ESSPER
Elastic and Scalable System for High-Performance Re-configurable Computing

RIKEN R-CCS PROCT

HPC・データセンター向け FPGAクラスタ拡張システム"ESSPER"の開発と今後の展望

佐野 健太郎

理化学研究所 計算科学研究センター (R-CCS)

May 25, 2022

理化学研究所 計算科学研究センター

チーム研究員公募中
R-CCS2105 or R-CCS2022 で検索

高性能計算 (HPC)分野の研究センター

- ✓ スーパーコンピュータ富岳の開発と運用
- ✓ スーパーコンピュータを用いた研究のための最先端インフラの整備と普及
- ✓ HPCに関する先端的研究

R-CCS

プロセッサ研究チーム

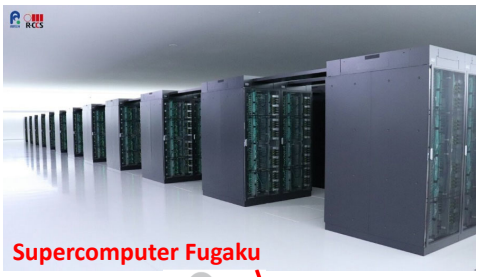
- ✓ 将来の高性能計算機アーキテクチャ
- ✓ 現在のHPCシステムの高度利用

PROCT

東北大学 理研R-CCS連携講座

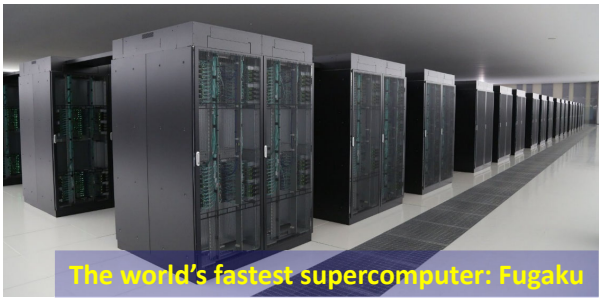
- ✓ 大学院情報科学研究科「先進的計算システム論講座」

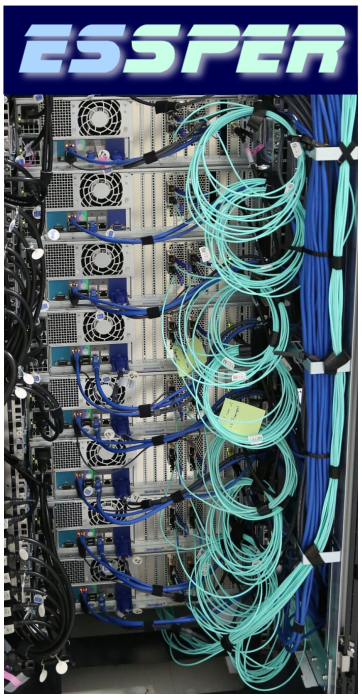
TOHOKU UNIVERSITY



本講演

- なぜ FPGA による高性能システム？
- HPC・データセンター向け
  FPGA システムはどうあるべきか？
- ESSPER：FPGAクラスタ概念実証システム
- ESSPERによる研究事例
- 今後の展望とまとめ





PoC FPGA Cluster System

なぜ FPGA による高性能システム？

[https://top500.org]

TOP500 World Ranking of Supercomputers

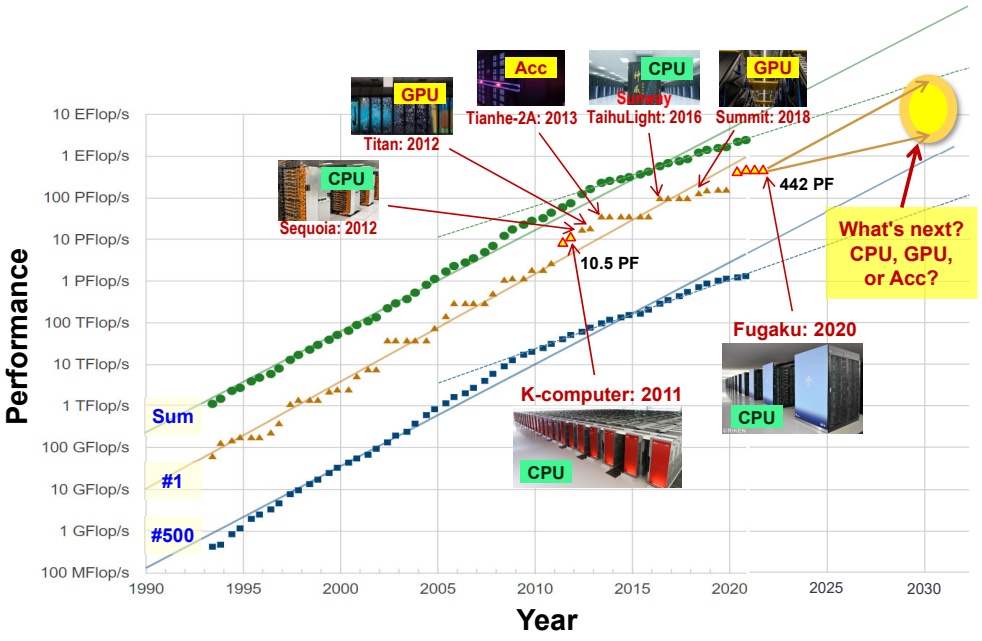
Ranking of HPL performance

- ✓ Linear algebra (LINPACK)
- ✓ Distributed-memory parallel computers

Supercomputer Fugaku

- ✓ #1 in TOP500
- ✓ #1 in HPCG, HPL-AI, Graph500
- ✓ #26 in Green500

問題は
性能だけではない



Constraint : System Power Consumption

Average power consumption

- ✓ in TOP10, TOP50, TOP500

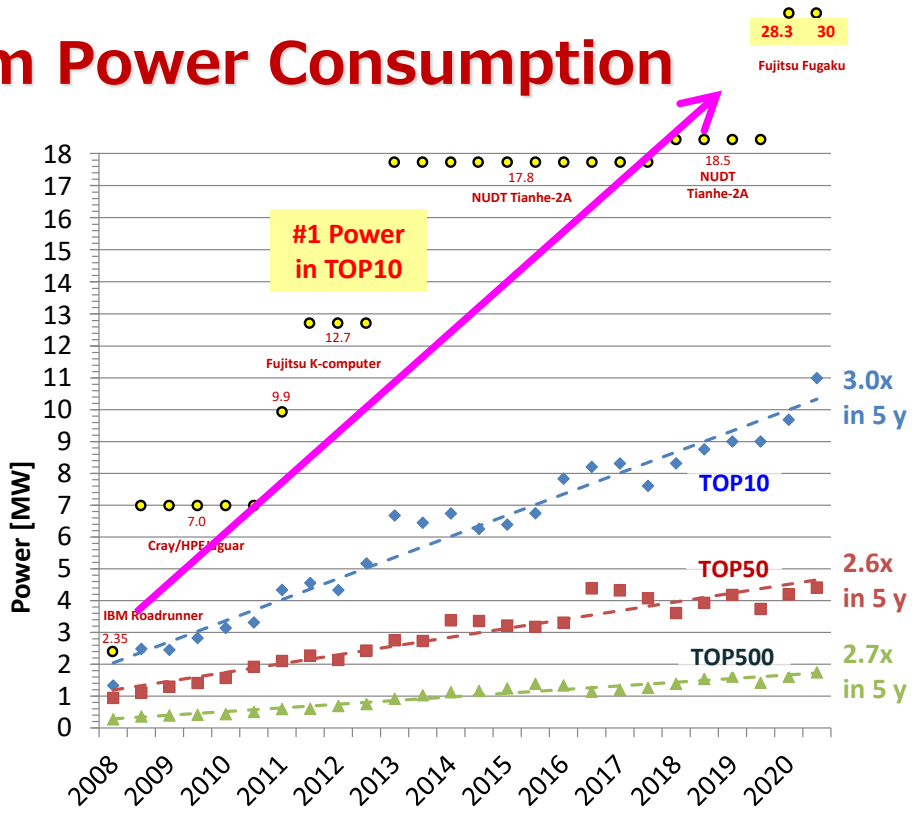
Needed to increase for higher system performance

- ✓ Limited improvement of performance per power

10s of MW for #1 systems

- ✓ **30MW** for 442PF (HPL) with 7,630,848 cores in Fugaku

増大するシステム電力が大問題
電気代、大丈夫?



