SNIA日本支部主催
「2024年度第1回最新技術動向講演会」

SDC 2024 参加報告

Ryosuke Tatsumi

注意事項

- 本スライドにて掲載している資料はSDC2024に参加登録した方のみアクセス可能な資料です (https://www.sniadeveloper.org/conference/)
- SNIA日本支部会員企業向けの報告に限ってのみSDCの主催であるSNIAより公開の許可を得ています
- SNIA日本支部会員企業以外への本スライドの配布を禁止します
- SDC2024の各講演については後日YouTubeにて公開予定です(一部のみ) (https://www.youtube.com/user/SNIAVideo)

目次

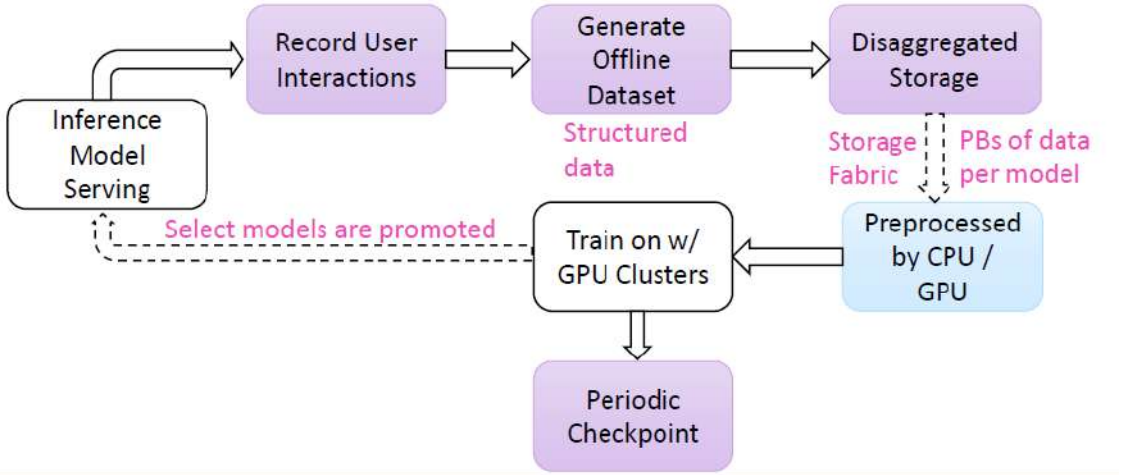
- What is the Role of Flash in Data Ingestion within the AI Pipeline?
- Storage for AI 101
- NVMe over CXL is much more than Just an SSD
- Accelerating GPU Server Access to Network Attached Disaggregated Storage Using Data Processing Unit

What is the Role of Flash in Data Ingestion within the AI Pipeline?

What is the Role of Flash in Data Ingestion within the AI Pipeline?

S. Sankaranarayanan, S. Rajgopa, S. Somandepalli (Micron)

- Deep Learning Recommendation Model のデータインジェクション時に求められるストレージの役割とベンチマーク
- Meta社のモデル：12trillion個のパラメータ、PBsのデータをトレーニングに使用。定期的なチェックポイント書き込み。
- トレンド：データセットのサイズは2年で2倍、必要なI/O帯域は4倍に増加。ストレージや前処理による電力消費が多い。
- 課題：ストレージの容量とIOスケーリングの両立、電力消費量削減



- Dataset storage size has grown 2x in 2 years
- IO bandwidth demand has grown 4x in 2 years
- A scalable architecture should meet storage and IO demands

- > 40% power is spent on Storage + Preprocess
- Less power for trainers ➔ inefficient use of compute
- Power is an important dimension in scaling

What is the Role of Flash in Data Ingestion within the AI Pipeline?

- HDDベースのFilesystem(Tectonic)のキャッシュとしてSSDを使用(Shift)
- データ前処理をInline化し、学習用のプロセッサに効率よくデータを供給する
- 容量とI/O性能のスケーリングを両立し、消費電力もHDDのみの場合に比べて、29%削減した

- Meta solves IO scaling with an SSD caching layer over Tectonic (HDD-based) called Shift
- Data in Shift is not durable, and Shift leverages underlying Meta's CacheLib
- Tectonic-Shift saves 29% of power relative to using HDDs alone

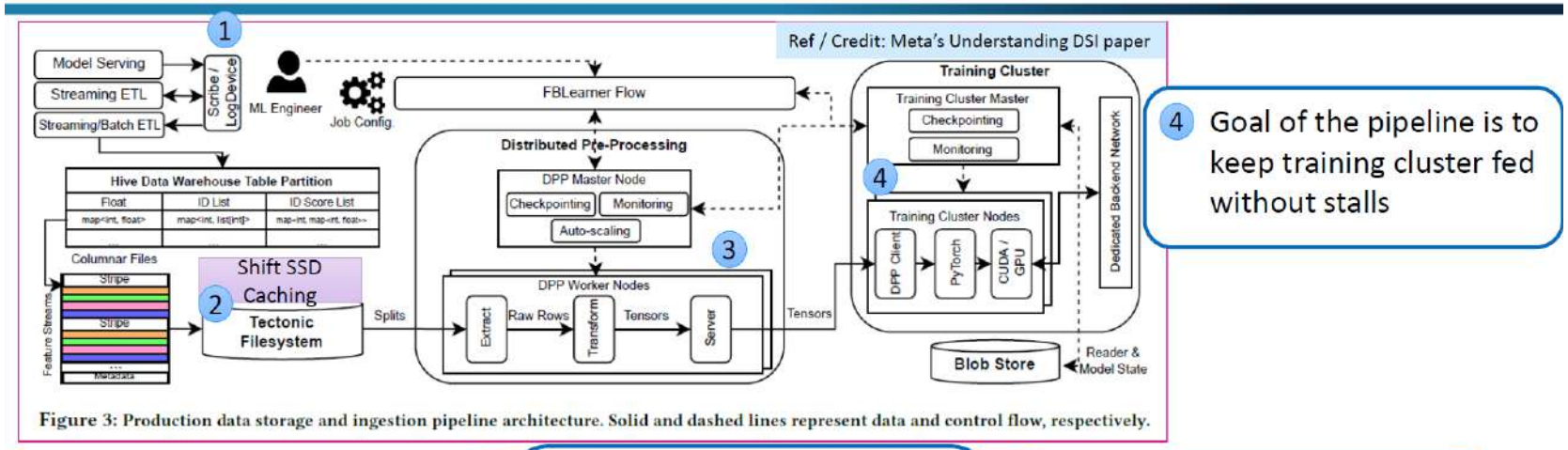


Figure 3: Production data storage and ingestion pipeline architecture. Solid and dashed lines represent data and control flow, respectively.

