Synopsys and Cerebras Systems

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

> *"In designing the Wafer-Scale Engine 2, we needed a trusted IP vendor to provide a comprehensive monitoring and sensing solution. The Synopsys DesignWare In-Chip temperature sensors and voltage monitors enabled us to understand the dynamic thermal and supply conditions of our WSE-2 in real time, which was important for power and performance optimization."*

~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).

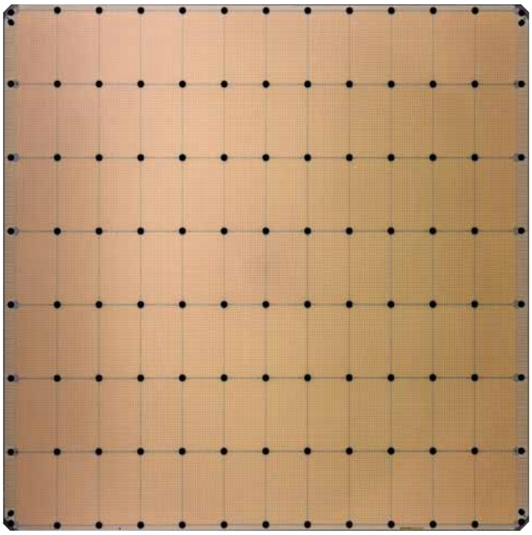
Synopsys Solution

Cerebras Systems selected the Synopsys DesignWare embedded in-chip temperature sensors and voltage monitors which were then distributed across the device. They monitored in real-time, polling at the fastest possible rate and keeping watermarks for highs and lows. The Cerebras wafer consumes approximately 20KW of power and they are able to cool that with their system, so the monitors and sensors were mainly being used to validate that they were staying in the safe zone, and for characterization where 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.

Key Benefits of Embedded In-Chip Monitors and Sensors for AI Devices

- Better manage thermal unpredictability with sensors closer to hotspots
- Optimize performance per Watt, maximize core utilization
- Maintain supply margin for critical logic operation
- Manage large-scale distribution of sensors
- Granular thermal management with in-cluster sensing and thermal throttling
- Accurate sensing of voltage margins to be optimized and monitoring of polling in real-time
- Assessment of localized process variation across die for voltage scaling and performance optimization
- Visibility of silicon health throughout each phase of the device lifecycle from design, test, production and in-field operation
- Monitoring of IR drops caused by bursty AI workloads
- Increased data throughput



Cerebras Systems second generation Wafer-Scale Engine on TSMC 7nm

Wafer-Scale Engine—Gen 2 & Monitor and Sensor Implementation

- 46,225 sq.mm silicon
- 2.6 trillion transistors
- 850,000 AI-optimized cores
- 40 GB on-chip memory
- 220 Pb/s bandwidth
- TSMC 7nm
- 8 temperature sensors and 8 voltage monitors (each with 16 voltage sense points) on each of the 84 WSE-2 die
- Per wafer that gave a total of 10,752 voltage sense points and 672 temperature sensors
- Location was the best compromise for geography and physical design restrictions

Expertise and Technical Support

Cerebras engaged with Synopsys to help bridge their resource gaps in embedded IP integration. The Synopsys DesignWare in-chip sensing and monitoring team did an excellent job all around. They understood the product and leveraged Synopsys' internal expertise to accelerate the development of the Wafer-Scale Engine project on both the 16nm and 7nm designs.

SYNOPSYS®
Silicon to Software™