超低消費電力型光エレクトロニクス実装技術開発事業
（光エレ実装PJ）

システム化技術
情報処理システム化技術
ラックスケール並列分散システム

2022年 2月 10日

技術研究組合　光電子融合基盤技術研究所（PETRA)

賣野　豊

アウトライン

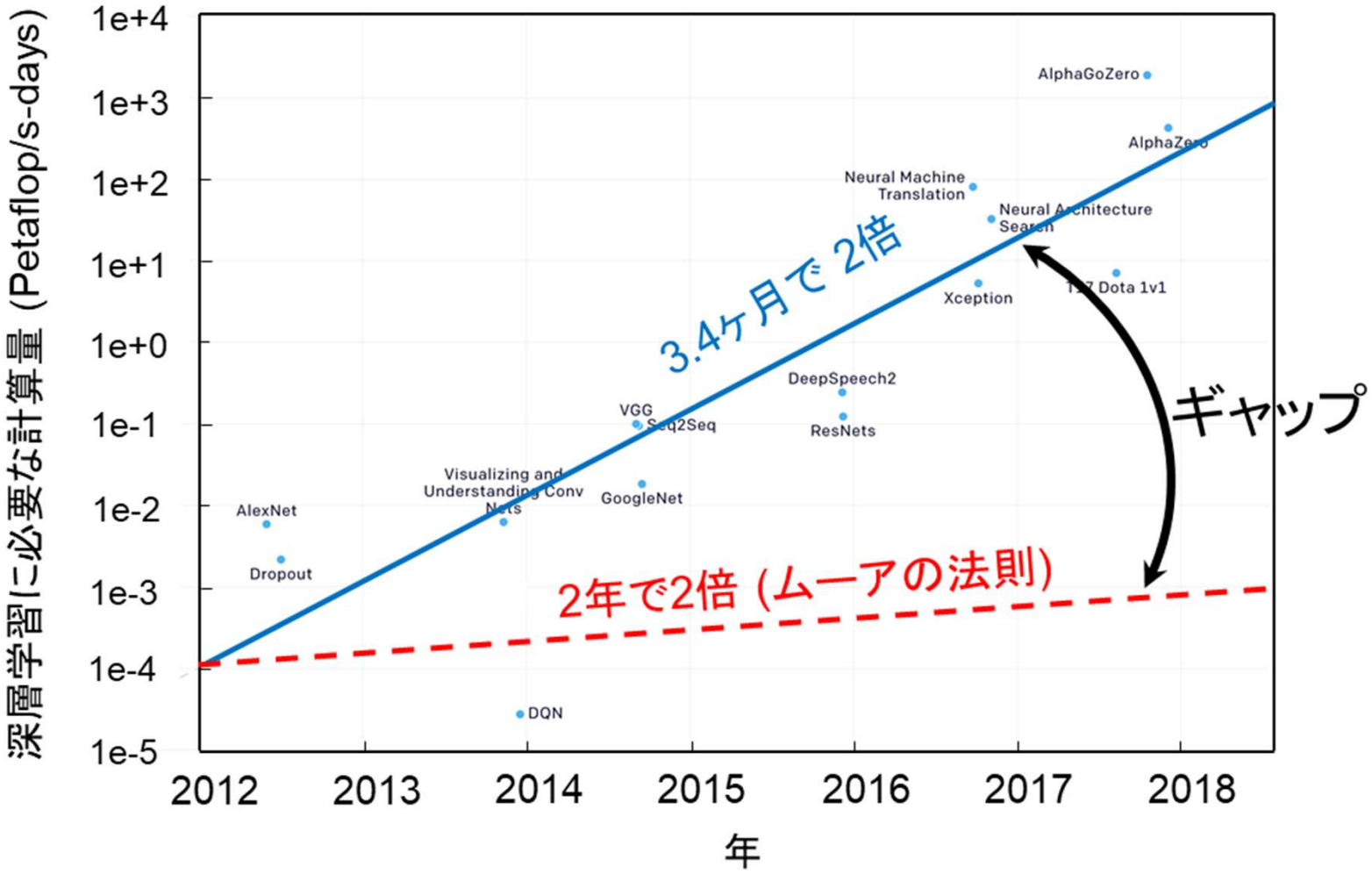
- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

深層学習のトレンド

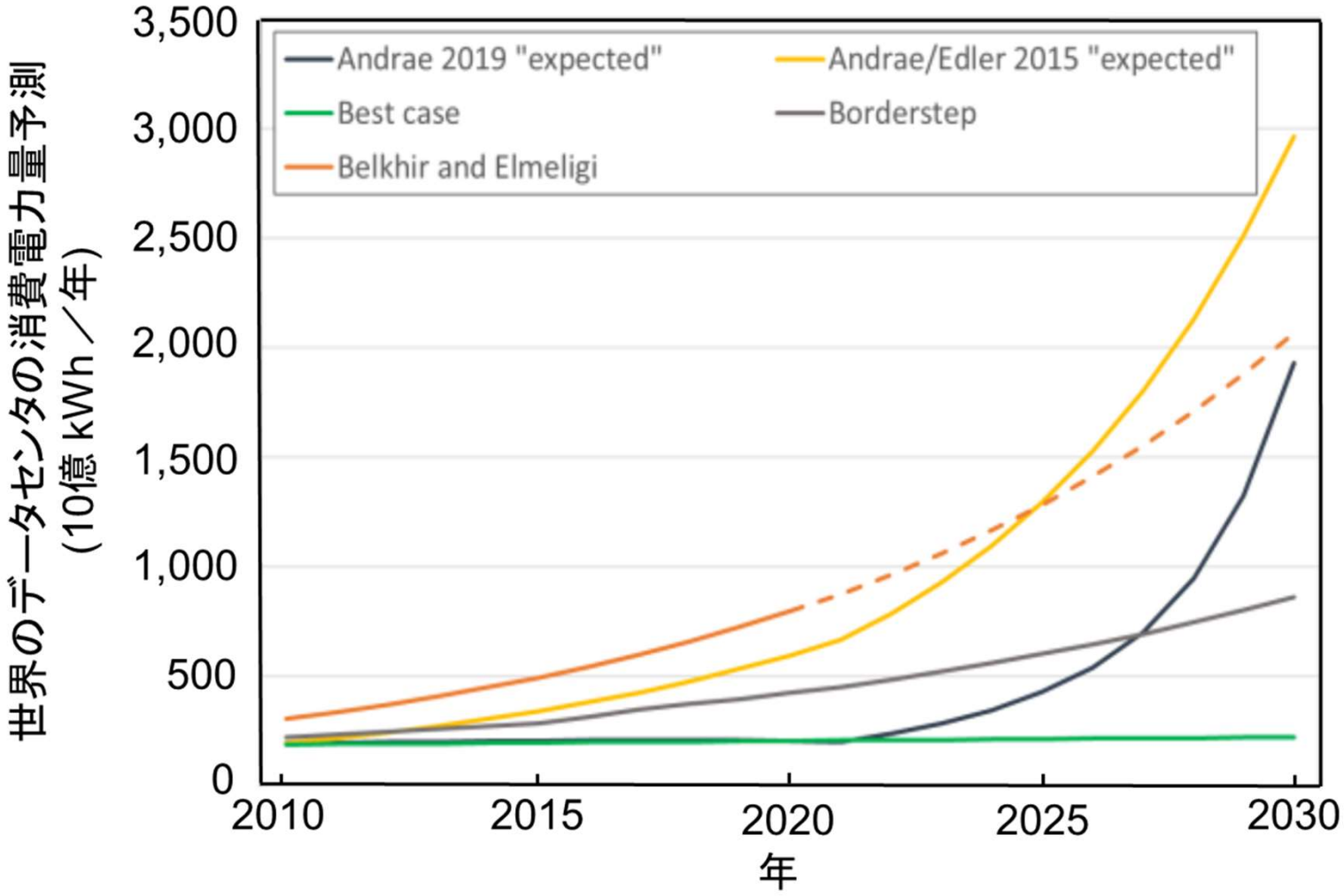
- 深層学習の計算量の伸びは、CPU性能の伸び（Moore's Law）より遥かに速い
- CPU計算能力の需要と供給のギャップは年々拡大
- 並列計算とヘテロジニアス計算（アクセラレータ）が、このギャップを埋める鍵



https://openai.com/blog/ai-and-compute/

データセンタの課題

データセンタの消費電力量は、
2020年代に急増すると予測されている



https://www.researchgate.net/publication/341427004_Energy_consumption_of_data_centers_worldwide_How_will_the_Internet_become_green

