WGS 1 万検体 を解析するための適切な Luster 設計とその運用

東京大学医科学研究所ヒトゲノム解析センター
シークエンスデータ情報処理分野
准教授　片山琴絵

ヒトゲノム解析センター（HGC）

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

✓東京大学 医科学研究所 ヒトゲノム解析センターはゲノム情報と医療情報に基づいた個別化ゲノム医療など、データサイエンスを広く推進

✓医学・生命科学研究に最適化したスーパーコンピュータ SHIROKANE を擁し、以下の事業を行い、また SHIROKANE を教育研究機関、民間企業の研究・事業活動に提供

- 個別化ゲノム医療のための次世代ゲノム医学研究の推進
- 個別化ゲノム医療のためのメディカルインフォマティクスの研究
- 倫理的・法的・社会的問題の研究による公共政策研究

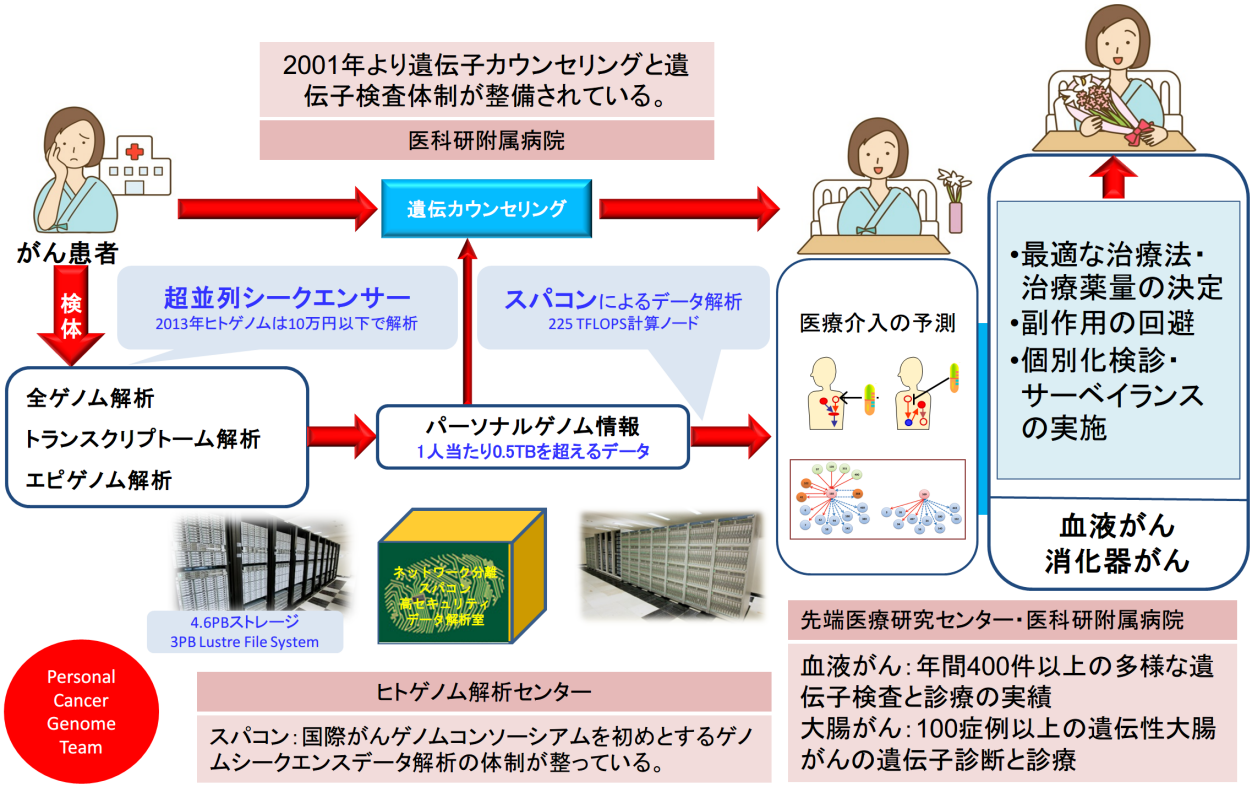
✓HGC スパコン SHIROKANE のミッションとして

- AMED 革新的がん医療実用化研究事業
- パーソナルゲノムに基づく個別化医療の推進

個別化医療の推進

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center the Institute of Medical Science, the University of Tokyo

✓個別化医療の第一歩は Whole Genome Sequencing であり、データ解析・解釈、メディカルインフォマティクスが鍵になる。HGC は病院と連携し、全ゲノム情報解析を SHIROKANE を用いて行ってる。



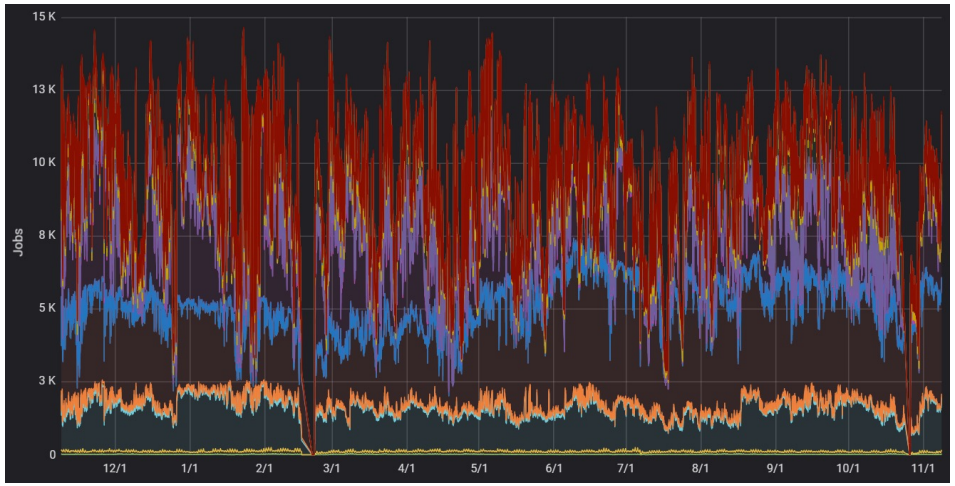
SHIROKANE の紹介

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

✓SHIROKANE は 2018 年 4 月稼働の Shirokane5 と 2022 年 4 月稼働の Shirokane6 が合わさったスーパーコンピュータシステムになる

✓SHIROKANE の特徴

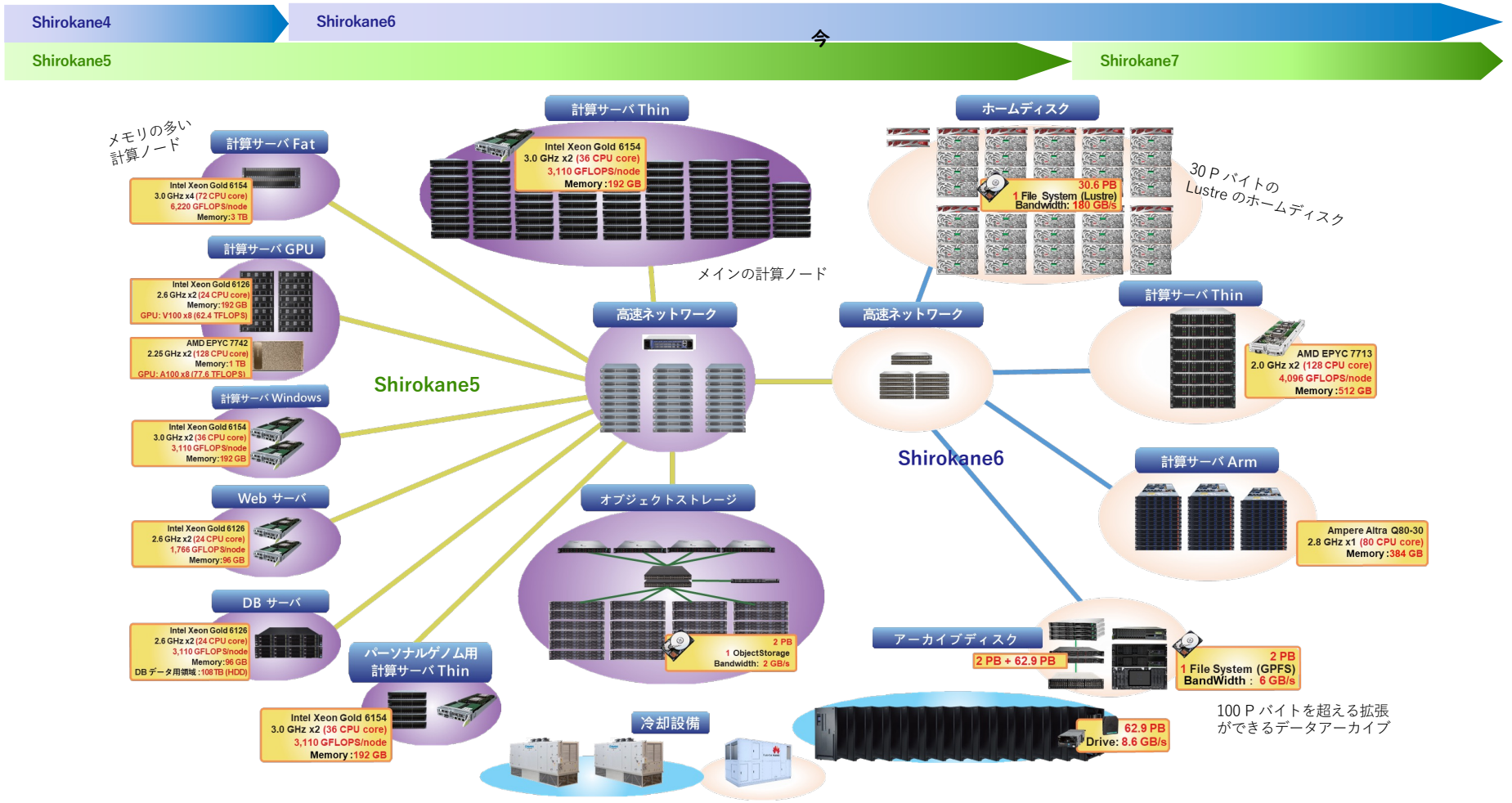
- 1.9 PFLOPS 余の総合理論演算性能になる 4 種類の計算機を用意
- 30 P バイト以上の高速ディスクアレイ装置であるホームディスクを用意
- 60 P バイト以上の長期保存するための巨大アーカイブディスクを用意
- 年間約 9,600 万件 (1日あたり約 25 万件)のジョブが実行



