭載した エネルギー効率の高い 高性能**RISC-V**プロセッサによる **ML**ワークロードの高速化

笠原栄二

エスペラント テクノロジーズ
eiji.kasahara@esperanto.ai

エスペラント・テクノロジーズ
AI革命をリードする**RISC-V**のファウンダリー企業

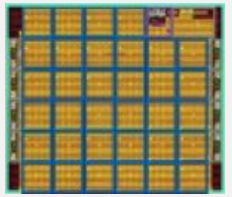
esperanto.ai

Industry Leadership:
Architecture, IP, AI, Silicon

チームには、RISCの共同発明者、CPU/IPのベテラン、データサイエンスとAIの専門家、マイクロアーキテクチャの専門家、低電力回路設計の専門家が含まれています

当社はRISC-Vの創立メンバーであり、CPU設計、AIシリコン開発、ソフトウェア＆ツール作成にRISC-Vをいち早く採用しています

当社のRISC-Vコアの性能と電力効率により、AIおよび非AIワークロードのデータセンターからエッジアプリケーションまで対応可能です

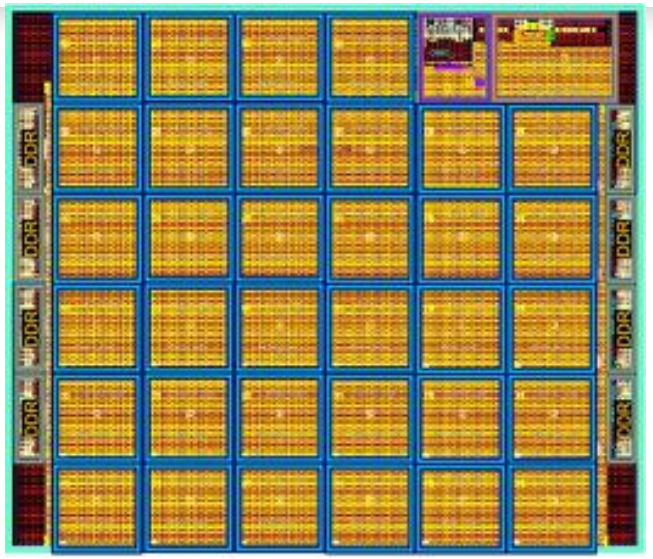


Esperanto社のRISC-VベースのET-SoC-1は、ホスト型およびセルフホスト型のシステムにおいて、主要なAIワークロードを20Wの電力で処理することができます。

Tier 1 投資家, ハイパースケーラー, エンタープライズ, そして SoC 企業に支持される

エスペラントのET-SoC-1は、世界最高性能の商用RISC-Vチップです

esperanto.ai



ET-SoC-1 TSMC 7nmで製造

- 240億個のトランジスター
- ダイの面積 : 570 mm^2
- 30,000個以上のバンプ

1088個 ET-Minion エネルギー効率に優れた64ビットRISC-Vプロセッサ

- それぞれがマルチスレッドで、ベクトル/テンソルユニットが付いている
- 代表的な動作 500MHz～1.5GHzを想定

4個 ET-Maxion 高性能64ビットRISC-Vアウトオブオーダープロセッサ

- 代表的な動作 500MHz～2GHzを想定

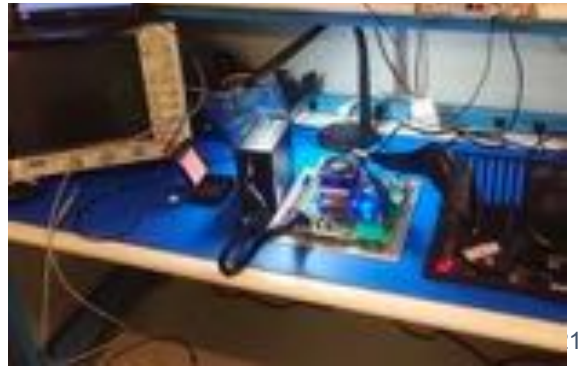
160 MB以上のオンダイSRAMと、チップ当たり最大32GBの外付けDRAM

１個 RISC-Vサービスプロセッサ セキュアブート、Root of Trust

消費電力は通常20ワット以下ですが、ソフトウエアにより最大値を上げる事ができる

パッケージ: 45x45mm PCBに2494個のボール

ステータス: シリコンは現在ブリングアップ・検証中



大規模データセンターにおけるMLレコメンドの要件と課題

今日、大規模なデータセンターにおけるレコメンドシステムの推論ワークロードのほとんどは、x86サーバで実行されています

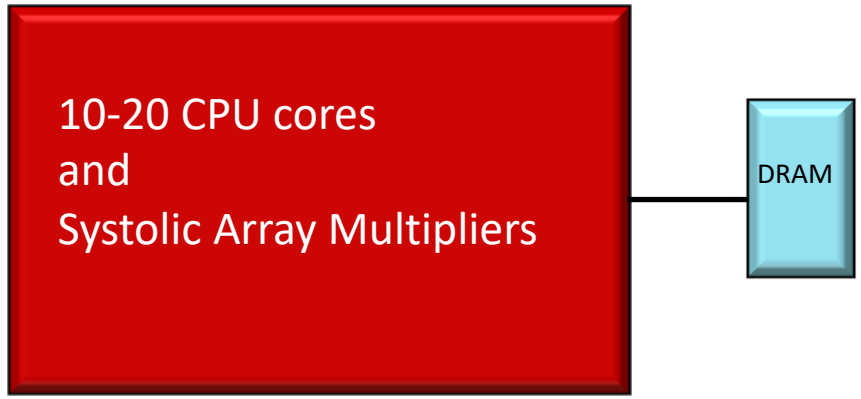
多くの場合、これらのサーバーにはアクセラレータカード用のスロットが用意されていますが、アクセラレータカードは重要な要件を満たす必要がある:

- **100TOPS～1000TOPS**のピークレートで、x86ホストCPU単体よりも高いパフォーマンスを実現
- カード1枚あたりの消費電力が**75～120ワット**と限られており、空冷式である必要がある
- **Int8**を強力にサポートするが、**FP16およびFP32**データもサポートしなければならない
- エンベッディング、ウェイト、アクティベーションのほとんどを格納できるアクセラレータカードのメモリ容量は**約100GB**
- **約100MB**のオンダイ・メモリ
- **密(dense)**な計算と**疎(sparse)**な計算の両方のワークロードを処理します.

