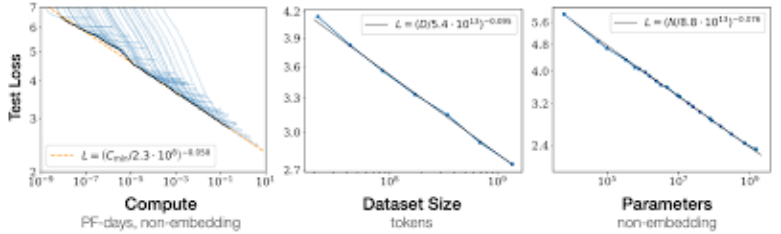
国内外の大規模言語モデルに関する取り組みについて

東京工業大学
学術国際情報センター
横田 理央

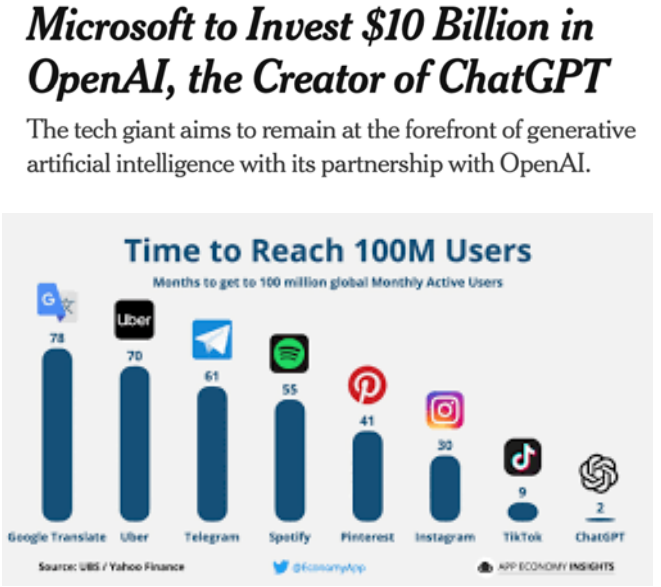
ChatGPT公開以前の流れ

2019/06/22 MicrosoftがOpenAIに10億ドル(当時 約1100億円)を投資
2020/01/23 OpenAIが言語生成モデルに関するScaling Lawの論文を発表
2020/03/28 OpenAIが言語生成モデルGPT3に関する論文を発表
2020/06/11 OpenAIがGPT3のAPIを公開
2020/09/22 OpenAIがMicrosoftに対してGPT3のソースコードを公開
2021/01/05 OpenAIが画像+言語モデルCLIPと画像生成モデルDALL-Eを発表
2021/06/29 GithubがGPT3にコードを追加学習したCodexを利用したCopilotを発表
2021/07/28 Free Software FoundationがGithub Copilotに対するライセンス上の懸念を表明
2021/09/10 Naverが言語生成モデルHyperClovaを発表
2022/02/22 DeepMindがコード生成モデルAlphaCodeを発表
2022/03/29 DeepMindが言語生成モデルChinchillaを発表
2022/05/23 Googleが画像生成モデルImagenを発表
2022/07/20 OpenAIが画像生成モデルDALL-E2を発表
2022/07/22 BigScience (HuggingFace,CNRS,GENCI)が多言語モデルBloomを発表
2022/08/04 精華大学が言語生成モデルGLM-130Bを発表
2022/08/22 Stability AIが画像生成モデルStable Diffusionを発表
2022/11/10 DeepMindが汎用モデルGatoを発表
2022/11/15 DeepMindが画像+言語モデルFlamingoを発表



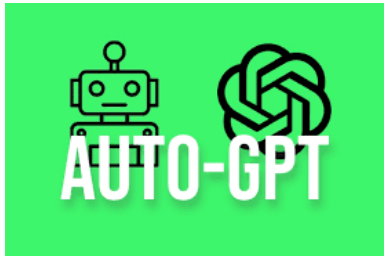
ChatGPT公開以降の時系列

2022/11/30 OpenAIがChatGPTを発表
2022/12/04 公開後5日で100万ユーザを突破
2022/12/05 Stack OverflowがChatGPTで生成された投稿を禁止
2022/12/21 GoogleがChatGPTの自社への脅威に対してCode Red(非常事態)を宣言
2023/01/23 MicrosoftがOpenAIに100億ドル(当時 約1.3兆円)を投資 →
2023/02/01 OpenAIが月額$20の有料サービスChatGPT Plusを開始
2023/02/02 公開後約２ヶ月で1億ユーザを突破 →
2023/02/03 自民党AIの進化と実装に関するプロジェクトチーム (第１回)
2023/02/06 GoogleがChatGPTに対抗して対話型AIのBardを限定公開
2023/02/07 Microsoftが検索エンジンBingにChatGPTを導入
2023/02/09 ChatGPTが米国の医師国家試験に合格できるレベルとの論文が発表される
2023/02/24 Metaが言語生成モデルLLaMAのソースコードを非商用ライセンスで公開
2023/03/01 OpenAIがChatGPTとWhisper(音声認識)のAPIを公開 (ユーザデータは学習に使わないことを宣言)
2023/03/09 MicrosoftがAzure OpenAI ServiceでChatGPTを提供開始
2023/03/10 Googleが言語生成モデルPaLM2を発表
2023/03/14 ChatGPT PlusでGPT-4が利用可能に →
2023/03/14 Anthropicが言語生成モデルClaudeを発表
2023/03/16 Microsoft 365 CopilotでWord,Outlook,TeamsなどからAIアシスタントが利用可能に

