



FPGAクラスタ試作システム ESSPERとそのFPGA間通信機構

理化学研究所 計算科学研究センター 佐野 健太郎

2021/2/26



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-CCS2022 公募」
で検索！

高性能計算 (HPC)分野を牽引する研究センター

- ✓ 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

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の探求
- ✓ 現在のHPCシステムの高度利用の探求
- ✓ 謝辞
 - Dr. Tomohiro Ueno, Dr. Takaaki Miyajima
 - Dr. Jens Huthmann, Dr. Atsushi Koshiba

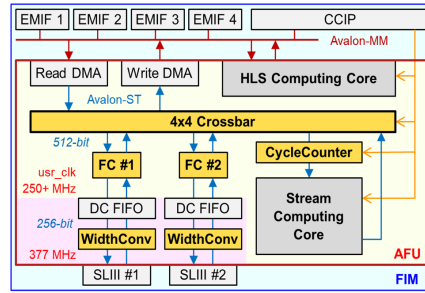


Supercomputer Fugaku

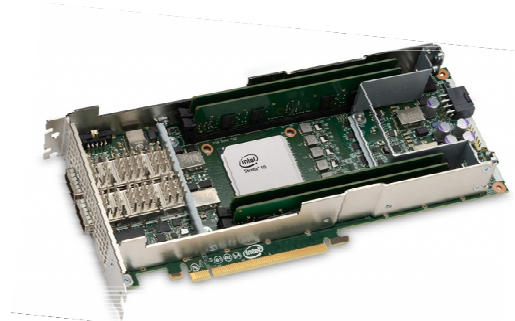


概要

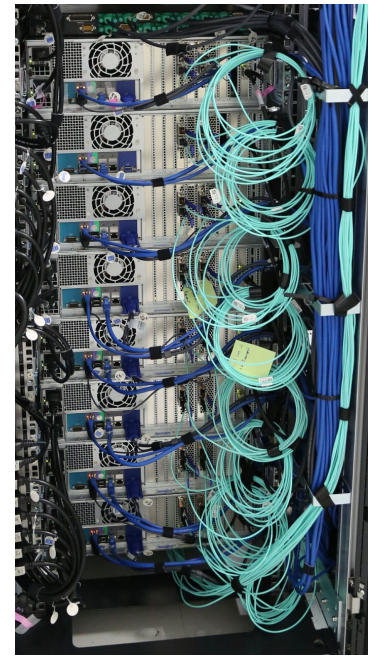
- 背景
- 高性能計算のための
FPGA クラスタの要件
- FPGAクラスタ実験システム
ESSPER
- 動作検証とデータ転送性能
- まとめ



AFU Shellの設計



Intel FPGA PAC D5005



FPGAクラスタ実験システム

背景

- **メニーコア等、既存の汎用プロセッサアーキテクチャの性能向上の鈍化**
 - ✓ 半導体スケーリングの減速、ダークシリコン、データ移動の非効率化
- **問題特化型のハードウェア機構による計算が有望**
 - ✓ カスタム / リコンフィギャラブルコンピューティング
 - ✓ **ハイエンドFPGAへの期待**（最先端半導体ノード、HBM2などの高速メモリ）
 - ✓ 大規模FPGAシステム利用の試み（産業、学術）
 - *Microsoft Catapult, AWS EC2 F1, Alibaba Cloud, Tencent Cloud, Huawei Cloud*
 - 筑波大学 *Cygnus*, Paderborn大学 *Noctua*
- **しかし、FPGAを搭載したクラスタシステムは主流には成り得ていない**
 - ✓ CPU-FPGA, FPGA間接続網のシステムアーキやプログラミングモデルに改善の余地
 - ✓ 既存のスパコンやデータセンターへFPGAを追加したいが、容易ではない（設置スペース、電力や冷却性能、インターフェース、資源管理の問題）

本研究の目的

既存FPGAクラスタシステムの課題

性能スケーラビリティ

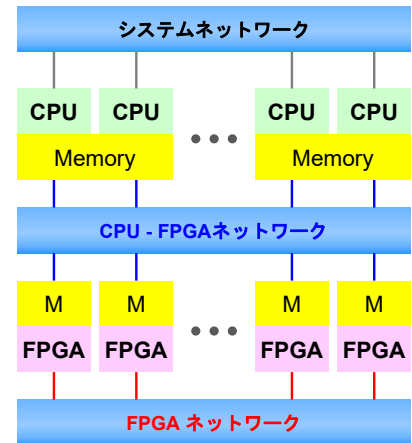
- ✓ FPGA間ネットワークが未成熟, 未整備であり規模に応じた性能向上が困難

柔軟性と運用可能性

- ✓ 大規模システムでの柔軟な資源利用が困難
- ✓ 様々な既存システムの拡張が困難

生産性と自由度の両立

- ✓ OpenCLでは所望のHW機構が実現困難。ある程度の抽象度で自由度の有る標準的研究基盤が必要



CPUクラスタとFPGAクラスタ
の疎結合アーキテクチャ

本研究

- ✓ これらの課題を解決する **スケーラブルかつ柔軟な高性能
FPGAクラスタの概念実証**のため、試作システム**ESSPER**を開発



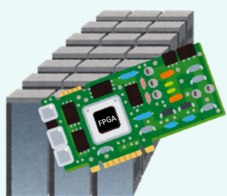
高性能計算のための FPGA クラスタの要件

FPGAクラスタの要件

可搬性、柔軟性、自由度の高い研究開発システム実現のために、以下の要件を設定：

要件1 既存の高性能計算機システムを容易に拡張可

- ✓ FPGA無しの既存システムを
FPGAにより容易に拡張
- ✓ 例えば
 - ・スーパーコンピュータ富岳
 - ・データセンター



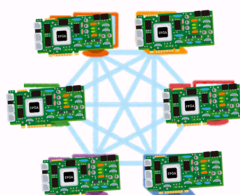
要件2 システム内のFPGA資源をどこからでも利用可

- ✓ システム内の無数のFPGAを
どのCPUからでも利用可とする
- ✓ 複数ユーザがシステムを共有し
部分的に利用する場合でも
FPGA資源の高い利用効率を実現



要件3 資源の部分的利用に適したFPGA間 接続網を持つ

- ✓ 要件2を考慮しながらも
FPGA間の広帯域/低遅延通信で
複数FPGAによる高性能計算
を可能とする



要件4 抽象化オブジェクトによりFPGA SoCを利用可

- ✓ 様々なハードウェア実装に対し
簡潔かつ統一したインタフェース
でCPU側コードを記述可とする
- ✓ 異なるFPGAボード、異なるShell
等のハードウェア抽象化

