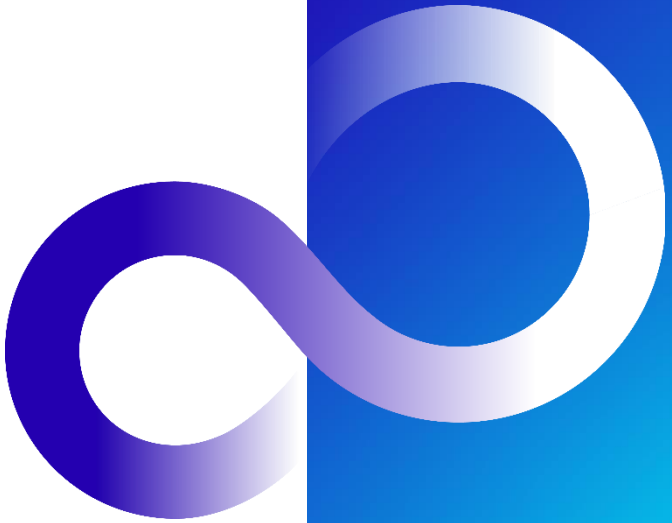
LUG JAPAN 2021

FUJITSU

次世代のHPCストレージ・ファイルシステム技術



2021年12月10日

Shinji Sumimoto, Ph.D.
FUJITSU LTD.

OpenSFS

Background

FUJITSU

- ストレージ、ファイルシステムへの要求が多様化
  - SNS, IoT, センサーデータ
  - クラウド利用の広まり
  - オンプレミス、エッジ・クラウド間データ連携、
  - リアルタイム性、メタデータ性能、データアクセス性能
- HPCとDLの融合：
  - 大容量データを利用したシミュレーション高速化
  - リアルタイムデータを用いたシミュレーション: データ同化
- メモリシステムとストレージシステムの境界消失
  - SCM, NVMe技術の一般化、キャッシュコヒーレンス領域の拡大
  - アクセス手法(命令 or API)の違いになりつつある
- 今後のHPCシステムを概観、ストレージ・ファイルシステムについて考察

次世代HPCシステムの方向性とこれからのストレージ

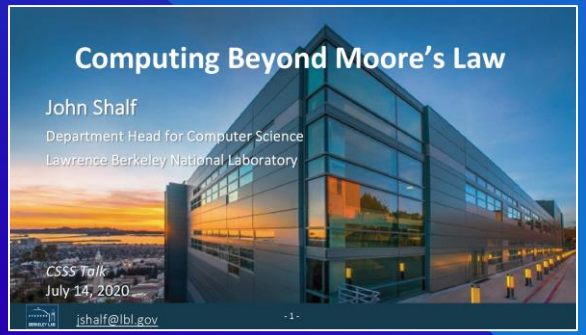
FUJITSU

- 次世代HPCシステムの方向性
  - Post Moore: Computing Beyond Moore’s Law
  - Disaggregated Computing
  - Cloud Native Computing
  - NGACI: Next-Generation Advanced Computing Infrastructure
- これからのストレージ・ファイルシステム

- 本講演で述べることは一個人の見解であり、現在所属している会社で開発中の製品とは直接関係ありません。

Post Moore : Computing Beyond Moore's Law

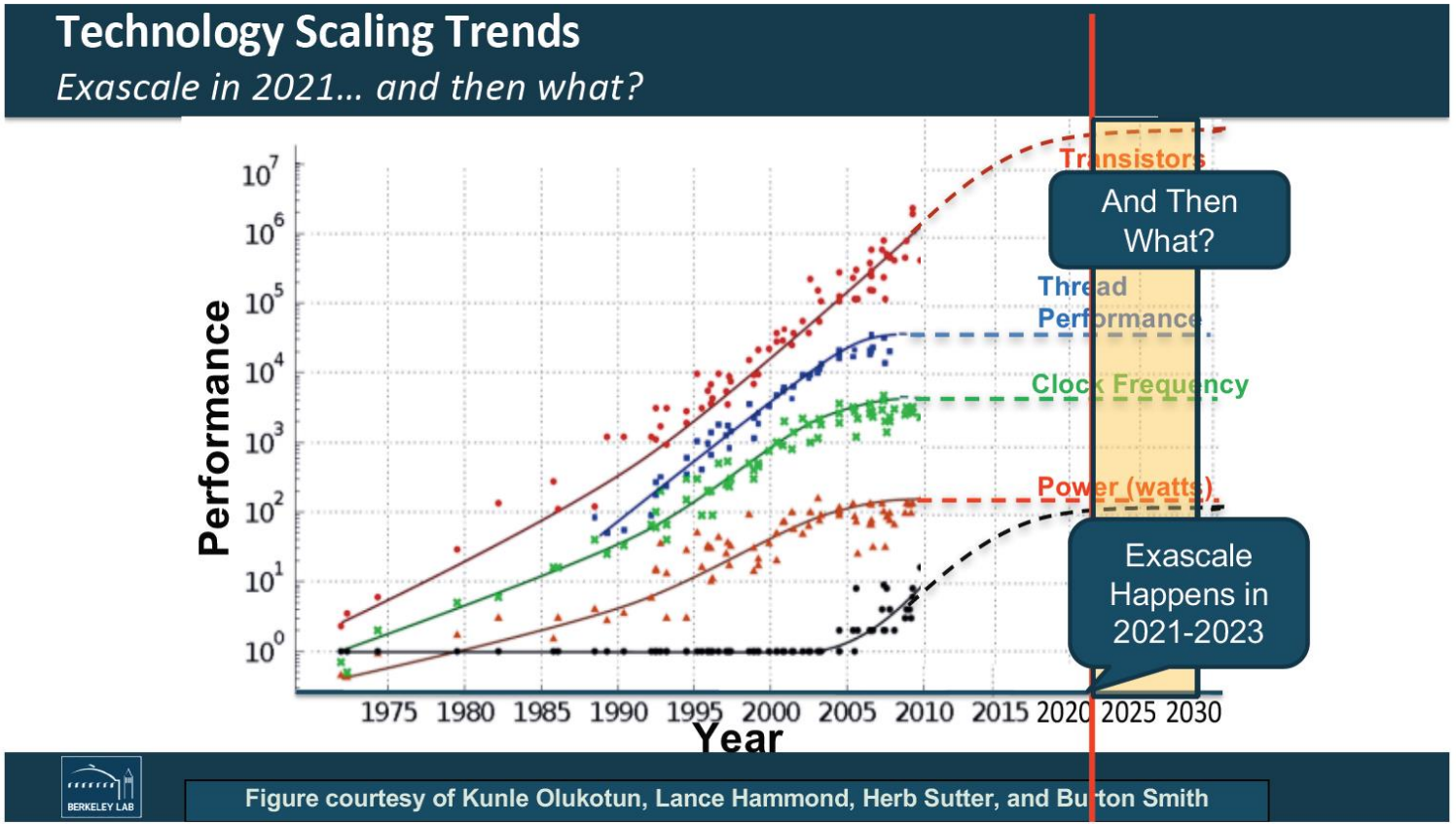
Computing Beyond Moore's Law
https://cs.lbl.gov/assets/CSSSP-Slides/20200714-Shalf.pdf



Computing Beyond Moore's Law

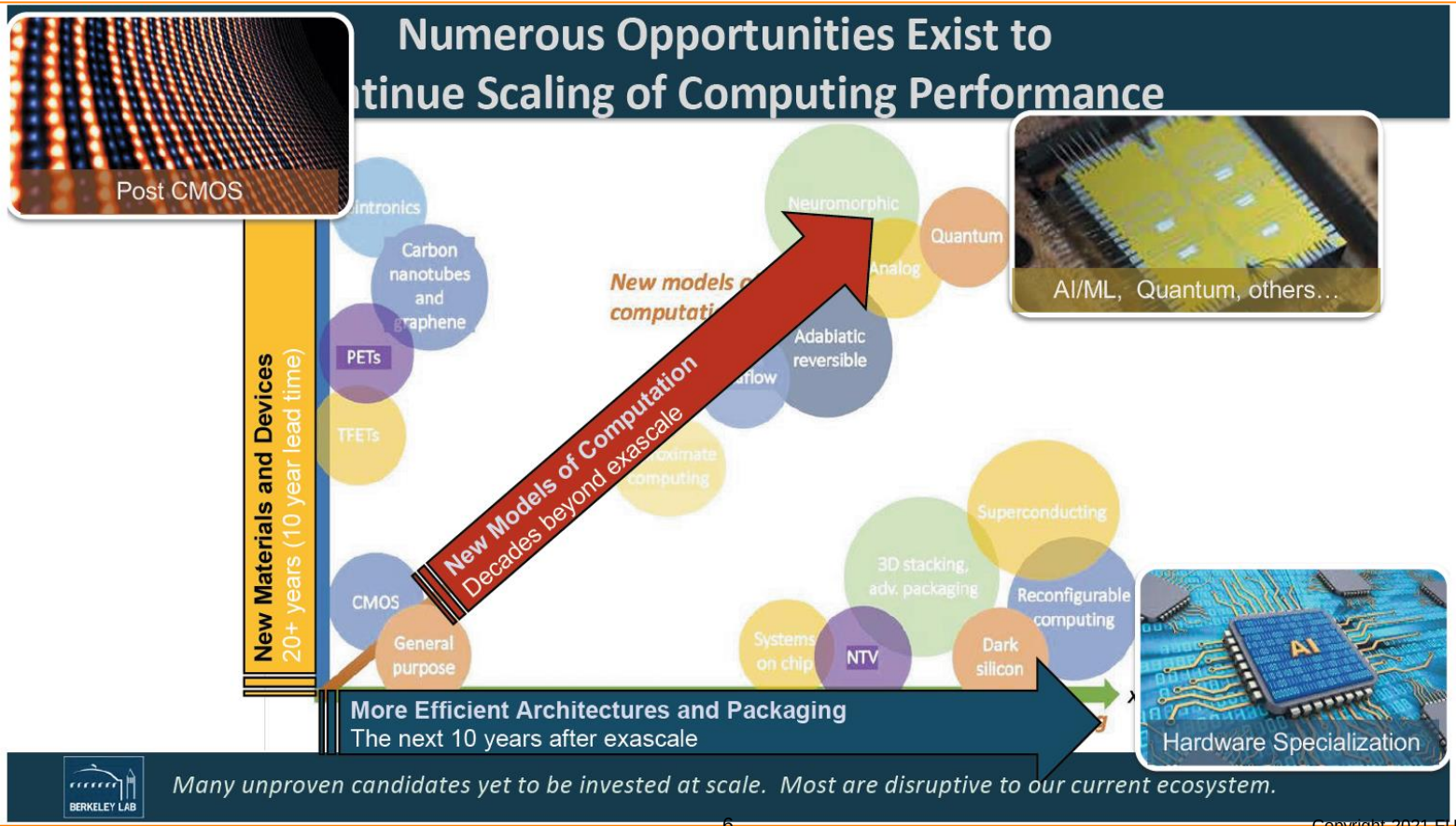
https://cs.lbl.gov/assets/CSSSP-Slides/20200714-Shalf.pdf

