SCALING STORAGE FOR EXASCALE

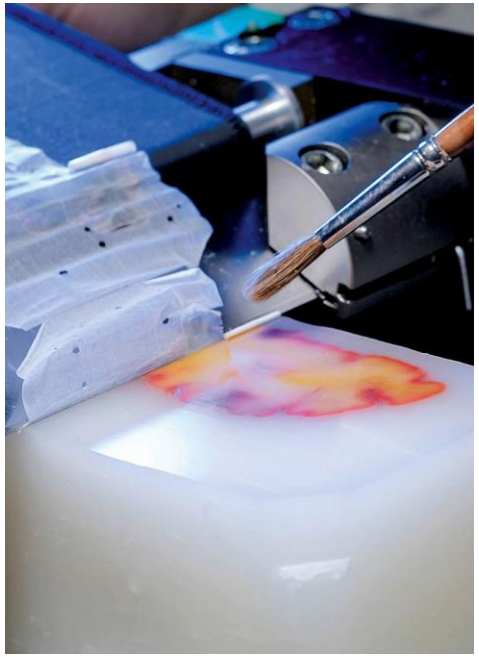
JUST6 & EXASTORE AT JUELICH

01. APRIL 2026 I STEPHAN GRAF, OLEG TSIGENOV, MARTIN LISCHWESKI & **SALEM EL SAYED**

FZJ AT FIRST GLANCE

Founded in 1956

Revenue:
€ 987 million in 2023



Research priorities:
information, energy, bioeconomy

Research campus with 14 institutes and 18 branch offices in Germany and abroad
7400 Employees
114 countries

Shareholders:
Federal Republic of Germany (90%), federal state of North Rhine-Westphalia (10%)

Member of the Helmholtz Association

JÜLICH SUPERCOMPUTING CENTRE

~300 Employees
(~280 FTE)

210 Scientists

22 PhD Students (+ externals)
27 Students (Bachelor/Master)

fz-juelich.de/jsc

also some JSC

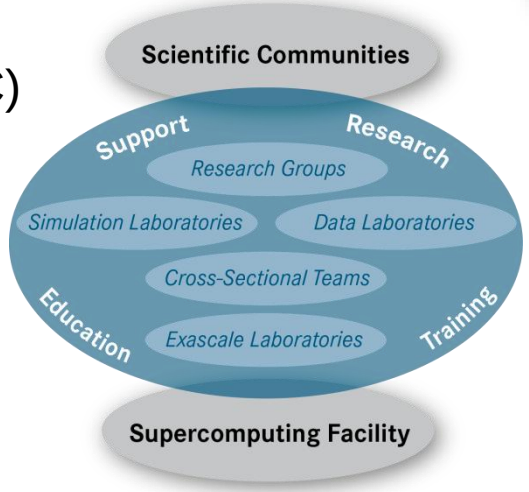
also some JSC

JSC

HPC, Cloud and Data Systems and Services Division at JSC

JÜLICH SUPERCOMPUTING CENTRE

- **Supercomputer operation for**
  - Centre – FZJ
  - Region – RWTH Aachen University
  - Germany – Gauss Centre for Supercomputing (GCS)
    John von Neumann Institute for Computing (NIC)
  - Europe – EuroHPC, PRACE, EU projects
- **Application support**
  - Unique support & research environment at JSC
  - Peer review support and coordination
- **R&D work**
  - Methods and algorithms, computational science, performance analysis and tools
  - Scientific Machine Learning and AI with HPC
  - Computer architectures, Co-Design, Exascale Labs together with IBM, Intel, NVIDIA
- **Education and training**



First Exascale supercomputer in Europe

DEEP

JÜLICH Forschungszentrum

JUPITER

Europa’s first ExaScale Supercomputer

JUPITER – HOSTING ENTITY DECISION 15.06.2022

tagesschau

Startseite ▸ Wirtschaft ▸ Technologie ▸ Hochleistungs-Rechner: Supercomputer "Jupiter" kommt nach Jülich

Hochleistungs-Rechner

Supercomputer "Jupiter" k

Stand: 15.06.2022 16:43 Uhr

**Das Forschungszentrum Jülich wird Standor
Exascale-Computers. "Jupiter" soll die Schal
Rechenoperationen in der Sekunde durchbr**

SPIEGEL Netzwelt

»Jupiter«

Jülich bekommt Europas ersten Exascale-Supercomputer

Das Forschungszentrum Jülich bekommt für eine halbe Milliarde Euro einen neuen Vorzeigerechner. Er soll helfen, Fragen zum Klimawandel zu beantworten – mit mehr als einer Trillion Rechenoperationen pro Sekunde.

15.06.2022, 16.52 Uhr

SIGN IN

The Register

{ HPC }

Germany to host Europe's first exascale supercomputer

Jupiter added to HPC solar system

Dan Robinson

Thu 16 Jun 2022 // 07:33 UTC

3

Germany will be the host of the first publicly known European exascale supercomputer, along with four other EU sites getting smaller but still powerful systems, the European High Performance Computing Joint Undertaking (EuroHPC JU) announced this week.

Germany will be the home of Jupiter, the "Joint Undertaking Pioneer for Innovative and Transformative Exascale Research." It should be switched on next year in a specially designed building on the campus of the Forschungszentrum Jülich research centre and operated by the Jülich Supercomputing Centre (JSC), alongside the existing Juwels and Jureca supercomputers.

PROCUREMENT BENCHMARK EFFORTS

A. Herten et al., "Application-Driven Exascale: The JUPITER Benchmark Suite," SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1-45, doi: 10.1109/SC41406.2024.00038.

Benchmark	Domain	Dwarfs
Amber*	MD	Particle, Spectral
Arbor	Neurosci.	Dense LA, Sparse LA, Unstructured Grid
Chroma-QCD	QCD	Sparse LA, Structured Grid, Monte Carlo
GROMACS	MD	Particle, Spectral
ICON	Climate	Sparse LA, Structured Grid
JUQCS	QC	Sparse LA
nekRS	CFD	Sparse LA, Spectral, Unstructured Grid
ParFlow*	Earth Sys.	Sparse LA, Structured Grid
PIConGPU	Plasma	Particle, Unstructured Grid
Quantum Espresso	Materials Sci.	Dense LA, Particle, Spectral
SOMA*	Polymer Sys.	Particle, Monte Carlo
MMoCLIP	AI (MM)	Dense LA
Megatron-LM	AI (LLM)	Dense LA
ResNet*	AI (Vision)	Dense LA
DynQCD	QCD	Sparse LA, Structured Grid, Monte Carlo
NAStJA	Biology	Structured Grid, Monte Carlo
Graph500	Graph	*Graph Traversal (D. 9)*
HPCG	CG	Sparse LA
HPL	LA	Dense LA
IOR	Filesys.	*Input/Output*
LinkTest	Network	*P2P, Topology*
OSU	Network	*Message Exchange, DMA*
STREAM	Memory	*Regular Access*

JUPITER CONTRACT ANNOUNCEMENT 3.10.2023

HPC wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast
- Events
- Solution Channels
- Job Bank
- About
- Subscribe

EU Grabs ARM for First ExaFLOP Supercomputer, x86 Misses Out

By Agam Shah

October 4, 2023

The configuration of Europe's first exascale supercomputer, Jupiter, has been finalized, and it is a win for Nvidia and a disappointment for x86 chip vendors Intel and AMD. The Jupiter supercomputer, which will cost €273 million to build, will pair SiPearl's Rhea processor, which is based on ARM architecture, with accelerator technology from Nvidia.

The supercomputer is being built by the European High-Performance Computing Joint Undertaking (EuroHPC JU) and a consortium including Eviden and ParTec. Eviden is an Atos business focusing on advanced computing initiatives tha

The Jülich Supercompu
(*Correction: Jülich Supe
about 600 km or 375 m

Specifically, the superc
CPUs, and the initial co
supercomputers in the
on ARM.

That is a big disappoint
invest €33 billion to buil
development initiatives
leaders in a bid to get m

Jülich's fastest system,
November 2021 and is
Europe are the third-ran
performance of 309 pet

Off The Wire Industry Headlines

October 13, 2023

- Coherent File Format Accelerates Time-to-Solution with OpenFOAM
- HealthyCloud Project Unveils Roadmap to Maximize Impact of Health Data and Research Across Europe
- NCSA Welcomes 2023-24 Fellows
- Berkeley Lab CS Area to Share Computing Expertise at SC23

October 12, 2023

- Samsung Electronics to Host AI Forum 2023 Highlighting AI and Computer Engineering Innovation
- PacBio Announces Complete Computational Workflow for Human Whole Genome Sequencing Data Analysis
- SiFive Announces Differentiated Solutions for Generative AI and ML Applications
- EQTC 2023: Europe's Quantum Sector to Showcase Successes and Its Roadmap for Global Leadership
- EuroHPC JU Announces Procurement Call for Upgrading Discoverer Supercomputer
- Los Alamos Partners with AirMettle for Efficient In-Storage Data Analysis
- Caltech Researchers Demonstrate Quantum Eraser to Combat Erasure Errors in Quantum Systems
- Research Results, Computational Expertise Flourish at

SIGN IN / UP

The Register

HPC

Atos subsidiary Eviden scores contract win in Europe's first exascale system

$526M Jupiter set to rule EU's tech orbit by 2024

Dan Robinson

Wed 4 Oct 2023 | 16:45 UTC

The EU's supercomputing initiative, the European High Performance Computing Joint Undertaking (EuroHPC JU), has awarded a procurement contract for Europe's first exascale system, with installation due to start in early 2024.

Known as Jupiter (Joint Undertaking Pioneer for Innovative and Transformative Exascale Research), the system was announced last year followed by a call for tender in January of this year.

EuroHPC JU said the procurement contract for Jupiter was awarded to a consortium comprising of Eviden, the professional services side of French IT giant Atos, and ParTec, a German supercomputing hardware company.

The project is expected to have a total cost of €273 million ($287 million) covering the build, delivery, installation, and maintenance of Jupiter, according to the EuroHPC.

However, Eviden put the overall project cost at €500 million ($526 million), saying that this is the figure for the entire project, including the system manufacturing and its

THENEXTPLATFORM

HOME | COMPUTE | STORE | CONNECT | CONTROL | CODE | AI | HPC | ENTERPRISE | HYPERSCALE | CLOUD

LATEST > Intel To Set Its FPGA Unit Free To Pursue Its Own Path • COMPUTE

Search ...

HOME > HPC > Details Emerge On Europe's First Exascale Supercomputer

DETAILS EMERGE ON EUROPE'S FIRST EXASCALE SUPERCOMPUTER

October 5, 2023 Timothy Prickett Morgan

me details are emerging on Europe's first exascale system, codenamed "Jupiter" and to be installed at the ch Supercomputing Center in Germany in 2024. There has been a lot of speculation about what Jupiter l include for its compute engines and networking and who will build and maintain the system. We now ow some of this and can infer some more from the statements that were made by the organizations rticipating in the Jupiter effort.

June 2022, the Forschungszentrum Jülich in Germany, which has played host to many supercomputers ce it was founded in 1987, was chosen to host the first of three European exascale-class percomputers to be funded through the EuroHPC Joint Undertaking and through the European national d state governments countries who are essentially paying to make sure these HPC and AI clusters are ere they want them. With Germany having the largest economy in Europe and being a heavy user of C thanks to its manufacturing focus, Jülich was the obvious place to park the first machine in Europe to ak the exaflops barrier.

at barrier is as much an economic one as it is a technical one. The six-year budget for Jupiter weighs in at 00 million, which is around $526.1 million at current exchange rates between the US dollar and the ropean euro. That is in the same ballpark price as what the "Frontier" exascale machine at Oak Ridge tional Laboratory and the "El Capitan" machine that is being installed right now at Lawrence Livermore tional Laboratory – both of which are based on a combination of AMD CPUs and GPUs and Hewlett ckard Enterprise's Slingshot variant of Ethernet with HPE as the prime contractor.

erybody knows that Jupiter was going to use SiPearl's first generation Arm processor based on the verse "Zeus" V1 core from Arm Ltd, which is codenamed "Rhea" by SiPearl and which is appropriate

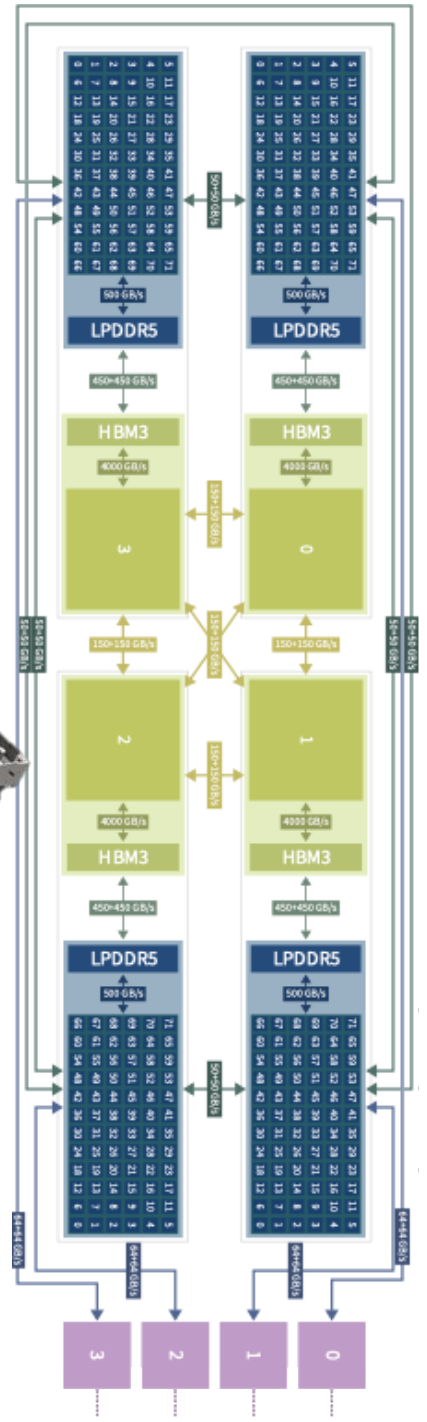
JUPITER – THE BOOSTER

HIGHLY-SCALABLE MODULE FOR HPC AND AI WORKLOADS

EVIDEN an atos business

NVIDIA

- **1** EXAFLOP/S (FP64, HPL)
- NVIDIA GRACE-HOPPER CG1
  - ~5900 compute nodes
  - **4×** CG1 chips per compute node
- NVIDIA MELLANOX NDR
  - **4×** NDR200 NICs per compute node
- BULLSEQUANA XH3000
  - Direct Liquid Cooled blades
  - **2× compute node per blade**

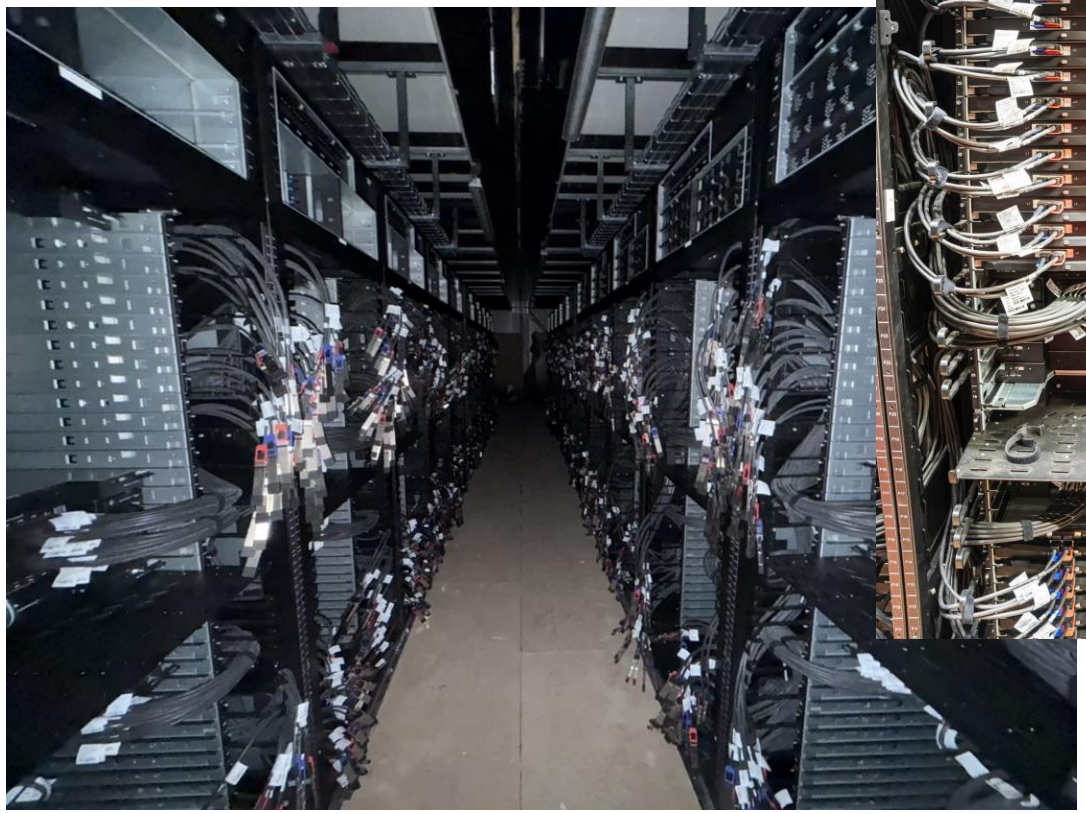


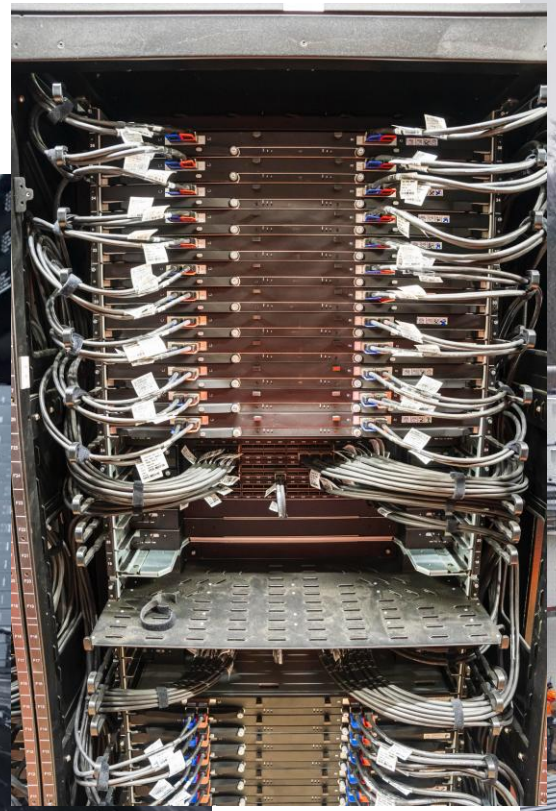
August 2024


Backup Cold Water Cooling: 1MW
Power (Campus): 2*60-80 MVA
Water (River Rur): up to 30 cbm/h
Network:
3,2 Tbit/s

JUPITER ASCENDING

SINCE JANUARY 2025






JUWELS Booster:
• 3744 NVIDIA A100 GPUs
• 44 PetaFLOP/s HPL
• 2,4 MW Peak, 1,1MW Average
• Direct-Liquid-Cooling
• 36°in, 42-44°out
• free cooling
• heat-reuse
Power (Datacenter): 15*2,5MVA
14 Free Cooling Towers, 1 Chiller
JUPITER Booster:
• ~24000 NVIDIA H100 GPUs
• 1 ExaFLOP/s HPL
• 17 MW Peak, 9-11MW Average
• Direct-Liquid-Cooling
• 36°in, 42-44°out
• free cooling
• heat-reuse
• Up to 30 cbm/h river water
• !!!Worst-case!!!
• Only on very-hot summer days
April 2025


September 2025

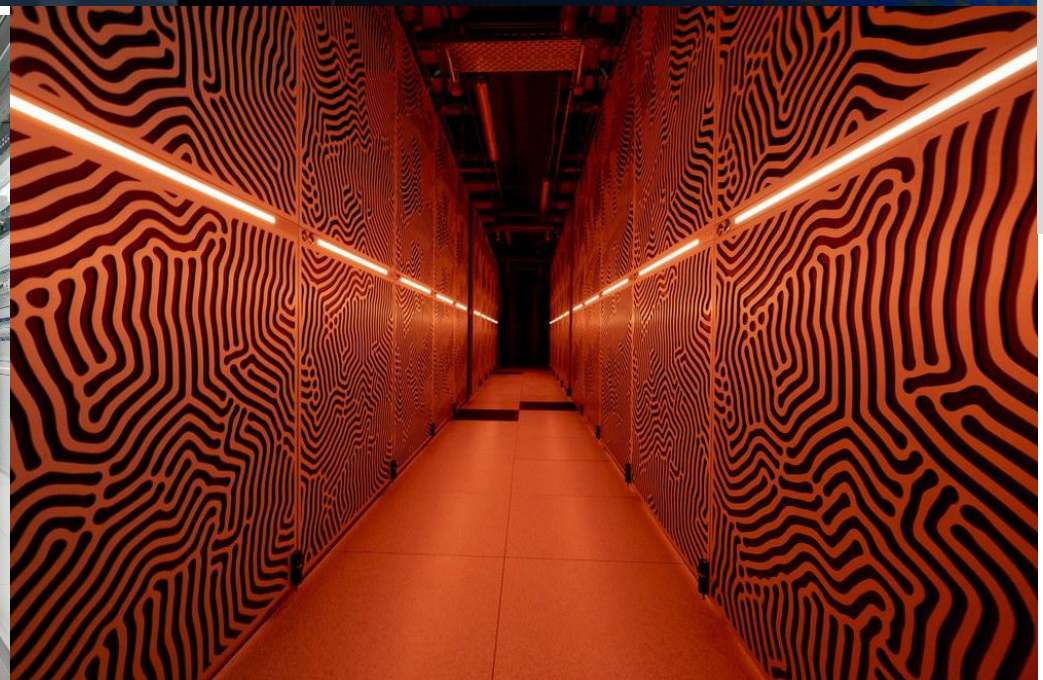
September 05, 2025: Inauguration of JUPITER