実行ジョブ数

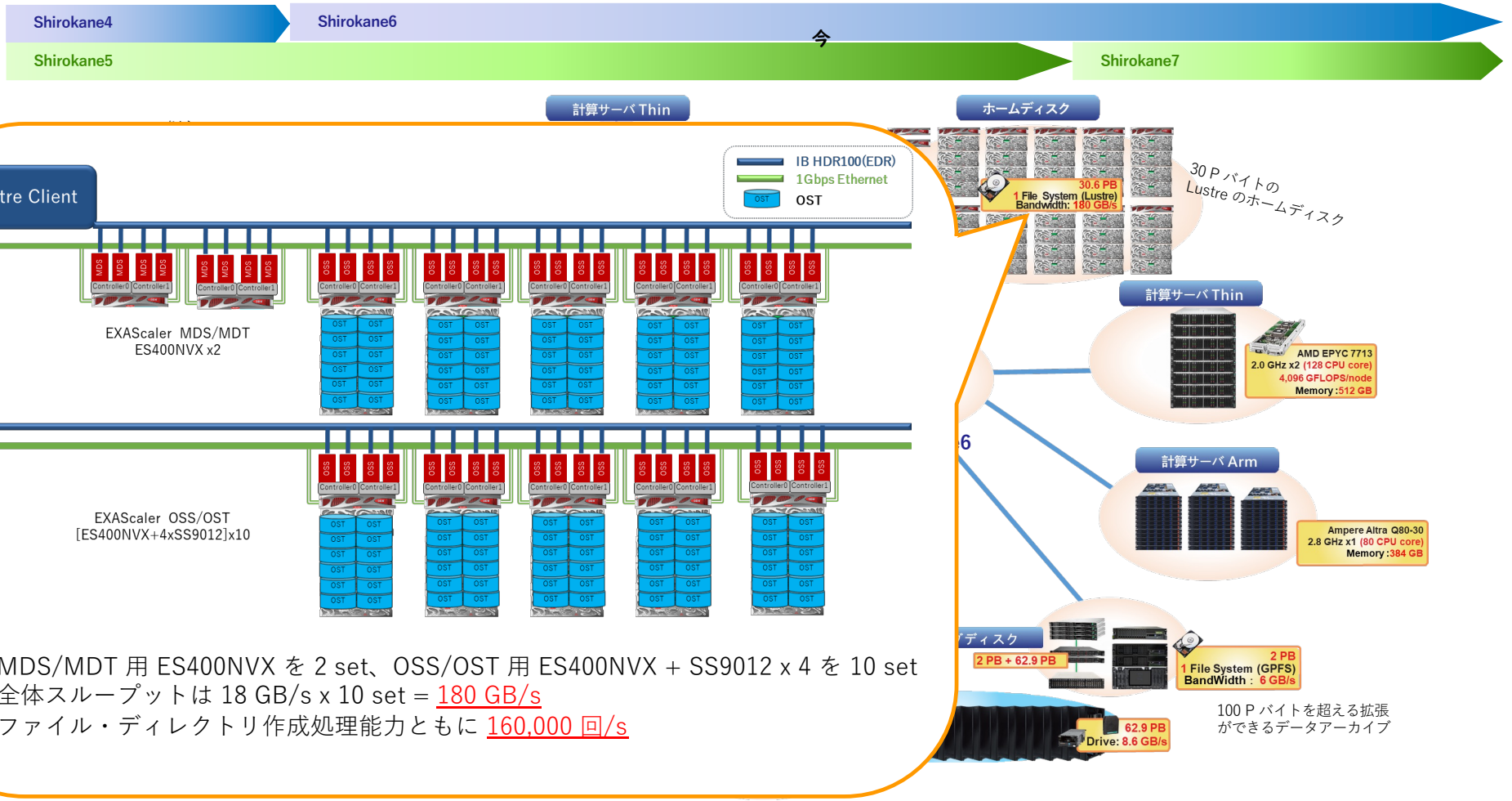
SHIROKANE の構成

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



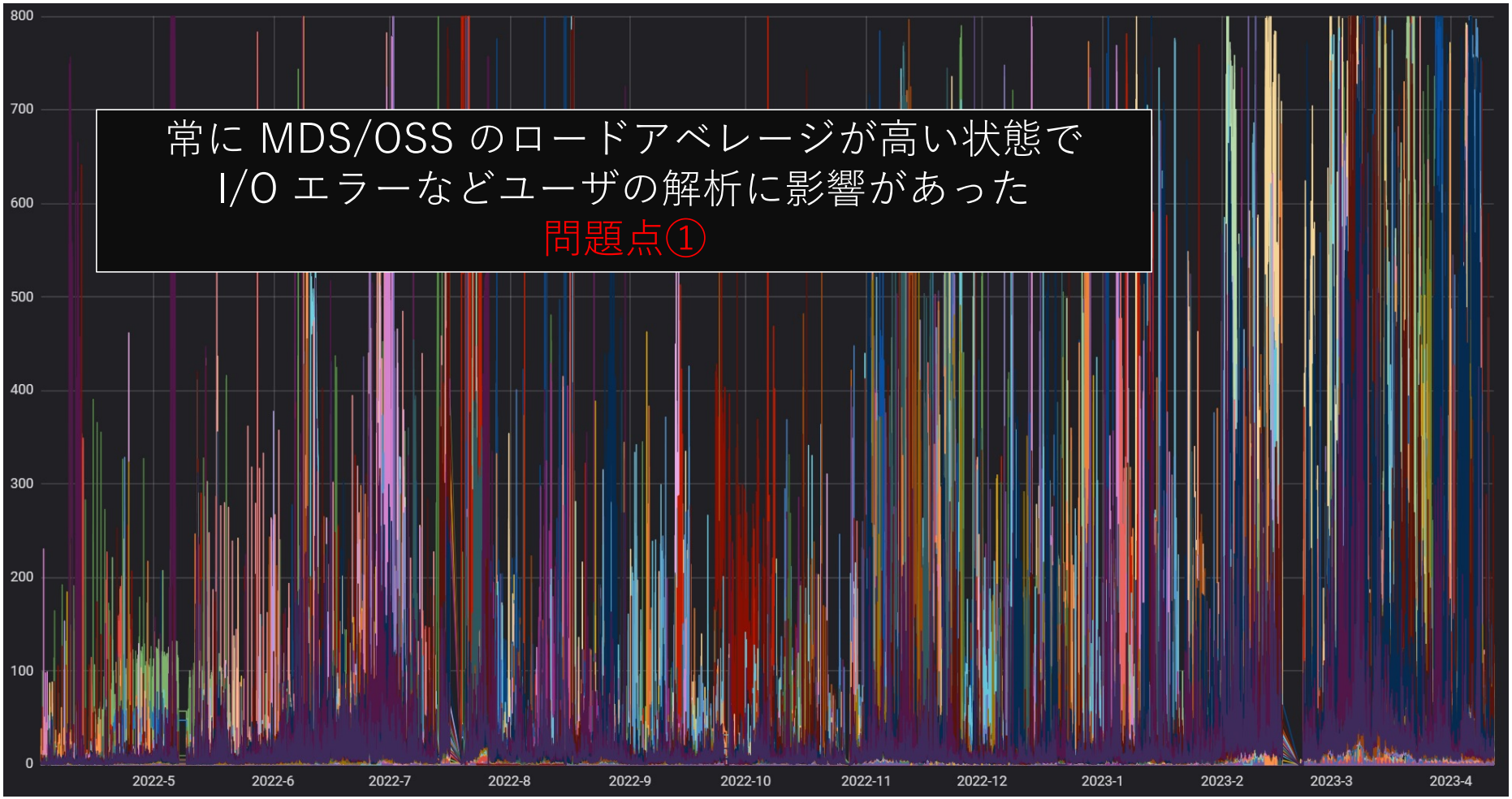
SHIROKANE の構成

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



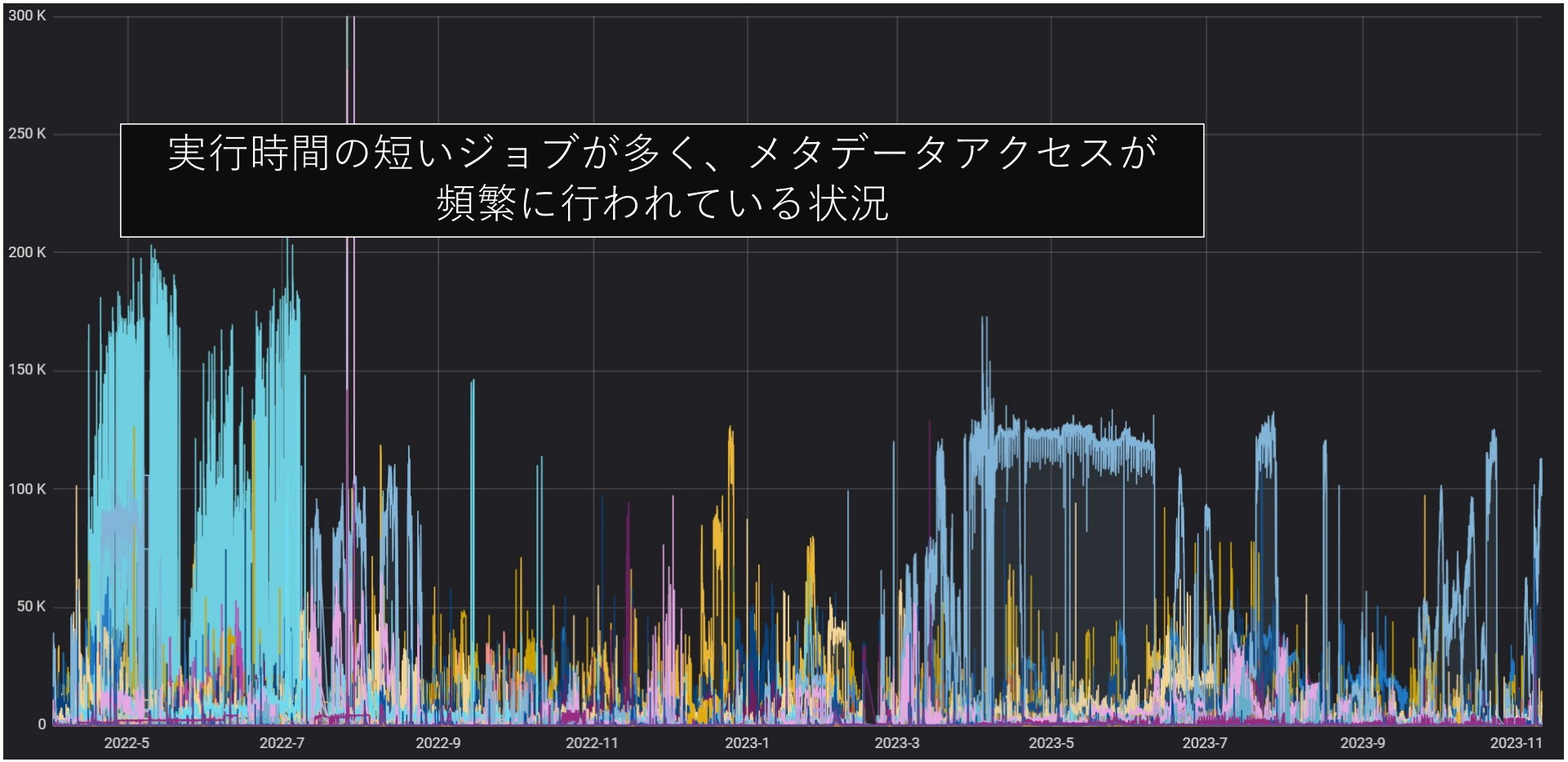
- MDS/MDT 用 ES400NVX を 2 set、OSS/OST 用 ES400NVX + SS9012 x 4 を 10 set
- 全体スループットは 18 GB/s x 10 set = 180 GB/s
- ファイル・ディレクトリ作成処理能力ともに 160,000 回/s

