2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA) | 978-1-6654-3333-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISCA52012.2021.00014

Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families

Industrial Product

Samuel Naffziger, Noah Beck, Thomas Burd, Kevin Lepak, Gabriel H. Loh, Mahesh Subramony, Sean White
Advanced Micro Devices, Inc.

Abstract— **For decades, Moore's Law has delivered the ability to integrate an exponentially increasing number of devices in the same silicon area at a roughly constant cost. This has enabled tremendous levels of integration, where the capabilities of computer systems that previously occupied entire rooms can now fit on a single integrated circuit.**

In recent times, the steady drum beat of Moore's Law has started to slow down. Whereas device density historically doubled every 18-24 months, the rate of recent silicon process advancements has declined. While improvements in device scaling continue, albeit at a reduced pace, the industry is simultaneously observing increases in manufacturing costs.

In response, the industry is now seeing a trend toward reversing direction on the traditional march toward more integration. Instead, multiple industry and academic groups are advocating that systems on chips (SoCs) be "disintegrated" into multiple smaller "chiplets." This paper details the technology challenges that motivated AMD to use chiplets, the technical solutions we developed for our products, and how we expanded the use of chiplets from individual processors to multiple product families.

Keywords: Chiplets, Moore's Law, Processors, Modular, Industry

I. Introduction

Moore's Law had set the pace for the semiconductor industry for decades with a reliable generation-upon-generation increase in transistor density and a corresponding reduction in cost per transistor [18]. One direct consequence of Moore's Law was the steady miniaturization and integration of complex computer systems into fewer components, from room-sized computers in the first half of last century to today's mobile and wearable devices.

In recent years, the pace of Moore's Law has slowed. While new silicon process nodes continue to be introduced, the cadence has decreased compared to historical rates. However, the challenges for the semiconductor industry span much more than having to wait a little longer for more transistors. Some of these challenges include rising manufacturing costs for upfront expenses like mask sets as well as cost per chip, increased complexity of design rules in leading-edge nodes, and the architectural challenges of meeting the relentless demand for more and more computational power. We will discuss these trends and challenges in greater detail in Section II, but it is the simultaneous combination of all these challenges that compelled the reassessment of the traditional integration story.

Over the last several years, AMD has been writing a new post-Moore's Law story that revises the historical trend of integrating more functionality per silicon chip and instead is *disintegrating* the traditional monolithic silicon chip into multiple smaller "chiplets." Section III explains the AMD chiplet approach and in particular how it addresses the challenges of a post-Moore's Law world.

While there has been a range of activities and research in partitioning systems of chips (SoCs) into multiple silicon die over the years [4][5][8][19][30], and the concept of multi-chip modules (MCMs) dates back even further [7][22][26][31], AMD is taking the theory of chiplet-based architectures and applying it to design real, high-volume, commercially-successful products. In Section IV, we explain how we started deploying the chiplet approach in the high-performance CPU server space. However, the post-Moore's Law challenges are not limited to the server market, and Sections V and VI discuss how coordinated chiplet designs enable significant reuse to effectively deliver solutions across a range of markets. Section VII reflects on some of the key learnings regarding what has made the chiplet approach such a success for AMD.

II. Chiplets Motivation

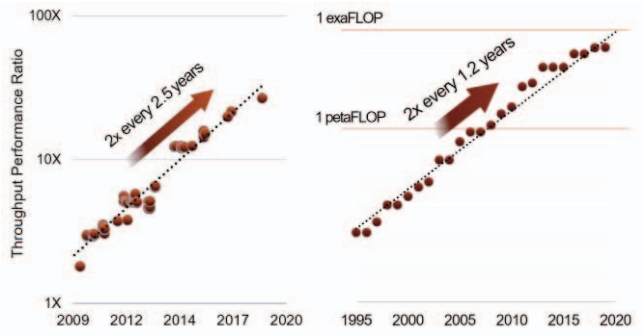
A. The Insatiable Demand for More Compute

The world's computational demands have been increasing exponentially over time. Figure 1 shows historical performance trends both at the component level [28] and the system level (using the world's fastest supercomputers as a representative example) [29]. The performance growth of the top supercomputers has been increasing at a rate faster than Moore's Law, approximately doubling in peak floating-point operations per second (FLOPS) every 1.2 years. With the advent of the world's first exascale supercomputers [13][15], the global computing ecosystem's desire for more computational power does not appear to be slowing down.