  密な行列の乗算による疎な行列のエンベッディングルックアップ
- 急速に進化するワークロードに対応するため、過度に特殊化されたハードウェアに依存するのではなく、**プログラマブル**であること

エスペラントによるMLレコメンドのための優れた手法

他のMLチップのアプローチ



1つの巨大なホットチップが電力予算を使い切る
限られたI/Oピンバジェットがメモリバンド幅を制限
シストリックアレイ乗算機への依存性

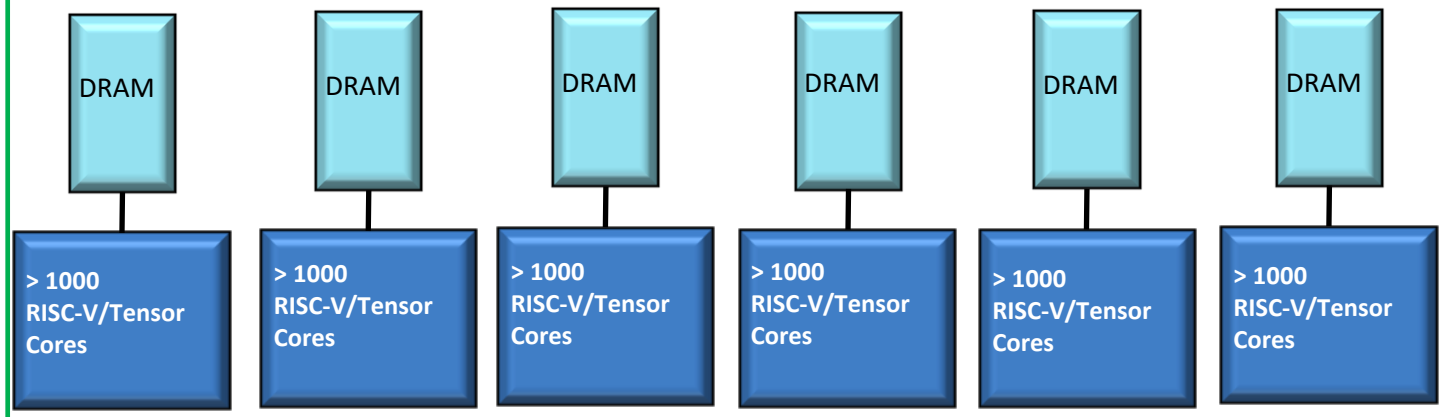
- ResNet50のハイスコアには最適
- 大規模なスパース・メモリではあまり効果がない

一握り（10〜20）のCPUコアのみ

- アレイのマルチプライヤーに問題が収まらない

場合のCPUコアによる**限定的な並列処理**
標準電圧：エネルギー効率が悪い

エスペラントの優れたアプローチ



パワーバジェット内に収まるように**複数の低電力チップ**を使用
性能、ピン、メモリ、バンド幅は、**チップが増えるほど大きくなります**
数千の汎用RISC-V/Tensorコア

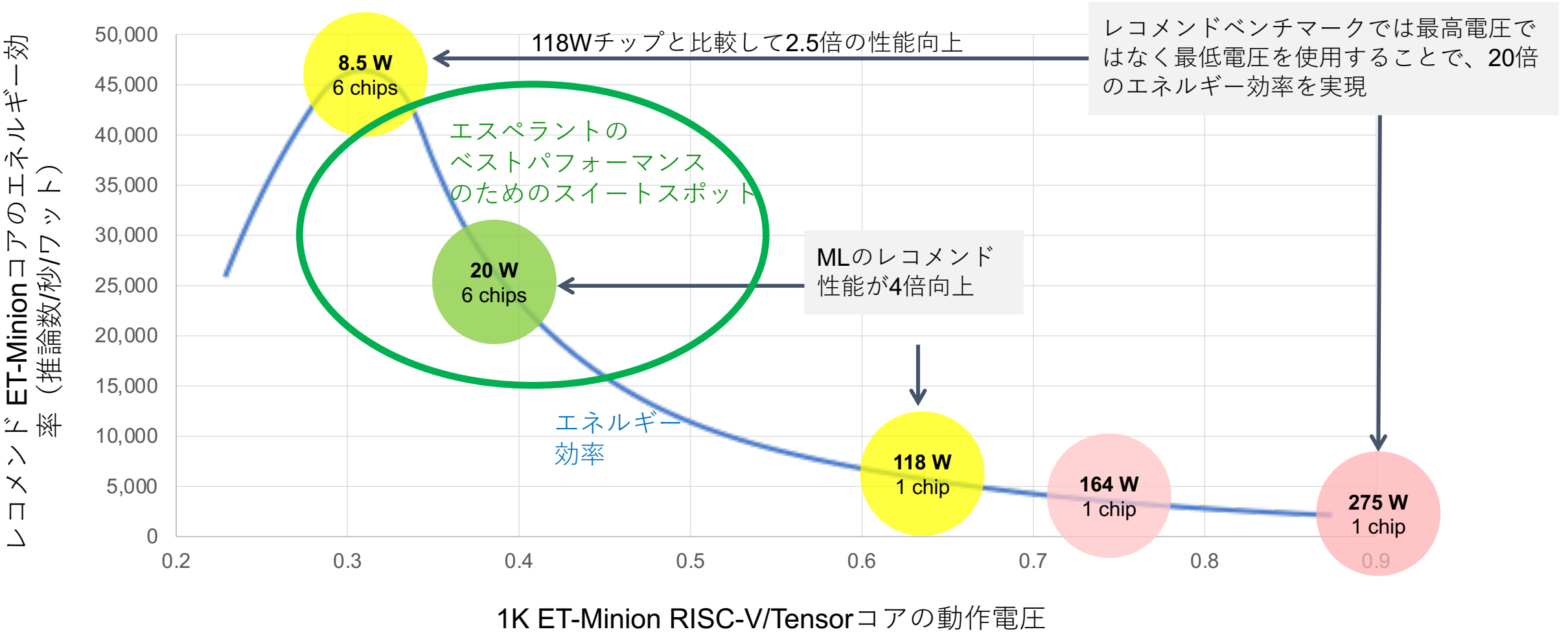
- 過度に特殊化した(例：systolic) HWよりも**はるかにプログラムしやすい**
- **スレッド**は大規模なスパースメモリーのレイテンシーを助ける

常に数千コアの**完全な並列処理**が可能
トランジスタの**低電圧動作はエネルギー効率が高い**

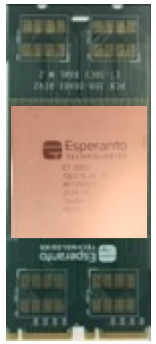
- 低電圧化で電力も削減
- **回路とアーキテクチャの両方の革新**を必要とする

