JLUG2020

FUJITSU

shaping tomorrow with you

LLIO/FEFS for the supercomputer Fugaku

Shinji Sumimoto, Ph.D.
FUJITSU LIMITED Dec. 4th, 2020

Copyright 2020 FUJITSU LIMITED

Outline of This Talk

FUJITSU

- The supercomputer Fugaku
  - Fugaku: Rank #1 Top500,HPCG,HPL-AI,Graph 500 @SC 20, Nov.2020

Benchmarks	Rank #1	Records	Rank #2	Records	#1/#2 Ratio
TOP500	Fugaku	442 PFLOPS	Summit	148.6	2.97
HPCG	Fugaku	16.0 PFLOPS	Summit	2.93	5.46
HPL-AI	Fugaku	2.00 EFLOPS	Summit	0.55	3.64
Graph500	Fugaku	102,956 GTEPS	Sunway	23,756	4.33

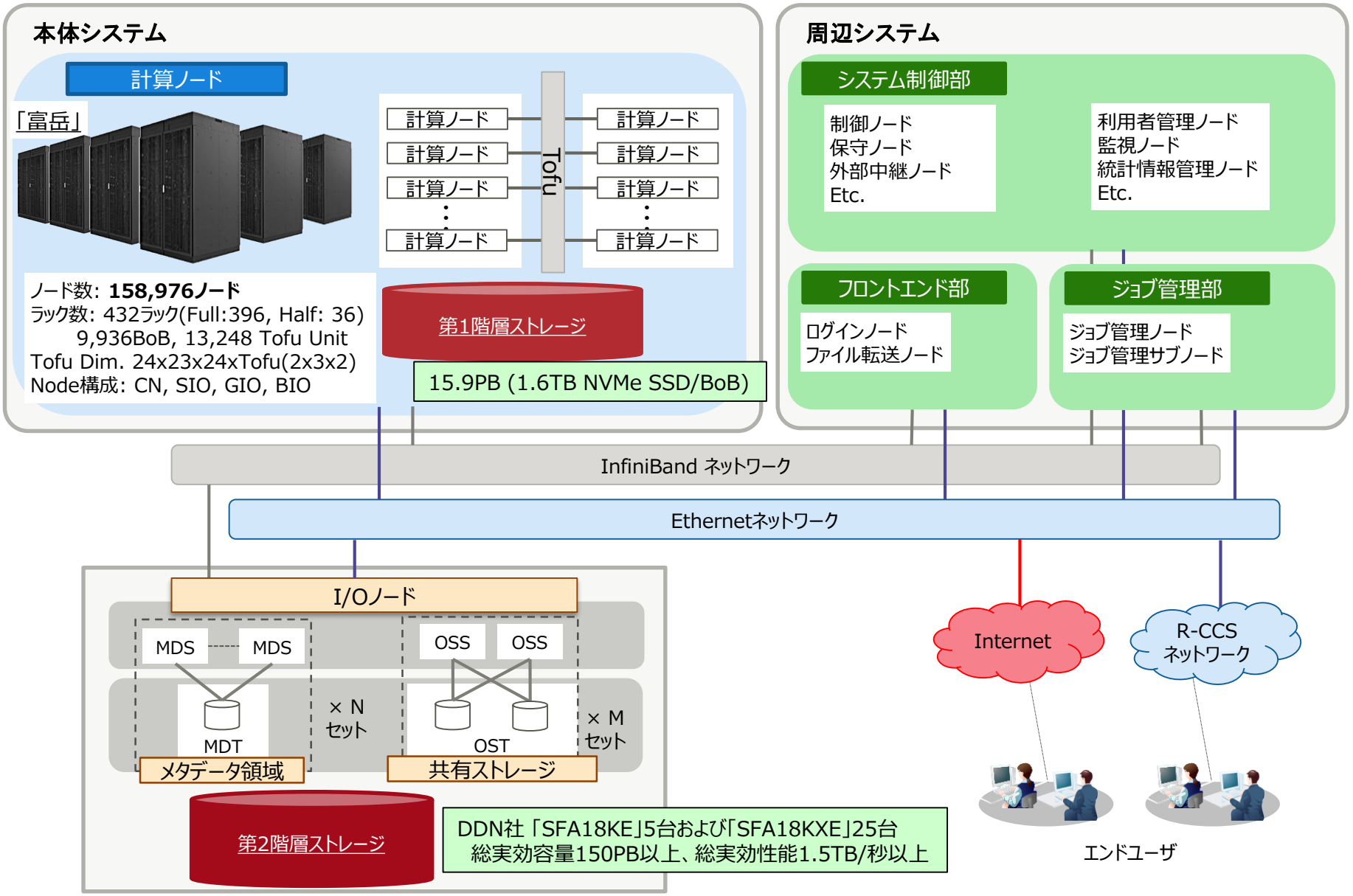
- Fugaku LLIO FEFS Update
  - Overview of the supercomputer Fugaku filesystem
  - Preliminary LLIO FEFS Performance Evaluation
- FEFS Lustre Status
  - Bug Fix Statistics Information

