AIチップ：世界の研究動向と東工大の研究戦略

2019年 10月 11日

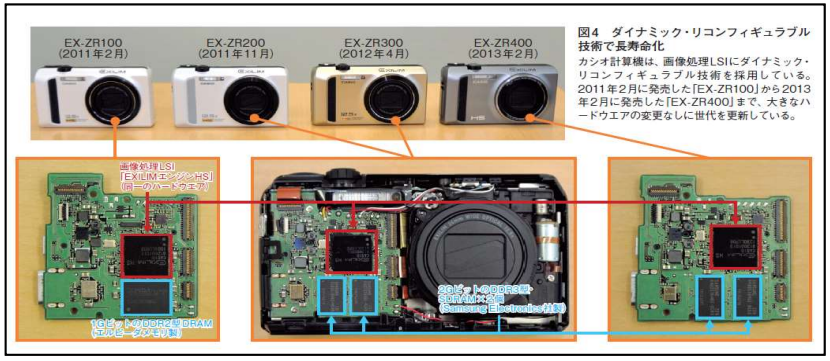
東京工業大学　科学技術創成研究院
AIコンピューティング研究ユニット (ArtIC)
本村 真人

Tokyo Tech

ARTIC

講演者の自己紹介

- 2011年3月まで、NECにおいてLSIアーキテクチャに関する研究開発と事業化に従事
  - 並列プロセッサ、メモリ-ロジック密結合型LSI、etc.
  - **動的再構成プロセッサ(DRP)**の研究開発と事業化
    - 入り口から出口まで: デジカメに搭載
    - **ルネサスエレ社が事業拡大中**
      ・AI処理向け技術開発中
  - ’91-92: MIT客員研究員
    - **データフローアーキテクチャ**

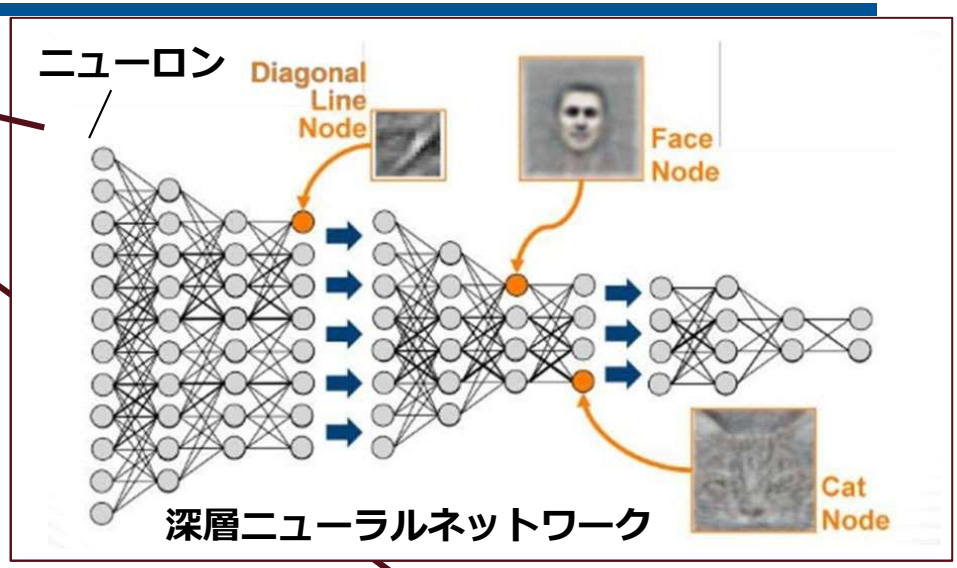


- 2011年4月に北大に転任
  - 人工知能処理向けLSI、リコンフィギュラブルHW技術などの研究

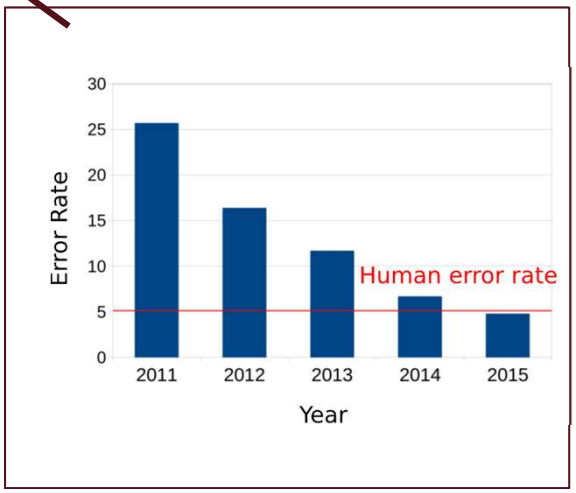
- 2019年4月に東工大に転任
  - AIコンピューティング研究ユニット(略称 ArtIC)を立ち上げ中

「人工知能(AI)革命」の急進

- 2012年: Googleの猫
- 2015年: 人間以上の画像認識能力を達成
- 2016年: アルファ碁
- 2018年: 自動翻訳技術が実用レベルに到達



日経新聞 9/24記事　（オリジナルは情報通信機構のデータ)

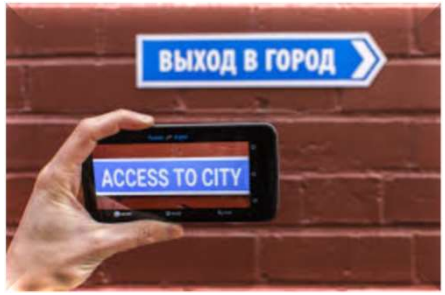


ImageNet コンペティション

人工知能、AI、ニューラルネット: 敬遠から熱狂へ大転換

AIが変える我々の暮らし

自動運転

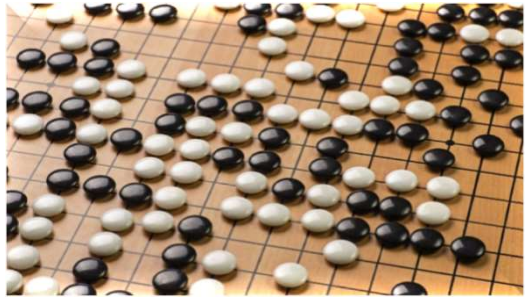


機械翻訳

自動通訳

ロボティクス

広告



問題解決・予測

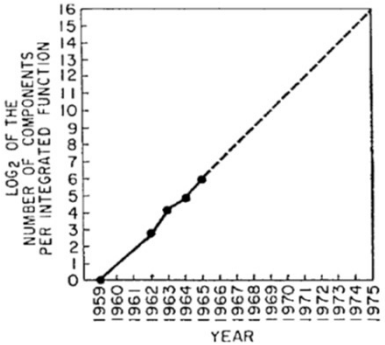
※ Google画像検索結果から写真をピックアップ

AI時代　:= AI処理ワークロードが爆発的に増大する時代

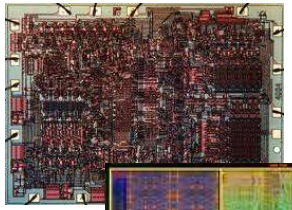
ビッグデータ ＋ AIコンピューティング

「ムーアの法則」とその終焉

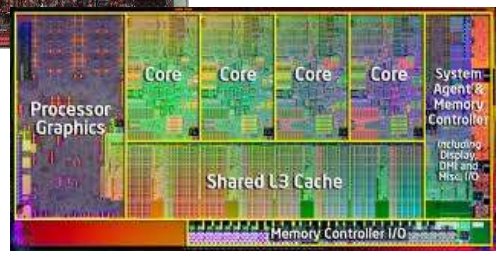
- 「ICの集積度は18か月で2倍になる」
  - インテル創設者ゴードン・ムーアが1965年予測
  - 50+年間、ICT社会の発展を支えた基本原則 (2^{35} 倍 = 160億倍)



Intel@1971

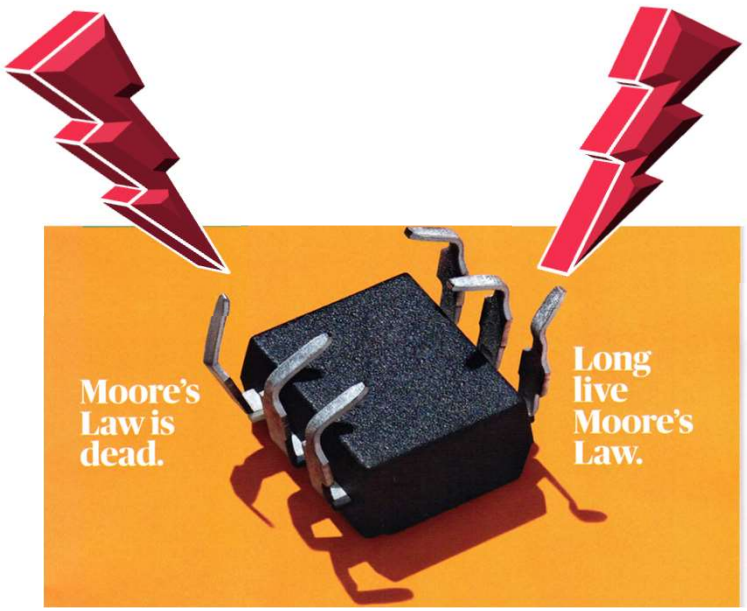


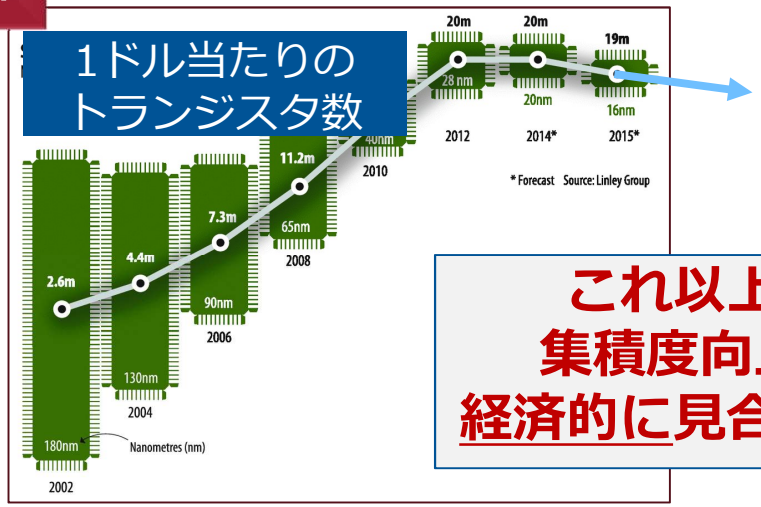
Intel@2012



電力限界

経済限界





これ以上の集積度向上は経済的に見合わない

What is Next Exponential ?

