ハイパフォーマンス ソフトウェア カンファレンス 春

ハイパフォーマンス・コンピューティングに価値を発揮するインテル® FPGA

インテル株式会社 プログラマブル・ソリューションズ営業本部
事業開発マネージャー

髙藤 良史

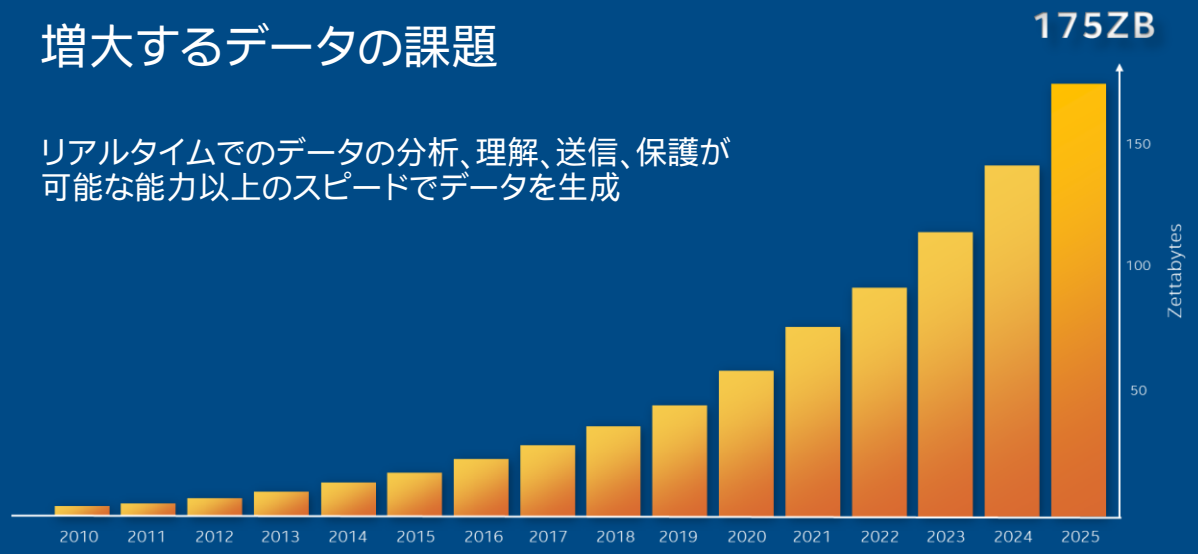
intel®

データの爆発的増加がけん引する ヘテロジニアス・コンピューティング

グローバルで作成されるデジタルデータ量の推移*

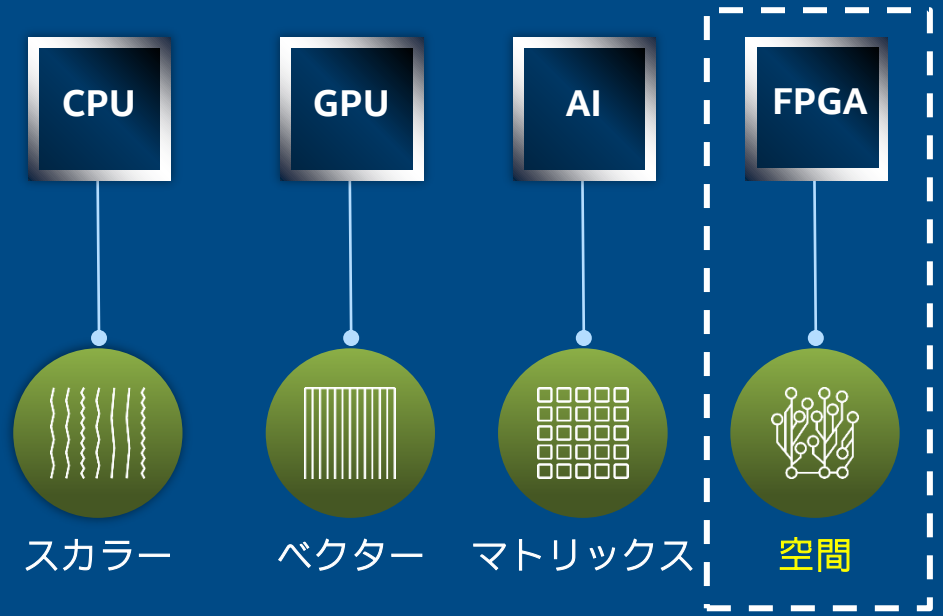
増大するデータの課題

リアルタイムでのデータの分析、理解、送信、保護が可能な能力以上のスピードでデータを生成



*Source: IDC Global DataSphere, May 2020

インテルのヘテロジニアス・アーキテクチャー戦略



FPGAとは？

- 動的に再プログラム可能なハードウェア回路を形成するシリコンデバイス
- 様々なワークロードに対応可能なデータパスを備え、処理速度が高く電力効率に優れた、低レイテンシーのサービスを提供

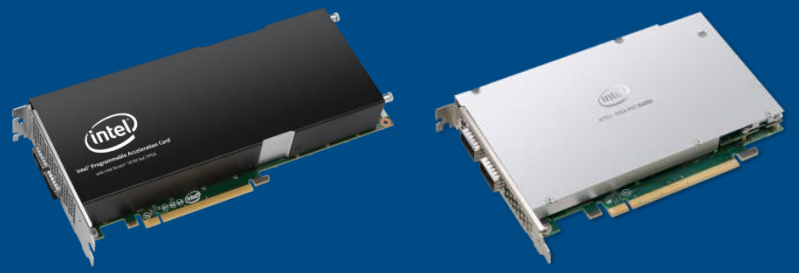
データ分析　**映像処理**　**ネットワーク**　**符号化処理**

組込み・アプライアンス向けのチップ製品

intel AGILEX™

intel STRATIX® 10

コンピューティング向けのカード製品



ヘテロジニアス・コンピューティングでの FPGA の役割

FPGA はヘテロジニアス・システムにおいて他のテクノロジーの補完的役割

- 他のアーキテクチャーの性能上のボトルネックにアプローチ

FPGA のターゲット

リアルタイム性が要求される処理

データが再利用される繰り返し処理

混合精度やベクターが使われる処理

サブブロックと I/O の間のストリーミング・データ処理

演算密度や電力効率を考慮すべき処理

インテル® Stratix®10 FPGA ユースケース

ホワイトペーパー: インテル® Stratix® 10 NX FPGA での WaveNet ニューラルネットワーク実行

自然な音声オーディオストリームをリアルタイムに生成できる最先端のボコーダーの WaveNet を、16 kHz で 256 並列実装

最適化された GPU ソリューションよりも 8 倍高いスループットで提供

インテル® Stratix® 10 NX FPGA によってニューラル・ネットワークを搭載した現実のアプリケーションに対して**高スループット、低レイテンシー**の推論を提供

(Myrtle.ai 社 ホワイトペーパー要約より)

	MYRTLE.AI WAVENET	NV-WAVENET
Platform	Intel® Stratix® 10 NX FPGA	NVIDIA* V100 GPU
Frequency (MHz)	240	1530
Numerical Precision	BFP16 / bfloat16	fp16
WaveNet Configuration	r=120, s=240, L=16, a=256, D=8	r=128, s=256, L=16, a=256, D=8
Operations per 1 second audio (GOPS/second)	65.03	60.36
Model Parameters (Millions)	2.08	1.99
Concurrent Voice Channels	256	32
Application TOPS	16.6	1.93
TDP Power (W)	215	250
Performance per Watt (GOPS/W)	77.4	7.7
Voice Channels per Watt (1/W)	1.19	0.128

Table 5. Performance Results for WaveNet Implementation at 16 kHz.

WhitePaper:
Text to Speech Synthesis Using Intel® Stratix® 10 NX FPGA

ホワイトペーパー: 2次元 FFT の高速化 インテル® Stratix® 10 MX での HBM2 とインテル® oneAPI の使用

なぜ 2次元 FFT を実装したのか?