The recent scramble to understand and find effective mitigations for the SARS-CoV-2019 virus provides a very concrete and societally-relevant example of just how much more could be accomplished if the world had more, and more powerful, computational resources on hand [1]. The thirst for

This paper is part of the Industry Track of ISCA 2021's program.

978-1-6654-3333-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ISCA52012.2021.00014



(a) (b)

Figure 1. (a) 2P server Specint®_rate2006 performance trend over time, (b) world's fastest supercomputers over time.

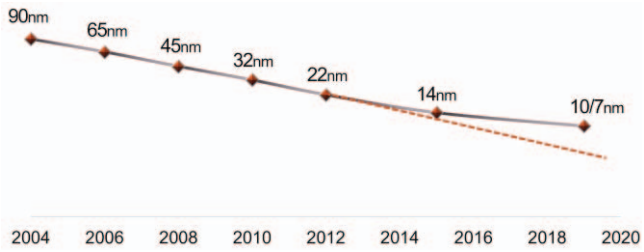


Figure 2. Process technology node introduction over time.

more compute is also clear in other areas such as seen with the explosion of machine learning (ML) and the massive computational demands that come with both the training of and inferencing with the latest algorithms and models. The number of parameters in the largest ML models over the past couple of years from GPT-1 [23] to DeepSpeed [24] has roughly been doubling every 0.2 years, and more recently the Switch Transformer model has broken the trillion parameter threshold [10]. Similarly, the amount of computational power required to train leading-edge ML models appears to be doubling approximately every 3.4 months, reflecting both increases in the model sizes and computational changes in the underlying algorithms [21]. We strongly believe that the need for more compute will continue for the foreseeable future, but the computing industry faces several challenges that we now discuss in more detail.

B. Moore's Law is Breaking Down

The original formulation of Moore's Law was not about processor performance but rather an observation about device density, and in particular that it was doubling approximately every year [18]. Over the past few decades, the rate has settled to be closer to a doubling every 18-24 months. However, for recent process nodes, the rate at which new technologies are introduced has slowed down. Figure 2 shows AMD internal estimates for the approximate introduction dates of major process nodes over approximately the last fifteen years [27]. From the 90nm node down to 22nm, the introduction

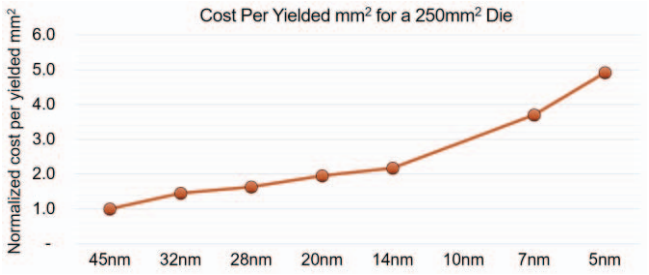


Figure 3. Normalized cost per chip versus technology node.

of new technologies followed a steady cadence of approximately one generation every two years. Starting with the 14nm node, we observed that the inter-node gap extended to approximately three years, and from there to the next major node the gap increased again to more than three years. The exact timing on the availability of future nodes is not yet entirely known, but it is unlikely that the industry will return to the predictable pace of yesterday's Moore's Law in a sustainable manner.

While not explicitly called out in Moore's original paper [11], an important economic consequence of Moore's Law was that each generation's device density increase came at roughly the same cost as the previous generation. Not only does device density increase exponentially, but the cost-per-transistor similarly exhibited a corresponding exponential decrease. However, over the past decade and more, the cost to manufacture an integrated chip has steadily been climbing, with a sharp increase in the latest generations due to reasons such as increased mask layers (e.g., for multiple patterning), more challenging and complex manufacturing (e.g., advanced metallurgy, new materials), and more. Figure 3 shows the cost for a yielded 250mm^2 die over time, normalized to the 45nm process node ("yielded" die costs includes the amortization of the expense for any defective chips) [28]. The 10nm process node in Figure 3 was omitted because AMD transitioned from 14nm technology directly to 7nm. So not only are processor manufacturers waiting longer for each new process node, but they must also pay more when the technology becomes available.