MDS/OSS の負荷状況



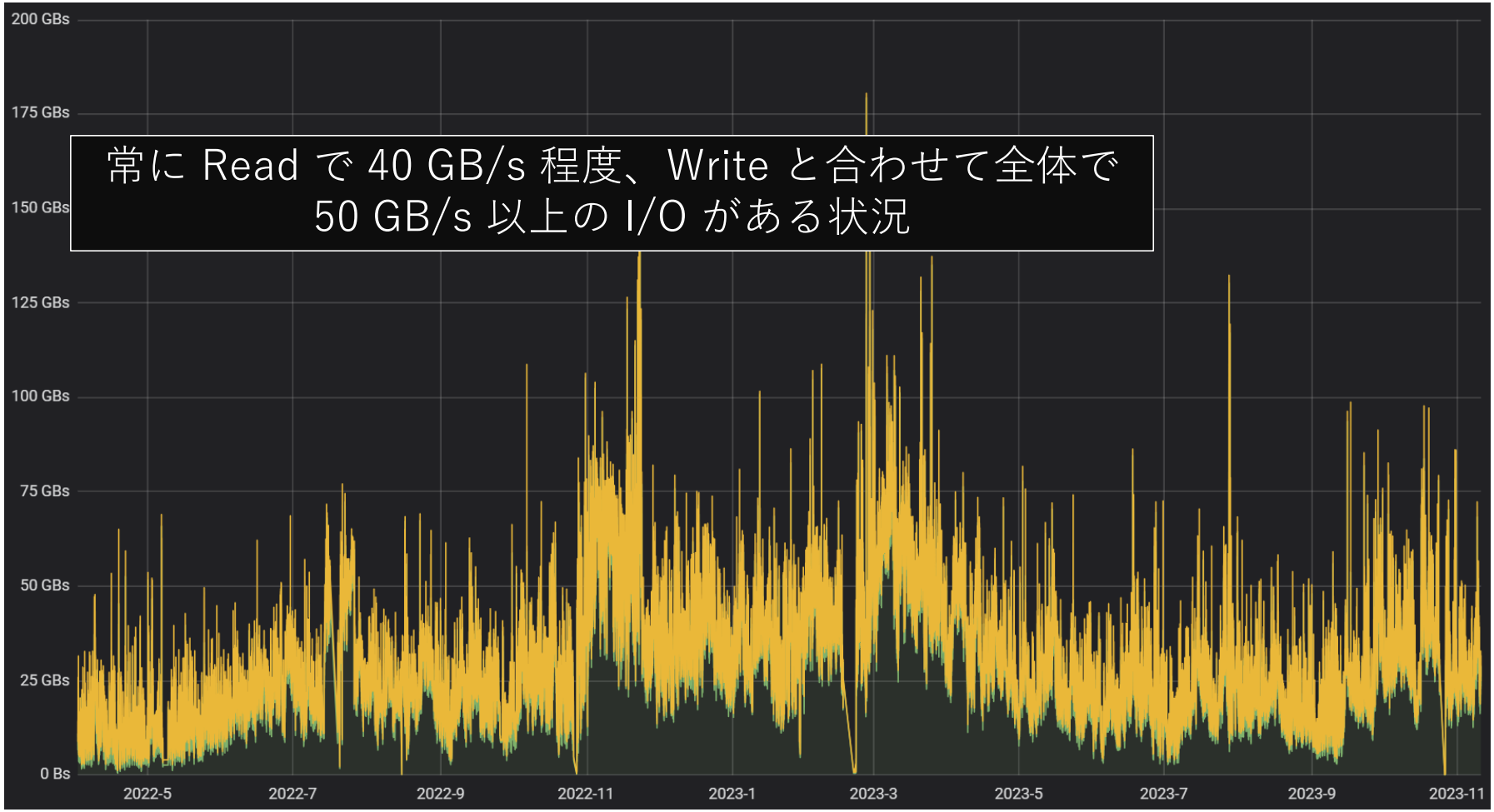
メタデータアクセス

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



実行時間の短いジョブが多く、メタデータアクセスが頻繁に行われている状況

アクセスバンド幅


200 GBs
175 GBs
150 GBs
125 GBs
100 GBs
75 GBs
50 GBs
25 GBs
0 Bs
2022-5
2022-7
2022-9
2022-11
2023-1
2023-3
2023-5
2023-7
2023-9
2023-11
常に Read で 40 GB/s 程度、Write と合わせて全体で
50 GB/s 以上の I/O がある状況

問題点①

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

✓状況

- 一部のユーザ が Lustre ファイルシステムへの I/O が多い大量の Array ジョブを実行し、OSS の過負荷を招いていた

✓問題点

- OSS の過負荷を発見した際に現地 SE で高負荷を招いているジョブを手作業で特定し、対象ジョブをサスペンドし、ユーザへ連絡していた
- 現地 SE が業務中では対応できるが、業務時間外の場合対応できていない

ターミナル — ssh

```
# iostat -mxt 1
.
Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz
await r_await w_await  svctm  %util
.
sfa0000           0.00     0.00  441.00    3.00   349.29     0.01  1611.17    62.68
134.97  135.89    0.33   2.25 100.00
sfa0001           0.00     0.00  142.00    0.00   111.00     0.00  1600.90     2.57
19.24   19.24    0.00   3.50  49.70
.
#
# grep read_bytes /proc/fs/lustre/obdfilter/rshare-OST0027/exports/*/stats | sort -
k7 -n
.
.
/proc/fs/lustre/obdfilter/rshare-OST0015/exports/172.28.4.29@o2ib/stats:read_bytes
5142 samples [bytes] 4096 4194304 21485920256
/proc/fs/lustre/obdfilter/rshare-OST0015/exports/172.28.5.132@o2ib/stats:read_bytes
5278 samples [bytes] 4194304 4194304 22137536512
```

ターミナル — ssh

```
# qstat -f -u '*' -q '*@rc010' -sr
queuename           qtype resv/used/tot. np_load arch     states
---------------------------------------------------------------------------------
mjobs.q@rc010i         BP  0/128/128   0.93   lx-amd64
 83183224 0.00000 sv_filt_20 user1    r   11/10/2023 08:57:33   6
 83183225 0.00000 sv_filt_20 user1    r   11/10/2023 08:57:33   6
 83183229 0.00000 sv_filt_20 user1    r   11/10/2023 09:27:27   6
 83183230 0.00000 sv_filt_20 user1    r   11/10/2023 09:27:28   6
 83158973 0.00000 Delly_rege user2    r   11/10/2023 10:46:03   1 498
 83161884 0.00000 Delly_rege user2    r   11/10/2023 11:35:43   1 480
.
.
# qmod -sj 83161884
# qhold 83161884
# qmod -rj 83161884
#
```

問題点①

✓対策

➢Lustre の QoS 機能を使用し、ユーザまたはグループ単位で I/O 制限する

- QoS で制限してしまうと常に全ユーザで制限がかかってしまうので適応しない方針とした

➢Lustre Monitaring 機能を使用し、ジョブスケジューラと連携したユーザ・ジョブ情報から原因となっているユーザを特定し、自動対応するスクリプトを作成する

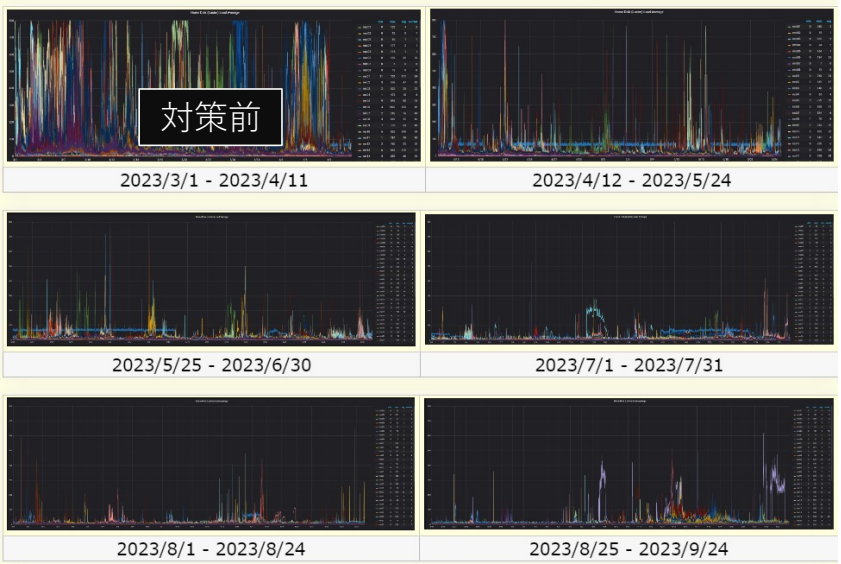
- MDS/OSS が過負荷になった場合に自動で以下を実施し、ユーザにアナウンス
  - 特定ユーザを特定し、ジョブをサスペンド
  - 2 回目は、実行待ちに戻し、ジョブをホールド
  - 3 回目は、ジョブを削除

✓効果

➢Lustre 過負荷の頻度が削減

➢アナウンスを実施することで過負荷を与えていたジョブ数が削減

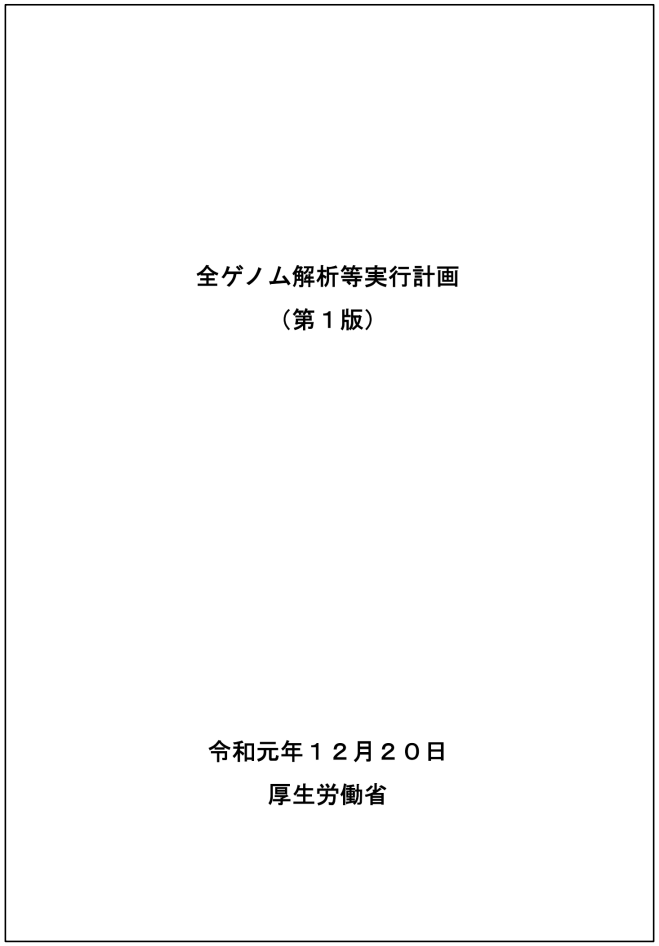
- ユーザへの教育ができた



全ゲノム解析等実行計画

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

✓がんの全ゲノム解析等は、一人ひとりにおける治療精度を格段に向上させ、治療法のない患者に新たな治療を提供するといったがん医療の発展や個別化医療 の推進など、がんの克服を目指したがん患者のより良い医療の推進のために実施する。
全ゲノム解析等により、がん医療への活用、日本人のがん全ゲノムデー タベースの構築、がんの本態解明、創薬等の産業利用を進めていく。