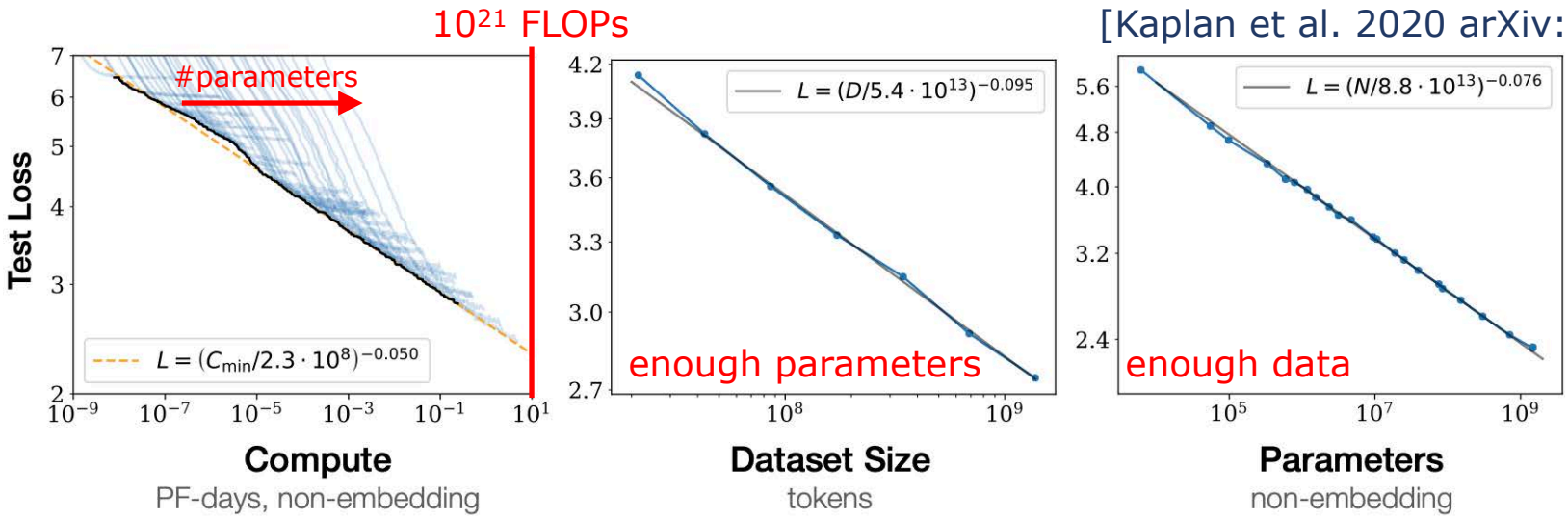


ChatGPT公開以降の時系列

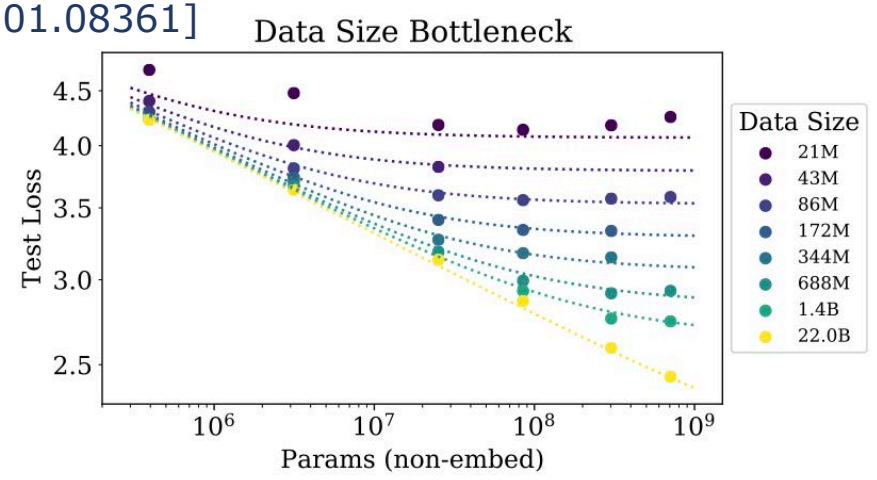
2023/03/20 ChatGPTから氏名,email,住所,カード番号などの個人情報が漏えい
2023/03/20 GitHubがGPT-4を搭載したCopilot Xを公開
2023/03/30 ChatGPTのAPIを利用したAI agentであるAuto-GPTが公開
2023/03/30 UC Berkeley,CMU,StanfordなどがGPT-4の会話データで微調整したVicunaを発表
2023/03/31 イタリアがChatGPTの利用を禁止 (4/28に撤回)
2023/05/19 G7広島サミットにて「広島AIプロセス」立ち上げ
2023/05/21 LLM勉強会 (第１回)
2023/06/01 「富岳」政策対応枠 利用開始
2023/06/05 TTI (Abu Dhabi)がFalcon-40Bのソースコードを商用ライセンスで公開 (09/06に180Bも)
2023/07/11 Anthropicが言語生成モデルClaude2を発表
2023/07/18 Metaが言語生成モデルLLaMA2のソースコードを限定的な商用ライセンスで公開
2023/09/07 Turingが画像+言語モデルHeronを公開
2023/10/02 AmazonがAnthropicに12億ドル(約1800億円)を投資
2023/10/03 産総研の生成AI開発支援プログラムに国立情報学研究所(NII)とELYZA社が採択
2023/10/26 Google, Microsoft, Anthropic, Open AI が AI Safety Fundに10億ドル(1500億円)を投資
2023/11/17 Sam AltmanがOpenAIから解任→Microsoftに移る→OpenAIに戻る
2023/12/06 AI Alliance立ち上げ：日本からは東大,慶應,SB Institute, sakana.ai, Sonyなどが参加
2023/12/07 GoogleがGemini, Alphacode2を発表



Transformerのスケール則(Scaling Law)



[Kaplan et al. 2020 arXiv:2001.08361]



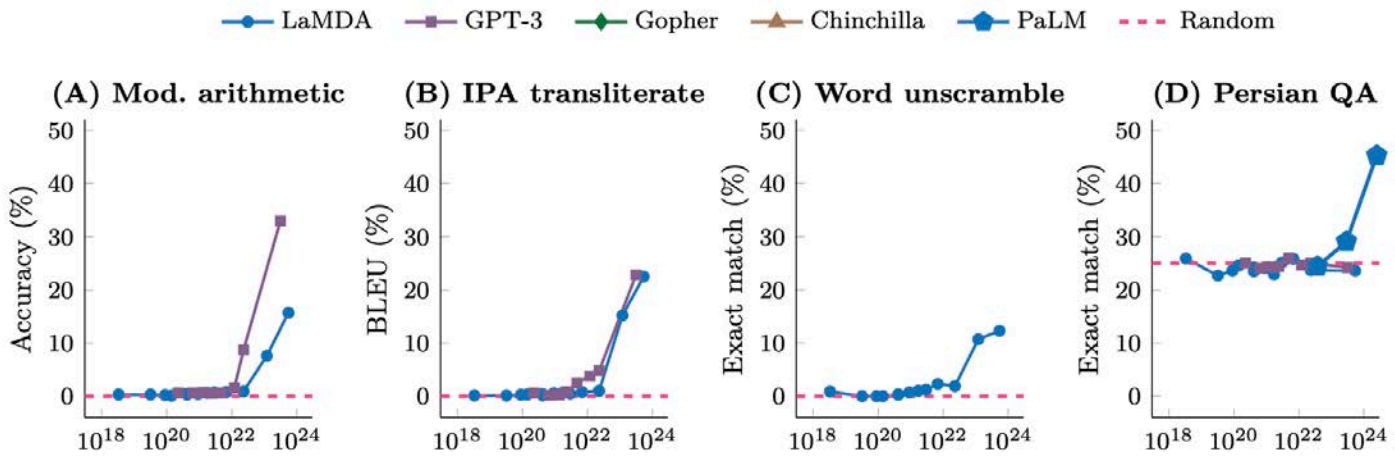
1:20のデータが
必要なことは
昔から知られていた

- パラメータ数が十分な場合、データ量に対してTest Lossはべき乗則に従って減少する
- データ量が十分な場合、バラメータ数に対してTest Lossはべき乗則に従って減少する
- 必要な計算資源 [FLOPs] = パラメータ数 x テータ量 [tokens] x 5.7
- どのくらいの計算資源(予算)を投入すればどのくらいの性能のものができるかが予想できる
- データ量はパラメータ数に比例させる必要がある （データ量：パラメータ数=20:1）
- GPT-3は データ量：パラメータ数が2:1 になっており圧倒的にデータが不足している
- 最近は推論コストを抑えるためにパラメータ数に対してデータ量を増やす傾向が顕著に

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

[Hoffmann et al. 2022 arXiv:2203.15556]

創発性(Emergence)