Copyright 2020 FUJITSU LIMITED

Overview of the supercomputer Fugaku

Copyright 2020 FUJITSU LIMITED

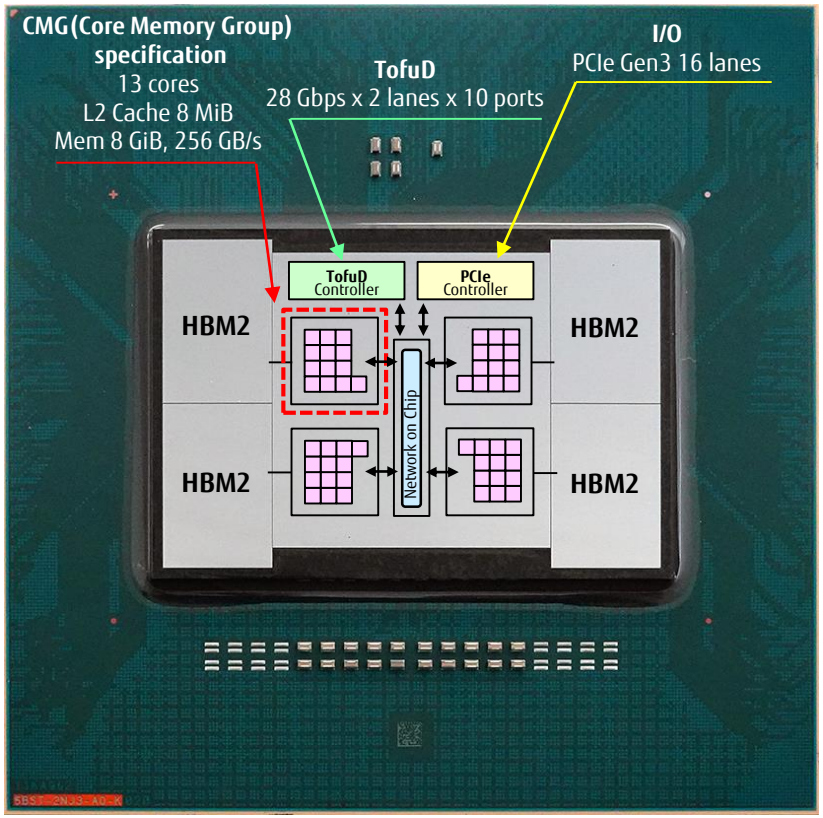
富岳システム全体構成



Copyright 2020 FUJITSU LIMITED

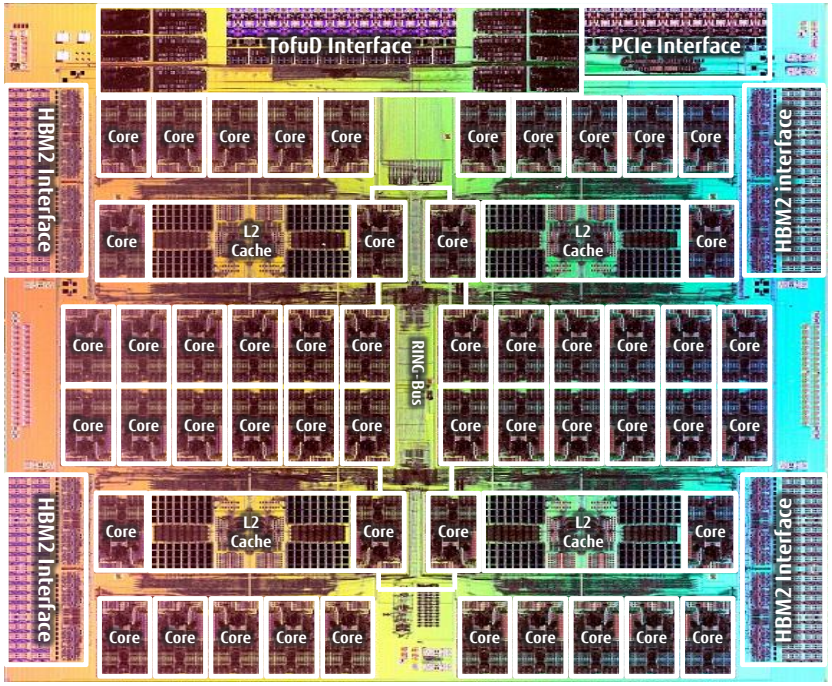
High Performance Arm CPU “A64FX”

FUJITSU



Architecture features

ISA	Armv8.2-A (AArch64 only) SVE (Scalable Vector Extension)
SIMD width	512-bit
Precision	FP64/32/16, INT64/32/16/8
# of cores	48 computing cores + 4 assistant cores (4 CMGs)
Memory	HBM2: Peak B/W 1024 GB/s
Interconnect	TofuD: 28 Gbps x 2 lanes x 10 ports



Copyright 2020 FUJITSU LIMITED

Software Stack for Fugaku

FUJITSU

- Fugaku system supports SBSA/SBBR
  - Keeping binary compatibility with the other Aarch64 based systems.

Fugaku Applications

FUJITSU Technical Computing Suite / RIKEN Advanced System Software

Management Software	Hierarchical File I/O Software	Programming Environment
System management for highly available & power saving operation	**Lightweight Layered IO-Accelerator LLIO**	XcalableMP
Job management for higher system utilization & power efficiency	**Lustre-based distributed file system FEFS**	MPI (Open MPI, MPICH)
		OpenMP, COARRAY, Math Libs.
		Compilers (C, C++, Fortran)
		Debugging and tuning tools

arm

Linux OS (RHEL)/ McKernel (Lightweight Kernel)

Fugaku System Hardware

Copyright 2020 FUJITSU LIMITED

K computer vs. Fugaku(1 Peta FLOPS System)

FUJITSU

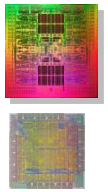
K computer

- 80x compute racks & 20x disk racks

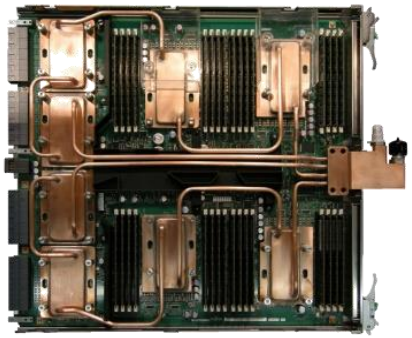
Fugaku

- 1x rack w/ SSDs

SPARC64 VIIIfx



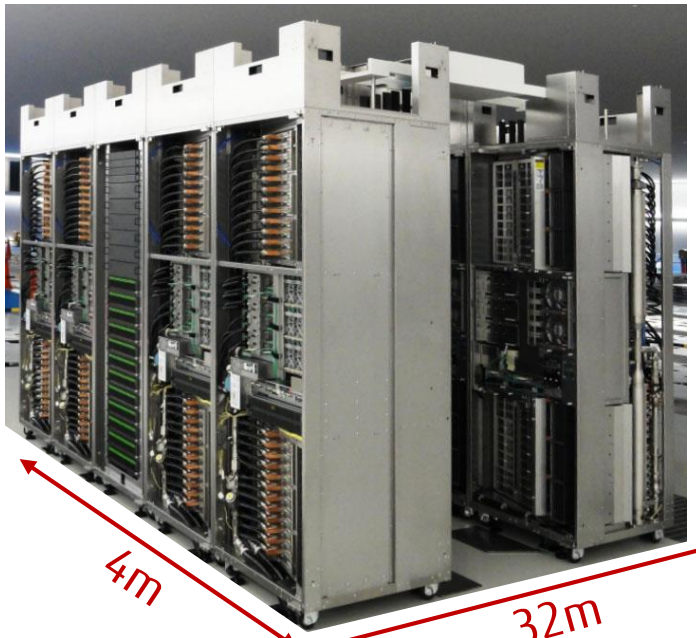
+ ICC

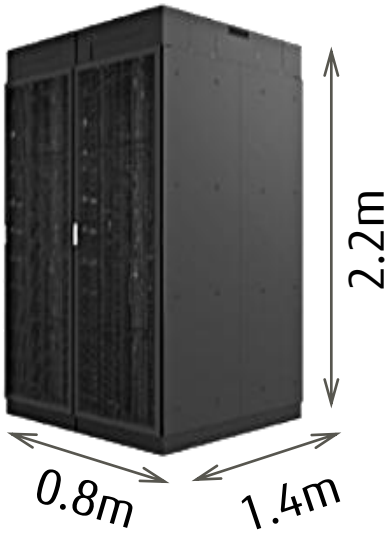


	K computer	Fugaku
Compute nodes	7,680(=96x80)	384
IO nodes	4,80(=6x80)	
Footprint (m^2)	128(=4x32)	1.1
	SPARC Linux	Arm Linux

A64FX

More applications as well as system software will come in collaboration with Open Source Community





