新技術説明会
New Technology Presentation Meetings!

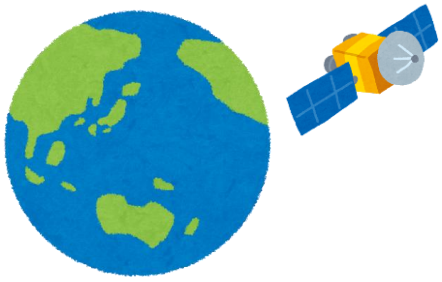
長寿命＆高耐性！
製品価値を高める次世代高効率AI

会津大学　コンピュータ理工学部 コンピュータ理工学科
上級准教授　富岡　洋一

2024年12月12日

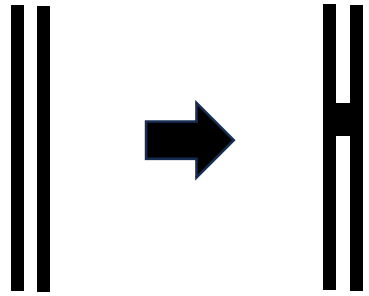
AIの活用と信頼性課題

- 様々な分野でAI活用が期待
  - ■インフラ・医療などのミッションクリティカルシステム
  - ■宇宙などの過酷な環境で運用される人工衛星
- **AI故障が人命に関わる深刻な誤動作や大きな損害**に至る

	自動運転	手術支援	人工衛星サービス
AI活用			
	• 車線、障害物認識	• 画像診断AI • 手術支援	• リモートセンシングAI
AI故障	人命に関わる事故		打ち上げ費用大 経済的損失大

