2020/12/02
第32回コンピュータシステム・シンポジウム（ComSys2020）

低遅延不揮発性メモリを用いたリソース分離型分散データストアとその応用

日本電気株式会社バイオメトリクス研究所主幹研究員
大阪大学サイバーメディアセンター招へい教授　吉川 隆士

合同会社リトルウイング代表社員/株式会社Diagence代表取締役社長　菅 真樹
kan@lwing.tech,kan@dia-gence.com

この発表の成果の一部は、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（JPNP16007）の結果得られたものです。

NEDO IoT PJ: 高速ストレージクラスメモリを用いた極低消費電力ヘテロジニアス分散ストレージサーバーシステムの研究開発

なぜ開発するのか

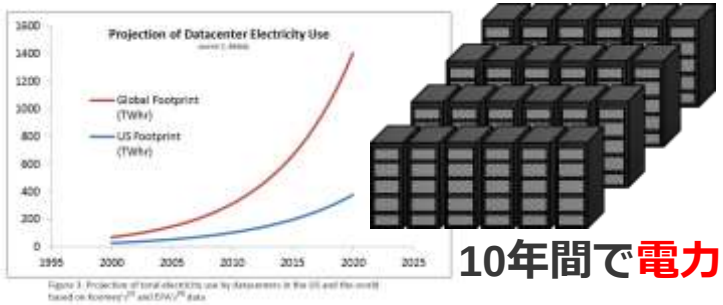
東京大学・東京工業大学・富士通・NEC（H28~）

【リアルタイム応答の多様なアプリケーションの出現】

インダストリ4.0

自動運転車

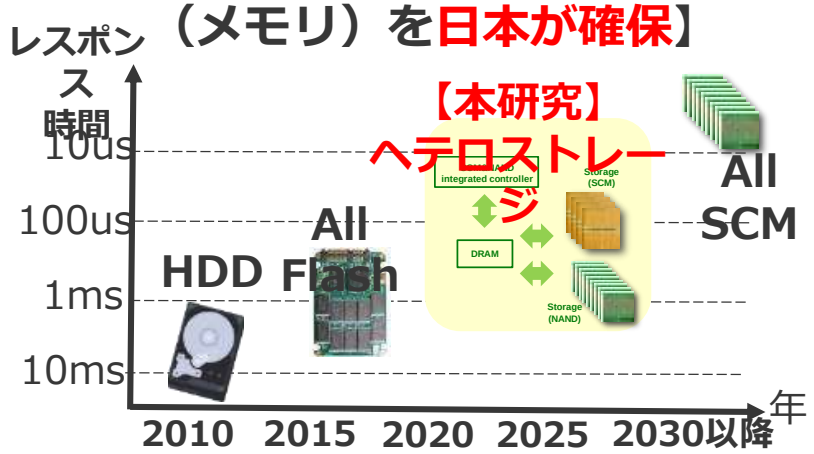
【データセンタの電力増大】



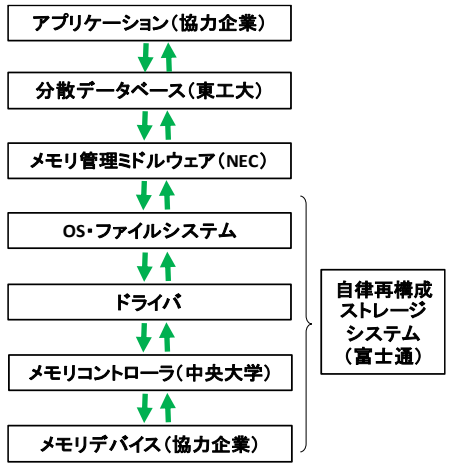
10年間で電力5倍

【広範な産業・社会応用の基盤技術】

【ITの根幹の基盤技術（メモリ）を日本が確保】



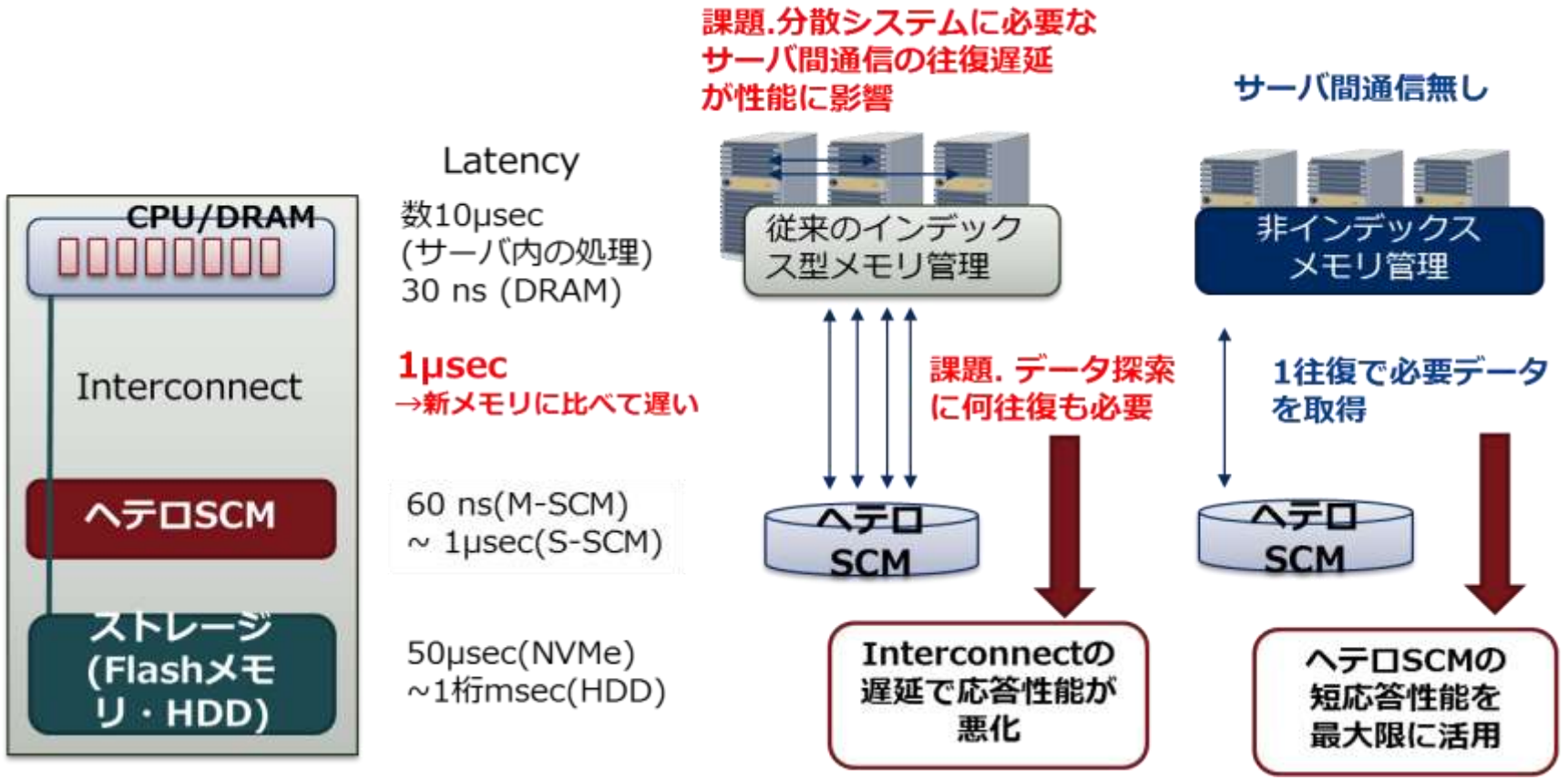
【多くの分野・階層の連携が必要】



RDStore: Resource Disaggregated Storage
高速次世代不揮発性メモリを活かした分散データストア

RDStore：次世代不揮発性メモリ+リソース分離向け分散データストアの研究

