Chiplets Gain Rapid Adoption:
Why Big Chips Are Getting Small

By Linley Gwennap
Principal Analyst

May 2021

www.linleygroup.com

Chiplets Gain Rapid Adoption

By Linley Gwennap, Principal Analyst, The Linley Group

Leading chip vendors such as AMD and Intel have adopted chiplet technology for several products. This technology can reduce cost for large 7nm designs by as much as 25%, according to our analysis; the savings are even greater at 5nm and beyond. We expect chiplets will be widely used for data-center processors and networking chips in these advanced nodes. Alphawave sponsored the creation of this white paper, but the opinions and analysis are those of the author.

As chip designers struggle with the slowdown in Moore's Law, many are taking a new approach known as chiplets. This approach divides a complex design, such as a high-end processor or networking chip, into several small die instead of one large monolithic die. Data-center products, which often have leading-edge transistor counts, have been early adopters of chiplets. AMD and Intel are already shipping multiple chiplet-based designs, and Nvidia is developing chiplet technology as well. We expect increased chiplet adoption.

Data-center customers are among the most demanding, requiring greater compute performance to deliver new cloud services and greater networking bandwidth to connect massive numbers of servers. To meet these compute needs, Intel and AMD race to cram more cores into their server processors. Nvidia's powerful GPUs have become popular for training huge AI models that can perform tasks beyond the ability of standard servers. Large data centers have pushed Ethernet speeds to 100Gbps and beyond while requiring switch chips with high port counts. High-end FPGA customers want products with ever-more gates for their leading-edge applications.

For decades, Moore's Law provided regular improvements in transistor technology that enabled chip vendors to satisfy these customer demands, but it is running out of steam. Doubling transistor density now takes three or four years instead of two. Each increase in density comes with a sharp rise in wafer cost, producing little or no reduction in cost per transistor, a key tenet of Moore's Law. Power and speed gains have also diminished with each new transistor node. In short, moving to the next node has become much more expensive while offering less benefit.

Chiplets provide an alternative way to create more advanced designs. By using two or more chips, a company can increase the design's transistor count beyond what a single chip can hold. It can use an older node for some of the chiplets to save cost while employing the leading-edge node where needed for optimal performance. For complex designs, this approach can reduce manufacturing cost. As designs move to 5nm and beyond, rising costs improve the economics of chiplets.

Chiplets in the Data Center

AMD was the first major vendor to introduce a chiplet architecture. Its original Epyc server processor, code-named Naples, launched in 2017 and featured four identical (homogeneous) chiplets in a single package to deliver a total of 32 CPU cores. In 2019, it upped the ante with its second Epyc design ("Rome"), which used eight CPU chiplets to implement 64 cores, more than twice as many as Intel's then-best processor. The Rome design adds a ninth chiplet that centralizes all the DRAM and I/O circuitry, as Figure 1 shows; this die uses a less expensive 14nm node while the CPU chiplets use 7nm transistors to improve speed and power. AMD retained the same chiplet configuration for its recent third-generation Epyc processor ("Milan").

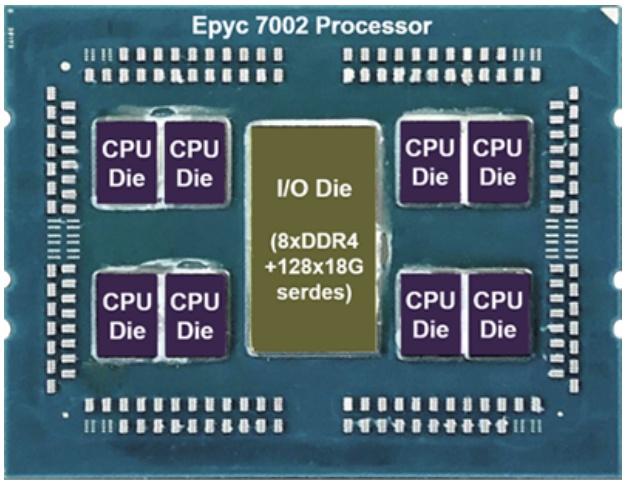

Figure 1. **AMD chiplet design.** The Epyc 7002 ("Rome") processor features eight 74mm² CPU die, each with eight x86 cores and 32MB of cache. The I/O die measures 410mm² and connects to DRAM as well as high-speed peripherals. The organic package measures 58x78mm. (Photo courtesy of AMD, overlay by The Linley Group)