Copyright 2020 FUJITSU LIMITED

The supercomputer Fugaku filesystem design

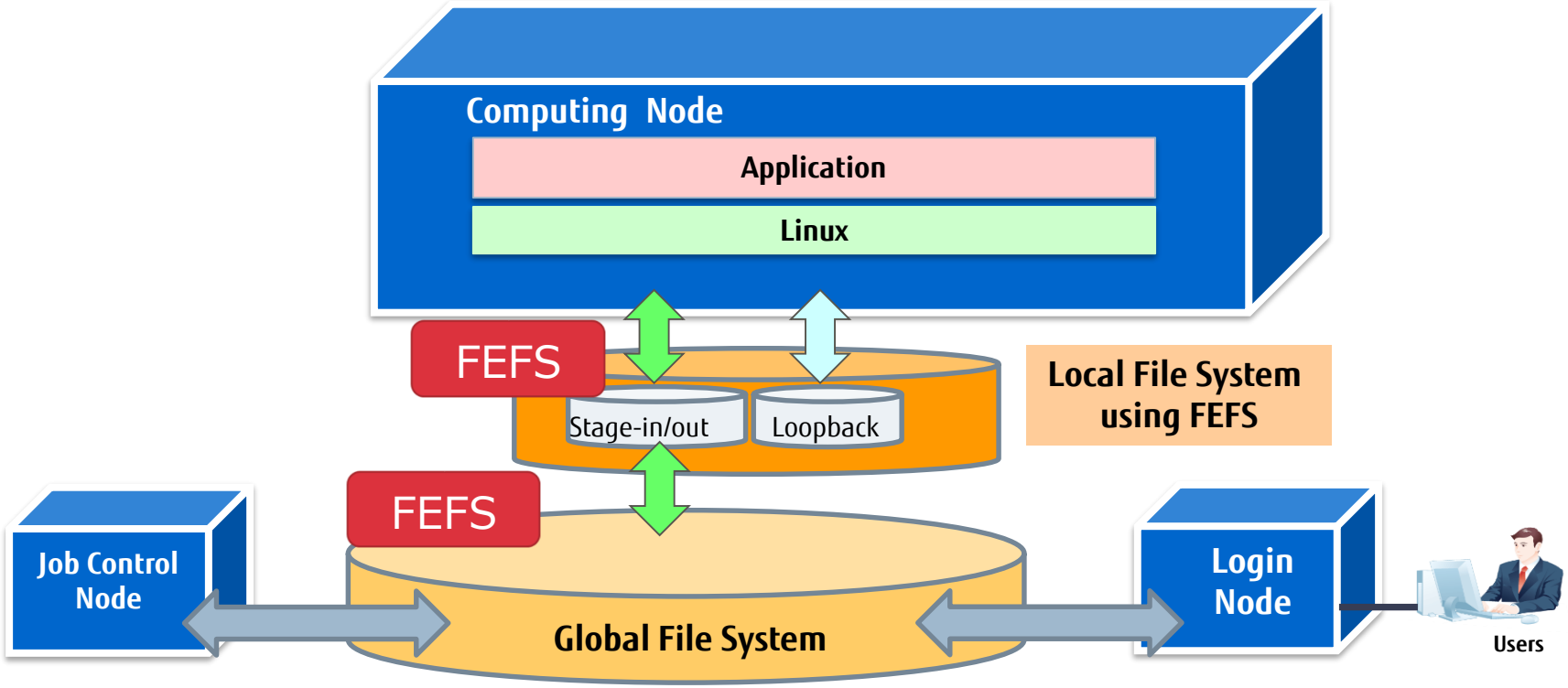
Copyright 2020 FUJITSU LIMITED

K computer File System Design

FUJITSU

JLUG2018

- Pros:
  - Stable Application Performance for Jobs
- Cons:
  - Requiring three times amount of storage which a job needs
  - Pre-defining file name of stage-in/out processing lacks of usability
  - Data-intensive application affects system usage to down because of waiting pre-staging-in/out processing



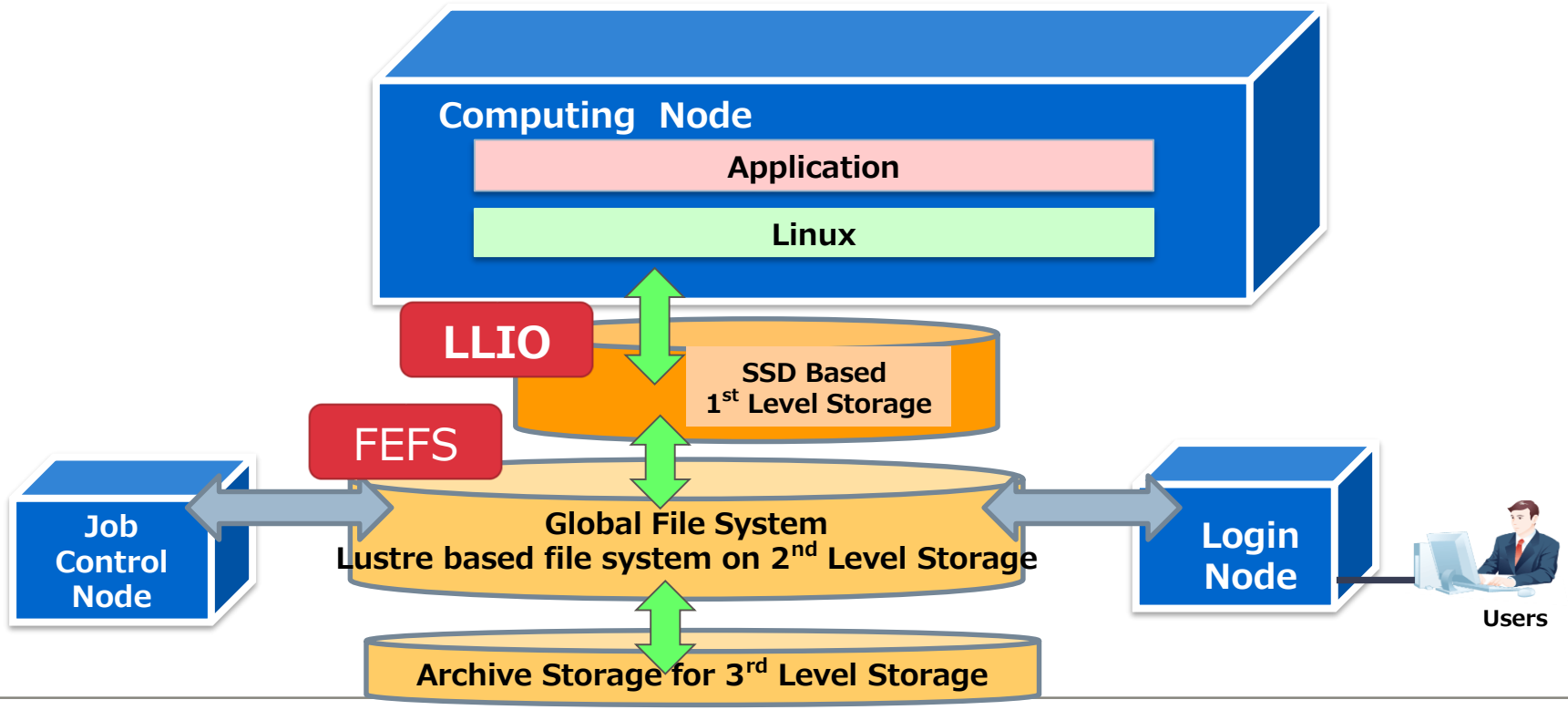
Copyright 2020 FUJITSU LIMITED

Fugaku File System Design

FUJITSU

JLUG2018

- Introducing three level hierarchical storage.
  - 1st level storage: Accelerating application file I/O performance (Local File System)
  - 2nd level storage: Sharing data using Lustre based file system (Global File System)
  - 3rd level storage: Archive Storage (Archive System)
- Accessing 1st level storage as file cache of global file system and local storage
  - File cache on computing node is also used as well as 1st level storage