FUJITSU



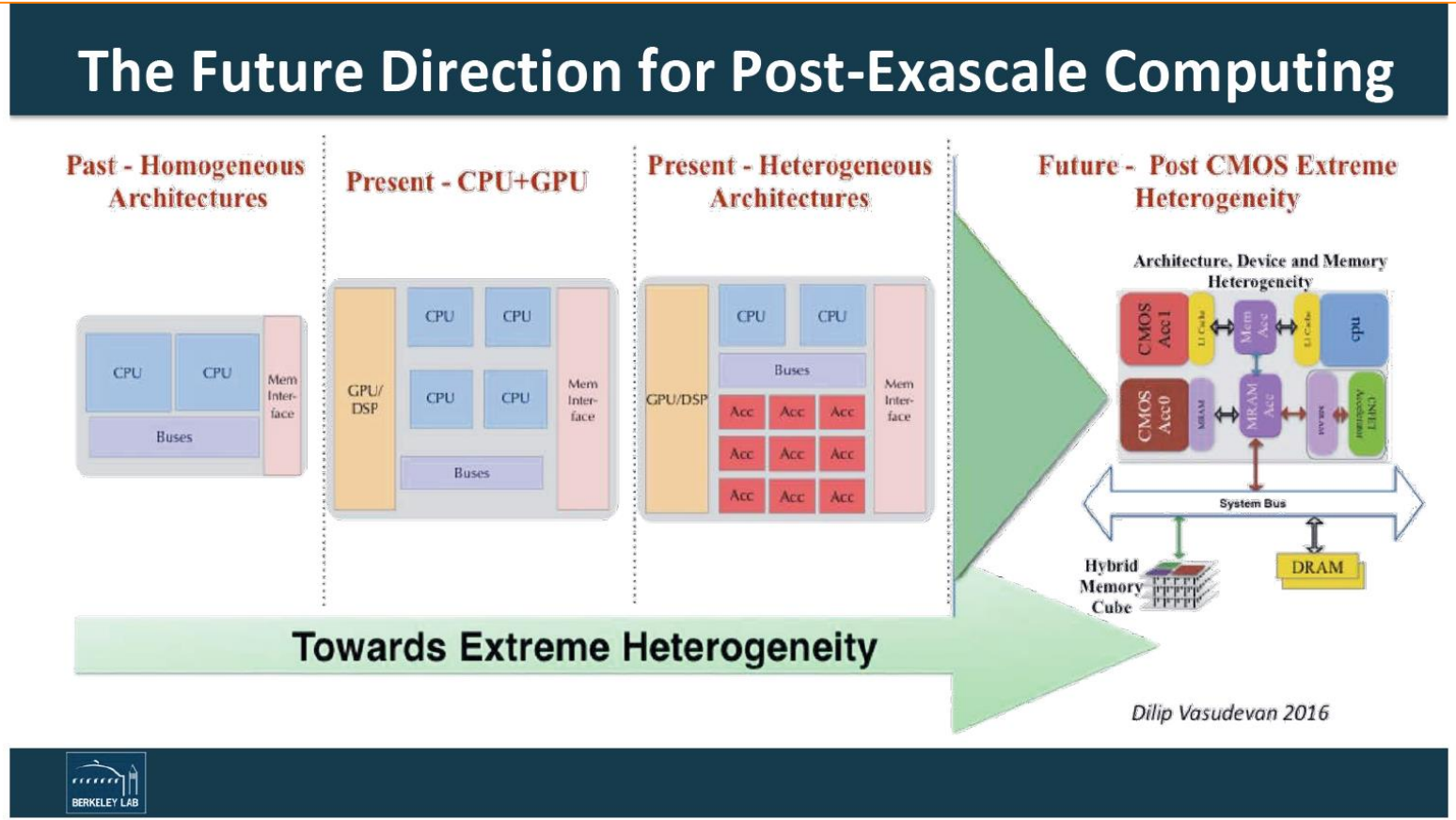
Computing Beyond Moore's Law

https://cs.lbl.gov/assets/CSSSP-Slides/20200714-Shalf.pdf



Computing Beyond Moore's Law

https://cs.lbl.gov/assets/CSSSP-Slides/20200714-Shalf.pdf



FUJITSU

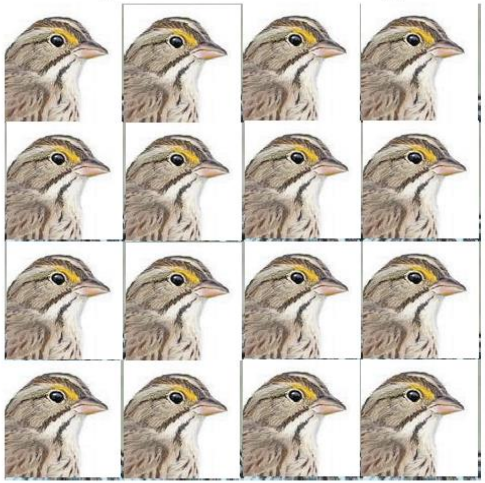
Specialization:

Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Xeon, Power

Many Lighter Weight
(post-Dennard scarcity)



KNL AMD, Cavium/Marvell, GPU

Many Different Specialized
(Post-Moore Scarcity)

Apple, Google, Amazon

BERKELEY LAB

Project 38
https://crd.lbl.gov/departments/computer-science/cag/research-2/project-2/

FUJITSU

Project 38 -- Background

DOD and DOE recognize the imperative to develop new mechanisms for engagement with the vendor community, particularly on architectural innovations with strategic value to USG HPC.

Project 38 (P38) is a set of vendor-agnostic architectural explorations involving DOD, the DOE Office of Science, and NNSA (these latter two organizations are referred to in this document as "DOE"). These explorations should accomplish the following:

- **Near-term goal:** *Quantify the performance value and identify the potential costs of specific architectural concepts against a limited set of applications of interest to both the DOE and DOD.*
- **Long-term goal:** *Develop an enduring capability for DOE and DOD to jointly explore architectural innovations and quantify their value.*
- **Stretch goal:** *Specification of a shared, purpose built architecture to drive future DOE-DOD collaborations and investments. (purpose-built HPC by 2025)*

BERKELEY LAB

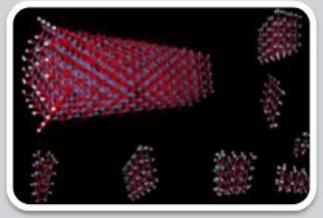
Project 38

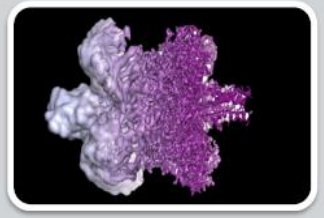
https://crd.lbl.gov/departments/computer-science/cag/research-2/project-2/

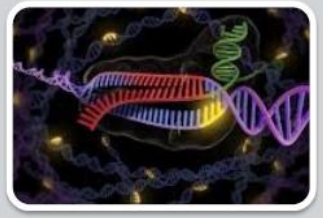
FUJITSU

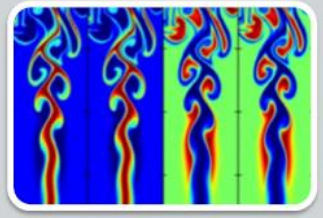
Architecture Specialization for Science

(hardware is design around the algorithms) can't design effective hardware without math









Materials	CryoEM Accelerator	Genomics Accelerator	Digital fluid Accelerator
Density Functional Theory (DFT)	LBNL detector	String matching	3D integration
Use O(n) algorithm	750 GB / sec	Hashing	Petascale *chip*
Dominated by FFTs	Custom ASIC near detector	2-8bit (ACTG)	1024-layers
FPGA or ASIC		FPGA solution	General / special HPC solution

BERKELEY LAB

- 10 -

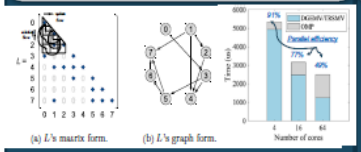
Project 38

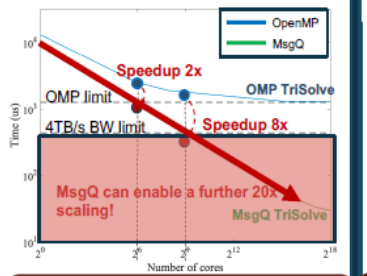
https://crd.lbl.gov/departments/computer-science/cag/research-2/project-2/

FUJITSU

Phase 1 Studies

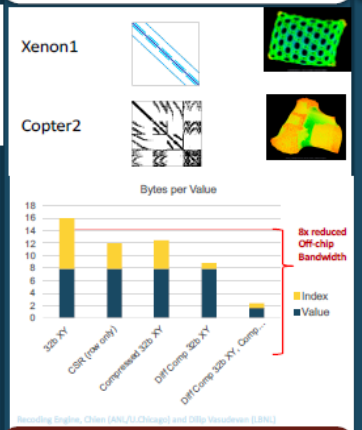
Message Queues (Sparse Matrix TRSMV)





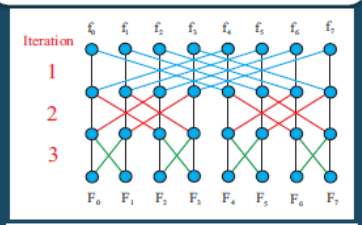
8x improved scaling (limited by memory BW) with room to grow

Recoding Engine (Sparse Matrix Compression)



2x-8x Reduction in Memory Bandwidth req.