二大潮流のシンクロニシティ

情報処理ハードウェアの革新

人工知能革命の急進

II

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

II

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 -

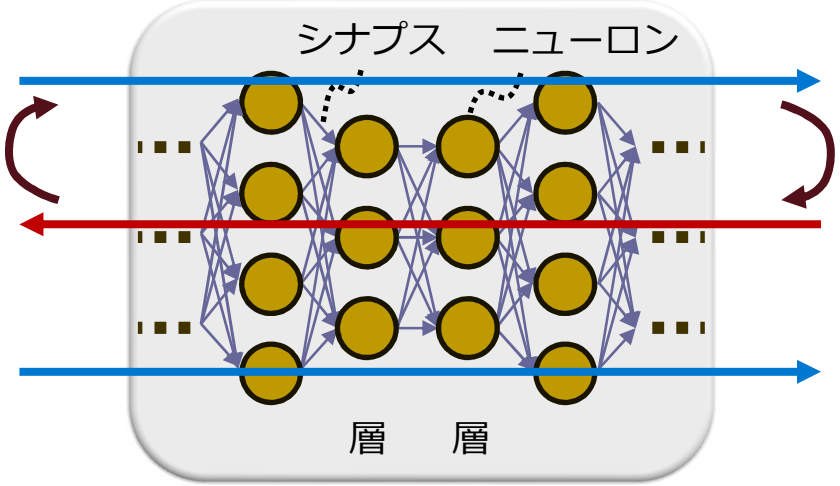
引く力　と　押す力

ディープラーニングの基本

深層(ディープ)ニューラルネットワーク

学習

推論



- **既知**データで**伝搬**と**逆伝搬**を反芻
- 所望の結果に向けてシナプス値調整

- **未知**データを**伝搬**させ結果を出力

正帰還ループにより
爆発的に発展

高速な計算ハードウェア

大量のデータ

オープン環境によるアルゴリズム開発

TensorFlow
Spark
Caffe
cuDNN
Chainer
arXiv.org

情報処理技術への影響: ソフトウェアの視点

ディープラーニングの本質 = 「誤差を最小にするように係数を少しづつ学習する作業を延々と繰り返すことで機能を獲得できる」という発見

– Before DL –

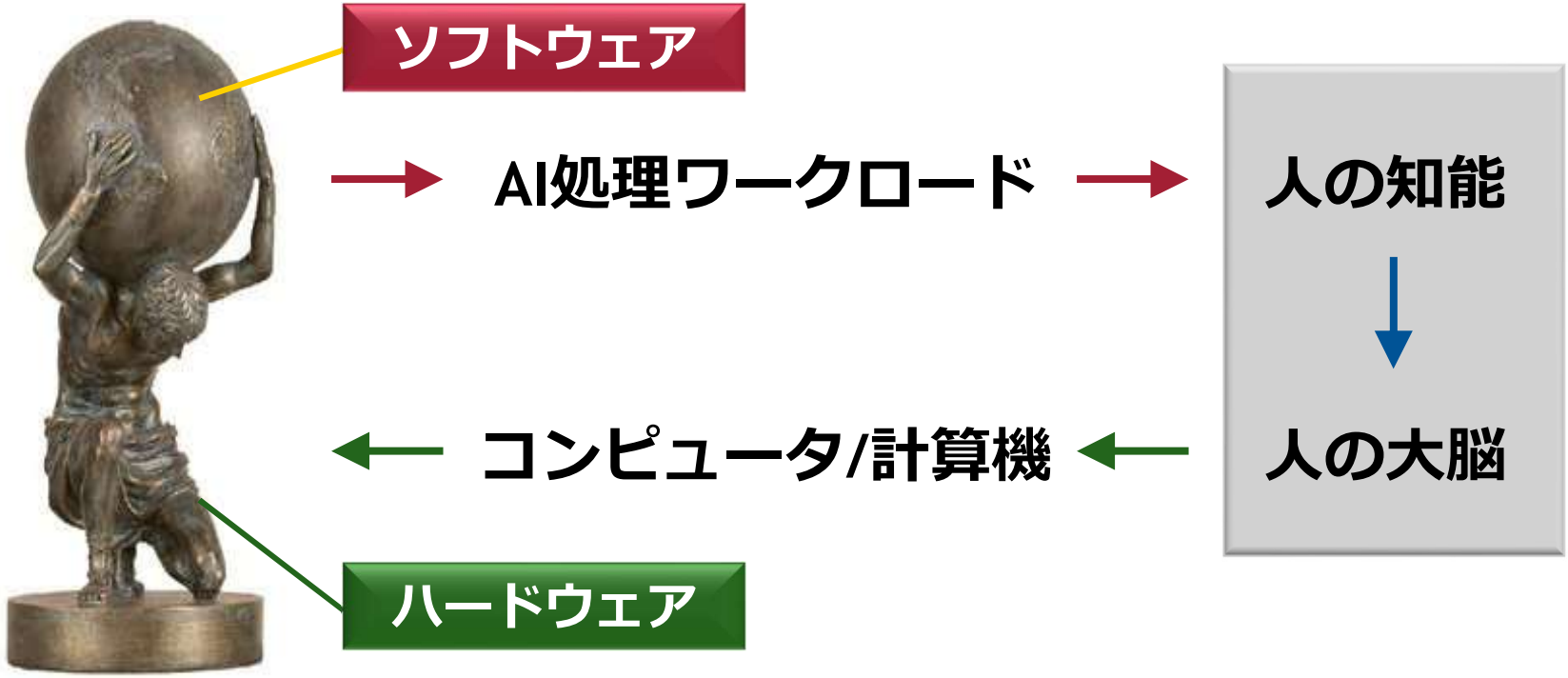
専門家の手による精緻な設計と組立

かなりの部分の処理がシフト

– After DL –

大量のデータが穿つことによる創発的造型

AI時代にハードウェアはどうあるべきか?



AIコンピューティングの命題

果たして何を計算するのか？
どう計算すればよいのか？

ワークロードと処理効率の相関

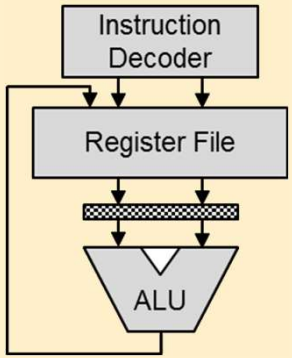
計算機と大脳のエネルギー効率比較

	高速計算機 (20万W)	大脳 (20W)
囲碁	仮想実行	ネイティブ実行

直観的・知能的な情報処理では、大脳の方が1万倍の高エネルギー効率

論理的・計算的な情報処理では、計算機の方が1億倍の高エネルギー効率

ハードウェアの基本構造



算術演算器

神経回路網

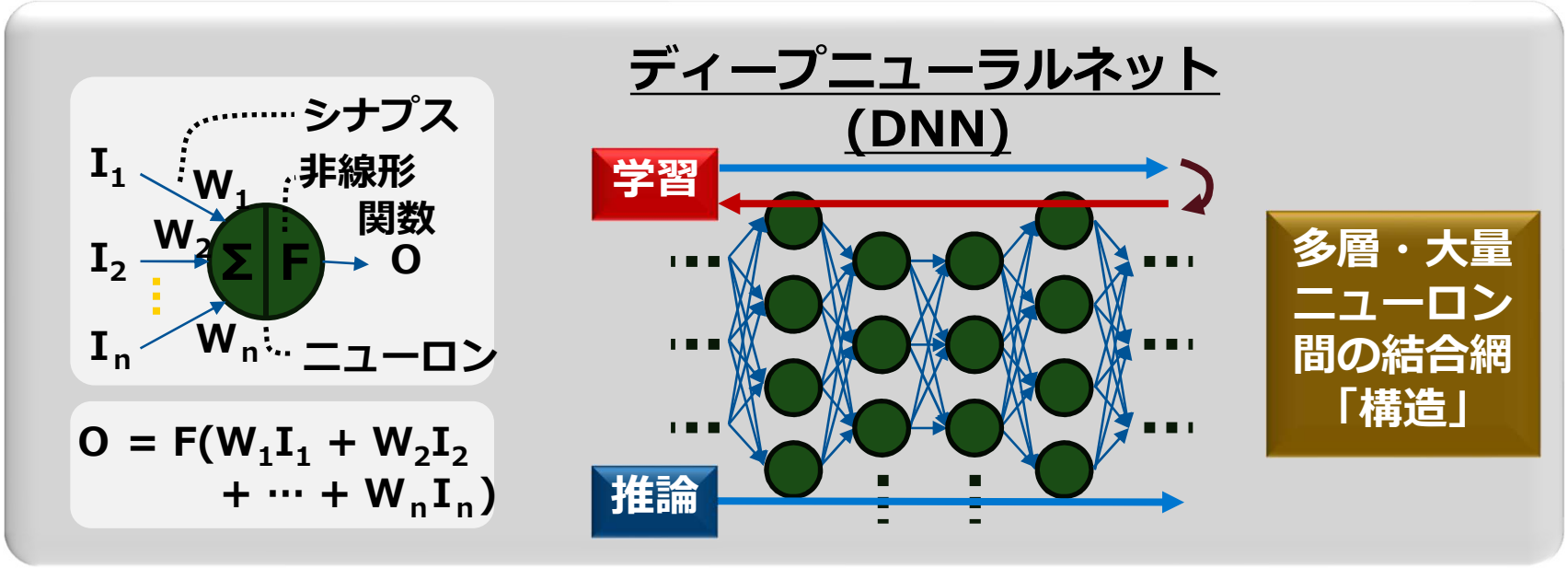
得意な情報処理領域ではそれぞれ相手を圧倒

↓

処理内容と処理構造の適合度合いが重要

AI処理ワークロードに寄り添うコンピュータが必要

AI処理 := 構造型情報処理



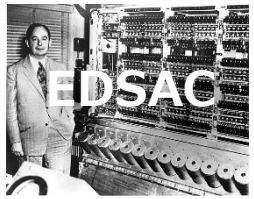
- 大規模なグラフ構造の並列処理
- DNN以外の多くのAI処理も同様に構造型
  - グラフ処理
  - 広い範囲の機械学習課題
    - データマイニング、分類・推定、ベイジアンネット、etc.
  - 組合せ最適化問題 (エネルギー関数の最小化で表現できる問題)