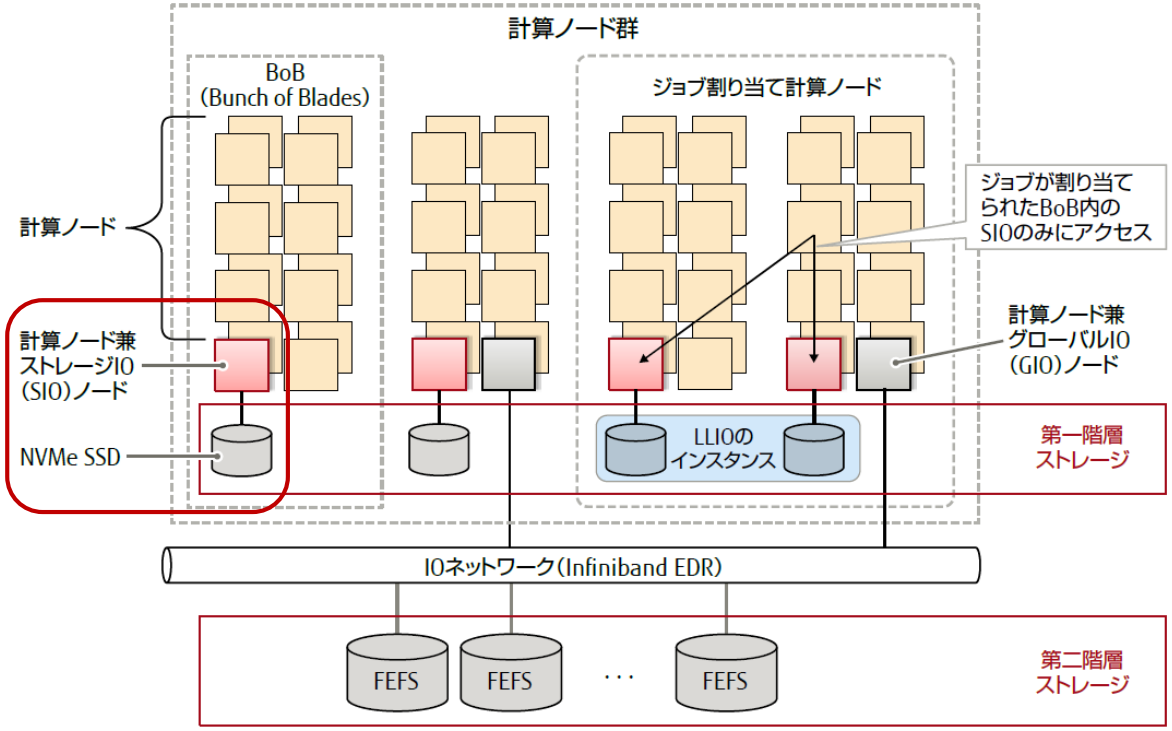
Copyright 2020 FUJITSU LIMITED

富岳の階層化ストレージ

FUJITSU

■ ユーザビリティの向上とアプリケーションのファイルIO最適化を実現

	ファイルシステム	目的	デバイス	特徴
第１階層ストレージ	LLIO 新規開発	ジョブ実行専用の高速領域	NVMe SSD	・用途に応じた3種のIO領域 ・16ノードごとにサーバSIOを配置
第２階層ストレージ	FEFS Lustreベース	大容量の共有領域	HDD	・「京」で開発したFEFSを改版



富士通テクニカルレビュー2020年No.3 ISSN 2435-6085 特集：スーパーコンピュータ「富岳」
https://www.fujitsu.com/jp/about/resources/publications/technicalreview/2020-03/

- ■ 第１階層：LLIO
  - ■ SSDによる高速化
    - BoBあたり1 SSD
      =システム全体 9,936 SSD
  - ■ 3種類のIO領域の提供
    - ノード内テンポラリ
    - 共有テンポラリ
    - 第２階層キャッシュ
- ■ 第２階層：FEFS
  - ■「京」実績のLustreベース
  - ■ 対故障性、保守性を改善

Copyright 2020 FUJITSU LIMITED

第１階層ファイルシステム：LLIO

FUJITSU

- LLIOの構成
  - BoB単位(16ノード毎)にLLIOサーバ SIO を配置。システム全体 9,936 SIO。
  - 一時領域およびキャッシュのため、サーバは非冗長構成。
- LLIOの処理概要
  - ジョブ単位に動的にLLIOを生成、マウントして利用。ジョブ終了後は解放。
  - 異常時はジョブ異常終了、影響をジョブに局所化。
- 各領域の処理概要
  - ノード内テンポラリ:BoB内SIOサーバ上でノード内の一時ファイルを管理
  - 共有テンポラリ：ジョブ内の複数SIOサーバ群でノード間の大規模データ共有
  - 第２階層キャッシュ:BoB内SIOサーバで第２階層FSデータをキャッシュ

領域	名前空間	サーバ	用途
ノード内テンポラリ	ノード内共有	同一BoB内SIO	一時ファイル
共有テンポラリ	ジョブ内共有	同一ジョブ内SIO群	一時ファイル
第２階層キャッシュ	第２階層ファイルシステム透過	同一ジョブ内SIO群	保存ファイル

Copyright 2020 FUJITSU LIMITED

第２階層ファイルシステム：FEFS

FUJITSU

■ 開発方針

- ■ Lustre 2.10 ベースで障害修正・コードエンハンス状況に応じてUpdate
- ■ コミュニティコードベース利用が基本：将来のUpdate障壁緩和

■ 開発・改善内容

- ■ 耐故障性向上
  - ●サーバ異常発生からファイルI/O 再開までの時間短縮：LNet MultiRail改善
- ■ 保守性向上
  - ●ファイルシステム復旧時間短縮：ファイルシステムパラメータ最適化
- ■ 負荷耐性向上
  - ●MDS高負荷防止: DNE採用によるメタデータ処理分散
- ■ アクセス公平性確保
  - ●QoS機能改善
- ■ 省メモリ・省CPU・OSノイズ排除
  - ●省資源で動作するためシステムパラメータ最適化

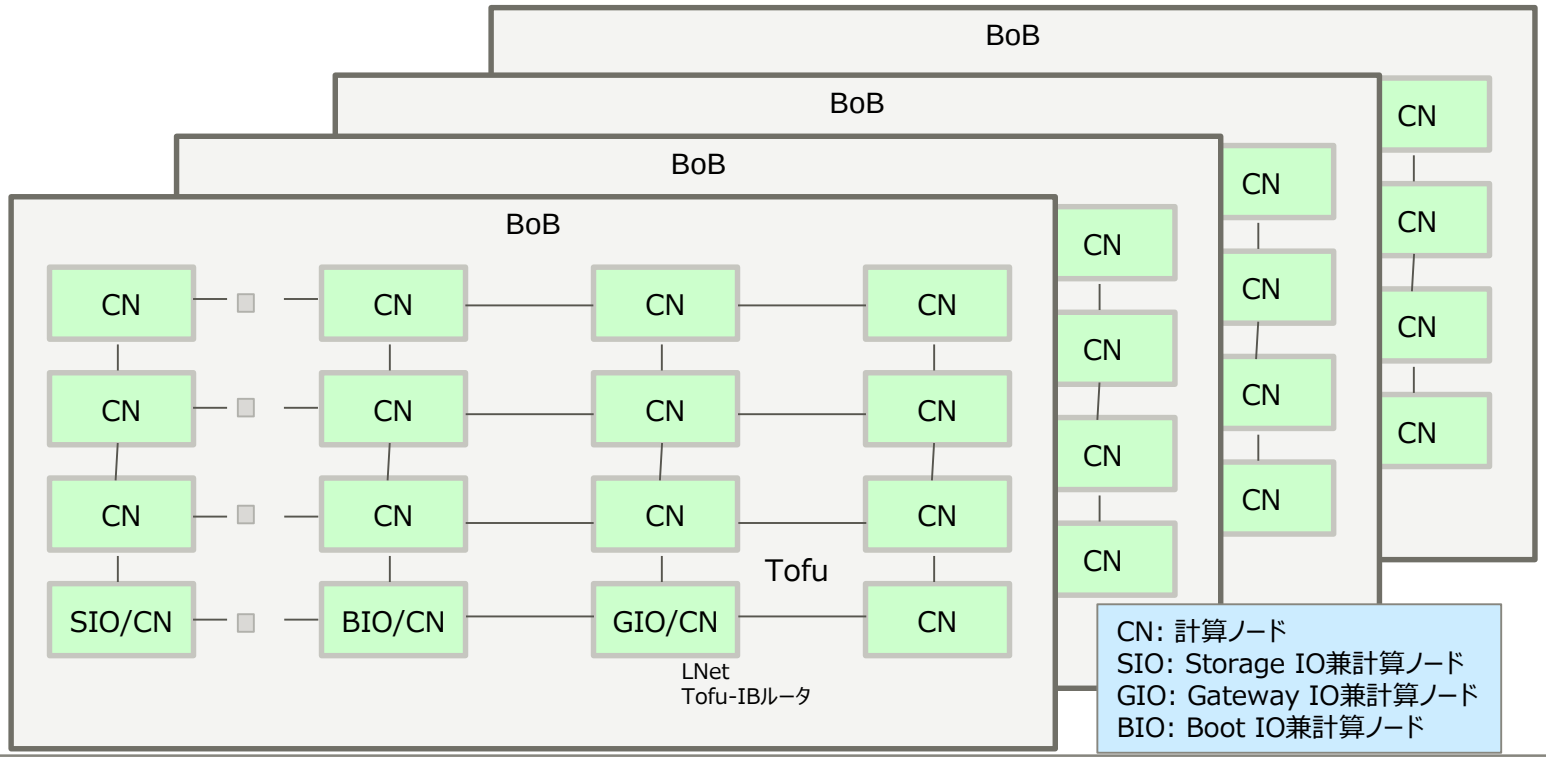
Copyright 2020 FUJITSU LIMITED

Performance of the First Layer: LLIO File System

Copyright 2020 FUJITSU LIMITED

LLIO 性能評価

- 目的：第１階層ストレージの多クライアントのスループット性能評価
- 評価環境：富岳環境　12万超のクライアント環境
  - ノード内テンポラリ　：158,963ノード（全9,936 SIO）
  - 共有テンポラリ、２階層キャッシュ：124,000ノード（7,750 SIO）
- 評価対象：プロセス毎ファイルのスループット性能



