IOWarp — CLIO, the Context Layer for I/O

PI TEAM Xian-He Sun, Anthony Kougkas, Gerd Heber, Jacob Hochhalter, Vivek Srikumar

PROJECT MEMBERS Jaime Cernuda, Luke Logan, Jie Ye, Hua Xu, Izzet Yildirim, Rajni Pawar, Shazzadul Islam, Daisy Quach, Matthew Larson, Joe Lee

NSF CSSI 2411318

A **multi-tiered data platform** that stores scientific data as **self-described content** and makes it **searchable by AI agents** from inside the storage layer. This year: **released** on Linux, Windows and macOS, **proven at 1,024 nodes**, and the memory substrate for scientific agents.

```
pip install iowarp-core
```

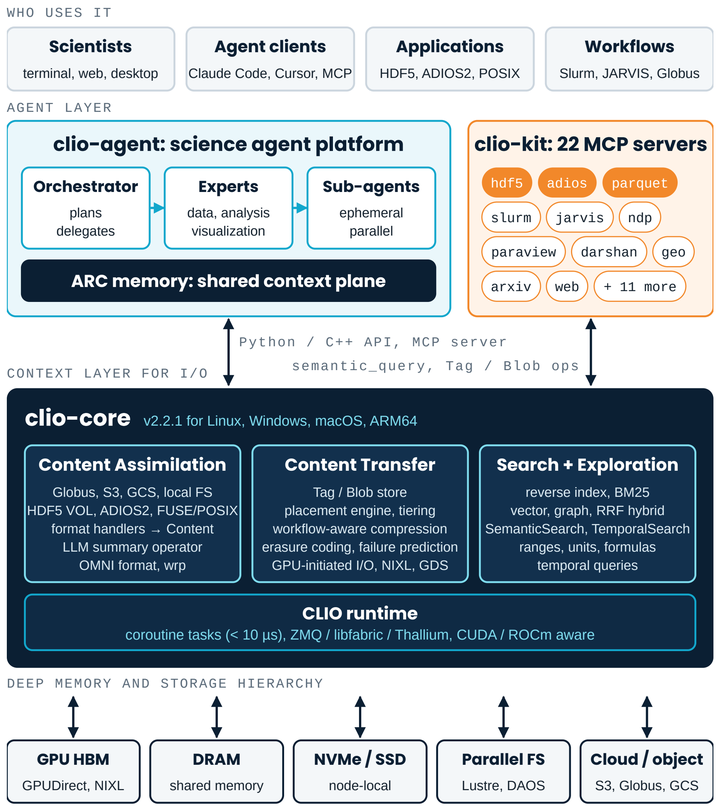
https://iowarp.ai

01 The Context Problem

Scientific workflows now interleave simulation, analytics, and AI, and they produce enormous, heterogeneous **context**: instrument data, simulation output, model weights, agent memory. The storage stack still treats all of it as opaque bytes, so agents guess at data with `glob` and `grep` and pay I/O, memory, and token cost at the point of reasoning.

- **Content, not bytes**: every blob carries a uniform data tag readable by any API.
- **Semantic search inside storage**: the layer that sees every write answers natural-language queries.
- **Tiered placement from GPU to cloud**: one hierarchy from HBM to object stores.
- **Agent memory as a storage problem**: context persists across sessions at logarithmic cost.

02 Architecture



Content = a blob plus a uniform data tag (type, structure, operators, provenance), tiered by the placement engine.

03 Get CLIO

clio-core `v2.2.1`

Tiered content store, adapters, semantic search. Linux, Windows, macOS, ARM64.

```
pip install iowarp-core  &&  clio_run start
```

clio-agent `v0.9`

The co-scientist: desktop, terminal, and container builds, ARC memory on clio-core.

```
https://iowarp.github.io/clio-agent/
```

clio-kit `v2.10`

22 MCP servers and 150+ tools for formats, schedulers, visualization, data platforms.

```
uv tool install clio-kit
```

YEAR 2 BY THE NUMBERS

109	48.4k	47	54
RELEASES	PYPI INSTALLS	CONTRIBUTORS	PRS UPSTREAM
27 core, 34 agent, 48 kit	six-month window	7,504 commits, 1,450 PRs	to 12 external projects

04 This Year: An Agentic-First Storage Layer

O(log n) was O(n)
Semantic + temporal search in storage
Hybrid BM25, vector, and graph retrieval with scientific operators.

