Cray I/O Solutions

Data proliferation in the global datasphere has exploded in the past few years — and is expected to grow more than tenfold, from 16.1 trillion gigabytes (GB) to 163 zettabytes (ZB), by 2025. The need for and use of data is changing too, from primarily business background data to highly life-critical applications.

Some of those uses include data-embedded systems in autonomous vehicles, robotics manufacturing, and connected homes. By 2025, the average person is expected to interact with connected devices nearly 5,000 times a day, and data stored by enterprises will reach 50% of total data stored globally. In cognitive/artificial intelligence (AI) applications alone — like machine learning and language processing — the projected growth of data subject to analysis will grow by a factor of 50, to 5.2 ZB by 2025; the amount of analyzed data that's touched by cognitive systems will grow by a factor of 100.

All that data and the insights it provides are being used to develop many of today's innovations in manufacturing, energy and life sciences. Data is everywhere, in constant use by millions of organizations, some with data-intensive applications and multiple users at multiple locations accessing it. The intense data demands associated with unique and varied industry needs require systems that can handle all the data, processing it in real time with low latency, flexibility and immediacy. The consequences of data that's unavailable, particularly in life-critical applications, can be disastrous.

DATA PROLIFERATION
Massive amounts of data are produced every day, and the volume will only increase. The need for and use of data is changing too, from primarily business background data to highly life-critical applications. You need quick, reliable access to all that data, and you need to process it in real time with low latency, flexibility and immediacy.

I/O CHALLENGES
In order to get good application results when processing large amounts of data, you need to understand and address the I/O characteristics of your data storage system. To avoid slow productivity and avoid frustrating users, your system needs to be efficient. Storage performance affects your time to discovery, which in turn affects your bottom line.

THE CLUSTERSTOR™ NXD SOLUTION
Cray's ClusterStor NXD Flash Accelerator Intelligent I/O Manager is designed to seamlessly handle small file I/O and large, sequential I/O acceleration for scale-out parallel file systems. Features include read persistence, write back, I/O histogram, performance statistics and dynamic flush.

Challenges

With the vast amount of data being generated and used, time to application results will be reduced without understanding and addressing the I/O characteristics of the data storage system. An inefficient data storage system will result in slow productivity, frustrate users and impede time to discovery — and time to profitability.

Adding to that challenge is **mixed workloads**. Traditional storage tiers struggle to perform when faced with random and mixed I/Os. Mixed I/O workloads are commonly found in data analytics, virtual simulation and scientific discovery. Hard drives provide the best efficiency value for large, predictable, streaming I/Os, but they are not optimized for small or random I/Os. Conversely, SSDs accelerate small I/Os efficiently, but do not provide the best value for large streaming I/Os. Most organizations and applications have a variation of mixed and random I/Os, requiring a more efficient solution.

Highest Performance Efficiency per Disk Drive

Cray HPC-Optimized Disk Drives

High-Density Enclosures

Another challenge is **technology constraints**. Hard drives, SSDs and burst buffers each present advantages and drawbacks when talking about handling massive amounts of data:

- Hard drives alone offer the best capacity and large-file sequential I/O value, but IOPS are typically only in the hundreds.

- SSDs offer significantly greater IOPS than hard drives, but they're not capacity-cost-efficient.

- Burst buffers can provide a faster storage resource to supercomputers, but traditional SSD-based, network-attached burst buffers only offer full performance less than 15% of the time, require data migration, and additional tiers need to be added to the architecture.

Solution

Cray's **ClusterStor NXD Flash Accelerator** Intelligent I/O Manager is designed to seamlessly handle small file I/O and large sequential I/O acceleration for scale out parallel file systems. The NXD portfolio of flash acceleration includes a number of unique modes that help accelerate I/O patterns. These modes work in conjunction with Cray's **NXD dynamic I/O analyzer**, which analyzes workloads in real time to accelerate applications that require storage performance.

NXD is a ClusterStor feature that combines new hardware configuration to provide a transparent and automatic capability to selectively accelerate the performance of specific I/O block sizes that are usually characterized as "small." The default sets this limit as equal to or less than 32 KB block size or less. The resulting benefit is a cost-efficient hybrid storage system requiring a minimum amount of SSD flash storage. The SSD flash storage accelerates the small-block I/O, thus reducing the client-to-storage I/O request and acknowledgement latency. This reduction in the I/O request latency directly reduces the time to complete job application. Larger block sizes are detected and written directly to the HDD pool:

- The **read persistence** feature uses advanced caching technologies to enhance small-block random I/O performance by identifying frequently accessed data. This data is copied into low-latency flash storage located in the ClusterStor storage appliance, specifically the ClusterStor L300N system.

- **Write back** helps in write-intensive workloads, allowing writes at full speed to the NXD storage tier, then gathering the data sets and writing it down to disk over a period of time.

- **I/O histogram** – Profiling specific application workloads from a storage point of view is critical. Understanding where data ultimately lands on the storage is key to set the correct tunable parameters for the NXD.

- **Performance statistics** – Real-time update of the cache device performance, including cache hits, cache usage level, cache space consumption, etc., is essential to keep the storage solution working at optimal performance.

- **Dynamic flush** – Data stored on the SSDs can be flushed to the HDD partition by the flush manager to free up space in expectation of high I/O requirements. The flushing "speed" can be determined by user-controlled parameter.

Multiple flash vendors offer all-flash and/ or hybrid (flash + disk) solutions to accelerate high-performance applications; however, it's important to recognize that the primary market for these solutions are databases, email and other enterprise applications in which the I/O profile is always small block.

NXD vs. Alternate Flash Caching Solutions

The Cray ClusterStor L300N HPC storage appliance is specifically designed for different data workloads (big data, random, sequential) and uses a fundamentally different approach to flash caching for parallel workloads. More importantly, the flash caching algorithm is designed to work with many different types of applications from a wide variety of industries, such as life sciences, rich media, financial services, energy, automotive and aerospace.

File System Integration

Cray ClusterStor NXD is designed to integrate with transparent file systems and applications to truly lower the total cost of ownership for customers trying to tackle the challenges of big data.

ClusterStor Storage for Your Most Data-Intensive Workloads

ClusterStor storage systems are built on Cray's robust enterprise-class storage designs from the device level up. Powered by industry-leading parallel file systems, ClusterStor storage systems help customers accelerate their most demanding data-intensive workloads at scale.

All Flash Array	NXD Flash Accelerator	All HDD Storage
SSD	SSD	
Very high mixed I/O workload efficiency Highest price/PB	High mixed I/O workload efficiency Low price/PB	Medium to low mixed I/O workload efficiency Lowest price/PB