Exploring Data Paths in HPC Systems using the IO500

Sarah Neuwirth*, Hariharan Devarajan (LLNL), Jay Lofstead (SNL)

*Johannes Gutenberg University Mainz, Germany

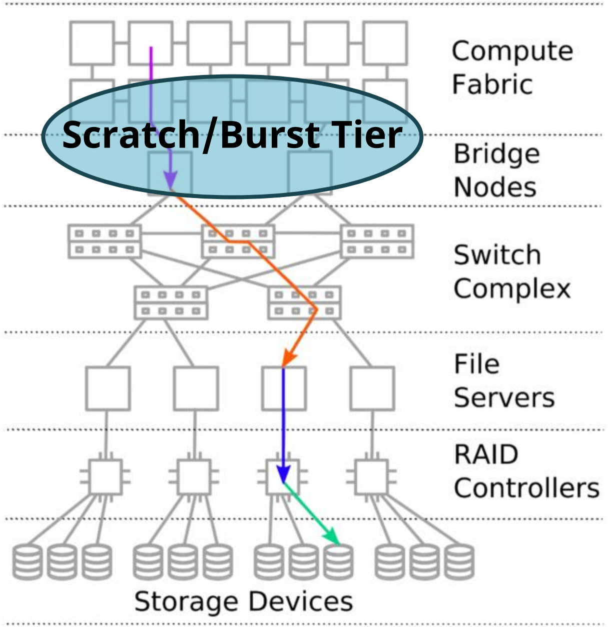
neuwirth@uni-mainz.de

HPC-IODC 2024 Workshop, ISC High Performance, May 2024

Motivation

Heterogeneous and Complex HPC Infrastructures

- HPC infrastructure *too complex*, humans are *overwhelmed*
- Complexity and scope increase the *urgency*
  - *New computational paradigms* (AI/ML apps vs. BSP-style HPC)
  - *New architectural directions* (e.g., IPU, RISC-V, data flow)
  - *Heterogeneity overall*: node architectures, within the system, storage and parallel file system during application design (e.g., ML within HPC applications)
  - *New operations paradigms* (e.g., cloud, container)
  - Simplistic approaches to increasing compute demand result in *unacceptable power costs*
- Difficult for humans to optimally adapt applications to systems and to detect and diagnose vulnerabilities



Carns, P., 2023. *HPC Storage: Adapting to Change*. Keynote at REX-IO’23 Workshop.

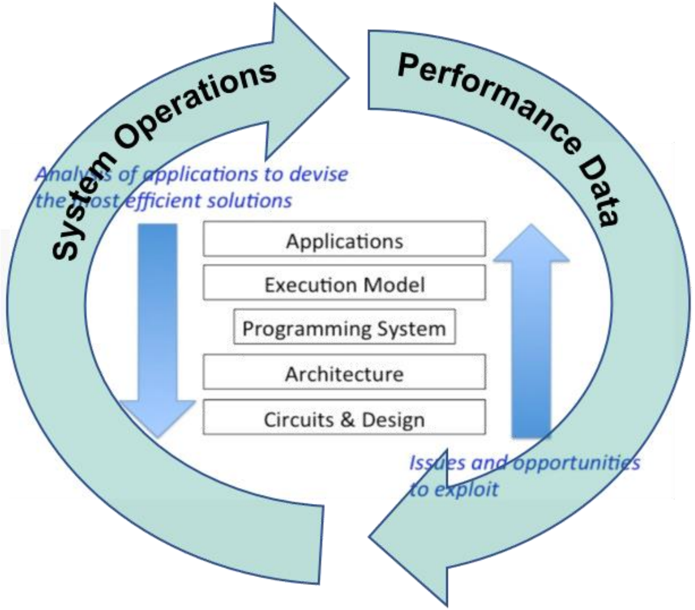
Ciorba, F., 2023. *Revolutionizing HPC Operations and Research*. Keynote at HPCMASPA’23 Workshop.

B. Settlemyer, G. Amvrosiadis, P. Carns and R. Ross, 2021. *It’s Time to Talk About HPC Storage: Perspectives on the Past and Future*, in Computing in Science & Engineering, vol. 23, no. 6, pp. 63-68.