全ゲノム解析等実行計画

（第１版）

令和元年１２月２０日

厚生労働省

全ゲノム解析等実行計画

「全ゲノム解析等実行計画」2022概要

目的
○これまでの先行解析においては、解析結果をより早期に日常診療へ導入し、新たな個別化医療等の推進を進めてきた。
○今後の本格解析においては、国民へ質の高い医療を届け、将来的な「がん・難病等の克服」を目指す。そのためには、戦略的なデータの蓄積を進め、それらを用いた研究・創薬等を促進することが重要であることから、本実行計画においては、全ゲノム解析等の解析結果を研究・創薬等に活用することを推進する。

	令和元年度～３年度	令和４年度	令和５年度	令和６年度	令和７年度～
解析フェーズ	先行解析（既存検体）	本格解析（新規患者の検体）			
実行計画	第１版	実行計画2022			
	○本格解析の方針決定と体制整備	○戦略的なデータの蓄積 ○解析結果の日常診療への早期導入 ○新たな個別化医療の実現 ⇒ 国民へ質の高い医療を届ける			
解析実績・予定	約19,200症例 ・がん領域(※1)：約13,700症例（新規患者 600症例を含む） ・難病領域(※2)：約5,500症例	○10万ゲノム規模を目指した解析のほか、マルチ・オミックス（網羅的な生体分子についての情報）解析を予定。			
患者還元	○患者還元体制の構築	○患者が、地域によらず、全ゲノム解析等の解析結果に基づく質の高い医療を受けられるようにする。			
情報基盤	○技術的課題の検証 ○統一パイプライン構築	○がん・難病に係る創薬推進等のため、臨床情報と全ゲノム解析の結果等の情報を連携させ搭載する情報基盤を構築し、その利活用に係る環境を整備する。			
事業実施組織	○本格解析に向けて事業実施組織に係る事項について検討	○令和４年度中に事業実施準備室を国立高度専門医療研究センター医療研究連携推進本部（JH:Japan Health Research Promotion Bureau）内に設置し、組織、構成等を検討する。 ○厚生労働省が主体となって、令和７年度からの事業実施組織の発足のため、令和５年度をめどに最も相応しい事業実施組織の組織形態を決定する。			
ELSI・PPI	○本格解析に向けてELSI・PPIに係る事項について検討	○事業実施組織にELSI部門を設置し、専門性を備えた人員を配置して、事業全体としてELSIに適切に配慮しつつ計画を実施するために必要な取り組みについて、検討、対応を行う。 ○事業実施組織に患者・市民参画部門を設置することに加え、本計画に参画する研究機関・医療機関においても患者・市民の視点を取り入れるための体制を設ける。			

※1 難治性のがん、希少がん（小児がん含む）、遺伝性がん(小児がん含む)等
※2 単一遺伝子性疾患、多因子疾患、診断困難な疾患

4

2023年7月26日 第16回 専門委員会 資料1 （全ゲノム解析等に係る検討状況等について）スライド4より

全ゲノム解析等実行計画

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center the Institute of Medical Science, the University of Tokyo

● 全ゲノムプロジェクト症例内訳とR4年度実施内容

公募の種類	がん種	代表機関・代表者		令和3年度		令和4年度体制・実施内容		
A班：患者還元班（体制構築班）	難治がん等	国立がん研究センター	角南久仁子	500症例（内新規の患者200症例）	計 9,900症例	代表：国立がん研究センター 分担：国立がん研究センター東病院 分担：成育医療研究センター	600症例＋α	600症例の内訳は、新規の患者400症例と、分担医療機関の新規の患者200症例。また、進捗状況に応じて、＋αとして、合わせて最大200症例を追加解析予定。
	難治がん等	静岡がんセンター	浦上研一	500症例（内新規の患者200症例）		代表：静岡がんセンター 分担：近畿大学病院	600症例＋α	
	難治がん等	がん研有明病院	上野貴之	500症例（内新規の患者200症例）		代表：がん研有明病院 分担：慶應義塾病院 分担：大阪大学病院	600症例＋α	
B班 患者還元班（領域別班）	消化器がん	東京大学	柴田龍弘	1,400症例		臨床情報の登録を行うとともに、蓄積された全ゲノムデータを用いた研究を行う。また、A班とも連携しB班全体としての成果をまとめる。		
	血液がん	京都大学	南谷泰仁	1,400症例				
	小児がん	東京大学	加藤元博	1,400症例				
	希少がん	東京大学	松田浩一	1,400症例				
	婦人科がん	がん研有明病院	森誠一	1,400症例				
	呼吸器がん他	国立がん研究センター	河野隆志	1,400症例				
C班：解析班		東京大学医科学研究所	井元清哉	A班、B班併せて、9,900症例の解析		臨床情報を収集するとともに、統一パイプラインによる解析及びレポート作成を行う。また、集中管理システムの構築、全ゲノム解析結果に基づいた臨床応用のための出口戦略の構築を行う。		

✓ 各班は連携し、臨床情報等の収集及び高度な横断的解析等を行う。
✓ 各班は、実施状況について「全ゲノム解析等の推進に関する専門委員会」に報告し、当該委員会の方針に沿って解析等を行う。

2022年7月7日 第10回 専門委員会 資料2－1（AMED提出資料）スライド２より

全ゲノム解析等実行計画

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center the Institute of Medical Science, the University of Tokyo

● **全ゲノムプロジェクト症例内訳とR4年度実施内容**

公募の種類	がん種	代表機関・代表者		令和3年度	令和4年度体制・実施内容
A班：患者還元（体制構築班）					現の医療症例。て、最大定。
B班 患者還元（領域別班）					行う。
	呼吸器がん他	国立がん研究センター	河野隆志		
C班：解析班		東京大学医科学研究所	井元清哉	A班、B班併せて、9,900症例の解析	臨床情報を収集するとともに、統一パイプラインによる解析及びレポート作成を行う。また、集中管理システムの構築、全ゲノム解析結果に基づいた臨床応用のための出口戦略の構築を行う。

9,900 症例の解析を **8 ヶ月**程度で実施。
解析領域として DDN Lustre ファイルシステムを
利用しており、解析完了するのに大いに役立った
(2023/11 時点では 12,500 症例以上)
1 症例あたり **800 GB** 以上
WGS (Fastq ファイル)：400 GB (Normal (x30)：80 GB, Tumor (x120)：320 GB
解析結果 (BAM ファイル、VCF ファイルなど)：400 GB

✓ **各班は連携し、臨床情報等の収集及び高度な横断的解析等を行う。**
✓ **各班は、実施状況について「全ゲノム解析等の推進に関する専門委員会」に報告し、当該委員会の方針に沿って解析等を行う。**

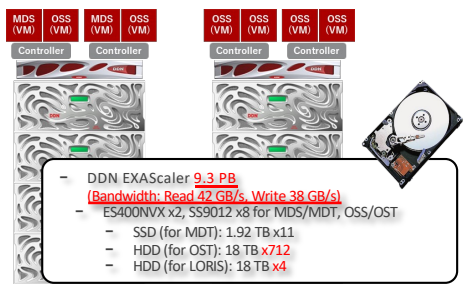
2022年7月7日 第10回 専門委員会 資料2－1（AMED提出資料）スライド２より

全ゲノム解析用システム

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

1次解析用システム（令和3年度に稼働）

1次解析用ストレージシステム

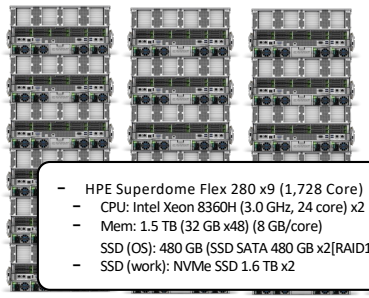


- DDN EXAScaler 9.3 PB
  (Bandwidth: Read 42 GB/s, Write 38 GB/s)
  - ES400NVX x2, SS9012 x8 for MDS/MDT, OSS/OST
    - SSD (for MDT): 1.92 TB x11
    - HDD (for OST): 18 TB x712
    - HDD (for LORIS): 18 TB x4

GPU 搭載サーバ

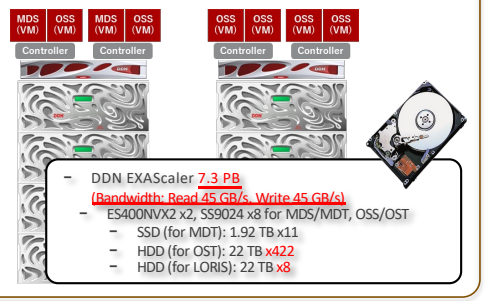
- NVIDIA DGX A100 x4
  - CPU: AMD EPYC 7742 (2.25 GHz, 64 core) x2
  - Mem: 1.0 TB (64 GB x16)
  - SSD (OS): 1.92 TB (NVMe SSD 1.92 TB x2[RAID1])
  - SSD (work): NVMe SSD 3.84 TB x4 [RAID0]

1次解析用サーバ



- HPE Superdome Flex 280 x9 (1,728 Core)
  - CPU: Intel Xeon 8360H (3.0 GHz, 24 core) x2
  - Mem: 1.5 TB (32 GB x48) (8 GB/core)
  - SSD (OS): 480 GB (SSD SATA 480 GB x2[RAID1])
  - SSD (work): NVMe SSD 1.6 TB x2

2次解析用ストレージシステム



- DDN EXAScaler 7.3 PB
  (Bandwidth: Read 45 GB/s, Write 45 GB/s)
  - ES400NVX2 x2, SS9024 x8 for MDS/MDT, OSS/OST
    - SSD (for MDT): 1.92 TB x11
    - HDD (for OST): 22 TB x422
    - HDD (for LORIS): 22 TB x8

2次解析用サーバ

- HPE ProLiant DL385 Gen11 x14 (1,792 Core)
  - CPU: AMD EPYC 9554 (3.1 GHz, 64 core) x2
  - Mem: 1.5 TB (64 GB x24) (12 GB/core)
  - SSD (OS): 480 GB (SSD SATA 480 GB x2[RAID1])
  - SSD (work): NVMe SSD 1.92 TB x2

マスタージョブ用サーバ

- HPE ProLiant DL385 Gen11 x14 (1,792 Core)
  - CPU: AMD EPYC 9554 (3.1 GHz, 64 core) x2
  - Mem: 384 GB (32 GB x12) (3 GB/core)
  - SSD (OS): 480 GB (SSD SATA 480 GB x2[RAID1])
  - SSD (work): NVMe SSD 1.92 TB x2