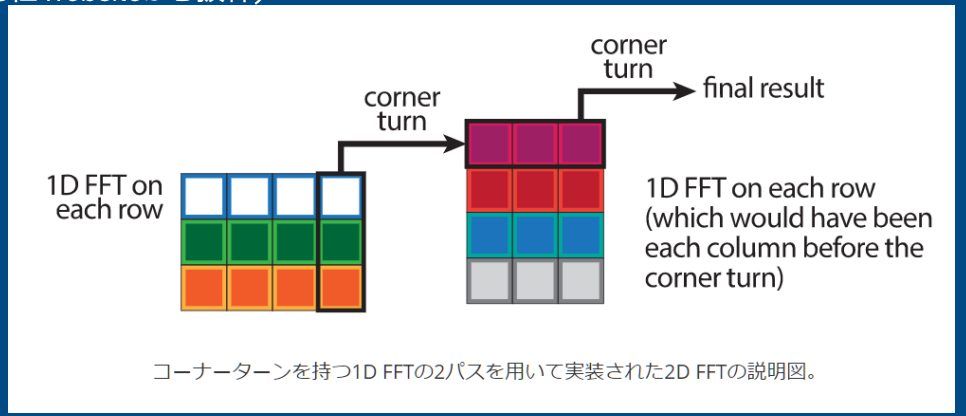
2D FFT は FPGA の IP ライブラリーに含まれていることが多く、プログラマーが自分で実装するものではありません。しかし、並列ハードウェア上での 2D FFT の一般的なインプリメンテーション戦略には、「コーナーターン」または「データ転置」ステップが含まれており、これは CPU や GPU 上で大きなパフォーマンスのボトルネックとなっています。

フリーコーナーターン

このデモで強調した洞察は、FPGA インプリメンテーションが計算と並行してデータ転置を実行できるため、レイテンシーの観点からほぼ「フリー」になるということです。

GPU アーキテクチャーでは、GPU 内部に十分なメモリーがなく、HBM2/DDR6 メモリーに触れることなく中間結果をパイプラインすることができないため、GPU では同じことができません。

(BittWare社 websiteから抜粋)



コーナーターンを持つ1D FFTの2パスを用いて実装された2D FFTの説明図。

WhitePaper (BittWare Website)
https://www.bittware.com/ja/resources/hbm2-2d-fft-oneapi/

HPC 向けインテル® FPGA 関連製品

インテル® FPGA ハイエンド・デバイス

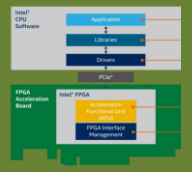
インテル® Agilex™ FPGA

インテル® Stratix® 10 FPGA

パートナーボード
- BittWare 社カード

インテル® FPGA SmartNIC/PAC ポートフォリオ

インテル® OFS
(Open FPGA Stack)



Silicom FPGA
SmartNIC N5010

FPGA SmartNIC
C5000X-PF

インテル® oneAPI プロダクト for FPGA

インテル®
oneAPI

インテル® oneAPI
DPC++ コンパイラー

インテル® Vtune
プロファイラー

インテル® FPGA ハイエンド・デバイス

インテル® FPGA

急激に変化する世界へフレキシビリティーを提供

intel AGILEX™ — データ・セントリック時代へ

intel STRATIX® — 最適化されたバンド幅

intel ARRIA® — ミッドレンジ FPGA & SoC

intel CYCLONE® / intel MAX® — 低コスト FPGA

インテル® Agilex™ FPGA：データ中心の世界に対応する FPGA

intel AGILEX™

データの処理

第 2 世代インテル® Hyperflex™ FPGA アーキテクチャー

平均 45% パフォーマンスが向上[1,3]

最大 40% 消費電力を削減[1,3]

最大 40TFLOPS の DSP 性能[2,3]

データの格納

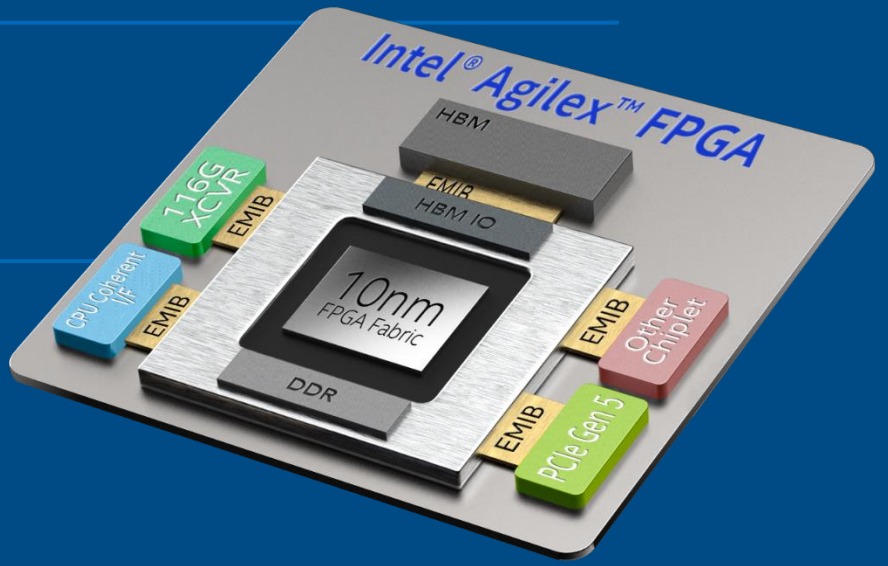
DDR5 と HBM

インテル® Optane ™ パーシステント・メモリーのサポート

データの転送

インテル® Xeon® プロセッサーとの接続性 Compute Express Link (CXL) と PCIe* Gen5

116G トランシーバー・データ・レート



[1] インテル ® Stratix ® 10 FPGA との比較
[2] FP16 コンフィグレーション
[3] 現在の推定値に基づく。構成の詳細については補足資料をご参照下さい。性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/(英語)を参照してください。

インテル® Agilex™ FPGA + インテル® Quartus® Prime 開発ソフトウェア 20.4

FPGA 性能の柔軟性を向上

約 2 倍 消費電力当たりのファブリック性能の向上

競合 7nm FPGA との比較

最大 **400** Gbps イーサネット

業界最速データレートの SerDes トランシーバー

intel AGILEX

業界最先端の FPGA が、5G、ネットワーク、クラウド、エッジにおける多様なワークロードの移行で革新的な適応性と俊敏性を提供

消費電力当たりのファブリック性能が、競合 7nm FPGA と比べて約 2 倍高く、データセンターとその先の次元で柔軟性と電力効率に優れた設計が可能

ビデオ処理アプリケーションにおけるビデオ IP 性能が、競合 7nm FPGA と比較して 50% 高速 (幾何平均)

高速ファイバー接続に不可欠な 高速 5G フロントホール・ゲートウェイ・アプリケーションを実現するファブリック性能を最大 49% 高速化*

* 前世代との比較

インテル® Stratix® 10 NX FPGA

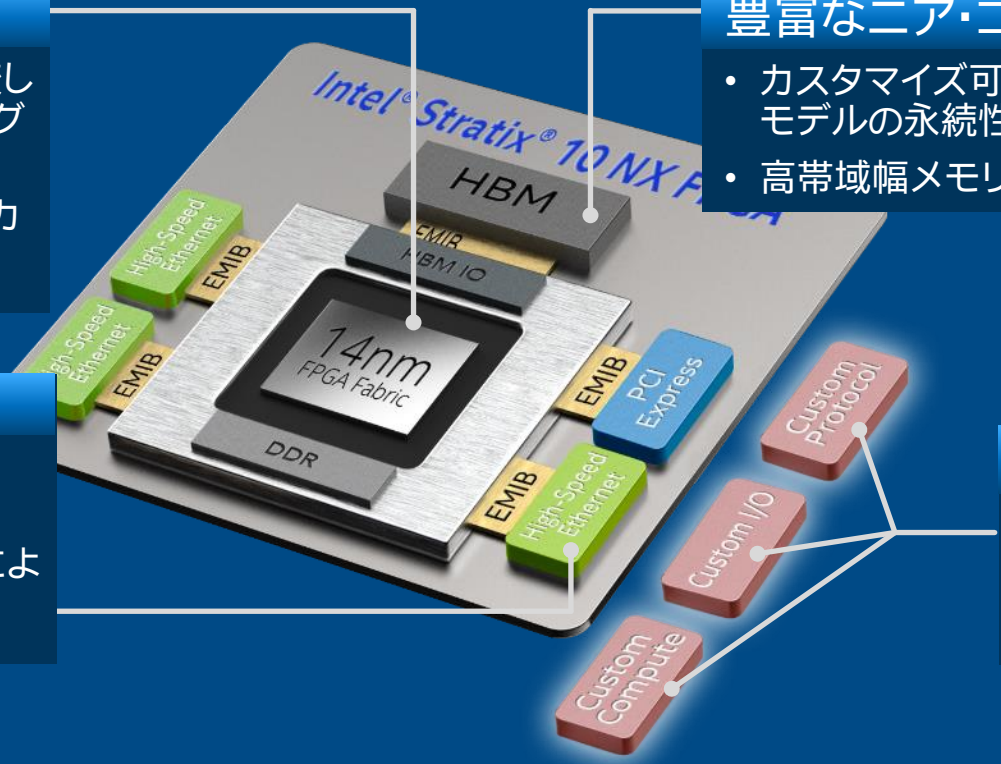
インテル初の AI に最適化された FPGA

高性能 AI Tensor ブロック

- 現行のインテル® Stratix® 10 MX FPGA と比較して、AI ワークロードの INT8 コンピューティング性能が最大 15 倍
- プログラマブルなハードウェアにより、AI のカスタム・ワークロードを実現