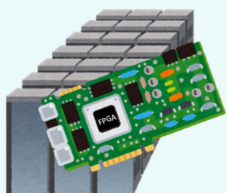


要件の実現方法

本研究の提案：

要件1 既存の高性能計算機システムを容易に拡張可

- ▶ **ホスト・FPGA間
汎用ネットワーク構成**
- ▶ **遠隔化ドライバ
(R-OPAE)**



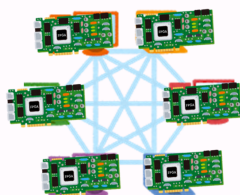
要件2 システム内のFPGA資源をどこからでも利用可

- ▶ **FPGA資源管理マネージャ**
- ▶ **遠隔化ドライバ
(R-OPAE)**



要件3 資源の部分的利用に適したFPGA間 接続網を持つ

- ▶ **Ethernetによる
FPGA間仮想回線交換NW**
- ▶ **FPGA Shell (SoC)開発
FIM, AFU Shell**



要件4 抽象化オブジェクトによりFPGA SoCを利用可

- ▶ **抽象化したFPGA API
(afushellクラス)**
- ▶ **Intel HLSやSPGenによる
開発フロー**





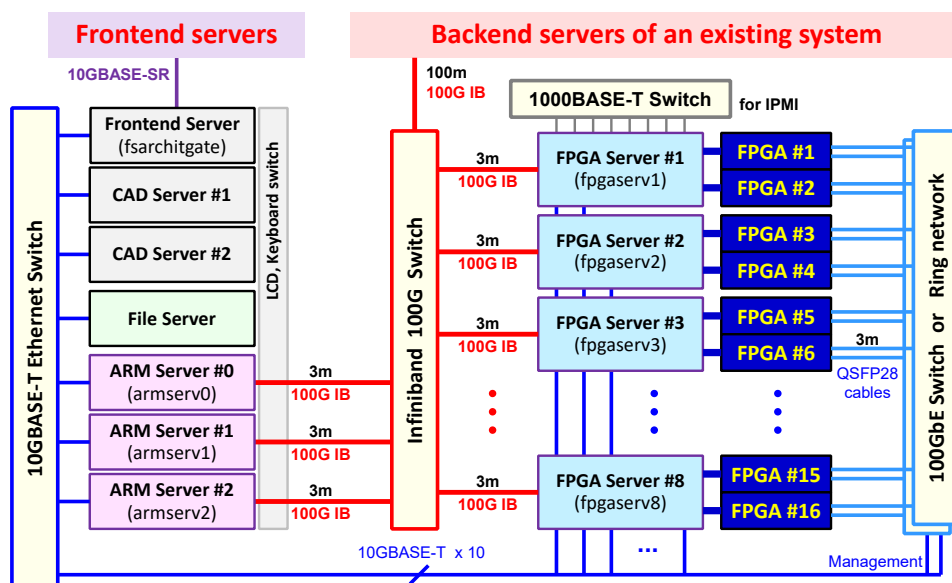
FPGAクラスタ 実験システム



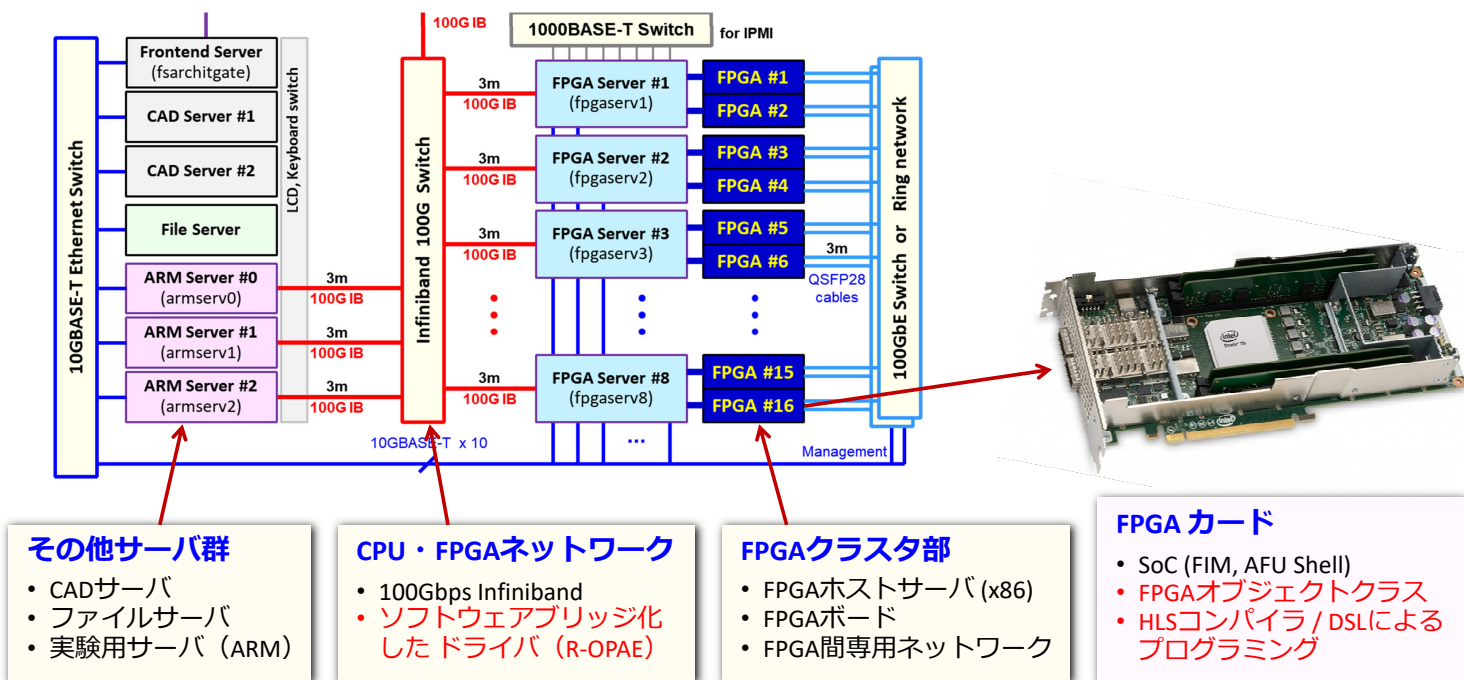
Elastic and Scalable System for High-Performance Reconfigurable Computing