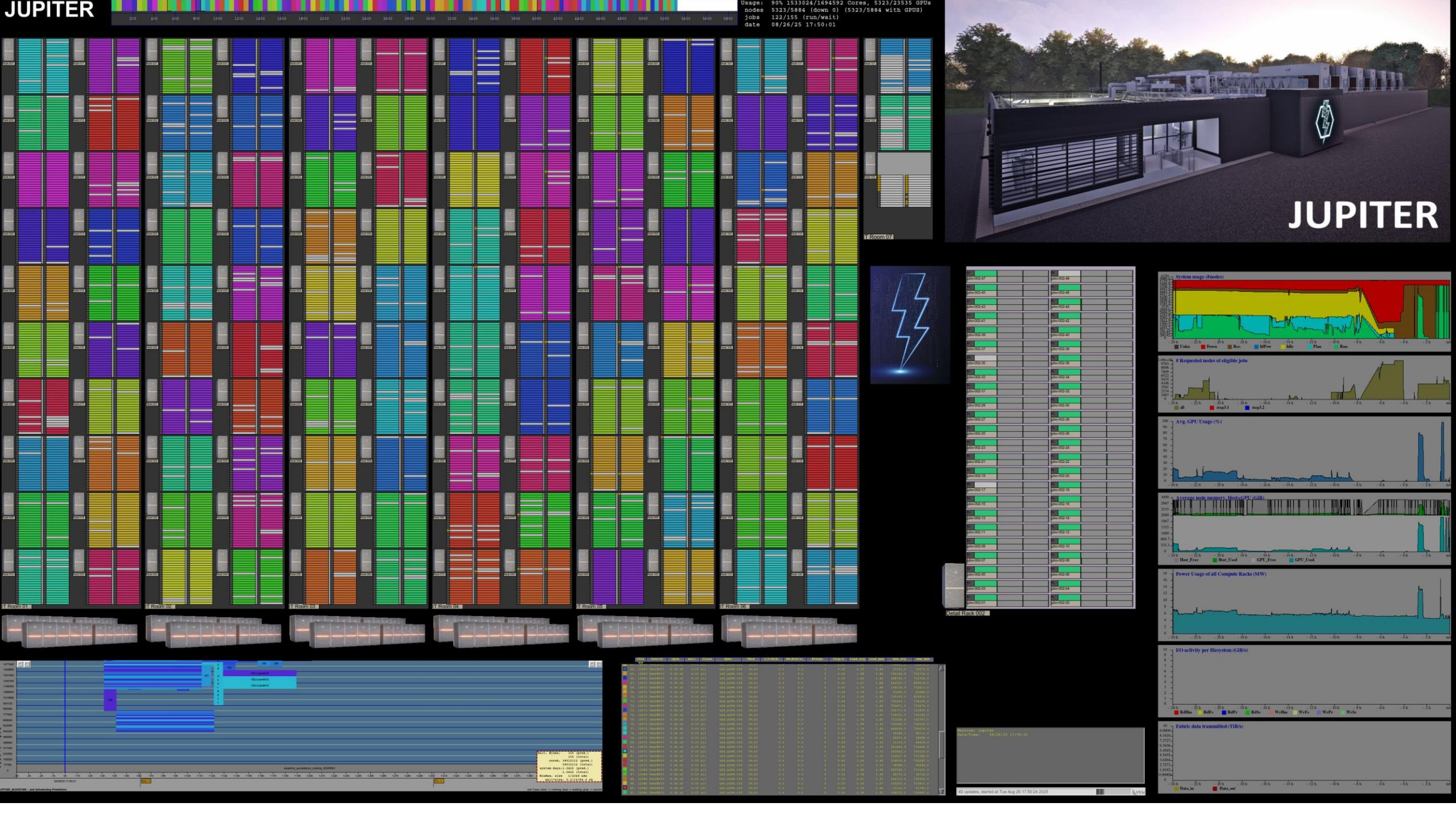
SUPERCOMPUTER

Launch of the Exascale Supercomputer JUPITER

JUPITER was unveiled to the public today at Forschungszentrum Jülich in front of high-ranking guests from politics, science, and industry. The German Chancellor Friedrich Merz and the Minister-President of North Rhine-Westphalia Hendrik Wüst attended the inauguration ceremony of Europe's fastest and most energy-efficient supercomputer.

Inauguration of JUPITER: Ceremonial act with Prof. Dr. Dr. Thomas Lippert, Director of Jülich Supercomputing Centre (JSC), Prof. Dr. Kristel Michielsen, Director of Jülich Supercomputing Centre (JSC), Ekaterina Zaharieva, Commissioner Startups, Research and Innovation, European Commission, Hendrik Wüst, Minister-President of North Rhine-Westphalia, Prof. Dr. Astrid Lambrecht, Chair of the Board of Directors at FZJ, Friedrich Merz, Federal Chancellor, Dorothee Bär, Federal Minister for Research, Technology and Space, Karsten Wildberger, Federal Ministry for Digital Transformation and Government Modernisation, Ina Brandes, Minister for Culture and Science of North Rhine-Westphalia,Prof. Dr. Laurens Kuipers, Member of the Board of Directors at FZJ. Fotos: Forschungszentrum Jülich / Kurt Steinhausen.




JUPITER
Usage: 90% 1533024/1694592 Cores, 5323/23535 GPUs
nodes 5323/5884 (down 0) (5323/5884 with GPUS)
jobs 122/155 (run/wait)
date 08/26/25 17:50:01
T Room 07
JUPITER
T Room 01
T Room 02
T Room 03
T Room 04
T Room 05
T Room 06
Detail Rack 002
System usage (#nodes)
Requested nodes of eligible jobs
Avg. GPU Usage (%)
Average node memory Host+GPU (GiB)
Power Usage of all Compute Racks (MW)
I/O activity per filesystem (GiB/s)
Fabric data transmitted (TiB/s)


The GREEN 500
CERTIFICATE
JUPITER Booster - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchi
Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux
EuroHPC/FZJ, Germany
is ranked
No. 14
among the World's TOP500 Supercomputers
with 63.316 GFlops/watts Performance
in the Green500 List published at the SC2025
Conference on November 18, 2025.
Congratulations from the Green
Wu-chun Feng
Virginia Tech
editors
Kirk Cameron
Virginia Tech
Erich Strohmaier
NERSC/Berkeley Lab
TOP 500
The List.
CERTIFICATE
JUPITER Booster - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip,
Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux
EuroHPC/FZJ, Germany
is ranked
No. 4
among the World's TOP500 Supercomputers
with 1000.00 PFlop/s Linpack Performance
in the 66th TOP500 List published at the SC2025
Conference on November 18, 2025.
Congratulations from the TOP500 Editors
Jack Dongarra
University of Tennessee
Horst Simon
NERSC/Berkeley Lab
Martin Meuer
Prometeus
JÜLICH
Forschungszentrum

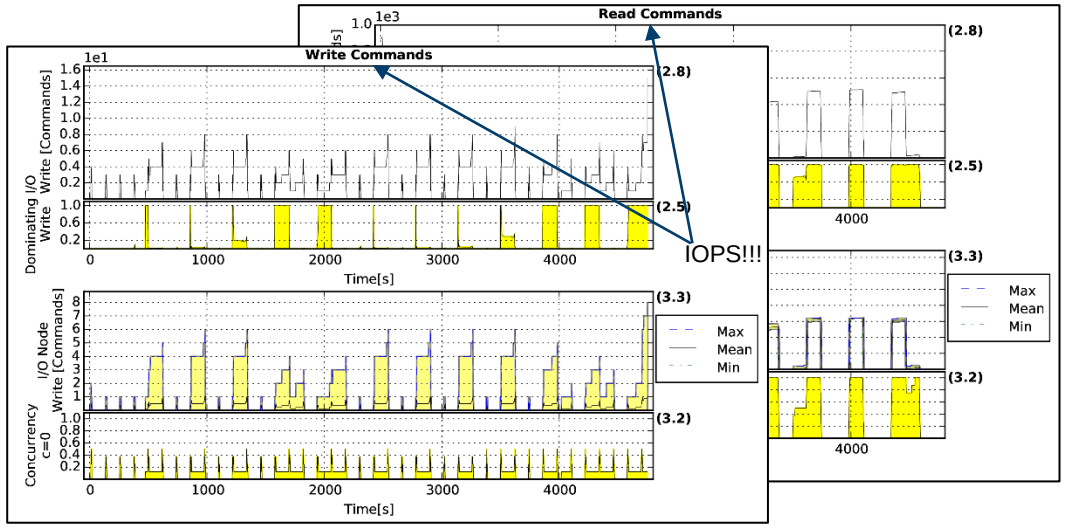
CHOOSING A STORAGE SYSTEM

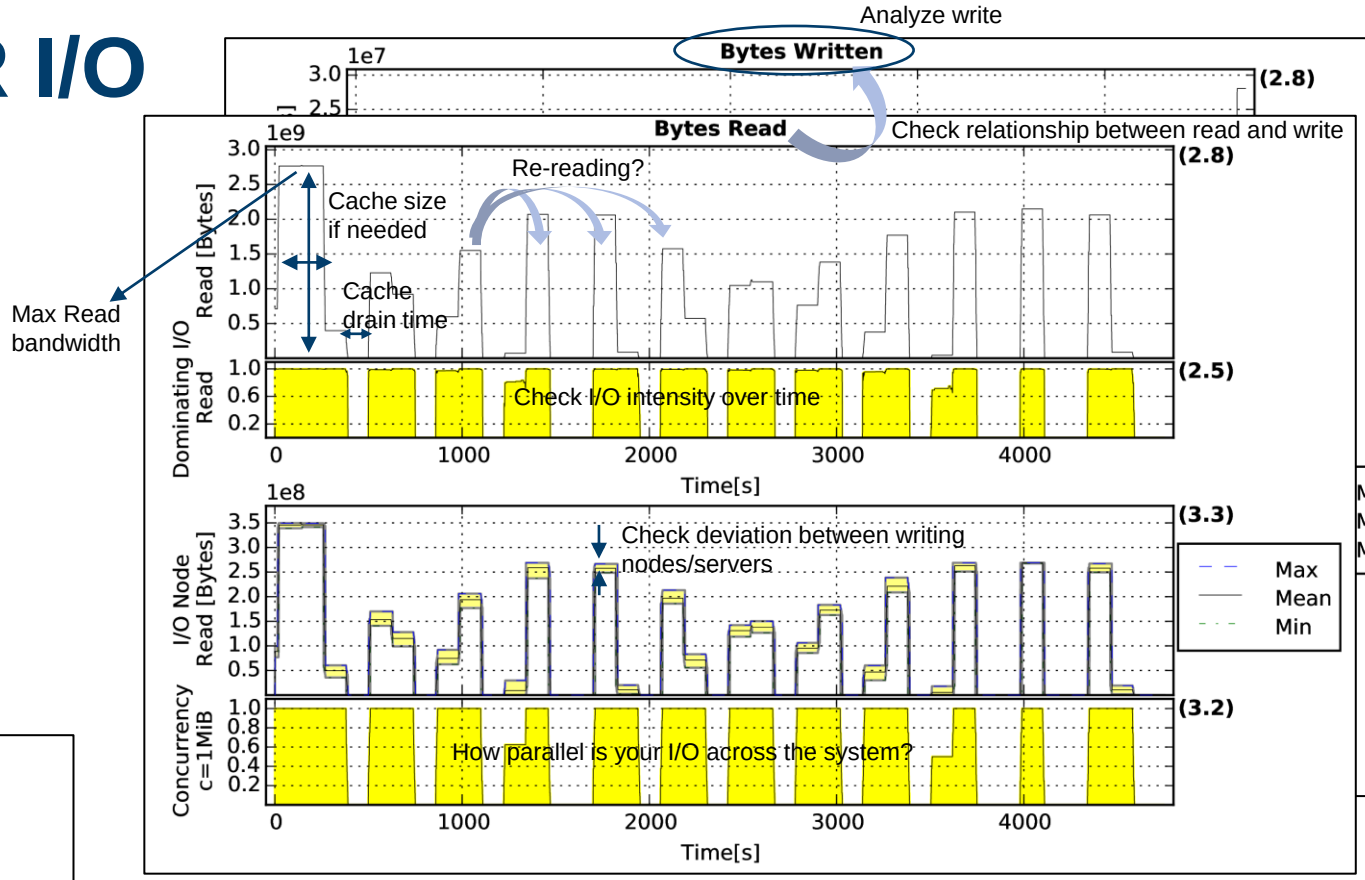
MEASURE & MATCH YOUR I/O

Without DATA it is just an opinion!

Hints:

- Do not over fit and add margins
- This was measured on your old system
- Make room for future expansion





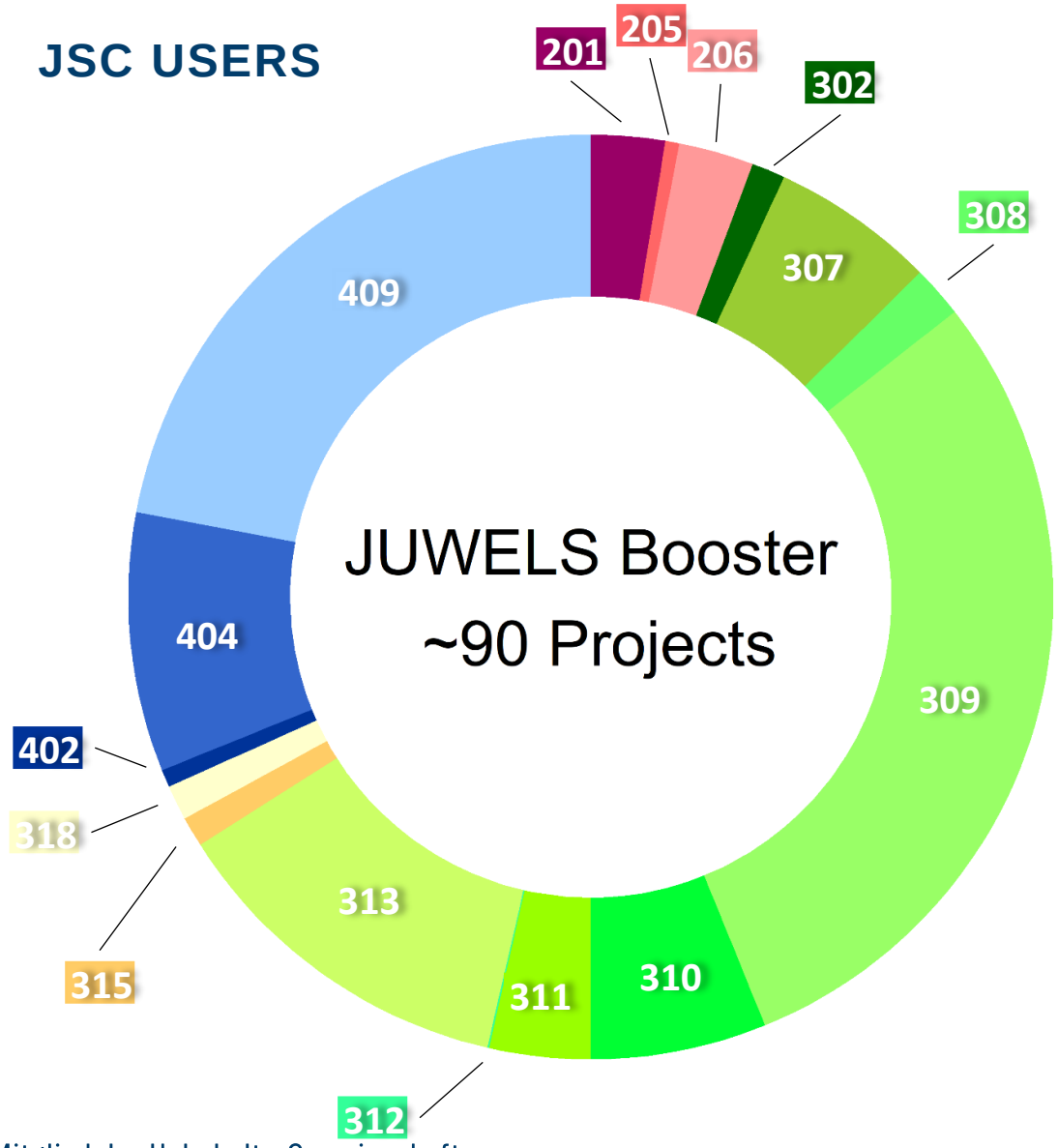
Besides performance:

- Capacity
- I/O interface
- Legacy infrastructure
- Retention time
- …

Images from:
S. El Sayed Mohamed, Analysis of I/O requirements of scientific applications. Jülich: Forschungszentrum Jülich GmbH, Zentralbibliothek, Verlag, [2018] [Online]. Available: https://nbn-resolving.org/urn:nbn:de:hbz:5:2-166546

WE DON'T HAVE THAT LUXURY

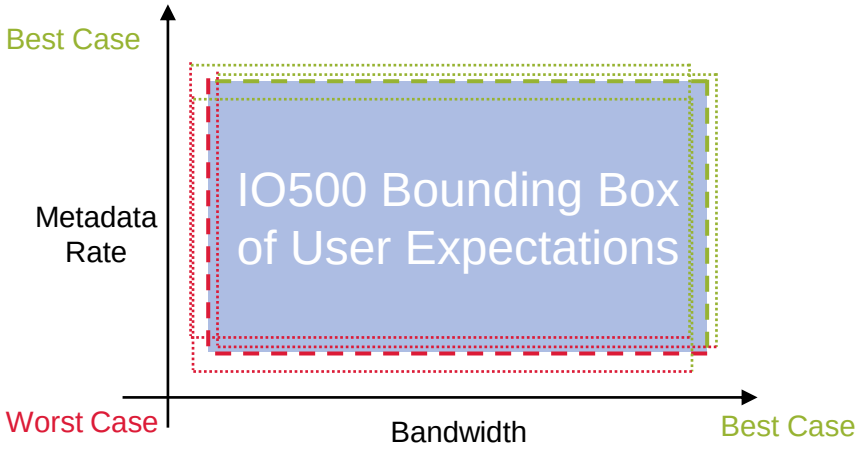
JSC USERS



Reseach Fields

- 201 Basic Biological and Medical Research
- 205 Medicine
- 206 Neurosciences
- 302 Chemical Solid State and Surface Research
- 307 Condensed Matter Physics
- 308 Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas
- 309 Particles, Nuclei and Fields
- 310 Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics
- 311 Astrophysics and Astronomy
- 312 Mathematics
- 313 Atmospheric Science, Oceanography and Climate Research
- 315 Geophysics and Geodesy
- 318 Water Research
- 402 Mechanics and Constructive Mechanical Engineering
- 404 Heat Energy Technology, Thermal Machines, Fluid Mechanics
- 409 Computer Science

BOUNDING BOX OF USER EXPECTATIONS



IO500

(HTTPS://IO500.ORG/ABOUT)

- **IOEasy**: Continues read/write of file per process + 4k random read
- **IOHard**: Interleaved read/write of 47008-byte blocks
- **MDEasy**: Write/stat/delete zero-byte files in a unique directory per process
- **MDHard**: Write/stat/read/delete small files (3901 bytes) in a shared directory
- Find: Finding relevant objects based on patterns

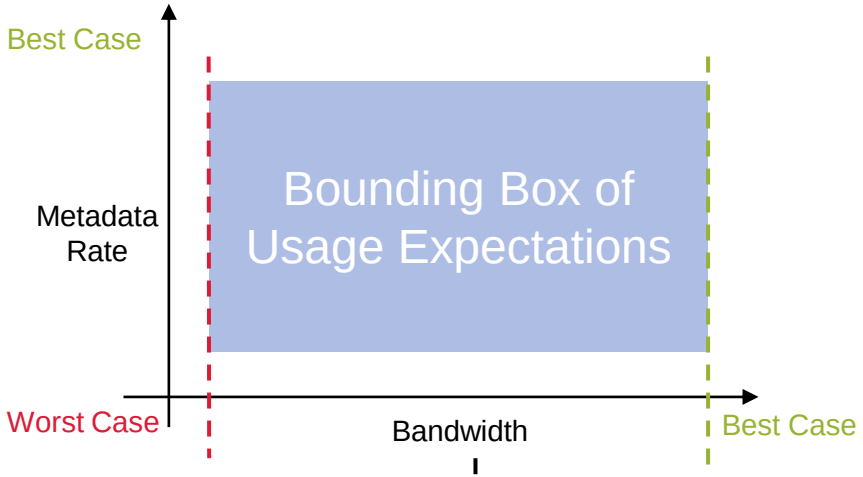
R. Liem et al. "User-Centric System Fault Identification Using IO500 Benchmark," 2021 IEEE/ACM Sixth International Parallel Data Systems Workshop (PDSW), St. Louis, MO, USA, 2021, pp. 35-40, doi: 10.1109/PDSW54622.2021.00011.