広帯域幅ネットワーキング

- 最大 57.8G の PAM4 トランシーバーとイーサネットのハード IP ブロックによる高い効率性
- 柔軟でカスタマイズ可能なインターコネクトにより、複数のノードにわたって拡張可能



豊富なニア・コンピューティング・メモリー

- カスタマイズ可能なメモリー階層によりモデルの永続性を実現
- 高帯域幅メモリー（HBM）内蔵

拡張性

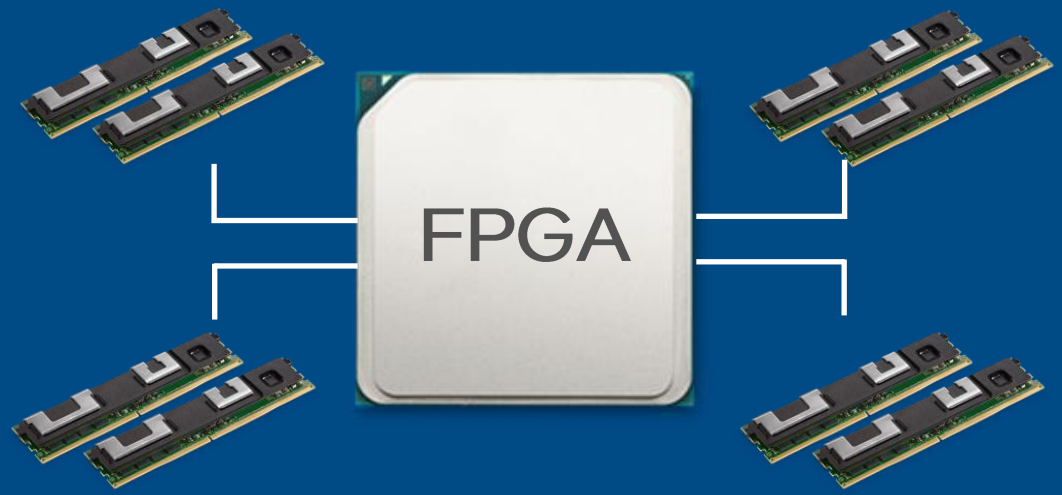
- インターフェイスのカスタマイズと ASIC による拡張を容易にするチップレット・アーキテクチャー

AI Tensor ブロック、ニアメモリー、ネットワーキングにより AI に最適化された高性能なハードウェアを実現

パフォーマンスの測定結果は、インテルによる推定値です。構成の詳細については、補足資料を参照してください。
性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/（英語）を参照してください。

インテル® FPGA + インテル® Optane™ パーシステント・メモリー

FPGA メモリー・サブシステムにより最大 **4TB** のインテル® Optane™ パーシステント・メモリーにアクセス可能

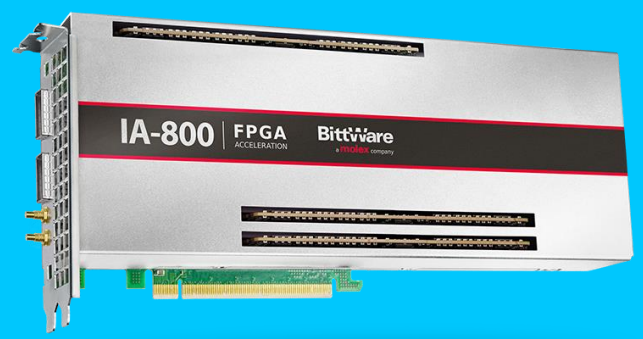


インテル® Optane™ パーシステント・メモリーによるコスト削減、処理能力の向上、高速化:

- TCO の削減[1]
- DDR4 DIMM よりも大容量
- NVMe* フラッシュ SSD よりも高速[2]
- アルゴリズムの高速化 (FPGA)

注 1: 1GB 当たりのコストに基づく、調査時点での価格。30ページを参照してください。
注 2: サードパーティーから提供された性能の測定結果。30ページを参照してください。性能やベンチマーク結果について、さらに詳しい情報をお知りになりたい場合は、http://www.intel.com/benchmarks/ (英語) を参照してください。

パートナー・プロダクト: BittWare IA-840F & 520NX



BittWare IA-840F
インテル® Agilex™
F-Series FPGA 搭載
FPGA Accelerator
PCIe カード

- インテル® Agilex™ AGF027 FPGA (2.6M LE)
- インテル® oneAPI サポート
- ハイパフォーマンス I/O
  - 3x QSFP-DD インターフェイスポート
  - PCIe Gen4 x16 ホストインターフェイス
  - さまざまなアプリケーション向けのMCIO 拡張ポート

intel. AGILEX™

IA-840F Datasheet PDF (bittware.com)

BittWare 520NX
インテル® Stratix® 10 NX FPGA搭載
FPGA Accelerator
PCIe カード

- インテル® Stratix 10 ™ NX2100 FPGA (2.1M LE)
- 8GB HBM2 DDR4 DIMM 2スロット, QDR-II+, インテル® Optane™
- ハイパフォーマンス I/O
  - 4x QSFP28 インターフェイスポート
  - PCIe Gen3 x16 ホストインターフェイス
  - さまざまなアプリケーション向けの OCuLink 拡張ポート

intel. STRATiX 10

520NX Datasheet PDF (bittware.com)

インテル® FPGA ハイエンド・デバイスのまとめ

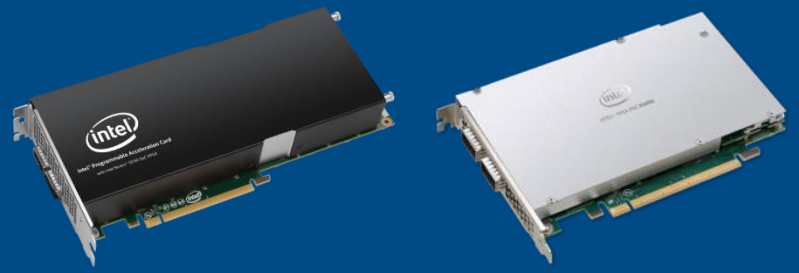
インテル® Agilex™ FPGA
- インテル 10nm SuperFin テクノロジー
- アーキテクチャーの革新
- 平均 45% 向上した高パフォーマンスの実現*

* インテル ® Stratix ® 10 FPGA との比較

インテル® Stratix® 10 FPGA
- インテル® Stratix® 10 NX
  - 高性能 AI Tensor ブロック
- 最大 4TB のインテル® Optane™ パーシステント・メモリーにアクセス可能

パートナーボード
- BittWare IA-840F
  - インテル® Agilex™ AGF027 FPGA 搭載
- BittWare 520NX
  - インテル® Stratix 10® NX2100 FPGA 搭載

インテル® FPGA SmartNIC/PAC ポートフォリオ



インテル® FPGA SmartNIC/PAC のポートフォリオ

クラウドとネットワークの変革

インテル® FPGA PAC D5005

インテル® PAC
インテル® Arria® 10 GX
FPGA搭載版

初の 商業量産
FPGA 搭載 SmartNIC

- アプリケーションとインフラストラクチャーのネットワーク機能仮想化 (NFV)
- 仮想無線エリア・ネットワーク (vRAN)

インテル® FPGA PAC N3000
ネットワーキング向け
SmartNIC

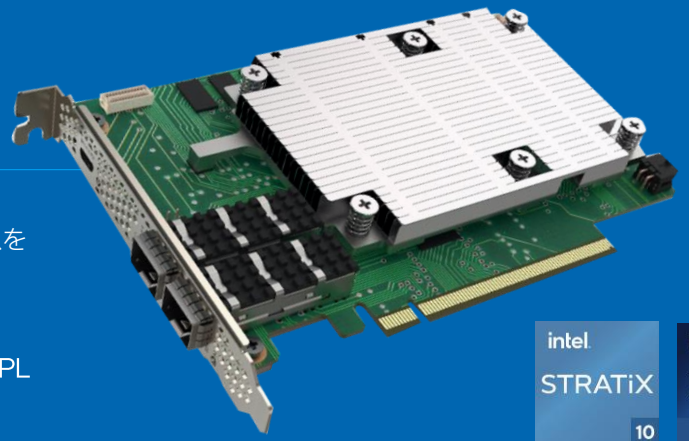
intel ARRiA 10

インテル®
イーサネット

verizon Rakuten NOKIA f5

Inventec FPGA SmartNIC C5020X (クラウド向け)

