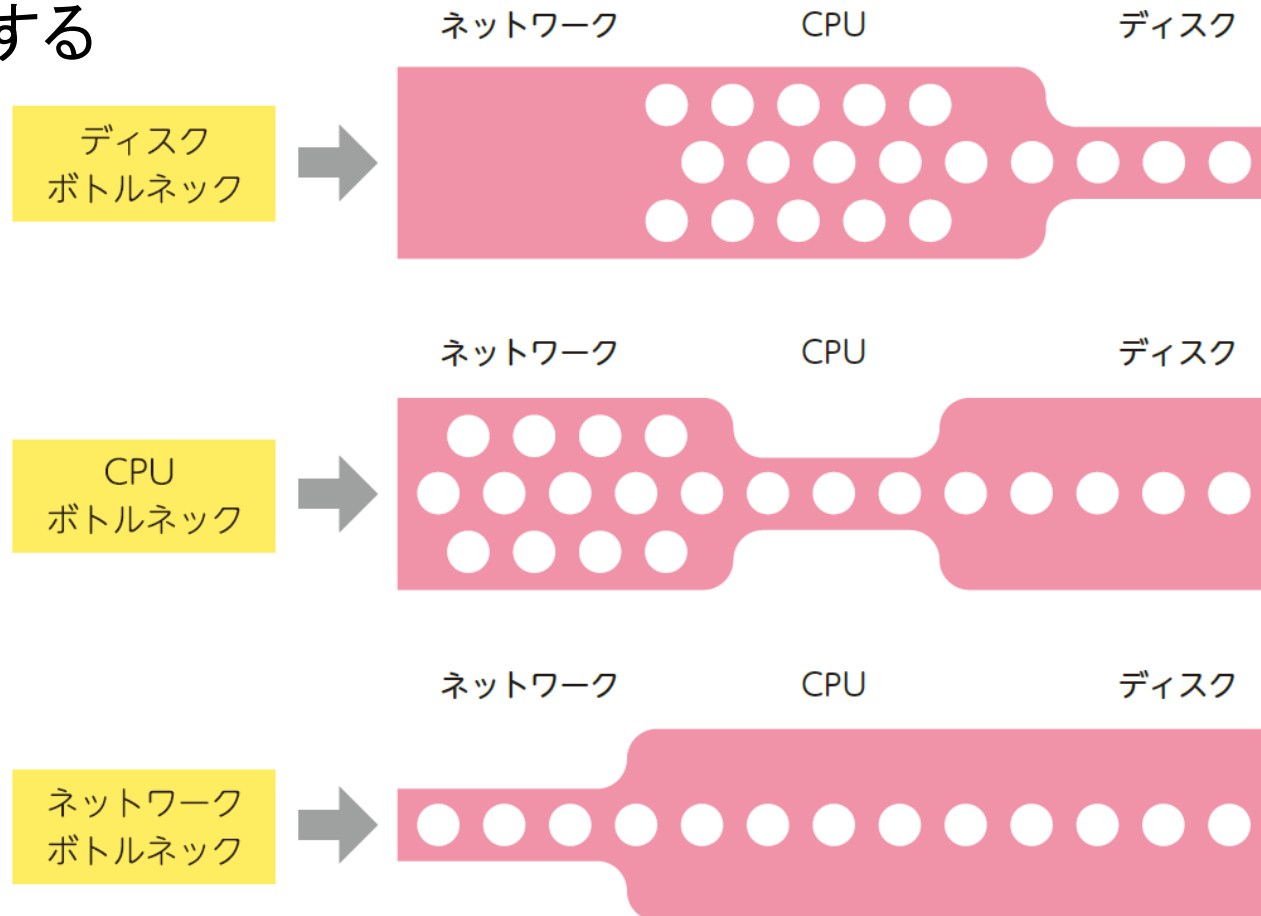


分散ファイルシステム とクラスター計算

性能問題とボトルネック

- ビッグデータの分析では大量のデータを扱うため、問題のほとんどは**性能問題**（分析バッチ処理が朝までに終わらない、BI製品の画面が重い、予測APIの応答速度が遅い）

処理時間の大部分を占めている処理「**ボトルネック**」を解析して、それを解消する

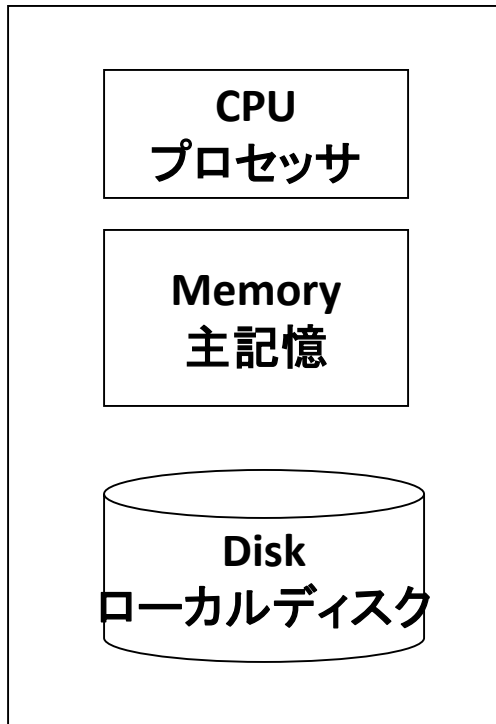


一般的な計算処理

単一のプロセッサ，メインメモリー，ローカルディスク



マルチコア
CPU



計算ノード

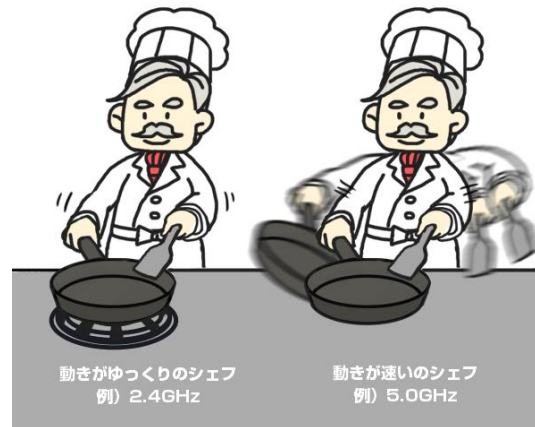
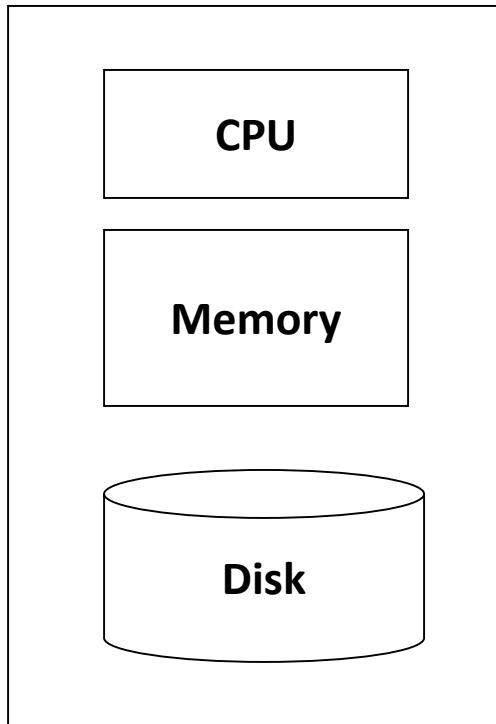
一般的な計算処理

