新技術説明会
New Technology Presentation Meetings!

締切を守りつつ精度を引き上げる時間適応型AI推論技術

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

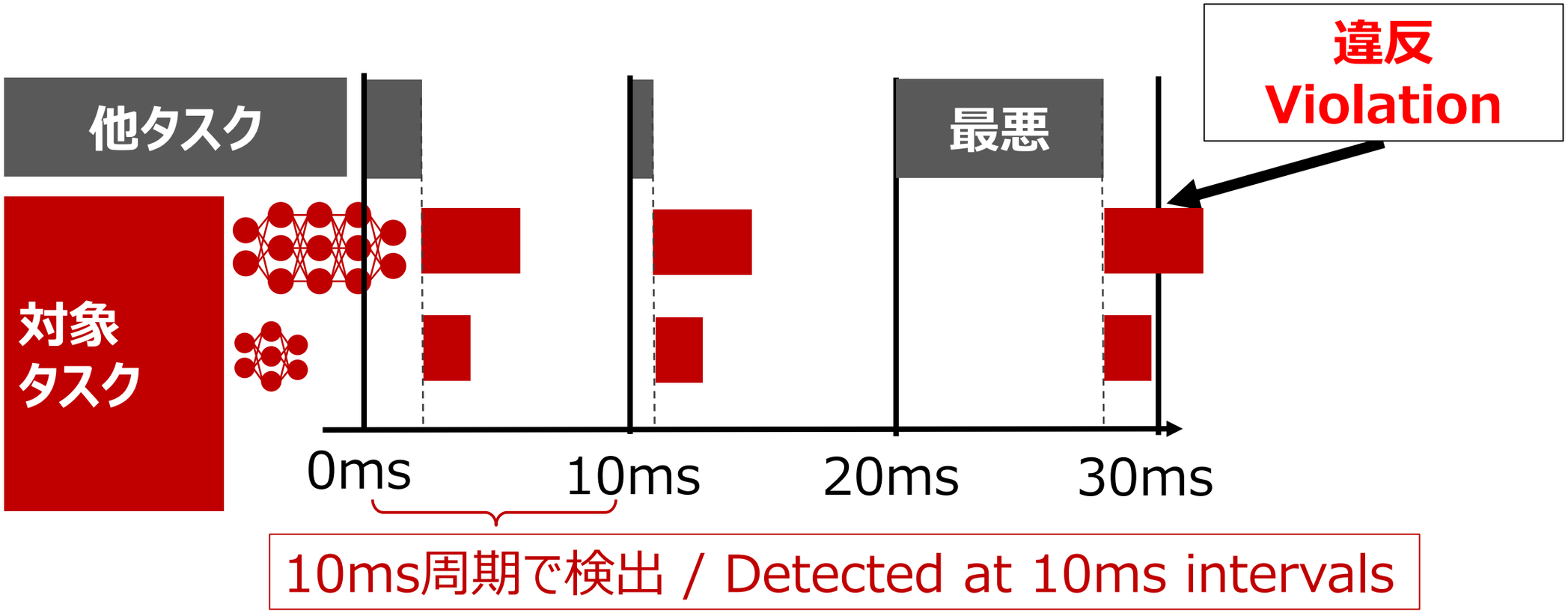
2026年1月20日

研究背景

- 社会的背景
  - 自動運転車やロボットは、周囲の状況を逐次理解しながら行動を決定
  - 判断が遅れると事故やタスク失敗につながる
  - 推論処理の時間制約は非常に厳しい
- 技術的課題
  - 現在のAIモデルは一般的には一定の計算量を前提として設計
  - 推論が遅くなると安全限界を超え、制御系全体の信頼性が損なわれる

