資料3

スーパーコンピュータ「富岳」政策対応枠における大規模言語モデル分散並列学習手法の開発について

Tokyo Tech

国立大学法人 東京工業大学
国立大学法人 東北大学
国立研究開発法人 理化学研究所
富士通株式会社
株式会社サイバーエージェント (2023年8月15日より参画)
国立大学法人東海国立大学機構 名古屋大学 (2023年8月15日より参画)
Kotoba Technologies Inc. (2023年8月15日より参画)

海外LLMの動向

Big Tech
- Google: Bard, Gemini
- Microsoft+OpenAI: GPT
- Meta: LLAMA, OPT
- Amazon+Anthropic: Claude

Startups
- EleutherAI: GPT-NeoX, Pythia
- TogetherAI: RedPajama
- Databricks: Dolly
- MosaicML: MPT
- StabilityAI: StableLM

Universities
- Stanford: Alpaca
- Tsinghua: GLM

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

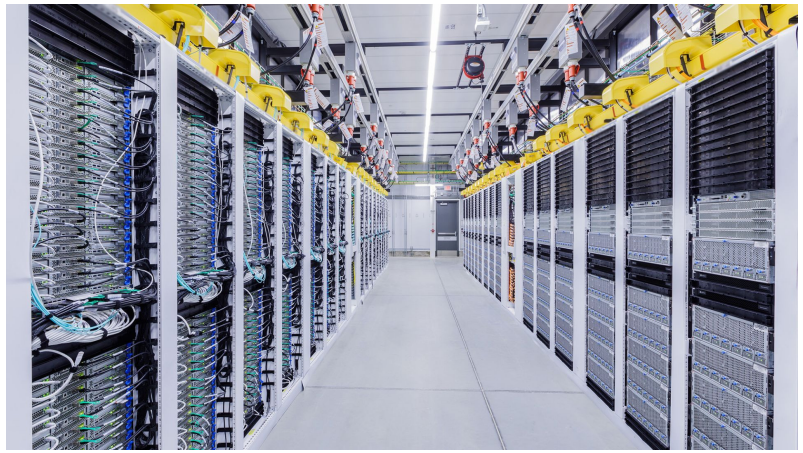
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Startups by Transformer paper authors
- Adept.AI: Ashish Vaswani, Niki Parmar (2021/11-)
- Character.AI: Noam Shazeer (2021/11-)
- Inceptive: Jakob Uszkoreit (2021/7-)
- Sakana.AI: Llion Jones (2023/8-)
- Cohere: Aidan N. Gomez (2019/9-)
- NEAR: Illia Polosukhin (2017/6-)

Huggingface Leaderboard
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Top500 Supercomputer

OpenAI (25,000 GPU?)



Frontier (37,888 GPU)

LUMI (20,480 GPU)

