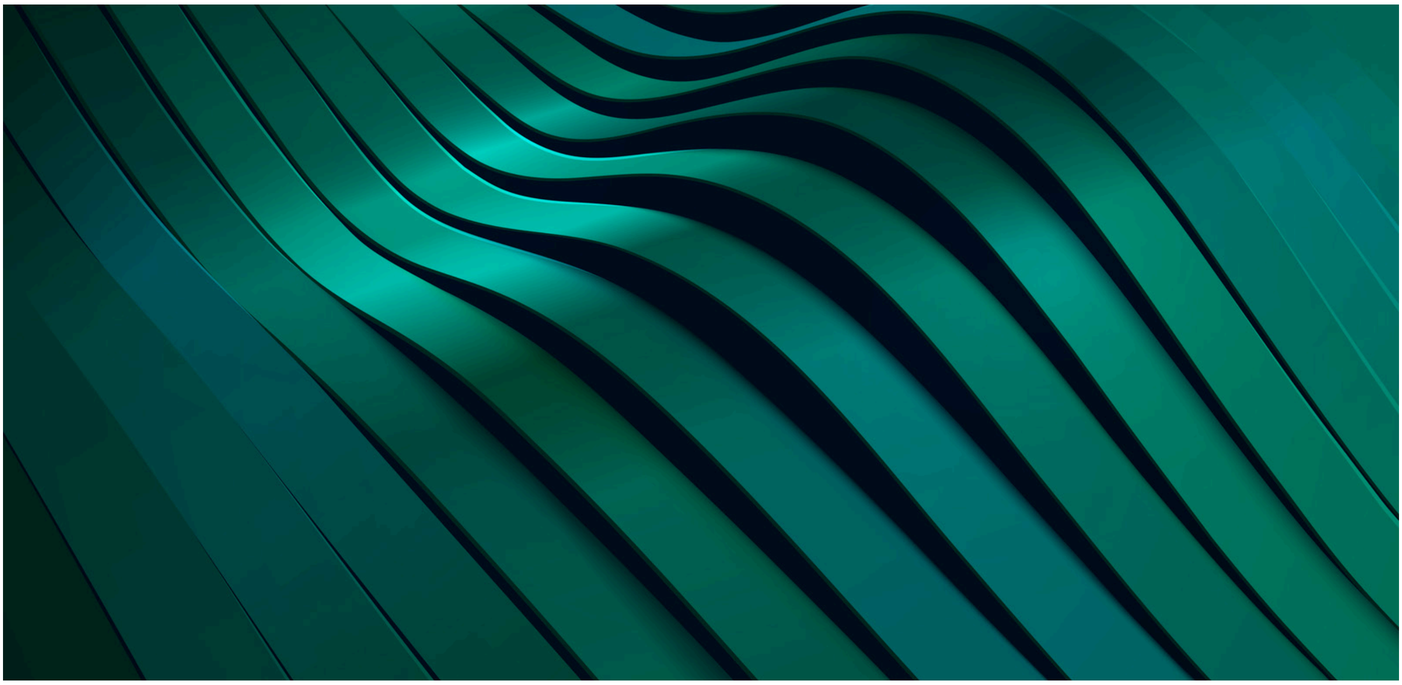
Technical white paper

How to accelerate the performance of your SAS 9.4 environment with HPE solutions for WEKA



Contents

List of figures

List of tables

Introduction

Complex business problems require sophisticated, high-end analytics and the ability to integrate massive data sources, including vast collections of text-based data. SAS® is the advanced and predictive analytics leader with industry-specific product sets for healthcare, banking, government, retail, telecommunications, aerospace, and marketing optimization or high-performance computing. SAS enables users to quickly prepare, explore and model multiple scenarios using data volumes never before possible. As a result, accurate and rapid insights are delivered in near-real-time (typically in minutes rather than hours). However, setting up File Systems to support large volumes of data with adequate performance, as well as ensuring sufficient storage space for the SAS temporary files, can be very challenging. Indeed, the most prevalent and noticeable cause of performance issues with SAS software is insufficient IO infrastructure bandwidth. In this white paper we'll discuss how HPE solutions for Weka provide the performance and scale to bring customers quicker time to insights on their business.

