Cerence™ Gains Superior Cost-Efficiency, Scalability, and Flexibility with WekaFS™

REQUIREMENTS AND CHALLENGES

- Ability to support the Cerence vision to be a cloud-focused company
- A POSIX-compliant solution that provides a gateway to object storage in the public cloud to reduce cost
- Automatic, seamless tiering to the public cloud for operational efficiency
- Superior performance with optimization for NVMe
- Compatibility with existing infrastructure for investment protection
- Modular scalability to accommodate anticipated growth

BENEFITS

- Ease of migration and a gateway to the public cloud
- Complete integration with POSIX infrastructure, much smaller datacenter footprint, and greater cost-efficiency
- Time and resource savings with simple, efficient automatic tiering
- The best performance and lowest latency of any candidate file system
- Use of both on-premises and cloud resources
- Flexible scalability for cost-optimized capacity planning expected expansion

We looked at our legacy architecture and instead of taking an evolutionary step and upgrading every component, we took the revolutionary approach.

Weka cost-effectively enables both the use of POSIX and object storage with performance and latency that is far superior to any other solution.

Bridget Collins, Chief Information Officer, Cerence Inc.

The Weka File System (WekaFS) delivers high-performance, low-latency data storage for Cerence Inc., enabling its work in conversational artificial intelligence (AI) and deep learning (DL) to perfect the way machines listen and speak. Cerence, headquartered in Burlington, MA, builds AI models for major automobile manufacturers worldwide. Its automotive cognitive assistance solutions power natural and intuitive interactions between automobiles, drivers and passengers, and the broader digital world. Cerence produces one of the world's most popular software platforms for building automotive virtual assistants — fueling AI for a World in Motion.

Cerence is building speech language models to improve the in-car experience, specifically using AI to deliver a deep understanding of human behavior, culture, and language. Its mission is to help make automotive transport safer and human interaction with the car more natural and comfortable. The compute infrastructure at Cerence supports multiple mixed workflows: speech language models; those for rapidly growing Natural Language Processing (NLP) and Natural Language Understanding (NLU) efforts; Text to Speech; and Predictive Keyboards — a pattern recognition effort.

THE CHALLENGE: FINDING A COST-OPTIMIZED WAY TO BRIDGE TO OBJECT STORAGE AND THE PUBLIC CLOUD FOR SCALABILITY

In 2019, Cerence spun off and became a separate entity from parent company Nuance Communications, a leading provider of conversational AI and user of the high-performance file system GPFS on massive storage arrays. Cerence determined that using its legacy file system would result in cost-prohibitive economics at scale and began to explore alternatives. The end-users at Cerence relied on a POSIX file system for their research workflows, but they had the added constraint of having to cost-effectively manage the mountains of data that would be collected. Therefore, using an expensive, monolithic infrastructure to provide the POSIX file server was not desirable as the cost would be prohibitive at scale.

The Cerence IT team decided to start from scratch and build something transformative that would scale cost-effectively and future-proof the new datacenter. While providing a POSIX file system was initially the predominant driver of architecting the system, the ability to leverage object storage now became another key factor. Therefore, the new storage solution had to meet several key criteria and be able to:

- support use of both the POSIX file system and object storage in a cost-effective way
- enable a hybrid implementation and tier data seamlessly from on-premises to the public cloud
- reduce datacenter footprint
- use modern storage technologies such as NVMe
- allow for modular growth and scale with the growing needs of the business.

THE SOLUTION: WekaFS ON HPE SERVERS

Cerence evaluated several other POSIX file systems, including GPFS, Lustre, and BeeGFS, but ultimately chose WekaFS (Weka) because it delivers the best performance and meets all the requirements with a single, software-defined storage architecture. Weka provides the simplest, most effective data platform integrating fast flash and object storage for cost savings, and delivers superior, consistent performance with no degradation even when the metadata size is large. This was significant because the previous file system suffered severe performance degradation when metadata sizes grew large, slowing down research to an unacceptably slow pace.

WekaFS is a fully parallel and distributed file system with a clean sheet design that leverages high-performance NVMe flash for the hot tier and cost-effective object storage for the cold tier. Both data and metadata are distributed across the entire storage infrastructure to ensure massively parallel access to the NVMe drives. Data is seamlessly tiered from on-premises to the hybrid or public cloud with Weka's internal tiering mechanism, making the optimum use of storage media for the best economics.

WekaFS has a two-tier architecture that takes NVMe flash and disk-based technologies and presents them as a single hybrid storage solution. The Cerence IT team decided to implement Weka in a converged mode with WekaFS running on GPU servers, creating a single namespace from all the locally attached NVMe

drives. The team is managing 900TB of data on the Weka file system to support the NLP and NLU workloads on a cluster consisting of 40 HPE Proliant DL360 servers, each with dual 25GbE networking adapters. The servers are interconnected with 4 switches for high availability (HA), performance, and redundancy. Each server has one network connection and two NVMe drives dedicated to WekaFS and one GPU card. Cerence IT is managing 3.2PB of data on object storage with SUSE Enterprise Storage™, with 900TB assigned to Weka, running on a cluster consisting of 9 HPE Apollo 4200 servers, each with twenty-four 14TB drives. The team also utilizes an HPE Apollo 6500 server with 8 GPUs for multi-GPU processing.

