AIコンピューティング研究ユニット
本村研　藤木研
紹介資料

25/11/12

総合研究院

AIコンピューティング研究ユニット (ArtIC)

情報通信系　本村研　藤木研

Institute of SCIENCE TOKYO

ARTIC

ArtICの成り立ち

情報通信系
本村・劉研究室 => 本村・藤木研究室
(23年2月まで)　　(23年11月から)

=
(イコール)

総合研究院
AIコンピューティング研究ユニット

2019年4月に発足

ARTIC

Artificially Intelligent Computing Research Unit
もう一つの意味: **ArtなIC(集積回路)** => ソフトとハードの総合技術・芸術

本村・藤木の自己紹介

本村

テクノロジストの時代

AI半導体 世界と競う

東京工業大教授
本村 真人氏

日経新聞
21/10/26日朝刊

’87 京大理学部修士
’96 京大工学博士
’87-’11 NEC研究所
’11-’18 北大
’19- 東工大
ArtIC立ち上げ

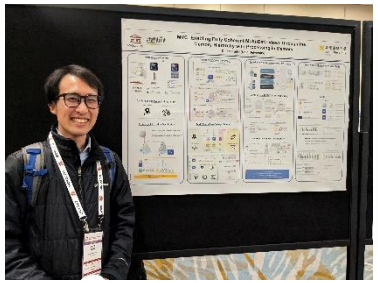
LSIのオリンピック ISSCCで次々に世界初のチップを発表

4千人参加の最高峰会議

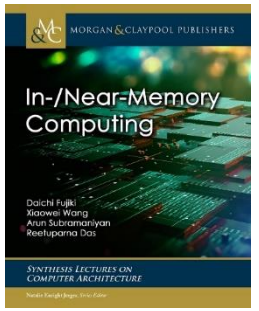
多くの国内外シンポジウムでAIハードウェアの招待講演

藤木

’16 慶應理工学部卒
’16 ミシガン大 Ph.D.課程入学
’22 ミシガン大 Ph.D.取得
’22-’23　慶應大学 助教
‘23.11月 東工大 准教授
ArtICに参加



計算機アーキテクチャの5大トップ国際会議で次々に論文発表。日本には稀有な存在

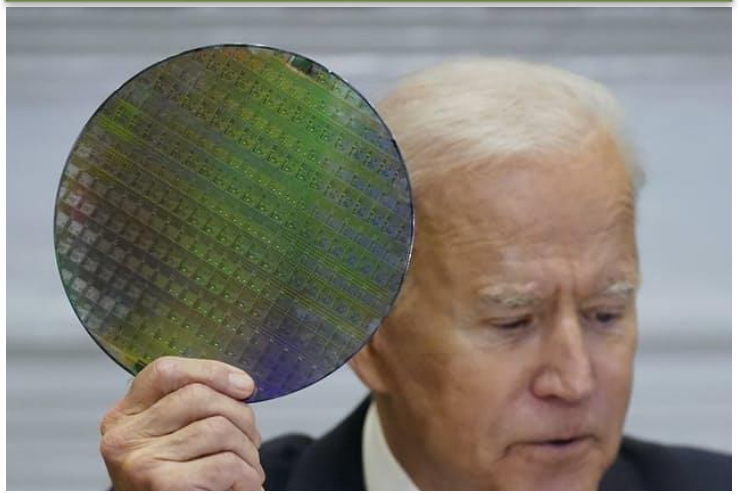


若くして既に著書も…

共同研究@Nvidia, Samsung, Meta FAIR

ArtIC 5年間の社会情勢の変化

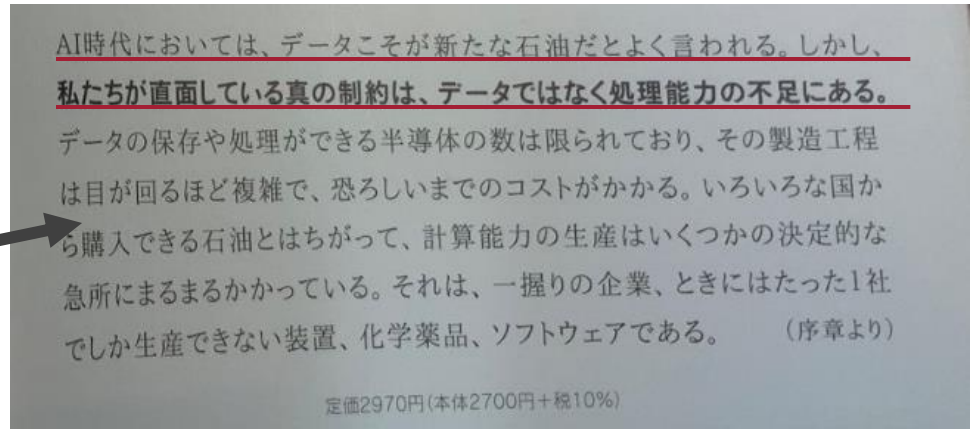
バイデン大統領 半導体 $50B 投資宣言 (2021)