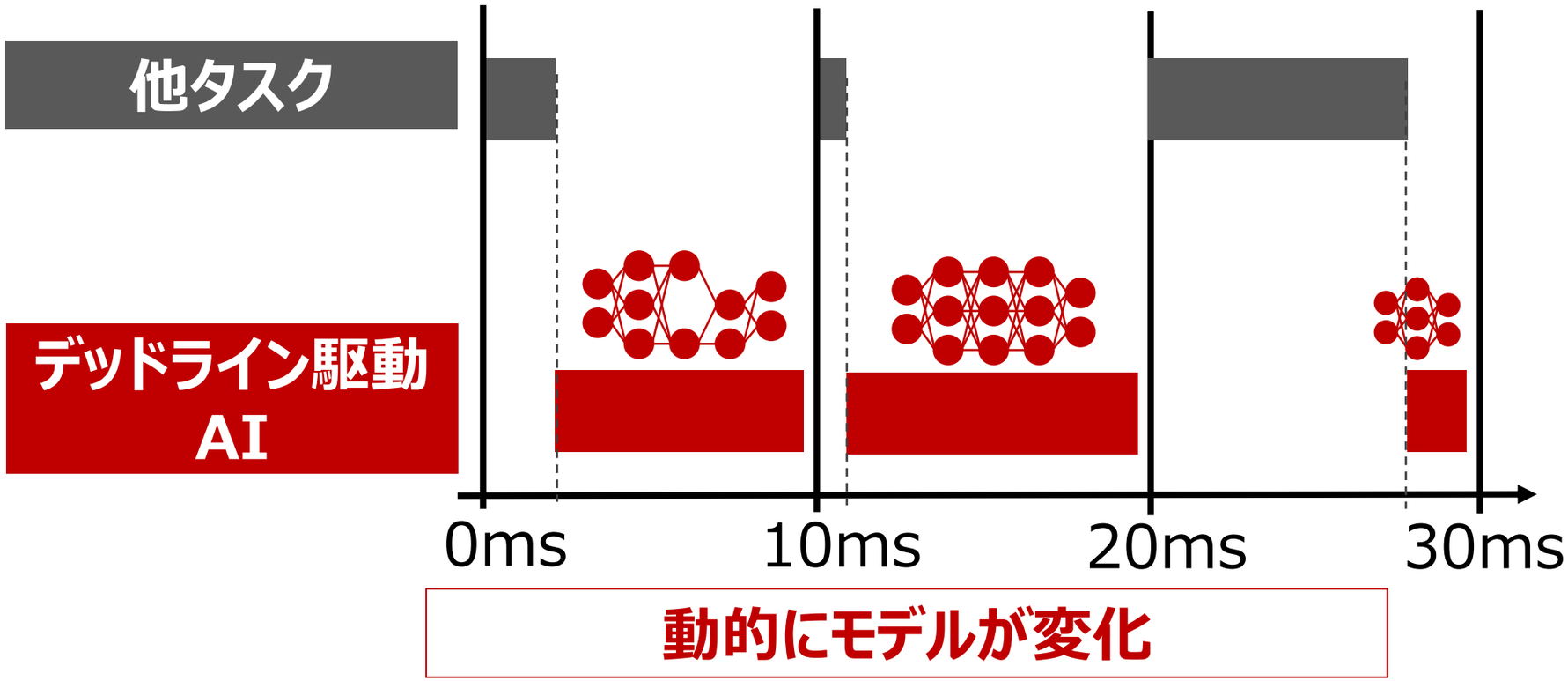
既存AIの問題点

- 他タスクの実行状況により、周期的タスクに対して許される処理時間は変化
- 最悪の場合の時間制約を考慮してAIモデルの小型化することで認識精度が低下



デッドライン駆動AI

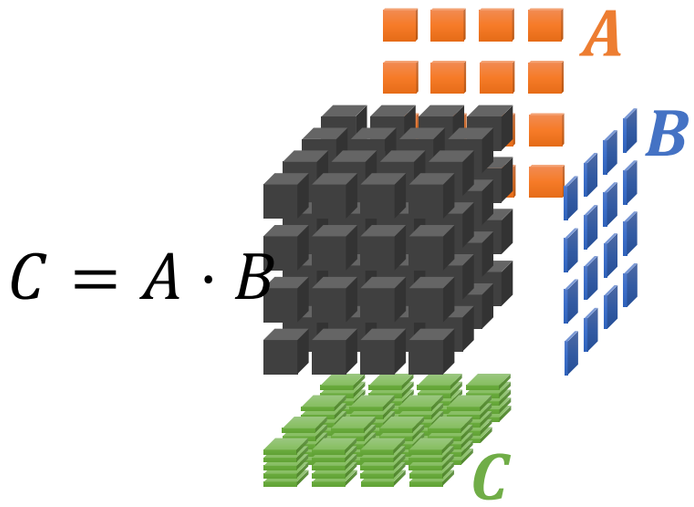
- デッドラインに応じて動的に計算コストを削減
- 締切を守りつつ許容時間の中で最大の精度を達成