課題：電力効率とスケーラビリティ

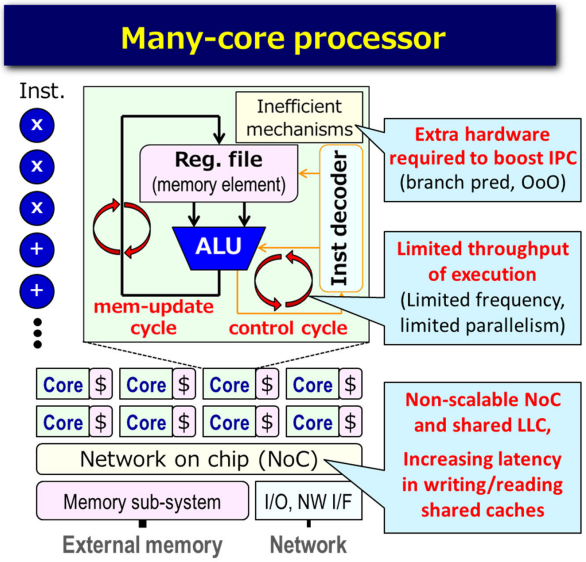
電力性能比

- ✓ ノイマン型プロセッサ（CPU, GPU）における
  演算以外の無駄な処理（スイッチングやデータ移動）
  が電力効率を悪化
  - ➢ 自動でIPCを向上させる機構（分岐予測機構、OoO）
  - ➢ 自動でメモリ遅延を隠蔽する機構（キャッシュメモリ）
  - ➢ **プログラマフレンドリな利点もあるが、もはや向上に限界**

スケーラビリティ

- ✓ 台数増に伴うオーバーヘッドが台数効果を制限
  - ➢ コア間やノード間の通信
  - ➢ 並列アルゴリズム・並列処理の大域的な同期
  - ➢ **既存方式では効率的スケールアップが困難、電力効率が低下**

解決策：データ駆動型の特化型ハードウェア



メニーコアプロセッサ

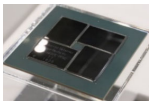
データ駆動型計算の実現方法

方法１　専用アーキテクチャのASIC開発

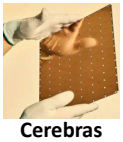
- ✓ **長所**　対象計算ドメインに最適化可（高性能・低電力）
- ✓ **短所**　適切なドメインは何？やり直しがきつい
  - 不適切なドメイン　→　汎用性の欠如 or 中途半端な性能・効率
  - 開発コスト(ハード、ソフト)・TTM

専用アーキは**開発コストを回収**できるアプリ向け。
その他のロングテイルをカバーするとCPUに近づく

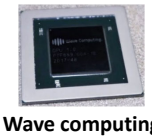

Graphcore


PFN MN-Core


Groq

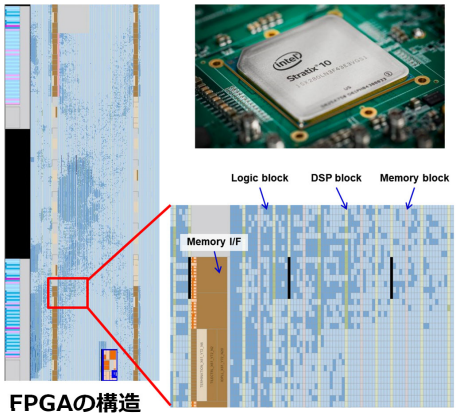

Cerebras

SambaNova

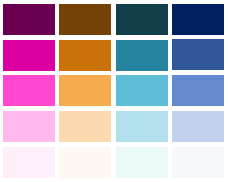

Wave computing

方法２　回路再構成可能デバイス FPGA の利用

- ✓ **長所**　汎用だが専用化したアーキを小NREで実装可
  - エコシステム（開発ツール、IPコア）がある
  - ASICより開発コストやTTMが小さい
  - 最先端の半導体テクノロジやXVCR
- ✓ **短所**　専用アーキの設計と実装はしんどい
  - **アーキテクチャの広範な知識が必須**
  - **高位合成の場合でも最適化にはハードの知識必要**
  - FPGA自身のオーバーヘッド（面積、電力、周波数）
  - FPGAベンダー依存となる（スペックや納期）



FPGAの構造



HPC・データセンター向けの FPGA システムはどうあるべきか？

FPGAを利用した高性能システム研究開発の課題

データ駆動型計算の実現方法

方法２　回路再構成可能デバイス FPGA の利用

- ✓ **長所**　汎用だが**専用化したアーキを小NREで実装可**
  - エコシステム（開発ツール、IPコア）がある
  - ASICより開発コストやTTMが小さい
  - 最先端の半導体テクノロジやXVCR
- ✓ **短所**　専用アーキの**設計と実装はしんどい**
  - **アーキテクチャの広範な知識が必須**
  - **高位合成の場合でも最適化にはハードの知識必要**
  - FPGA自身のオーバーヘッド（面積、電力、周波数）
  - FPGAベンダー依存となる（スペックや納期）

→ ✓ **要件1　様々な専用アーキを実装可**
- ＋複数FPGAを使用して計算性能やサイズをスケールできるシステムを提供
- ＋富岳などの既存システムをFPGAで容易に拡張する方法を提供

→ ✓ **要件2　対象問題に対しコストを抑えた専用アーキ開発やプログラミングが可**
- ＋様々な実装方法をサポート
- ＋システムのHW, SWスタックを提供
- ＋検証済の基盤HWシステムを提供（PCIe I/F, メモリ I/F, FPGA間Network）

誰を対象に、どのような研究開発を行うためのシステムにするか？（コンセプト）

- ✓ **ベンダー提供の技術に付加価値を載せる研究開発をすべき：**　アプリ、アーキ、開発ツール、システムスタック
- ✓ OpenCLやoneAPIなどのプログラミングのみでなく、HDL～HLSにより自由かつ楽にアーキを実装できること

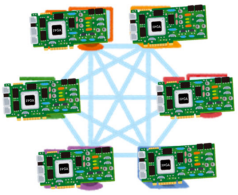
概念実証システムの方向性

✓ **要件1　様々な専用アーキを実装可**
＋複数FPGAを使用して計算性能や
サイズをスケールできるシステムを提供
＋富岳などの既存システムをFPGAで
容易に拡張する方法を提供

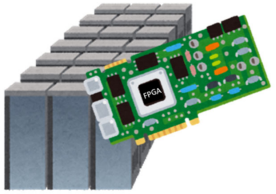
✓ **要件2　対象問題に対しコストを抑えた**
専用アーキ開発やプログラミングが可
＋様々な実装方法をサポート
＋システムのHW, SWスタックを提供
＋検証済の基盤HWシステムを提供
（PCIe I/F, メモリ I/F, FPGA間Network）

目標：以下を提供する概念実証システム
- ✓ **十分な専用化**が可能な実装手段
- ✓ 実装を楽にする**共通基盤のハード・ソフト**

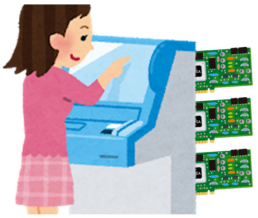
FPGA直結網をサポートする
カスタム可能なFPGA上SoC



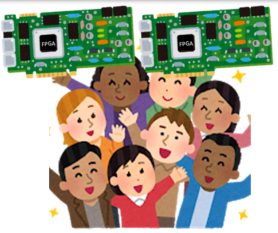
ネットワーク越しにFPGAを
利用可とする遠隔化ドライバ



回路記述言語や高位合成など
様々な実装方法に対応



相互利用を促進する
HW, SWシステムスタック



ESSPER
Elastic and Scalable System for High-Performance Re-configurable Computing