For its Agilex FPGA, Intel employs chiplets but takes a different approach. Agilex keeps the main portion of the design—including the programmable logic, CPU subsystem, and DRAM controller—on a single die. The chiplets (which Intel calls tiles) implement additional I/O connections such as high-speed serdes, PCIe Gen5, high-bandwidth memory (HBM2), and Optane persistent memory. This approach allows the company to offer Agilex products with various I/O combinations simply by changing the installed chiplets. Intel can even create customer-specific chiplets to add special features. The company can add new chiplets, such as 112Gbps serdes or HBM3, even after the base die has reached production.

Barefoot Networks, now part of Intel, used chiplets for its 400G Ethernet switch chip called Tofino 2. Like Agilex, this product keeps the central logic on a single die but divides its 32 Ethernet ports among four chiplets. This division simplifies routing the I/O signals in the package. We believe the Ethernet chiplets employ 16nm technology, reducing cost relative to keeping them on the main 7nm die. The company could also develop and test the high-speed Ethernet circuitry before completing the rest of the complex design, decoupling the development risk.

Other chiplet designs in production include the Huawei Ascend 910, which comprises a compute die and a separate I/O die. It uses a silicon substrate to connect the two die along with four HBM2 stacks. Nvidia's next-generation GPU (alternatively known as Lovelace or Ampere Next) is rumored to use chiplets. Intel plans to move to chiplets in future processors, optimizing the technology node for CPU, GPU, and I/O circuitry. Its Ponte Vecchio GPU, which targets data centers and supercomputers, divides across several chiplets, although Intel hasn't disclosed details of the 2022 product. Ultimately, the company hopes to reduce processor-design time by implementing individual function blocks on chiplets, creating new products through a mix-and-match process.

Benefits of Chiplets

Dividing a large chip into smaller chiplets reduces manufacturing cost by improving yield. A traditional yield model assumes that defects scatter randomly across a wafer, and that a defect anywhere on the die renders it unusable. Therefore, a large die is much more likely to contain a defect than a small die. A reticle-size 700mm^2 design (the largest possible) generally yields around 30% good die, whereas a 150mm^2 die has about 80% yield. This yield improvement saves considerable cost, even accounting for the greater number of chiplets.

To improve the yield of large die, some vendors include redundant circuitry that can accommodate certain defects. For example, a block of SRAM could have extra rows that replace any failed rows. This approach adds die area but reduces the "effective area" that is vulnerable to defects. A chiplet design could remove redundant circuitry, reducing die area, while still improving yield.

Further cost savings come from creating different (heterogeneous) chiplets using different manufacturing nodes, which isn't possible in a monolithic design. For example, 7nm transistors are cheaper than 16nm transistors for densely packed logic and memory, but I/O interfaces typically have analog circuitry and other large features that don't benefit from the smaller node. For this reason, many chiplet designs segregate I/O functions into a separate die manufactured in an older node. Some logic circuitry, such as accelerators, may not need to run at the same maximum clock rate as the main processor and thus can be fabricated in an intermediate node. Using older process technology can reduce the manufacturing cost of those chiplets by as much as 50%.

A company can reduce design time and tapeout fees by reusing chiplets in multiple products. For example, AMD used the same chiplet design in its first-generation Epyc and Ryzen products; the PC processor used a single chiplet, whereas the server processor contained up to four chiplets. Furthermore, AMD can easily offer a broad range of Epyc core counts by varying the number of chiplets in the package. In contrast, Intel typically tapes out three Xeon chips, each with different core counts, to cover the full range of models in each generation. Similarly, Barefoot can scale the port count of its switch by changing the number of Ethernet chiplets.

Dividing a design into chiplets also divides the design risk. With a monolithic design, a faulty I/O interface can prevent the product from launching. To reduce this risk, Barefoot moved its bleeding-edge Ethernet design onto a separate chiplet, enabling it to develop and test this circuitry independently of the main logic chip. Although its Agilix FPGA is already shipping, Intel plans to develop new chiplets to upgrade the product's I/O capabilities over time.