OUR BOUNDING BOX OF EXPECTATIONS

Free to choose:
- Number of nodes
- Number of processes per node

Restrictions:
- Run at least 1min
- Avoid caching

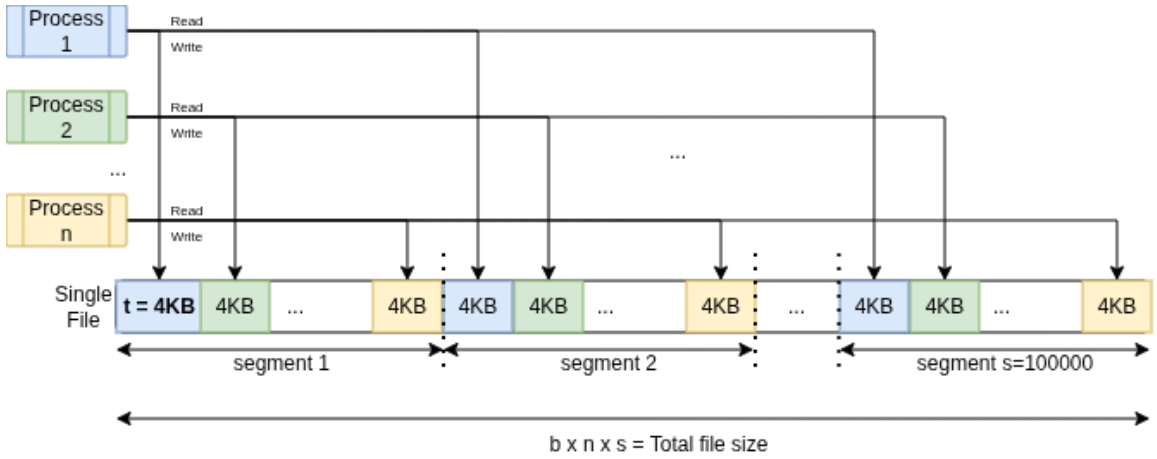
File system block size:
- 16MB
- 512KB



Common IOR Parameters	
-C	reorderTasksConstant
-Q	taskPerNodeOffset (choose in relation to --distribution from slurm)
-g	intraTestBarriers
-a	API
-i	Repititions

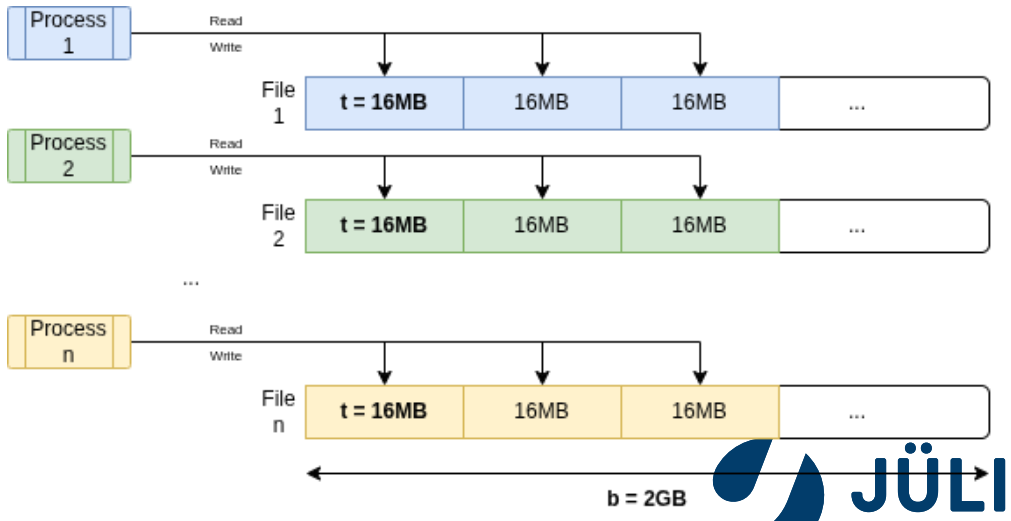
IOR Hard

ior -C -Q 1 -g ***-t 4k -b 4k –s 100000*** -a POSIX -i 3



IOR Easy

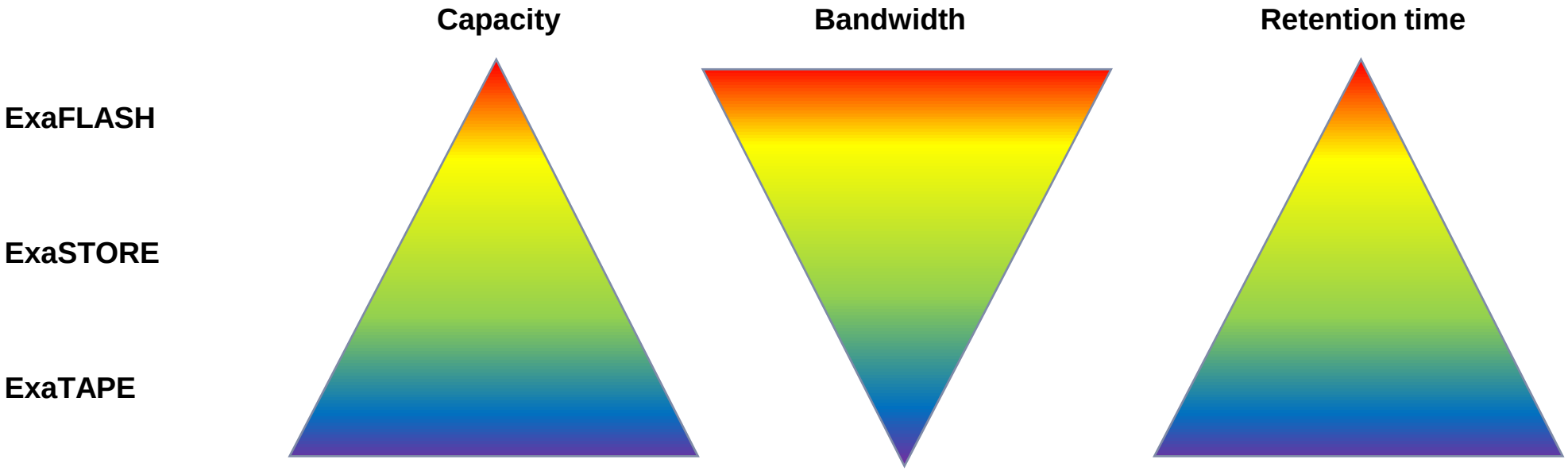
ior -C -Q 1 -g ***-t 16m -b 2g (-s 1)*** -a POSIX -i 3



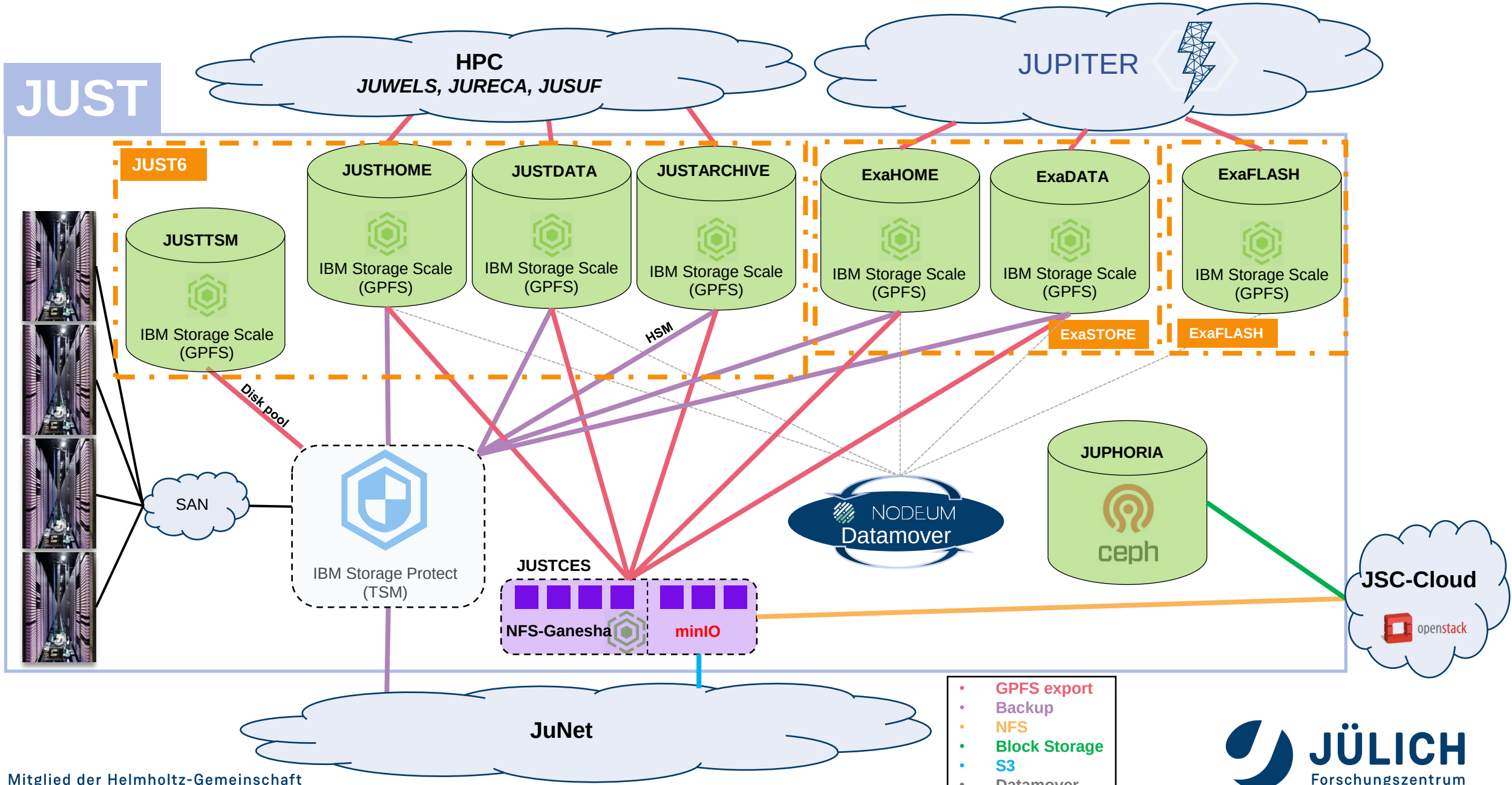
JUST

JUelich STorage

TIERED STORAGE OFFERING FOR JUPITER

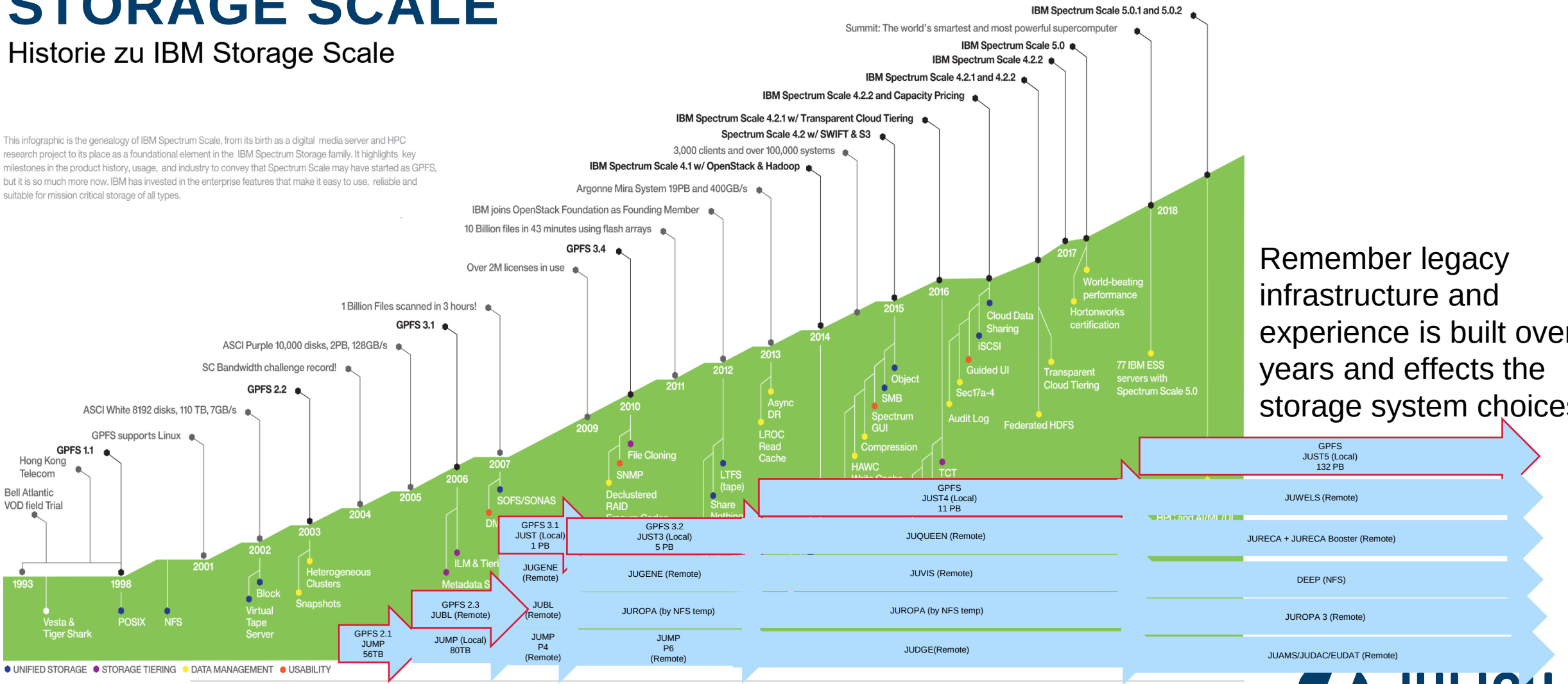


- ExaFLASH: IBM SSS6K cluster (29 PB raw capacity on NVMe)
- ExaSTORE: IBM SSS6K + ESS3500 cluster (308 PB raw capacity on 22 TB HDD)
- ExaTAPE: Tape storage (Backup + GPFS&TSM-HSM as part of JUST6, 2x TS4500, ~370 PB)


JUST
HPC
JUWELS, JURECA, JUSUF
JUPITER
JUST6
JUSTHOME
JUSTDATA
JUSTARCHIVE
ExaHOME
ExaDATA
ExaFLASH
JUSTTSM
IBM Storage Scale (GPFS)
HSM
ExaSTORE
ExaFLASH
Disk pool
SAN
IBM Storage Protect (TSM)
JUSTCES
NFS-Ganesha
minIO
NODEUM Datamover
JUPHORIA
ceph
JSC-Cloud
openstack
JuNet
GPFS export
Backup
NFS
Block Storage
S3
Datamover

HISTORY OF ~~GPFS~~ / ~~SPECTRUM SCALE~~ / STORAGE SCALE

Historie zu IBM Storage Scale



Remember legacy infrastructure and experience is built over years and effects the storage system choices

Images from:
Christof Schirra, Storage Technical Specialist, cschirra@de.ibm.com. Title "IBM Storage Scale: Konzepte", presentation, Storage Strategy Days 2026.

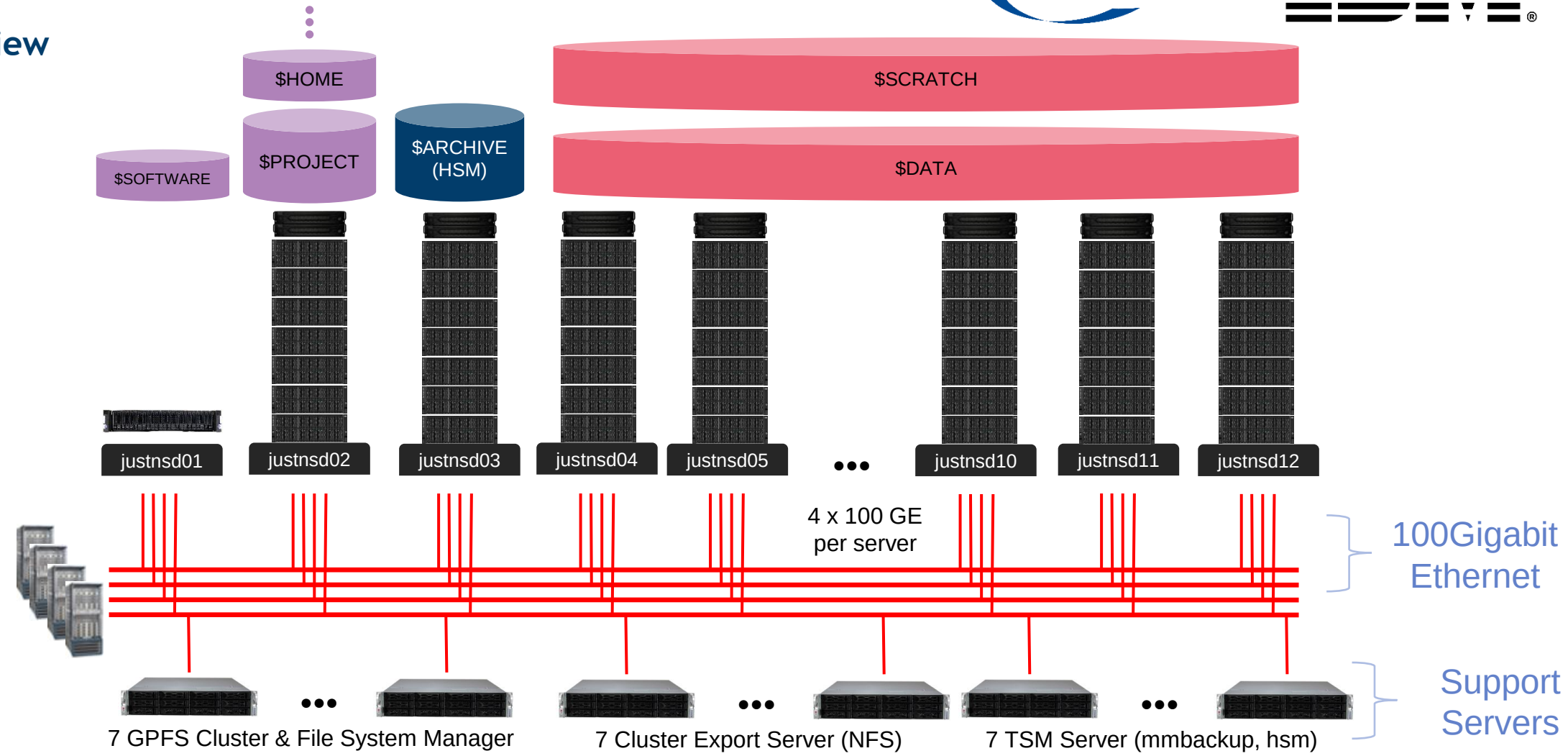
STORAGE SCALE 6000

Storage Scale System 6000

- 2 x AMD EPYC Genoa 9454 CPU 48c, 3,8GHz
- 1024GB DDR5 memory in 16 x 64
- 2 x CX-7 NDR200 Dual Port
- 4 x SAS4 x8 G4 Adapter
- 7 enclosures (91 x 22TB HDD drives) and a raw capacity of 14 PB