4 Goal of the pipeline is to keep training cluster fed without stalls

Meta's DLRM Pipeline
Data stored in columnar format in a distributed filesystem
Significant portion of DLRM data is read out of SSD caching layer
Inline preprocessing of data includes decompression, decryption, transformation, and data filtering
Preprocessing is self-contained within a mini-batch operated by a DPP worker

1 User interactions are logged → growth in training dataset

2 Data compressed and stored in chunk store
- SSD Cache serve AI data → leverages data locality across RM jobs → Help tackle scaling challenge

3 Pre-processing is inline
- Splits are independent and self-contained work items managed by preprocessing worker nodes

What is the Role of Flash in Data Ingestion within the AI Pipeline?

- DLRMのストレージワークロードの解析
- 大サイズブロック(512KB以上)のシーケンシャルアクセスが多い
- 前処理ではライトが発生、トレーニングでは読み込み。ライトはほぼシーケンシャル。リードは部分的にシーケンシャル
- 時折、大きなI/Oが発生。需要を満たすにはフラッシュストレージのような高速なストレージが必要

DLRM Storage Trace Analysis - Results

Storage Trace	DLRM Preprocessing w/ GPU	DLRM Preprocessing w/ CPU	DLRM Training on GPU
Experimental Setup	(1) Preprocessed with 8 GPUs	Preprocessed with 2 64-core CPUs	Trained with 8 GPUs: batch size = 8K, # of batches = 64014
What's in storage?	Criteo click dataset in Gen. 4 drive	Criteo click dataset in Gen. 4 drive	Preprocessed dataset in 2 Gen. 4 drives (RAID0)
Run time (secs)	1900	5181	445
% Read Volume (#)	72 (7.7M)	55 (17M)	100 (469K)
Perf. (MBpS)	(4) $1500\text{-}6000_{Read}$ 3000_{Write}	$500\text{-}6000_{Read}$ $1800\text{-}3000_{Write}$	454_{Read}
QD	$250_{mean} \rightarrow 10_{mean}$	1-11	4-5
Read Payload (KB)	(2) $512_{90\%}$	$512_{89\%}$	$512_{71\%}$
Read – Sequential Volume %	(4) 43-55	50-90 (in large portions of the trace)	68
Write Payload (KB)	(2) $1280_{65\%}$	$1280_{40\%}$	N/A
Write – Sequential Volume %	(3) 85-95	90-99 (in large portions of the trace)	N/A

1. Preprocessing is an offline task
2. Read and write payloads are large
3. Writes during preprocessing are sequential
4. Reads on certain portions of the workload are highly sequential
5. Performance demands from storage are time-variant

SDC 24

SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

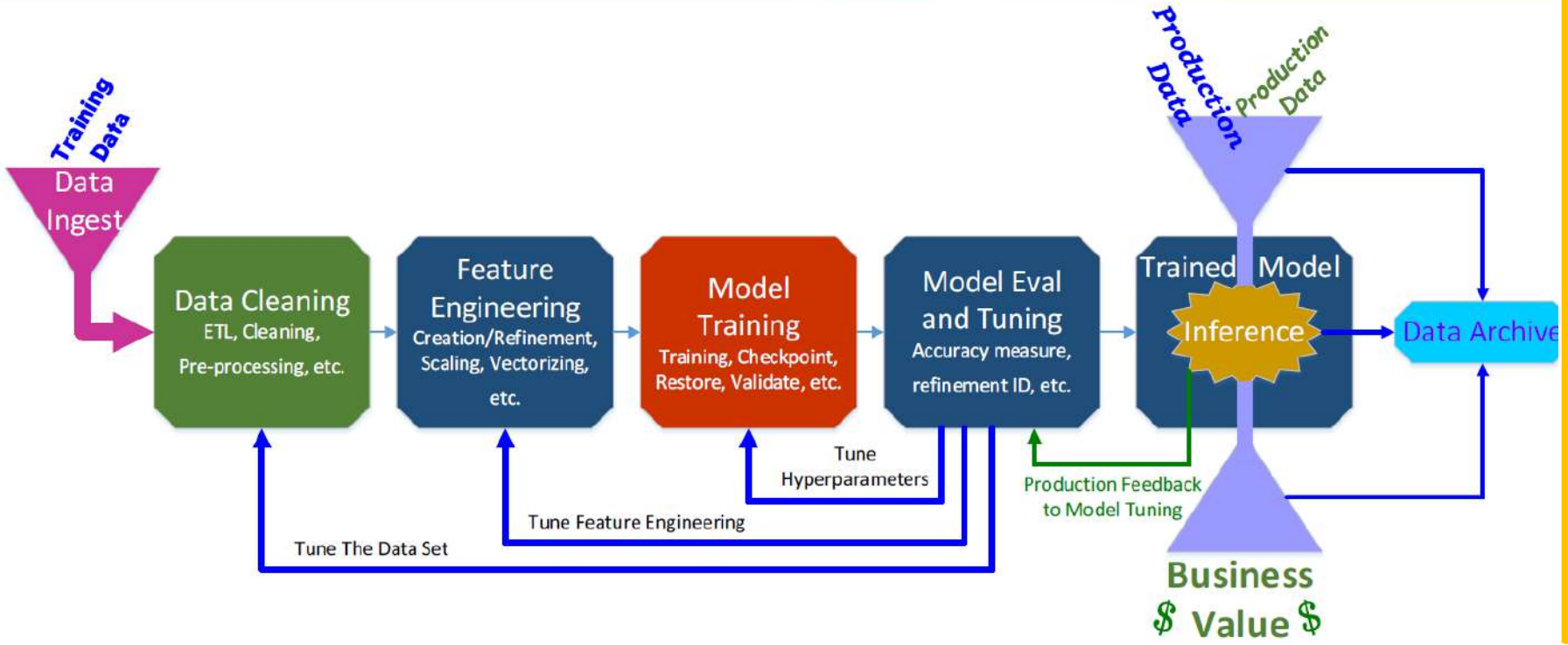
SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

