Building Better AI Chips

By Linley Gwennap
Principal Analyst

August 2020

The Linley Group

www.linleygroup.com

Building Better AI Chips

By Linley Gwennap, Principal Analyst, The Linley Group

As progressing to 7nm and beyond becomes ever more complex and expensive, GlobalFoundries is taking a different approach to improving performance by enhancing its 12nm node with lower operating voltages and new IP blocks. The changes are particularly effective for AI (neural-network) accelerators. The new 12LP+ technology builds on the success that the foundry's customers have already achieved in AI acceleration. GlobalFoundries sponsored this white paper, but the opinions and analysis are those of the author.

Introduction

Moore's Law may not be dead, but it is rapidly running out of steam. After 50 years of consistent progress, achieving the next node is becoming more and more difficult. Over the past decade, lithography costs have been rising, particularly with the recent introduction of EUV. The move to 3D (FinFET) transistors further raised cost, starting with the 16nm node. As a result, the historically rapid decrease in transistor cost has flattened out, with only slow progress over the past few nodes, as Figure 1 shows. The initial tapeout cost for a new chip design has also skyrocketed from $1 million for 28nm to about $10 million for 7nm.

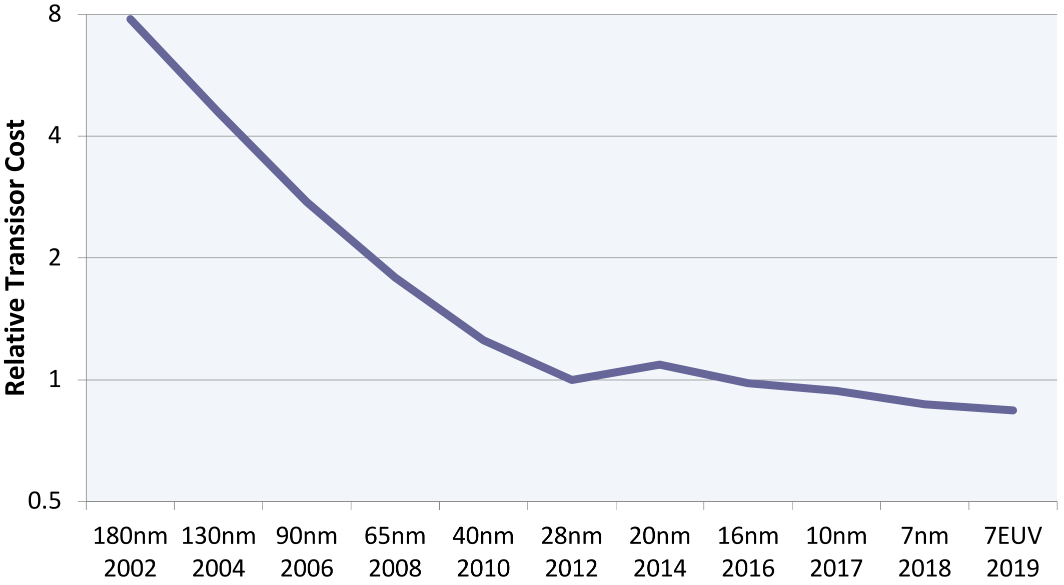

Figure 1. Moore's Law slows down. Transistor cost fell about 40% per node prior to 28nm, but since then the cost is falling only 10% per node. (Source: The Linley Group)

Some chip companies are willing to pay these higher costs to achieve better performance and power efficiency for their design, but these benefits are also slowing. Intel processors soared from 1.0GHz in 2002 to 3.8GHz in 2005, but in the past decade the top clock speed rose only 3% per year. Other processor designers have seen similar difficulties: Arm CPU speeds have risen about 6% per year since 2014. Part of the problem has been that most designs already operate well below 1.0V, leaving little room to further reduce voltage and thus power. Given these tradeoffs, many companies aren't pushing their chip designs to the 7nm node and beyond.

To assist these companies, GlobalFoundries (GF) has enhanced its 12nm technology to improve performance and power efficiency, creating a new 12LP+ process. The changes are particularly effective for AI (neural-network) accelerators. For example, neural networks frequently employ the multiply-accumulate (MAC) function, so GF redesigned its 12nm MAC unit to improve power efficiency by 65%. A new SRAM cell optimized for the sequential data accesses often found in neural networks doubles power efficiency. In addition, a new dual work-function metal gate cuts the supply voltage to reduce power by another 50%.