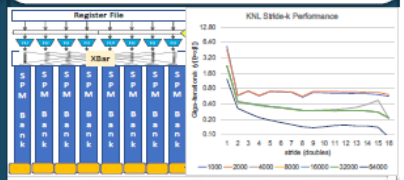
Fixed Function Accelerators (FFT as example)



1GHz @ 14nm
100 GB/s off-chip memory
~1.5TFLOPs uses 4.5mm²
1TB/s off-chip memory
~15TFLOPs uses 47mm²

Up to 20x-100x less area consumption for same perf. (viable FFTx/BLAS)

Gather/Scatter Scratchpads (for Tensor Contractions)



Bandwidth waste for loading the t3 or v in inner loop	Bandwidth waste for the entire application
55%	36%
100%	65.4%
700%	457.8%
154%	100.7%
166%	109%

55%-700% reduction in data movement

BERKELEY LAB

Disaggregated Computing

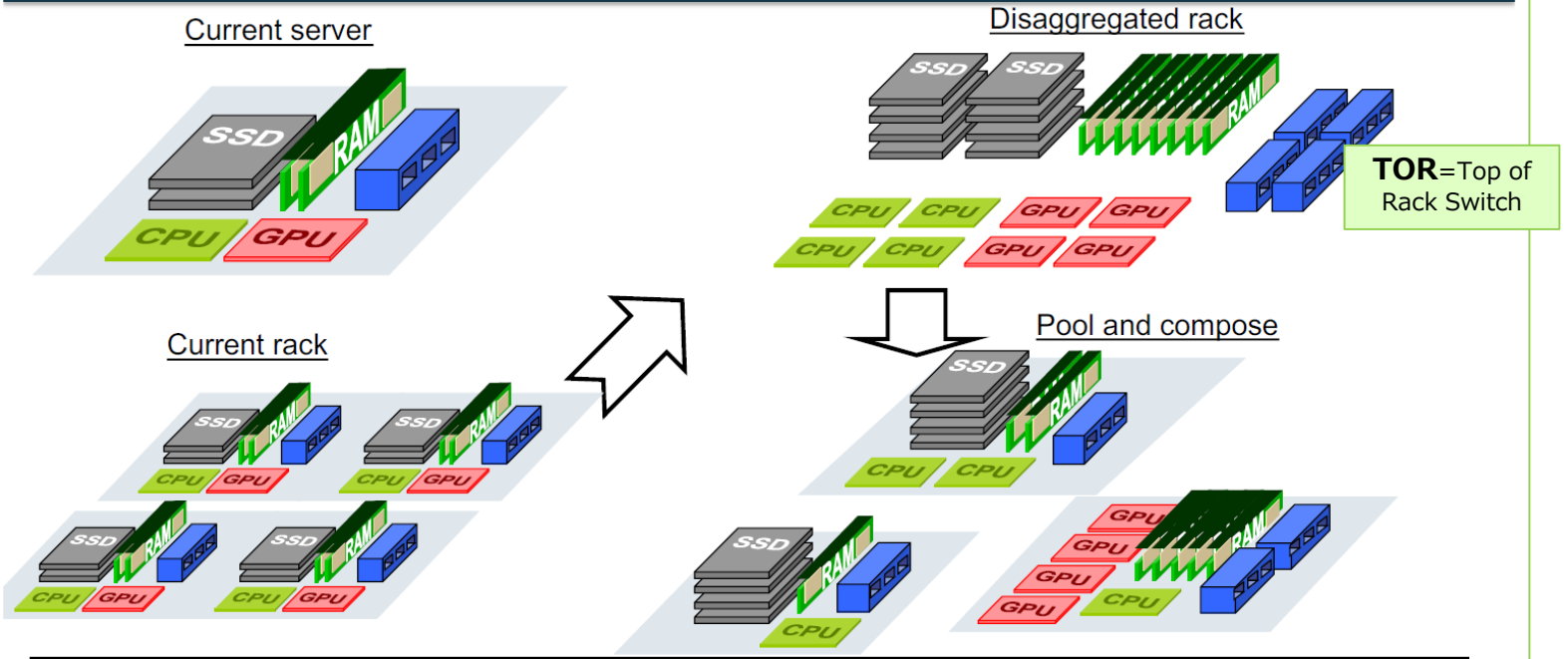
https://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf

Disaggregated Computing

https://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf

FUJITSU

Disaggregated Node/Rack Architecture



Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)
But this is significantly inferior to RAM bandwidth (100 GB/s – 1 TB/s)

Disaggregated Computing

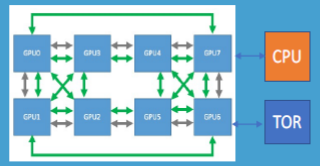
https://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf

FUJITSU

Diverse Node Configurations for Datacenter Workloads

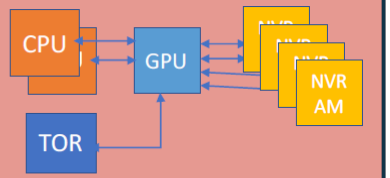
Training

- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)

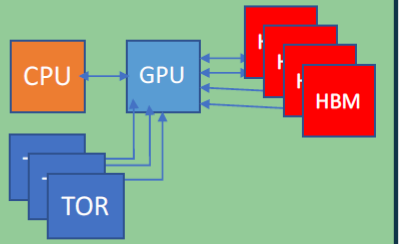


Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

Graph Analytics

- 16 links HBM
- 8 links TOR
- 1 Link CPU



BERKELEY LAB

TOR | NVRAM | GPU | HBM | CPU

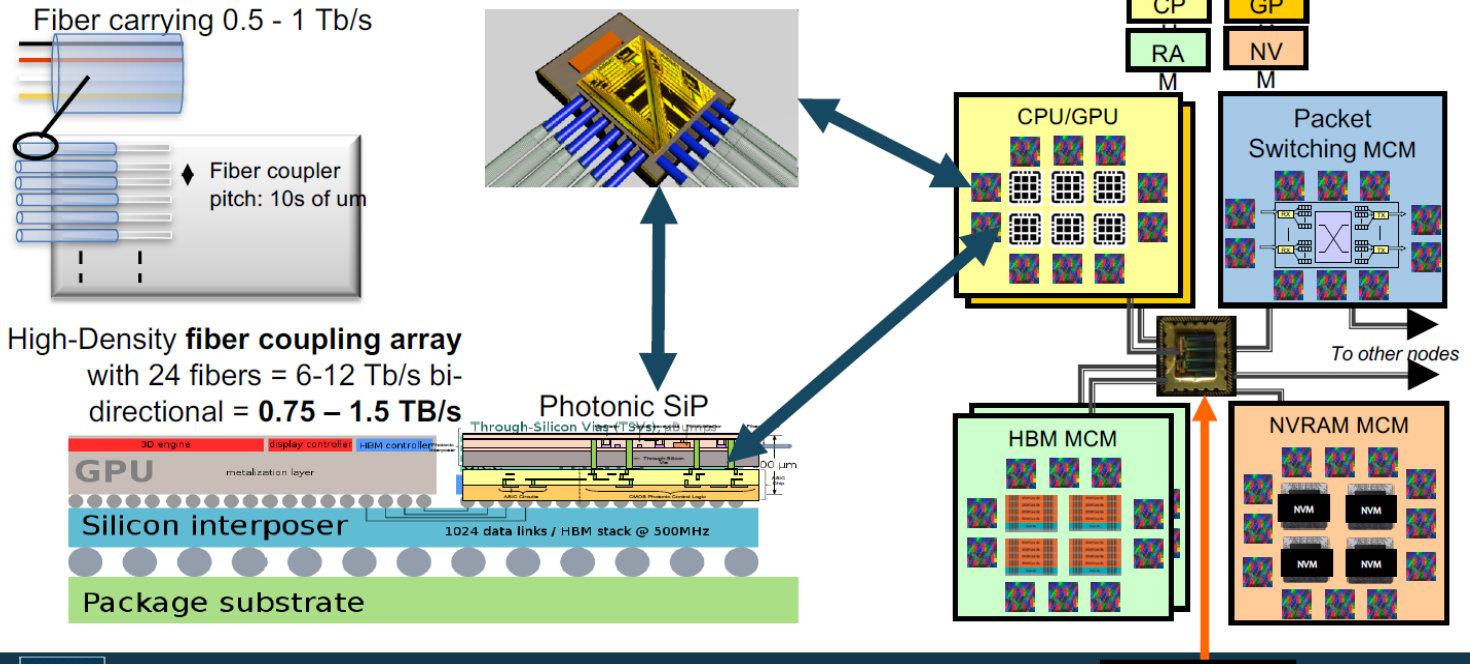
TOR=Top of Rack Switch

Disaggregated Computing

https://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf

FUJITSU

Photonic MCM (Multi-Chip Module)