良く参照されるこちらの図では10^{22}を境に創発性が現れているように見える

実際はタスクによって創発性が現れる演算量はバラバラであり、10^{23}でも解けないタスクはまだまだ無数にある

演算量に対する性能向上がべき乗則に従うということに矛盾していないか？
→べき乗則に従うのはTest Lossであって、特定のタスクの性能ではない

原著論文ではもっと幅広いタスクやモデルに対して調べた結果が載っており
それを見ると10^{20}から10^{24}(1万倍)の範囲でばらついている
→この論文に載っていないような難しいタスクは無数にある

片対数グラフだから急峻な立ち上がりのように見えているだけでは？
→スケール則のグラフも片対数なのに直線(べき乗則)になっている

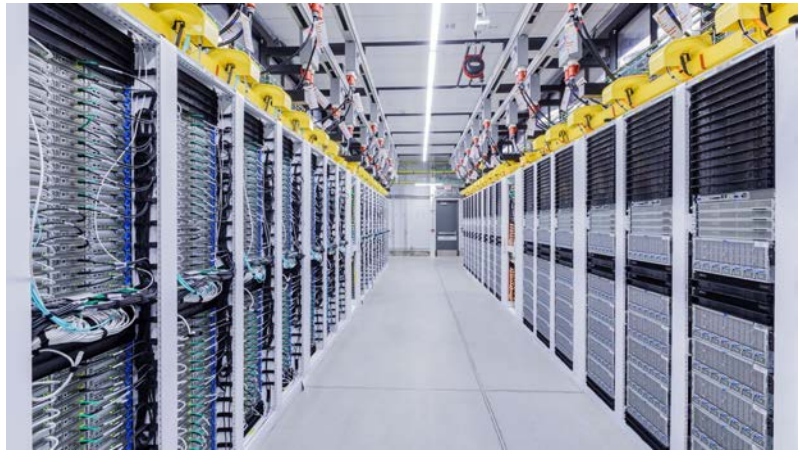
OpenAI, Googleの視点からみると
スケール則 (2019) → 資金調達や社内での正当化をするのに好都合
創発性 (2023) → 後続の企業を諦めさせるのに好都合

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

GPT-4(3x10^{24} FLOPs)をスパコンで学習するには？

理論ピークの半分がでた場合の概算

OpenAI (A100x25,000) 90日



Frontier (MI250Xx37,888) 75日

LUMI (MI250Xx20,480) 140日

Aurora (PVCx63,744) 30日?

Fugaku (A64FXx158,976) 700日

ABCI (A100x960+V100x4352) 770日

深層学習においては、国内の計算資源は国外のものと比べて一桁少ない

→計算資源・ノウハウ・人材の観点からも海外との連携が喫緊の課題

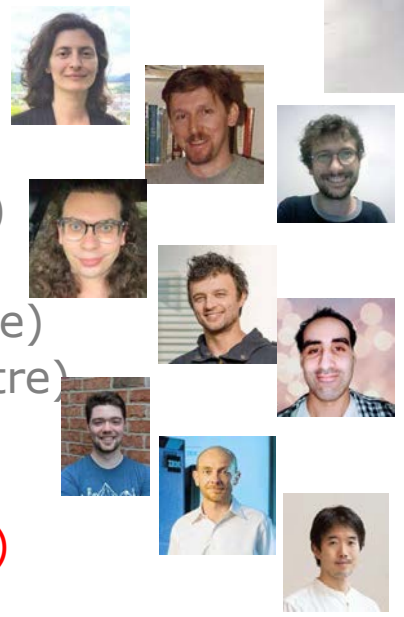
海外との連携 その１（INCITE）

Type: New
Title: "Scalable Foundational Models for Transferable Generalist AI"

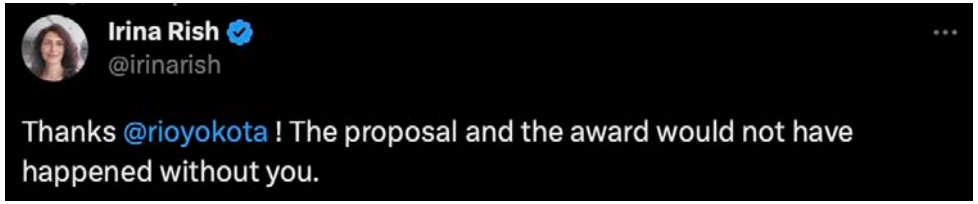
Principal Investigator: Irina Rish, University of Montreal, Mila - Quebec AI Institute

6M GPU hours on Summit

- Project Members
  - Irina Rish (University of Montreal)
  - Sergey Panitkin (University of Montreal)
  - Guillaume Dumas (University of Montreal)
  - Stella Biderman (EleutherAI)
  - Jenia Jitsev (Jülich Supercomputing Centre)
  - Mehdi Cherti (Jülich Supercomputing Centre)
  - Quentin Anthony (Ohio State University)
  - Guillermo Cecchi (IBM Research)
  - Rio Yokota (Tokyo Institute of Technology)

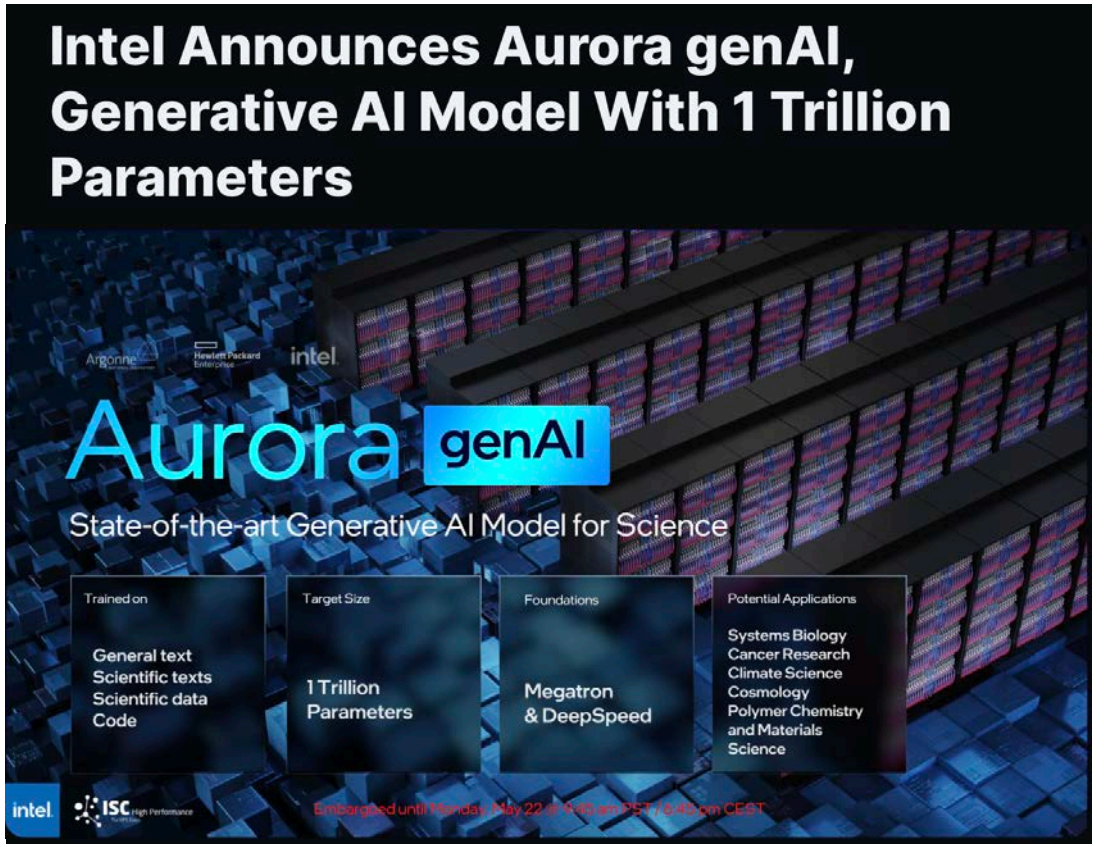


	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion



海外との連携 その２（Aurora GenAI）

- 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: Architecture & Compute Performance
  - Rio Yokota (Tokyo Tech.) & Jeyan Thiyagalingam (RAL) &
  - Shantenu Jha (BNL) & Juan Durillo (Liebniz LRZ)
- 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