アーキテクチャ: 手続き型から構造型へ転換

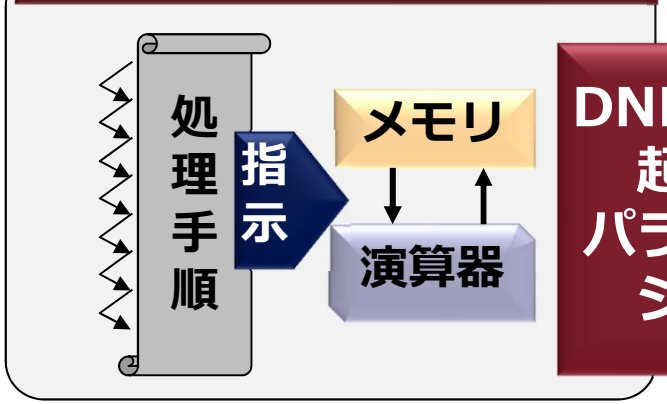
「手続き型」情報処理

「コントロールフロー型」
「フォンノイマン型」

1949



- 処理手順をプログラム
- 逐次ネイティブに実行



情報処理の王道

DNN興隆で起きるパラダイムシフト

「構造型」情報処理

「データフロー型」
「リコンフィギュラブルHW型」

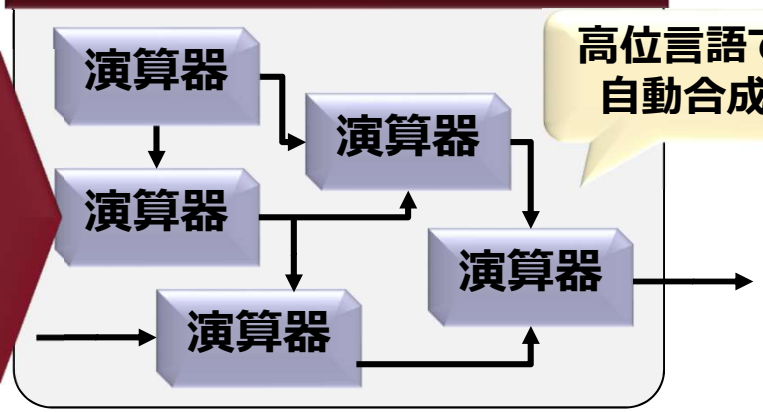
1946

接続を手動で切替

進化

- 処理構造をプログラム
- 並列ネイティブに実行

高位言語で自動合成



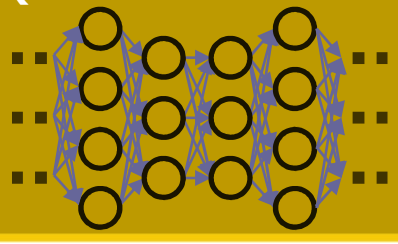
左脳 ← アナロジー → 右脳

情報処理ハードウェアへの影響

大規模データ

- ニューラルネット構造をハードウェアに写像
- 入力データを流し込み結果を抽出
- 膨大な単純演算=>大量の並列性が存在

AIコンピュータ（アクセラレータ）



メモリ

- 重み係数構造情報を格納
- プロセッサに直結
- 大量・高速の読み出し

有用な知見
予知予測、法則性、特異値、分類、etc.

高性能 & 低電力

データセンター

学習　推論

組込み機器

学習　推論

計算機アーキテクチャ研究の黄金期

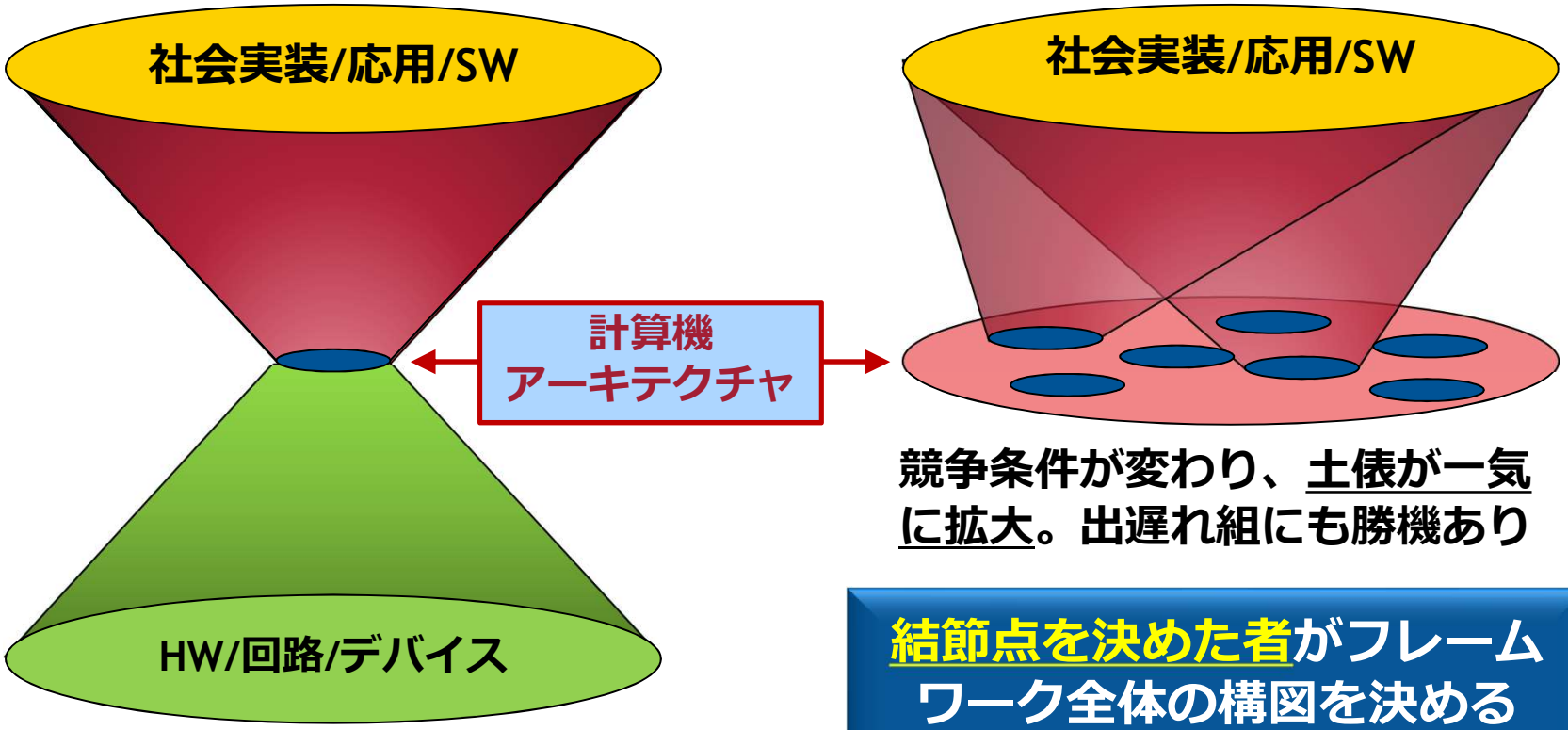
これまでの情報処理フレームワーク
(手続型情報処理)

新しいAIコンピューティングのフレームワーク
(構造型情報処理)

