

AIコンピューティング研究ユニット の設立趣旨

2019/04/23

東京工業大学 科学技術創成研究院

ユニットリーダー

本村 真人

(2019年3月まで北大に所属)



Tokyo Tech

今後 数10年: コンピューティングの新時代

AIコンピューティング = 情報処理ハードウェアの革新

人工知能革命の急進

||

ポストノイマン時代

情報処理ハードウェアに
変化のチャンスが到来

- Pull -

ムーアの法則の終焉

||

ポストムーア時代

情報処理ハードウェアが
変化せざるを得ない

- Push -

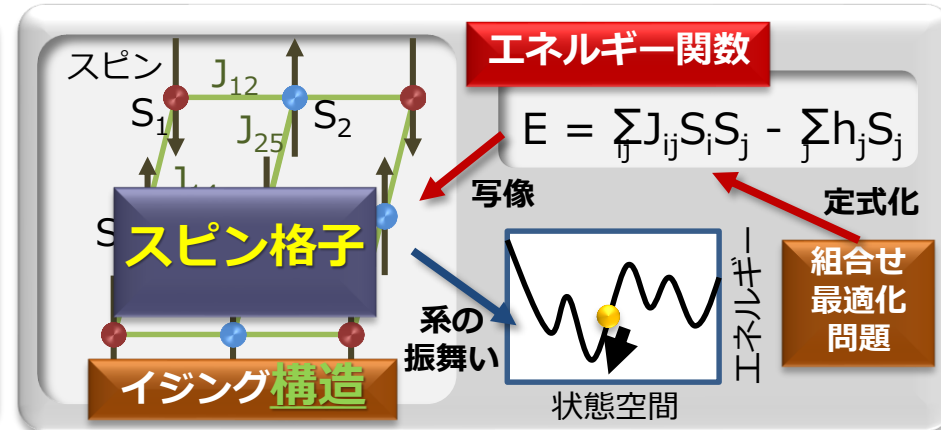
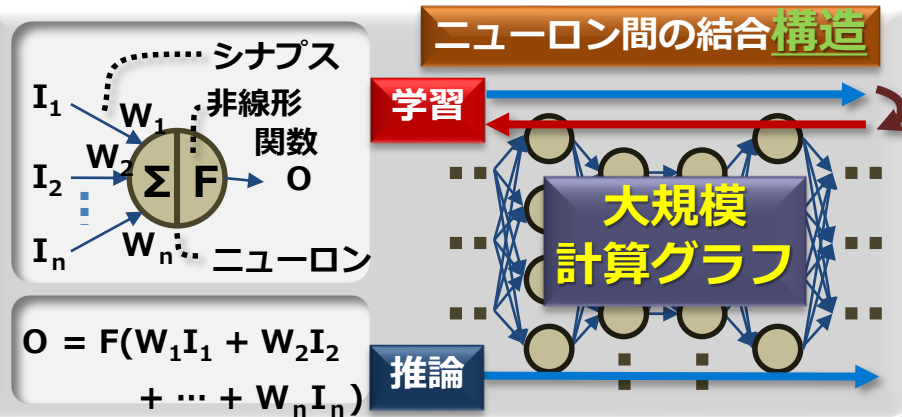
異なるメカニズム、同一の方向性

AIコンピューティング

=> 構造型情報処理アーキテクチャ

深層ニューラルネット(DNN)

アニーリング計算

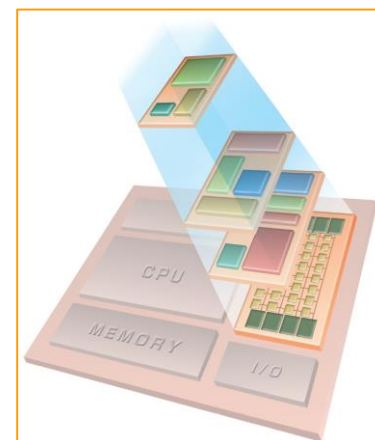


大規模、静的、単純演算ノードからなるデータ処理「構造」

構造の空間展開 (HWネイティブ並列)

構造の時間的切替 (HWの仮想化)

