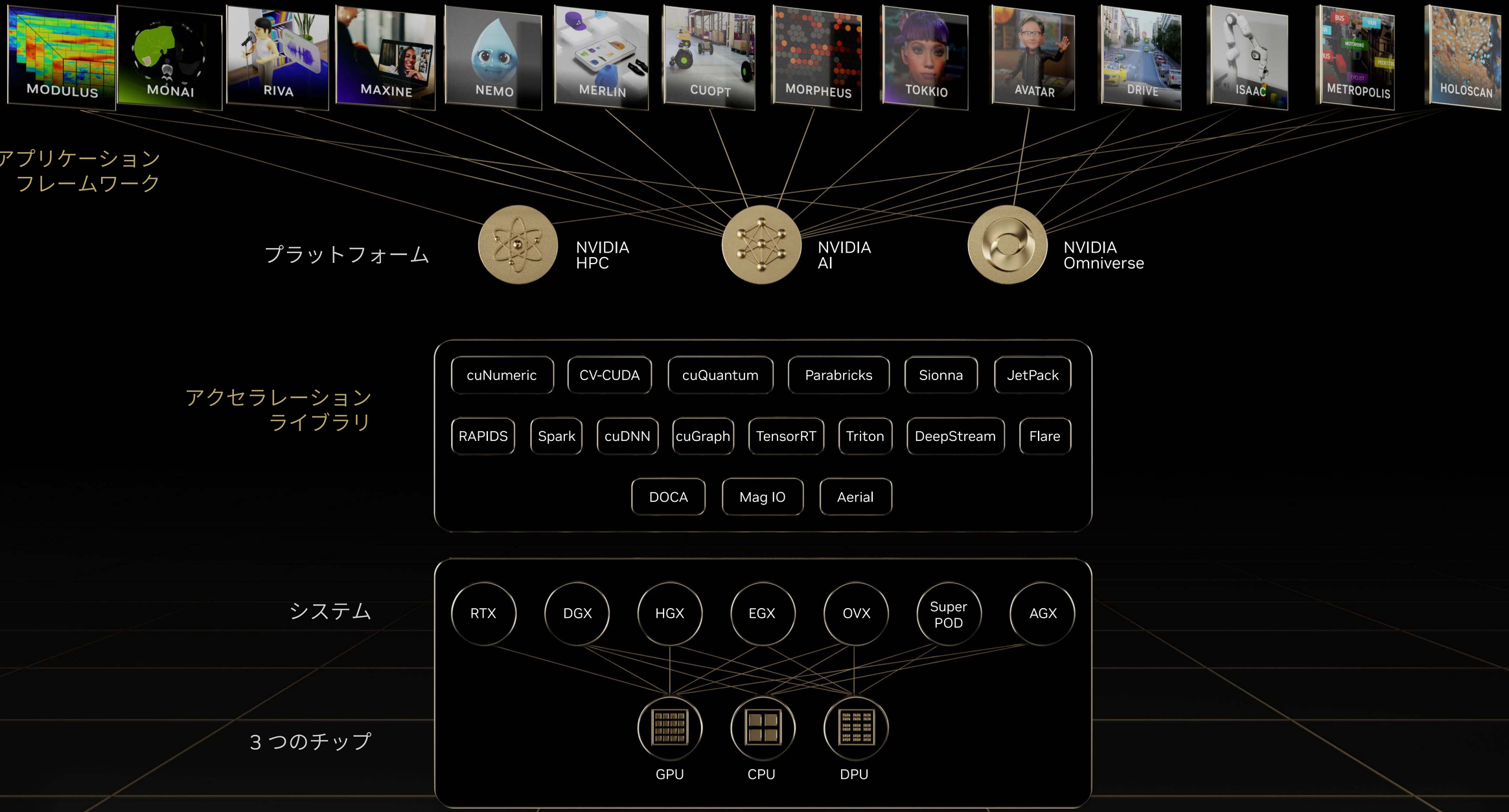
生成 AI 技術やビジネスの最新動向

エンタープライズ事業本部　事業本部長

井﨑 武士

AI コンピューティング カンパニー

- 創業　1993年
- 創業者 / CEO　ジェンスン ファン
- 従業員数　22,500人
- 2023年度売上 270 億ドル
- 時価総額　1兆ドル


AI アプリケーション
フレームワーク
MODULUS
MONAI
RIVA
MAXINE
NEMO
MERLIN
CUOPT
MORPHEUS
TOKKIO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
プラットフォーム
NVIDIA
HPC
NVIDIA
AI
NVIDIA
Omniverse
アクセラレーション
ライブラリ
cuNumeric
CV-CUDA
cuQuantum
Parabricks
Sionna
JetPack
RAPIDS
Spark
cuDNN
cuGraph
TensorRT
Triton
DeepStream
Flare
DOCA
Mag IO
Aerial
システム
RTX
DGX
HGX
EGX
OVX
Super
POD
AGX
3 つのチップ
GPU
CPU
DPU

生成AI の衝撃

iPhone moment

Stable Diffusion



Water front city landscape, sunshining, photo realistic, summer, 4k, rule of thirds, extreme detail, detailed drawing, trending artstation, hd, realistic lighting, by Alphonse Mucha, Greg Rutkowski, sharp focus, backlit

ChatGPT

DI

プロの経済アナリストの立場で答えてください。AIの活用による経済効果を分野ごとにリストアップしてください

AIの活用による経済効果は、さまざまな分野で期待されています。以下に、いくつかの分野とそれに伴う経済効果をリストアップします。

1. 製造業:
   - 自動化とロボット化による生産性向上
   - プロセスの最適化による品質向上
   - 機械学習による予測メンテナンスによる機械の故
   - データ分析による在庫最適化と生産計画の最適
2. 金融業:
   - 高速なデータ処理とパターン認識によるリスク管理
   - モデルベースの予測による株式市場や為替市場の
   - ロボアドバイザーによる個人投資家の助言やポート
   - 不正検知によるセキュリティの向上
3. 医療業界:
   - パターン認識と画像解析による病気の早期検出
   - データマイニングによる臨床試験の最適化と新薬
   - ヘルスケアデータの分析による予防医療の向上と
   - ロボット手術による手術の精度向上と手術時間の
4. 小売業:

DI

クルマの車種を分類するVITを使ったPyTorchのスクリプト書いて

以下は、PyTorchを使用してVITを使ったクルマの車種分類スクリプトの例です。このスクリプトは、CIFAR-10データセットで訓練されたもので、VITのパラメータは事前に学習されたものを使用しています。

python Copy code

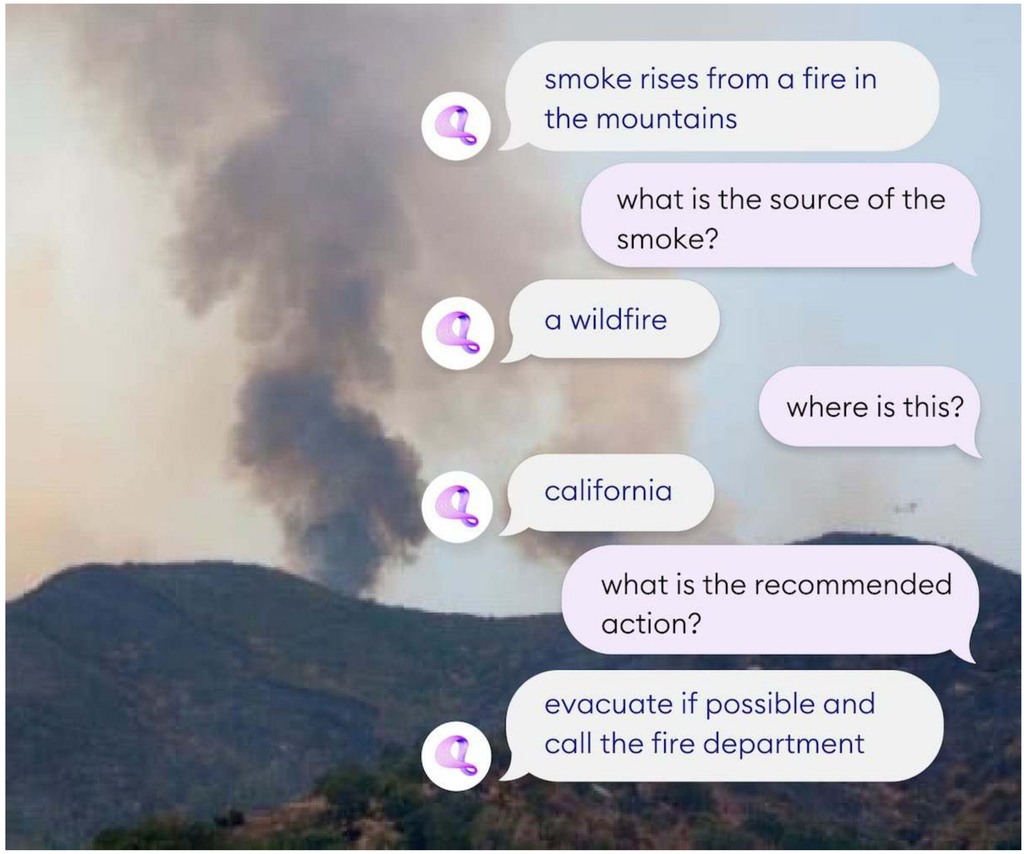
```
import torch
import torch.nn as nn
import torch.optim as optim
import torchvision.datasets as datasets
import torchvision.transforms as transforms
from torch.utils.data import DataLoader
from einops.layers.torch import Rearrange

# Define the VIT model architecture
class ViT(nn.Module):
    def __init__(self, image_size=32, patch_size=4, num_classes=10, dim=768, depth=12, hea
        super(ViT, self).__init__()

        # Calculate the number of patches
        self.num_patches = (image_size // patch_size) ** 2
        self.patch_dim = 3 * patch_size ** 2
```