単一のプロセッサ，メインメモリー，ローカルディスク

CPUのクロック周波数



マルチコア
CPU



<https://www.pc-koubou.jp/magazine/23926>

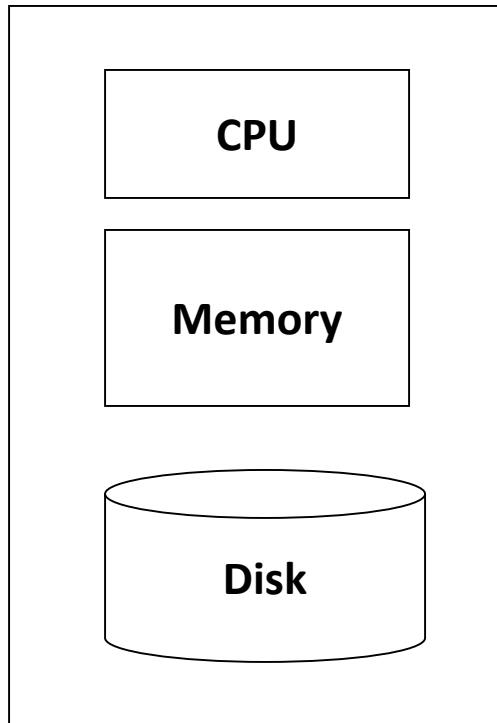
計算ノード

一般的な計算処理

単一のプロセッサ，メインメモリー，ローカルディスク



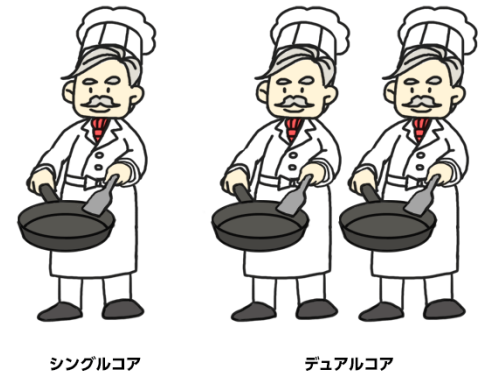
マルチコア
CPU



計算ノード

CPUのクロック周波数

CPUのコア(Core)数



<https://www.pc-koubou.jp/magazine/23926>

“Classical” Machine Learning &
Data Mining, Statistics, ...

small dataの世界

一般的な計算処理 + アクセラレータ

- コアよりも性能や機能が低い演算装置を**多数配置**
- 基本的な演算(積, 和など)を高い電力効率で実行
- ホストCPUから操作

例) GPGPU (General Purpose computation
on Graphic Processing Unit)



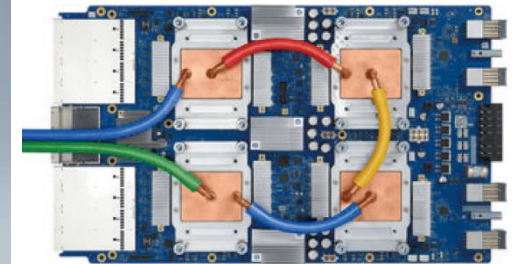
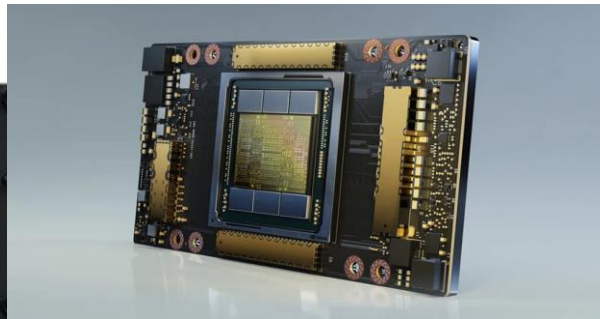
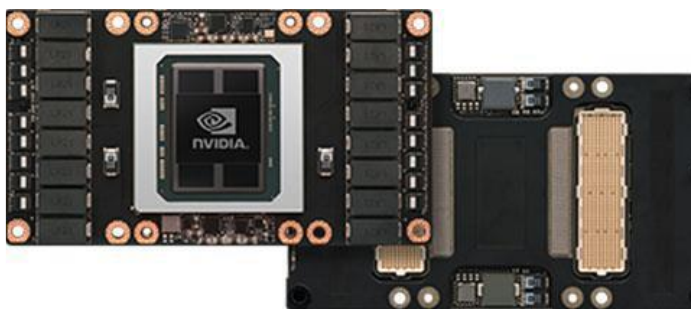
一般的な計算処理 + アクセラレータ

- コアよりも性能や機能が低い演算装置を多数配置
- 基本的な演算(積, 和など)を高い電力効率で実行
- ホストCPUから操作

例) GPGPU (General Purpose computation
on Graphic Processing Unit)

もともとグラフィック処理用プロセッサGPUをシミュレーションや機械学習の計算に使用

例) NVIDIA Tesla P100: 3584 CUDAコア A100: 6912 CUDA cores



<https://gdep-sol.co.jp/gpu-products/nvidia-gpu/>

Deep Learning, +

一般的な計算処理 + アクセラレータ

Different Kinds of Parallelism



CPU - Task Parallelism



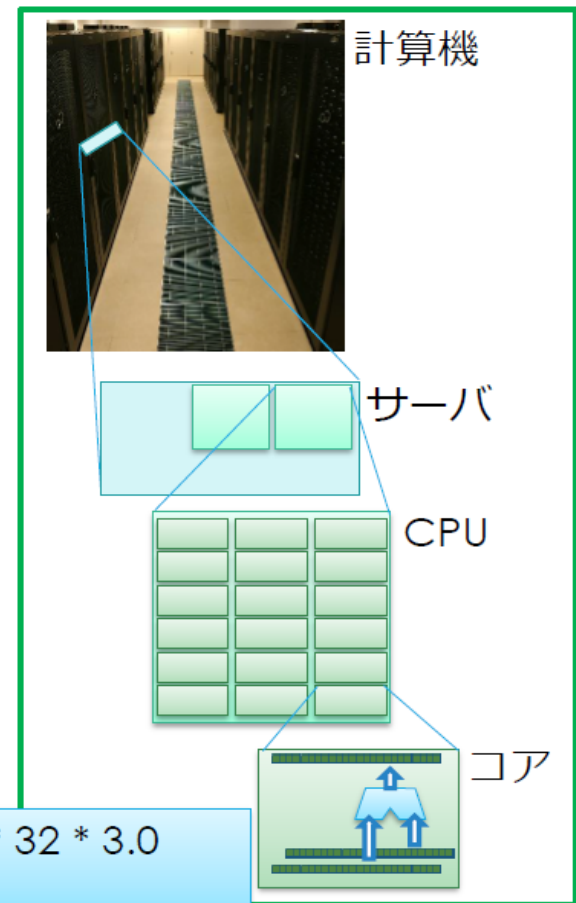
GPU - Data Parallelism

専用の並列計算機(スーパーコンピュータ)