アウトライン

- 背景
  - データセンタの動向と課題
- **目標とアプローチ**
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

光エレ実装PJにおける目標とアプローチ

- 目標

  計算ノード間を光配線で接続した光配線サーバにより、
  従来型の電気配線サーバに比べて、
  サーバの消費電力量（エネルギー）を30%以上削減する

- アプローチ
  - サーバで実行されるワークロード単位で考える
  - 消費電力量＝サーバの消費電力×ワークロードの実行時間
  - サーバは、デバイスからアプリまで、
    多層の技術レイヤ、多数のパーツで構成される
  - 消費電力は足し算　⇒　各レイヤ／パーツ毎に削減
  - 実行時間は最大値　⇒　システム全体の協調設計で削減

ラックスケール並列分散システムの全体像

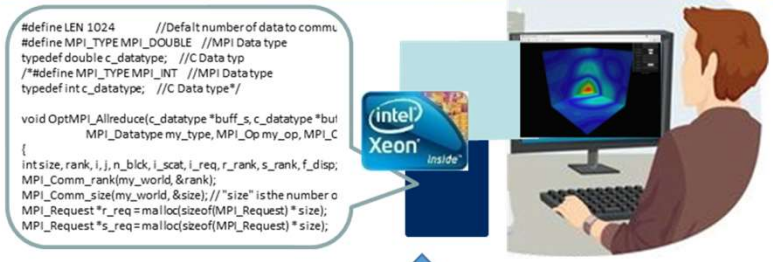
並列分散処理
アプリケーション

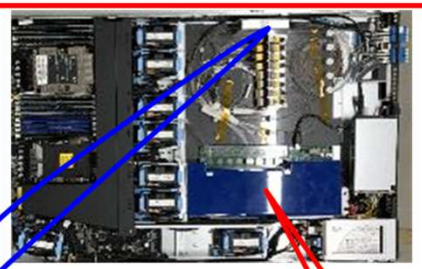
サーバ・システム

リンク＆ルーティング
（レイヤ２＆３）

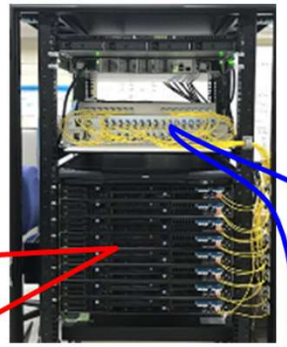
波長多重ルーティング
ネットワーク
（レイヤ1）

デバイス

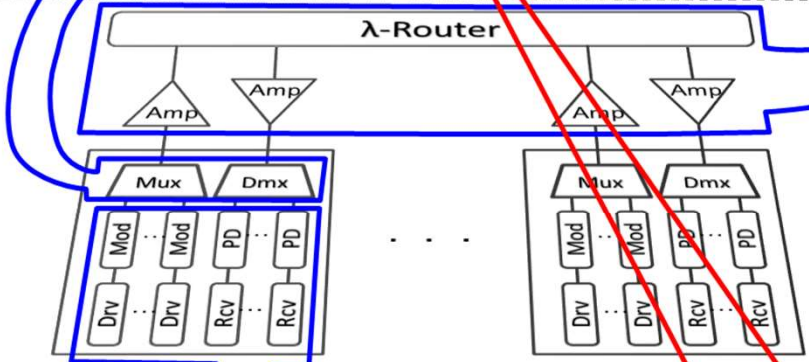


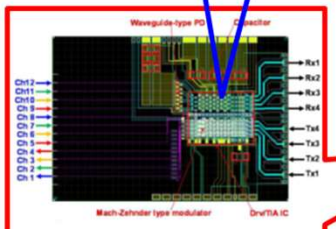


光電子融合サーバボード

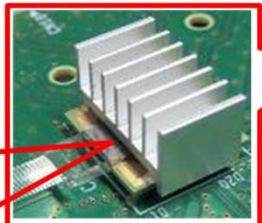


ラックサーバ





シリフォト・トランシーバ・チップ

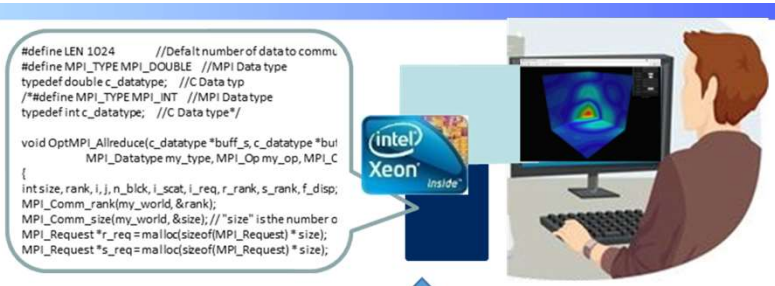


光電子集積インターポーザ

FPGAカード