The 12LP+ technology builds on the success that GF customers have already achieved in AI acceleration. One startup has built a chip that delivers an industry-leading AI performance of 820 trillion operations per second (TOPS) using 12LP technology. Another 12LP customer achieved industry-leading power efficiency among data-center chips on the popular ResNet-50 inference benchmark. At the other end of the scale, a chip using GF 22nm technology achieves impressive AI performance while consuming just 50mW. These and other customers combine unique and innovative logic designs with GF manufacturing technology to achieve these leading marks.

Smaller Transistors, Bigger Problems

Lithography has been a critical cost driver in recent nodes. Deep ultraviolet (DUV) lithography reached a limit in the 28nm node. To achieve further progress, the industry moved to costly double patterning for 22nm and even more costly quadruple patterning for 10nm. At 7nm, fabs started to use extreme ultraviolet (EUV), but this technology requires new (expensive) masks, new resists, and new steppers that weigh 180 tons and cost more than $100 million. FinFETs require additional process steps to form the 3D transistors. The 7nm node introduces a new material (cobalt) for vias. Each node also adds another metal layer to the stack (now up to 14 layers in TSMC 5nm), adding several more process steps.

Each new process step increases wafer cost, and expensive lithography tools must be amortized across all wafers. As a result, wafer cost has been rising rapidly since the 28nm node, erasing most of the potential decrease in transistor cost. As the name implies, double patterning requires twice as many process steps, and quadruple patterning even more. EUV steppers eliminate multiple patterning, but their greater equipment cost and lower throughput mean that an EUV layer costs three times as much as a DUV layer. EUV masks must be constructed using special materials that block near-X-ray light, and they require very fine details. As a result, tapeout cost (which includes building a complete set of masks) is rising rapidly as EUV gains adoption.

These heroic efforts continue to reduce transistor area by about 50% per node, as demanded by Moore's Law. Because smaller transistors require fewer electrons to switch states, they consume less power and can also switch faster. As transistors have shrunk, however, most designers have simply packed more functions into their chips, keeping die area about the same. Thus, the metal connections between the transistors are still the same length. Worse, these connections have become thinner with each node, increasing their resistance. For complex high-end processors, the power required to push signals through this interconnect now far outweighs the switching power of the transistors, minimizing the benefit of transistor shrinks. At 7nm, many designers see little or no gain in clock speed and perhaps 10% improvement in power efficiency from the previous node.

The situation isn't likely to improve in future nodes. Although 5nm employs single-patterning EUV, that approach isn't sufficient for the next node. One option is double patterning EUV, which again doubles cost for these layers. To avoid this problem, equipment makers are working on a new technology called high-NA EUV that can create smaller features in a single pass. But this equipment will be even more expensive than current EUV steppers, and the technology requires new resist materials that are still in development. That node will also move to a new transistor technology (GAAFET) that will require additional process steps, further raising cost and design complexity. The process of resolving all of these issues is likely to delay the introduction of 3nm and future nodes.

Design Smarter, Not Smaller

Instead of continuing down this ever-shrinking rabbit hole, GlobalFoundries decided to enhance its cost-effective 12nm process to deliver better performance and power efficiency. In particular, the company focused on the hot market for AI-enhanced chips, ranging from dedicated AI accelerators for servers to microcontrollers that integrate tiny AI engines. Despite the varying end applications, these chips all need the same thing: maximum power efficiency for common AI operations.

The most popular AI applications today implement convolutional neural networks (CNNs). As the name implies, CNNs primarily perform convolution functions, which repeatedly multiply a fixed weight by an incoming activation value and add the product to an accumulator. To streamline this operation, GF focused on two things: fetching the activation value from SRAM and efficiently computing the MAC operation.

General-purpose processors typically use SRAM for cache or other on-chip memories that must respond quickly for any access pattern. Thus, foundries optimize their SRAM designs for random accesses. These SRAM arrays fetch several values (e.g., a cache line) at once, then use a multiplexor (mux) to select the desired value, discarding the others. Convolutions, however, operate on very large arrays, so the data is typically processed sequentially.

GF designed a new SRAM that reads and latches four values at once, then uses a mux to select the desired value. The latch slows down the first access, but if the second access is sequential, it can immediate read the next value from the latch without accessing the array again. Thus, a series of sequential reads can eliminate three of four accesses, greatly reducing the power required by the SRAM array. For a typical CNN, this approach reduces the SRAM power by about 50%.