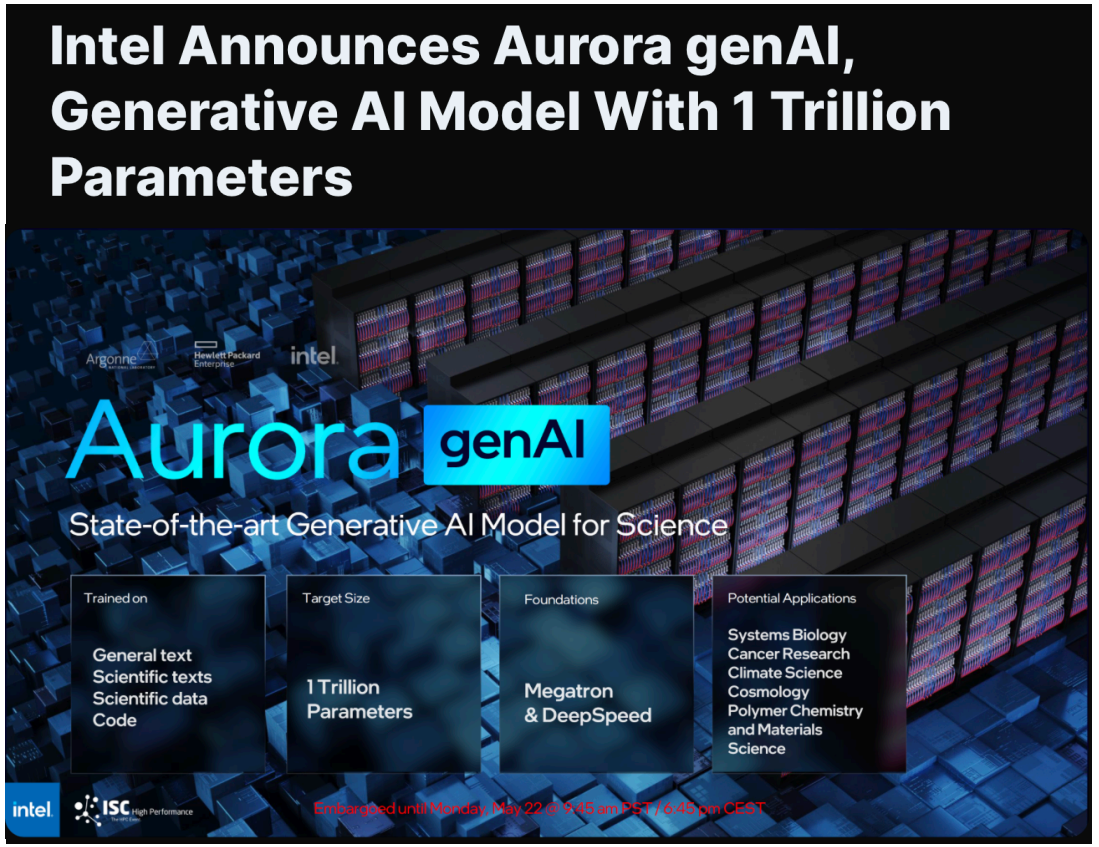
Aurora (63,744 GPU)

Fugaku (158,976 CPU)

El Capitan (20,000> GPU)

Aurora GenAI Project

- Data: General text
  - Neeraj Kumar (PNNL) & Andrew McNaughton (PNNL)
- Data: Biological / medical
  - Arvind Ramanathan (Argonne) & Miguel Vazquez (BSC)
- Data: Chemistry / materials
  - Eliu Huerta (Argonne) & Gihan Pinipitiya (PNNL)
- Data: Physics
  - Salman Habib (Argonne), Paolo Calafiura (LBL)
- Data: Climate / environment
  - Po-Lun Ma (PNNL)
- Models: Evaluation, alignment, safety, and ethics
  - Bo Li (UIUC) & Prasanna Balaprakash (ORNL)
- Models: Downstream instruct tuning
  - Venkat Vishwanath (Argonne) & Väinö Hatanpää (CSC)
- Models: Pretraining runtime mixing monitoring
  - Shantenu Jha (BNL) & Juan Durillo (Liebniz LRZ)
- Models: Inference / optimization & architecture / performance
  - Rio Yokota (Tokyo Tech.) & Jeyan Thiyagalingam (RAL)
- AI for Computer Science
  - Valerie Taylor (Argonne) & Pete Beckman (Argonne)
- TPC coordination, strategies, and policy
  - Charlie Catlett (Argonne) & David Martin (Argonne)



国内LLMの動向

Industry
- Cyberagent
- ELYZA
- Line
- PFN
- Rinna
- StabilityAI

Universities/Labs
- NII(LLM勉強会): MDX
- 理研(GPT-Fugaku): 「富岳」
- 産総研: ABCI
- 松尾研
- NICT