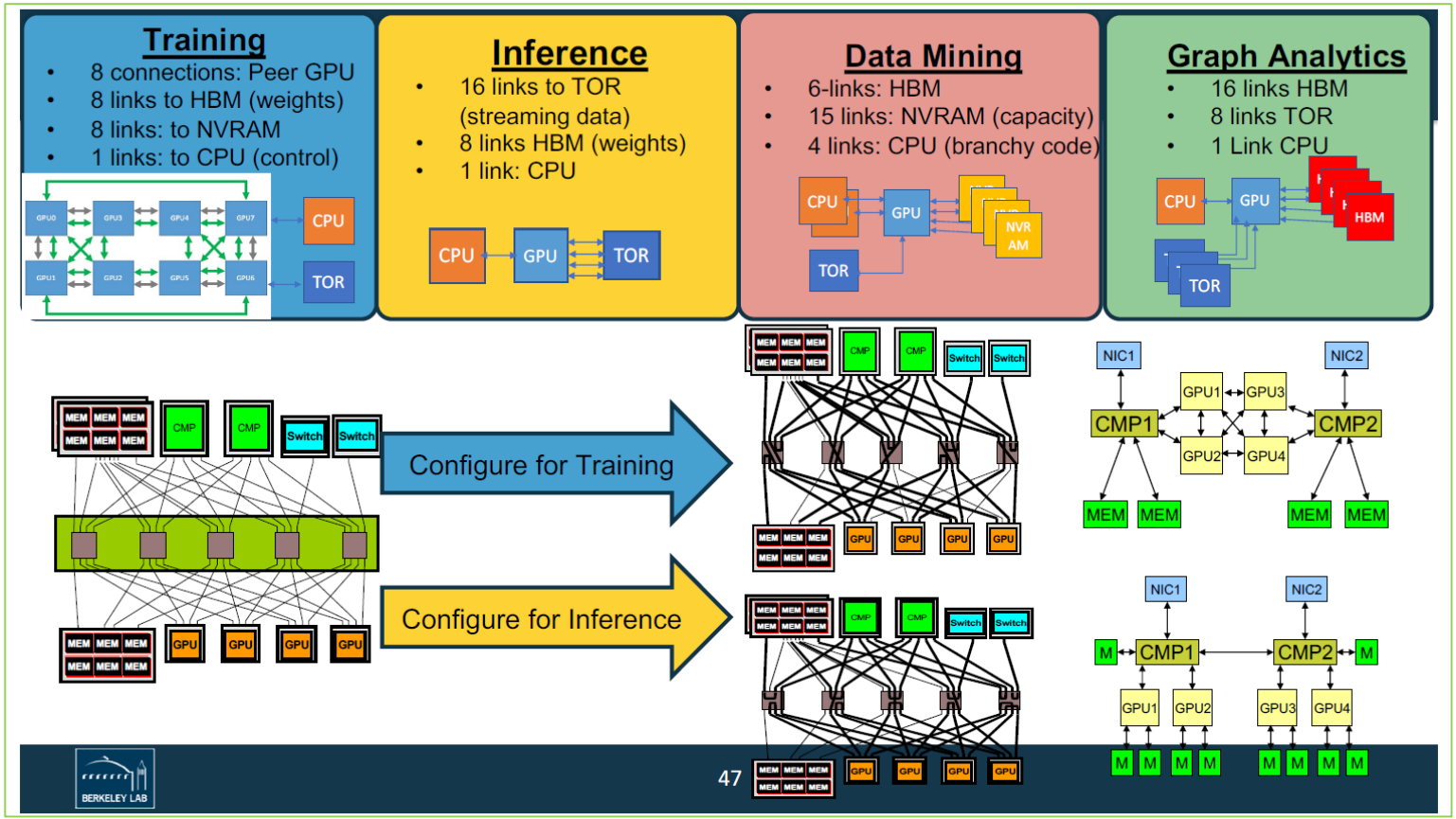


BERKELEY LAB

46

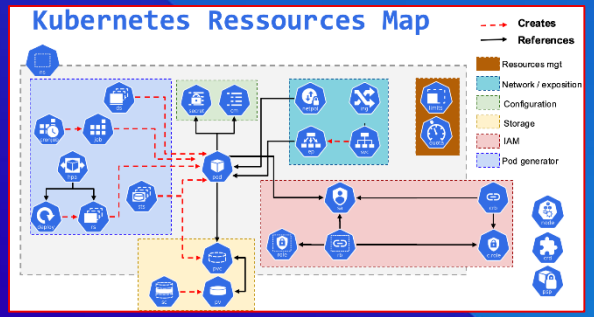
Disaggregated Computing

https://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf



Cloud Native Computing

https://www.cncf.io/



Cloud Native Computing: Cloud Nativeとは？

FUJITSU

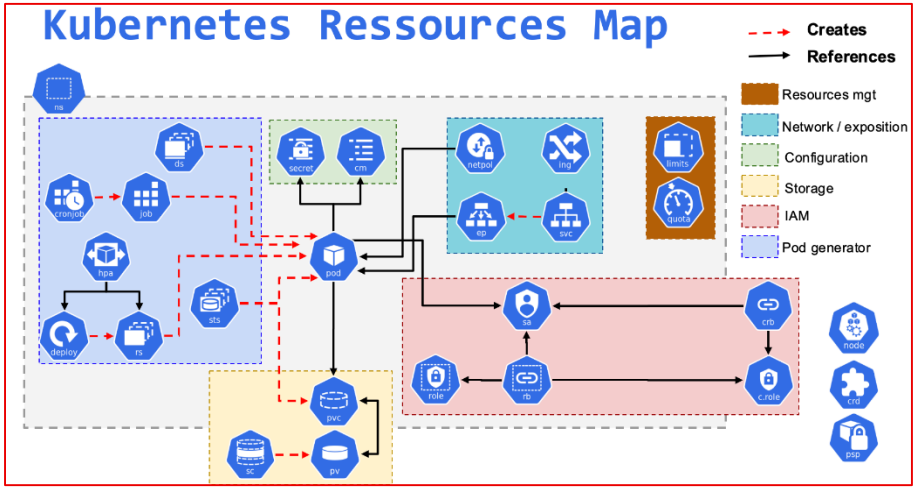
- **CNCF Cloud Native Definition v1.0**
  - クラウドネイティブ技術は、パブリッククラウド、プライベートクラウド、ハイブリッドクラウドなどの近代的でダイナミックな環境において、スケーラブルなアプリケーションを構築および実行するための能力を組織にもたらします。 このアプローチの代表例に、**コンテナ、サービスメッシュ、マイクロサービス、イミュータブルインフラストラクチャ、および宣言型API**があります。
  - これらの手法により、**回復性、管理力、および可観測性のある疎結合システム**が実現します。 これらを堅牢な自動化と組み合わせることで、エンジニアはインパクトのある変更を最小限の労力で頻繁かつ予測どおりに行うことができます。
  - Cloud Native Computing Foundationは、オープンソースでベンダー中立プロジェクトのエコシステムを育成・維持して、このパラダイムの採用を促進したいと考えてます。 私たちは最先端のパターンを民主化し、これらのイノベーションを誰もが利用できるようにします。

  - https://github.com/cncf/toc/blob/main/DEFINITION.md
  - https://12factor.net/

Cloud Native Computing

FUJITSU

- Server Less
  - On demand Provisioning
  - Automatic Deploy and Scaling
- Run Anywhere by Container, Docker
  - Micro Services, Isolated Processes, Event(Message) Driven
- Loosely Coupled (vs. Tightly Coupled)
  - Orchestration, Fault Tolerance, Automatic Scaling



NGACI: Next-Generation Advanced Computing Infrastructure

https://sites.google.com/view/ngaci/home

NGACI White Paper

https://www.mext.go.jp/content/20210618-mxt-jyohoka01-000016078-01.pdf

FUJITSU

NGACI活動の紹介

- NGACI: Next-Generation Advanced Computing Infrastructure
  - 概要と活動目的

今後の高性能計算機の持続的な発展を考えるにあたり、AIやビッグデータ技術とのさらなる融合、Society5.0といった新しい応用分野への展開など、さらなる発展も期待されますが、ムーアの法則の終焉など多くの技術的課題が待ち受けていることも事実です。本活動(NGACI)は、将来の高性能計算環境として、また共用計算機資源としてどのような技術的課題があり、どのような研究開発が必要なのか、コミュニティーとしてどのような活動をしていくべきなのかなどに関して、オープンに意見交換をしつつそれをWhite Paperとしてまとめることで本分野の発展に寄与することを目的としています。

https://sites.google.com/view/ngaci/home

  - これまでの実績
    - 本活動に登録して頂いているコミュニティのメンバー数：104人
    - 7回の全体ミーティングと3回のセミナーを実施
    - 4つのWGにより将来のシステム像や課題を集中的に議論
      - アーキテクチャWG、システムソフトWG、アプリ/ライブラリWG、システム運用WG
  - White Paperについて
    - 1.0.0版（164ページ）を公開中（https://sites.google.com/view/ngaci/home）

NGACI white paper

2021/5/27　次世代計算基盤検討部会　2

White Paperの執筆協力者

所属等は2020年11月時点

- 取りまとめ: 近藤(東大・理研)
- アーキテクチャWG
  - WGリーダ: 三輪(電通大), 佐野(理研), 谷本(九大)
  - WGメンバ: 安島(富士通), 井口(北陸先端大), 井上(九大), 江川(電機大),岡本(Spin Memory) 小野(九大), 鯉渕(NII), 児玉(理研), 小林(筑波大), 小松(東北大), 佐藤(東北大), 塩見(京大), 田邊(東大), 中里(会津大), 吉川(富士通研), 福本(富士通研), 星(NEC), 三好(わさらぼ), 宮島(理研)
- システムソフトWG
  - WGリーダ: 佐藤(理研), 佐藤(豊橋技科大)
  - WGメンバ: 合田(NII), 遠藤(東工大), 小柴(理研), 小松(東北大), 坂本(東大), 高野(産総研), 滝沢(東北大), 辻(理研), ゲローフイ(理研), 中島(富士通研), 深井(理研), 山本(理研), 和田(明星大)
- アプリケーション・ライブラリ・アルゴリズムWG
  - WGリーダ: 深沢(京大), 今村(理研), 中島(東大・理研)
  - WGメンバ: 岩下(北大), 小野(九大), 笠置(富士通研), 片桐(名大),白幡(富士通研) 住元(富士通研), 高橋(筑波大), 寺尾(理研), 長坂(富士通研), 椋木(理研), 村上(都立大)
- システム運用WG
  - WGリーダ: 塙(東大), 野村(東工大)
  - WGメンバ: 大島(名大), 實本(理研), 庄司(理研), 滝澤(産総研), 竹房(NII), 藤原(NII), 三浦(理研)