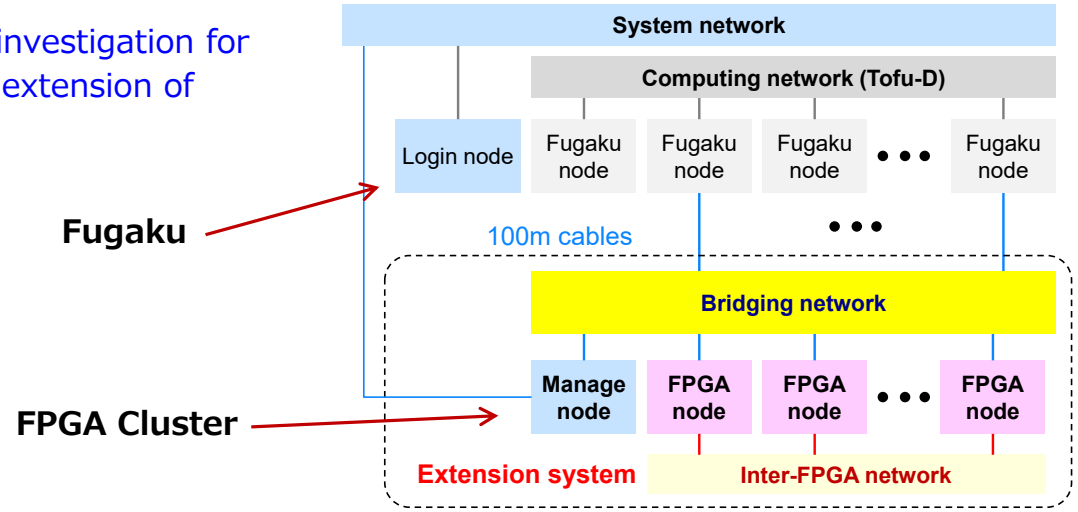
ESSPER：FPGAクラスタ概念実証システム

ESSPER
Elastic and Scalable System for High-Performance Re-configurable Computing

Architecture of ESSPER

Goal

- ✓ Technical investigation for functional extension of Fugaku.

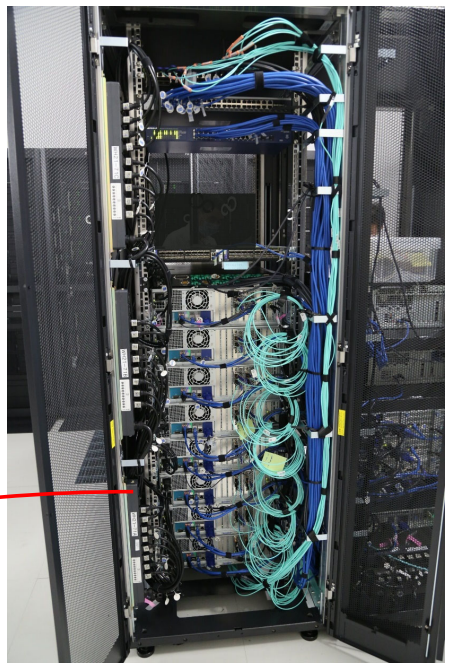


ESSPER
Elastic and Scalable System for High-Performance Re-configurable Computing

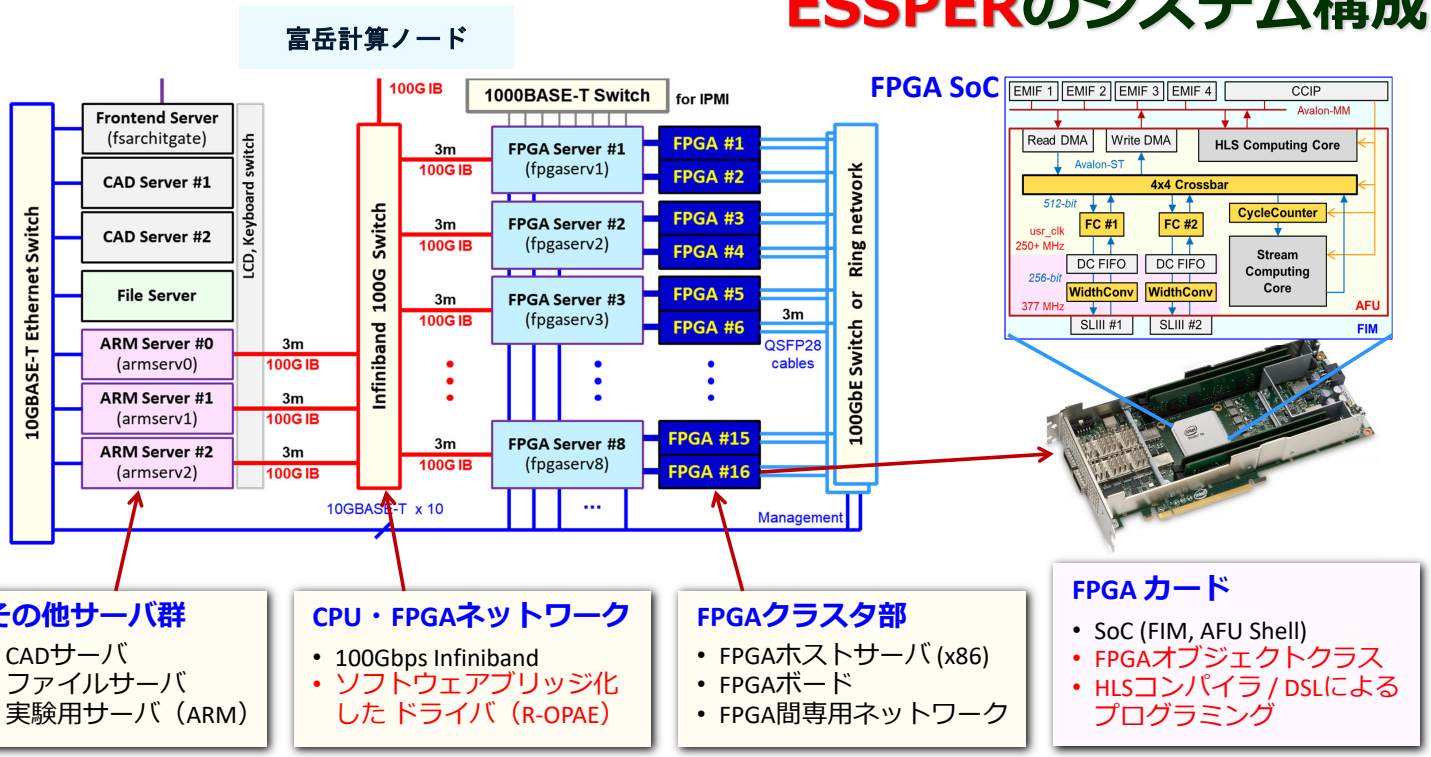
付加価値検討のための富岳との接続実験

FPGAクラスタ
実験試作機
ESSPER (別室)

「拡張により富岳に付加価値を与える方式の検討課題」
(FY2017～2020)



ESSPERのシステム構成



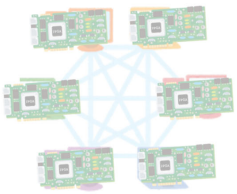
その他サーバ群
- CADサーバ
- ファイルサーバ
- 実験用サーバ（ARM）

CPU・FPGAネットワーク
- 100Gbps Infiniband
- ソフトウェアブリッジ化したドライバ（R-OPAE）

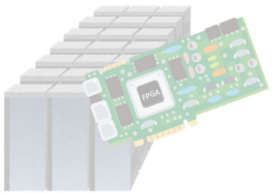
FPGAクラスタ部
- FPGAホストサーバ(x86)
- FPGAボード
- FPGA間専用ネットワーク

