Why is my I/O Slow?

Rethinking Storage Systems for Explainability



July 13th, 2026

Hariharan Devarajan, Computer Scientist
Center for Applied Scientific Computing (CASC)

Lawrence Livermore National Laboratory
LLNL-CFPRES-2021258
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.
National Nuclear Security Administration

Why Is My I/O Slow?

To answer, we must first understand **who** is asking and **why**.



Different People. Different Questions.

SCIENTISTS & DATA ANALYSTS

They have an **I/O budget**

↓

Their real question:

"How can I stay within my I/O budget and actually do science?"

SYSTEMS PEOPLE

We need to **explain, diagnose, and optimize**

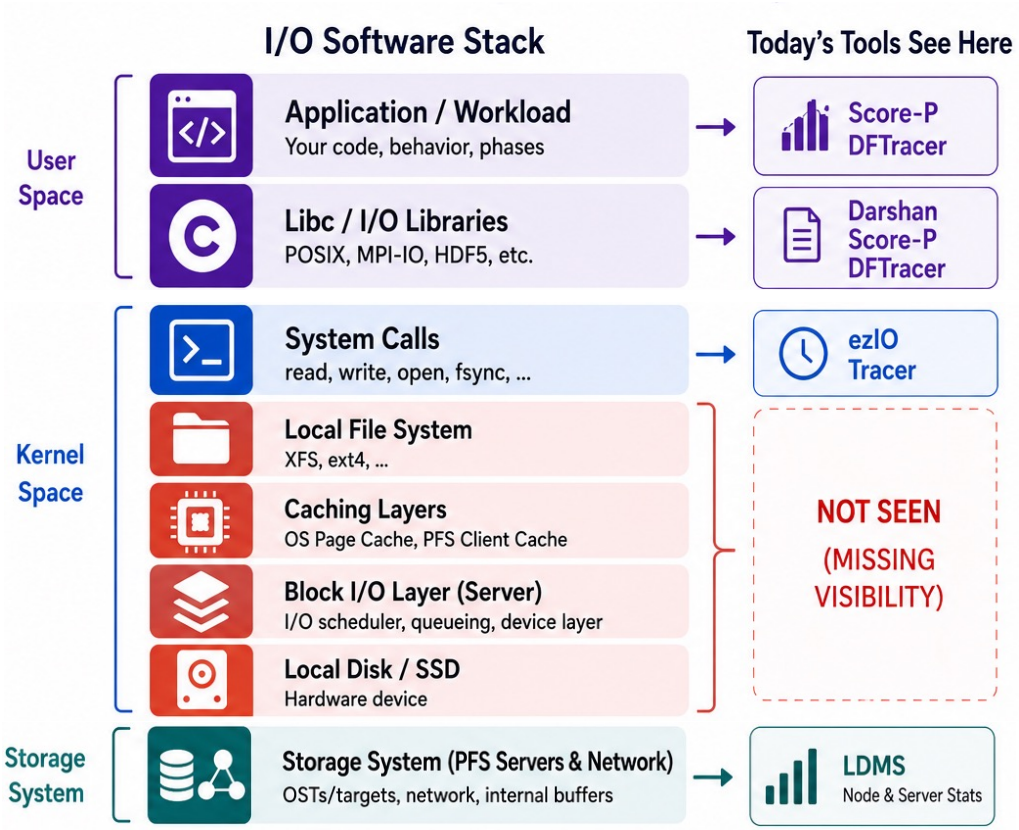
↓

Key Challenge:

Complex storage stack with many layers

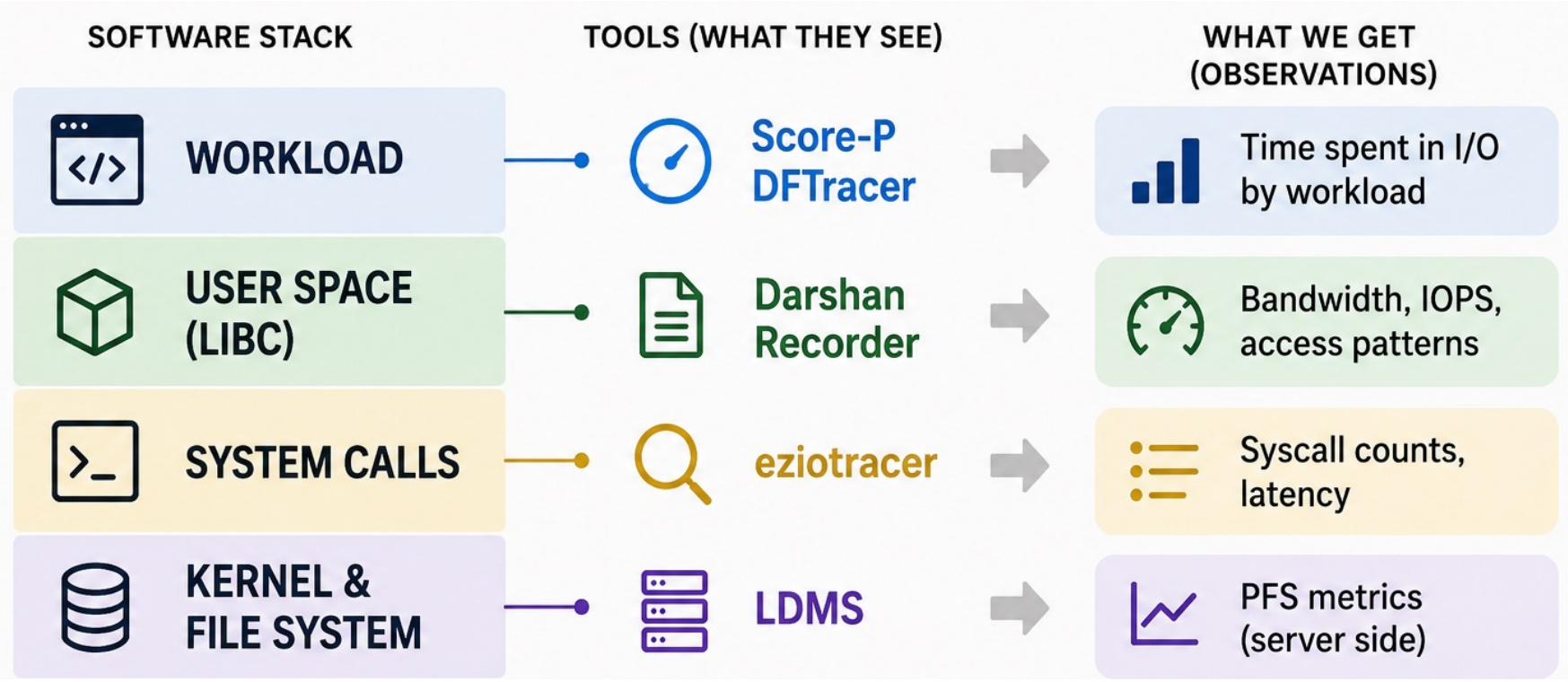
Can we really Answer, “Why my I/O is slow?”

What do tools actually see today?



What today's tools tell us?

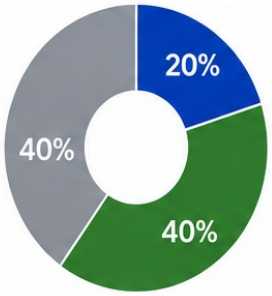
They show what happens, not whether its good or bad.



