I/O characterization of large-scale applications with Darshan

Philip Carns and Robert Ross (PI), Argonne National Laboratory
{carns,rross}@mcs.anl.gov

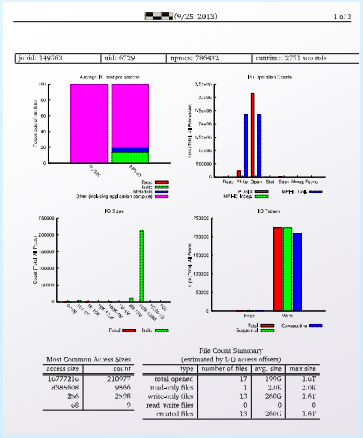
Overview

Darshan is a lightweight, scalable I/O characterization tool that transparently captures I/O access pattern information from production applications.

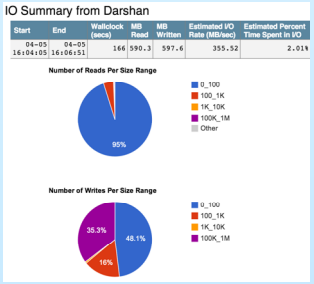
Darshan automatically collects production data on both the Mira IBM Blue Gene/Q at the Argonne Leadership Computing Facility and the Hopper Cray XE6 at the National Energy Research Scientific Computing Center. Data collected with Darshan can be used to tune applications, understand system usage, and guide I/O research activity. The most recent version of Darshan (2.2.8) was released in September 2013. It is portable across a wide variety of architectures, compilers, and MPI implementations.

http://www.mcs.anl.gov/darshan/

Darshan produces a separate characterization log for each instrumented application. Command line utilities can be used to produce graphical summaries. This example shows a 768,384 process turbulence application on Mira.



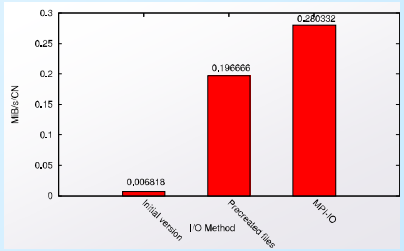
Darshan records a concise collection of access pattern, timing, and performance data. This example application used MPI-IO collectives to transform ~2.5 million write operations into ~200 thousand optimized file system accesses.



Darshan can also be integrated into existing system diagnostic tools. This screenshot from the NERSC web portal provides feedback on I/O behavior to users as soon as a job is completed.

Application performance engineering

Darshan fundamentally changes the approach to I/O performance engineering on large-scale systems. When an I/O performance problem is observed, the scientists and I/O experts involved can immediately refer to the Darshan report for initial diagnosis. Prior to the introduction of Darshan, the first step was to either rerun the application with additional instrumentation (costly in terms of CPU time) or to inspect the source code of the application (costly in terms of manpower).



This example shows the per compute node I/O performance improvement in a combustion physics application on Intrepid that was tuned based on feedback from Darshan. Darshan characterization indicated that a file creation bottleneck was the limiting factor in performance.

Examples of other applications that have been tuned with the help of Darshan include:

- HSCD (combustion physics)
- FLASH (astrophysics)
- NekCEM (electromagnetics)
- PHASTA (unstructured meshes)
- pF3D (plasma physics)

Metric: Over 100 million small write operations to shared files, no use of MPI-IO collectives.	
Jobs analyzed	261,890
Jobs matching metric	220
Users matching metric	11

Data from Darshan can be used not only to investigate specific applications, but also to simply find candidate applications that could benefit from additional tuning. This table shows an example of Darshan log filtering on Hopper, in this case identifying candidate applications with poor write behavior. The most extreme example was issuing over 5 billion independent writes of less than 100 bytes each to shared files.

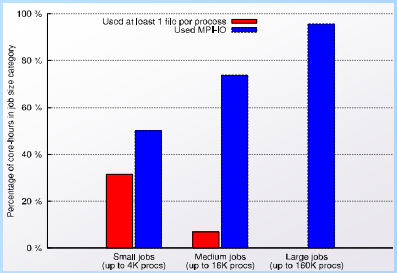
Key Darshan collaborators:

- Kevin Harms (ALCF)
- Yushu Yao (NERSC)
- Katie Antypas (NERSC)

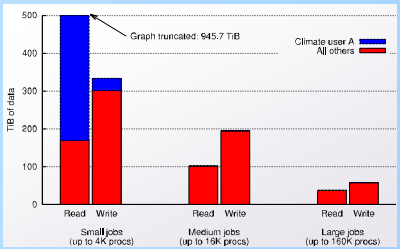
Understanding system usage

Darshan data can be analyzed in aggregate to better understand how systems are being used. This information helps to guide procurement decisions and I/O research.

Data collected on Intrepid illustrates how MPI-IO usage becomes more prevalent in large scale applications, while file-per-process access patterns become less prevalent.



Data collected on Intrepid illustrates that I/O access is generally write intensive at all scales. This example also shows how a single application or user (labeled "Climate A" here) can dominate I/O traffic on a production machine.



Contributing to I/O research

Darshan includes anonymization tools to aid in sharing data within the I/O research community. The ALCF I/O Data Repository provides over 150 thousand production Darshan logs for public use.

http://www.mcs.anl.gov/research/projects/darshan/data/

ALCF I/O Data Repository Statistics	
Unique log files	152,167
Core-hours instrumented	721 million
Data read	25.2 petabytes
Data written	5.7 petabytes