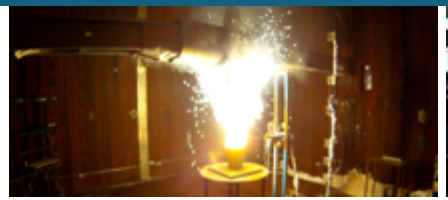
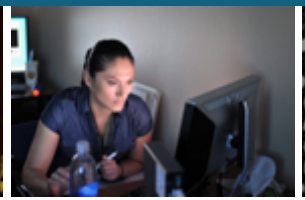
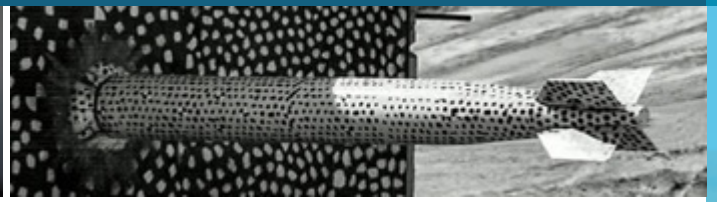
Sandia National Laboratories

SAND2020-13874PE

Monitoring, Metrics, and Analysis Integration







12/9/2020 Tri-lab CCE Meeting

Ben Schwaller & Jim Brandt

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.

Statement of Purpose

The Monitoring, Metrics, and Analytics Integration project increases the efficiency of NNSA HPC centers and aids future planning utilizing monitoring and analysis. Specifically, the project will:

- Deploy data collection and analysis infrastructure across the HPC center (clusters, applications, facilities, etc.)
- Develop portable analysis techniques (ML and otherwise) that can be applied to data gathered at multiple facilities to
- Derive figures of merit (FoMs) from monitored data that can guide and optimize decisions by resource managers, applications, administrators, and management.

FY21 Planned Activities (from ASC IP)

Explore the use of emerging technologies that would further enhance current monitoring infrastructures.

Continue working with NCSA to deploy performant analytic and visualization frameworks across tri-lab infrastructures.

Develop and deploy ML and statistical techniques along with application resource aware scheduling and resource allocation to improve production HPC workflows.

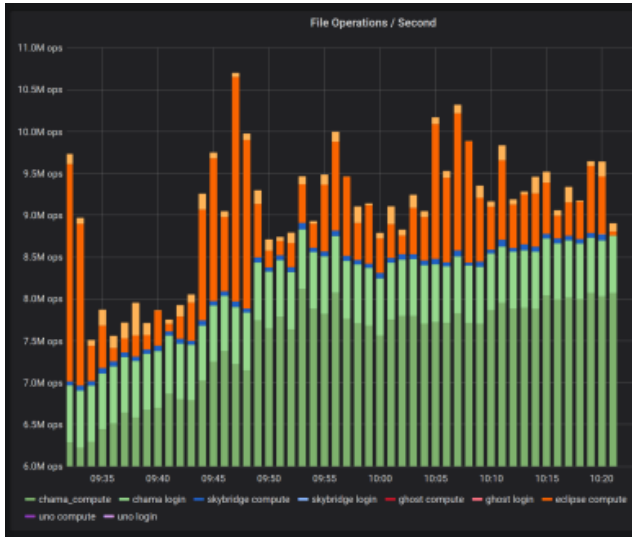
FY20 Meeting / Collaboration Accomplishments

CCE MMAI Monthly Meetings

Established confluence repository at LLNL to enable secure tri-lab participation and share analysis and visualization methods

Modified meetings to a presentation style with one or two presenters to discuss latest and greatest advancements from each of the lab communities on relevant topics

- SNL:
  - Grafana-based visualization and analysis pipeline for HPC system data (1)
  - LDMS pub-sub (Streams) capability and effort to fuse system and application data
  - ASC FY21 L2 milestone – “Integrated System and Application Continuous Performance Monitoring and Analysis Capability”
- LANL:
  - Using Machine Learning to identify application behavioral characteristics and anomalies related to those
- LLNL/SNL:
  - Demonstration of SPOT: a performance analysis orchestration layer
    - SPOT is developed at LLNL and is currently a collaborative effort between LLNL and SNL that is a potential collaboration point for the entire tri-lab