Let us look at an example.
Are these behaviors Good or Bad?

WHAT WE OBSERVE

1024 parallel reads on a Lustre file system



- 20% of reads 4–8 GB/s per process
- 40% of reads 1–2 GB/s per process
- 40% of reads 500 MB/s – 1 GB/s per process

Wide and variable performance range.

No clarity.

LITERATURE WILL GUESS

"Reads > 1 GB/s are probably served by RAM.

Rest by the parallel file system."



RAM

> 1 GB/s



Parallel File System

≤ 1 GB/s

WHAT IS THE EVIDENCE?

No direct evidence of where reads are served from.

RAM performance can be 1–8 GB/s — too wide to be conclusive.

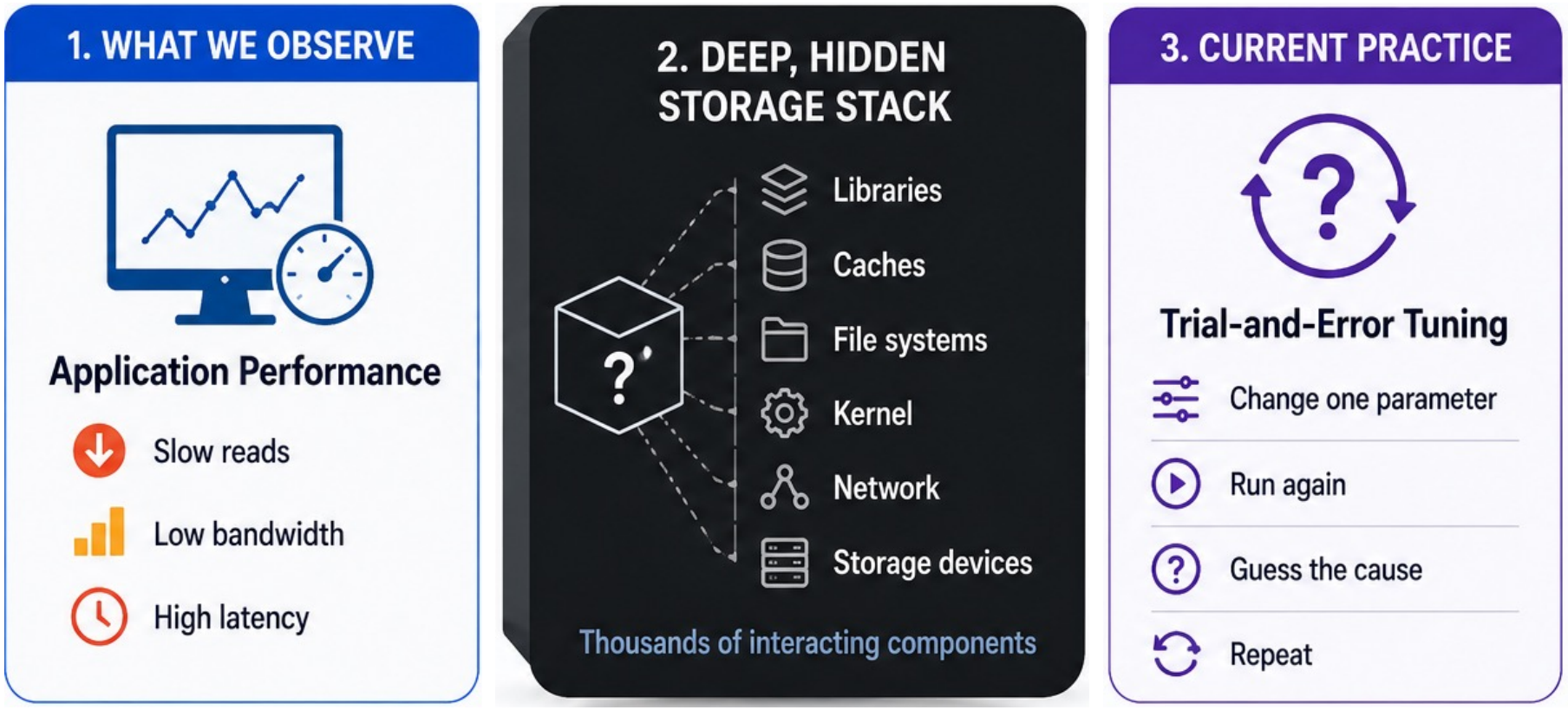
So... good or bad?

We don't know.

We need data to know the truth.

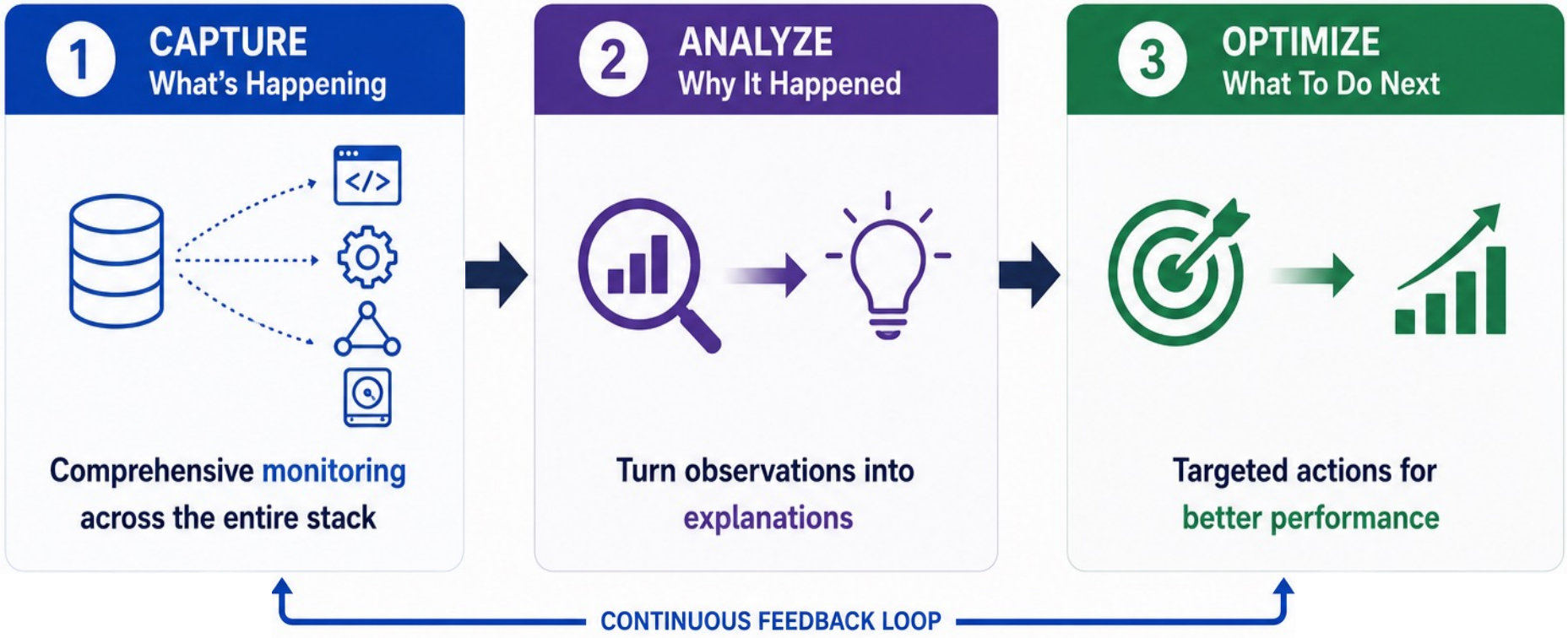
We need to look beyond Observations

Observations show symptoms, not causes.