Copyright 2020 FUJITSU LIMITED

LLIO性能:ノード内テンポラリ

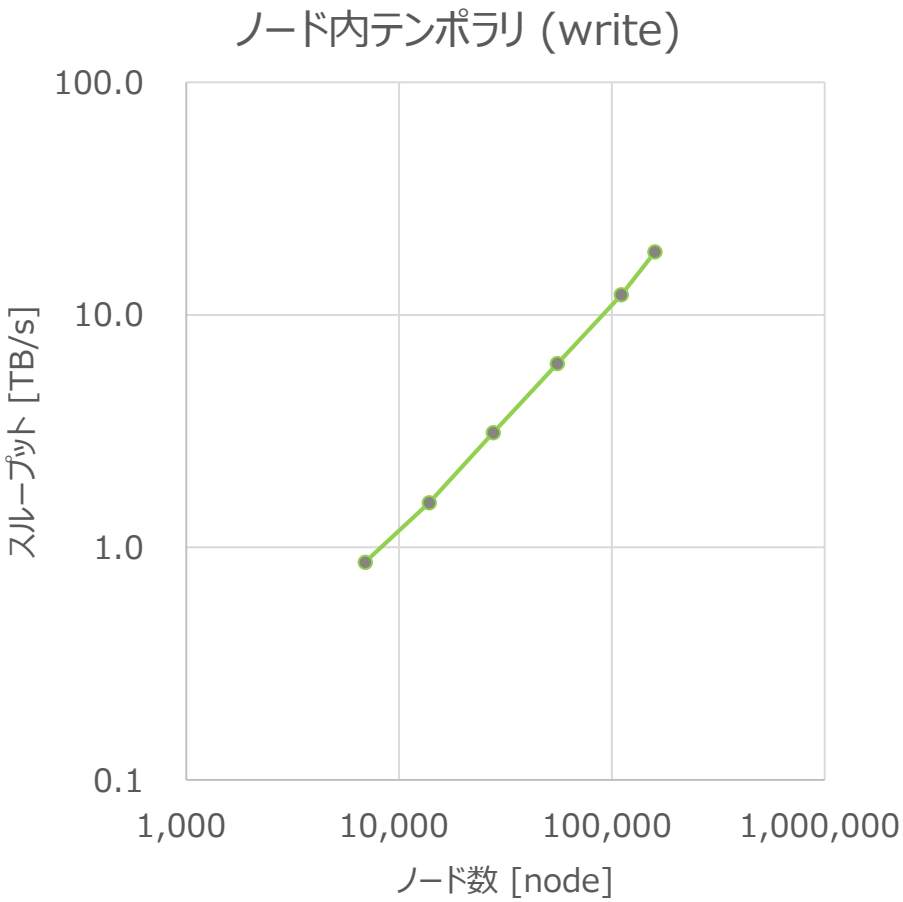
FUJITSU

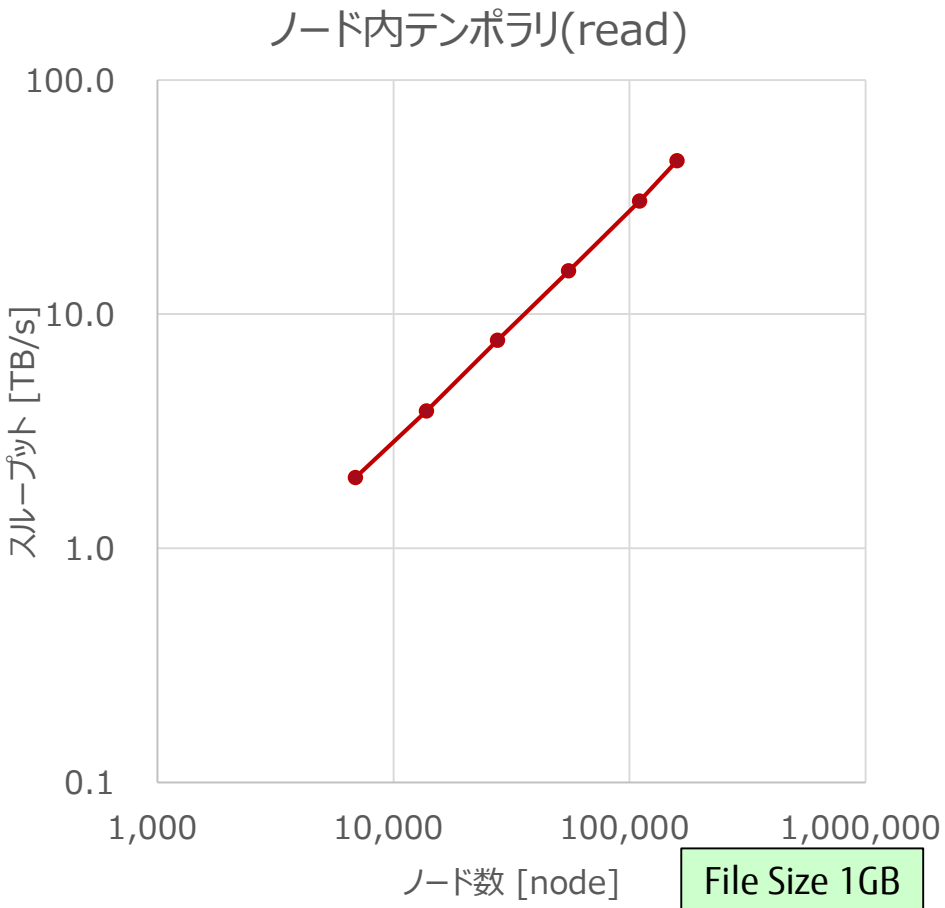
- IOR WRITE(File/Proc): 18.6 TB/s@158,963 nodes
- IOR READ(File/Proc): 45.2 TB/s@158,963 nodes