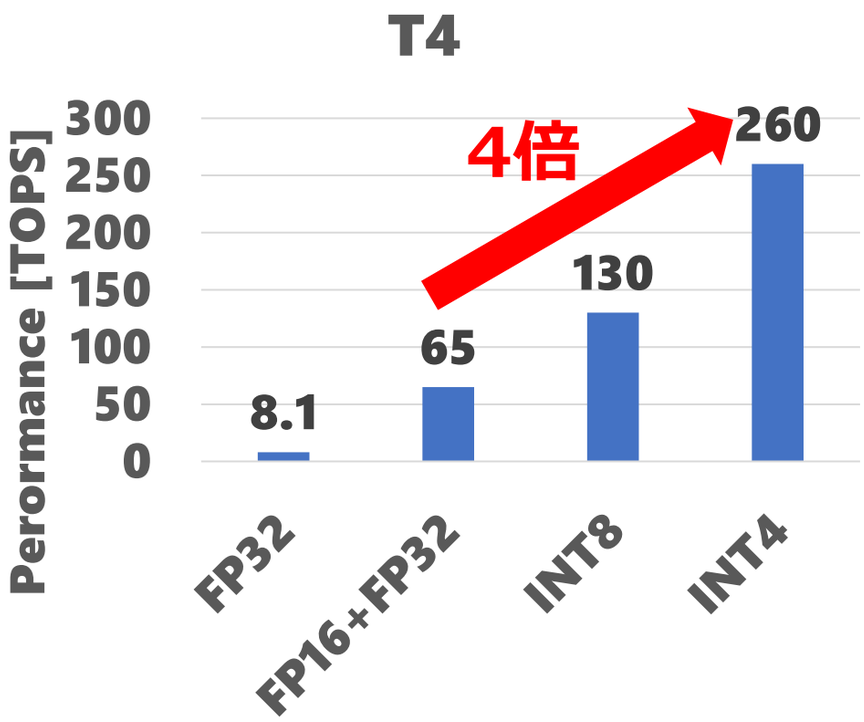


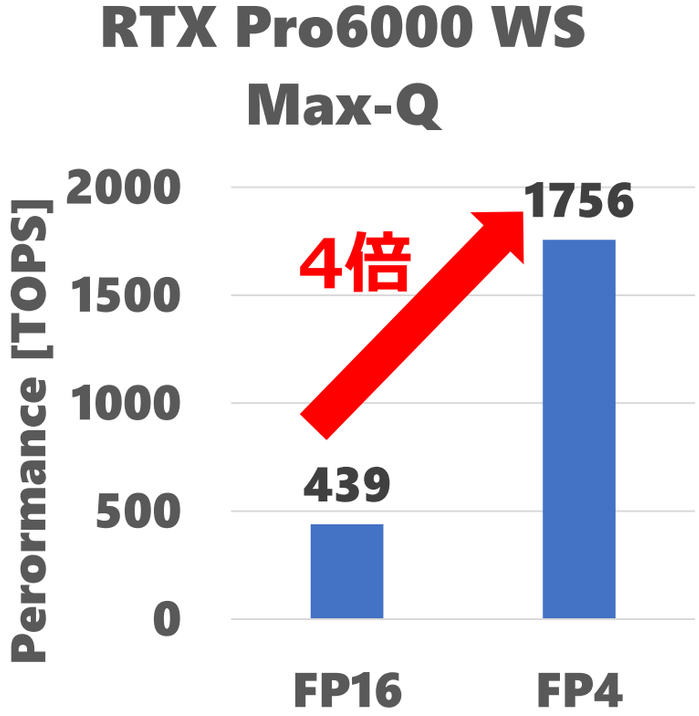
データ型毎の計算性能

- より低ビット演算器はより面積が小さい
- ハードウェアによっては、高ビット演算器を複数の低ビット演算器に分割して並列計算する機能がある

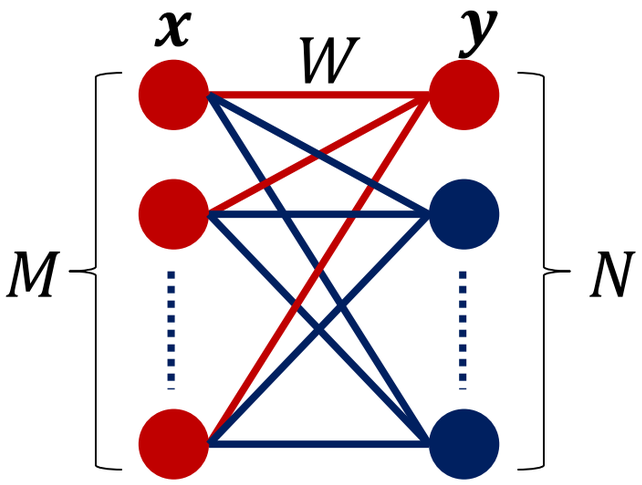
Nvidia Tensor Core







ニューラルネットワークの計算

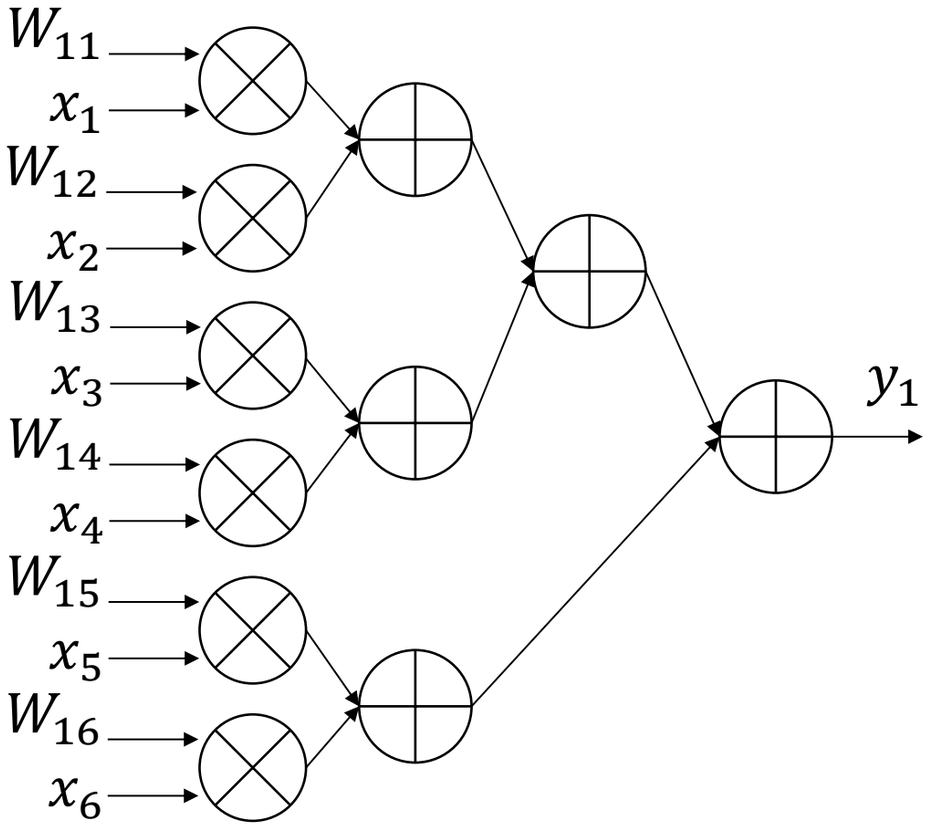


$$y_1 = W_{11}x_1 + W_{12}x_2 + \cdots + W_{1M}x_M$$

$$\boldsymbol{y} = g(W\boldsymbol{x})$$

$$= g\left(\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1M} \\ W_{21} & W_{22} & \cdots & W_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & \cdots & W_{NM} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}\right)$$

