Machine Learning-Driven Full-Flow Chip Design Automation

By Rod Metcalfe, ML Product Manager, Cadence

To enable the semiconductor industry to continue growing, the chip design process must become more efficient. With the availability of massive, cloud-enabled, distributed computing and advancements in machine learning computer science, the next chip design automation revolution is now possible. The Cadence® Cerebrus™ Intelligent Chip Explorer utilizes both of these technologies, based on the industry-leading Cadence digital full flow, to deliver better power, performance, and area (PPA) more quickly. Engineering teams now are able to scale and become more productive using the Cadence Cerebrus reinforcement learning engine to meet the challenges of increasingly large and more complex system-on-chip (SoC) designs.
Introduction

New applications and technology are driving demand for even more compute and functionality in the devices we use every day. The semiconductor industry is experiencing strong growth based on drivers like 5G, autonomous driving, hyperscale compute, industrial IoT (IIoT), and many others. All of this has resulted in an ever-increasing number of chip design projects. Additionally, SoC designs are quickly migrating to new process nodes, and rapidly growing in size and complexity.

Each new process node offers improvements in power and performance, and enables more transistors to be squeezed onto a typical die. To ensure new products are competitive, design teams must utilize these process-node advantages to deliver more features at a higher performance and lower power. A good example of this is GPU cores, which a few years ago had to support a 1080p resolution, now are required to support 4K, with 8K being available on many high-end mobile devices. In addition, the process node requirements are increasing, with more voltage and temperature corners to consider as part of design closure and signoff. This all represents an explosion in size and complexity, leaving engineering teams struggling to keep up. The whole chip design process must become more automated, improving the productivity of engineering teams, so new products can be delivered to market on schedule.

Improving productivity has a long history in chip design (Figure 1). When chip design first started, each transistor was created and connected manually in a full custom layout editor, which was a time-consuming process. To improve efficiency, digital chip design moved to a standard cell and schematic netlist methodology. This enabled engineers to implement digital logic designs much more quickly, but creating a schematic netlist by hand took a lot of effort. As desktop Unix workstations became available, each engineer suddenly had access to a lot more compute power, so RTL synthesis became practical. Chip designers could capture digital logic functions using high-level languages like VHDL and Verilog and easily synthesize a netlist of millions of gates. However, this massive leap forward in productivity presented the next problem, how to layout millions of standard cells. So along with RTL synthesis, automated place-and-route systems were developed. Now large netlists could be quickly implemented, delivering another significant productivity improvement.

Today’s Challenge

Although design technology has become much more sophisticated over the decades, the basic chip design flow has remained the same, synthesis followed by place and route, with PPA-related challenges. In contrast, today’s biggest design challenge is an industry shortage of skilled design and implementation engineers. Current teams are overloaded, which impacts the ability of companies to bring new products to market. The reality is that future chips must be produced faster with more automation.

Fortunately, during the past few years, some key technologies have become available that will enable the next big leap forward in chip design productivity: Engineering teams now have access to massive compute power, either on premises or using cloud resources, and machine learning computer science has made significant progress, and is now ready and available for electronic design automation purposes. Both these technologies have enabled the next revolution in chip design—automated, machine learning-driven flow optimization.
Intelligent Chip Explorer

Cadence Cerebrus is built on these massive compute and machine learning architectures and utilizes the complete Cadence digital full flow solution. Cadence Cerebrus uses a unique reinforcement learning engine to deliver better design PPA results. By using a completely automated, machine learning-driven, RTL-to-GDS full-flow optimization technology, Cadence Cerebrus can deliver these better PPA results more quickly than a manually tuned flow, thereby improving engineering team productivity. Cadence Cerebrus uses the latest scalable distributed computing technology resources, either on-premises or in the cloud, to enable efficient and scalable chip implementation for the ever-increasing size and complexity of current SoC designs.

