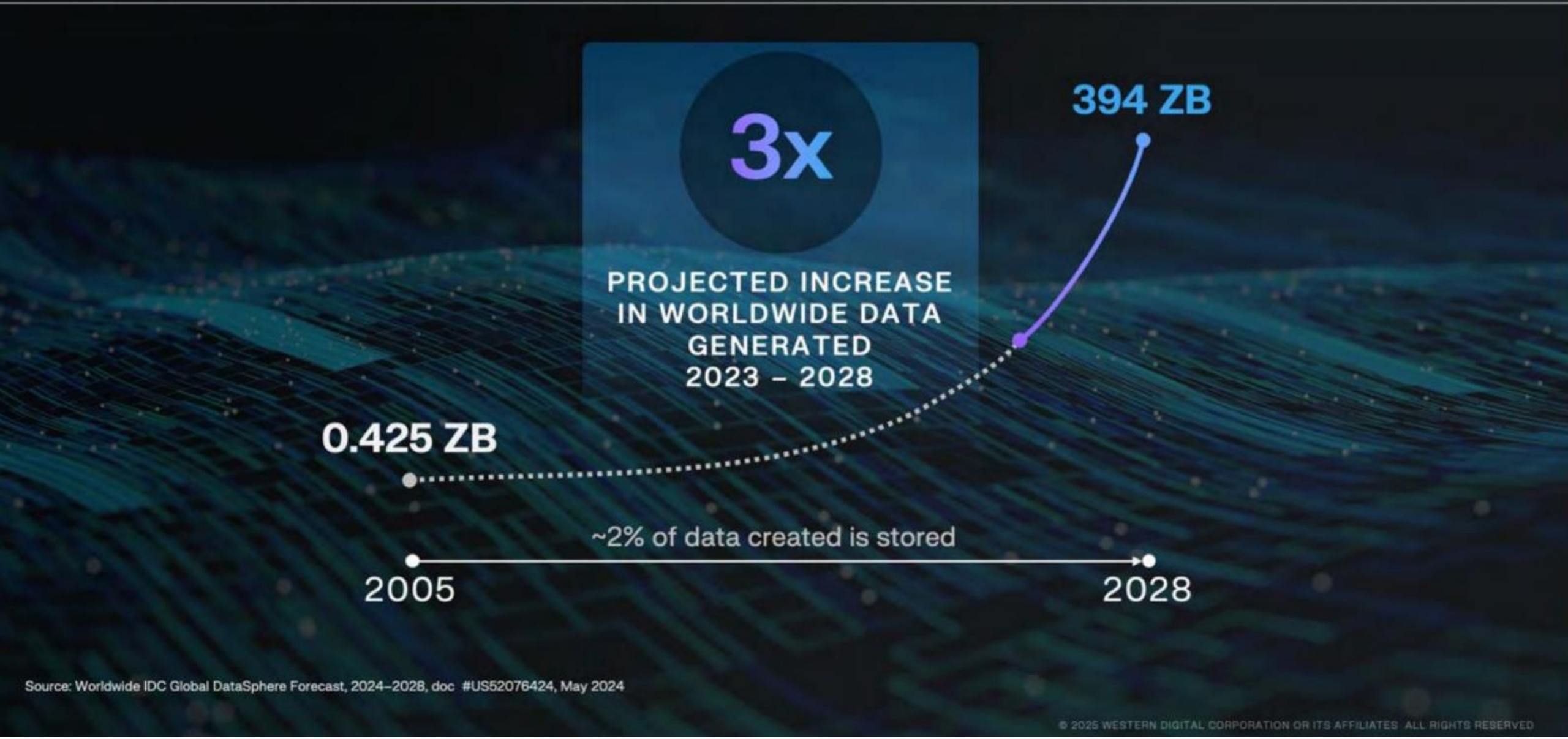
Международная конференция
«Суперкомпьютерные дни в России 2025»

Аспекты эффективности работы параллельных файловых систем на базе высокопроизводительных коммуникационных сетей