生成AI 事例

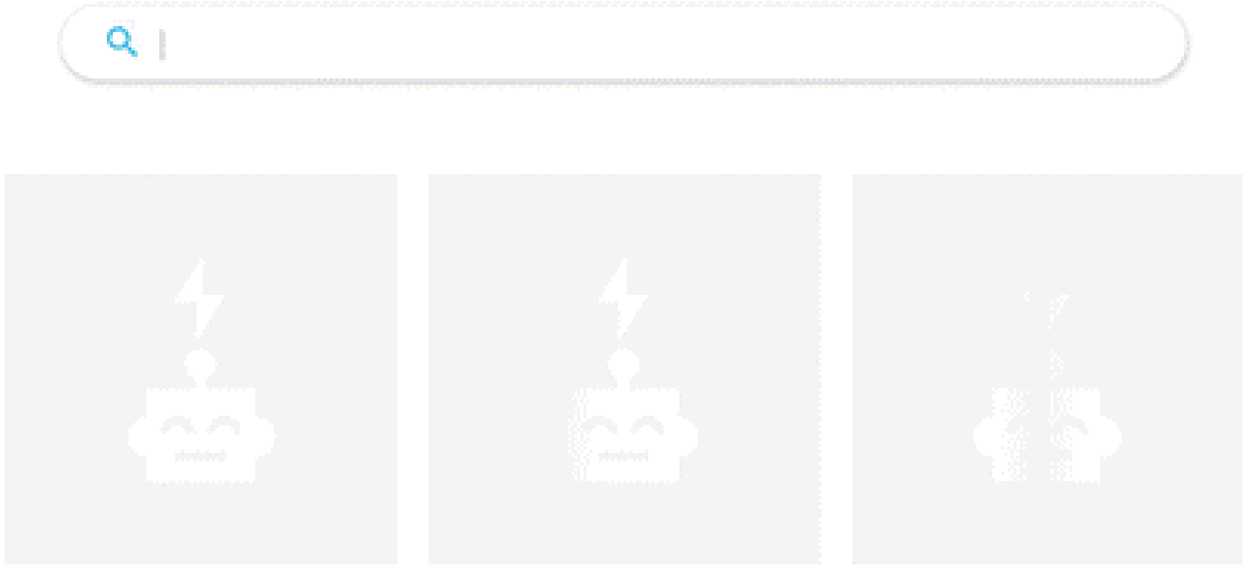
Insight Face
独自の顔画像認識技術とMidjurneyを組み合わせて
顔のスワッピングを行う

Briefly
電話会議の文字起こしを元に洗練された文書や会議後の
成果物を生成する

生成AI 事例

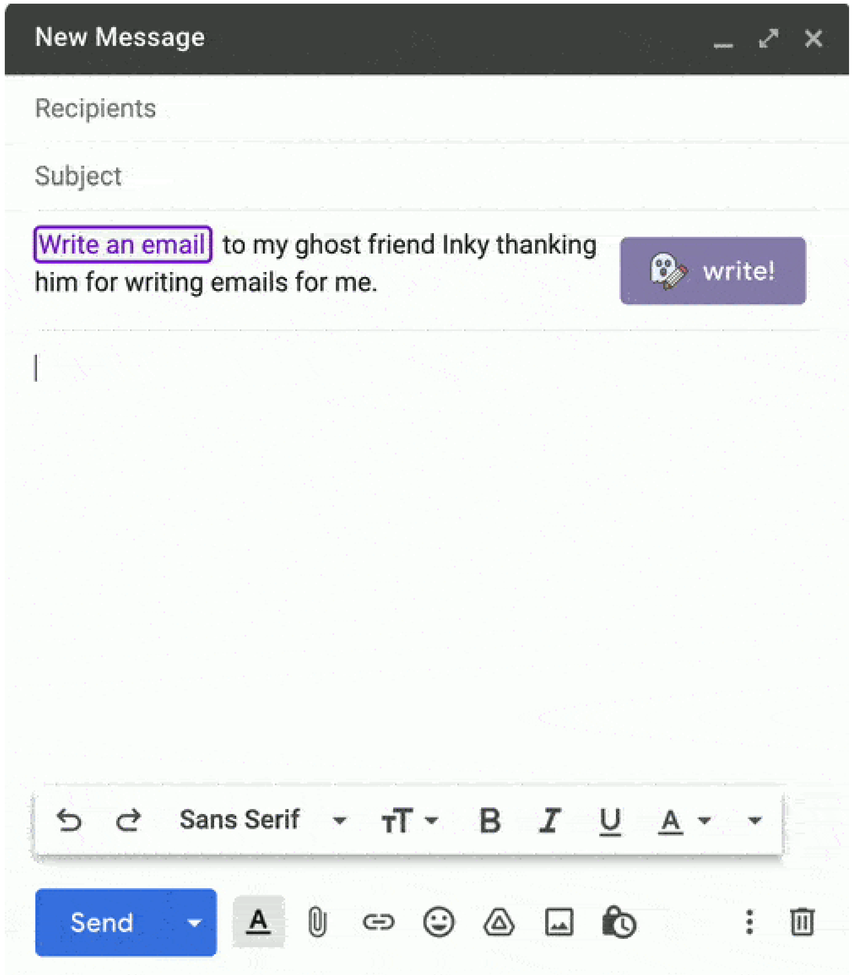


chooch
画像データに関するQ&AをChat形式で行う



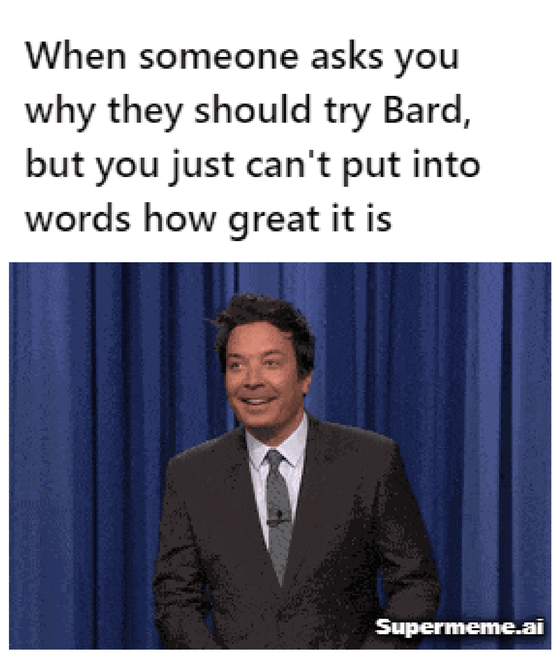
BeautifulAI
入力されたプロンプトに応じて、プレゼンテーションスライドを作成する