Hariharan Devarajan, and Kathryn Mohror. "TailorFS: An Adaptive file system to support dynamic I/O requirements of HPC workloads." In 2024 IEEE 36th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), pp. 81-92. IEEE, 2024.

What do we need for Explainable Systems?



Outline

The roadmap for explainable storage systems.



1

Holistic Monitoring for Explanations

Capture complete visibility across the storage stack to understand what's happening.



2

Scalable Analysis for Deep Insights

Turn massive traces into causal insights that uncover the real root causes.



3

Intelligent Optimization Engine

Translate insights into targeted actions for measurable and portable impact.



4

The Future of Storage Systems

Emerging trends and innovations shaping self-explaining, self-adjusting storage.



5

Conclusions & Call to Action

Key takeaways and how we can build this future—together.

Outline

The roadmap for explainable storage systems.



1

Holistic Monitoring for Explanations

Capture complete visibility across the storage stack to understand what's happening.



2

Scalable Analysis for Deep Insights

Turn massive traces into causal insights that uncover the real root causes.



3

Intelligent Optimization Engine

Translate insights into targeted actions for measurable and portable impact.



4

The Future of Storage Systems

Emerging trends and innovations shaping self-explaining, self-adjusting storage.



5

Conclusions & Call to Action

Key takeaways and how we can build this future—together.

Monitoring Solution: eBPF technology

Capture what matters, Securely and Efficiently

1 EXPLORE NEW TOOLS
Access More Data

Deep visibility across user space and kernel space.

- APIs
- I/O Libraries
- MPI
- App Transforms
- ...
- OS Page Cache
- File Systems
- Block Layer & Drivers
- Network Stack
- ...

From high-level APIs to low-level devices—observe internal transformations, not just calls.

2 BEYOND USER SPACE
Special Privileges

Secure. Efficient. Production-ready.

eBPF

- Secure by design
- Safe observability
- No change to user logic
- Respect permissions
- Low overhead

Deep visibility without compromising security or performance.

3 FIND THE RIGHT COUNTERS
Explain Performance

The right counters. The right why.

- Discover relevant counters across the I/O stack
- Focus on causality, not collection
- Literature per stack to know what matters

I/O IS MULTI-DIMENSIONAL

- Memory
- Compute
- Comm
- Data Ops
- I/O & Devices

Understand the full pipeline to explain performance and drive the right optimization.

DataCrumbs

An eBPF-based tool for full-stack monitoring

WHAT WE CAPTURE

Applications & Runtimes
I/O patterns, context, phases

Libraries & Middleware
HDF5, MPI-IO, filters, aggregations

File Systems & Kernel
Page cache, block layer, drivers, network stack

Storage Devices
NVMe, SSD, disk stats, latencies

WHAT PROBLEMS WE SOLVE

Answer the "Why"
Go beyond API-level observations

Full-Stack Visibility
See the entire I/O path, not just the API boundary

Explain Performance Variability
Understand cache hits vs. misses, contention, queueing, device behavior

Find True Bottlenecks
Attribute time and cost to the right layer and component

THE BIG PICTURE



- Capture the right data
- Connect the dots
- Explain I/O performance

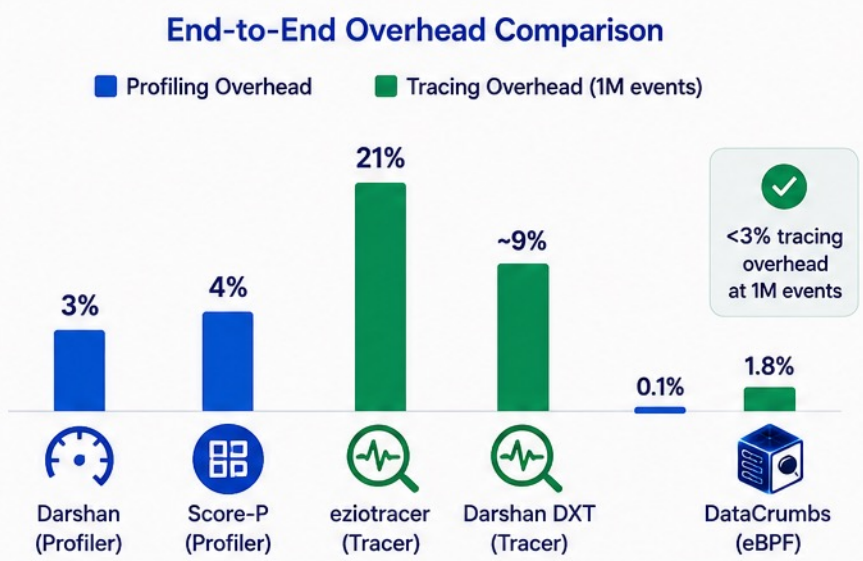
Understand, optimize, and accelerate discovery.

Sarah Neuwirth, Hariharan Devarajan, Chen Wang, and Jay Lofstead. "XIO: Toward eXplainable I/O for HPC Systems." In Proceedings of the 37th International Conference on Scalable Scientific Data Management, pp. 1-6. 2025.

DataCrumbs Results

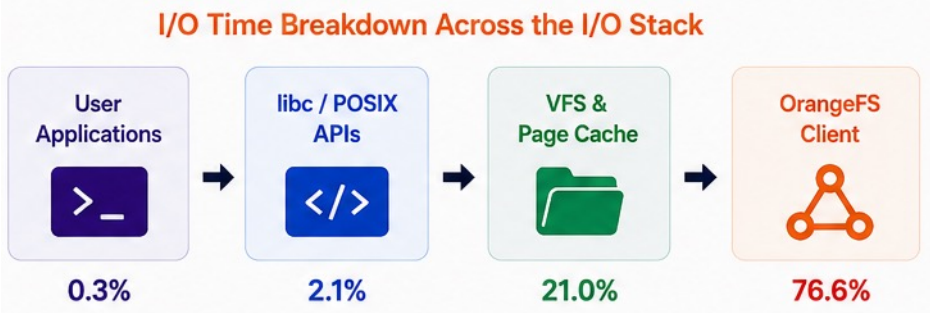
Deep visibility with minimal overhead

1. LOW OVERHEAD



Up to 9× lower tracing overhead vs. Darshan DXT while capturing deep, cross-layer data.