ストレージ分散型のためハードウェア性能に比例した性能達成

ORNL Summit: (物理素性能)
WRITE: 9.7 TB/s
READ:26.7 TB/s

World Record
on Single System





File Size 1GB

Collaborative work with RIKEN on the supercomputer Fugaku

Copyright 2020 FUJITSU LIMITED

LLIO性能:共有テンポラリ

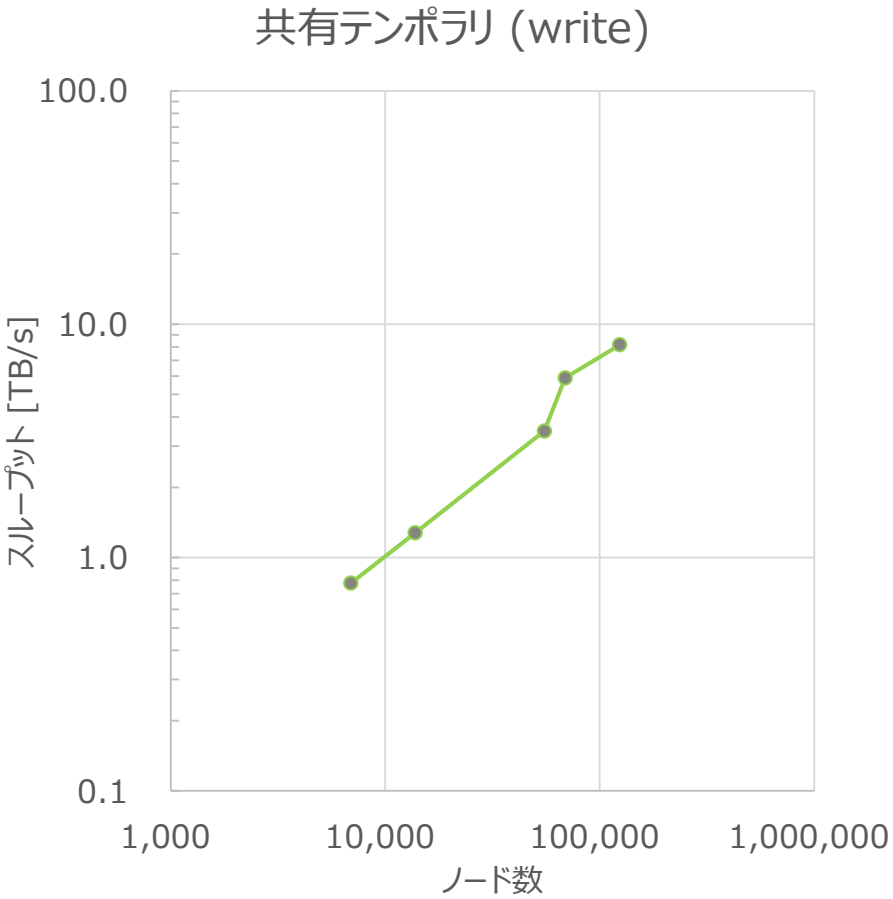
FUJITSU

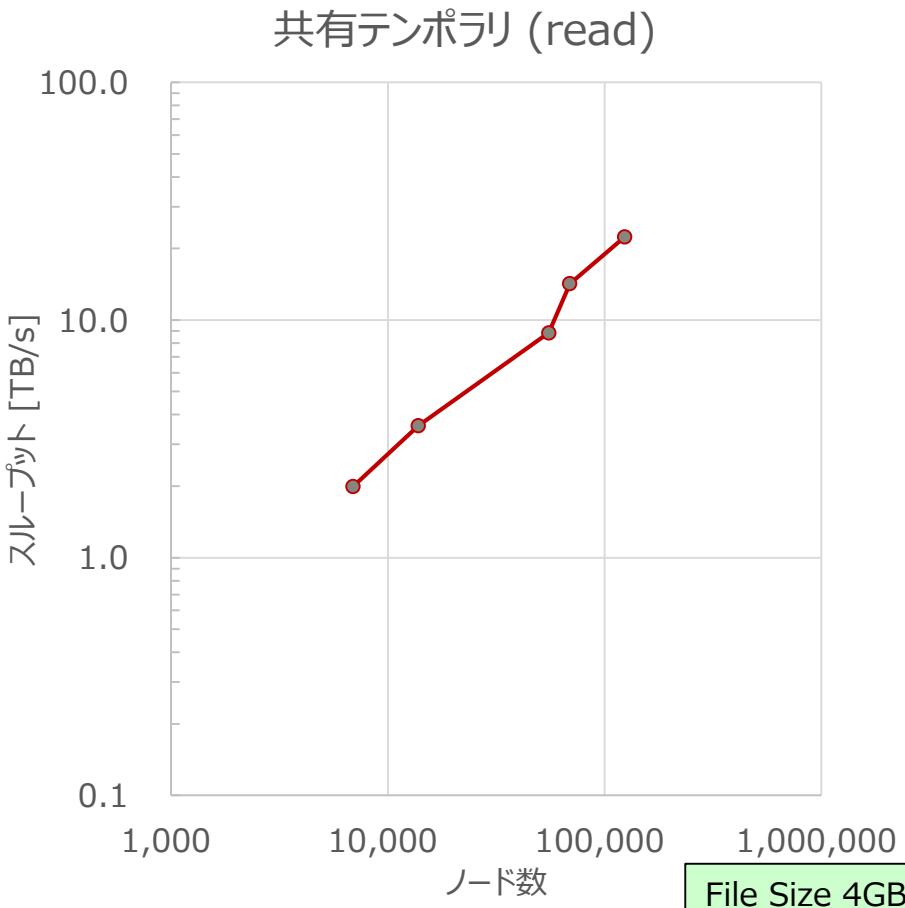
- IOR WRITE(File/Proc): 8.16 TB/s@124,000 nodes
- IOR READ(File/Proc): 22.4 TB/s@124,000 nodes

