PCクラスターワークショップ 2024

インテル®Gaudi®アクセラレーターにおけるOSSの活用事例

インテル株式会社　戸谷 大介

注意事項および免責条項

性能は、使用状況、構成、その他の要因によって異なります。詳細については、https://www.Intel.com/PerformanceIndex/ (英語) を参照してください。

性能の測定結果は、構成に示されている日付時点のテストに基づいています。また、現在公開中のすべてのアップデートが適用されているとは限りません。構成の詳細については、補足資料を参照してください。絶対的なセキュリティーを提供できる製品またはコンポーネントはありません。

インテルのテクノロジーを使用するには、対応したハードウェア、ソフトウェア、またはサービスの有効化が必要となる場合があります。

各アクセラレーターの利用可否は SKU ごとに異なります。詳細については、担当のインテル販売代理店までお問い合わせください。

実際のコストや結果は異なる場合があります。

©2024 Intel Corporation. Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® AI 向け製品ポートフォリオ

Software	oneAPI, OpenVINO, PyTorch, deepspeed, TensorFlow, XGBoost, ONNX RUNTIME, pandas, SynapseAI	
AI 専用アクセラレーター	Intel Gaudi	ディープラーニングの学習および推論向けの専用アクセラレーター
汎用アクセラレーター(GPU)	Intel Data Center GPU Flex Series	クラウドゲーミング, VDI, メディア処理、リアルタイム動画配信
	Intel Data Center GPU Max Series	並列計算、HPC, AI学習・推論
汎用プロセッサ	Intel Xeon, Intel Ethernet	リアルタイム、メディアスループット、低遅延、Sparsity のある AI 推論
	Intel Xeon, Intel Ethernet	小-中規模 AI モデルの学習、ファイン・チューニング
	Intel Core Ultra, Intel Xeon D, Intel GPU, Intel Arc Graphics, Intel Core, Intel Ethernet	エッジおよびうネットワーク設備側での AI 推論
	Intel Core, Intel Core Ultra, Intel Arc Graphics	PC などクライアント側でのAI 推論

©2024 Intel Corporation. 無断での引用、転載を禁じます。

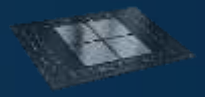
インテル　データセンター　ロードマップ

Efficient -Core
CPU

Performance-Core
CPU

AI
専用アクセラレータ

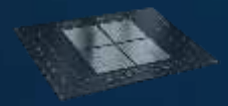
HPC & AI
GPU



インテル® Xeon® 6 プロセッサー
with E-core
コードネーム Sierra Forest

次世代インテル® Xeon® プロセッサー
with E-core
コードネーム Clearwater Forest

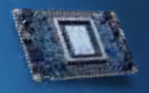
第四世代 インテル® Xeon®
スケーラブル・プロセッサー



第五世代 インテル® Xeon®
プロセッサー
コードネーム Emerald Rapids

インテル® Xeon® 6 プロセッサー
with P-core
コードネーム Granite Rapids

インテル® Xeon®
CPU マックス・シリーズ



Habana®
Gaudi® 2

Habana®
Gaudi® 3

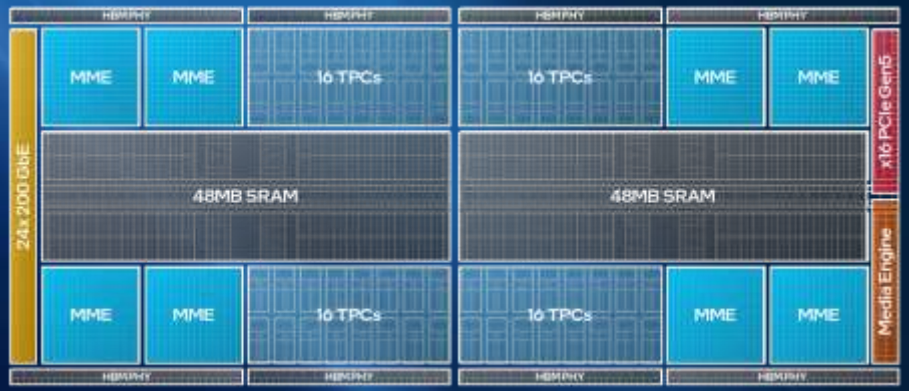
インテル® Data Center GPU マックス・シリーズ

次世代 GPU
コードネーム Falcon Shores

2023 ——— 2025+

©2024 Intel Corporation. 無断での引用、転載を禁じます。

生成 AI のパフォーマンス
と
生産性に特化した
アーキテクチャー設計



AI 専用の設計

効率とパフォーマンスを
大幅に加速

64
テンソル・
プロセッサー・
コア数 (第 5 世代)