- インテル® Xeon® D プロセッサー + FPGA プラットフォーム、ハードウェア・プログラマブル・データパスを提供
- 対象: ストレージ / 仮想スイッチのワークロード
- クラウド向けインテル® FPGA SmartNIC C5000X-PL プラットフォームが基盤



intel STRATiX 10

intel XEON

intel STRATiX 10

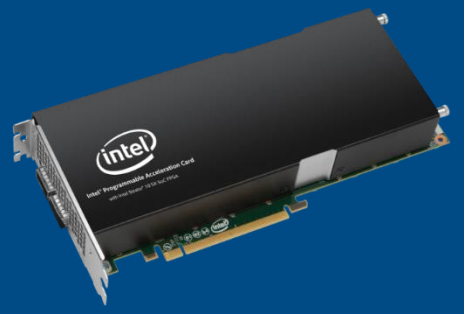
インテル® イーサネット
インテル® Tofino™ スイッチ

Silicom FPGA SmartNIC N5010 (ネットワーキング向け)

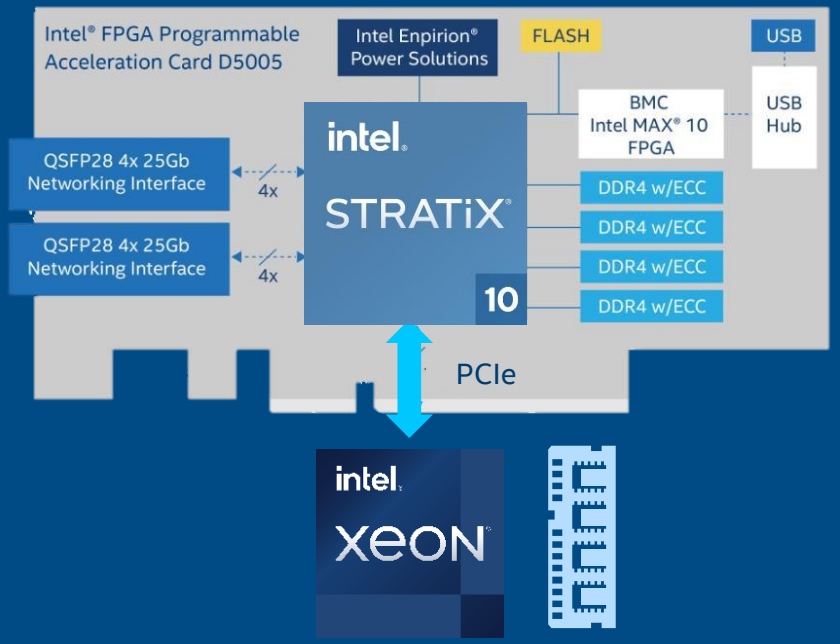
- ハードウェア・プログラマブル 4x100GE FPGA アクセラレーションを実装する初の SmartNIC: 次世代の IA ベースのサーバーと Tofino ベースのスイッチにより、5G コア・ネットワーク (UPF)、アクセス・ゲートウェイ (BNG、AGF)、セキュリティー機能 (ファイアウォール、IPsec) のパフォーマンスとスケーリングのニーズに対応。
- 機能 / アクセス・ゲートウェイ機能、その他のワークロード