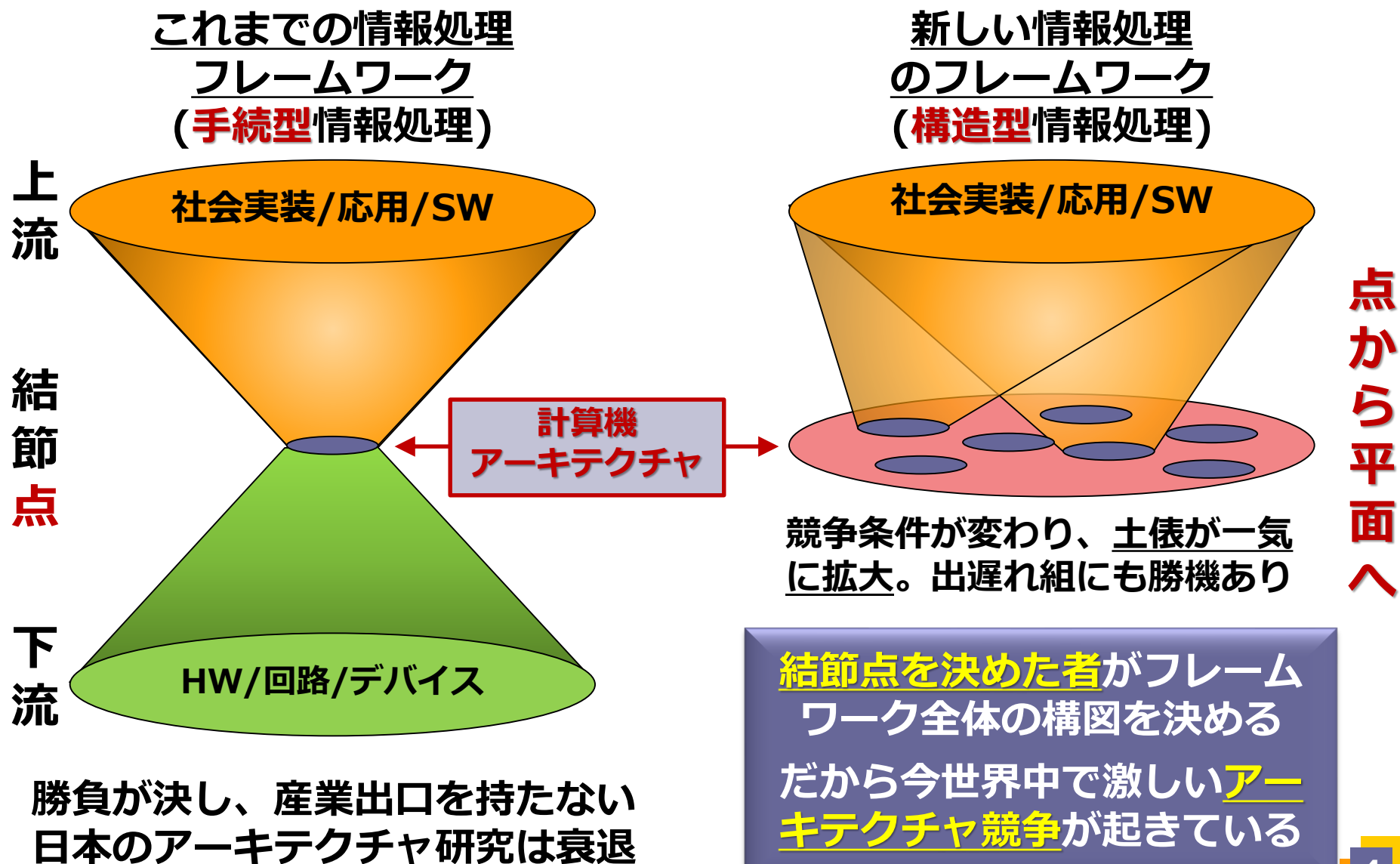
構造型情報処理アーキテクチャ



一定サイズの
HWリソースで
任意サイズの
問題を情報処理

= リンコンフィギュラブルハードウェア

計算機アーキテクチャ研究の黄金期



AIチップ：世界の状況を見ると

□ トップ国際会議 ISSCC及びVLSI シンポジウムでのAIチップ関連の発表（'16～'18）

- 殆どは大学・公的研究機関
- 米国が圧倒的な強さを発揮
- 韓国はKAIST 1 極、台湾は国立大学 3 極
- 中国勃興の兆し
- 日本の存在感低し

本村ユニットリーダーの
北大での実績に関しては
P13-14を参照

ISSCC

米国	欧州	韓国	台湾	中国	日本
14	2	8	4	0	1

KAIST

NCTU
NTU
NTHU

北大

VLSIシンポジウム

米国	欧州	韓国	台湾	中国	日本
10	1	2	3	4	3

清華大

北大
ルネ
東芝

日本の産官学はどう取り組むべきか

AIコンピューティング: 必要な方策

方策1: AIアーキテクチャ-プロジェクト

- システムアウェアなAI-HW研究
- 産学連携プロジェクト群で応用-アルゴリズム-アーキテクチャ縦断エコシステムの形成

方策2: AIチップ-ファーム

- 先進アーキテクチャを低コスト・短TATでHW化・実証するためのLSI設計試作環境の構築
- システムアウェアHWプレイヤーの育成・放牧

方策3: AIコンピューティング-センター

- 既存のAIセンターはAIアルゴリズムが中心
- システムアウェアAI-HWの**中核拠点**を形成

東工大に本研究ユニットを新設する狙い

経産省・NEDO

文科省・JST

文科省・JST

経産省・NEDO

高効率・高速処理を可能とするAIチップ・次世代コンピューティングの技術開発事業 平成30年度概算事業額 100.0億円 (新規)

事業目的・目標

- ① AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。
- ② AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。

事業イメージ

① AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。

科学技術振興機構 平成30年度 JST戦略的創造推進事業(CREST) 「コンピューティング基盤」提案募集 (第1期)

Society5.0を支える 革新的コンピューティング技術

研究総括 坂井 修一 (さかい しゅういち) 東京大学 情報理工学系研究科 教授 2018年4月12日、25日(東京)

本領域の研究ターゲット

高性能化 (高スループット、低レイテンシ)、低コスト化

低消費電力化、低消費エネルギー化

信頼性、セキュリティ、プライバシー

主要な技術的課題: 計算機・メモリ・ストレージ・ネットワーク・セキュリティ・プライバシー・信頼性・低消費電力化、低消費エネルギー化

手段: スーパーコンピュータ、SIMD命令、GPU、AIチップ、量子計算、ニューロモρφic計算

量的イノベーションから質的イノベーションへ

1980 1990 2000 2010 2020 2030 2040 2050

微細化の限界

持続可能なスマート社会のための情報基盤

AIチップ開発加速のためのイノベーション推進事業 平成30年度概算事業額 26.0億円 (新規)

事業目的・目標

- ① AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。
- ② AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。

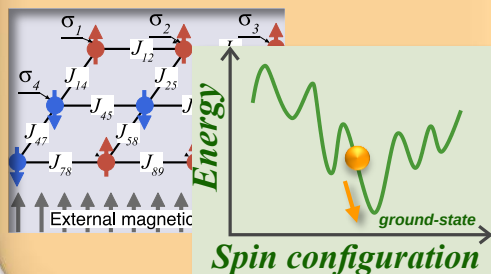
事業イメージ

① AIチップの開発・実証・評価を促進するため、産学官連携によるAIチップの開発・実証・評価を促進する。

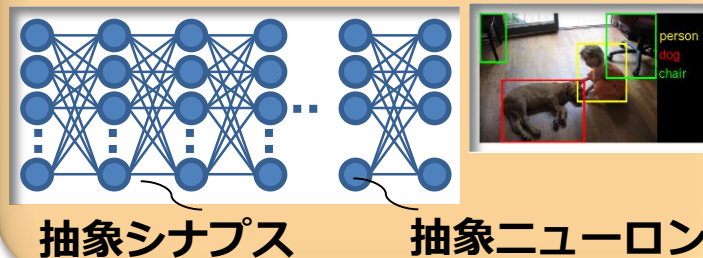
本ユニットの研究ターゲット (1)