上流

結節点

下流



点から平面へ

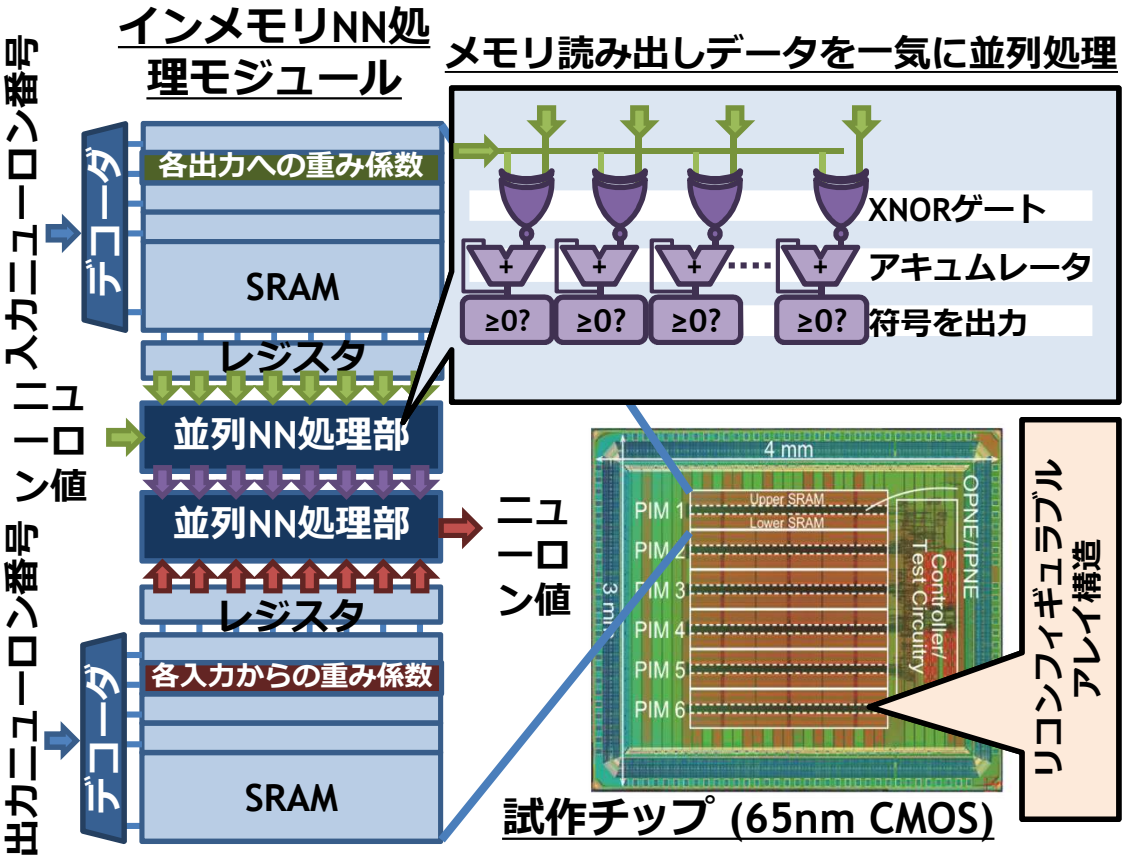
競争条件が変わり、土俵が一気に拡大。出遅れ組にも勝機あり

勝負が決し、産業出口を持たない日本のアーキテクチャ研究は衰退

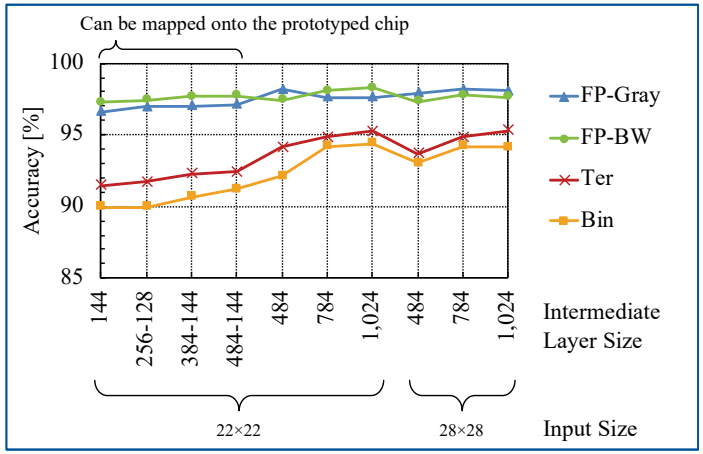
結節点を決めた者がフレームワーク全体の構図を決める

だから今世界中で激しいアーキテクチャ競争が起きている

実例-1: 本村G 二値化DNNチップ



認識精度評価



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

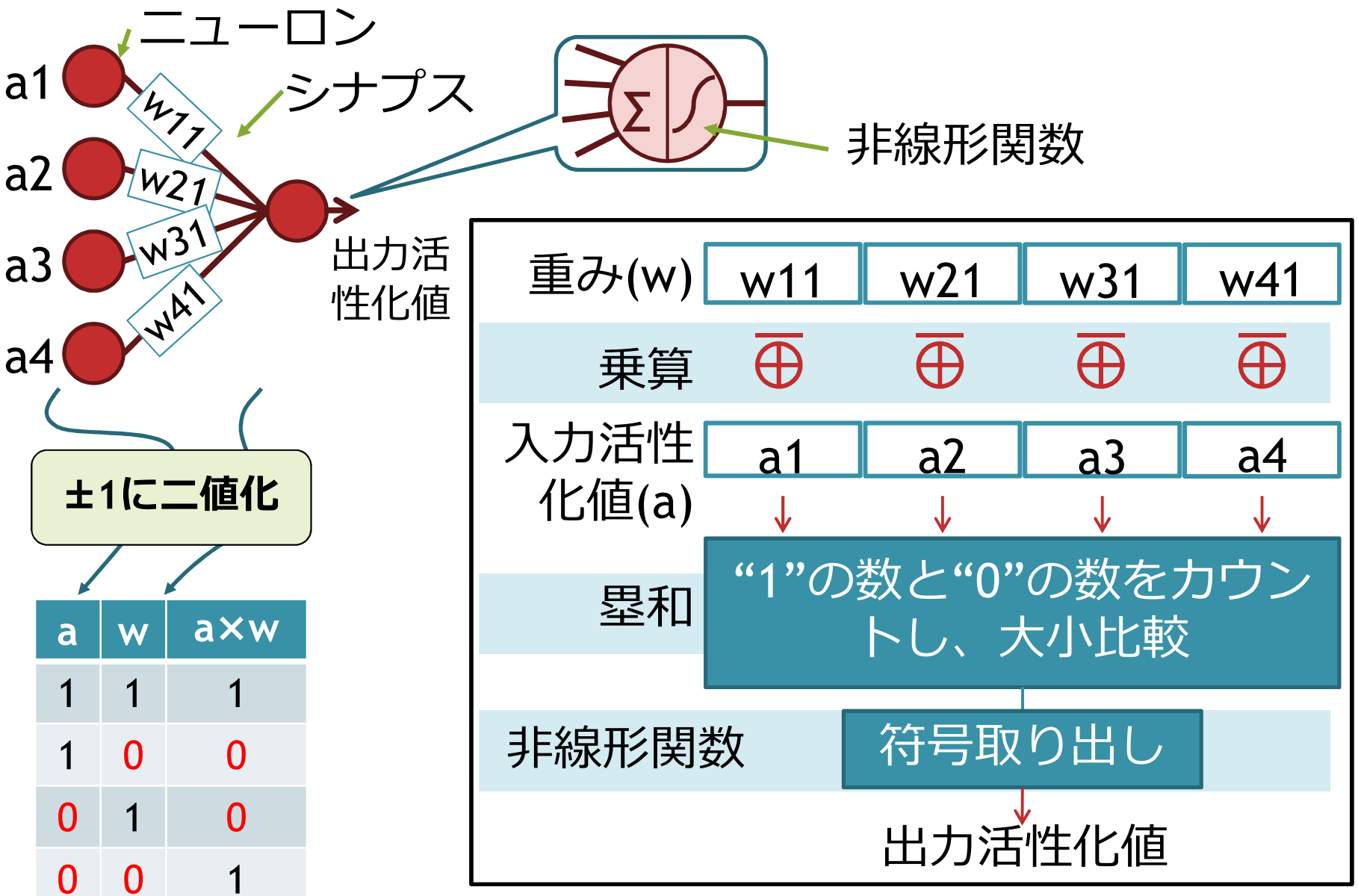
CPU等との比較 (手書き文字認識)

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

VLSIシンポジウムで発表 (17年6月)

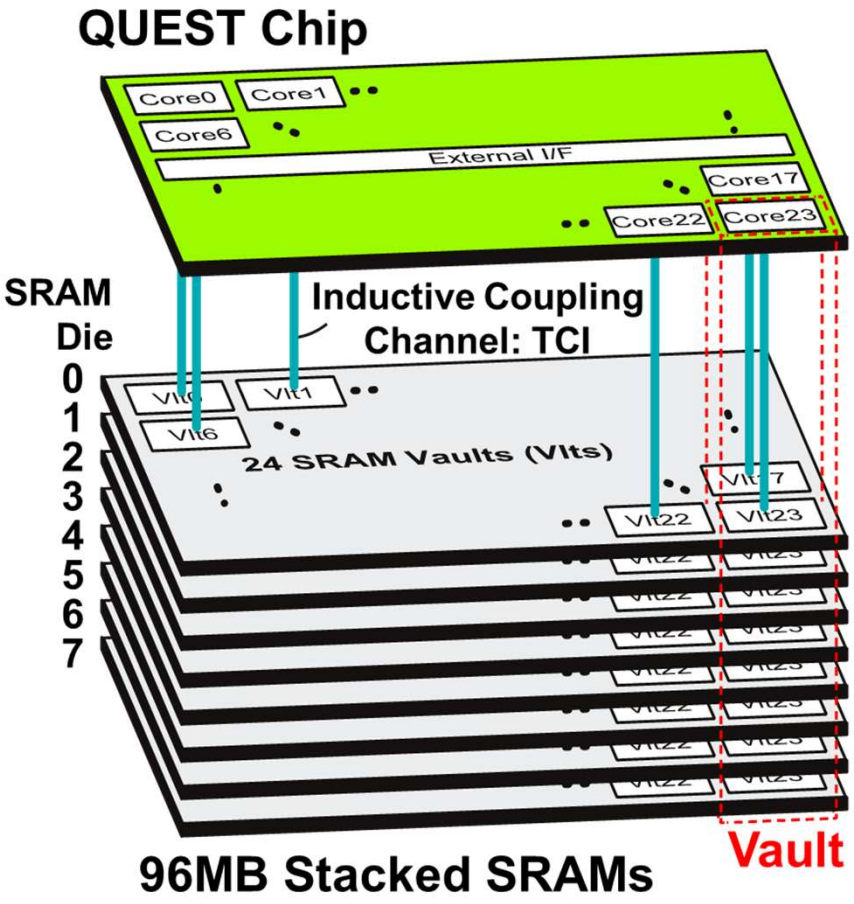
世界初の二値化・三値化DNNチップ。**GPU比3K倍**のエネルギー効率を実証。日経新聞等掲載

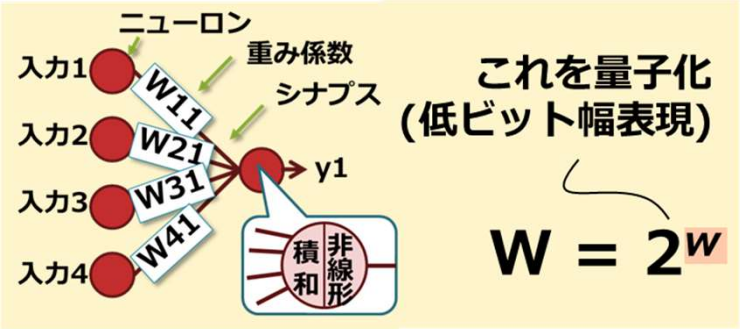
補足: 二値化ニューラルネットとは

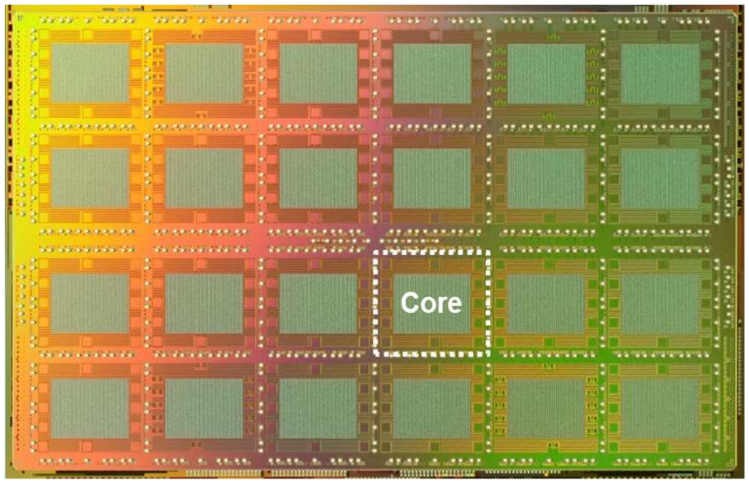


a	w	a×w
1	1	1
1	0	0
0	1	0
0	0	1

実例-2: 本村G 対数量子化-3次元DNNチップ







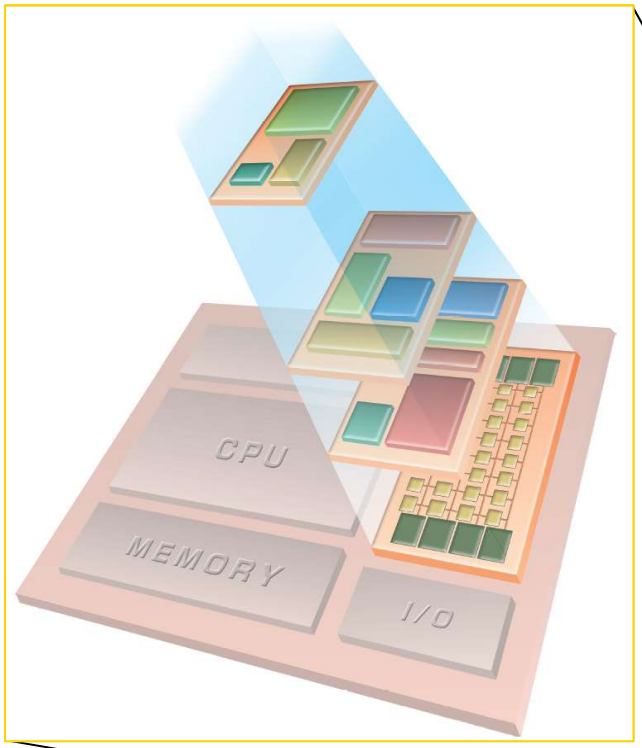
- **対数量子化**近似DNN推論エンジン
- **枝刈りによる不規則 & スパースNNがターゲット**
- 7.49TOPS(バイナリ時)