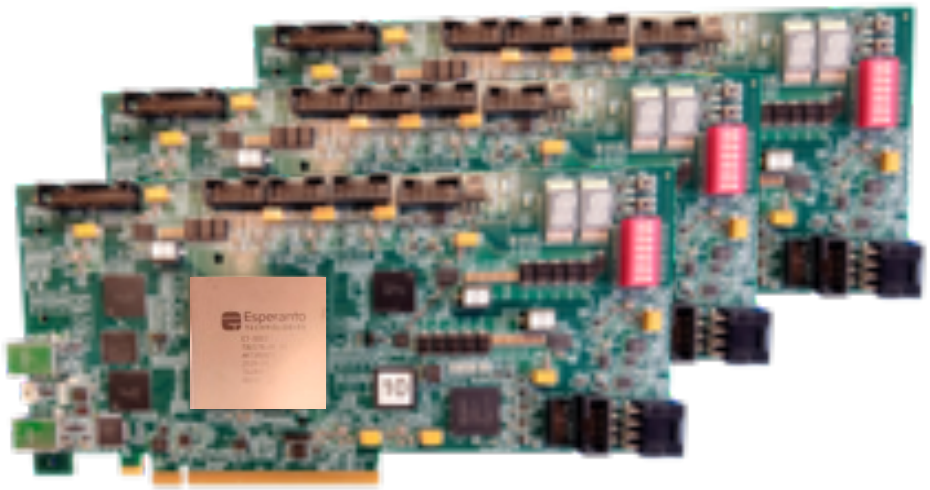
課題：最高のMLレコメンデーションパフォーマンスを120ワット制限の単一のアクセラレータカードに投入するには？

6つのEsperantoチップが連携し、120ワットで最高のML性能を実現



ET-SOC-1 AI/MLアクセラレータカードのポートフォリオ





Dual M.2 and Glacier Point v2 Cards:

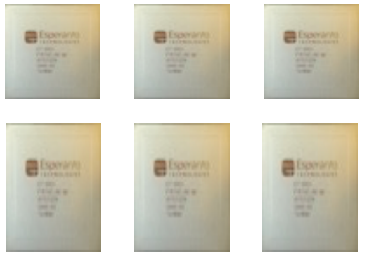
- OCP compliant
- Implemented using Dual M.2 form factors
- Fits up to six ET-SoC-1 chips per card GP v2 card
- 6,558 total RISC-V cores
- Up to 192 GB of DRAM
- Up to 822 GB/s DRAM bandwidth
- ~120 W total power consumption

PCI Express Cards:

- PCI-Small (HHHL) @ 20W, 16GB RAM
- Fits up to four ET-SoC-1 chips per PCI-Large (FHML) @ 100W, 64GB RAM
- Fits up to six ET-SoC-1 chips per PCI-Xlarge (FHFL) @ 140W, 96 GB RAM
- ET-SoC-1 power can be configured from 10W to to 60W per chip

ET-SoC-1は、既存のOCPデータセンターで大規模に展開することができます

x2 x4 x8



OCP Glacier Point v2 accelerator card holds:

- 6 Esperanto ET-SoC-1 chips

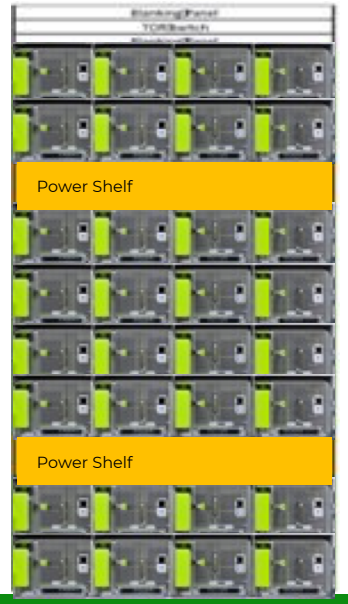
Yosemite v2 Sled holds:

- 2 Glacier Point v2 accelerator cards
- 12 ET-SoC-1 chips

Yosemite v2 Cubby holds:

- 4 Yosemite Sleds
- 8 accelerator cards
- 48 ET-SoC-1 chips

Top of Rack Switch



Rack with Yosemite v2 holds:

- 8 Yosemite v2 Cubbies
- 64 accelerator cards
- 384 ET-SoC-1 chips

Example OCP Data Center holds:

- @ 30 sq. ft. per OCP rack
- Estimated 4K-20K racks / data center
- Millions of ET-SoC-1 chips

ベクトル・テンソルユニットをカスタム搭載した省電力RISC-V CPU ET-Minion

esperanto.ai

ET-Minionはカスタムメイドの64ビットRISC-Vプロセッサです

- 低電圧動作が可能なようにアーキテクチャと回路を最適化
- 低ゲート/ステージのインオーダー・パイプラインにより、低電圧時のMHzを向上
- 2つのハードウェアスレッドを実行可能
- ソフトウェアで設定可能なL1データキャッシュおよびスクラッチパッド

各Minionには、MLに最適化されたベクトル/テンソルユニットが搭載されています

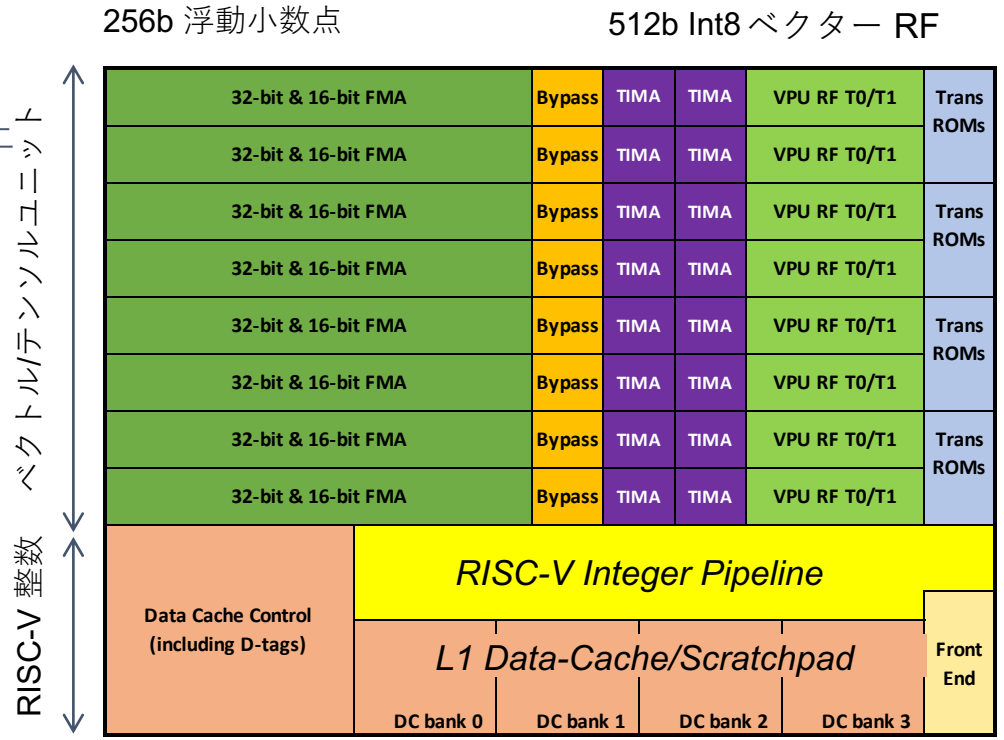
- 1サイクルあたり512ビット幅の整数
- 128個の8ビット整数演算/サイクル、32ビット整数に累積
- 256ビット幅の浮動小数点演算/サイクル
- 16個の32ビット単精度演算/サイクル
- 32個の16ビット半精度演算/サイクル