TSMC熊本工場(2022-)

ラピダス千歳工場(2023-)

Chip War (2022)



AIの勃興と国際情勢の緊迫化

↓

半導体・集積回路・チップ(すなわちハードウエア)が、技術的にも社会的にも注目の的へ

AIコンピューティング研究の背景

急ピッチで進むAI技術の社会応用

自然言語・自動
対話型AI

スマート
ロボット

自律航行
ドローン

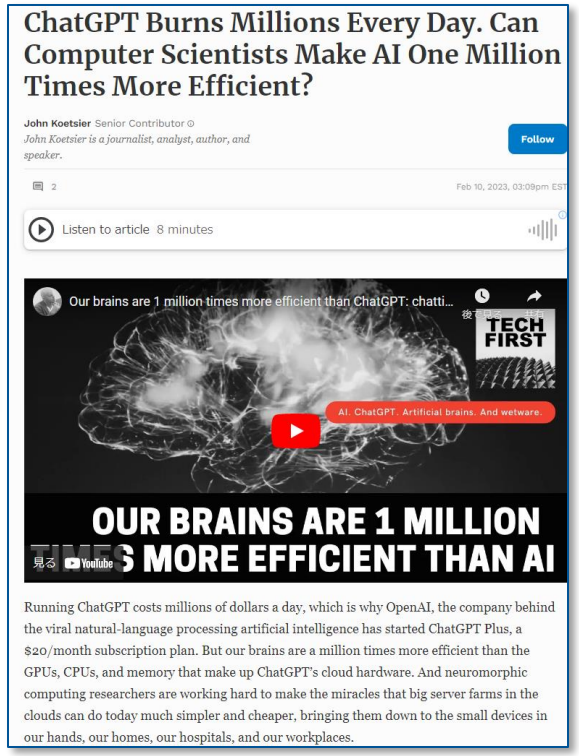
スマート社会
インフラ

AI処理の急拡大 => エネルギー消費問題の深刻化

より豊かで低環境負荷なスマート社会を実現したい

↓

「AIコンピューティング」の処理効率を向上する
情報処理アーキテクチャの技術革新が必要



ChatGPT Burns Millions Every Day. Can Computer Scientists Make AI One Million Times More Efficient?

John Koetsier Senior Contributor
John Koetsier is a journalist, analyst, author, and speaker.

Follow

Feb 10, 2023, 03:09pm EST

Listen to article 8 minutes

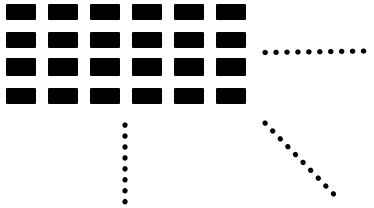
Running ChatGPT costs millions of dollars a day, which is why OpenAI, the company behind the viral natural-language processing artificial intelligence has started ChatGPT Plus, a $20/month subscription plan. But our brains are a million times more efficient than the GPUs, CPUs, and memory that make up ChatGPT's cloud hardware. And neuromorphic computing researchers are working hard to make the miracles that big server farms in the clouds can do today much simpler and cheaper, bringing them down to the small devices in our hands, our homes, our hospitals, and our workplaces.

(Forbes)

ChatGPTの学習には
一般家庭の数百年分も
の消費電力が必要
(Stanford大学報告書)

一つの視点: 大脳とGPU

GPU(数万モジュール)



演算性能	メモリ容量
数 Exa (10^6 Tera) Ops/sec	Peta (10^6 Giga) Byte

消費電力

数MW

GPU(単体モジュール)