C. Bigger Chips Will Not Save Us

If the rate of introduction of new, denser silicon technologies is slowing down, one possible approach is to instead build larger chips to achieve the desired higher overall device counts. In fact, this approach has already been utilized in some segments, as Figure 4 shows the die-size trends over time for GPUs and server CPUs [27]. At the high end of the market, the higher average selling prices of products can offset the higher manufacturing costs of larger chips. However, the industry is now running up against the lithographic reticle

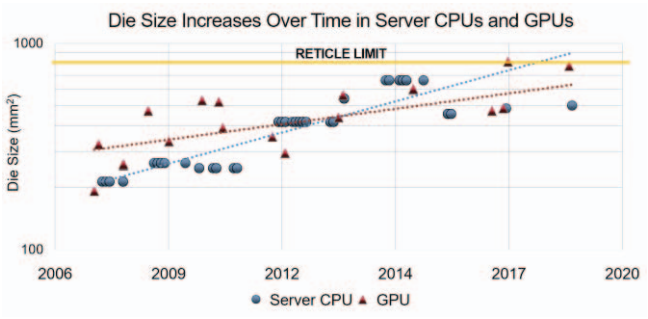


Figure 4. Die size increases for large SoCs over time; die sizes are reaching the reticle limit and further increases are not feasible.

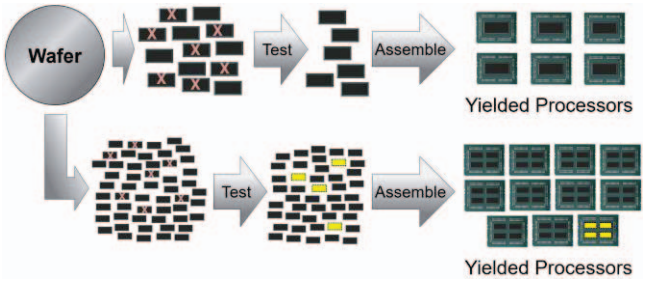


Figure 5. Illustrative construction of processors using (top) monolithic die and (bottom) reassembled chiplets. The light/yellow chiplets represent chiplets that can run at higher clock speeds.

limit, which is a practical ceiling on how large silicon die can be manufactured.

D. The Product Portfolio Multiplier

The above discussions regarding the challenges of leading-edge silicon technologies apply not only to a single chip design, but rather the impacts may be multiplied over a company's broader portfolio of products. As an example, consider a server CPU lineup that consists of five products with 16, 24, 32, 48, and 64 cores. Each of these products could represent a separate tapeout, which comes with their own mask sets, yield and cost profiles, etc. Beyond the silicon costs, there are many additional upfront costs that must be accounted for on a per-SoC basis. For example, each unique chip requires its own physical design (e.g., floorplanning, power delivery, clocking), test and debug, validation, firmware, power and thermal management optimization, etc. Given a finite engineering budget, a possible consequence of rising costs in a post-Moore's Law world could be the reduction in the number of products that a company could offer at a time when customer demand for more and differentiated products continues to grow.

III. Chiplets to Extend Moore's Law

The overall problem statement is that semiconductor companies need to continue delivering products with a larger number of transistors to provide customers with more functionality and computing capability at market-friendly price points, but the unraveling of Moore's Law delays the availability of the new process nodes that would deliver the additional devices, and the costs are also increasing.