2021/5/27 次世代計算基盤検討部会 3

2028年頃に実現可能な次世代システムの予測

- 次世代システムの構成としていくつかのアーキテクチャタイプを検討
  - 汎用システム型
    - メニーコアCPU型：富岳の構成の延長として考えられるシステム
    - メニーコアCPU＆GPU混載型：GPUとホストCPUで構成（現在多くのシステムでも採用）
    - ベクトルプロセッサ混載型：ベクトルプロセッサとホストCPUで構成（例： SX-Aurora TSUBASA ）
  - 専用システム混載型（ムーアの法則減速により重要な検討事項に）
    - CPU拡張型
      - ISA（SIMD）の専用的な命令をCPUに拡張機能として搭載（例：Intel AMXやARM SVEのFMMLAなど）
      - BFloat16やINT8、INT4などの応用に特化したデータ型の導入
    - アクセラレータ主体型/ヘテロジニアス型
      - システム搭載方式：チップ内拡張(SoCやMCM)、ノード内拡張、ラック間疎結合
      - アクセラレータ構成方式：専用、準専用、汎用
    - Processing-In-memory主体型
      - 演算器とメモリの密接実装によるメモリアクセスの高バンド幅化と低遅延化
    - 新計算原理の混載

NGACI White Paper

https://www.mext.go.jp/content/20210618-mxt-jyohoka01-000016078-01.pdf

FUJITSU

2028年のメニーコア型システムの予測性能

- 最も積極的な予測でも最大1.8 EFLOPS (富岳の性能の3.37倍)

	30MW			40MW			50MW			
←システムの電力制約の仮定										
←CPUの電力バジェット	60%	70%	80%	60%	70%	80%	60%	70%	80%	
ソケット数	46620	54390	62160	62160	72520	82880	77700	90650	103600	158,976
総コア数	3.3×10^6	3.8×10^6	4.4×10^6	4.4×10^6	5.1×10^6	5.8×10^6	5.4×10^6	6.3×10^6	7.3×10^6	8.3×10^6
PFLOPS	815	950	1086	1086	1267	1448	1358	1584	1810	537
DDR 総BW (PB/s)	102	120	137	137	160	182	171	200	228	—
HBM 総BW (PB/s)	307	358	410	410	478	547	512	598	683	163
DDR 総容量 (PB)	17	20	23	23	27	31	29	34	39	—
HBM 総容量 (PB)	4	5	5	5	6	7	7	8	9	4.85
インジェクションBW (Tb/s)	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	0.33
総I/O性能 (TB/s)	34	34	34	34	34	34	34	34	34	
総ストレージ容量 (EB)	3.45	3.45	3.45	3.45	3.45	3.45	3.45	3.45	3.45	

参考)富岳の諸元

2021/5/27

アクセラレータの構成法式

- 専用アクセラレータ
  - 特定の計算問題または計算ドメインのみを高速に処理可能な構成方式
  - 最も高性能・高電力効率であるが、柔軟性やプログラマビリティが低い
  - 例)ディープニューラルネットワークアクセラレータ、FFTアクセラレータ
- 準専用アクセラレータ
  - 複数の計算モデル・計算問題を高速に処理可能＋ある程度のプログラマビリティを持つ構成
  - 専用アクセラレータに比べ若干性能や電力効率で劣る
  - 例)データフロープロセッサ、コンボリューションプロセッサ
- 汎用アクセラレータ
  - 特定のアクセラレータを再構成可能デバイス上に構成する方式
  - 例)FPGA, Coarse-Grain Reconfigurable Arrays (CGRA)
- 特定計算ドメインの候補
  - 疎行列計算、N体問題、FFT、ソート、グラフ処理、脳型計算、量子計算シミュレーション、エージェントシミュレーション

2021/5/27 次世代計算基盤検討部会 9

2028年のGPU混載型システムの予測性能

- 最も積極的な予測で最大18.0 EFLOPS (富岳の性能の33.5倍)

保守的な予測の場合
(NVIDIA GPUの性能トレンドから外挿)

	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	5.3×10^8	6.2×10^8	7.1×10^8	7.1×10^8	8.3×10^8	9.4×10^8	8.8×10^8	1.0×10^9	1.2×10^9
PFLOPS	1279	1492	1706	1706	1940	2474	2132	2487	2843
HBM 総BW (PB/s)	91	107	122	122	143	163	153	178	204
HBM 総容量 (PB)	1	1	2	2	2	2	2	3	3

積極的な予測の場合
(CPUとの性能比のトレンドから外挿)

	30MW			40MW			50MW		
	60%	70%	80%	60%	70%	80%	60%	70%	80%
GPU 数	50661	59104	67548	67548	78806	90064	84435	98508	112580
総コア数	3.4×10^9	3.9×10^9	4.5×10^9	4.5×10^9	5.2×10^9	6.0×10^9	5.6×10^9	6.5×10^9	7.5×10^9
PFLOPS	8083	9431	10778	10778	12574	14371	13472	15718	17963
HBM 総BW (PB/s)	334	390	445	445	520	594	557	650	743
HBM 総容量 (PB)	4	5	6	6	7	8	8	9	10

2021/5/27 次世代計算基盤検討部会 8

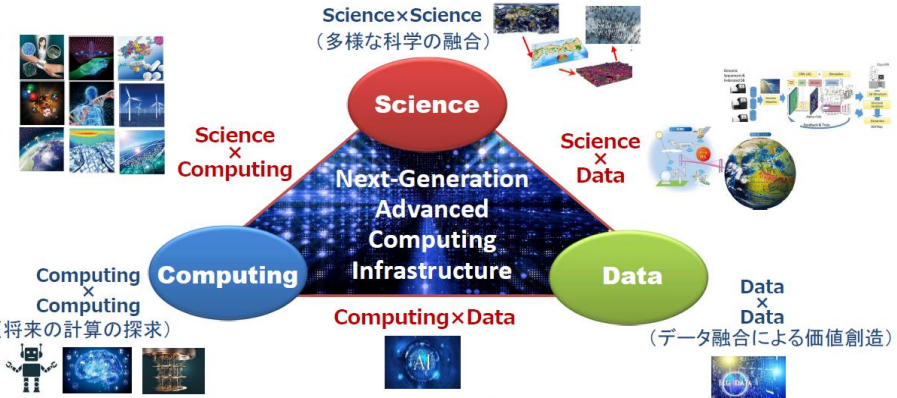
NGACI White Paper

https://www.mext.go.jp/content/20210618-mxt-jyohoka01-000016078-01.pdf

FUJITSU

次世代先端的計算基盤の可能性

- 様々な科学（物理・工学・化学・生物・計算科学・情報科学など）の共創の場の創生へ



2021/5/27　次世代計算基盤検討部会　17

次世代型運用への要求

- 新利用形態への対応
  - 観測データやセンサデータ、外部データベースを直接取り込みながらリアルタイム処理
    - 例）地震観測データによるデータ同化、ゲリラ豪雨予測、ビッグデータによる異常検知
  - 高信頼性・高セキュリティへの要求
- データアーカイブ・流通
  - 重要データの保護、ディザスタリカバリ → 複数拠点でのデータ連携
- 設備・管理
  - 省エネ運用、電力変動対応、冷却設備の負荷変動大への対応
  - 外気導入や湖水・海水の利用を含めた次世代冷却技術
- ユーザ利用・課金モデル
  - Service Level Agreement (SLA)の定義
  - 省エネ実行に対するユーザのインセンティブの明確化と課金モデルへの反映