多数のプロセッサと特殊なハードウェア

例) 九州大学のスーパーコンピュータ IIO
(サブシステムA)

- システムを構成するサーバ数 : **2,000** 台
- サーバあたりのCPU数 : **2** 個
- CPUあたりのコア数 : **18** 個
- コアあたりの最大同時演算数 : **32**
- CPUのクロック周波数 : **3.0** GHz



6912000 GFLOPS (ギガ フロップス) = $2000 * 2 * 18 * 32 * 3.0$
一秒間に 6912兆回の演算が出来る

Kilo = 10^3 , Mega = 10^6 , Giga = 10^9 , Tera = 10^{12} , Peta = 10^{15} , Exa = 10^{18}

HUMAN BRAIN PROJECT

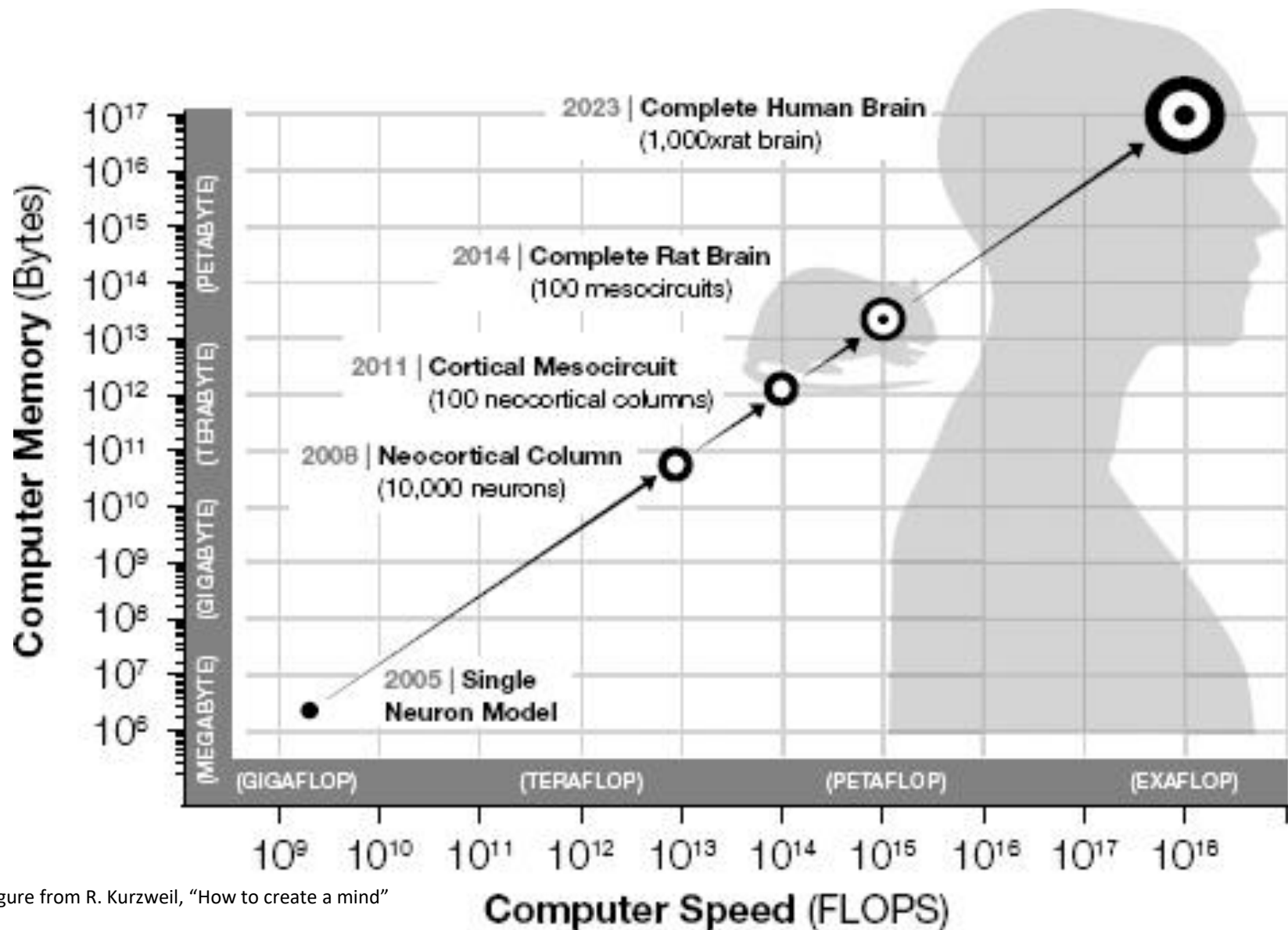


Figure from R. Kurzweil, "How to create a mind"

スパコン

歴代1位の一覧

ランク付け年月		設置国	ベンダ	名称
2019年	11月	 アメリカ合衆国	IBM	サミット
	6月			
2018年	11月			
	6月			
2017年	11月	 中華人民共和国	NRCPC	神威・太湖之光
	6月			
2016年	11月			
	6月			
2015年	11月		NUDT	天河二号
	6月			
2014年	11月			
	6月			
2013年	11月			
	6月			
2012年	11月	 アメリカ合衆国	Cray	タイタン
	6月		IBM	セコイア
2011年	11月	 日本	富士通	京
	6月			

<https://ja.wikipedia.org/wiki/TOP500>

スパコン

歴代1位の一覧

<https://ja.wikipedia.org/wiki/TOP500>

ランク付け年月		設置国	ベンダ	名称
2021年	6月	 日本	富士通	富岳
2020年	11月			
	6月			
2019年	11月	 アメリカ合衆国	IBM	サミット
	6月			
2018年	11月			
	6月			
2017年	11月			
	6月			
2016年	11月		NRCP	神威・太湖之光
	6月			
2015年	11月	 中華人民共和国		
	6月			
2014年	11月		NUDT	天河二号
	6月			
2013年	11月			
	6月			
2012年	11月	 アメリカ合衆国	Cray	タイタン
	6月		IBM	セコイア
2011年	11月	 日本	富士通	京
	6月			
2010年	11月	 中華人民共和国	NUDT	天河一号A
	6月			