A. The Chiplet Approach

The overall idea with chiplets is to take what would normally be a monolithic, single-die SoC, and then partition it into multiple smaller die or "chiplets" and then "reintegrate" them with some form of in-package interconnect to enable the collective to operate as a single, logical SoC. The reason why this approach can economically make sense is that the cost of silicon is not a linear function of chip area. For example, a chip with $T/2$ transistors may cost considerably less than half the cost of a chip with T transistors. In general, if an SoC with T transistors can be partitioned into n separate chiplets, such that the combination of those n chiplets provides the equivalent functionality of the original T-transistor SoC while the sum of the costs of the individual chiplets plus any additional costs for reintegration (e.g., additional packaging expenses) still comes in lower than the cost of a monolithic T-transistor SoC, then a chiplet implementation for this SoC may be worthwhile.

For very large SoCs, for example those approaching the reticle limit (Figure 4), the yield rates of such large chips can be economically very challenging. The top of Figure 5 shows an abstract depiction of a conventional manufacturing flow using monolithic SoCs. Each chip is constructed on the wafer using standard lithographic procedures to build the front-end transistors and the back-end metal layers. After chip construction, each SoC die undergoes a test procedure to determine functionality, also commonly called "known good die" (KGD) testing. Each unrepairable manufacturing fault can result in an entire SoC's worth of silicon being discarded (thereby burdening the remaining functional parts with the added costs). The SoC die that are functional can then be assembled with the final packaging solution, which results in some number of yielded processors that can now be sold.

With a chiplet-based approach, the lower portion of Figure 5 shows how the same hypothetical SoC has been partitioned into chiplets where each chiplet has approximately one quarter of the original SoC functionality (e.g., one fourth the core count). Each chiplet is manufactured using the same standard lithographic procedures as in the monolithic case to produce to a larger number of smaller chiplets. The individual chiplets then undergo KGD testing. Now, for the same fault distribution as in the monolithic case, each potential defect results in discarding only approximately one-fourth of the amount of silicon. The chiplets can be individually tested and then reassembled and packaged into the complete final SoCs. The overall result is that each wafer can yield a significantly larger number of functional SoCs.

Beyond functional testing, individual chiplets can also be tested for maximum performance (e.g., clock speed) [6][14]. Figure 5 shows a few light-colored/yellow chiplets that indicate samples that can achieve higher frequencies, for example, due to parametric variations in devices across the wafer. These faster chiplets can be identified, collected together, and assembled into premium parts with all fast cores. In contrast, a monolithic chip may only have a fraction of its cores that fall within the region of the wafer with faster transistors, and as such it becomes statistically far less likely to find a monolithic chip with all fast cores. Beyond raw functional yield rates, chiplet-based assembly can also increase the supply of higher-performing products.

In addition to the yield and cost argument illustrated in Figure 5, chiplets have other potential benefits. Early in the introduction of a new process technology node, the yield rates are often lower than compared to a more mature process. As such, trying to build large SoCs early in the lifetime of a new technology can be even more economically challenging. However, by utilizing a collection of smaller chiplets that yield at substantially higher rates, this can enable one to transition to a new process node earlier than would make sense with a conventional monolithic design.

Another advantage of chiplets is that the lithographic reticle limit only applies to individual chips, but multiple chiplets can be assembled such that their cumulative silicon area exceeds that possible with monolithic designs. Commercial examples of these and other advantages will be further illustrated in later sections.

B. Chiplets are Not a Free Lunch

While the chiplet approach to SoC construction has a lot of potential advantages, it also introduces some new costs and complexities. A chiplet design requires more engineering work upfront to appropriately partition the SoC into the right number and right kinds of chiplets. There are a combinatorial number of possibilities for partitioning an SoC, but not all may satisfy cost constraints, performance requirements, ease of IP and silicon reuse, and more.

Chiplets also require new inter-chiplet communication paths [6]. Compared to on-chip metal, these interconnects involve longer routes with potentially higher impedances, lower available bandwidth, higher power consumption, and/or higher latency. The interconnect overhead may also include circuits for crossing voltage and timing domains, protocol conversions, and/or serializer-deserializers (SerDes), and these circuits all represent additional power and silicon area overheads that would not have been present in a monolithic design.