2021/5/27　次世代計算基盤検討部会　14

技術課題と研究開発ロードマップ（抜粋）

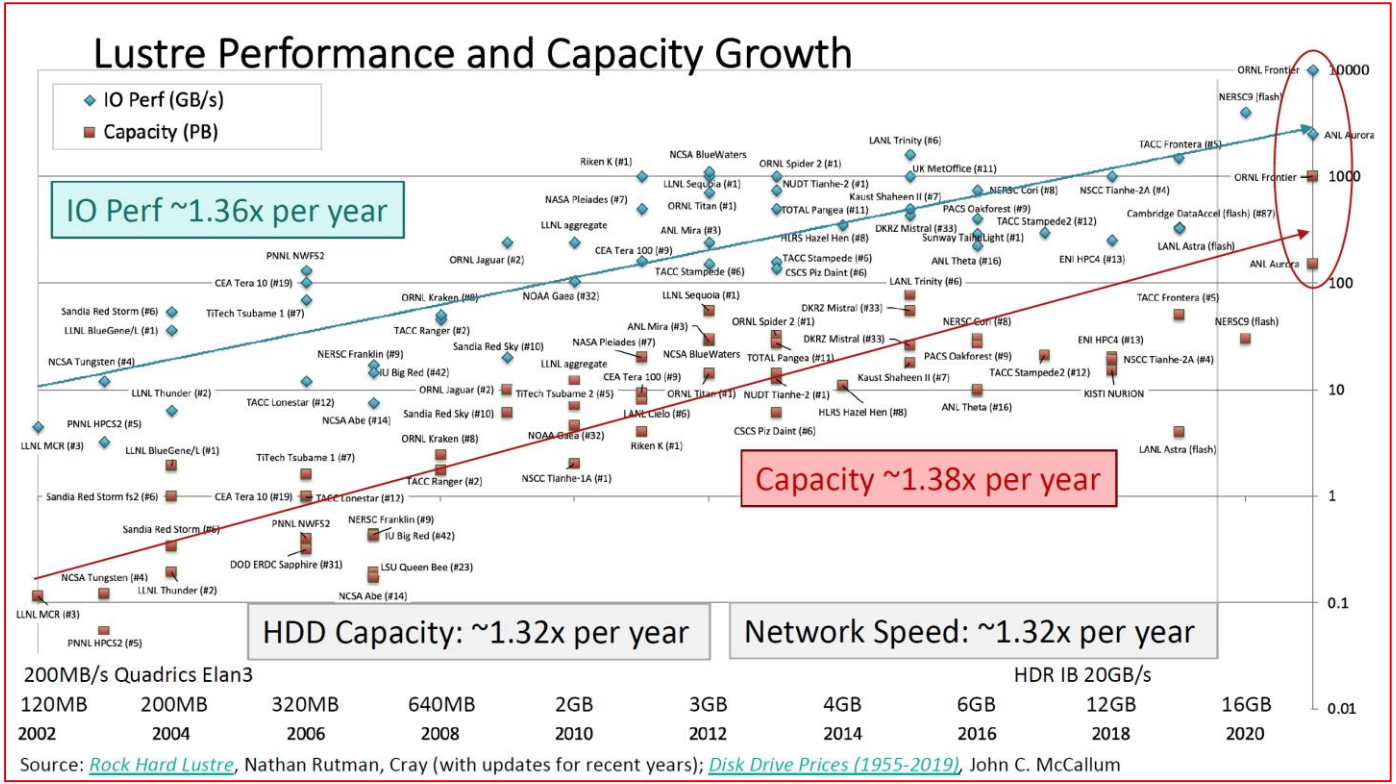
- デバイス・アーキテクチャ
  - 電力効率の改善（350Wでノード当たり35TFLOPSの達成は簡単ではない）
  - アクセラレータアーキテクチャの検討
  - 積層メモリの広帯域化と大容量化、ロジックとメモリの3次元積層化
- システムソフトウェア
  - コンテナ化されたアプリケーション／ユニカーネルの効率的な実行
  - データフレームワークの導入
  - ディスアグリゲーション、外部資源（クラウドやIoT）との連携
- ライブラリ・アルゴリズム
  - 新アーキテクチャや非均質なプロセッサへの対応
  - 混合精度演算の考慮
  - 従来型計算機と量子計算のハイブリッド利用

2021/5/27　次世代計算基盤検討部会　15

2028 年頃におけるCPU・ネットワーク・ストレージの予測性能

白書 P114-115(30MW, 80%の値より)	予測性能
メニーコアCPU型 総コア数	4.4x10E06 cores
総演算性能	1086 PFLOPS
DDR Mem 総バンド幅	137 PB/s
HBM Mem 総バンド幅	410 PB/s
総DDR 総容量	23 PB
総HBM 総容量	5 PB
NW インジェクションバンド幅	1/6 TB/s
総Storage 容量	3.45 EB
総ストレージスループット	34 TB/s

From ISC19 Workshop: Lustre – The Next 20 Years



https://hps.vi4io.org/_media/events/2019/hpc-iodc-lustre_next_20_years-dilger.pdf
2028年ストレージ見積もり:2018 年のSummit のストレージ性能（2.5TB/s, 250PB）を外挿

次世代ストレージへの要求と課題

FUJITSU

- システム処理能力がエクサスケールシステム以上
  - より大量のデータ処理を高速に実行する必要がある
  - エクサスケールの一桁以上の容量と性能(150PB->3.4EB, 1.5TB/s->34TB/s)
- より電力制約が求められる
  - 計算能力とデータ量が増加しても電力増加は最小限
  - どのように電力使用を絞るべきか
- 多種多様なアーキテクチャに対してデータを供給する必要がある
  - 多種多様なデータ供給手法が想定される
  - どのようなストレージアーキテクチャが必要か？

FUJITSU

次世代のHPCシステムと ストレージ・ファイルシステム

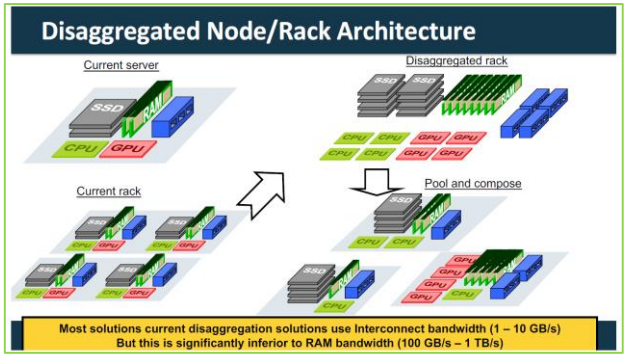
Keywords: Extreme Heterogeneity, Disaggregated, Cloud Native

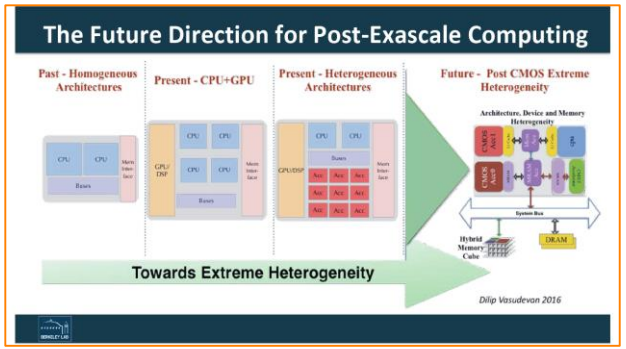
= Dynamically Configured Heterogeneous System

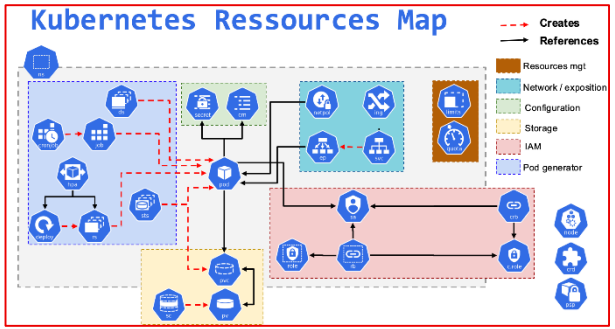
次世代のHPCシステムの方向性

FUJITSU

- 異種加速機構混合システム
  - 汎用CPU,GPU,FPGA, 専用加速機構, QC, NC,etc.
- 動的再構成による最適化システム
  - アプリケーション毎に加速機構構成変更
  - システムソフトウェア：On Demand再構成





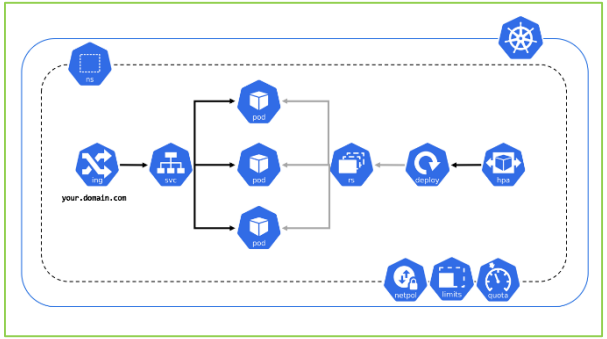


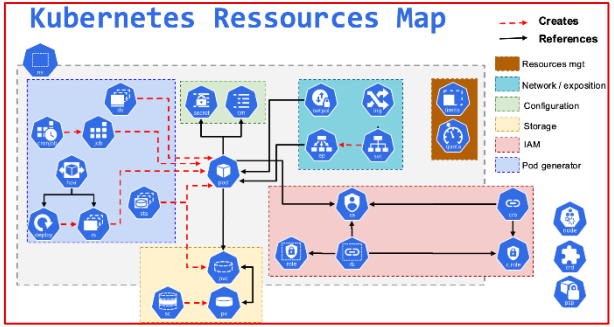
次世代のHPCシステムソフトウェアの方向性

FUJITSU

- マイクロサービスによる動的再構成
  - 異種ハードウェア、アプリケーション毎に最適化
  - 動的スケールアウトによる対負荷耐性向上
  - 再Deployによる耐故障性維持
- ユーザへの見え方は様々な形態に個別設定
  - 伝統的なbatch system,Web インターフェイス
  - Jupiter Notebook等のモダンインターフェイス
  - Vscodeなどの統合開発環境　など
- 提供サービスも様々な形態
  - 計算機センターサービス
  - IoTセンサー・リアルタイム処理サービス
  - イベントドリブン型処理サービス
  - 社会インフラと連動した動態解析サービス等

https://github.com/kubernetes/community/blob/master/icons/docs/k8s-exposed-pod.png



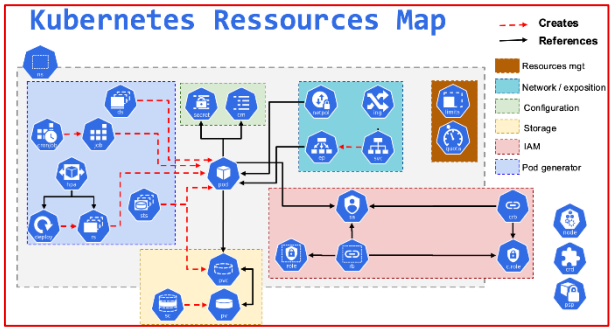


次世代のHPCシステムでのファイルシステムの方向性

FUJITSU

- マイクロサービスによる動的再構成
  - アプリケーション毎に動的構成最適化
  - 動的スケールアウトによる対負荷耐性向上
  - 再Deployによる耐故障性維持
- ユーザへの見え方は様々な形態に個別設定
  - ファイルAPI, メモリ、ストリーム、DB
  - Permanent, Temporary
  - プロセス内、グループプロセス内
- 提供サービスも様々な形態
  - メタデータ、ファイルデータ、データベース、トランザクション、キャッシュ、分散レプリカ他
  - ストリーム、On Memory、ブロック、アーカイブ、圧縮、暗号化他
  - 他クラウドサービス、広域ファイルシステム連携他

https://github.com/kubernetes/community/blob/master/icons/docs/k8s-exposed-pod.png