生成AI 事例



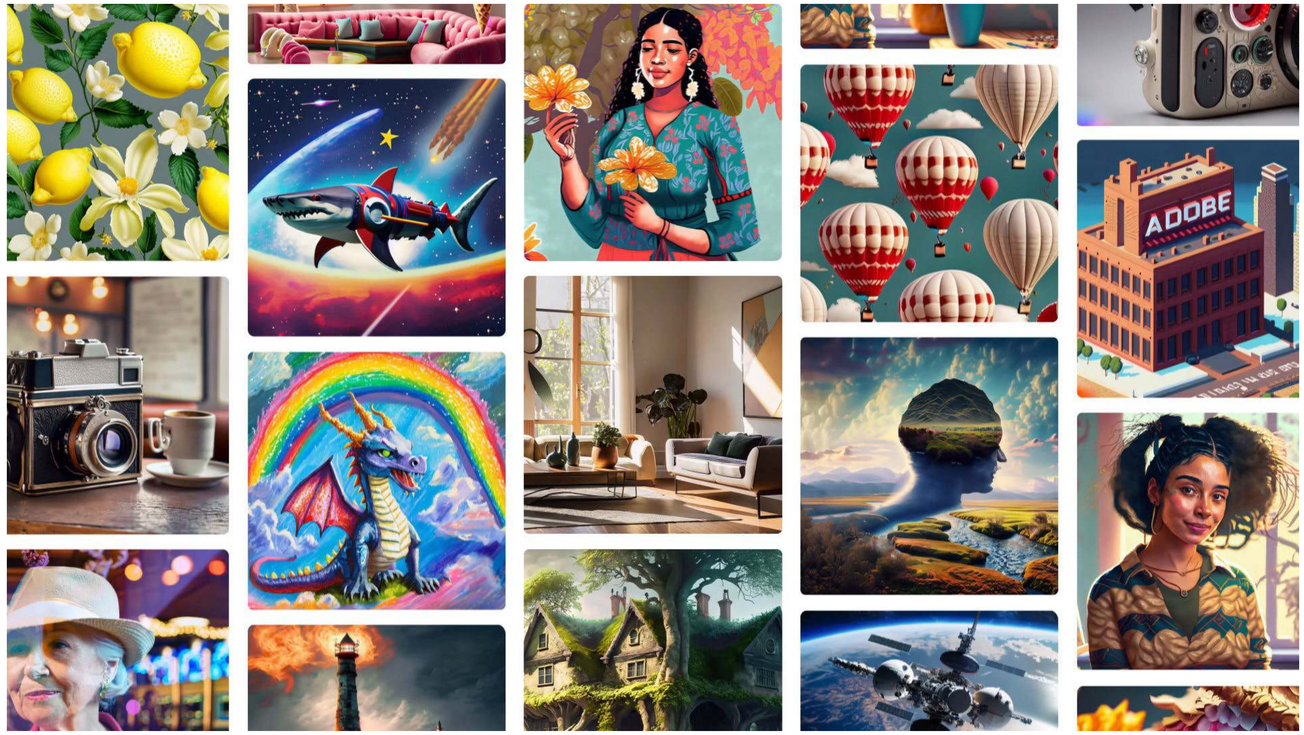
Ghostwrite
ChatGPTを活用した、e-mail 作成アシスタント

入力文章
Have you tried ChatGPT or Bard? Lots of fun! Try it!



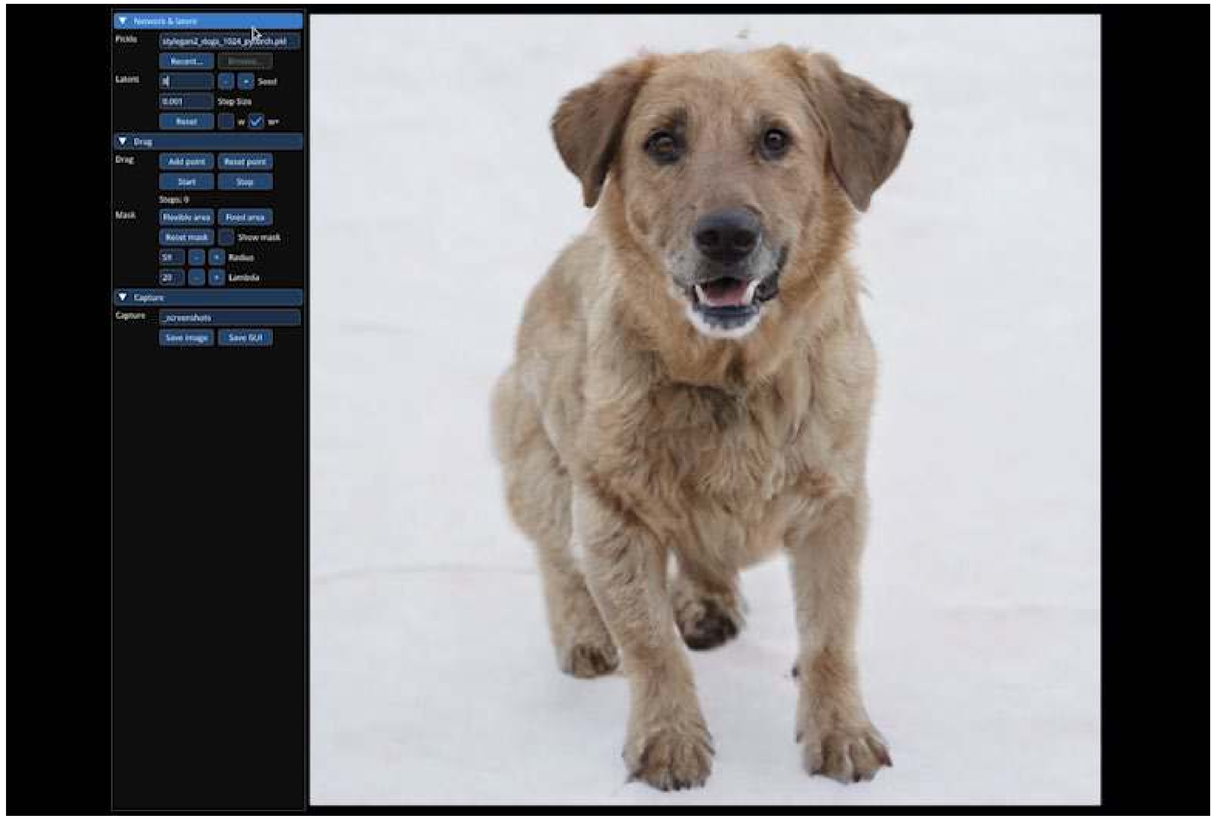
Super Meme
入力文章に応じたミーム（静止画やGIF）を生成
110以上の言語に対応