C. Ballard (HPE), C. Carlson (AMD)

- AIワークロードのストレージ要件のポイントを整理
- トレーニングに使用されるGPUの利用率を最大化するために、各フェーズで必要なストレージ要件が異なる
- 1種類のストレージではなく、用途に応じて適切なストレージを提供するプラットフォームが必要

Storage Phases of AI one perspective



Extensive use of:

- Data Scientists
- Compute Resources
- Storage Resources
- GPU Resources

With a goal of:

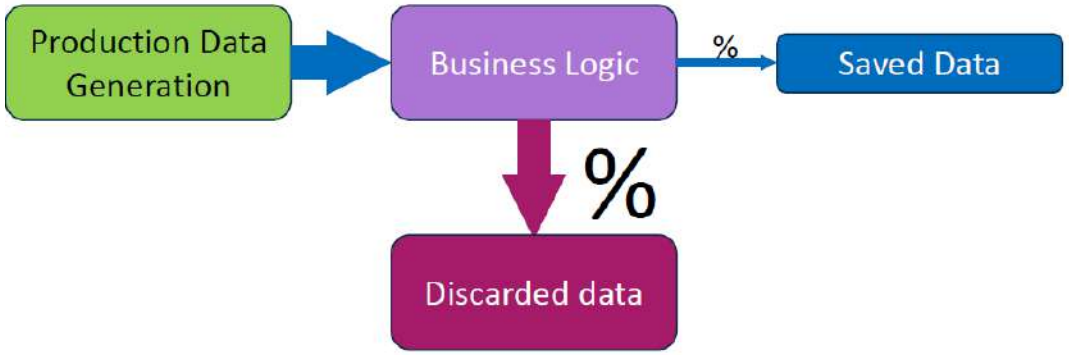
- Generating a Trained Model

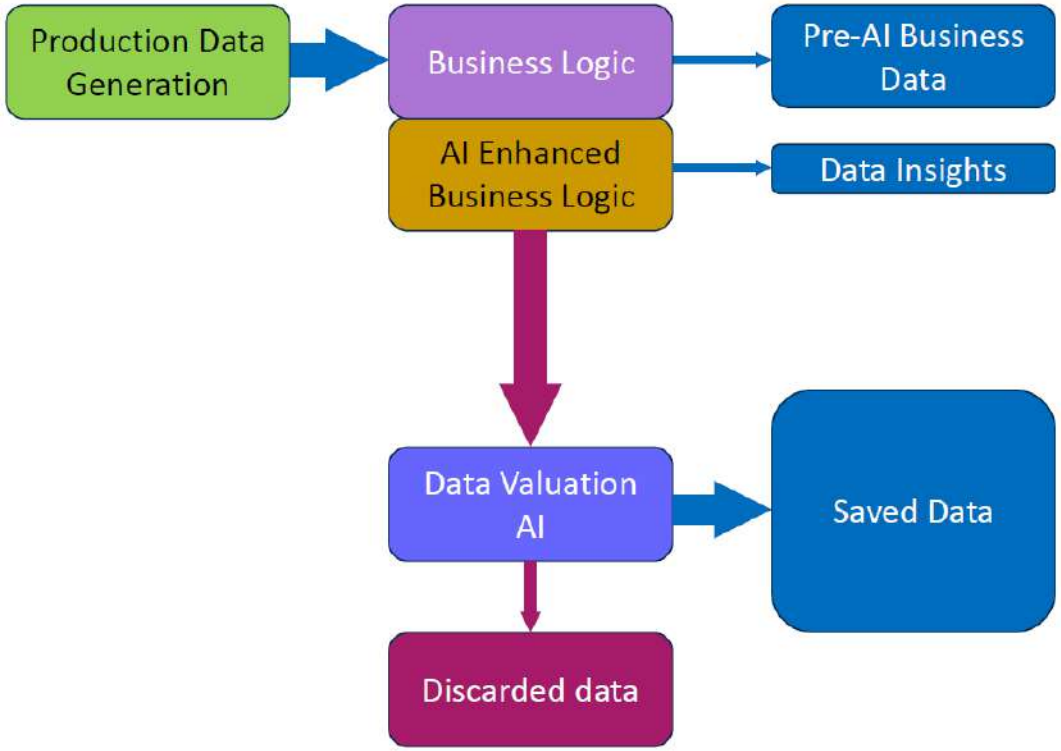
Not generating business value unless your business is selling foundational models (e.g., LLMs)

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- AIをビジネスに活用するためには、Data Ingestionが大事
- ある企業の例：AIを活用する以前、生成されたデータをビジネスロジックに流し、ビジネス指標を図るKPIのみを保存
- ビジネスロジックのデータをAIを使って学習し、新たなビジネスチャンスを生み出すことに活用
- 保存されるデータ量や、AIワークロードのために転送するデータ量が爆発的に増えた
- ストレージの要件を見直す必要性