2. REAL WORKLOAD RESULT: MONTAGE ON ORANGEFS



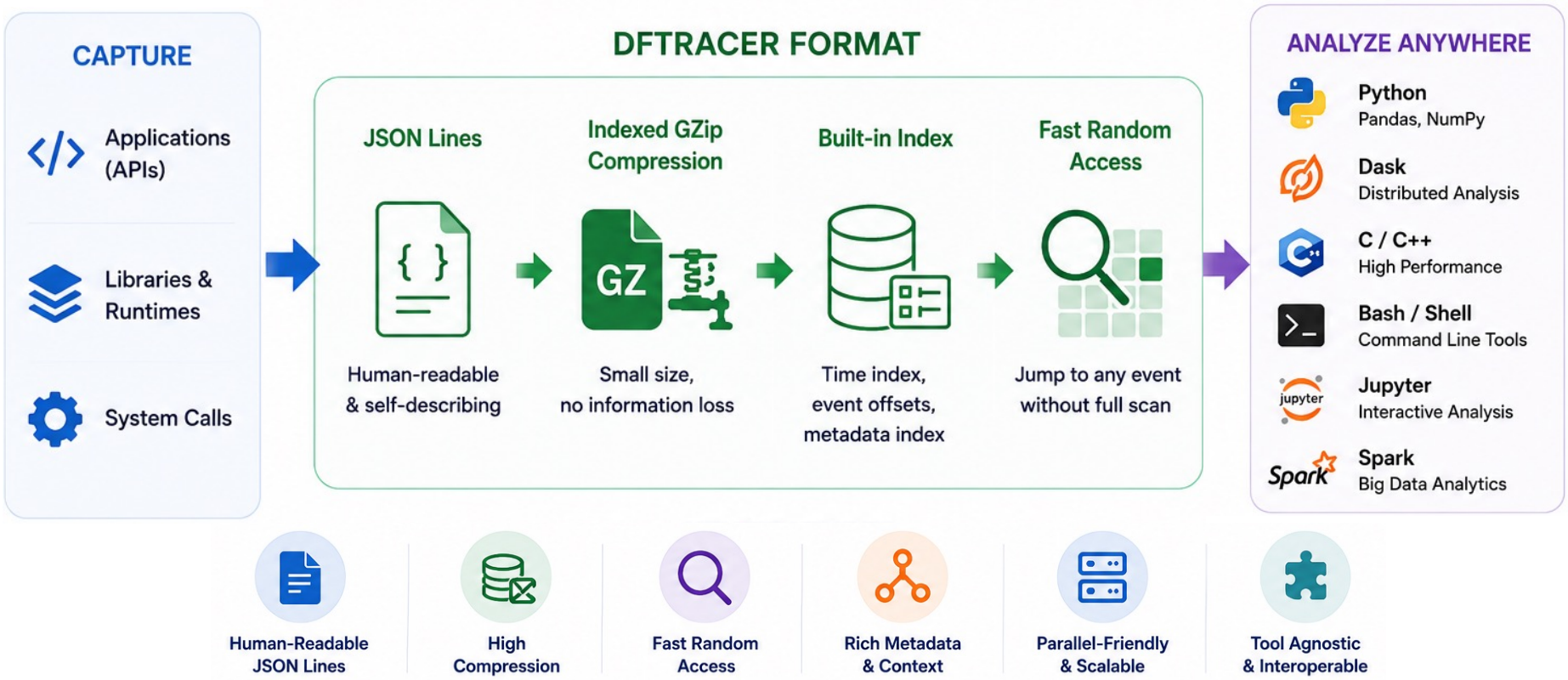
OrangeFS client layer dominates (76.6%)
DataCrumbs pinpoints where time is spent.

Key Insights Revealed by DataCrumbs

- Metadata ops & directory lookups add hidden latency.
- Small I/O sizes & file distribution patterns amplify overheads.
- Actionable insights guide optimizations at the right layer.

DFTracer Format

Built for Efficient Storage and Analysis

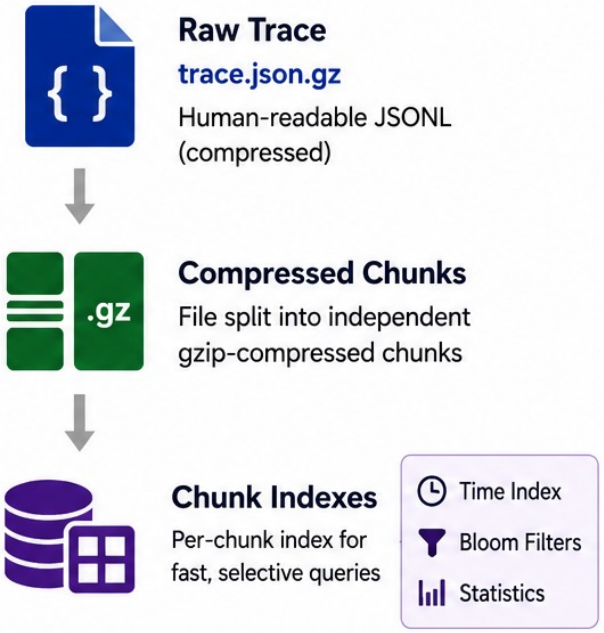


Hariharan Devarajan, Loïc Pottier, Kaushik Velusamy, Huihuo Zheng, Izzet Yildirim, Olga Kogiou, Weikuan Yu et al. "Dftracer: An analysis-friendly data flow tracer for ai-driven workflows." In SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1-24. IEEE, 2024.

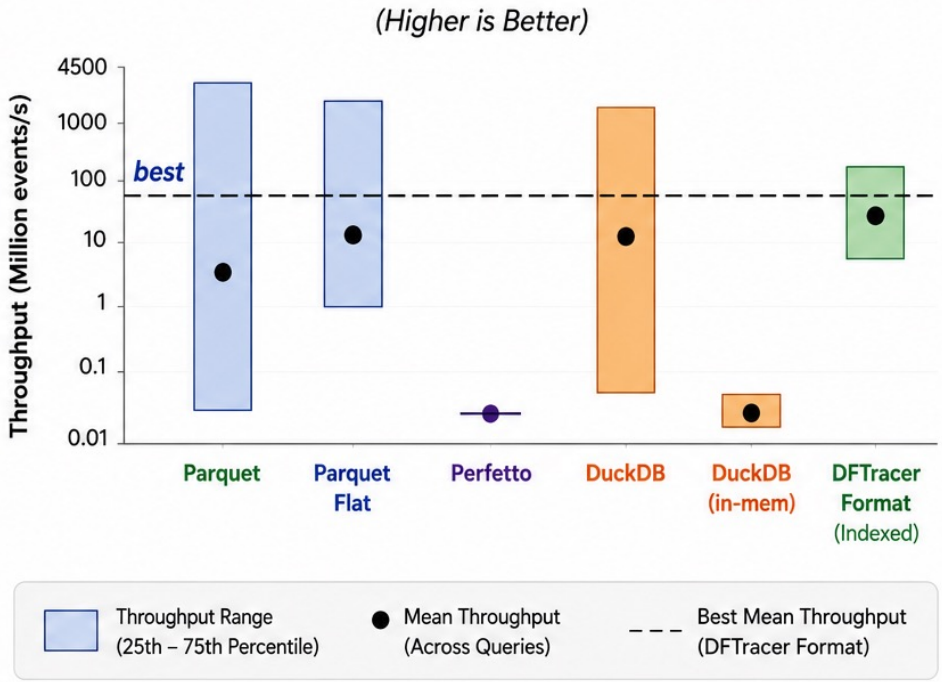
DFTracer Format Query Capabilities

Fast and Selective analysis

How DFTracer Format is Organized



Query Performance: Cross-query Throughput Spread



Ray A. O. Sinurat, William Nixon, Haryadi S. Gunawi, Nikoli Dryden, and Hariharan Devarajan. 2026. HORATIO: Bridging Management and Analysis of Traces at Scale. Paper accepted to the 38th International Conference on Scalable Scientific Data Management (SSDBM 2026), San Diego, CA, August 11–13, 2026.

Outline

The roadmap for explainable storage systems.



1

Holistic Monitoring for Explanations

Capture complete visibility across the storage stack to understand what's happening.