演算性能	メモリ容量 (iPhone並み)
100 Tera Ops/sec	32 Giga Byte

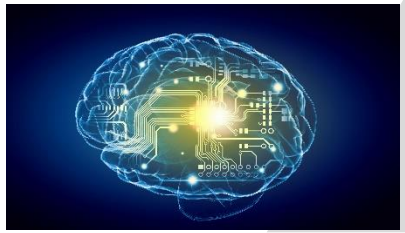
消費電力

400W超

数十倍

数万倍

人間の大脳



演算性能 (iPhone並み)	メモリ容量
2 Tera Ops /sec	Peta (10^6 Giga) Byte

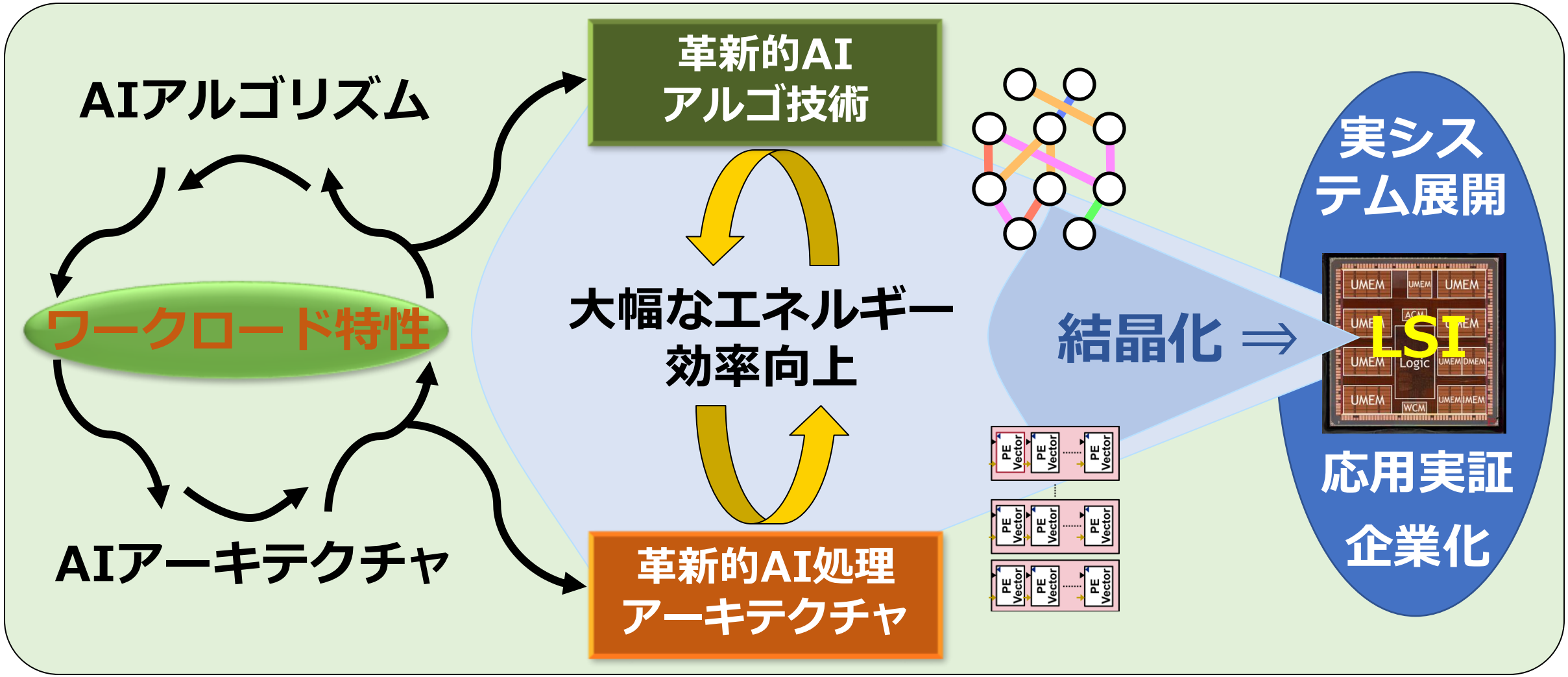
数十倍

～20W

消費電力

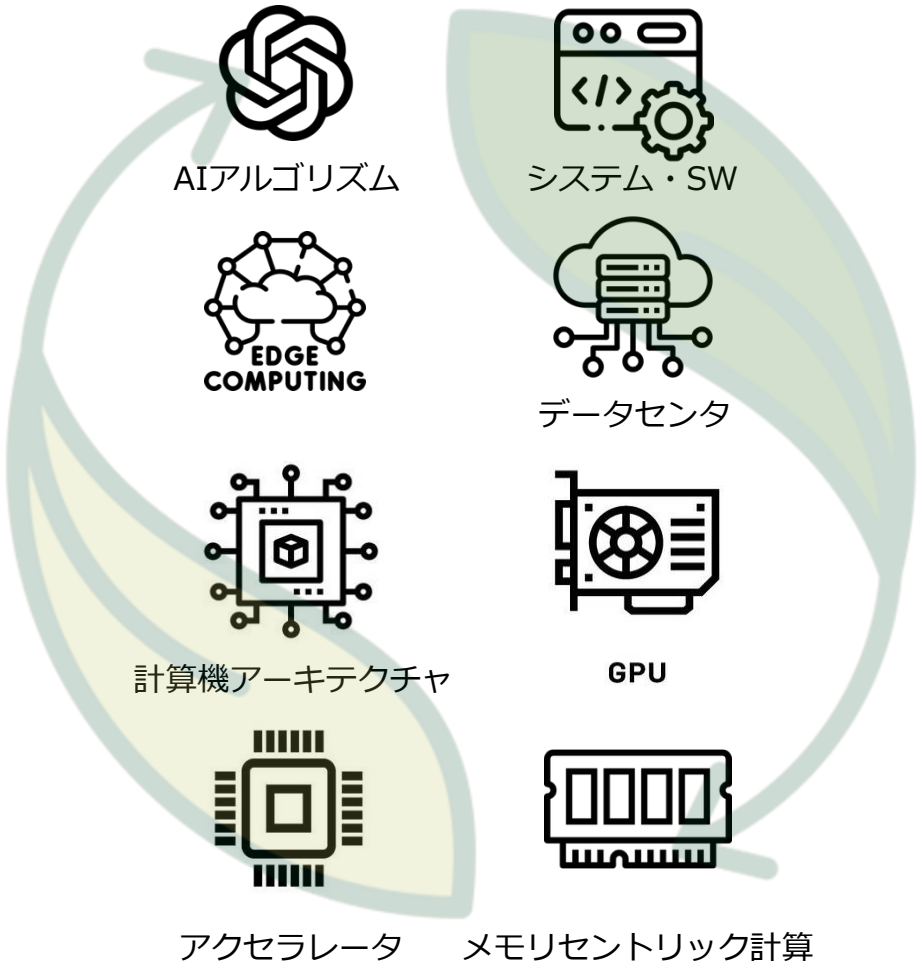
工夫すべきところは山ほどある
↓
研究しよう！

ArtICの研究スタイル



SW-HWの両面にわたって研究を推進

ArtICの研究スタイル



エッジ計算からサーバーまで様々なフォームファクタを俯瞰し、1チップの開発に閉じない、効率と柔軟性を両立する次世代の計算機基盤を実現

・データ移動コストに注目したメモリ内計算技術とそのシステム応用

・AI向けメモリ最適化コプロセッサ

・HW/SWコデザインを通した高効率システム

・AIワークロードを意識した時空間展開型コンピューティング

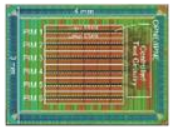
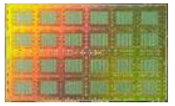
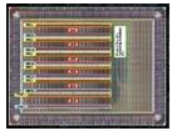
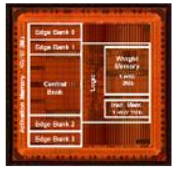
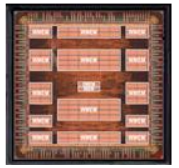