To better understand the importance of the File System in the SAS 9.4 performance, let's see the typical SAS 9.4 Workloads and how SAS 9.4 manages IO.

SAS 9.4 Workload characteristics

With SAS 9.4, it is common to have many instances running at any given time. Generally, each SAS 9.4 user starts their own SAS 9.4 session for each SAS 9.4 job/application they are running.

The SAS 9.4 software IO and workload characteristics are:

- SAS 9.4 sessions tend to perform large sequential reads and writes. Some SAS 9.4 applications (e.g., BI) perform random access to data, but for the most part, the SAS 9.4 Workloads are predominately large sequential IO requests with high volumes of data.
- Block sizes of 32K, 64K, or even 128K.
- SAS 9.4 does not pre-allocate storage when SAS 9.4 initializes or when performing writes to a file. When SAS 9.4 creates a file, it allocates a small amount of storage, but as the file grows during a SAS 9.4 task, SAS 9.4 extends the amount of storage needed.
- Reading and writing of data are done via the OS file cache. Therefore, SAS 9.4 does not use direct IO by default.[1]
- A SAS 9.4 job, especially if the job is long, can create a large number of temporary files. The size of the files may range from very small (under 1 MB) to very large (in the 10s of GB).
- SAS 9.4 creates standard OS files for its datastore (SAS 9.4 datasets, indexes, etc.).
- SAS 9.4 sessions are single-writer threads, and they can be multi-read threads.

File system for SAS 9.4 characteristics

SAS 9.4 data sets and associated files are built within the confines of the underlying operating system (OS) and are just File System files. Therefore, they can be managed by file management utilities that are a part of the OS or third-party products.

The critical aspect of any File System supporting SAS 9.4 applications is to provide the sustained IO bandwidth needed by the SAS 9.4 jobs.

Recommended IO throughput[2]

As a rule of thumb, 100–150 MB/sec per application core is the minimum throughput that SAS 9.4 recommends for the storage system. This value ensures that data are delivered fast enough to keep the CPU fully utilized and to avoid having the CPU idle while waiting for incoming data. For example, the File System throughput required for a single host machine with 24 CPU cores is 24 Core x 150 MB/sec/core = 3.6 GB/sec minimum.

Let's see more in detail the throughput for each SAS 9.4 File System:

- **SAS 9.4 Data.** It stores persistent data created by the SAS 9.4 ETL processes. This File System read and write ratio typically ranges from 80/20 to 60/40. SAS 9.4 recommends providing a minimum sustained IO bandwidth of 100 MB/sec from storage to each SAS 9.4 data File System for normal SAS 9.4 usage and up to 150 MB/sec for heavy statistics and analytics operations.

[1] Since SAS uses the OS's file cache to read and write data, the maximum IO throughput rate is restricted by how fast the OS's file cache can process the data.
[2] SAS Paper SAS6761-2016 Best Practices for Configuring Your I/O Subsystem for SAS9 Applications.

- **SAS 9.4 Work.** It is the scratch working space for the single thread SAS 9.4 jobs. It is used to perform the working storage activity of single-threaded SAS 9.4 procedures. This File System read and write ratio is typically 50/50. SAS 9.4 recommends providing a minimum sustained IO bandwidth of 100 MB/sec from storage to each Work File System for normal SAS 9.4 usage and up to 150 for heavy statistics and analytics operations.
- **UTILLOC.** It is the scratch working space for the single and multi-threaded SAS 9.4 procedures. UTILLOC read and write ratio is typically 50/50. It is generally heavily used. SAS 9.4 recommends providing a minimum sustained IO bandwidth of 100 MB/sec from storage to each UTILLOC file space for normal SAS 9.4 usage, and up to 150 MB/sec for heavy statistics and analytics operations.