新しいマルチサイクルテンソル命令

- 1つのテンソル命令で最大512サイクル（または最大32K演算）の実行が可能
- 命令フェッチの帯域幅を減らし、電力を削減
- テンソル命令実行中はRISC-V整数パイプラインをスリープ状態にする

ベクトルの超越関数命令（指数関数、対数、および三角関数など）

動作範囲 300MHz～2GHz

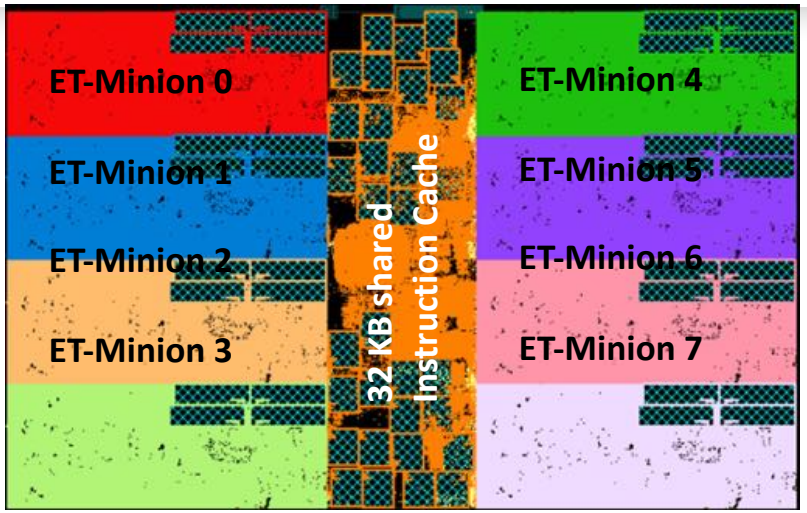


低電圧動作に最適化されたET-Minion RISC-Vコアおよびベクトル/テンソルユニットにより、エネルギー効率が向上

エネルギー効率の高いML動作に最適化
各ET-Minionは、1GHzあたり128 Int8 GOPSのピークを提供することができます

8個のミニオンが “ネイバーフッド(neighborhoods)” を形成する

esperanto.ai



ネイバーフッド・コアが密接に連携

- 8コアの物理的近接性を活かしたアーキテクチャの改善
- ほぼ常に高度な並列コードが実行されていることを利用

同じコードを実行するコアからの最適化

- 8つのET-Minionが1つの大きな命令キャッシュを共有することで、多くの個別の命令キャッシュよりもエネルギー効率が高くなります
- Cooperative loads (協調ロード) により、L2キャッシュへのメモリトラフィックを大幅に削減

新しい命令で連携を効率化

- 新しいTensor命令は、命令フェッチのバンド幅を劇的に削減する
- グループ内のローカル同期を高速化する新命令
- 新しいSend-to-Neighbor（ネイバーへの送信）命令
- 新しいReceive-from-Neighbor（ネイバーからの受信）命令