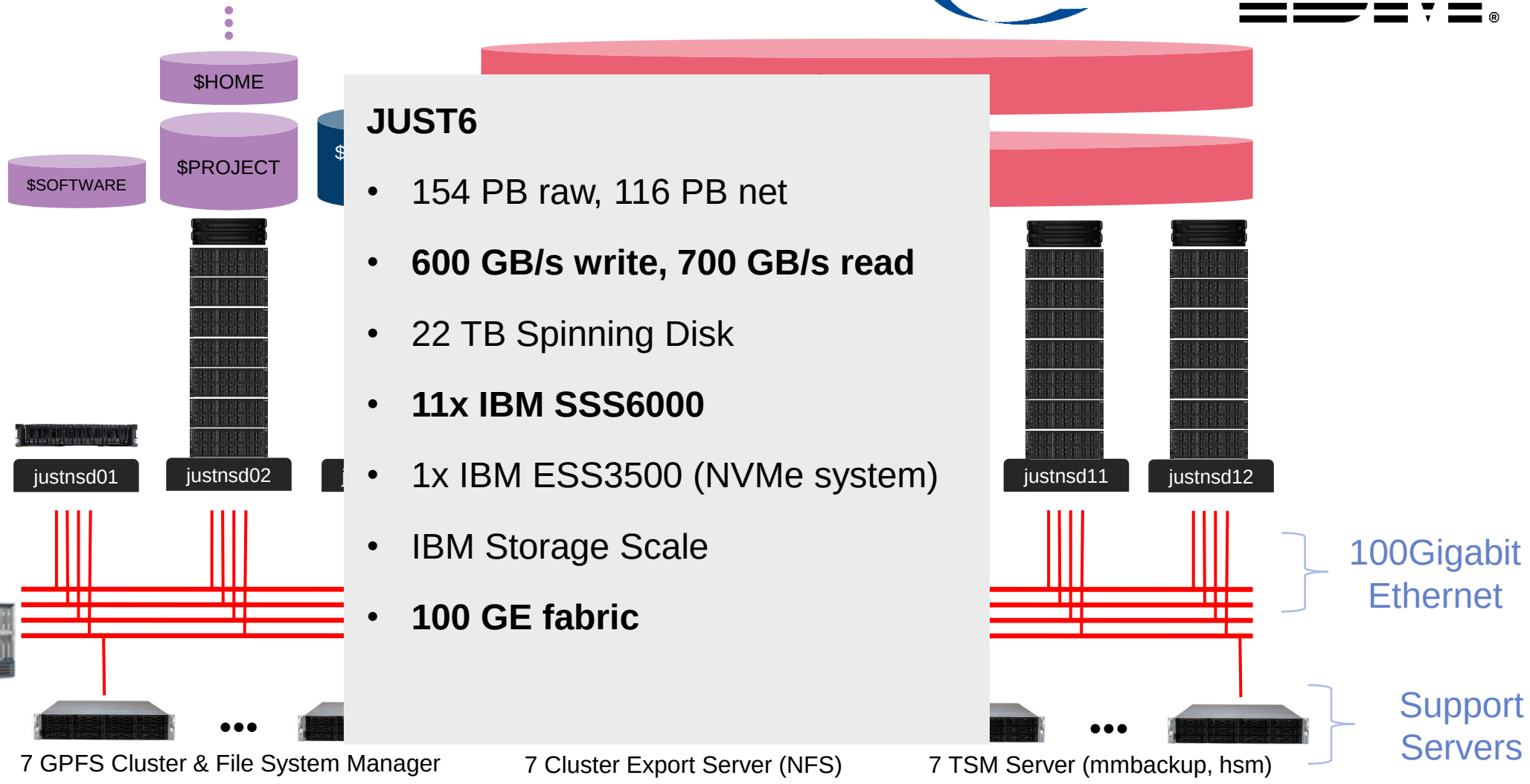
JUST6

Physical view



JUST6

Physical view

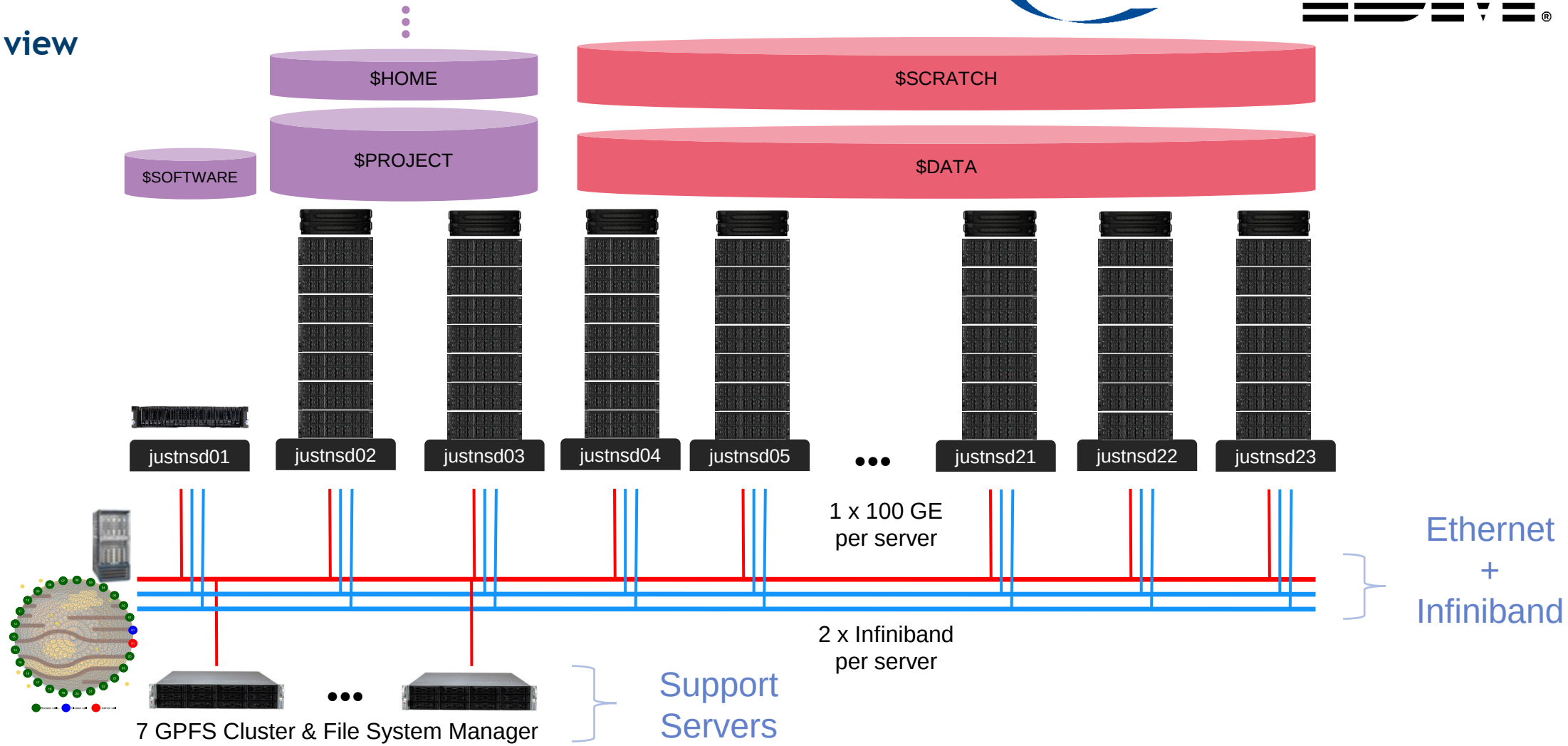


JUST6

- 154 PB raw, 116 PB net
- **600 GB/s write, 700 GB/s read**
- 22 TB Spinning Disk
- **11x IBM SSS6000**
- 1x IBM ESS3500 (NVMe system)
- IBM Storage Scale
- **100 GE fabric**

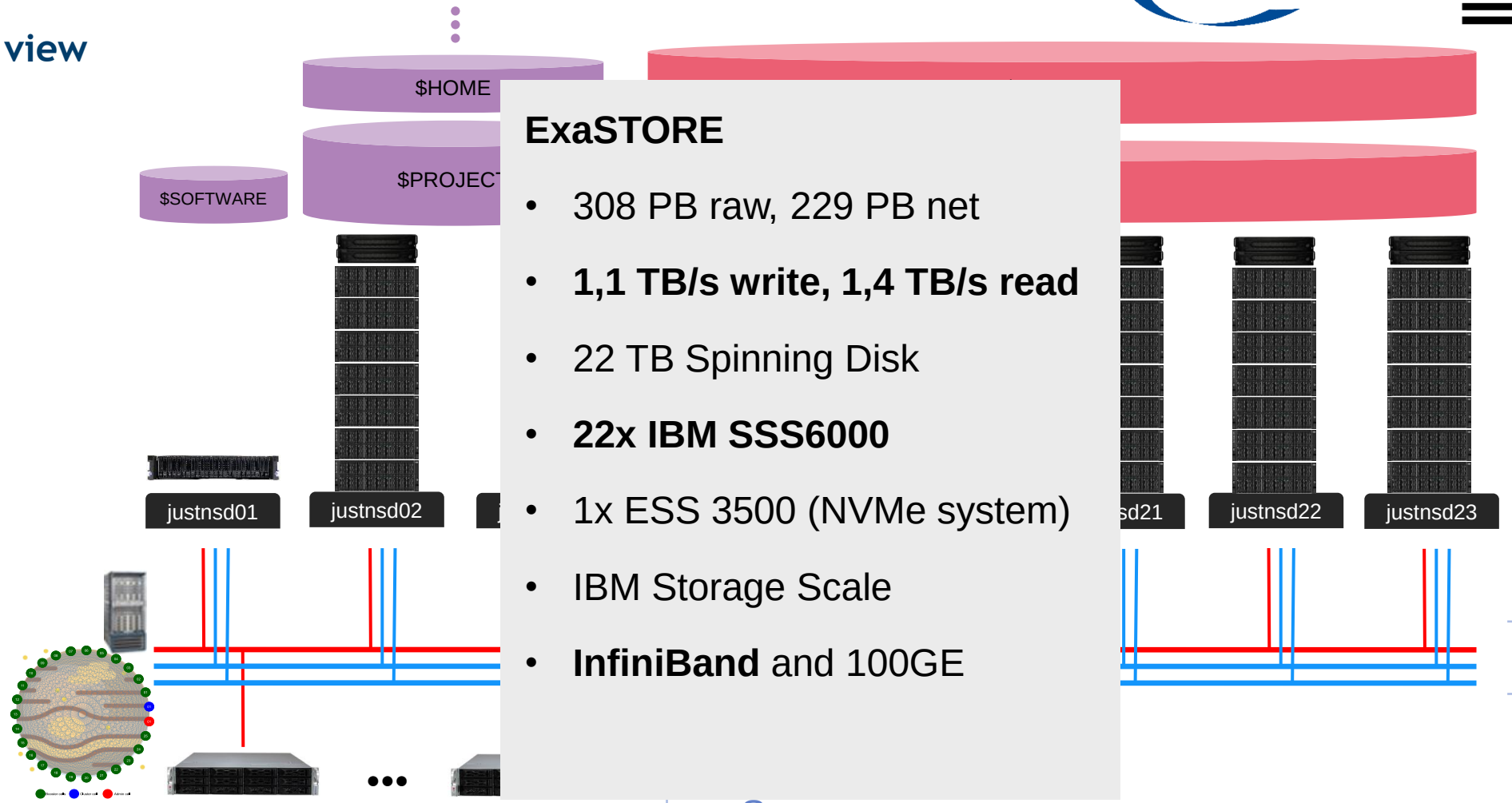
EXASTORE

Physical view



EXASTORE

Physical view



ExaSTORE

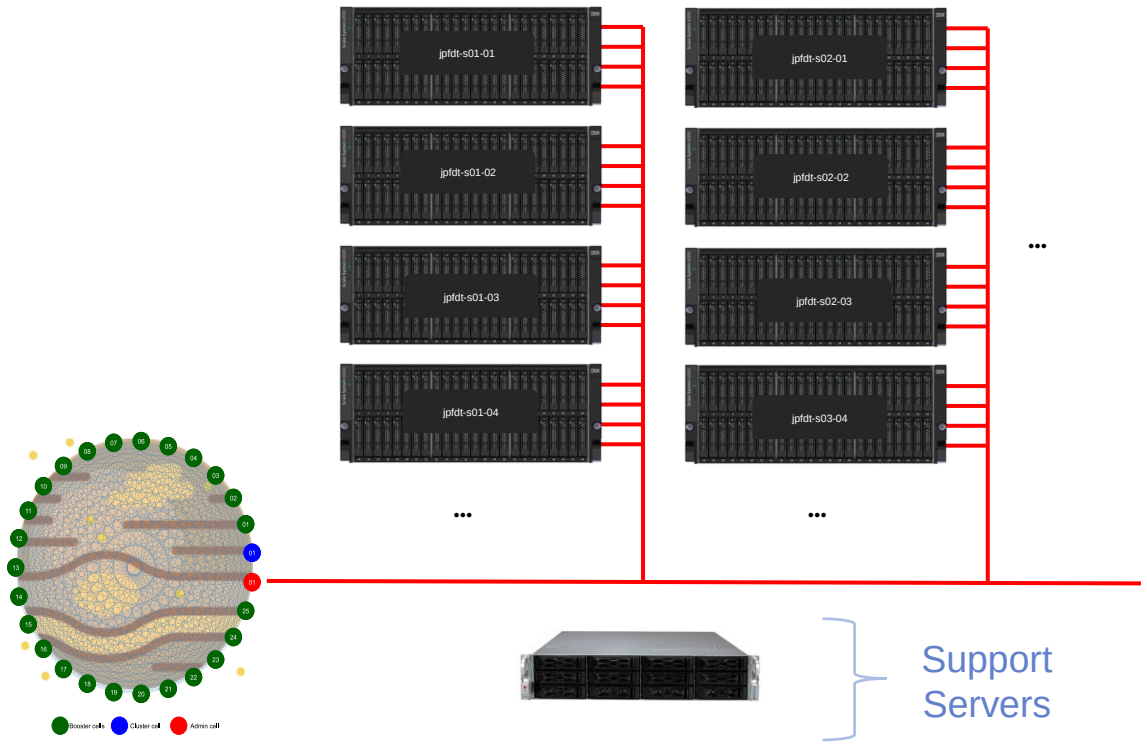
- 308 PB raw, 229 PB net
- **1,1 TB/s write, 1,4 TB/s read**
- 22 TB Spinning Disk
- **22x IBM SSS6000**
- 1x ESS 3500 (NVMe system)
- IBM Storage Scale
- **InfiniBand** and 100GE

EXAFLASH

Physical View

ExaFLASH

- 29 PB raw, 21PB net
- **2,1 TB/s write, 3,1 TB/s read**
- 30 TB NVME
- **20x IBM SSS6000**
- IBM Storage Scale
- **InfiniBand**



ACCEPTANCE !

ACCEPTANCE TEST LIST

- ❑IOR Easy
  - ❑Each Building Block
    - ❑IOR Easy 16MB blocksize file system
    - ❑IOR Easy 512KB blocksize file system
  - ❑All Building Blocks
    - ❑IOR Easy 16MB blocksize file system
    - ❑IOR Easy 512KB blocksize file system
- ❑IOR Hard
  - ❑Each Building Block
    - ❑IOR Hard 16MB blocksize file system
    - ❑IOR Hard 512KB blocksize file system
  - ❑All Building Blocks
    - ❑IOR Hard 16MB blocksize file system
    - ❑IOR Hard 512KB blocksize file system
- ❑MDTest
  - ❑Each Building Block
    - ❑MDTest 16MB block size file system
    - ❑MDTest 512KB block size file system
  - ❑All Building Block
    - ❑MDTest 16MB block size file system
    - ❑MDTest 512KB block size file system

DISCLAIMERS

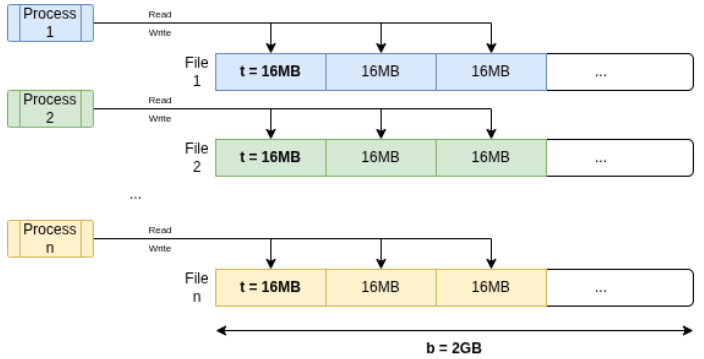
- **This is a summary:** Not all acceptance events can/should be listed here. The following points are a subset of the hundreds of tests and benchmarks. The examples have been extracted from hundreds of emails with long subject lines containing many `Re:` and `Aw:`. Additionally, any meeting minutes and acceptance documents available have been used.
- **Based on real events**: However some events have been dramatized for effect, others we wish were not true. The persons involved are real and their pain has been scientifically proven to exist.
- **Depends on remembering**: This one too was in a galaxy far far away! At least it feels that way. Events are partially pulled out of our Random Access Memory. Expect events not to be in the correct order or missing curtail details. Ask us or better yet, one of our trusted IBM colleagues.

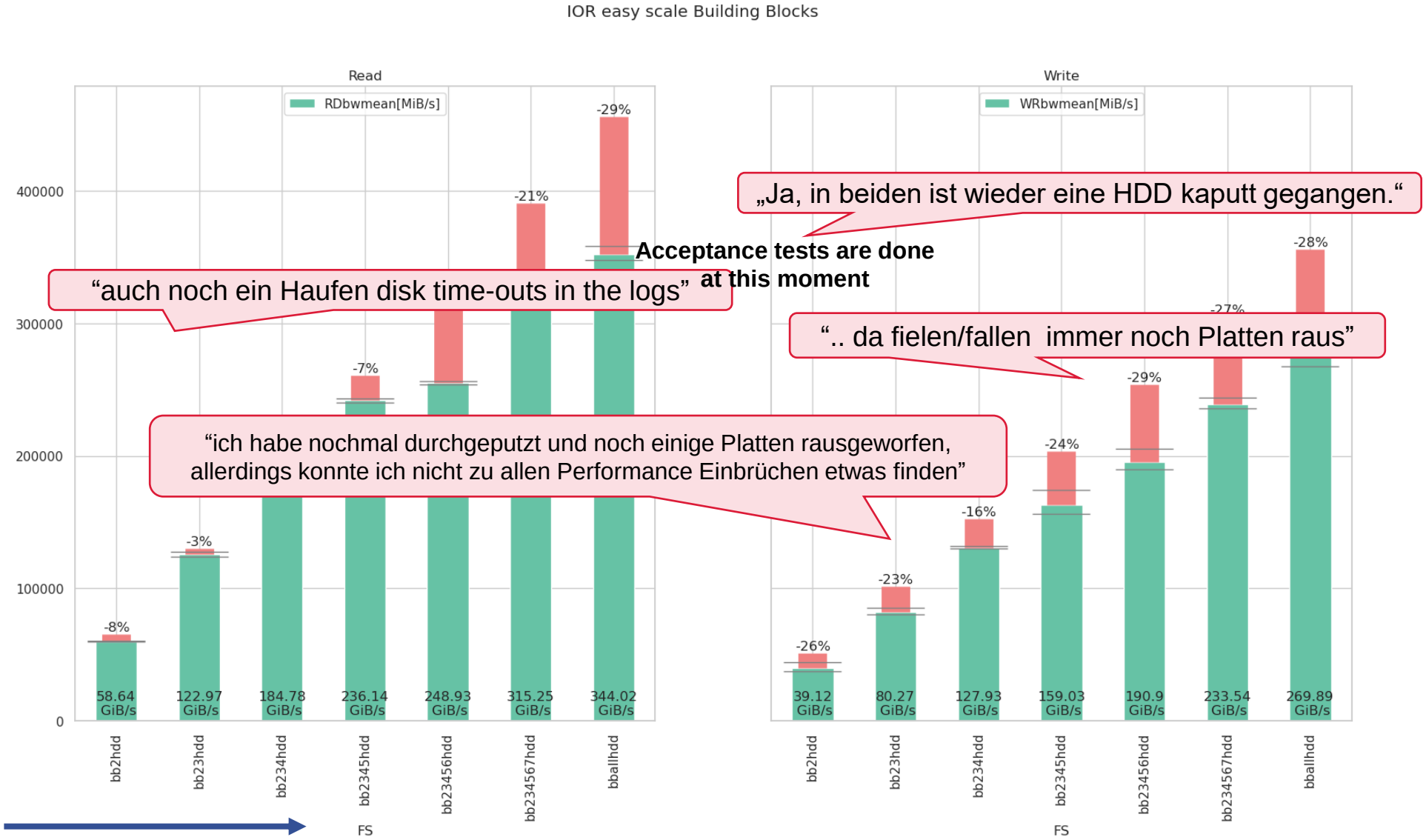
Warning:

You might not be able to keep track of all the details in following tests. We couldn't either and we had more than 20min

IOR EASY - JUST6 PHASE 1

7 x SSS6000 Building Blocks





„Ja, in beiden ist wieder eine HDD kaputt gegangen.“

Acceptance tests are done at this moment

“auch noch ein Haufen disk time-outs in the logs”

“.. da fielen/fallen immer noch Platten raus”

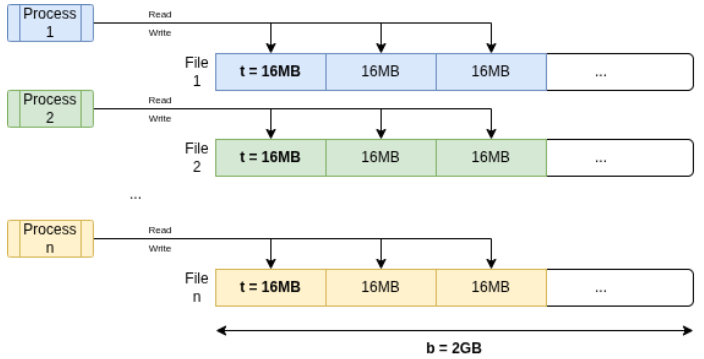
“ich habe nochmal durchgeputzt und noch einige Platten rausgeworfen, allerdings konnte ich nicht zu allen Performance Einbrüchen etwas finden”