人工知能(AI)応用の急速な拡大
「**コントロール**駆動から**データ**駆動へ」
計算機アーキテクチャの革命

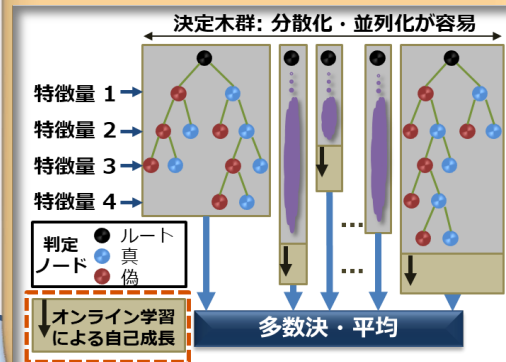
組合せ最適化問題 →
スピン格子のエネルギー最小化問題



大量データの学習 →
強力な推論・識別・予測能力



説明性・制御性の高さと
低学習負荷の両立



深層ニューラルネット
・ディープラーニング

アニーリング計算機
(非量子)

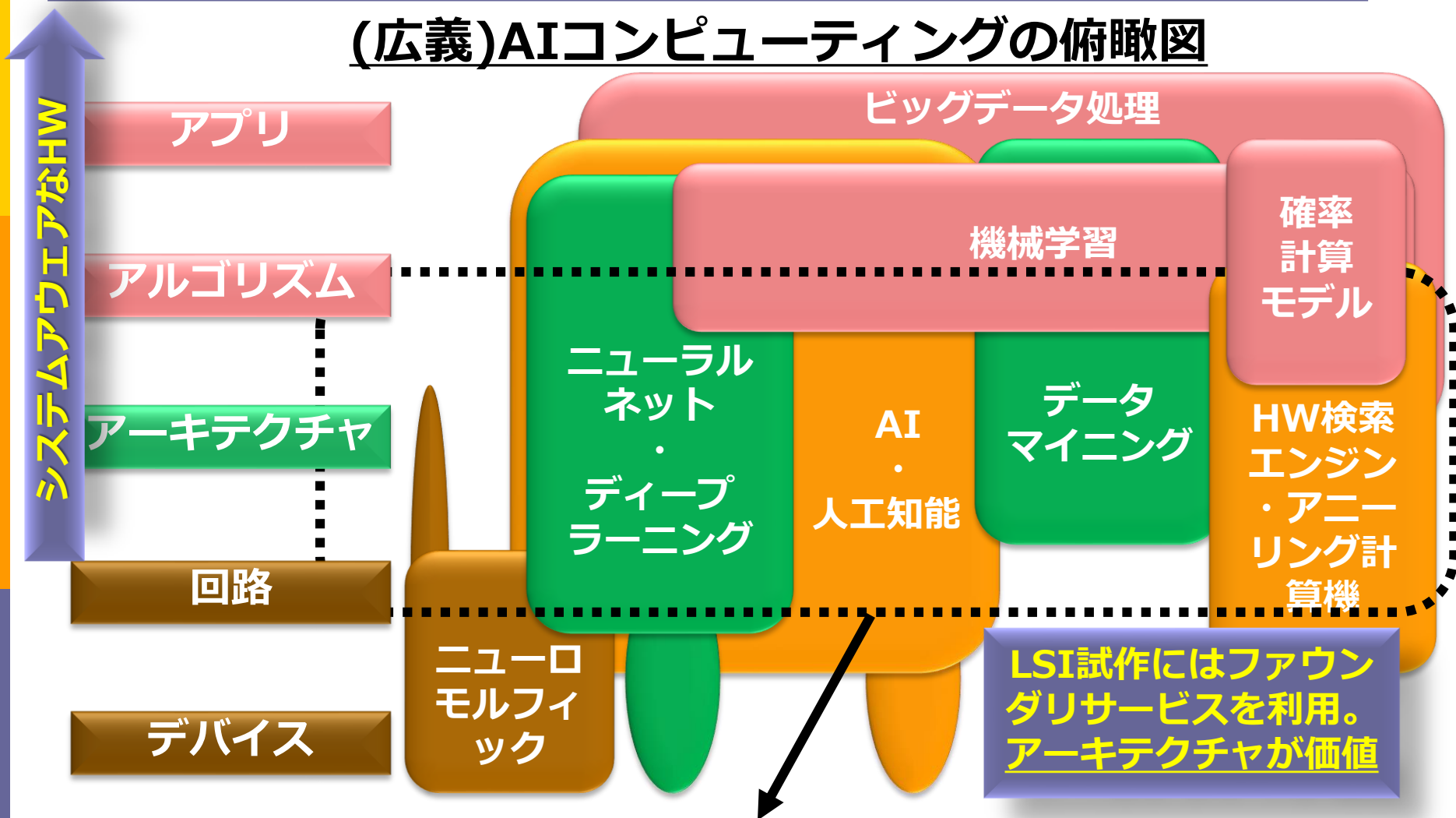
統計的機械学習
(アンサンブル学習等)