フルスケールで性能向上余地あり

ジョブ内共有ファイルシステムとして京より一ケタ高い性能を実現

World Record on Single System





File Size 4GB

Collaborative work with RIKEN on the supercomputer Fugaku

Copyright 2020 FUJITSU LIMITED

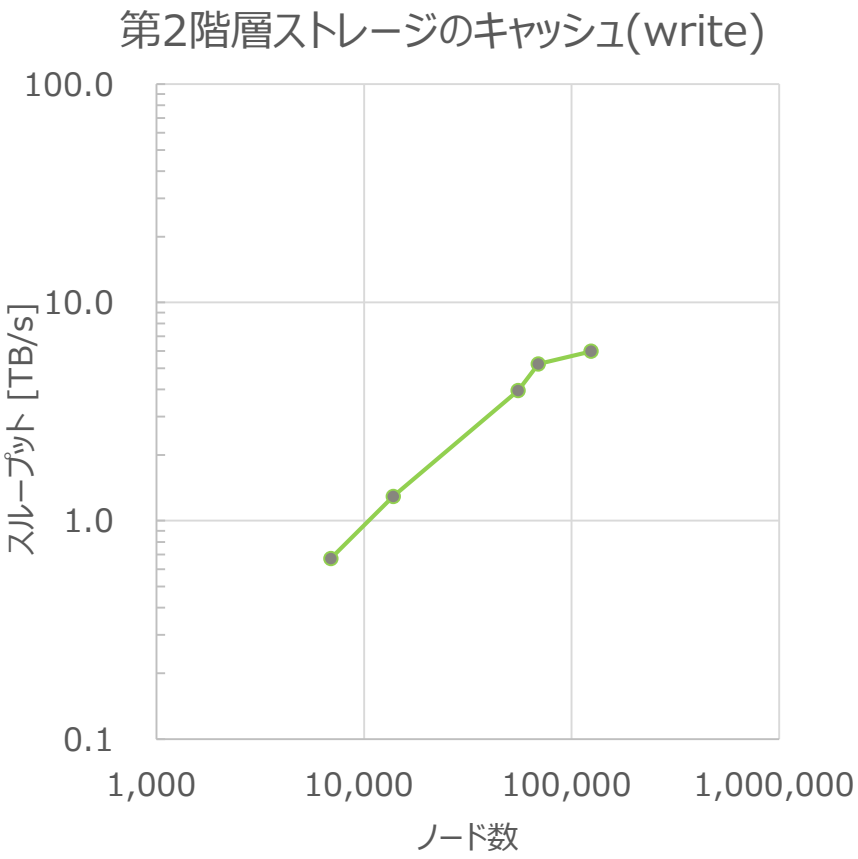
LLIO性能:第 2 階層キャッシュ

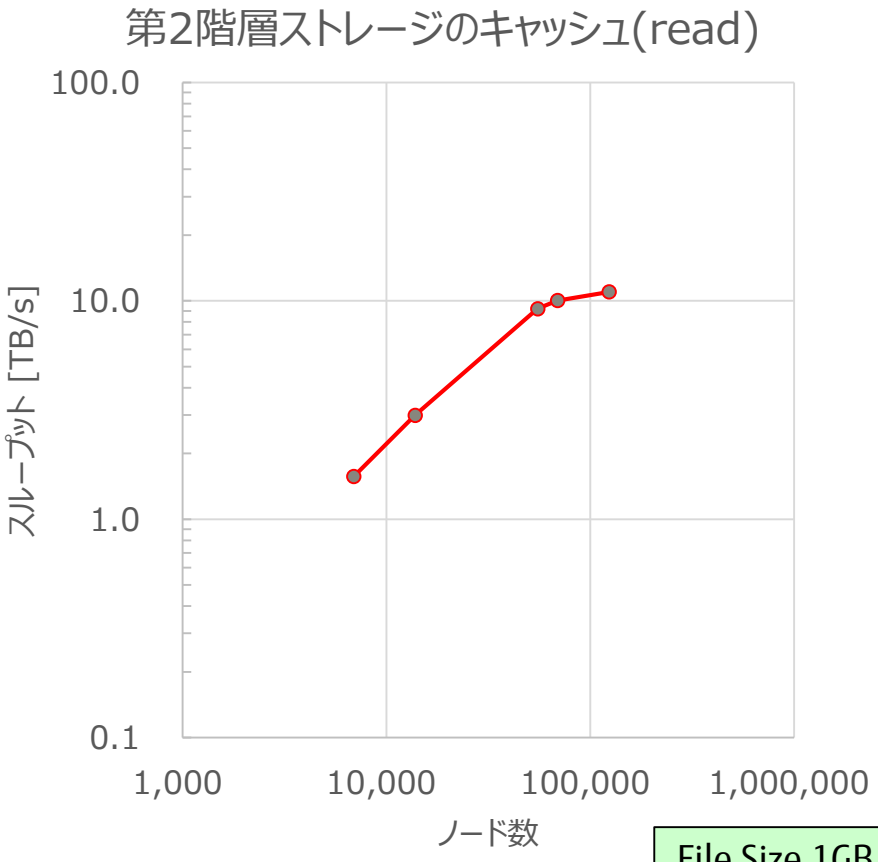
FUJITSU

- IOR WRITE(File/Proc):　5.97 TB/s@124,000 nodes
- IOR READ(File/Proc):　11.0 TB/s@124,000 nodes

フルスケールで性能向上余地あり

共有テンポラリ性能を目標に最適化実施中





File Size 1GB

Collaborative work with RIKEN on the supercomputer Fugaku

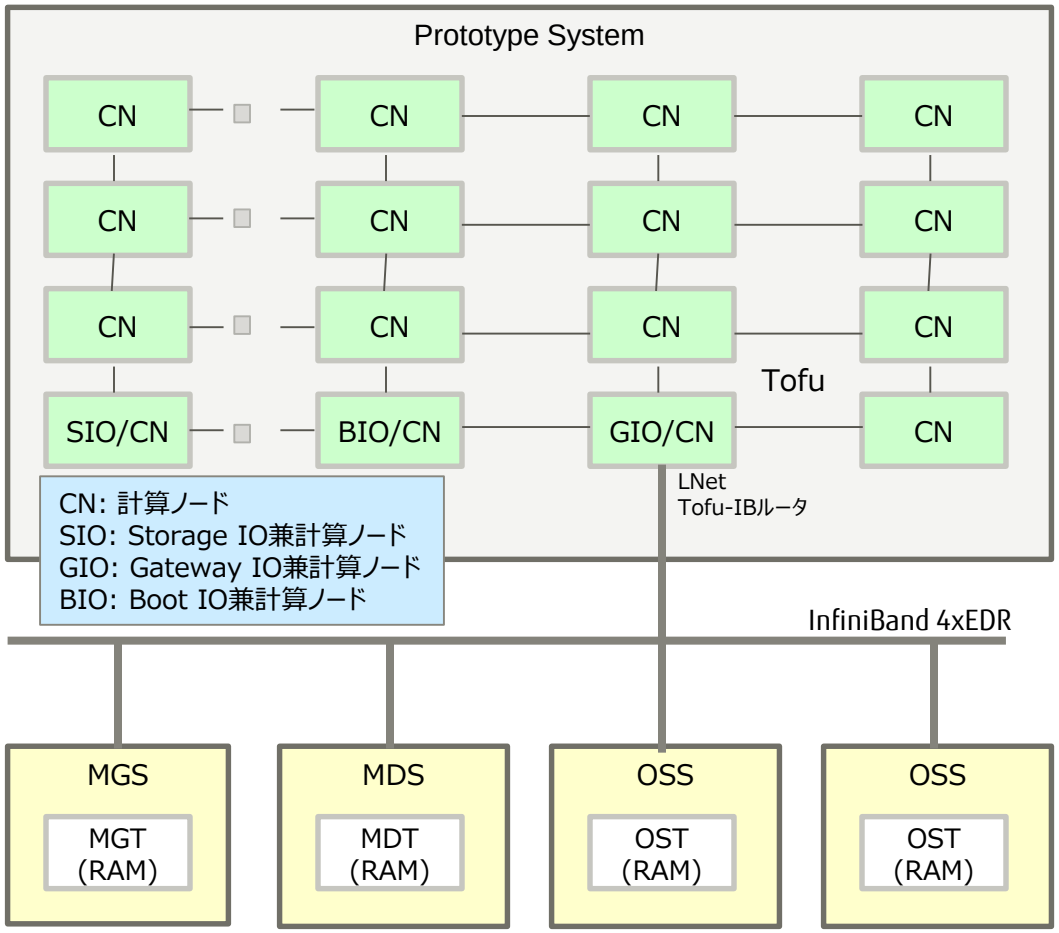
Copyright 2020 FUJITSU LIMITED

Performance of the Second Layer: FEFS File System

Copyright 2020 FUJITSU LIMITED

FEFS基礎性能評価

- 目的：第２階層ストレージの全系性能推定
  - LNet Tofu-IBルータ利用: LNetルーティング環境下における性能検証
- 評価環境：プロトタイプシステム(1BoB, RAM Disk)
- 性能評価: 1GIOあたり性能
  - システム全体:828 GIO