カードのコンセプトの違い

D5005

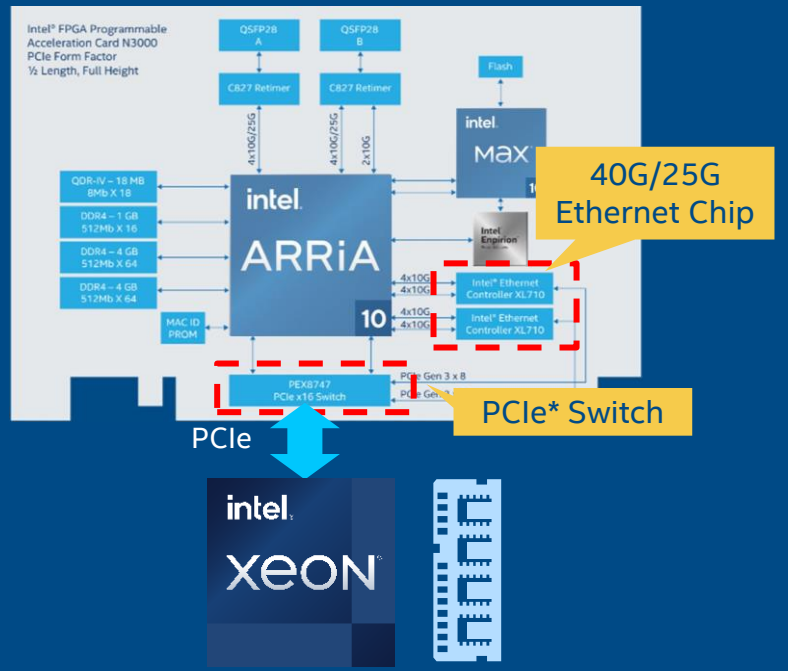


データセンター
ワークロード向け



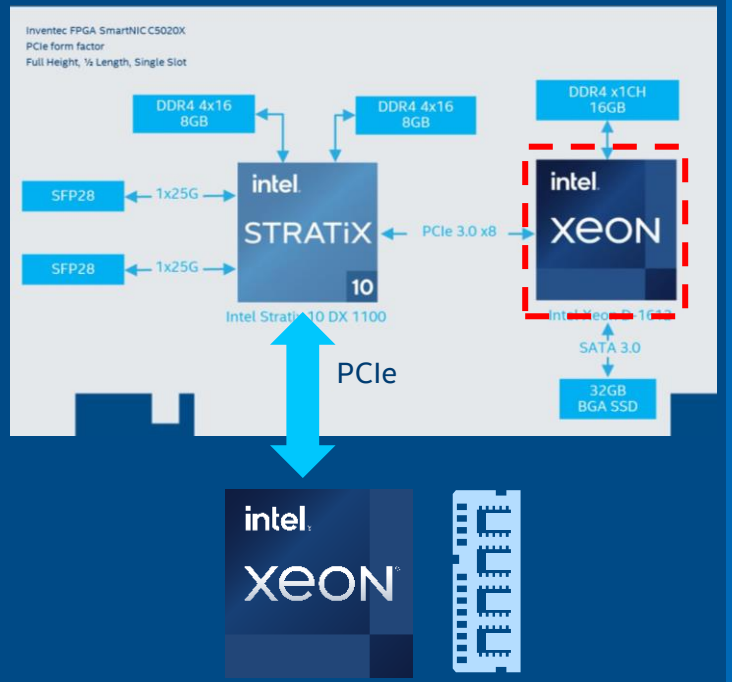
N3000

ネットワーク、NFV 向け



C5010X

クラウドインフラ向け



インテル® オープン FPGA スタック 概要

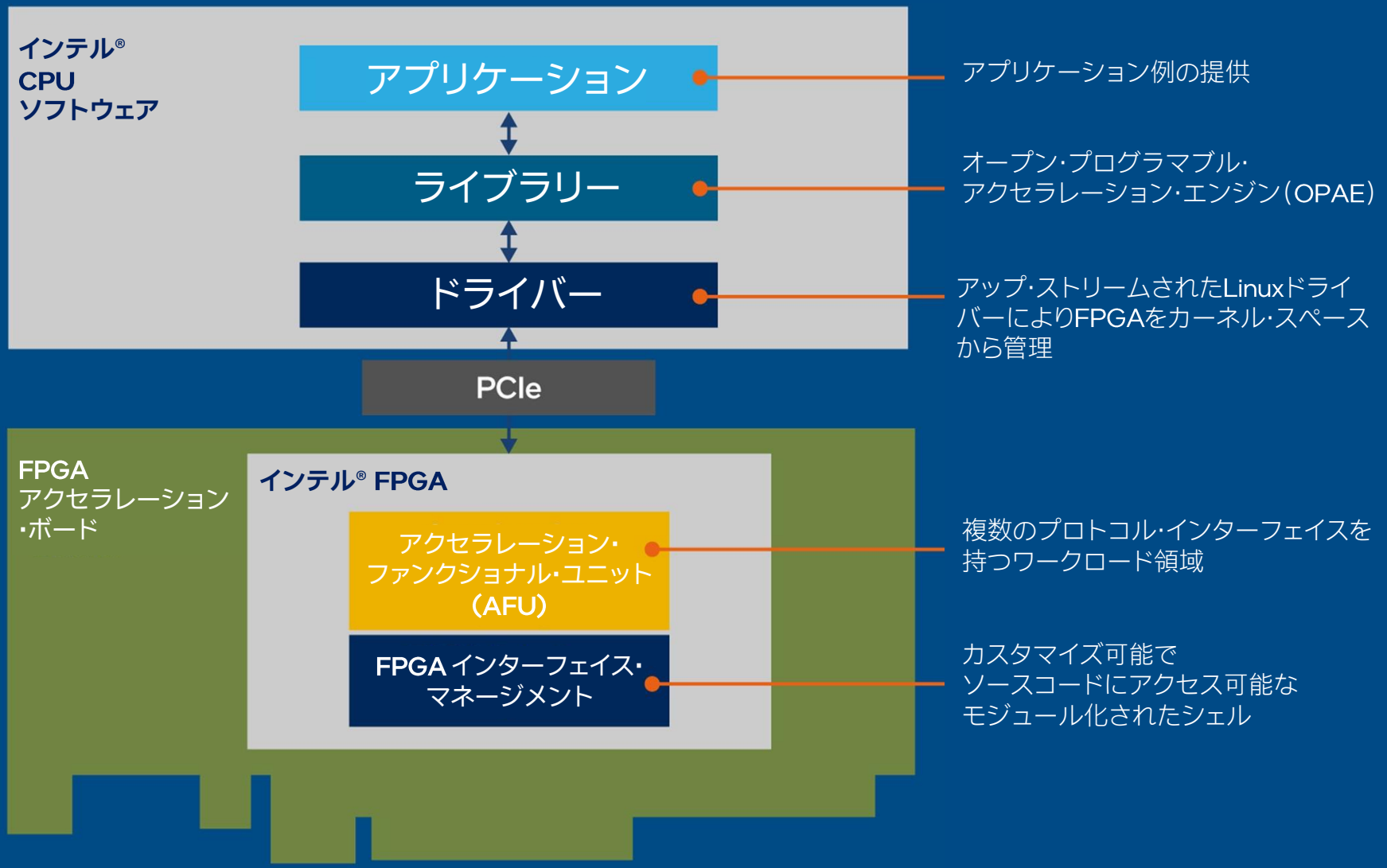
ソフトウェアの拡張性

- ベアメタルと仮想化使用モデルの拡張性
  - ホストのリセットや再初期化が不要のリモート・アップデート
- 標準ソフトウェアのアプリケーション・フレームワークへの統合

ハードウェアの拡張性

▪アプリケーション固有のシェルを構築するモジュール式の組み合わせ可能なハードウェア

▪標準化による効率的な再利用と、プラットフォーム・デザインや移植性を提供するエコシステム



Silicom FPGA SmartNIC N5010

高性能ネットワーキング・アクセラレーション・カード

- インテル® DPDK および OPAE を介して再構成可能

インテル® Stratix® 10 FPGA 搭載

- 高速イーサネット・サポート：4x100GbE / 4x4x25GbE
- 2x PCIe* Gen 4 x16（16物理）
- 32GB DDR4メモリー、144MB QDR4、8GB HBM（4GB/スタック）
- 監視制御用BMC（PLDM）
- ¾ レングス、フルハイト、デュアルスロット PCIe* カード

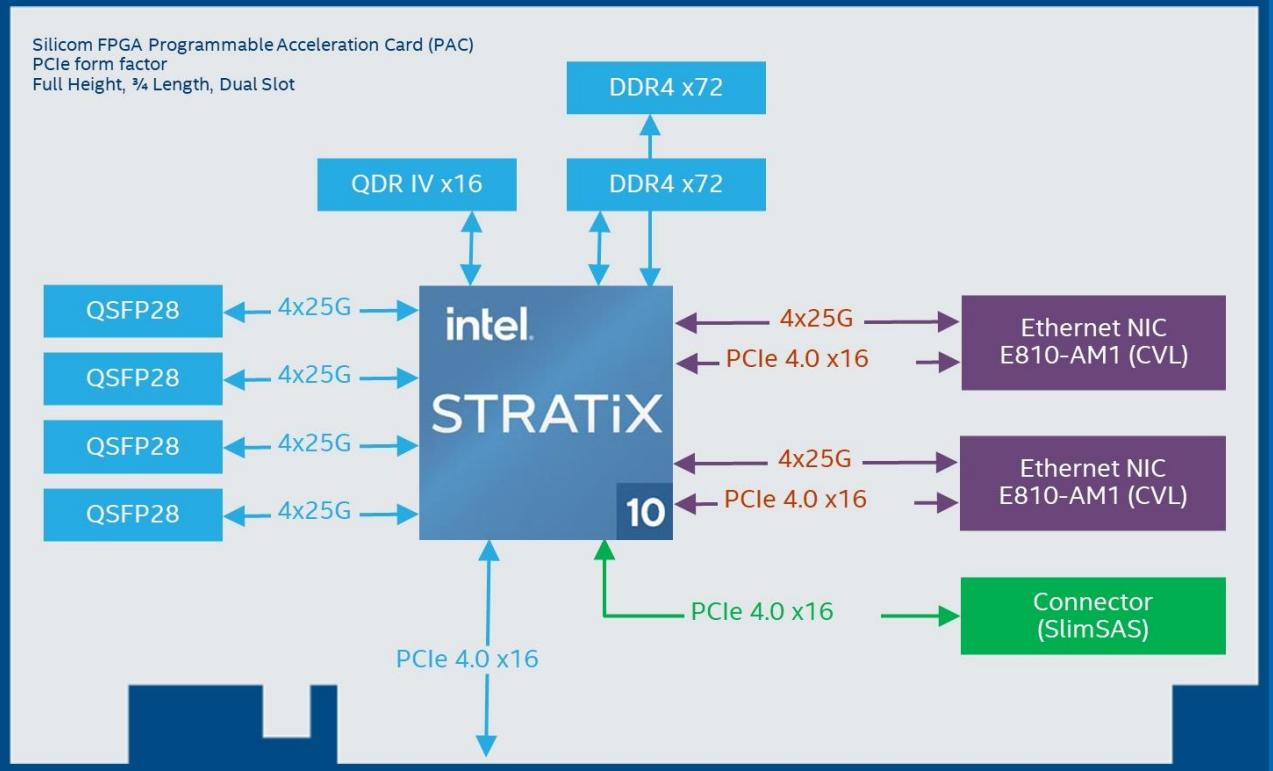
インテル® イーサネット・コントローラー E810 のデュアルサポート

- 幅広い OS サポートとより簡単なシステム統合

インテルとサードパーティによる高速化ワークロード

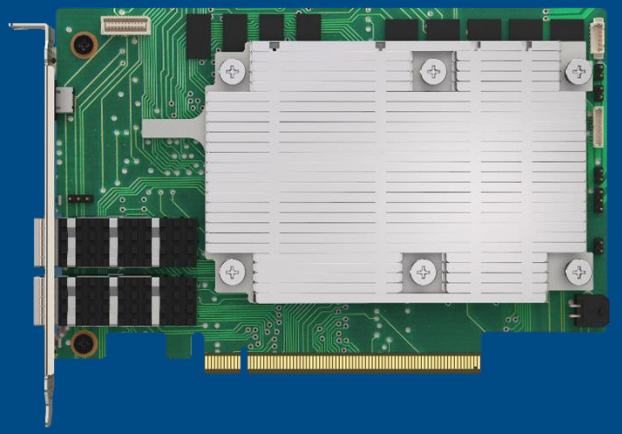
- vBNG: H-QoS、分類、ポリシング、スケジューリング
- vEPC 5G NGCN、4G/5G UPF、AGF
- IPSec、vFW
- VPP