構造型情報処理アーキテクチャ
として共通基盤化

アルゴリズム理解 ⇒ アーキテクチャ研究 ⇒ ハードウェア実現
- FPGA実装評価及び本格的なLSIの研究開発 -

本ユニットの研究ターゲット (2)

(広義)AIコンピューティングの俯瞰図



これら広範囲のAIコンピューティング群を加速する
アーキテクチャプラットフォームの研究を推進する

体制構築・研究推進シナリオ（発足時点）

項目	年度	2019	2020	2021	2022	2023	2024年以降
研究ユニット 体制構築		拠点整備	研究活動の拡大				
		准教授・ 助教採用	国内外連携・拠点活動				
		研究スタッフ採用	研究スタッフ拡充				
科研費 基盤(S) ・ 計算機システム分野 研究 (代表)		知能コンピューティングを加速 する自己学習型・革新的アーキ テクチャ基盤技術の創出					SW-HW融合 型研究の深耕 により、より 大きな研究展 開を狙う
NEDO 革新的AIエッ ジ事業 研究 (分担)		動的再構成技術を活用した組み 込みAIシステムの研究開発					
JSTコンピューティン グ基盤 CREST 研究 (代表)		学習/数理モデルに基づく時空間展開型 アーキテクチャの創出と応用					

補足: CREST研究構想の概略

研究体制・研究テーマ

Society 5.0を支える
革新的コンピューティング

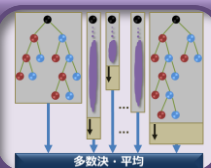
機械学習G

分散・信頼性志向機械学習理論

機械学習応用展開

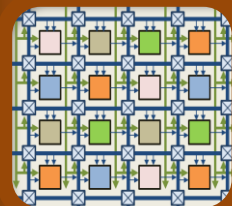
アーキテクチャG

アンサンブル・
オンライン学習
エッジ
アクセラレータ

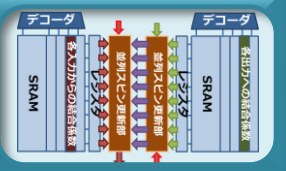


多数決・平均

時空間展開型
エッジ知能
コンピュータ



多目的
エネルギー
最小化
エンジン



社会応用アルゴリズムG

アニーリング計算モデル・実問題適用

最適化問題応用展開

トポロ
ジー理論

統計物
理理論

数理科学G

センサデータをリ
アルタイム&自動
で認識

自発学習-自律的な
機能の獲得

制約充足型問題の
最適解を瞬時に
導出

安全性・信頼性・プライバシー
の担保

低エネルギー・低コストの実現

前半3年間

後半2.5年間

Pj 期間終了後、更に展開

達成目標: アンサンブル・オンライン学習エッジアクセラレータと多目的エネルギー最小化エンジンの技術確立

達成目標: 時空間展開型・エッジ知能コンピュータにより、最先端ソフトウェア(SW)解に比べて二桁のエネルギー効率向上を実現

まとめ

今後も発展を続ける**AIコンピューティング**(DNN、アンサンブル学習、アニーリングマシン、グラフ処理、ニューロモルフィックHW等)を支えるアーキテクチャ**プラットフォーム**の創出を目指す。

DNNとアンサンブル学習、DNNとニューロモルフィック等、複数処理にまたがる**共通基盤コンピューティングアーキテクチャ**を構築する。

構造型情報処理に適した**リコンフィギュラブル**アーキテクチャ、知能コンピューティングの学習メカニズムを支える**エネルギー最小化原理**を効率よく処理するハードウェア技術等の研究を推進する。