Motivation

Holistic Monitoring and Operational Data Analytics

- Continuous and holistic *monitoring*, *archiving*, and *analysis* of operational and performance data open up interactivity with applications, system software, and hardware through
  - Automated feedback
  - Dynamic analysis of workloads and application demands, architecture and resource state
  - Actionable analytics and adaptive response
- Enable *efficient HPC operations*



Gentile, A., 2021. *Enabling Application and System Data Fusion*. Keynote at MODA'21 Workshop.

Ciorba, F., 2023. *Revolutionizing HPC Operations and Research*. Keynote at HPCMASPA'23 Workshop.

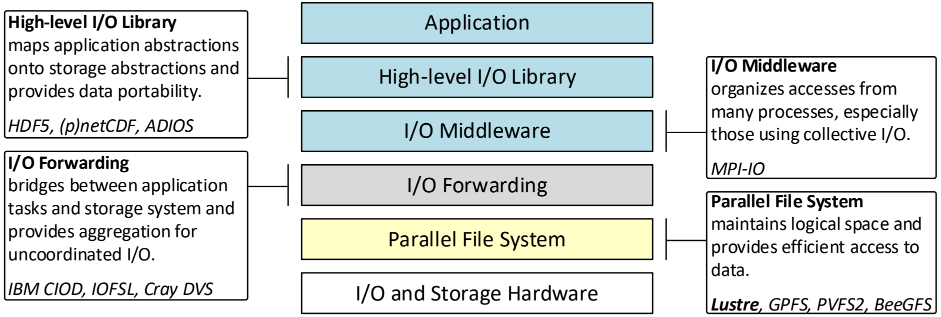
Dagstuhl Seminar 23171, 2023. *Driving HPC Operations With Holistic Monitoring and Operational Data Analytics*. https://www.dagstuhl.de/23171

Tracking the Data Trail

Tracking the Data Trail

Software Architectures for Parallel I/O

- Characterizing and understanding I/O behavior is critical => *increasingly complex I/O stack*
  - More diverse applications, computational frameworks, etc.
  - Emerging hardware and storage paradigms
- Understanding and re-envisioning I/O stands to benefit numerous HPC stakeholders:
  - Application scientists: Improved I/O performance ⇒ decreased time to scientific discovery
  - Admins: Inform decisions related to procuring new systems
  - Researchers: Optimizing storage system and I/O library designs



Application

High-level I/O Library

I/O Middleware

I/O Forwarding

Parallel File System

I/O and Storage Hardware

High-level I/O Library maps application abstractions onto storage abstractions and provides data portability.

HDF5, (p)netCDF, ADIOS

I/O Forwarding bridges between application tasks and storage system and provides aggregation for uncoordinated I/O.

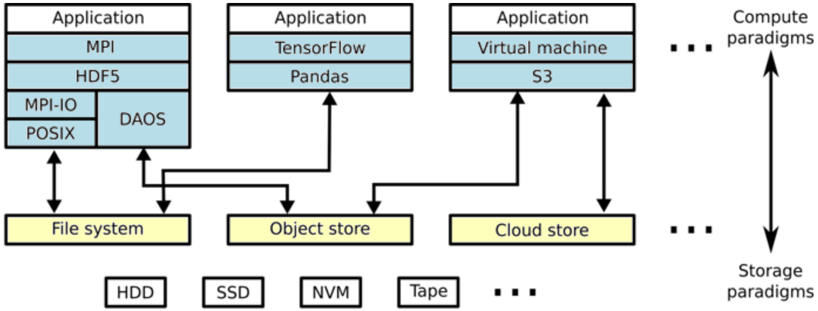
IBM CIOD, IOFSL, Cray DVS

I/O Middleware organizes accesses from many processes, especially those using collective I/O.

MPI-IO

Parallel File System maintains logical space and provides efficient access to data.

***Lustre**, GPFS, PVFS2, BeeGFS*



Tracking the Data Trail

I/O Performance Factors and Metrics

Factors Potentially Affecting Reproducibility of I/O Performance:

Application

- Number of processes
- Request sizes
- Access patterns
- I/O operation
- Data volume

Network

- Message sizes
- Network topology
- Network paths
- Network type

File System

- Type of file system
- Disk types
- Stripe sizes
- File hierarchy
- Shared access

Multiple Tools for I/O Performance Analysis:

- May be a problem when users need to change the tool and want to ensure the measurement continuity and comparability
- There is no easy way to verify metrics consistency between tools

=> *Mango-IO first attempt to provides tools-agnostic metrics calculation*

Liem, Radita, Sebastian Oeste, Jay Lofstead, and Julian Kunkel. *Mango-IO: I/O Metrics Consistency Analysis*. In 2023 IEEE International Conference on Cluster Computing Workshops (CLUSTER Workshops), pp. 18-24. IEEE, 2023.