Base
Hard NIC
Dual Host



Inventec FPGA SmartNIC C5020X

クラウド･サービス･プロバイダー向けSmartNIC



高性能ネットワーキングおよびストレージ向けアクセラレーション･カード

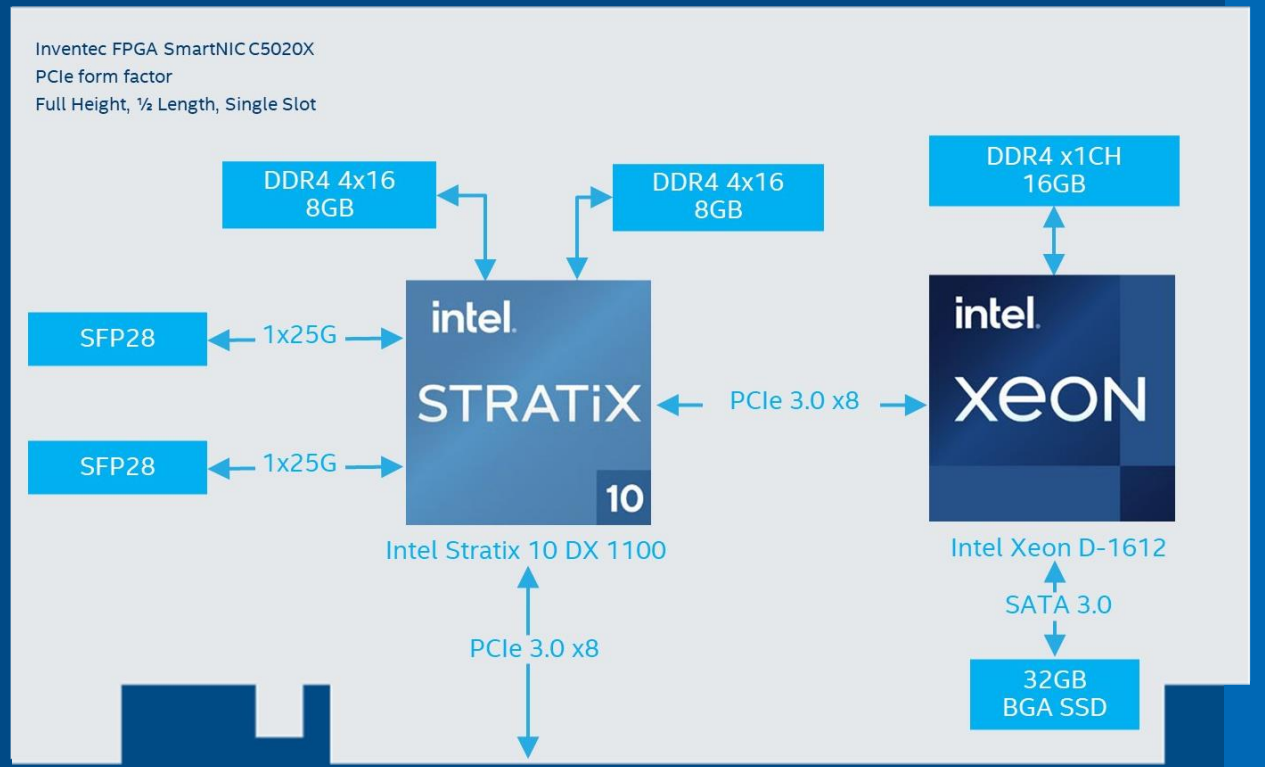
- インテル® DPDK、SPDKおよびOPAEを介してプログラム可能

インテル® Stratix 10 FPGA および Xeon® D SoC 搭載

- 高速イーサネットをサポート: 50G / 25G
- PCIe* Gen 4 x8
- 16GB DDR4 メモリー(FPGA用)
- 16GB DDR4 メモリー(インテル® Xeon® D用)
- 監視制御用 BMC (PLDM)
- ハーフレングス、フルハイト PCIe* カード

インテルとサードパーティによる高速化ワークロード

- DPDK による OVS
- SPDK による NVMEoF 等
- セキュリティー

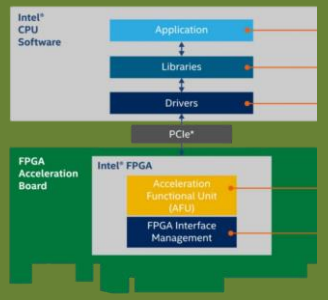


クラウド向け インテル® FPGA SmartNIC プラットフォームに基づく

FPGA SmartNIC/PAC ポートフォリオのまとめ

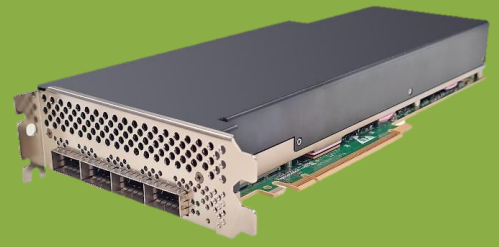
インテル® OFS
(Open FPGA Stack)

- ソフトウェアの拡張性と
  ハードウェアの拡張性を提供
- オープン・プログラマブル・
  アクセラレーション・エンジン
  (OPAE)
- カスタマイズ可能で
  ソースコードにアクセス可能な
  モジュール化されたシェル



Silicom FPGA
SmartNIC N5010

- ハードウェア・プログラマブル
  4x100G FPGA SmartNIC
  - 2x PCIe* Gen 4 x16
  - 32GB DDR4メモリー、144MB QDR4、8GB HBM
  - イーサネット・コントローラー E810



FPGA SmartNIC
C5000Xプラットフォーム

- インテル® Xeon® D プロセッサー
  SoC + FPGA プラットフォーム
- インテルとサードパーティによる
  高速化ワークロード
  - OVS、NVMEoF、セキュリティー

インテル® oneAPI プロダクト for FPGA

インテル® oneAPI プロダクト

インダストリー・イニシアチブ

oneAPI

開発の早期化

- ホスト側でのエミュレーション、レポートの各ステージでのコンパイル手法により開発期間を削減
- VTune™ プロファイラーにより強化されたデバッグおよびランタイム解析手法

拡張可能なコード

- アーキテクチャー、ベンダーを越えたコードの再利用

敷居の低さ

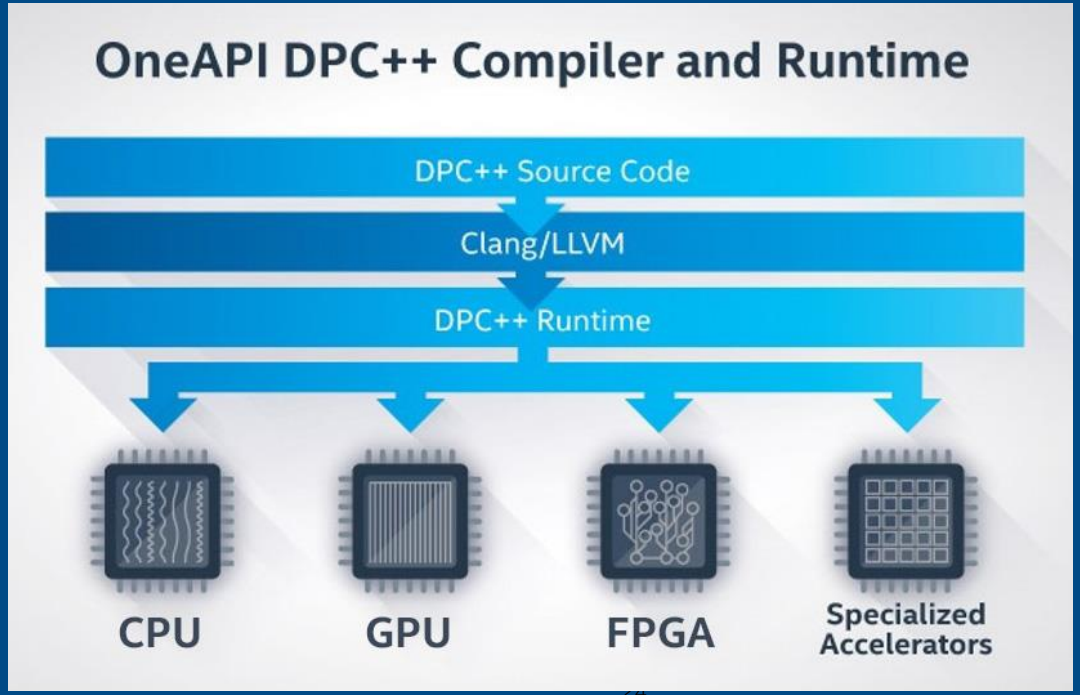
- 標準的でシーケンシャルなプログラミング言語モデルの活用により導入、トレーニングの時間を改善
- Eclipse, VS, VS Codeといった各種IDEをサポート

