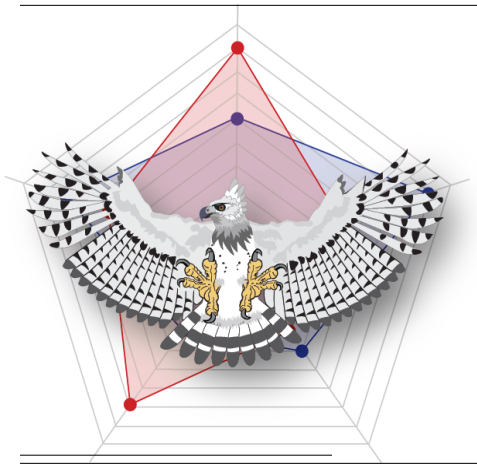


EMC3 and USRC Newsletter

Fall 2019 Edition

Inside this issue:

Introduction

Welcome to this inaugural **Efficient Mission-Centric Computing Consortium** bi-annual newsletter! The consortium launched in the Fall of 2018 and we couldn't be more excited about the number of participants who have joined, be it academic, systems and component developers/providers and HPC users/purchasers, in collaboratively pursuing mission-centric efficiencies. We look forward to welcoming more members over the next year as EMC3 grows and we continue to collaboratively pursuing greater efficient co-design activities.

EMC3 is both a logo and a mission statement. The word efficient addresses the fact that just simply spending money to attain more power/cooling/flops is a poor long-term strategy. Eventually, more efficient HPC solutions will be needed. Unfortunately, the co-design of application and hardware has largely become about fitting your applications to the latest hardware trends in industry. We seek real co-design where our current and emerging applications/workflows and the hardware/systems software/environment are synergistic. Mission-Centric also has meaning. While we at LANL can map some of our mission requirements into recent hyper trends in industry, such as Machine Learning/Deep Learning/AI, we still have a very large and critical need for simulations. The simulations that drive us are far too large to fit onto any system, involve very unstructured/irregular memory access, have many physics packages running simultaneously, and run at the largest scales for many months. By banding together with several large, potential US Government and Industrial Base members who have similar needs, we can pursue more efficient compute, with more useable ops per watt or per dollar. The third "C," consortium, is all about working together.

Through EMC3, like-minded technology developers/providers and users can:

- Foster initiatives to pursue technology development for common, novel solutions that will serve future Extreme Scale Mission-Centric Computing needs.

CONTACT US
usrc-contact@lanl.gov

Co-Executive Director for Partnering
Beth Kaspar
bethk@lanl.gov
(505) 309-1729

Co-Executive Director for Technical Operations
Nathan DeBardeleben, Ph.D.
ndebard@lanl.gov
(505) 412-1069

Los Alamos Research Park
4200 W. Jemez Rd.
Los Alamos, NM 87544

usrc.lanl.gov/emc3.php

This newsletter has been approved for unlimited release and has been assigned LA-UR-19-31291. This work has been authored by an employee of Triad National Security, LLC which operates Los Alamos National Laboratory under Contract No. 89233218CNA000001 with the U.S. Department of Energy/National Nuclear Security Administration.

- Provide a unified advocacy body to provide market guidance.
- Focus on applying efficient computing architectures, system components, and environments toward the real-world mix of mission-centric work. This focus on efficiency seeks to improve application performance, workflows, and code efforts for the better exploitation of current and future systems/workflow software and computer platforms, while maintaining a proper balance of compute, memory size, memory bandwidth/latency, network, and I/O throughout.

As referenced on the previous page, the initiative is located at USRC, which is right across the street from LANL. USRC is a great collaboration space where we can cultivate strategic collaborations to make collective missions workloads more tractable, and frankly, to work on the hard mission problems, whether or not they are aligned with industry trends. NMC is a formal collaboration among LANL, UNM, NMT and NMSU, providing the collaboration space and other collaboration services for USRC and EMC3. In short, think of USRC as a place to collaborate, and EMC3 as the collaborative consortium.

As we start our second year, we are already planning our summer USRC engagements. We also plan to issue at least two newsletters like this one, highlighting work being done at USRC by and for LANL and its EMC3 partners. As a heads up, we are considering an annual retreat for face-to-face time, potentially around end of summer, when we have student showcase and symposia to display results. There is also a potential annual visit to the sites of our premier and potential collaborators, where that makes sense. If interested please email or call. Planning is underway!

I hope you enjoy this first issue!