既存システムをFPGAで 拡張する評価システム

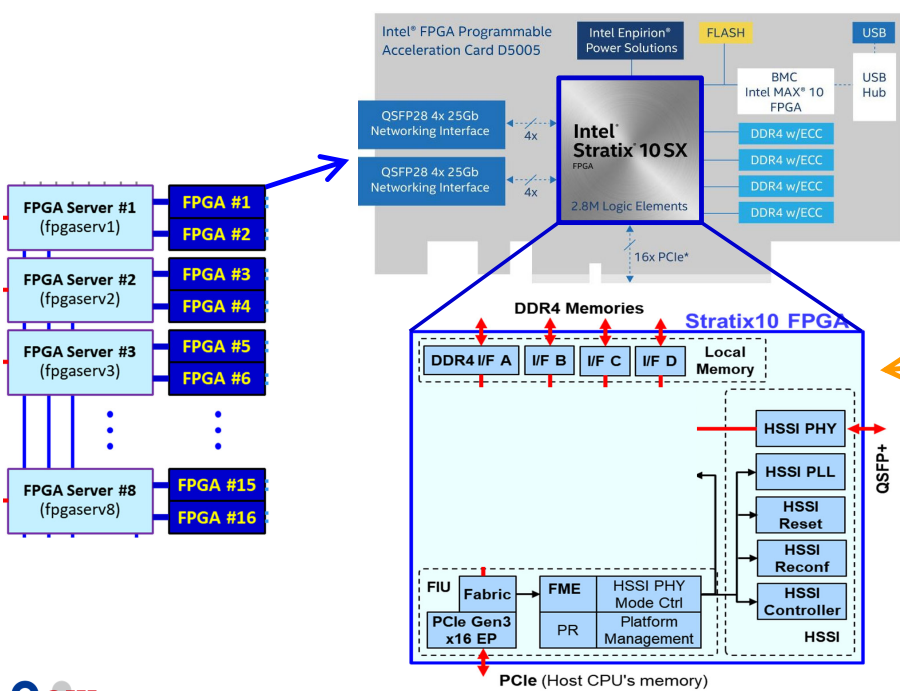
-  FPGAクラスタ部
-  CPU・FPGA
汎用広帯域ネットワーク
-  その他サーバ群



ESSPERのシステム構成



FPGAカードとFPGA SoC



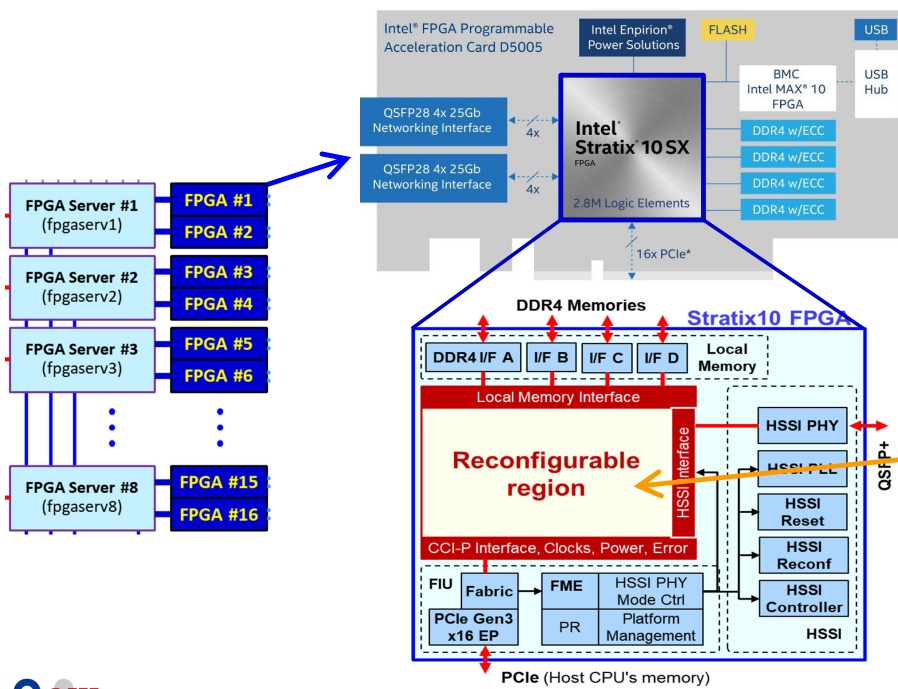
Intel FPGA PAC D5005

- ✓ *Stratix 10 FPGA (14nm)*
- ✓ *2753K LEs, 229 Mb BRAMs*
- ✓ *5760 FP DSPs (7TF @ 600MHz)*
- ✓ 8GB DDR4 x 4ch
- ✓ PCIe Gen3 x16
- ✓ 2x QSFP28 (100Gb/s)

FIM (FPGA Interface Manager)

- ✓ インタフェースを含む固定領域

FPGAカードとFPGA SoC



Intel FPGA PAC D5005

- ✓ *Stratix 10 FPGA (14nm)*
- ✓ *2753K LEs, 229 Mb BRAMs*
- ✓ *5760 FP DSPs (7TF @ 600MHz)*
- ✓ 8GB DDR4 x 4ch
- ✓ PCIe Gen3 x16
- ✓ 2x QSFP28 (100Gb/s)

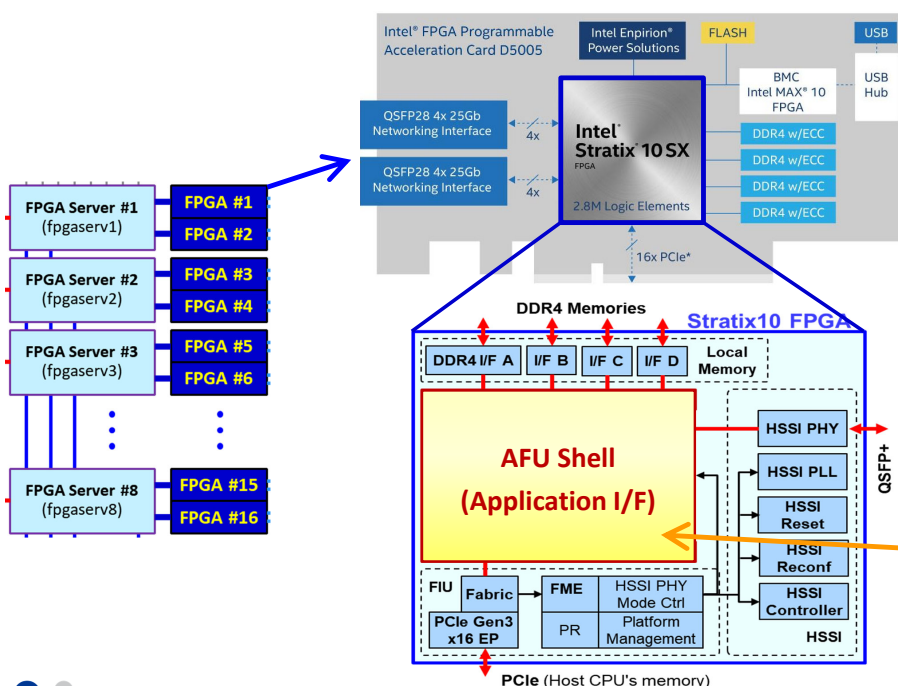
FIM (FPGA Interface Manager)