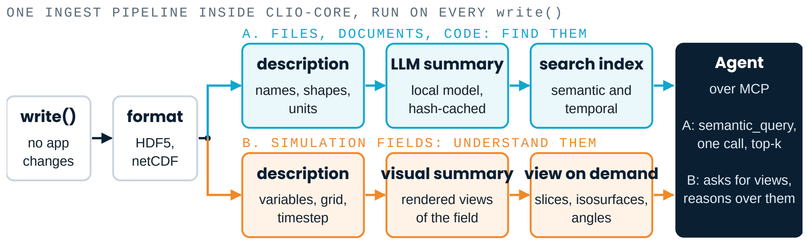
3× I/O, representative workflows
GPU-initiated I/O
GPUs issue I/O in-kernel; no synchronous HBM-to-DRAM copy.

30% end-to-end workflow time
Workflow-aware compression
Placement runs compression on unburdened nodes.

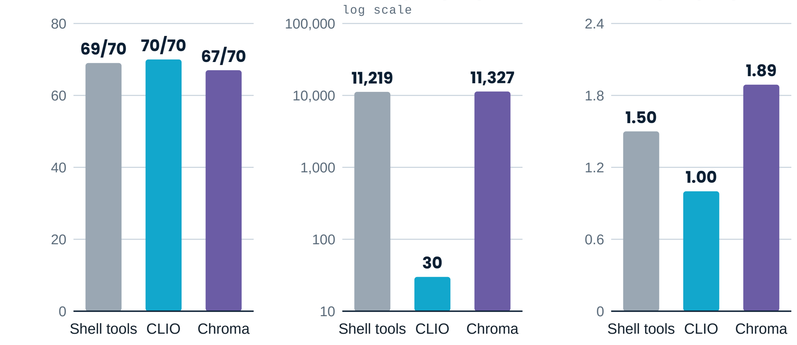
0 application code changes
Transparent ingestion
ADIOS2, HDF5 VOL, FUSE/POSIX, Globus, S3 adapters.

05 Agents Find and Understand Data

clio-core attaches meaning to content at write time, through operators inside the ingest path. Files get an LLM summary and a search index, so an agent finds them with one call instead of a glob-and-grep loop. Scientific fields, which fit no context window, get rendered views the agent can choose and reason over. One pipeline serves both, with no per-application training.



A. FINDING FILES: ACROPOLIS (ESCIENCE '26), 70-QUERY CMIP6 BENCHMARK



- Claude Code with shell tools only, with the clio-core search tool added, and Chroma's context-1 RAG agent. **70/70 solved, one call and 30 tokens per query** against 69/70, 1.5 calls and 11,219 tokens: no LLM in the lookup loop.

B. UNDERSTANDING SIMULATIONS: VIGIL (SC '26)

4
simulations explored, every domain question answered correctly

~1,500 tokens
per rendered view of a field that fits no context window

0
per-application training or application code changes

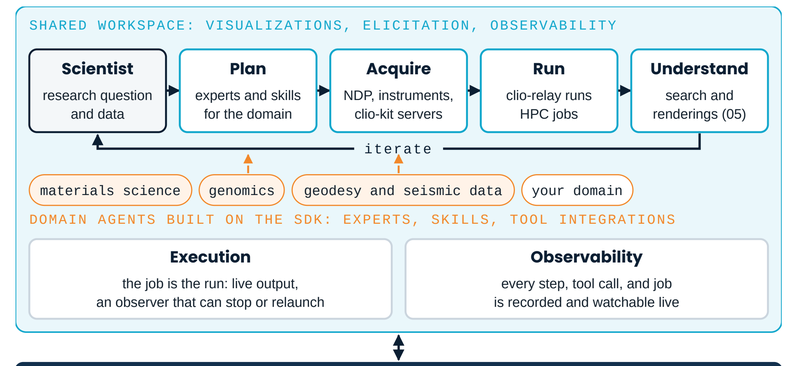
- **The agent decides what it sees.** Same prompt and tools across LBM, LAMMPS, Gray-Scott, and Xcompact3d.

06 Community and Next Steps

- **5 publications this year** resulting from this project, across SC, HPDC, and eScience.
- **Collaborations:** Argonne ALCF, National Data Platform, Oak Ridge, NVIDIA, NCSA, materials science at Utah.
- **Year 3:** the CLIO co-scientist for materials science, biology, and seismic data; advisory board at SC '26, tutorials, user group; CLIO inside vLLM and llama.cpp for KV-cache management.

Looking for new collaborations!