FPGA カード
- SoC (FIM, AFU Shell)
- FPGAオブジェクトクラス
- HLSコンパイラ / DSLによるプログラミング

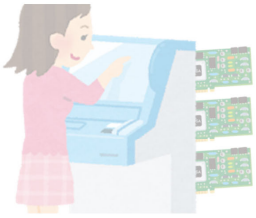
FPGA直結網をサポートするカスタム可能なFPGA上SoC



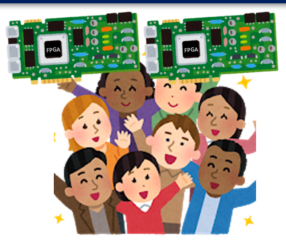
ネットワーク越しにFPGAを利用可とする遠隔化ドライバ



回路記述言語や高位合成など様々な実装方法に対応



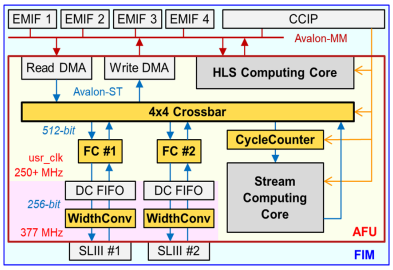
相互利用を促進するHW, SWシステムスタック

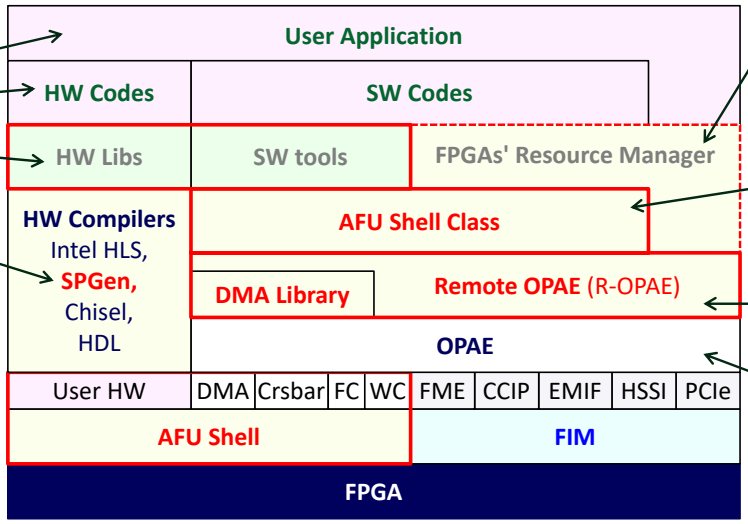


ESSPERのソフトウェアスタック

共同研究を期待

- Applications
- Libraries for HW and SW
- Tools / system software
- Parallelization techniques with multi FPGAs





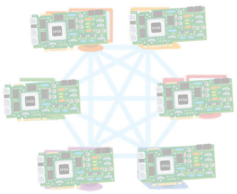
FPGA資源管理サービス
- Search and allocate resources of multiple FPGAs
- FPGA network management / control

AFU Shell class (抽象化)
- Object of AFU shell
- Abstraction of HW

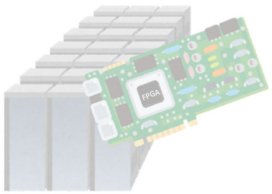
R-OPAE (遠隔化ドライバ)
- Software bridge using Infiniband Verbs

OPAE
- Low-level driver

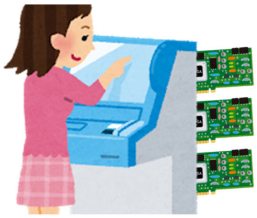
FPGA直結網をサポートする カスタム可能なFPGA上SoC



ネットワーク越しにFPGAを 利用可とする遠隔化ドライバ



回路記述言語や高位合成など 様々な実装方法に対応

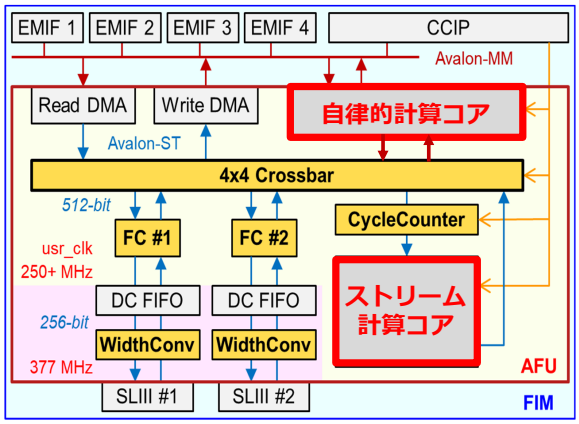


相互利用を促進する HW, SWシステムスタック

ハードウェアのプログラミング

計算コアを実装し AFU Shellに組込む

- ✓ **自律的計算コア** DDR4メモリバスに接続され自らデータを読み書きし動作
- ✓ **ストリーム計算コア** クロスバに接続され供給されたデータストリームに対し動作



コアのプログラミング方法

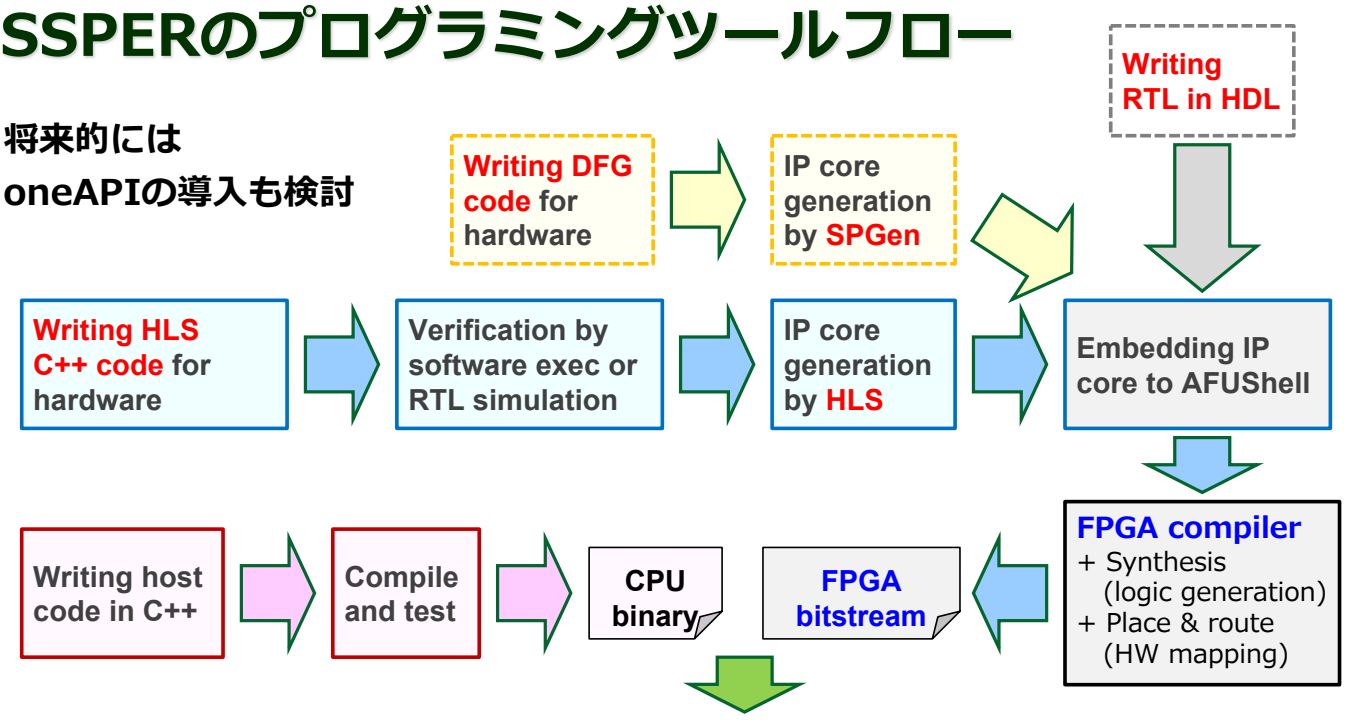
- ✓ **ソフトウェア指向** **HLS** **アルゴリズムをC/C++で記述**（Intel HLS）
- ✓ **ハードウェア指向** **HDL** **ハードウェア構造を記述**（Verilog-HDL, VHDL, Chisel, etc.）
- ✓ **その他** **DSL** **ドメイン特化型言語等**（ストリーム計算コアコンパイラ：SPGen）

ローレベルだがOpenCLより自由度が高く、ネットワークも用意済

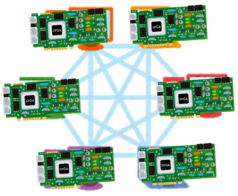
HLS: High-level synthesis, **Chisel**: Scala-based language for RTL, **SPGen** : Stream processor generator