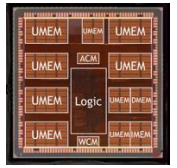
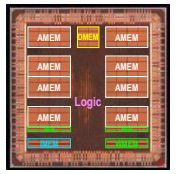
・HWによる高効率実行に最適化したAIアルゴリズム

大規模生成モデル・ゲノムプロセッシング・プライバシープリザーブド計算・リアルタイムデータベースなどの重要アプリの効率化へ

SW-HWの両面にわたって研究を推進

ArtICの世界最先端AIチップ実績

太字下線部は**世界初の技術**

	半導体プロセス
□**二値化**DNNアクセラレータチップ ■**VLSIシンポジウム2017**	 65nm
□**対数量子化**・**三次元積層**DNNアクセラレータチップ ■**ISSCC2018, JSSC2018**	 40nm
□**完全結合型・全並列型**デジタルアニーリングチップ ■**ISSCC2020, JSSC2020**	 65nm
□**シフト演算型**・直積アレイ型DNNアクセラレータチップ ■**Hot Chips 2021**	 40nm
□**ランダム重み固定型**DNNアクセラレータチップ ■**ISSCC2022、ICML2022**	 40nm
□完全結合型**メタモルフィックアニーリング**チップ ■**ISSCC2023**	 40nm
□超低電力・**ビット累進型**DNNアクセラレータチップ ■**VLSIシンポジウム2023**	 40nm
□**三重・不規則スパース性対応**DNNアクセラレータチップ ■**A-SSCC2024**	 40nm

最先端のAIアルゴリズムに注目し**アーキテクチャ・イノベーション**にこだわって研究を推進.