次世代HPCストレージ・ファイルシステム課題と方向性

FUJITSU

- クラウドネイティブ・マイクロサービス化
  - 既存サービスの再構成と段階的な移行
- エクサスケールの一桁以上の容量と性能
  - 性能は新デバイスの活用、新しいストレージアーキ
- どのように電力使用を絞るべきか？
  - 新デバイスとストレージアーキ
- どのようなストレージアーキテクチャが必要か？
  - 冗長なデータ移動を抑制する
    - Computational Storage, Host CPU Processing Bypass Offload
  - 主記憶の電力を抑制しながら大量データ処理
    - Integration of Storage Class Memory
    - データ保持電力の排除: RPMA(Remote Persistent Memory Access)
  - 一桁多いデータ処理
    - Computational Storage, 蓄積型からStreaming型への転換

Cloud Native, Offload 事例

FUJITSU

- Cloud Edition for Lustre software on AWS, Azure, GCP, OCI
  - CloudにおけるManaged Service
  - AWSにおいてはS3 storageのcacheとして動作
- HDF5: DAOS with Optane SCM, k8s連携
  - HDF as a service:
- FuzzBall:
  - Computing Service based on Kubernates
- Computational Storage
  - FPGA
  - Computation in SSD Controller
- RPMA+CXLによるSystem Wide Offload

HDF as a service

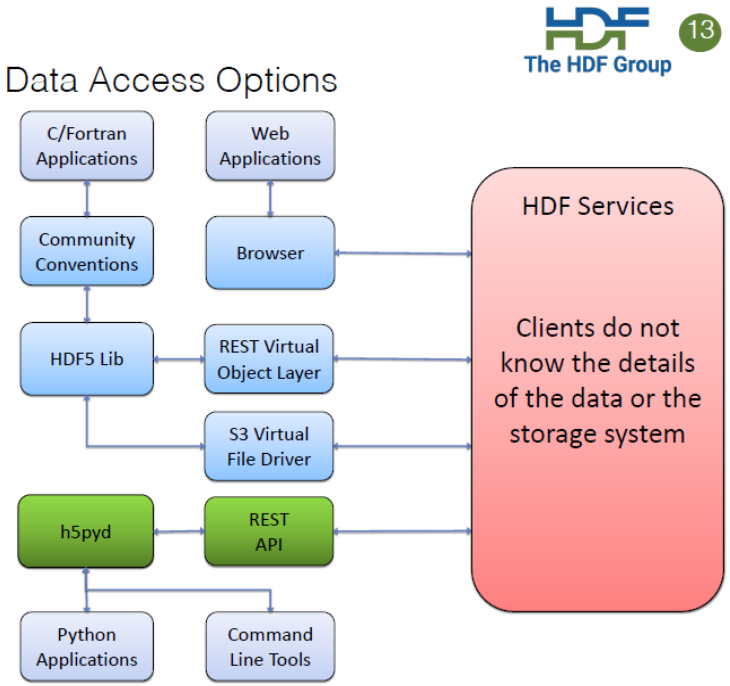
https://www.hdfgroup.org/wp-content/uploads/2019/09/ServerBasedHDF.pdf

FUJITSU

Architecture

Client SDKs for Python and C are drop-in replacements for libraries used with local files.

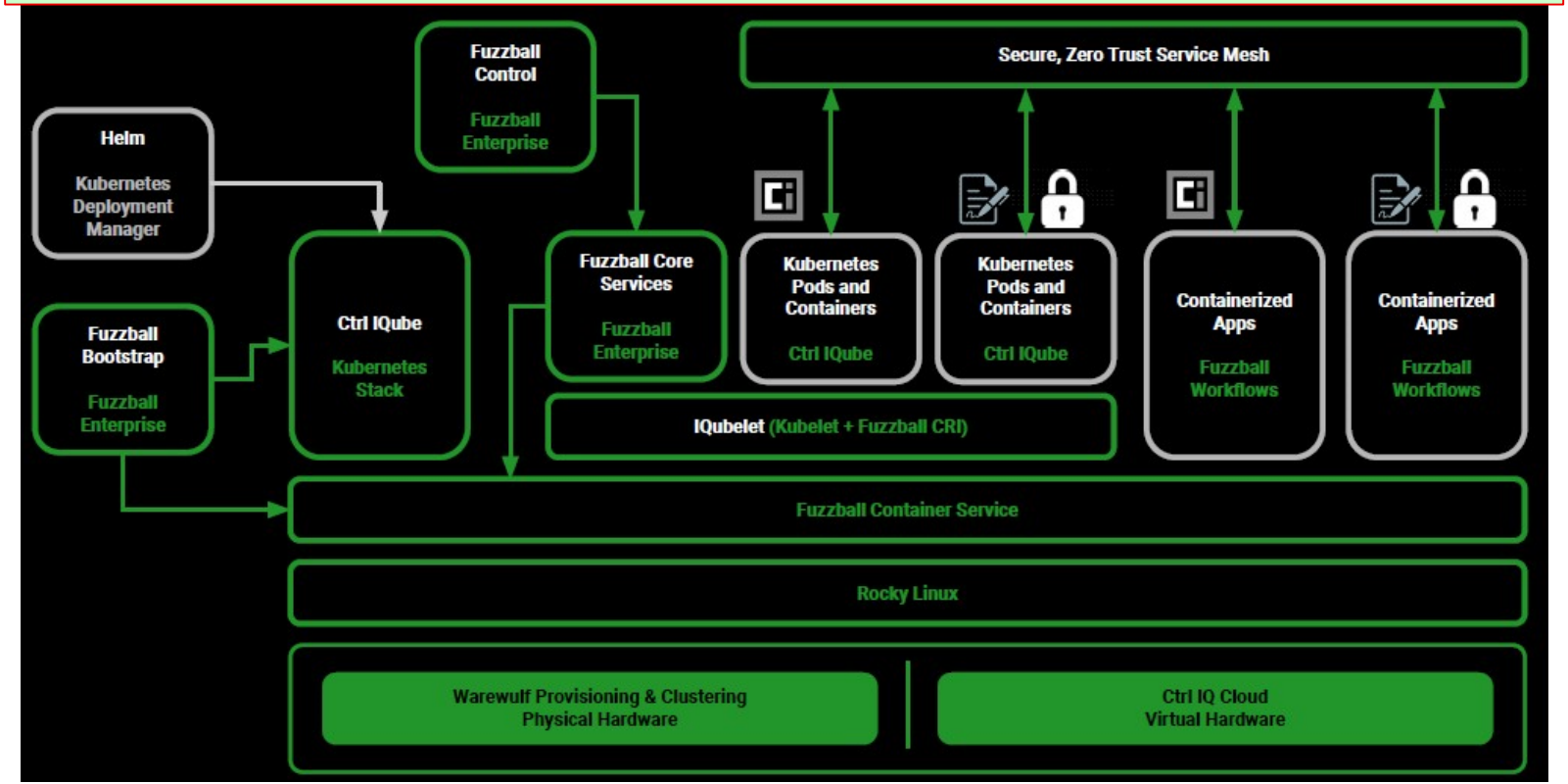
No significant code change to access local and cloud based data.



Fuzzball: the Ctrl IQ software stacks

https://www.nextplatform.com/2021/02/18/why-the-founder-of-centos-is-building-the-next-platform/

Greg Kurtzer: one of the co-founders of the CentOS Linux distribution,
the Singularity container environment for HPC workloads, the founder of the new Rocky Linux distribution



Computational Storage : FPGA

FUJITSU

Slides from FMS2018 https://www.flashmemorysummit.com/

Computational Storage: Alibaba Cloud

- HTAP: Hybrid Transactional/Analytical Processing
- No lag for analytics, low cost, unified storage
- ScaleFlux compute @ data enables:
  - HW accelerated real-time analytical processing from transactional data
  - Intelligent data management at the storage layer
  - Massive data movement, power, and latency reduction
  - Fast time-to-value with programmable HW

Alibaba Cloud POLARDB

"By bringing compute to the data, ScaleFlux is **transforming** the way we are architecting our **Flash storage infrastructure**.

We're looking to fully utilize the values of Computational Storage in order to **cost-effectively** scale **real-time analytics** across exploding transactional data sets, all the while delivering the **most responsive, cloud-native** user experience."