В.Е. Машков, В.В. Стегайлов

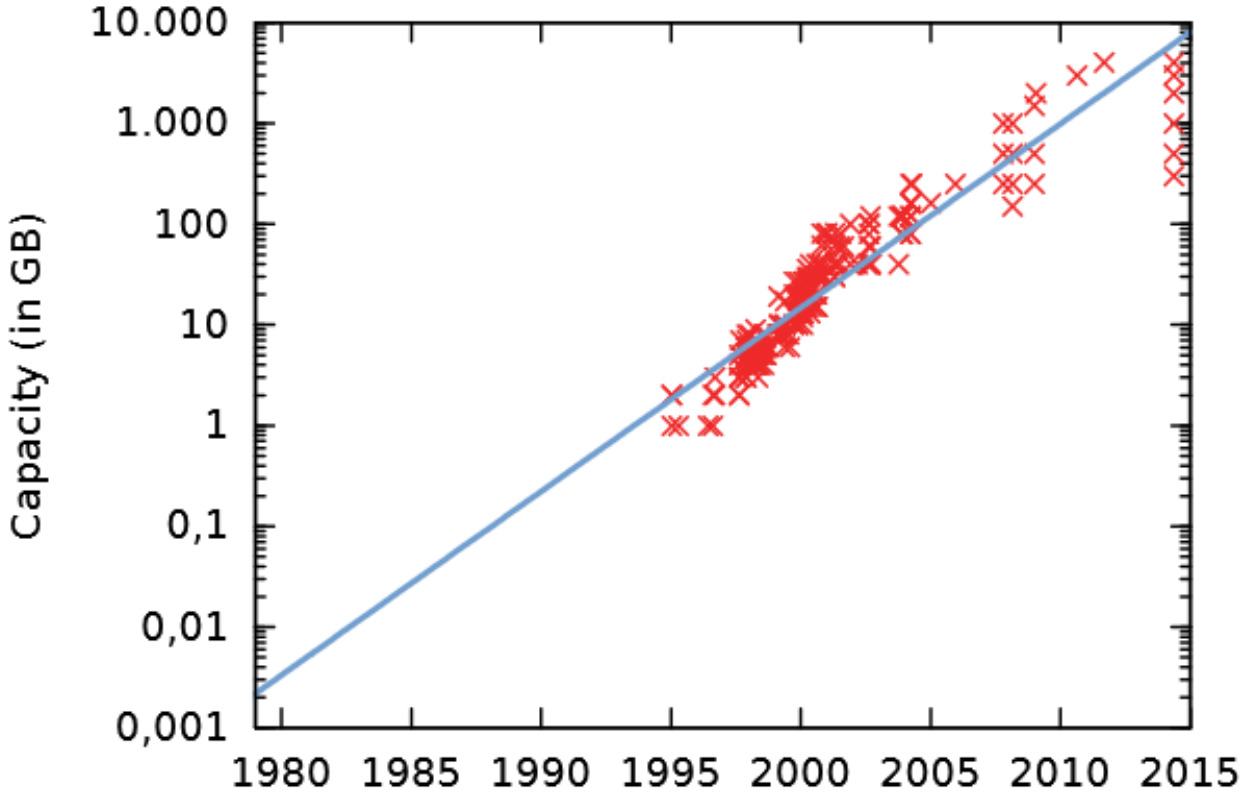
29 сентября 2025 г.

Data Generated Will See Massive Growth

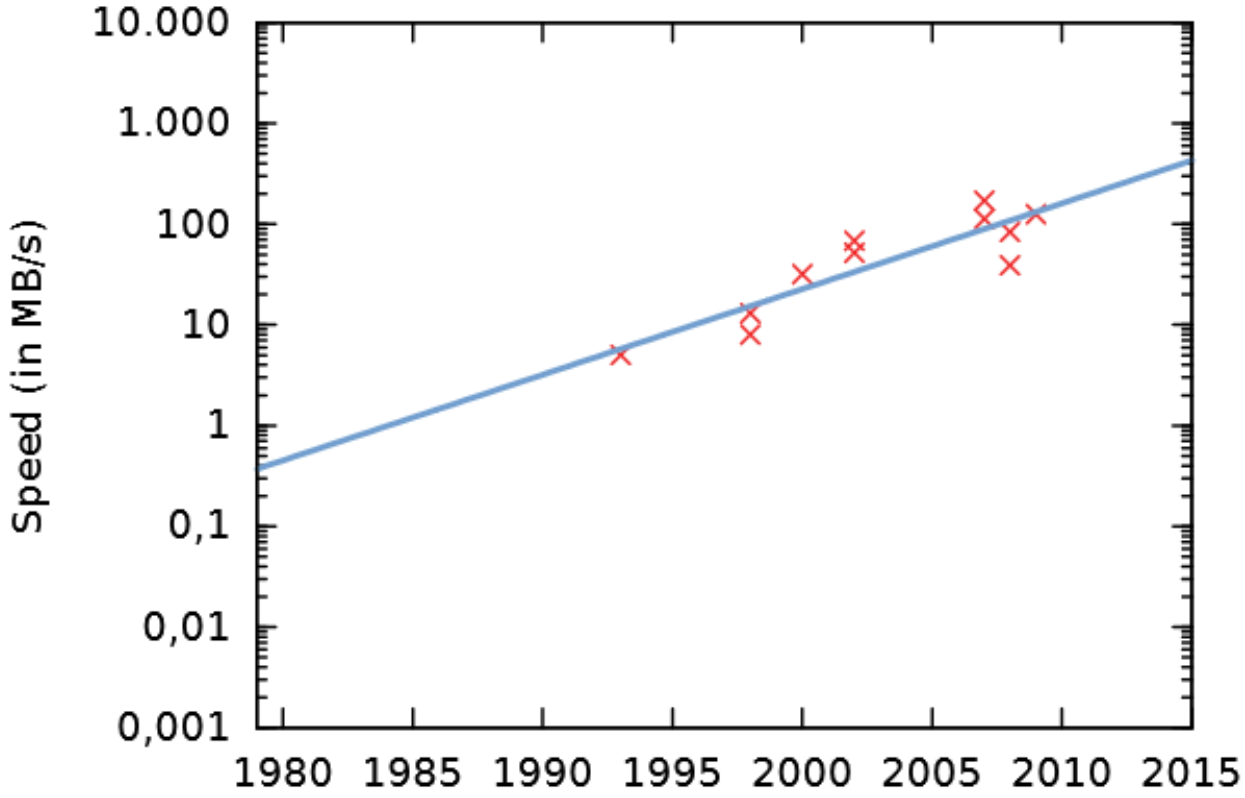


Source: Worldwide IDC Global DataSphere Forecast, 2024–2028, doc #US52076424, May 2024

Развитие технологий хранения данных: рост емкости хранения/скорости записи данных