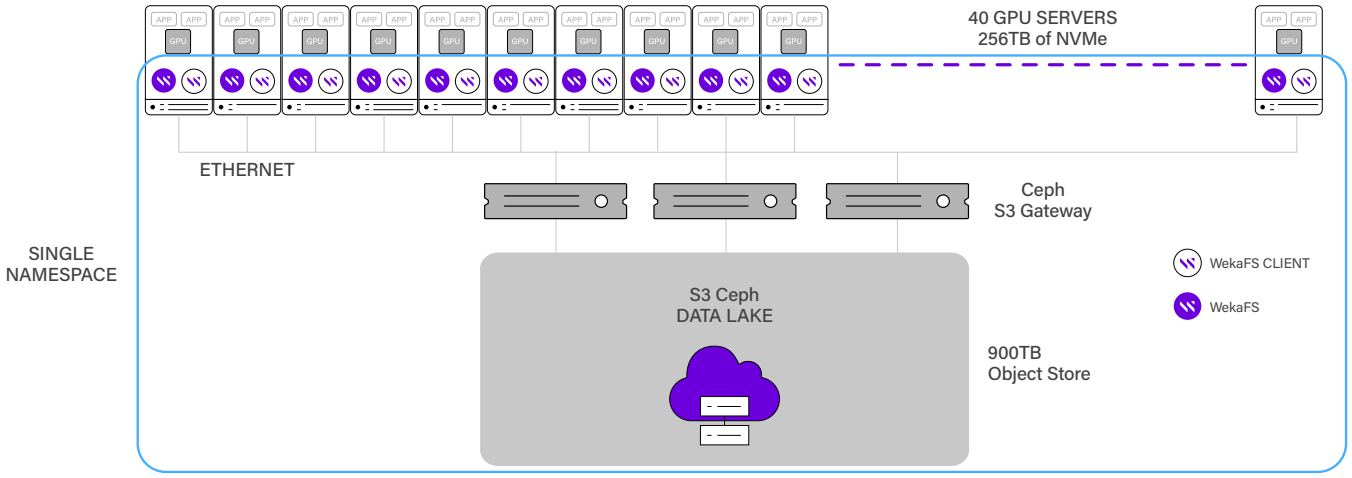

Figure 1 - Cerence infrastructure with tiering from on-premises to a private cloud

THE WEKA INNOVATION NETWORK™ (WIN) IS THE GATEWAY TO INNOVATION

HighFens, a WIN Accelerator partner, provided technology orchestration expertise. The team at HighFens has a long-standing relationship with Cerence and a long track record of building scalable environments for on-premises, public cloud, or hybrid implementations. HighFens trusted Weka as the backbone of the system because the WekaFS architecture fit the customers' needs, offered a smaller footprint running in a converged mode, and also provided a modular way to grow and expand to a hybrid infrastructure that utilizes the public cloud as a data lake.

The Weka Innovation Network (WIN) plays an important role in building out the solution. HighFens has strong partnerships with Weka and HPE, a WIN Innovator partner that played an essential part in the recommendation for Cerence to choose Weka as its high-performance storage system. The pre-validated and engineered Reference Architectures (RAs) with HPE's premium Apollo and ProLiant server product lines with WekaFS are an ideal combination for AI workloads. In addition, Weka's network stack leverages Intel's open-source Data Plane Development Kit (DPDK) technology. Leveraging DPDK, Weka accelerates and improves networking performance for data access. In addition, Mellanox networking products have been a critical component in the systems under test for many record-breaking results on industry-standard benchmarks such as the IO-500, SPEC SFS, and the STAC-M3 "tick analytics" benchmark.

BENEFITS AND ROI

- Cost-efficiency: the new environment is much smaller, with superior performance without massive amounts of hardware
- Improved Resource Utilization: with WekaFS, there is investment protection of existing infrastructure and full utilization of GPU cycles
- Automated Tiering: Cerence can now move data up and down the stack without legacy labor and resource-intensive manual tiering, all while accelerating end-user performance
- Flexibility: WekaFS provides a bridge between POSIX and object storage as well as a gateway to the public cloud
- Performance: having a Tier 0 solution that is much faster and with lower latency provides Cerence with unprecedented performance
- Modular Scalability: Cerence can now scale cost-effectively at a pace that is adaptable to business demands.

For more information or to locate a partner in the WIN Global Partner Program, go to: https://www.weka.io/partners.

Weka Innovation Network | Hewlett Packard Enterprise | HighFens Inc

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: +1 844.392.0665 E: info@weka.io www.weka.io

W04CS202007