Available Now

インテル® oneAPI DPC++ コンパイラー

並列プログラミングの生産性とパフォーマンス

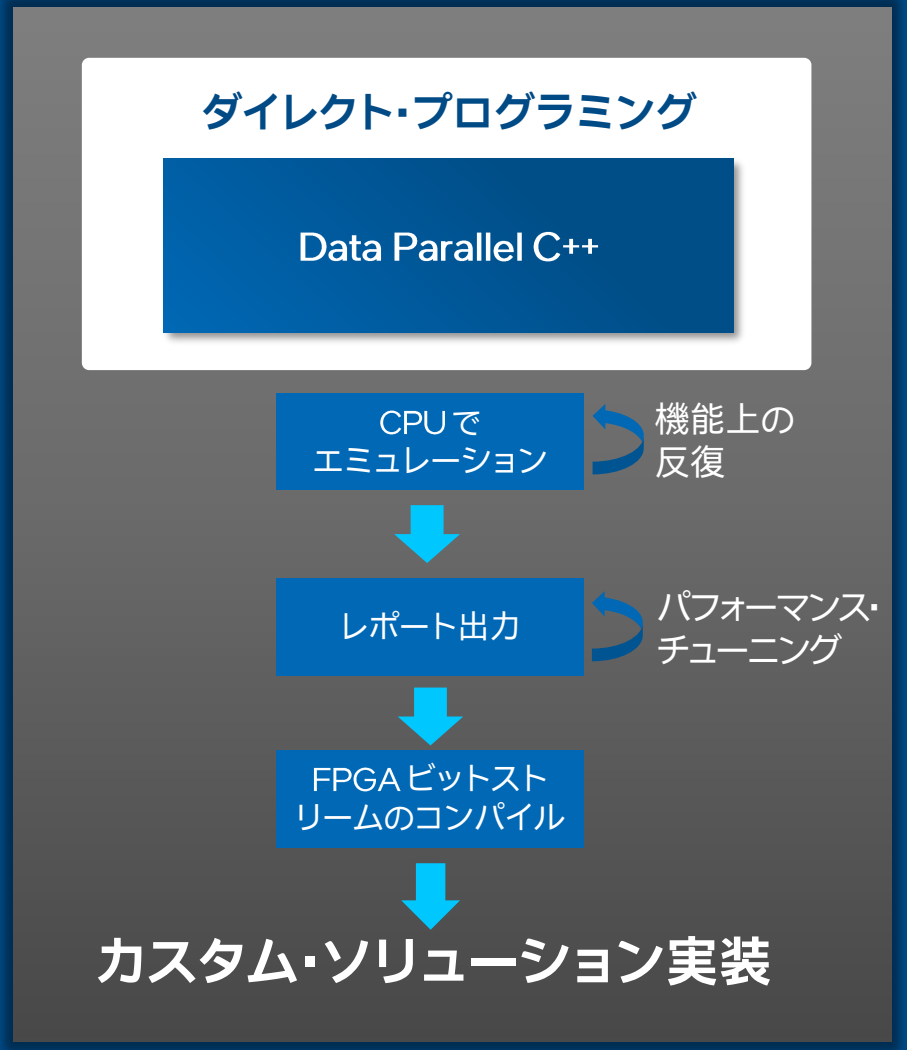
- CPUとアクセラレーターの両方で並列プログラミングの生産性を引き出すことが可能
  - 単一アーキテクチャーの独自言語に代わる、業界間で共通のオープン言語
- ベースは最新規格の C++ と SYCL*
  - 広く採用されている慣れ親しんだ C や C++ の構造により、C++ の生産性メリットを活用
- アーキテクチャーと高性能コンパイラーにおけるインテルの数十年にわたる経験を基に構築



FPGA 向け DPC++ コーディング

経験豊富な FPGA 開発者向け

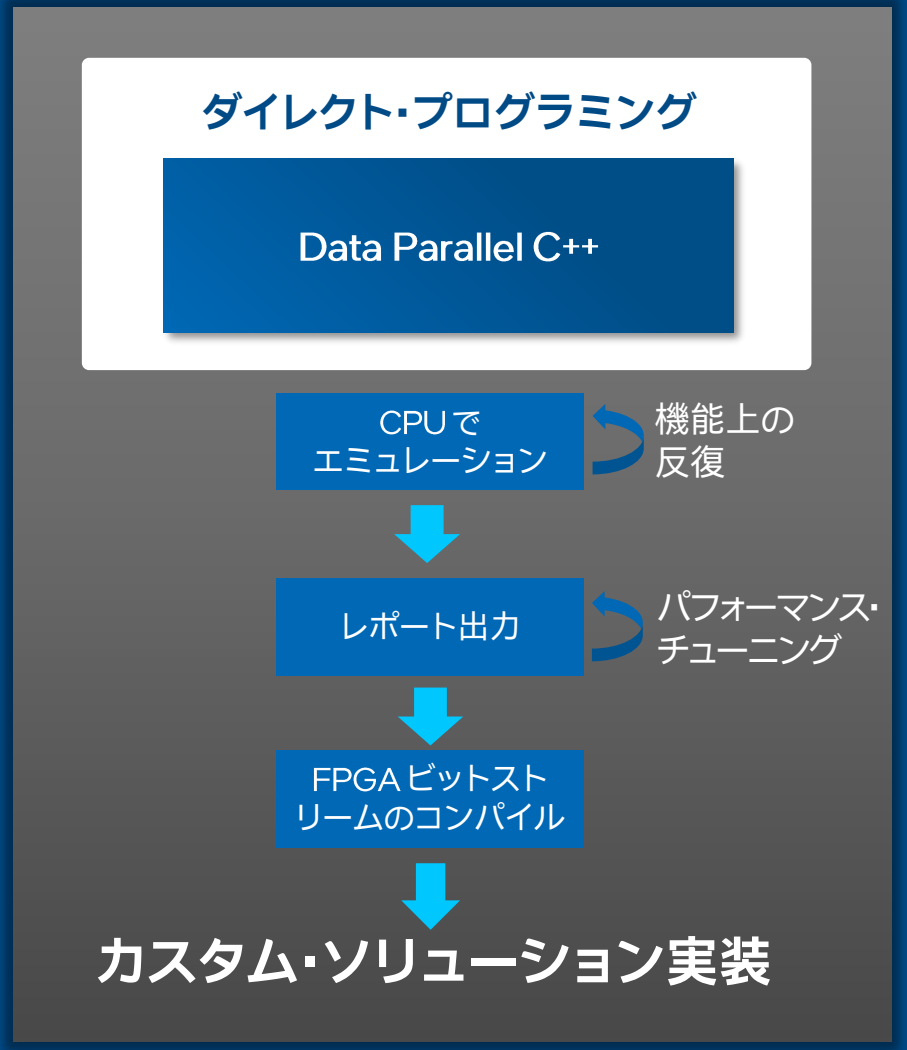
- 使いやすさ
  - 経験豊富な FPGA ユーザーは、Data Parallel C++ (DPC++) を使用してプログラミング・モデルの効率性を有効活用
- リアルタイム処理
  - 確定的な低レイテンシーと高スループットにより、データ処理を高速化
- ランタイム分析のサポート
  - ランタイム時にプロファイリング・データを収集して、インテル® VTune™ プロファイラーが CPU と FPGA の相互作用を分析
- デバイス固有の最適化
  - 経験豊富な FPGA 開発者を対象として、DPC++ コードを FPGA 向けに最適化するトレーニング・クラスを提供



FPGA 向け DPC++ コーディング

経験豊富な FPGA 開発者向け

- 使いやすさ
  - 経験豊富な FPGA ユーザーは、Data Parallel C++ (DPC++) を使用してプログラミング・モデルの効率性を有効活用
- リアルタイム処理
  - 確定的な低レイテンシーと高スループットにより、データ処理を高速化
- ランタイム分析のサポート
  - ランタイム時にプロファイリング・データを収集して、インテル® VTune™ プロファイラーが CPU と FPGA の相互作用を分析
- デバイス固有の最適化
  - 経験豊富な FPGA 開発者を対象として、DPC++ コードを FPGA 向けに最適化するトレーニング・クラスを提供