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

実例-3: ルネサスエレ社 DRP-AIチップ

DRP: 動的再構成プロセッサ

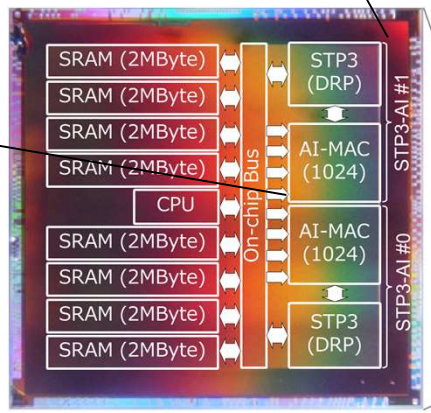


(2018年10月ルネサスエレ社広報)

VLSIシンポジウムで発表

(18年6月)

日本企業初のAIチップ学会発表(北大も共著で参加)。リコンフィギュラブルハードウェアAIチップの**世界的代表例**



- ユーザカスタマイズAI処理
- 実応用性能

を訴求

AIチップ: 世界の状況

- **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KAIST１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

ISSCC

米国	欧州	韓国	台湾	中国	日本
14	2	8	4	0	1

KAIST

NCTU
NTU
NTHU

本村G

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

本村G
ルネ
東芝

- **チップ設計試作支援の存在感**
  - **米国** DARPAのCRAFTプロジェクト
  - **台湾** CIC (Chip Impl. Center)
    - 両社ともTSMC16nm最先端CMOSによるチップ試作(~2億円)をフルサポート
      => **ただ同然で試作**

ISSCC2018: 機械学習分野の技術的特徴

	セッション名 →	NM.	Deep Learning and …				Ext.	Computation in Memory …				
	↓ 技術分類項目	7.4	13.2	13.3	13.4	13.5	21.2	31.1	31.2	31.3	31.4	31.5
対象ML技術	深層NN		✔	✔	✔	✔	✔	✔			✔	✔
	Binary		✔	✔		✔		✔			✔	✔
	量子化		Log	線形								
	その他ML	NM							SVM	HD		
	特徴	学習	汎用	汎用					学習	学習		
メモリ	インメモリ					✔		✔	✔		✔	✔
	メモリ積層		磁界									
回路/Tech.	Mixed-Signal	✔				✔	✔	✔		✔	✔	✔
	Non CMOS									CNT	ReR	
	発表機関	Georgia	北大	KAIST	KAIST	Stanford	Columbia	MIT	Illinoi	Stanford	NTHU	NTHU

凡例

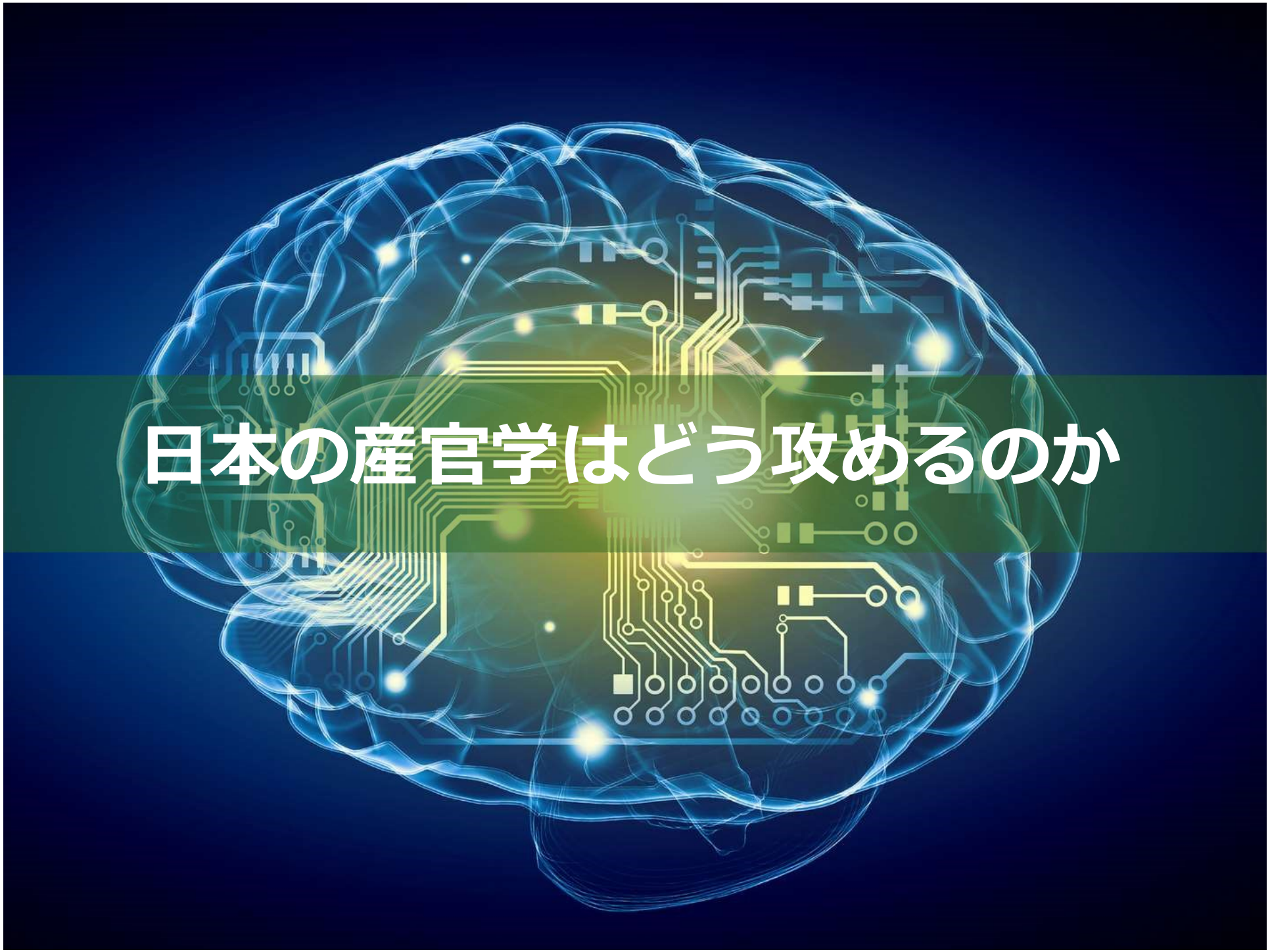
NN: ニューラルネット
NM: ニューロモルフィック
学習: 強化学習
Log: 対数量子化
線形: 線形量子化

汎用: CNN, RNN, MLP, 等の種々のDNNに対応可能
磁界: 磁界結合データ転送
SVM: サポートベクトルマシーン
HD: ハイパーディメンショナルコンピューティング
CNT: カーボンナノチューブFET
ReR: 抵抗変化型メモリ(ReRAM)

ISSCC2019: 機械学習分野の技術的特徴

	セッション名 →	Machine Learning							ML & …			… & Com-in-M		
	↓ 技術分類項目	7.1	7.2	7.3	7.4	7.5	7.6	7.7	14.2	14.3	14.4	24.1	24.4	24.5
対象ML技術	深層NN	✔	✔		✔	✔		✔	✔	✔	✔	✔	✔	✔
	低ビット量子化					✔					✔	✔	✔	✔
	スパース化	✔				✔		✔						
	その他		CNN/ Image	CNN/ Track	汎用	汎用 /FFT		汎用			Conv	Conv	Conv	Conv
	その他ML						NM							
	学習対応				RL		新	DL						
	システム集積化	モバイル	車載	ビデオ						マイコン				
回路	インメモリ					✔			✔		✔	✔	✔	✔
	Mixed-Signal										✔	✔	✔	✔
	新規デバイス									ReRAM		ReRAM		
	発表機関	サムスン	東芝	ミシガン大	KAIST	清華大	ソウル大	KAIST	ミシガン大	スタンフォード	テキサス大	国立精華大	東南大	国立精華大

- プロセッシング視点: スパース化・学習対応・システム集積化
- 回路技術視点: 低ビット精度・インメモリ・ミックスドシグナル



日本の産官学はどう攻めるのか

私見

AIコンピューティング: プロモーション方策

- 方策1: AIアーキテクチャ-プロジェクト
- 方策2: AIチップ-ファーム
- 方策3: AIコンピューティング-センター

AIコンピューティング: プロモーション方策

- **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ハードウェアの研究
    - 製造技術に依存しないファブレス型研究
  - **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**

- **方策3: AIコンピューティング-センター**

経産省: AIチップ・次世代コンピューティング関連事業 (NEDO, 2018～)