内積を計算する回路の例



残差量子化 [Yvinec+23]

$$x^{\mathrm{FP16}} = \begin{bmatrix} x^{\mathrm{FP16}}[1] \\ \vdots \\ x^{\mathrm{FP16}}[n] \end{bmatrix}$$

分解

第一次量子化項 $\frac{1}{\alpha_1} x_1^{\mathrm{INT4}} + \Delta x_1^{\mathrm{FP16}}$

分解

第二次量子化項 $\frac{1}{\alpha_1} x_1^{\mathrm{INT4}} + \frac{1}{\alpha_2} x_2^{\mathrm{INT4}} + \Delta x_2^{\mathrm{FP16}}$

$$w^{\mathrm{FP16}} = \begin{bmatrix} w^{\mathrm{FP16}}[1] \\ \vdots \\ w^{\mathrm{FP16}}[n] \end{bmatrix}$$

分解

$\frac{1}{\beta_1} w_1^{\mathrm{INT4}} + \Delta w_1^{\mathrm{FP16}}$

分解

$\frac{1}{\beta_1} w_1^{\mathrm{INT4}} + \frac{1}{\beta_2} w_2^{\mathrm{INT4}} + \Delta w_2^{\mathrm{FP16}}$

残差量子化による内積計算

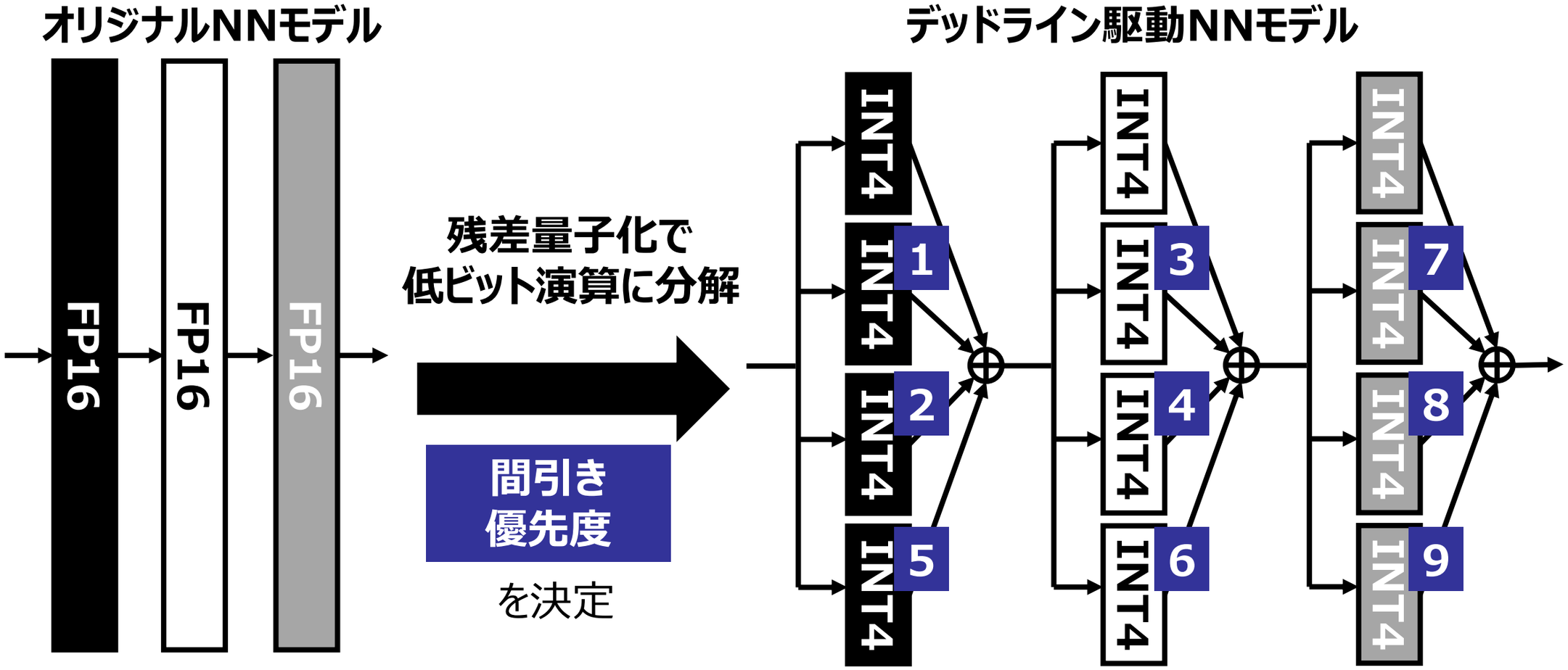
$$\boldsymbol{x}^{\mathrm{FP16}} \cdot \boldsymbol{w}^{\mathrm{FP16}} \approx \left(\frac{1}{\alpha_1}\boldsymbol{x}_1^{\mathrm{INT4}} + \frac{1}{\alpha_2}\boldsymbol{x}_2^{\mathrm{INT4}}\right)\left(\frac{1}{\beta_1}\boldsymbol{w}_1^{\mathrm{INT4}} + \frac{1}{\beta_2}\boldsymbol{w}_2^{\mathrm{INT4}}\right)$$