The summary of the main SAS 9.4 File Systems is reported in the table below:

Table 1. SAS 9.4 File Systems characteristics

	Data	Work	UTILLOC
Purpose	Persistent data for SAS 9.4 output files	Scratch working space for single-threaded procedures	Scratch working space for multi-threaded procedures
Read/Write ratio	80/20 to 60/40	50/50	50/50
Min sustained throughput required	100 MB/sec/core	100–150 MB/sec/core	100–150 MB/sec/core

Other recommended File Systems aspects

On top of the high IO throughput requirement, a shared File System for SAS 9.4 should provide the following features:

- Efficient handling of File System Metadata
- Optimal large sequential block IO
- File system data retention in a local file cache (in memory)
- Transparency for access, location, concurrency, replication, etc.

Why Weka File System is a good choice for SAS 9.4?

WEKA is a modern Data Platform containing the WekaFS Filesystem (formally WekaIO Matrix), which is a flash-optimized, software-only, high-performance File System designed for GPU, AI, and Analytics workloads with large, diverse data sets and high-compute requirements. It provides a high-performance, scalable parallel File System architected for massive throughput, high-IOPS, low-latency, Exabyte scalability, as well as advanced data movement capabilities to facilitate hybrid cloud environments and multi-protocol access to all data sets.

WEKA File Systems overview

The WEKA File System has a POSIX interface to present all data in a single namespace. In addition, it provides support for the NVIDIA® GDS (GPUDirect Storage) and additional interfaces like NFS and SMB. It also has a CSI interface to provide persistent storage to containerized applications.

Figure 1 below provides an overview of the software architecture.

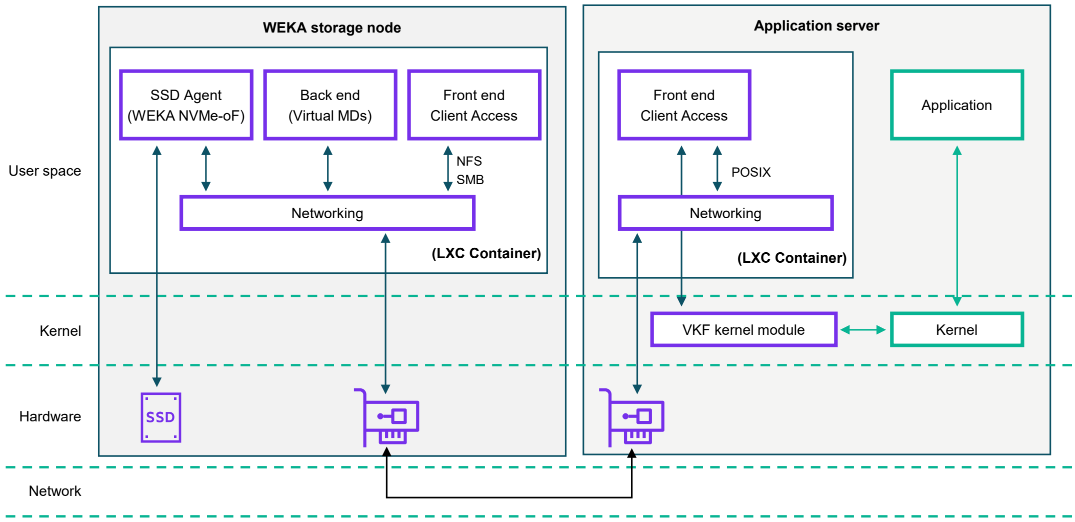


Figure 1. WEKA internal architecture

Apart from the WEKA Virtual File System (VFS) kernel driver, all WEKA core components run in a Linux® container (LXC) in user space. Running into WEKA's own RTOS (Real-Time Operating System) eliminates time-sharing and other kernel-specific dependencies. Indeed, WEKA manages its assigned resources (CPU cores, memory regions, network interface cards, and SSDs) to provide process scheduling, memory management and to control the IO and networking stacks. Furthermore, by not relying on the Linux kernel, WekaFS effectively utilizes a zero-copy architecture with much more predictable latencies. The additional advantage to run inside an LXC container is the improved isolation from other server processes.

