I/O Performance Measuring – White box test and Black box test-

U.S. Department of Energy Best Practices Workshop on File Systems & Archives
San Francisco, CA September 26-27, 2011 Position Paper

FUJITA, Naoyuki
JAXA: Japan Aerospace Exploration Agency
fujita@chofu.jaxa.jp

SOMEYA, Kazuhiro
JAXA: Japan Aerospace Exploration Agency
someya.kazuhiro@jaxa.jp

ABSTRACT

The complexity of the component of file system/storage system (Thereafter, called the system.) is given to one of the reasons that the I/O performance measurement doesn't generalize. Here, let's think about the scene that discusses the I/O performance. Two cases are greatly thought. One is characteristic grasp of the system, and another is comparison between systems. We insist on using the white box test and the black box test properly in this paper. It is necessary to understand a detailed characteristic of the system by the white box test to guide an appropriate I/O operation to the system user and for the I/O tuning. On the other hand, when other systems and one system are compared, you should start from black box test comparison for a constructive discussion because of each system has each design and architecture.

INTRODUCTION

CPU benchmarking is widely discussed and some major benchmark suites[1,2] exist. However, I/O benchmarking is not more general than that of CPU. Therefore, generally speaking, I/O performance measuring and discussion are difficult. In this paper, we insist on using the white box test and the black box test properly.

As you know, white box test is a test done under the design and architecture of the system is understood. On the other hand, black box test is a test done without requiring them. In case of this time, design and architecture is a component, and the connection relationship of the system such as file systems and the storage devices.

WHITE BOX TEST EXAMPLE

This chapter shows examples of the white box test strategy and result. One result is a system that was operating in 2000(Thereafter, called 2000 System), another one is a system that has been operated since 2010(Therefore called 2010 System).

White box test strategy

As a number of nodes increases in HPC system, the system design and architecture becomes complex and changes it's characteristic. We propose layered benchmark as white box test, and show some results. Layered means device level(Measuring Point 1), local file system level(Measuring Point 2), network file system level(Measuring Point 3), and FORTRAN level(Measuring Point 4). Fig. 1 shows measuring points on recent HPC System.

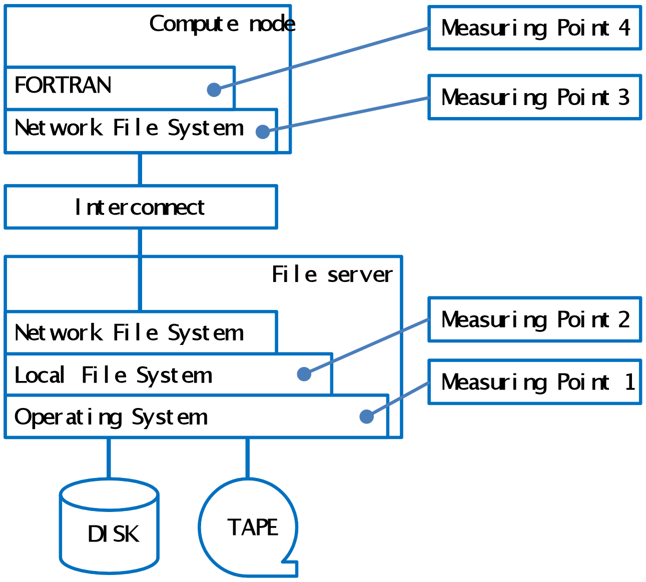
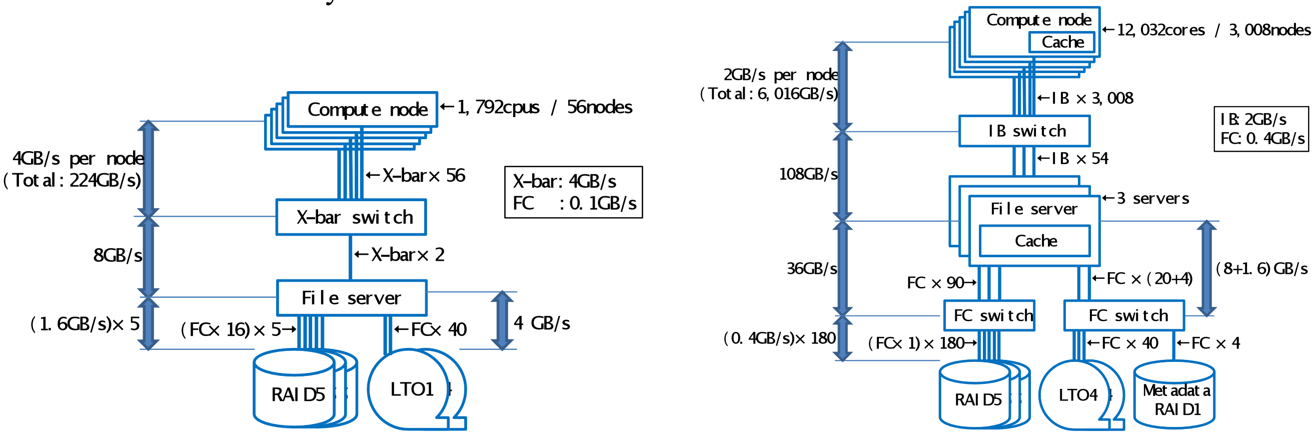