32個のET-Minion CPUと4MBメモリが “Minion Shire”を形成

1つの Minion Shire に32個のET-Minion RISC-Vコアを搭載

- 8コアの "neighborhoods"を4つ配置

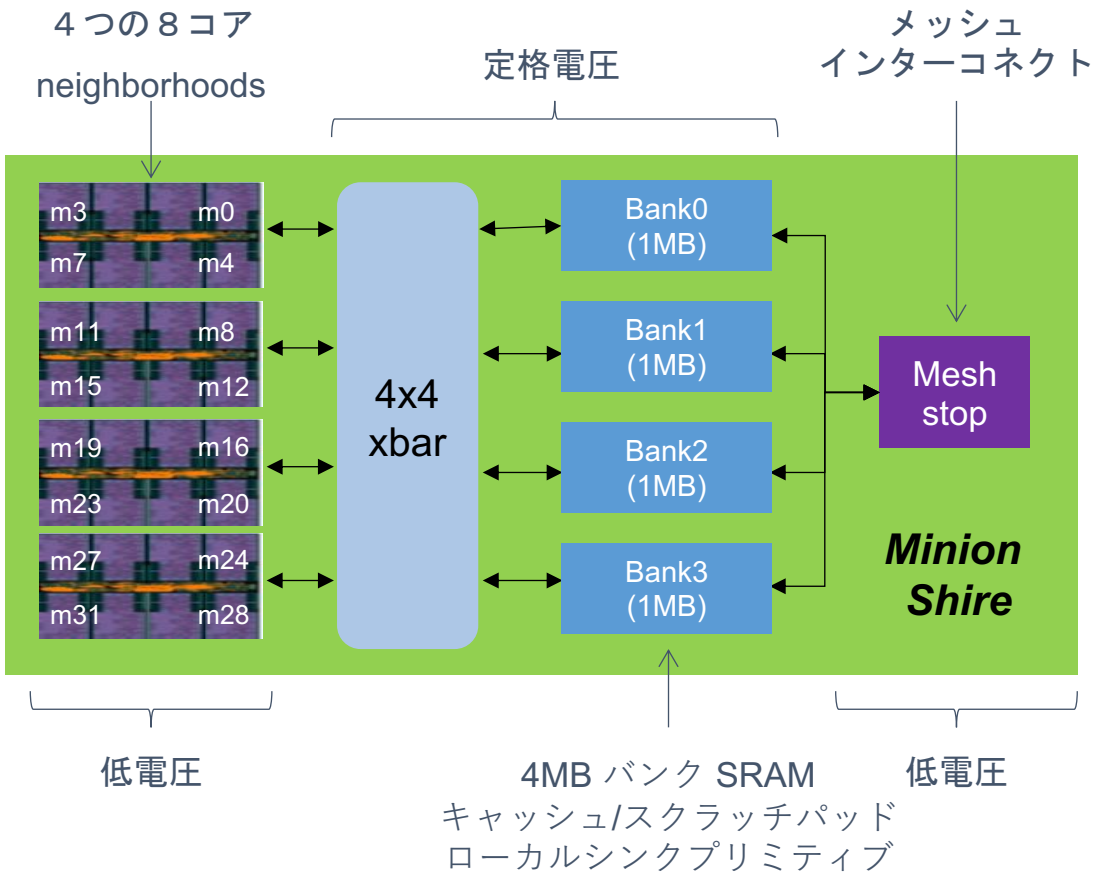
ソフトウェアで構成可能なメモリ階層

- L1データキャッシュはスクラッチパッドとしても設定可能
- 4つの1MB SRAMバンクは、プライベートL2、共有L3、およびスクラッチパッドとして分割可能

Shiresはメッシュネットワークで接続されています

新しい同期プリミティブ

- 高速・ローカル・アトミック
- 高速・ローカル・バリア
- 高速ローカルクレジットカウンタ
- プロセッサ間の割り込みをサポート



ET-SoC-1:フルチップ内部のブロック図

esperanto.ai

34 Minion Shires
- 1088 ET-Minions

8 Memory Shires
- LPDDR4x DRAM コントローラ

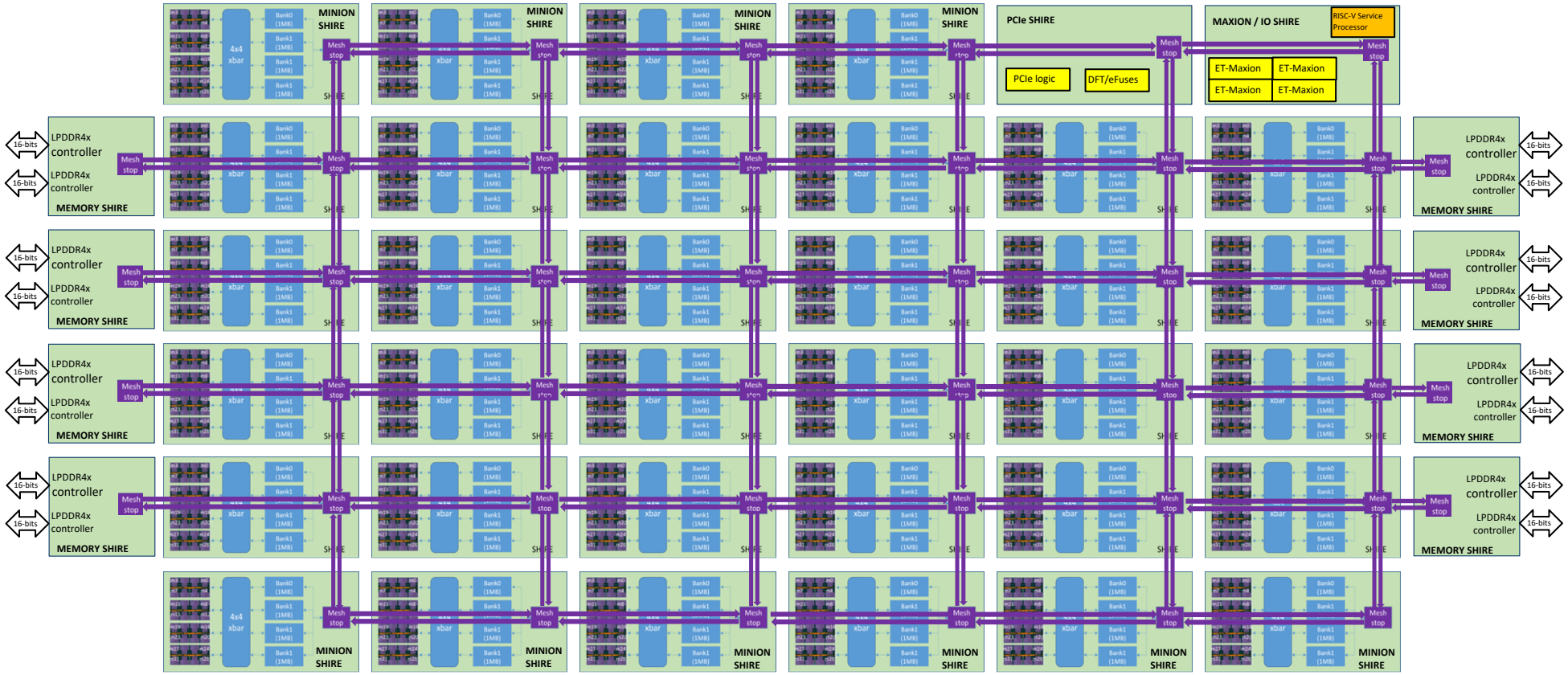
1 Maxion / IO Shire
- 4 ET-Maxions
- 1 RISC-V Service Processor

PCIe Shire

160 MB of オンダイ SRAM

x8 PCIe Gen4

Secure Root of Trust



エスペラントのML最適化されたベクトル/テンソル拡張

RISC-V Vectorの仕様が整う前に開発され、機械学習に最適化されている

ベクトル、テンソル、マルチコアの協調動作を新たに採用し、最小のダイエリアと最大の効率で高性能を実現

ML指向のマルチレベルのレジスタ/キャッシュ/スクラッチパッド階層との協調動作

サポートするデータ型は、Int8、Int32、FP16、FP32

- ベクトル演算ではInt32、FP32をサポート; テンソル演算ではInt8、FP16、FP32
- 記憶形式としてBF16をサポート; BF16データに対する演算はFP32で実行される

すべての命令エンコーディングはRISC-Vの規則に準拠

conventions

Vector ops
- Convert
- Compute
- Compare
- Gather/Scatter
- Load/Store
- Mask
- Transcendental