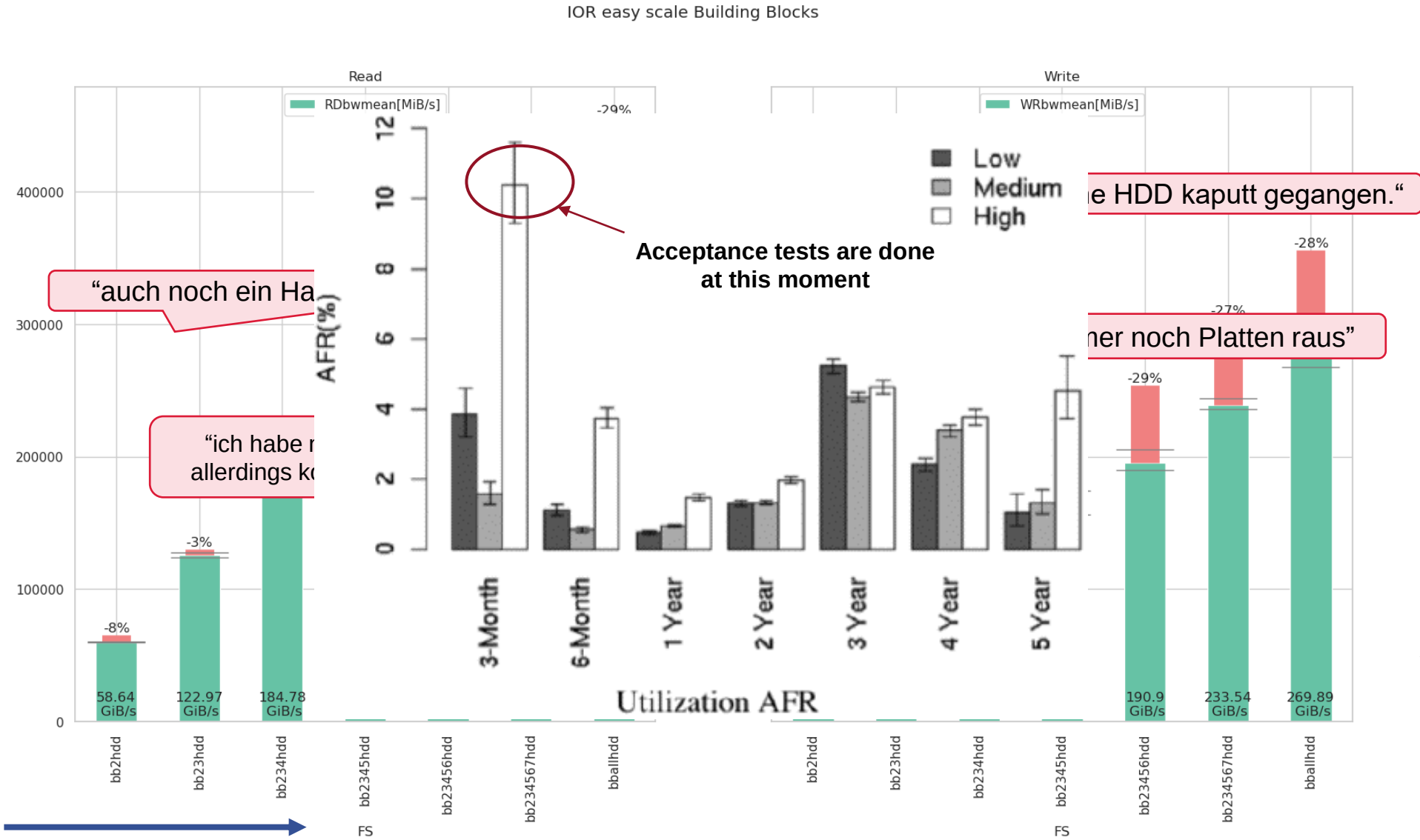
Burn IN

JÜLICH Forschungszentrum

IOR EASY - JUST6 PHASE 1

7 x SSS6000 Building Blocks





Acceptance tests are done at this moment

“auch noch ein Ha

“ich habe r
allerdings k

e HDD kaputt gegangen.“

ner noch Platten raus”

Burn IN

Increasing number of building blocks

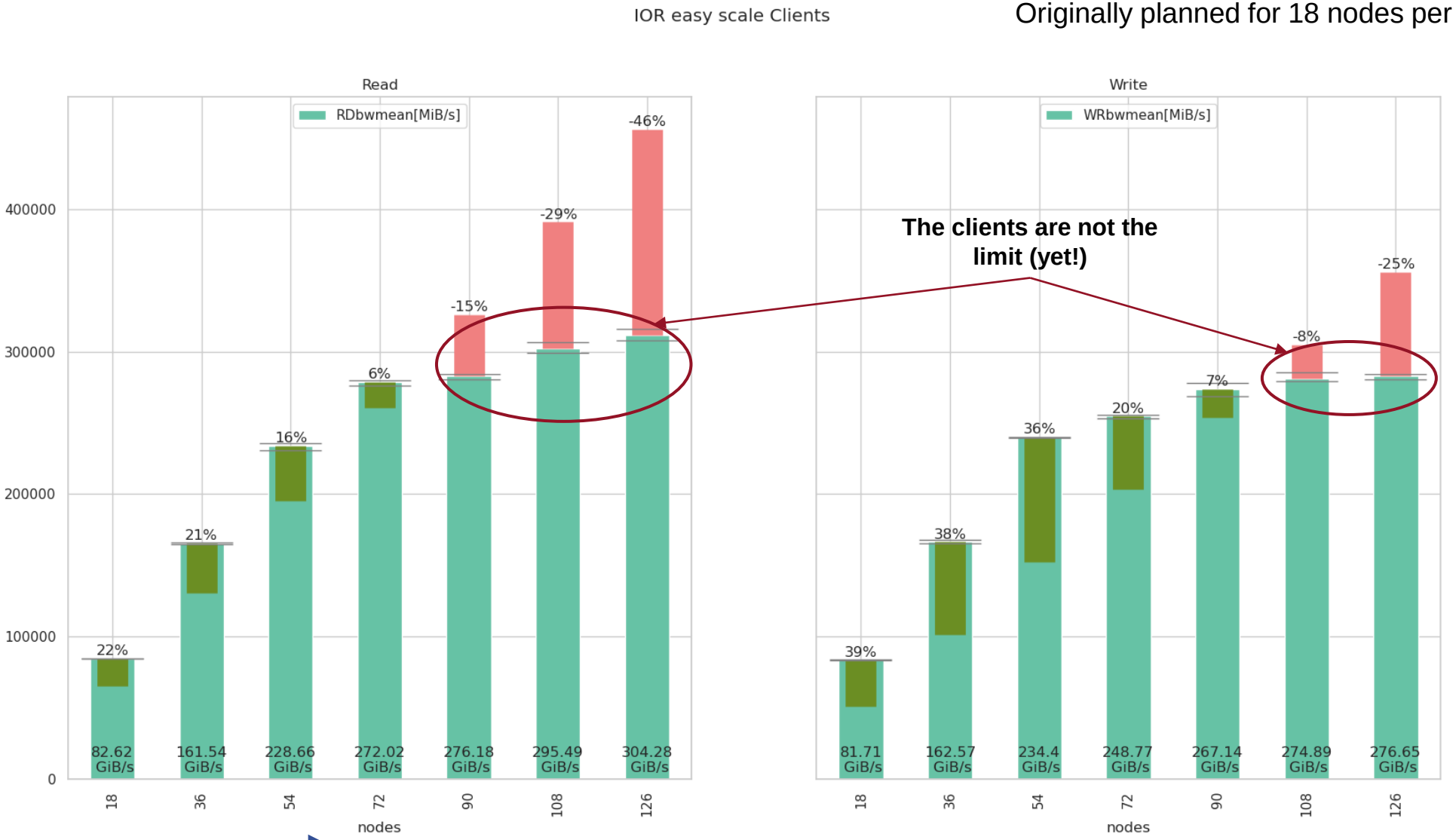
E. Pinheiro, et al. "Failure trends in a large disk drive population". 2007. In Proceedings of the 5th USENIX conference on File and Storage Technologies (FAST '07). USENIX Association, USA, 2.

JÜLICH Forschungszentrum

IOR EASY – JUST6 PHASE 1

Don’t forget the clients - 7 x SSS6000 Building Blocks

Originally planned for 18 nodes per Building Block (126 for 7)

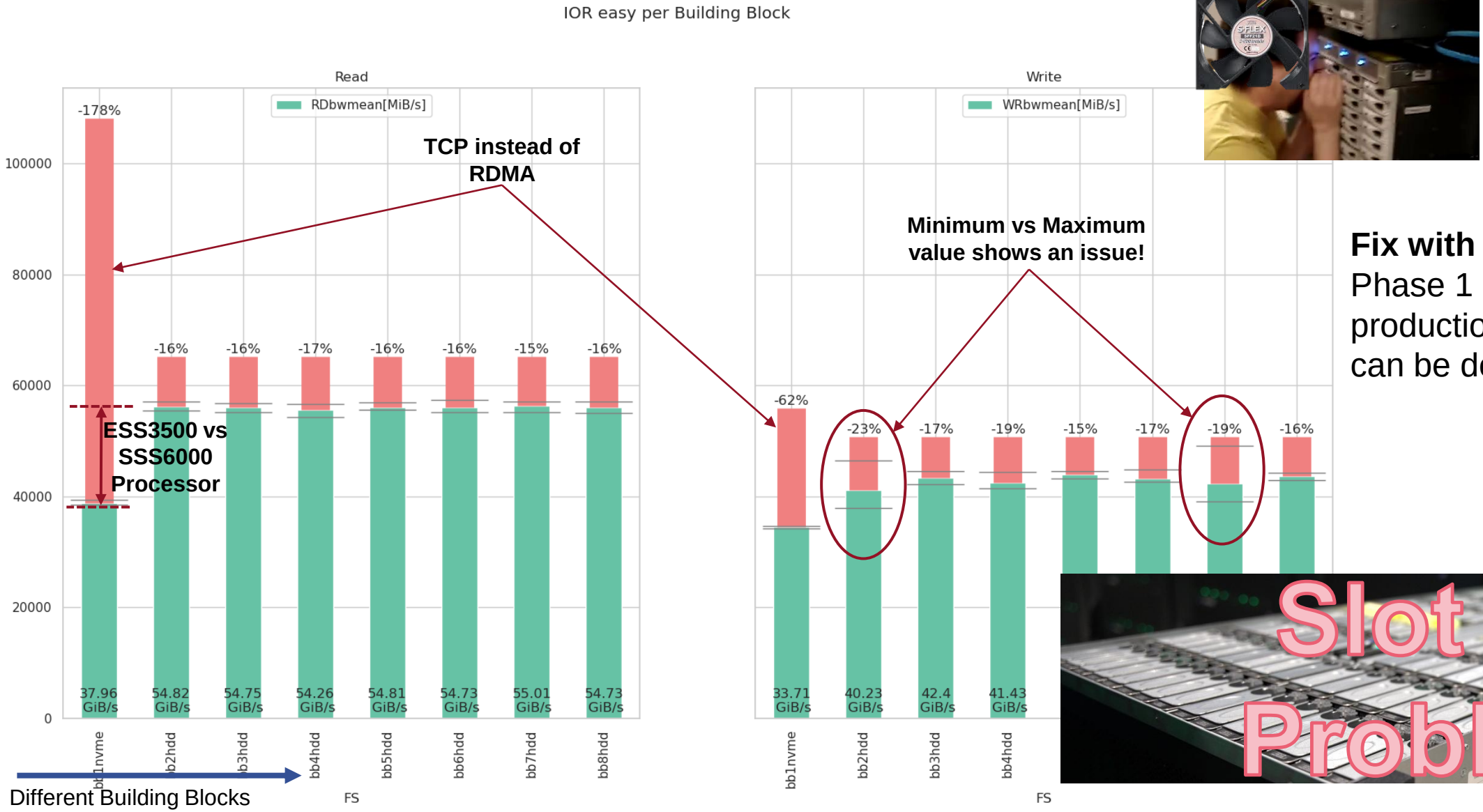


Increasing number of client nodes

JÜLICH Forschungszentrum

IOR EASY – JUST6 PHASE 1

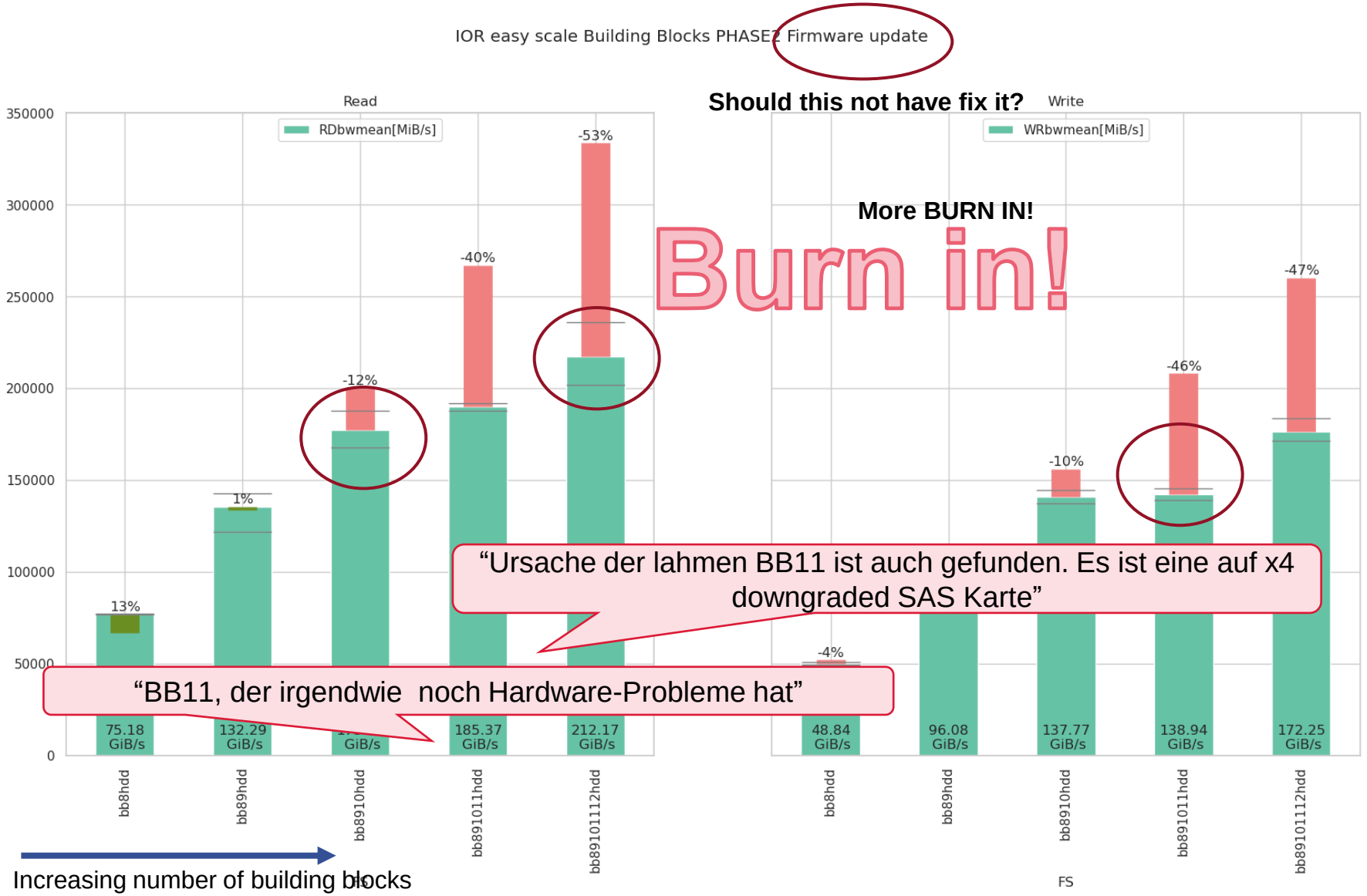
Let's test individual BB - 7 x SSS6000 Building Blocks



Fix with Firmware
Phase 1 goes into production before fix can be deployed!

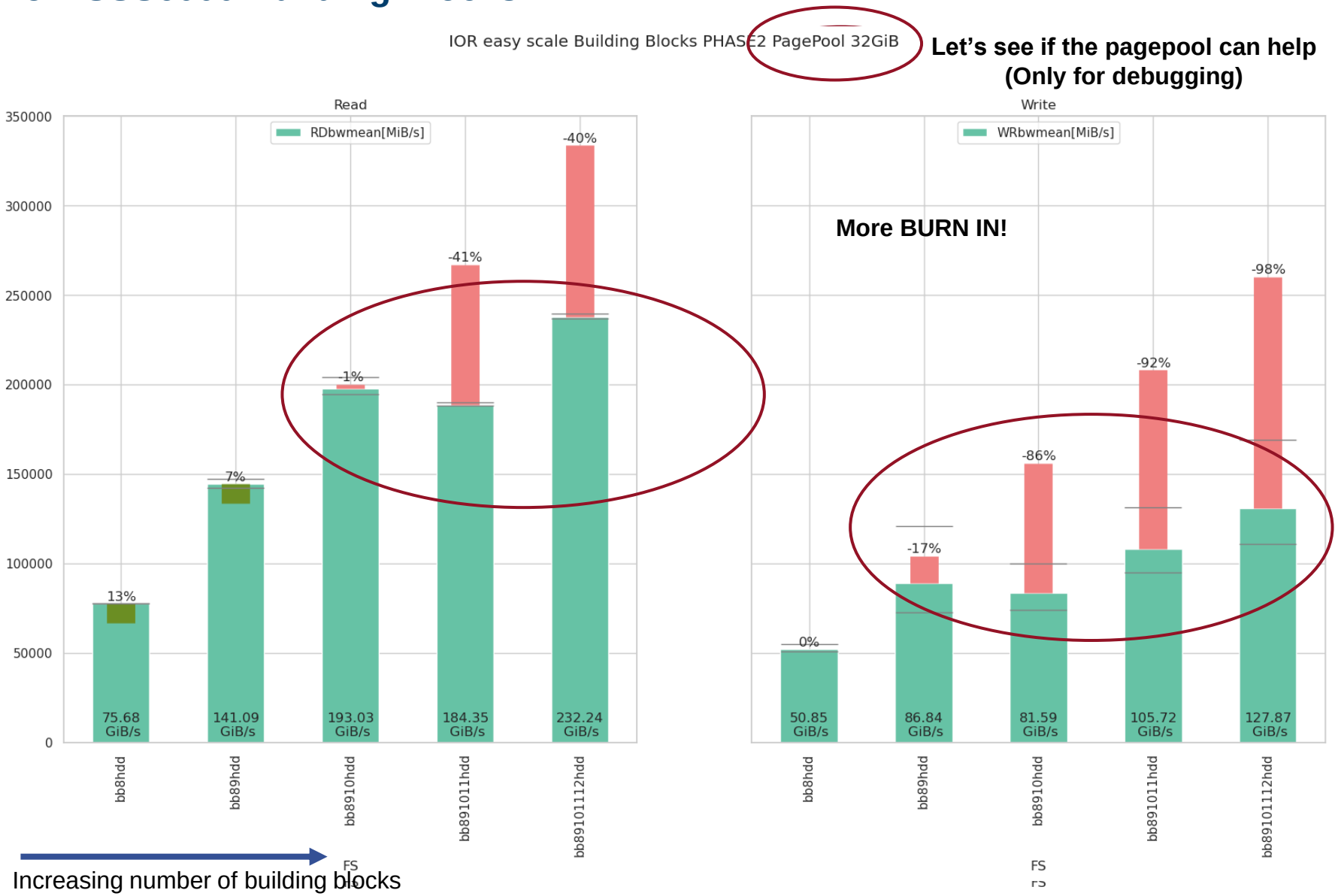
IOR EASY – JUST6 PHASE 2

5 x SSS6000 Building Blocks



IOR EASY – JUST6 PHASE 2

5 x SSS6000 Building Blocks



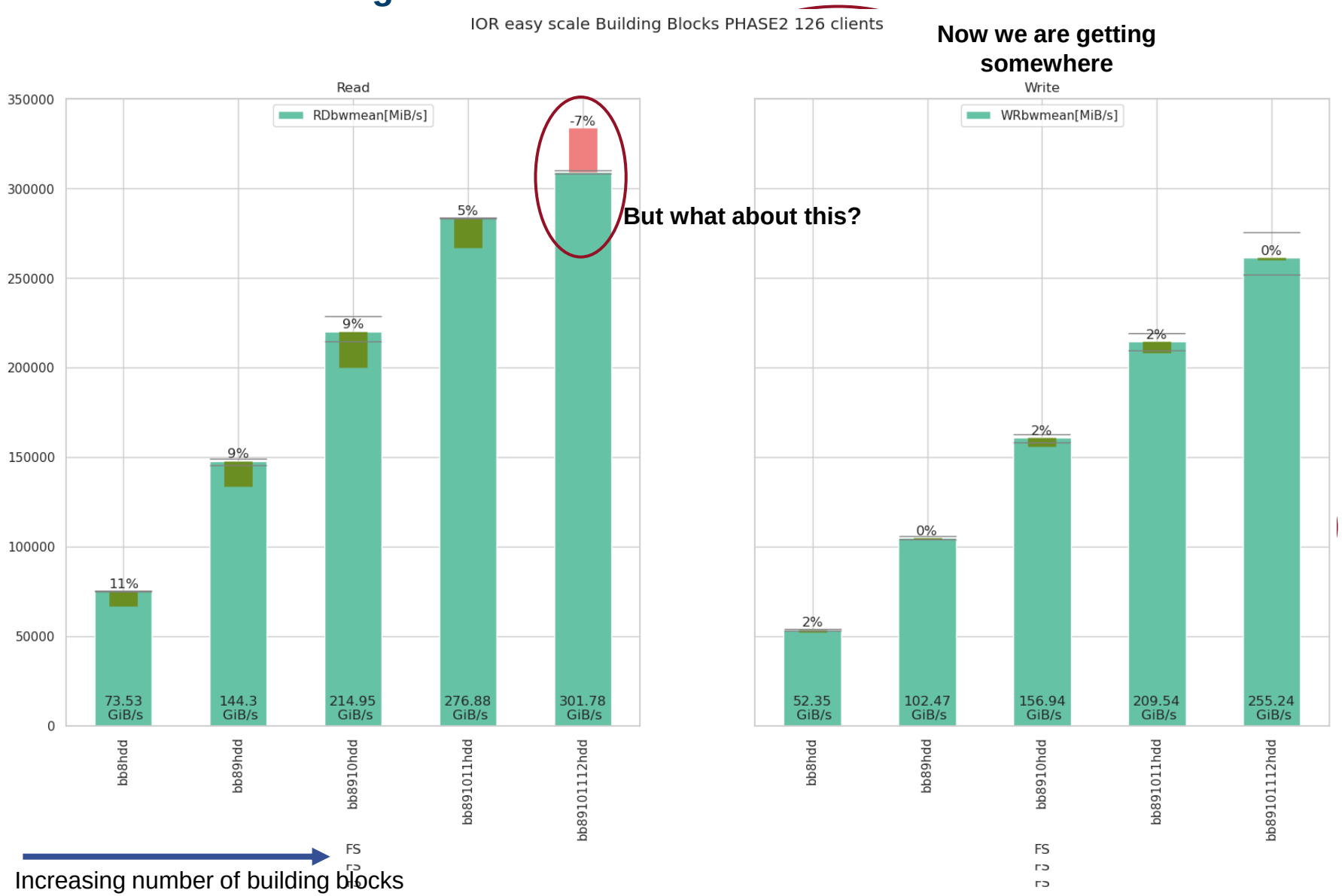
Let's see if the pagepool can help (Only for debugging)