- ✓ インタフェースを含む固定領域

AFU (Acceleration Function Unit)

- ✓ 再構成可能領域

FPGAカードとFPGA SoC



Intel FPGA PAC D5005

- ✓ *Stratix 10 FPGA (14nm)*
- ✓ *2753K LEs, 229 Mb BRAMs*
- ✓ *5760 FP DSPs (7TF @ 600MHz)*
- ✓ 8GB DDR4 x 4ch
- ✓ PCIe Gen3 x16
- ✓ 2x QSFP28 (100Gb/s)

FIM (FPGA Interface Manager)

- ✓ インタフェースを含む固定領域

AFU (Acceleration Function Unit)

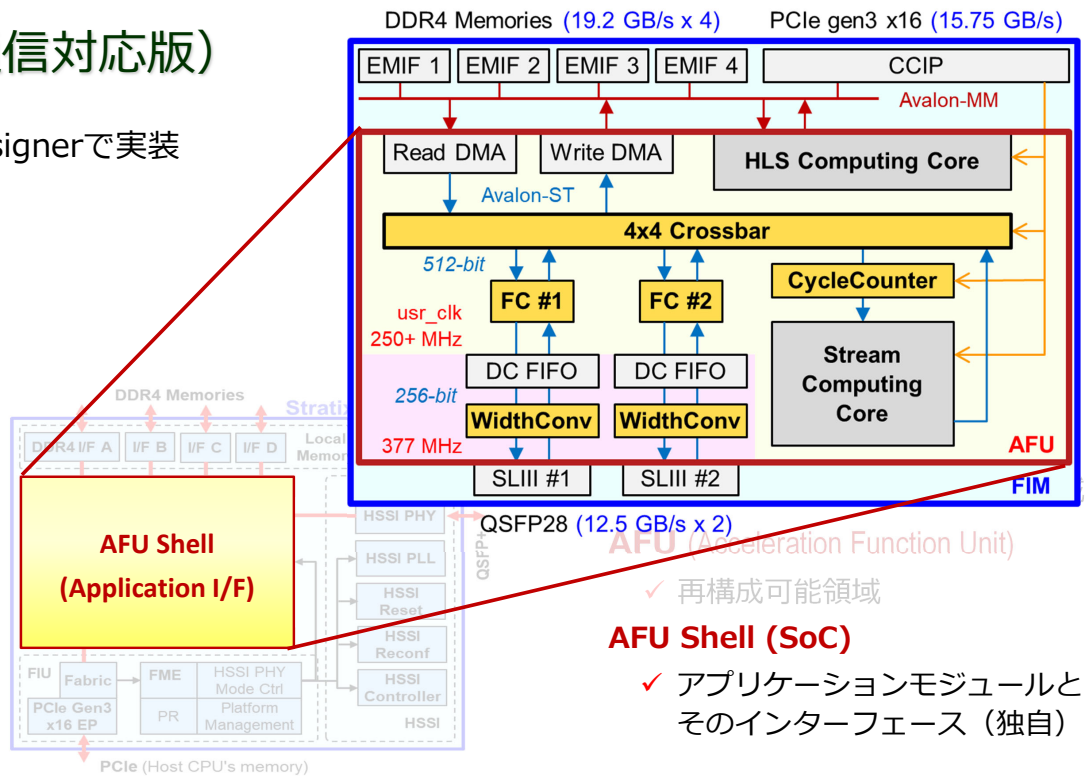
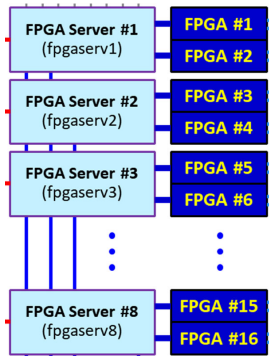
- ✓ 再構成可能領域

AFU Shell (SoC)

- ✓ アプリケーションモジュールとそのインターフェース (独自)

AFU Shell (高速シリアル通信対応版)

✓ Intel Platform Designerで実装



ハードウェアのプログラミング

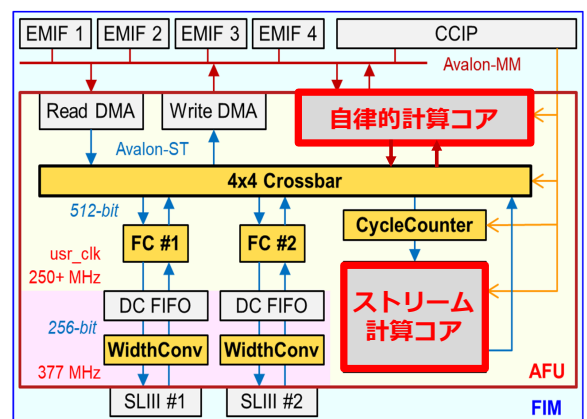
計算コアを実装し AFU Shellに組込む

- ✓ **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

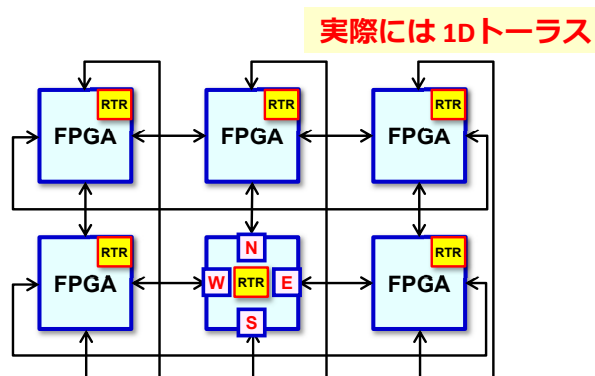
- ✓ **ソフトウェア指向** **HLS** アルゴリズムをC/C++で記述 (Intel HLS)
- ✓ **ハードウェア指向** **HDL** ハードウェア構造を記述 (Verilog-HDL, VHDL, Chisel, etc.)
- ✓ **その他** **DSL** ドメイン特化型言語等 (ストリーム計算コアコンパイラ : SPGen)

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



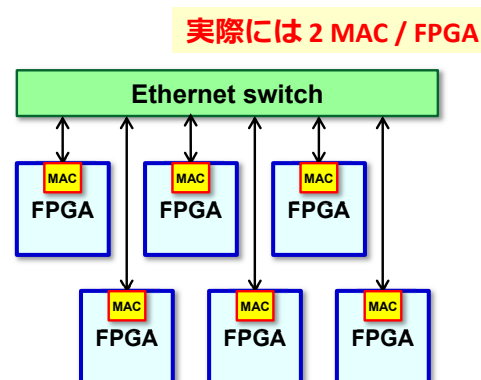
HLS: High-level synthesis, **Chisel**: Scala-based language for RTL, **SPGen**: Stream processor generator