以て、現在の人工知能よりも格段に**自律性・安全性・エネルギー効率・コスト効率**が高い知能コンピューティングを支える**ハードウェア基盤実現**を目指す。

日本発の技術として、**FPGAを凌駕・置換する次世代HW基盤**を実現するとともに、**上位システムレイヤとの積極的な連携**を通じてSW分野の世界的エコシステムにも積極的に参画する。

2019年4月
に発足
(北大から転籍)
2020年4月
に本格稼働

発足に合わせて准教授・
助教ポストを公募。任期
は10年(5年x2)。業績に
より任期なしに転任

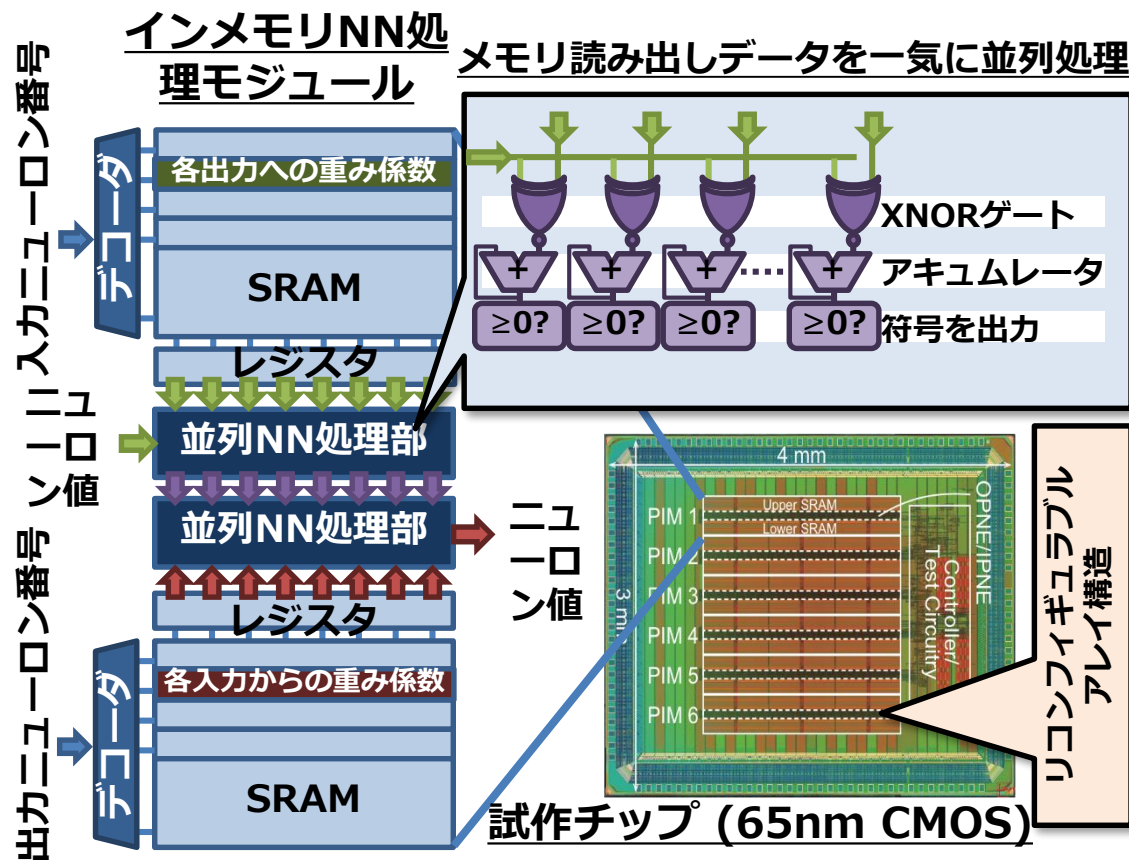
すずかけ台
キャンパス
J3棟 17F



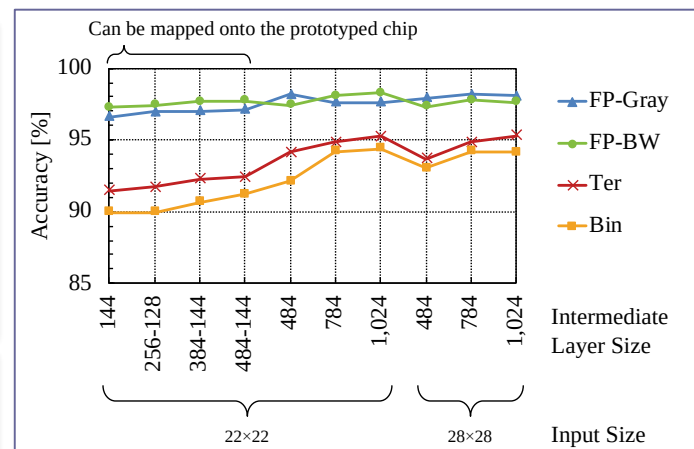
ユニットリーダー(本村教授) 情報

- 前所属: 北海道大学 情報科学研究科 <http://lalsie.ist.hokudai.ac.jp/jp/>
- 講演スライド例 (公開版)
 - NEDOセミナー講演: 「AIコンピューティングがアーキテクチャにもたらすもの」 (2018/10/31) <https://www.nedo.go.jp/content/100885737.pdf>
 - CRDSシンポジウム講演: 「AI応用がもたらすプロセッサLSIのゲームチェンジ」 (2017/03/07) https://www.jst.go.jp/crds/sympo/20170307/pdf/20170307_06.pdf