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年度予算額 100.0億円（新規）

商務情報政策局 情報産業課
03-3501-6944

事業の内容

事業目的・概要

- IoT社会の到来で急増した情報を活用するためには、革新的なセンサ技術などで効率的に情報を活用するだけでなく、ネットワークの末端（エッジ）側で中心的な情報処理を行うエッジコンピューティング等、従来のサーバー（クラウド）集約型から情報処理の分散化を実現することが不可欠です。
- 半導体の開発指標たるムーアの法則の終焉が叫ばれ、既存技術の延長は限界を迎えつつあります。他方、エッジ側でAI処理を実現するため、小型かつ省エネルギーながら高度な処理能力をもった専用チップと、それを用いたコンピューティング技術が必要です。また、クラウド側でも増加が著しいデータの処理電力を劇的に低減するため、従来の延長線上にない新しい技術の実現が求められます。
- 本事業は、エッジ側で超低消費電力AIコンピューティングや、新原理により高速化と低消費電力化を両立する次世代コンピューティング等、ソフトだけではなくハードと一体化した技術開発を実施。ポストムーア時代における我が国のベンチャーを含む情報産業の競争力強化、再興を目的とします。

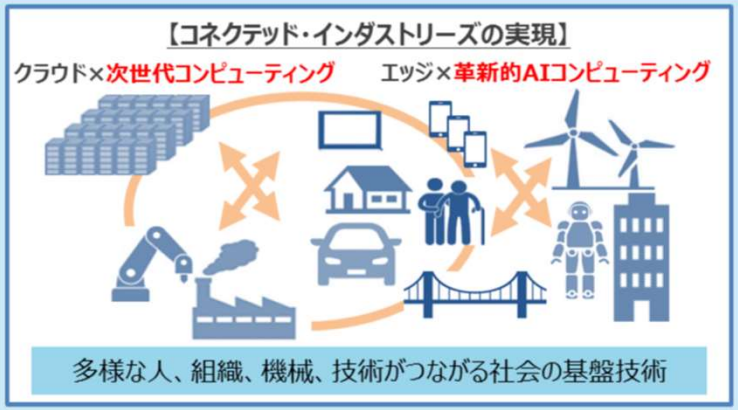
成果目標

- 平成30年度から最長で平成39年度までの10年間の事業であり、 IoT社会をエッジからクラウドまで高度化する基盤技術を確立、省電力化を実現します。(平成49年度において約4,900万t/年のCO2削減を目指します。)

条件（対象者、対象行為、補助率等）

国 → 交付金 → (研)新エネルギー・産業技術総合開発機構（NEDO） → 委託 → 大学・研究機関・民間企業等

事業イメージ



革新的AIエッジコンピューティング技術の開発

- エッジ側では電力等の制限が厳しく、革新的AIチップを用いたエッジAIコンピューティングの省エネ化に関する開発を実施。
- エッジAIコンピューティングの開発では、良質なデータを用いた用途毎の擦り合わせが重要。ソフトとハードを一体化する技術開発等を実施。

次世代コンピューティング技術の開発

- 中期的には高速化と省エネ化を実現するコンピューティング技術開発を実施する。
- 長期的には、現状を打破する破壊的イノベーションの創造に向けた新原理コンピューティングの技術の開発等を実施する。

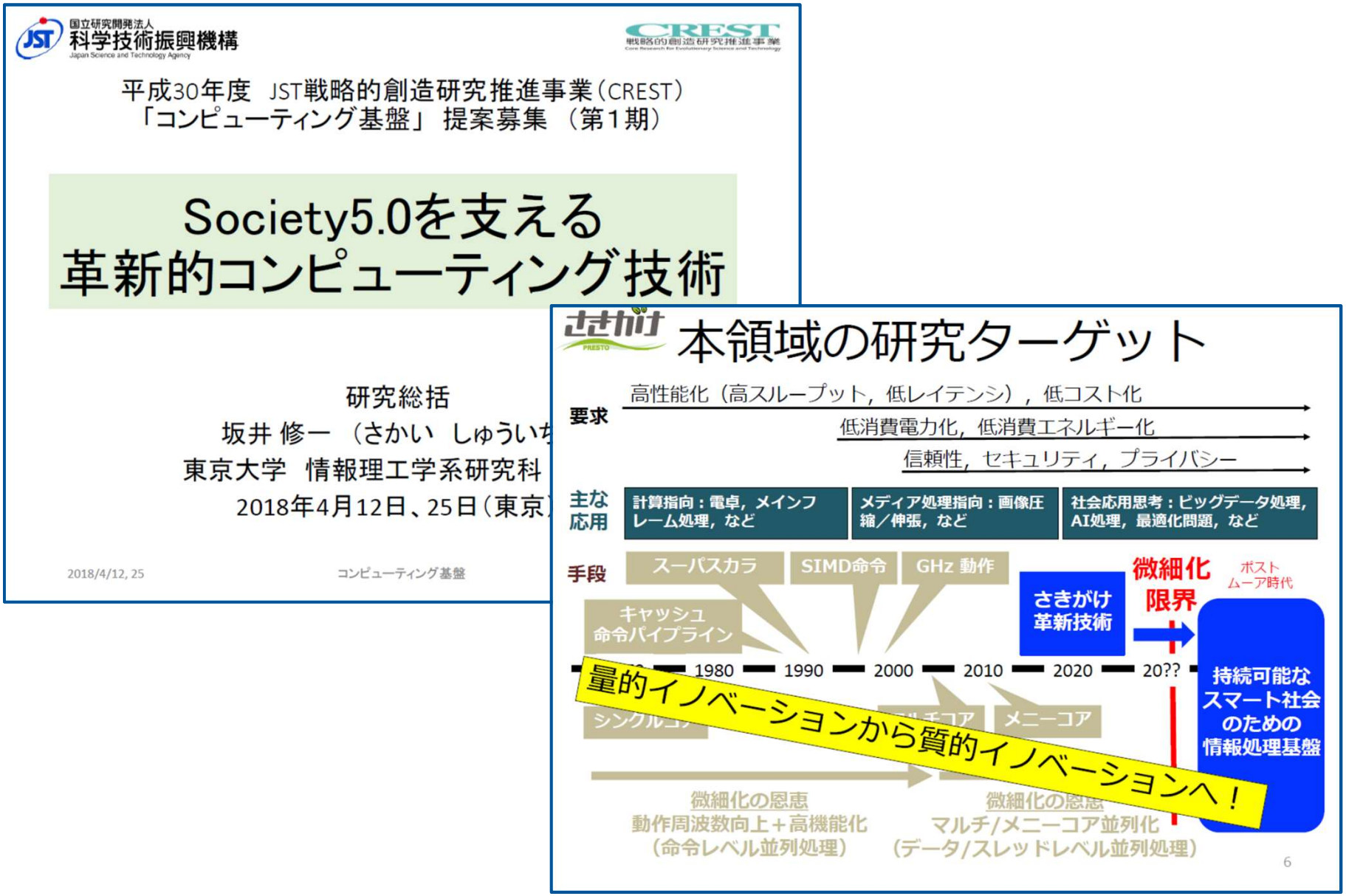
高度なIoT社会を実現する横断的技術開発

- 大量のデータの効率的かつ高度な利活用を実現する情報の収集、蓄積、解析、セキュリティ等に関する横断的技術開発を実施する。

NEDO 「革新的AIエッジコンピューティング技術の開発」採択12プロジェクト

テーマのキーワード	事業者
エッジHPC、OS	イーソル、名古屋大
超広範囲センシングAIエッジ技術	沖電気、会津大、ALSOK、ジャパンマリンユナイテッド
5G、高度自律的学習機能	KDDI、アラヤ
進化型・低消費電力AI エッジLSI	ソシオネクスト、Architek、豊田自動織機
動的再構成技術、組み込みAI	ルネサスエレクトロニクス、三菱電機、SOINN、北大
完全自動運転SoC	東京大、埼玉大、アクセル、ティアフォー
不揮発省電力FPGA	NEC
エッジビジョンAI、超軽量化技術	フィックスターズ
動的多分岐・結合トレース型AI HW	NSITEXE、ユーリカ、日立、東工大
横断的なセキュリティ評価	産総研、ECSEC、コネクテックジャパン、IIJイノベーションインスティチュート
エッジヘビーAI向けSoC	Preferred Networks、神戸大
セキュアオープンアーキテクチャ	日立、産総研、慶應大

文科省: 革新的コンピューティング戦略目標 => CREST/さきがけ (JST, 2018 ～)



AIコンピューティング: プロモーション方策

- **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ハードウェアの研究
    - 製造技術に依存しないファブレス型研究
  - 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

NEDO事業

産総研・東大 AIチップ設計拠点

- **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プレイヤの**育成・放牧**

東大

システムデザイン研究センタ (d.lab)

- **方策3: AIコンピューティング-センター**

AIコンピューティング: プロモーション方策

- **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ハードウェアの研究
    - 製造技術に依存しないファブレス型研究
  - 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・応用にフォーカス
  - 上位レイヤと連携するAIハードウェア**中核拠点**が必要

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

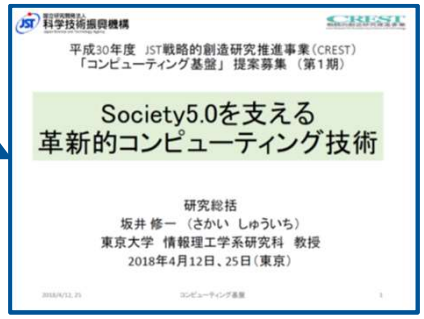
- **方策1: AIアーキテクチャ-プロジェクト**
  - システムアウェアなAI-HW研究
  - 産学連携プロジェクト群で応用-アルゴリズム-アーキテクチャ縦断エコシステムの形成
- **方策2: AIチップ-ファーム**
  - 先進アーキテクチャを低コスト・短TATでHW化・実証するためのLSI設計試作環境の構築
  - システムアウェアHWプレイヤの育成・放牧
- **方策3: AIコンピューティング-センター**
  - 既存のAIセンターはAIアルゴリズムが中心
  - システムアウェアAI-HWの**中核拠点**を形成