8
行列乗算
エンジン数

メモリー拡大による LLM の効率性とコスト・パフォーマンスの向上

128GB
HBM 容量
3.7TB/秒の
帯域幅

96MB
SRAM
12.8TB/秒の
SRAM 帯域幅

超大規模システムに拡張できる柔軟なオンチップ・ネットワーキング

オープン・スタンダード vs.
専用 InfiniBand

24x 200 GbE
業界標準の RoCE
イーサネット・
ポート

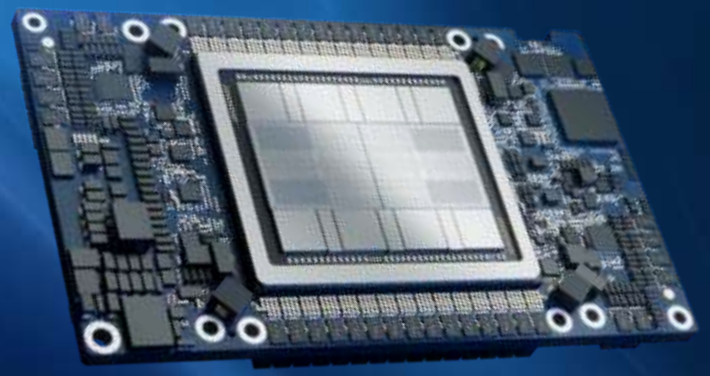
PCIe 5
x16

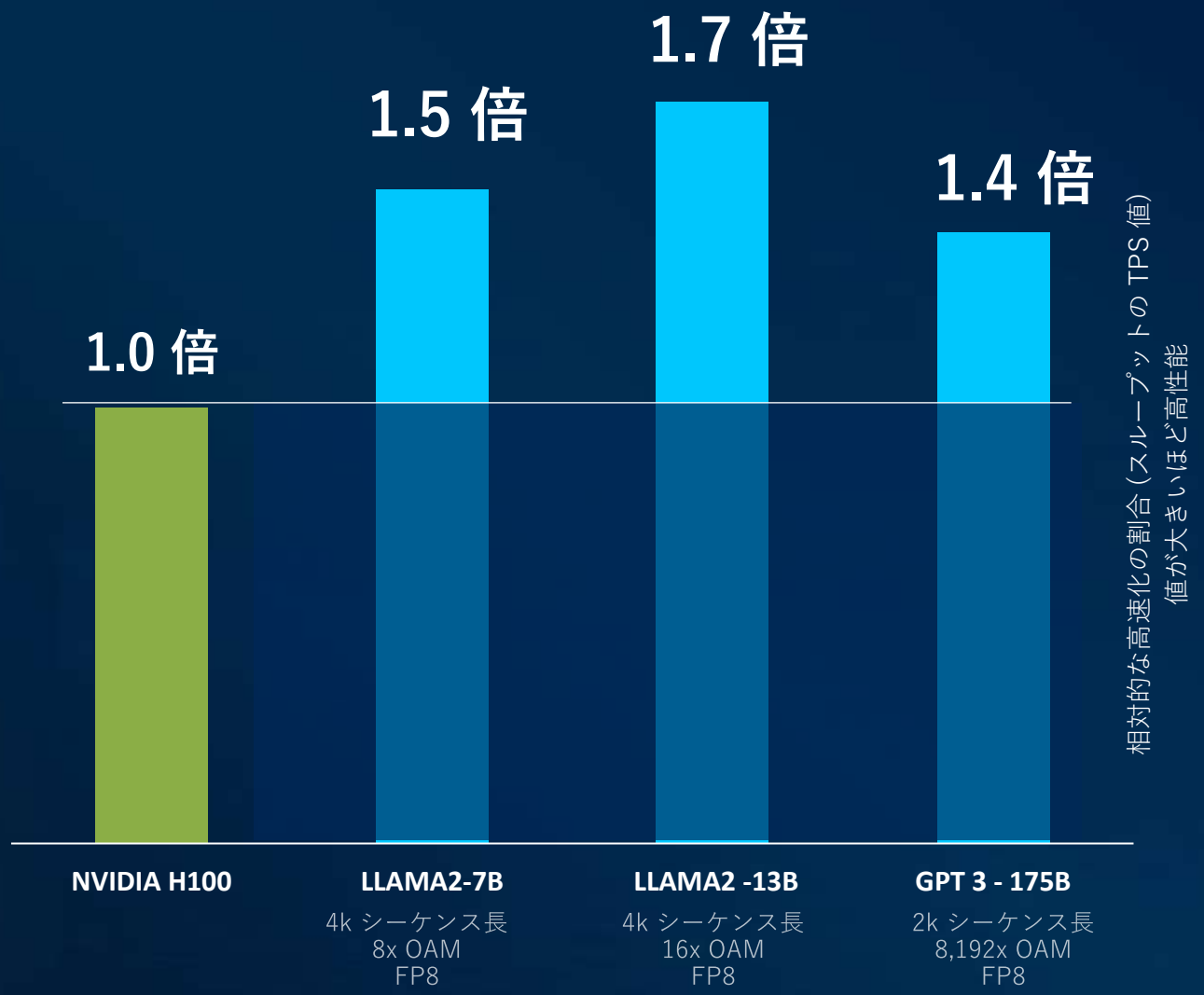
©2024 Intel Corporation. 無断での引用、転載を禁じます。

intel GAUDI

1.5 倍高速の学習処理時間

広く利用されている大規模言語モデルを実行し、インテル® Gaudi® 3 アクセラレーターと NVIDIA H100 の平均の推定パフォーマンスを比較*





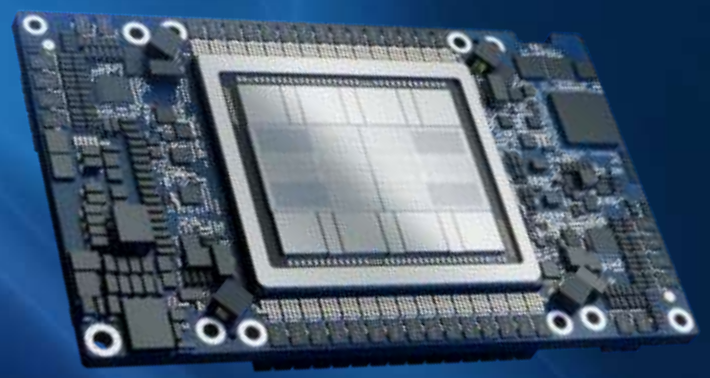
* NVIDIA H100 との比較は、https://developer.nvidia.com/deep-learning-performance-training-inference/training → 「Large Language Model」タブに記載されている 2024年3月28日時点の数値と、LLAMA2-7B、LLAMA2-13B、GPT3-175B を実行した場合のインテル® Gaudi® 3 アクセラレーターの推定値 (2024年3月28日時点) に基づきます。結果は異なる場合があります。

