Solutions Brief

Navigating the Data Deluge With High Performance Advanced Cluster Architecture

keepertechnology

In This Issue:

Machine Learning Dilemma

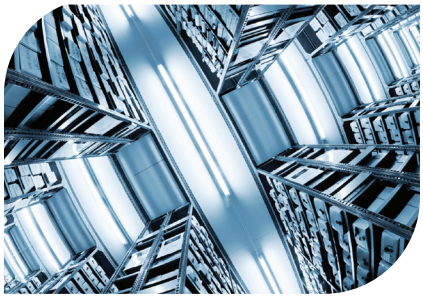
A 'Data Deluge' results when the amount of new data generated surpasses an organization's power to manage it, the analysts' capacity to analyze it, and the researchers' ability to deduce any useful conclusions from it.

Over two exabytes of data is generated each and every day in nearly every imaginable way. This copious outcome is fueled by factors that compound each other. Data sensors are exploding: cameras, phones, digital assistants (Alexa, Siri, etc.), cars etc. Massive growth in image resolution now allows for 3D image resolution at 20 times the size of 2D. Both the increasing resolution of data collected and its frequency has generated this data explosion.

Electronic versions of traditional objects such as aircraft engines, wind turbines, and heavy equipment are now joined by digital doppelgangers of everything from toothbrushes and traffic lights to entire shops and factories. Even humans have begun developing these copies in the electronic frontier.

The National Football League will design a digital avatar for every player to be added to their platform, thus creating real-time location data, speed, and acceleration for every player during every play on every inch of the field. This and much more has contributed to 90% of the world's data being generated in the last two years alone.

Taking on the Challenges of Data Deluge



How is an organization supposed to survive their own Data Deluge? Every institution in every industry will face this challenge. Those that are best equipped and prepared to handle it will have the best chance to survive. And those that embrace it and take advantage will have the best chance to thrive.

Whether generated by the healthcare industry, the government or even professional sports leagues, the huge collection of data is not being created simply because it can through advances in technology, but in an effort to improve the health, welfare and safety of the individuals served. How the industry will accomplish solving these data complexities is constantly evolving today, but regardless it will necessitate leveraging artificial intelligence and machine learning. In order to meet accelerating advancements in AI technologies, Hardware configurations must also evolve rapidly to collect, compute, and process all this data.

Both government and private entities have been using data analytics for years with excellent results. The ability to identify and quantify complex relationships has saved countless lives and great amounts of resources. At the other end of the spectrum, quantitative analysis has made others into billionaires.

While traditional data analytics continue to get more complex, they ultimately give way to more advanced AI systems. And when the data is so large and so complex that humans can no longer keep up with the AI systems, machine learning is needed to allow systems to continue to evolve and keep up with growing data needs. As the data tsunami continues to explode, however, a powerful solution using intelligent machines for data access, process, storage and analysis is needed to keep up with evolving capabilities and requirements. The ultimate goal is to make sense of it all.

Projected Data Collection Requirements

INDUSTRY SECTOR	FUTURE DATA CHALLENGE EXAMPLE
Professional Sports	Real-time body telemetry with enhanced statistical modeling requirements
Bio-medical/Genomics	Expedited human genome sequencing needed for precision medicine
Self-Driving Cars	Autonomous vehicles will generate 2-4 TB per hour, for each vehicle[3]
Military/Battlefield	The military surveillance sensor market expected to reach $32 Billion in 5 years
Digital Data-sphere	By 2025, IDC predicts global data repositories will surpass a trillion Gigabytes

> “In 2021, artificial intelligence augmentation will create $2.9 trillion of business value and 6.2 billion hours of worker productivity globally.” Gartner, Inc.

A Powerful Solution

Artificial intelligence and machine learning can be a powerful solution for challenges facing the government today. A study at the Harvard Kennedy School identified several areas that can benefit from AI/ML:

Resource Allocation
- Administrative support is needed to speed up task completion
- Inquiry response times are long due to insufficient support

Large Datasets
- Dataset is too large for employees to work with efficiently
- Internal and external datasets can be combined to enhance outputs and insights
- Data is highly structured with years of history

Experts Shortage
- Basic questions can be answered, freeing up time for experts
- Niche issues can be learned to support experts in research

Predictable Scenario
- Situation is predictable based on historical data
- Prediction will help with time-sensitive responses

Procedural
- Task is repetitive in nature
- Inputs/outputs have binary answer

Diverse Data
- Data includes visual/spatial and auditory/linguistic information
- Qualitative and quantitative data needs to be summarized regularly

The Keeper Solution

The engineers at Keeper Technology have decades of experience helping government and commercial customers manage some of the largest data repositories. Most of these are part of high velocity data environments where data rarely stands still. High-speed ingest, time-critical processing, complex data analytics, and digital dissemination typically characterize such environments. Keeper Technology also has years of experience working with and protecting data in some of the most secure environments of the Intelligence Community today.