Chiplets can enable a "more than Moore" gain in transistor count by instantiating more transistors than could fit on a single die. Xilinx has used approach since 2011, when it combined four midsized chiplets that together delivered twice as many gates as the then-largest monolithic FPGA. AMD's Rome product incorporates 40 billion transistors across nine chiplets, whereas Intel's contemporary Skylake Xeon, a monolithic design, has only 8 billion. For many leading-edge products, however, power dissipation (TDP) limits the number of transistors before the design reaches the maximum die size.

Chiplet Cost Study

Designers can implement chiplets in many different ways. Some use homogeneous chiplets, while others segregate compute and I/O functions into different chiplets. Silicon substrates offer dense routing and greater bandwidth between chiplets, but organic substrates cost less. As a simple case study, let's look at a hypothetical processor that could be broken into four homogeneous chiplets. The monolithic version requires 600mm^2 in the 7nm node and an expensive 60x60mm organic BGA package with many routing layers to handle the large number of I/Os. The design includes a sizable memory with redundant rows, leaving an effective area of 80%.

	Monolithic	Diff	Chiplet
Wafer Cost (7nm)	$9,350	1x	$9,350
Total Die Size	600mm2	1.1x	660mm2
Single Die Size	600mm2		165mm2
Gross Die per Wafer	96		387
Defect Rate (per cm^2)	0.2	1x	0.2
Effective Area	80%	1x	80%
Estimated Yield (Dingwall)	43%		78%
Net Die per Wafer	42		300
Single Die Cost	$224		$31
Total Die Cost	$224		$124
Total Test Cost	$10	1.2x	$12
Package and Packaging	$160	1.25x	$200
Packaging Loss	1%	4x	4%
Total Manufacturing Cost	$398		$347

Table 1. Chiplet cost comparison. This comparison assumes a large die (600mm^2) with little redundancy (80% effective area) and a large (60x60mm) organic package is split into four identical chiplets. The chiplets reduce total die cost but require a more expensive package, producing a net savings of 13%. (Source: The Linley Group estimates)

Dividing this design could result in four 150mm^2 die, but chiplets require extra area for the die-to-die connections, which require much larger drivers than on-die signals; we estimate 10% overhead for this purpose. Even so, the yield of the smaller chips is nearly twice that of the large monolithic die, as Table 1 shows, resulting in a $100 savings in total die cost.

The test cost is slightly higher, however, due to the overhead of testing four chips instead of one. The package, already expensive, sees a sizable cost increase for several reasons. The total area of the chiplets is 10% larger, as noted above, and the package requires some space between the chiplets, so it grows to 60x80mm. Spreading the I/Os across a larger package reduces the number of layers needed to route these signals, but the new die-to-die signals add routing layers; we assume these changes offset, and the layer count remains the same. Finally, the assembly cost will be higher for the multidie package, as will the assembly loss. These packaging costs offset about half of the die-cost savings, leaving a net gain of 13%.

Extending this cost model to additional examples, we see the greatest savings for large die with little or no redundancy. In the 7nm node, chiplets reduce cost for monolithic designs with an effective area of greater than 400mm^2, as Figure 2a shows. For highly regular processors in which identical cores consume 50% or more of the die area, chiplets generally increase manufacturing cost, according to our model. In this case, the redundant cores increase yield on the big die, reducing the yield gain for chiplets.