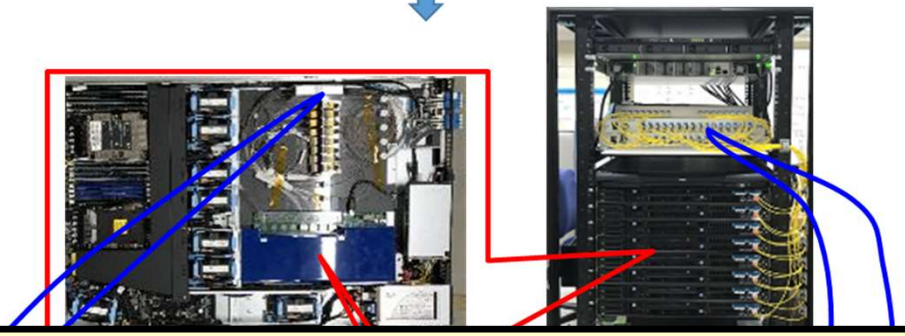
ラックスケール並列分散システムの全体像

並列分散処理
アプリケーション



サーバ・システム

リンク＆ルーティング
（レイヤ２＆３）

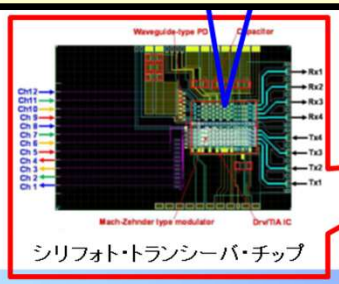


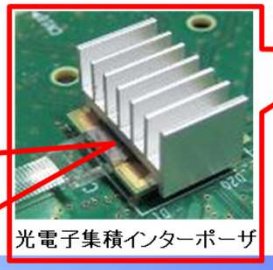
データ移動を加速するアクセラレータとして、

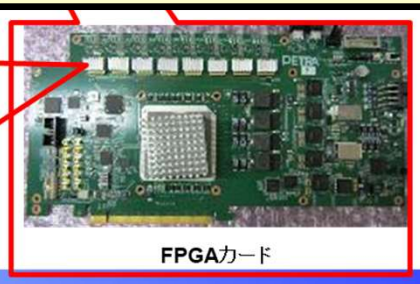
FPGAカードを搭載した8台の計算ノードで構成される

並列 & ヘテロジニアス計算システム

デバイス


シリフォト・トランシーバ・チップ


光電子集積インターポーザ

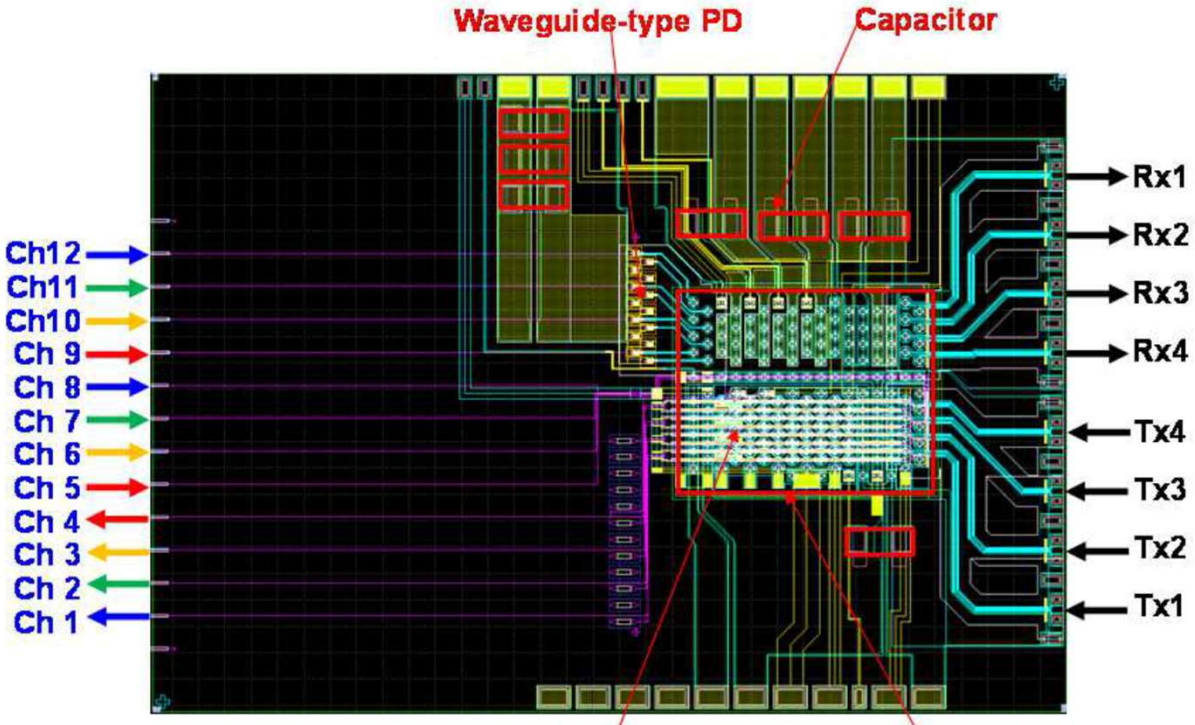

FPGAカード

アウトライン

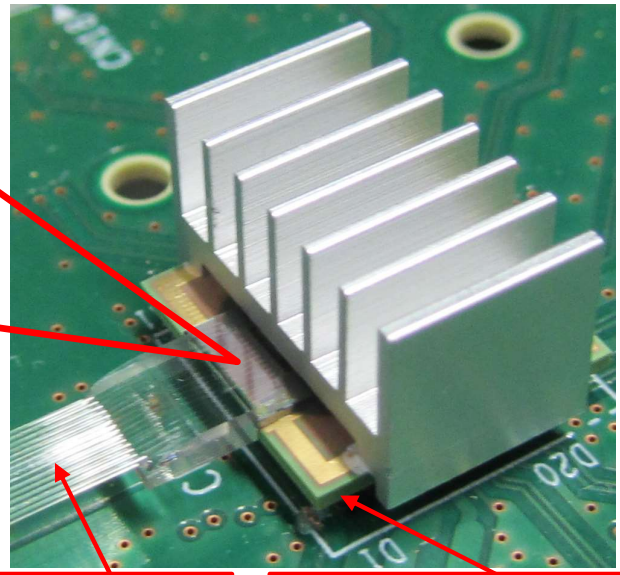
- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

光電子集積インターポーザ (EOM)

- シリコンフォトニクス・トランシーバ・チップ（5×7 mm^2）
- 高帯域密度＆小粒度（25Gbps×4ch. ／ 12mm角）
- C帯DWDM対応、12芯SMF-MTコネクタ付
- オンボード・オプティクス：FPGAカードとソケット接続（着脱可）



シリフォト・トランシーバ・チップのレイアウト

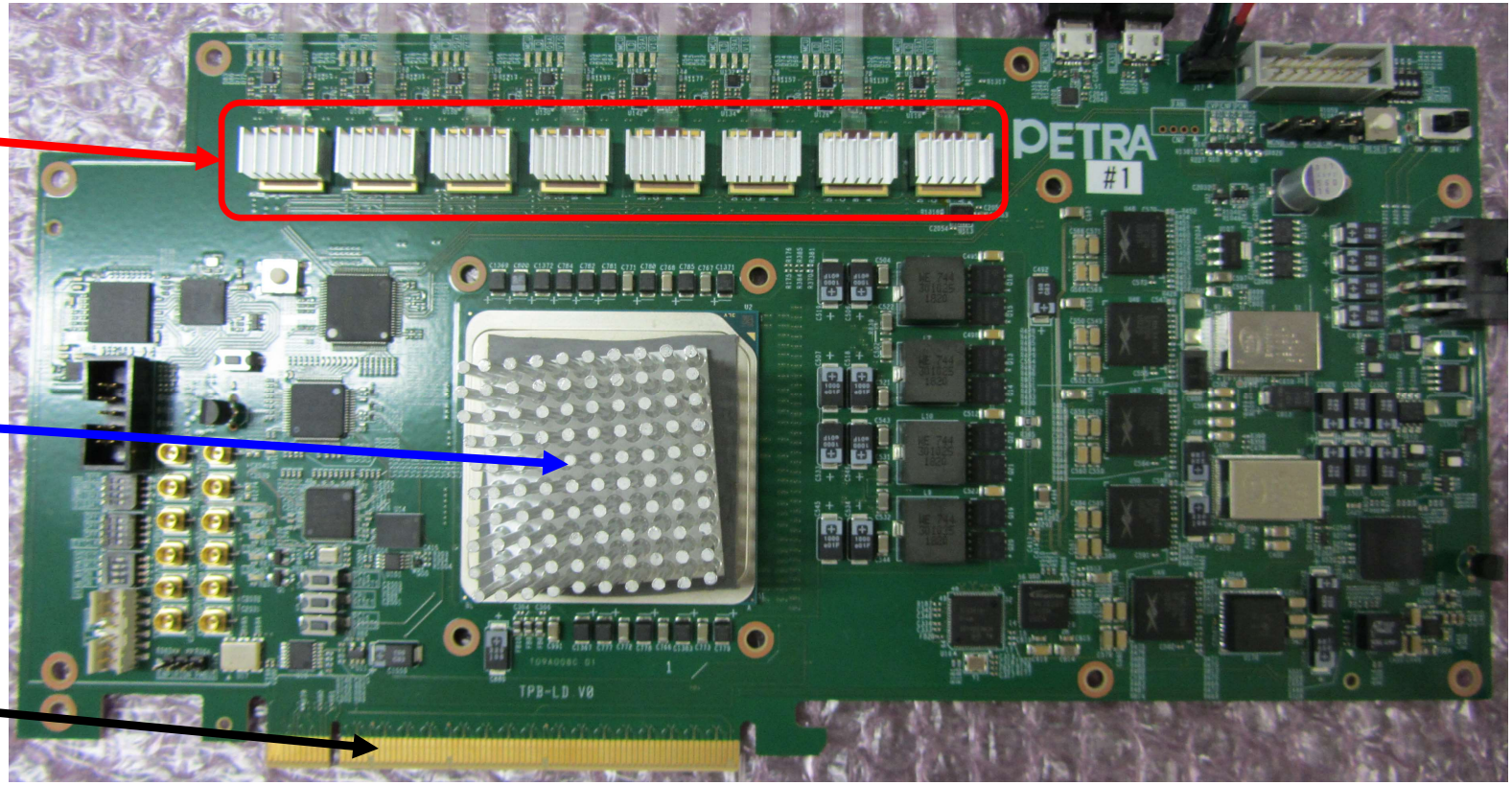


光電子集積インターポーザ

カスタムFPGAカード

- Intel Stratix10 MX FPGA:
  - 低クロック周波数(～200MHz) ⇒ 低消費電力
  - HBM2メモリ: 819 Gbpsのメモリ帯域
- EOM×8: 800 Gbpsのネットワーク帯域