目玉となる成果に絞って**チップ実装・実機デモ**.

チップ化は成果の極く一部であり、**アルゴリズム-アーキテクチャ融合型**の研究が、研究活動の大半.

ArtICの世界最先端アーキテクチャ研究実績

- In-ReRAMベクタ計算アーキテクチャ
  - ASPLOS 2018
- オートマトン型ゲノムプロセッサアーキテクチャ
  - ISCA 2018
- In-SRAM SIMT計算アーキテクチャ
  - ISCA 2019
- GPU内SpMMデータ変換コプロセッサ
  - SC 2019
- DP型ゲノムプロセッサアーキテクチャ
  - MICRO 2020
- 複数階層型インメモリ計算アーキテクチャ
  - MICRO 2022
- ビュー生成によるコヒーレントPIMアーキテクチャ
  - MICRO 2023
- モバイル向けIn-SRAM PIM ISA
  - HPCA 2025
- GNNアクセラレータ
  - ISCA 2025

データ移動コストを大幅に低減するPIM/IMC技術に注目しアーキテクチャ・イノベーションにこだわって研究を推進.

コンパイラやシステム設計など、計算機スタックを縦断的に分析し実装・及びシミュレーション評価.

ゲノム分析などのAI以外の重要アプリにも着目し、アルゴリズム-アーキテクチャ融合型の研究を実施.

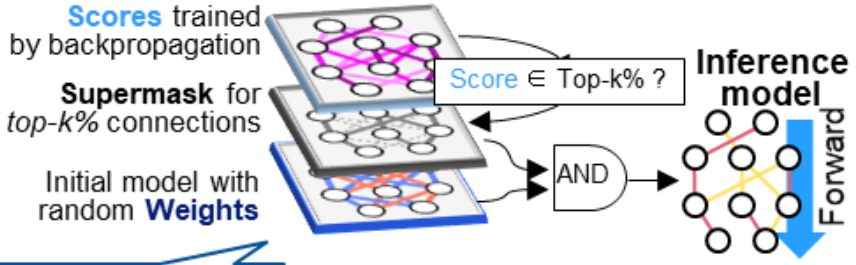
Fixed-Random-Weight DNN Chip @ ISSCC2022

Algorithm

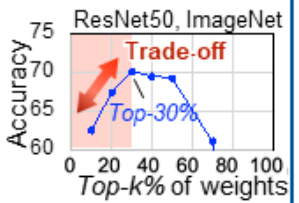
Background for Deep Learning

- Requirements for external memory access reduction
- The advent of a new efficient algorithm

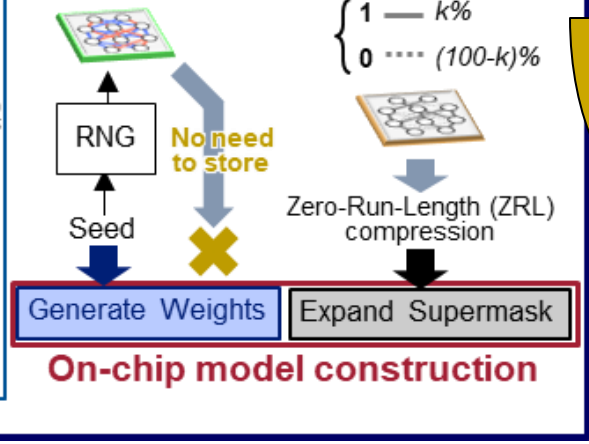
Hidden Network (HNN) [V. Ramanujan+, CVPR2020]