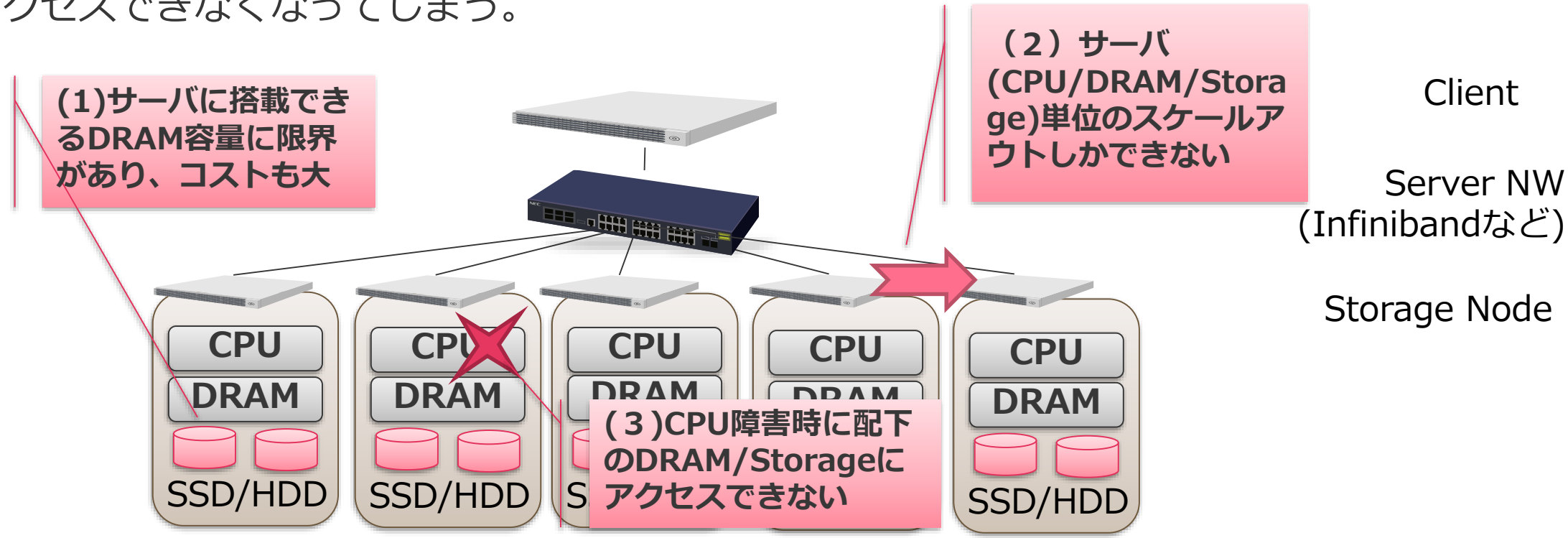
RDStore: Resource Disaggregated Storage
NVMe over ExpEther上のSoftware Defined Storage



”ExpEtherとNVMeデバイスを活用した分散ストレージRDStoreの紹介” (SNIA-J エクストリームストレージ研究会, 2016)
“リソース分離アーキテクチャにおける分散キーバリューストア”(SWoPP2015)より引用

現在のクラスタ型インメモリKVSの課題

- 近年インメモリ型KVS(NoSQL)のような高速・低レイテンシな共有データアクセスを必要とするデータストアが様々な領域で利用されています.
  - Webサービス，機械学習処理（Parameter Serverなど), …etc.
- **（インメモリ）分散KVSの３つの課題**
  1. **コスト対容量：**DRAMのコスト、サーバ単位のスケールアウトによりコスト増
  2. **スケーラビリティ：**不足リソースに関係なく、サーバ単位のスケールアウト
  3. **耐障害性：**故障するコンポーネントと関係なく、サーバのDRAM/記憶デバイスすべてにアクセスできなくなってしまう。



本研究の課題と解決アプローチ

リソース分離アーキテクチャと高速不揮発性デバイスを活用した、新しい分散KVSシステムにより３つの課題を解決する

（１）コスト対容量

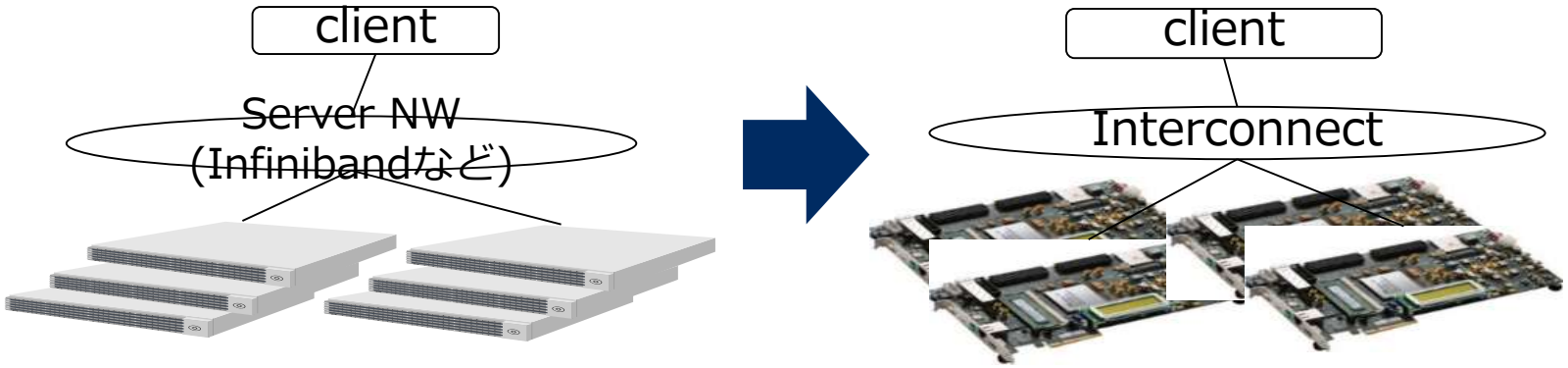
- **サーバ＋DRAM**の代わりに**NVMe Flashデバイスを直接拡張する**（さらなる高速化のために次世代不揮発性記憶デバイスを用いる方向性もありえる）