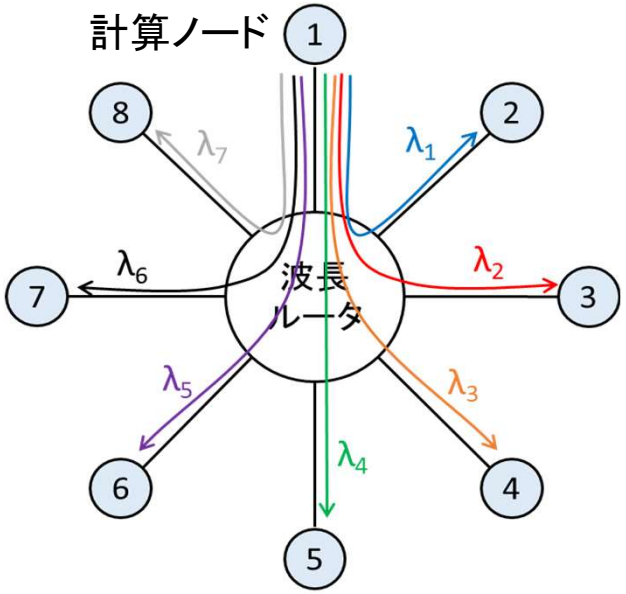
バランス



アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

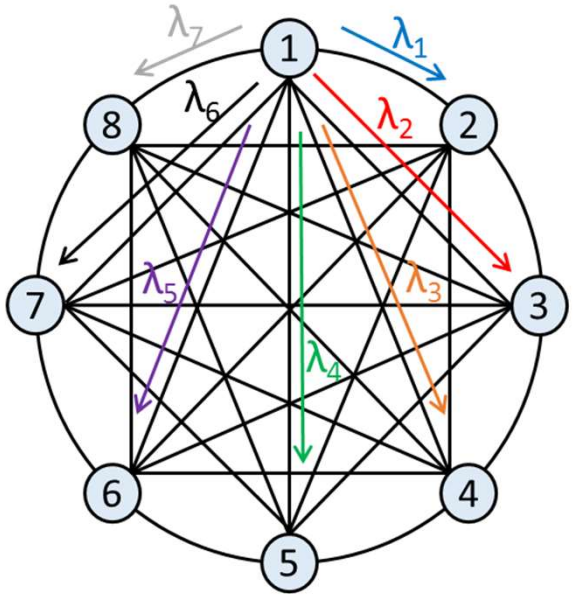
波長ルーティング・ネットワーク構成と動作原理



物理トポロジー
（スター型）

入力ポート＼波長	出力ポート 1	2	3	4	5	6	7	8
1		λ1	λ2	λ3	λ4	λ5	λ6	λ7
2	λ7		λ1	λ2	λ3	λ4	λ5	λ6
3	λ6	λ7		λ1	λ2	λ3	λ4	λ5
4	λ5	λ6	λ7		λ1	λ2	λ3	λ4
5	λ4	λ5	λ6	λ7		λ1	λ2	λ3
6	λ3	λ4	λ5	λ6	λ7		λ1	λ2
7	λ2	λ3	λ4	λ5	λ6	λ7		λ1
8	λ1	λ2	λ3	λ4	λ5	λ6	λ7	

波長ルータの周期的な
ルーティングテーブル

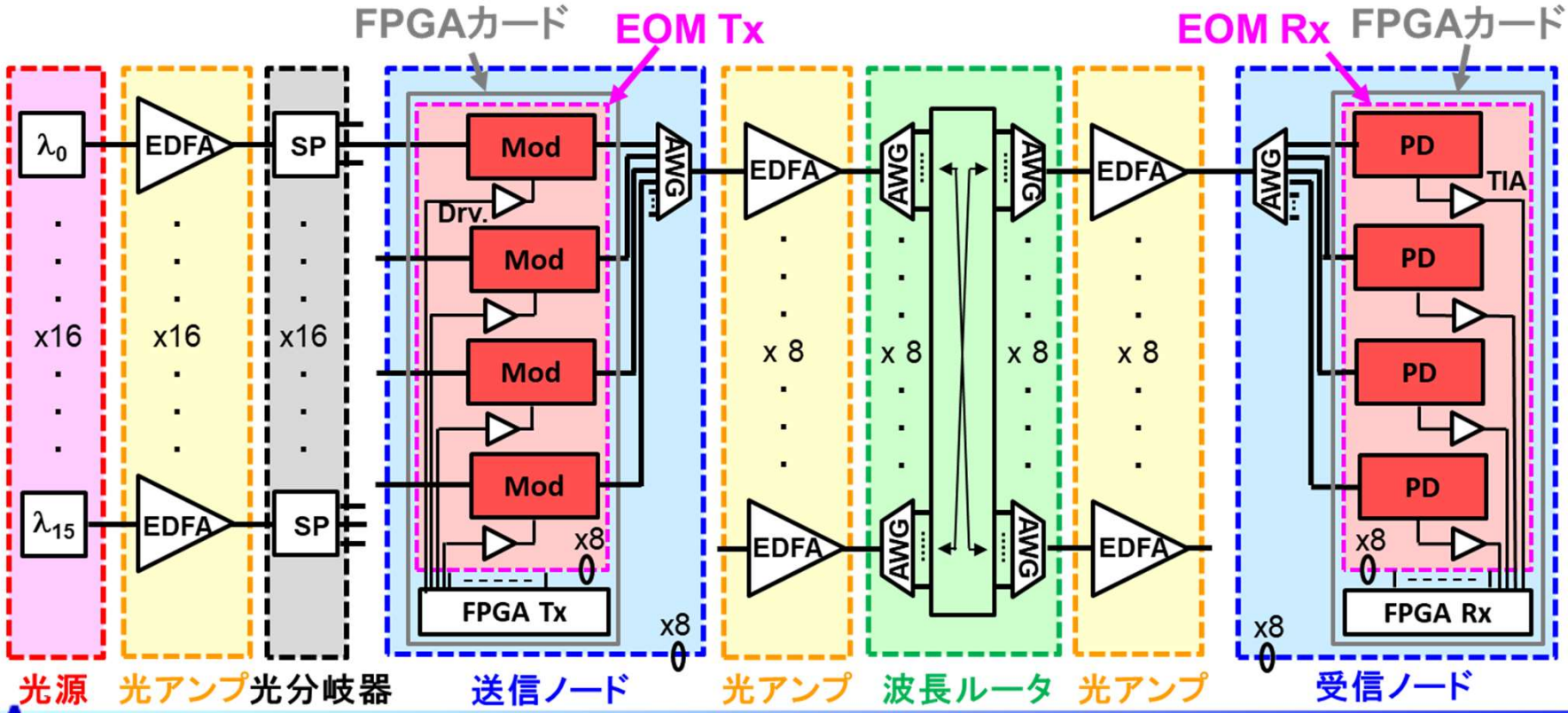


論理トポロジー
（フルメッシュ）

- ✓ 物理トポロジーは、パッシブな波長ルータをハブとしたスター型。
- ✓ 波長ルータでは、入力ポートと光の波長に応じて、周期的に出力ポートが決定される
- ✓ 論理トポロジーは、全てのノード間が常時直接接続（フルメッシュ・ネットワーク）
  ⇒ パケット衝突無し ⇒ 複雑なパケット処理不要
  ⇒ マルチ・ホップ不要 ⇒ 低遅延通信可