- 略歴: '87年京都大学理学部修士、'96年同博士(工学)。'87年よりNECにてリコンフィギュラブルハードウェア、オンチップマルチプロセッサ等の研究開発と事業化に従事。'92年MIT客員研究員。'11年より北海道大学教授。リコンフィギュラブルアーキテクチャ/人工知能向けハードウェアアーキテクチャの研究などに従事。'19年4月に東工大に着任。IEICE/IPSJ/IEEEに所属。'92年IEEE JSSC Best Paper Award、'99年IPSJ年間最優秀論文、'11年IEICE業績賞、'18年ISSCC Silkroad Awardを各受賞。
- 東工大の関連情報
 - 科学技術創成研究院 (AIコンピューティング研究ユニットが所属-本籍)
 - <https://www.iir.titech.ac.jp/2019/04/10/info-76/>
 - 未来産業技術研究所 (研究室として所属)
 - http://www.first.iir.titech.ac.jp/news/2019/detail_593.html
 - <http://www.first.iir.titech.ac.jp/member/core2.html>
 - 情報通信系・コース (教育を担当: 採用准教授・助教の担当系・コースは未定)
 - https://educ.titech.ac.jp/ict/faculty/research_lab/
 - <https://educ.titech.ac.jp/ict/publications/file/pamphlet2019j.pdf>