What their data ingest WAS before using AI

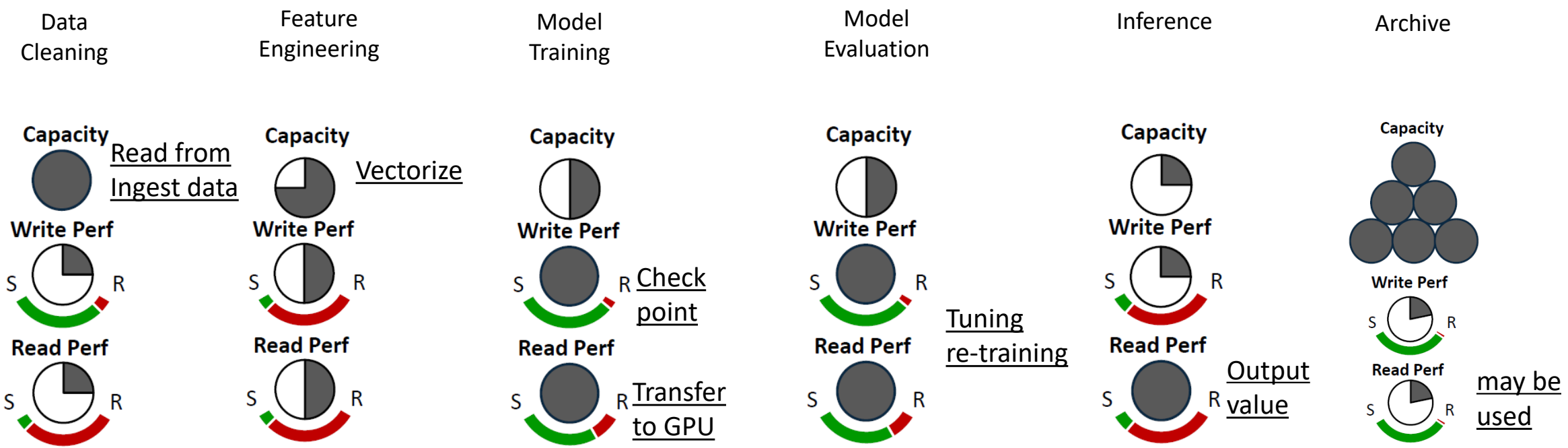




SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- 各パイプラインフェーズにおけるストレージワークロードの整理
- モデルのトレーニング時には、トレーニングデータの読み込みおよびチェックポイントの書き込みのために、高いストレージ性能が必要
- ビジネス価値創出に直結する推論では、高いランダムリード性能が要求される
- アーカイブデータは、膨大な量だが、後に別の学習のために再利用されるかもしれない



https://www.snia.org/educational-library/storage-requirements-ai-2024

SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- 様々なフェーズの要件に対応するAIワークロードの性能を測定するためのベンチマークが利用できる
- トレーニング、推論、ストレージワークロード
- 学習アルゴリズムのベンチマーク

Calculating Performance

- Benchmarking
  - Publicly available AI benchmarks are available through ML Commons
  - Multiple categories
    - MLPerf Training
    - MLPerf Inference
      - Mobile
      - Tiny
      - Datacenter
      - Edge
    - MLPerf Storage
    - AlgoPerf: Training Algorithms Benchmark Results

SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- アクセラレータ
  - SDXI：SNIAがまとめている新しいデータ転送方式の標準規格。データ転送に色んな機能を追加していく。
  - CS：SNIAとNVMeで進めているストレージデバイスと計算リソースを統合し、ストレージサイドでアプリケーションを動作させるための標準プラットフォーム。
  - GPU：並列演算処理によりAI等のワークロードを高速化。HBMと呼ばれる広帯域メモリを使用

Accelerators – SDXI

- SDXI is a standard data mover being developed by SNIA
- Future versions of SDXI are looking to p additional functions
  - Encryption/decryption
  - Compression/decompression

Accelerators – Computational Storag

- Computation Storage defined by both SNIA and NVMe
  - Open platform for adding computational functions to storage devices
  - Moves the computation closer to the data
  - Typical functions could be
    - Encryption/decryption
    - Compression/decompression
    - Data filtering
    - Data preparation for training

Accelerators - GPUs

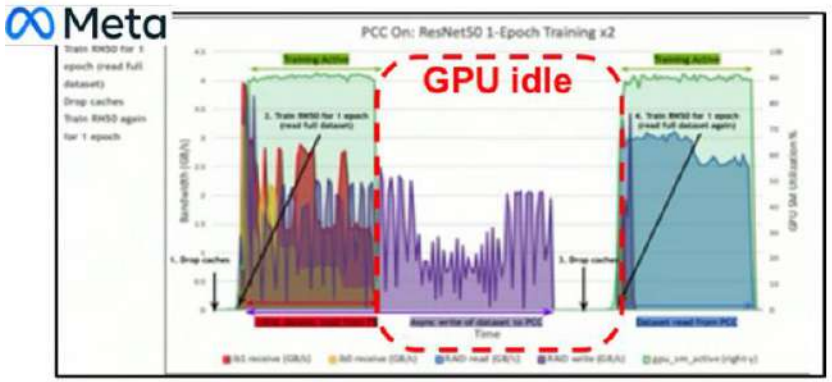
- Parallel operations
  - AI calculations can be made highly parallel
    - Typically they are multiple similar calculations acrross a matrix
    - This is the type of calculation that GPUs are designed to handle in a massively parallel fashion
    - CPUs typically can only do a single calculation at a time
  - Not only do parallel operations reduce the computation time dramatically, but they also make it more energy efficient
- HBM - High speed memory typically found on datacenter GPUs

SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- “システムの性能は最も遅い部分の性能で決まる”　遅い = ネットワーク、ストレージ
- GPUが遊んでいる = お金を浪費している
- Checkpoint時のデータ書き込みによって、GPUの稼働率が低下
- Checkpoint時にはNetworkに大きなトラフィックが発生

- Remember, you are only as fast as your slowest part
  - Due to inherent latency and device constraints… Network and storage components are often the slowest components in a system
  - Storage devices typically have slower access times
  - Networks are limited by latency
- The goal… Keep the GPUs fed!