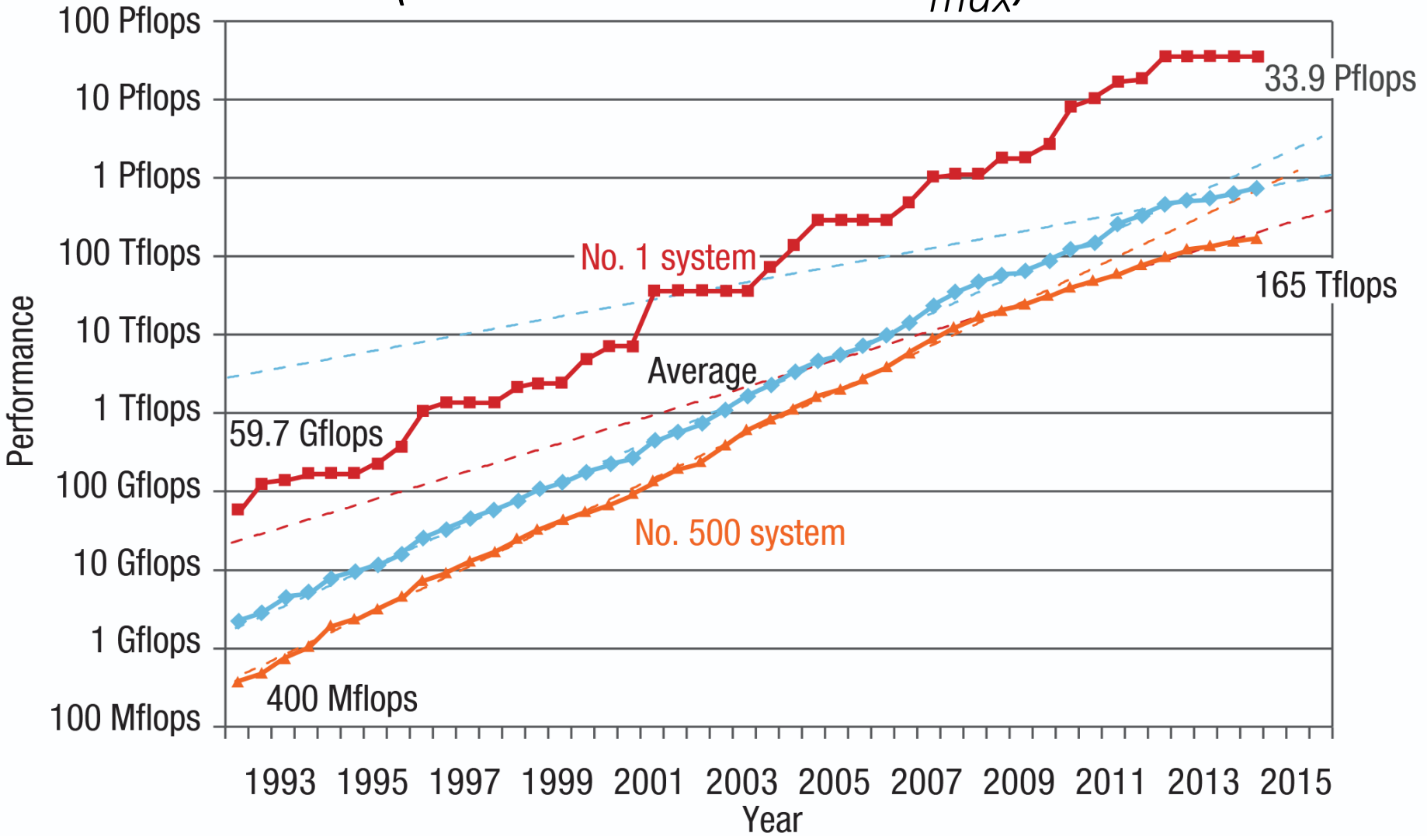
a) Capacity over the years



b) Speed over the years

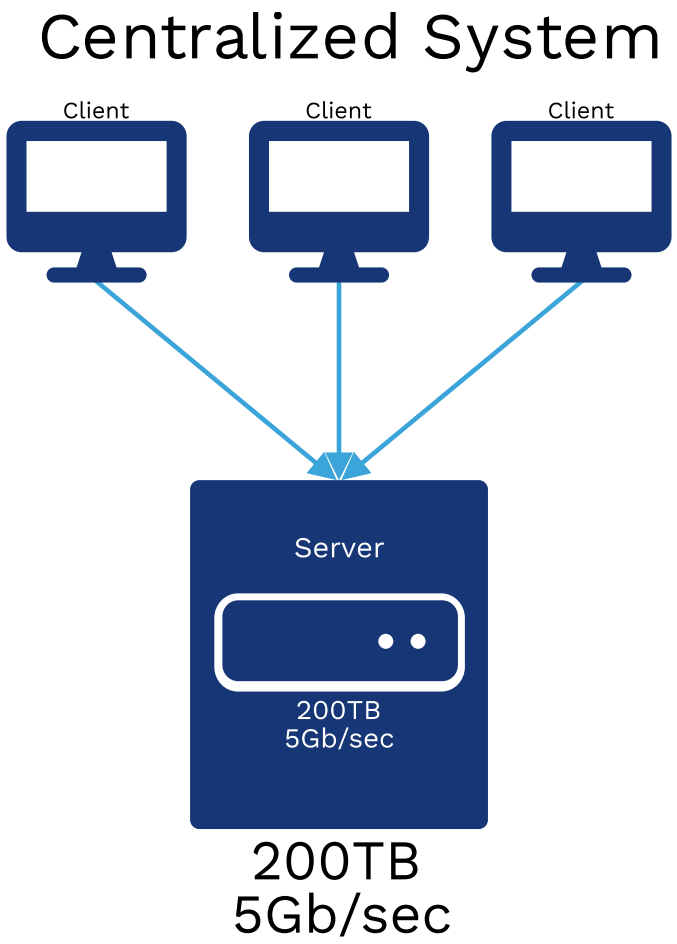
Kunkel et al. 2014, DOI: 10.14529/jsfi140106