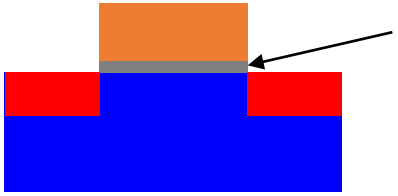
故障・誤動作の要因

放熱不良により配線がショート



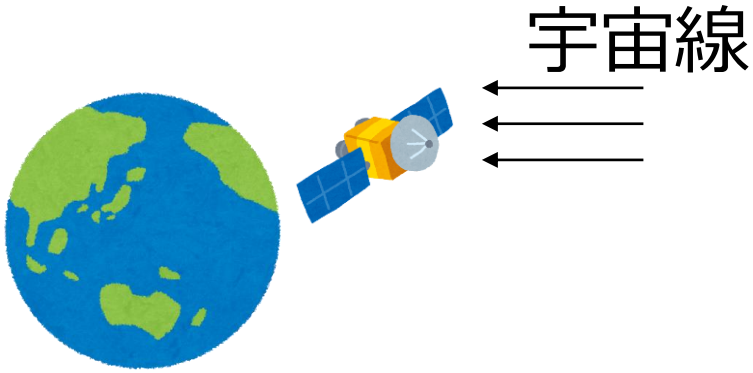
経年劣化・摩耗による遅延増大

MOSFET

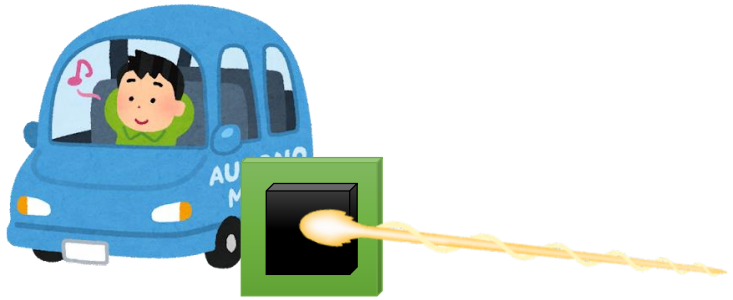


酸化膜中の欠陥に起因徐々に閾値電圧が増加し、遅延が増大

宇宙線による一時的・永続的故障



レーザーによるレジスタ・メモリデータの破壊

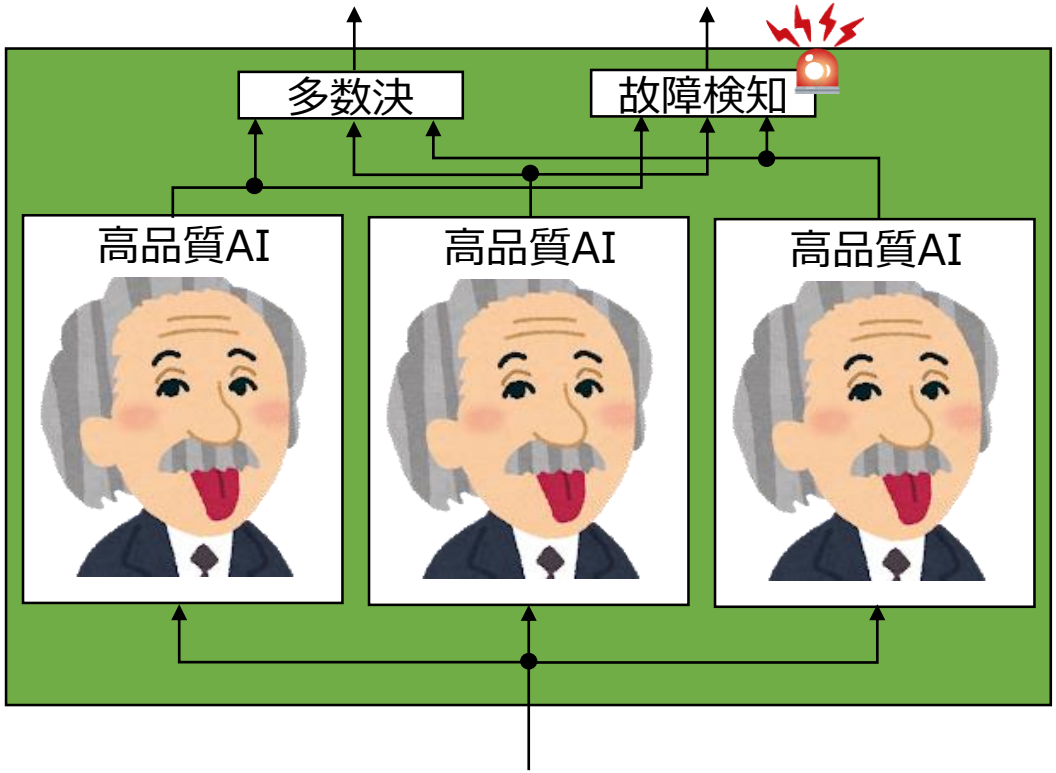


自動運転における故障の影響例

Donkey Car Simulatorを使用して生成

従来技術とその問題点

- 実用化されている技術として、Dual/Triple Modular Redundancy (DMR/TMR) 等の冗長化がある
- 一部モジュールの故障に頑健になるが、**消費電力や計算コストが増加する問題**があるため、冗長化による耐故障AIは広くは普及していない
- 製品コストや電力に厳しい制約があるエッジデバイスでは、計算負荷の高いAI処理の冗長化は難しいのが現状



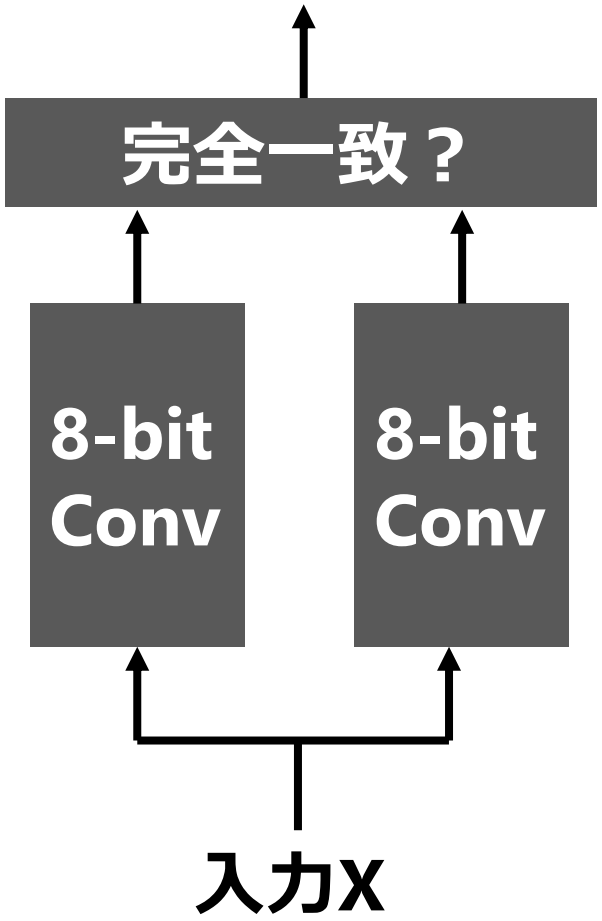
新技術の特徴・従来技術との比較

- 従来技術：**冗長化（DMR, TMR）**
  - AI推論処理に要する計算コスト、消費電力が2倍、3倍に増加してしまう問題
  - 各AI回路が利用可能な電力が減少し性能劣化
  - 活用先は自動運転等の高信頼性が必要な製品に限定
- 提案技術：**近似故障検出、耐故障アンサンブル**
  - AIの故障検出に要する消費電力が2～3割削減でき、耐故障AIの推論に要する計算コストが半分程度まで削減できることから、低コスト・電力制約が厳しい製品の信頼性向上に貢献できる