Meta’s @scale Jun’24 Credit: NVidia

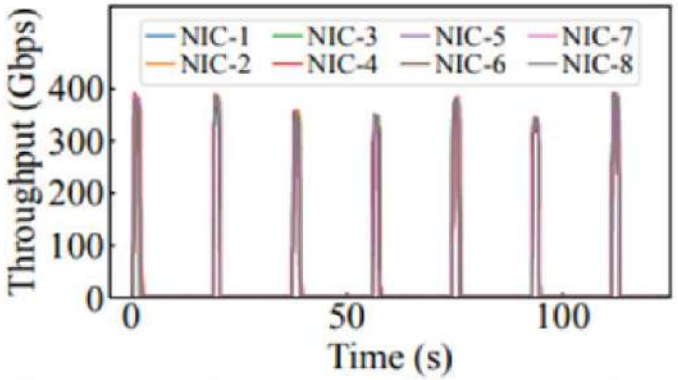
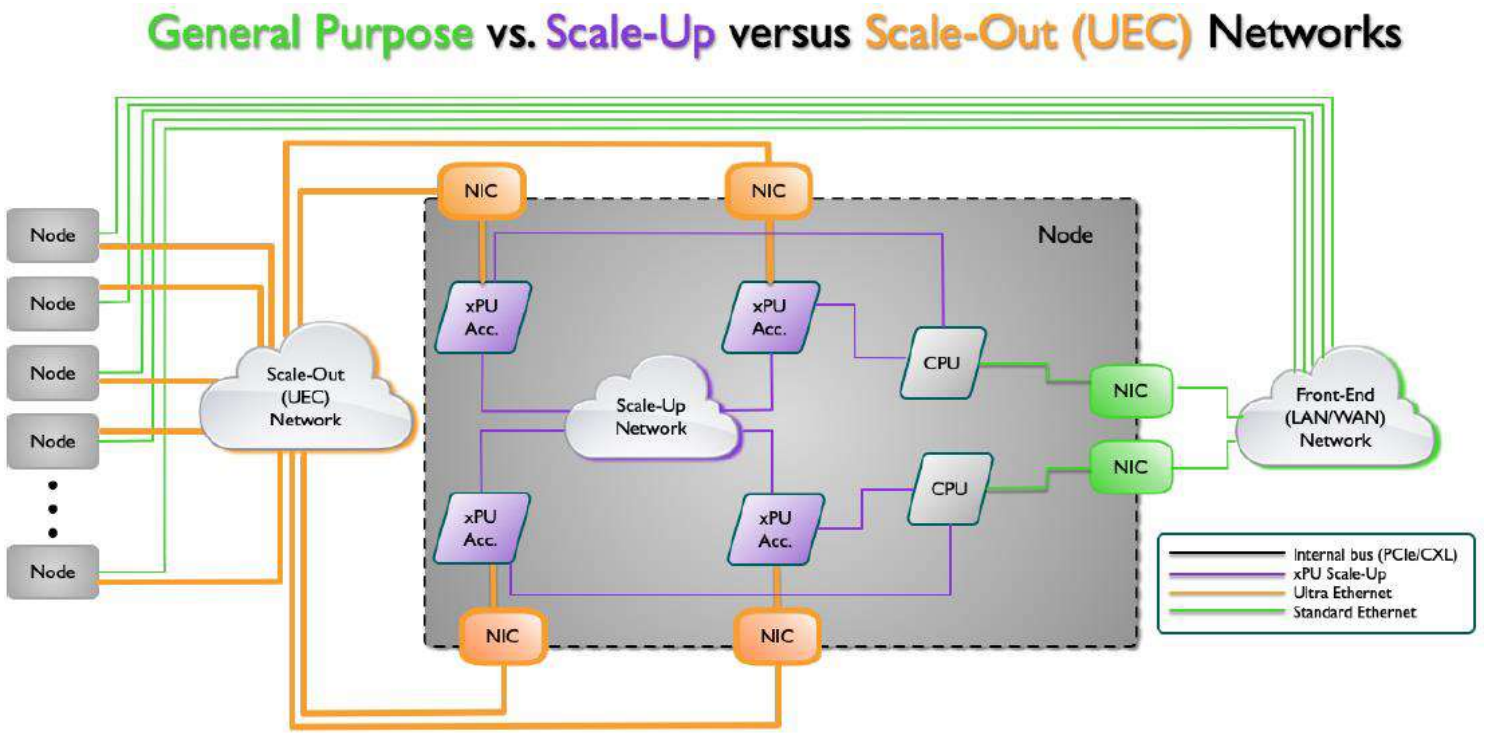


Figure 2: NIC egress traffic pattern during production model training.

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- スケーラブルで広帯域、低遅延なバックエンドネットワーク(Scale-Out Network)
- Ultra Ethernet consortium：Linux Foundationと提携して開発中のHPC/AIワークロード向けの高速ネットワーク技術
- 数100万ノードの拡張性、最新の輻輳制御、低遅延、エラー訂正、セキュリティなどの機能をもつ
- 年内に仕様完成の見通し



SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- AIワークロード向けの3種類のストレージタイプ（クラウド、オブジェクト、ブロック）
- ModelデータはObjectやCloudにいれるのが一般的
- Blockは低遅延が要求されるCheckpointの処理に向いている
- CXLのメモリプールによって、ストレージが抱える性能、スケーラビリティ、信頼性等の問題にアプローチできる

- Three types of storage typically used for AI
  - Cloud
  - Object
  - Block
- Model data (input and output) typically stored in the cloud or on object storage
- Block storage often (but not always) used for checkpointing
  - Low latency/high performance
- Additional Storage Functions could be provided by CXL attached memory pools
  - Allows a tiered memory where some, non-immediate use data, could be stored in a CXL pool
- CXL and other new memory architectures could provide relief to the existing memory bottleneck (or the “memory wall”)

SNIA

Storage for AI 101 A Primer on AI Workloads and Their Storage Requirements

- SNIAではIOトレースのデータを集めている
- AI向けのトレースはまだないがこれから拡充していきたい

SNIA Data pattern repository

- SNIA has an IO trace repository used extensively for research
  - The SNIA I/O Traces, Tools, and Analysis repository, IOTTA https://iotta.snia.org
- The repository does not yet have AI Storage workload traces
  - A gap SNIA would like to fill
- Please consider sharing any IO trace data you have with SNIA IOTTA so we can start building a repository for AI traces

NVMe over CXL is much more than Just an SSD

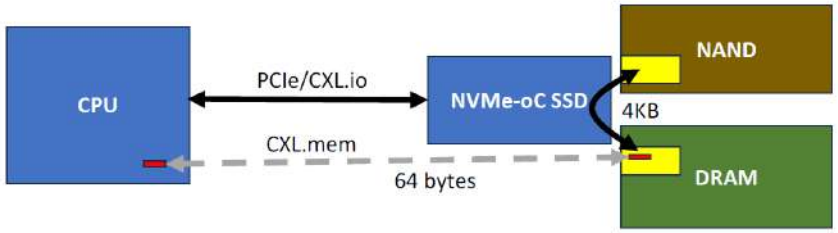
SNIA

NVMe Over CXL is Much More Than Just a SSD

B. Gervasi (Wolley)

- メモリとストレージが別々のインターフェースであり、本来必要なデータ以上のデータを読み込む必要があった
- NVMe over CXLにより統合することで、NVMeプロトコルを通じてコントローラメモリバッファ（CMB）をCXLスペースに配置し、データの必要な部分のみを効率的に取得
- メモリやネットワークのボトルネックを解消し、システム性能を改善する

The NVMe Over CXL Solution: Only grab the FLITs you need



Dual-function device

Appears to the system as an NVMe SSD and as Host Directed Memory (HDM)