intel
©2024 Intel Corporation. 無断での引用、転載を禁じます。

intel GAUDI

2倍高速の推論性能

広く利用されている大規模言語モデルを実行し、インテル® Gaudi® 3 アクセラレーターと NVIDIA H100 の平均の推定パフォーマンスを比較*



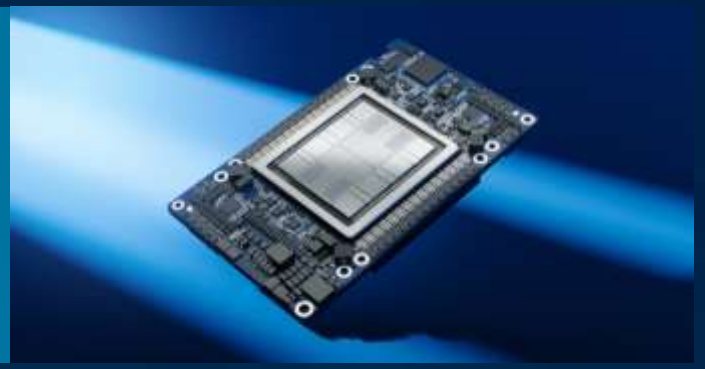
Nvidiaの性能に関する情報源： Overview — tensorrt_llm documentation (nvidia.github.io), 2024年5月 報告されている数字はGPUあたりのもの。
Habana LabsによるインテルGaudi 3予測、2024年4月。

©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Gaudi® 3 アクセラレーター

製品ライン

アクセラレータカード
HL-325L OAM準拠



ユニバーサルベースボード
HLB-325

PCIe CEM
HL-338 アドインカード

©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Gaudi® 3 AI アクセラレーター

製品スケジュール

1H 2024 – サンプル供給開始

インテル® Gaudi® 3
空冷向け

2024年 第1 四半期

インテル® Gaudi® 3
液冷向け

2024年第2 四半期

2H 2024 – 量産開始

インテル® Gaudi® 3
空冷向け

2024年第3 四半期

インテル® Gaudi® 3
液冷向け

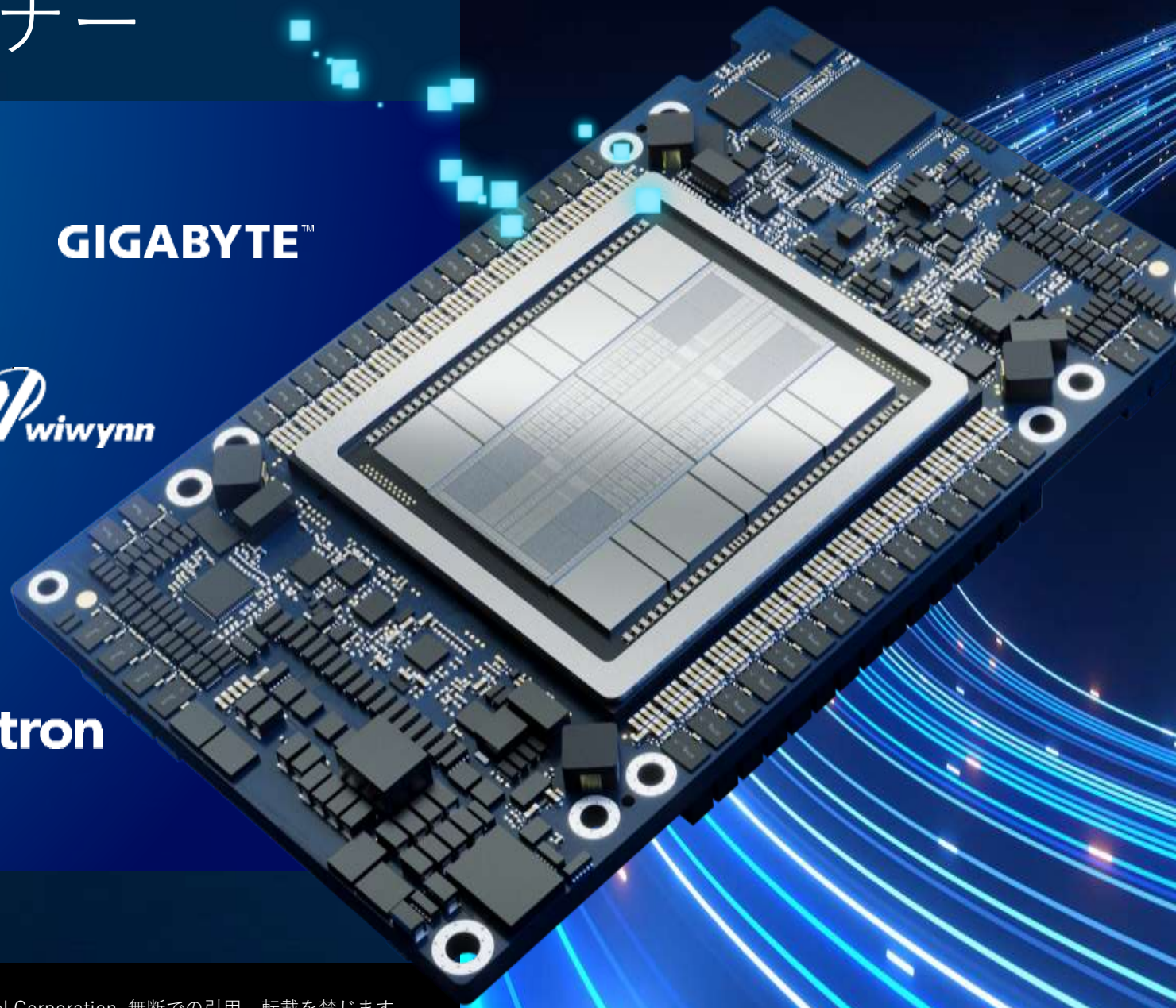
2024年第4 四半期

注釈: 液浸冷却も可能(システム OxM による開発に依拠)