（２）スケーラビリティ

- **リソース分離アーキテクチャの採用**による，サーバ（CPU）とデバイスを個別にスケールアウト
- サーバ通信レスの排他制御

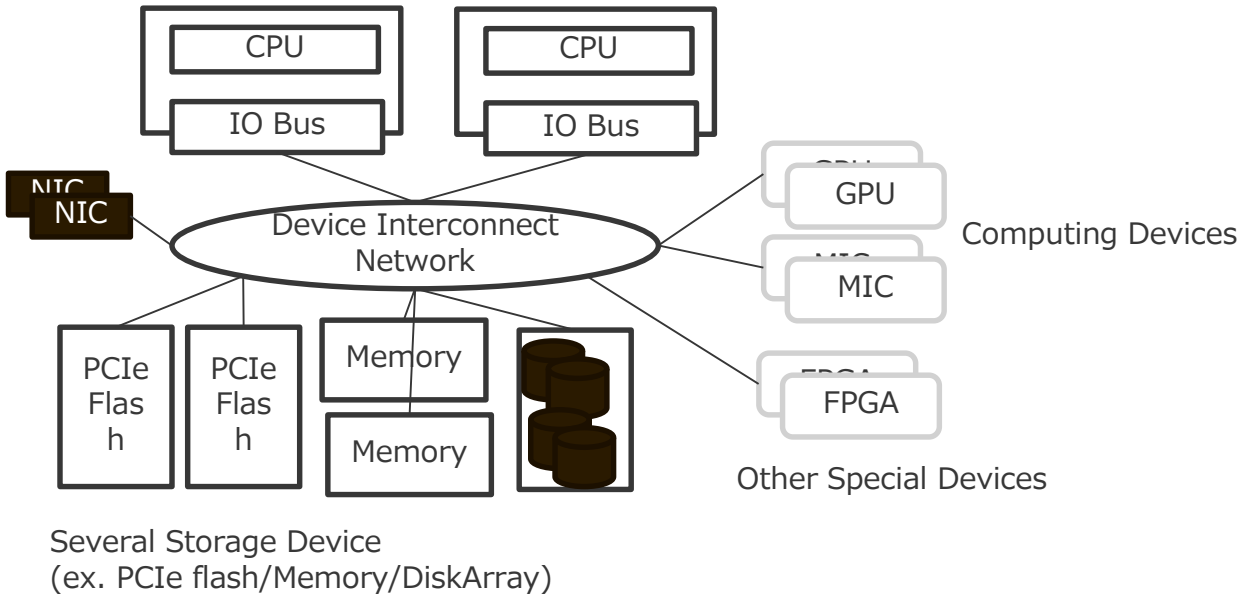
（３）耐障害性

- **リソース分離アーキテクチャの採用**による，サーバ障害とデバイス障害の分離



【参考】リソース分離アーキテクチャ

- IO**リソース**（デバイス）をサーバから**分離**し、CPUはインターコネクトNWを経由してIOリソースを利用。CPU間でIOリソースを共有。
  - 代表例として**Intel Rack Scale Architecture**があげられる。
- 下記のトレンドから構想されている。
  - データセンタの高密度化，省電力化の要求
    - 高密度なマイクロサーバ製品（HP,AMDなど）
  - IOデバイスの性能・機能が強化（PCI-e Flash, GPU, FPGAなど）



リソース分離アーキテクチャの特徴：

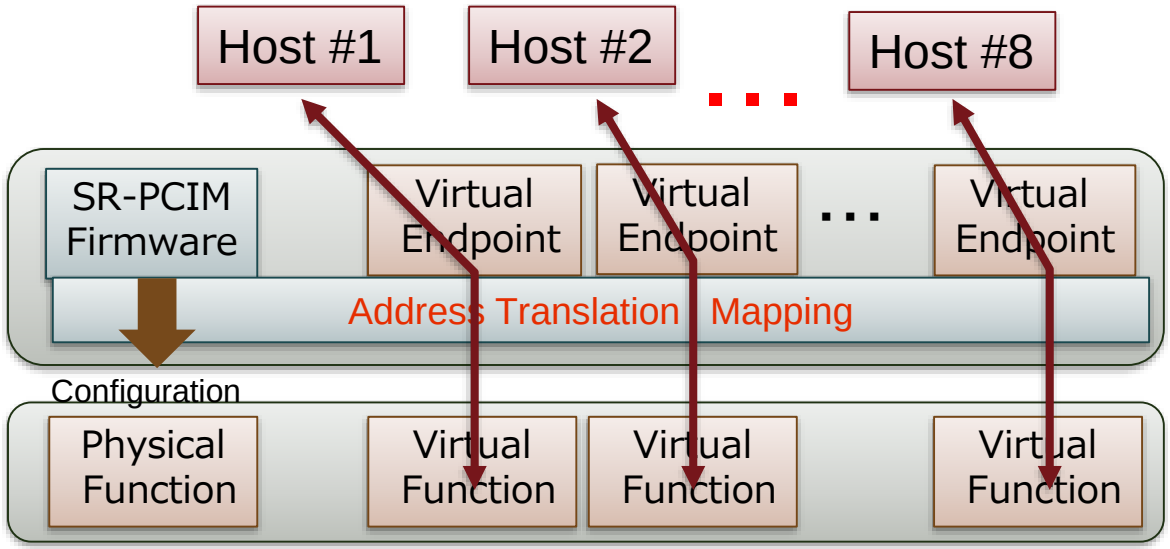
（１）デバイス単位でシステムを容易に拡張できる

（２）CPUとIOデバイス障害の分離

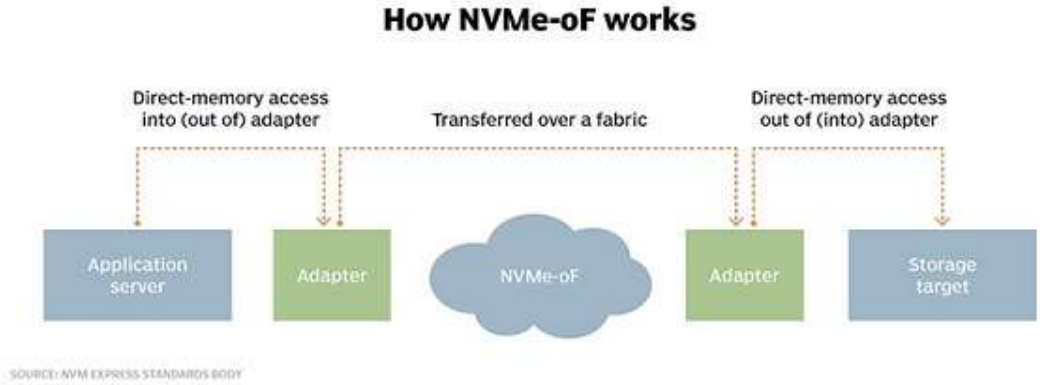
（３）デバイスをホスト間で共有できる

共有ストレージの実現方法と低レイテンシの活用

PCIe over Ethernet(ExpEther)