To run the network stack and NVMe from user-space, WekaFS leverages acceleration technologies like DPDK (Data Plane Development Kit), SPDK (NVMe only), and SR-IOV (Single Root IO Virtualization/For some Ethernet only) as well as running its own proprietary high-performance network protocol that DOES NOT require ROCE, thereby simplifying networks configuration. WekaFS software can also leverage RDMA for NVIDIA GPUDirect Storage to provide a direct path between storage and GPUs, eliminating IO bottlenecks for data-intensive AI applications. These technologies and Weka's innovative software help WekaFS maximize performance over InfiniBand or Ethernet networks.

All these aspects are entirely transparent for the application as WekaFS is fully POSIX compliant, so applications won't need to be re-written to take advantage of WekaFS performance.

The WEKA Virtual File System (VFS) kernel driver, which provides the POSIX File System interface to applications, is the only WEKA component that runs in the kernel. Running in the kernel provides significantly higher performance than what can be achieved using a FUSE user-space driver. In addition, it allows applications that require full POSIX compatibility to run on a shared storage system. The WEKA VFS driver enables WekaFS to support full POSIX semantics while running the WEKA client component in user space as a process in its containerized LXC environment.

The WEKA system client is installed on the application servers and enables file access to WEKA File Systems. It intercepts and performs all File System operations and communicates in parallel with all of the WEKA storage nodes for data and metadata to provide the best optimal performance that the hardware can provide. In addition, WekaFS offers advanced capabilities such as byte-range locks, and it is tightly integrated with the Linux operating system page cache.

Protocols

On top of POSIX, WekaFS has full multi-protocol and data-sharing capability (all data can be accessed through all protocols) across a variety of protocols:

- Full POSIX for local File System support
- NVIDIA GPUDirect Storage (GDS) for GPU acceleration
- NFS for Linux
- SMB for Windows
- S3 for Object access

Adaptive caching

The WEKA File System provides a unique advanced caching capability called adaptive caching. This capability allows users to fully leverage the performance advantages of Linux data caching (page cache) and metadata caching (dentry cache) while ensuring full coherency across the shared storage cluster.

WekaFS supports leveraging Linux page cache—typically reserved for direct-attached storage (DAS) or file services run over block storage—on a shared networked File System while maintaining complete data consistency.

WekaFS provides the same capability for metadata caching, also known as Linux dentry cache. As a result, a client can leverage the local metadata cache for a directory, reducing latency significantly.

Metadata

In addition to data throughput optimization, WekaFS optimizes access to the metadata, distributing it across the entire storage infrastructure to ensure massively parallel access to NVMe drives. So, metadata performance scales linearly with the size of the File System.

Data and node resilience

Weka utilizes a patented distributed data protection coding scheme that increases resiliency as the number of nodes in the cluster scale. It delivers the scalability and durability of erasure coding but without the performance penalty. Indeed, WEKA's rebuild time gets faster and more resilient as the system scales because every node in the cluster participates in the rebuild process. Furthermore, WekaFS equally distributes data and metadata across logical nodes that span failure domains (FD).

WEKA Filesystems for SAS 9.4 test results

Recently, a joint group of HPE and WEKA engineers has tested WEKA as File Systems with SAS 9.4. These tests aimed to measure the throughput of the WEKA File Systems for SAS 9.4. The scenario used was the SAS 9.4 Mixed Analytics workload suite scenario.[3] It consists of typical analytic jobs designed to replicate light to heavy workloads. These jobs are launched via a script that includes time delays to simulate scheduled jobs and interactive users launching at different times.

