SAND2022-16948D

Sandia National Laboratories

SNL ATDM* Software Ecosystem *Then and Now*
Operating Systems and On-Node Runtime

Stephen Olivier (PI), Ron Brightwell (PM),
Matthew Dosanjh, Kurt Ferreira, Scott Levy,
Kevin Pedretti, Andrew Younge

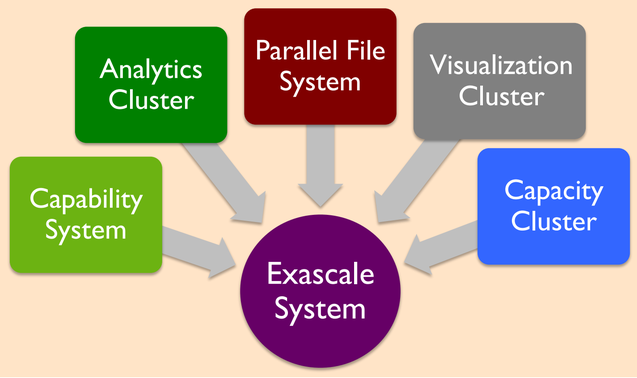
Key Thrusts

- **Containers and virtualization** technologies for productivity, portability, and performance
- **Application characterization** to understand MPI usage and response to system noise
- **Runtime systems** for on-node tasking, centered on the SNL Qthreads multithreading library
- **Standards body participation**, encompassing work in the MPI Forum and the OpenMP Language Committee
- **Preparing for future platforms**, especially NNSA ASC Advanced Technology Systems (ATS) and Vanguard Advanced Architecture Prototype Systems (AAPS)

Interactions

- **Applications:** Representative applications incorporating simulation, analytics, visualization, and/or tool components
- **Resource Management:** OS/runtime mechanisms, policies, and interfaces to enable more effective resource management
- **Testbeds:** Novel platforms able to test custom OS/R stacks

The growing complexity of applications and platforms requires a flexible system software stack, not a one-size-fits all solution.



Then & Now: Containers

- **Then:** Adoption of Containers in HPC had been limited, in part due to security limitations of frameworks like Docker.
- **Now:** Our community building efforts and collaborations with vendors and other DOE labs have made container solutions such as Podman, Apptainer, and CharlieCloud available and useful on our mission DOE platforms.
- **Now:** We have containerized full mission DOE apps such as NALU and SPARC to run at scale on machines like our Astra supercomputer (up to 2048 nodes / 114,688 containers).

Recent Publications:

R. Priedhorsky et al., "Minimizing privilege for building HPC containers", *SC21*, Nov. 2021. IEEE.

L. Stephey et al., "Scaling Podman on Perlmutter: Embracing a community-supported container", *2022 CANOPIE-HPC Workshop* at *SC22*, Nov. 2022. IEEE.

Then & Now: Standards

- **Then:** MPI 3.1 and OpenMP 4.5 were released in 2015. Threading in MPI incurred high overhead costs. GPU support in OpenMP was improved from initial support in 4.0 but still raw.
- **Now:** Partitioned Communication in MPI 4.0 allows for low overhead MPI communication by multiple threads.
- **Now:** OpenMP 5.0 – 5.2 have added unified shared memory support, a descriptive loop construct, and many other capabilities for GPUs and CPUs.

Recent Publications:

M. G. F. Dosanjh et al., "Implementation and evaluation of MPI 4.0 partitioned communication libraries", *Parallel Computing*, vol. 108, Dec. 2021. Elsevier.

J. Ciesko and S. L. Olivier, "Characterizing the performance of task reductions in OpenMP 5.X implementations", *2022 International Workshop on OpenMP*, LNCS vol. 13527, Sept. 2022. Springer.

Then & Now: MPI Analysis

- **Then:** MPI usage in applications was not well understood, despite its crucial impact on application performance and design of future HPC interconnects.
- **Now:** Our recent extensions to the LoGOPSim simulator analyze the behavior of MPI message matching with metrics such as queue depth and quantifies their impact on performance.
- **Now:** Using traces of real applications, we evaluate their MPI usage, optimize where possible, and influence vendor hardware and software design decisions.

Recent Publications:

K.B. Ferreira and S. Levy, "Evaluating MPI resource usage summary statistics", *Parallel Computing*, Vol. 108, Dec. 2021. Elsevier.

K. B. Ferreira et al., "Hardware MPI message matching: Insights into MPI matching behavior to inform design", *Concurrency and Computation. Practice and Experience*, 32 (3), Feb 2020. Wiley.

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.
SAND No. SAND2023-XXXX D

* Advanced Technology Development and Mitigation (ATDM) is a subprogram of the NNSA Advanced Simulation and Computing (ASC) program. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of two U.S. Department of Energy organizations (Office of Science and the National Nuclear Security Administration) responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, in support of the nation's exascale computing imperative.