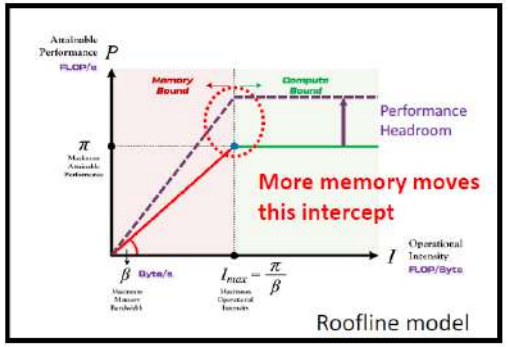
NVMe is just a cache protocol between NAND and DRAM

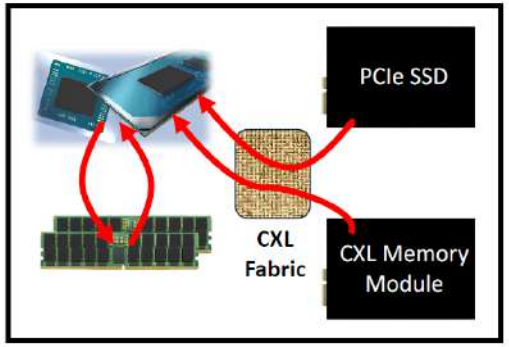
NVMe-oC places the controller memory buffer (CMB) in CXL space (HDM)

Processor grabs only the FLITs needed using CXL.mem

The rest of the CMB data (on average, 97%) remains where it is

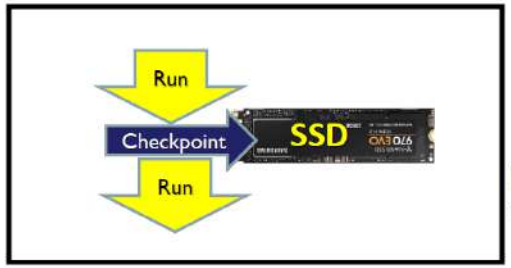
This cache management scheme is expanded to create Virtual HDM





NVMe-oC addresses the memory wall which limits AI

Always let the Host decide where data belongs



NVMe-oC reduces wasted data traffic over the fabric by 30x or more

NVMe-oC supports persistence, allowing checkpoint elimination

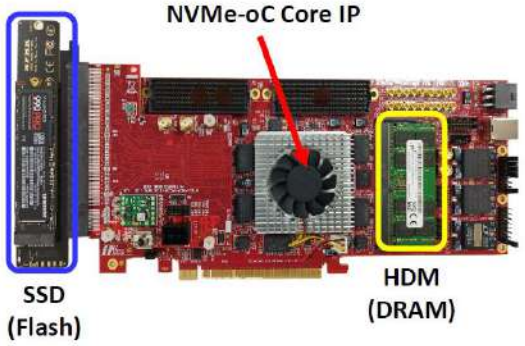
SNIA

NVMe Over CXL is Much More Than Just a SSD

- 実際にHWを作って、性能検証を実施（初歩的な実装でありレイテンシの影響が大きい）
- アイドル状態でもわずかに性能は向上する
- システムに付加をかけたときの性能低下が約半分に抑えられている
- SSDとメモリ間の通信をCPUの処理から分離したことで、CPUが高負荷の状態でも高い性能を維持

NVMe-oC Demonstration Platform

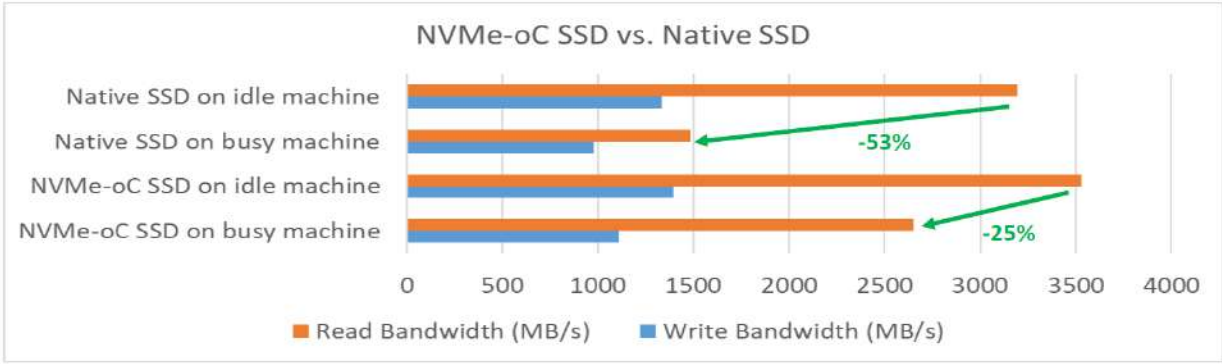
Device	
Host Interface	CXL 1.1/2.0 Gen3x8
HDM	16GB (DDR4-2000)
SSD	128GB ~ 1TB
System Clock	250MHz
NVMe	2.0
Operation Mode	Memory / Storage



Demonstrating Virtual HDM mode using NVMe-oC

Host	
CPU	Intel Granite Rapids, 2 processors, 288-cpu
Memory	128GB DRAM 6400MT
OS	Fedora release 40 (Forty)
Kernel	6.9.5

NVMe-oC Versus Traditional SSD Impact of Traffic Reduction



STREAM (memory benchmark) with 256 background threads

Results:

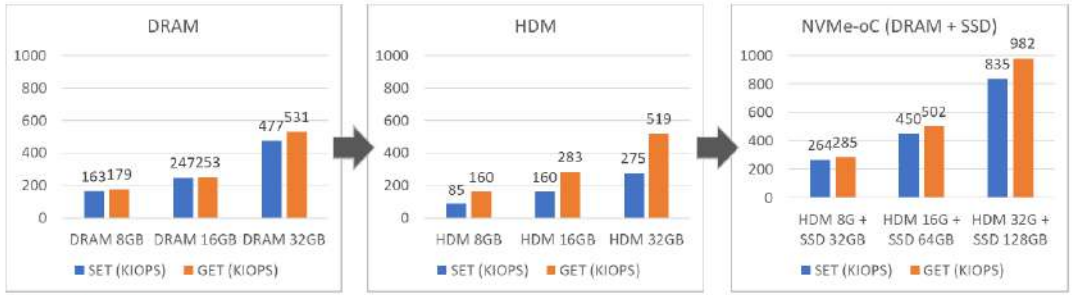
2.8X reduced impact on read performance

SNIA

NVMe Over CXL is Much More Than Just a SSD

- HDMとSSDを組み合わせると、メモリフットプリントが拡大しより高い性能を出すことができる
- SSDによって容量を増やしているので、DRAMを追加するよりも安価(90%コストダウン)
- 圧縮性能のベンチマークでは、多重度を増やすことでDRAMと同等の性能を実現

