ANL/ALCF/TM-13/1

ALCF I/O Data Repository

Argonne Leadership Computing Facility
Technical Memorandum ANL/ALCF/TM-13/1

About Argonne National Laboratory

Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, see www.anl.gov.

Availability of This Report

This report is available, at no cost, at http://www.osti.gov/bridge. It is also available on paper to the U.S. Department of Energy and its contractors, for a processing fee, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone (865) 576-8401
fax (865) 576-5728
reports@adonis.osti.gov

ANL/ALCF/TM-13/1

ALCF I/O Data Repository

by
Philip H. Carns
Argonne Leadership Computing Facility

February 2013

This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

Contents

ALCF I/O Data Repository

by
Philip H. Carns

Abstract

The Argonne Leadership Computing Facility (ALCF) I/O Data Repository is a collection of anonymized production I/O activity logs from the ALCF. This report provides an overview of the repository, including how to interpret the data and how to acknowledge its use.

1. Data Formats

The ALCF I/O Data Repository includes two distinct types of data. The first is application characterization data collected using the Darshan characterization tool. The second is file server block device usage data collected using the input/output statistics (iostat) utility.

A. Darshan Logs

Darshan is an application-level I/O characterization tool. It intercepts I/O function calls in user space and records access pattern information before the I/O operations are interpreted by the operating system or file system. Darshan *does not* record a complete trace of all I/O operations. Instead, it records a fixed-size collection of statistics for each file that is opened by the application. This includes information such as the number of write operations, the amount of time consumed by I/O operations, and the file system type. Darshan generates a separate log file for each job. The design and capabilities of the Darshan utility are described in greater detail in our previous work [1].

Darshan log files are stored in a compressed binary format. The following tools are provided as part of Darshan [2] to assist in parsing and interpreting the log files:

1. `darshan-parser`: command-line utility that converts entire log files into human-readable text format. More information about the darshan-parser output format can be found in the online Darshan Documentation [3].

2. `darshan-job-summary.pl`: command-line utility that produces a summary of the I/O behavior of an application in PDF format. The

summary includes several graphs and tables that can be used as a starting point for further analysis.
3. `darshan-logutils.h`: a C language application program interface (API) for use by analysis utilities.

The following fields have been anonymized in each log file and replaced with strings of numbers:

- job id
- uid
- exe (command line)
- file name suffix
- project name (annotated in the metadata portion of the header for some logs)

The anonymization has been performed in a consistent manner so that log files can still be grouped consistently by uid or by project.

B. Block Device Logs

File server block device logs are the second source of data available in the ALCF I/O Data Repository. In contrast to the Darshan data, the block device logs are *not* application-specific. They record all traffic on a file system, whether it is generated by an application, by a combination of applications, or by a system maintenance activity. The data includes bandwidth, number of bytes read and written, device utilization, and average response times across all block devices in the storage system.

The block device data was collected using the `iostat` command-line tool [4]. We used a small set of wrappers (known as *iostat-mon*) to gather iostat data on each file server. Data was collected every 60 seconds, logged in a compact format, and then post-processed to produce aggregate summaries. All local disk activity was filtered out.

Each iostat log file in the repository contains 24 hours of aggregate block device statistics from a file system. The data is stored in gzipped, tab-delimited text format. A header at the beginning of each file describes the format in detail. Each row represents a 60-second time interval. The first column is the time stamp of that interval in Unix seconds format. The following columns are summed across all devices:

- r/s: reads per second
- w/s: writes per second
- rMB/s: read MiB per second
- wMB/s: write MiB per second
- MB_r: total MiB read

- MB_w: total MiB written

The following columns are averaged across all devices:

- svctm: average service time in ms
- %util: percent utilization

2. System Descriptions

The data repository consists exclusively of data captured from the Intrepid Blue Gene/P system at Argonne National Laboratory. Intrepid is a 163,840-core production system with 80 TiB of RAM and a peak performance of 557 TFlops. The primary high-performance storage system employs 128 file servers running both a Parallel Virtual File System (PVFS) and a General Parallel File System (GPFS). Data is stored on 16 DataDirect Networks S2A9900 storage area networks (SANs). The storage system has a total capacity of 5.2 PiB and a peak I/O rate of approximately 78~GiB/s. The architecture and scalability of this storage system have been analyzed in detail in a previous study [5].