In addition to the inter-chip communication interfaces, other circuits and functionality may also need to be replicated on a per-chiplet basis. Some examples include test and debug

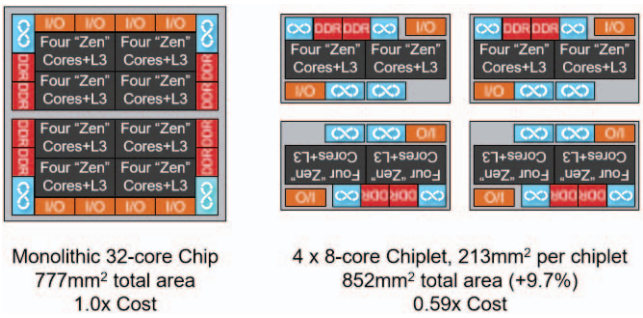


Figure 6. Hypothetical monolithic 32-core chip compared to an assembly of four eight-core chiplets.

interfaces (e.g., for testing individual die prior to assembly), clock generation and distribution, power management, on-chip temperature sensors, certain types of I/O (e.g., USB, SATA), and more. As a result, the total silicon area of a single monolithic T-transistor SoC, Area(SoC(T)), will typically be less than the total area of n chiplets where each chiplet has a die area of Area(SoC(T/n)+K), where K represents the additional overheads discussed above (e.g., inter-chip interfaces, test circuitry). However, using more total silicon area for n separate chiplets may still result in lower total cost compared to a monolithic approach (Section III-A).

C. Why Chiplets Now?

The fundamental idea of SoC partitioning is not new, and multi-chip module (MCM) technologies have been around for decades [7][9][22][26][31]. However, many past MCM applications have been sequestered to relatively narrower use cases and markets. What is now new is that the post-Moore's Law challenges discussed in Section II are changing the semiconductor industry landscape and creating pressures toward adopting chiplet approaches even for mainstream, high-volume markets.

IV. Case Study: AMD EPYC™ Processors

AMD EPYC™ server CPUs are our first products to utilize a modern chiplet-based design methodology. In this section, we discuss two generations of AMD EPYC™ processors to highlight the evolution in the approach and share some of the design decisions and benefits.

A. First-generation AMD EPYC™ Processor

AMD market analysis and our product definition process set a target of 32 cores for our first-generation AMD EPYC™ processors, formerly codenamed "Naples," to compete in the server CPU market. In addition to raw core count, we also targeted an SoC supporting eight DDR4 memory channels and 128 lanes of PCIe® gen3 I/O to provide industry-leading memory and I/O bandwidths, both of which are frequently high priorities for server, cloud, and enterprise use cases.

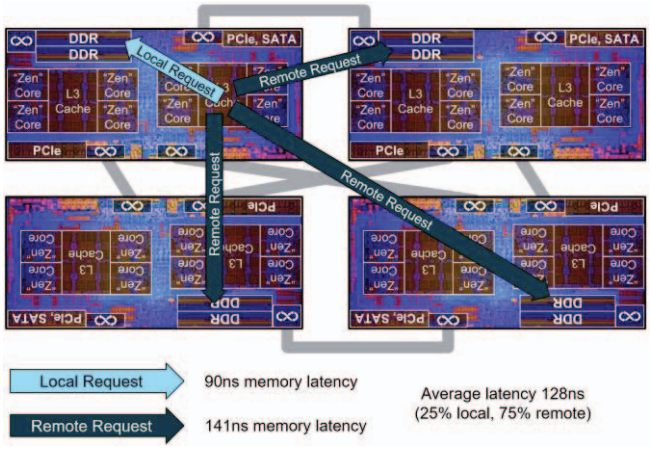


Figure 7. Connectivity of the four "Zeppelin" chiplets and memory latencies for local, remote, and average requests.

Figure 6(a) shows a schematic of a hypothetical monolithic 32-core processor. Based on our internal analysis and product planning exercises, such a processor would have required 777mm^2 of die area in a 14nm process. While still within the reticle limit and therefore technically manufacturable, such a large chip would have been very costly and put the product in a potentially uncompetitive position.