Tracking the Data Trail
Example: Darshan I/O Characterization Tool

❶ **Blue Waters, Mira, and Theta popular Darshan log sources used for research:**

- https://bluewaters.ncsa.illinois.edu/data-sets
- https://reports.alcf.anl.gov/data/
- ftp://ftp.mcs.anl.gov/pub/darshan/data

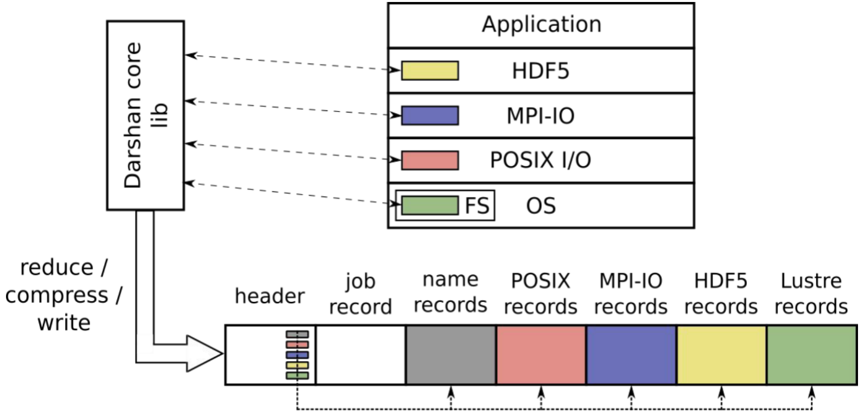
❷ **Some open questions:**

- How relevant are the logs to current systems?
- How do we know the integrity of the logs?

❸ **Community statements:**

- Darshan is one of the first tools to be deactivated in the event of I/O problems.
- Darshan cannot grasp the complexity of state-of-the-art parallel storage systems.

Darshan I/O Characterization Tool



Snyder, S., 2022. *Darshan: Enabling Insights into HPC I/O Behavior*. ECP Community BoF Days.

What are the implications of these questions and observations?

Tracking the Data Trail

Key Questions and Goals of our Project

❶ How does the HPC compute and storage ecosystem look like?

- Overview of modern HPC infrastructures and architectures from tier-1 to tier-3
- Identification of new storage tiers and all possible data paths

❷ Which monitoring infrastructures for I/O and data are deployed globally?

- Identification of the world's most popular monitoring software and toolchains

❸ To what extent are current monitoring infrastructures capable of supporting heterogeneous I/O and storage architectures?

- Identification of shortcomings in widely used monitoring software
- Development of concepts to better support complex parallel I/O and storage systems

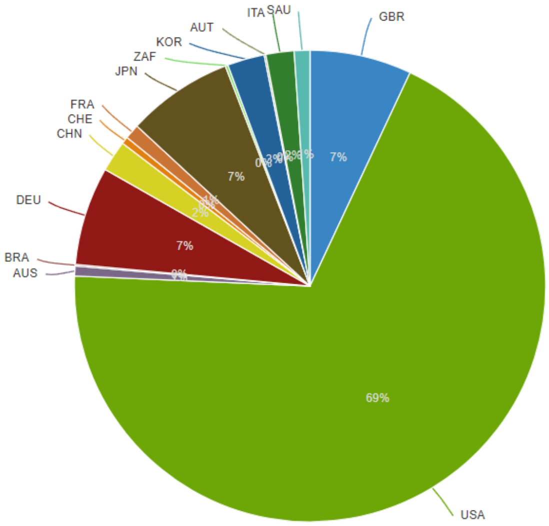
Preliminary Analysis

Preliminary Analysis
Glance at the TOP500

- **Nworld** - Position within the TOP500 ranking
- **Manufacturer** - Manufacturer or vendor
- **Computer** - Type indicated by manufacturer or vendor
- **Installation** Site - Customer
- **Location** - Location and country
- **Year** - Year of installation/last major update
- **Field of Application**
- **#Proc.** - Number of processors (Cores)
- **Rmax** - Maximal LINPACK performance achieved
- **Rpeak** - Theoretical peak performance
- **Nmax** - Problem size for achieving Rmax
- **N1/2** - Problem size for achieving half of Rmax

Preliminary Analysis
Comprehensive Data Center List @ VI4IO

https://www.vi4io.org/



VI4IO Goals:

- Document storage system design
  - Offer long-term storage system design archive, …
    => *Comprehensive Data Center List* (CDCL)
  - …including benchmarks => *IO500 Benchmark and List*
- Build community
- Incubator for IO/Storage related efforts

Comprehensive Data Center List:

- Detailed information about data centers including the top storage systems (file systems, tape libraries)
- Most complete release in 2021

IO500 Data Analysis

IO500 Data Analysis

Goals and Mission of the IO500

Lofstead, Jay. "*Meaningful Measurements? IO500's 5th Year's Search for Meaning*." Invited talk at the High Performance Storage Workshop (HPS) at IPDPS 2021.

Mission:

(1) Provide a competitive list to justify compute time

(2) Gather best practices for different storage system designs

(3) Document various storage systems

(4) Friendly cooperation and competition

(5) Provide a way to justify using compute time to run the benchmarks

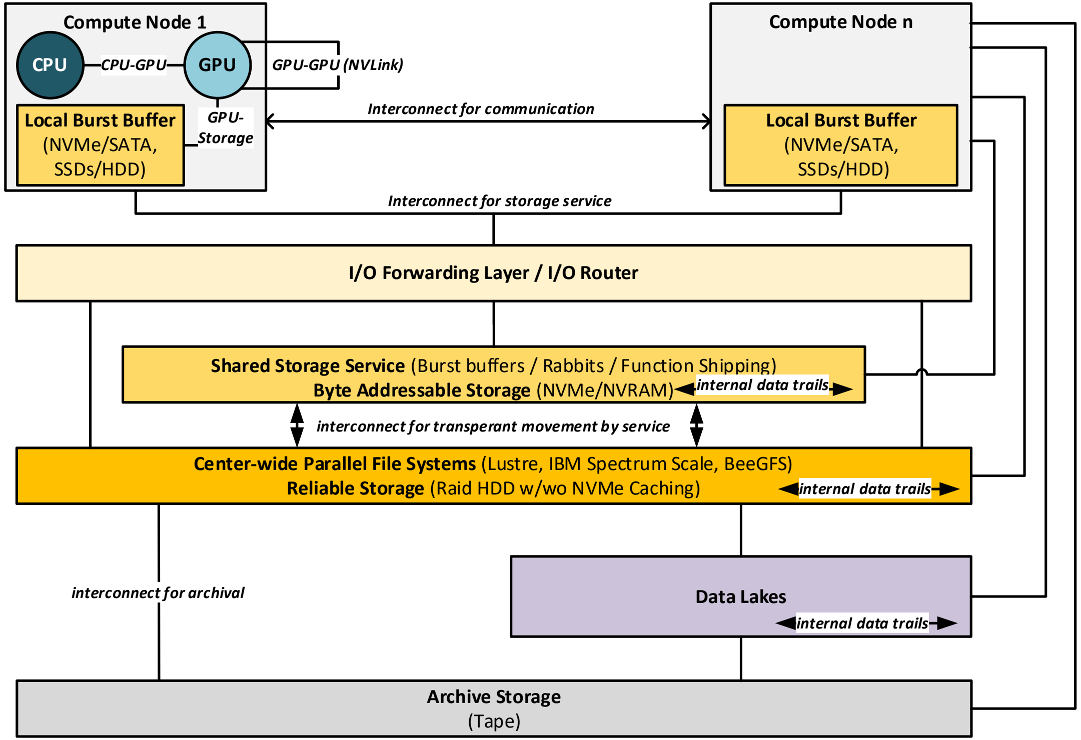
=> ***Use accepted benchmarks with generally accepted configurations (for the hard setup)***

Can we use the IO500 for our data trail project?

IO500 Data Analysis

Project: Tracking the Data Trail in HPC Systems

Joint work-in-progress with Hariharan Devarajan (LLNL) and Jay Lofstead (SNL)



Compute Node 1

CPU —*CPU-GPU*— GPU

GPU-GPU (NVLink)

GPU-Storage

Local Burst Buffer (NVMe/SATA, SSDs/HDD)

Interconnect for communication

Compute Node n

Local Burst Buffer (NVMe/SATA, SSDs/HDD)