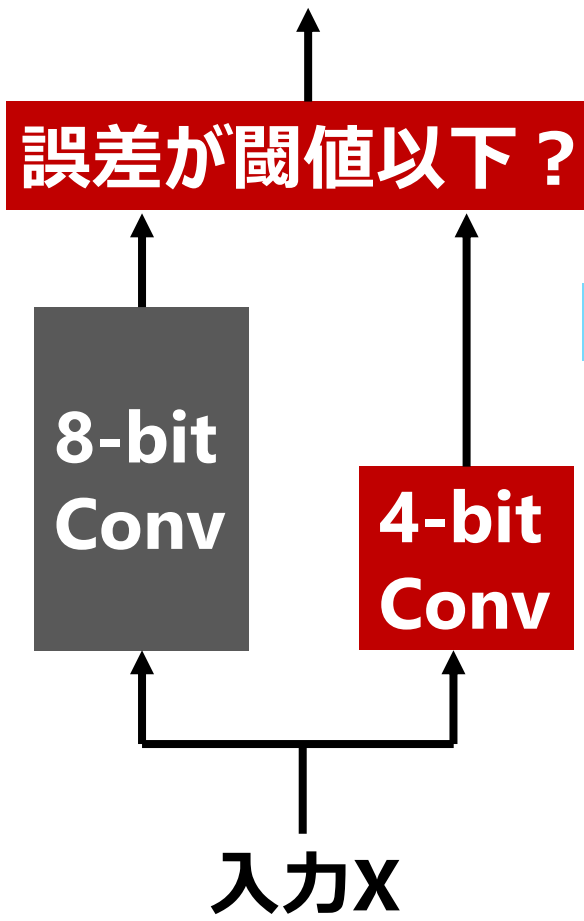
近似故障検出の基本的なアイデア

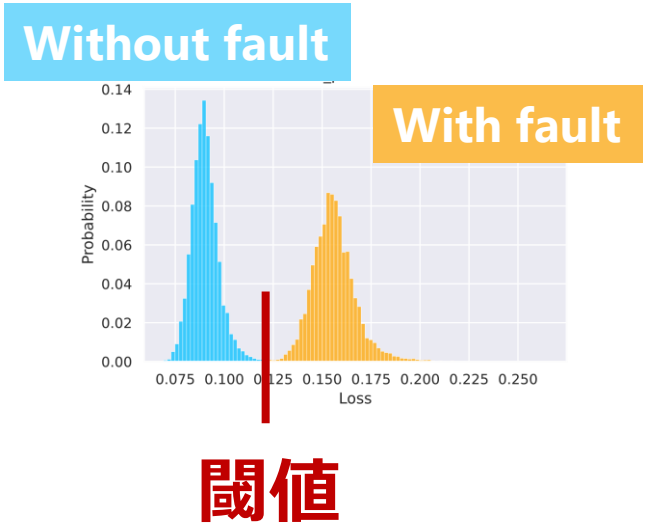
DMR: Dual Modular Redundancy　　**Conv: 畳み込み層**