The first-generation AMD EPYC™ processor utilized a design with four identical chiplets, codenamed "Zeppelin," shown in Figure 6(b) [3][6]. Each chiplet provides eight "Zen" CPU cores, two channels of DDR4 memory, and 32 lanes of PCIe such that the combination of the four chiplets meets our product requirements. Additional die area was also required to implement our Infinity Fabric™ interconnect ∞ between the four chiplets [6] as well as other per-chip circuitry as discussed in Section III-B. As a result, each chiplet had a die area of 213mm^2 in a 14nm process, for a total aggregate die area of 4×213mm^2 = 852mm^2. This represents a ~10% die area overhead compared to the hypothetical monolithic 32-core chip. Based on AMD-internal yield modeling using historical defect density data for a mature process technology, we estimated that the final cost of the quad-chiplet design is only approximately 0.59× of the monolithic approach despite consuming approximately 10% more total silicon.

Beyond the cost savings for the 32-core product, the chiplet approach also provides a flexible platform for reusing the chiplets across multiple product offerings. For example, after testing individual chiplets, some may have four or fewer cores that have been rendered inoperable due to manufacturing defects [20]. However, these four "harvested" chiplets (each still with four functional cores) can be assembled into a 16-core processor. Due to the chiplet-based design, this 16-core processor can also be "fully featured" with all eight DDR4 memory channels and 128 PCIe I/O lanes. Network and storage-oriented use cases may only need a moderate number of CPU cores to keep the storage and/or networking fully utilized; any additional cores beyond that represent unnecessary costs for the end user.

Harvesting is highly advantageous for several reasons. The first is that we can utilize a higher fraction of the total number of chips per wafer even in the face of some manufacturing defects. The second is that with just a single chiplet design (i.e., one mask set, one tapeout), we can deliver multiple different products that traditionally would have required multiple separate unique SoCs. Third, the chiplet approach makes it more practical to offer products with a full complement of memory and I/O capabilities. In contrast, a dedicated 16-core monolithic SoC might not by itself be able to tolerate the cost burden of the additional die area needed for so many memory channels and I/O lanes. However, the lower cost of the individual chiplets combined with the ability to amortize the additional memory and I/O interfaces over multiple products can make this an economically viable approach while enabling valuable product differentiation to meet customer needs.

The multi-chiplet design of the first-generation AMD EPYC™ processor introduces additional interconnect latency when chiplets need to communicate across the Infinity Fabric™ on-package (IFOP) interconnect, which are implemented as point-to-point links directly on the organic package substrate [6]. The IFOP links utilize custom high-speed SerDes circuits. Compared to SerDes for off-package I/O like PCIe gen3, which consumes approximately 11pJ per bit, the IFOP SerDes have been carefully optimized for shorter package substrate route lengths and achieves a power efficiency of ~2pJ per bit. Transmitting data over the IFOP links still represents a power overhead compared to a monolithic chip, where on-chip interconnect power is typically much less than 1pJ per bit, with the exact power cost depending on the route length and other factors.

For the eight CPU cores on a given chiplet, only two out of the eight total DDR4 memory channels are resident on the same chiplet. This means that in the absence of non-uniform memory access (NUMA) data management and thread pinning, some memory requests must be serviced by "remote" memory channels, as shown in Figure 7.

Based on AMD internal testing with requests that generate DRAM page misses, the typical latency for a memory request to a local memory channel (i.e., on the same chiplet) was measured to be 90ns, whereas accesses to remote memory channels (i.e., different chiplet, same socket) incur a latency of 141ns. The additional latency is due to the round-trip combination of the IFOP links and additional hops through each chiplet's local data fabric (also known as a network-on-chip or NoC). For a memory access pattern uniformly interleaved across the eight memory channels, this results in an average memory latency of 128ns. These intra-socket NUMA effects

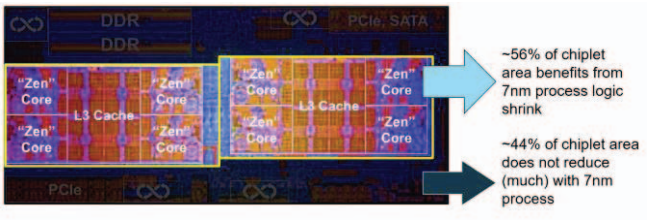


Figure 8. A significant fraction of the "Zeppelin" chiplet area is dominated by DDR PHYs, SerDes, and other I/O that hardly reduce in size from moving to a 7nm process.