スパコン

歴代1位の一覧

2022年5月、[TOP500](#)で、
1.102エクサFLOPSを達成
し、[富岳](#)を抜き世界1位の
スーパーコンピュータとなっ
た

[Aurora](#)は2023年6月22日に完成
アメリカで2番目の[エクサスケール](#)
コンピュータとなる
2exaFLOPS/s(毎秒200京回の計算
に相当)

<https://ja.wikipedia.org/wiki/TOP500>

ランク付け年月		設置国	ベンダ	名称
2022年	6月	 アメリカ合衆国	HPE	Frontier
2021年	11月	 日本	富士通	富岳
	6月			
2020年	11月			
	6月			
2019年	11月	 アメリカ合衆国	IBM	サミット
	6月			
2018年	11月			
	6月			
2017年	11月	 中華人民共和国	NRCPC	神威・太湖之光
	6月			
2016年	11月			
	6月			
2015年	11月		NUDT	天河二号
	6月			
2014年	11月			
	6月			
2013年	11月			
	6月			

スパコン

歴代1位の一覧

El Capitan (エル・キャピタン) は、2019年8月に開発が発表され、2024年に完成し、2024年11月のTOP500で2.79 エクサFLOPSを達成し世界最高速となったスーパーコンピュータ

CPUとGPUのコア数合計は1,103万9,616基を搭載している

ランク付け年月		設置国	ベンダ	名称
2024年	11月	 アメリカ合衆国	HPE	El Capitan
	6月			Frontier
2023年	11月			
	6月			
2022年	11月			
	6月			
2021年	11月	 日本	富士通	富岳
	6月			
2020年	11月			
	6月			
2019年	11月		 アメリカ合衆国	サミット
	6月			
2018年	11月			
	6月			
2017年	11月		NRCPC	神威・太湖之光
	6月			
2016年	11月			
	6月			

スパコン ランキング 性能

2020年 富岳

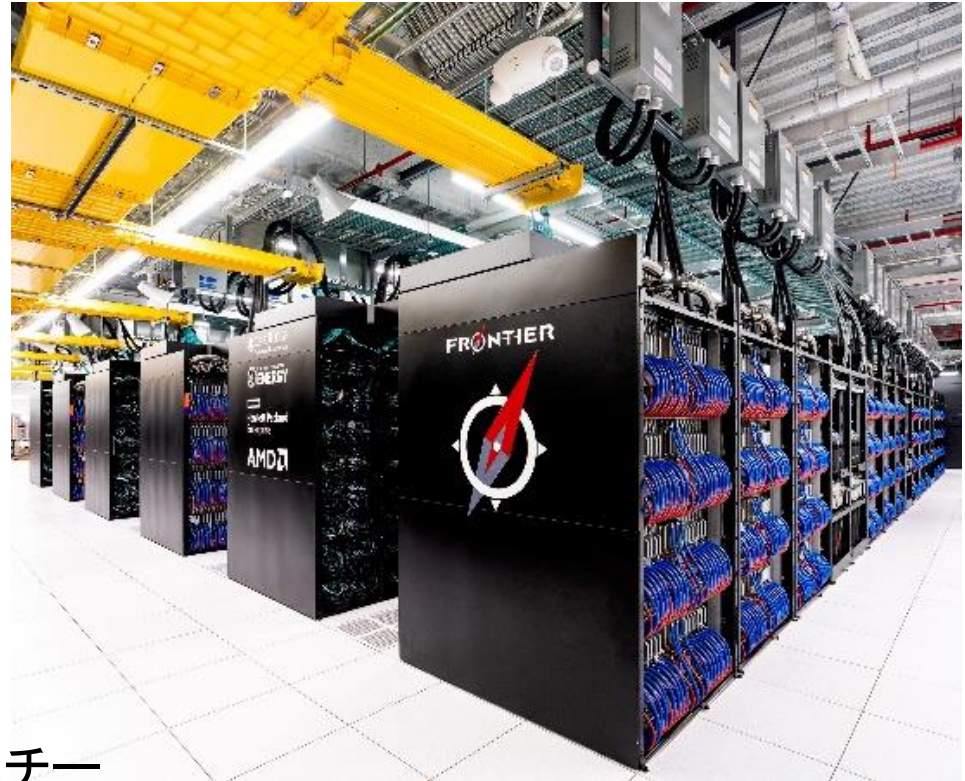
順位	Rmax Rpeak (PFLOPS)	名称	コンピュータ設計 プロセッサ, 接続	ベンダー	場所 国, 年
1	442.010 537.212	富岳	A64FX 48C 2.2GHz Tofu interconnect D	富士通	RIKEN 日本, 2020
2	148.600 200.795	Summit	IBM Power System AC922 Power9 22C + Tesla V100, Mellanox dual-rail EDR InfiniBand	IBM	オークリッジ国立研究所 アメリカ合衆国, 2018
3	94.640 125.712	Sierra	IBM Power System S922LC Power9 22C + Tesla V100, Mellanox dual-rail EDR InfiniBand	IBM	ローレンス・リバモア国立研 アメリカ合衆国, 2018
4	93.015 125.436	神威・太湖之光	Sunway MPP SW26010, Sunway	NRCPC	国家超級計算無錫中心 (英語版) 中国, 2016
5▲	64.590 89.795	Perlmutter	HPE Cray EX235n AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	HPE	ローレンス・バークレー国立 アメリカ合衆国, 2021

性能指標
(Peta FLOPS)

<https://ja.wikipedia.org/wiki/TOP500>

講義の内容について

SUPERCOMPUTER



ゲノム科学の研究をやっているある研究チームが、普通のパソコンでやったら30年もかかる計算をサミットで1時間で終わった

[https://en.wikipedia.org/wiki/Frontier_\(supercomputer\)](https://en.wikipedia.org/wiki/Frontier_(supercomputer))

スーパーコンピュータ と PC、スマートフォンの CPU の演算性能





	スマートフォン (Snapdragon 865)	PC (Intel Core i9 10980XE)	スーパーコンピュータ ITO サブシステム A (Intel Xeon Gold 6154, Skylake-SP)	スーパーコンピュータ 富岳 (Fujitsu A64FX) 2020年6月時点で世界最速
				
CPU数	1	1	4,000	158,976
クロック周波数	1.8GHz, 2.4GHz, 2.84GHz	3.0GHz	3.0GHz	2.0GHz
コア数	計 8	18	18	48
コアあたり最大同時演算数	8?	32	32	32
総理論演算性能	137.92 GFLOPS	1,728 GFLOPS	6,912,000 GFLOPS	488,374,272 GFLOPS