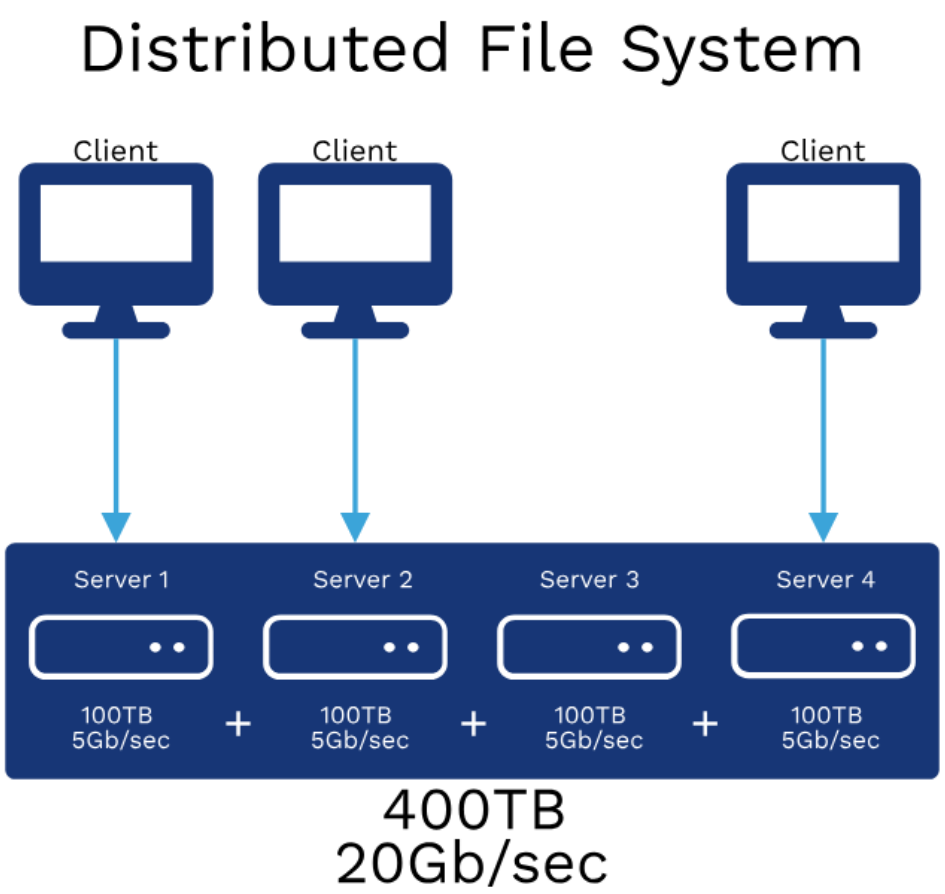
Производительность суперкомпьютеров (по TOP500 HPL R_{max})



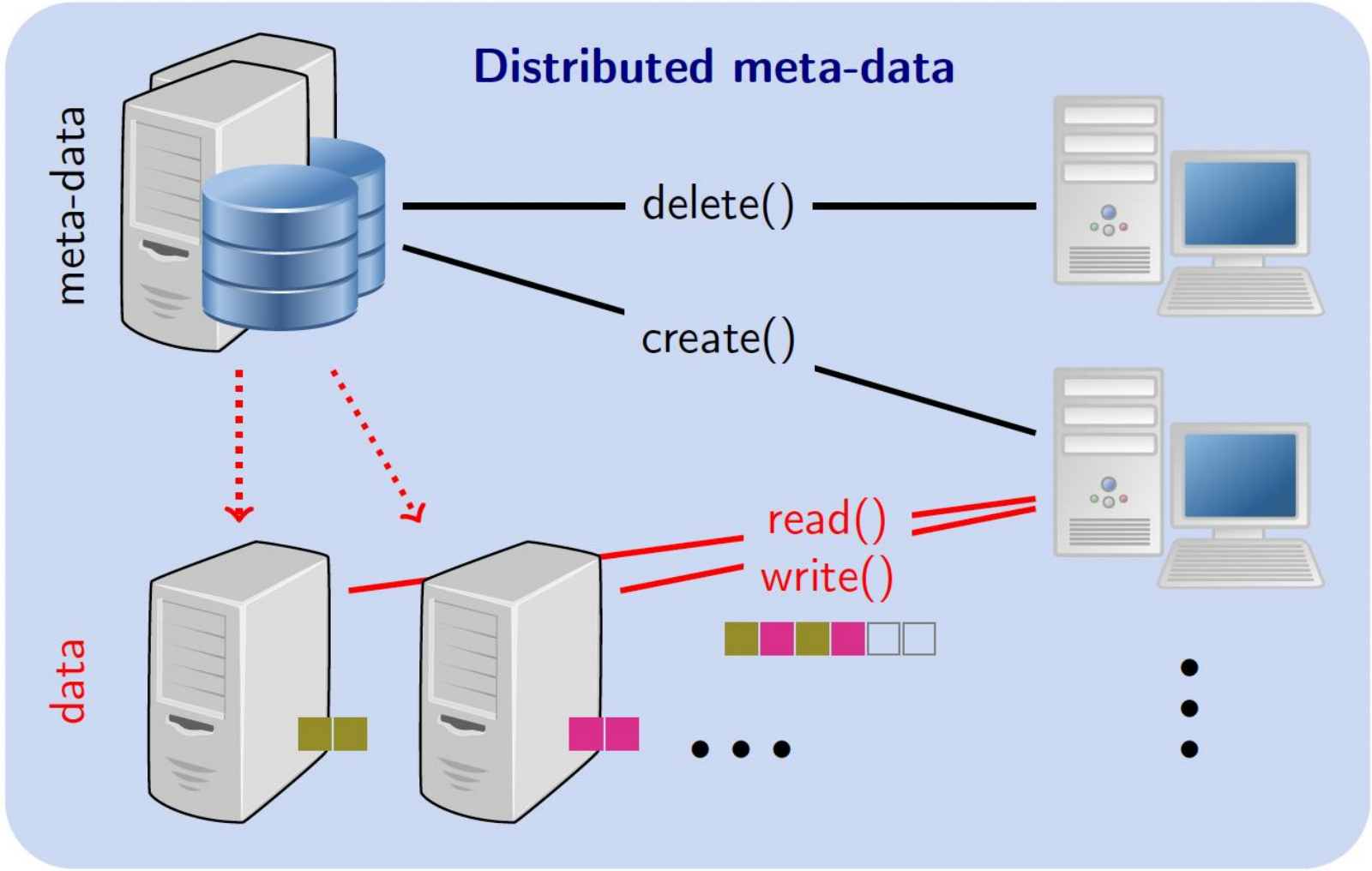
Strohmaier et al., 2015, DOI: 10.1109/MC.2015.338