The test showcased the WekaFS and SAS 9.4 running on the environment described in the table below:

Table 2. Test environment description

WEKA	SAS 9.4
7x HPE ProLiant DL360 with	8x HPE ProLiant DL360 with
• 2x Intel® Xeon® Gold 6126 CPU @ 2.60 GHz with 12 Core each	• 2x Intel Xeon Gold 6126 CPU @ 2.60 GHz with 12 Core each
• 192 GB RAM	• 192 GB RAM
• 6x 3.84 TB SSD disk	• 100 Gb/s Ethernet
• 100 Gb/s Ethernet	

Test scenario

The SAS 9.4 Mixed Analytics workload used for this testing had the following characteristics:

- 50% CPU intensive and 50% IO intensive jobs
- Testing used SAS 9.4
- Utilized SAS 9.4 procedure that includes DATA step, PROC RISK, PROC LOGISTIC, PROC GLM (general linear model), PROC REQ, PROC SQL, PROC MEANS, PROC SUMMARY, PROC FREQ, and PROC SORT
- SAS 9.4 program input sizes up to 50 GB per job
- Input data types are text, SAS 9.4 data set, and SAS 9.4 transport files
- Memory use per job is up to 1 GB
- Job runtimes were varied (short and long-running tasks)

[3] The SAS 9.4 Mixed Analytics workload suite scenario uses real-world data volumes and structures of a typical SAS customer. The scenario simulates the types of jobs received from various SAS 9.4 clients such as Display Manager, batch, SAS 9.4 Data Integration Studio, SAS 9.4 Enterprise Miner™, SAS 9.4 Add-In for Microsoft Office, SAS 9.4 Enterprise Guide®, and SAS 9.4 Studio.

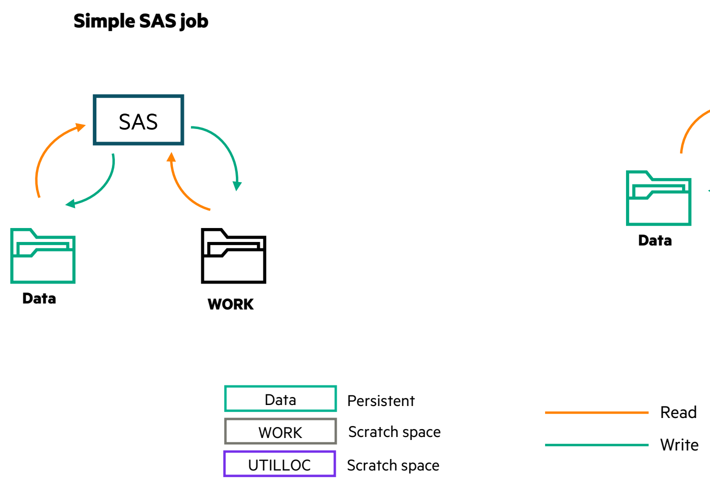


Figure 2. Example of SAS 9.4 Workloads

The environment was queried every 5 seconds during the tests. The results were then aggregated for the best 10 intervals (50 seconds), 24 intervals (2 minutes), 120 intervals (10 minutes), and 240 intervals (20 minutes). This was done because the IO demand starts high and tails steadily off during the duration of the tests. In fact, for a number of intervals during the later stages of the tests, only CPU intensive jobs remained active, and there was no IO being performed.

The figure below shows the average throughput for each of those intervals versus the total number of jobs.