intel.
©2024 Intel Corporation. 無断での引用、転載を禁じます。

システム開発パートナー

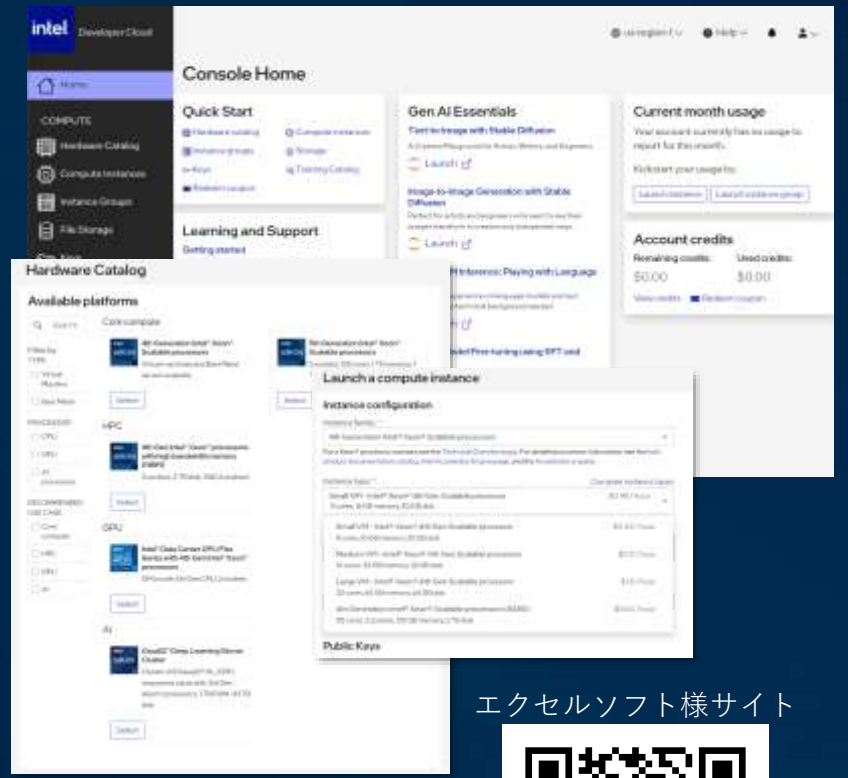
ASUS
DELL Technologies
GIGABYTE
Hewlett Packard Enterprise
ingrasys
wiwynn
Inventec
Lenovo
QCT
SUPERMICRO
wistron