外国製

	model_name	平均点 ↓	MARC-ja	JSTS-pearson	JNLI	JSQuAD-F1	JCommonsenseQ
17	gpt-4	0.9048	0.9781	0.8902	0.7654	0.9492	0.941
8	stabilityai/StableBeluga2	0.771	0.9655	0.6629	0.4306	0.9092	0.8865
18	gpt-3.5-turbo	0.7698	0.954	0.8354	0.62	0.8407	0.5987
7	stabilityai/StableBeluga-13B	0.682	0.9515	0.7243	0.3135	0.87	0.5505
11	mosaicml/mpt-30b-instruct	0.4469	0.8511	0.03051	0.145	0.8154	0.3923
20	mosaicml/mpt-7b-instruct	0.377	0.8495	0.0574	0.2399	0.545	0.193
21	rinna/japanese-gpt-neox-3.6b-instruction-ppo	0.356	0.9558	0.2661	0.145	0.2202	0.193
4	line-corporation/japanese-large-lm-3.6b	0.3553	0.8463	0.3874	0.145	0.2046	0.193
1	matsuo-lab/weblab-10b-instruction-sft	0.3433	0.853	-0.09396	0.145	0.6192	0.193
6	stabilityai/japanese-stablelm-instruct-alpha-7b	0.3315	0.7989	0.03619	0.145	0.4842	0.193
19	rinna/japanese-gpt-neox-3.6b-instruction-sft	0.3141	0.9625	0.1303	0.145	0.1407	0.1921
3	stockmark/gpt-neox-japanese-1.4b	0.2681	0.8544	-0.001339	0.145	0.1493	0.193
5	line-corporation/japanese-large-lm-1.7b	0.2679	0.8544	-0.04536	0.145	0.1922	0.193
2	meta-llama/Llama-2-7b-chat-hf	0.2652	0.8976	0	0.145	0.09017	0.193
10	Salesforce/xgen-7b-8k-inst	0.2416	0.8212	-0.07148	0.145	0.1201	0.193
12	cyberagent/open-calm-7b	0.2411	0.8546	-0.01269	0.145	0.03199	0.1868

Weights & Biases Leaderboard
http://wandb.me/nejumi

Stability AI Leaderboard
https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable

LLM勉強会

主催者：黒橋（国立情報学研究所）

基盤センター：北海道大学情報基盤センター、東北大学サイバーサイエンスセンター、
東京大学情報基盤センター、東京工業大学学術国際情報センター、
名古屋大学情報基盤センター、京都大学学術情報メディアセンター、
大阪大学サイバーメディアセンター、九州大学情報基盤研究開発センター

大学の研究室：東北大学乾研究室、東北大学鈴木研究室、東京大学今泉研究室、
東京大学大関研究室、東京大学川原研究室、東京大学鶴岡研究室、
東京大学松尾研究室、東京大学宮尾研究室、東京大学谷中研究室、
東京大学吉永研究室、東京大学医療AI開発学講座、早稲田大学河原研究室、
東京工業大学岡崎研究室、東京工業大学横田研究室、お茶の水女子大学小林研究室、
名古屋大学武田・笹野研究室、京都大学黒橋研究室、大阪大学鬼塚研究室、
奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室、
愛媛大学人工知能研究室

研究所：理化学研究所AIP、理化学研究所GRP、産業技術総合研究所、国立情報学研究所、
情報通信研究機構、科学技術振興機構、情報・システム研究機構、

企業：NTT、LINE/ヤフー、レトリバ、サイバーエージェント、富士通、
Microsoft、Studio Ousia、プレシジョン、ZENKIGEN、Legalscape、
Turing、AWS、みらい翻訳、Megagon Labs、ストックマーク、
matsuri technologies、ファーストアカウンティング、東芝、Preferred Networks、
オムロンサイニックエックス、トヨタ、NTT Communications、バオバブ、Polaris.AI、
Stability AI Japan、マネーフォワード、メルカリ、NVIDIA、アステラス製薬株式会社、
パスコ、朝日新聞社、楽天、ELYZA、ベルシステム24、Lightblue、Intel

mdx

ABCI
AI Bridging Cloud Infrastructure

MDX: A100 x 128 x 60 days
ABCI: A100 x 480 x 60 days

フレームワーク: Megatron-DeepSpeed、GPT-Neox、
LLM-Foundry

日本語データ: Wikipedia: 1.4Bトークン (1.3M文書)
mC4 (ウェブコーパス): 136Bトークン (75M文書)
Common Crawl全量: 1Tトークン? (1B文書?)
JST J-STAGE (論文): 3Bトークン程度 (5.5M文書)
NDL WARP (ウェブアーカイブ): 1Tトークン以上?

モデルサイズ：1.3B、7B、13B、175B

「富岳」政策対応枠

目的：スーパーコンピュータ「富岳」を活用した大規模言語モデル分散学習手法の開発

成果物：研究開発の成果物を公開
- アカデミアや企業が幅広く使える大規模言語モデルの構築環境を整備やノウハウの共有
- 検証実験の過程で構築された大規模言語モデル（基盤モデル）の公開

→国内におけるAIの研究力向上に貢献し、学術および産業の両面で「富岳」の活用価値を高めることを目指す

参画組織：東京工業大学、東北大学、富士通株式会社、理化学研究所
2023年8月15日より参画:
株式会社サイバーエージェント、名古屋大学、Kotoba Technologies Inc.

