Intel and Qihoo 360 Internet Portal Datacenter - Big Data Storage Optimization Case Study

The adoption of cloud computing creates many challenges and opportunities in big data management and storage. To resolve this, many independent software vendors and system integrators are working closely with worldwide IT solution vendors to use the latest technology and make improvements in big data storage. This paper discusses one such collaboration between Qihoo 360 Technology Co. Ltd.* and Intel to optimize the storage infrastructure in their Internet Portal Datacenter (IPDC). The solution that was chosen reduced the required storage space by almost two-thirds and increased performance by more than 10x.

The IDPC Business Requirements Of Qihoo 360

As the preeminent provider of internet and mobile phone security products and services for the People's Republic of China, Qihoo 360 focuses on providing free security solutions to internet users. By September 2012, Qihoo 360 became one of the biggest internet security companies in China with a user penetration of 95% and personal computing products and services that reached 442 million active users per month. In addition to *360 Safe Guard** and *360 Internet Security*,* Qihoo 360 also has recently released many new products including *360 Cloud**, *360 Browser**, and *360 Search**. With a rapidly growing business and user base, Qihoo 360 faced a great deal of pressure to support their increasing data storage capability. Take *360 Cloud* for example, with this free product, users can get 18 gigabytes (GB) of initial free storage space, which can be expanded further through participation in promotional campaigns or through lotteries. Besides providing free storage for over 120 million users, 360 Cloud continually upgrades its services, releasing new features such as a file safe box, online video playing, group sharing, and offline downloading. All these features not only require a large amount of data storage, but they also require a large amount of data analysis processing capability.

IDPC scaling has become a limiting factor due to Qihoo 360's fast growing business model requiring ever larger data storage solutions, analytical abilities, and an increased need for data reliability. Simply expanding the IDPC server pool with additional servers will not provide a cost effective solution that can meet the technical requirements that Qihoo 360 eagerly needs.

Collaboration Between Qihoo 360 And Intel

Intel continues improving the capabilities of its Intel® Xeon®, Intel® Core™, and Intel® Atom™ processor families to satisfy the real-time storage of data and its associated processing requirements. The hardware as well as software products are continually refined to provide increased benefits in the cloud storage segment. Intel can provide the total solution for modern IPDC needs, with integrated software and hardware solutions for the cloud storage stack including the network, processor, and storage components.

Currently, most of IPDC hardware components used by Qihoo 360 are Intel® architecture based products, and many of its software solutions are also developed on Intel® architectures. This provides Qihoo 360 the opportunity to implement a new strategy to improve big data processing and storage utilizing Intel provided hardware and software solutions. Since the beginning of 2013, Qihoo 360 and Intel have worked together using the Intel® Intelligent Storage Acceleration Library (Intel® ISA-L), to optimize the storage infrastructure of Qihoo 360's IPDC. Intel® ISA-L assists by providing increased computing capability and a reduction in real physical storage. Facing a variety of diverse requirements, Intel® ISA-L smoothly integrates with previous Qihoo 360 solutions, including Hadoop*, Cassandra*, Openstack* and Swift*.

Intel® Intelligent Storage Acceleration Library (Intel® ISA-L)

Intel® ISA-L accelerates many storage specific algorithms, extracting more performance out of the storage infrastructure. It includes functions that implement a general Reed-Solomon type encoding for blocks of data that helps protect against erasure of whole blocks. The general library for Intel® ISA-L contains an expanded set of functions used for data protection, hashing, encryption, etc., common to the needs of storage customers building everything from enterprise storage systems to small office NAS appliances. Intel® ISA-L assists original equipment manufacturers (OEMs) and independent software vendors (ISVs) by focusing on storage to gain better performance on Intel® architecture products, reducing the cost of performance optimization.