ESSPERのプログラミングツールフロー

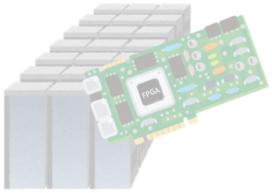
- **将来的には oneAPIの導入も検討**



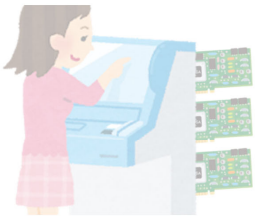
FPGA直結網をサポートする
カスタム可能なFPGA上SoC



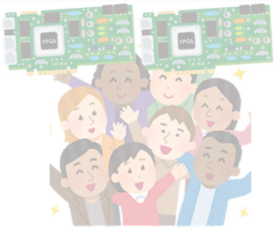
ネットワーク越しにFPGAを
利用可とする遠隔化ドライバ



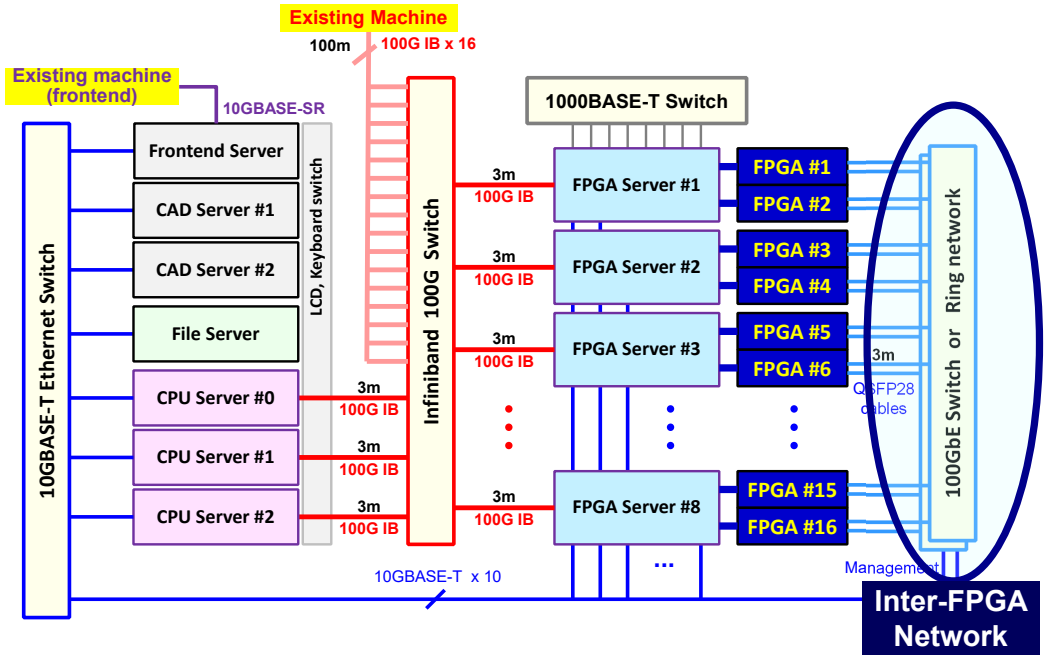
回路記述言語や高位合成など
様々な実装方法に対応

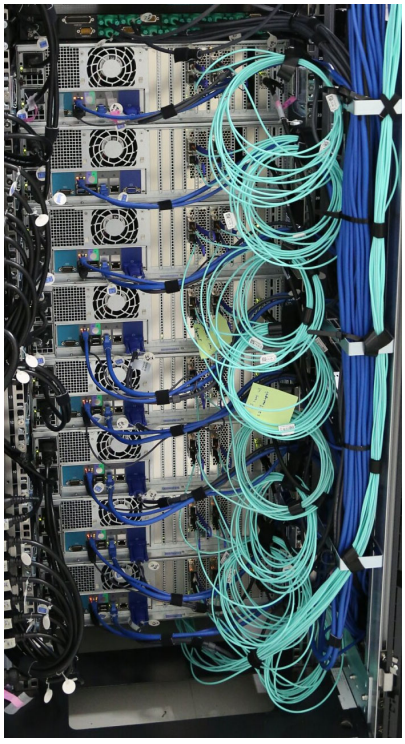


相互利用を促進する
HW, SWシステムスタック

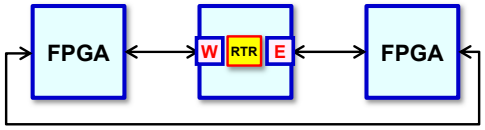


FPGA間の専用ネットワーク

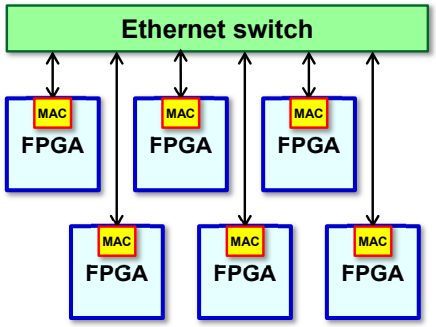




2種類のFPGA間ネットワーク



実際には 2 MAC / FPGA



直接網（1D トーラス）

利点) オーバーヘッド小
（低遅延、固定遅延）, 容易な利用

欠点) 資源割り当てに制約,
自前の実装が必要
（最新技術への対応コスト大）

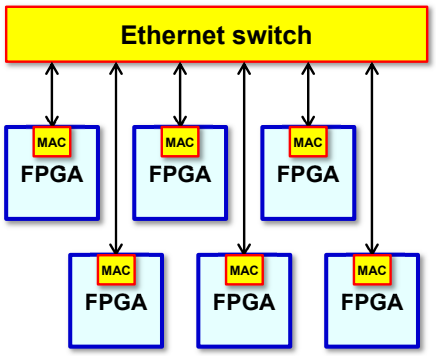
間接網（100G Ethernet）

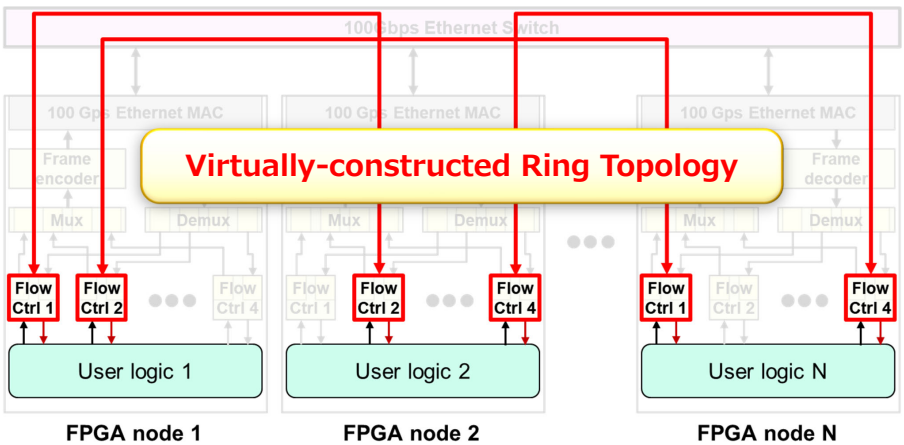
利点) 資源割り当ての柔軟性,
最新技術への対応が容易

欠点) Ethernet frameのオーバーヘッド
（高遅延、遅延の変動）,
フロー制御の問題,
高性能スイッチは高価

Ethernet上に仮想的な回線交換網を実現

- **Inter-FPGA connection should also be flexible.**
  - ✓ ***VCSN: Virtualized circuit switching network***