「富岳」でGPTを学習するには？

GPT-4: $3x10^{25}$FLOPs (予測値)

GPT-3.5 (ChatGPT): $3x10^{24}$FLOPs (予測値)

GPT-3: $3x10^{23}$FLOPs

「富岳」:
FP32 6.76TFLOP/s x 158,976 = 1.07 EFLOP/s (理論ピーク性能)

GPT-4: 328 days x 4
GPT-3.5: 32 days x 4
GPT-3: 3.3 days x 4

実行性能を考慮すると

OpenAI:
BF16 312 TFLOP/s x 25,000 = 7.8 EFLOP/s (理論ピーク性能)

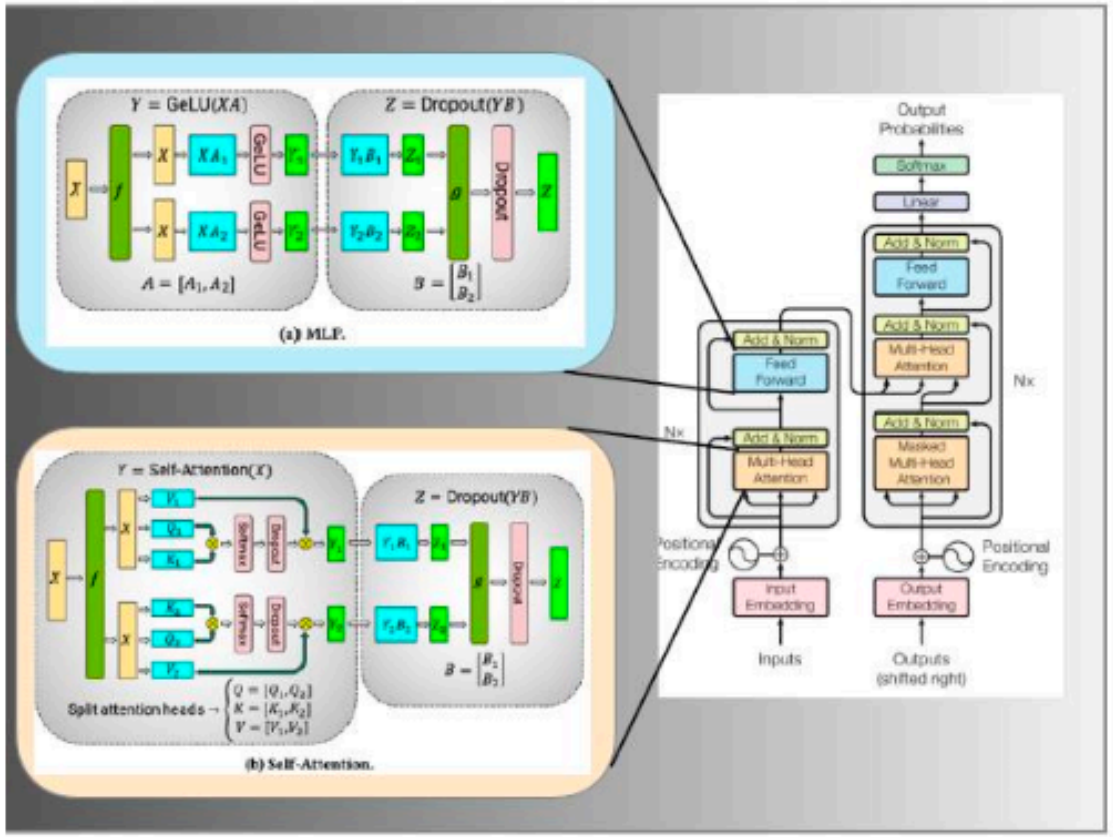
GPT-4: 45 days x 2
GPT-3.5: 4.5 days x 2
GPT-3: 11 hours x 2