Параллельность — не «приятная опция», а необходимость





Что такое ПФС, их преимущества и недостатки

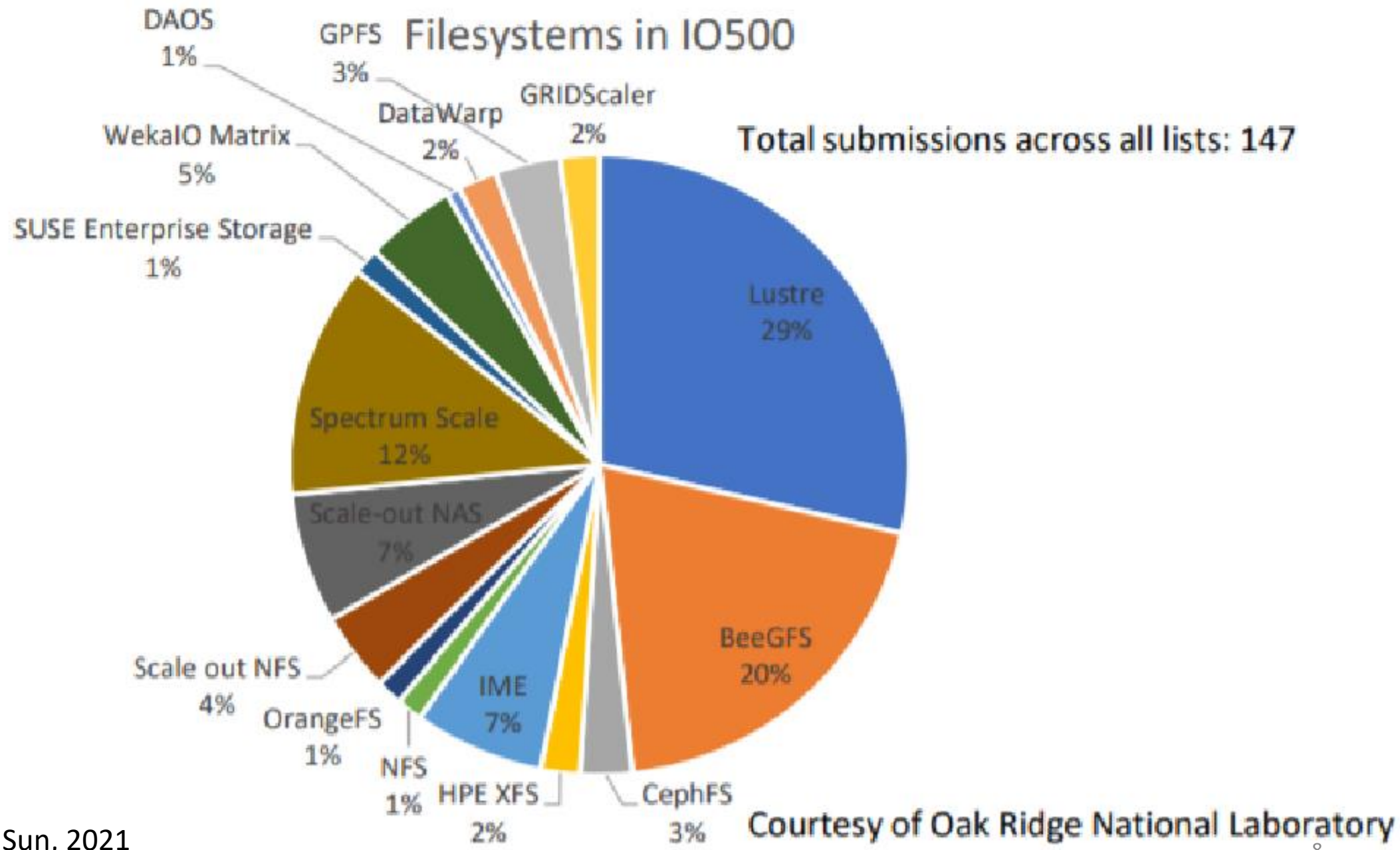


Преимущества:

- *Масштабирование пропускной способности*
- *Параллелизм доступа*
- *Единое пространство имён (POSIX)*
- *Гибкая настройка*
- *Высокая доступность и ремонтопригодность*
- *Интеграция с RDMA/высокоскоростными сетями*

Недостатки:

- *Узкие места метаданных и мелких файлов*
- *Накладные расходы*
- *Зависимость от сети*
- *Операционные издержки*
- *Зависимость от настроек сети*
- *Трудности смешанных профилей*



P Shah, J Ye, XH Sun, 2021

Как сравнивают производительность?