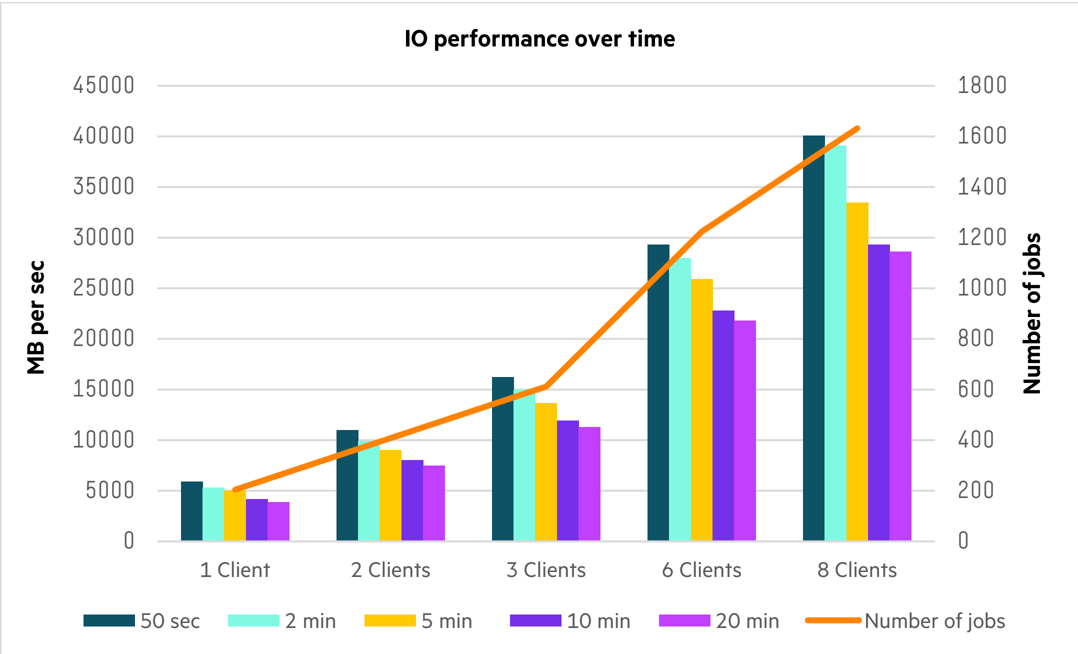


Figure 3. Test results

The graph legend tells the number of 60 user runs (288 SAS 9.4 jobs) being performs on each compute server (Client). A legend of 1 Client means that two x 30-user runs were run simultaneously on a single server. 2 Clients means that four 30-user runs were run simultaneously, 2 on each server. 8 Clients means two x 30-user runs on each of 8 servers. Each Client ran 288 SAS 9.4 jobs. This means that for the 1 Client, a total of 288 SAS 9.4 jobs were run, and for 2 Clients, 576 jobs were run. Likewise, 3 Clients ran a total of 864 SAS 9.4 jobs, etc. The largest number of SAS 9.4 jobs occurred with the 8 Client run. In that run, there were 2304 total SAS 9.4 jobs run across all environments, with their IO demand being satisfied by the WEKA storage system.

Conclusions

The testing shows that a seven-node WEKA cluster provides a sustainable throughput of up to 40 GB/sec. This means such a cluster can support up to 260 cores of a SAS 9.4 system. Moreover, due to the WEKA distributed architecture and single namespace, additional performance can be simply obtained by progressively adding nodes.

HPE Engineering and WEKA engineering jointly developed and validated an optimized HPE Solution for WEKA based on the latest HPE ProLiant DL360 Gen10 Plus and HPE ProLiant DL325 Gen10 Plus Servers. The task-optimized HPE ProLiant DL360 and HPE ProLiant DL325 server families provide a very efficient and cost-effective hardware platform to combine with WekaFS to support the most demanding AI and analytics workloads. HPE ProLiant DL servers support all the hardware required for maximizing WEKA performance over the fabric, including NVMe drives and fast InfiniBand or Ethernet network adapters.

So, for enterprises that need to accommodate large SAS 9.4 IO-intensive workloads, the HPE solution for WEKA on the HPE ProLiant DL server family is a storage solution capable of providing high throughput, scalability, easy-to-use, high availability, and resiliency.

Why HPE

The key benefits of the combination of HPE solutions for WEKA software with HPE ProLiant DL servers are:

- It delivers a simple, high-performance, consumer-grade user experience for large SAS 9.4 environments, both compute and storage
- It is easy to operate and tune
- HPE Solutions for WEKA can provide over 50% better NVMe density to optimize data center footprint
- HPE is a single-step shop for the whole hardware and software solution
- HPE has a dedicated solutions team focusing on SAS 9.4 analytics, unlike several competitors
- HPE produces reference architectures and collaterals regularly

Learn more at

HPE Solutions for WEKA

HPE Solutions for Weka QuickSpecs

HPE ProLiant DL325 Gen10 Plus server product brief

HPE ProLiant DL360 Gen10 Plus server QuickSpecs

SAS 9.4 Paper SAS6761-2016 Best Practices for Configuring Your I/O Subsystem for SAS9 Applications

WekaFS architecture white paper

Hewlett Packard Enterprise

a50006102ENW, Rev. 1