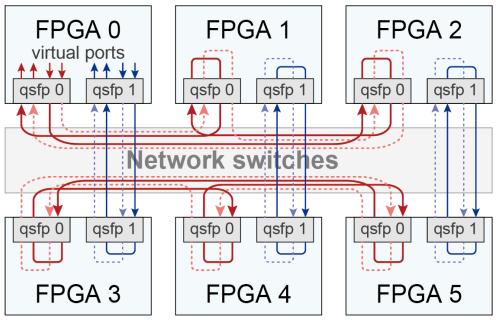
100G Ethernet switches

Pros) Flexibility, cutting-edge technology

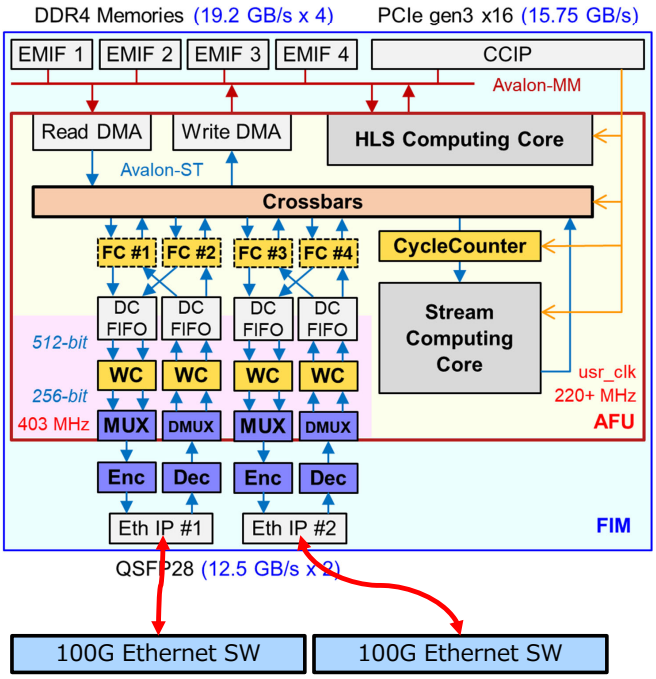
Cons) Overhead of ethernet frames, higher and variable latency, difficulty in flow-control and use

間接網に対する仮想回線交換方式（VCSN）の実装

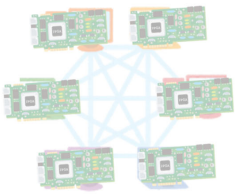
- **Indirect network : 100G Ethernet**
  - ✓ Implementation completed, under verification (2, 4, and 8 virtual ports per Eth MAC)
  - ✓ Higher throughput than Direct network
  - ✓ Developing system software to manage VCSN



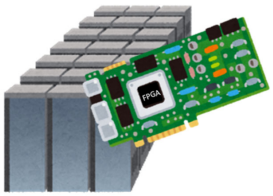
b. 2D torus (bi-directional)



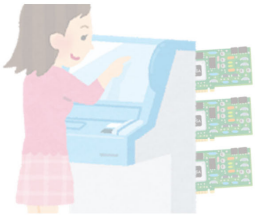
FPGA直結網をサポートする
カスタム可能なFPGA上SoC



ネットワーク越しにFPGAを
利用可とする遠隔化ドライバ

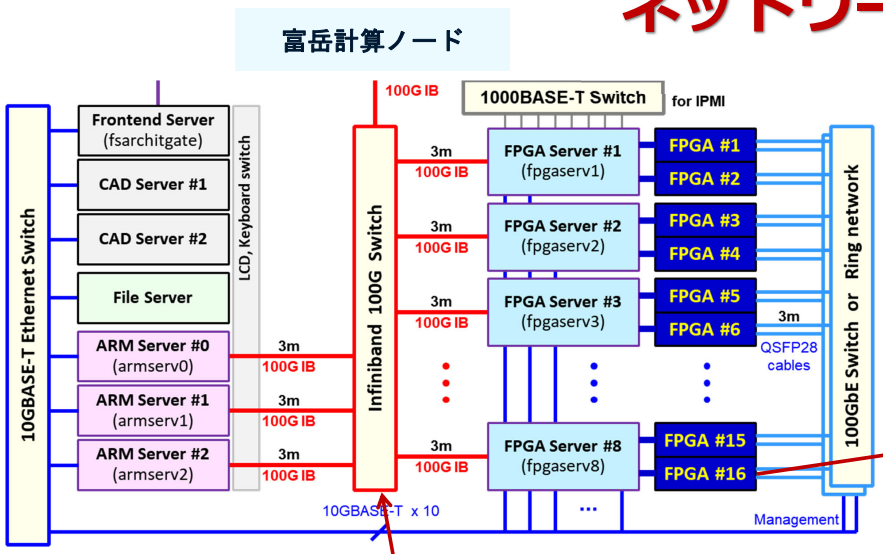


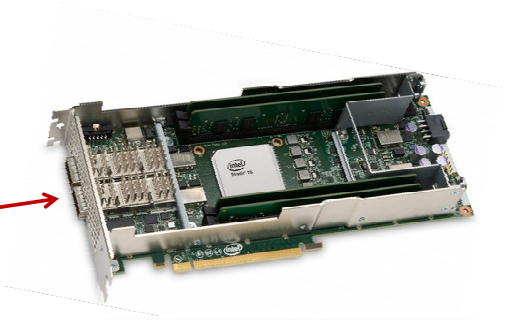
回路記述言語や高位合成など
様々な実装方法に対応



相互利用を促進する
HW, SWシステムスタック

ネットワーク越しのFPGA利用





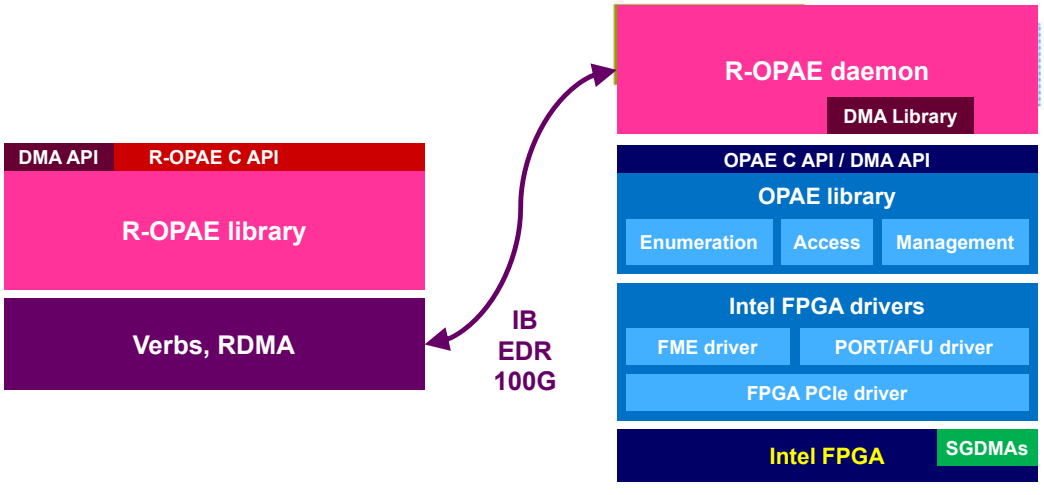
CPU・FPGAネットワーク
- 100Gbps Infiniband
- ソフトウェアブリッジ化したドライバ（R-OPAE）

FPGA カード
- SoC (FIM, AFU Shell)
- FPGAオブジェクトクラス
- HLSコンパイラ / DSLによるプログラミング

Remote-OPAE (ソフトウェアブリッジによる遠隔化)