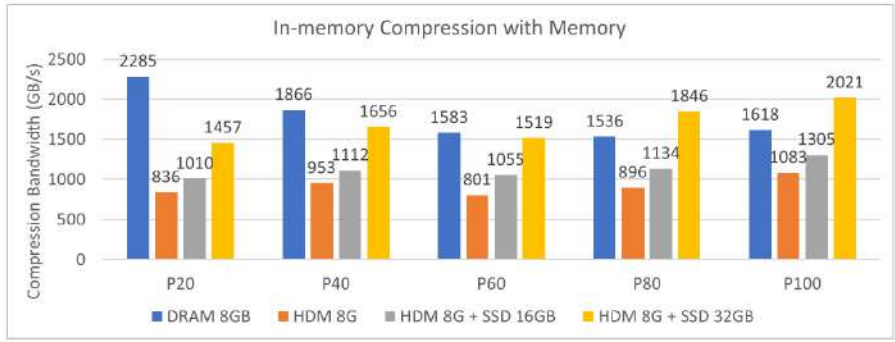
Virtual HDM Mode Redis Performance Versus HDM
Impact of Large Memory Footprint



Unmodified Redis
(In-memory Key Value Store)

Results:
4X memory capacity
2X performance
90% cost reduction

Virtual HDM Mode Compression Performance Versus HDM
Impact of Higher Thread Count



LZ77 lossless data compression method

Results:
4X memory compression
Same performance as DRAM
More compression threads executed

SNIA

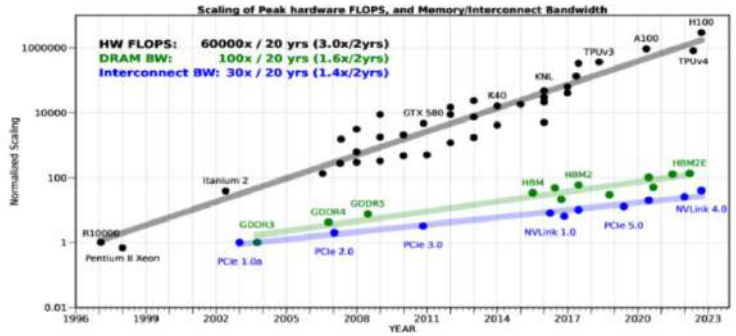
Accelerating GPU Server Access to Network Attached Disaggregated Storage Using Data Processing Unit

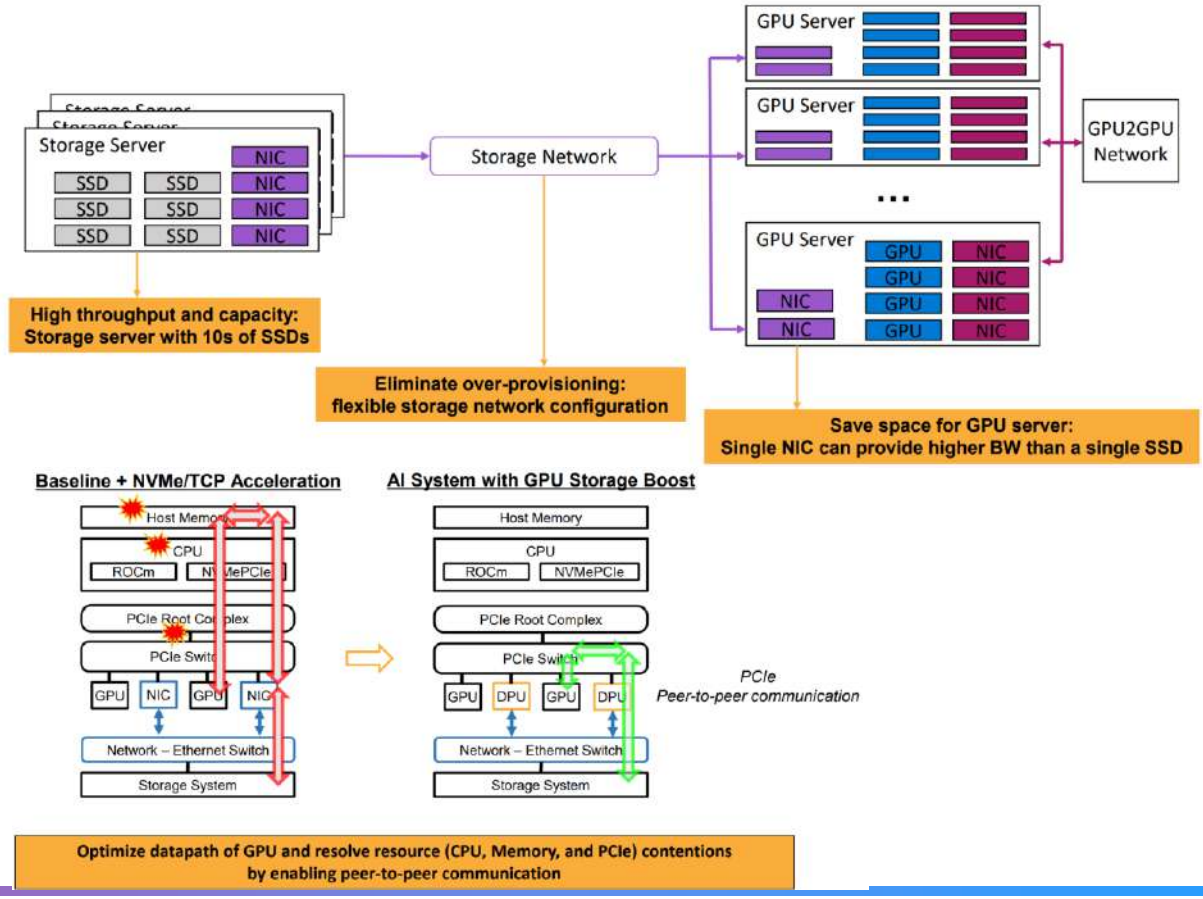
SNIA

Accelerating GPU Server Access to Network Attached Disaggregated Storage Using Data Processing Unit