生成AI 事例



Adobe.comより

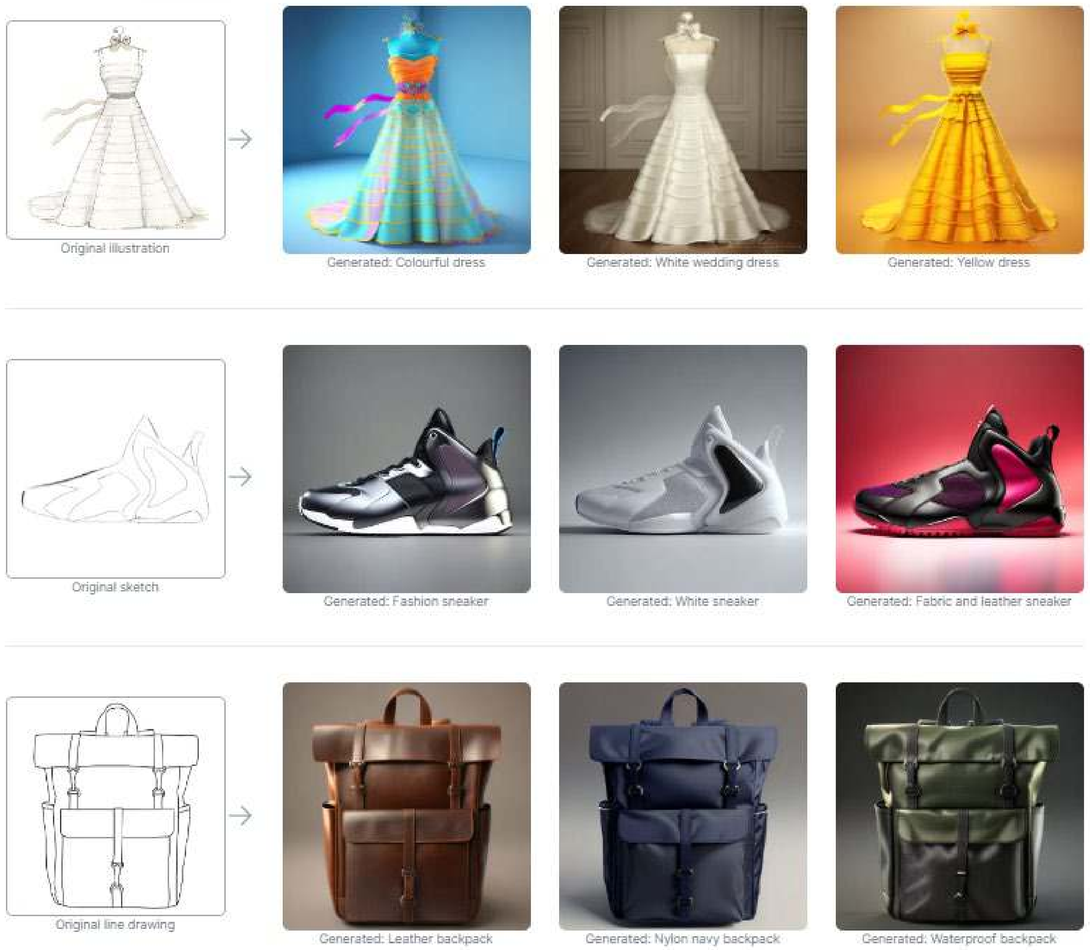
Adobe Firefly
ベータリリース



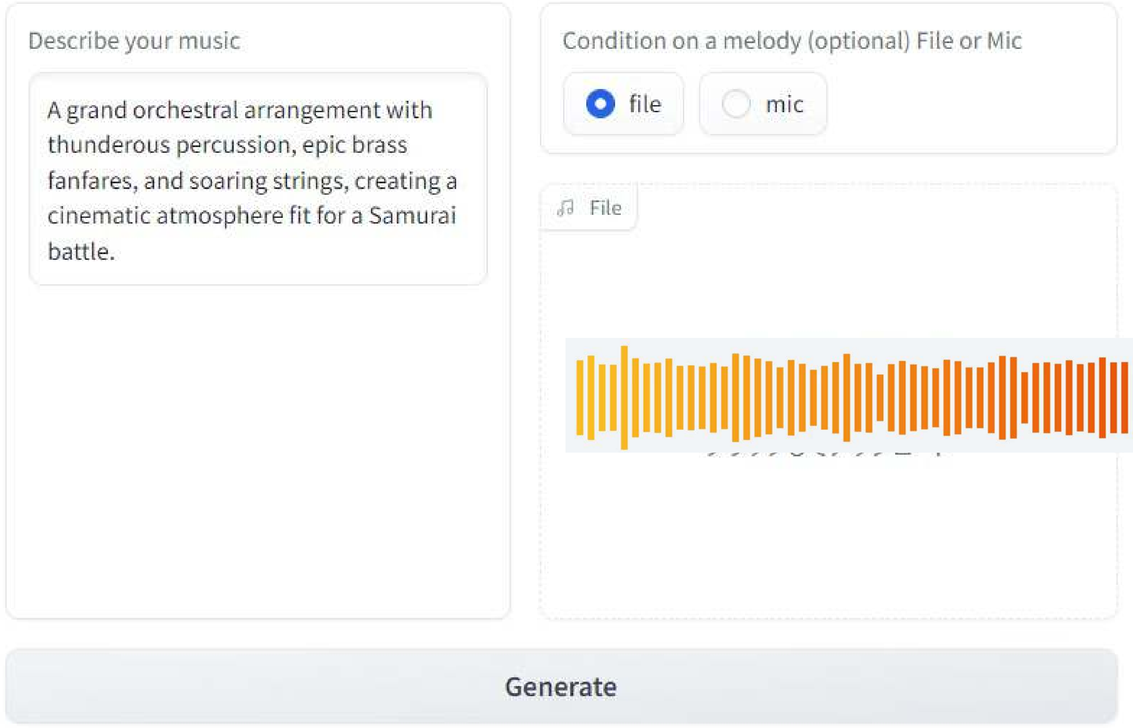
DragGAN

“DragYourGAN:InteractivePoint-basedManipulationontheGenerative ImageManifold”, XINGANGPAN, et.al, SIGGRAPH’23

生成AI 事例



newarc.ai

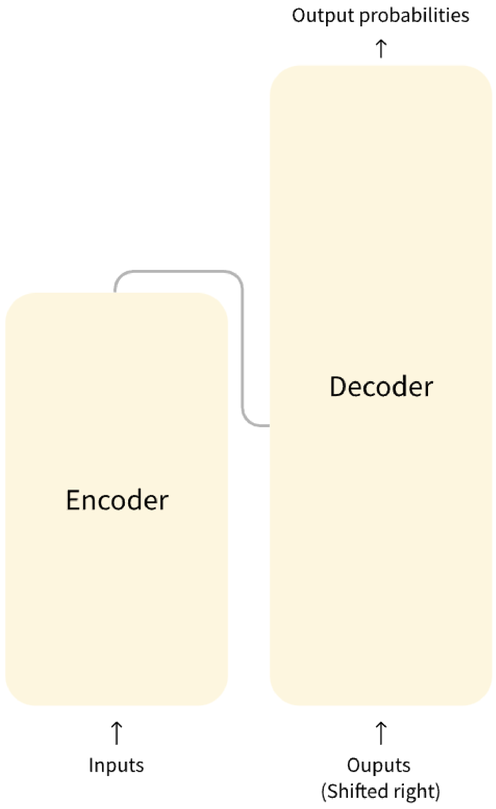


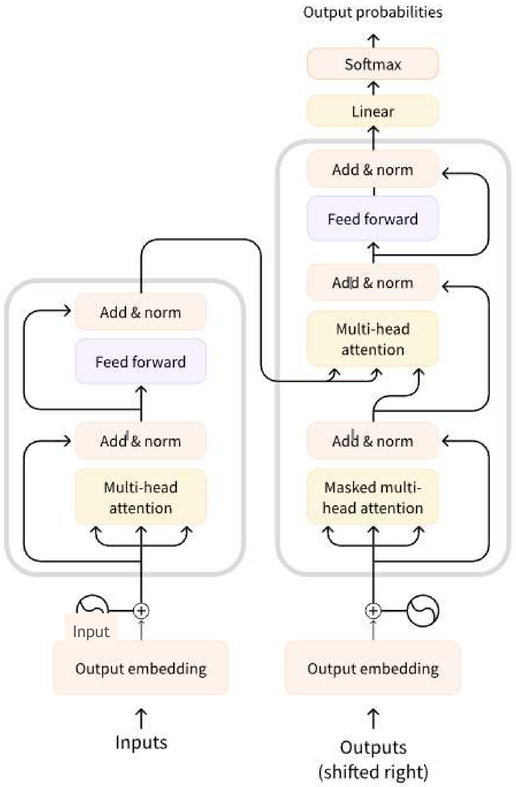
A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a Samurai battle.