2次解析用システム（現在構築中）

SHIROKANE

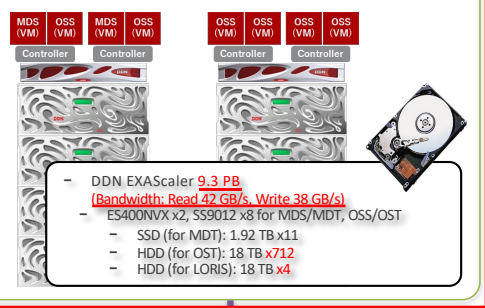
解析用途として、SHIROKANE の計算機リソース（一部）を利用

全ゲノム解析用システム

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

1次解析用システム（令和3年度に稼働）

1次解析用ストレージシステム

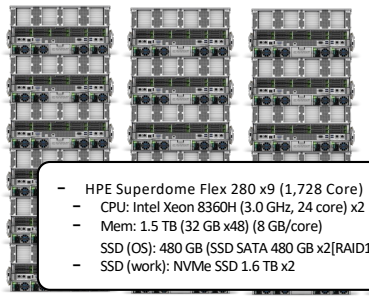


- DDN EXAScaler 9.3 PB
  (Bandwidth: Read 42 GB/s, Write 38 GB/s)
  - ES400NVX x2, SS9012 x8 for MDS/MDT, OSS/OST
    - SSD (for MDT): 1.92 TB x11
    - HDD (for OST): 18 TB x712
    - HDD (for LORIS): 18 TB x4

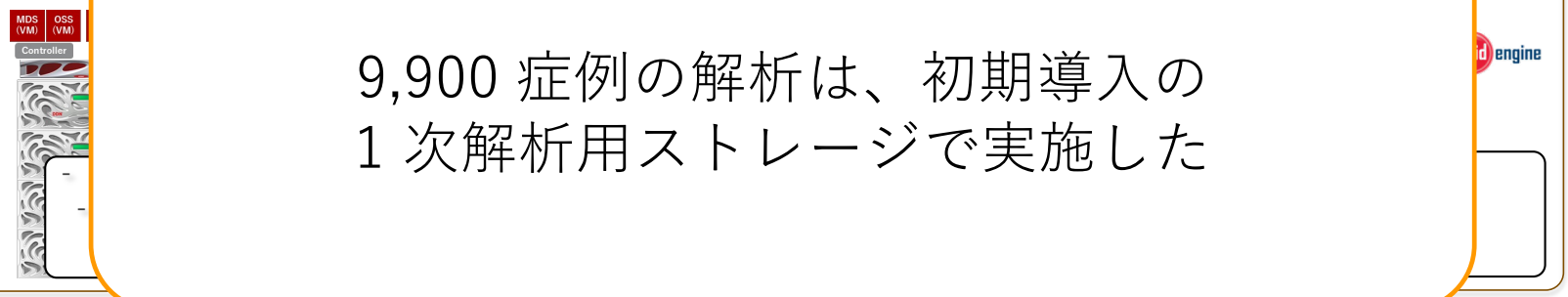
GPU 搭載サーバ

- NVIDIA DGX A100 x4
  - CPU: AMD EPYC 7742 (2.25 GHz, 64 core) x2
  - Mem: 1.0 TB (64 GB x16)
  - SSD (OS): 1.92 TB (NVMe SSD 1.92 TB x2[RAID1])
  - SSD (work): NVMe SSD 3.84 TB x4 [RAID0]

1次解析用サーバ



- HPE Superdome Flex 280 x9 (1,728 Core)
  - CPU: Intel Xeon 8360H (3.0 GHz, 24 core) x2
  - Mem: 1.5 TB (32 GB x48) (8 GB/core)
  - SSD (OS): 480 GB (SSD SATA 480 GB x2[RAID1])
  - SSD (work): NVMe SSD 1.6 TB x2