Original Big Data Storage Solution

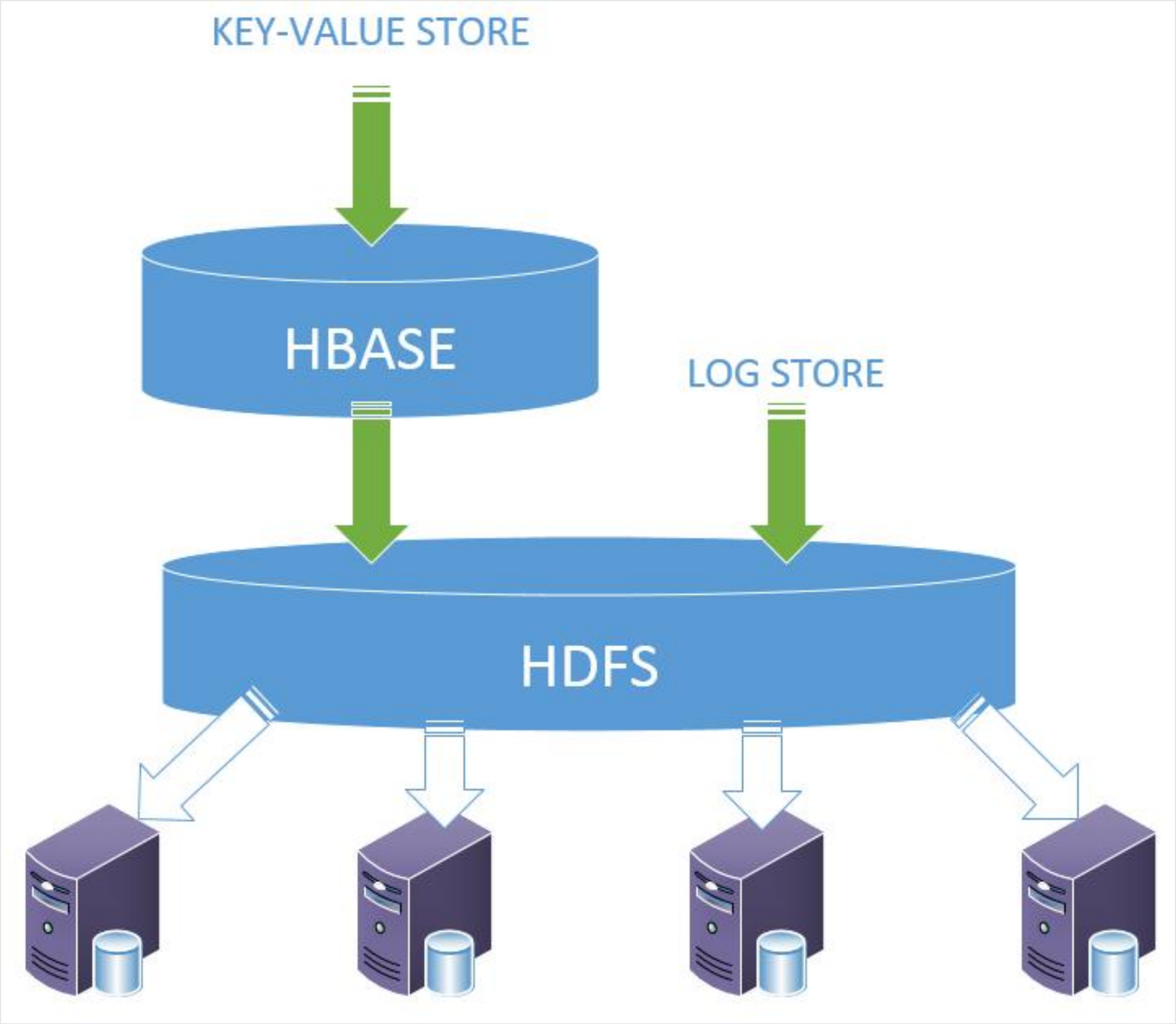


Diagram 1 Original Hadoop Big Data Storage Solution

Originally, Qihoo 360 used an open source Hadoop solution for its IPDC storage as shown in Diagram 1. This solution used an HBase database management system with an HDFS file system for the backup of the key-value store. The log store contained the log files from various business departments and was used for various analysis operations.

The key-value store and the log store combined exceeded 40 petabytes of storage space spread across thousands of servers. Hundreds of terabytes of data were added every day for the key-value store and the log store, which increased the demand for additional servers in their Hadoop cluster. The Hadoop solution used a 3-copy policy for data protection, but at the cost of requiring triple the storage space. Dealing logistically with this data redundancy scheme required an ever increasing volume of servers which created a significant challenge for Qihoo 360.