Music gen

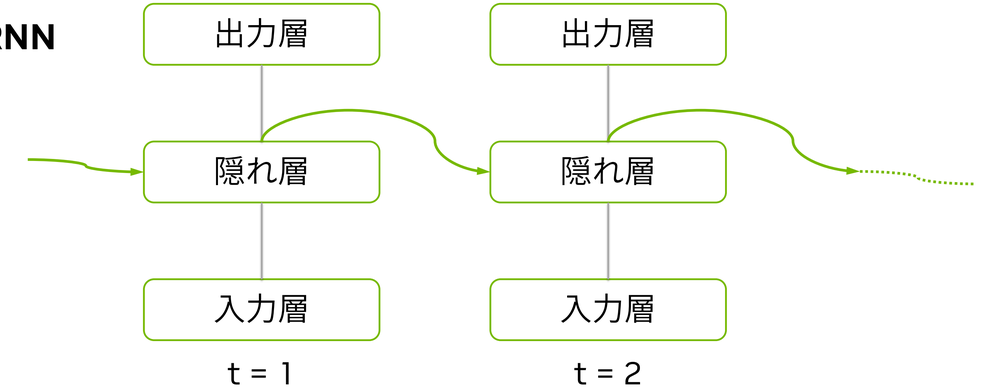
トランスフォーマー

アーキテクチャ

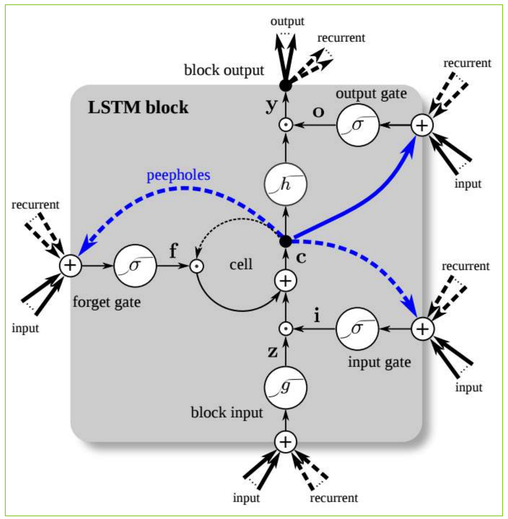




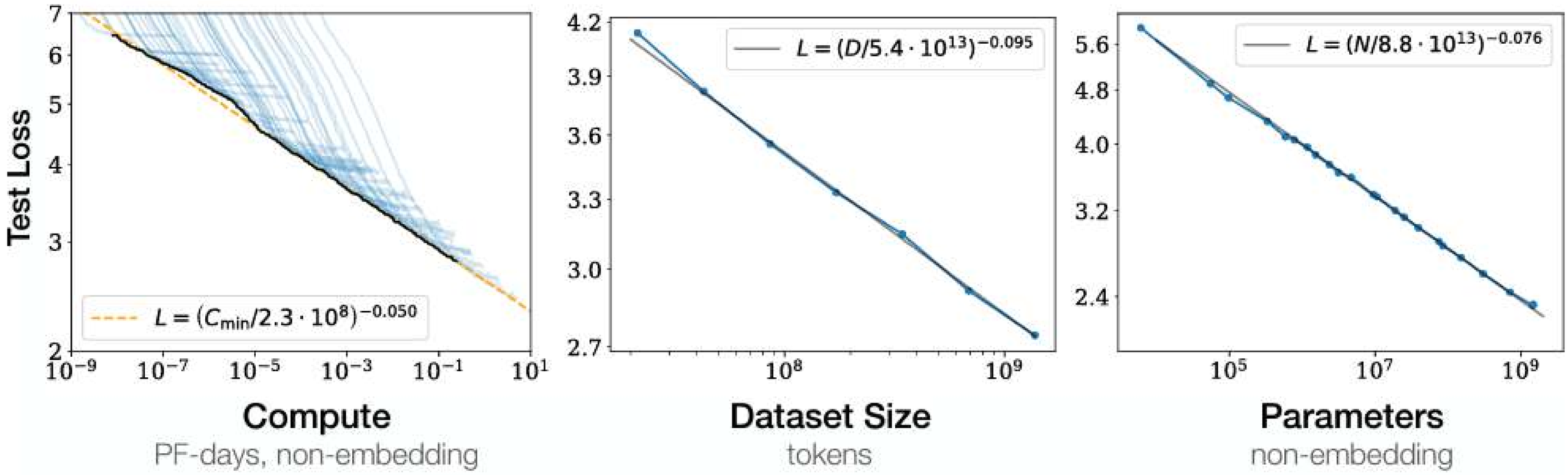
RNN



LSTM



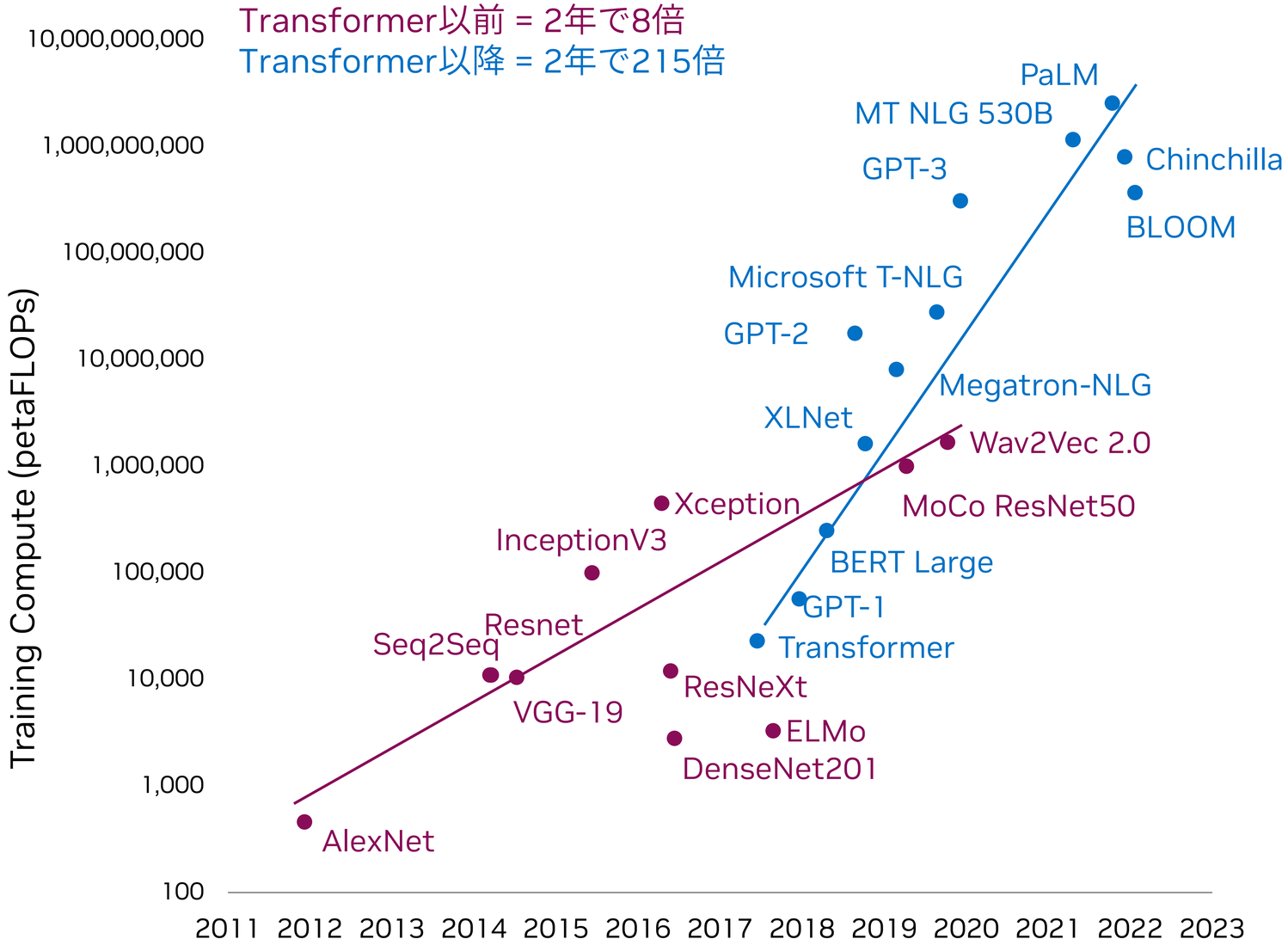
自然言語モデルのスケール則



Scaling laws for neural language model, Jared Kaplan, et al., Jan 2020

指数関数的に増大するLLMモデルサイズ

LLMが必要な計算能力は指数関数的に伸びている



代表的なLLMと学習

モデル	発表時期	モデルサイズ(B)	事前学習 データサイズ	ハードウェア	学習時間
OPT (Meta)	May-2022	175	180B Token	992 80G A100	-
GLM (清華大学)	Oct-2022	130	400B Token	768 40G A100	60日
BLOOM (BigScience)	Nov-2022	176	366B Token	384 80G A100	105日
LLaMA (Meta)	Feb-2023	65	1.4T Token	2048 80G A100	21日
MT-NLG (MS/NVIDIA)	Jan-2022	530	270B Token	4480 80G A100	-

NVIDIA H100

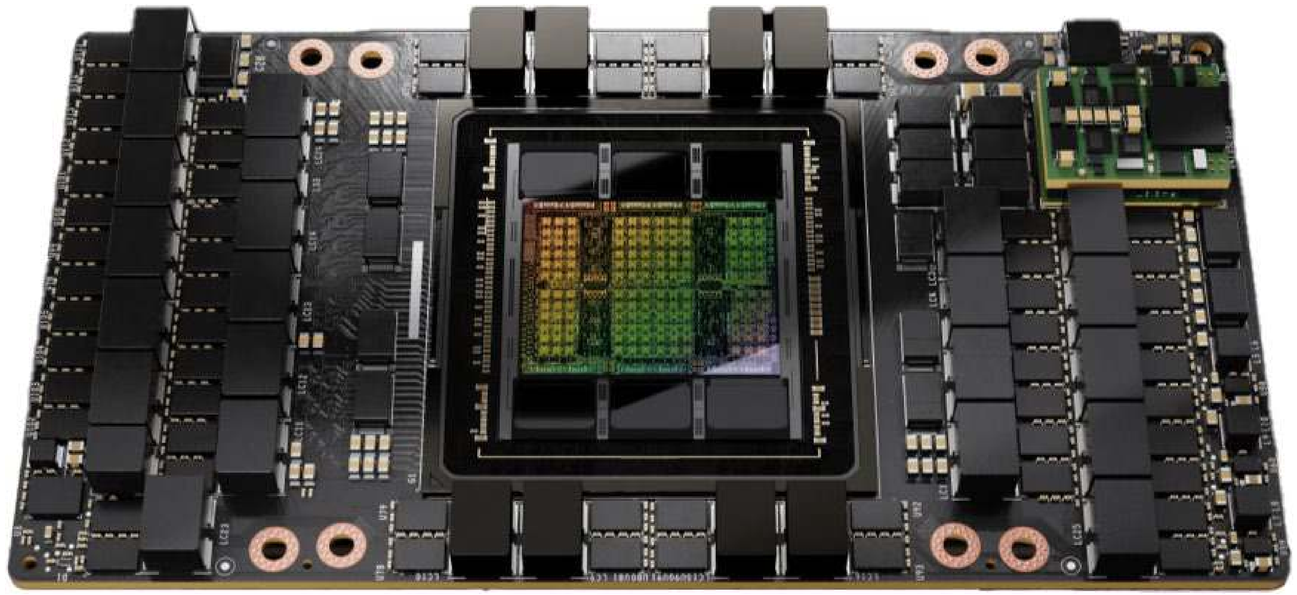
世界の AI インフラを支える新たなエンジン

最高の AI/HPC 性能
4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3X)
3TB/s (1.5X), 80GB HBM3 memory