9,900 症例の解析は、初期導入の
1 次解析用ストレージで実施した

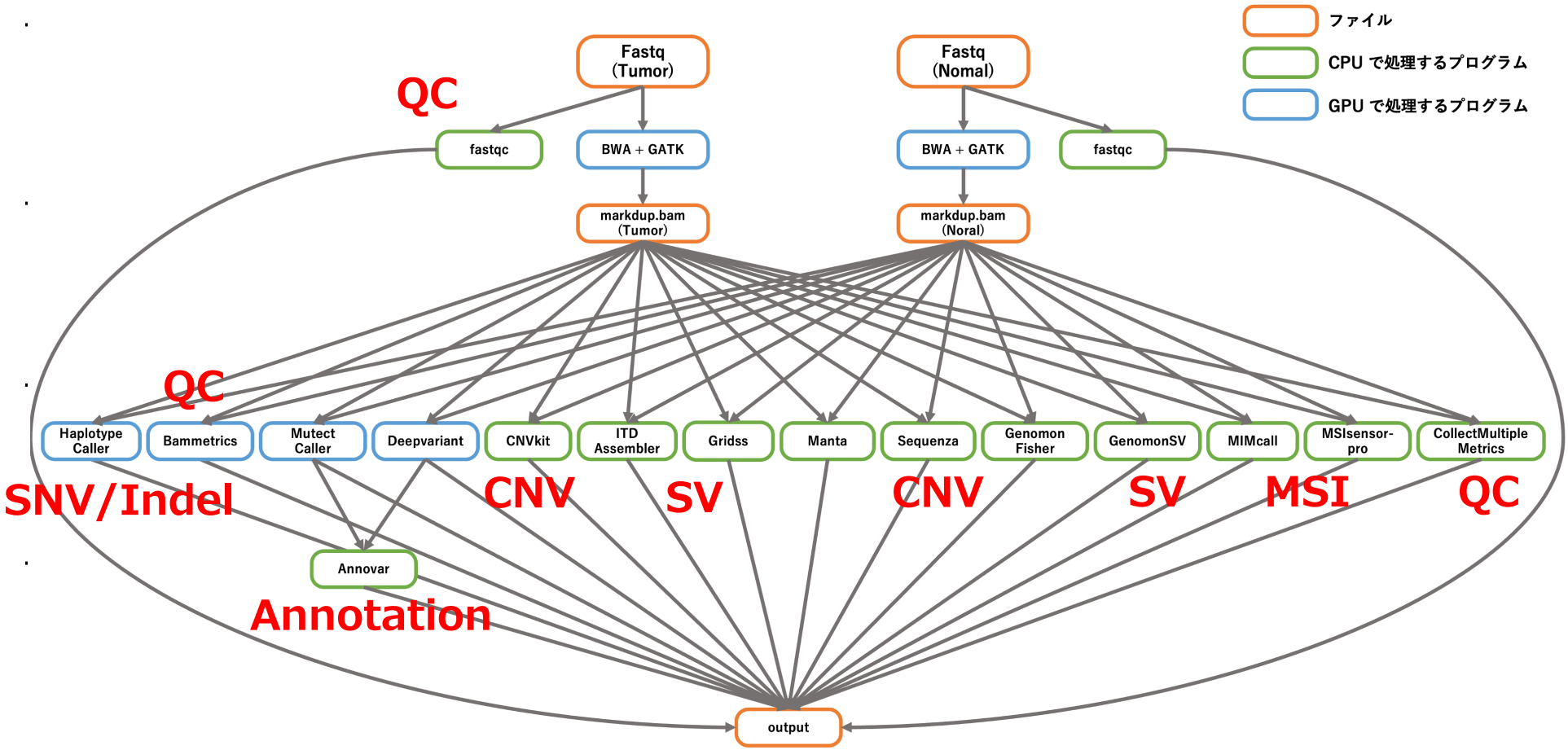
2次解析用システム（現在構築中）

SHIROKANE

解析用途として、
SHIROKANE の
計算機リソース
（一部）を利用

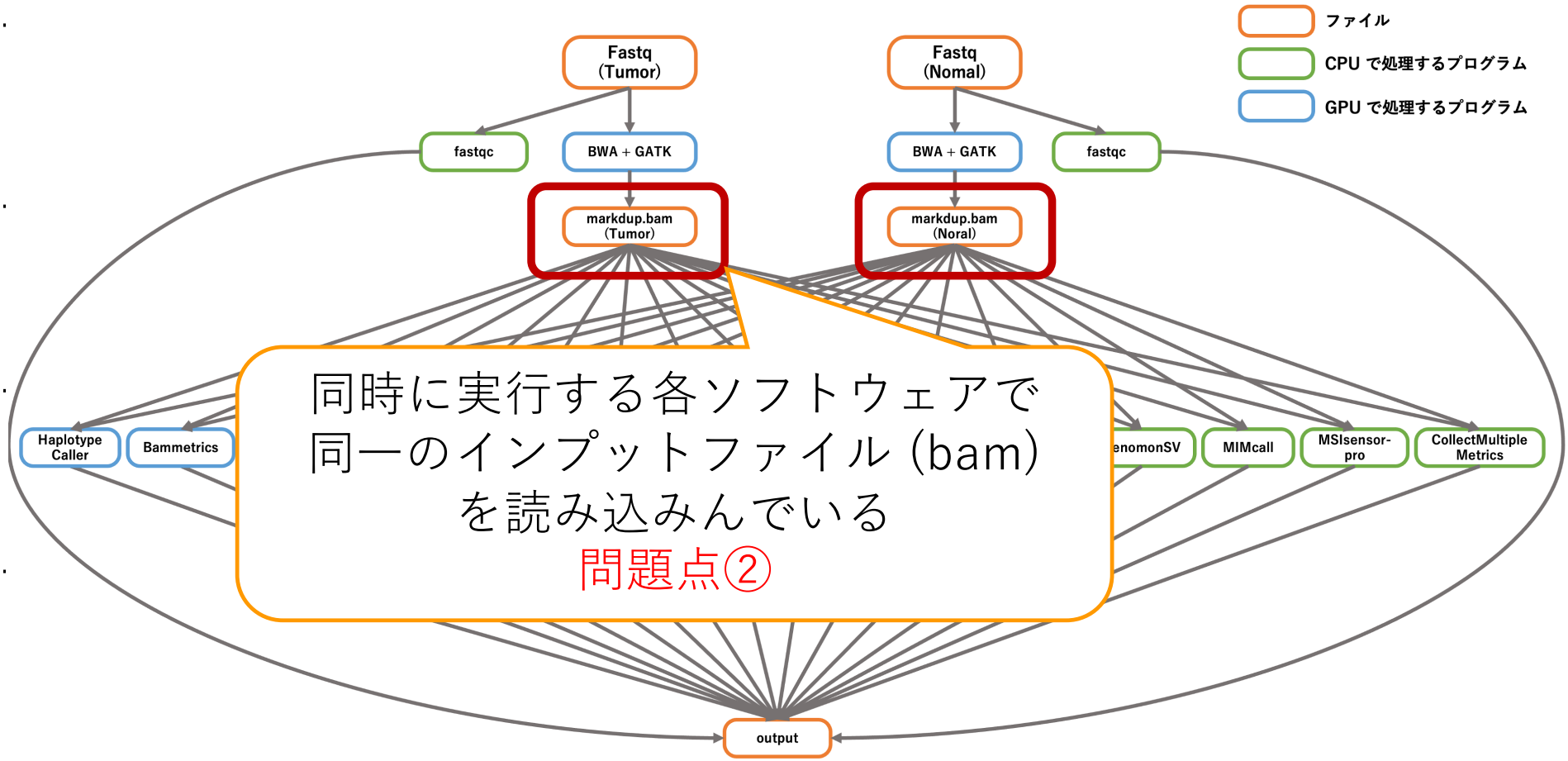
解析用統一パイプライン

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



解析用統一パイプライン

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



実行タイムライン

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

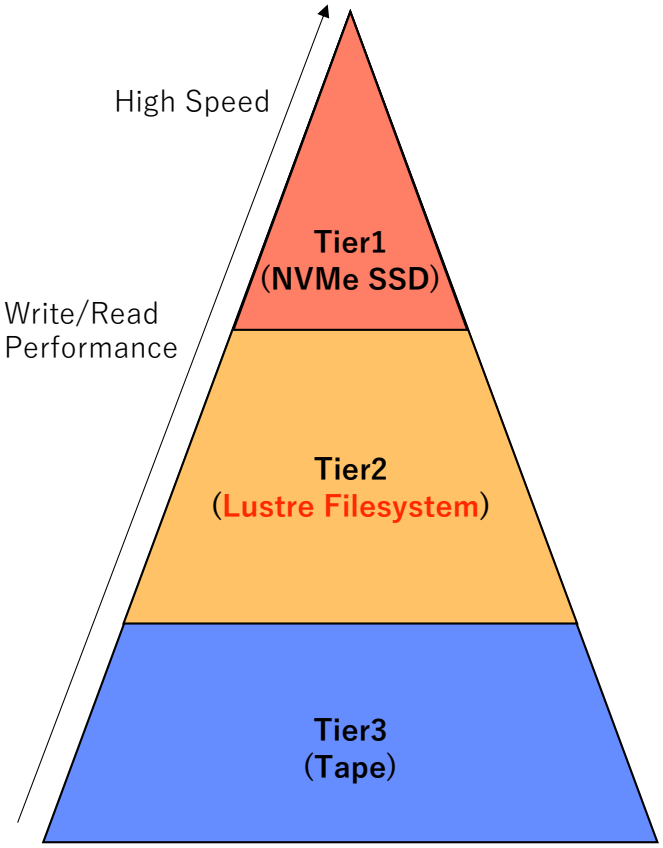
Step name	CPU/GPU hours																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
run_CaGMeJ_1.3.1.sh	1																										
qsub_CaGMeJ.sh	1																										
parabricks_fq2bam	4	4	4	4																							
parabricks_bammetrics					2																						
CollectMultipleMetrics					2	2	2	2																			
parabricks_mutect					2	2	2																				
parabricks_haplotypecaller					2	2	2	2																			
parabricks_deepvariant					2	2	2	2																			
parabricks_cnvkit					12																						
cnvkit_graphics					1																						
cnvkit_compare					4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4			
manta					6	6	6	6																			
gridss					8	8	8	8	8	8	8	8	8	8	8	8	8	8	8								
genomon_pipeline					1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
mutect_annovar								8																			
deepvariant_annovar									16																		
msisensor_pro					1	1	1	1	1																		
NCM_pileup					2																						
NCM_run						1																					
facets_pileup					24																						
facets_R					1																						
mimcall					1																						
mimcall_result						1																					
multiqc									1																		
chord																									1		
chord_summary																									1		
CPUs	2	0	0	0	95	56	54	32	33	15	15	15	14	14	14	14	14	14	14	6	11	11	11	11	9	7	6
GPUs	4	4	4	4	8	6	6	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Memory(GB)	225	94	94	94	949	515	503	304	198	166	166	166	158	158	158	158	158	158	158	145	184	184	184	184	182	170	42
Jobs	4	2	2	2	74	44	42	12	9	5	5	5	4	4	4	4	4	4	4	3	3	3	3	3	4	2	1

実行タイムライン

Step name	CPU/GPU hours																														
	1	2	3	4	5	6	7	8	9																						
run_CaGMeJ_1.3.1.sh	1																														
qsub_CaGMeJ.sh	1																														
parabricks_fq2bam	4	4	4	4																											
parabricks_bammetrics					2																										
CollectMultipleMetrics					2	2	2	2																							
parabricks_mutect					2	2	2																								
parabricks_haplotypecaller					2	2	2																								
parabricks_deepvariant					2	2																									
parabricks_cnvkit					12																										
cnvkit_graphics					1																										
cnvkit_compare					4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4				
manta					6	6	6	6																							
gridss					8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8										
genomon_pipeline					1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
mutect_annovar								8																							
deepvariant_annovar									16																						
msisensor_pro					1	1	1	1	1																						
NCM_pileup					2																										
NCM_run						1																									
facets_pileup					24																										
facets_R					1																										
mimcall					1																										
mimcall_result						1																									
multiqc									1																						
chord																												1		
chord_summary																												1			
CPUs	2	0	0	0	95	56	54	32	33	15	15	15	14	14	14	14	14	14	14	6	11	11	11	11	9	7	6				
GPUs	4	4	4	4	8	6	6	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
Memory(GB)	225	94	94	94	949	515	503	304	198	166	166	166	158	158	158	158	158	158	158	145	184	184	184	184	182	170	42				
Jobs	4	2	2	2	74	44	42	12	9	5	5	5	4	4	4	4	4	4	4	3	3	3	3	3	4	2	1				