クロック周波数はほぼ同じ



基本的に CPU数やコア数で演算性能を稼ぐ

アクセラレータを加えた性能

	スーパーコンピュータ ITO		スーパーコンピュータ Summit	スーパーコンピュータ 富岳 (Fujitsu A64FX)
	サブシステム A 	サブシステム B 	 世界2位 https://www.ibm.com/thought-leadership/summit-supercomputer/	 世界1位 https://www.fujitsu.com/jp/about/businesspolicy/tech/fugaku/
CPU数	4,000	256	9,216	158,976
クロック周波数	3.0GHz	2.3GHz	3.07GHz	2.0GHz
コア数	18	18	22	48
コア当たり 最大同時演算数	32	32	8	32
CPU理論演算性能	6.9 PFLOPS	0.3 PFLOPS	5.0 PFLOPS	488 PFLOPS
アクセラレータ数	0	512	27,648	0
アクセラレータ 理論演算性能	0	2.7 PFLOPS	193.5 PFLOPS	0
総演算性能		9.9 PFLOPS	199 PFLOPS	488 PFLOPS

ITOの後継として導入 新スーパーコンピュータシステム玄界(げんかい)
(2024年10月運用開始) 富士通社のFUJITSU Server PRIMERGYシリーズを中核とするシステム

新スーパーコンピュータシステム玄界の概要

総理論演算性能は約 13 PFLOPS

ノードグループA 1,024ノード

CPU: Intel Xeon (Sapphire Rapids, 60core) × 2 / node

RAM: 512 GiB / node

ノードグループB 38ノード

CPU: Intel Xeon (Sapphire Rapids, 60core) × 2 / node

RAM: 1 TiB / node

GPU: NVIDIA H100 (SXM) x 4 / node

SSD: 12.8TB / node

ノードグループC 2ノード

CPU: Intel Xeon (Sapphire Rapids, 56core) × 2 / node

RAM: 8 TiB / node

GPU: NVIDIA H100 (SXM) x 8 / node

SSD: 15.3TB / node

ストレージ

HDD: 55.2 PB

SSD: 0.7 PB

<https://www.kyushu-u.ac.jp/ja/notices/view/2699/>



クラスター計算 (Cluster Computing)

- INTERNET + 大規模なウェブサービスの普及
- ◆ 現代のインターネットアプリケーションでは、**巨大なデータを迅速に管理する**ことが求められるようになった
- ◆ 多くのアプリケーションでは、**データは極めて規則的であり、並列化を活用できる**十分な余地がある

クラスター計算 (Cluster Computing)

- INTERNET + 大規模なウェブサービスの普及
- ◆ 現代のインターネットアプリケーションでは、**巨大なデータを迅速に管理する**ことが求められるようになった
- ◆ 多くのアプリケーションでは、**データは極めて規則的であり、並列化を活用できる**十分な余地がある
- 例：
 1. **ウェブページを重要度に応じてランキングする** (次元が数百億におよぶ行列とベクトルの乗算の繰り返しが生じる)
 2. **ソーシャルネットワークサイトで友達のネットワークを検索する** (数億のノード (個人) と数十億の枝 (友達関係) を扱う)

クラスター計算 (Cluster Computing)

QUIZ

- 20,000,000,000+ web pages x 20KB = 400+ TB
- Assume 1 computer reads 30-35 MB/sec from disk
- How long it would take to read the web (単一計算ノードで)
 - 1 day to read the web
 - 4 months to read the web
 - 2 years to read the web

クラスター計算 (Cluster Computing)

QUIZ

- 20,000,000,000+ web pages x 20KB = 400+ TB
- Assume 1 computer reads 30-35 MB/sec from disk
- How long it would take to read the web (単一計算ノードで)
 - 1 day to read the web
 - 4 months to read the web
 - 2 years to read the web

クラスター計算 (Cluster Computing)

QUIZ

- How many pages to print all web pages on the Internet
 - 200,000,000
 - 200,000,000,000
 - 200,000,000,000,000

クラスター計算 (Cluster Computing)

QUIZ

- How many pages to print all web pages on the Internet
 - 200,000,000
 - 200,000,000,000
 - 200,000,000,000,000

クラスター計算 (Cluster Computing)

- 新しい計算システム＋新世代のプログラミングシステムの先駆け →

独立した計算ノード(プロセッサ, 主記憶, ディスク)のクラスター

- 有利な点:
 - 並列化を活用できる
 - 信頼性(障害対応)
 - 計算ノードはありふれたハードウェアであり, 専用並列計算機と比べてコストを大きく削減することができる