一般的なDMR



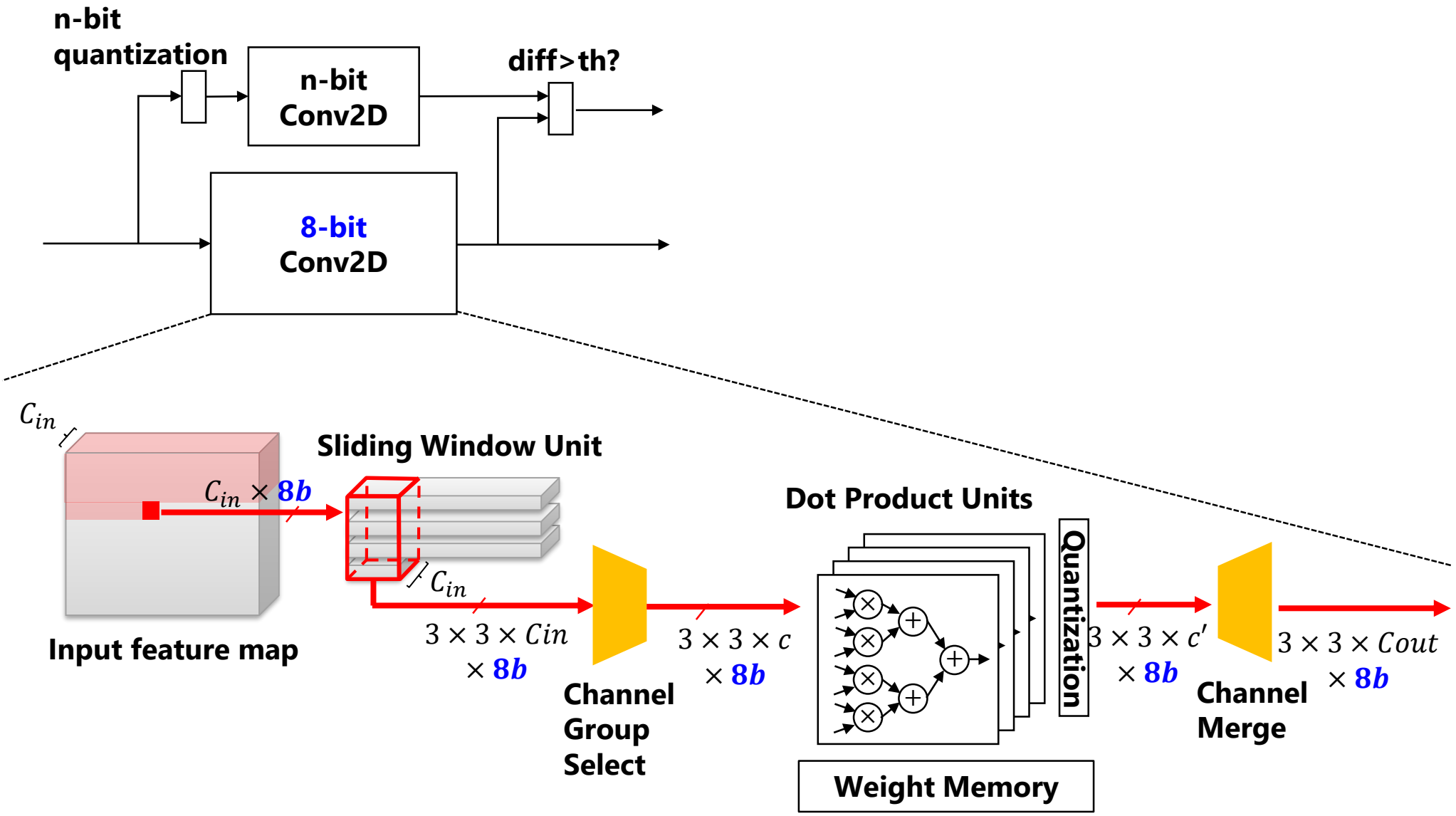
近似DMR





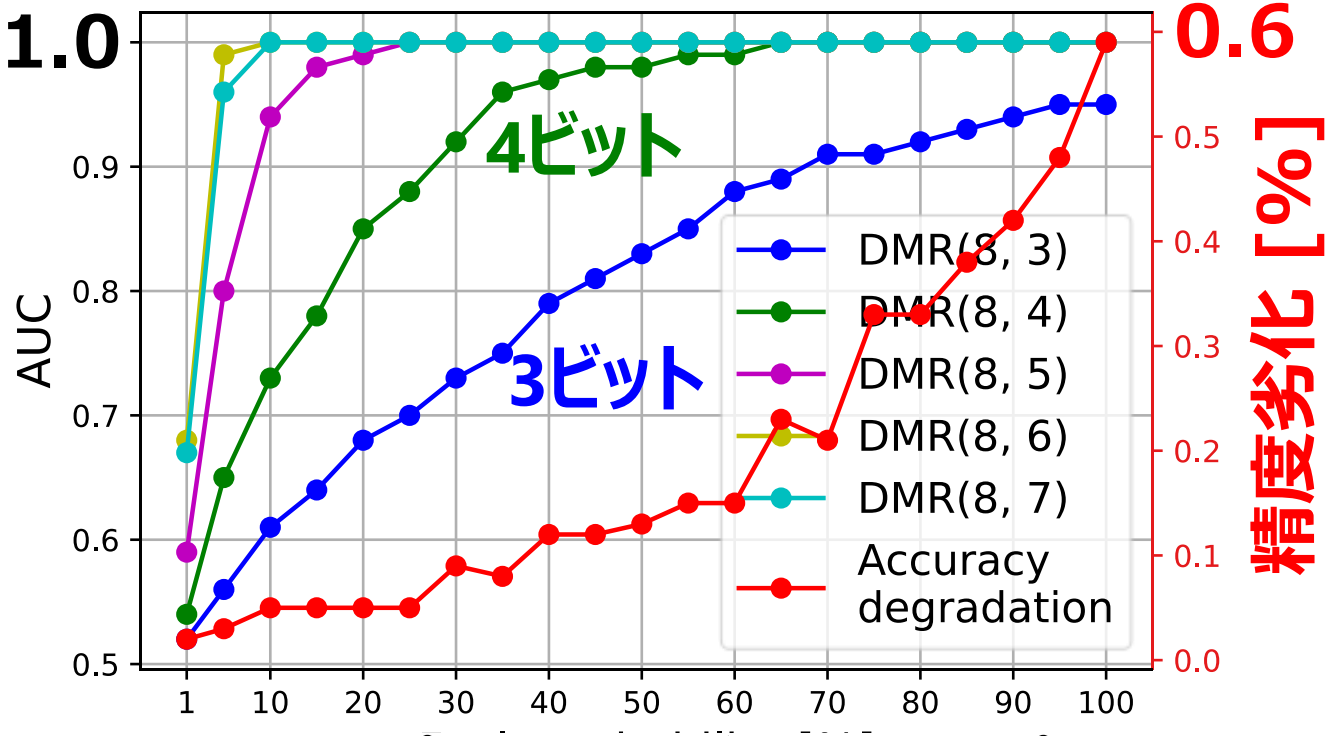
閾値

近似DMR回路例


n-bit
quantization
n-bit
Conv2D
diff>th?
8-bit
Conv2D
C_{in}
$C_{in} \times 8b$
Sliding Window Unit
C_{in}
$3 \times 3 \times Cin$
$\times 8b$
Input feature map
Channel
Group
Select
$3 \times 3 \times c$
$\times 8b$
Dot Product Units
Quantization
Weight Memory
$3 \times 3 \times c'$
$\times 8b$
Channel
Merge
$3 \times 3 \times Cout$
$\times 8b$

