Quantifying HPC and I/O System Balance

A. Uselton, J. Hick ({acuselton,jhick}@lbnl.gov)

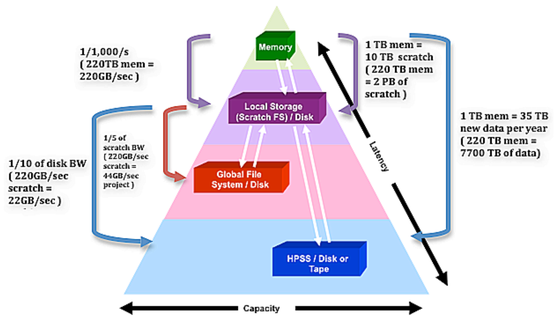
A. Gamst (acgamst@math.ucsd.edu)

Overview

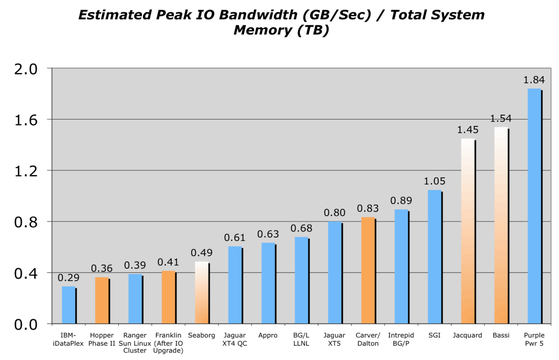
There is a balance point in the design of a high performance computer (HPC) system at which the contribution to performance of the mass storage (I/O) infrastructure is neither too large nor too small. That balance point has commonly been identified heuristically. Our contribution quantifies the balance point by examining the relative costs and impact of compute resources versus I/O resources.

If storage is a bottleneck, improving the I/O capability can raise system utilization, increasing throughput. The balance point is where the cost of increasing throughput by adding I/O capability is the same as the cost of doing so by adding nodes.

Heuristics for System Balance



a) Heuristics



b) $1GB/s$ per TB

Figure 1

Memory capacity is the key determinant of necessary I/O bandwidth and capacity [2]. Figure 1(a) presents the heuristic that the file system should be able to move all of memory in about 1000 seconds. Figure 1(b) presents this heuristic as applied to several HPC systems.

A Model for System Balance

Simplifying assumptions:

- The cost model for compute and I/O capability is linear.
- A job is either doing computation or I/O, and its run time is the sum of its compute time and its I/O time.
- A job carries out its I/O at the optimum rate for the storage system.

Definitions:

- P_n is the cost of a node
- P_r is the cost of a "unit" of bandwidth
- The cost of the system is $nP_n + rP_r$
- U is the system utilization - the fraction of node-seconds spent in computation

Results:

- The marginal cost of adding compute capability by adding nodes is P_n/U
- The marginal cost of adding compute capability by adding I/O capability is $\frac{P_r}{N}\frac{R}{1-U}$ (see [1])
- The system balance is the ratio of these two $B = \frac{NP_n}{RP_r}\frac{1-U}{U}$

Carver: A Case Study

Carver (see Figure 1b)

Budget for compute	85%	for I/O	15%
Total mamory	$30TB$	target bandwidth	$30GB/s$
Actual bandwidth	$25GB/s$	fraction of target	0.83

The Carver IBM Dataplex cluster at NERSC was provisioned with approximately 15% of its budget dedicated to I/O infrastructure[3], which gave it about 83% of the target suggested by the heuristic in Figure 1a.

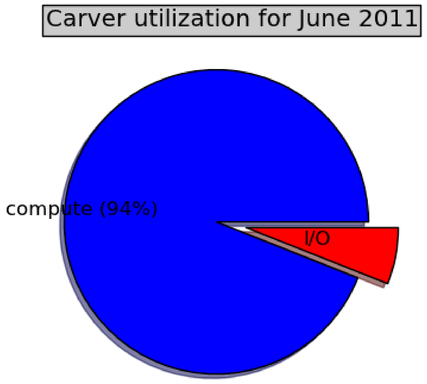


Figure 2

Carver runs the Integrated Performance Monitoring (IPM) library [5], which provides the utilization U. That results in a balance factor $B = 2.5$.

Challenges

The nodes of a job may not divide their time between computation and I/O as cleanly as represented here. Write-back cache, staged I/O, and asynchronous I/O operations all attempt to hide I/O delays from the application, improving utilization.

Time may be spent in neither computation nor I/O, such as communication delays. Communication may compete for bandwidth with I/O.

The model assumes that all I/O proceeds at its optimum rate, but that is not what most systems observe. Thus the marginal cost of increasing bandwidth may be significantly higher.

The delay caused by I/O may depend on the latency of metadata operations. The model and analysis will want to track all forms of delay brought on by file system activity.

These challenges will be addressed as the work goes forward. In some cases it only requires an obvious, if awkward, additional detail in the model. In other cases the data collected to characterize the workload may need to be extended.

Future Work

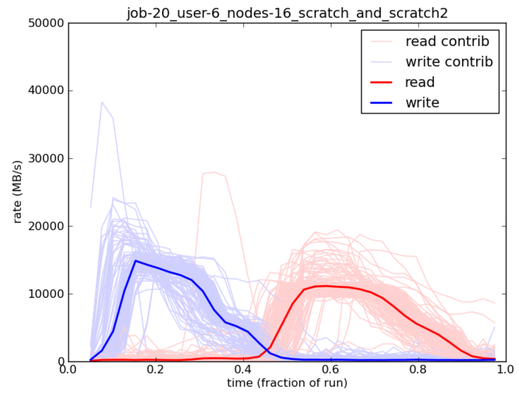


Figure 3. The average I/O behavior of many runs of a job.

Most HPC systems do not have IPM or other direct measures of the utilization. An alternative strategy infers the utilization U from file system server monitoring data. The Lustre Monitoring Tool (LMT) [4] collects server-side data, which is anonymous with respect to the jobs running on the compute nodes. Nevertheless, it is often possible to infer the job from the I/O pattern.

Figure 3 shows an estimate for the average I/O behavior during 175 instances of a job running the IOR[6] file system benchmark. The dark colors represent the average behavior, blue for writes, and red for reads. The lighter colors show the individual instances.
When most jobs, accounting for most I/O activity, have been given this treatment, the result is a comprehensive, job-by-job characterization of the actual I/O workload, which will allow us to calculate the utilization U and therefore the balance factor B. Note that the balance factor would vary over time as the workload varies.

References

[1] A. Uselton, Quantifying HPC and I/O Sytem Balance, Technical Report, In preparation
[2] LLNL Internal Study on Memory Capacity to Archive Data 2008 - 2009
[3] J. Broughton, Private communication, We would like to thank Jeff for discussions on the procurement budgeting process.
[4] A. Uselton, Deploying Server-side File System Monitoring at NERSC, Cray User Group Conference, Atlanta, GA, 2009
[5] D. Skinner Integrated Performance Monitoring: A Portable Profiling Infrastructure for Parallel Applications Proc. ISC2005: International Supercomputing Conference Heidelberg, Germany, 2005
[6] H. Shan and J. Shalf, Using IOR to Analyze the I/O Performance of HPC Platforms, Cray Users Group Meeting (CUG) 2007, Seattle, Washington, May 2007

Funding

This work was supported in part by the National Energy Research Scientific Computing Center (NERSC), under Contract No. DE-AC02-05CH11231.