Interconnect for storage service

I/O Forwarding Layer / I/O Router

Shared Storage Service (Burst buffers / Rabbits / Function Shipping)
Byte Addressable Storage (NVMe/NVRAM) *internal data trails*

interconnect for transperant movement by service

Center-wide Parallel File Systems (Lustre, IBM Spectrum Scale, BeeGFS)
Reliable Storage (Raid HDD w/wo NVMe Caching) *internal data trails*

interconnect for archival

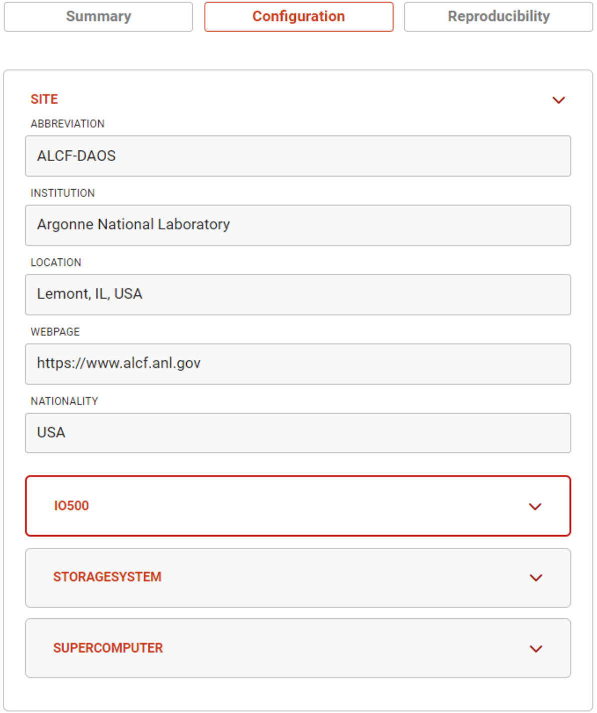
Data Lakes
internal data trails

Archive Storage
(Tape)

IO500 Data Analysis

System Insights: Configuration & Benchmark Results

- IO500 test conditions:
  - Number of nodes, #processes per node, exclusive
- File system information:
  - Type, vendor, software, mount options, capacity, etc.
- Distinguishes between client-side storage and data servers:
  - STORAGESYSTEM
  - SUPERCOMPUTER
- Effort to collect different interconnects for storage (metadata and data) and clients => *not mandatory*