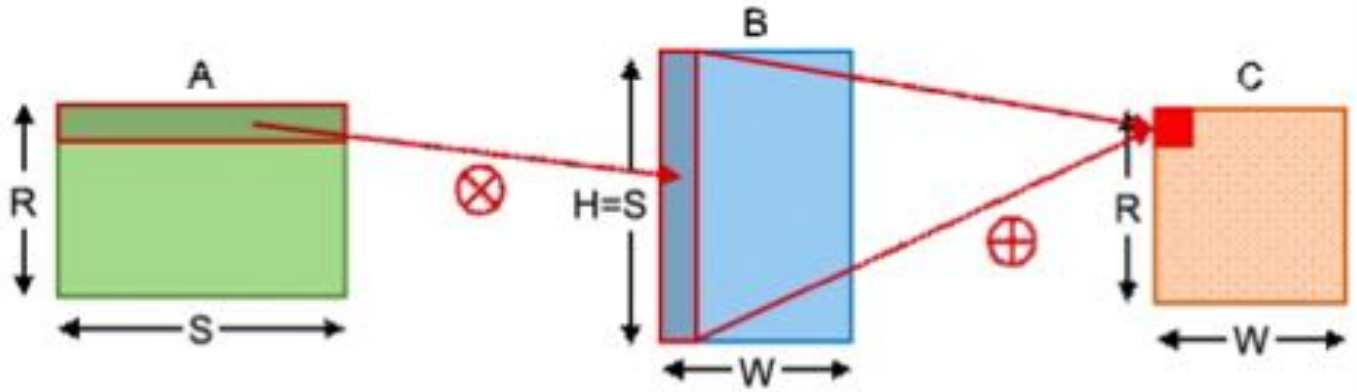
Tensor ops
- Broadcast
- Compute
- Load/Store (with optional interleave/transpose)
- Reduce
- Transform

Cache Control ops
- Atomics
- Barrier and Credit
- Cache evict and flush
- Cache lock
- Cache partition
- Cache prefetch
- Scratchpad reservation

