Synopsys and Cerebras Systems

DesignWare In-Chip Temperature Sensors and Voltage Monitors deployed in Cerebras Systems WSE-2 chip

~Dhiraj Mallick, Senior Vice President, Hardware Engineering and Operations, Cerebras Systems

Project Overview

The Cerebras Systems Wafer-Scale Engine 2 (WSE-2) is by far the largest silicon product available, with a total silicon area of 46,225mm². It utilizes the maximum square of silicon that can be made out of a 300mm diameter wafer. The square of silicon contains 84 die that are 550mm² each. These die were stitched together using proprietary layers of interconnect, making a continuous compute fabric. By developing this interconnect on a single piece of silicon, Cerebras were able to connect the equivalent of 84 die and significantly lower the communication overhead and physical connections within the systems.

Challenges

Giant models need massive memory, compute, and massive communication to tie it all together. Trying to provide this with thousands of small devices, turns the scaling of all 3 of these into distributed problems that are inter-dependent. As model size grows, Cerebras needed to do more partitioning of the model onto more chips, and do more fine-grained coordination and more synchronization, The challenge is one of distribution complexity to get them all to work together to solve a single large neural network problem. And this complexity grows dramatically with cluster size and becomes overwhelming as the network grows. Cerebras have spent the last year figuring out how to overcome these challenges and the result is the second-generation Wafer-Scale Engine (WSE-2).

Synopsys Solution

Cerebras Systems selected the Synopsys Designati89 e[illegible] here they ramped the temperature up and down.

Having the monitors and sensors distributed in large number across the device allowed Cerebras to measure variation across the tiles in the wafer, so they put as many hooks in as possible to allow access. They also conducted in-cluster thermal throttling, due to the general architecture being distributed. As part of the process they created GUI's showing heat maps and statistical variations and measurements over time which created valuable data about the health of the silicon throughout each phase of the device lifecycle from design, test, production and in-field operation.