$$= \frac{\boldsymbol{x}_1^{\mathrm{INT4}}\boldsymbol{w}_1^{\mathrm{INT4}}}{\alpha_1\beta_1} + \frac{\boldsymbol{x}_1^{\mathrm{INT4}}\boldsymbol{w}_2^{\mathrm{INT4}}}{\alpha_1\beta_2} + \frac{\boldsymbol{x}_2^{\mathrm{INT4}}\boldsymbol{w}_1^{\mathrm{INT4}}}{\alpha_2\beta_1} + \frac{\boldsymbol{x}_2^{\mathrm{INT4}}\boldsymbol{w}_2^{\mathrm{INT4}}}{\alpha_2\beta_2}$$

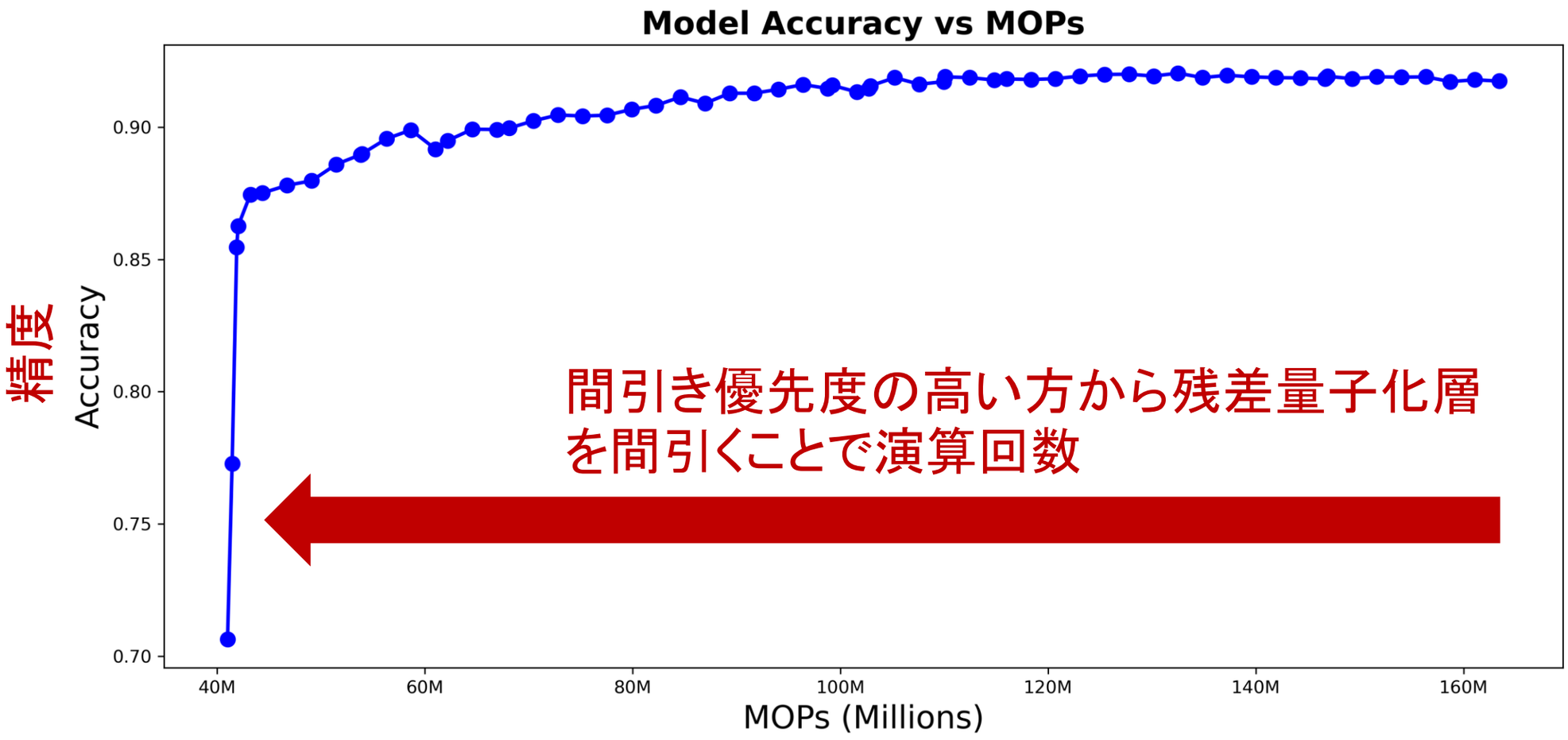
INT4の計算性能がFP16の４倍の場合、実行時間はあまり変わらない

デッドライン駆動AIの生成

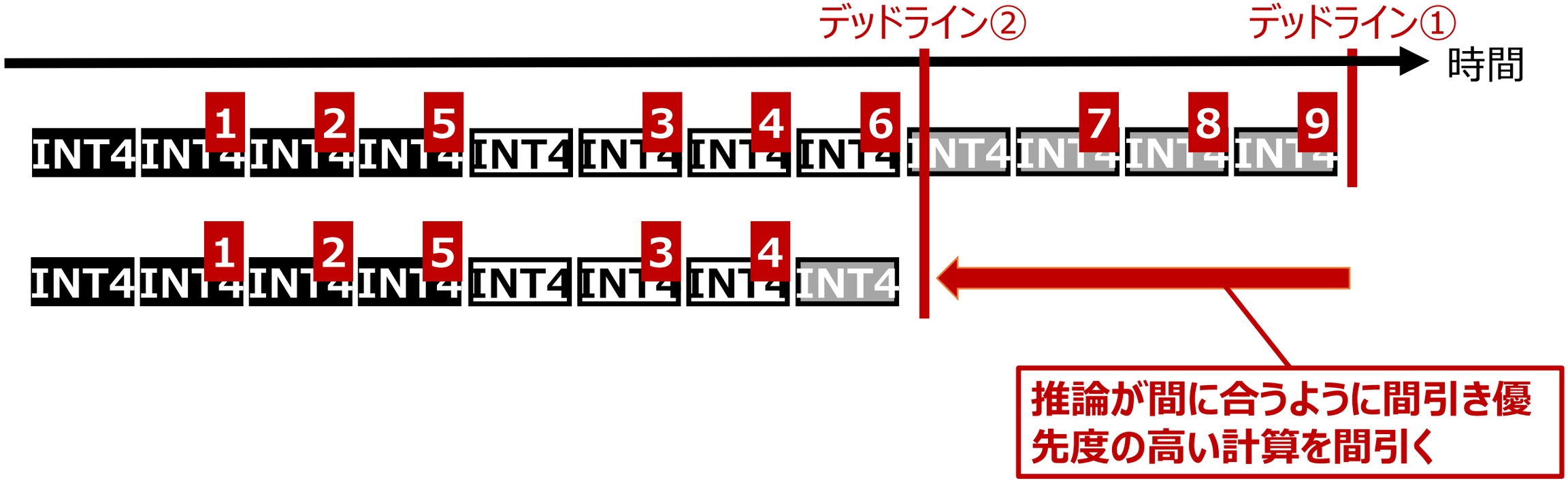
- 浮動小数点型のニューラルネットワークの各層を複数のINT4型の層にデッドラインに応じて重要度の低いINT4層から計算を省略