企業

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

大学・研究所

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

ベンチマークのtrain dataで学習

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

Rakuda
https://yuzuai.jp/benchmark

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

日本語のベンチマーク性能

	model_name	AVG ↓	MARC-ja-balanced	JNLI-balanced	JSQuAD-F1	JCommonsenseQA
7	gpt-4	0.8978	0.9594	0.7416	0.9492	0.941
2	llm-jp/llm-jp-13b-instruct-full-jaster-dolly-	0.8302	0.5476	0.9051	0.9626	0.9053
13	stabilityai/StableBeluga2	0.8262	0.944	0.5651	0.9092	0.8865
4	anthropic.claude-v1	0.7566	0.9028	0.7324	0.6371	0.7542
10	gpt-3.5-turbo	0.7485	0.9186	0.636	0.8407	0.5987
14	anthropic.claude-v2	0.7189	0.8597	0.6344	0.7022	0.6792
12	stabilityai/StableBeluga-13B	0.7029	0.9525	0.4388	0.87	0.5505
11	mosaicml/mpt-30b-instruct	0.6004	0.8604	0.3333	0.8154	0.3923
9	lightblue/open orca_stx	0.5858	0.5	0.4232	0.8898	0.5299
8	Xwin-LM/Xwin-LM-70B-V0.1	0.5525	0.7643	0.3333	0.8693	0.2431
5	matsuo-lab/weblab-10b-instruction-	0.4951	0.5004	0.3371	0.8532	0.2895
1	stabilityai/japanese-stablelm-instruct-	0.4621	0.6509	0.3333	0.5808	0.2833
33	rinna/japanese-gpt-neox-3.6b-instruction-ppo	0.4248	0.9524	0.3333	0.2202	0.193
30	Xwin-LM/Xwin-LM-7B-V0.1	0.4147	0.4997	0.3333	0.6327	0.193

外国製

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

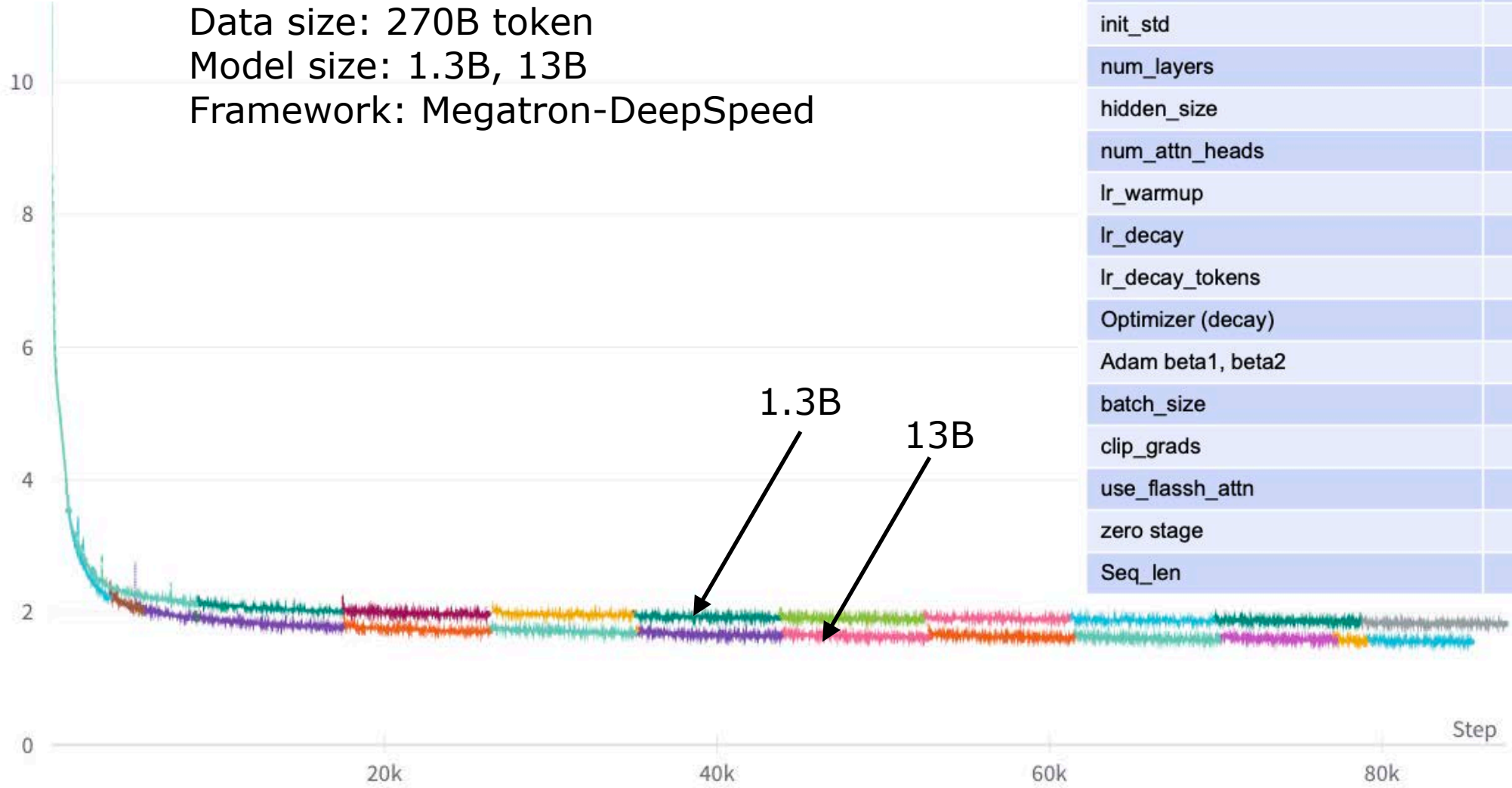
ABCI
AI Bridging Cloud Infrastructure

MDX: A100 x 128 x 60 days
ABCI: A100 x 480 x 60 days
フレームワーク: Megatron-DeepSpeed
モデルサイズ：1.3B、13B、175B

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)