NVMe over Fabrics



Ethernet/InfinibandなどのNW越しにNVMeアクセスする規格

一部製品が出てきている段階

東芝、PureStorage(DirectFlash)、NVMesh(Excelero)、E8(2019/7/31にAmazonに買収）、WestanDegital(OpenFlex)

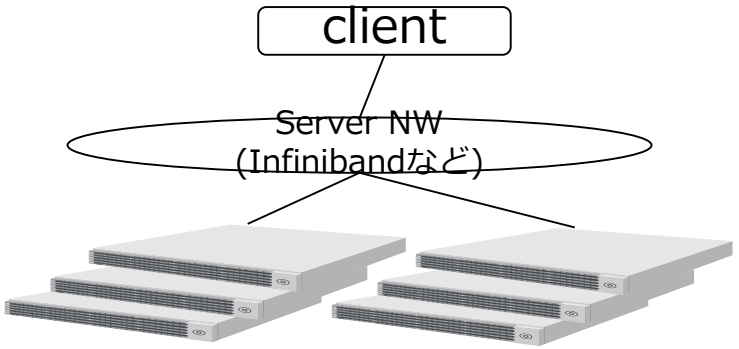
双方とも通信オーバヘッドは数us~数10us.
高速不揮発性メモリの低遅延（現在は10usオーダ、将来1usオーダが期待される）

本研究では、オーバヘッドを極力少なくするための、軽量ソフトウェアストレージを設計
各クライアントが直接ExpEtherやNVMeOFで接続されたNVMeデバイスを利用

RDStoreのターゲット(１)

DRAM容量はスペース・コスト面で限界
安価なHW構成では，DRAMレベルの性能をそもそも引き出せていない

- しかし，近年のデータ量増大に対し**膨大な容量**を必要とされ，インメモリ型KVSを実現する計算機クラスタでは膨大な数の計算機が必要になり**非常に高価なHWコスト**となります.
  - また，高性能なシステムを実現するには，RDMAなどが利用可能な高速なサーバ間NW(Infiniband, 10G/40G Ethernet)などが必要で，更に高価
  - **DBMS/NoSQL，KernelレイヤのSWおよびNWが性能を決定**
    - MICA[H.Lim, NSDI2014],RamCloud[D.Ongaro, SOSP2011]などのDRAM KVS＋高価なNWでも**1～２桁μsecの遅延**
    - OSS(Riak)+安価なNW構成ではインメモリ動作では**msの遅延**

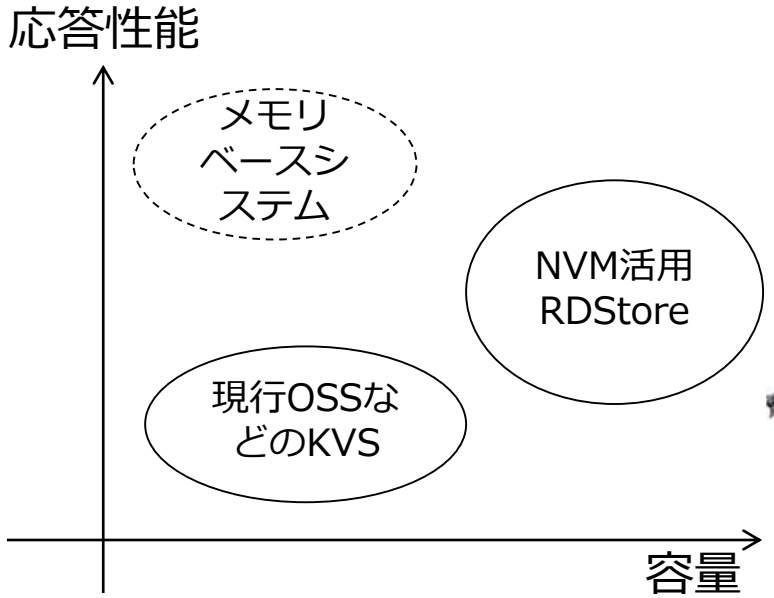


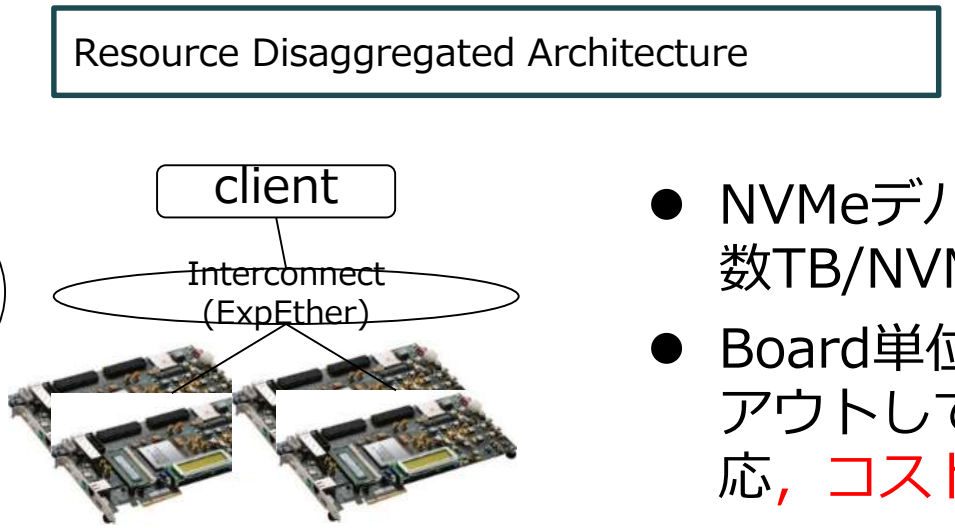
- 当時のState of the art研究(RamCloud, MICAなど)で10~60μsec(8bytes)程度のレイテンシ
- サーバ搭載メモリ容量(256GB-1TB/server程度)で容量が制約される.
- サーバ単位のスケールアウトのためコスト増加