Gary Grider, Beth Kaspar and Nathan DeBardeleben

EMC3 2019 Year in Review

FY 2019 was a very eventful year for EMC3. On the partnership front, we started many collaborations:

- Marvell – acceleration of Marvell TX4 ARM processor/memory complex, enhancing core computing, memory and chip interconnect capabilities.
- DDN – exploration of massively parallel related failure management.
- Cray – prototyping large-scale file system metadata management.
- Mellanox – exploring utilization of processing in the network fabric.
- nCorium – exploring high-bandwidth memory system computing offloads.
- Eideticom – exploring high-bandwidth I/O system computing offloads.

We explored a variety of forms of collaboration, including co-mentored students funded by industry partners and co-funded by LANL, LANL-funded technology investments, and LANL-funded early technology test vehicles for some of our members. We also explored collaborations with other technology providers and technology consumer institutions. In the projects section and appendix A of the newsletter there is a lot of good technical and collaboration work that is ongoing.

**Please go to the USRC/EMC3 website for press releases:* https://usrc.lanl.gov/emc3-news.php

Member Update

As you can see, many technology-providing industry partners have begun collaborations with USRC via EMC3. Collaborations can take on many forms, and we thought it would be useful to describe some of those forms:

- Projects can be a mix of proprietary and publishable
- Jointly mentored interns are a great way to foster real collaboration. The responsibility to mentor a student gives rise to regular communications, from summer-project to long-term thesis time frames.
- Interns could be funded by Industry partner, or split-funded with LANL, depending on topic. An intern could be employee of LANL, or of an industry or University partner.
- Visit/sabbatical at USRC.
- Co-fund R&D.
- Leverage LANL testbeds and early access equipment/software.
- Assist with providing a unified advocacy to provide market guidance.
- Provide data/benchmarks/etc. as another market pull mechanism.

We are seeking a few more HPC technology collaborations, especially like minded HPC consumer collaborations. If interested, please contact Nathan Debardeleben or Beth Kaspar.

Events: 2019 Intern Showcase

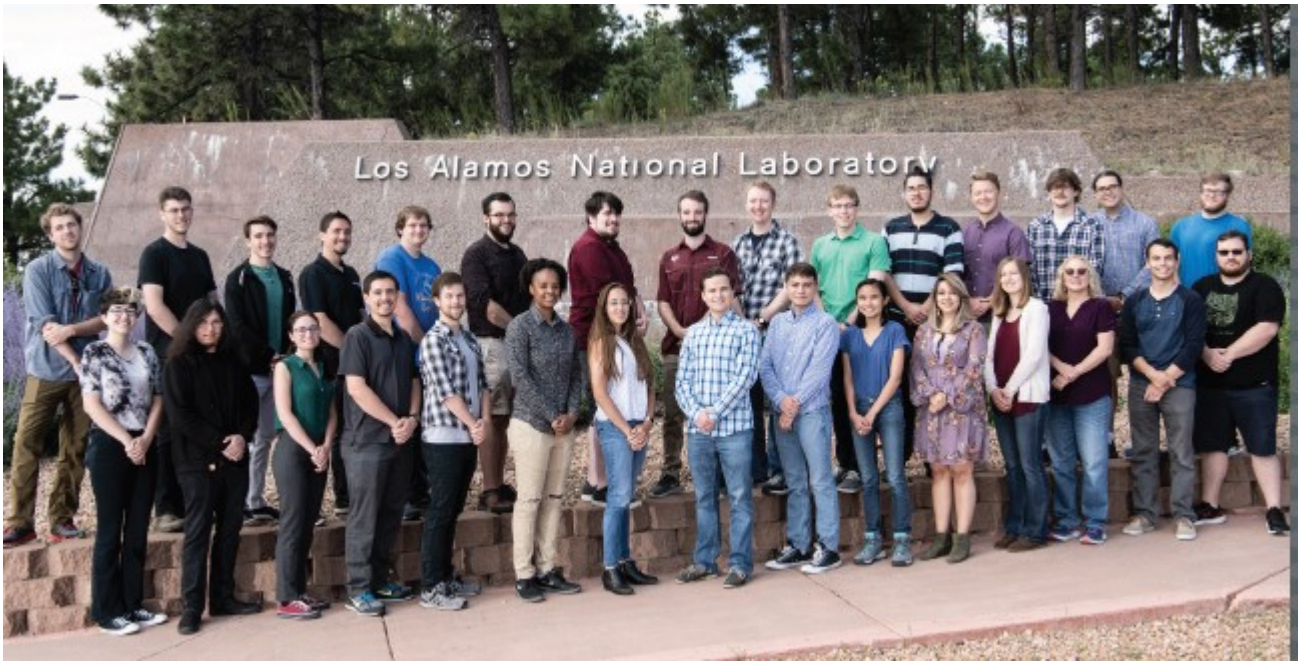
LANL hosted two significant HPC Intern Showcases in 2019.