SITE

ABBREVIATION: ALCF-DAOS

INSTITUTION: Argonne National Laboratory

LOCATION: Lemont, IL, USA

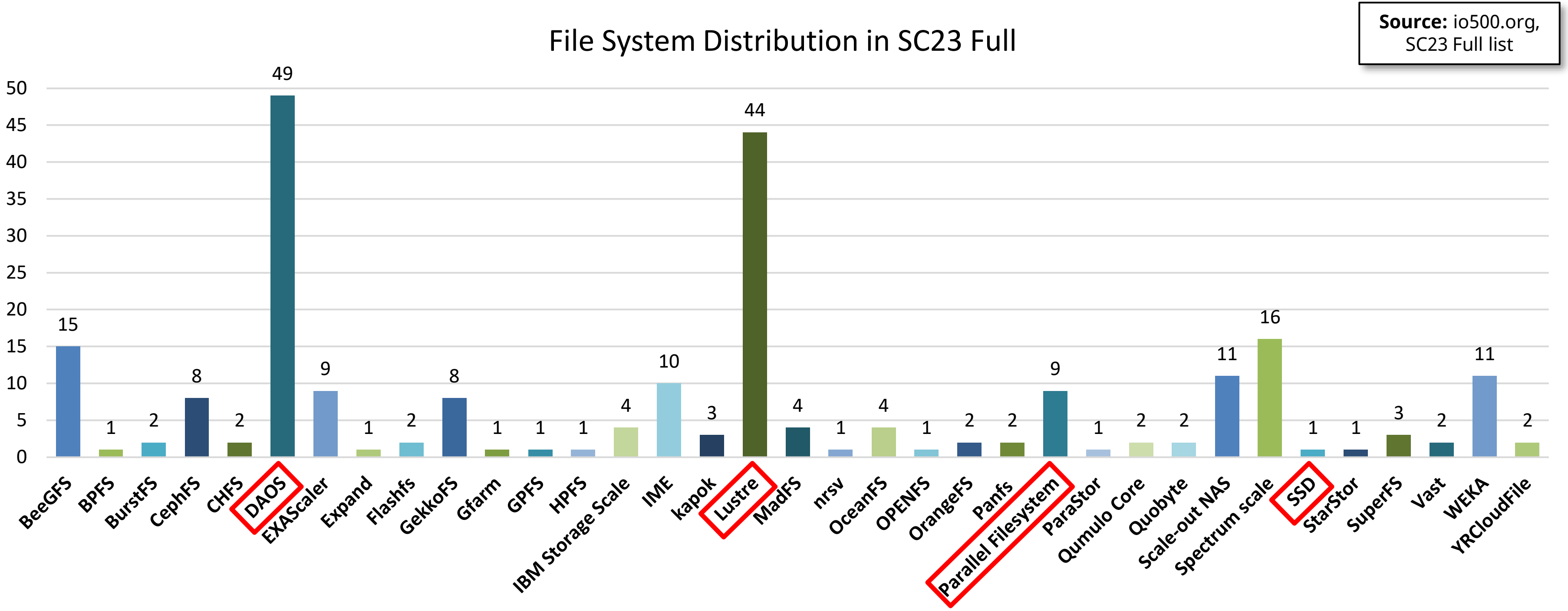
WEBPAGE: https://www.alcf.anl.gov

NATIONALITY: USA

IO500 Data Analysis

System Insights: File System Statistics

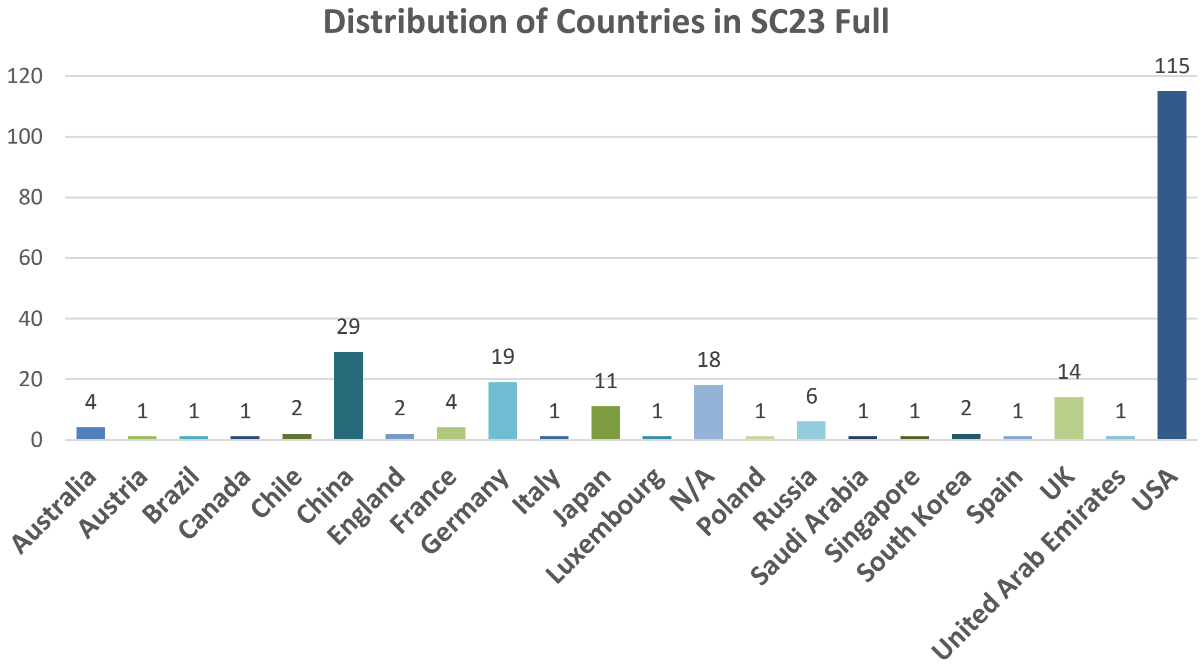
Source: io500.org, SC23 Full list



IO500 Data Analysis

System Insights: Distribution of Submissions

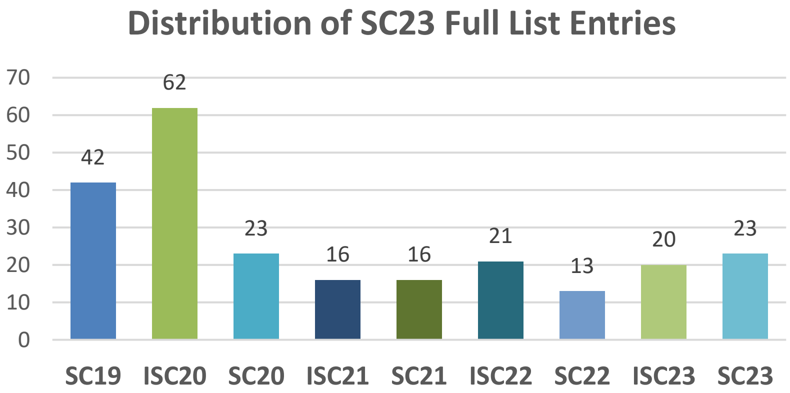
- 2 TOP10 systems
- Less than 10 TOP500 systems in total
- US submissions dominating
- Asia and Europe have contributed equally
- No submissions from Africa
- 91 unique submitting sites
  - Mostly industry testbeds
  - Some university clusters
  - Almost no major HPC sites
- 236 entries in SC23 Full



IO500 Data Analysis

System Insights: Additional Statistics

- 159 submissions provided information_ds_network
  - 69x Ethernet, 62x Infiniband, 28x Omnipath
- 43 submisisons provided information_client_interconnect_type
  - 1x Aries, 15x Ethernet, 26 Infiniband, 1x Omnipath
- Submissions:



IO500 Data Analysis

Surprising Findings

- CSV file does not contain all the information available on the IO500 website
  - E.g., Burst buffers and persistent storage cannot be distinguished in the file