LLM勉強会

Data size: 270B token
Model size: 1.3B, 13B
Framework: Megatron-DeepSpeed

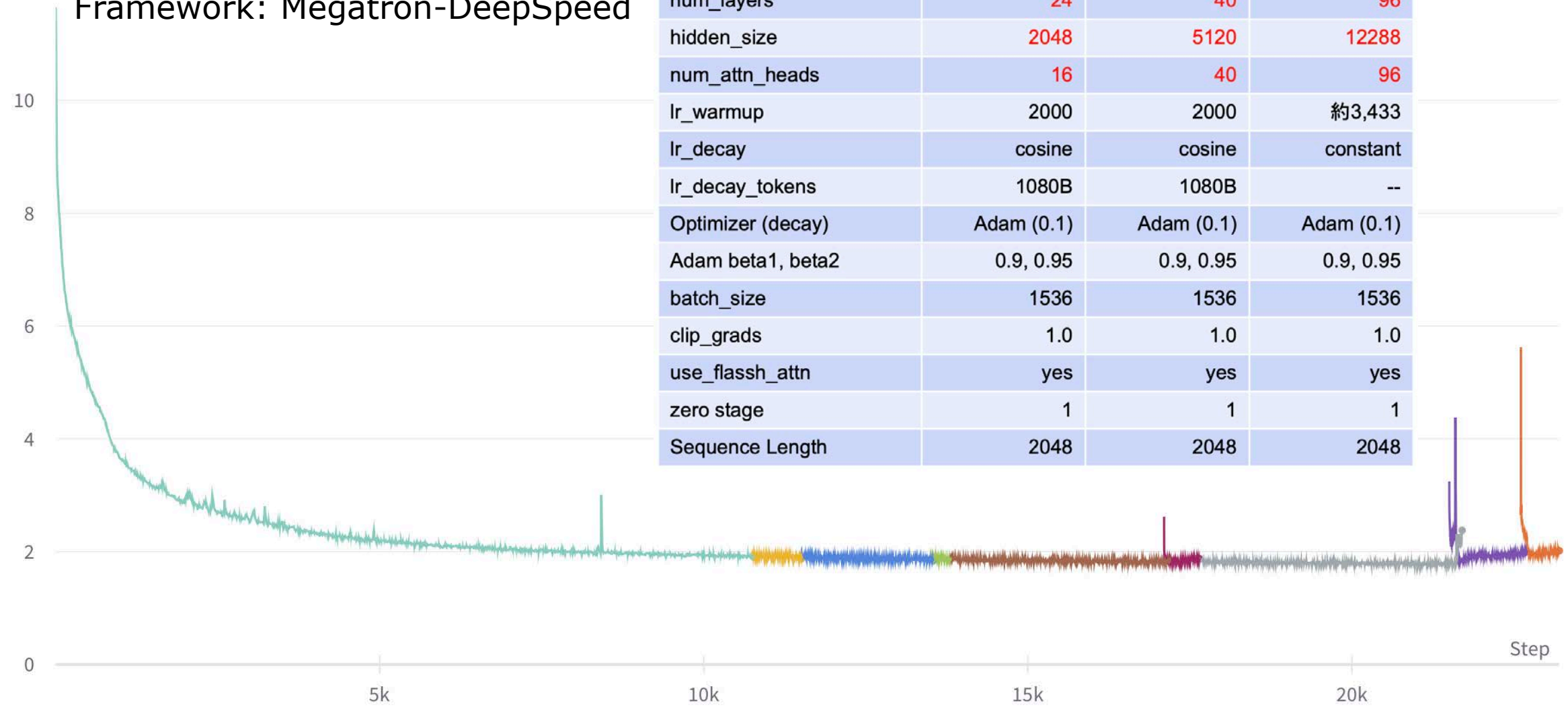


	1.3B	13B
lr (min_lr)	2.0e-4 (1.0e-6)	1.0e-4 (1.0e-6)
init_std	0.013	0.008
num_layers	24	40
hidden_size	2048	5120
num_attn_heads	16	40
lr_warmup	2000	2000
lr_decay	cosine	cosine
lr_decay_tokens	1080B	1080B
Optimizer (decay)	Adam (0.1)	Adam (0.1)
Adam beta1, beta2	0.9, 0.95	0.9, 0.95
batch_size	1536	1536
clip_grads	1.0	1.0
use_flassh_attn	yes	yes
zero stage	1	1
Seq_len	2048	2048

LLM勉強会

Data size: 70B token
Model size: 175B
Framework: Megatron-DeepSpeed

	1.3B	13B	175B
lr (min_lr)	2.0e-4 (1.0e-6)	1.0e-4 (1.0e-6)	0.6e-4 (1.0e-6)
init_std	0.013	0.008	0.005
num_layers	24	40	96
hidden_size	2048	5120	12288
num_attn_heads	16	40	96
lr_warmup	2000	2000	約3,433
lr_decay	cosine	cosine	constant
lr_decay_tokens	1080B	1080B	--
Optimizer (decay)	Adam (0.1)	Adam (0.1)	Adam (0.1)
Adam beta1, beta2	0.9, 0.95	0.9, 0.95	0.9, 0.95
batch_size	1536	1536	1536
clip_grads	1.0	1.0	1.0
use_flassh_attn	yes	yes	yes
zero stage	1	1	1
Sequence Length	2048	2048	2048



産総研＋東工大(岡崎研＋横田研)

LLaMA-2からの継続学習

パラメータ数：7B, 13B, 70B

データ：岡崎研CC, Wikipediaなど

フレームワーク：MDS

語彙拡張：あり、なし

日本語:英語＝50:50,90:10

評価：日英LM Evaluation Harness

Model	日本語	XLSUM_ja	MATH (mgsm_ja)	MC	NLI	QA	RC
Llama2-7b	0.3495	0.0183	0.0760	0.3852	0.443	0.3825	0.7917
拡張なし (2K)	0.3588	0.0194	0.0880	*0.3762*	*0.3782*	0.4569	0.8339
拡張なし (4K)	0.3635	0.0234	0.0800	*0.3128*	*0.4127*	0.5164	0.8355

Model	英語	gsm8k	squad2	triviaqa	hellaswag	openbookqa	xwinograd_en
Llama2-7b	0.4883	0.1372	0.3208	0.6278	0.5855	0.3560	0.9027
拡張なし(2K)	0.4402	0.0773	0.2583	0.5265	0.5495	0.3380	0.8916
拡張なし(4K)	0.4351	0.0871	0.2998	0.4778	0.5388	0.3200	0.8868

Model	日本語	XLSUM_ja	MATH (mgsm_ja)	MC	NLI	QA	RC
Llama2-13B	0.4210	0.0192	0.1320	0.6819	0.4065	0.4314	0.8551
拡張あり(5K)	0.4745	0.1394	0.1200	0.7659	*0.3854*	0.5519	0.8846
拡張あり(10K)	0.4832	0.1404	0.1560	0.7525	*0.4052*	0.5572	0.8878

Model	英語	gsm8k	squad2	triviaqa	hellaswag	openbookqa	xwinograd_en
Llama2-13B	0.5381	0.2342	0.3672	0.7258	0.6146	0.3740	0.9126
拡張あり (5K)	0.4693	0.1478	0.3050	0.5674	0.5633	0.3360	0.8963
拡張あり (10K)	0.4680	0.1592	0.3339	0.5342	0.5550	0.3300	0.8959