2

Scalable Analysis for Deep Insights

Turn massive traces into causal insights that uncover the real root causes.



3

Intelligent Optimization Engine

Translate insights into targeted actions for measurable and portable impact.



4

The Future of Storage Systems

Emerging trends and innovations shaping self-explaining, self-adjusting storage.



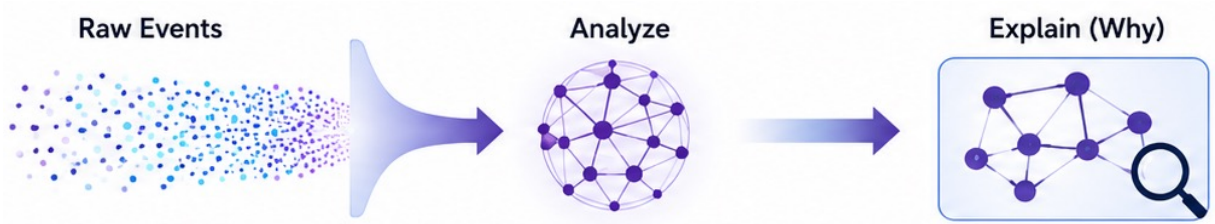
5

Conclusions & Call to Action

Key takeaways and how we can build this future—together.

Analysis Pipeline for I/O Explainability

From massive data to actionable explanations.



1 HANDLE DATA EXPLOSION

Capture at Massive Scale

Efficient Analysis Infrastructure

Find What Matters

Goal: Analyze billions of events efficiently without losing fidelity.

2 DATA-DRIVEN CAUSAL MODELING

Go beyond correlation

Discover what causes what

Understand impact & hidden factors

Outcome: Know what impacts I/O performance and what we're missing.

3 HUMAN-IN-THE-LOOP AI REASONING

AI Reasons Over Data

Human in the Loop

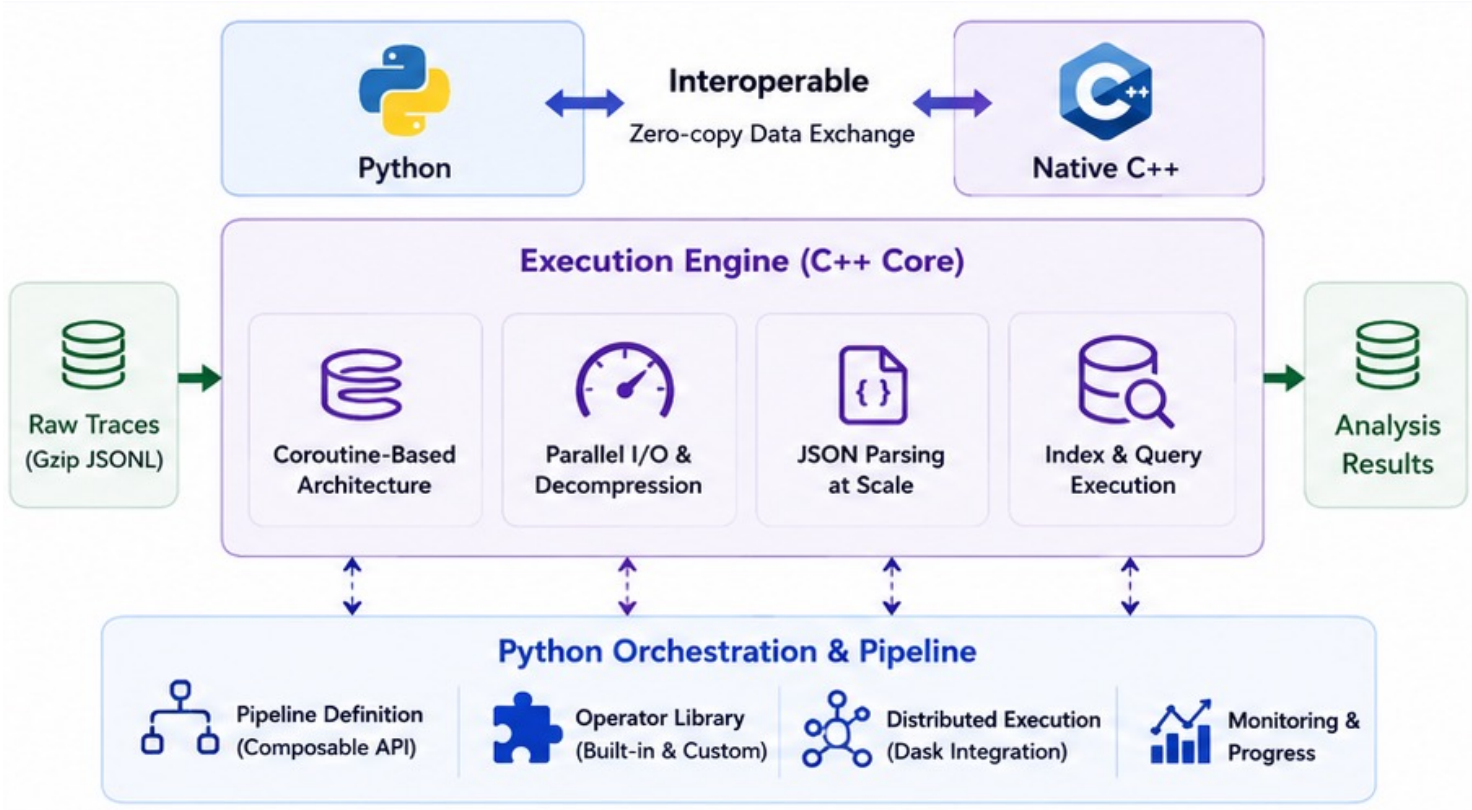
Insights to Actions

Optimize with Confidence

Outcome: Human expertise + AI reasoning to understand and act with confidence.

HORATIO Execution Engine and Analysis Pipeline.

High Performance, Interoperable, and Extensible





Ray Sinurat



Haryadi Gunawi

Ray A. O. Sinurat, William Nixon, Haryadi S. Gunawi, Nikoli Dryden, and Hariharan Devarajan. 2026. HORATIO: Bridging Management and Analysis of Traces at Scale. Paper accepted to the 38th International Conference on Scalable Scientific Data Management (SSDBM 2026), San Diego, CA, August 11–13, 2026.