  - Explanation for different columns in the CSV file not available on the website
  - Information about the site such as location, nationality, etc. omitted
  - Type of file system (parallel, object store, …) only available on the website
- Only minimal information about file systems is collected
  - Link to documentation or further information would be helpful
- A lot of information is not mandatory
- No information about monitoring infrastructure collected (not even as optional information)

IO500 Data Analysis

[*Our*] *Wish List for IO500 and CDCL*

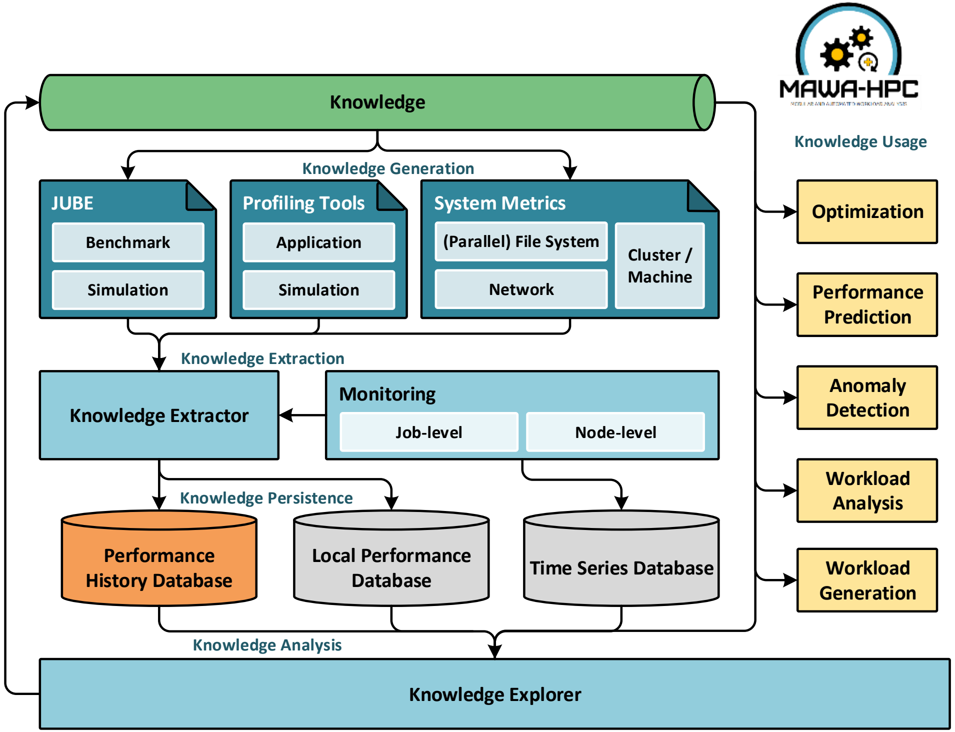
- Make CDCL submission mandatory for new list entries and link them?
- Document evolution of storage / HPC systems
  - Include information about what changed from previous submission (combined with CDCL?)
- Information about type of system, storage tiers, …, including:
  - Production systems versus testbeds?
  - HPC cluster versus cloud system?
  - Local versus shared burst buffer?
  - Vendor system, leadership system, tier-1/tier-2/tier-3 system?
  - Monitoring system deployed? If so, which one?
- Provide sample submission form to understand what is collected
  => *and include all information in the CSV file for further data analysis*
- Attract more leadership and production systems of the TOP500 to submit