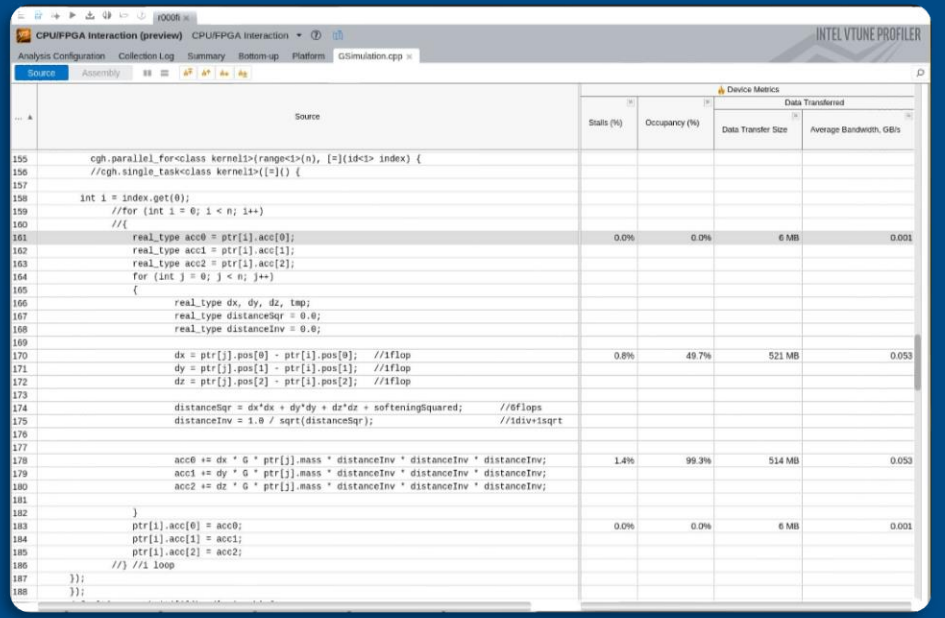
インテル® Vtune プロファイラー
ランタイム解析

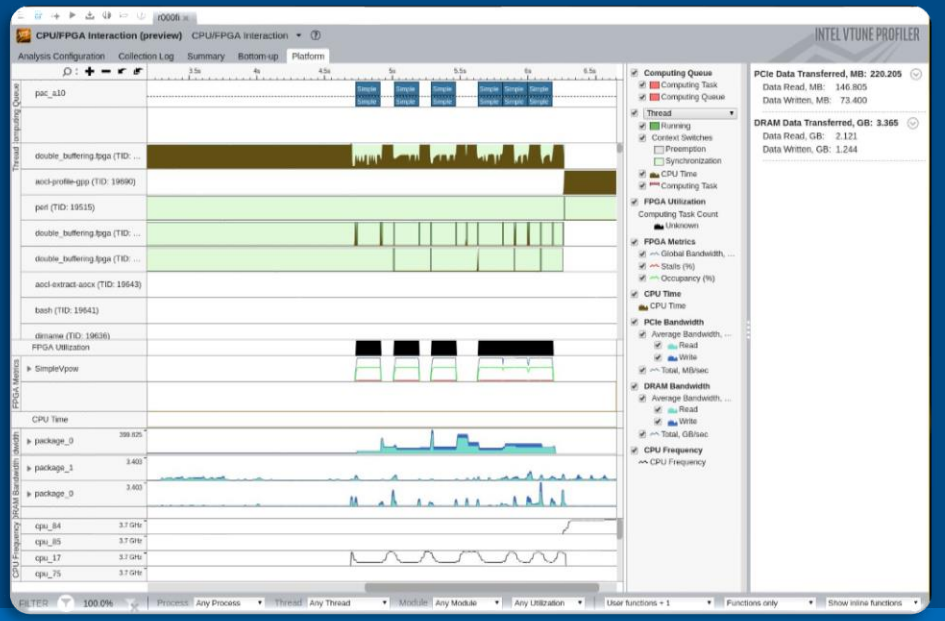
Data Parallel C++ (DPC++) コードの解析
- DPC++ コード上で処理時間がかかっているラインの確認

インテル の CPU, GPU & FPGA 向けのチューニング
- サポートされたハードウェア・アクセラレーターの最適化
- メモリーとパイプアクセスに関する詳細な統計情報を表示
  (ソースビュー形式 および タイムライン形式)

オフロードの最適化
- CPU/FPGA インタラクション・ビューを利用して
  パフォーマンス情報を確認
- カーネル プログラムの実行プロセス全体をグラフィカルに表示
  (ホスト側とデバイス側、両方のイベントを表示)

多様なパフォーマンス・プロファイル
- CPU, GPU, FPGA, スレッド,メモリー,キャッシュ, ストレージなど





インテル® oneAPI プロダクト for FPGA のまとめ

インテル® oneAPI

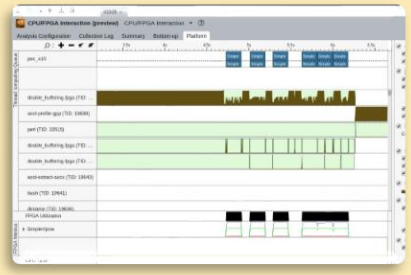
- 複数のアーキテクチャーを対象とし、言語、開発環境、デバッグ環境を統一
- さまざまなアーキテクチャー (xPU) をまたいでアクセラレーション実装に向けたパーティショニング・アルゴリズムの複雑さを解消

インテル® oneAPI
DPC++ コンパイラー

- 単一アーキテクチャーの独自言語に代わる、業界間で共通のオープン言語
- ベースは最新規格の C++ と SYCL*
- アーキテクチャーと高性能コンパイラーにおけるインテルの数十年にわたる経験を基に構築

インテル® Vtune プロファイラー

- ランタイム時にプロファイリング・データを収集して、インテル® VTune™ プロファイラーが CPU と FPGA の相互作用を分析
  - DPC++ コードの解析
  - オフロードの最適化
  - 多様なパフォーマンス・プロファイル



全体のまとめ

インテル® FPGA ハイエンド・デバイス

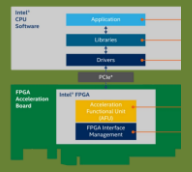
インテル® Agilex™ FPGA

インテル® Stratix® 10 FPGA

パートナーボード
- BittWare社カード

インテル® FPGA SmartNIC/PAC ポートフォリオ

インテル® OFS
(Open FPGA Stack)



Silicom FPGA
SmartNIC N5010

FPGA SmartNIC
C5000X-PF

インテル® oneAPI プロダクト for FPGA

インテル®
oneAPI

インテル® oneAPI
DPC++ コンパイラー

インテル® Vtune
プロファイラー

免責事項

インテルは、明示されているか否かにかかわらず、いかなる保証もいたしません。ここにいう保証には、商品適格性、特定目的への適合性、および非侵害性の黙示の保証、ならびに履行の過程、取引の過程、または取引での使用から生じるあらゆる保証を含みますが、これらに限定されるわけではありません。

性能の測定結果は、システム構成に記載された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。

結果は推定またはシミュレーションに基づいています。

実際のコストや結果は異なる場合があります。

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。

本資料に記載されているインテル製品に関する侵害行為または法的調査に関連して、本資料を使用または使用を促すことはできません。本資料を使用することにより、お客様は、インテルに対し、本資料で開示された内容を含む特許クレームで、その後に作成したものについて、非独占的かつロイヤルティー無料の実施権を許諾することに同意することになります。

本資料は、(明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず) いかなる知的財産権のライセンスも許諾するものではありません。

本資料に記載されているインテル製品には、エラッタと呼ばれる設計上の不具合が含まれている可能性があり、公表されている仕様とは異なる動作をする場合があります。 現在確認済みのエラッタについては、インテルまでお問い合わせください。

インテルは、サードパーティーのデータについて管理や監査を行っていません。 ほかの情報も参考にしてデータの正確さを評価してください。

Intel、インテル、Intel ロゴ、Agilex、Arria、Cyclone、Optane、MAX、Stratix、VTune、Xeon は、アメリカ合衆国および / またはその他の国における 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。

前提条件 (12 ページ)

1. TCO の削減は 1GB 当たりのメモリー価格で算出 (2020年9月28日時点)。インテル® Optane™ パーシステント・メモリー 100 シリーズ 128GB @487.66 米ドル (ShopBLT)。DDR4 ECC DIMM 128GB DDR4-2400、998.00 米ドル (memory.net)。
2. さまざまなメモリータイプについての相対パフォーマンス (揮発性と不揮発性の比較、ブロックアクセスとバイトアクセスの比較)。インテル® Optane™ パーシステント・メモリーは、NVMe* フラッシュ SSD に比べてパフォーマンスが高速。https://glesys.se/blogg/benchmarking-intel-optane-dc-persistent-memory (英語)


intel