Increasing number of building blocks

JÜLICH Forschungszentrum

IOR EASY – JUST6 PHASE 2

5 x SSS6000 Building Blocks



Now we are getting somewhere

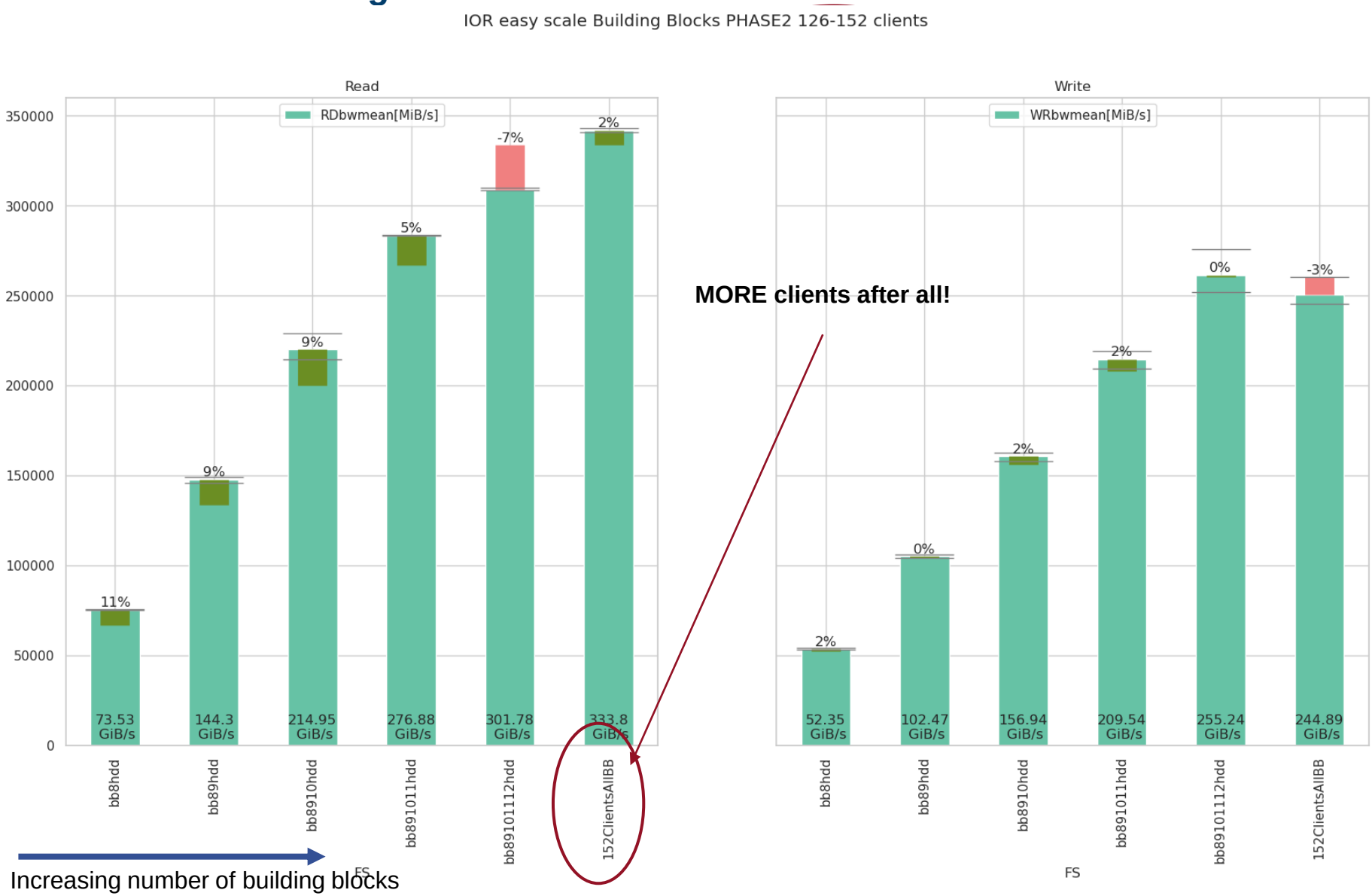
But what about this?

Increasing number of building blocks

JÜLICH Forschungszentrum

IOR EASY – JUST6 PHASE 2

5 x SSS6000 Building Blocks



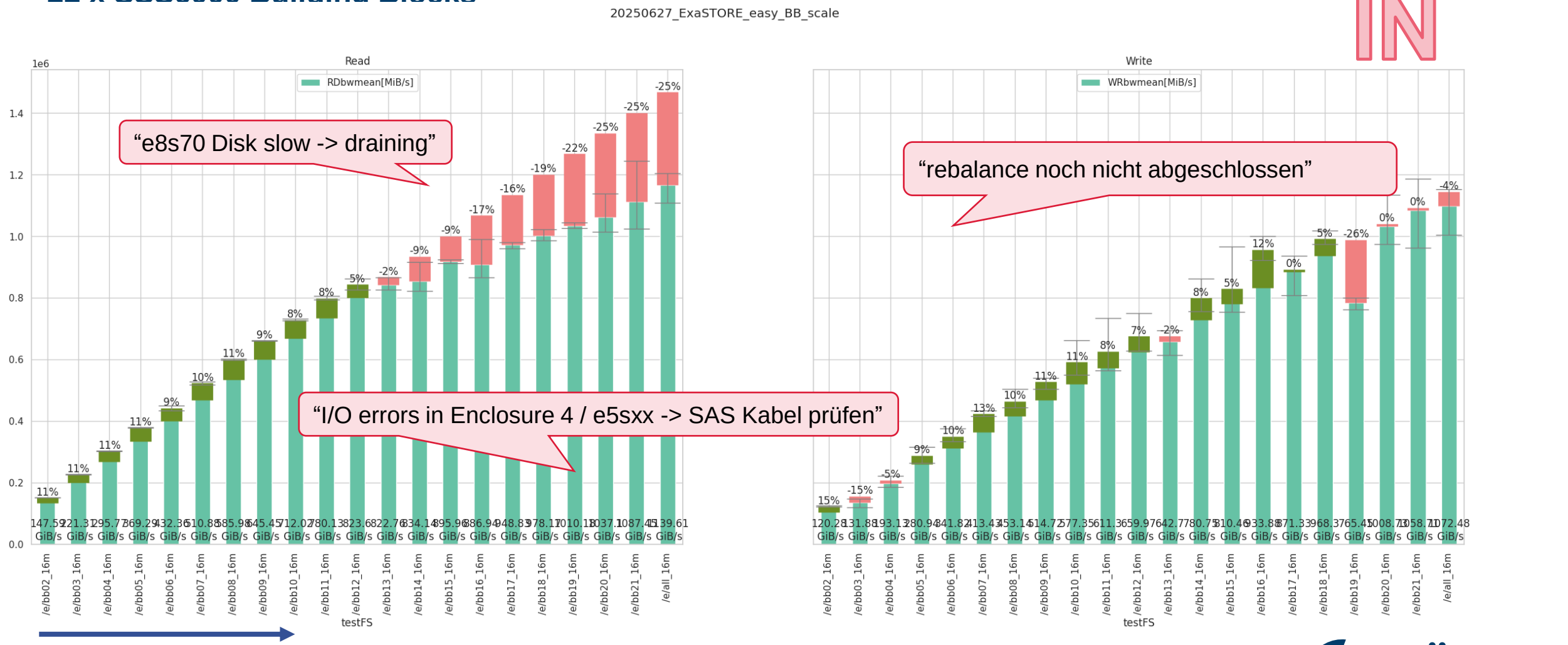
Increasing number of building blocks

JÜLICH Forschungszentrum

IOR EASY – EXASTORE

22 x SSS6000 Building Blocks

Burn IN

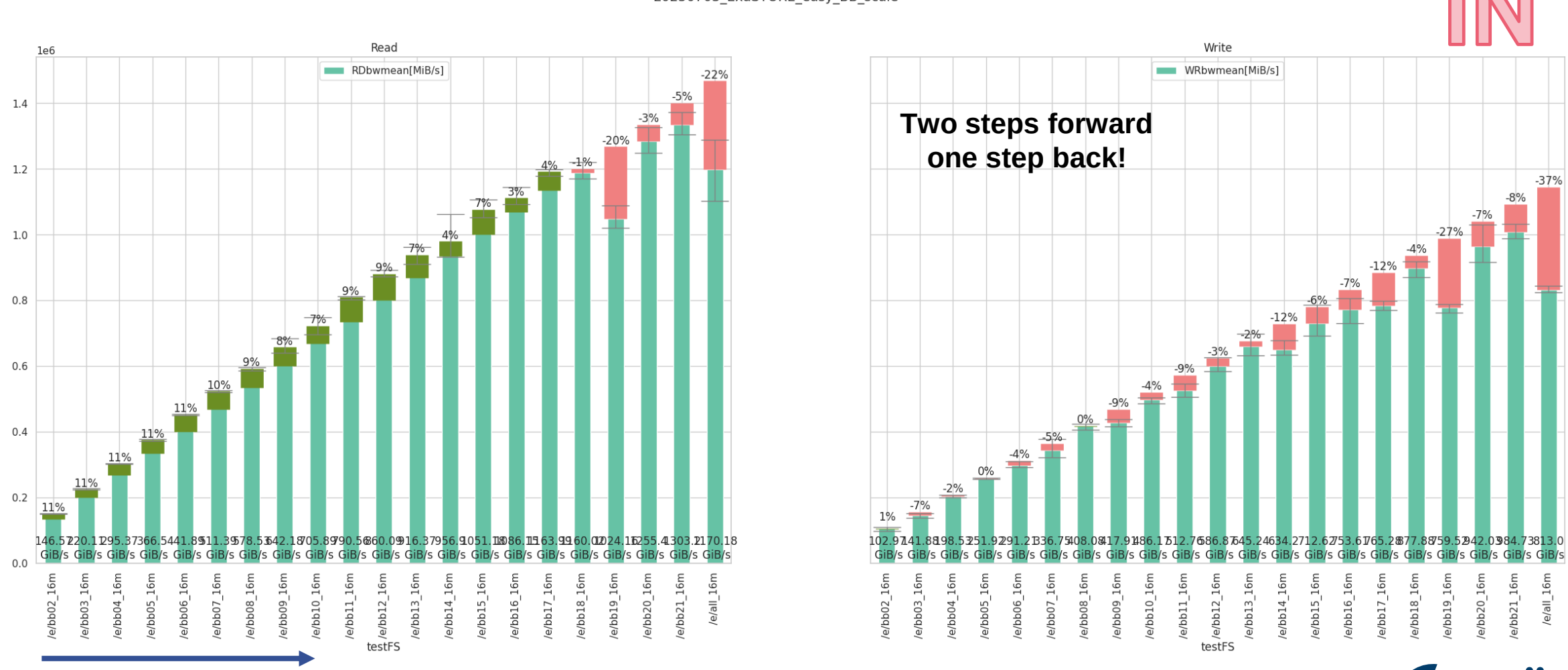


Increasing number of Building Blocks

IOR EASY – EXASTORE

22 x SSS6000 Building Blocks

Burn IN



Increasing number of Building Blocks

IOR EASY – EXASTORE

Better clients, more problems - 22 x SSS6000 Building Blocks

„ ... irgendein HPC node hat da seinen mount verloren“

„Wieder mounts verloren“

„ bb04 sah sehr gut aus, zumindest bis ein Knoten die mounts verloren hat,` “

- IPoIB Problems:

1. **Compute expelling Storage nodes (and vice versa):**

   Temp solutions:

   Ping all compute and storage nodes before mounting

   Increase ARP cache and extend Garbage collector thresholds

   Increase timeout till nodes are expelled (`failureDetectionTime`)

   Final solution: Hope the stabilization of the IB network goes well

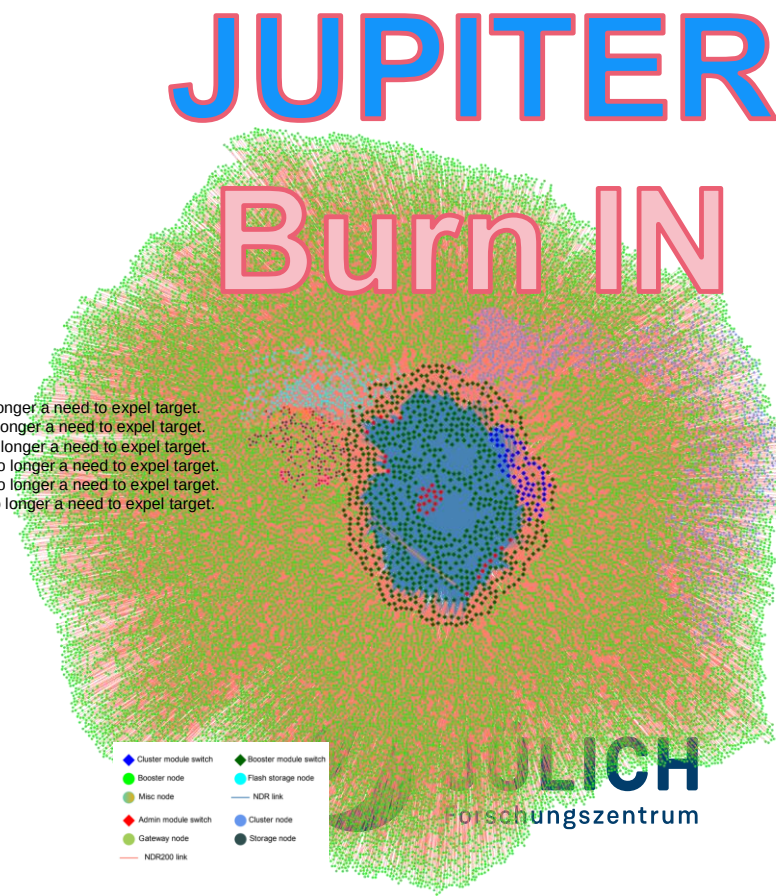
2. **Compute expelling compute:**

```
[N] Expel 10.128.16.8   (jpbo-001-08-interconnect-1.jupiter.internal in j request from 10.128.19.137 (jpbo-019-41-interconnect-1.jupiter.internal in j. Expelling 10.128.16.8   (jpbo-001-08-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
[N] Expel 10.128.18.84  (jpbo-013-20-interconnect-1.jupiter.internal in j request from 10.128.16.91  (jpbo-002-43-interconnect-1.jupiter.internal in j. Expelling 10.128.18.84  (jpbo-013-20-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
[N] Expel 10.128.18.90  (jpbo-013-26-interconnect-1.jupiter.internal in j request from 10.128.17.196 (jpbo-010-20-interconnect-1.jupiter.internal in j. Expelling 10.128.18.90  (jpbo-013-26-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
[N] Expel 10.128.16.116 (jpbo-003-20-interconnect-1.jupiter.internal in j request from 10.128.26.162 (jpbo-035-02-interconnect-1.jupiter.internal in j. Expelling 10.128.16.116 (jpbo-003-20-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
[N] Expel 10.128.18.104 (jpbo-013-40-interconnect-1.jupiter.internal in j request from 10.128.16.230 (jpbo-005-38-interconnect-1.jupiter.internal in j. Expelling 10.128.18.104 (jpbo-013-40-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
[N] Expel 10.128.18.106 (jpbo-013-42-interconnect-1.jupiter.internal in j request from 10.128.18.59  (jpbo-012-43-interconnect-1.jupiter.internal in j. Expelling 10.128.18.106 (jpbo-013-42-interconnect-1.jupiter.internal in j based on no longer a need to expel target.
...
```

   Temp solution:

   Avoid inter client node communication (`preferDesignatedMnode=yes`)

   Final solution: Hope the stabilization of the IB network goes well



IOR EASY - EXASTORE

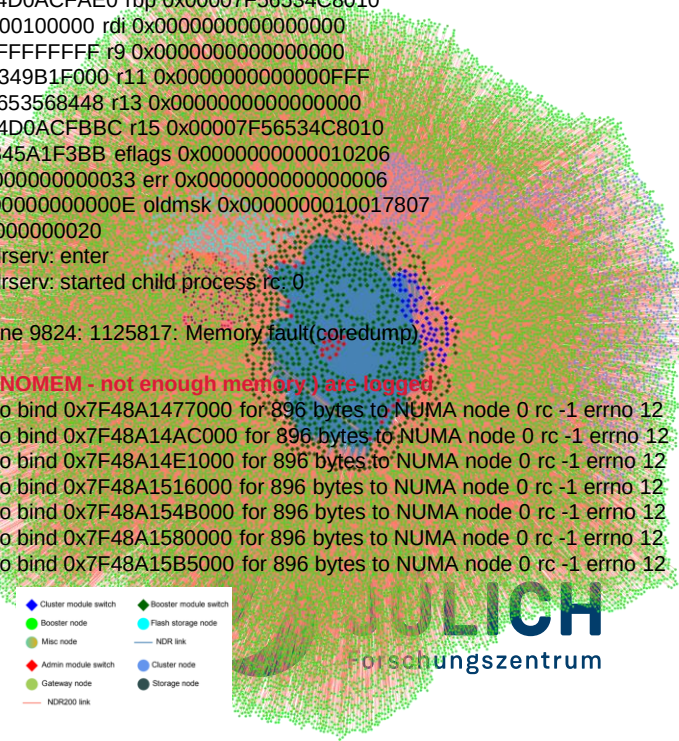
Too many clients for Scale? - 22 x SSS6000 Building Blocks

Remember:

5884 Nodes with NVIDIA Mellanox NDR **4×** NDR200 NICs

- **Problem:** At 3-4k nodes client manager hits (ENOMEM – not enough memory) Managers start to cycle
- **Temporary Solution:**
  - `maxReceiverThreads=128` [quorum]
  - `maxTcpConnsPerNodeConn=1`
  - `proactiveReconnect=no`
  - **`verbsPorts=mlx5_0/1** ~~*mlx5_1/1 mlx5_2/1 mlx5_3/1*~~`
- **Final Solution (Untested):**
  - `PF_LOG_BAR_SIZE=10` (Recommended by Nvidia)

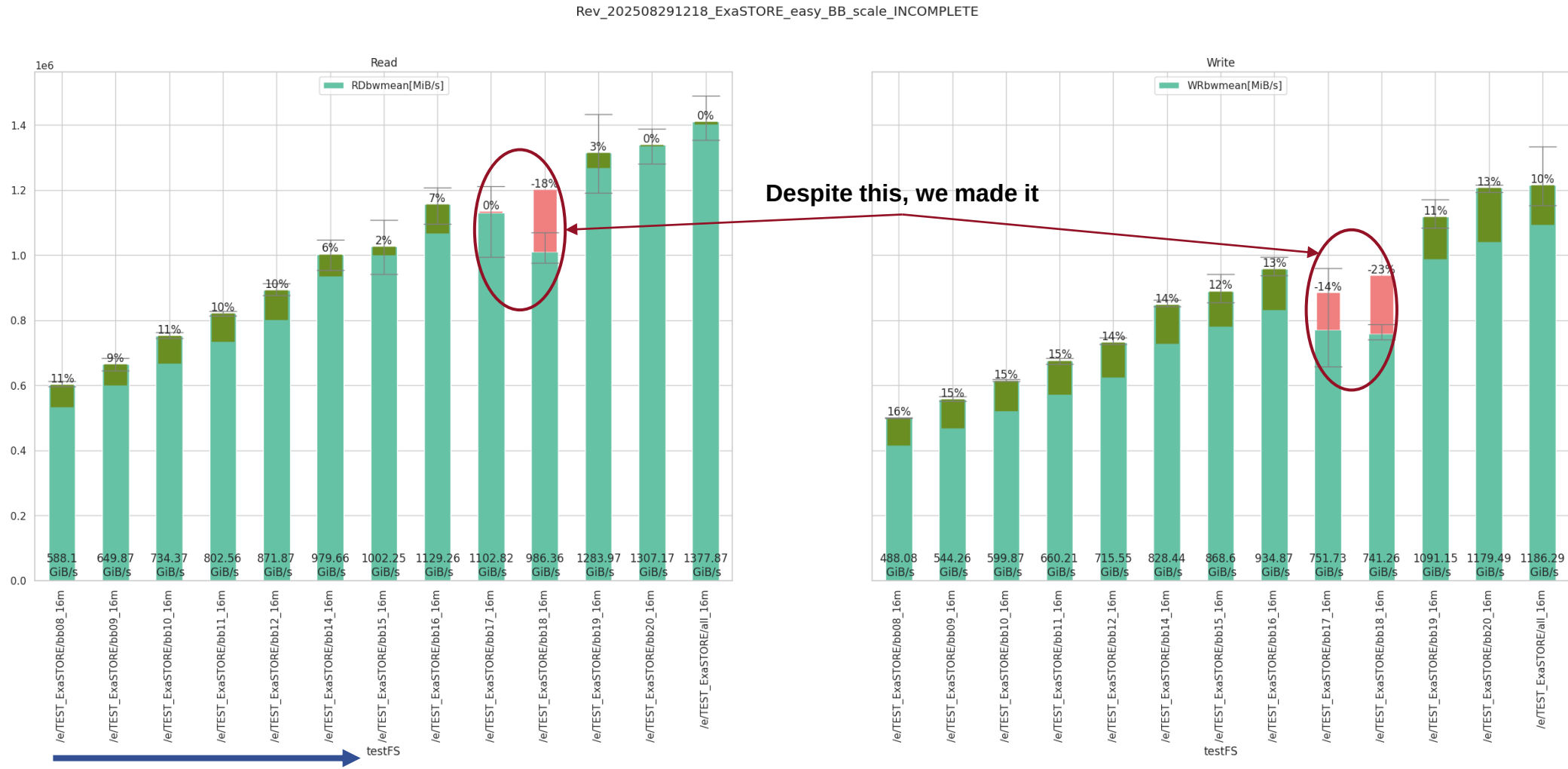
2025-07-21_13:11:38.476+0200: [E] Signal 11 at location 0x55E345A1F3BB in process 1125817, link reg 0xFFFFFFFFFF
2025-07-21_13:11:38.476+0200: [I] rax 0x0000000000000000 rbx 0x0000000000000000
2025-07-21_13:11:38.476+0200: [I] rcx 0x0000000000000000 rdx 0x0000000000000000
2025-07-21_13:11:38.476+0200: [I] rsp 0x00007F54D0ACFAE0 rbp 0x00007F56534C8010
2025-07-21_13:11:38.476+0200: [I] rsi 0x0000000000100000 rdi 0x0000000000000000
2025-07-21_13:11:38.476+0200: [I] r8 0x00000000FFFFFFFF r9 0x0000000000000000
2025-07-21_13:11:38.476+0200: [I] r10 0x000055E349B1F000 r11 0x0000000000000FFF
2025-07-21_13:11:38.476+0200: [I] r12 0x00007F5653568448 r13 0x0000000000000000
2025-07-21_13:11:38.476+0200: [I] r14 0x00007F54D0ACFBBC r15 0x00007F56534C8010
2025-07-21_13:11:38.476+0200: [I] rip 0x000055E345A1F3BB eflags 0x0000000000010206
2025-07-21_13:11:38.476+0200: [I] csgsfs 0x002B000000000033 err 0x0000000000000006
2025-07-21_13:11:38.476+0200: [I] trapno 0x000000000000000E oldmsk 0x0000000010017807
2025-07-21_13:11:38.476+0200: [I] cr2 0x0000000000000020
2025-07-21_13:11:38.480+0200: [N] Starting mmsdrserv: enter
2025-07-21_13:11:38.480+0200: [N] Starting mmsdrserv: started child process rc: 0
malloc(131072) for operator new() failed.
/usr/lpp/mmfs/bin/runmmfs[494]: numactlWrapper: line 9824: 1125817: Memory fault(coredump)
no further stack information.
before that a lot of NUMA warning with err 12 (ENOMEM - not enough memory) are logged
2025-07-21_13:11:26.734+0200: [W] NUMA failed to bind 0x7F48A1477000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:26.742+0200: [W] NUMA failed to bind 0x7F48A14AC000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:26.750+0200: [W] NUMA failed to bind 0x7F48A14E1000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:26.759+0200: [W] NUMA failed to bind 0x7F48A1516000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:27.288+0200: [W] NUMA failed to bind 0x7F48A154B000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:27.289+0200: [W] NUMA failed to bind 0x7F48A1580000 for 896 bytes to NUMA node 0 rc -1 errno 12
2025-07-21_13:11:27.291+0200: [W] NUMA failed to bind 0x7F48A15B5000 for 896 bytes to NUMA node 0 rc -1 errno 12
…