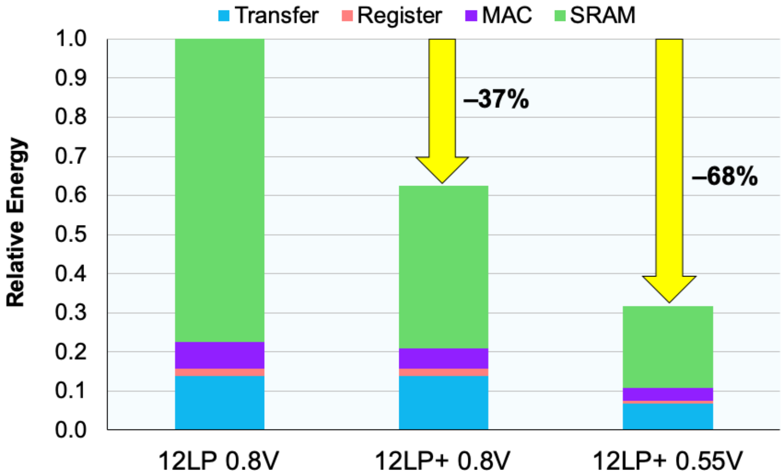

Figure 2. Energy reduction in 12LP+. A combination of new circuit designs and lower voltage reduces the energy for a typical CNN operation by nearly 70% relative to the previous 12LP technology. (Source: GlobalFoundries)

Two of the challenges required with low-voltage operation are device mismatching and the voltage margin required for SRAM operation. For 12LP+, GF implemented separate gate stacks for the logic devices and the SRAM cells. The two stacks have different work functions that are tuned to reduce mismatching and minimize the voltage margin. This technique enables dropping the SRAM supply voltage from 0.7V to 0.55V, reducing power by another factor of two.

While memory contributes the biggest portion of the power in a typical CNN operation, the other major contributor is the MAC unit, as Figure 2 shows. In talking to customers, GF found that, unlike general-purpose CPUs that optimize for single-thread performance and multi-GHz clock speeds, AI accelerators process highly parallel workloads and operate at around 1GHz to maximize power efficiency. Therefore, it designed a new multiplier and adder optimized for a lower clock speed, enabling a 25% reduction in power.

Taken together, these optimizations provide a 37% reduction in power at the same supply voltage and a 68% reduction when taking advantage of the dual-work-function gates to reduce the supply voltage. In other words, the core of the convolution function, which can consume 90% or more of the compute cycles in a CNN, operates at three times the power efficiency relative using standard logic blocks on the older 12LP process.

Powering the AI Leaders

The new technology builds on the proven success of GF's 12LP process, which powers industry-leading AI products. For example, Silicon Valley startup Groq has developed a new architectural approach to accelerating neural networks that combines hundreds of function units in a single core. The massive design includes 220MB of SRAM and more than 200,000 MAC units. Groq adopted 12LP to keep such a large design within a 300W power budget. At an initial speed of 1.0GHz, the chip achieves a peak throughput of 820 trillion operations per second (TOPS) for INT8 data, surpassing all other announced accelerators.

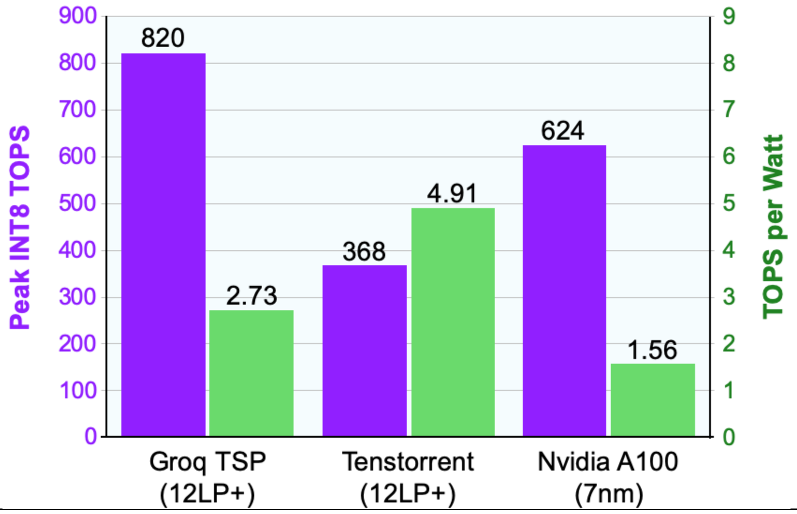

Figure 3. Comparison of high-end AI accelerators. Groq's TSP accelerator delivers greater performance (measured in trillions of operations per second, or TOPS) than Nvidia's new A100 product while using less power. Tenstorrent targets a lower performance point but delivers three times the power efficiency (TOPS/W) of Nvidia's accelerator. (Source: vendor data)

Tenstorrent, a Canadian startup, also accelerates inferencing but chose a different design target: the 75W power limit for a bus-powered PCIe card. Its first chip features 120 independent cores that each include 1MB of SRAM and about 500 MAC units. This approach still requires lots of SRAM and MAC units. At a preliminary speed of 1.3GHz, the chip can deliver 368 TOPS. The 12LP technology helps Tenstorrent achieve 4.9 TOPS per watt, the best efficiency rating among data-center products, as Figure 3 shows.