近似DMRによる故障検出精度

- 故障検出精度をArea Under Curve (**AUC**)で評価
  - 擬陽性と偽陰性の両方が減少すると1に近づく
- ResNet-20の2層を除いて、8ビットと4ビット畳み込み層による近似DMRで**認識精度が1%低下する前にAUC=1**達成



※グラフはResNet-20@CIFAR-10による評価
※ConvNeXt@IMAGENETも検証済

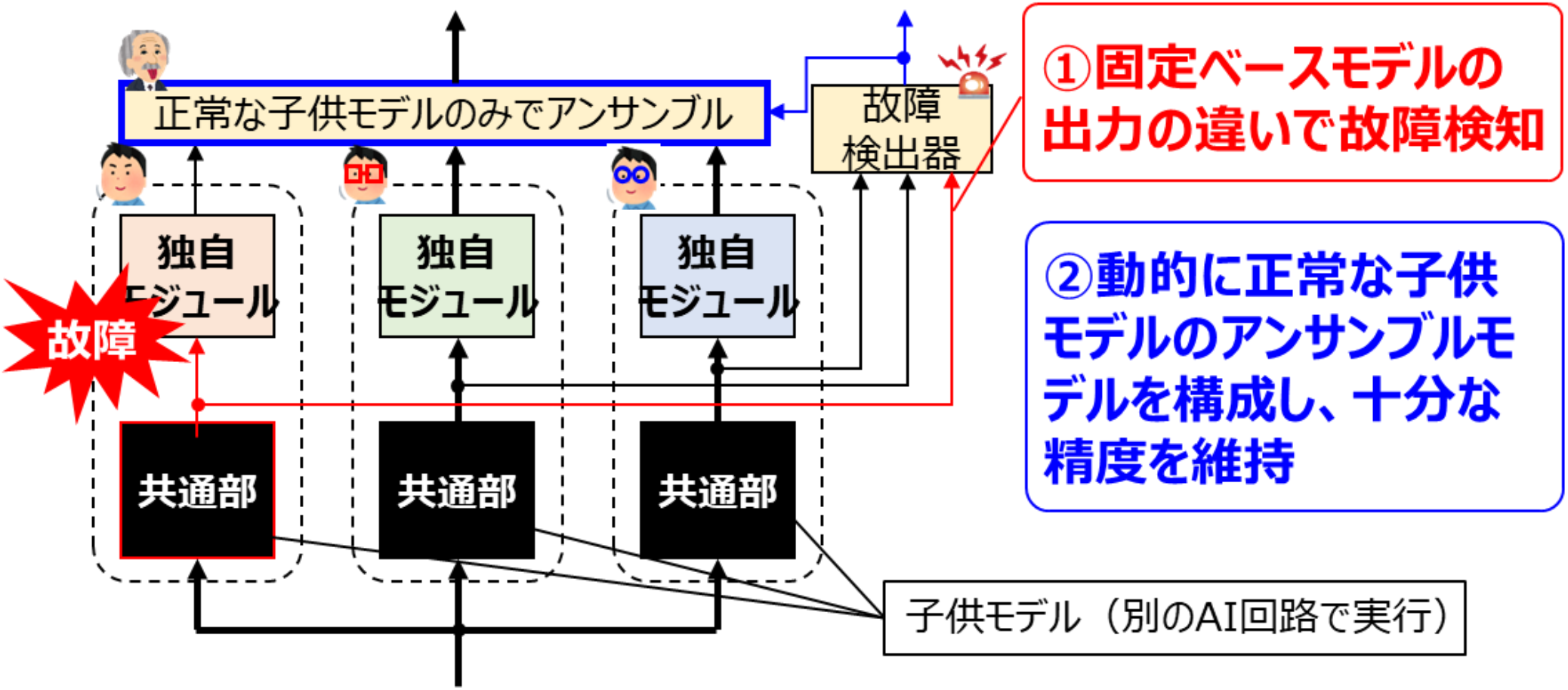
近似DMRによる消費電力削減効果

- 8ビットだけで構成した従来のDMRに対して、8ビットと4ビットによる近似DMRで**消費電力を約27%削減**

FPGA: Xilinx Kintex KU5P
モデル: ResNet-20
データセット: CIFAR-10

アンサンブル耐故障AI

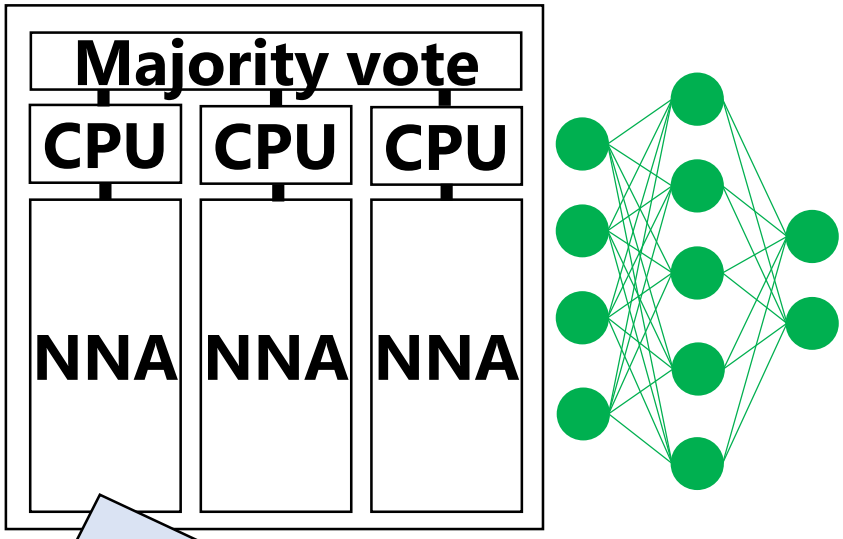
- 故障発生時にも、子供モデルの相関・共通計算に基づき故障を検知
- 正常な子供モデルのアンサンブルにより十分な精度を維持