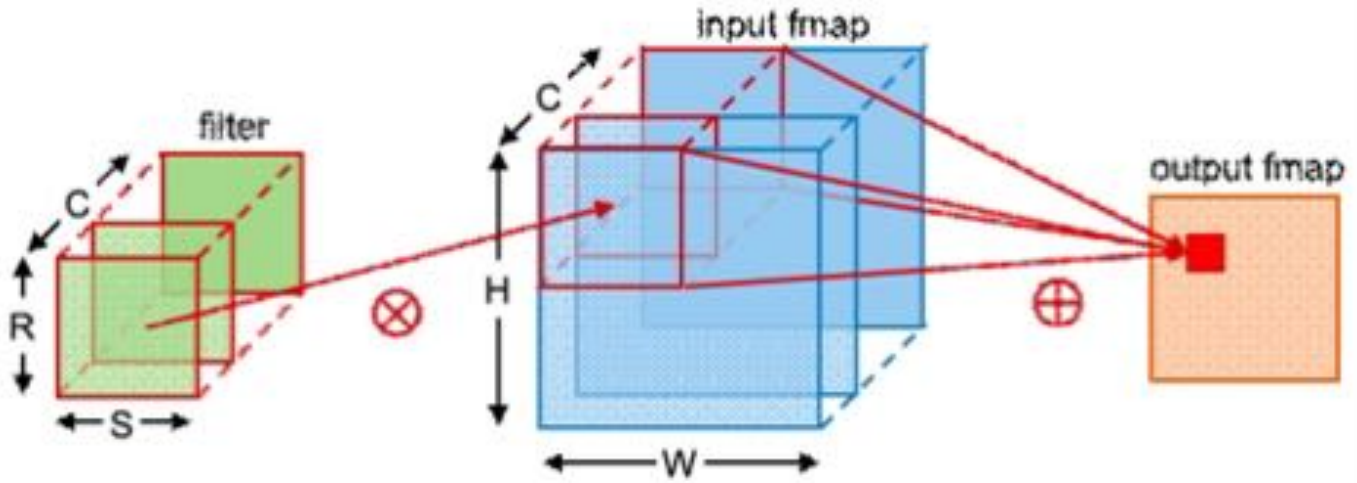
テンソルの乗算演算

esperanto.ai

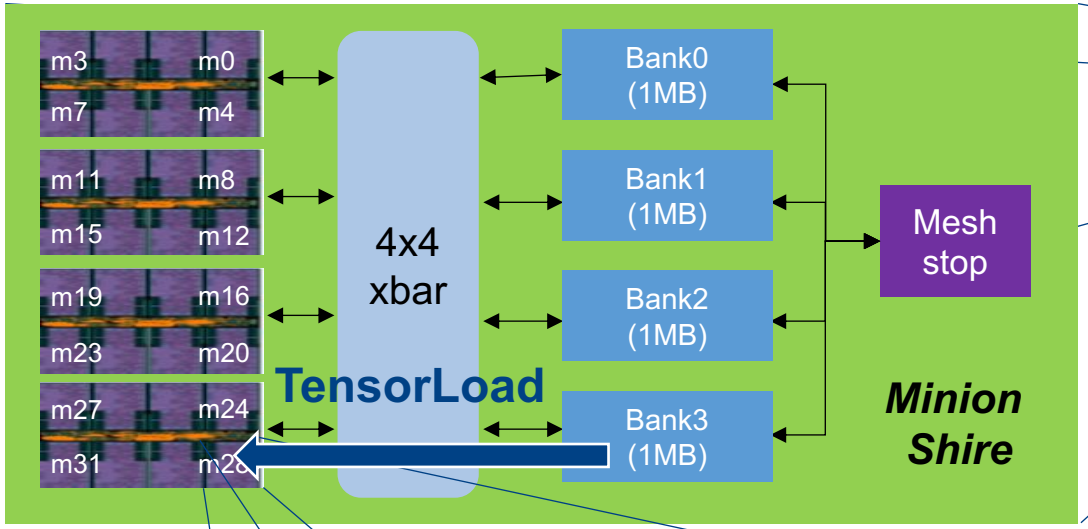
行列AとBの2次元行列乗算

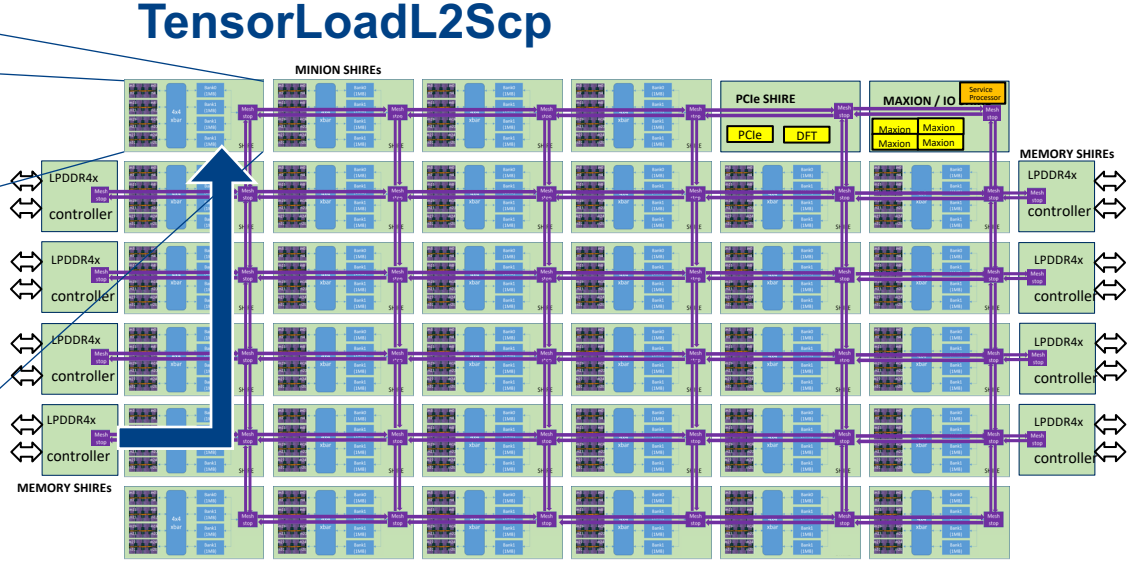


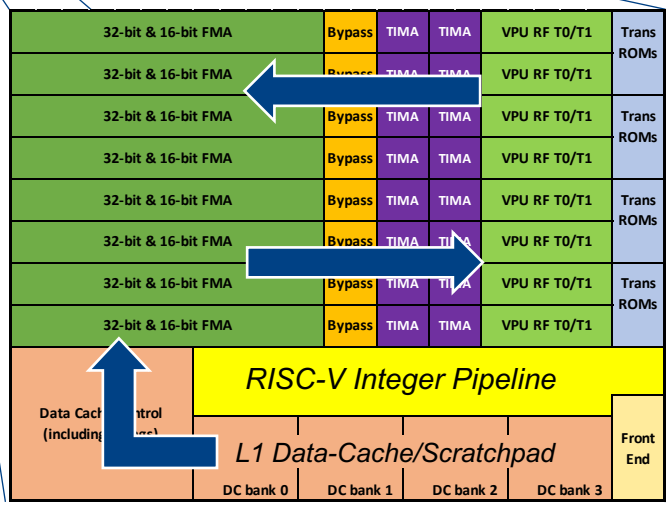
テンソルAとBの3次元テンソル乗算



データ移動のためのカスタム命令







TensorFMA

- TensorLoadL2Scpは、入力データをメモリ（常駐している場合は分散型L3）からShireローカルのスクラッチパッドメモリに転送します
- TensorLoadは、各TensorFMA操作用の入力データのブロックをMinionのスクラッチパッドに移動させる
- TensorFMAの結果はVPUのレジスタファイルに格納される

TensorFMA16A32 Operation は512サイクルで100%の使用率で動作

TensorFMAが行列の乗算を実行 A x B [+]=C

A行列とB行列は、TensorLoadによってMinionスクラッチパッドまたはTensorBレジスタファイルに配置されます

- 最大64バイトのA行列が最大16行まで
- 最大64バイトのB行列が最大16行まで

Cマトリックスは、VPUのRF

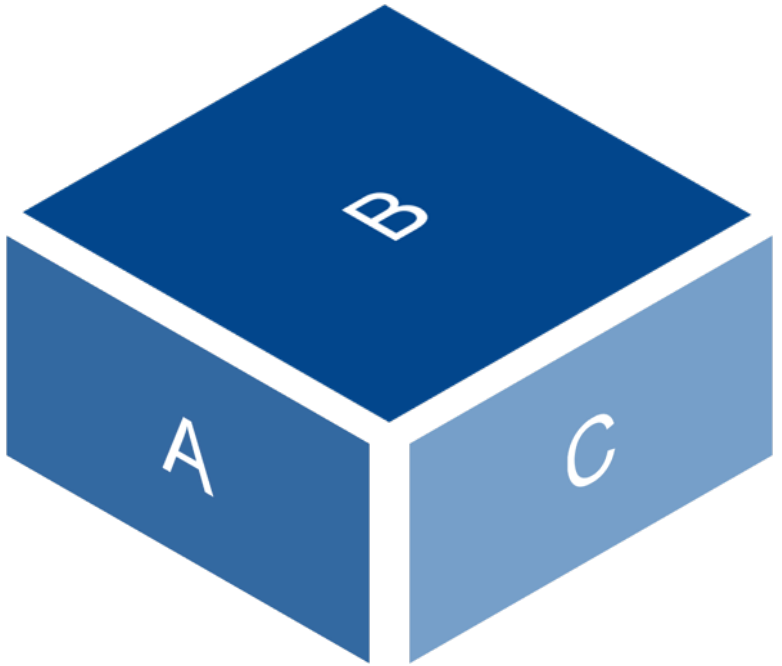
FP16 mul / FP32 addの最大タイルサイズは、16 x 32 x 16です