RDStoreのターゲット（２）

**高速不揮発性メモリをうまく活用することで，
コスト対容量に優れた分散KVS/NoSQLが実現可能では？**

- 一方，flashなどの不揮発性メモリを活用したPCI-e SSD(Fusion IO, NVMexpressなど)は近年普及が進み，コスト対効果に優れたストレージアクセスが可能
  - NVMexpressは，デバイスアクセスレベルで**２桁μsecオーダ[Intel P3700]のカタログスペックは4KB SEQ READで20μsec**の性能
  - 次世代不揮発性メモリ（PRAM, MRAMなど)では**約100nsec~１桁μsec**が期待される
  - **→デバイスアクセス性能のポテンシャルとしては，通常のインメモリNoSQL程度の性能はFlashベースでも充分実現可能では？**



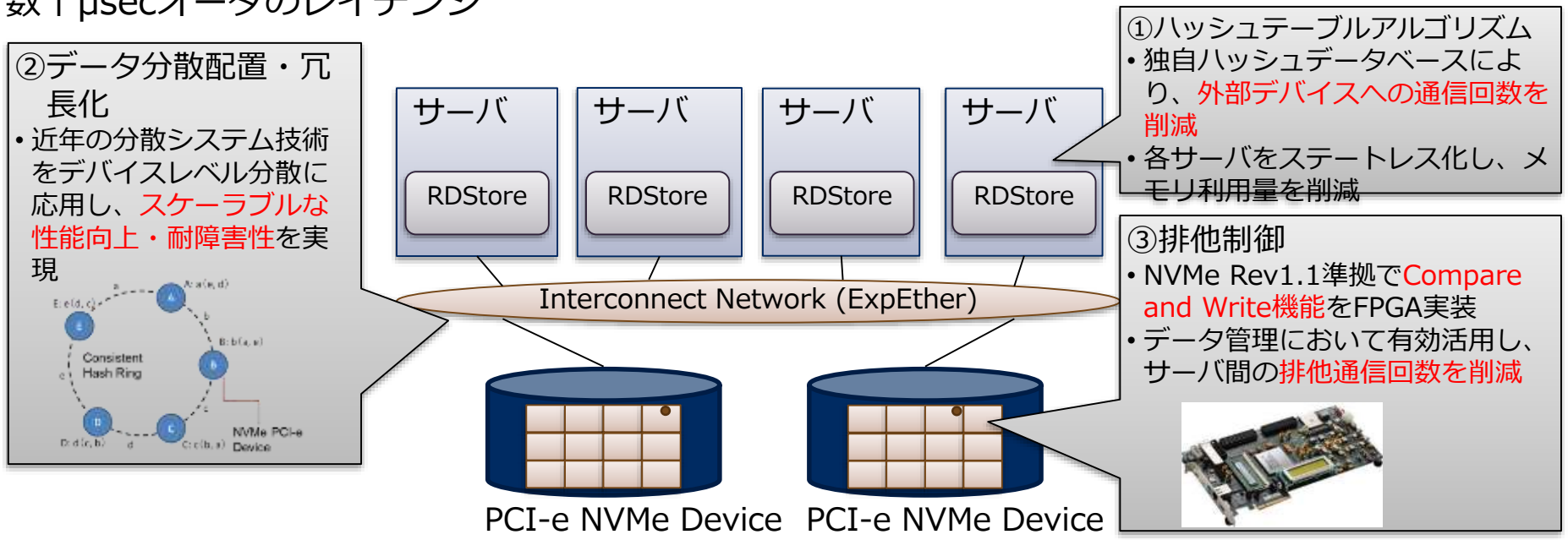


- NVMeデバイス容量は数TB/NVMe Board.
- Board単位でスケールアウトして大容量対応，コスト低減

RDStoreの特徴と効果

サーバ・デバイスを自在に組合せ、柔軟性・耐障害性・性能を兼ね備えた分散データストア（KVS）を実現

- サーバ・I/Oデバイスをそれぞれスケールアウト可能なソフトウェア分散ストレージシステム(Software Defined Storage)
  - 各サーバが全ての最新データに対し，高速にアクセス可能
  - ExpEther（弊社独自技術）によって高速なPCI-e 記憶デバイスを低レイテンシで共有
- 可用性，耐障害性，柔軟なスケールアウト
  - メタデータサーバなどの特別なノードが無く，単一障害点を排除．各サーバはステートレスのためサーバ障害発生時にデータロス0、サーバ増減も瞬時に可能
  - 複数PCI-e記憶デバイスに複製を保持することで耐障害性を実現，記憶デバイスの動的な増減も可能
- KVSとして数十μsecオーダのレイテンシ



アーキテクチャ概略

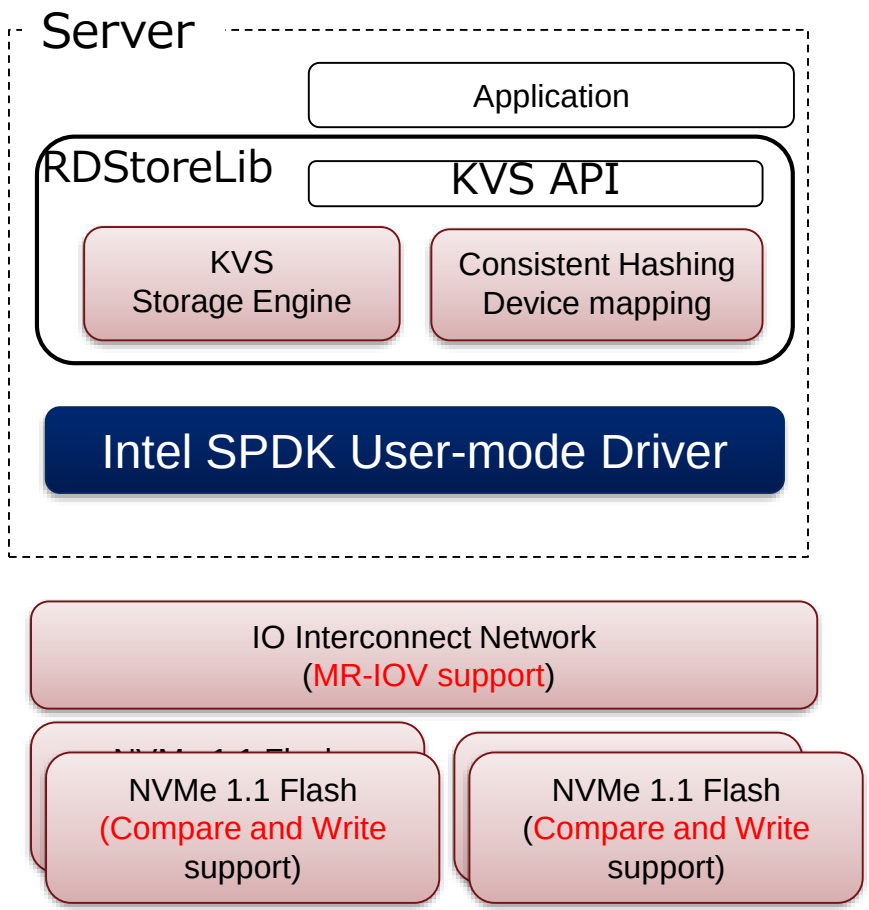
ソフト・ハード両方に工夫を入れ込んだアーキテクチャ

ハードウェアアーキテクチャ

- **リソース分離アーキテクチャ**
  - インターコネクトネットワークを介して**NVMeデバイスを共有**（MR-IOV）
- **NVMeデバイス側でCompare and Write機能**をサポート(NVMe 1.1準拠)