想定するアーキテクチャ

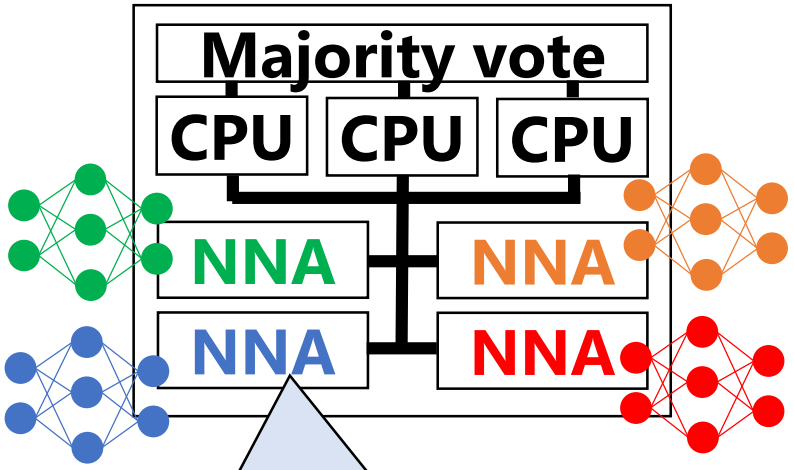
NNA: Neural Network Accelerator

TMRベース SoCアーキテクチャ



大型モデルの推論処理のために
高性能なNNAが必要

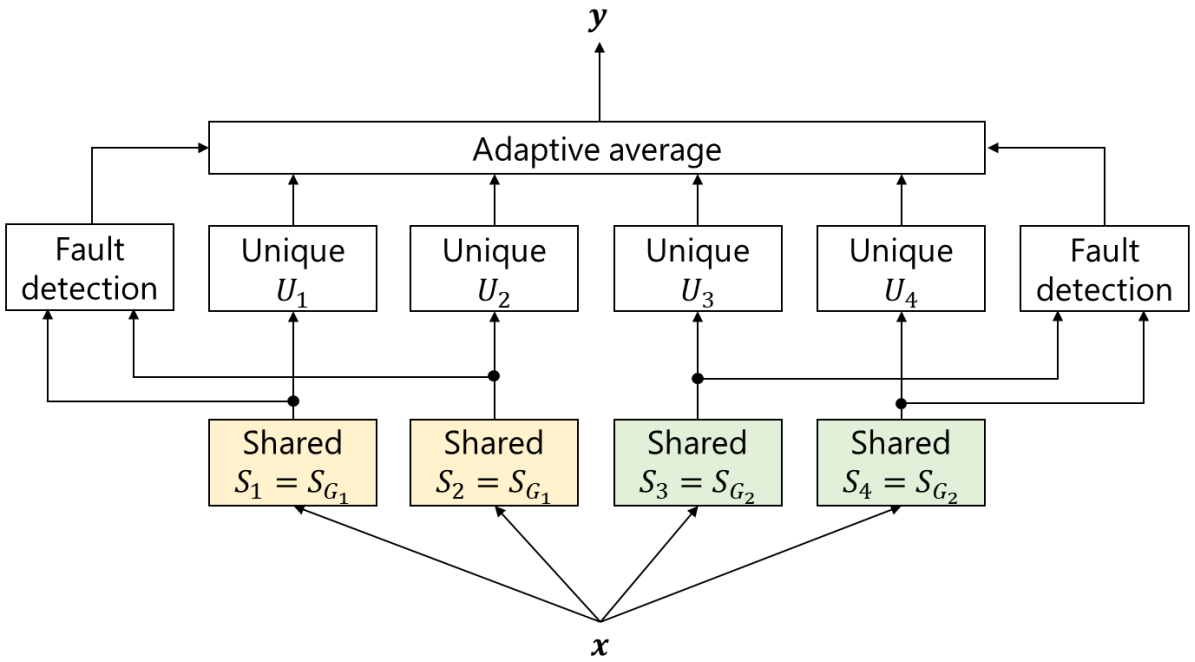
耐故障アンサンブルベース SoCアーキテクチャ

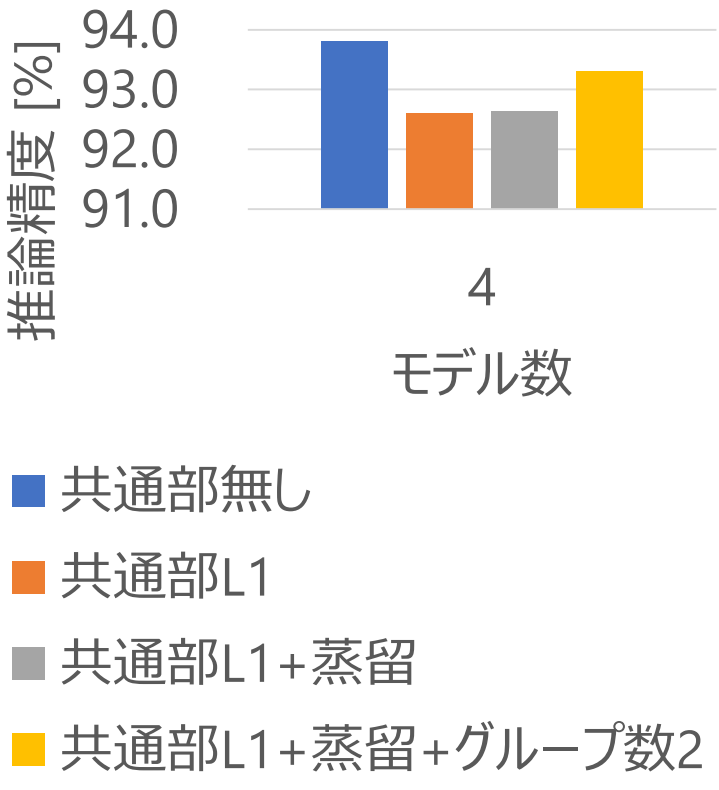


小型モデルの推論処理で
十分なので低消費電力

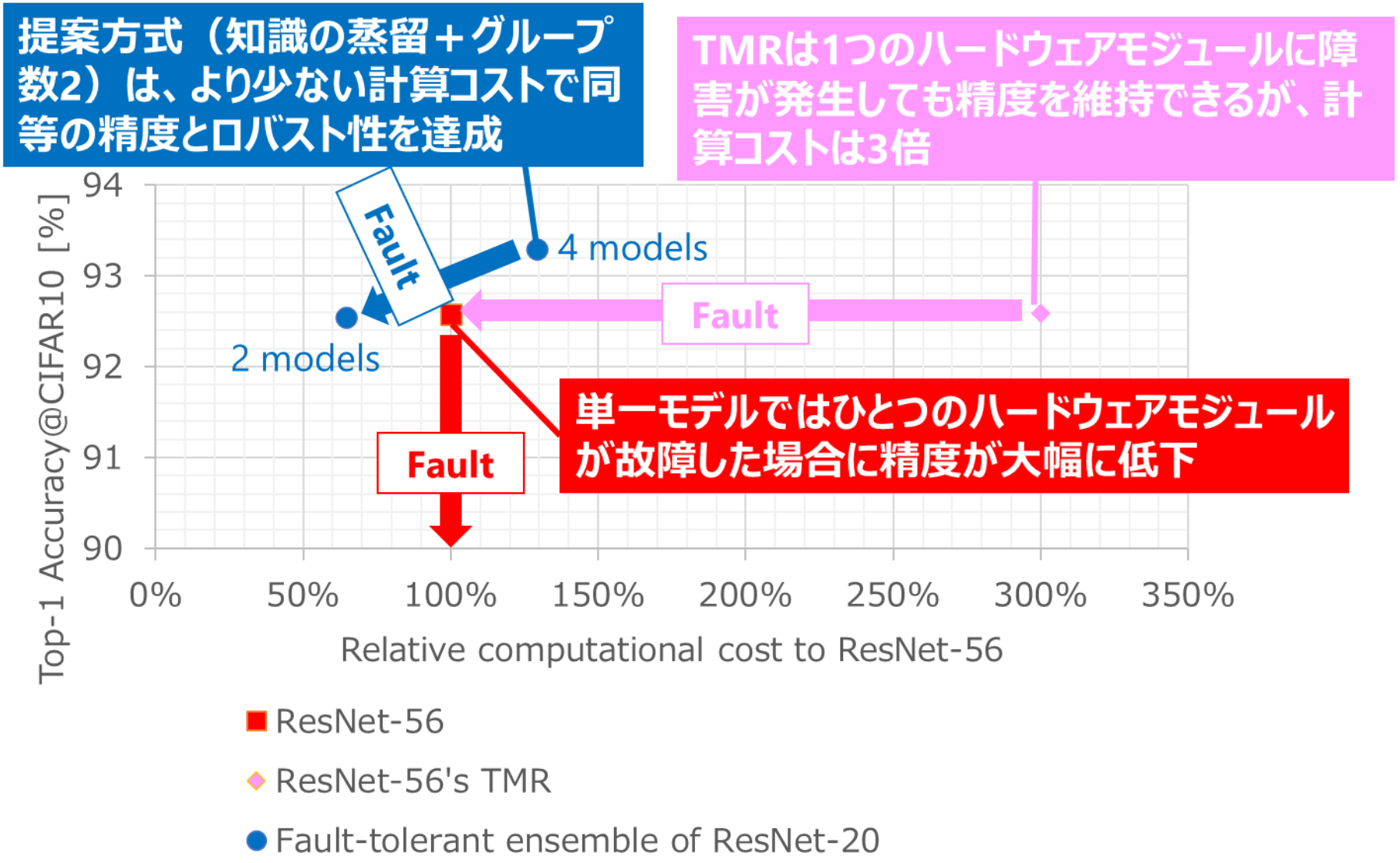
学習方法

- 課題：共通部のために子供モデルの多様性が低下しアンサンブルモデル精度も低下
- 学習方法：
  - 多様性の高い子供モデルの知識を伝達（知識の蒸留）
  - 共通部を複数グループに分ける

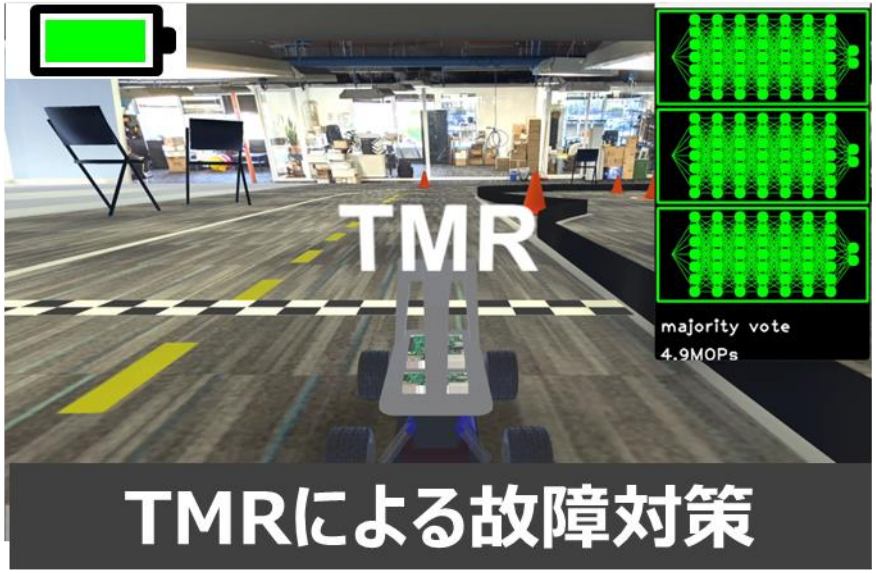




ResNet20@CIFAR-10における耐故障アンサンブルAIの効果



耐故障アンサンブルAIデモ



提案技術