開発するFPGA間ネットワーク



直接網 (1D / 2D 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)

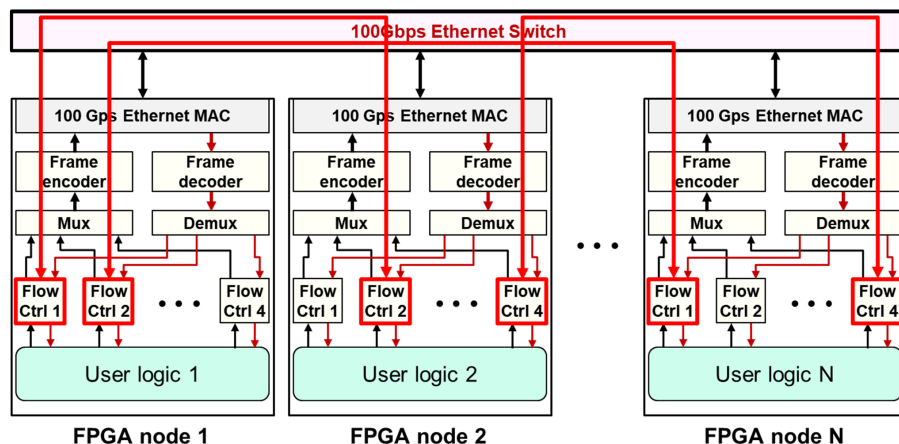
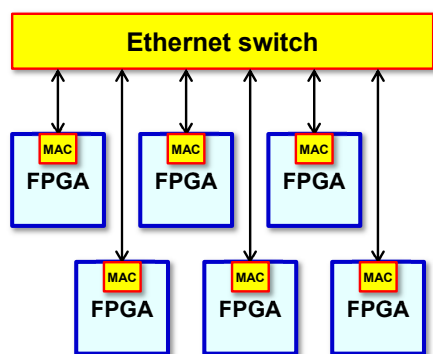


間接網 (100G Ethernet)

- 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.
- ✓ **Virtualized circuit switching network** : form any topology with any groups of FPGAs.



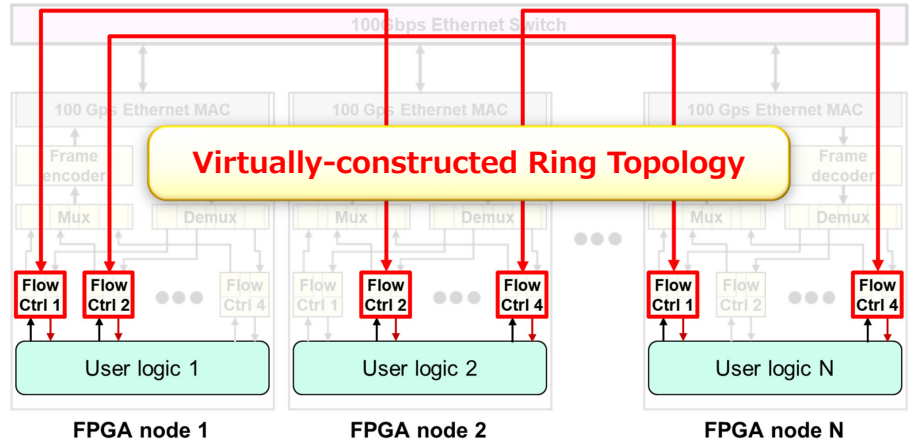
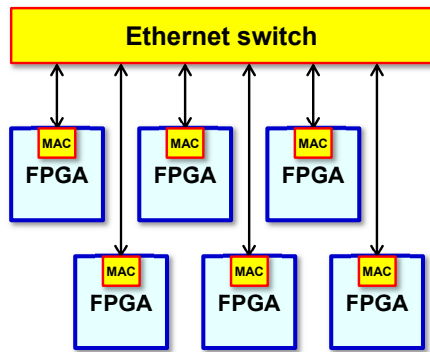
100G Ethernet switches

- Pros)** Flexibility, cutting-edge technology
- Cons)** Overhead of ethernet frames, higher and variable latency, difficulty in flow-control and use

Ethernet上に仮想的な回線交換網を実現

- Inter-FPGA connection should also be flexible.

✓ **Virtualized circuit switching network** : form any topology with any groups of FPGAs.



100G Ethernet switches

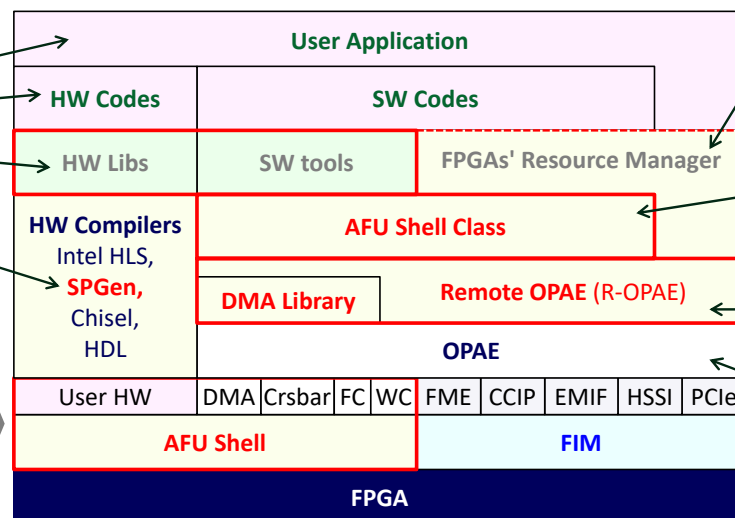
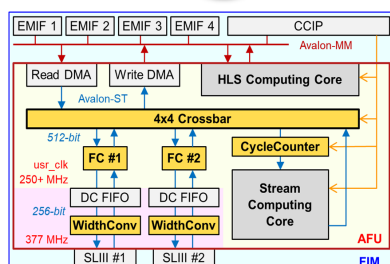
Pros) Flexibility, cutting-edge technology