HORATIO provides speedup compared to Dask

Complete where **Pure Dask** cannot finish

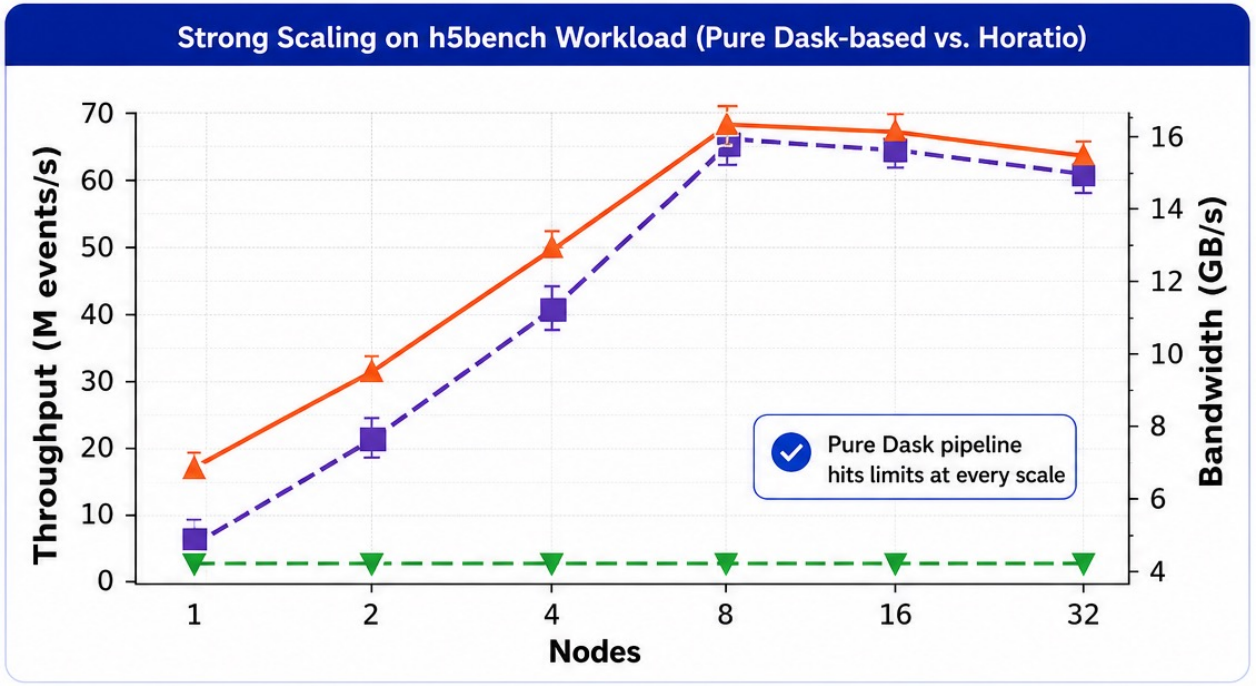
Up to **16×** more throughput

Strong scaling across nodes

Horatio (Thpt.)
Throughput (M events/s)

Horatio (BW.)
Bandwidth (GB/s)

Pure Dask Pipeline
Throughput (M events/s)



Up to **16×** higher throughput

80–83× faster analysis engine

Scales to **32 nodes** (MPI/C++)

Completes where Pure Dask hits OOM

Ray A. O. Sinurat, William Nixon, Haryadi S. Gunawi, Nikoli Dryden, and Hariharan Devarajan. 2026. HORATIO: Bridging Management and Analysis of Traces at Scale. Paper accepted to the 38th International Conference on Scalable Scientific Data Management (SSDBM 2026), San Diego, CA, August 11–13, 2026.

Outline

The roadmap for explainable storage systems.



1

Holistic Monitoring for Explanations

Capture complete visibility across the storage stack to understand what's happening.



2

Scalable Analysis for Deep Insights

Turn massive traces into causal insights that uncover the real root causes.



3

Intelligent Optimization Engine

Translate insights into targeted actions for measurable and portable impact.



4

The Future of Storage Systems

Emerging trends and innovations shaping self-explaining, self-adjusting storage.



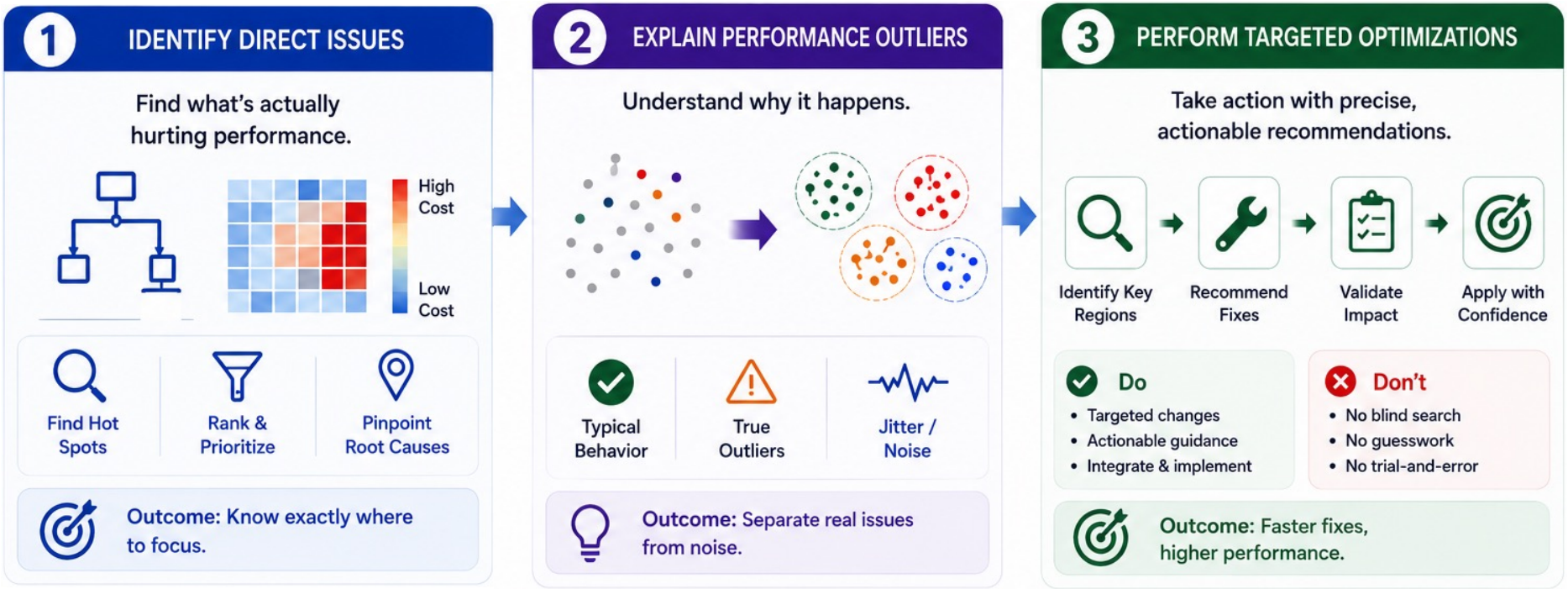
5

Conclusions & Call to Action

Key takeaways and how we can build this future—together.

Optimization Strategies

From Trail-and-Error to Targeted Optimizations



ClarIOn

From Deep I/O Observability to Actionable, Targeted Optimizations

1 DISCOVER & UNDERSTAND ROOT CAUSES

Deep I/O Data (DataCrumbs) → ClarIOn Analysis → Actionable Insight → Better Performance

- Find Root Causes
- Isolate True Issues
- Compare Execution Paths
- Recommend Targeted Fixes

2 DELIVER IMPACT

- **Clear Explanations**
  Know exactly why performance is slow
- **Faster Root-Cause Analysis**
  Go from “slow” to “here is the cause”
- **Smarter Optimizations**
  Fix the right issue, on the right layer
- **Higher Performance**
  Speedups up to orders of magnitude
- **Portable & Reusable**
  Insights transfer across workloads
- **Empower Developers & Users**
  Make deep I/O expertise accessible