3. Data Sets

The data sets are available for download via file transfer protocol (FTP) [6]. The `intrepid` subdirectory contains application Darshan logs organized into a directory tree by year/month/day.

Note: The Darshan logs do not cover 100% of all applications that execute on Intrepid.

Although Darshan is enabled by default for all users, a number of factors may prevent a log from being captured. Darshan only instruments Message Passing Interface (MPI) applications and it only produces a log if the application successfully calls the `MPI_Finalize()` function. Users may also opt out of using the Darshan tool. The coverage rate of Darshan varies anywhere from 20% to 80% from week to week.

The `intrepid-iostat/summaries` directory contains iostat block device logs named by the year, month, and day. The log files record traffic on the high-performance GPFS and PVFS storage volumes. They *do not* include any home directory or archival traffic.

The logs that cover the time range of January 1, 2010 to March 30, 2010 were analyzed in detail by Carns et al. [7]. Their study provides several examples of how to analyze Darshan and iostat data.

The `wget` utility may be helpful for downloading many log files at once. For example, to download all March 2010 Darshan logs from the Intrepid collection, you would use the following command:

```
wget -r ftp://ftp.mcs.anl.gov/pub/darshan/data/intrepid/2010/3
```

Other FTP clients may also be able to perform recursive downloads in a similar manner.

4. Acknowledgment and Citation of the ALCF I/O Data Repository

Please include the following acknowledgment in any publications that use data from the ALCF I/O data repository:

> This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

We also suggest citing the following publications as needed:

- Example system I/O study and description of data collection methodology: *K. Harms, W. Allcock, C. Bacon, S. Lang, R. Latham, and R. Ross. 2011. "Understanding and improving computational science storage access through continuous characterization," ACM Transactions on Storage (TOS). 7(3):8*. [7]
- Initial design of the Darshan characterization tool *P. Carns, R. Latham, R. Ross, K. Iskra, S. Lang, and K. Riley. 2009. "24/7 characterization of petascale I/O workloads," in Proceedings of 2009 Workshop on Interfaces and Architectures for Scientific Data Storage, September*. [1]

References

[1] P. Carns, R. Latham, R. Ross, K. Iskra, S. Lang, and K. Riley. "24/7 characterization of petascale I/O workloads" in *Proceedings of 2009 Workshop on Interfaces and Architectures for Scientific Data Storage*. September 2009.

[2] "Darshan download page." http://www.mcs.anl.gov/research/projects/darshan/download/

[3] "Guide to darshan-parser output." http://www.mcs.anl.gov/research/projects/darshan/docs/darshan-util. html# guide to darshan parser output

[4] S. Godard. "SYSSTAT utilities home page." 2010. http://pagesperso-orange.fr/sebastien.godard/

[5] S. Lang, P. Carns, R. Latham, R. Ross, K. Harms, and W. Allcock, "I/O performance challenges at leadership scale" in *SC '09: Proceedings of the*

Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA: ACM. 2009. pp. 1–12.

[6] “FTP site: Darshan data.” ftp://ftp.mcs.anl.gov/pub/darshan/data/

[7] P. Carns, K. Harms, W. Allcock, C. Bacon, S. Lang, R. Latham, and R. Ross. 2011. “Understanding and improving computational science storage access through continuous characterization.” *ACM Transactions on Storage (TOS)*. 7(3):8.

Argonne Leadership Computing Facility
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 240
Argonne, IL 60439-4847

www.anl.gov