このオペレーションは512サイクルで実行されます

- 512 FMA x 8 レーン x 4 FLOPs/レーン = 16,384 FLOPs

パフォーマンスと電力効率を最大化するために、最大のタイルサイズで作業したいと思います

将来の派生製品は、将来のワークロードに応じて必要な拡張が可能



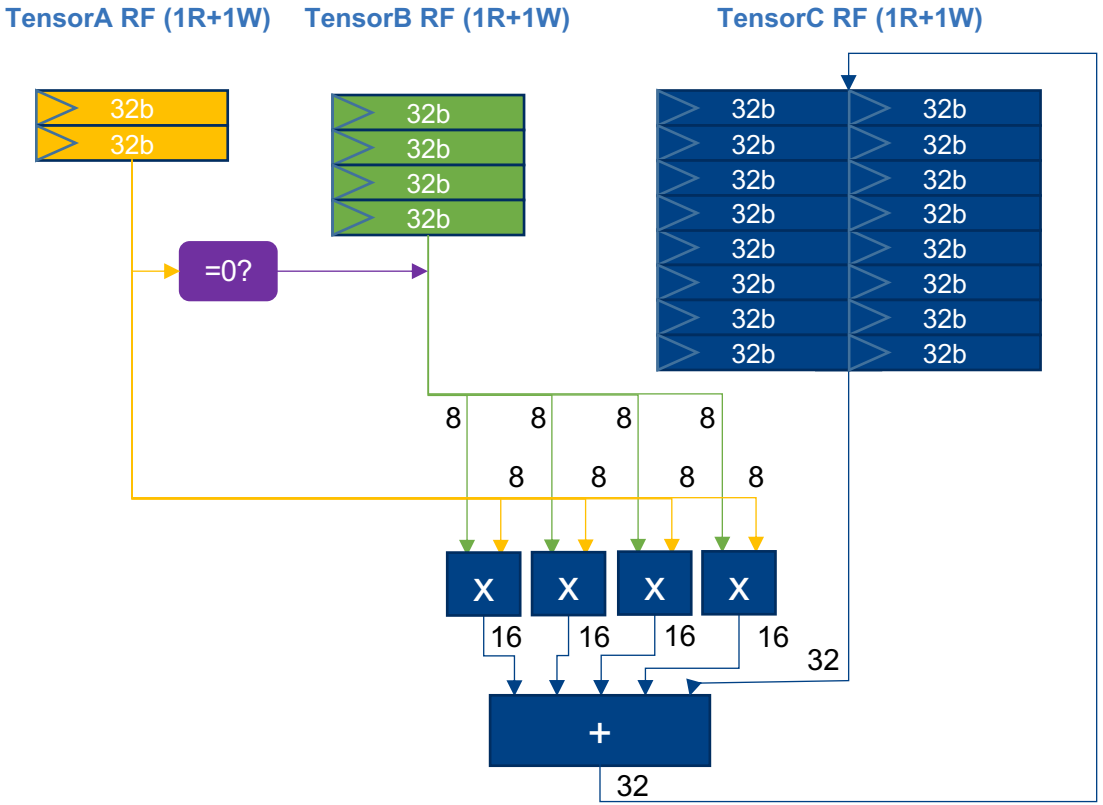
独立したInt8 TIMAユニットは、FP16に比べて4倍のスループット

TIMA = Tensor Integer Multiply-Add(テンソル 整数 乗算加算)

メインVPUのレジスタファイルとは別のプライベートレジスタファイルにより、テンソル演算時のダイナミックスイッチングを最小化

ダイナミックパワーのさらなる削減のために、メインバイパスマルチプレクサのトグルを行わない

Tensor-Aの値がゼロの場合、Tensor-Bのレジスタファイルはゲートされ、以前の値が保存される



Coordination Ops: Atomic BarriersとCreditsについて すべての ET-Minionsを連携させ、効率化

- ET-SoC-1を含むメニーコアチップから良い結果を得るためには、1,000個以上のRISC-Vコア間で実行の流れを効率的に調整することが重要です
- エスペラントは、MinionsとShiresにカスタム命令とそれに付随する特殊レジスタを追加し、高速で効率的な調整をサポートしました
  - **Barriers**: 各ET-Minion Shireには、バリアカウンターの特殊レジスタセットが含まれています; アトミック操作は、Minionのタスクを追跡し、進捗状況を同期させるために使用されます
  - **Credit counters**: スレッドごとのローカルクレジットカウンタは、リモートエージェントによって更新され、カスタム命令によってチェックされ、共有リソースへのアクセスなどの進捗を許可します
- 1つのET-Minion Shireは、スレッドおよびShireレベルの特別なレジスタへのストアを使用して、他の32のShireの操作を調整するように割り当てられます

エスペラントのソフトウェア開発環境

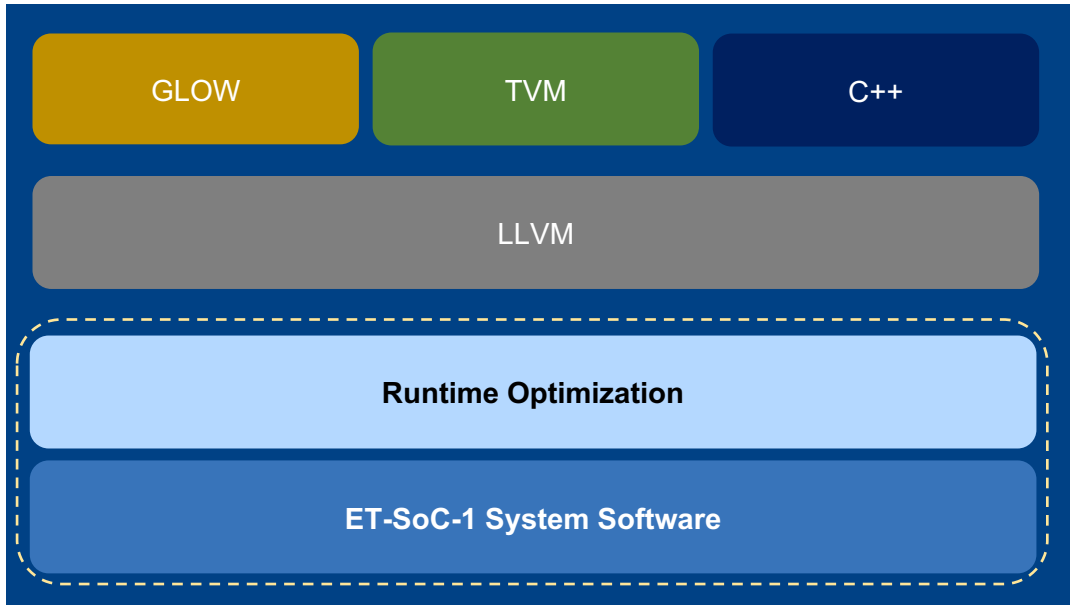
esperanto.ai

RISC-V

Software / Tools エコシステム

esperanto

Software / Tools 開発



Development Tools

- Console / Debugger
- Performance Monitor
- ET IP Models
- Benchmark Tools
- Virtual Platform (OVP)

Management Utilities

- Diagnostics
- Firmware Updater

Esperantoは、SWとツールのソリューションに重点的に投資し、RISC-Vのオープンソース・エコシステムを拡大し、Esperantoの顧客により多くの選択肢を提供しています

サマリー

esperanto.ai

Esperanto社の低電圧技術は、ワットあたりの性能が最も高い、差別化されたRISC-Vプロセッサを提供

- エネルギー効率が重要!
- ベストパフォーマンス・パー・ワット 一定のワット数で最高のパフォーマンスを発揮
- データセンターの推論ワークロード（特にMLレコメンデーション）にエネルギー効率の高いアクセラレーションを提供するソリューション

Esperanto Vector/Tensor Extensionsは、低消費電力で高いML性能を実現

- 低電圧ベクトル/テンソルユニットは、最小限の面積と電力で整数および浮動小数点の推論をサポート
- これらのユニットは、マルチレベルのレジスタ/キャッシュ/スクラッチパッド階層と連動し、多様なアルゴリズムに対応

Esperanto ET-SoC-1は拡張性の高いスケーラブルな設計

- MLレコメンドを効率的に実行
- 数千個の汎用RISC-Vコアは、他の多くの高度な並列計算タスクに適用可能
- モジュール式のアプローチにより、クラウドからエッジ、そして他の半導体プロセスへと設計を拡張可能

Early Access Program for qualified customers beginning later in 2021 (for info, contact: chips@esperanto.ai)

Thank you!

ET-SOC-1を評価したい方は、下記までお問い合わせください。

World Wide: chips＠esperanto.ai
日本語: takashi.murayama@esperanto.ai