WekaIO Overview

Omar Quijano
Department Head, SCS-ISS

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Agenda

- Background
  - History
  - Overview
  - Performance
- Operations
  - What worked well
  - What requires improvement
- Challenges
  - Current Challenges
  - Future Requirements
- Q/A

History

Two WekaIO file systems in LCLS:

- Home for controls and data systems infrastructure
  - Homes for 2000+ users, software repos
  - Mostly accessed via NFS using automount
  - 300 TB, 16 x supermicro nodes with 5 x NVME, 100 Gb Ethernet
  - Replacement for old standard Linux-native, JBOD-based, ZFS-based NFS
- Fast feedback storage layer for science data generated at the LCLS beamlines
  - Raw data from the detector, some users generated data
  - Mostly accessed natively
  - 450 TB, 16 x supermicro nodes with 5 x NVME, IB HDR100
  - Replacement for Lustre

Our old NFS infrastructure was slow and fragile, Lustre required significant effort to set up

Overview

- **Overall Mission:** Provide a robust, high-performance, and scalable storage foundation for critical research, HPC, and operational workloads.

- **Environment Snapshot:**
  - 6 Weka Clusters
  - 2 Administrative Domains: PCDSN & S3DF
  - 100 servers total: 2,186 CPUs (logical), 22.5TB RAM, 1,450GB/s aggregate network bandwidth
  - Total Drive Capacity: 1,270 NVMe drives, 12.1PB configured capacity, 8.3PB available
  - Primary Protocols: POSIX, NFSv4, S3
  - Key Use Cases: HPC, Home Directories, Boot/Root FS, Raw Data Ingestion, Kubernetes, Scratch Space, VMWare Datastore.

- **Technology Stack:**
  - Weka Versions:
    - Backends: 4.4.10.150
    - Clients: 4.4.8.53/4.4.10.150
  - Backends: RHEL 8.6 and Rocky 9.4; Intel Xeon Silver & AMD EPYC CPUs
  - Networking: ConnectX-6 100GbE; FFB has Ethernet & InfiniBand (HDR)
  - Storage: High-performance NVMe (Samsung, Micron, WD, Kioxia)

- **Strategic Importance:** Foundational to data-intensive operations and research.

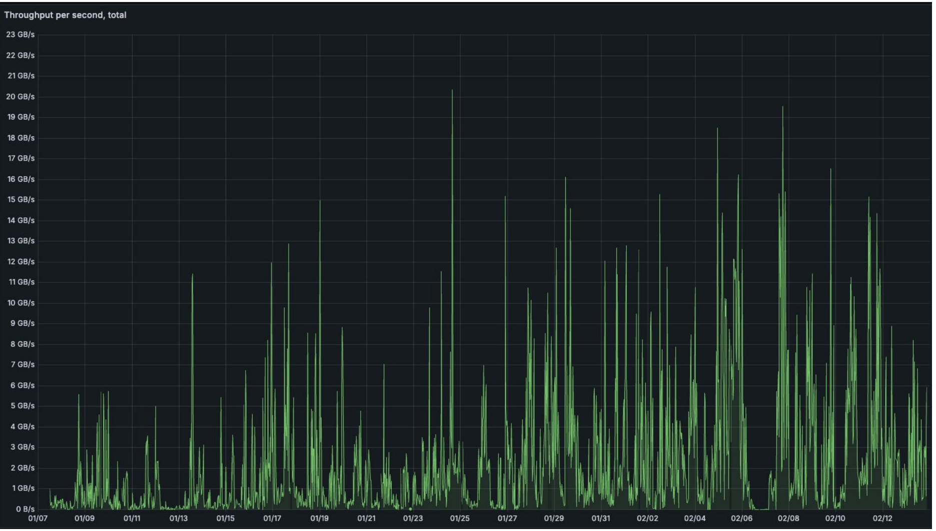
S3DF Storage Summary

Cluster Name	Version	Utilization (consumed/ provisioned)	Total Capacity	Usage
sdfdata	4.4.10.150	2.5 PB / 4.7 PB	6 PB	Tiered file system (Weka - managed Ceph OBS cluster) for scientific and experimental data
	18.2.2	54 PiB / 66 PiB	90 PiB	**One of the Largest Single CEPH Clusters (HDD)**
sdfhome	4.4.10.150	434 TB / 900 TB	1.4 PB	User home directories, group (community), and software space
sdfscratch	4.4.10.150	688 TB / 1.0 PB	1.7 PB	Scratch space for high performance workloads
sdfk8s	4.4.10.150	466 TB / 1.0 PB	2.2 PB	Persistent storage for Kubernetes
slac-ffb	4.4.10.150	830 TB / 1.0 PB	1.0 PB	LCLS Fast Feedback Cluster
WEKACDS	4.4.10.150	89 TB / 149 TB	332 TB	LCLS home, group, software, and diskless Cluster