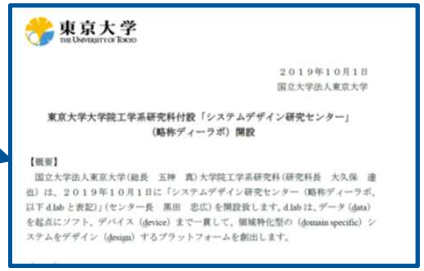
経産省・NEDO

文科省・JST

経産省・NEDO

東京大学





AIコンピューティング研究ユニット
@東京工業大学の狙い

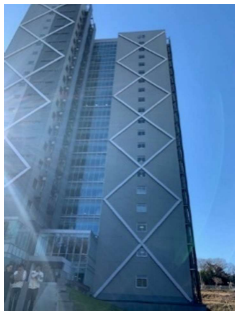
AIコンピューティング研究ユニット: ArtIC

2019年4月に発足

2020年4月からフルメンバで活動予定

10/1に阪大から劉(Yu)准教授を迎え，研究活動を本格化．**研究スタッフも募集中**

すずかけ台キャンパス J3棟 17F



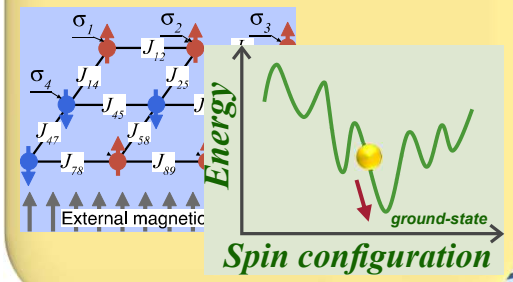
http://www.artic.iir.titech.ac.jp

助教，特任助教: 近日着任予定

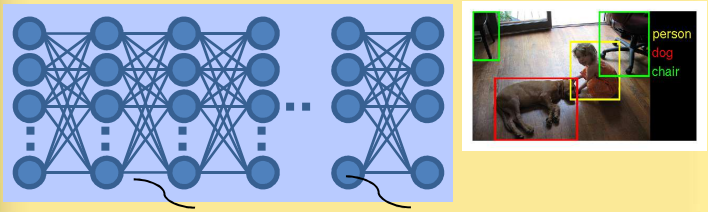
ArtIC: 研究ターゲット

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

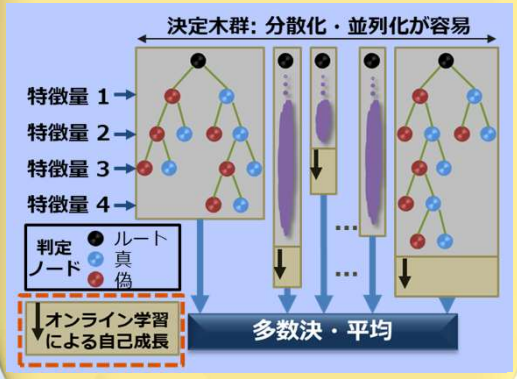


大量データの学習 → 強力な推論・識別・予測能力



抽象シナプス　抽象ニューロン

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

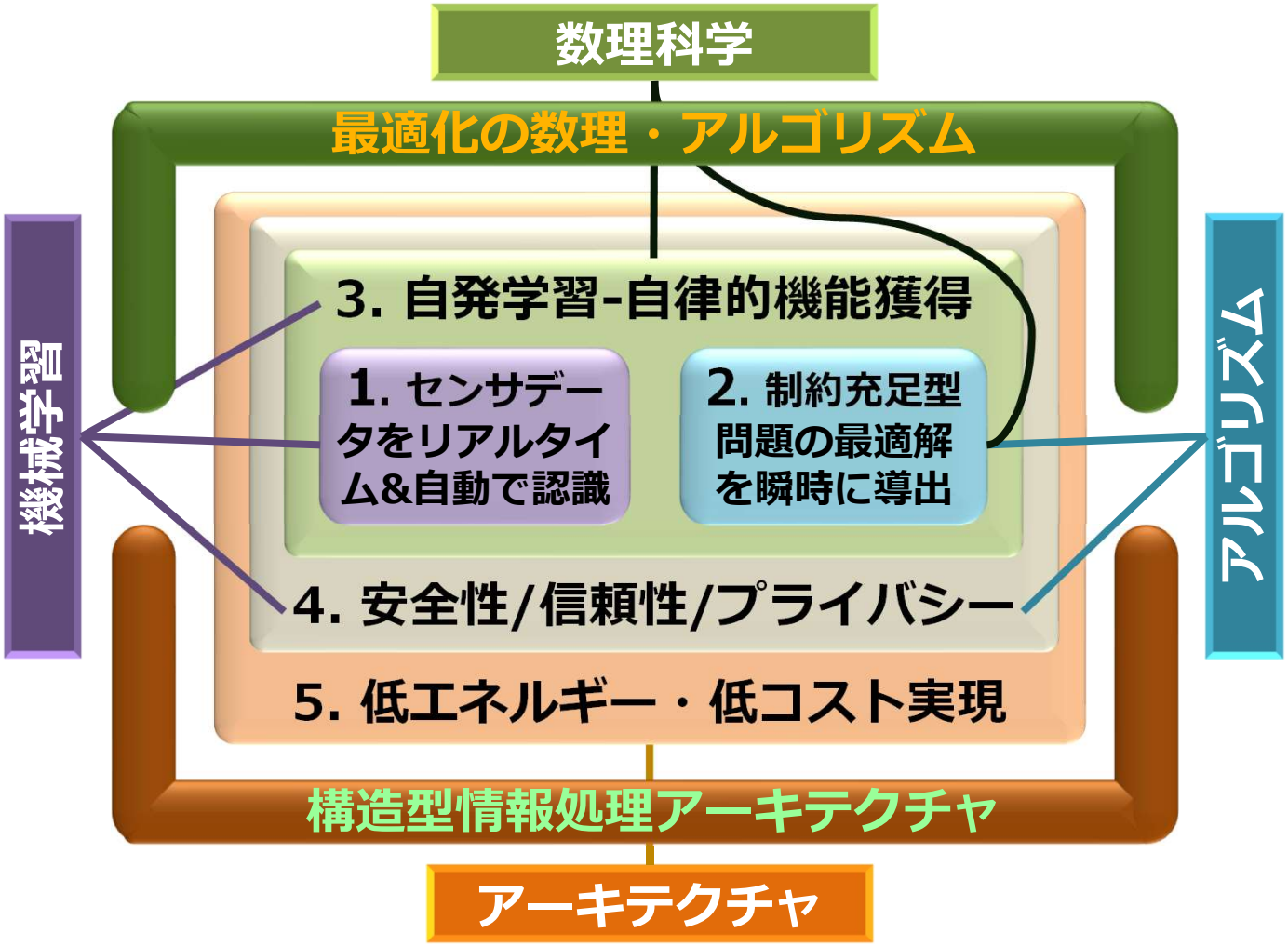
アルゴリズム理解 => アーキテクチャ研究 => ハードウェア実現

ArtIC: 研究アプローチ (Society5.0観点で)