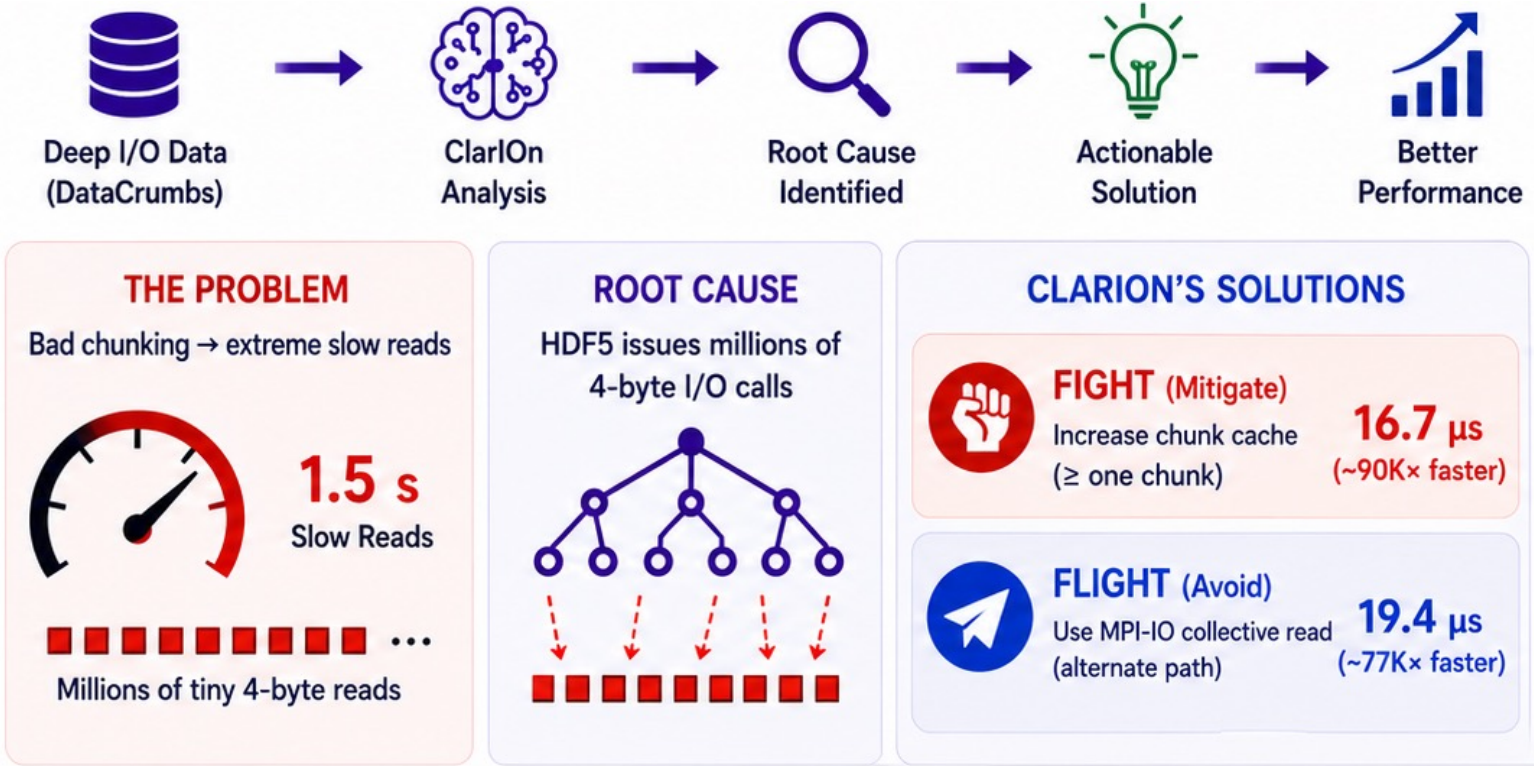
Arta Salimiparsa



Suren Byna

ClarIOn Use Case: HDF5 Chunking

Explaining anomalous behavior for highlighting the correct fix.



Outline

The roadmap for explainable storage systems.



1

Holistic Monitoring for Explanations

Capture complete visibility across the storage stack to understand what's happening.



2

Scalable Analysis for Deep Insights

Turn massive traces into causal insights that uncover the real root causes.



3

Intelligent Optimization Engine

Translate insights into targeted actions for measurable and portable impact.



4

The Future of Storage Systems

Emerging trends and innovations shaping self-explaining, self-adjusting storage.

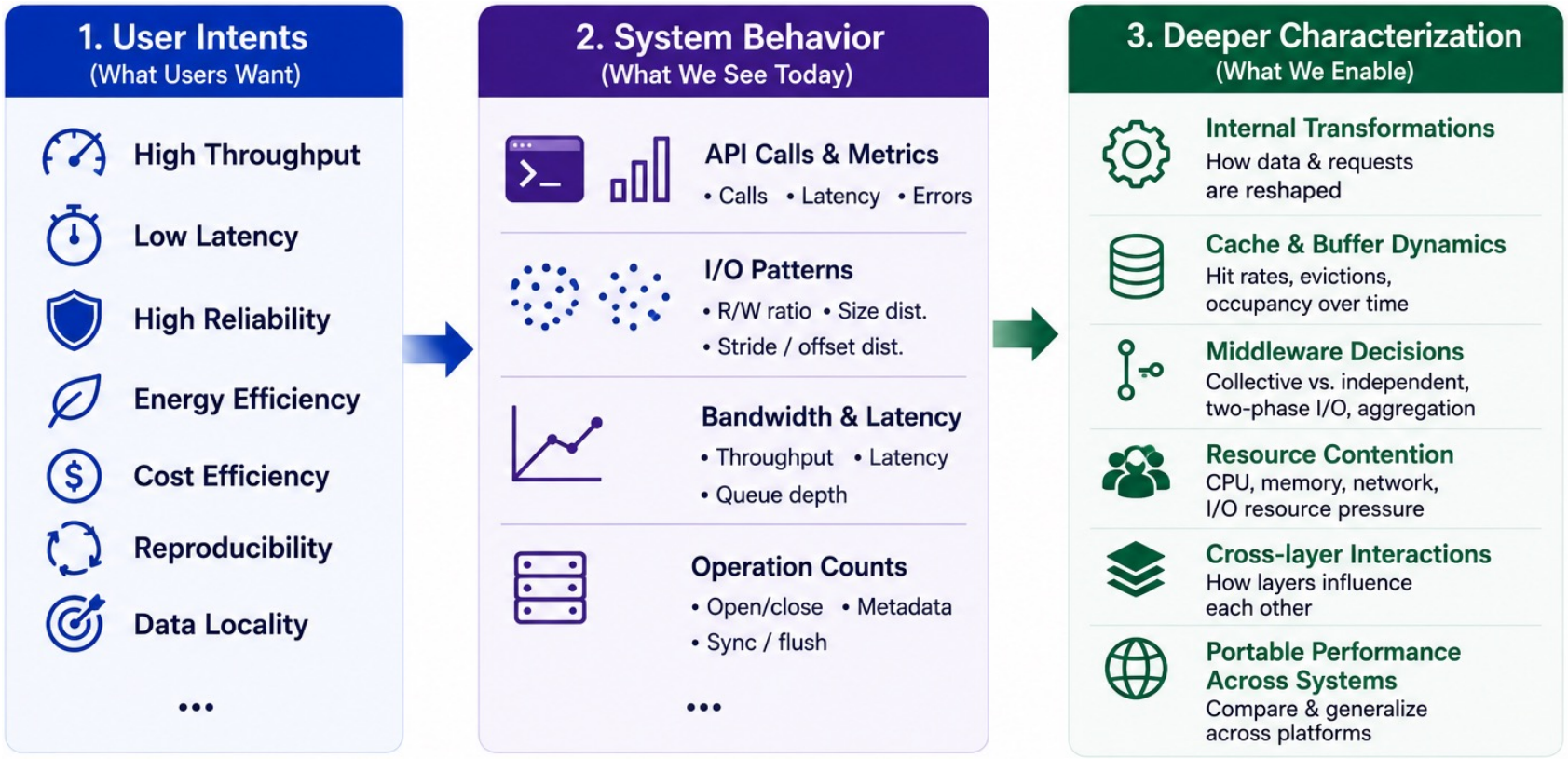


5

Conclusions & Call to Action

Key takeaways and how we can build this future—together.