E. Nurvitadhi (MangoBoost), C. Carlson

- LLMの急激な成長により、GPUの技術が加速。
- GPUのメモリだけでは処理しきれない　ストレージやメモリの性能が不足　CPU処理がボトルネック
- DPUを用いたStorage Disaggregated StorageやGPU Directの登場

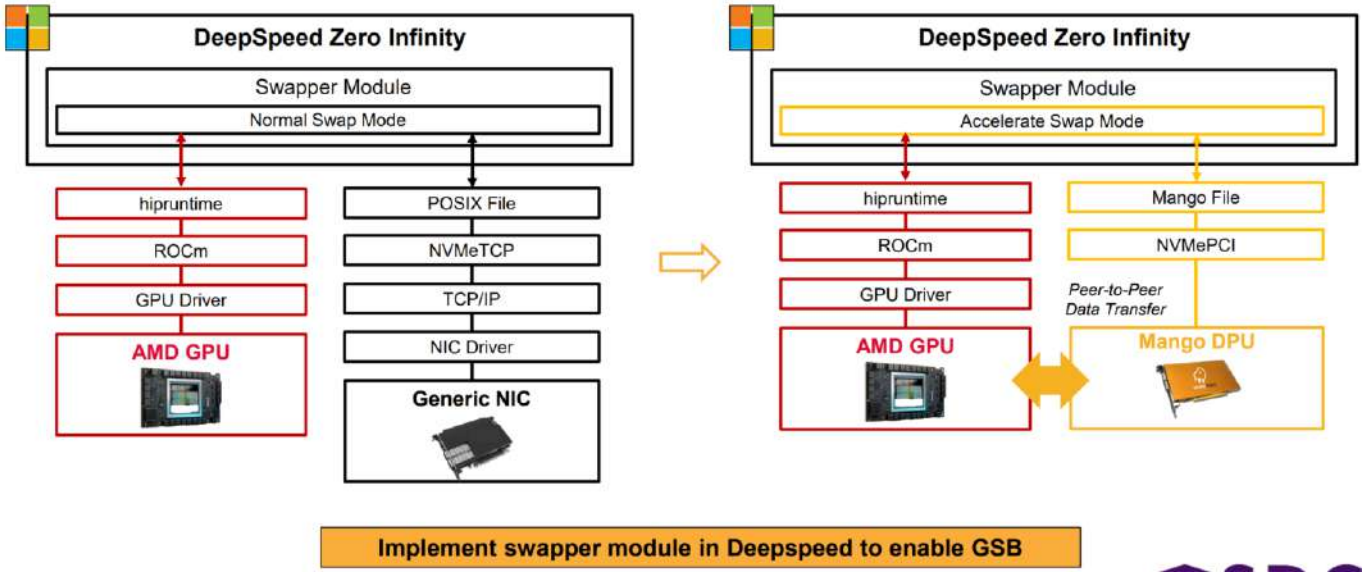




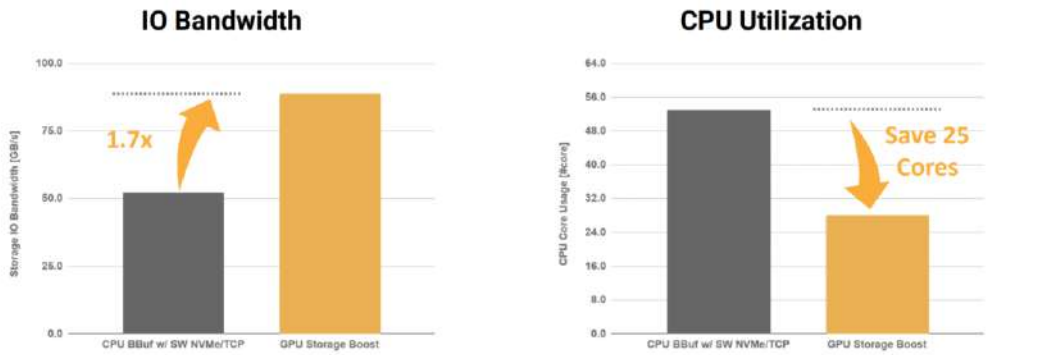
SNIA

Accelerating GPU Server Access to Network Attached Disaggregated Storage Using Data Processing Unit

- FPGAベースのSmartNICに、FilesystemとGPU Direct Storage機能を実装
- FIOベンチマークで、IO Bandwidthが1.7倍、CPU負荷が25％軽減



Implement swapper module in Deepspeed to enable GSB



Provide higher bandwidth and lower cpu utilization in the state-of-the-art AI training framework

Accelerating GPU Server Access to Network Attached Disaggregated Storage Using Data Processing Unit

- AMDのGPU RadeonとInstinct：コンシューマ向けとDC向け
- ROCm：オープンソースの開発用ツール
- ROCｍ：CUDA Library相当のLibrary
- 2016に1.0がリリースされて、最新版は6.2でプロファイリングツールが強化

GPU

- AMD GPUs come in two classes
  - Radeon – Consumer GPUs
    - Primarily used for gaming, but can be used for AI/HPC
  - Instinct – Data center GPUs
    - CDNA – Architecture Designed for AI and HPC applications
    - HBM – Includes High Bandwidth Memory
    - Infinity Fabric – High speed interconnect
- ROCm Development Platform
  - Open source
  - Supports Instinct and Radeon GPUs

ROCm

ROCm Libraries

CUDA Library	ROCm Library	Description
cuBLAS	rocBLAS	Basic Linear Algebra Subroutines
cuFFT	rocFFT	Fast Fourier Transfer Library
cuSPARSE	rocSPARSE	Sparse BLAS + SPMV
cuSolver	rocSolver	Lapack Library
AMG-X	rocALUTION	Sparse iterative solvers & preconditioners with Geometric & Algebraic MultiGrid
Thrust	rocThrust	C++ parallel algorithms library
CUB	rocPRIM	Low Level Optimized Parallel Primitives
cuDNN	MIOpen	Deep learning Solver Library
cuRAND	rocRAND	Random Number Generator Library
EIGEN	EIGEN	C++ template library for linear algebra: matrices, vectors, numerical solvers
NCCL	RCCL	Communications Primitives Library based on the MPI equivalents

SNIA

SNIA日本支部主催
「2024年度第1回最新技術動向講演会」

SDC 2024 参加報告

Ryosuke Tatsumi