TRANSFORMER モデルへの最適化
6X faster on largest transformer models

高い稼働率とセキュリティ
7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

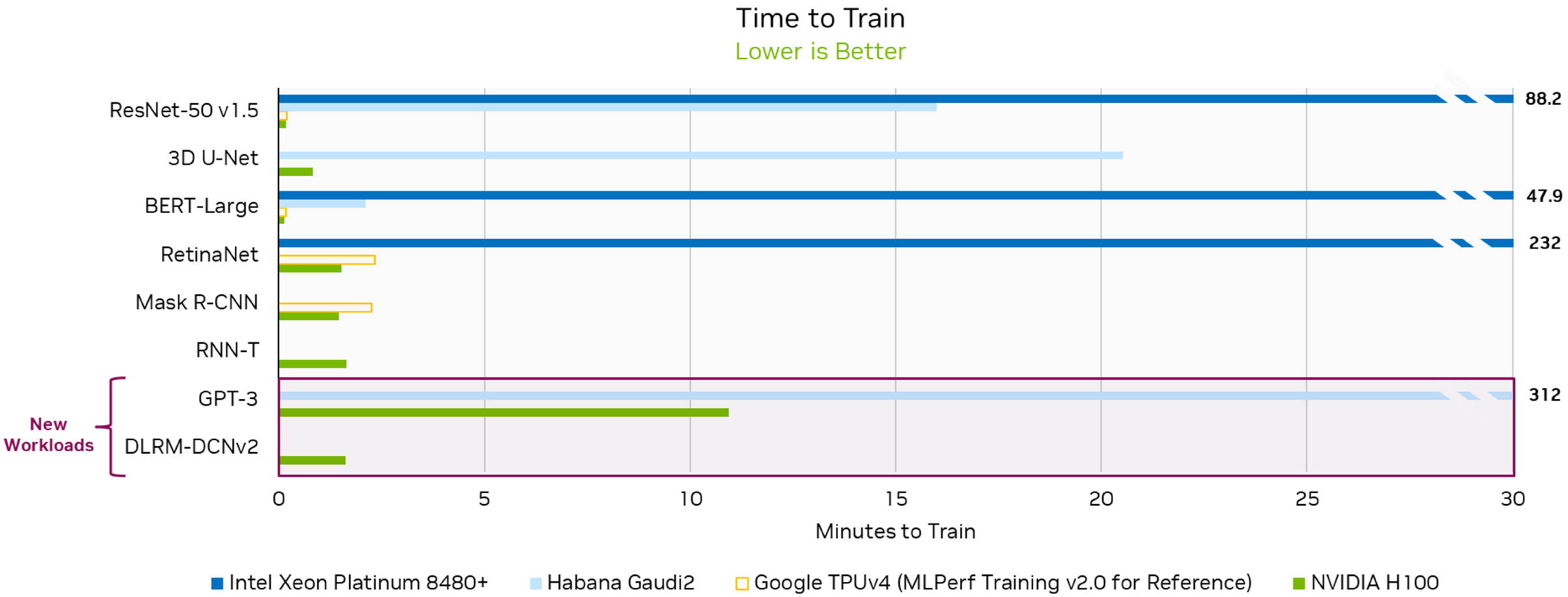
史上最速でスケーラブルなインターコネクト
900 GB/s GPU-2-GPU connectivity (1.5X) | 128GB/s PCI Gen5



Custom TSMC 4N Process | 4.9 TB/s Total External B/W

MLPerf Training v3.0

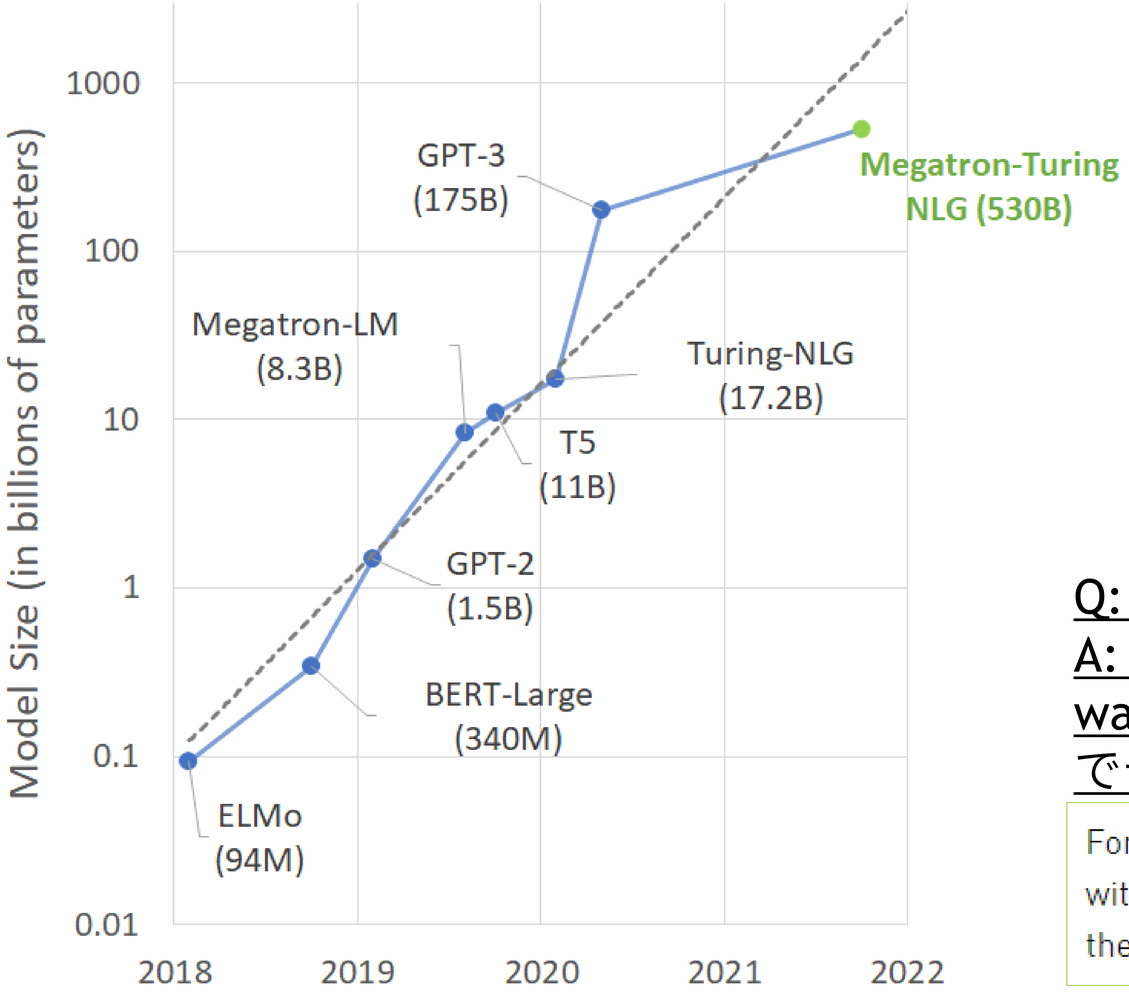
全てのワークロードにおいて最速を記録



MLPerf Training v3.0. Fastest time to train on each benchmark by each submitter's platform | Format: Chip count, Submitter, MLPerf-ID | ResNet-50 v1.5: 3584x NVIDIA+CoreWeave 3.0-2002, 8x Intel-HabanaLabs 3.0-2017, 32x Intel 3.0-2011 | 3D U-Net: 768x NVIDIA 3.0-2075, 8x Intel-Habana Labs 3.0-2016 | BERT-Large: 3,072x NVIDIA+CoreWeave 3.0-2001, 64x Intel-HabanaLabs 3.0-2015, 32x Intel 3.0-2011, 4,096x Google 2.0-2012 | RetinaNet: 768x NVIDIA 3.0-2077, 32x Intel 3.0-2011, 2,048x Google 2.0-2010 | Mask R-CNN: 384x NVIDIA 3.0-2066, 2,048x Google 2.0-2010 | RNN-T: 512x NVIDIA 3.0-2070 | GPT-3: 3,584x NVIDIA+CoreWeave 3.0-2003, 384x Intel-HabanaLabs 3.0-2014 | DLRM-dcnv2: 128x NVIDIA 3.0-2065.