Better translation from user's storage intents to system behaviors



Hariharan Devarajan and Kathryn Mohror. "Mimir: Extending I/O interfaces to express user intent for complex workloads in HPC." In 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 178-188. IEEE, 2023.

More Fluid, Configurable Storage Systems



Understanding the stack unlocks powerful configurations and helps us use them better.

Deeper Visibility Enables



Understand internal algorithms and policies



Quantify impact of each configuration choice



Discover ineffective or hidden interactions



Enable safe, portable defaults with escape hatches

Hariharan Devarajan. Towards a Self-Programmable Storage Solution in Extreme-Scale Environments. Illinois Institute of Technology, 2021.

Invention of Self-Adjusting Middleware

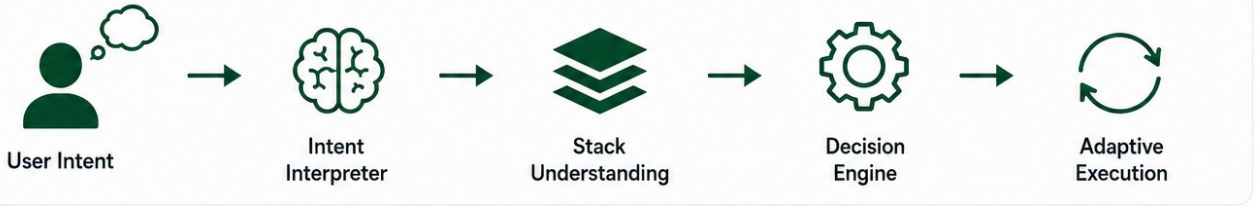
With visibility and configurability, middleware closes the loop and makes optimal decisions for you—dynamically.

What Self-Adjusting Middleware Can Do

- Map intents to the best configuration and data placement
- Adapt to changing system state and workload phases
- Balance trade-offs (throughput, latency, energy, cost, reliability)
- Reconfigure on the fly with minimal disruption
- Learn from outcomes and improve over time

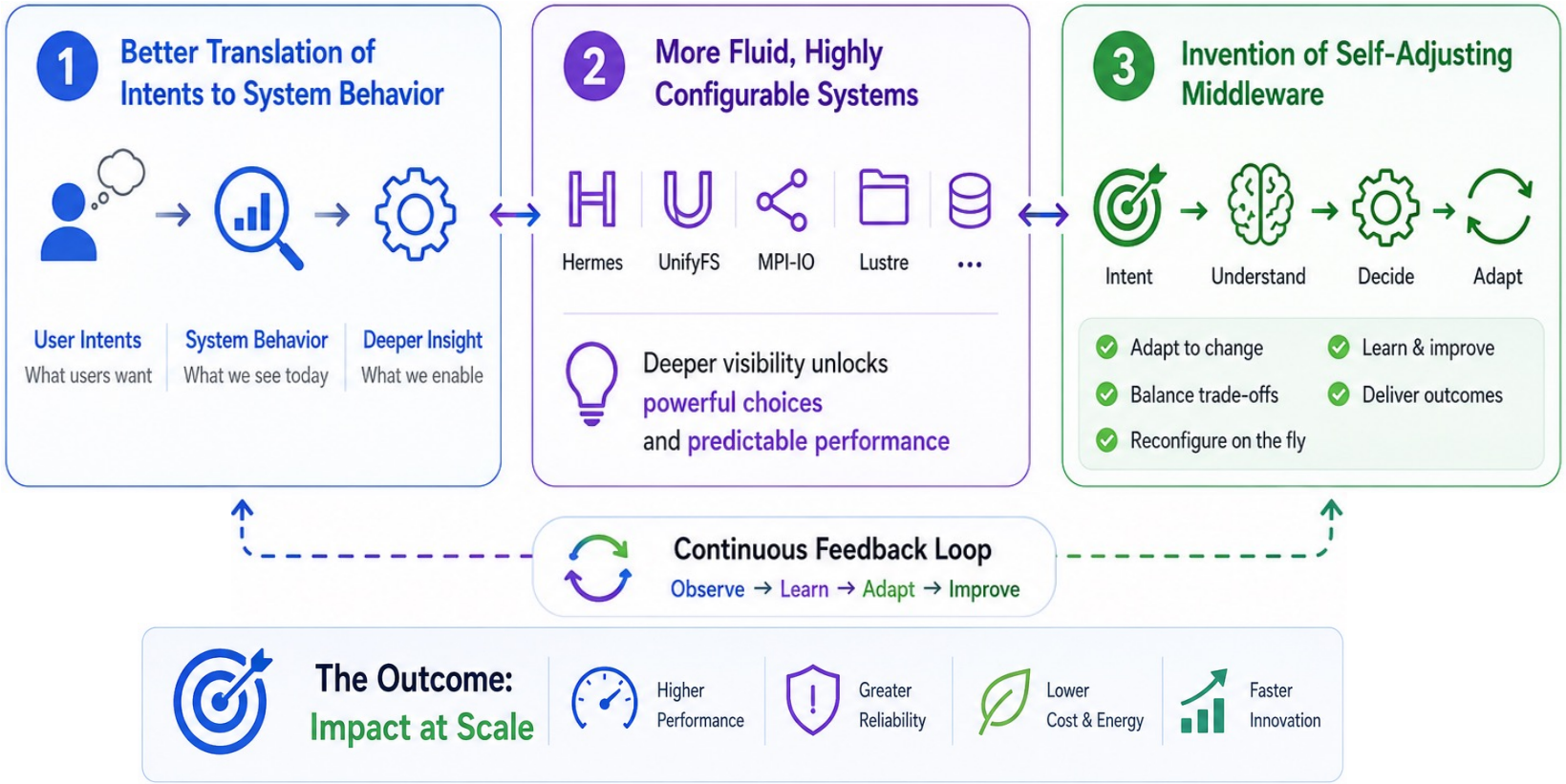
Examples of Dynamic Decisions

Chunking Strategy | Aggregator Count | Buffer & Cache Policies | Prefetching & Read-Ahead | I/O Path Selection



Tightly integrated continuous loop

From intent to impact.



Call for Action

From observing I/O to explaining and optimizing it.

1 Look Beyond Observation for Optimizations

From what happened...
to why it happened.

- Close visibility gaps
- Build complete pictures
- Enable targeted optimizations

Aim for complete, explainable performance.

2 Look Beyond Traditional Tools and Find New Capabilities

- Go beyond APIs
- Use new capabilities
- Understand the stack
- Optimize with knowledge

Stop at "what" and discover the "why".

3 Collaborate for Next-Generation Monitoring and Systems

- Work as a community
- Partner with system administrators
- Deploy and iterate together
- Shape next-generation systems

Stronger together.
Better tools, better systems, better science.

Great Research Happens Together



Conclusions

Towards Explainable, Optimizable, and Self Adjusting Storage system