(1)

Center-wide visualization of Lustre file ops/second
Different colors represent different clusters

Contract with the National Center for Supercomputing Applications (NCSA)

Collaborated with SNL to implement and deploy the following on an analytics cluster with production CTS1 system data

- HTML-based metric browser along with drill-down filter capability to draw attention to behaviors of interest (e.g., low CPU utilization, high Lustre activity)
- Vitess-based in-memory backing store to facilitate low latency queries

Explored use of Machine Learning for application characteristic clustering and anomaly detection

- Obtained promising results but required a large number of Blue Waters nodes to process data

SNL University Collaborations

Boston University:

- Use of Machine Learning for detecting and diagnosing anomalous behavior in HPC applications
  - Began integrating models into production monitoring and analysis framework at SNL
- Use of High-Speed network statistics to drive improved network-intensive application placement

University of Central Florida:

- Use of characterizations of application Lustre resource utilization to drive scheduling and resource allocation decisions

Northeastern University:

- How to characterize High-Speed network and Lustre resource utilization on HPC systems - feeds into more intelligent scheduling and resource allocation methods

University of Illinois Urbana-Champaign:

- Use of Machine Learning to drive more efficient application execution through automated run-time adjustment of traffic injection

University Publications

Boston University:

- Machine Learning-Based Performance Analytics on High-Performance Computing Systems at SC20 Workshop (Machine Learning for Computer Systems), Nov. 2020

University of Central Florida:

- Towards workload-adaptive scheduling for HPC clusters, In Proceedings of the Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) Workshop at IEEE Cluster 2020, Sept. 14, 2020

University of Illinois Urbana-Champaign:

- Measuring Congestion in High-Performance Datacenter Networks, The 17th USENIX Symposium on networked Systems Design and Implementation (NSDI), Feb. 2020

FY21 Meeting / Collaboration Planned Activities

MMAI Meeting Plans

Continue Monthly CCE MMAI meeting format

- Foster communication between the labs on current and new HPC monitoring related endeavors

Evaluate SPOT container deployment strategies for each site and identify tri-lab collaboration points

- Need to address security concerns related to running a web-based server in containers on a per-lab basis

In-person face to face meeting in the Summer 2021 timeframe

- Subject to global pandemic

Collaboration Plans

Continue University collaborations on use of ML and other statistical methods on monitoring data to improve HPC operations

NCSA contract renewed for FY21

- Migration of System Administrator portal to Grafana with per-user authentication and data ingest from Kafka using common data format
- Continued work on ML-based clustering and anomaly detection

FY20 LDMS Accomplishments

Lightweight Distributed Metric Service (LDMS)

General Availability release of LDMS v4.3

- Packaged by TOSS maintainers at LLNL and included in TOSS 2020 Q1
- Deployed at LLNL, SNL, and LANL
  - v3.x.y is still mainstream on CTS1 platforms at LANL
  - SNL is in the process of upgrading to v4.3
- Deployed on Cori at NERSC

Community contribution growth

- LLNL, NERSC, LANL, SNL, HPE/Cray, Boston University, University of New Mexico, New Mexico State University in addition to Open Grid Computing