インフィニバンドネットワーク越しに、遠隔のFPGAを利用可能

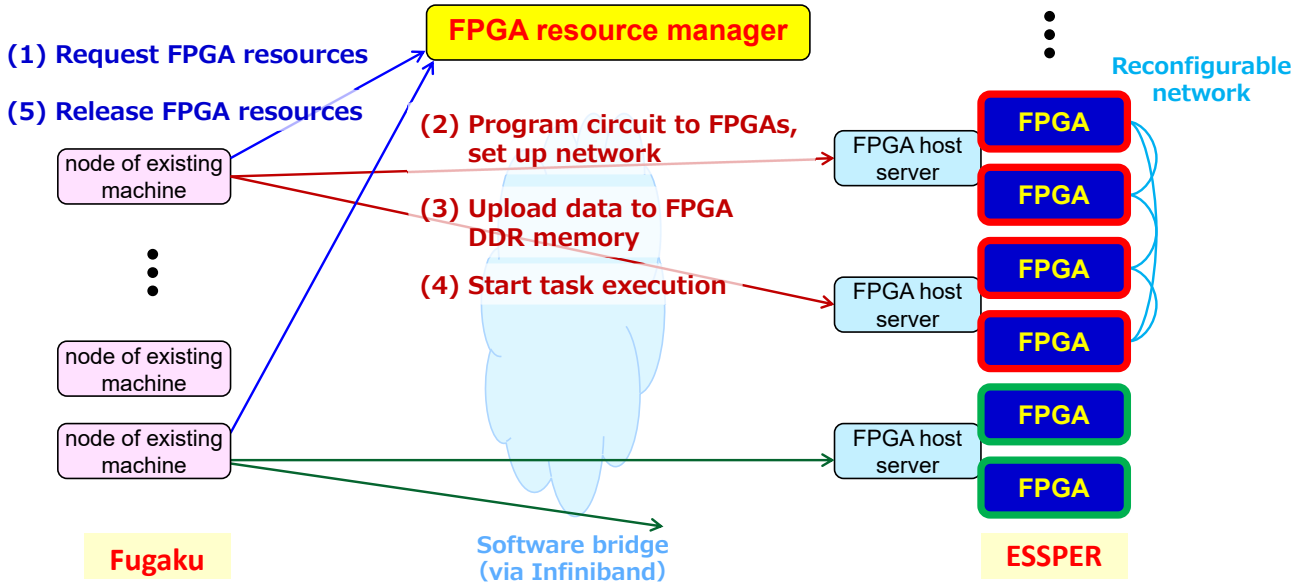
- ✓ Intel FPGA PAC のドライバ：OPAE（ローカル）
- ✓ あたかもローカルデバイスのように **遠隔*FPGA*を利用可**
- ✓ 空きFPGA資源を自由に利用可（= 運用の柔軟性）
- ✓ ベンダ非依存 異なるCPUアーキや環境からも利用可（= 可搬性・拡張性）



Host server **FPGA-gateway server** (x86)

FPGA資源管理サービス

✓ R-OPAEと併せて**資源分散(Resource disaggregation)**を実現



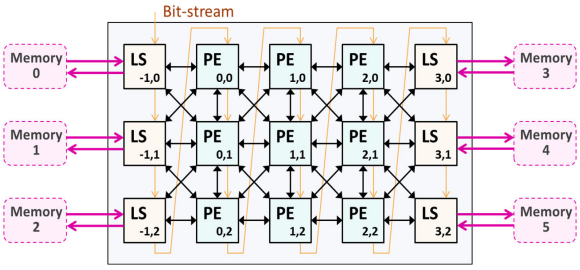
ESSPERによる研究事例

on-going project

粗粒度再構成アレイCGRAの設計空間探索（理研）

HPC向け粗粒度再構成可能アレイ(CGRA)のFPGAエミュレータ / オーバーレイを開発

- ✓ 理研 R-CCS プロセッサ研究チーム
- ✓ **ASIC実装向けのCGRA**について設計空間を探索
  - ➢ 複数FPGA実装でマルチチップ構成も評価
  - ➢ FIFO, Mux, ALUなどのライブラリモジュールを組み合わせて、様々な構成を実装可
- ✓ **CGRAコンパイラ**も併せて研究開発（東大）
  - ➢ OpenMPコードの一部をアレイに配置配線可
  - ➢ ベンチマーク評価（Stencil, Convolution, FFT）
- ✓ 初期実装完了（SystemVerilog）
  - ➢ RTLシミュレーション検証済
  - ➢ FPGAへ実装中



CGRAの全体構成（アレイサイズはパラメータにより可変）

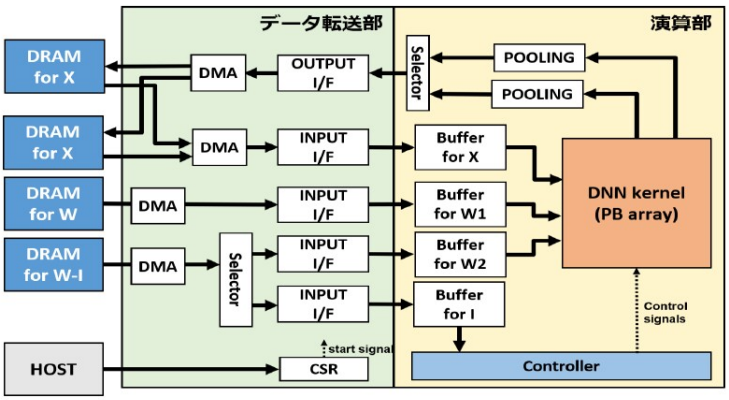
計算式のアレイへのマッピング例（16x16, 8x16）

on-going project

CNN推論専用アーキテクチャReNA（熊本大学）

エッジ向け推論専用アーキテクチャReNAをASIC からESSPERに移植

- ✓ 熊本大学　飯田教授らのグループ
- ✓ **複数FPGAによる高スケーラブル推論**が目標
  - ➢ 演算の並列化と専用ネットワーク接続を介したFPGA間通信
- ✓ **64x64のシストリックアレイ**
  - ➢ 積和演算器 x 64^2 = 8192 並列
  - ➢ 畳み込み計算と全結合計算をのPBアレイへのマッピングを工夫
  - ➢ 回路を変更せずに様々なモデルに対応可能
- ✓ 単一FPGA向け初期実装は完了
  - ➢ Verilog HDL



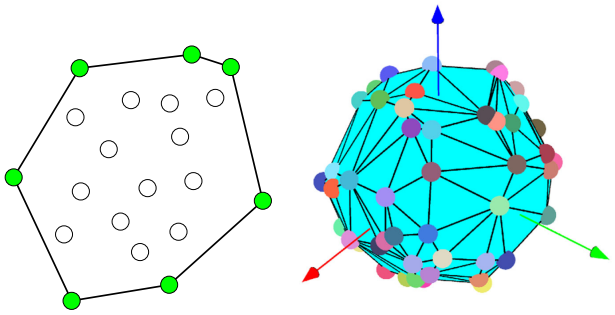
ReNAのアーキテクチャ

on-going project

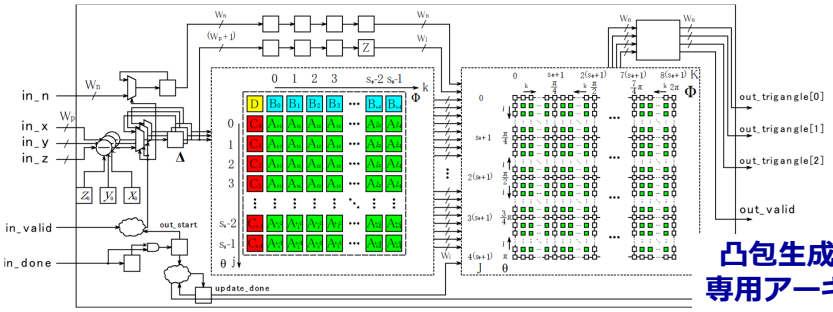
凸包生成専用アーキテクチャ（長崎大学）

点群データの凸包生成処理を複数FPGAで高速化

- ✓ 長崎大学　柴田教授らのグループ
- ✓ 凸包の応用分野
  - ➢ ドロネー図の構築/面積推定/ レジストレーション/画像処理等
  - ➢ 物体の衝突検出 / ゲームの当たり判定
  - ➢ 経路計画における移動体と障害物の近似
  - ➢ 物理シミュレータ
  - ➢ 点群のリアルタイムレンダリング
- ✓ **GPUよりも高スループット・低遅延**の パイプライン処理
- ✓ 初期実装完了（SystemVerilog）
  - ➢ ソフトウェア実装Qhullとの性能比較