©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Tiber™ デベロッパー・クラウド

ソフトウェア開発者向けのクラウド・コンピューティング・サービス



エクセルソフト様サイト

インテル® Gaudi® 3 アクセラレーターは2024年後半に提供開始

インテル® Tiber™ デベロッパー・クラウド - インテル・ソフトウェア開発製品 正式販売代理店 | XLsoft エクセルソフト

インテルのブースで無料クーポン($250)を配布中。ぜひお立ち寄りください。

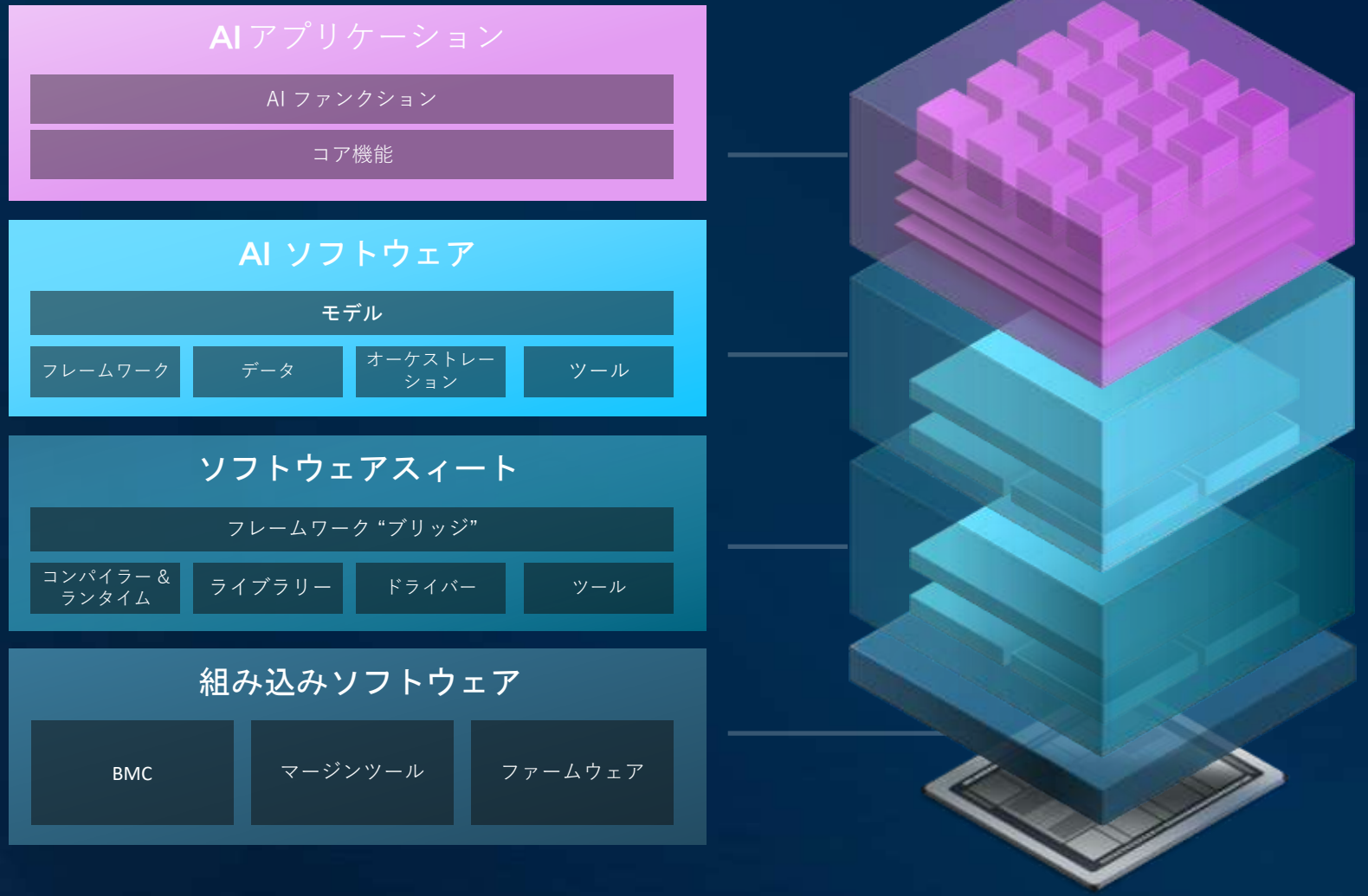
エクセルソフト株式会社様が導入支援(日本語での購入サポート、日本円での請求書払いなどの対応)

