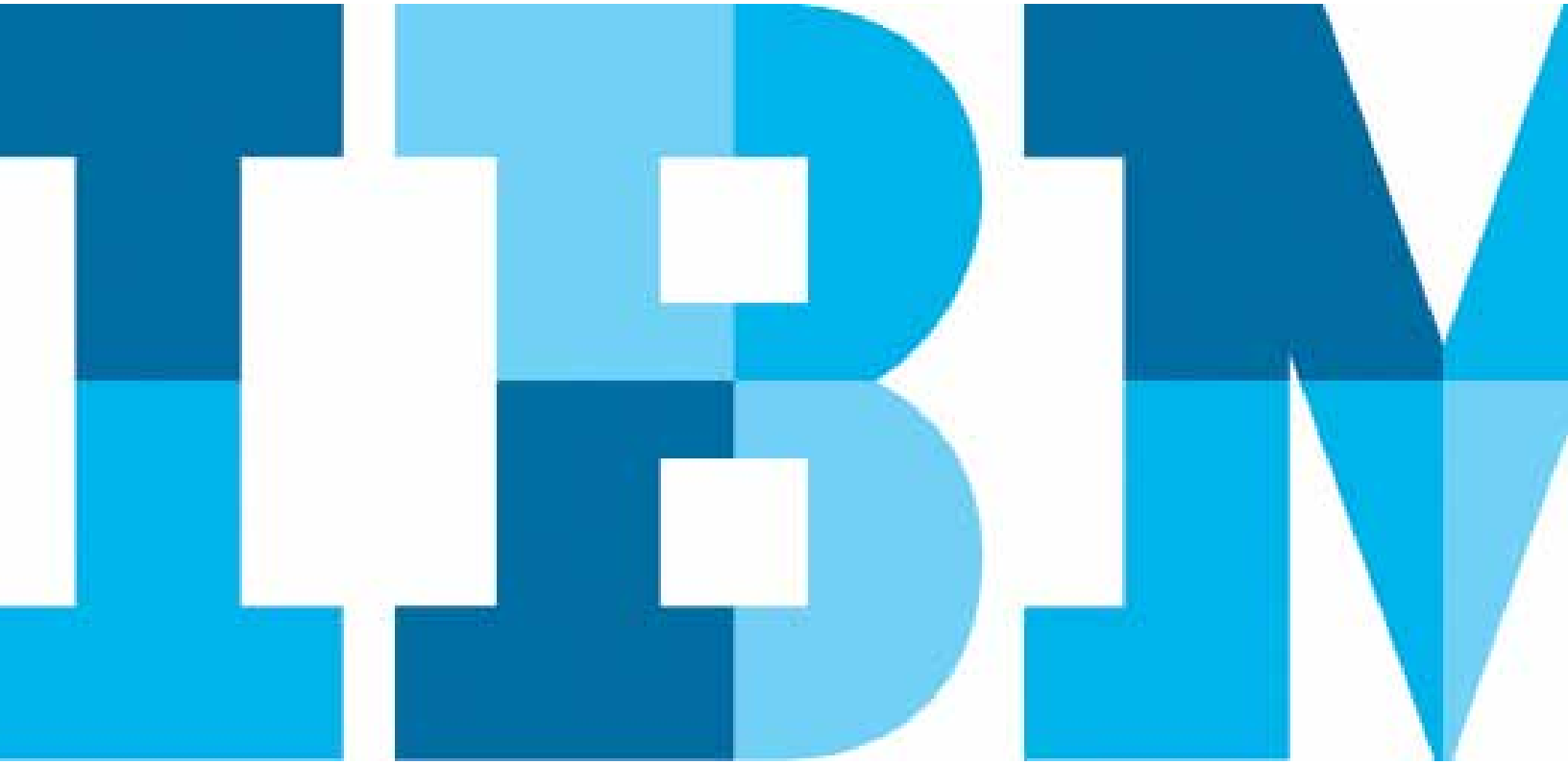
IBM Systems and Technology Group
Thought Leadership White Paper

April 2013

Technical computing benefits with IBM FlashSystem



IBM

from the rest of the data is a simple task in most parallel file systems. Using FlashSystem storage for metadata I/O dramatically reduces the time required for file system maintenance operations and processing jobs that create many small files. Separating the metadata onto FlashSystem storage will also accelerate the entire file system as the small block I/O metadata operations no longer interfere with the large streaming accesses that many HPC workloads generate. Traditional disks perform well for large block sequential accesses but mixing in small block metadata I/O disrupts this streaming and drops the disk bandwidth utilization dramatically.

Data

Using solid state storage to accelerate the data in an HPC system is a recent development and is a rapidly expanding use case for solid state storage. In the past, solid state storage technology was too expensive to be seriously considered for the large datasets that many HPC jobs deal with. The entry of high capacity flash systems at reasonable price points has changed this. High capacity solid state storage deployments are becoming more and more common and are expanding the field of problems that HPC systems can solve. Many HPC workloads have been designed assuming that the latency and I/O per second of the disks is so poor that only large block access should be used to grab enough data to fill the compute nodes memory to process locally and then stream out the results. This design only uses the disks for streaming bandwidth; however, it limits the problem sets to those that can easily be partitioned and processed independently on different compute nodes.

Solid state storage offer tremendous streaming bandwidth in a dense form factor and fit the traditional HPC software model. However, flash storage also support a high enough random I/O per second rate that this same bandwidth can be achieved with small block random I/O. This is a tremendous resource to HPC architectures as problems that require each compute node to access the entire dataset during processing can be accelerated dramatically. The alternative approaches of using massive memory proprietary supercomputers or using heavy remote direct memory access (RDMA) traffic to pool the memory of a large number of nodes both compare unfavorably to the flash storage approach. Using FlashSystem to hold the working set simplifies the architecture and allows the cluster to benefit from cost advantages of Flash over RAM including the dramatic power and space savings of flash. The only piece that has been missing is flash storage system designed with enough capacity, bandwidth and a small enough footprint to unlock the potential of flash for complex large working set HPC problems.

IBM FlashSystem

FlashSystem is the ideal implementation of flash for a technical computing environment. It uses enterprise class SLC and eMLC Flash with high write endurance to handle the heavy workloads that will be thrown at it. It provides QDR InfiniBand ports with shared access to up to 24 TB of flash per 1U chassis and can seamlessly integrate into an HPC environment. All of this is done in a chassis using less than 400 watts of power while still providing 5 GBps of data throughput. Additional scaling is as easy as adding an additional 1U chassis to reach any performance and capacity level desired. One PB of flash can fit into a single floor tile and weigh less than 1,400 lbs. This capacity density is not new to HPC but the performance/Watt and performance/GB is unprecedented.