モデルがGPUに載らない？

分散学習の必要性



MT-NLGは530Bパラメータ

- 単純計算で 2,120 GB (in FP32)
- モデルをすべてメモリにロードするだけで、

  8xA100 (80GB) サーバが、3 台強必要
- ワーキングメモリも当然必要

Q: 実際どう扱っているのか？
A: 280xA100 (=35 nodes) で 8-way tensor parallel & 35-way pipeline parallel のモデル並列＆4480GPUのクラスタでデータ並列

> For example, for the 530 billion model, each model replica spans 280 NVIDIA A100 GPUs, with 8-way tensor-slicing within a node and 35-way pipeline parallelism across nodes. We then use data parallelism from DeepSpeed to scale out further to thousands of GPUs.

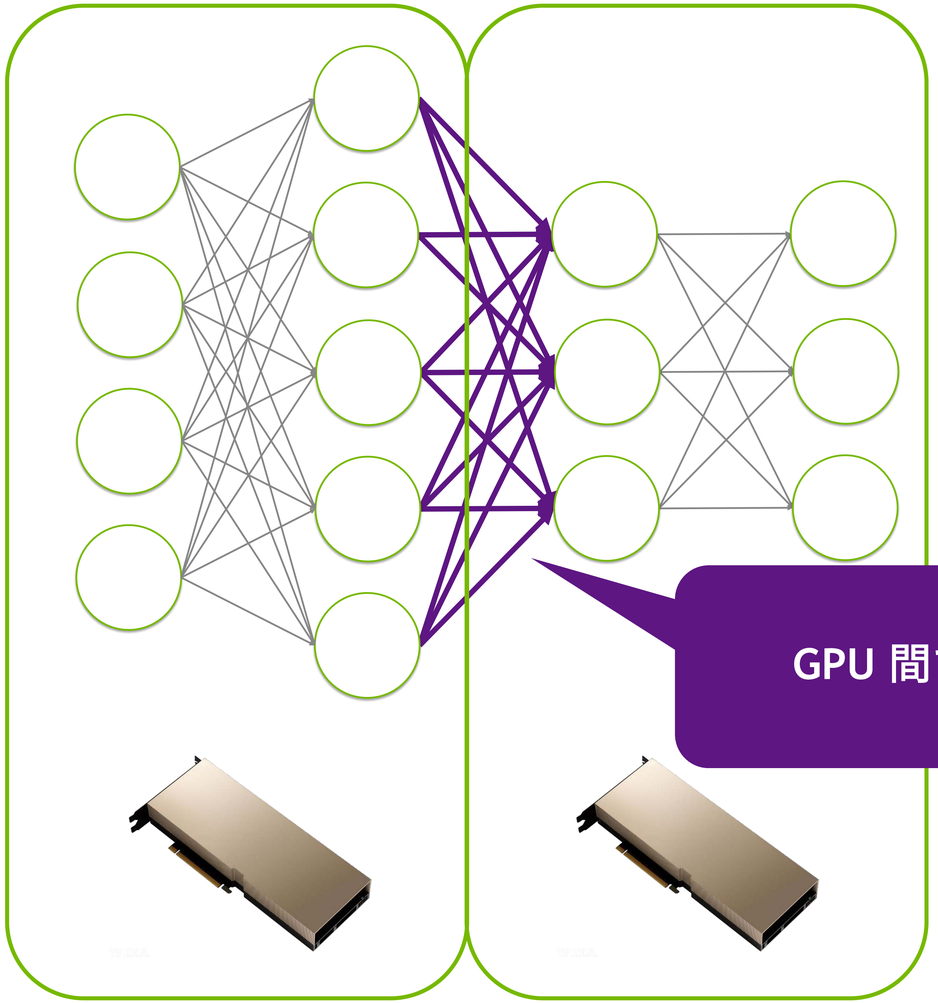
https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

NVIDIA

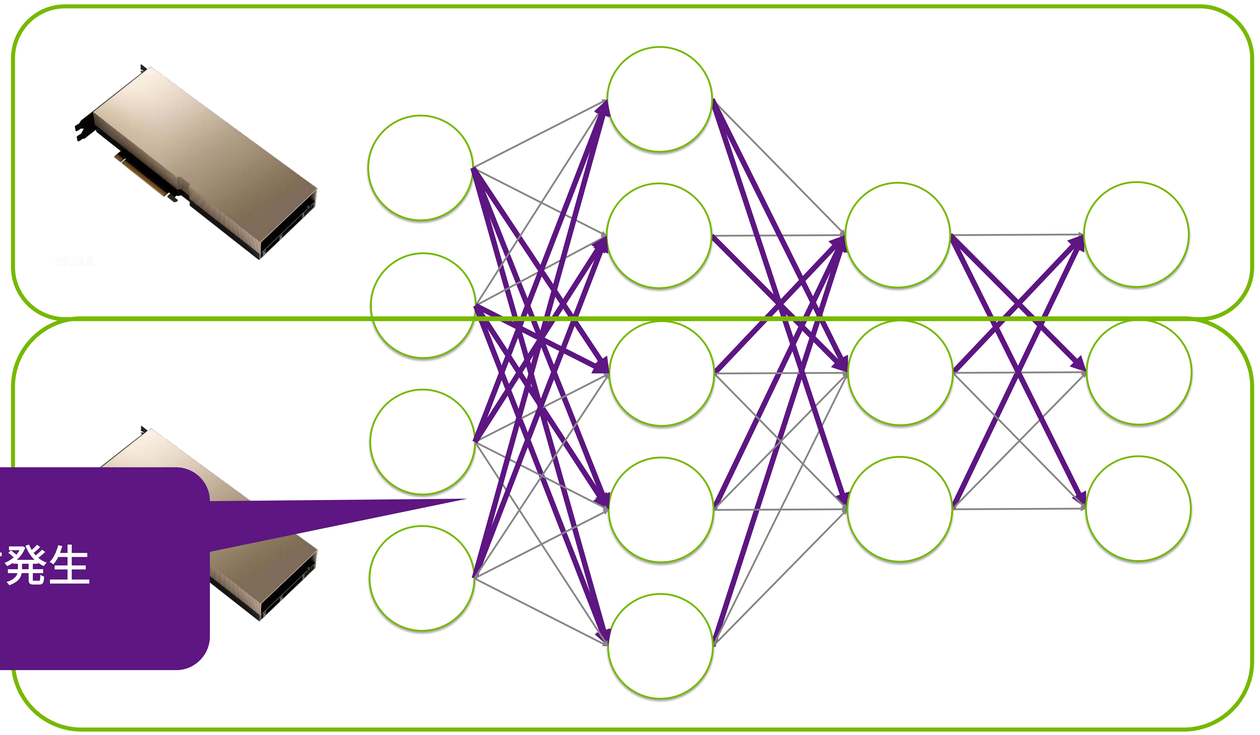
モデル並列の考え方

大規模なモデルを扱えるサイズに分割する

レイヤーごとに別デバイスへ割り当て
(pipeline parallel とも)



各デバイスでレイヤーの一部を担当
(tensor parallel とも)