©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Gaudi® 3 アクセラレーター

高い生成AI性能のための エンド・ツー・エンドの AIソフトウェア・スタック

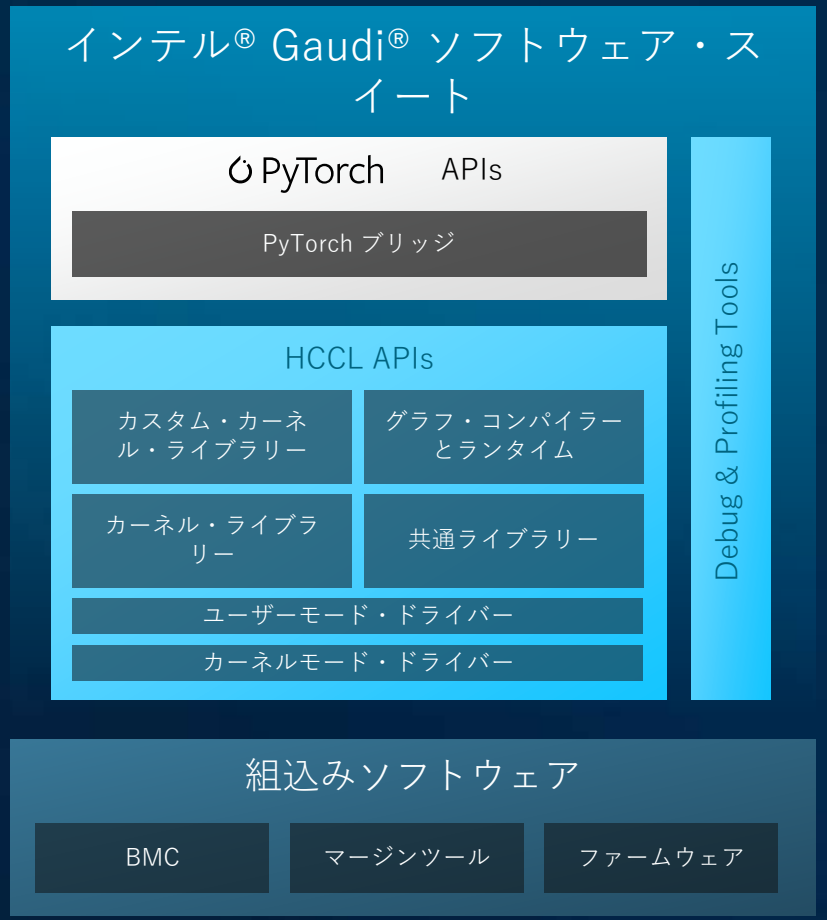
カーネルから アプリケーションまで

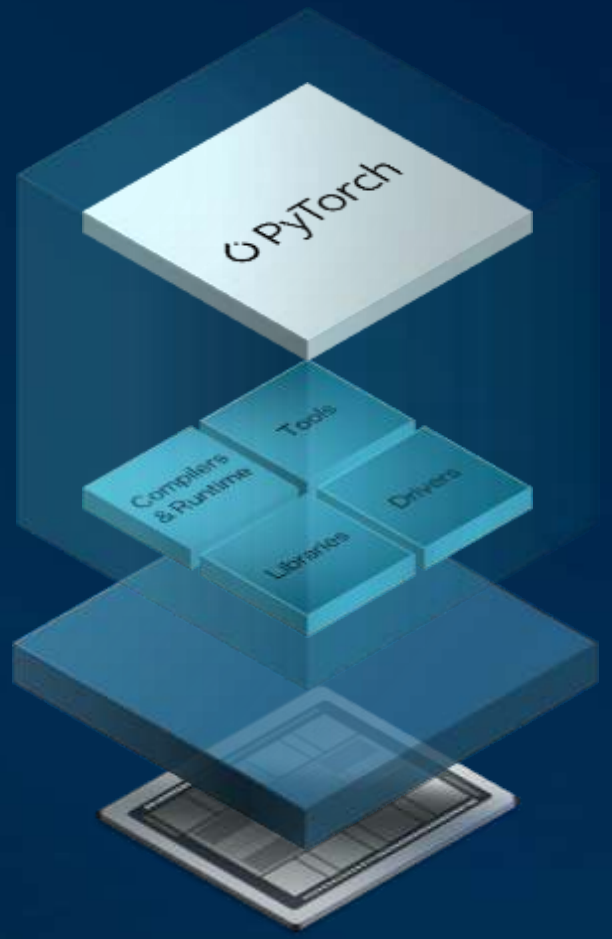


©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Gaudi® ソフトウェア・スイート

インテル® Gaudi® ソフトウェア上で実行される何百もの Gen AI モデルへのエコシステム・アクセスにより、開発が容易になります





©2024 Intel Corporation. 無断での引用、転載を禁じます。

主なモデル実行方法

全ての方法にてステップバイステップのドキュメントが用意

GPU からの移行

手動での作業または、Intel® Gaudi PyTorch パッケージに含まれるGPU Migration Toolkit を使って僅かなステップにて、GPU向けのモデルをインテル® Gaudi® AI アクセラレーターに移行可能

GitHub 上の最適化済みモデルを利用

完全に検証済みの PyTorch 上の参照モデルを提供。性能評価だけでなく、詳細な設定の検証、独自のカスタムを始める上での出発点として最適

*20以上のモデルが利用可能**

Hugging Face 上の最適化済みモデルを利用

使い慣れた Hugging Face のインターフェースが利用可能なモデルを提供

15 以上のタスクと30以上の最適化済みモデルが利用可能*

*2024年6月1日時点での状況。最新の対応状況は、 https://github.com/HabanaAI/Model-References 、および https://github.com/huggingface/optimum-habana をご参照ください

©2024 Intel Corporation. 無断での引用、転載を禁じます。

主なモデル実行方法

全ての方法にてステップバイステップのドキュメントが用意

GPU からの移行

手動での作業または、Intel® Gaudi PyTorch パッケージに含まれるGPU Migration Toolkit を使って僅かなステップにて、GPU向けのモデルをインテル® Gaudi® AI アクセラレーターに移行可能

GitHub 上の最適化済みモデルを利用

完全に検証済みの PyTorch 上の参照モデルを提供。性能評価だけでなく、詳細な設定の検証、独自のカスタムを始める上での出発点として最適

*20以上のモデルが利用可能**

Hugging Face 上の最適化済みモデルを利用

使い慣れた Hugging Face のインターフェースが利用可能なモデルを提供

15 以上のタスクと30以上の最適化済みモデルが利用可能*

*2024年6月1日時点での状況。最新の対応状況は、 https://github.com/HabanaAI/Model-References 、および https://github.com/huggingface/optimum-habana をご参照ください

©2024 Intel Corporation. 無断での引用、転載を禁じます。

主なモデル実行方法

全ての方法にてステップバイステップのドキュメントが用意

GPU からの移行

手動での作業または、Intel® Gaudi PyTorch パッケージに含まれるGPU Migration Toolkit を使って僅かなステップにて、GPU向けのモデルをインテル® Gaudi® AI アクセラレーターに移行可能

GitHub 上の最適化済みモデルを利用

完全に検証済みの PyTorch 上の参照モデルを提供。性能評価だけでなく、詳細な設定の検証、独自のカスタムを始める上での出発点として最適

*20以上のモデルが利用可能**

Hugging Face 上の最適化済みモデルを利用

使い慣れた Hugging Face のインターフェースが利用可能なモデルを提供

15 以上のタスクと30以上の最適化済みモデルが利用可能*

*2024年6月1日時点での状況。最新の対応状況は、 https://github.com/HabanaAI/Model-References 、および https://github.com/huggingface/optimum-habana をご参照ください

©2024 Intel Corporation. 無断での引用、転載を禁じます。

Hugging Face Optimum Habana

元のコード

```
from transformers import Trainer, TrainingArguments

training_args = TrainingArguments(
  # training arguments...
)
  :
  :
# Initialize our Trainer
trainer = Trainer(
    model=model,
    args=training_args,  # Original training arguments.
    train_dataset=train_dataset if training_args.do_train
else None,
    eval_dataset=eval_dataset if training_args.do_eval else
None,
    compute_metrics=compute_metrics,
    tokenizer=tokenizer,
    data_collator=data_collator,
)
```

Gaudi

```
from optimum.habana import GaudiConfig, GaudiTrainer,
GaudiTrainingArguments

training_args = GaudiTrainingArguments(
  # training arguments...
 use_habana=True,
 use_lazy_mode=True,  # whether to use lazy or eager mode
 gaudi_config_name=path_to_gaudi_config,
)
  :
  :
# Initialize our Trainer
trainer = GaudiTrainer(
    model=model,
    args=training_args,  # Original training arguments.
    train_dataset=train_dataset if training_args.do_train
else None,
    eval_dataset=eval_dataset if training_args.do_eval else
None,
    compute_metrics=compute_metrics,
    tokenizer=tokenizer,
    data_collator=data_collator,
)
```

intel.
©2024 Intel Corporation. 無断での引用、転載を禁じます。

主なモデル実行方法

全ての方法にてステップバイステップのドキュメントが用意

GPU からの移行

手動での作業または、Intel® Gaudi PyTorch パッケージに含まれるGPU Migration Toolkit を使って僅かなステップにて、GPU向けのモデルをインテル® Gaudi® AI アクセラレーターに移行可能