S3DF Storage Performance

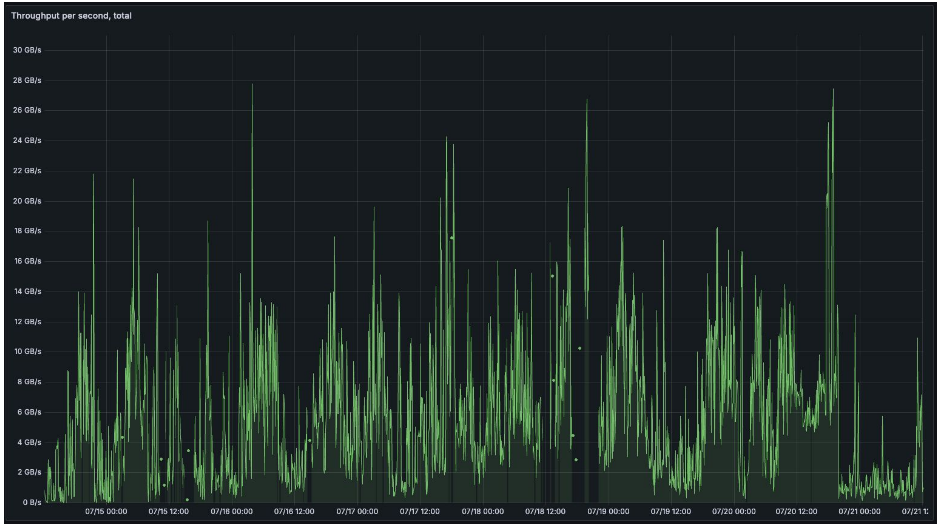
Weka + Ceph

January Results: Throughput per Second



Peak: 20.5 GiB/s

July 2025 Plot: Throughput per Second



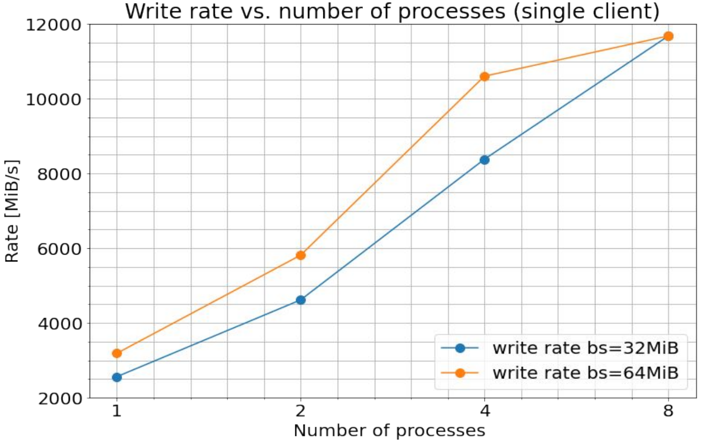
Peak: 28 GiB/s

Operation Efficiency

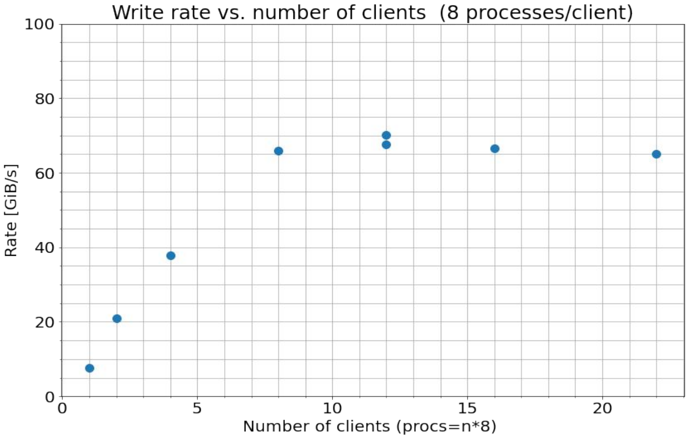
What has worked Well:

- Strong partnership
  - Great support from the weka team for installation, setup and maintenance (on-site engineer)
- Great reliability, performance and scalability
- Simplicity (eg intuitive/powerful command line) and flexibility (eg file system resizing)
- S3 Protocol Gateway: allow customers to read/write to the SDFData file systems

Single client, multiple writers



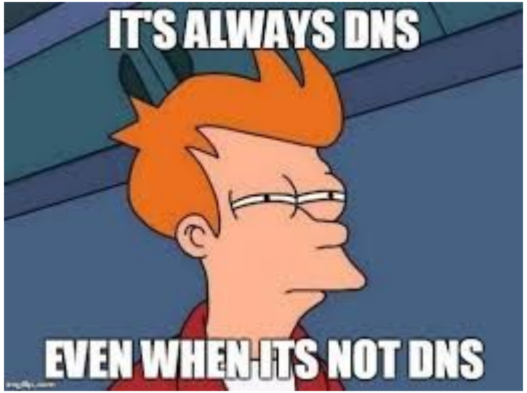
Multiple clients, 8 writers per client



Operational Efficiency

What requires Improvement:

- Better NDU upgrade process
- Reviewing the resources required to operate WEKA, especially from the perspective of client resource allocation (CPU/RAM) - multi cluster single client memory 5GiB per container.
- Client upgrade coordination
- SLURM + Weka core pinning
- Feature request and implementation timeline



Challenges

Current:

- Improving the upgrade process to have zero business impact
- Support for Intel E810 network cards
- Improve observability and alerting (still getting false-positive)
- Coordinating the deployment of the WEKA Kubernetes operator and aligning configuration across clusters, namespaces, and teams.
- Implementing per PVC snapshots for WEKA without requiring per-filesystem PVCs.
- Continue next-generation hardware planning.
- Moving current backend servers to the new storage network

Challenges

Future:

- LCLS-HE Data rates up to 5Tibps (Max Peak)
- Identify the stripe width configuration that ensures predictable performance and linear scalability as the cluster grows beyond N nodes.
- Implementing secure NFSv4/Kerberos authentication to meet security requirements while minimizing operational overhead.
- Evaluating TLC vs. QLC trade-offs to balance endurance, performance, and cost for future scaling.
- Developing a disaster recovery plan for SDFData
- Ensuring support for specific network adapters with modern secure OSs.

Q/A

SDFData