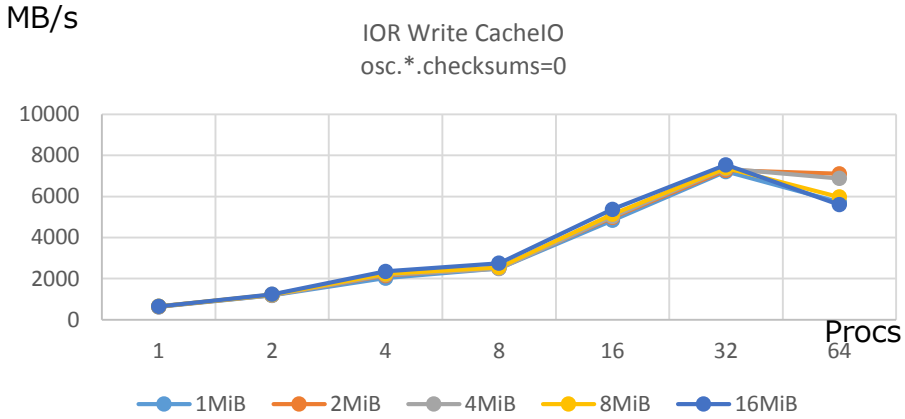


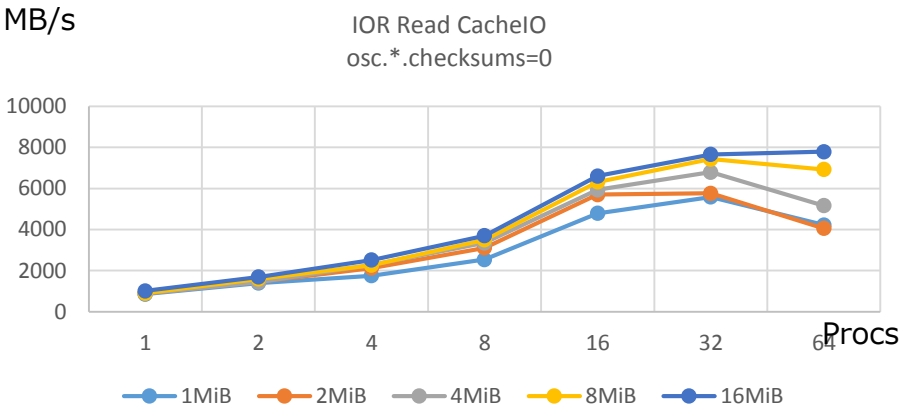
Copyright 2020 FUJITSU LIMITED

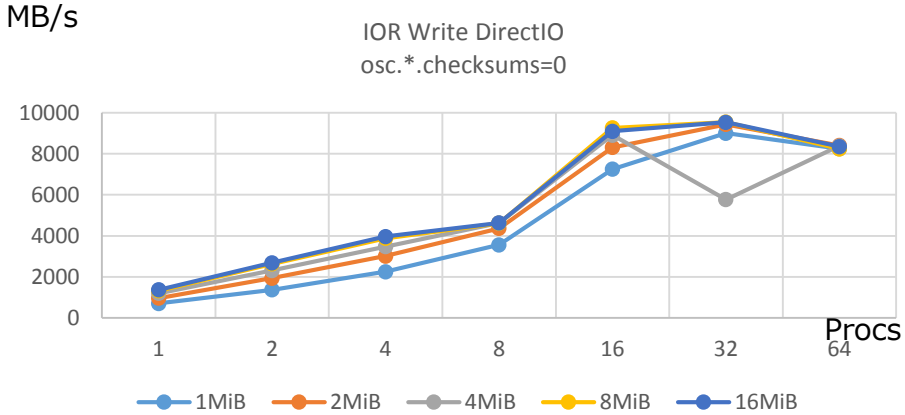
FEFSの基礎性能

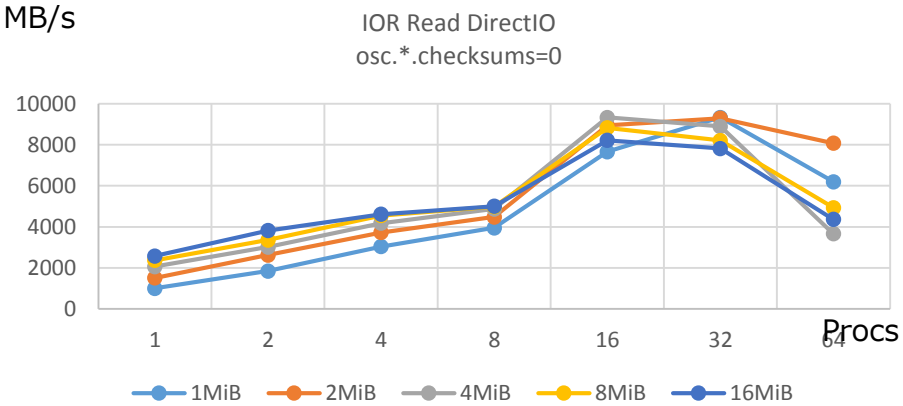
- Cache IO, DirectIO共に8GB/s以上の転送性能
- DirectIOではIB性能MAXである10GB/sの転送性能

⇒第２階層ストレージ目標1.5TB/sに対し、GIOあたり 8GB/sを確認。
1500GB/s ÷（828GIO x 8GB/s）＝ GIO全体の23%で充足見込み









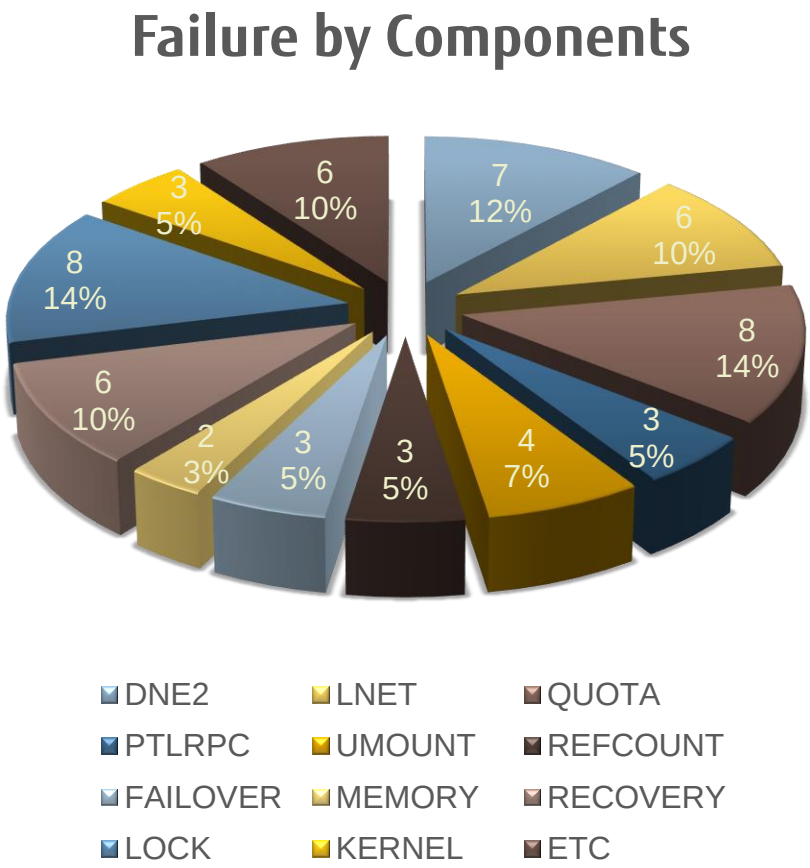
第２階層ストレージ：DDN社「SFA18KE」5台および「SFA18KXE」25台
総実効容量150PB以上、総実効性能1.5TB/秒以上

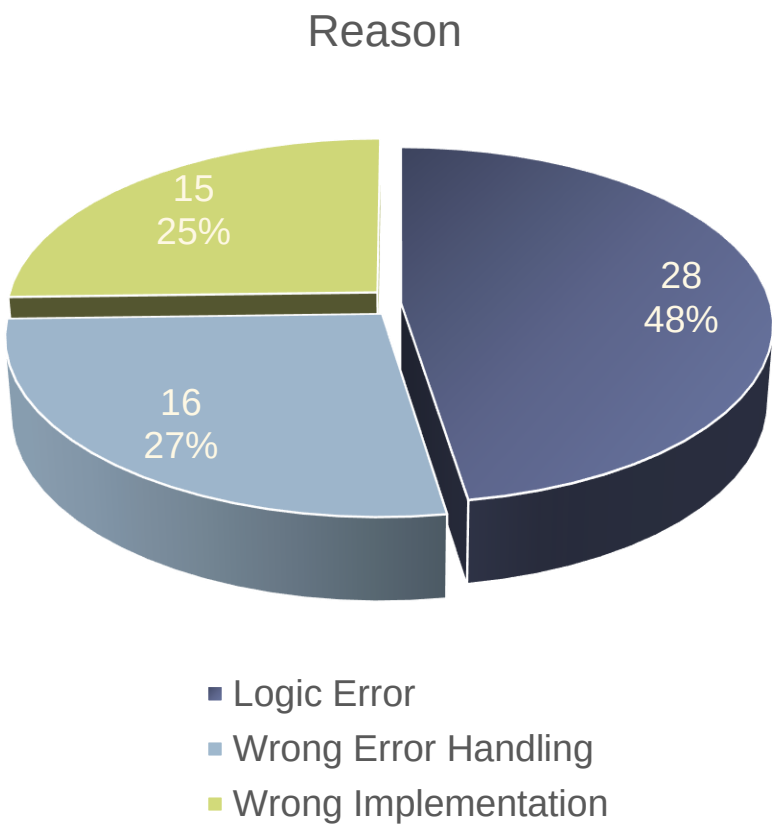
Copyright 2020 FUJITSU LIMITED

FUJITSU

Failure Analysis on FEFS (Lustre 2.10.8)

- Total Fix 59: 5 in LU-ticket (Known bugs)





Copyright 2020 FUJITSU LIMITED

おわりに

FUJITSU

- まとめ
  - プロセス毎ファイルのスループット素性能について、十分性を確認。
  - 2021年度の富岳共用開始に向けたアプリケーション評価を進め、更なる性能向上・安定化を目指す。
- コミュニティへのフィードバック
  - 富岳を優先しており、Lustreコミュニティへのフィードバックは遅延
  - 来年以降、コミュニティへのBug fix、フィードバックに貢献予定

Copyright 2020 FUJITSU LIMITED

FUJITSU

shaping tomorrow with you