Fig. 1 HPC System I/O measuring points

White box test results

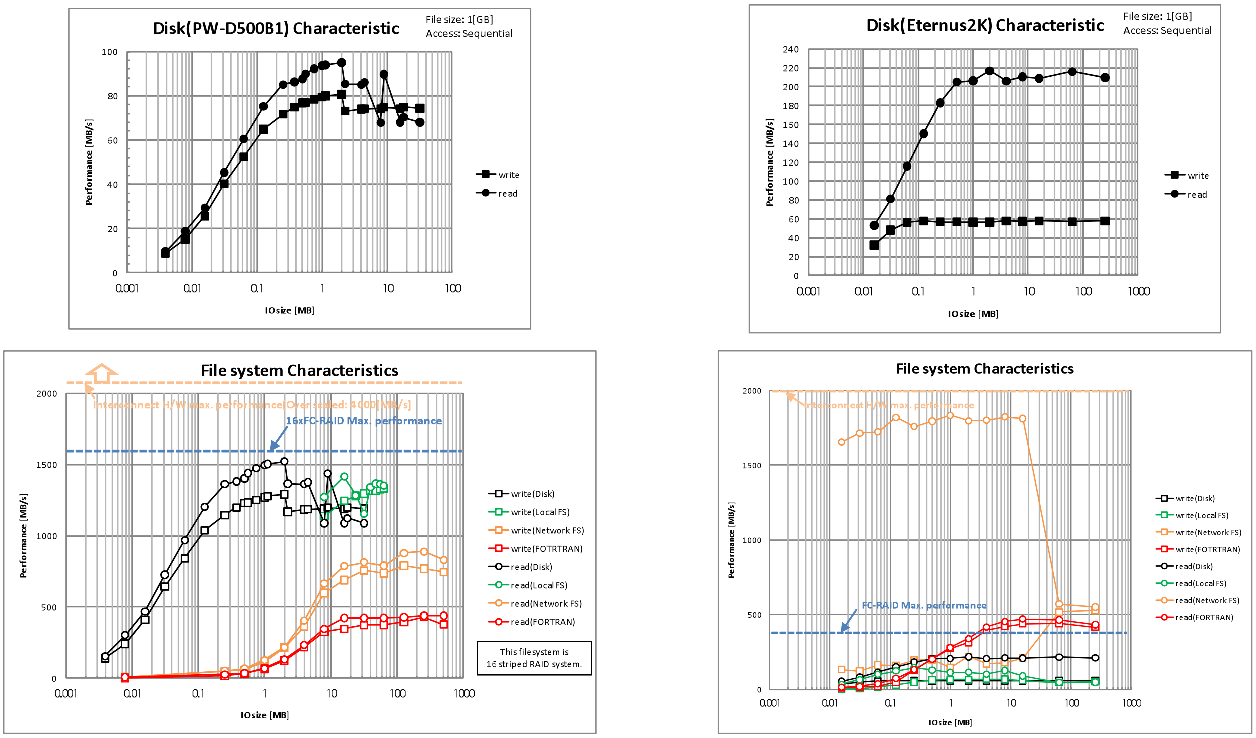
Fig. 2 shows 2000 and 2010 system configuration chart. Each system has Compute nodes, Interconnect, which are X-bar switch and IB switch, and File server(s). So there is no change in a basic composition. As a partial change point, 2010 System has clustered file servers and storage devices are attached via FC-SAN switches. Fig. 3 shows the result of the layered benchmark of these two systems. There are some bottleneck points in a system. To analyze a bottleneck, we aim at file system cache, interconnect bandwidth, and DAS/SAN bandwidth and its connection relationship design.

Device level benchmark showed similar result between two designs except disk write performance. But the characteristic of network file system level was very different. In this case, it depends on file system cache effect.



(a) 2000 System (b) 2010 System

Fig. 2 File system and storage system configuration



(a) 2000 System (b) 2010 System

Fig. 3 Layered benchmark results

BLACK BOX TEST EXAMPLE

This chapter shows examples of the black box test strategy and result.

Black box test strategy

As we said, each system has each design and architecture, so the comparison of simple I/O performance is not significant. But when we discuss about I/O performance, especially compare with several systems, first of all, it should take a general view of a rough performance. A common tool to measure the file system performance that is appropriate for the measurement of large-scale storage doesn't exist, and the performance measurement tool is made individually in each system and the performance is evaluated individually. In addition, as a peculiar operation to the file system will be needed, it is difficult to compare it with the performance measurement result in another file system. Then, we model the measurement tool and the measurement item, and propose the method of simply diagnosing the characteristic of the large-scale storage system based on the result of a measurement that uses the tool[3].