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

ソフトウェアスタック開発状況と共同研究の可能性

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

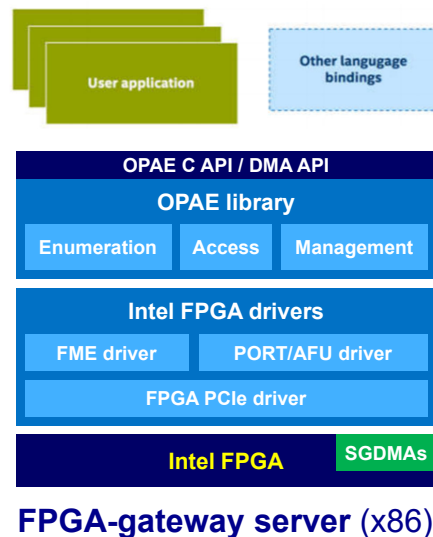
OPAe

- Low-level driver

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

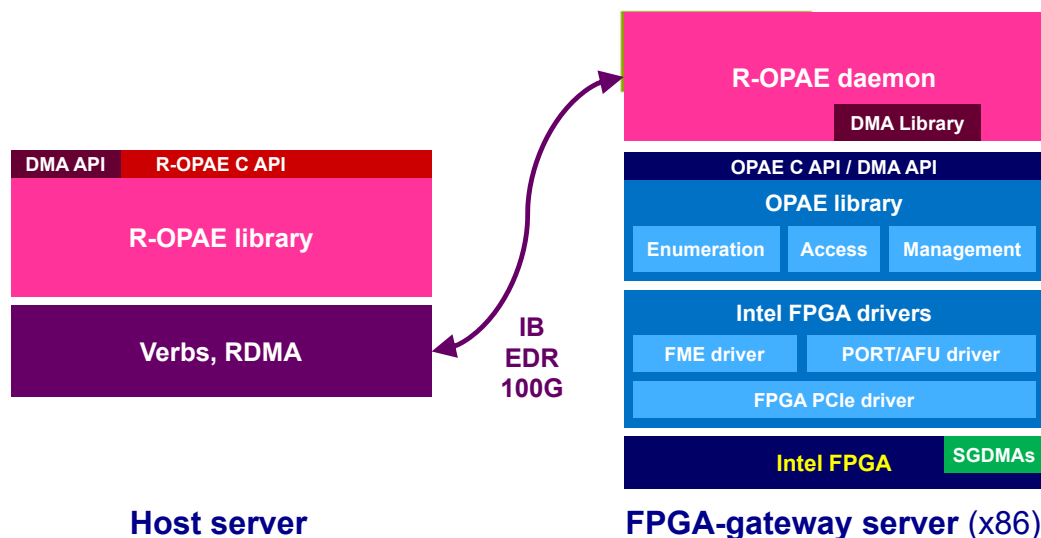
- ✓ Intel FPGA PAC のドライバ : OPAE (ローカル)



Remote-OPAE (ソフトウェアブリッジによる遠隔化)

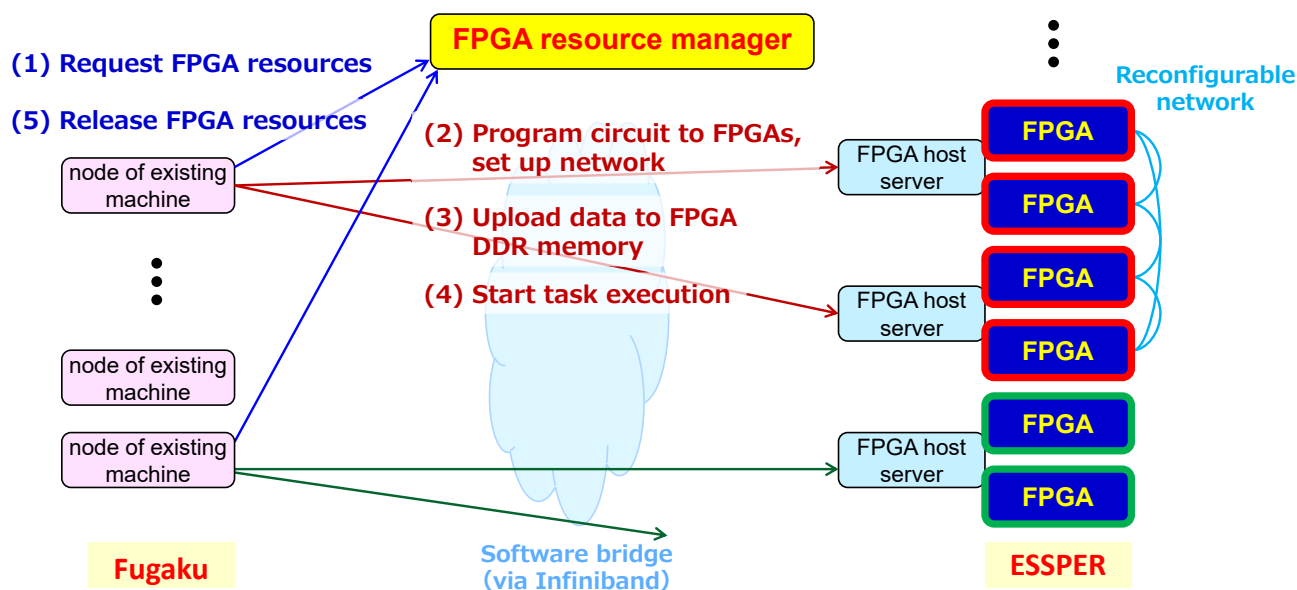
インフィニバンドネットワーク越しに、遠隔のFPGAを利用可能

- ✓ 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)



FPGA資源管理サービス

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



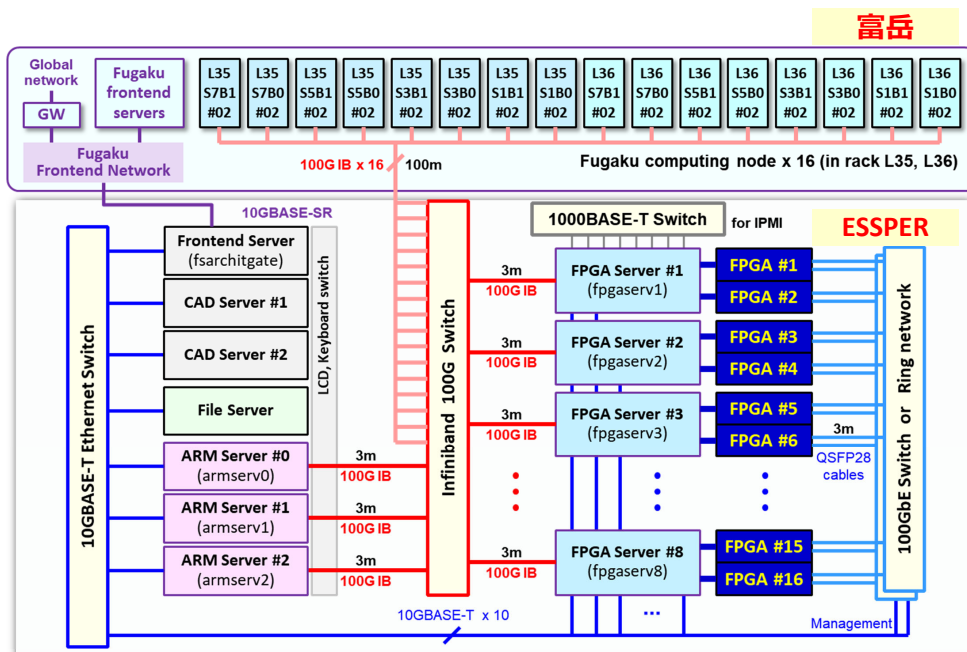
動作検証とデータ転送性能

提案するFPGAクラスタと システムスタックを実装

- ✓ FIM, AFU Shell (SL3, Eth)
- ✓ R-OPAE
- ✓ afushell_class
- ✓ 資源管理マネージャ(開発中)