Optimized Big Data Storage Solution

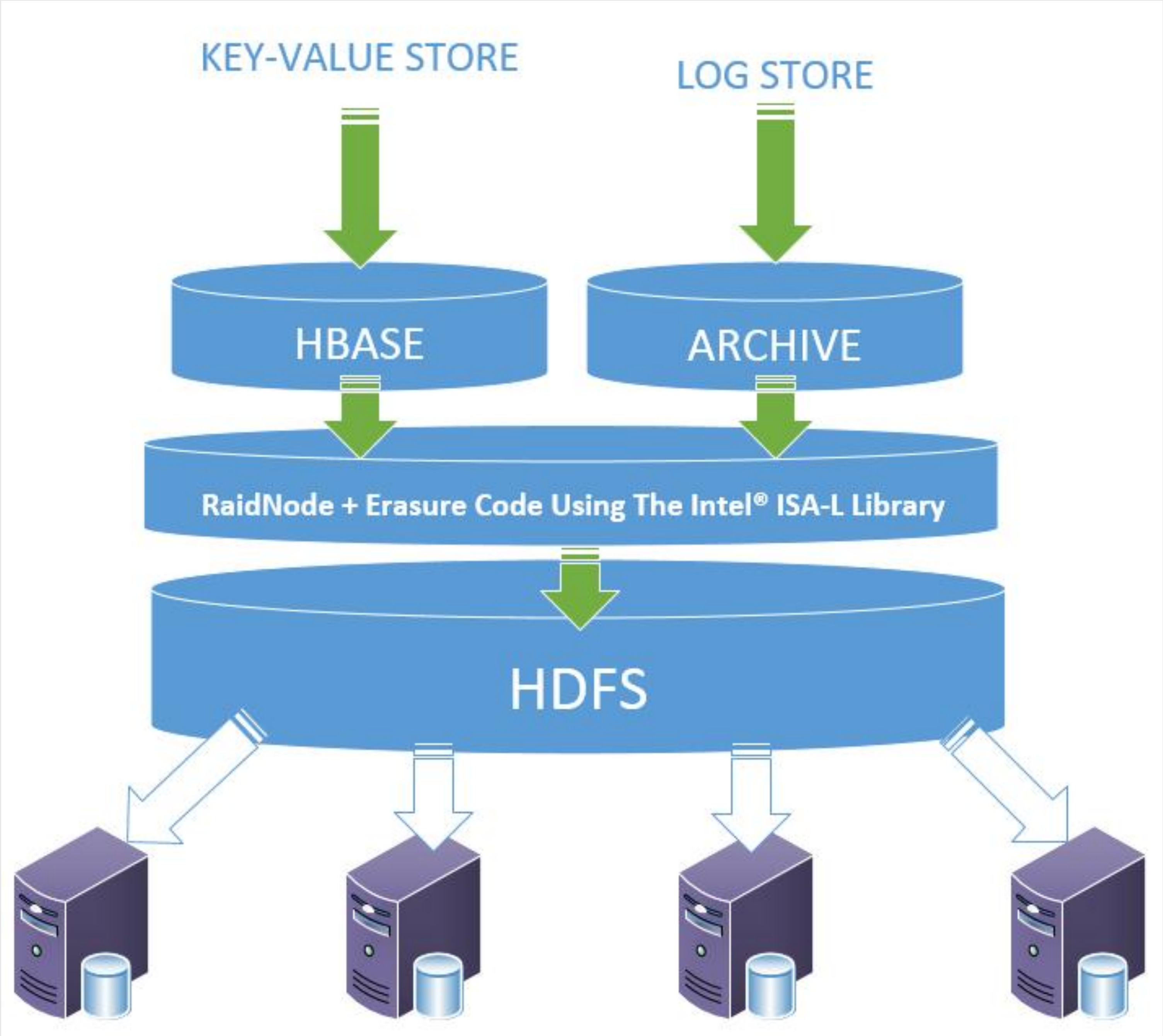
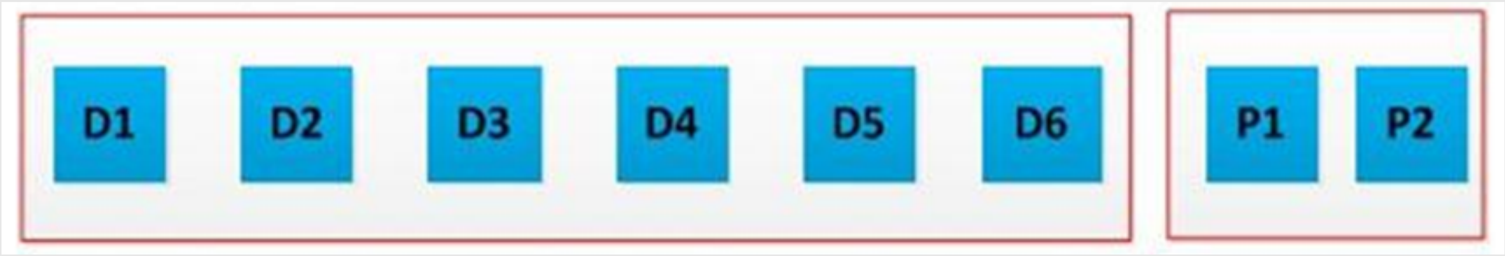