ソフトウェアアーキテクチャ

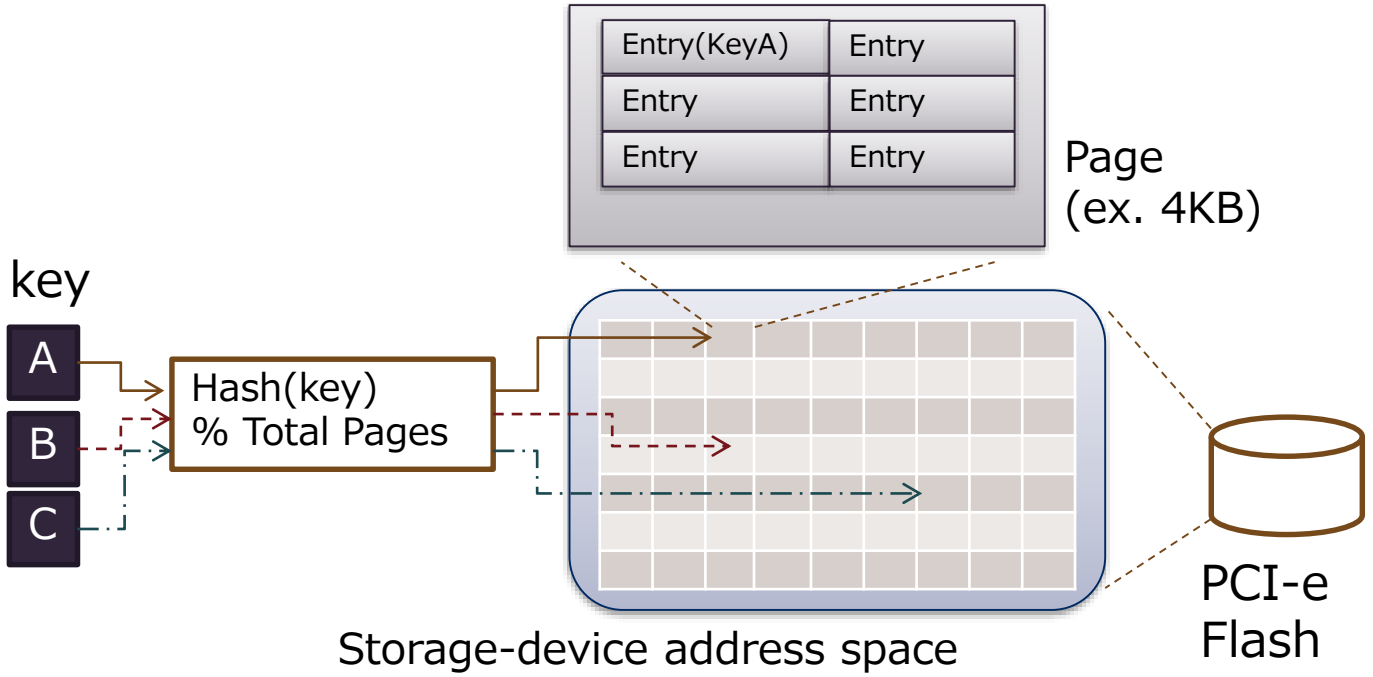
- **ステートレスなKVSストレージエンジン**
  - サーバメモリにデータ保持管理のためのメタデータを持たない
- **可能な限りサーバ通信レス**
  - 排他制御にNVMeデバイス機能を活用
- デバイス間のデータ分散，複製制御はConsistent Hashingベースの手法で分散
- SPDK User-mode Driverでファイルシステムレスでデバイスを直接利用
  - ファイルシステムやカーネル・ユーザ空間のメモリコピーオーバヘッド回避



①ハッシュテーブルアルゴリズム

ホストメモリにメタデータを保持しないことにより，サーバ間通信を排除

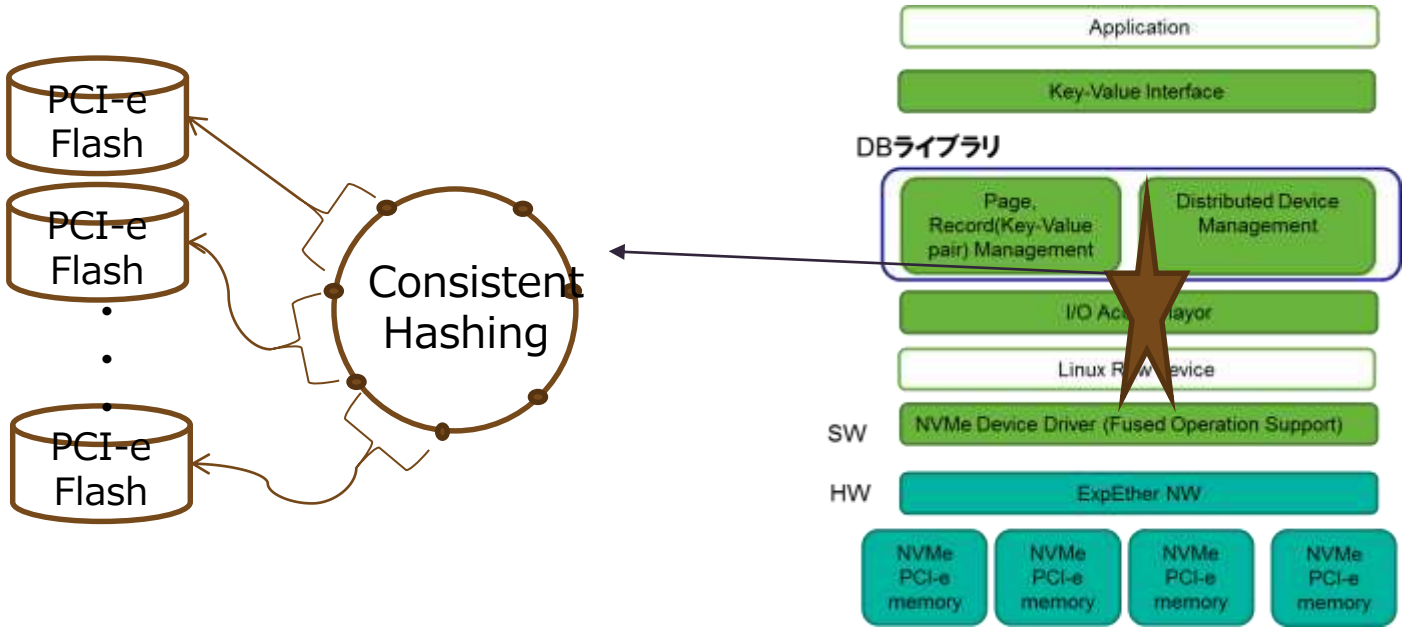
- ステートレスなハッシュテーブル
  - オープンアドレス法と類似した方法で実現
- Keyとハッシュ値から決定されるページ（４KB)に，複数のKey/Valueエントリを格納.
  - ホストからはページ（４KB)ブロックアクセスで取得し，ホストで該当データを取得
  - 更新もページ単位のCompare and Writeで実現（同一Pageのアクセス以外はアクセス競合無し）
  - ページからあふれた場合には再ハッシュ.
- 各ホストが独立して，静的な情報（ハッシュ関数と総Page数）だけでデータの配置場所を算出できる