Outlook and Vision

Outlook and Vision

Holistic Performance Engineering and Analysis

- ***Idea:*** Design and implement standardized and tool-independent approach for HPC workload and application analysis
- Support and integration of various community tools, increasing the compatibility and coverage of different use cases
- Intuitive performance modeling and visualization so that users without prior knowledge can understand the results
- ***Goal:*** Establish a ***performance history database*** to categorize systems, workload behaviors, and characteristic patterns for different science domains



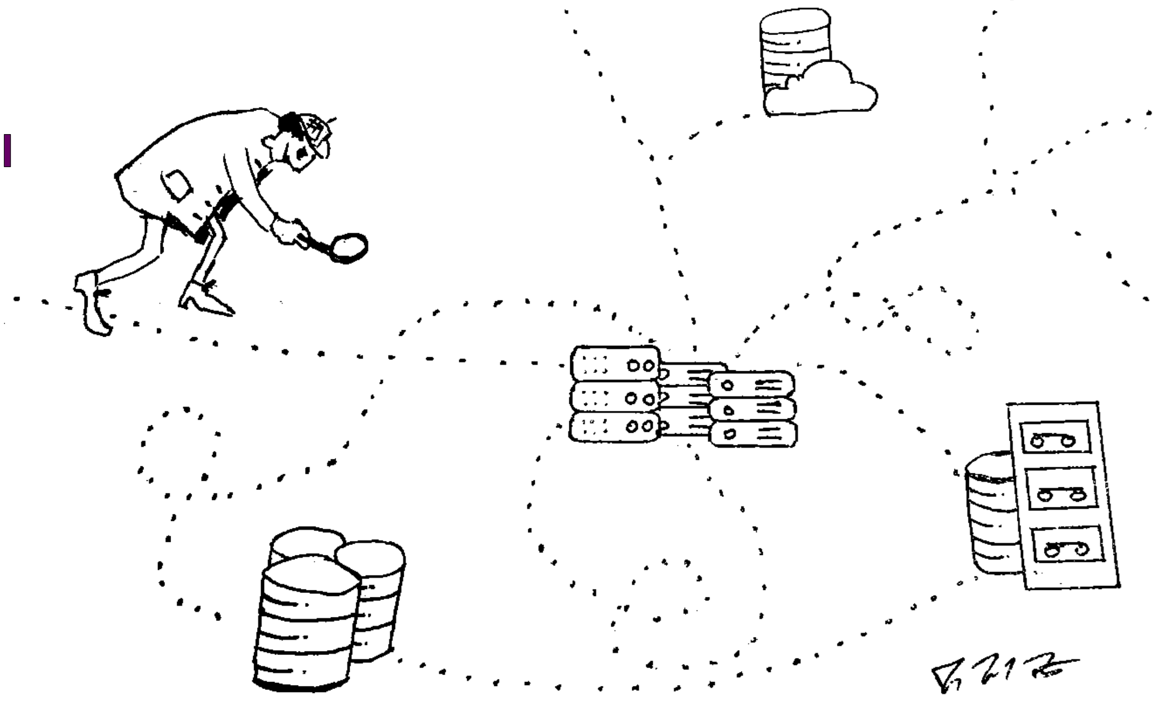
Outlook and Vision

Toward traceroute for Data Trails?

How can the differences between modern monitoring infrastructures and the actual data trails be reconciled?

↓

What if there would be *traceroute* for parallel storage and I/O architectures?



© Joey White-Swift

Outlook and Vision

Data Trails Community Survey

- Please help us with complementing the information collected by IO500 and CDCL!
- Which systems should take part?
  - **Tier-1:** Top-tier supercomputers with the highest performance for large-scale, national or international projects.
  - **Tier-2:** Mid-level systems for regional or institutional use, balancing cost and performance.
  - **Tier-3:** Entry-level clusters for smaller scale, departmental use, and less intensive computational tasks.

=> Basically everyone! We want a diverse and global mix! ☺

>> https://forms.gle/uPUQLYaciT41q7of6 <<

SCAN ME

Thank you for your Attention!