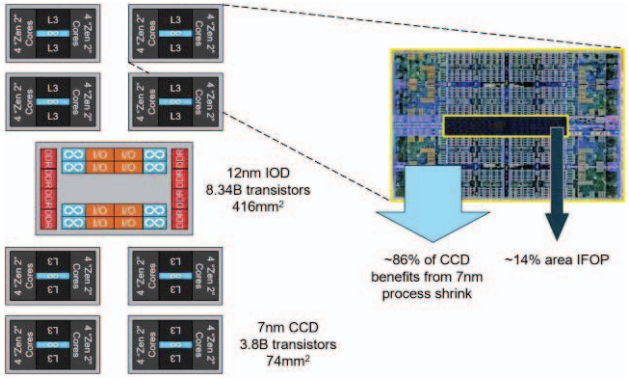


Figure 9. (a) Second-generation AMD EPYC™ processor consisting of a 12nm IOD die and up to eight 7nm CCDs, (b) CCD detail showing majority of chiplet area occupied by cores and L3 cache.

are one of the items (among others) that our second-generation AMD EPYC™ processor addresses.

B. Second-generation AMD EPYC™ Processor

The first-generation AMD EPYC™ processor was very well received by the market, and we set our sights to be even more aggressive for our second generation of server CPU products [20]. The timing of the second-generation AMD EPYC™ processors also aligned with the dawn of 7nm silicon technology, which provided both benefits and challenges that required new innovations in our chiplet methodology.

1) 7nm Benefits and Cost Challenges: Compared to the 14nm process technology used for the first-generation AMD EPYC™ processor, the 7nm process that was becoming available was very promising from a device perspective. Based on AMD-internal analysis of a 14nm product ported to a 7nm node with a similar implementation flow and design methodology, we projected that we could obtain a doubling in transistor density for the core logic. The 7nm devices also delivered significant improvements in transistor speed and power efficiency. From the same study, we projected that at the same power consumption, the 7nm version could deliver over 25% more performance, or for the same performance, power could be reduced by approximately one half. Starting from our previous 32-core design point, the increase in device density with a corresponding power-efficiency improvement meant that a 64-core product might be within striking distance. However, Figure 3 showed that the transition to the 7nm technology node also comes with increasing die cost, and so further innovation was required.

2) Hybrid Multi-die Architecture: We initially started by analyzing the "Zeppelin" chiplet and considered a hypothetical 7nm version of it. We found that such an approach would be challenging to make work with the higher cost of the latest silicon technology. In particular, high-performance server products demand a lot of memory and I/O, which occupy a significant fraction of the first-generation chiplet. Unfortunately, most of these structures do not scale well with shrinking device geometries either due to the analog devices or from being limited by the bump pitch of the external I/O connections. As shown in Figure 8, the CPU cores, L3 cache, and other logic account for approximately 56% of the "Zeppelin" chiplet area. Rather than halving the die size from going to 7nm, we would only achieve approximately a 28% reduction (i.e., 56%/2 + 44%=72%). When considering the relative cost increase of transitioning from 14nm to 7nm, this 28% die size reduction might not be sufficient to even reach a cost break-even point compared to the original 14nm "Zeppelin" chiplet.