②データ分散配置・複製

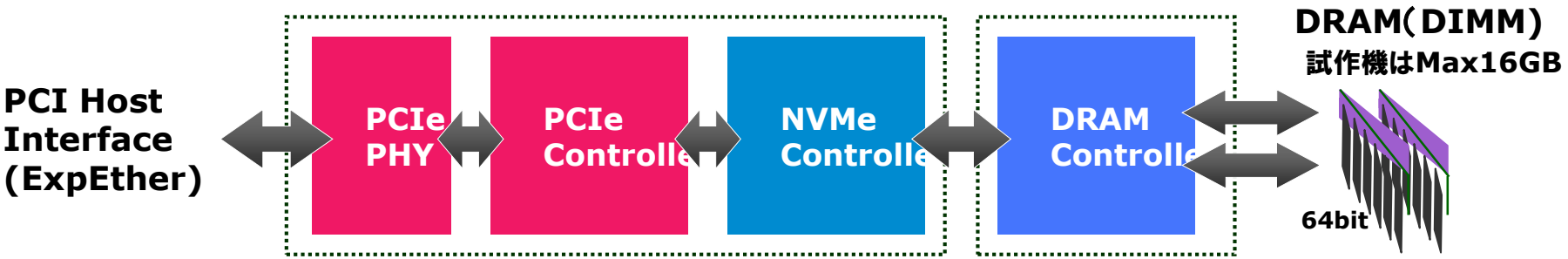
クラスタ向け分散ストレージのスキームをデバイス単位に適用

- Consistent Hashingベースの方法で，デバイスへのデータ分散配置，レプリカ作成先を決定.
  - Amazon Dynamoに類似．サーバでなく，相手がNVMeデバイス.
- Consistency保証のためUpdate命令はNVMe Rev1.1 Fused operations機能により，Compare and Writeで行う
  - 複数のデバイスに書き込み・読み込みを行うことで高信頼性を確保
  - 不整合時は多数決で解決



③排他制御：NVMe 1.1以降準拠のCompare and Write機能を試作して検証

- 現状，SR-IOVサポートし，NVMe1.1のoption機能であるFusedOperationをサポートするデバイスは無いため自作.
- ボード
  - Xilinx社製評価ボード VC709
- FPGA
  - PCI Express R2.0 < 5.0Gbps x 8 lane >
    SR-IOVサポートし、MR-EE(V2.0a)とのインターワークを考慮
    PF数：1　VF数：8 （Virtex-7制限により当面はmax 6VFs）
  - NVM Express
    PCI SSD規格であるNVM Express 1.0d をサポート
  - DRAM Controller
    メモリ読み書き制御 （DDR3-1066 インタフェース ECCなし）
  - CAS命令（Fused Operation, NVMe1.1)実装



NVMe 1.1に基づくFused Operationを用いた排他

サーバ間で2Phase commitのような手法をとった場合と比較して，サーバ間通信の削減が期待できる.

2 Phase Commit

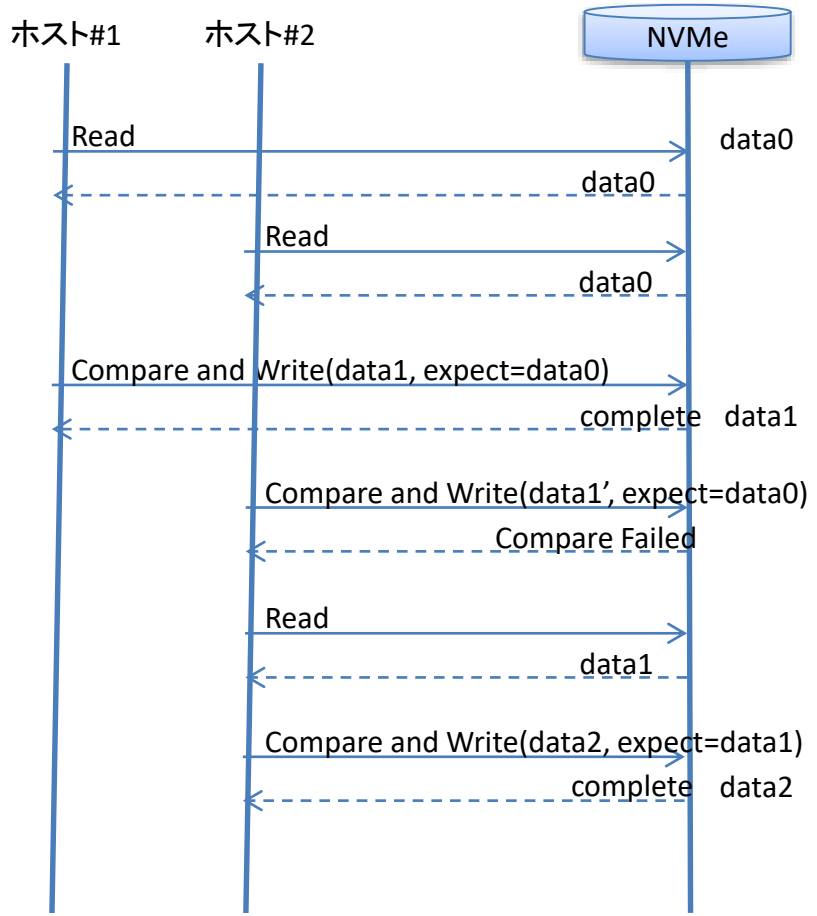
- NW通信回数が多い
  - Read,Request-to-Prepare,Write,Commit
  - サーバ間通信が頻発するため，Infinibandなどの低遅延NWを必要とする
- Coordinator負荷を考慮する必要がある
  - CPUコストなども必要である.
  - Coordinator分散・障害など.
    - Coordinatorログ分の書き込みコストも増える

Key毎に担当ノードを割り当てる方法

- サーバ間通信が１hop余計に必要.