- GM of Alibaba Cloud

10X Transactional-Analytical Processing, **Half** the Flash Capacity

ScaleFlux Flash Memory Summit 2018 12

Computational Storage: Arm

FUJITSU

Slides from FMS2018 https://www.flashmemorysummit.com/

Flash Memory Summit

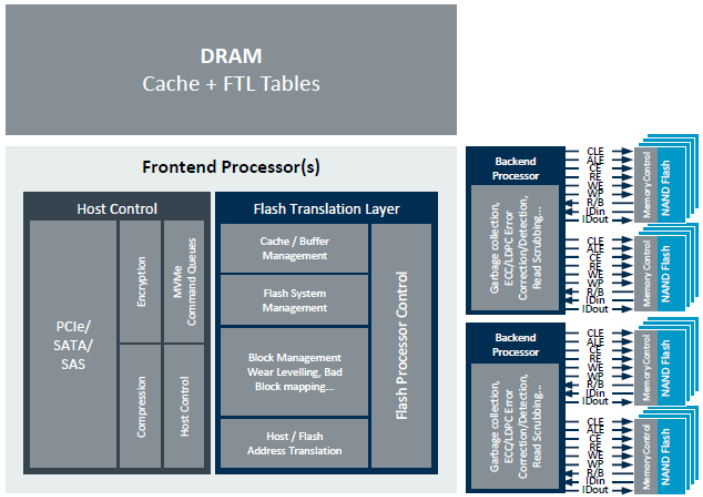
Compute in SSD controllers

Compute:

- Frontend: Host I/F + Flash Translation Layer
  - Cortex-R or Cortex-A series
- Backend: Flash management
  - Cortex-R or Cortex-M series
- Accelerators:
  - Encryption, LDPC,…
  - Arm NEON, ML, FPGA…

Memory: DRAM ~1GB for each 1TB of flash
Storage: 256GB to 64TB… flash storage
Interfaces: PCIe/SATA/SAS…

SSD SoC Functionality:



Flash Memory Summit 2018
Santa Clara, CA

arm

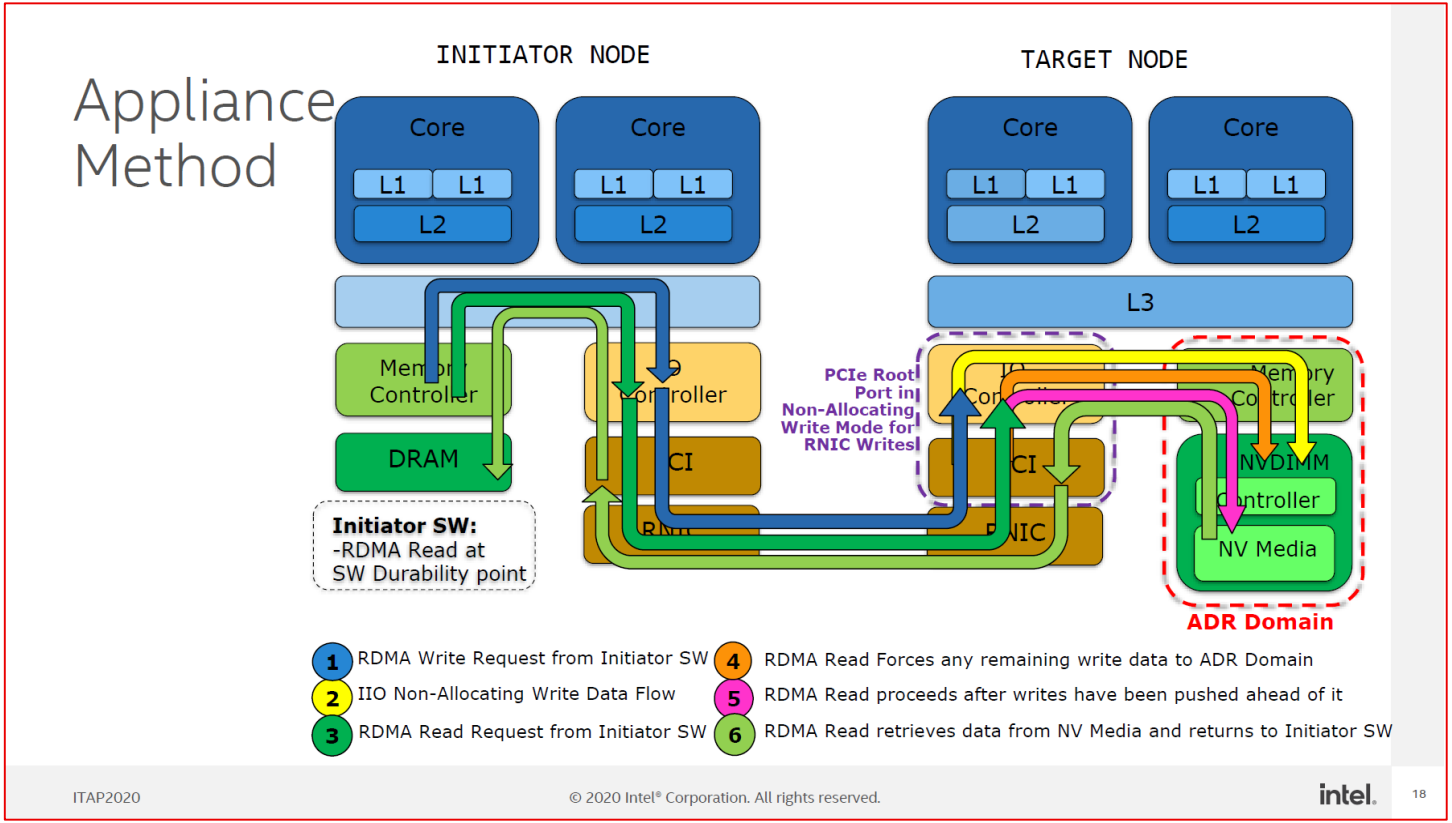
8

RPMA: Remote persistent Memory Access

https://github.com/pmem/rpma

FUJITSU

IEEE Cluster 2020 WS: ITAP2020-Session2-2-PMoRDMA.pdf



CXL: Compute Express Link

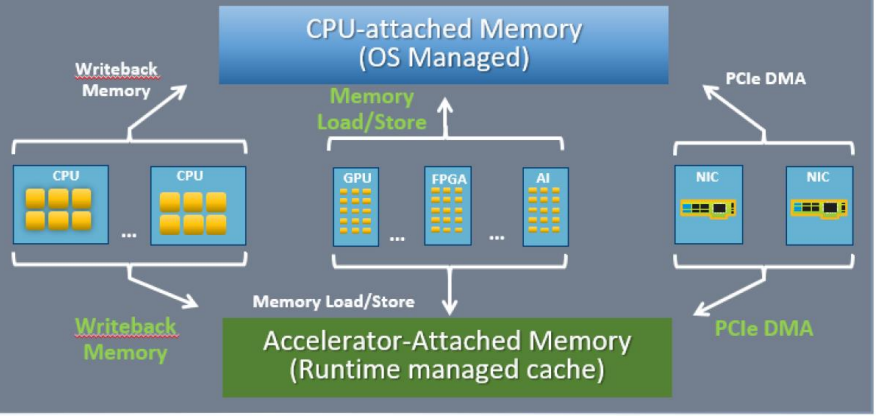
FUJITSU

Slides from OFA VWS2020 https://www.openfabrics.org/wp-content/uploads/2020-workshop-presentations/403-CXL-OFA-Virtual-Workshop-2020_FINAL.pdf

HETEROGENEOUS COMPUTING REVISITED – WITH CXL

- CXL enables a more fluid and flexible memory model
- Single, common, memory address space across processors and devices

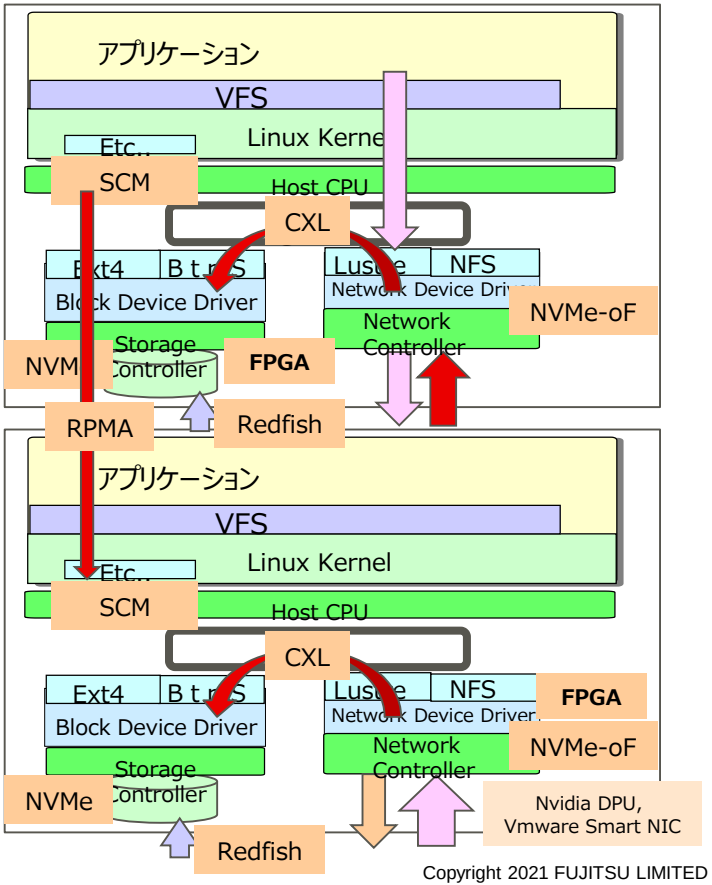
- More efficient population and update of operands
- More efficient extraction of results
- Memory resource "borrowing"
- User/Kernel level data access and data movement
- Low latency to memory, host to device and device to host



17

システム内機能オフロード化

- OS and Host CPU Processing Bypass+Offload
  - ユーザレベルクライアント：カーネルーユーザコピー排除
  - SCMファイルマップで直接Read/Write/Atomic Operation実行
- Computational Storage
  - アプリケーションの前処理をStorage Controller(+FPGA)実行
  - 前処理後のデータを主記憶に転送
  - FPGA処理によるOn the fly高機能データ転送
    - RAID, 暗号化、符号化、疎行列処理
- Computational Network Transfer
  - 転置、分配他通信処理をNIC/HCA+FPGAで実行
    - MPI集団通信実行
    - On the fly高機能データ転送、転置、Pack/Unpack、Strideデータ転送
    - NVidia DPU, VMware Smart NIC
- 冗長なコピーの削減:
  - ストレージ(他ノード)からユーザメモリ直結転送



まとめ

FUJITSU

- 次世代HPCシステムの方向性
  - Post Moore: Computing Beyond Moore’s Law
  - Disaggregated Computing
  - Cloud Native Computing
  - NGACI: Next-Generation Advanced Computing Infrastructure

- これからのシステム・ストレージ・ファイルシステム
  - Keywords: Extreme Heterogeneity, Disaggregated, Cloud Native
    = Dynamically Configured Heterogeneous System

Thank you