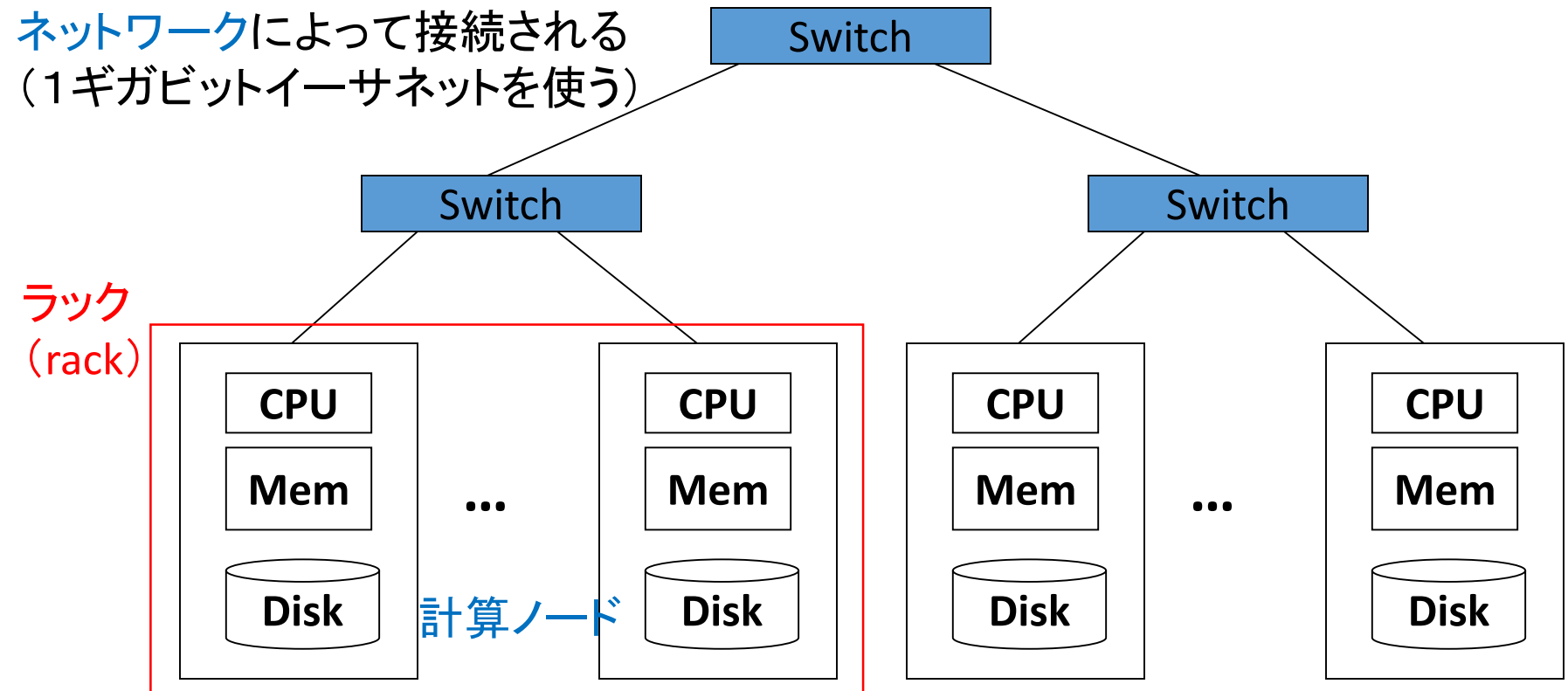


<http://bit.ly/Shh0RO>

計算ノードクラスターの物理的な構成

ラックは異なるレベルのネットワークやスイッチを使って接続される (2-10 Gbps backbone between racks)

ラックに格納されたノードは、
ネットワークによって接続される
(1ギガビットイーサネットを使う)



1つのラックに16-64個の計算ノードが配置される



In 2011 it was guestimated that Google had 1M machines, <http://bit.ly/Shh0RO>

構成要素の障害

- 計算ノードや相互接続ネットワークなどシステムを構成する要素が多くなればなるほど、システムが任意の時点で動かなくなる可能性が高くなる
- 障害の主要な要因
 1. ノードの損失 (例えば, ハードディスクの損傷)
 2. ラック全体の損失 (他のノードや外界と接続しているネットワークの障害)
 - One server(node) may stay up 3 years (1,000 days)
 - If you have 1,000 servers, expect to loose 1/day
 - People estimated Google had ~1M machines in 2011
 - 1,000 machines fail every day!

構成要素の障害 → 対策

- 重要な計算途中, 1つの計算ノードが障害を起こすたびに, **処理を中断し構成要素を再起動する**のでは, 計算を完了させることはできない. . .



構成要素の障害 → 対策

- 重要な計算途中, 1つの計算ノードが障害を起こすたびに, **処理を中断し構成要素を再起動する**のでは, 計算を完了させることはできない. . .

- 対策

- (1) **ファイルを冗長に保存する**

- * 複数の計算ノード上でファイルを複製する



構成要素の障害 → 対策

- 重要な計算途中, 1つの計算ノードが障害を起こすたびに, **処理を中断し構成要素を再起動する**のでは, 計算を完了させることはできない...



- 対策

- (1) **ファイルを冗長に保存する**

- * 複数の計算ノード上でファイルを複製する

- (2) **計算をタスクに分割する**

- * どれか1つのタスクの実行が止まっても, 他のタスクに影響を与えることなく再開できる (MapReduceで実現)

分散ファイルシステム (Distributed File System, DFS)

大規模ファイルシステムの特徴

- ファイルサイズは巨大 (TBレベル)

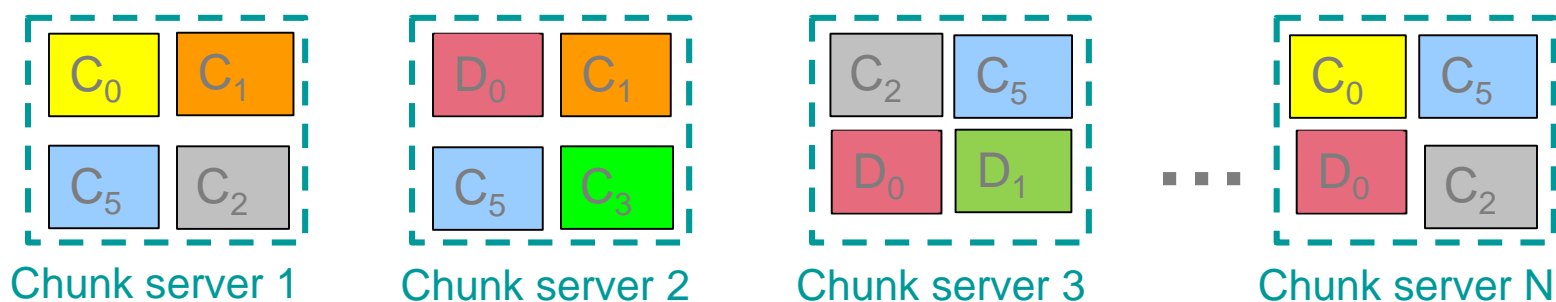
分散ファイルシステム (Distributed File System, DFS)

大規模ファイルシステムの特徴

- ファイルサイズは巨大 (TBレベル)
- ファイルは**チャンク(chunk)** に分割される

チャンクサイズ: 通常128MB

各チャンクを**3つに複製**, 3つの**異なる計算ノード**に置かれる
チャンクの複製を保持するノードは, **別のラック**に置かれる



Chunk server は Compute server としても使う

Bring computation directly to the data!

分散ファイルシステム (Distributed File System, DFS)

大規模ファイルシステムの特徴

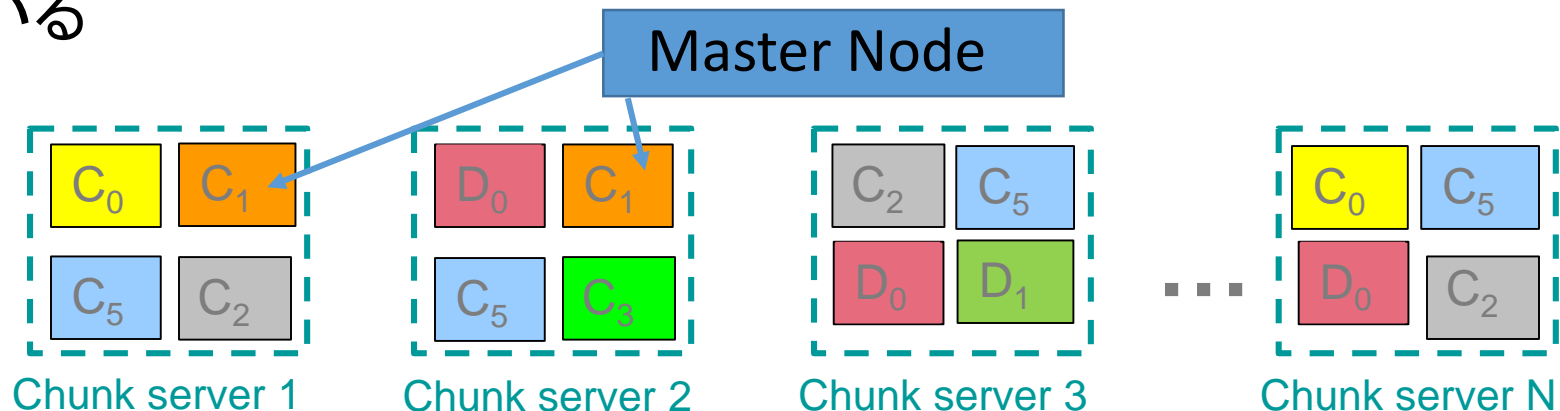
- マスターノード (Master Node, Name Node)