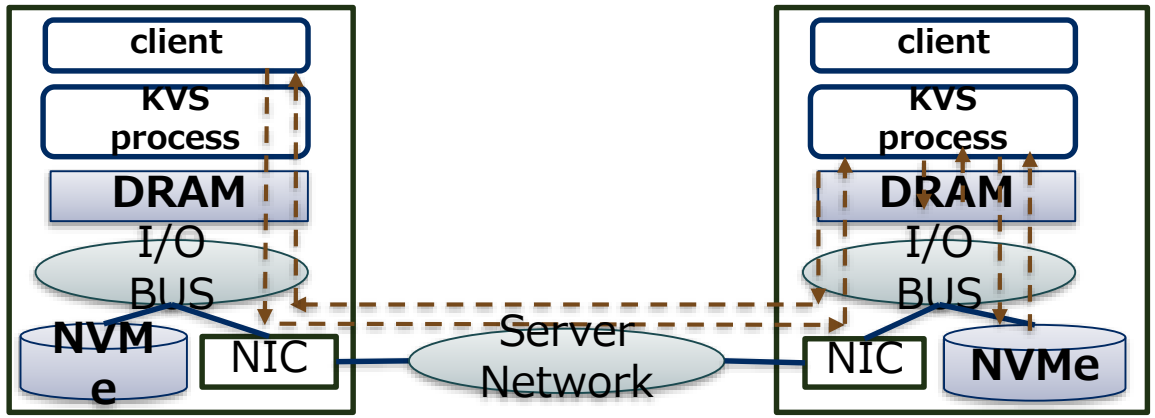
提案手法（CAS利用)

- サーバ間通信は必ずしも低遅延のNWを必要としない
- 通信回数そのものが少ない（2回）



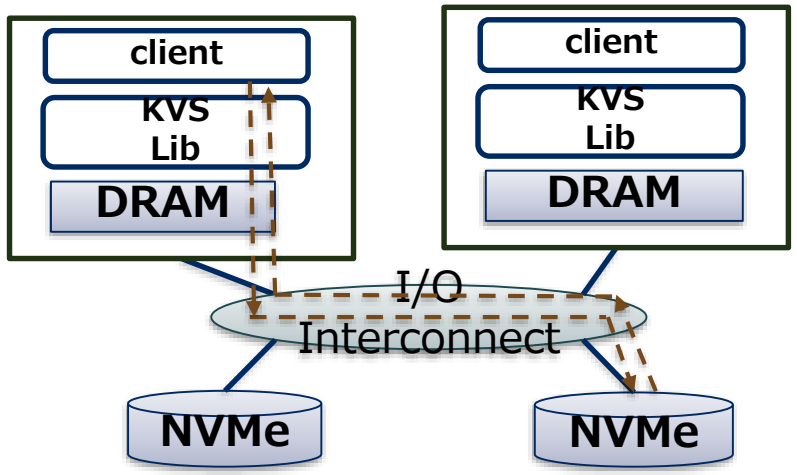
評価実験（１）：Memcachedとの比較

RDStoreは，Memcachedのようなサーバ型KVSの代わりに，NVMeデバイスをI/O Interconnect経由で接続．この差について比較評価を行う．



←従来(Memcached)

- データストアが計算機で出来ているため、サーバ間のネットワーク処理、CPU内での処理、DRAMおよびNVMeへのアクセス処理が、データを応答するために必要

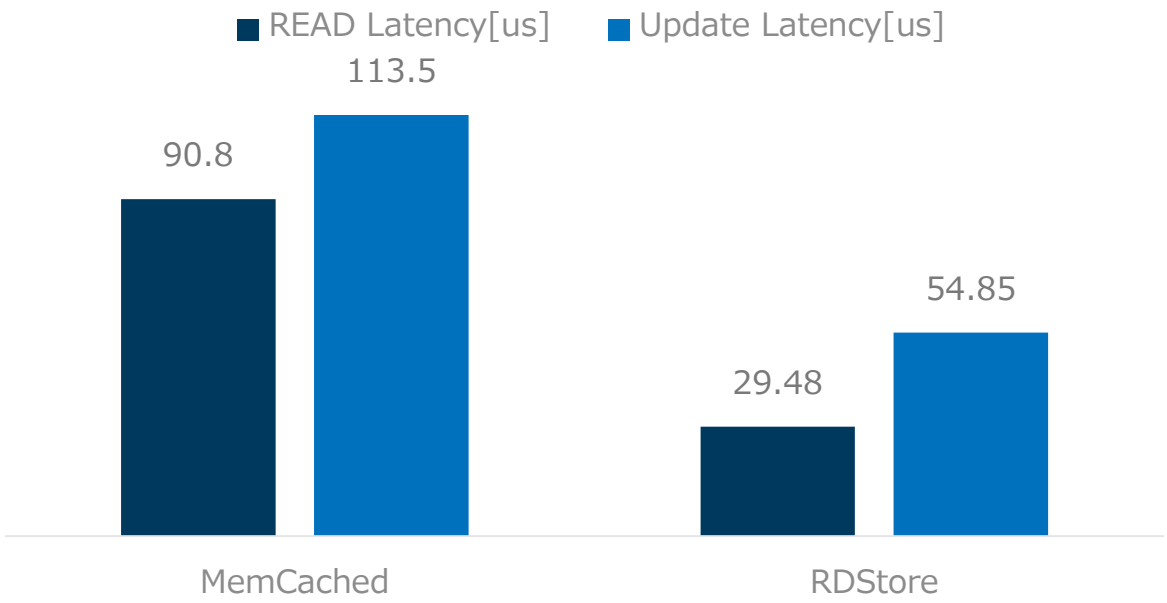


←提案（RDStore)

- インターコネクトネットワークを共有し、直接デバイスに対してアクセスするため、クライアント計算機がDMAコールをするだけでよい

RDStore: 評価結果

- **Intel Optaneを利用**
  - **次世代不揮発性メモリとソフト/アーキの工夫で、リモートのサーバ＋DRAM以上の性能を実現可能**
- Memcached: Data Access to DRAM
  - 1 Server(client)-1 server(Memcached), connected through 10Gbit Ethernet
- RDStore: Data Access to External NVMe(Intel Optane)
  - 1 server-1 device, 40Gbit Ethernet (ExpEther: PCI Express over Ethernet)



Benchmark Application:
YCSB 4KB Random Read/Write

まとめ

- ネットワーク経由で接続する次世代不揮発性メモリの活用を目指した分散データストア（RDStore）の紹介をしました
- 将来高性能（低レイテンシ）なデバイスの出現を想定して、可能な限り低オーバヘッドになるように設計・試作し性能評価しました

この発表の成果の一部は、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（JPNP16007）の結果得られたものです。