The first, the Ultrascale Systems Research Center (USRC) Symposium, provided an opportunity for bachelor's, master's or graduate-program computer scientists and engineers to conduct meaningful research in their chosen career field. The intent was to broaden expertise and prepare interns for careers in their field. Our host for this symposium is the New Mexico Consortium (NMC), which is a non-profit corporation formed by the three New Mexico Universities under a teaming agreement with The University of California (UC) to partner with Los Alamos National Laboratory (LANL) to advance scientific research and education in New Mexico. This year, we had 9 presentations and 14 posters participating in areas ranging from "Deep I/O: Smart Networks for Fast Storage" to "Differential Privacy for Supercomputer Sensor Data." Interns worked with their LANL and EMC3 Industry Partner mentors on investigations for LANL and EMC3 partner topics in many areas of computing.

The second event was the HPC Intern Mini-Showcase, which provided an opportunity for interns in the high-performance computing field to present their research. This is an excellent forum for interns to network and make professional contacts. This year, we had 12 presentations and 25 posters participating in areas ranging from "Using Containers to Build Complex Software Applications" to "Flow Monitoring for Security and Reliability."

**Please go to the USRC/EMC3 website for details on the Intern Posters and Abstracts*

https://usrc.lanl.gov/student-symposiums.php



2019 class of NMC Student Interns.

Maximizing CPU Memory Architectures:

Under their Spring 2019 contract, **Marvell** is advancing key areas of next-generation CPU architectures and maximizing memory efficiencies and overall performance of large scale supercomputers. Drivers focus on more efficient usable cycles for our most complex multi-physics applications. LANL engineers and scientists are bringing to the table extreme-scale, complex simulation and workflow expertise to provide Marvell with insights to ensure the new architectures support the need for much more efficient, and faster predictive extreme-scale simulation.

**See Press Release at:*

https://usrc.lanl.gov/emc3-news.php

Increasing Operational Availability:

LANL's **CoFactor**, **Co**rrelated **Fa**ilure **C**onsultation **T**ool for **O**perational **R**eliability, is well underway with encouraging early results. Under this EMC3 effort, LANL and DDN computer scientists are evaluating new failure prediction methods and technologies for large-scale, catastrophic file-systems data loss scenarios. This could greatly benefit operational reliability for large, long-running simulation runs. Currently, the tool can currently generate failure streams based on LANL's failure data, and the traces are ingested by the simulator developed in collaboration with University of Chicago and DDN. This effort will continue with a more research-and-development focus, with the goal of expanding it to various large-scale, complex systems. Because systems are becoming so large and complex, it is necessary to have a decision tool to assist in designing the architectures of these largescale platforms. This tool should, therefore, be of value in the exa-scale era. This effort is ongoing through FY20, and more features will be added along with more refinement in the model.

Improved Data Query and Response:

Our data is useless, if we can't wrangle and query it efficiently, but queries can also have a devastating impact on the performance of ongoing computations/operations, which must themselves have unimpeded storage/file-system access to avoid wasting computational resources. Furthermore, security requires that results of user-queries to file systems must be constrained to include only data that a given user is allowed to see, including even the filenames. It is also useful to consider the differences in types of queries done by users versus by systems/data management professionals, and try to accommodate both areas in a comprehensive indexing capability.

The R&D100-award-winning Grand Unified File Index (GUFI) is an index that holds file-system metadata (e.g. filenames, access and creation dates, file attributes, extended attributes, etc.) pulled from huge file/data storage systems using full and incremental index updates. This allows rapid searches that do not have to impact the file-systems themselves, supporting both users and system/data managers. GUFI is a very fast software solution that offers speed and security, while minimizing impact on supercomputing resources. GUFI can be applied to a variety of file and archive systems (tape archives, Parallel File-systems (PFS), and others), making it a ubiquitous solution for indexing that supports arbitrary queries, and unifies information from all the places where a file might reside. It's a one-stop shop for file-system metadata, secure for the use of both storage administrators and users.