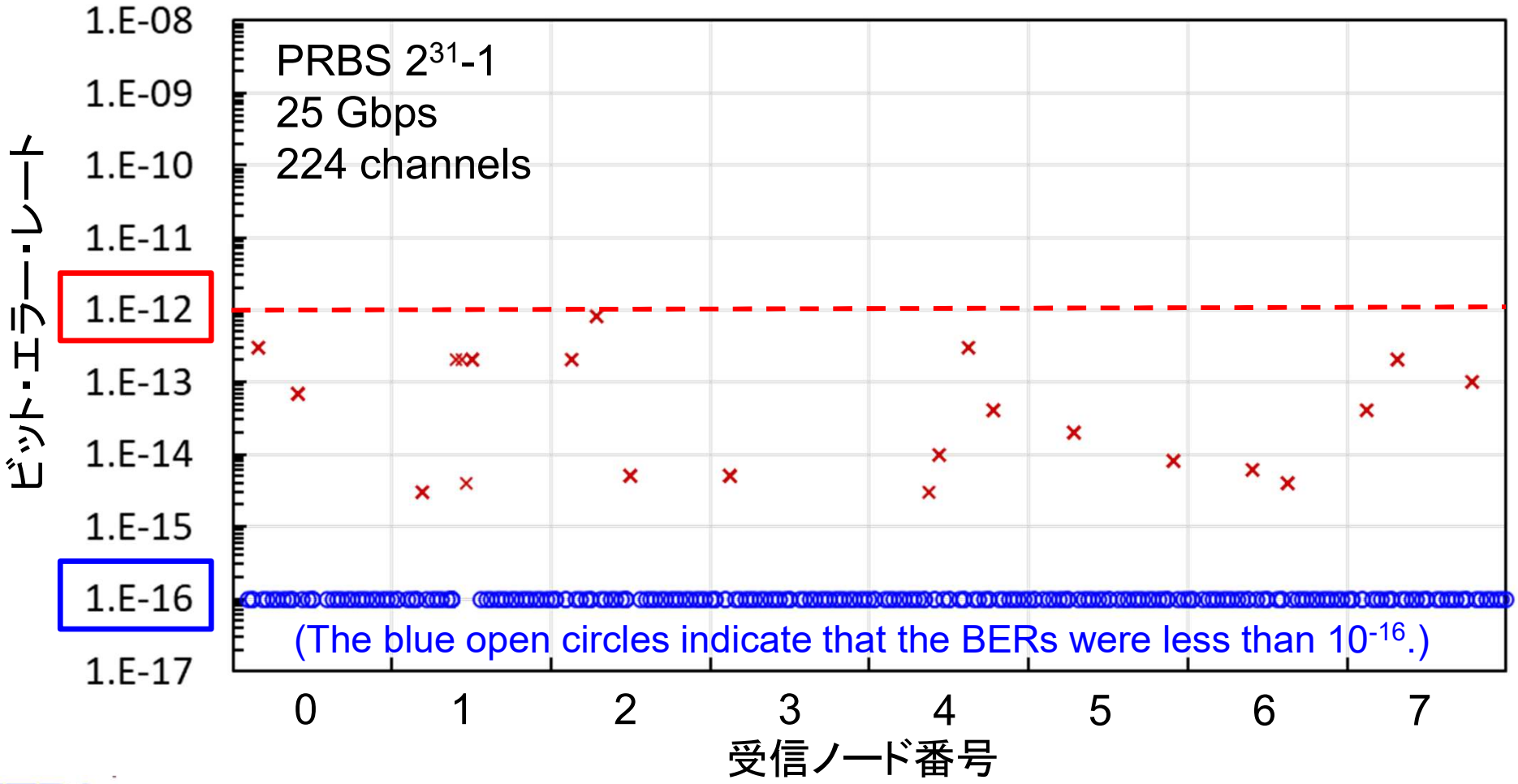
波長ルーティングネットワークの構成要素

- オフチップ波長多重光源
  ⇒ 集中管理により、高精度波長制御、高エネルギー効率、高信頼性
- C帯EDFA光アンプ ⇒ 高エネルギー効率、低ノイズ
- 波長ルータ ⇒ パッシブ・デバイス（消費電力無し、EO/OE変換無し）
- 送受信ノード以外の全コンポーネントがビット・レート、変調方式に無依存
  ⇒ 送受信ノード（トランシーバ）の世代交代に対応可能



波長ルーティング・ネットワークの全チャンネル・エラーフリー動作

8台のFPGA間を接続する全224chで、BER < 10^{-12} を確認
（その内の91%は、BER < 10^{-16}）
⇒ トータル帯域 5.6Tbpsのエラー・フリー動作を実証
⇒ （電気・光共に）高い信号品質と高い均一性を実証