IO500

Home About Steering Lists BoFs Rules Running Submission News Graphs Contact

YOU ARE HERE LISTS / SC24 / Production LIST

Production SC24 List

Customize Download

Production | 10 Node Production | Research | 10 Node Research | Full | Historical

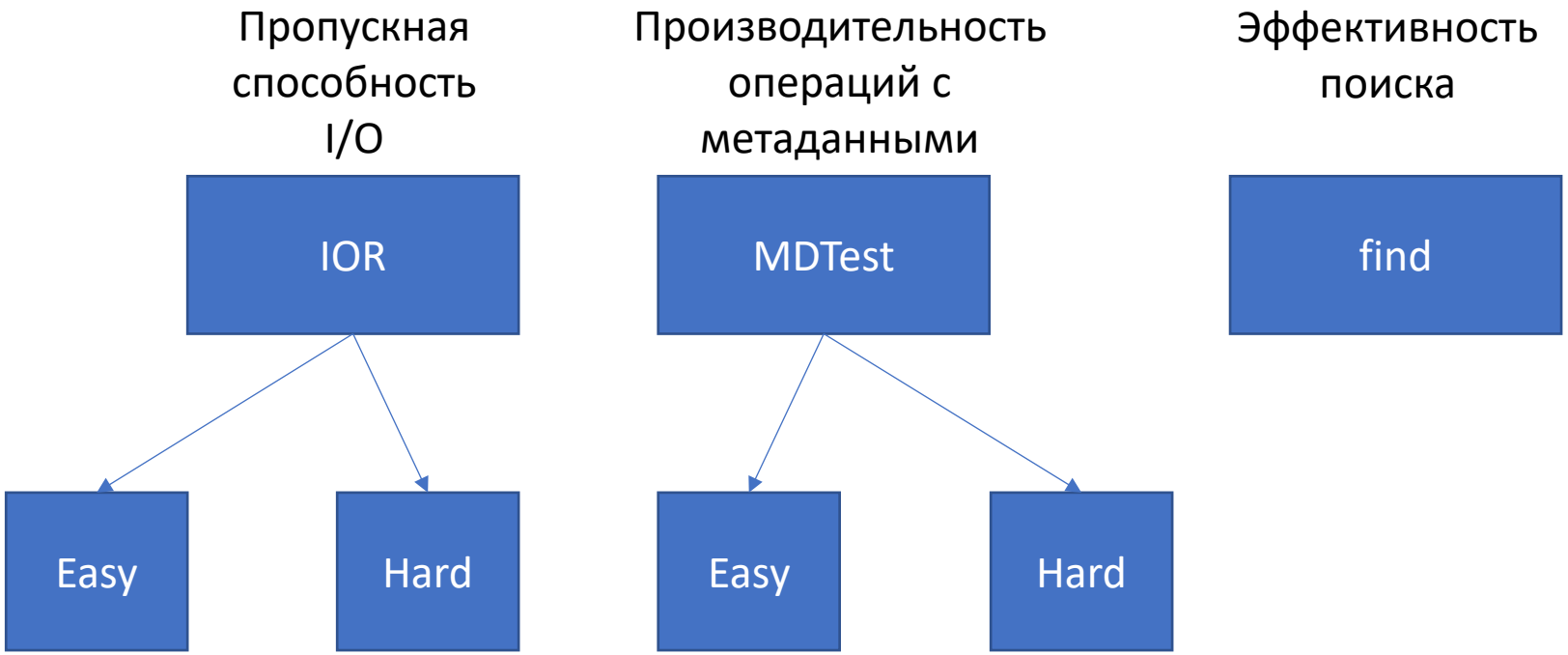
Ranking of production system submissions. This is a subset of the Full List of submissions, showing only one highest-scoring result per storage system. Submitters who want a submission that is currently on the Research List to be on the Production List should contact the IO500 Steering Committee.

# ↑	INFORMATION							IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	✓
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	✓
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	✓
4	SC24	MSKCC	IRIS	WekaIO	WekaIO	261	27,144	665.49	252.54	1,753.69	✓
5	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	✓
6	SC24	SoftBank Corp	CHIE-3	DDN	EXAScaler	240	26,880	500.20	331.66	754.41	✓
7	SC24	Danish Centre for AI innovation AS	GEFION	DDN	EXAScaler	128	12,288	368.56	209.06	649.73	✓
8	ISC24	Zuse Institute Berlin	Lise	Megware	DAOS	10	960	324.54	65.01	1,620.13	✓
9	SC24	University of Florida	HiPerGator AI	DDN	EXAScaler	10	640	243.61	124.89	475.20	✓
10	ISC22	China Telecom Research Institute	CTPAI	CTCLOUD	DAOS	10	200	187.84	25.29	1,395.01	-

Аспекты IO500:

- Прозрачность и воспроизводимость
- Сообщество и развитие
- Полнота оценки
- Публикация результатов и элемент «Соревнования»

