SANDIA REPORT
SAND2019-12564
Printed October 2019

Sandia National Laboratories

Figures of merit for production HPC

Benjamin A Allan

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods/

ABSTRACT

This report summarizes a set of figures of merit of interest in monitoring the hardware and hardware usage in a Sandia high performance computing (HPC) center. These figures are computable from high frequency monitoring data and other non-metric data and may aid administrators and customer support personnel in their decision processes. The figures are derived from interviews of the HPC center staff. The figures are in many cases simplistic data reductions, but they are our initial targets in creating dashboards that turn voluminous monitoring data into actionable information. Because simplistic reductions may obscure as well as reveal the situation under study, we also document the necessary 'drill-down' and 'exploration' views needed to make the data better understood quickly. These figures of merit may be compared to dashboarding tools documented by other HPC centers.

ACKNOWLEDGEMENTS

We thank the staff of Sandia's production HPC department for their survey input.

CONTENTS

This page left blank

ACRONYMS AND DEFINITIONS

Abbreviation	Definition
BW	Bandwidth (data moved/second)
FoM	Figure of merit
HPC	high performance computing
HW	hardware
IB, HSN	Infiniband (or OmniPath) high speed network
LDMS	Lightweight Distributed Metric Service
MB, GB	Megabyte, gigabyte (2^n values)
NFS	Network file system
RAM	Main memory (random access memory)

1. SURVEY RESULTS TABULATION

We conducted a local survey of HPC system administrators to see what their *first* figures of merit would be, and we present the results here for our peers working on monitoring systems development. Most of those surveyed stressed as part of their responses that they would have additional and more complicated responses if they saw an infrastructure deployed which deals efficiently with these. Thus, the set presented here is not considered *complete* in any sense.

The results presented here are grouped by the scope of information: compute job, system (a single cluster), or center (all clusters sharing the same network resources). Omitted from the results is the idea of results computed across a specific set of related jobs with a single underlying simulation model run broken in pieces by the dictate of maximum job wall-clock time per submission; we assume a reasonable dashboard implementation will be able to composite such a run set into a single job report.

Tabulated below is a description of the information desired by support staff. While most information items require a numeric value (figure of merit) to be computed, some are lists of items with a logical rather than numeric status. In these cases, the underlying figure is obvious. Consult the author for further details.

Table 1-1. Center Scope Figures of Merit

Figure of Merit	Drill-down	Exploration
Per user aggregate read/write flow to NFS	list of jobs/logins by that user sorted by volume contributed	NFS rate histograms and timeseries from job data
Count of down/unusual network link status by system	Per system list of odd links not accounted for by down nodes or known maintenance in progress.	Link error histograms & timeseries (per link or collective)
Aggregate read/write/operations volume to Lustre per target	list of user/logins per metric sorted by volume	Lustre activity by user & job
Per user aggregate read/write volume to Lustre per target	list of jobs/logins writing by that user sorted by volume	Lustre activity by job
Per user %idle list of jobs by that user	list of jobs by that user sorted by %idle	single-job stats page
% link bandwidth saturation, error counts in storage IB subnet	histogram %time vs %used and time series for BW	job list active during current, peak intervals
Per user login node %user	list of nodes with user hogging CPU persistently	per-node history and name of binaries in use

Administrators want all the system scope figures in Table 1-2 to be presented only after they have passed a "what is new" filter. Any item that is known to be already being addressed should be hidden or sorted to the bottom of the list. Marking the item "in work" for the dashboard hide or sort should be automatic from an associated break-fix information database, not a per-client user interface setting.

Table 1-2. System Scope Figures of Merit

Figure of Merit	Drill-down	Exploration
Peak die temperature per node (node list beyond high limit)	1 hour, 24 hour, 1 week time history by node, rack map	related metrics (power, fan, enclosing rack)
Fan failures list	1 hour, 24 hour, 1 week time history by node, rack map	related metrics (power, temperature, enclosing rack)
Power supply failures list	1 hour, 24 hour, 1 week time history by node, rack map	related metrics (fan, temperature, enclosing rack)
Ethernet switch errors beyond high limit threshold (list)	histogram (log(10)x24 hourly bins)	time history, related metrics e.g. temperature
IB switch disappearance from data list	system connectivity map of error/missing	time history of link errors and other status metrics
Lustre gateway traffic BW, memory, and CPU load uniformities	Node histograms	time history of link errors and other status metrics
Count of IB links with status regularly below threshold but non-zero	Link list	time history of link errors and other status metrics

The job scope figures presented in Table 1-3 are of primarily administrative (rather than application developer) interest. Those that are network flow oriented are most interesting if presented as a fraction of the maximum possible flow given the system topology, but the upperbound is in many cases dynamic and difficult to determine in a live system.

Table 1-3. Job Scope Figures of Merit

Figure of Merit	Drill-down	Exploration
user-retraining-needed (a boolean = [oom or no HSN usage or high NFS usage or high Ethernet usage or high proc_running or (%mem low and all %mem bandwidth low and %proc_running low)])	binaries used all contributing	all contributing FoM
Aggregate Lustre usages per target (read and write MB/s, operations/minute)	histogram %node time vs log2 bandwidth used/node (100 x log2(NIC rating) always)	node subsets with similar histogram, time charts (line, heat) (lustre, NIC) in similarity groups
HSN usage (GB/node/hour deducting lustre traffic)	histogram %node time vs log2 bandwidth used/node (100 x log2(NIC rating) always)	node subsets with similar histogram, time charts (line, heat) (lustre, NIC) in similarity groups
Ethernet usage (MB/node/hour)	histogram %node time vs log2 bandwidth used/node (100 x log2(NIC rating) always)	node subsets with similar histogram, time charts (line, heat) (lustre, NIC) in similarity groups
% idle	histogram %core-time vs %idle (100x100 scales always)	node subsets with similar histogram, time charts (line, heat) in similarity groups
% proc_running/core	histogram %node time vs %proc_running (100x2.1 always)	node subsets with similar histogram, time charts (line, heat) in similarity groups
% memory bandwidth used at bottleneck and HW location	L1/L2/L3 histogram %node-time vs %bandwidth (100x100 scales always)	for i in (L1,L2,L3) show node subsets with similar histogram, time charts (line, heat) in similarity groups
% RAM used	histogram %node-time vs %RAM (100x100 scales always)	node subsets with similar histogram, time charts (line, heat) in similarity groups
% top binary used	histogram %core time vs top N binaries (100xN); default N=5	bigger N or node subsets with similar histogram, Gantt charts by node similarity group

REFERENCES

[1] *Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications*, A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop, A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden, M. Rajan, M. Showerman, J. Stevenson, N. Taerat, and T. Tucker, IEEE/ACM Int'l. Conf. for High Performance Storage, Networking, and Analysis (SC14) New Orleans, LA. Nov 2014.

DISTRIBUTION

Click here, then press delete to remove this guidance statement.
Required. Must be on an odd-numbered page. SAND Reports submitted through R&A are automatically sent to the Technical Library; however, it still needs to be included on the distribution. Ensure a blank odd-numbered page is inserted prior to the back cover.

Email—Internal

Name	Org.	Sandia Email Address
James Brandt	9328	brandt@sandia.gov
Technical Library	01177	libref@sandia.gov

This page left blank

This page left blank

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.