Key Features

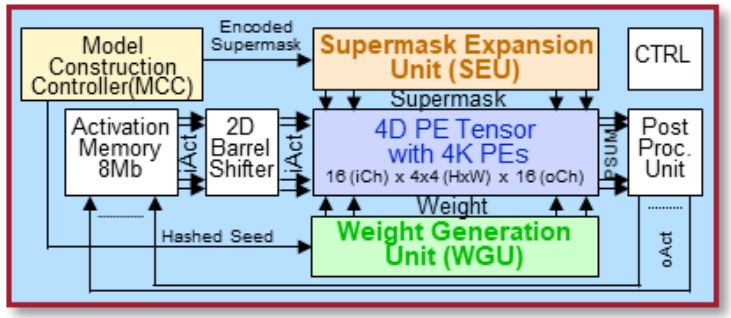


Architecture Implication



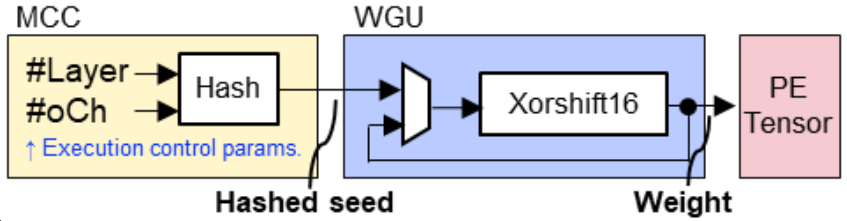
Architecture

Hiddenite Chip

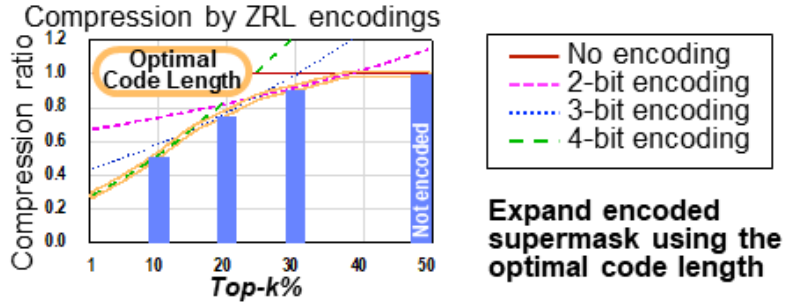


On-chip Weight Generation

Hashed seeds eliminate the need to store weights without accuracy degradation

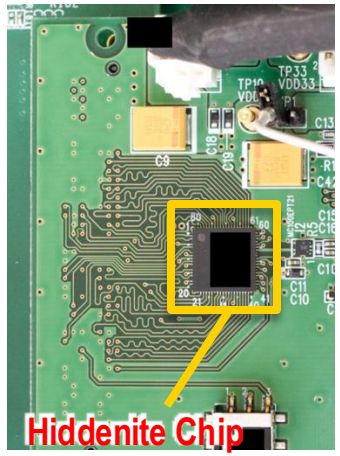


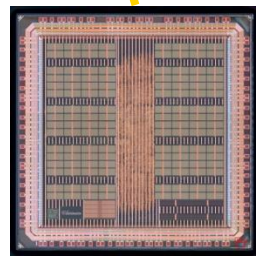
On-chip Supermask Expansion



Expand encoded supermask using the optimal code length

Real Chip & Demo System





ArtIC人員構成

		本村研	藤木研	
教員	教授	本村 真人		全体運営、ディープラーニング
	准教授		藤木 大地	アーキテクチャ
	助教	金子 竜也		立ち上げ
	特任助教/ポスドク			
秘書		橋本，土屋		
学生	博士	9名		
	修士	5名		
	学部	3名		