GitHub 上の最適化済みモデルを利用

完全に検証済みの PyTorch 上の参照モデルを提供。性能評価だけでなく、詳細な設定の検証、独自のカスタムを始める上での出発点として最適

*20以上のモデルが利用可能**

Hugging Face 上の最適化済みモデルを利用

使い慣れた Hugging Face のインターフェースが利用可能なモデルを提供

15 以上のタスクと30以上の最適化済みモデルが利用可能*

*2024年6月1日時点での状況。最新の対応状況は、 https://github.com/HabanaAI/Model-References 、および https://github.com/huggingface/optimum-habana をご参照ください

©2024 Intel Corporation. 無断での引用、転載を禁じます。

PyTorch での移行手順 (マニュアルの場合)

1 事前の準備: CUDA のコールを削除

2 インテル® Gaudi ® 対応 Torch ライブラリのインポート

```
# Import Habana Torch Library
import habana_frameworks.torch.core as htcore

# neural network model
class SimpleModel(nn.Module):
...
```

3 学習ループ部分のコードの修正

```
# training loop
def train(net,criterion,optimizer,trainloader,device):
...
        loss.backward()
        htcore.mark_step()

        optimizer.step()
        htcore.mark_step()

def main():
...
    # Target the Gaudi HPU device
    device = torch.device("hpu")
```

4 Autocast の設定変更

```
with torch.autocast(device_type="hpu", dtype=torch.bfloat16):
      output = model(input)
      loss = loss_fn(output, target)
loss.backward()
```

5 (オプション) 分散学習用の設定変更

```
import habana_frameworks.torch.distributed.hccl
torch.distributed.init_process_group(backend='hccl')

...
# Use with PyTorch DDP Hook
ddp_model = DDP(model)

(model) loss_fn = nn.MSELoss()
optimizer = optim.SGD(ddp_model.parameters(), lr=0.001)

optimizer.zero_grad()
outputs = ddp_model(torch.randn(20, 10).to(device))
```

intel.
©2024 Intel Corporation. 無断での引用、転載を禁じます。

PyTorch での移行手順
(GPU Migration Toolkit 利用の場合)

1 Pre Work: Removing CUDA calls

2
```
import torch

# Import Habana Torch Library

import habana_frameworks.torch.core as htcore

import habana_frameworks.torch.gpu_migration

# neural network model

class SimpleModel(nn.Module):

...
```

3
```
# training loop
def train(net,criterion,optimizer,trainloader,device):
...
        loss.backward()
        htcore.mark_step()

        optimizer.step()call to trigger execution
        htcore.mark_step()
```

```
def main():
...
    device = torch.device("hpu")
```

4 Autocast For Mixed Precision

```
with torch.autocast(device_type="hpu", dtype=torch.bfloat16):
    output = model(input)
    loss = loss_fn(output, target)
loss.backward()
```

わずか 2 ステップ
作業完了

```
# Use with PyTorch DDP Hook
ddp_model = DDP(model)

(model) loss_fn = nn.MSELoss()
optimizer = optim.SGD(ddp_model.parameters(), lr=0.001)

optimizer.zero_grad()
outputs = ddp_model(torch.randn(20, 10).to(device))
```

©2024 Intel Corporation. 無断での引用、転載を禁じます。

GPU 依存のコードの自動変換

GPU Migration API によって以下のような変更ログを出力

- torch.cuda
- NVIDIA GPU 関連のパラメーターを持つ Torch API 例: torch.randn(device=”cuda”)
- NVIDIA Apex (PyTorch Extension)
- NVIDIA Pynvml (Management Library)

hpu_match

GPU Migration Toolkit が GPU 依存の引数を Gaudi 対応のものに変更

```
x = torch.ones(1, device="cuda")
GPU Migration changes the argument `device` from "cuda" to "hpu".
```

hpu_modified

GPU 特有の機能を Habana対応のコールに変更

```
torch.cuda.FloatTensor
torch.FloatTensor and torch.Tensor.to(“hpu”)
```

hpu_mismatch

インテル® Gaudi ® にて実装されていない GPU 依存のコードに対して、
GPU Migration Toolkit が Inactive なコールもしくは例外を上げる

Inactive call:

```
torch.cuda.empty_cache()
```

NotImplementedError Exception:

```
torch.cuda.comm.broadcast
```

注：hpu_mismatch が発生した場合は、ログファイルを生成し、ミスマッチした機能を手動で修正してトレーニングを再実行してください
変換されたCUDA コールのリストとHPUとの互換性や API の詳細についてはこちらを参照してください。
https://docs.habana.ai/en/latest/PyTorch/PyTorch_Model_Porting/GPU_Migration_Toolkit/Intel_Gaudi_Migration_APIs.html

intel.
©2024 Intel Corporation. 無断での引用、転載を禁じます。

OSS ソフトウェア/モデル 継続的に対応

フレームワーク＆ライブラリー

- PyTorch FSDPをサポート
- vLLM (v1.16でプレビューサポート)
- Hugging Face TGI (性能改善)
- Optimum Habana: Speculative Sampling

AI モデル

- Optimum Habana： Llama3, Qwen2, Gemma, SVD, Whisper, Phi, ControlNet, Dreambooth, SDXL
- Megatron-DeepSpeed： Mixtral 4x7B/8x7B, Llama2 70B FP8
- 性能改善: Llama2 7B/13B/70B FP8/BF16 (学習), Llama2 7B/70B FP8 (推論), Mixtral 4x7B/8x7B (推論), etc.