Combining this experience with years of developing turn-key and cost-effective solutions, Keeper Technology has developed a computing cluster that is ideal for AI/ML and data analytics. The converged infrastructure design combines processing, storage, and data protection in minimal rack space.

The latest Intel processors and high-speed memory provide the optimum platform for the analytic engine whether its Hadoop, HPCC Systems, Spark, Splunk, or other data analytics package. The only way to fully leverage the incredibly high performance of the processors is to ensure they're not starved for data. The Keeper Cluster can feed massive amounts of data to the processors from the tightly coupled NVMe storage layer. Tying the cluster together is a very low latency InfiniBand fabric. The salient features are:

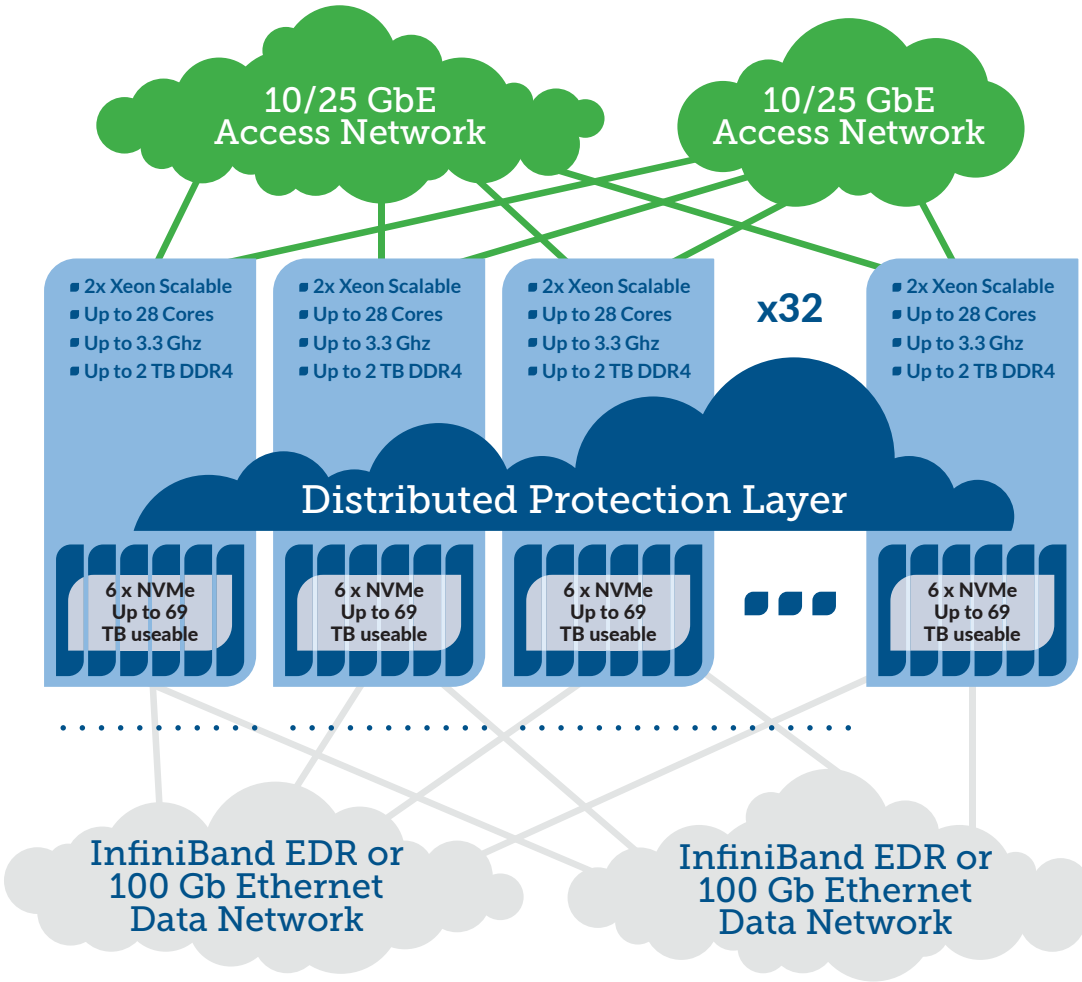
Processing	The cluster is made up of 32 redundant processing nodes with no single point of failure. Up to 1,792 processing cores (over 3,500 threads) running at speeds as high as 3.3 Ghz (base frequency) provide the most processing power for a system in its class.
Memory	With a maximum of 64 TB of aggregate memory, even the largest data sets can remain resident in memory. Larger searches, faster training, more detailed models: Ultimately, quicker and better results.
Storage	Up to 2.2 PB of usable storage is included in the cluster. The NVMe storage is fully protected against drive and node failures with configurable protection schemes (mirroring, single parity, dual parity, erasure coding) spreading data across multiple nodes. Data can be dedicated to a processing element or shared across multiple nodes or the entire cluster. With 500 GB/s access to storage and 10's of Millions of IOPs the application can run at maximum performance without having to worry about being starved for data.
Networking	To ensure data flows freely between nodes, the cluster is based on a fully redundant InfiniBand backplane with ~150ns latency between nodes. The network adapters also provide protocol offload to keep the processors focused on the task at hand.
Access	The independent access network is also fully redundant and provides 80 GB/s of connectivity to your network core.
Density	The whole cluster takes less than one half of a standard 19" rack.