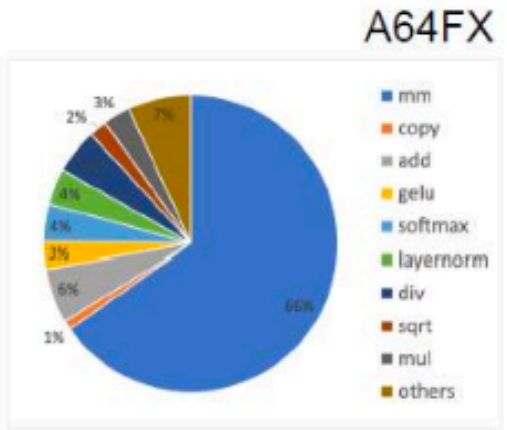
Googleは現在この5倍の資源でGeminiを学習中

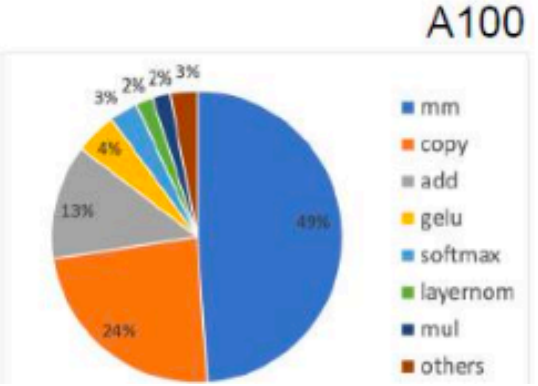
TransformerのA64FX向け性能最適化(演算部分)

- FLOPsの99%は小規模な密行列積
  - → A64FXでは66%、A100では49%の時間がこの部分に費やされている
  - → 現時点で理論ピークの1/3の性能になっており大幅に向上できる可能性がある

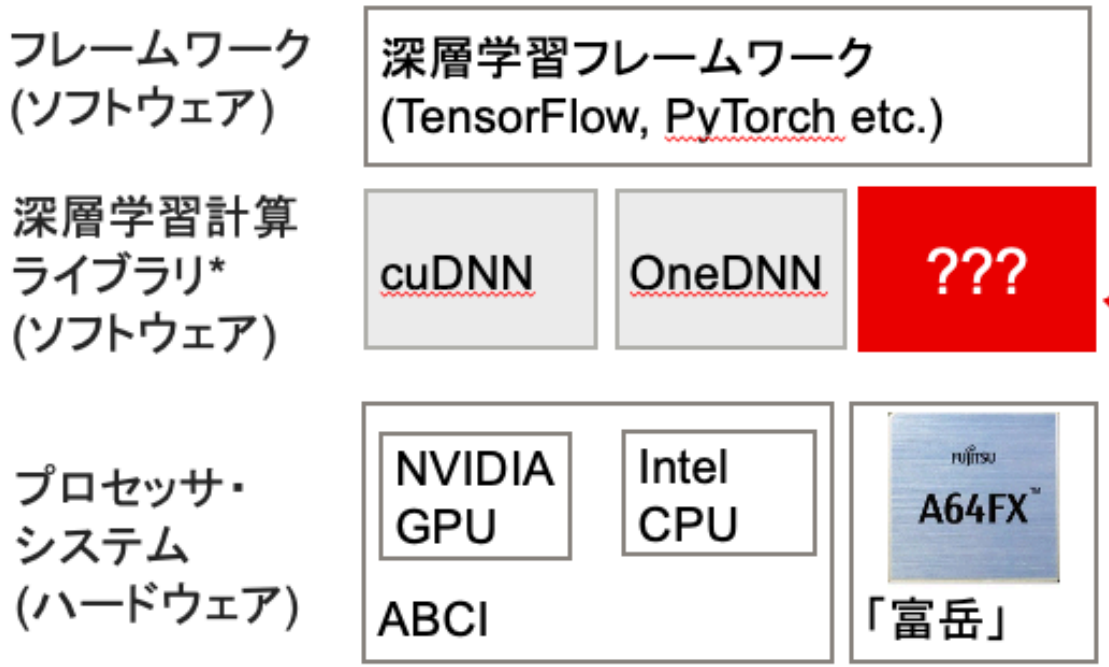
ある共通の入力 x に対しそれぞれの変換行列を適用して
$Q=XW_Q$, $K=XW_K$, $V=XW_V$ を用意する。
自分自身の要素との注目度合いを抽出する。