同一ファイルをインプットした
ジョブが同時に実行される
問題点②

ストレージの使い方



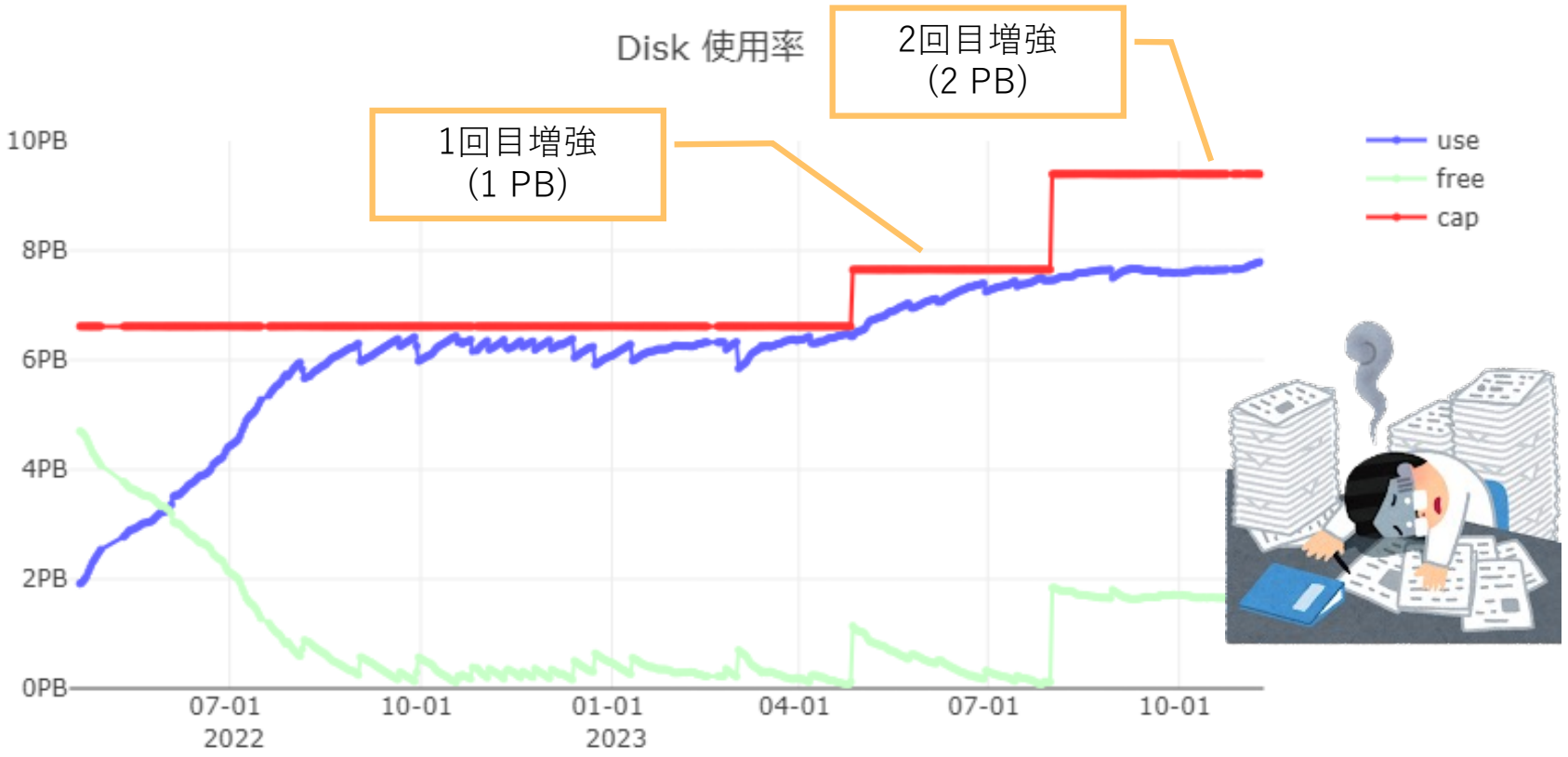
・・・統一パイプラインの一部ソフトウェアの
スクラッチ領域として使用

・・・統一パイプラインの全ソフトウェアの
インプット/アウトプット領域として使用
スクラッチ領域を指定可能なソフトウェアは
少なく、解析に重要な領域

・・・解析が終了した検体置き場として使用

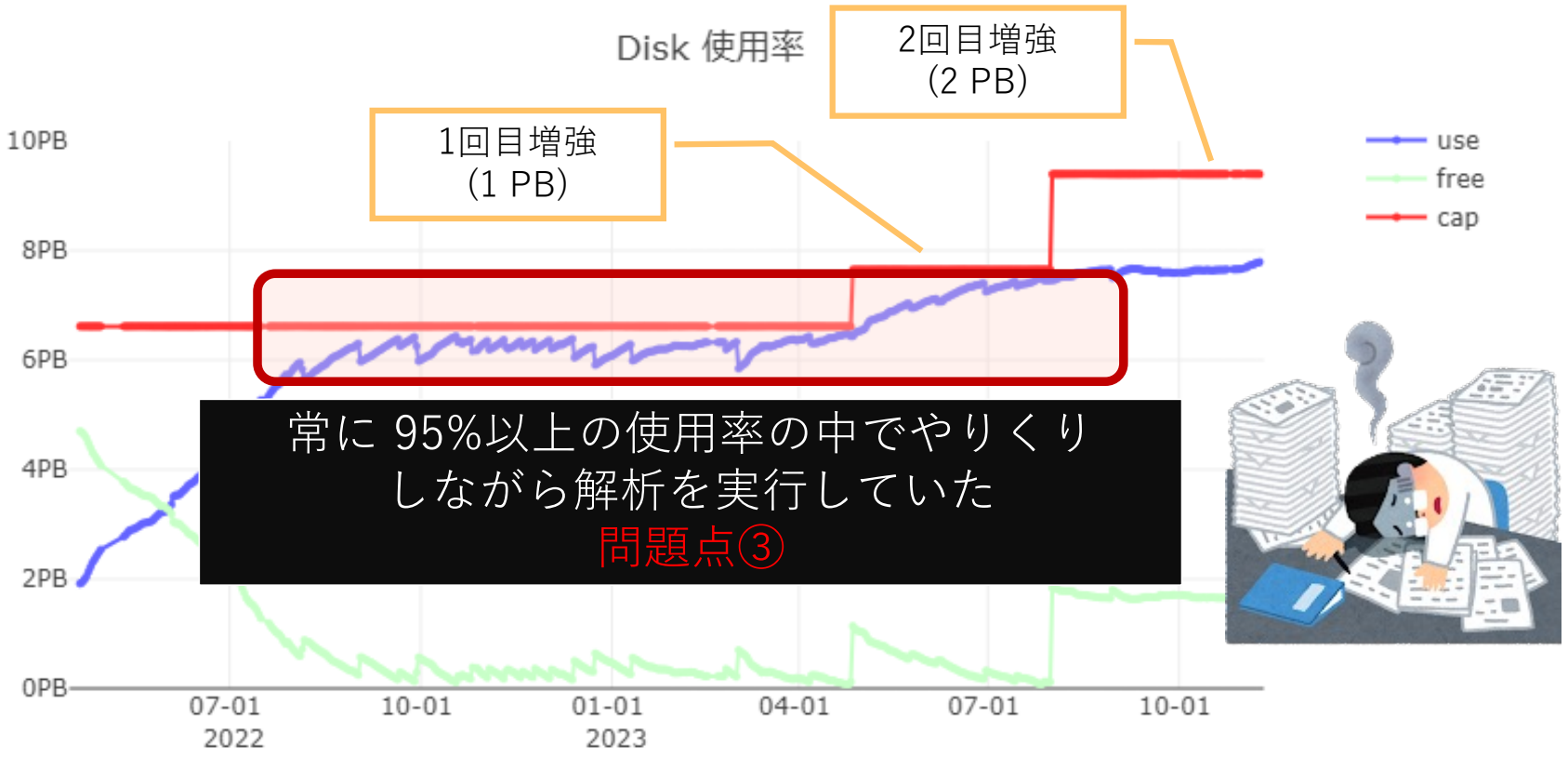
革新がんのデータ量推移

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



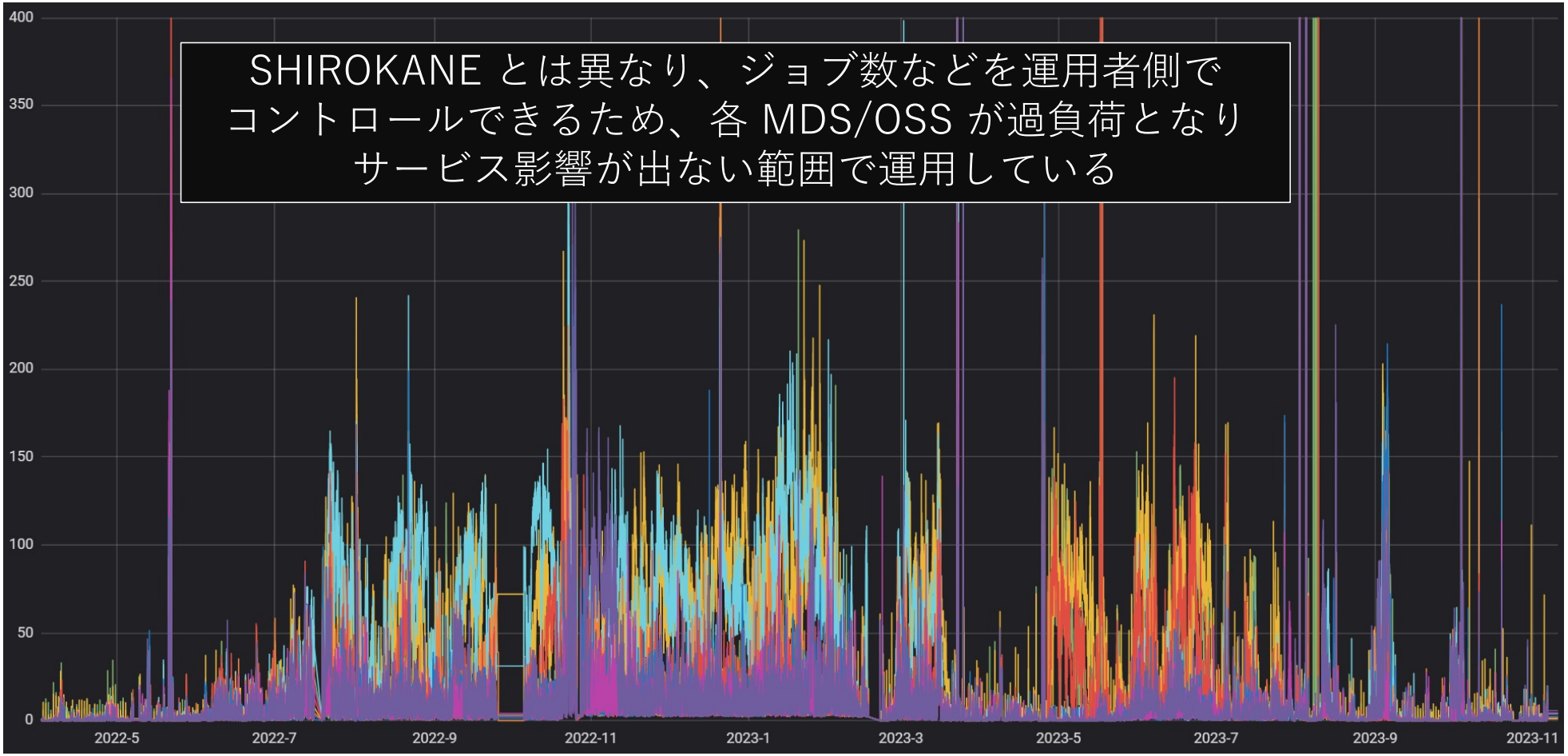
革新がんのデータ量推移

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

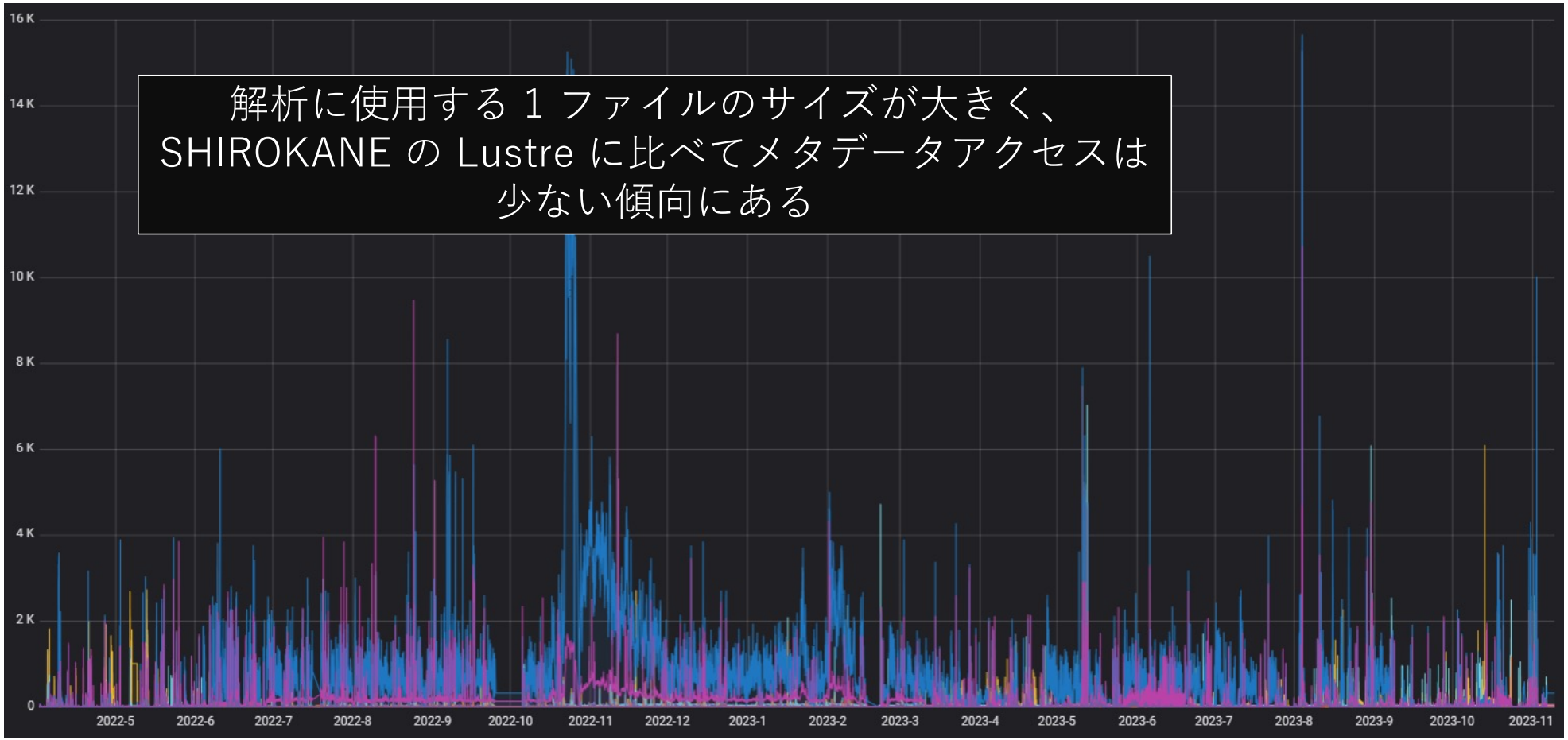


革新がん MDS/OSS の負荷状況

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

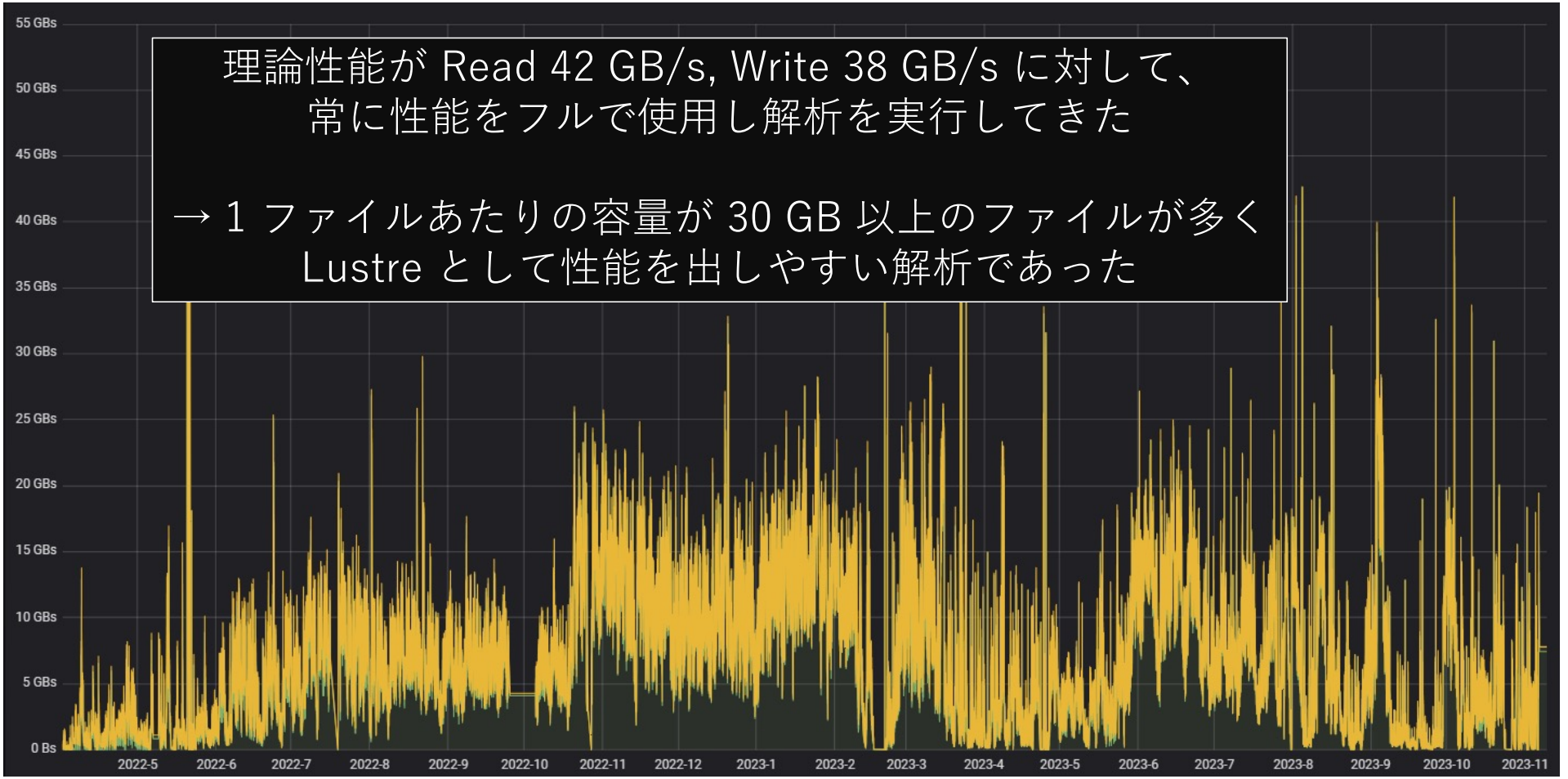


Lustre のメタデータアクセス



Lustre のアクセスバンド幅

東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo



問題点②

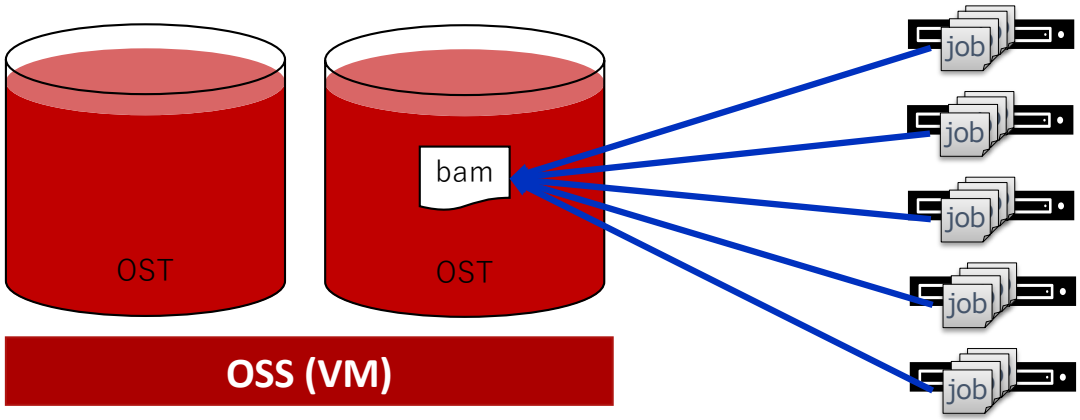
東京大学 THE UNIVERSITY OF TOKYO
医科学研究所 THE INSTITUTE OF MEDICAL SCIENCE
Human Genome Center
the Institute of Medical Science, the University of Tokyo

✓状況

- 統一パイプラインではアライメント後の処理 (変異コール) は、同じ bam ファイルをインプットとして実行する
- ストライプカウントは「1」としている

✓問題点

- 同一の bam ファイルを複数のジョブで読み込むことで、実データが格納された OST への I/O が増え、OST をマウントしている OSS で過負荷を引き起こす



問題点②

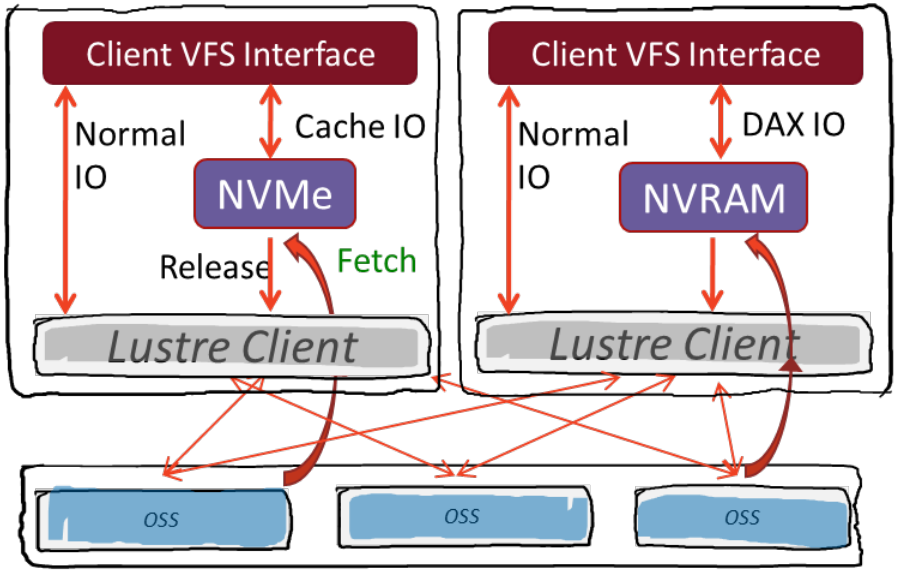
✓対策案

➢bam ファイルのストライプカウントを複数の OST でストライプするか検討中

- ストライプカウントをいくつにするか未検討。Lustre コミュニティとしてアイデアがあれば教えてほしい