Структура набора тестов IO500



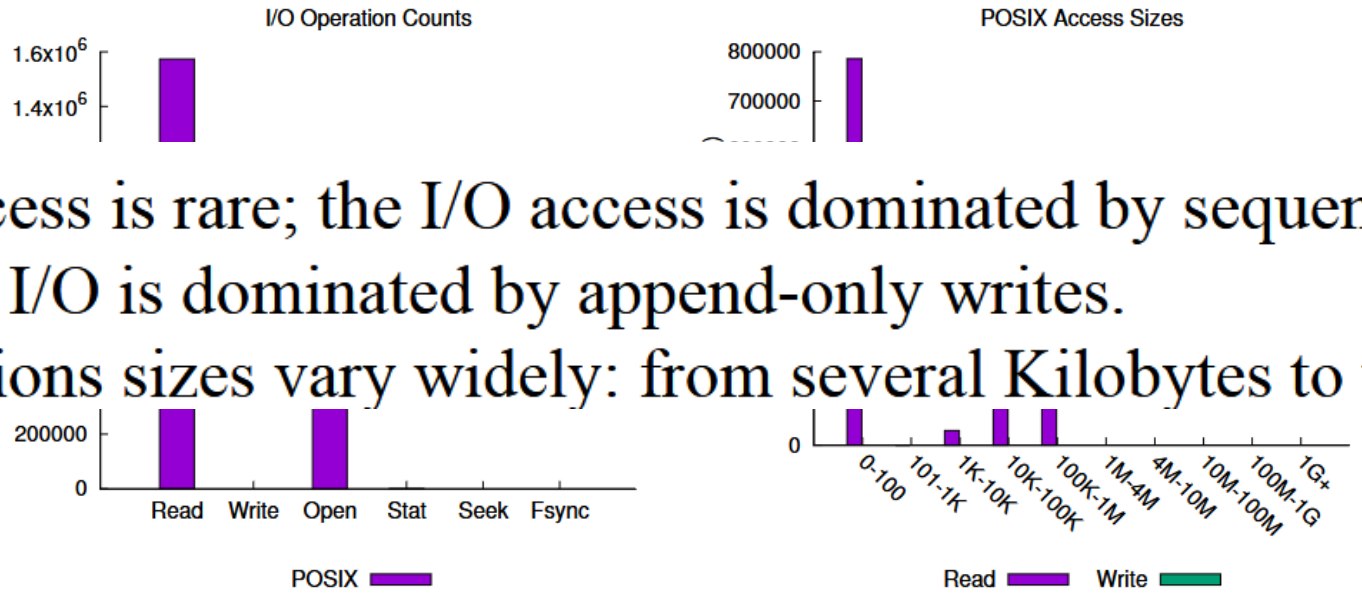
- **Easy — тесты для больших потоковых операций и простых сценариев.**
- **Hard — тесты для интенсивного, часто случайного доступа и работы с более сложными операциями**

Узкое место ПФС — метаданные

Масштабирование сводится к трём идеям:

1. **Делить один горячий каталог** на части и обслуживать их несколькими серверами метаданных одновременно (Lustre DNE2)
2. **Разносить каталоги по разным серверам**, чтобы нагрузка распределялась по дереву (BeeGFS)
3. **Кэшировать результаты** запросов метаданных, чтобы повторные обращения шли быстрее (IBM Spectrum Scale)

Какие операции ввода-вывода характерны для задач HPC и ML?



(a) Ops vs. I/O Operations (b) Operation Count vs. File Sizes

- Random access is rare; the I/O access is dominated by sequential read/write.
- Application I/O is dominated by append-only writes.
- I/O transactions sizes vary widely: from several Kilobytes to tens of megabytes.

Figure 4: TensorFlow ImageNet data reader I/O pattern

Shan & Shalf 2007. Using IOR to Analyze the I/O Performance for HPC Platforms.
Fahim Chowdhury et al 2019 Association for Computing Machinery https://doi.org/10.1145/3337821.3337902

International Journal of
HIGH PERFORMANCE
COMPUTING APPLICATIONS

Special Issue Paper

Refining HPCToolkit for application performance analysis at exascale

The International Journal of High Performance Computing Applications
2024, Vol. 38(6) 612–632

DOI: 10.1177/10943420241277839
journals.sagepub.com/home/hpc

Laksono Adhianto[1,*], Jonathon Anderson[1*], Robert Matthew Barnett[1,*], Dragana Grbic[1,*], Vladimir Indic[2,*], Mark Krentel[1,*], Yumeng Liu[1,*], Srđan Milaković[1,*], Wileam Phan[1,*] and John Mellor-Crummey[1]

Abstract

As part of the US Department of Energy's Exascale Computing Project (ECP), Rice University has been refining its HPCToolkit performance tools to better support measurement and analysis of applications executing on exascale supercomputers. To efficiently collect performance measurements of GPU-accelerated applications, HPCToolkit employs novel non-blocking data structures to communicate performance measurements between tool threads and

compute nodes, even without using *hpcrun* to collect performance data.

The root cause behind these crashes is the limited performance of Frontier's Orion filesystem (Oak Ridge Leadership Computing Facility, 2024)—a shared Lustre filesystem that stores LAMMPS, HPCToolkit, and their library dependencies. Before an application such as LAMMPS can begin execution, the dynamic linker ld.so must first load all of the application's library dependencies into the application's address space. When more than a few thousand compute nodes simultaneously attempt to read library dependencies out of the shared filesystem, the fil-

IO500 SCORES

IO500 SCORE	32,165.93
IO500 BW	10,066.09 GiB/s
IO500 MD	102,785.41 kIOP/s

INFORMATION

CLIENT NODES	300
CLIENT TOTAL PROCS	62,400

IOR & FIND

EASY WRITE	20,693.63 GiB/s
EASY READ	12,122.87 GiB/s
HARD WRITE	4,216.34 GiB/s
HARD READ	9,706.55 GiB/s
FIND	229,672.10 kIOP/s

METADATA

EASY WRITE	60,985.13 kIOP/s
EASY STAT	225,295.35 kIOP/s
EASY DELETE	57,648.44 kIOP/s
HARD WRITE	33,827.19 kIOP/s
HARD READ	141,467.16 kIOP/s
HARD STAT	230,086.03 kIOP/s
HARD DELETE	62,196.78 kIOP/s

io500.org 10 Node Production SC24 List

Высокопроизводительные сети как транспорт ПФС (IB / RoCE / OPA / «Ангара»)