- * ファイルのチャンクの場所についてのデータ (メタデータ)

- * マスターノード自体も複製される

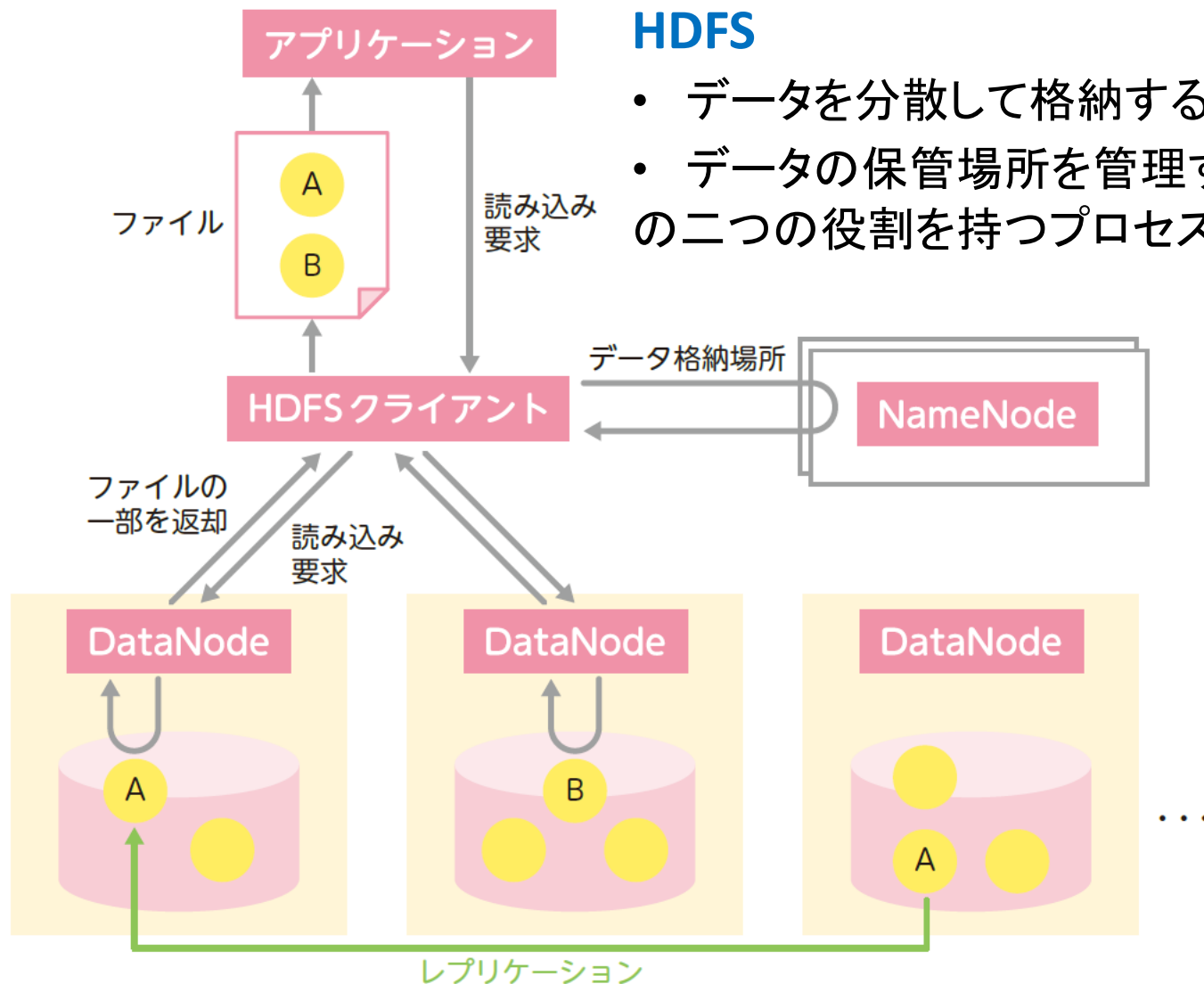
- * ファイルシステム全体のディレクトリーは、どこに複製があるかを知っている

- * ディレクトリーそのものも複製することが可能, DFSの使用者は、ディレクトリーの複製がどこにあるかを知ることができるようになっている



分散ファイルシステム (Distributed File System, DFS)

■ HDFSの構成



代表的な製品はHadoopプロジェクトの一部である
HDFS

- データを分散して格納する **DataNode**
 - データの保管場所を管理する **NameNode**
- の二つの役割を持つプロセスから構成されます

分散ファイルシステム (Distributed File System, DFS)