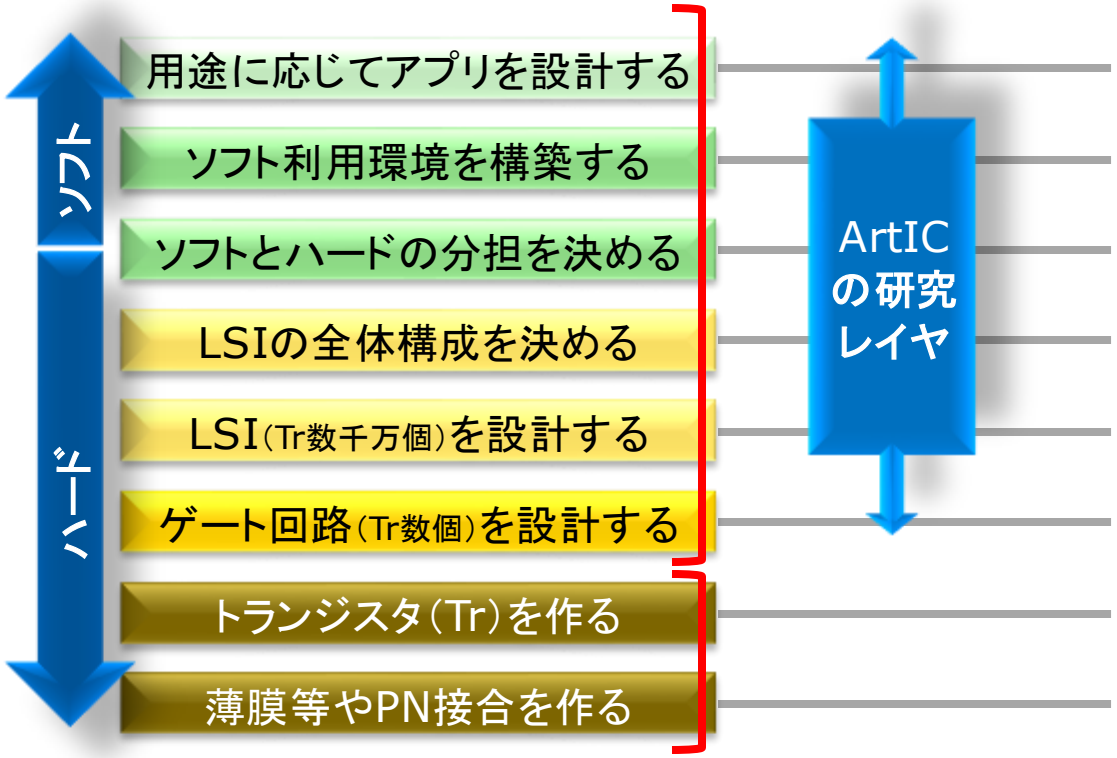
ユニット一体運営

Thiem van Chu,
川村一志　のお二人が転出済み

- 機械学習アルゴリズムや，計算機アーキテクチャ，ハードウェア設計などに興味を持つ皆さんのArtICへの参加を歓迎します
- 二研究室で共同運営しており，研究ユニット内に垣根はありません
  - どちらを志望しても全く差はありません
  - ArtIC教員全員で協力して，丁寧な指導と居心地よい環境づくりとを心掛けています

研究対象の技術レイヤ

技術レイヤ



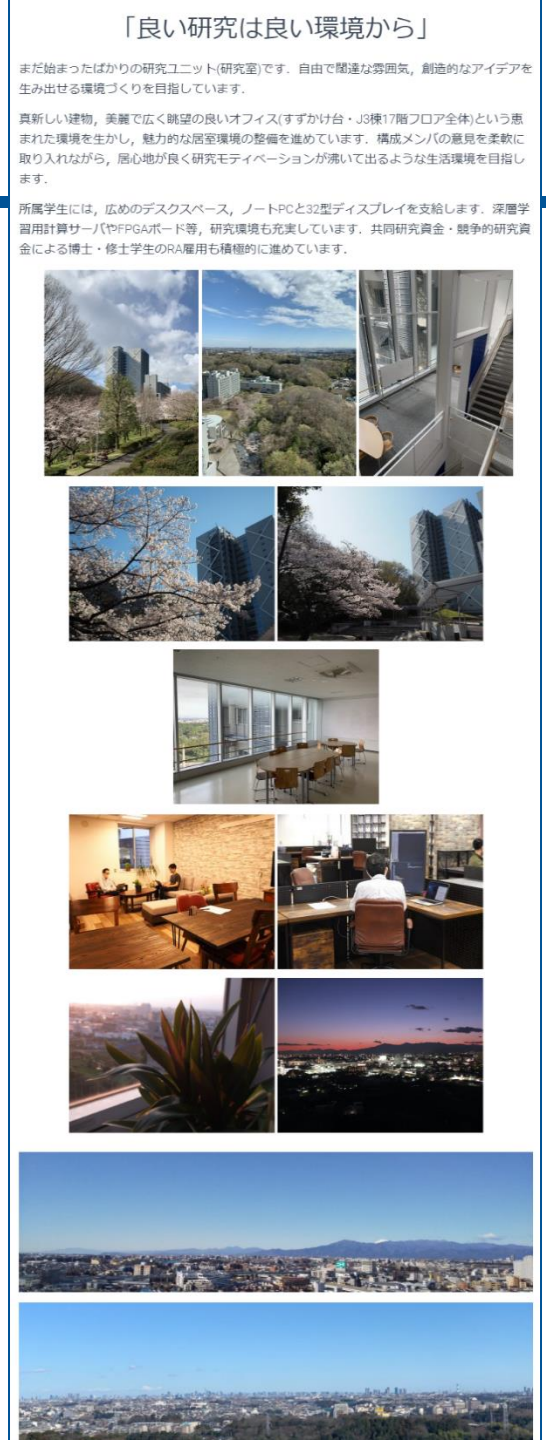
身に付くスキル・知識（例）

SW-HWの両方に興味がある人向き

研究室生活

ArtICホームページ =>

- 研究に集中できるよう、学部4年生から**全員RA**雇用しています
- 研究環境(オフィス，計算機，実験評価)の整備には力を入れています
- 導入教育 (～3月)
  - 輪講: アーキテクチャ，機械学習
  - 実習: ディープラーニング，FPGA
- 研究テーマ配属
  - 4年生 4月頃。本人の希望に応じてテーマ調整
- 研究の進め方
  - 全体会議: 週1回 (バーチャル)
  - 4グループ毎の研究報告: 週1回 (バーチャル)
  - その他，適宜個別に打合せ・議論 (対面)
  - コアタイムは設けていません
- **修士の6割(過去実績)が博士進学**、4割が企業就職
  博士の大半は企業研究職に就職