Figure 2 shows how Cadence Cerebrus can improve engineering team productivity. The red area shows the current manual, iterative, flow development process. Designers create an initial flow and run the design to generate some results. Based on these results, expert designers decide what flow changes to make, and then run the flow again to generate updated results. These flow iterations continue until acceptable PPA is achieved or until the design team runs out of time and must accept the current results. This requires a lot of engineering effort and is generally an inefficient use of compute resources. Even if more engineers are added to the team, the PPA may not improve that much. By adopting Cadence Cerebrus automation, the green area moves to the left. Cadence Cerebrus can use automated reinforcement learning-driven, full-flow optimization to generate better PPA more quickly, improving the engineering team's productivity and making more effective use of the available compute resources.

Reinforcement Learning Engine

The Cadence Cerebrus reinforcement learning engine samples design data in real time, so it can make optimization decisions as the flow is running. This enables Cadence Cerebrus to immediately stop flow runs that are not converging on improved PPA results and reallocate the compute resources to alternative flow configurations. This is a much more efficient use of distributed compute resources compared to manual flow tuning, when the results are only reviewed at the end of each run.

A vast amount of design data is analysed by the Cadence Cerebrus reinforcement learning engine during full-flow optimization. As the reinforcement learning process proceeds, a machine learning model is created, capturing the design data analysis. This Cadence Cerebrus machine learning model can be used as a starting point for future design flow optimization, allowing data to be easily reused between projects, thus saving significant compute time and delivering improved PPA even more quickly.

Implementation Flow Optimization

Figure 3 shows a typical 5nm, high-performance, 3.5GHz CPU design. Here Cadence Cerebrus was used to automatically optimize the implementation flow to improve power and performance. The results are significant, with a 420HMz performance improvement, and also good power reduction. This was achieved by one engineer using Cadence Cerebrus for about two weeks. If manual, iterative flow tuning had been used, it would have taken many engineers a few months to complete, with no guarantee these PPA results would have been achieved.
Floorplan Optimization

Figure 4 shows how Cadence Cerebrus can use built-in high-level design optimization capabilities. In this 12nm, 2GHz CPU design, Cadence Cerebrus was used to concurrently optimize the floorplan and implementation flow for better power and performance. Using floorplan optimization, Cadence Cerebrus was able to dynamically change the size and aspect ratio of the floorplan, and used the Cadence Innovus™ Implementation System’s mixed placer technology to optimally place the macros in the resized floorplan, completely automatically. This resulted in over 200MHz better CPU performance, with a good reduction in leakage power. Cadence Cerebrus was able to come up with an optimized floorplan much more quickly than a manual, iterative approach could achieve.

Figure 3: Example Cadence Cerebrus results for a 5nm, high-performance CPU design

Figure 4: Example Cadence Cerebrus floorplan optimization for a CPU
Conclusion

To enable stretched engineering teams to implement the ever-increasing number of new 5G, autonomous driving, hyperscale compute, and IIoT-driven products, more automated chip design is urgently required. Recent advances in readily accessible distributed computing and machine learning computer science provide the necessary technologies for the next chip design productivity breakthrough.

Cadence Cerebrus Intelligent Chip Explorer utilizes this massively distributed compute power and a unique reinforcement learning engine, combined with the Cadence digital full flow solution, to deliver better PPA more quickly. Cadence Cerebrus automation enables current engineering teams to scale more efficiently and boosts productivity, so more designs can be implemented concurrently.

In addition to automated implementation flow optimization, Cadence Cerebrus has the capability to explore high-level design optimizations, such as dynamically resizing and shaping a floorplan to improve PPA much more efficiently than a manual approach. All design learnings are stored in a reinforcement learning model that can easily be used in future design projects to optimize the flow even more quickly.

Cadence Cerebrus offers a revolution in chip design productivity, which will allow the semiconductor industry to continue growing and delivering the new SoC product features and capabilities we all expect in our ever more connected world.

Cadence Cerebrus, the Future of Intelligent Chip Design