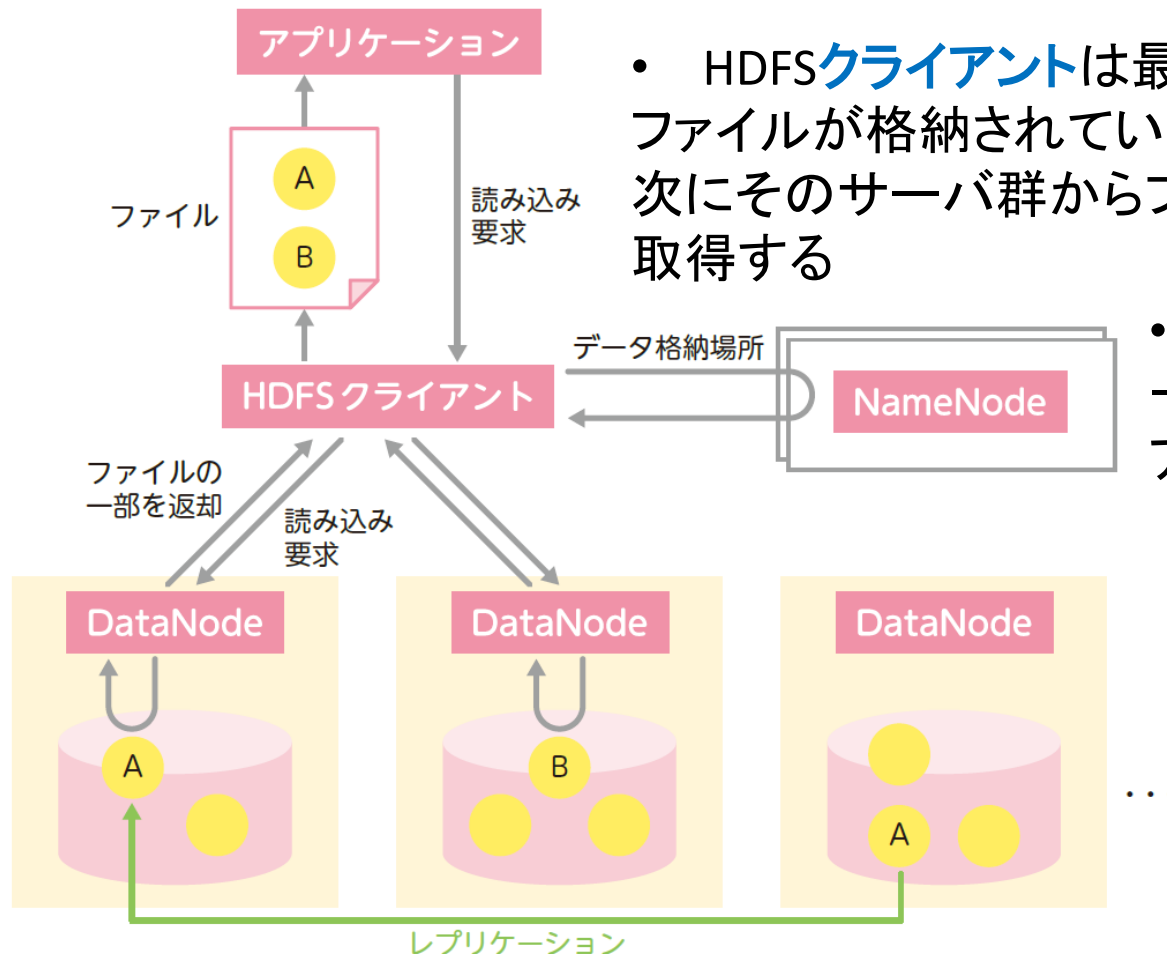
データへのアクセス方法

- アプリケーションからファイルにアクセスする場合は **HDFS クライアント** を使います

- HDFSクライアント** は最初に **NameNode** に対してファイルが格納されている **DataNode** 群を問い合わせ、次にそのサーバ群からファイルを構成するデータを取得する

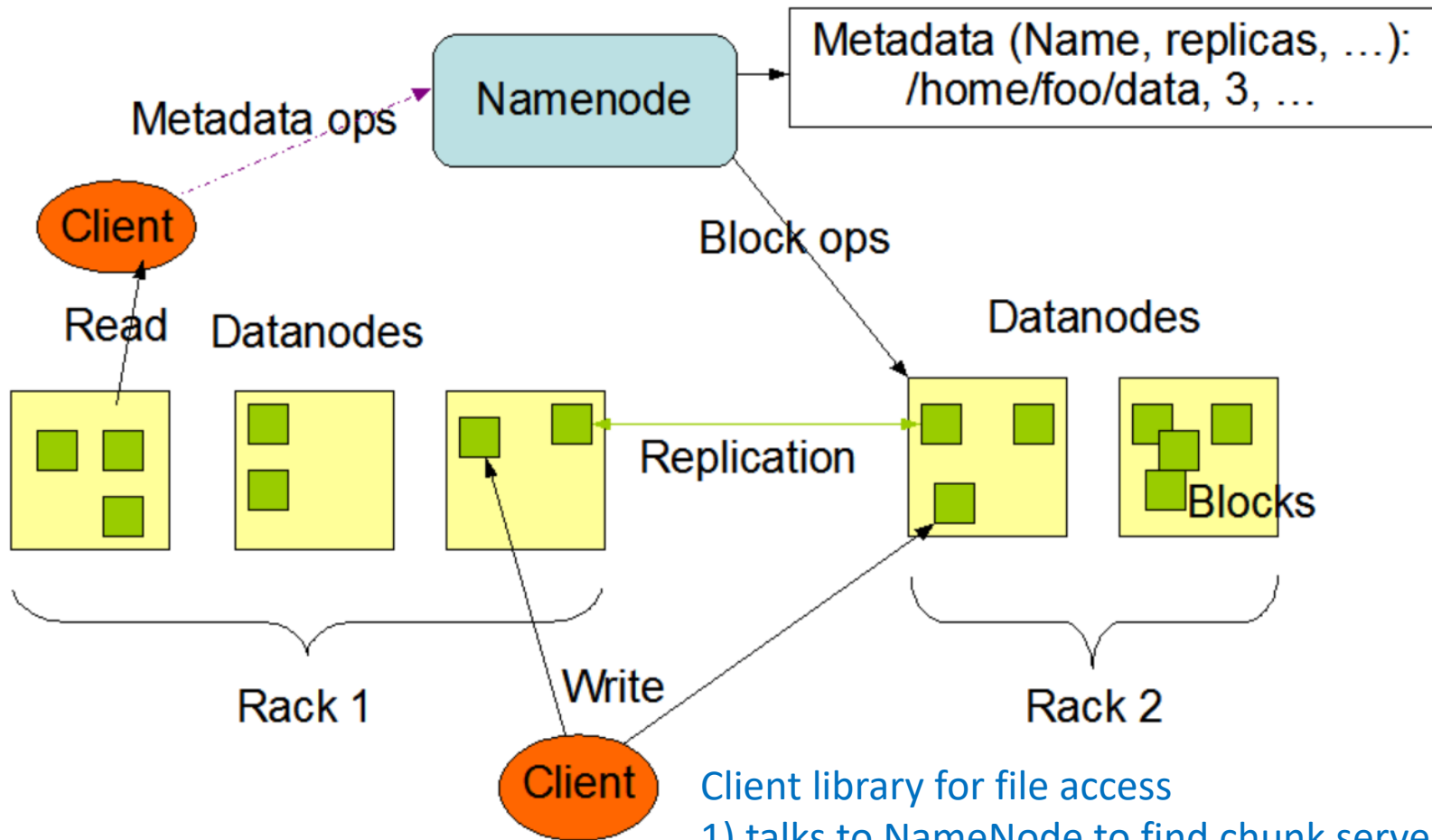
- 最後に **HDFSクライアント** 上で一つのファイルに統合し、アプリケーションに返す

■ HDFSの構成



分散ファイルシステム (Distributed File System, DFS)

HDFS Architecture



Client library for file access

1) talks to NameNode to find chunk servers

2) Connects directly to chunk servers to access data

References

Leskovec et al., Mining of Massive Datasets, 3ed., CUP, 2020.

渡部徹太郎, 図解即戦力 ビッグデータ分析, 技術評論社, 2021.