富岳に接続し実験

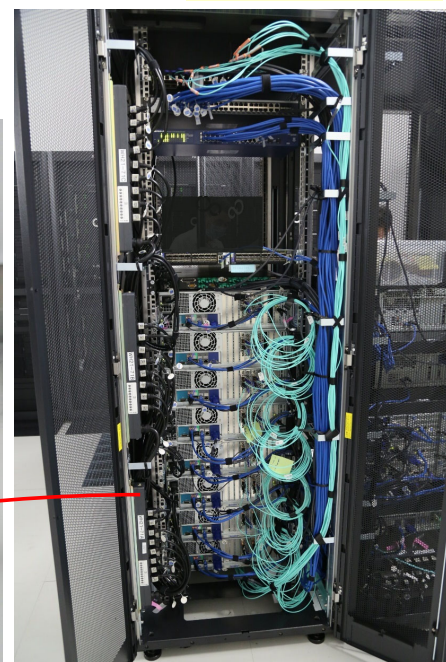
- ✓ 富岳に付加価値を
与える方式の検討課題
(FY2017～2020)
- ✓ 100m IB EDR ケーブル



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」

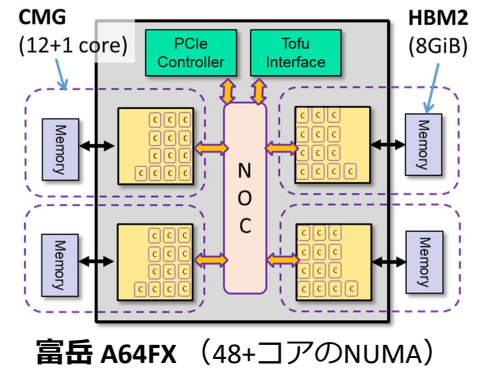


動作検証と性能測定

使用したサーバ

- ✓ FPGA サーバ Intel Xeon Gold5122 (3.6GHz, 4 コア) x2
- ✓ ARM サーバ Cavium ThunderX2 (2.0GHz, 28 コア) x2
- ✓ 富岳ノード Fujitsu A64FX (2.0GHz, 48+2 コア) x 1

x86, ARM, 富岳 Linux環境でafushellクラス、R-OPAE等をリンク可



動作検証 (OPAE, R-OPAE)

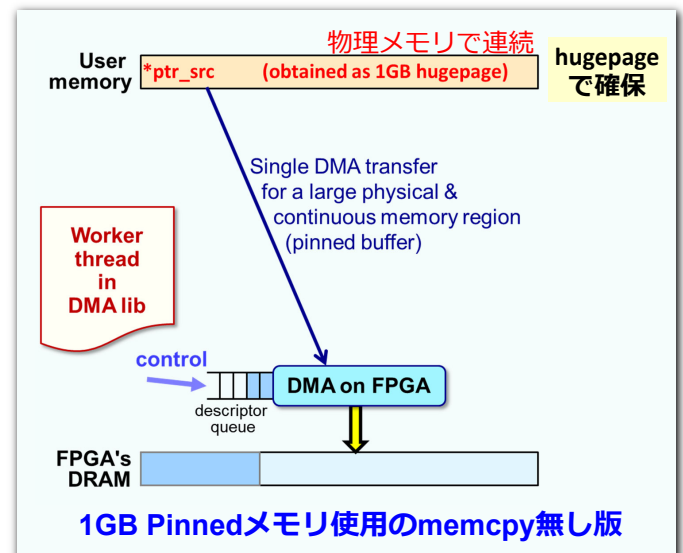
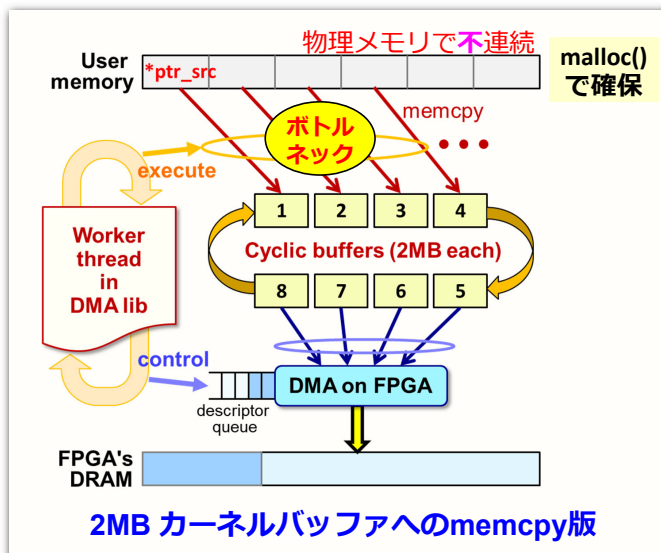
- ✓ FPGAアプリを実行するのに必要な動作を確認
 - AFUの部分再構成 (2秒程度)
 - CPU・FPGA間データ転送
 - FPGA上モジュールの制御

FPGAアプリ実行の一連の流れ

- ✓ IBネットワークを介して利用可能なFPGAを探索
- ✓ FPGAへAFU Shellのbit streamを書込み
- ✓ AFU Shellのデバイスをオープン
- ✓ **FPGA DDR4へデータをアップロード**
- ✓ FPGAの計算コアを駆動
- ✓ **FPGA DDR4からデータをダウンロード**
- ✓ AFU Shellのデバイスをクローズ