Objective

It aims at the thing that the following two points can be measured generally in a short time.

(1)Checkup of installed system

Whether the performance at which it aimed when the system administrator installed the system has gone out is examined.

(2) Routine physical examination under operation

Grasp of performance in aspect of user. The operation performance is measured.

Diagnostic model

(1)Checkup of installed system

-Maximum I/O bandwidth performance

Read performance immediately after Write. It is assumed that data are in the client cash.

-Minimum I/O bandwidth performance

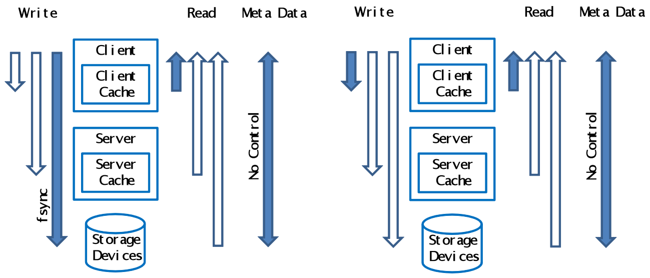
Fsync is assumed after Write and the multiplication cash assumes all things forwarded to the real storage device.

-Meta data access performance

The presence of cash is not considered (Because the cache management cannot be controlled in the black box test).

(2) Routine physical examination under operation

This diagnosis tool is regularly made to work while really operating it, and the state grasp is enabled. In this case, it is assumed to gather the maximum performance (cash hit performance) from the viewpoint of the user aspect. An enough prior confirmation by the system administrator is necessary to make the measurement tool work regularly. Moreover, customizing the measurement tool (measurement downsizing etc.) might be needed. The measurement model is shown in Fig. 4.



(1)Checkup of installation (2)Routine examination

Fig. 4 Measurement model

Measurement tool and item

Using IOR.

(a)Large-scale data transfer (Throughput performance: Constant amount of file for each process, large I/O length)

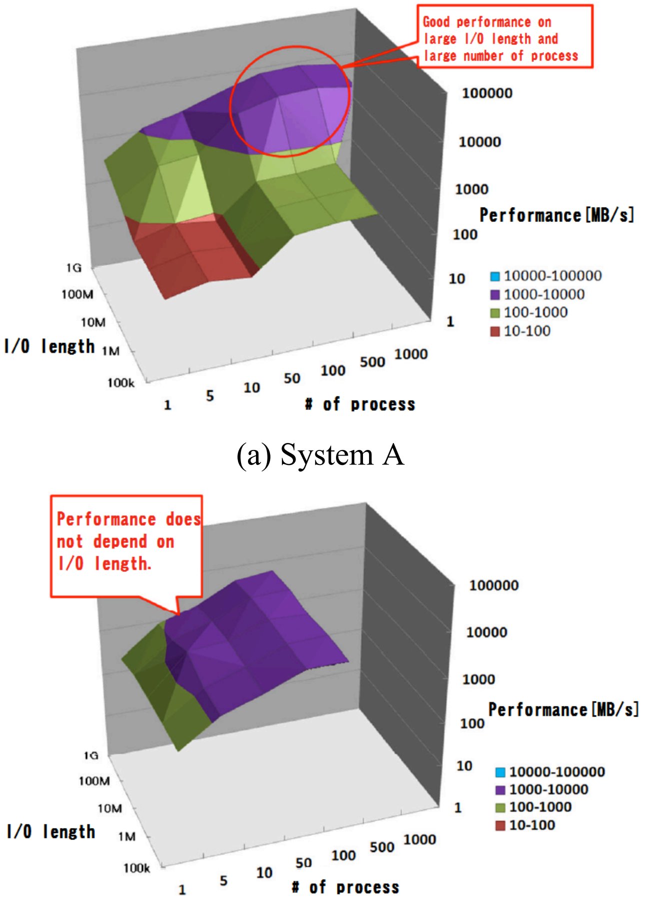
(b) Constant volume of data (Throughput performance: Small file size, small I/O length)

Using mdtest.

(3) Meta data access (response performance)

Black box test results

Fig. 5 shows the result example of large-scale data transfer on System A and B.


Good performance on large I/O length and large number of process

(a) System A

Performance does not depend on I/O length.

(b) System B

Fig. 5 Large-scale data transfer results

CONCLUSIONS

"What should be measured?"

Each layer benchmark should be done when we want to know the characteristics of the system. The Layers are device level, local file system level, network file system level, and FORTRAN level. This kind of measurement will be done as a white box test.

Modeled measurement item and tool should be used, when we want to compare several systems. The result is a starting point of the discussion. This kind of measurement will be called black box test.

Both white box test and black box test should be used when we manage file system and storage system.

REFERENCES

1. Standard Performance Evaluation Corporation http://www.spec.org/benchmarks.html
2. TOP500 supercomputer Sites, http://www.top500.org/

Scientific System Society, "Large-scale storage system working group report," 2011,(in Japanese)