Model	日本語	XLSUM_ja	MATH (mgsm_ja)	MC	NLI	QA	RC
Llama2-70b	0.5606	0.0208	0.3680	0.8704	0.6794	0.5160	0.9090
拡張 (4K)	0.5998	0.1519	0.3680	0.9026	*0.6797*	0.5795	0.9169
拡張 (6K)	0.6098	0.1568	0.3680	0.9312	*0.6685*	0.6181	0.9159
拡張 (8K)	0.6061	0.1522	0.3960	0.9160	*0.6222*	0.6321	0.9181
拡張なし (4K)	0.5674	0.0232	0.3800	0.9071	*0.6005*	0.5708	0.9230

産総研＋東工大(岡崎研＋横田研)

日本語と英語の混合率

Setting	Iteration	Average	MC	NLI	QA	RC	MATH	XLSUM
Llama-2-chat		0.3766	0.5255	0.4665	0.3439	0.8045	0.1040	0.0154
50:50	5500	0.4006	0.5880	0.4566	0.3632	0.8176	0.0800	0.0979
90:10	5500	**0.4051**	**0.5907**	0.4478	**0.3754**	**0.8182**	**0.1040**	0.0944

Setting	Iteration	Average	GSM8K	TriviaQA	HellaSwag	OpenBookQA	Winogrande	SQuAD
Llama-2-chat		0.4741	0.1827	0.5566	0.5882	0.3340	0.8817	0.3016
50:50	5500	0.4677	0.1758	0.5774	0.5659	0.3320	0.8924	0.2627
90:10	5500	0.4719	0.1690	0.5701	0.5644	0.3320	0.8984	0.2972

追加語彙数の影響

Setting	Iteration	Average	MC	NLI	QA	RC	MATH	XLSUM
Llama-2-chat		0.3766	0.5255	0.4665	0.3439	0.8045	0.1040	0.0154
16K	6500	0.**4073**	**0.5862**	**0.4601**	0.3805	**0.8161**	**0.1040**	0.0970
32K	6500	0.4007	0.5764	0.4514	0.3894	0.8137	0.0720	0.1013

Setting	Iteration	Average	GSM8K	TriviaQA	HellaSwag	OpenBookQA	Winogrande	SQuAD
Llama-2-chat		0.4741	0.1827	0.5566	0.5882	0.3340	0.8817	0.3016
16K	6500	0.4773	0.1766	0.5770	0.5706	0.3420	0.9006	0.2971
32K	6500	0.4667	0.1500	0.5800	0.5800	0.3400	0.8900	0.2600

「富岳」政策対応枠

課題名：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 (hp230254)

実施機関：2023/5/24 - 2024/3/31

参画組織：
東京工業大学 (全体統括、並列化、学習の実施)
東北大学（学習データ収集・精製）
富士通株式会社（高速化）
理化学研究所（並列化・高速化）
株式会社サイバーエージェント（学習データ収集）
名古屋大学（性能計測）
Kotoba Technologies Inc. （モデル評価）