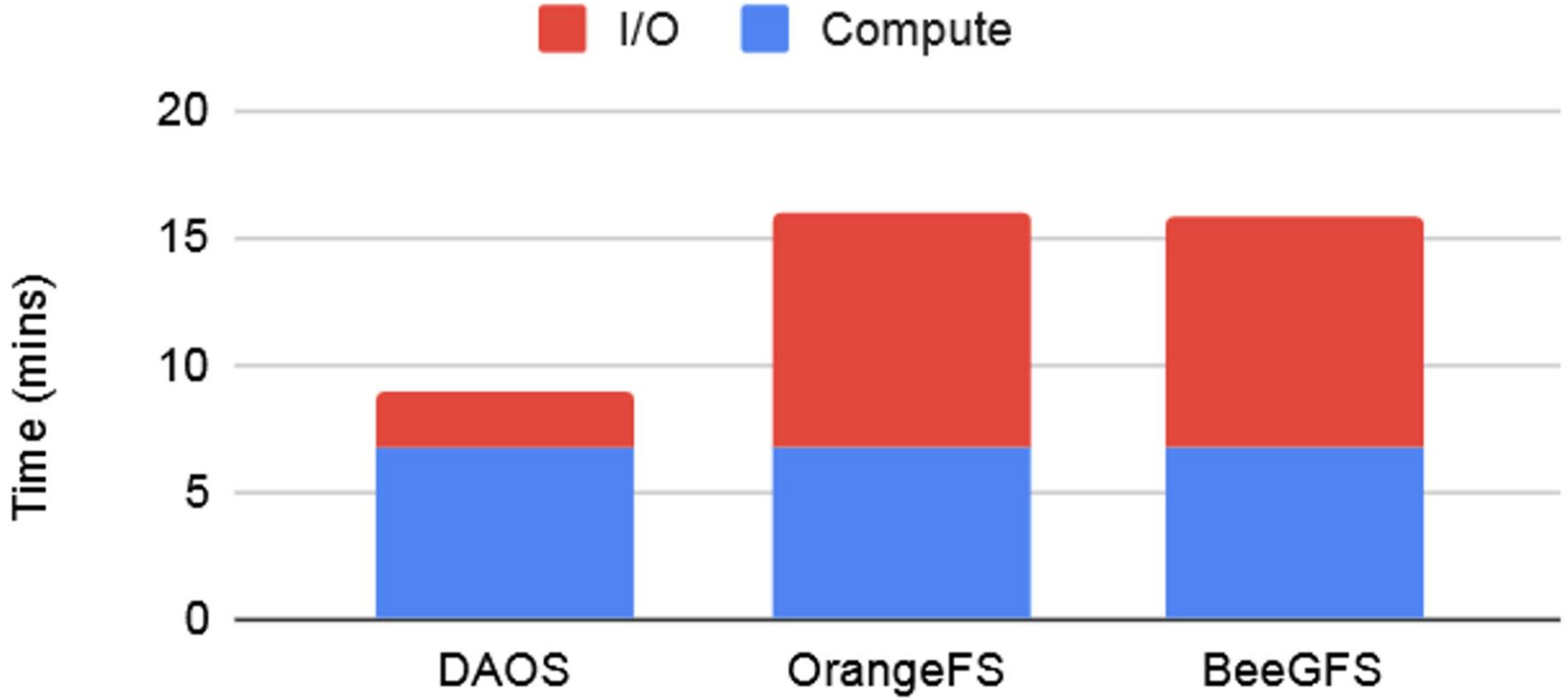
InfiniBand: индустриальный интерконнект, кредитная *lossless* передача, аппаратный RDMA; типовые топологии Clos/Fat-Tree, Dragonfly.

Omni-Path (OPA / OPX) — преемник Intel OPA от Cornelis Networks; собственный стек OPX, RDMA. Топологии Clos/Fat-Tree (в т.ч. large-radix), поддержка адаптивной маршрутизации.

Ethernet (RoCE): сопоставимая с IB задержка и полоса достижимы только при строго lossless-настройке

«Ангара»: интерконнект разработки АО «НИЦЭВТ», топология многомерный-тор продемонстрированы тесты BeeGFS/IO500 поверх прототипа TCP/IP-стека (EoA) (RSD 2022)

Сравнение результатов распределения времени при решении типичной задачи машинного обучения



(a) CosmicTagger for various storage systems

Luke Logan et al “An Evaluation of DAOS for Simulation and Deep Learning HPC Workloads” 2023

Модель настройки системы «ПФС-сеть»

$y = F(p; e, W) + \varepsilon$, где:

y – целевые метрики (BW, IOPS, p99 и т.д.)

e – параметры сети (Задержки, BW, топология)

p - параметры ПФС: страйпинг (size, count), размещение/пулы и т.д.

W – профиль нагрузки

ε – шум, вариативность

Цель: компромисс (Парето)

Заключение:

- Ввод-вывод данных – один из главных факторов, ограничивающих производительность HPC и ML.
- Обычные файловые системы не справляются с интенсивными I/O-нагрузками.
- ПФС позволяют повысить эффективность работы кластеров, но требуют сложной настройки.
- Бенчмарки (например, IO500) помогают объективно оценить производительность различных решений.
- Роль управления метаданными растет с масштабированием и является «узким местом»
- Начаты работы по развитию модели для оптимизации взаимосвязи системы «ПФС – сеть»