Transformerの性能を引き出すための専用ライブラリが必要



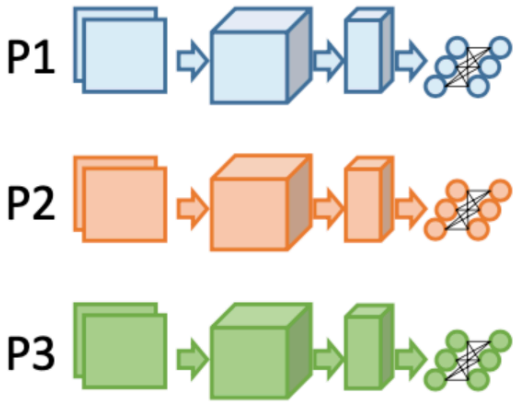
「富岳」搭載CPU (A64FX) 向けに
高速化された深層学習計算ライブラリ*が
存在しなかった

課題：深層学習計算ライブラリの移植**

*ライブラリ：特定の計算を高速に行うソフトウェア
**移植：ソースコード(プログラム)を書き換えること

分散並列学習

データ並列



データは分散

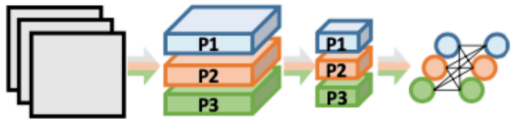
モデルは冗長

勾配の集約通信

課題：バッチサイズの
増大に伴う汎化性能の低下

解決策：正則化・最適化
手法の工夫

テンソル並列



データは冗長

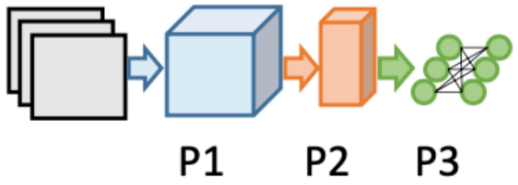
モデルは分散

活性の集約通信

課題：通信頻度の増加

解決策：通信のオーバーラップ

パイプライン並列



データは冗長

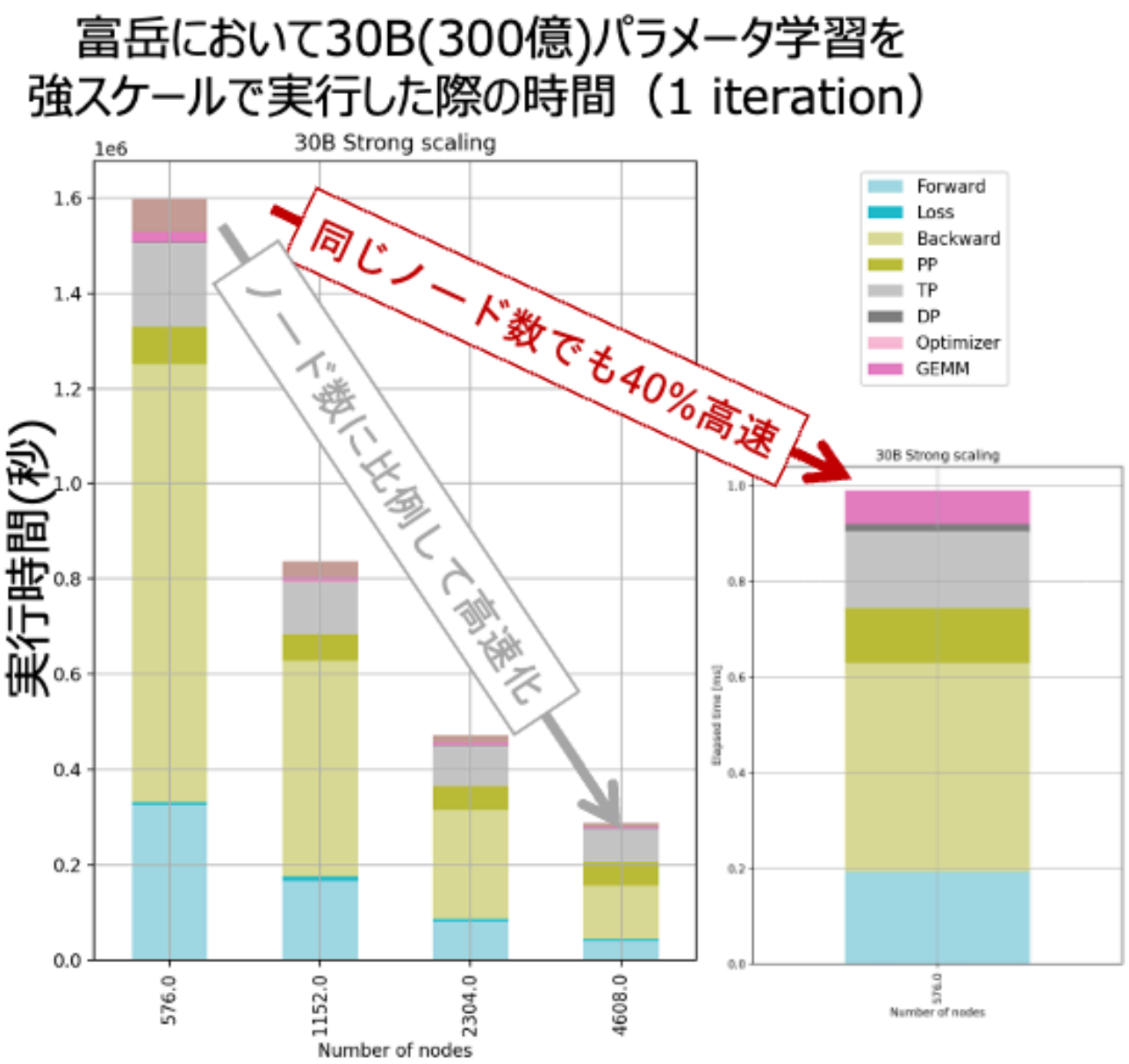
モデルは分散

活性のI対I通信

課題：パイプラインバブル

解決策：パイプライン
の工夫

Transformerの「富岳」上での高速化



５月に研究を開始した時点では理論ピーク性能の10%だったものが9月現在は25%に

→目標は50%

「富岳」のTofuネットワークは高速であるため
ノード数に比例して高速化
→テンソル並列、パイプライン並列を駆使することでデータ並列数を抑えている

日本語データ

コーパス	ソース	サイズ(GB)	補足説明
CyberAgent CommonCrawl (ja)	Common Crawl	1,300	サイバーエージェント社が独自に構築
mc4 (ja)	Common Cwawl	97	
日本語Webコーパス	東北大学	63	東北大学が独自に構築
Oscar (ja)	Common Crawl	37	
Wiki-40B(ja)	日本語Wikipedia	2	
WMT22 (en-ja)	Multiple	9	対訳データ（翻訳能力＋英語からの知識移転）
News (ja, en-ja)	毎日、朝日、日経、読売新聞	~1.5/年	商用利用には別途契約が必要（現在交渉中）