- 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

富士通提供資料

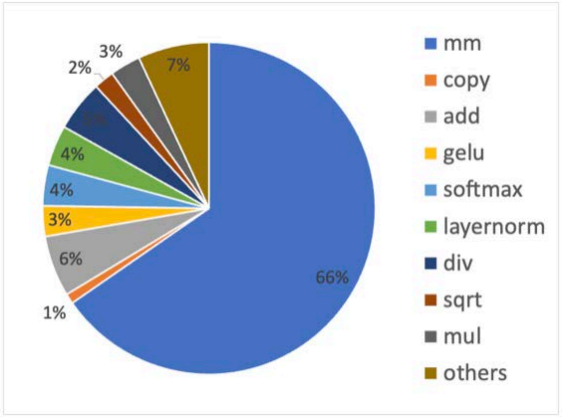
「富岳」単一ノード上での性能

演算の99%は大量の小規模な密行列積が占める
- GEMMの塊なのでHPL並の実効性能が出る実アプリ
- A64FXでは66%、A100では49%の時間が行列積

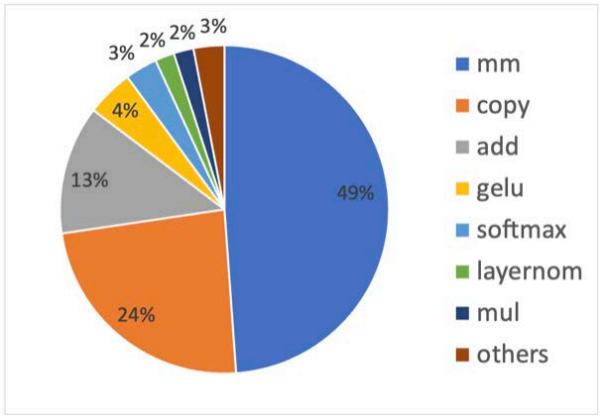
モデルサイズが大きくなるほどノード毎のFLOPsが増加

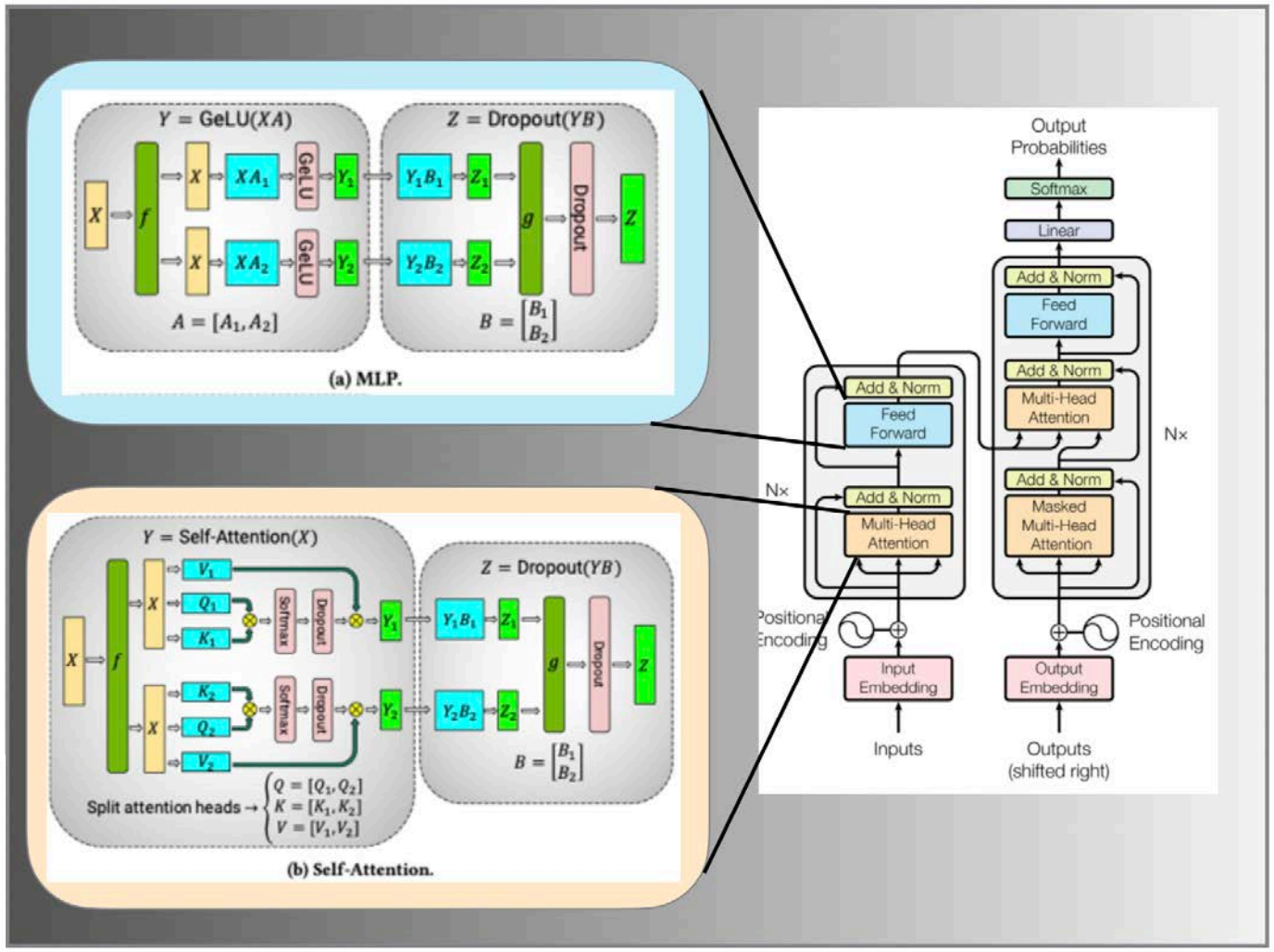
Model size	Hidden size	Number of layers	Number of parameters (billion)	Model-parallel size	Number of GPUs	Batch size	Achieved teraFlOPs per GPU	Percentage of theoretical peak FLOPs	Achieved aggregate petaFLOPs
1.7B	2304	24	1.7	1	32	512	137	44%	4.4
3.6B	3072	30	3.6	2	64	512	138	44%	8.8
7.5B	4096	36	7.5	4	128	512	142	46%	18.2
18B	6144	40	18.4	8	256	1024	135	43%	34.6
39B	8192	48	39.1	16	512	1536	138	44%	70.8
76B	10240	60	76.1	32	1024	1792	140	45%	143.8
145B	12288	80	145.6	64	1536	2304	148	47%	227.1
310B	16384	96	310.1	128	1920	2160	155	50%	297.4
530B	20480	105	529.6	280	2520	2520	163	52%	410.2
1T	25600	128	1008.0	512	3072	3072	163	52%	502.0

A64FX CPU



A100 GPU



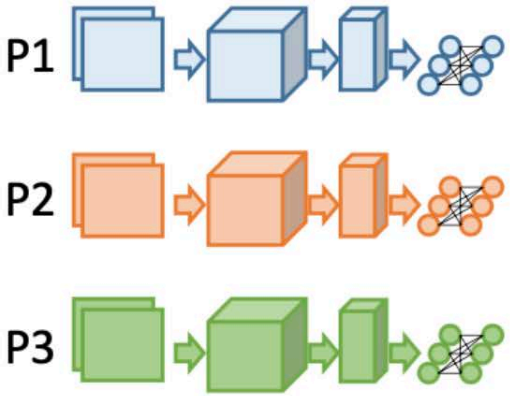


入力はtoken数(sequence長)×hidden size×batch size

1 batchの推論においても数千次元のGEMMが演算の大半

「富岳」上での分散並列学習

データ並列 (DP)



データは分散

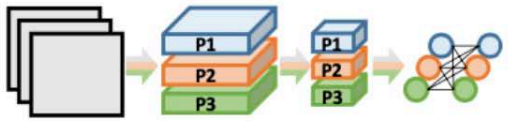
モデルは冗長

勾配のAllReduce

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 (TP)
or
ZeRO (FSDP)



データは冗長

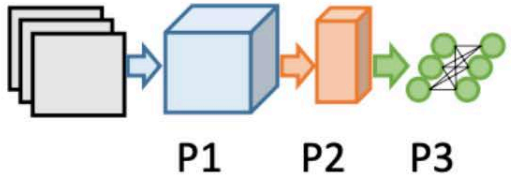
モデルは分散

活性のAllReduce
or
パラメータのAllGather

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(PP)