Nvidia, which has the greatest share in this market, recently delivered its A100 accelerator based on the new Ampere architecture. Ampere introduces many innovative features and boosts peak performance to 624 TOPS, beating every announced chip except Groq's. But despite a shrink to 7nm technology, the A100 requires 400W TDP, 33% higher than the previous 12nm product. To fit even this increased power budget, Nvidia had to reduce the clock speed relative to the 12nm product and disable 15% of the cores on the die, an unusual tactic that could indicate the chip's power was considerably higher than simulated. As a result, the A100 badly lags the Groq and Tenstorrent chips in performance per watt, despite its smaller transistors.

GlobalFoundries also supports customers developing low-power chips for embedded systems, many of which are also adding AI capability. These products are more cost focused than data-center accelerators, so they typically use older nodes. Innovative startups such as GreenWaves and Perceive have chosen GF's 22FDX process, which employs silicon-on-insulator (FD-SOI) technology to conserve power without the greater cost of FinFET nodes. FD-SOI enables adaptive back-bias, which allows the designer to vary the body bias depending on the state of the chip. For example, in sleep mode, applying reverse bias lowers leakage current by up to 10x, greatly extending battery life. But when the device is active, applying forward bias maximizes performance.

The GreenWaves GAP9 is a RISC-V microcontroller that includes a small neural-network accelerator that can operate at just 50mW and achieve 34x the power efficiency of a standard microcontroller for AI workloads. Perceive has created completely new AI algorithms to run on its Ergo chip at 70mW. With the assistance of the FD-SOI technology, Ergo is rated at an industry-leading 55 TOPS/W. For even greater efficiency, 22FDX also supports analog in-memory computing; the foundry has partnered with researchers at Imec to develop a test chip using this technique that achieves up to 2,900 TOPS/W.

Better Than 7nm

Moore's Law is showing its age. Although the industry continues to find new ways to fabricate smaller transistors, the technology to do so is increasingly expensive, negating most of the cost advantage. Supply voltage is nearing fundamental limits, preventing reductions that would ease power. The switching speed and energy reductions from smaller transistors are swamped by the difficulty of pushing signals through ever-thinner metal lines. Thus, leading-edge foundries will be increasingly challenged to deliver meaningful progress in cost, speed, or power simply by shrinking transistors.

Processor designers have already started adapting to this new environment by creating more specialized designs. A good example is the emerging trend of building AI-specific accelerators to offload standard CPUs and GPUs. Foundries can follow suit by creating application-specific versions of their technology. Instead of simply shrinking the transistor and metal stack, these versions can apply optimized function-block and circuit designs to better fit the needs of particular product types.

GlobalFoundries has taken this route with its 12nm node to create 12LP+ technology for AI accelerators. Optimizations include a dual-work-function gate that enables a sizable voltage reduction along with a burst-optimized SRAM and a low-power MAC design. Taken together, these optimizations improve power efficiency for a typical convolution operation by 3x. This improvement is much greater than what could be achieved by simply porting an existing design from 12nm to 7nm at another foundry, and the design and tapeout costs are also lower than for 7nm.

Customers are already achieving impressive results using GF technology. Using the 12LP process, Groq and Tenstorrent lead all data-center accelerators in AI performance and power efficiency. Perceive and GreenWaves use GF 22FDX technology to reduce power and improve efficiency in client devices, helping to move AI processing to the edge. GF also provides silicon-photonics technology to connect the data center to the edge, completing its end-to-end AI play. These examples show how GlobalFoundries helps customers achieve industry-leading performance without the high cost of 7nm manufacturing. The new 12LP+ enhancements offer even greater gains.

Linley Gwennap is principal analyst at The Linley Group and editor-in-chief of Microprocessor Report. *The Linley Group offers the most-comprehensive analysis of microprocessors and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth articles cover topics including embedded processors, mobile processors, server processors, AI accelerators, IoT processors, processor-IP cores, and Ethernet chips. For more information, see our website at www.linleygroup.com.*