➢EXA6 の新機能である Hot Nodes が使えないか検討中

- クライアント側でキャッシュを持つことでデータを再利用できることで OSS 負荷の軽減、統一パイプラインのスループット向上を期待
- ジョブスケジューラに Altair Grid Engine を使用しているが、キャッシュを持っているクライアントに対象ジョブを Dispatch する方法を検討中。
  Lustre コミュニティとしてアイデアがあれば教えてほしい



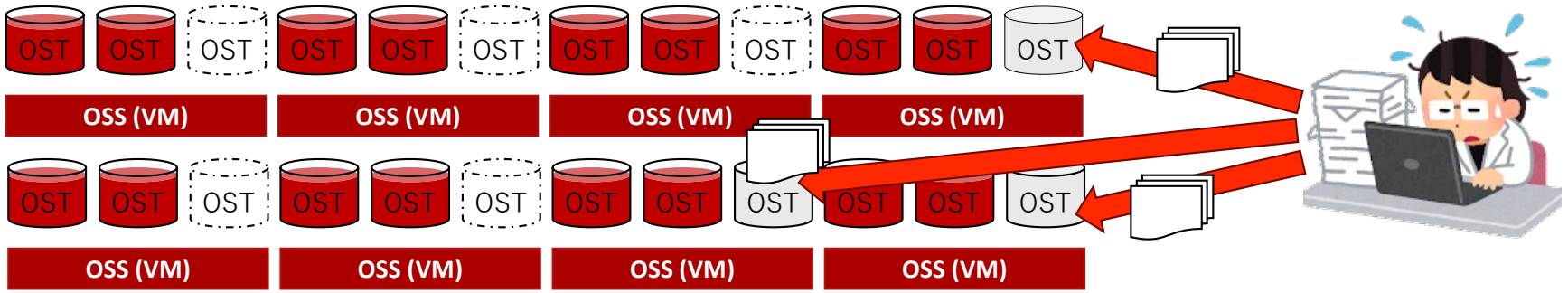
苦労話 (問題点③)

✓HDD 1 回目増強時の状況

- ➢バランスよく OST を増設できず、OSS の OST マウント数に偏りが出てしまった
- ➢既設の OST の使用率は 95% 以上とかなりひっ迫していた

✓問題点

- ➢Lustre の以下のデフォルト仕様により、増設した OST にのみ新規の I/O が偏り、対象 OST をマウントしている OSS が高負荷となり I/O エラーが多発し、解析スループットが 40% 程度に低下した
  - 100 - (最も使用量の少ない OST/最も使用量の多い OST) * 100 が 17% 以上であれば使用量の少ない OST にファイルを書き込む



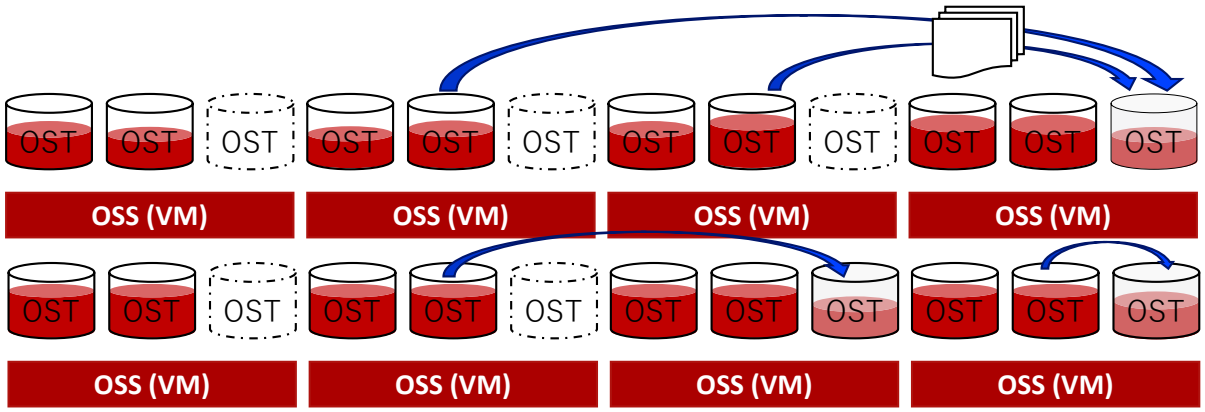
苦労話 (問題点③)

✓対策

➢使用率の高い OST から使用率の低い OST へ実データを移動

- lfs migrate コマンドで Read 予定のないデータを選択し移動

➢既設 OST に書き込まれた解析済みの検体データ (400 GB) を徐々に削除



✓教訓

➢OST の使用率が高い状況で HDD 増強する場合は、増設する新規 OST の数は OSS で均等になるように設計してから実施しないと運用上苦しくなる

- 予算の関係上から難しい場合もあるが極力意識する必要がある

Luster コミュニティへのお願い

✓アクセスするケースごとにストライプカウントの推奨値があれば教えてほしい

✓lfs find コマンドを OS 標準の find コマンドと同等のオプションが使えるともう少し便利になりそう

謝辞

✓株式会社日立製作所の
- シロカネエキスパート（SE）の皆様
- 革新がん解析チームの皆様

小寺 晋平 様
門間 則和 様
加藤 誉運 様
高木 公介 様
飯沼 和也 様
高岡 春佳 様
竹内 祐哉 様

樫村 洋平 様→会場にいます
関　隆博 様
黒田 哲　 様
三吉 直紀 様
大久保 哲 様
川田 直斗 様
大平 勇人 様
喜多見 将 様
小宮山 萌 様
小松 清隆 様
引田 真陽 様
孕石 裕昭 様