Donkey Car Simulatorを使用して生成

AI製品の価値を高める

- 故障を検知して安全に停止できる
- 一部故障しても動き続ける
  - 完全に故障する前にお知らせできる

- 製品に対する信頼・安心感を高める




想定される用途

- 医療機器
- パーソナルモビリティ
- 家電
- 工場等で使用する異常検知器
- ドローン、ロボット
- ロケット、人工衛星

消費電力や計算コストのオーバーヘッドが少ないので
信頼性向上も新しい選択肢になるのでは？

現状と実用化に向けた課題

- 近似故障検出
  - 発展として故障時に出力を復元して動作し続けるところまで開発済
- 耐故障アンサンブル
  - 学習アルゴリズムを強化し、更に高い精度を達成済
- 両技術に共通する課題
  - 公開データセットに対する評価のみで実問題に対して未適用
  - 実問題に対しては、問題の特徴を捉えた最適化が必要になる可能性がある

企業への期待と貢献

- 技術の活用
  - あらゆる製品にAIが導入される時代が近づいている中、多方面からAIの信頼性を向上することが必要
- 共同研究
  - 特定課題向けの技術拡張
  - SoC設計に関する技術のある企業との連携
- 本格導入にあたっては技術指導も可能

本技術に関する知的財産権①

- 発明の名称　：ニューラルネットワークの近似故障検出手法及び回路
- 出願番号　：特願2023-181283
- 出願人　：公立大学法人会津大学
- 発明者　：富岡 洋一、齋藤　寛

本技術に関する知的財産権②

- 発明の名称　：ニューラルネットワークの
　　耐故障アンサンブル技術
- 出願番号　：特願2024-002563
- 出願人　：公立大学法人会津大学
- 発明者　：富岡 洋一、齋藤　寛

産学連携の経歴

- 共同研究実績
  - 電機メーカー
  - 半導体メーカ
  - 電子部品メーカー
- 企業との共同プロジェクト実績
  - 2018年-2020年　NEDO事業（委託）
  - 2020年-2022年　NEDO事業（再委託）
  - 2022年-2025年　JSTさきがけ事業に採択
    （産学連携を目指し、研究中）

お問い合わせ先

会津大学
産学官連携コーディネーター　石橋　史朗

ＴＥＬ　０２４２－３７－２７７６
ＦＡＸ　０２４２－３７－２７７８
e-mail　ubic-adm@ubic-u-aizu.jp