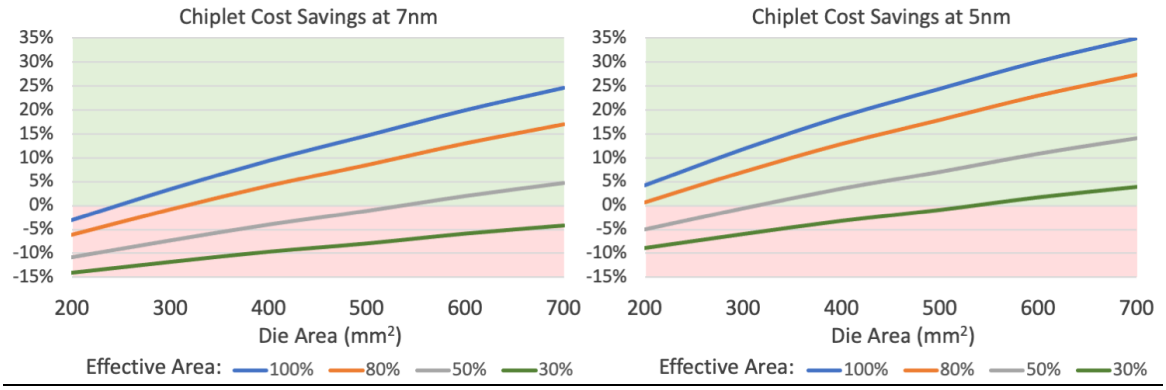

Figure 2. Chiplet cost scenarios. Chiplets are most cost effective for large die with little redundancy. In the 7nm node, the crossover point for a design with 100% effective area is about 400mm^2, whereas for 5nm, the crossover is below 300mm^2. (Source: The Linley Group analysis)

In more advanced manufacturing nodes, chiplets provide advantages across a broader range of designs. At 5nm, for example, wafer costs nearly double to $17,000. Thus, the savings from shifting to smaller die are greater, making it easier to offset the increase in package cost. According to our model, the net cost savings are about 10% greater for 5nm than 7nm, meaning that chiplets could reduce cost for die as small as 200mm^2. Even for a processor with 50% effective area, cost savings occur above 300mm^2. Although wafer costs aren't yet available for 3nm, cost savings will certainly rise again, pushing the breakeven point for chiplets below 150mm^2.

Conclusion

Our cost analysis shows a sizable benefit of chiplet architectures for large data-center chips. In 7nm, we estimate cost savings of up to 20% for breaking up very large die and lesser savings for die as small as 400mm^2. Rising wafer costs enhance the value of the chiplet approach, pushing the potential savings for very large die to 30% in 5nm and possibly 40% in 3nm. In these advanced nodes, chiplets will be attractive for even moderate-sized die of 300mm^2 or less, although the savings will naturally be smaller for these less expensive designs. This analysis excludes other potential benefits, such as cutting manufacturing cost by building part of the design in a trailing node, or reducing design cost by reusing chiplets in multiple products.

This analysis is consistent with chiplet deployment to date. Most chiplet products in production would be reticle sized (or larger) if implemented as a single 7nm chip, making this approach best suited to expensive data-center chips. Some vendors refer to their PC processors as chiplet designs, but they merely copackage the main processor and the south bridge, something Intel has done for several years. In 2022, we expect most PC GPUs to adopt chiplets along with some midrange networking chips and FPGAs. Companies designing internal ASICs will also begin adopting the technology.

Chiplets aren't appropriate for all designs. PC and smartphone processors typically measure 150mm^2 or less, so they won't benefit. Vendors such as Intel and Nvidia improve yield by reducing the core count of certain product models to accommodate defective cores; this approach also reduces the cost benefit of chiplets. Heterogeneous chiplet designs (such as Agilex and Ascend) can actually increase the number of expensive tapeouts, although some of these tapeouts often shift to older, lower-cost nodes. Reusing chiplets in multiple products can offset the extra tapeout costs, but given the differing requirements across product segments and across generations, we have seen few reuse examples so far.

Many data-center chips, however, fall in the sweet spot for chiplets. The technology has already seen considerable interest among the leading vendors, with AMD, Intel, and Nvidia all shipping or at least developing chiplet-based products. As these vendors have demonstrated, the benefits go beyond cost savings to include building designs that are bigger than any monolithic chip, reducing schedule risk by decoupling new technologies, and offering flexible product configurations. Other companies building large designs in advanced nodes (including ASICs) should assess this new approach to determine if they can benefit from chiplet technology.

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 microprocessor and SoC design. We analyze not only the business strategy but also the internal technology. Our Linley Processor Conferences allow attendees to hear about the latest chip products and technologies. For more information, see our web site at www.linleygroup.com.*