Diagram 2 Optimized Hadoop Big Data Storage Solution

To solve the big data storage requirements, Qihoo 360 optimized the IPDC architecture by adding several additional components. A reduction in the log storage space was accomplished with the help of Hadoop Archive which packages many small files into a single large file. Erasure code was implemented which allows the data to be broken into many smaller pieces, along with parity bits, and stored across many servers. This reduced the required disk capacity while maintaining redundancy through RAID striping. RaidNode was also implemented to manage the health of the parity files generated by the erasure code. Lastly Intel® ISA-L was implemented for performance improvements including optimization of the erasure code functions, which requires increased overhead. Overall these changes resulted in improved storage efficiency and cost savings for Qihoo 360.

Prior to optimizations (Diagram 1), if Qihoo 360 had 10 GB of data that needed to be stored, then 30 GB of actual space was required to accommodate the 3-copy redundancy solution. By implementing RaidNode and erasure code (Diagram 2), the 10 GB of data with data protection only requires 13 GB of space, reducing storage requirements by nearly two-thirds as compared to the previous 3-copy solution.

Here is a general example of data striping:



6 data blocks generate 2 CRC blocks, and the data blocks are recoverable if up to 2 blocks are lost. (This is configurable.)

Implementing erasure code provided data protection and a reduction of storage space as compared to 3-copy, but at a cost to performance due to additional processing overhead.

Qihoo 360 took advantage of combining Intel® architecture with Intel® ISA-L providing a 50x improvement in the encoding/decoding performance of the erasure code compared to using Java*. This solution is transparent to the normal operations of the HBase* cluster.

Processor	Intel® Xeon® Processor E5-2630 @ 2.30 Gigahertz	
Redundancy	Erasure code using a 10+4 data stripe with Java JDK 1.6	Erasure code using a 10+4 data stripe with Intel® ISA-L version 2.8
File System	Hadoop/HDFS version 2.0	
Encoding (Single-Node on a Single-Core) @ 100% CPU Utilization	**30 Megabytes per second**	**1.5 Gigabytes per second**
Decoding (Single-Node on a Single-Core) @ 100% CPU Utilization	**31 Megabytes per second**	**1.6 Gigabytes per second**

Summary

Intel® ISA-L provides an opportunity to help customers gain better performance from Intel processors with a lower investment in development. The Intel® architecture based storage solution implemented by Qihoo 360 satisfied their entire functional requirements of Qihoo 360 while reducing system complexity, and scaling with future increases in network speed.

Acknowledgements

As a result of this successful collaboration with Intel, Qihoo 360 said the following: "*Qihoo 360 focuses to provide high quality services freely for China internet users to resolve various internet security issues. With the fast growth of our user base and our business product lines, we need to build up a powerful and stable IPDC to provide strong support for our business. We are glad that we engaged with Intel to apply Intel® Intelligent Storage Acceleration Library to improve the data storage and analysis capability of our IPDC, which will help us to maintain our tech leadership in industry and to grasp the new opportunities to grow our business.*"

The author would like recognize the following individuals for their contributions to this article.

Bruce Chen, Yanfeng Mu, Zhongyu Li, Taylor Kidd, Quoc-Thai Le, Pujiang He, and Belinda Liviero.

The Author: David Mulnix is a software engineer and has been with Intel Corporation for over 15 years. His areas of focus have included software automation, server power and performance analysis, and cloud security.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information. The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web Site http://www.intel.com/.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance.

*Other names and brands may be claimed as the property of others.