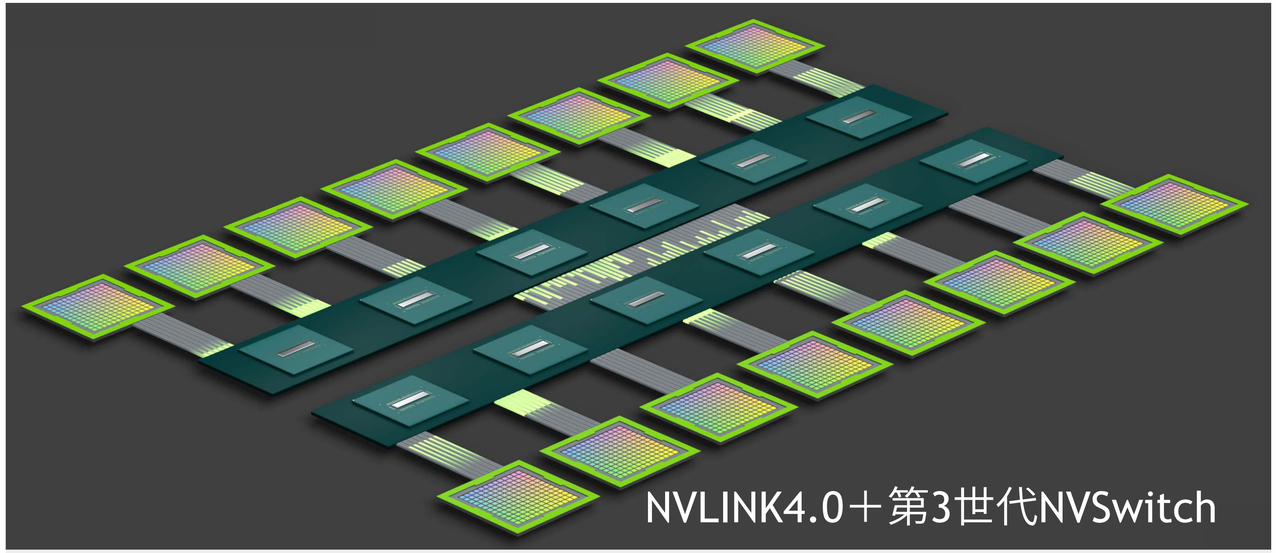


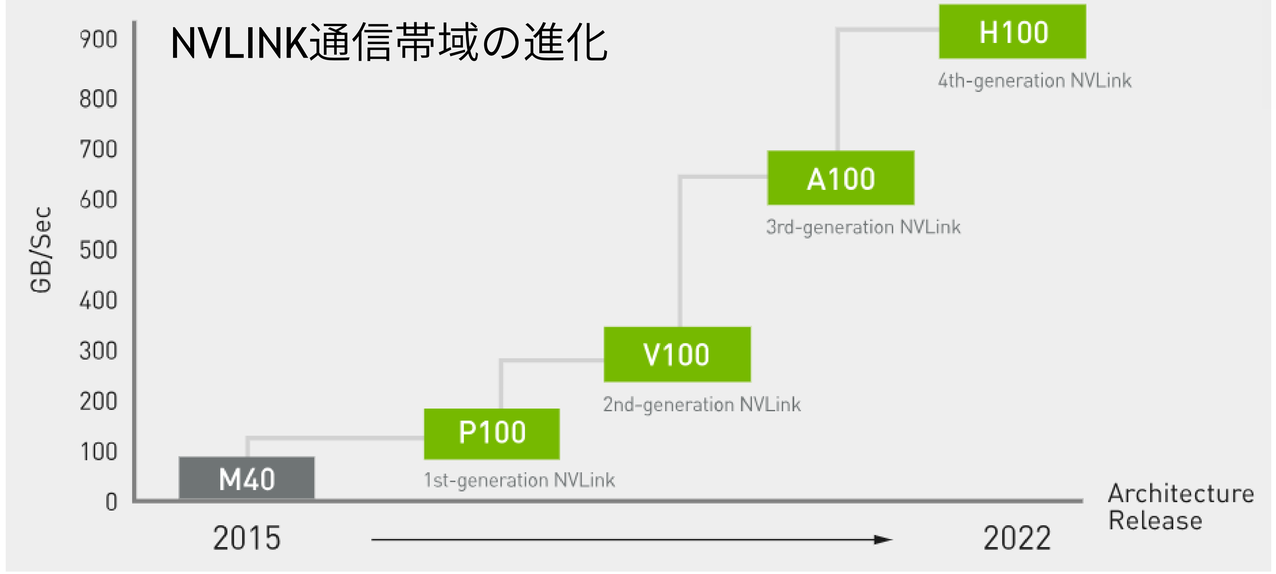
GPU 間での通信発生

その他、畳み込みのチャネルを
別デバイスに割り当てるなどなど……

GPU-GPU インターコネクト/ノード間通信

ノード内





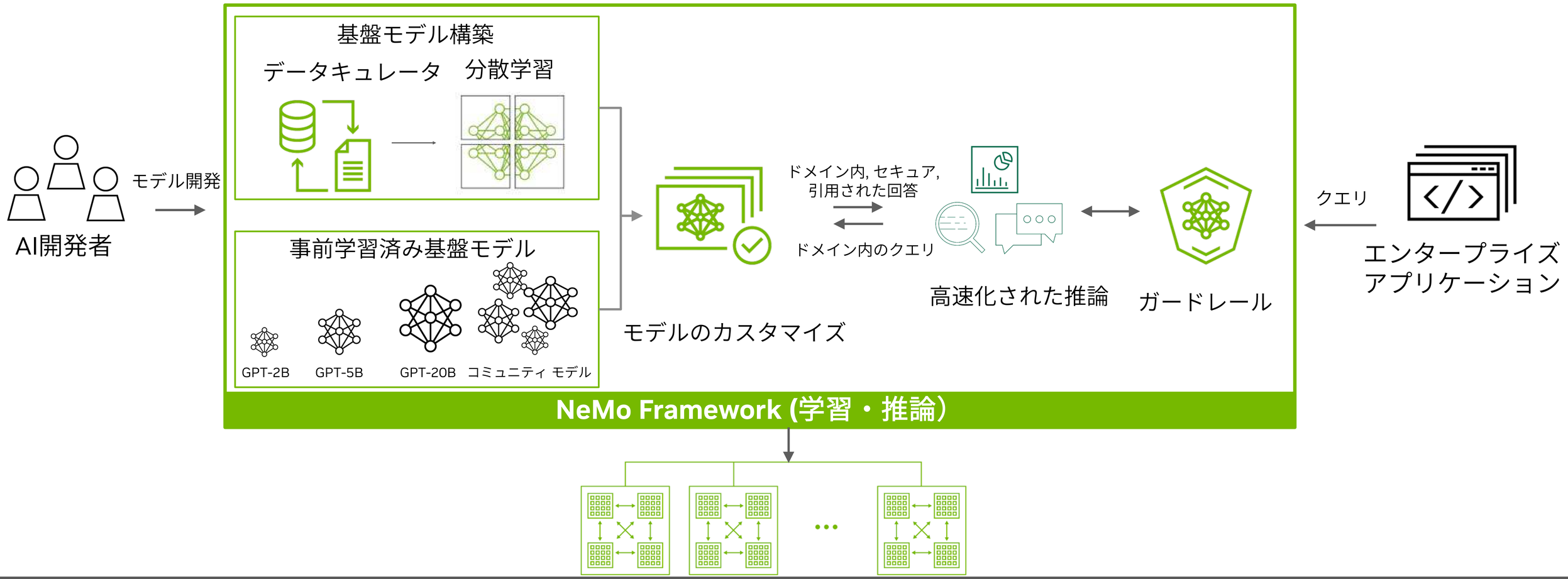
ノード間

ConnectX-7 Infiniband アダプター

NDR 400Gb/s InfiniBand
PCIe Gen5 最大x32 レーン
GPUDirect® RDMA
GPUDirect Storage
In-Network Computing

NeMo Framework

会話、言語、画像の生成AIモデルを構築、カスタマイズ、デプロイするためのエンド・ツー・エンドのワークフロー



マルチモダリティ対応

言語、画像の生成AIモデル構築

データキュレーション @ Scale

抽出、重複、大規模非構造化データのフィルタ情報, @ scale

最適化された学習

数千のノードで並列化されたモデルと学習データによる高速化された学習とスループット

モデルのカスタマイズ

P-tuning, SFT, Adapters, RLHF, AliBiによるカスタマイズ

どこにでもどの規模にも展開可能

あらゆる場所に拡張可能な最適化された推論実行

ガードレール

安全、セキュリティ要求を満たすアプリケーションをNeMoガードレールで実現

サポート

NVIDIA AI Enterprise と顧客専門家によりプロジェクトをオントラックに

NeMo Framework 性能 – 学習

	3000億トークンを学習するための日数 (A100) – BF16			
	800 GPU (100 DGX A100)	**480 GPU (60 DGX A100)**	**160 GPU (20 DGX A100)**	**64 GPU (8 DGX A100)**
GPT-3: 126M	0.07	0.12	0.37	0.92
GPT-3: 5B	0.8	1.3	3.9	9.8
GPT-3: 20B	3.6	6	18.1	45.3
GPT-3: 40B	6.6	10.9	32.8	82
GPT-3: 175B	28	46.7	140	349.9