Continued bi-weekly Users Group meetings

LDMS New Features

- Pub-Sub interface (Streams) which enables asynchronous push of arbitrary data over LDMS communication infrastructure
- Sampler & related Store plugins
  - LLNL Lustre
  - Infiniband switch fabric
  - Job Characterization
  - Multi-tenancy Slurm
  - Job-scheduler Aware PAPI
  - System Hardware Performance Counter
  - NVIDIA GPU
- Cray Aries uGNI Resiliency Improvements
- Improved resiliency in the presence of node failure
  - Robust reference counting
- Support for Set Destroy
- LDMS Daemon Scalability and Performance Improvements
  - Reduce lock contention (use events and workers) so only workers take a lock (not a reader)
- Maestro LDMS Cluster Management Service
  - Dynamic load balancing among aggregators
  - Monitoring aggregator to detect failure and facilitate recovery
- TOSS Packaging Support
- Pandas/SciPy Analysis Infrastructure with Visualization Plugins
- Grafana Visualization Infrastructure
  - Data formatting based on display type
- Distributed Scalable Object Store
- OmniPath RDMA transport capability

LDMSCON 2020

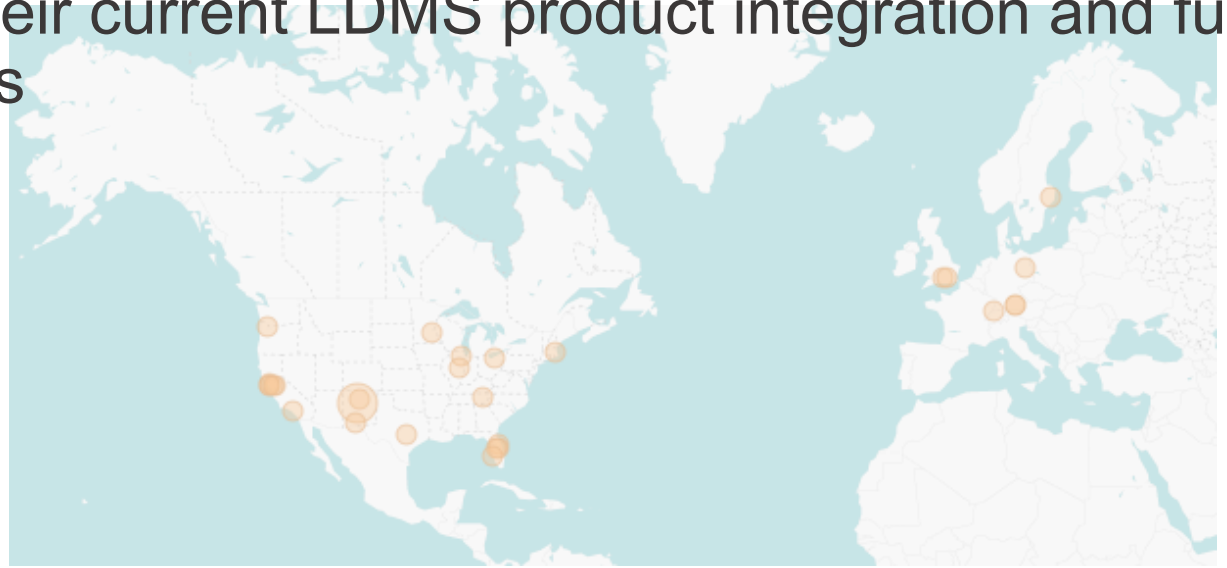
Virtual LDMS Users Conference (August 4-6, 2020)

Was attended by 60 representatives from 20 institutions from the US and Europe.

There were 8 user presentations and 6 tutorials.

Developer sessions were well attended and very interactive.

HPE described their current LDMS product integration and future plans in a keynote address



FY21 LDMS Planned Activities

LDMS Plans

Community-driven file-based configuration implementation

Load balancing aggregators with automatic fail-over

Distributed high-performance object store plugin

Facilitate deployment of LDMS HPE/Cray Shasta Platform

Code releases as new features or bug fixes are added

New planned samplers

- Migrate current samplers to be ARM-compatible and add additional ARM architecture samplers
- AMD CPU and GPU hardware performance counter
- Mellanox IB sampler
- Aries latency sampler
- Others as needed

ASC L2 Milestone: Integrated System and Application Continuous Performance Monitoring and Analysis Capability

Data Flow Diagram

Applications dynamically and irregularly inject data into the LDMS transport

LDMS continuously and regularly collects and transports full system data