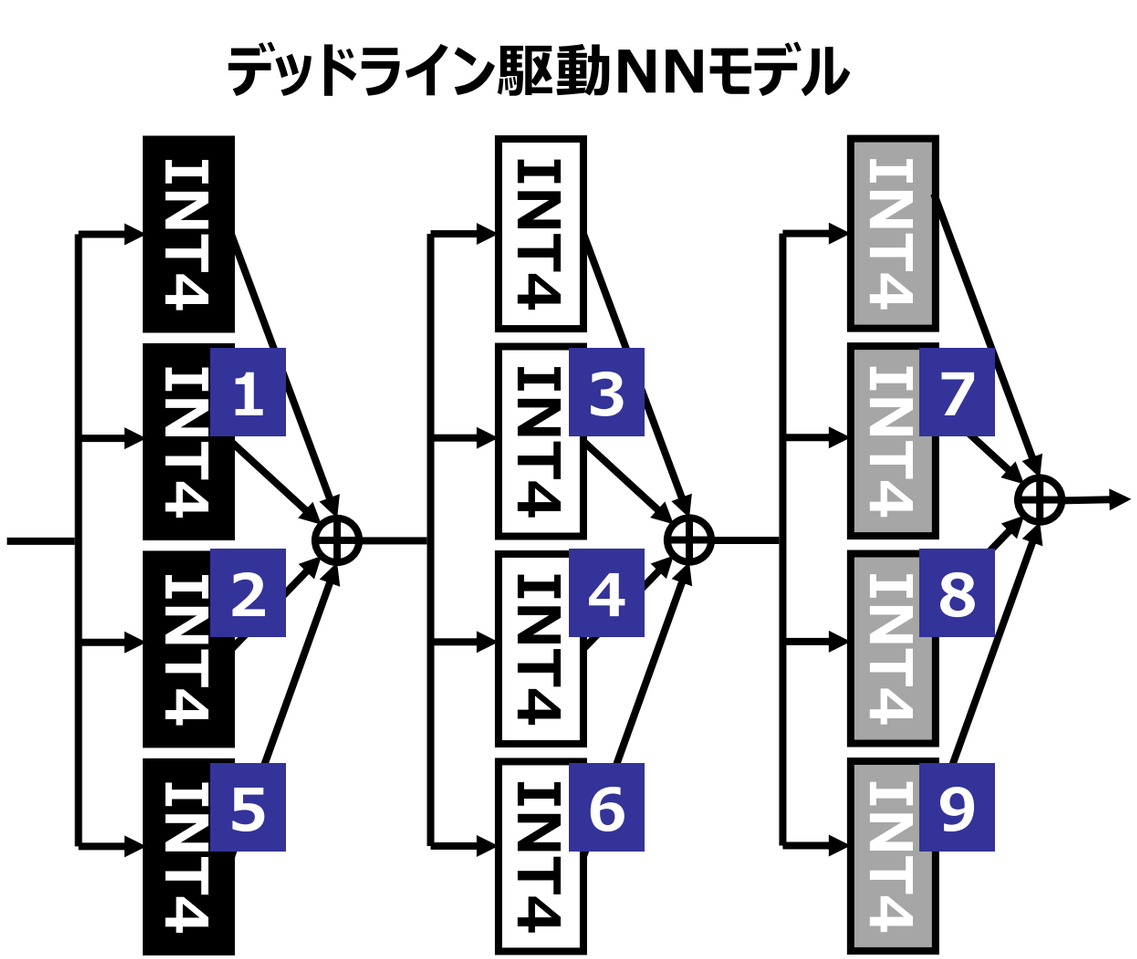
精度と演算回数の関係

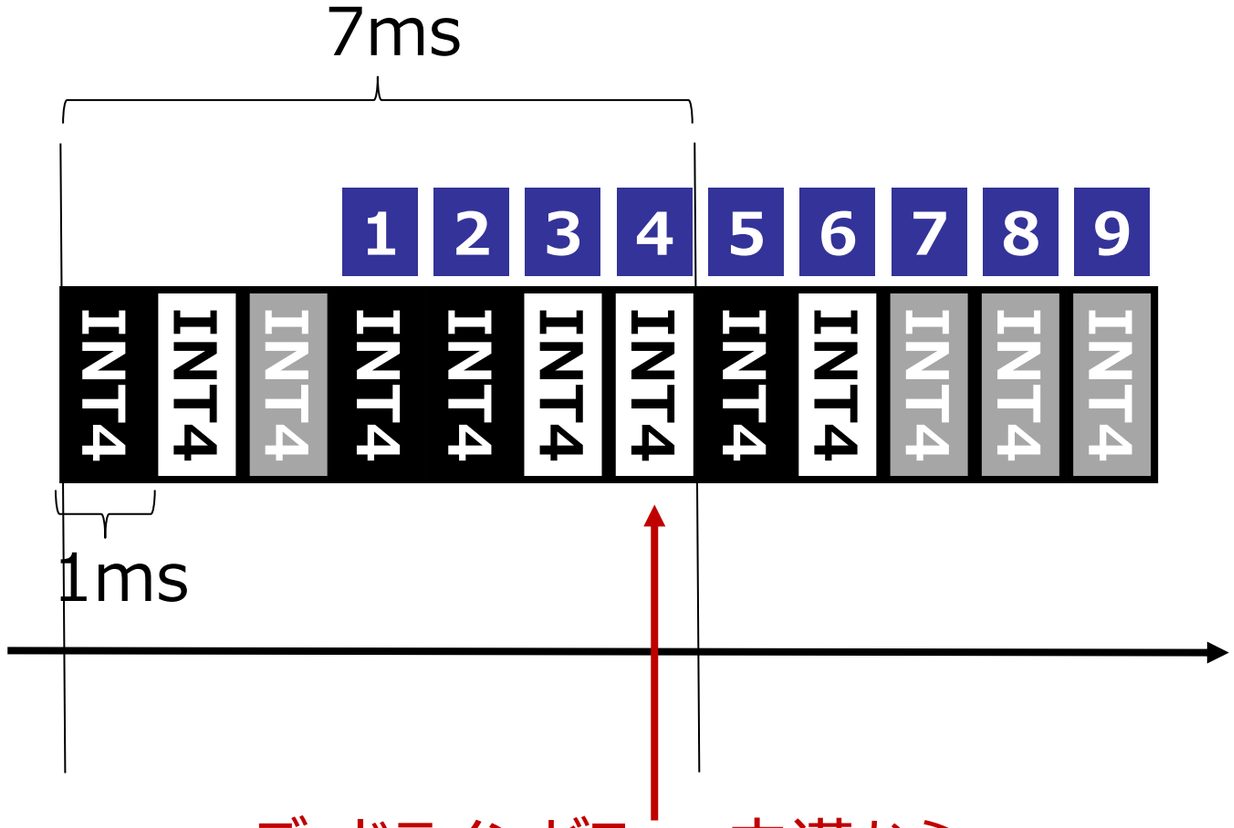

Model Accuracy vs MOPs
精度
Accuracy
0.90
0.85
0.80
0.75
0.70
40M
60M
80M
100M
120M
140M
160M
MOPs (Millions)
間引き優先度の高い方から残差量子化層
を間引くことで演算回数
演算回数

推論時間短縮の仕組み


デッドライン②
デッドライン①
時間
INT4 1 INT4 2 INT4 5 INT4 3 INT4 4 INT4 6 INT4 7 INT4 8 INT4 9
INT4 1 INT4 2 INT4 5 INT4 3 INT4 4 INT4
推論が間に合うように間引き優先度の高い計算を間引く