- 深層ニューラルネット -
- より広範な機械学習 -

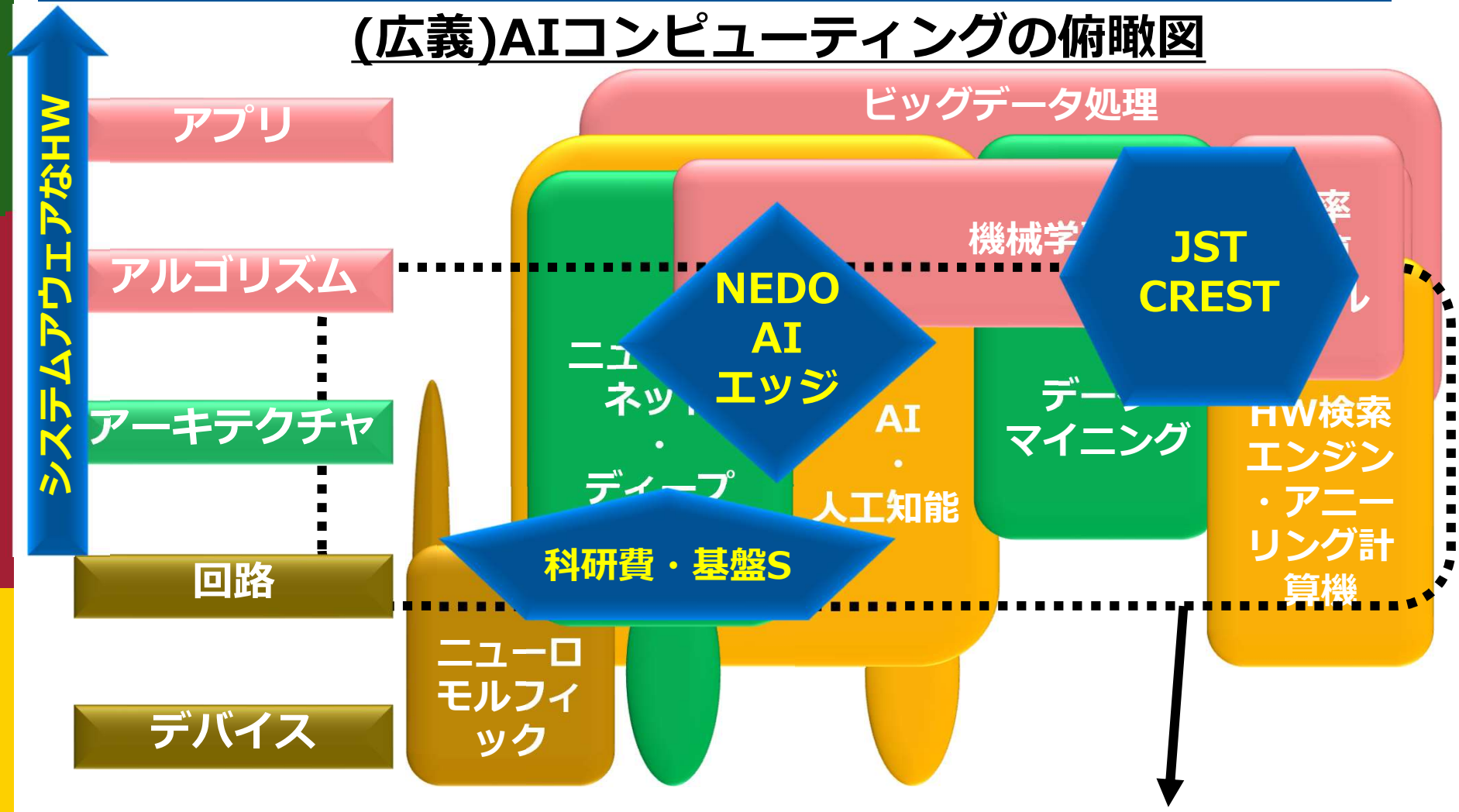
AIコンピューティング

- 組合せ最適化・列挙 -
- アニーリング計算 -
- グラフマイニング -



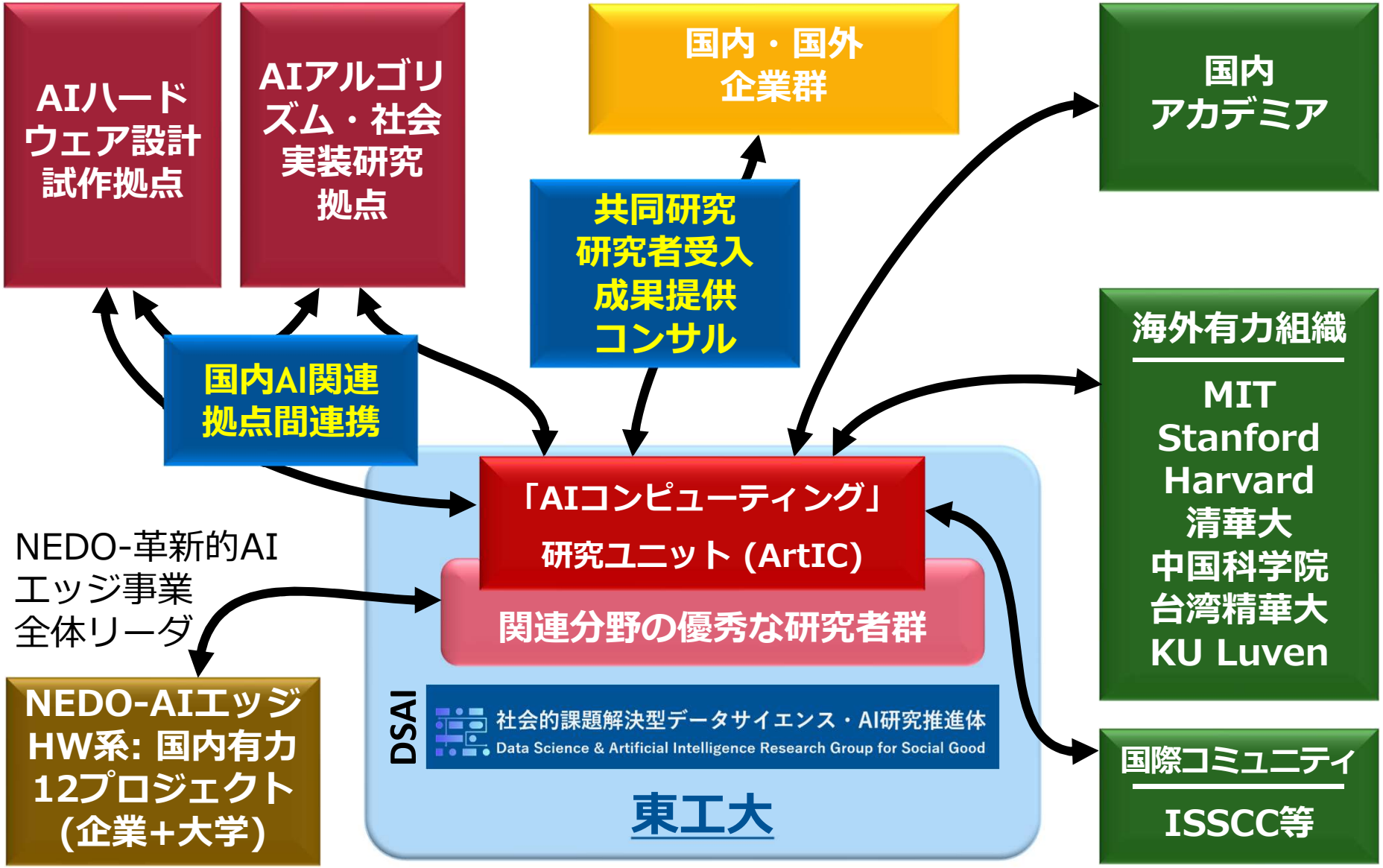
ArtIC: 研究分野・研究レイヤ

(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

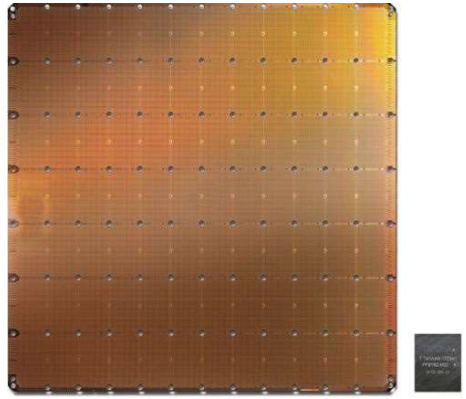
ArtIC: 外部連携構想



注目すべき研究トレンド

- **DNNアクセラレータは更なる変革期へ**
  - 推論単体:
    - 疎結合性の活用が進行中
    - 非畳み込み型DNNの勃興
  - 学習アルゴリズム-アーキテクチャ協創
    - バックプロパゲーション離脱の試み
  - エッジ学習・オンライン学習への取り組み
  - アテンション技術のアーキテクチャ展開

Cerebras社:
Wafer Scale Engine



21.5 cm

- 構造型情報処理アプローチ
- All SRAM

- **DNNアクセラレータ as a モジュール**
  - 上位システムアーキテクチャへの展開
  - 従来機械学習手法との融合・適材適所
  - Explainable AI

- **非線形・確率的コンピューティングへの展開**
  - 最適化アクセラレータ・アニーリングチップ
    - 非量子，集積回路ベースの研究活性化
    - 日本のアクティビティが高い分野

AI := コンピューティングのゲームチェンジャ

- **AIチップへのゴールドラッシュ**
  - 計算機アーキテクチャ研究者がこの分野に集結
    - Re-Inventing Computer

- **What is Next Exponential (指数関数的増大)?**
  - ムーアの法則は続くのか？
  - メモリ容量/チップ，CPU性能/チップの次はあるのか？
  - **私見: ニューロン数/チップ ないしは AI性能/チップが次の指標に**

- **日本は周回遅れ**
  - システムアウェアハードウェアへのR＆D投資が中抜け状態
  - しかし、**本来の技術蓄積は厚い**
    - AI分野は元来日本の得意分野

- **「遅い?」=> Yes. 「遅すぎ?」=> No.**
  - 現在のAIブームは「氷山の一角」
  - 変化の激しさ => 後発者利益を取り得る分野
  - ゲームチェンジ => **日本にとっての再チャンス**

AIコンピューティング: 日本の活きる道

- アイデアの先回り
  - 程よいジャンプ感
  - 飛ぶ距離と飛ぶ方向を間違わない
  - 集合知の強みを活かす
- 出口 = 組み込みシステム
  - アジェンダ設定の「戦略」
  - 勝てる枠組みを作る・使う
- 蓄積材活用
  - 人材・知財
    - 10－20年前のアイデアがようやくmake sense
  - 時間を買う

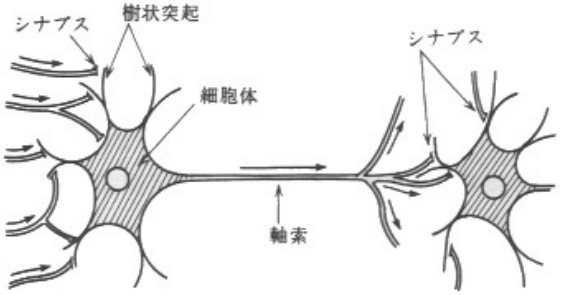
日本からより多くの産学プレイヤーの「AIコンピューティング・AIチップ」分野への参画を期待

Tokyo Tech

ARTIC

脳模倣型情報処理とディープラーニング

シナプス/ニューロン



http://www.si.hirosaki-u.ac.jp/~ei_sys/neuro/neuron.html

スパイキングニューロン

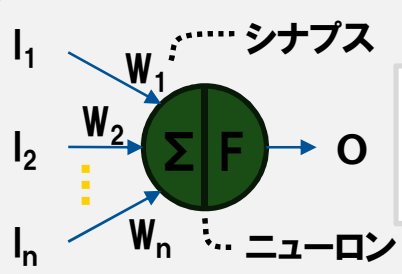
- 生理学モデルをなるべく**忠実に**電子回路化
- パルス列で信号を表現
- アナログ的に蓄積=>発火

=> **ニューロモルフィック**

=> 脳模倣・エミュレーション

積和型ニューロン

- 計算機上で処理しやすいように**抽象化**したモデル
- 単なる**乗算と加算**



活性化関数

$$O = F(W_1 I_1 + W_2 I_2 + \cdots + W_n I_n)$$

=> **ニューラルネット**

=> 機械学習の分類器としても広く利用されている

脳模倣型情報処理

“鳥”を目指す

別物/別時間軸

ディープラーニング

“飛行機”を目指す

脳模倣型情報処理とディープラーニング

脳模倣型情報処理

“鳥”を目指す

何が本質?
実現可能?

ディープラーニング

“飛行機”を目指す

“鳥”	比較項目	“飛行機”
自然界が作り上げたお手本	位置づけ	鳥に学び、「飛ぶこと」の本質を残し、工学的に実現方法な枠組みを構築
◎	エネルギー効率	△
×	大量輸送/高速輸送	◎

飛行機が社会にもたらした変革の本質

AIコンピューティング: 米国＋中国が圧倒

必ずしもチップ実装まで行わない
ハードウエア方式レベルの研究も含む

- **米国**
  - NVIDA, Google, Apple,,,
  - AIハードウェア系スタートアップ多数
  - DARPA: ERI(Electronics Resurgence Initiative) 2018～
    - 年間**200億円**を超える予算
      - **リコンフィギュラブル**ハードウェア
      - **ドメインスペシフィック** チップ
      - HW設計のモジュール化・スマート化
      - **オープンソース**ハードウェア
      - 3次元集積 System-on-a-chip
      - 非ノイマンコンピューティングデバイス
- **中国**
  - **年間1,000億円規模の**AIハードウェア国家Pj
    - 清華大・中国科学院にAIハードウェアセンタ
  - 一説には約30社のAIチップスタートアップ企業が存在

AIチップ設計拠点への期待

- チップ化の重要性
  - アイデアを形に (プロトタイピング)
  - リアリティによる説得力
  - 現実制約との折り合い
- 活動範囲
  - 設計検証ツール・インフラ
  - IPコアの提供
  - 設計サービス
  - 試作サービス
- as a Serviceの視点から構築
- AIエッジ等、他プロジェクトとの連携

ハードウェアなんて分からない

ファンダリ利用のターンキー

チップ化して初めて競争の土俵に上がれる

AI応用の専門家、AIアルゴリズムの専門家をチップ化に呼び込むには?

AIチップ設計拠点

↓

AIチップ設計・試作・サービス拠点