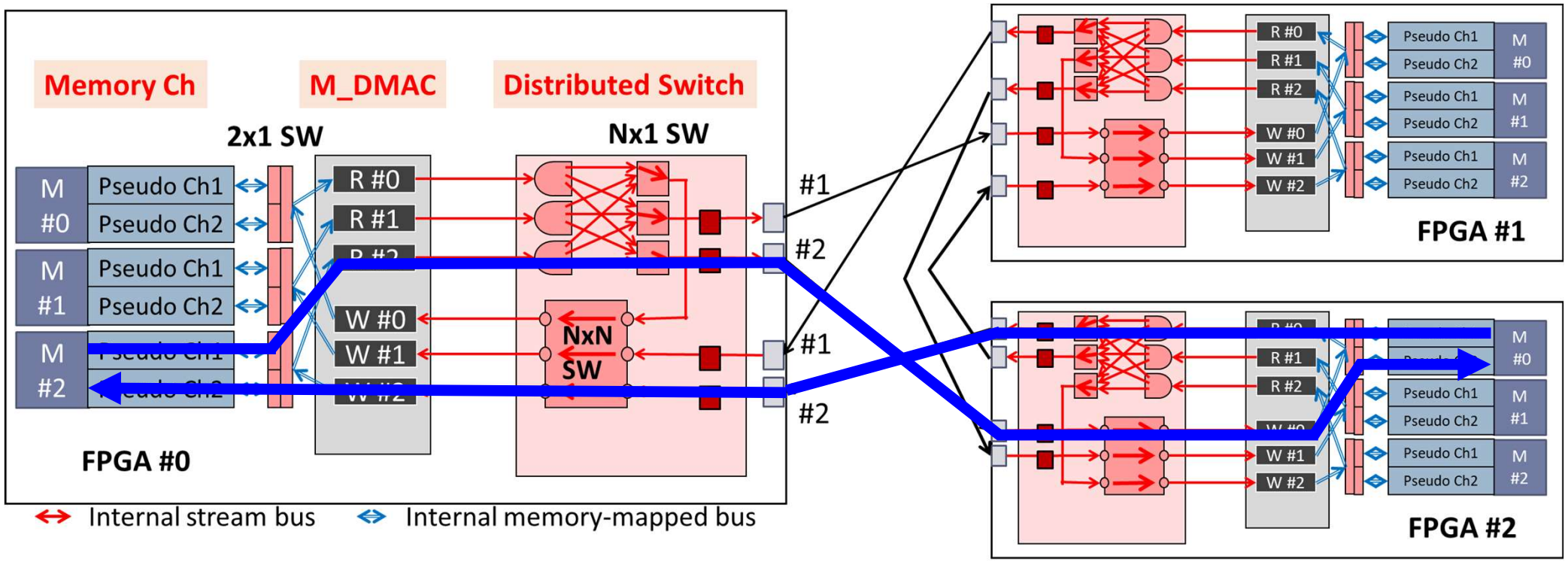


アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

OPTWEB：軽量なFPGA間ネットワーク・アーキテクチャ

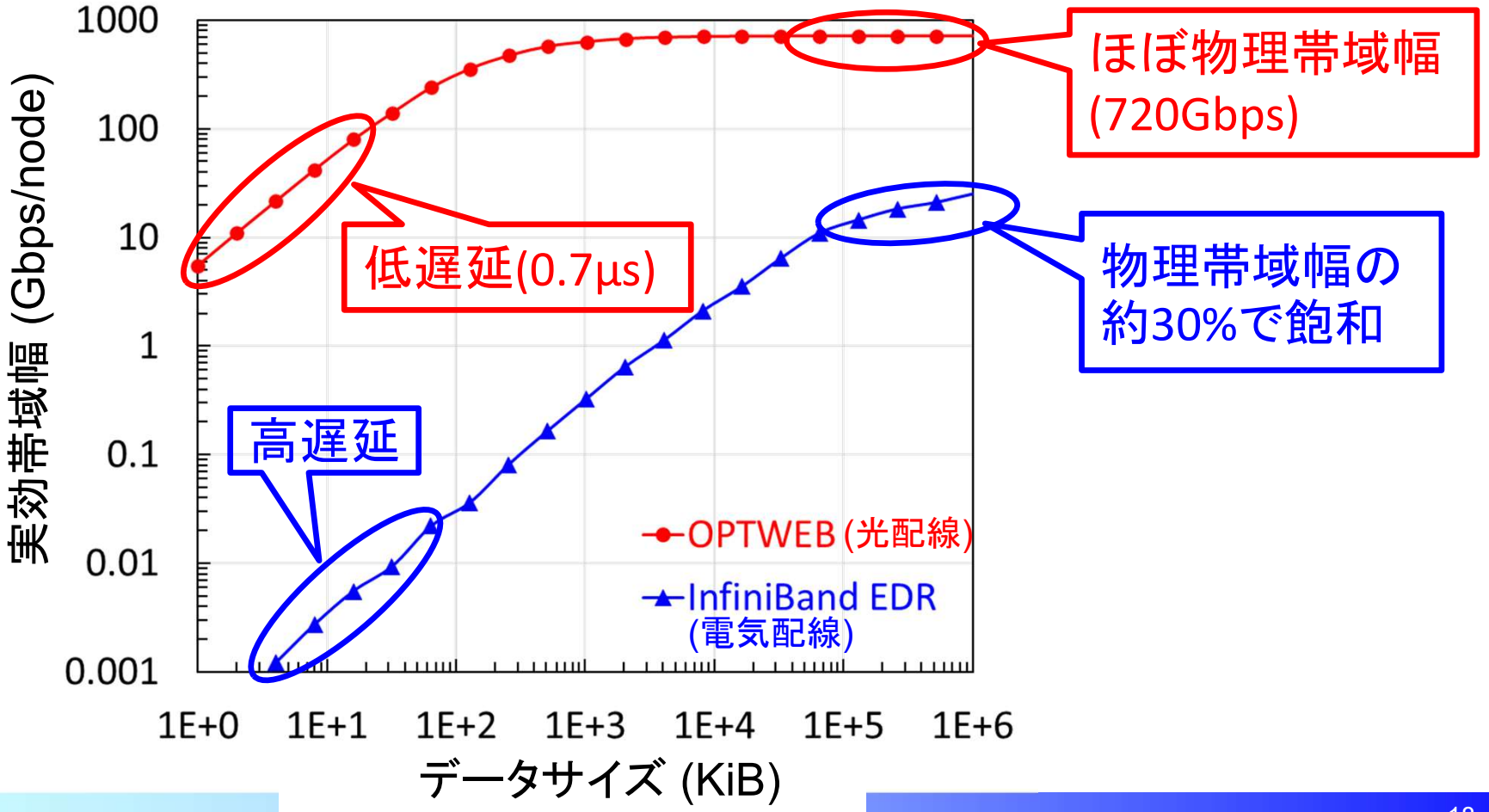
- FPGA間のmemory-to-memoryの常時接続専用パスを構築
- **例えば、FPGA #0のmemory #2とFPGA #2のmemory #0を直結**
- FPGA間のリモート・ダイレクト・メモリ・アクセス(RDMA)を実現