データは冗長

モデルは分散

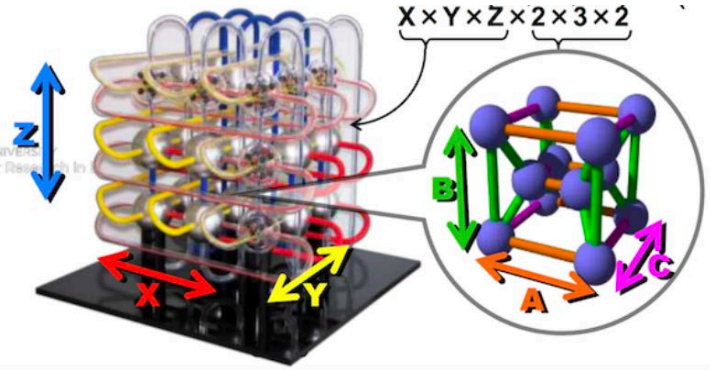
活性のSendRecv

課題：パイプラインバブル

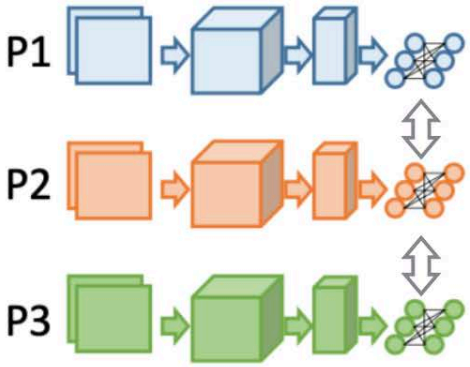
解決策：パイプライン
の工夫

ハイブリッド並列化
並列数=DP x TP x PP x FSDP

Tofuの６次元トーラスを
DP,TP,PP,FSDPにどのように
マッピングするか？



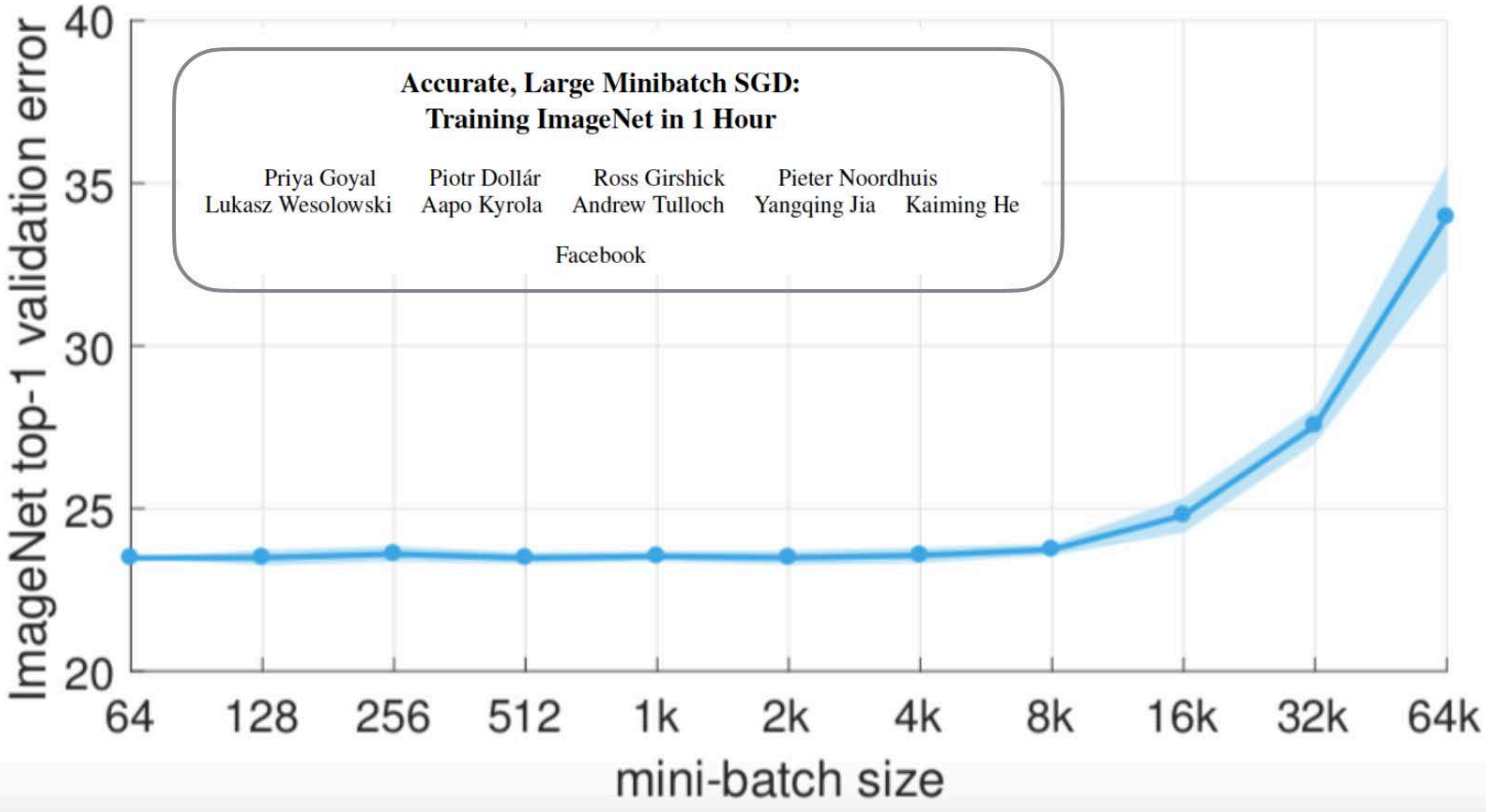
データ並列の問題



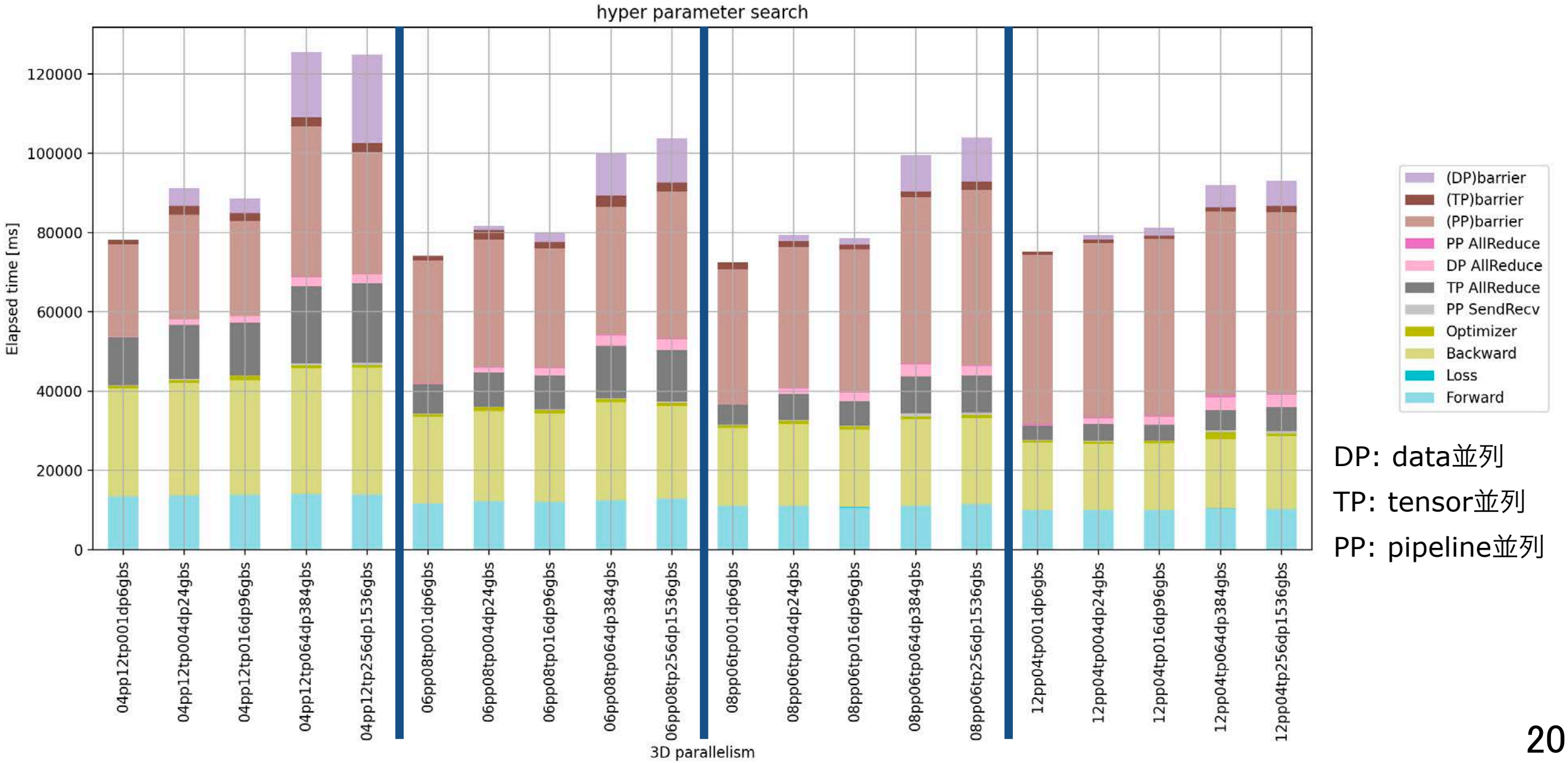
バッチサイズ ＝ GPUあたりのバッチサイズ

×

GPU数



「富岳」上での30Bモデルのデータ弱スケーリング



DP: data並列

TP: tensor並列

PP: pipeline並列

今後のLLMに関して知っておくべきこと

- パラメータ数至上主義の時代は終わっている
- データの質と量こそが重要

- Microsoftが所有するGPU数は40万規模
- 世界最大級のスパコンAuroraは6万程度

- Transformerの演算の99%は行列積
- 1batchの推論でも行列積
- Matrix engineがなければ話にならない