LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-812894

Livermore Computing Integrates Advanced Cognitive Simulation Resource

I. Karlin, A. Bertsch

July 28, 2020

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Livermore Computing Integrates Advanced Cognitive Simulation Resource

Summary: With the integration of the first production Cerebras CS-1 system into Lassen, Livermore Computing (LC) is preparing the way for advanced cognitive simulation (CogSim) capabilities integrated into leadership HPC systems to support Artificial Intelligence (AI) driven computational needs of the Stockpile Stewardship Program.

Over the last few weeks LC staff in collaboration with Cerebras integrated the first production Cerebras CS-1 system into Lassen. This pairing creates the world's first computer system design for cognitive simulation (CogSim) workloads. CogSim workloads combine traditional HPC simulations and Artificial Intelligence (AI). With this new design the HPC part will run on Lassen's GPU nodes and the AI portion can offloaded to the CS-1.

LLNL researchers are collaborating with Cerebras in an Artificial Intelligence Center of Excellence to prepare for delivery. The team has two AI models selected from the director's initiative AI LDRD, running on CS-1 systems. With the CS-1 integrated into Lassen the next steps will be working to have integrated CogSim simulations run on the combined system.

For LLNL this is a first-of-its-kind disaggregated system with two unique compute elements on different physical nodes. It will allow us to understand if these types of systems are beneficial, the bandwidths needed to connect the heterogenous components and the advantages and disadvantages of this design. For example, often the AI component of a CogSim simulation can run asynchronously with the HPC portion. Providing unique compute for each allows further acceleration of workloads.

Integration of CS-1 into Lassen required close collaboration between Cerebras and Livermore Computing. The integrated system has been designed to allow up to 1.2Tb/s of data to flow from the Lassen Mellanox EDR InfiniBand high performance interconnect through the CS-1 Worker nodes, over the 100Gb ethernet CS-1 interconnect, and into the CS-1. This data could either be sourced from the Lassen parallel filesystem or directly from a running HPC job on Lassen. Eventually CS-1 Worker node function will be replaced by running the Cerebras software directly on Lassen HPC nodes for a fully integrated CogSim experience.

First users gained access to the CS-1 on July 1st. A briefing of their initial results, hardware features, and installation was presented to LC on July 23rd. A second user training on July 22nd was successful and an expanded user base has begun exploring other applications on the system.

- *Ian Karlin and Adam Bertsch*

Livermore Computing Center staff prepare to integrate the Cerebras CS-1 compute unit with the Lassen HPC system.