High performance is a natural result of our turn-key architecture.

Architecture Differentiators

- Scale-out performance across multiple NVMe servers
- Non invasive cluster-wide data protection
- Local latency at data center scale

32 Nodes with Data Protection & Redundant Networks



32 Node Cluster:

- 64 x Xeon Scalable
- Up to 1,792 Cores
- Up to 3.3 Ghz
- Up to 64 TB
- Up to 2.2 PB Useable

Executive Summary

For many agencies and industries, the data deluge is here. This can be an unfortunate situation for those who are unprepared or cannot react quickly enough. For them, data and information will be lost, most of it potentially irreplaceable. But along with the deluge comes incredible opportunities for insight and knowledge. The agencies that can take advantage of it will open the door to a new world of capabilities.

Fortunately, a variety of software packages, many of them open source, are available to help tame the deluge of data and extract the valuable knowledge contained within. Some of the most common are:

- Hadoop
- HPCC Systems
- Spark
- Splunk
- TensorFlow

Systems such as these can bring the power of artificial intelligence out of the lab and into real-world situations. They make it significantly easier to extract information and glean knowledge from massive amounts of raw data.

The Keeper solution is a scalable cluster with flexible configurations. It is designed specifically for running the kinds of AI and machine learning processes required to truly harness the torrent of data that new technologies are creating. The data deluge is here. The only way to thrive is to embrace it and take advantage of it.

Competitive Assessment

- Ran the customer application 4 times faster
- Fit in one quarter the physical space (less than half a rack compared to two full racks)
- Has twice the usable storage
- Costs approximately 40% less

ABOUT Dutch Ridge

Dutch Ridge Consulting Group (DRCG) is a Department of Veterans Affairs (VA) Center for Veterans Enterprise (CVE) certified, Service-Disabled Veteran-Owned Small Business (SDVOSB) with primary offices in Beaver, Pennsylvania and Ashburn, Virginia. DRCG delivers mission focused expertise in Cybersecurity Engineering and Operations; IT Solutioning; Program Management; Policy, Planning, Communications and Compliance Support; Cyber Threat Intelligence; Workflow Solutioning; Insider Threat Prevention and Detection; Criminal History Record Information (CHRI) collection and Professional Business and Management Consulting Services. DRCG's technical approach optimizes client investments by leveraging expertise in managing growth and transformation of existing IT environments. DRCG's understanding and knowledge of the federal contracting landscape enables its management team to enhance, not replace, agency investments in both personnel and technology that have already been made.

ABOUT Keeper Technology

Keeper Technology is a strategic IT partner that helps organizations navigate between business needs and technology solutions in order to advance and support their mission, reduce costs, and reduce risks. We leverage our team of highly certified engineers and architects that use proven methodologies to design and implement innovative solutions. Founded in 2005, Keeper Technology is a Virginia-based small business and systems integrator helping to solve today's most challenging data storage and processing issues. Keeper Technology offers unique big data storage platforms in order ensure reliable and agile data access, management, and protection.

keepertechnology | 21740 Beaumeade Circle | Suite 150 | Ashburn, VA 20147
P [571] 333 2725 | F [703] 738 7231 | solutions@keepertech.com | www.keepertech.com