修論・卒論 タイトル名 (2024年度)

- **修士論文**
  - 制約付き組合せ最適化問題の解探索を高効率に行うアニーリングプロセッサ、兵藤
  - 説明の後方互換性を考慮した勾配ブースティング決定木の再訓練法、山倉
  - 強い宝くじ仮説とグラフ分割を用いたGNN 処理効率化の研究、伊藤
  - 乱数部分ネットワーク内におけるコンパクトな強い宝くじの研究、大塚
  - 平均場近似の高精度化に基づく高性能な二次無制約二値最適化手法、黒木
  - エッジデバイス向けのニューラルネットワーク圧縮手法、塩田
- **卒業論文**
  - 鍵共有不要な可変長データPIR、中森
  - PIMによる動的量子化を用いた大規模言語モデル推論の効率化、松島
  - ニューラルネットワークのエッジ向け継続学習手法、石橋

修論タイトル名 (2023年度)

- **修士論文**
  - A Highly Accurate and Parallel Vision MLP FPGA Accelerator based on FP7/8 SIMD Operations and efficient dataflow design、安永
  - イジング計算機を対象とした動的パラメータ調整の研究、井上
  - バケッティング・データ構造による自己位置推定機構のメモリ削減及び高速化、市川
  - 負荷均等配分を目指した高並列疎行列積アーキテクチャの研究、永原
  - 局所鋭敏性ハッシング機構を用いた超次元コンピューティングの研究、渡邊

修論・卒論 タイトル名 (2022年度)

- **修士論文**
  - 全並列アニーリングの解探索性能を向上させる動的なスピン反転機構の研究、小此木
  - 局所解脱出を容易にするアニーリング手法とそのアクセラレータ設計、神保
  - 乱数重みニューラルネットワークにおける精度・サイズトレードオフの向上に関する研究、大越
  - 高効率な量子化決定森推論アクセラレータのためのモデル最適化手法の研究、北島
- **卒業論文**
  - 強い宝くじ仮説に基づく超軽量物体検出ネットワーク、大塚
  - 同変性ネットワークに基づく自律走行向け強化学習手法、塩田
  - 表形式データを対象とした決定木とニューラルネットワークの融合型機械学習手法の研究、山倉
  - 2スピン同時フリップを並列試行可能なシミュレーテッド アニーリング手法、兵頭
  - 組合せ最適化問題のアニーリング解法に関する難易度評価、四元

最後に…

- ArtICホームページを確認ください　(東工大 ArtIC)
- 機械学習アルゴリズムや，計算機アーキテクチャ，ハードウェア設計に興味を持つ皆さんのArtICへの参加を歓迎します
  - 特別な知識は求めません。この分野の研究に対する意欲を期待します
  - 4年生前半までに基礎知識が身に付くよう、輪講や研修を行います
- 二研究室で共同運営しており，研究ユニット内に垣根はありません
  - どちらを志望しても全く差はありません
  - ArtIC教員全員で協力して，丁寧な指導と居心地よい環境づくりとを心掛けています
- 一線級の国際会議で発表できるグループです
- 実戦的な研究活動を主体としています
- 実社会で役立つ考え方・スキル・知識を身に着けることができます
- 産学連携、大学間連携，国家プロジェクト参画を活発に進めています


J2
J3

フロア案内
FLOOR INFORMATION

17F

1703	共通事務室1
情報工学系 Department of Computer Science	
山村研究室 Yamamura Lab.	
1706	教員室(山村) Professor M.Yamamura
1707	研究室
1710	学生室
小野 功研究室 Isao Ono Lab.	
1704	教員室(小野 功) Associate Professor Isao Ono
1705	学生室
1709	研究室
科学技術創成研究院 Institute of Innovative Research	
AIコンピューティング研究ユニット AI Computing Unit	
1711	ユニット控室
1712	学生室
1713	教員室(本村) Professor M.Motomura
1714	教員室(劉) Associate Professor J.Yu
1715	秘書室・会議室
1716	教員・研究員室


1711-171
1713·1714·1715
715
ARTIC


1711-1716
1713・1714・1715
1715
ArtIC
秘書室
押


ニトリ


ANSEL ADAMS