Ask your IBM contact what parameters are best for your setup

IOR EASY - EXASTORE

21 x SSS6000 Building Blocks (1 BB already in production)



Despite this, we made it

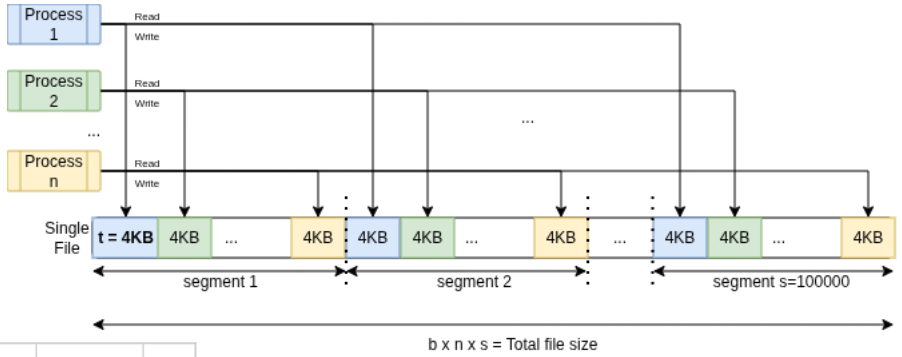
Increasing number of Building Blocks

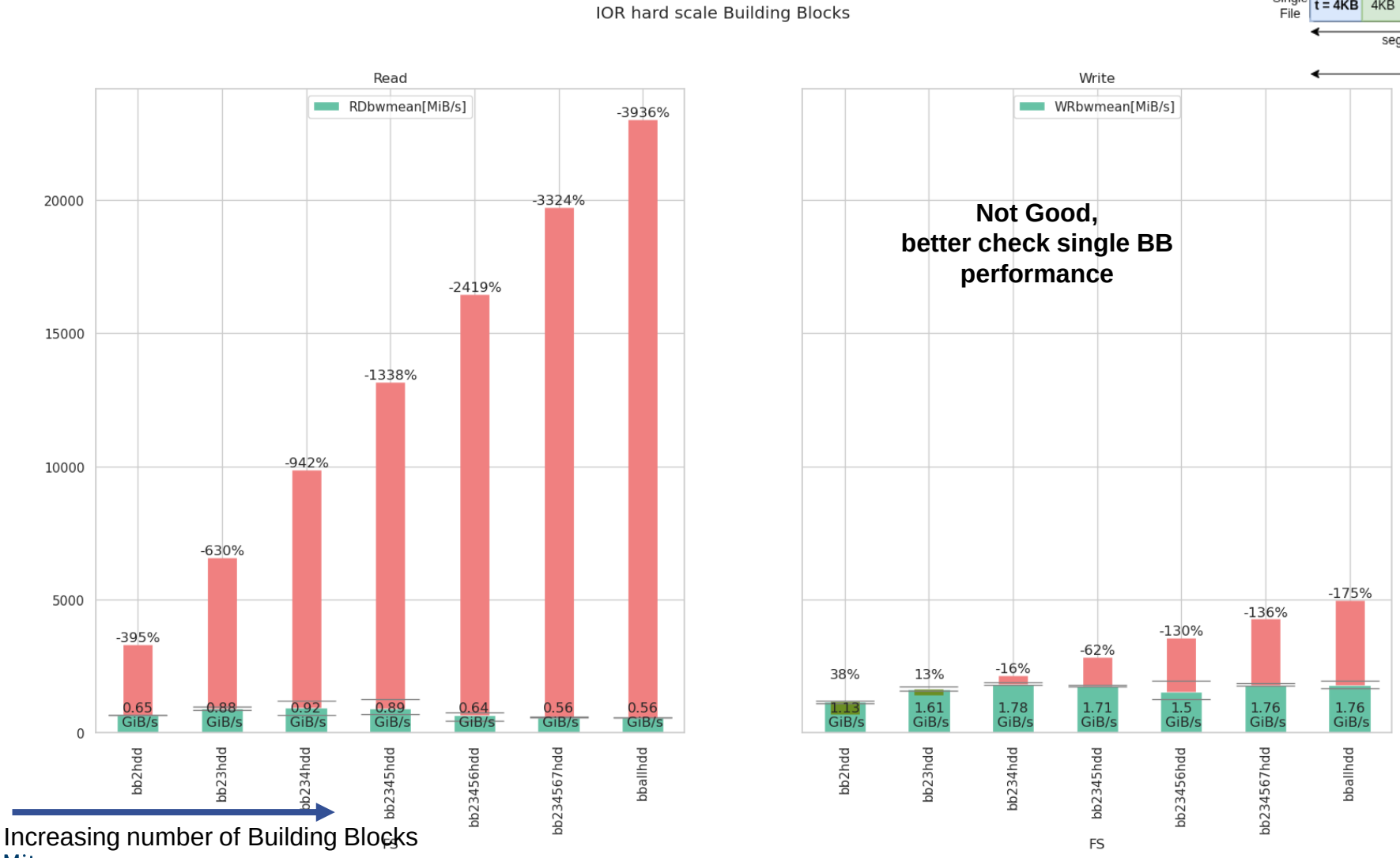
ACCEPTANCE TEST LIST – HOW FAR ARE WE?

- ✓ IOR Easy
  - ✓ Each Building Block
    - ✓ IOR Easy 16MB blocksize file system
    - ❑ IOR Easy 512KB blocksize file system
  - ✓ All Building Blocks
    - ✓ IOR Easy 16MB blocksize file system
    - ❑ IOR Easy 512KB blocksize file system
- ❑ IOR Hard
  - ❑ Each Building Block
    - ❑ IOR Hard 16MB blocksize file system
    - ❑ IOR Hard 512KB blocksize file system
  - ❑ All Building Blocks
    - ❑ IOR Hard 16MB blocksize file system
    - ❑ IOR Hard 512KB blocksize file system
- ❑ MDTest
  - ❑ Each Building Block
    - ❑ MDTest 16MB block size file system
    - ❑ MDTest 512KB block size file system
  - ❑ All Building Block
    - ❑ MDTest 16MB block size file system
    - ❑ MDTest 512KB block size file system

IOR HARD - JUST6 PHASE 1

7 x SSS6000 Building Blocks

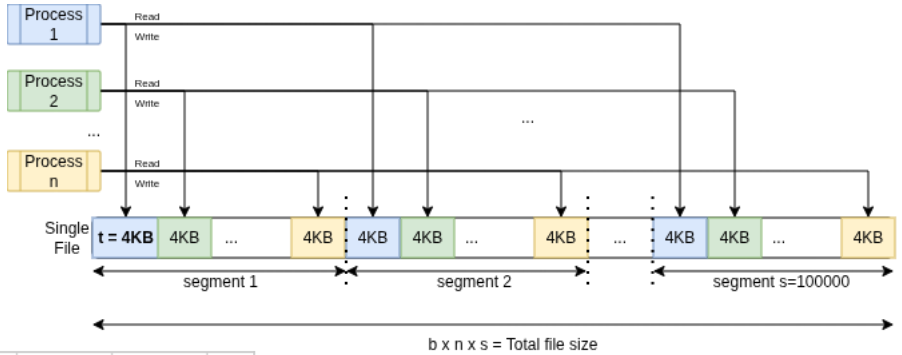


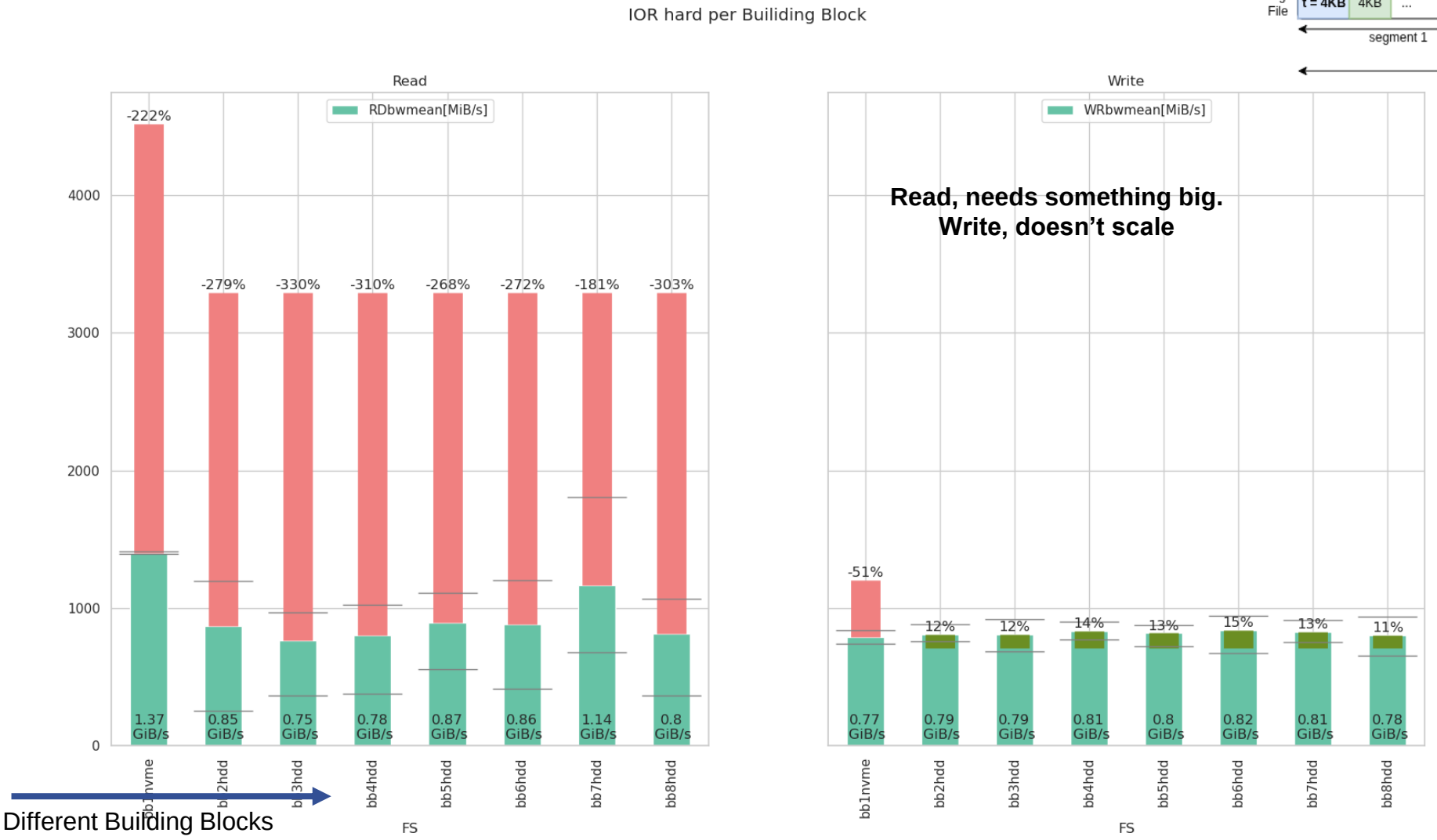


JÜLICH Forschungszentrum

IOR HARD - JUST6 PHASE 1

Single BB Tests - 7 x SSS6000 Building Blocks



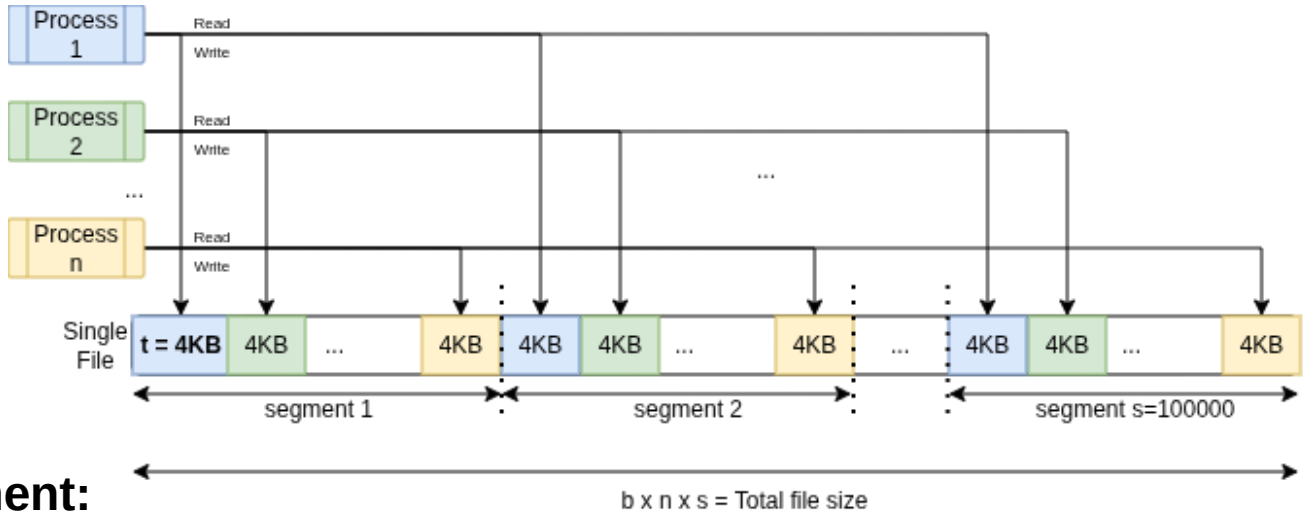


Software issue that requires:

- `finegrainwritesharing`and `finegrainreadshraing`.
- Many many many parameter changes
- Additional efix and updates to GPFS

IOR HARD - JUST6 PHASE 1

Why it is harder than you think? - GPFS Tokens



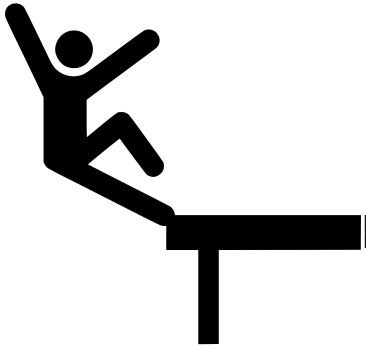
GPFS implements a distributed token management:

- Each Node (GPFS daemon) has to acquire a token for its regions
- The acquisition requires contacting token server or a client node that holds the token for this region

REQUIRED TO BE POSIX COMPLYANT!

Solution, let's not be:

- `finegrainwritesharing` and `finegrainreadshraing` IOR
  = No token management for that operation
  (**I know what I'm doing**)

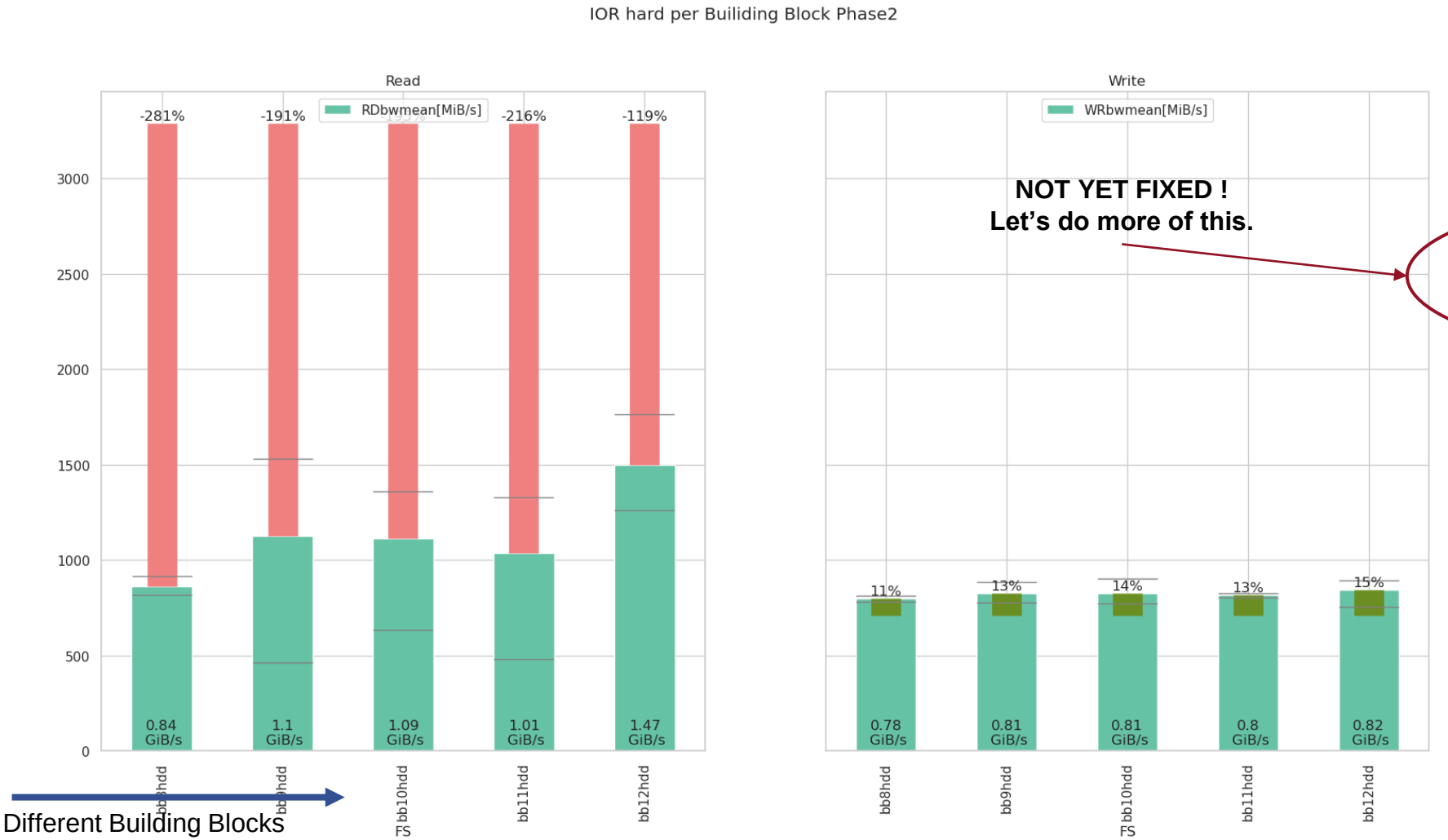


POSIX Safety

Let's revisit in **Phase 2.**

IOR HARD - JUST6 PHASE 2

Single BB Tests - 5 x SSS6000 Building Blocks



Different Building Blocks

Software issue that requires:

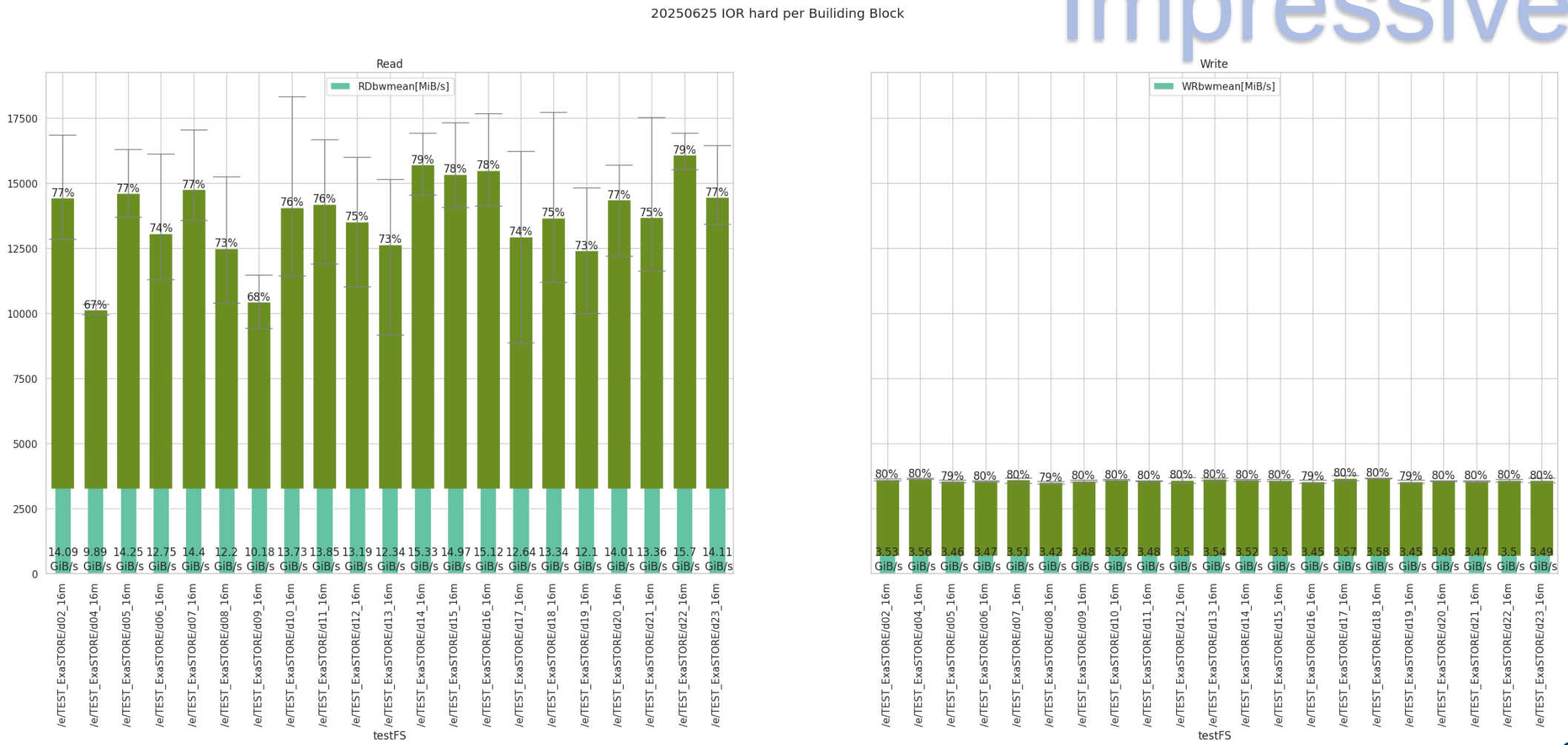
- `finegrainwritesharing`and `finegrainreadshraing`.
- Many many many parameter changes
- Additional efix and updates to GPFS

Let's revisit in **ExaSTORE**

IOR HARD - EXASTORE

22 x SSS6000 Building Blocks

Impressive!



Different Building Blocks

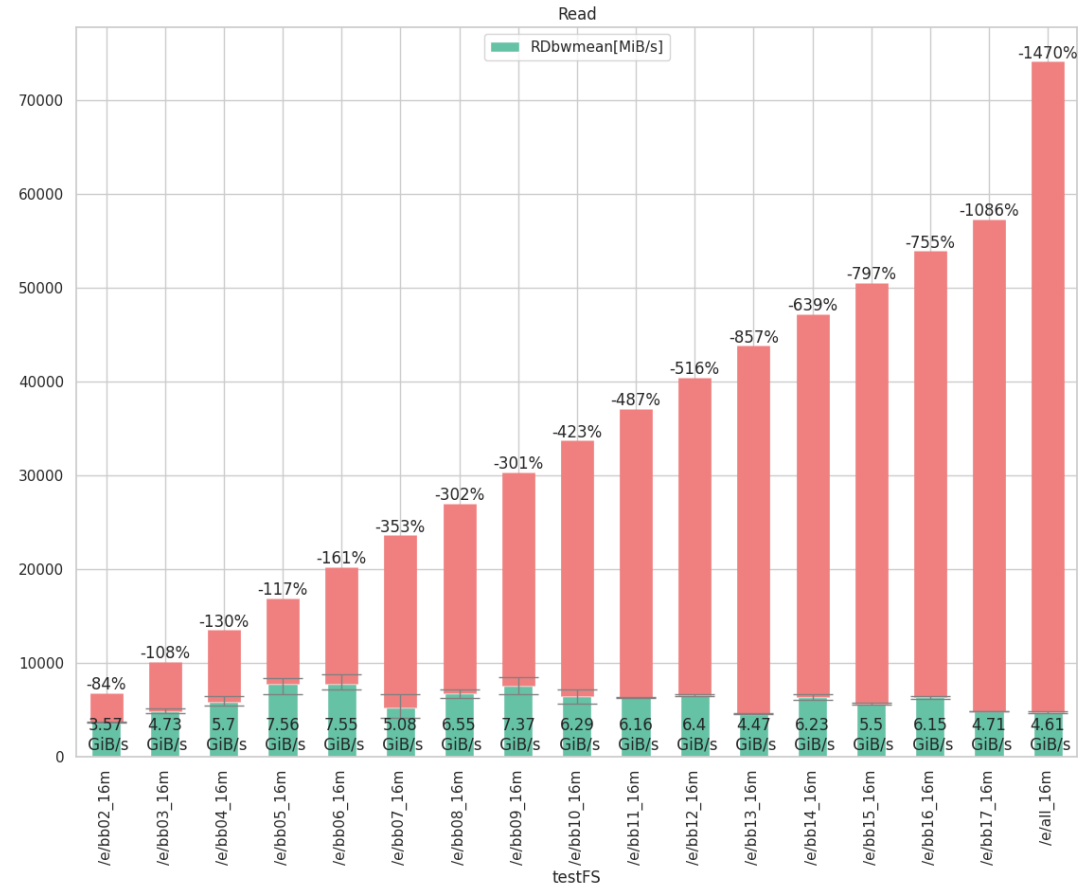
But does it scale?

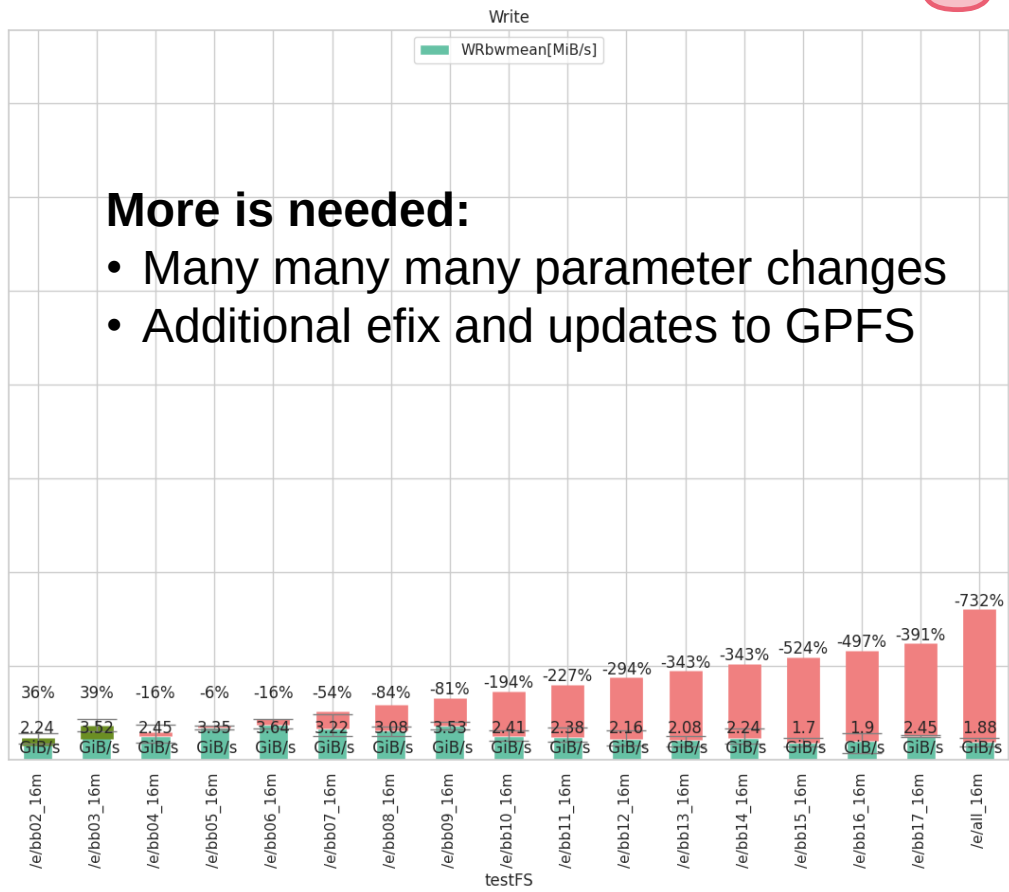
IOR HARD - EXASTORE

Scale test - 22 x SSS6000 Building Blocks

20250629_IOR hard scale Building Blocks INCOMPLETE

Not enough





More is needed:

- Many many many parameter changes
- Additional efix and updates to GPFS

Different Building Blocks

Did it work?

IOR HARD – EXASTORE

22 x SSS6000 Building Blocks

Still working on it

Cliff hanger!

PRESENT ENDEAVORS

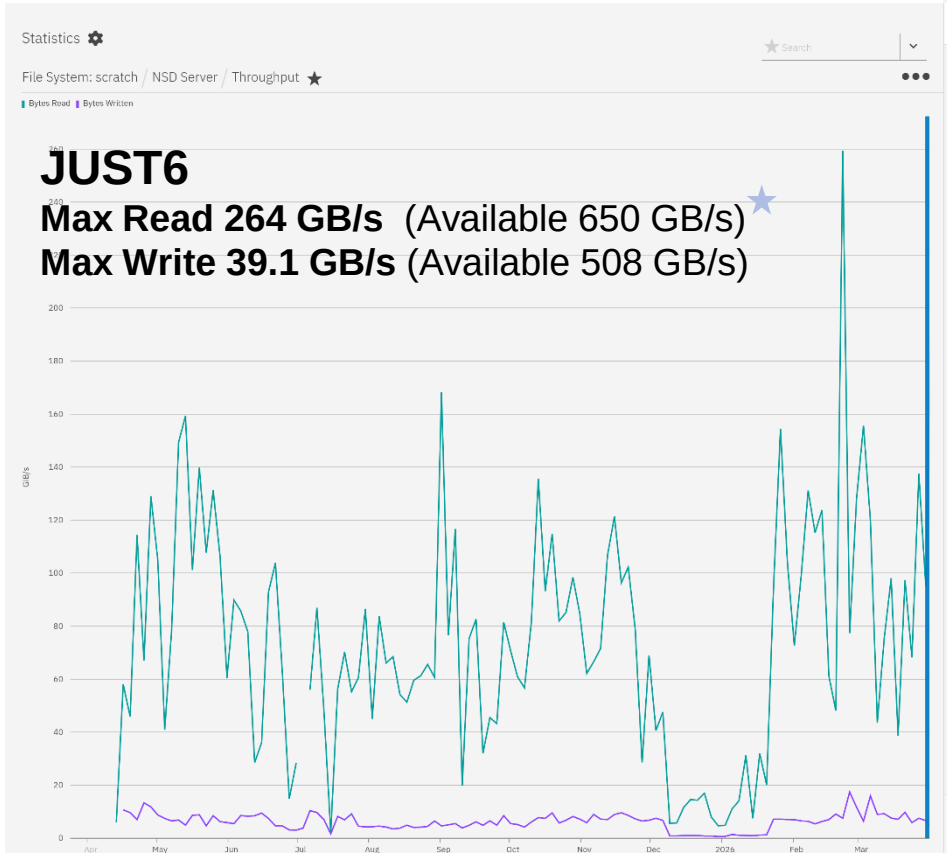
- IOR Hard scaling tests
- IO500 on ExaSTORE
- ExaFLASH Acceptance tests
- IO500 on ExaFLASH
- Continuous Benchmarking of I/O performance
- …

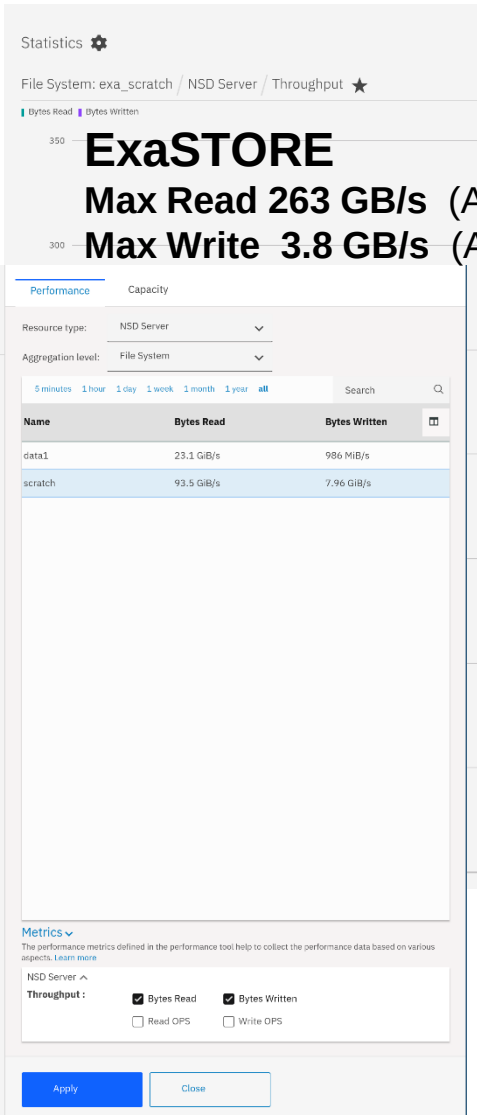
BUT !

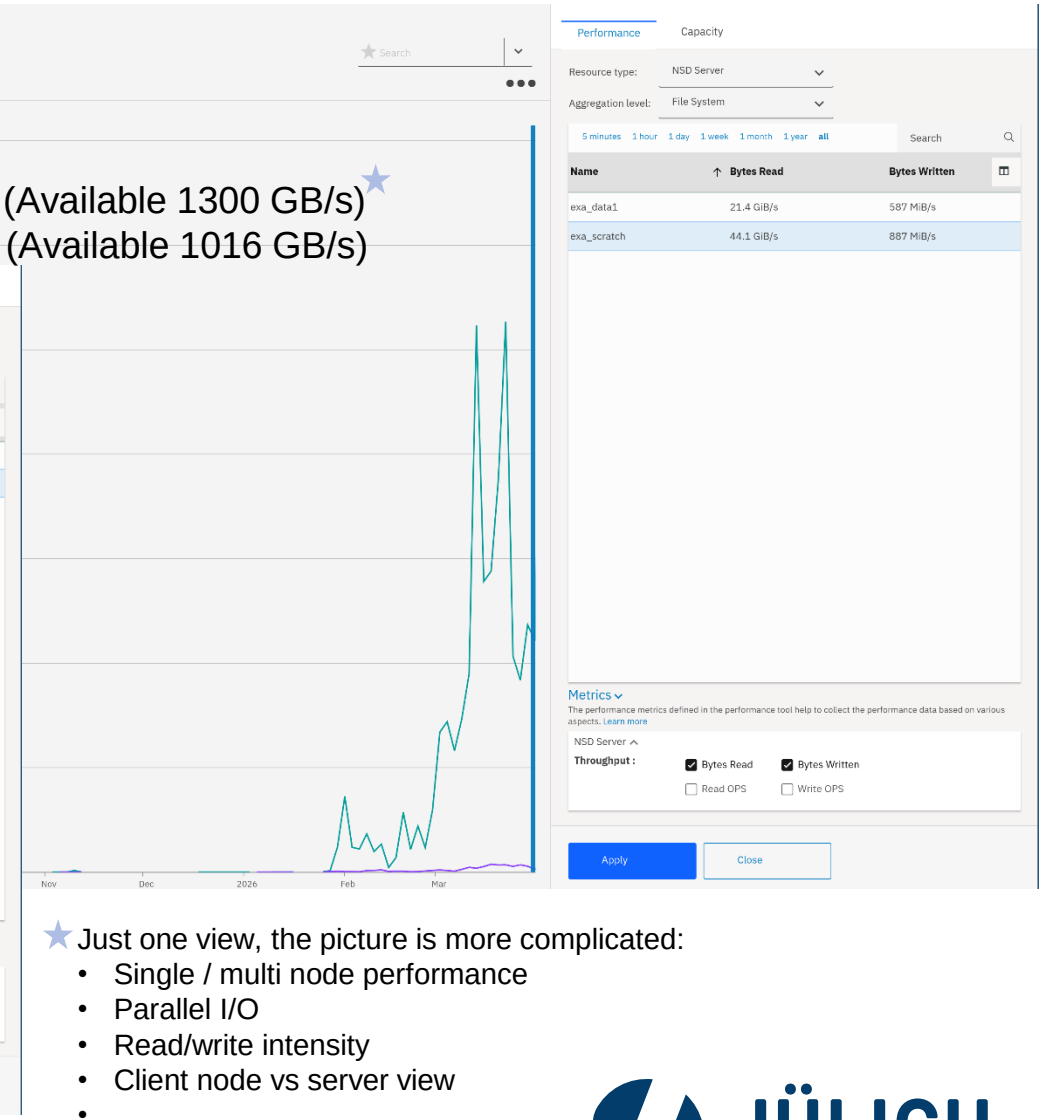
Who cares about I/O?

ARE USERS DOING I/O?

I/O reported by GPFS







★ Just one view, the picture is more complicated:

- Single / multi node performance
- Parallel I/O
- Read/write intensity
- Client node vs server view
- …

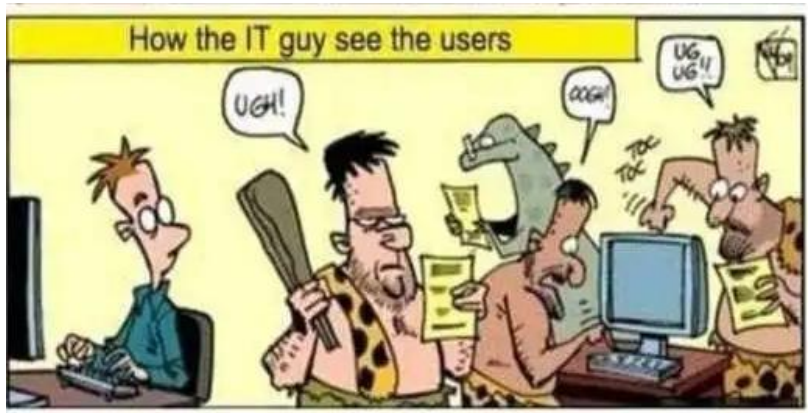
AN OBSERVED MISMATCH

USERS

- Regularly claim they need more I/O performance

Admins

- Monitoring shows little I/O performed



WHICH IS IT?

Or better,
how do we find the truth?

REMEMBER - PROCUREMENT BENCHMARK

Benchmark	Domain	Dwarfs
Amber*	MD	Particle, Spectral
Arbor	Neurosci.	Dense LA, Sparse LA, Unstructured Grid
Chroma-QCD	QCD	Sparse LA, Structured Grid, Monte Carlo
GROMACS	MD	Particle, Spectral
ICON	Climate	Sparse LA, Structured Grid
JUQCS	QC	Sparse LA
nekRS	CFD	Sparse LA, Spectral, Unstructured Grid
ParFlow*	Earth Sys.	Sparse LA, Structured Grid
PIConGPU	Plasma	Particle, Unstructured Grid
Quantum Espresso	Materials Sci.	Dense LA, Particle, Spectral
SOMA*	Polymer Sys.	Particle, Monte Carlo
MMoCLIP	AI (MM)	Dense LA
Megatron-LM	AI (LLM)	Dense LA
ResNet*	AI (Vision)	Dense LA
DynQCD	QCD	Sparse LA, Structured Grid, Monte Carlo
NAStJA	Biology	Structured Grid, Monte Carlo
Graph500	Graph	*Graph Traversal (D. 9)*
HPCG	CG	Sparse LA
HPL	LA	Dense LA
IOR	Filesys.	*Input/Output*
LinkTest	Network	*P2P, Topology*
OSU	Network	*Message Exchange, DMA*
STREAM	Memory	*Regular Access*

A. Herten et al., "Application-Driven Exascale: The JUPITER Benchmark Suite," SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1-45, doi: 10.1109/SC41406.2024.00038.

Just one side of the story.

The only I/O benchmark and it's synthetic!

★ Procurement of a compute system

IT’S A DISCUSSION !

Do we need to understand the users requirement?

Another,
Cliff hanger!

THANKS !

GREETINGS FROM THE JSC STORAGE TEAM