A summer-student joint project with Cray explored the use of GUFI accelerations for the use of storage administrators, leading to interesting follow-on discussions.

**See Press Releases at:*

www.lanl.gov/discover/news-release-archive/2018/November/1119-rd-100-awards.php

**See all current projects in Appendix A*

Fun Facts about EMC3 and USRC

- Number of staff — 45
- Number of visiting professors — 1
- Number of LANL interns — 33
- Number of NMC interns — 24
- Number of Universities/Colleges interns came from — 31
- Machine room space — About 1100 square feet
- Racks — 45, including an area for locked protection
- Computers — Around 600 total compute systems
- Power & Cooling — 1 MW power and cooling
- 10 Gbit internet connectivity to ESNet
- Hosted 3 EMC3 industry collaborator events in 2019

Looking Forward into 2020

In 2020, we will to continue to grow EMC3 in order to influence future supercomputing hardware and system software at all lifecycle stages. The focus will remain on nurturing architecture, component, workflow, infrastructure, and applications-algorithm areas that can improve the overall efficiency of our supercomputers. Together, as a consortium, we can pursue greater efficiency for systems that feature very unstructured/irregular memory access, and have many simultaneously running scientific packages that run scale for many months. Because we are focusing on all lifecycle stages, large integrators and subsystem vendors are not the only targeted vendors and participants. Small tech companies with technologies that might be adopted for future computers are also important targets. Those technology could be AI or even some other technique. If you think your company can help, please contact us. We are seeking a few more HPC technology collaborators, especially like-minded HPC consumer collaborators. We'd love to have you join.

Appendix A: Current Projects

Data Management – Fast Storage Servers/Compute on/near storage/network for byte ops
ZFS Performance
NVMe/DIMM Offloads
Storage Area Networking/storage related offloads for Data Movement
Separate data and metadata SANs for Lustre
Stilts – Short-Term, Intermediate and Long-Term Scaffolding for Campaign Storage
Analysis ready Intelligent Storage and Storage Networking
DeltaFS
Intelligent KV-SSD
HXHIM High Dimensional Hierarchical Indexing Middleware
GUFI Grand Unified File Indexing
Parallel File System Evaluation/Characterization
Data Protection Management
New data protection schemes
CoFactor **Co**rrelated **Fa**ilure **C**onsultation **T**ool for **O**perational **R**eliability
Data Driven Data Center
Production Monitoring/Monitoring Analysis
Advanced network evaluations
Logan – LOG ANomalie detection
Revere (job outcome prediction for early alerts)
Archival Storage
DNA storage evaluations
Marchive (MarFS Archive)
Intelligent Interconnects
Use of intelligence in the Interconnect nic/switch/etc, targeted at compute/analysis assists
Standardize interface to smart nic
Compute frameworks/runtime/memory-processor shaping
Marvell TX4 acceleration
Arm processor enhancements
UCX/OpenShmem maintenance and enhancement activities
Application exploitation of current/near future processor/memory technology
Containerization
CharlieCloud
Resilient Computing
Operational data analysis
SaNSA (Supercomputer and Node State Architecture
Testing frameworks/fault injection tools

**To see recent publications please go to: https://usrc.lanl.gov/publications.php*

Appendix B: Acronyms

ACM - Association of Computing Machinery
ADDRL— Average DRAM Demand Read Latency
AI — Artificial Intelligence
Bin FI— Binary Fault Injector
CoFACTOR— Correlated Failure Consultation Tool for Operational Reliability
CPU— Computer Processing Unit
EMC3 - Efficient Mission Centric Computing Consortium
EPI—Energy per Instruction
ESNet— Energy Sciences Network
FI— Fault Injector
GPU – Graphics Processing Unit
GUFI— Grand Unified File Index
HPC – High Performance Computing
ICAC— International Conference on Autonomic Computing
IEEE – Institute of Electrical and Electronics Engineers
LANL — Los Alamos National Laboratory
MPI— Message Passing Interface
MTTDL— Mean Time to Data Loss
NMC — New Mexico Consortium
NMSU — New Mexico State University
RAID – Redundant Array of Inexpensive Disks
SMART— Self-Monitoring Analysis and Reporting Technology
TSM2 – Tall and Skinny Matrix Multiplication
TX4— ThunderX 4
UC— University of California
UNM — University of New Mexico
USRC —- Ultrascale System Research Center
ZFS RAID— Z File System Redundant Array of Inexpensive Disks