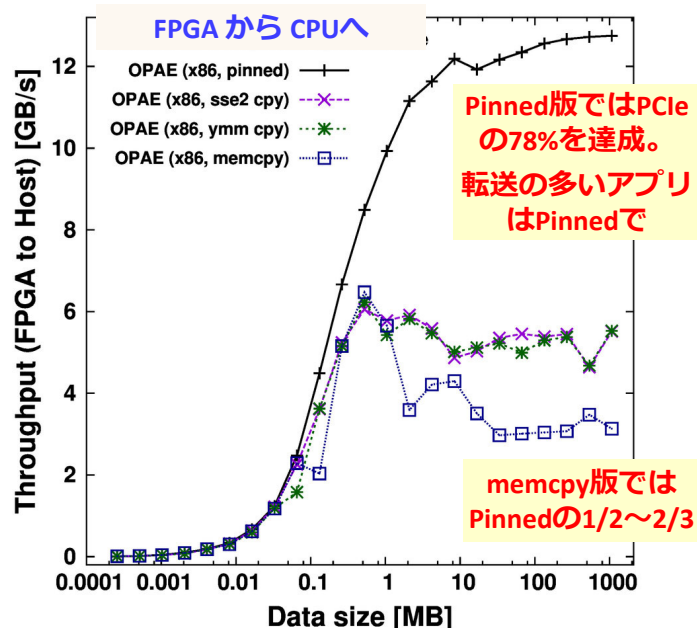
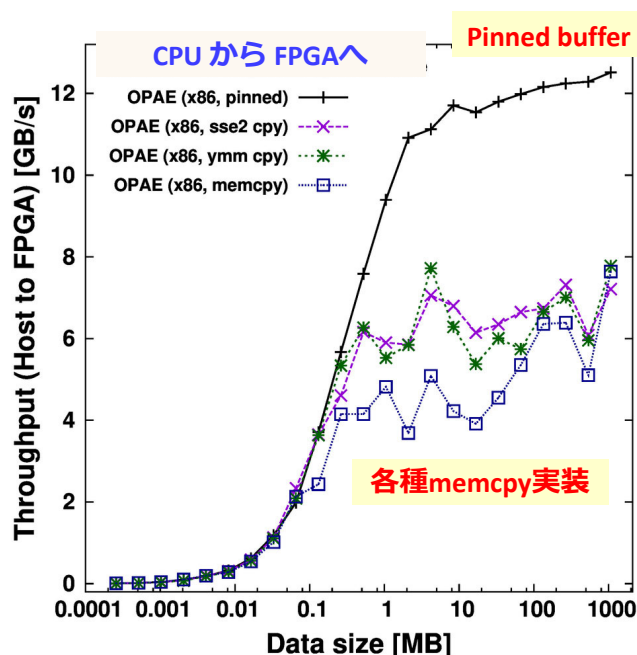
CPU・FPGA間データ転送性能の評価

- ローカルFPGAへの転送 (OPAE) vs. IB越しのリモートFPGAへの転送 (R-OPAE)
- ローカル転送 (OPAE) では、様々なDMA制御ライブラリ実装で評価



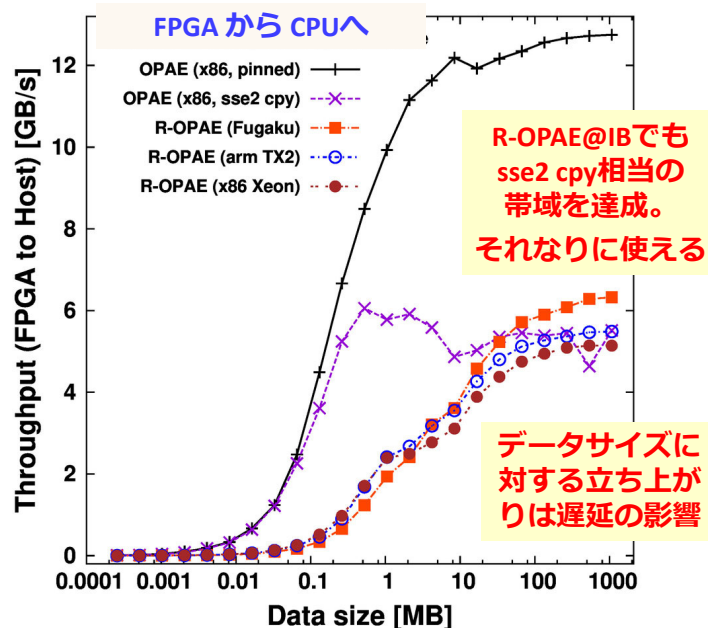
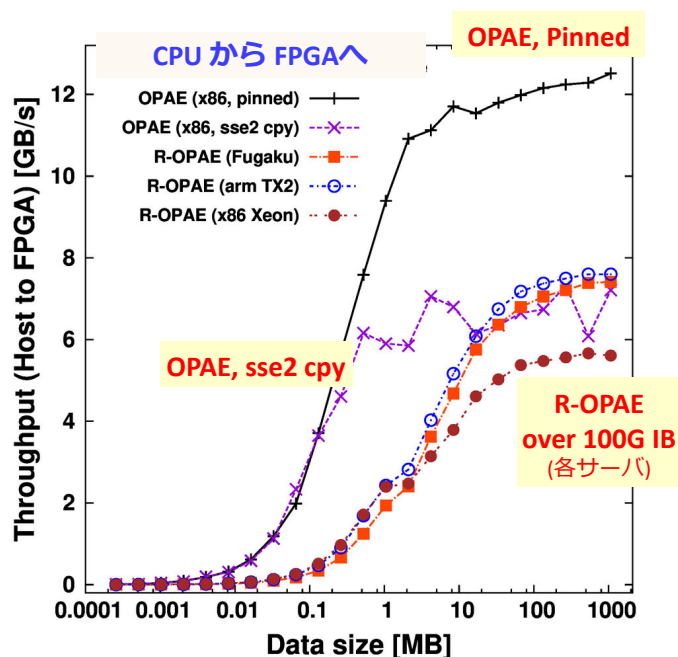
ローカルFPGAとのDMA転送

• PCIe Gen3 x16	16 GB/s
• DDR4-2400 (300MHz)	19.2 GB/s
• 100Gbps IB EDR	12.5 GB/s



R-OPAeとOPAe の転送性能

• PCIe Gen3 x16	16 GB/s
• DDR4-2400 (300MHz)	19.2 GB/s
• 100Gbps IB EDR	12.5 GB/s



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

開発内容 ハードウェア・ソフトウェアスタックが完成、運用中

- ✓ **R-OPAE、FPGA間ネットワーク**、AFU Shell、afushellクラス、他
- ✓ 富岳との接続実験。動作検証とデータ転送性能評価。 **「R-OPAEは使える」**
- ✓ アプリ開発に注力中
- ✓ システムを利用した共同研究を実施中（5大学 > 順次拡大）

今後の課題

- ✓ アプリ開発フローの自動化、複数FPGAプログラミング方式、
システムソフト（資源管理/タスクスケジューリング）、高性能アプリ開発

共同研究・産学連携を拡大

- ✓ FPGAクラスタ向けアプリを募集中
- ✓ システムやIPの産業利用など？ > **Intel PAC OEM企業の皆様、他の皆様**

求む、研究員！
「R-CCS2022 公募」
で検索



HBM2搭載 Stratix10 DX
向けAFU Shellを開発中