Twitter (ja)	東北大学	10~	言語モデルの事前学習には適さない（短文）

今後のスケジュール

	~2023/7	~2023/9	~2023/11	~2024/1	~2024/3	2024/4~
LLM高速化	Transformerの高速化 (効率15%) (済)	Transformerの高速化 (効率30%)	Transformerの高速化 (効率50%)			
LLM事前学習	深層学習フレームワークの富岳への移植 (済)	300Mパラメータのモデルの学習	1.3B, 7B, 13Bパラメータのモデルの学習	30Bパラメータのモデルの学習	175Bパラメータのモデルの学習	
LLM事後学習				LLM勉強会で事後学習		
公開						ソースコード、モデルの公開

次世代計算基盤への期待

NVIDIA H100
FP64: 34 TFLOP/s
FP32: 67 TFLOP/s
FP16: 1,979 TFLOP/s
FP8: 3,958 TFLOP/s
Memory Bandwidth: 3.35 TB/s

AMD MI250X
FP64: 95.7 TFLOP/s
FP32: 95.7 TFLOP/s
FP16: 383 TFLOP/s
Memory Bandwidth: 3.2 TB/s

Intel Ponte Vecchio
FP64: 52 TFLOP/s
FP32: 52 TFLOP/s
FP16: 832 TFLOP/s
Memory Bandwidth: 12.8 TB/s

- FP16の性能で圧倒的な差がついている
  →深層学習では実効性能においてこの差がそのままでている
- 深層学習の推論が最も大きな市場
  →チップメーカーは低精度行列演算で差をつけてくる
- 推論ではメモリ帯域がネックになる
  →低精度行列演算が十分に速ければの話
- メモリ帯域と低精度行列演算は相反する要件ではない
  →メモリ帯域の確保は大前提で、演算性能が勝負どころ

Fujitsu A64FX
FP64: 3.38 TFLOP/s
FP32: 6.76 TFLOP/s
FP16: 13.5 TFLOP/s
Memory Bandwidth: 1.024 TB/s