07 CLIO Agent: The Co-Scientist



ARC memory: persistent laboratory memory
context, artifacts, and lineage retained across sessions

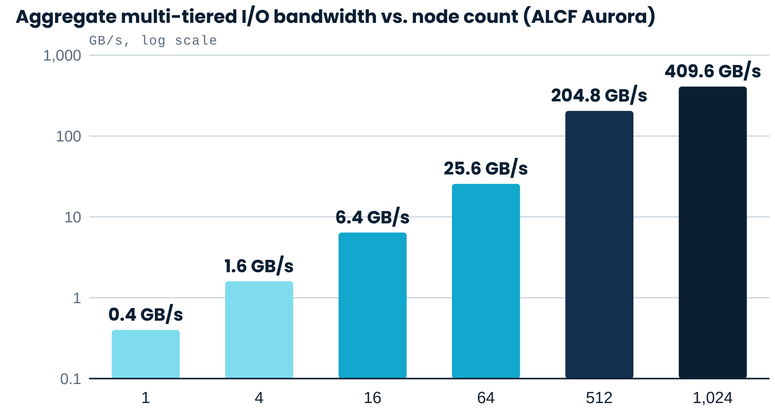
clio-core
tiered from DRAM to parallel file system and object store

- **Research questions and data as input.** The agent acquires, processes, and interprets data with the scientist, plans experiments, and iterates.
- **A domain-agnostic agentic SDK.** Materials science, genomics, and geodesy agents are composed from reusable experts, skills, and tools.
- **Execution, not only planning.** clio-relay submits and supervises HPC jobs, with live output and an observer that can stop or relaunch them.
- **A shared workspace.** Interactive visualizations, elicitation when a decision is the scientist's, full observability of every step and job.
- **Persistent laboratory memory.** ARC keeps context, artifacts, and lineage across sessions, tiered on clio-core.

08 Proven at Leadership Scale

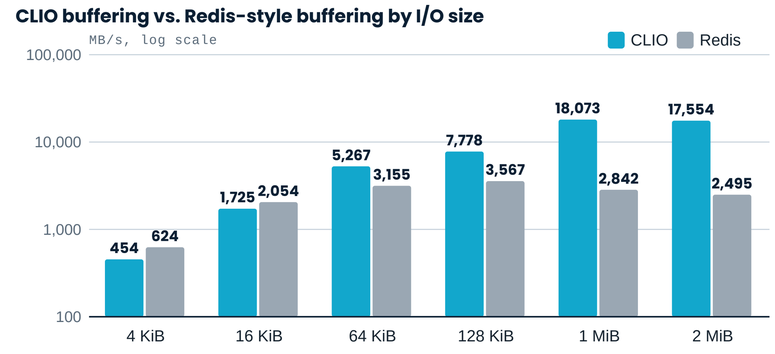
MACHINE	ALCF Aurora	EVENT	INCITE '26 GPU hackathon
SCALABILITY DEMO	512 nodes	REAL WORKLOADS	up to 128 nodes
PLATFORM EVALUATION	1,024 nodes	ALSO RUNS ON	Polaris, DeltaAI, Ares

Aggregate multi-tiered I/O bandwidth vs. node count (ALCF Aurora)



- **409.6 GB/s at 1,024 nodes**, linear at 0.4 GB/s per node from one node to the full run.

CLIO buffering vs. Redis-style buffering by I/O size



- **Up to 7× Redis** at 1 to 2 MiB objects; Redis stays ahead below 64 KiB. Capacity spills across tiers instead of stopping at DRAM.

LINKS

- **iowarp.ai** — project site
- **clio-core** — source and releases
- **clio-agent** — download
- **clio-kit** — MCP servers and tools

REFERENCES

R. P. Pawar, L. Logan, J. Cernuda Garcia, A. Kougkas, X.-H. Sun. **Acropolis: A Portable, AI-Driven Semantic Search System for HPC.** IEEE eScience 2026.

H. Xu, J. Cernuda, A. Gainaru, S. Klasky, A. Kougkas, X.-H. Sun. **Trigger, Render, Reason: Multimodal LLM Agents for Science-Aware Simulation Monitoring.** SC 2026.

I. Yildirim, X.-H. Sun, A. Kougkas. **AgentStage: Exploiting LLM Thinking for Data Staging in Scientific Agents.** IEEE eScience 2026.

K. Bateman, J. Cernuda Garcia, L. Logan, B. Nicolae, F. Cappello, X.-H. Sun, A. Kougkas. **Agent Error Simulator: Fault Injection for Testing Recovery in Agentic Workflows.** IEEE eScience 2026.

J. Ye, A. Maurya, K. Chitty-Venkata, B. Nicolae, A. Kougkas, X.-H. Sun. **PKAS: Predictive KVCache-Aware Scheduling for Faster LLM and Transformer Inferences.** ACM HPDC 2026.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. **2411318** and **2411319**. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

Computing resources: Argonne Leadership Computing Facility (Aurora and Polaris); NCSA Delta and DeltaAI through the ACCESS program; the Ares cluster at Illinois Tech.

Related efforts and collaborators: National Data Platform (NSF), EarthScope (NSF), Oak Ridge National Laboratory, NVIDIA, The HDF Group, Kitware ParaView.

Gnosis Research Center @ ILLINOIS TECH