©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Gaudi® 2 アクセラレーター採用事例：Stability AI

stability.ai

- インテルが構築する、4,000 個のインテル® Gaudi® 2 アクセラレーターを搭載した AI スーパーコンピューターの中心顧客
- Stable Diffusion 3のMultimodal Diffusion Transformer model (MMDiT) を学習
  - 学習時間が H100-80GB より 50% 高速
  - 学習時間が A100-80GB より 300% 高速
- Stable Beluga 2.5 (70B 言語モデル)
  - 推論速度が A100 よりも 28% 高速
- 高い性能と優れた TCO を、非常にわずかなコード変更で実現 (コード移行に要した時間は 1 日以内)

2B Multimodal Diffusion Transformer (MMDiT) における学習スループット

Device	Attention	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec (100-MA)
Gaudi2	FusedSDPA	2	16	32	512	1,254
Gaudi2	FusedSDPA	2	16	16	256	927
H100-80GB	xFormers	2	16	16	256	595
A100-80GB	xFormers	2	16	16	256	381

Device	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec	Images / sec / device
Gaudi2	32	256	16	4096	12,654	49.4
A100-80GB	32	256	16	4096	3,992	15.6

https://stability.ai/news/building-new-ai-supercomputer
https://stability.ai/news/putting-the-ai-supercomputer-to-work

intel.
©2024 Intel Corporation. 無断での引用、転載を禁じます。

インテル® Tiber™ デベロッパー・クラウドを利用してコスト削減

コスト削減

- 他社のクラウドサービスと比較してベアメタルのAIアクセラレーターやGPUが平均20-40%割安
- 3,6,12ヶ月のELA契約では10-30%ディスカウント

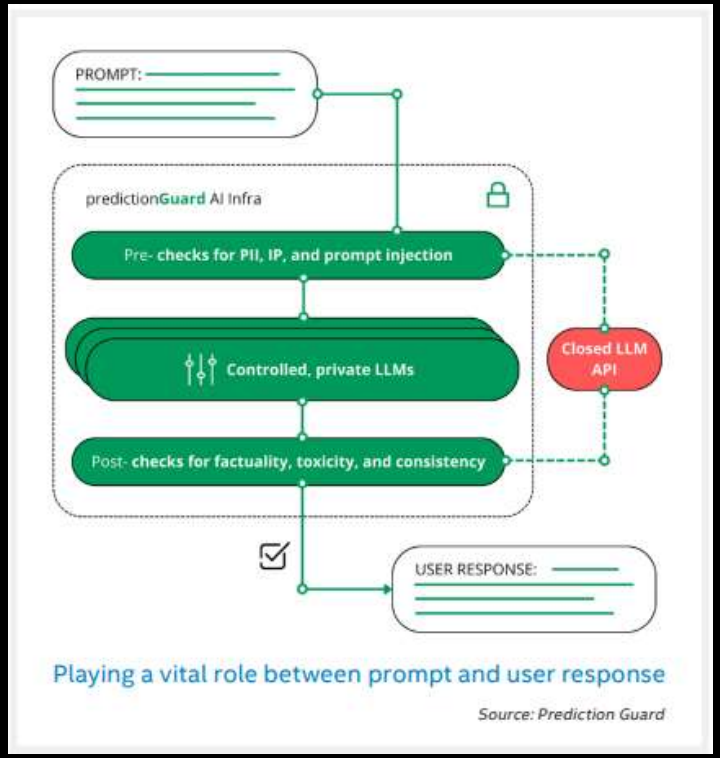
他社にはない特徴

- インテル® Gaudi2® アクセラレーターを利用可能(マルチノードも可)
- CPU,GPU,アクセラレータを混在させて利用可能
- それぞれのHWに最適化されたソフトウェア・スタックが上ですぐに利用可能

採用事例

prediction Guard

- Prediction Guardは出力コンテンツの事実性や有害性を判定、入力のPPIやPIを検出する機能を持ったLLMのAPIサービスを提供
- GPUを使用した他のプロバイダからIDCのインテル® Gaudi®2 アクセラレーターに移行することで80%のコスト削減＆レイテンシーを半減
- **Intel® Liftoff for Startups**プログラムが提供するIDCへのアクセス、技術サポートや共同マーケティングを活用



Beewant codee DIGITAL CORTEX Dive Hasty ÎSÍMA
moonshot seekr sitemana Weaviate Starwit Technologies

©2024 Intel Corporation. 無断での引用、転載を禁じます。

まとめ

- インテル®は生成AI向けの新しい選択肢としてAI専用アクセラレーターGaudi®を提供
- 大規模な学習や推論に必要なOSSはサポートされており、継続して性能や機能を改善
- インテル® Tiber™ デベロッパー・クラウド上でインテル® Gaudi®アクセラレーターを評価可能
- インテル®ブースではGaudi®3の展示やインテル® Tiber®デベロッパー・クラウドの紹介しています。ぜひお立ちよりください。

©2024 Intel Corporation. 無断での引用、転載を禁じます。

Intel、 インテル、 Intel ロゴ、 その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。
その他の社名、 製品名などは、 一般に各社の表示、 商標または登録商標です。
©2024 Intel Corporation. 無断での引用、転載を禁じます。