凸包：点集合を囲む最小の凸集合



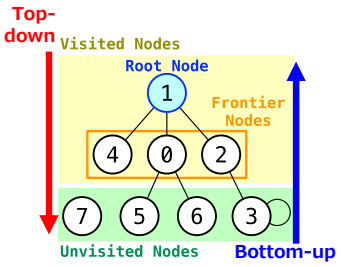
凸包生成専用アーキ

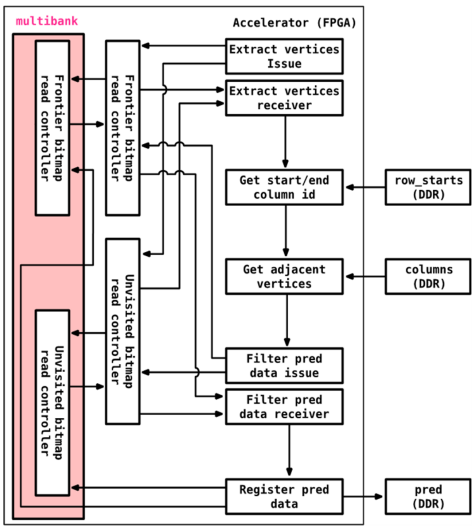
on-going project

高位合成によるグラフ幅優先探索専用ハード（広島市立大学）

幅優先探索アクセラレータHyGTAを開発（Graph500）

- ✓ 広島市立大学　谷川講師らのグループ
- ✓ **Hy**brid **G**raph **T**raversal **A**ccelerator
  - ➢ **Top-down** と **Bottom-up** を組み合わせた **Hybrid Graph Traversal**を高速化
- ✓ **高位合成**で実装し、**FPGA**上で動作と評価
  - ➢ 別途アーキテクチャシミュレータを開発
- ✓ **パイプライン処理、遅延隠蔽、高効率メモリシステム**を研究
  - ➢ 幅優先探索パイプライン
  - ➢ 隣接節点キャッシュ
  - ➢ マルチバンク ビットマップ （訪問済接点記録）
  - ➢ 幅優先探索特有のメモリ参照パタンを有効利用





幅優先探索専用 アーキテクチャ HyGTA

Graph500のランキング（2021年11月 BFS）

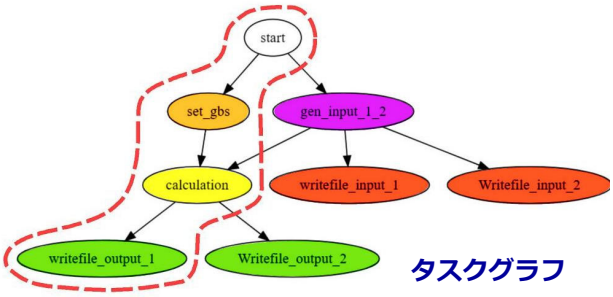
RANK	MACHINE	SCALE	GTEPS
32	ENIAD （FPGA）	26	783.75

on-going project

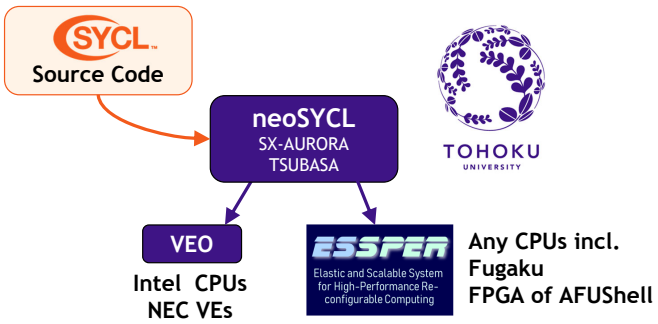
独自SYCLによるFPGAタスクオフロード（東北大学）

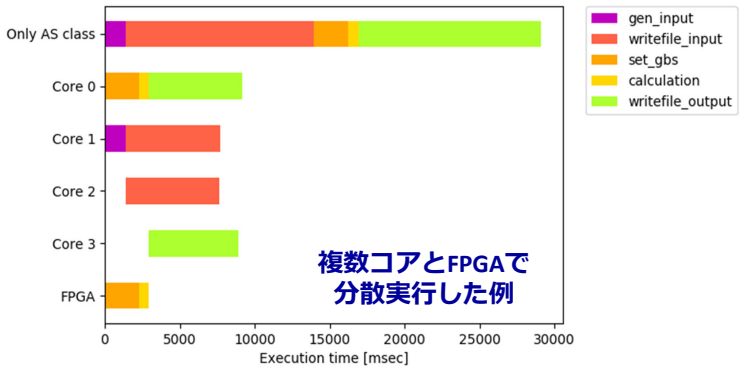
neoSYCL：SYCL実装の一つ

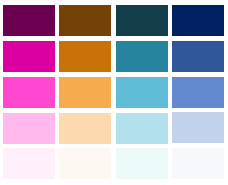
- ✓ 東北大学 滝沢教授らのグループ
- ✓ NEC Vector Processor向け
- ✓ ESSPERのFPGA & AFUShell向けにも
  タスクオフロードやスケジューラを研究開発
- ✓ 富岳からFPGAにオフロード可



タスクグラフ



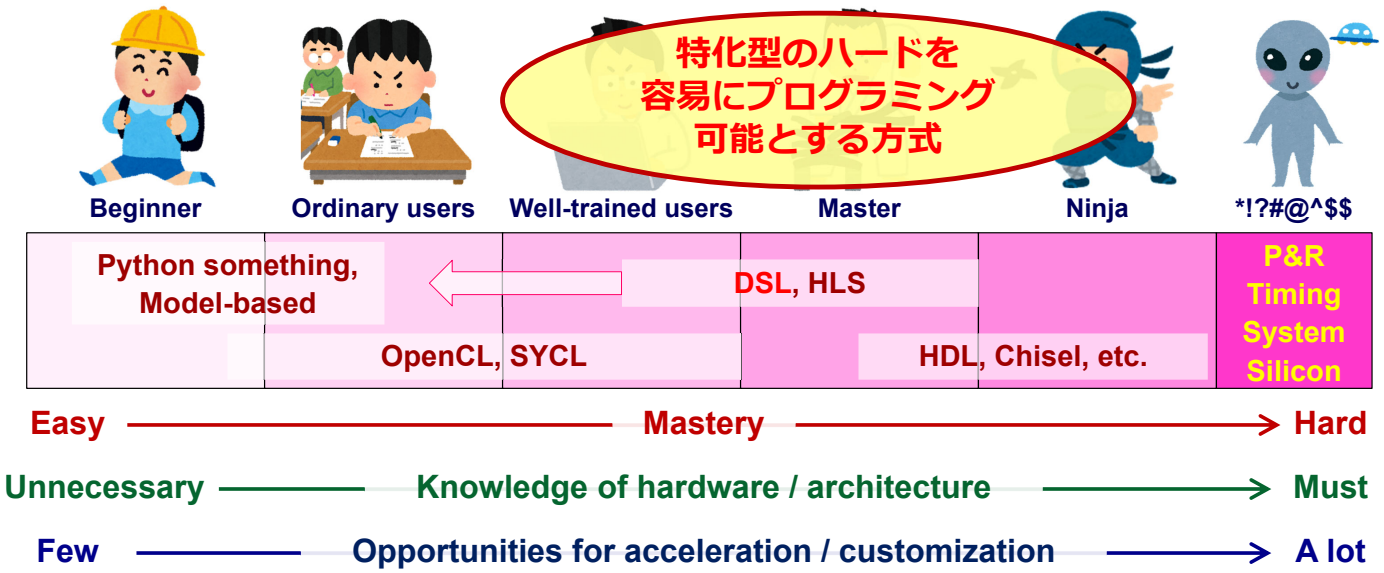




FPGAクラスタの
今後の展望

HPC向けFPGAシステムの研究領域（プログラミング視点）

プログラミングの難易度と自由度



おわりに

本研究　スケーラブルかつ柔軟な高性能FPGAクラスタの概念実証システム **ESSPER** を開発

研究インフラ ESSPER が完成
- ✓ ハードウェア記述言語 や 高位合成 によるハードウェア実装が可能
- ✓ 富岳と接続され、富岳からFPGAを利用可（R-OPAE）
- ✓ 様々な研究：　CGRA, CNN, Convex, Graph, neoSYCL, etc.

カスタム可能システムスタックが利用可
- ✓ ハードウェア
- ✓ ソフトウェア

今後の課題
- ✓ アプリ開発フローの自動化、ノウハウの蓄積（記述例）
- ✓ 複数FPGAのプログラミング方式、oneAPIの導入
- ✓ システムソフト（CPU側と協調したタスクスケジューリング）
- ✓ 高性能アプリ開発（量子誤り訂正、ニアセンサ、他）

次世代FPGAクラスタ開発を計画中

共同研究・産業連携を募集中

研究員も募集中！ R-CCS2105 or R-CCS2022