デッドラインに応じた実行ルール

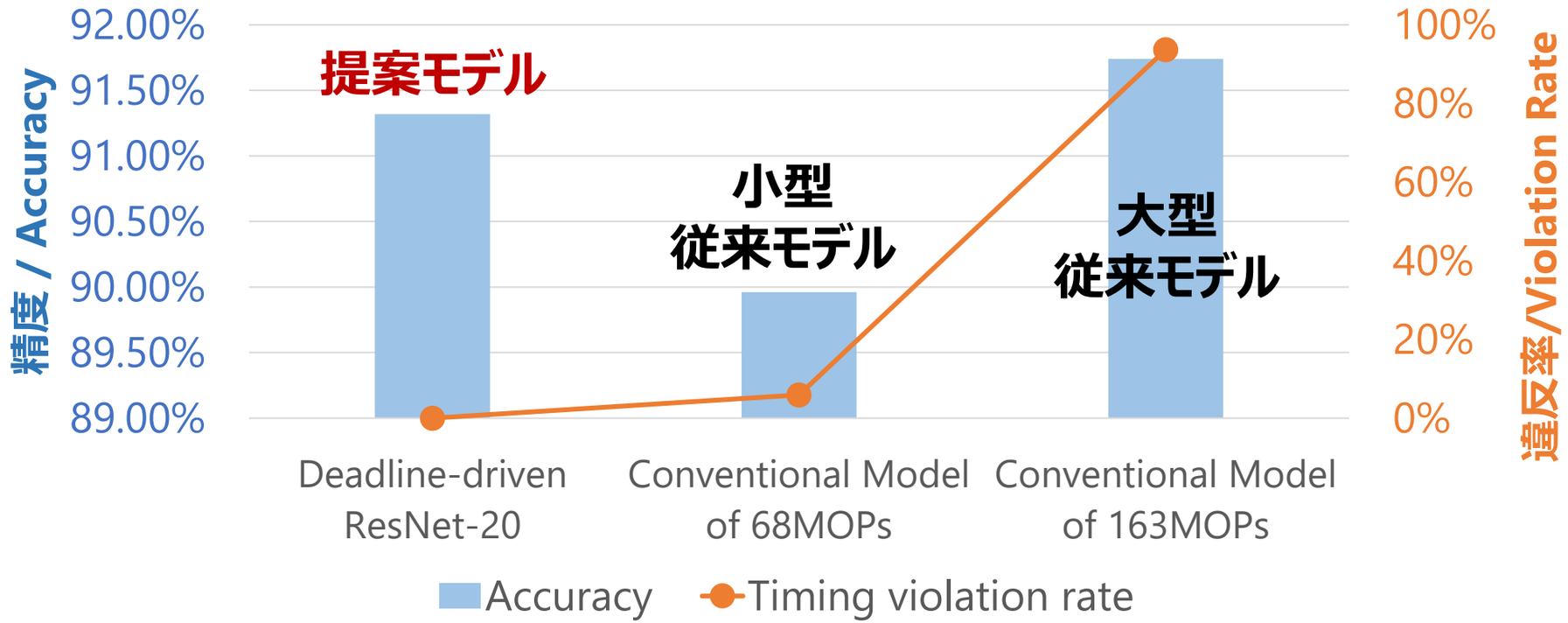




デッドラインが7ms未満なら
実行できない
＝デッドラインが7ms以上なら実行

評価結果

- 時間制約付きデータセット
  - CIFAR-10データセットに計算コスト制約を追加
- デッドライン駆動推論により違反が無しで、高精度の推論を実現



従来技術とその問題点

現在の一定の計算量を前提としたモデル設計では

- 精度と最悪実行時間の制約を満たすためには
  多くのモデル学習時間、労力を要する
- **精度か実行時間のどちらかの妥協**が必要

となる可能性がある

新技術の特徴・従来技術との比較

高精度モデルをデッドライン駆動モデルに変換することで

- **高精度を維持したまま、**
- **厳しい時間制約を達成すること**

が可能となった。これにより、

- **システムの安全性を高めると共に、**
- **モデル効率化の労力を削減できる**

想定される用途

- 自動運転、ロボティクス等、厳しい時間制約を有するリアルタイムシステムに適用することで安全性向上が期待できる

実用化に向けた課題

- 現在、デッドライン駆動モデル変換までは開発済み。
- CPU, GPU等の各プラットフォームにおいて、低ビット残差量子化層の実行を高速化する部分が未対応

社会実装への道筋

時期	取り組む課題や明らかにしたい原理等	社会実装へ取り組みについて記載
現在	・畳み込みニューラルネットワークを対象に、デッドライン駆動AIの原理と変換手法を確立	
1年後	・評価ツールの実現（元モデルとデッドライン駆動モデルに対して、各システム負荷におけるモデルの実行時間のばらつき評価を自動化）	• GAPファンドの獲得、ツールセットの準備、デモンストレーション実施 • JSTのCREST事業へ応募しh発展研究のための資金獲得を目指す
4年後	・LLM等、適応可能なニューラルネットワークを拡張 ・CPU, GPUでの実行最適化	• 共同研究先の応用を踏まえた最適化方針を決定
6年後	・デッドライン駆動AIの実応用向け統合基盤の確立	• ライセンスビジネスの立ち上げと産業界への展開

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

- 発明の名称　：ニューラルネットワークモデル装置及びニューラルネットワークモデルプログラム
- 出願番号　：特願2025-136266
- 出願人　：公立大学法人会津大学
- 発明者　：富岡洋一、奥山 祐市、愼 重弼、アルヴィ アリ ハイダー

産学連携の経歴

- 共同研究実績
  - 電機
  - 半導体
  - 電子部品
  - 宇宙産業
- 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）
  - 2025年-2027年　JAXA宇宙戦略基金（共同機関）

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