補足: 北大での研究実績 (1) BRein Memory



認識精度評価



13 layers, 4.2K neurons, and 0.8M Synapses, 1.4TOPS at 0.6W

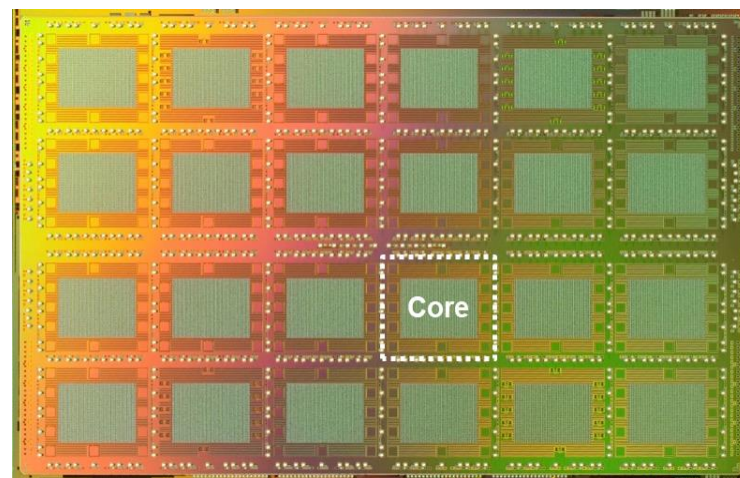
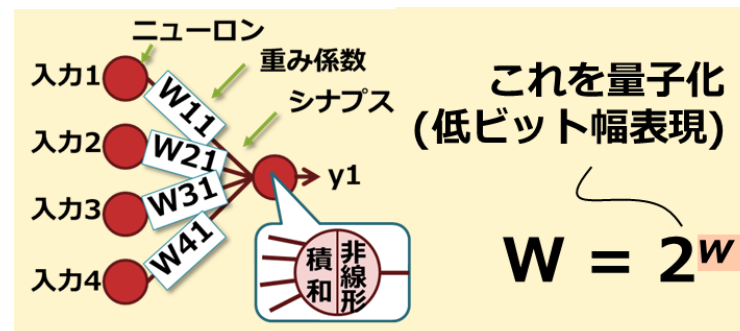
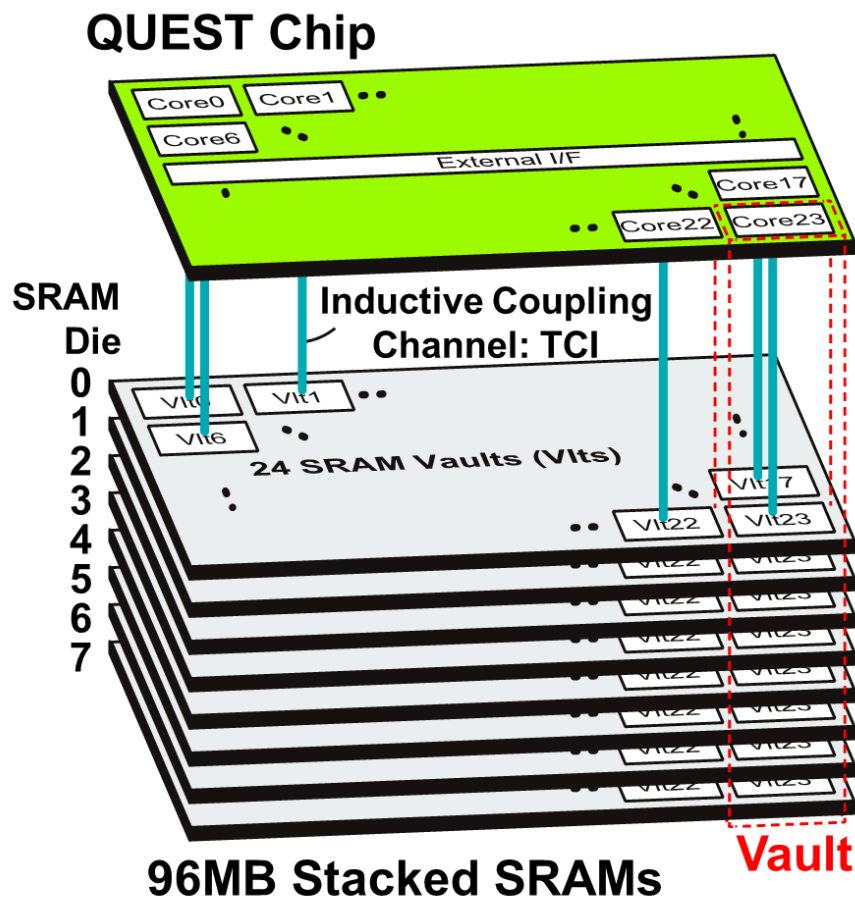
VLSIシンポジウム
で発表 (17年6月)

CPU等との比較 (手書き文字認識)

	BRein Memory	FPGA	GPU	CPU
エネルギー効率 対CPU相対比	2.6万	33	9	1

世界初の二値化・三値化 DNNチップ。GPU比3K倍のエネルギー効率を実証。日経新聞等掲載

補足: 北大での研究実績 (2) QUEST



- **対数量子化**近似DNN推論エンジン
- **枝刈りによる不規則 & スパース NNがターゲット**
- 7.49TOPS(バイナリ時)

ISSCCで発表 (18年2月)
世界初の対数量子化・3次元積層DNNチップ。Silkroad Award受賞。日経新聞等掲載