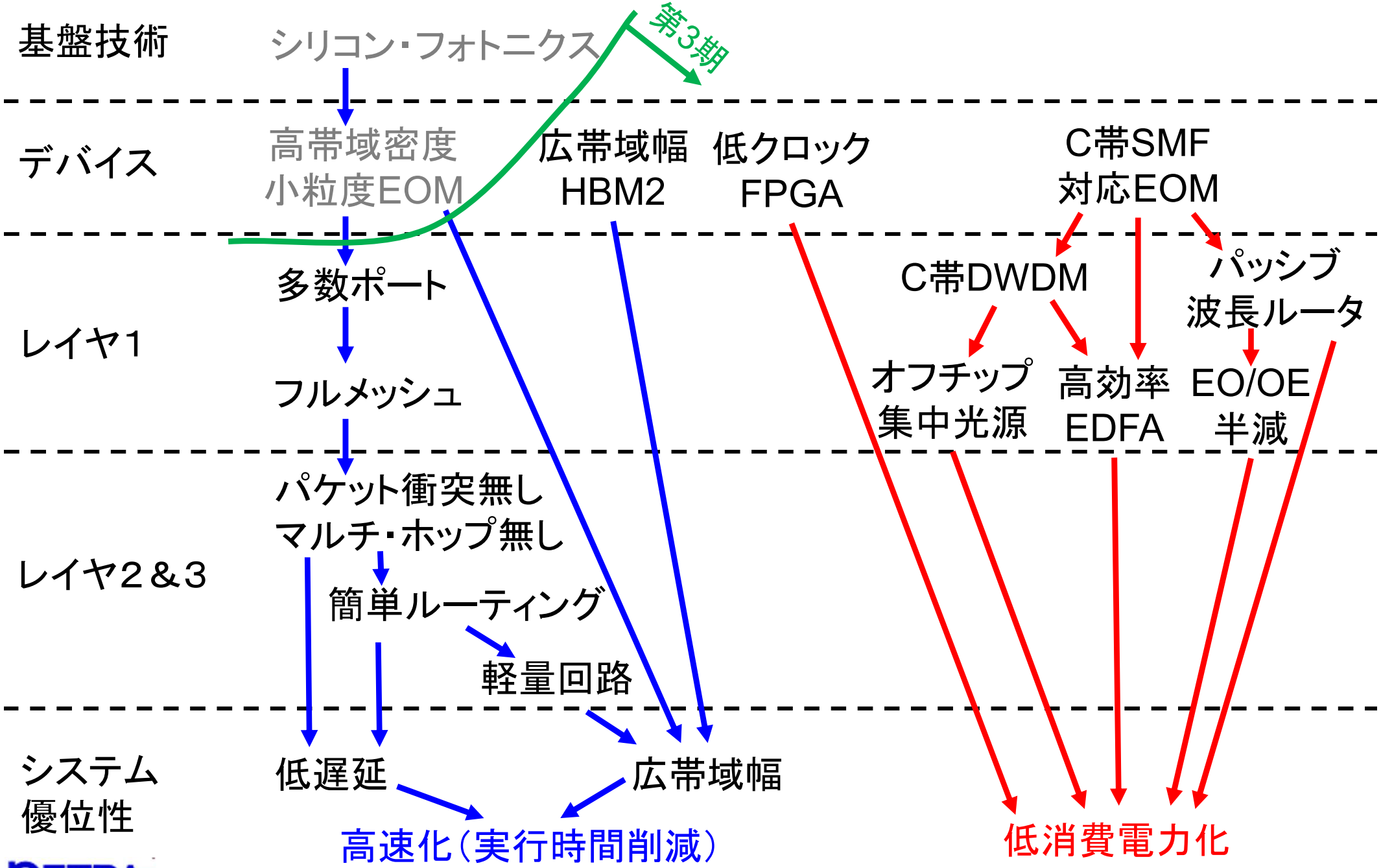
3台のFPGA間をOPTWEBで接続した場合の構成（実際は8台間を接続）

8ノード間Alltoall通信の実効帯域幅の測定

- (実効帯域幅) = (データサイズ) ／ (通信時間)
- (通信時間) ≈ (遅延時間) + (データサイズ) / (物理帯域幅)
- OPTWEB: フルメッシュ・ネットワーク ＋ 簡単なネットワーク制御
  - ⇒ 小さなデータに対して: 低遅延 ⇒ 比較的広い実効帯域幅
  - ⇒ 大きなデータに対して: ほぼ物理帯域幅に近い実効帯域幅



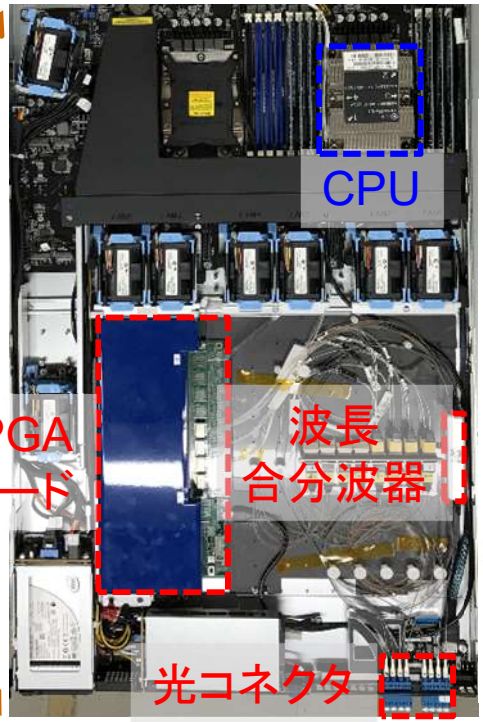
低消費電力量化に向けたアプローチ（ここまでのまとめ）



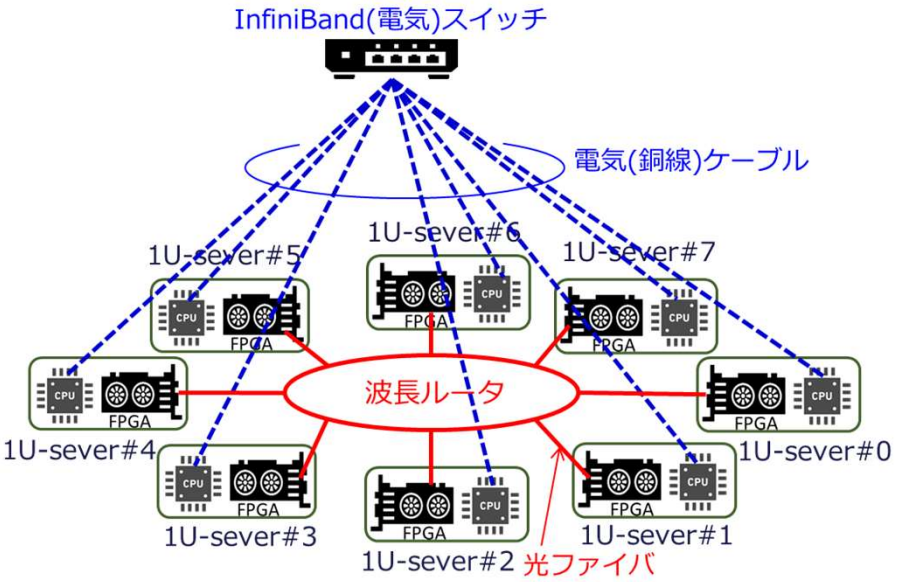
アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

8ノード・ラックサーバ・システム



光電子融合サーバボード



8ノード間の電気／光接続

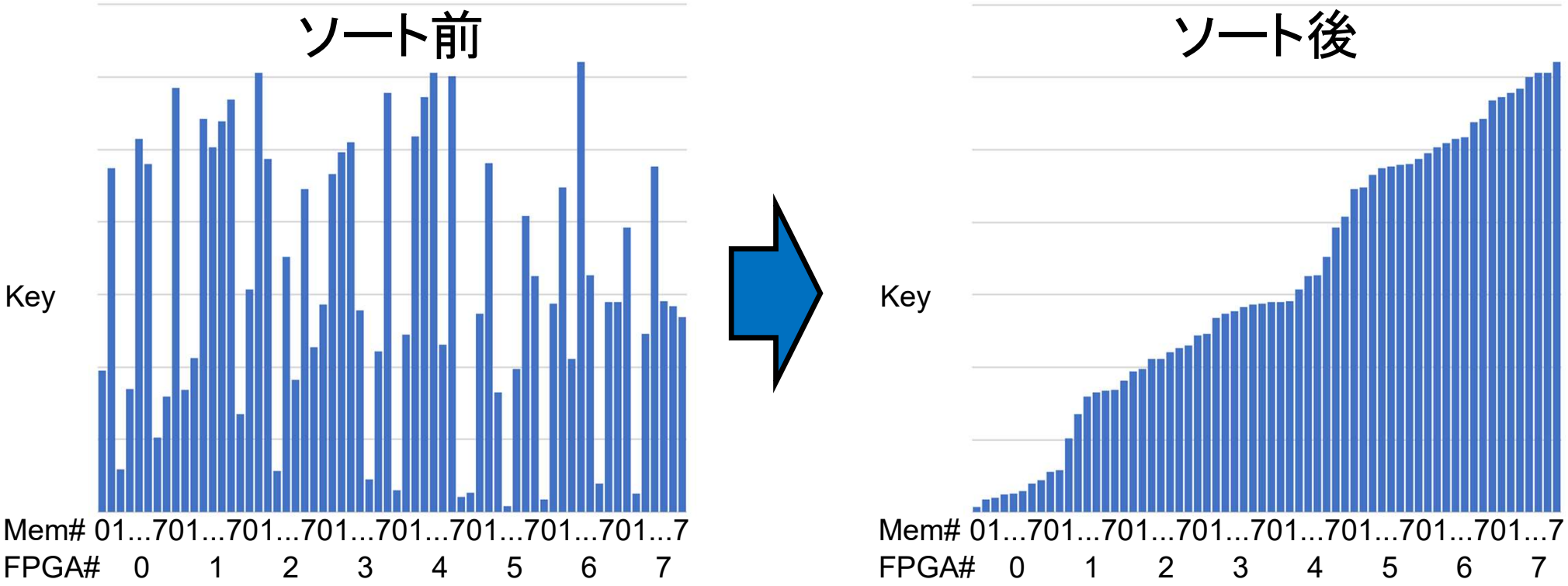
- 各光電子融合サーバボード（1Uサーバ）にCPUとFPGAカードを搭載
- CPU間は、電気配線（100-Gbps DACケーブル ＋ InfiniBand EDR）で接続
- FPGA間は、光配線（波長ルーティング＆OPTWEB）で接続
- 16波長多重　⇒　1Uサーバのフロントパネルの光コネクタ数を1/16に削減
  ⇒　空冷の気流を妨げない

アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

並列分散ソータ

- 8ノード・ラックサーバ・システムに並列分散ソータを実装
- 32bit整数キーを8台のFPGAの8ヶ所のメモリ領域にランダムに格納
- 4bitソート×8回で、32bitのソーティング処理
- 各ソートは、パイプライン処理により、1回のメモリアクセス(Read/Write)
- グローバルソートのAlltoall通信は、波長ルーティング＋OPTWEBで高速化

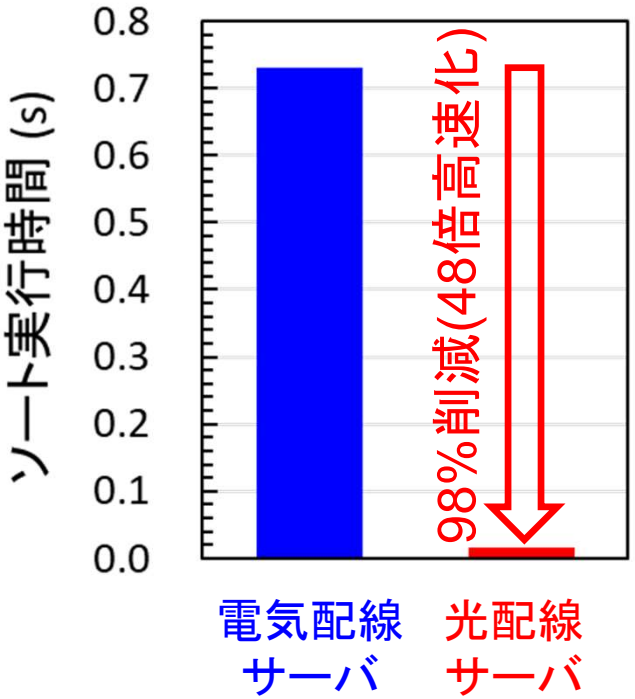


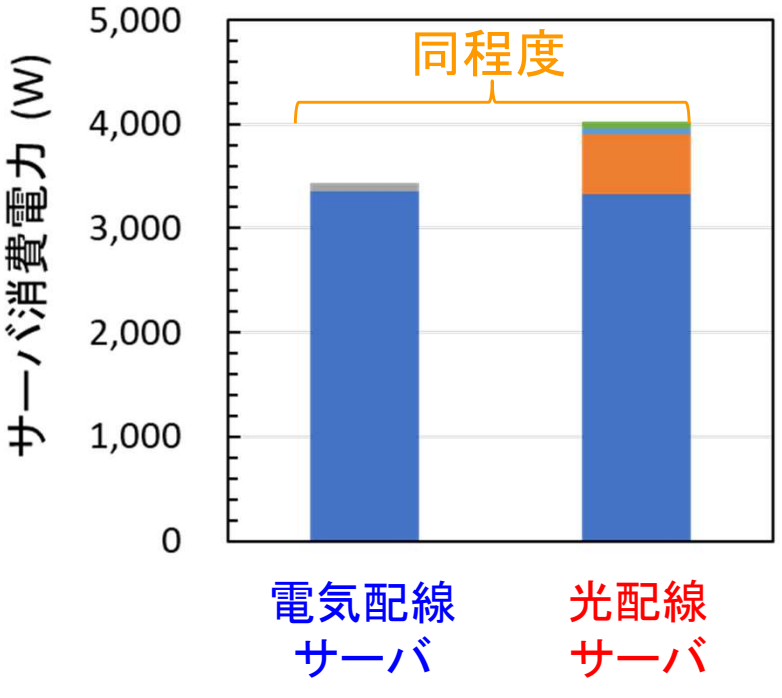
ソート処理のイメージ

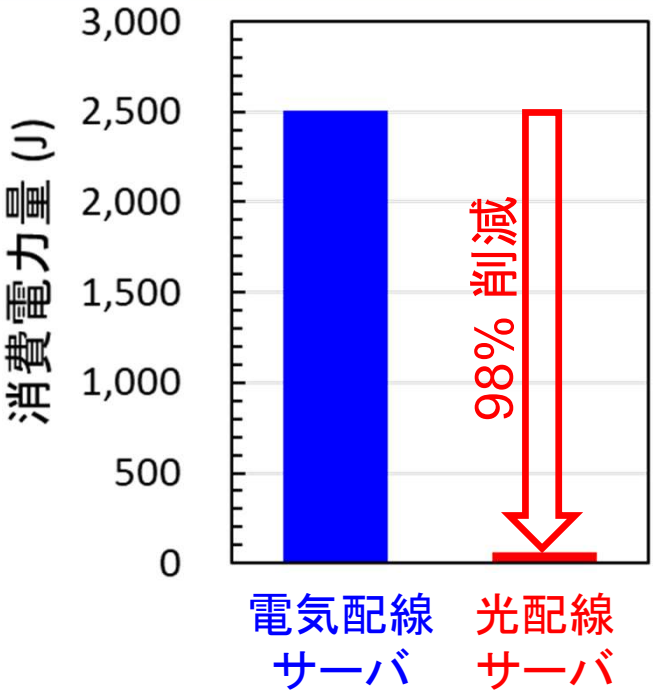
アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- **システム性能評価**
- まとめ

実行時間／消費電力／消費電力量の評価







Key数：2^{32}(約40億)

- 光配線サーバにより、電気配線サーバに比べて、
  ソート実行時間の98%削減（48倍の高速化）を実証
- 光配線サーバと電気配線サーバの消費電力は、ほぼ同じ
- 光配線サーバにより、電気配線サーバに比べて、
  消費電力量（実行時間×消費電力）の98%削減を実証

アウトライン

- 背景
  - データセンタの動向と課題
- 目標とアプローチ
- 各技術レイヤにおける設計と試作
  - デバイス
  - 物理ネットワーク（レイヤ１）
  - リンク＆ルーティング（レイヤ２＆３）
  - サーバ
  - アプリケーション
- システム性能評価
- まとめ

まとめ

デバイスからアプリまでの協調設計により、高速・低消費電力量サーバを開発

- デバイス
  - ✓ DWDM対応高帯域密度小粒度EOM
  - ✓ HBM2メモリ搭載の800GbpsカスタムFPGAカード
- 物理ネットワーク（レイヤ1）
  - ✓ 電力消費無、OE/EO変換半減のフルメッシュ波長ルーティング
  - ✓ 全224チャンネル、総帯域幅5.6Tbpsでエラー・フリー動作を実証
- リンク＆ルーティング（レイヤ2＆3）
  - ✓ OPTWEB：軽量なFPGA間ネットワーク・アーキテクチャ
  - ✓ FPGA間のリモート・ダイレクト・メモリ・アクセス
  - ✓ 低遅延（0.7μs）・広実効帯域幅（720Gbps）のAlltoall通信を実証
- サーバ／アプリケーション
  - ✓ 8ノード（電気配線／光配線）ラックサーバ・システム
  - ✓ 並列分散ソータを実装
  - ✓ 光配線サーバにより、電気配線サーバに比べて、
    48倍の高速化と98%の消費電力量削減を同時に実証

謝　辞

- この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務（JPNP13004）の結果得られたものです。

- FPGA間ネットワーク（OPTWEB）および分散計数ソートのFPGAへの実装にご尽力頂いた、NECプラットフォームズ株式会社の藤原博志氏、秋吉賢治氏、日本システムウエア株式会社の石田好延氏、阪本貴仁氏に感謝致します。

ご清聴、ありがとうございました。