Instead, the second-generation AMD EPYC™ processor, formerly codenamed "Rome," utilizes a dual-chiplet approach. The first chiplet, called the I/O die (IOD), was implemented in a mature and cost-effective 12nm process. The IOD has a size of 416mm^2 with 8.34 billion transistors, and it contains the full complement of eight DDR4 memory channels, 128 lanes of PCIe gen4 I/O, other I/O such as USB and SATA, the SoC data fabric, and other system-level functionality. The second chiplet, the core-complex die (CCD), was implemented in the leading edge 7nm node. Each CCD is only ~74mm^2 in size, leading to very good yield rates even in the early days of a new process node. Figure 9(a) shows how one IOD can be assembled with up to eight CCDs. Each CCD provides eight "Zen 2" CPU cores, and so all together this arrangement can enable an impressive 64 cores in a single socket. The CCD attempts to utilize as much of the high-performance, but more costly, 7nm silicon for the functions that benefit from the advanced devices, namely the eight CPU cores and the L3 cache. Only a small portion of the CCD area is consumed for the IFOP, which is placed centrally on the chiplet to minimize distance from the L3 cache banks. Figure 9(b) shows how the CCD improves the utilization of the 7nm chiplet so that the CPU cores and L3 caches now account for 86% of the total chiplet area.

3) Packaging Technology Decisions: AMD was among the first companies to commercially introduce silicon interposer technologies starting with the AMD Radeon™ R9 "Fury" GPUs with high-bandwidth memory (HBM) in 2015 [16]. A

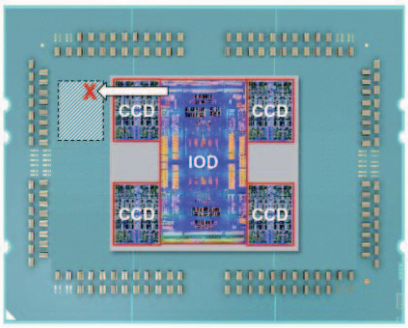


Figure 10. Hypothetical silicon interposer-based chiplet architecture. Maximum interposer size and unbuffered interposer routing limits configuration to four CCDs.

natural question for our chiplet-based products is why we chose to use package substrate routing rather than the higher-density interconnects enabled by silicon interposers. There are several factors that drove the decision to not use silicon interposers for our chiplet-based processors. The first is the communication requirements of our chiplets. With eight CCDs and eight memory channels, on average each chiplet's IFOP only needs to handle approximately one DDR4 channel's worth of bandwidth. Using DDR4-2933 as an example, a single channel would correspond to ~23.5 GB/s of peak bandwidth. Even accounting for some load imbalance across the CCDs, a single CCD's IFOP would still be expected to observe no more than a few tens of GB/s of traffic, and in fact each link can support approximately 55GB/s of effective bandwidth. Point-to-point links in the package substrate routing layers are more than sufficient to handle this modest level of bandwidth. In contrast, a single HBM stack can deliver hundreds of GB/s of memory bandwidth, which far exceeds the capabilities of the organic package substrate, and this is why HBM-enabled GPU products need a higher-bandwidth solution such as silicon interposers [2][16][17].

The second factor against silicon interposers for our chiplet-based processors is the reach of the interposer-based interconnects. While interposers can provide great signal density for very high bandwidths, the lengths of the signals are limited and as such constrain the connections to edge-to-edge links. The reach of interposer-based interconnects can in principle be extended using wider metal routes and greater spacing between routes, but this would decrease the effective bandwidth per interface because fewer total routes could be supported for a fixed width of routing tracks. This argument also applies to silicon bridge technologies [12]. The next subsection describes the challenges of providing sufficient IFOP bandwidth across the package substrate. Figure 10 illustrates a hypothetical interposer-based processor design. The edge-connectivity constraint would limit the architecture to only four CCDs, which would render the product concept to be far less compelling. Even if interconnect reach was not a limiting factor, the IOD and the eight CCDs would require so much area that the underlying interposer would greatly exceed the reticle limit (while a passive interposer does not contain any transistors, the metal layers are still lithographically created and therefore must stay within the same reticle field constraints). Figure 10 shows the placement where an additional CCD would have to be, which is both outside the boundary of a maximum-sized interposer and too far for the unbuffered interposer routes to reach while supporting required bandwidths. Recent advancements in silicon interposer manufacturing have enabled reticle stitching to create very large interposers [11], but such an approach would have been cost prohibitive for this market segment. Last, the silicon interposer itself adds more cost to the overall solution. A CCD with the twice the core count could have been used, but that would have resulted in lower yield and decreased configurability. For all these reasons, routing IFOP directly across the package substrate was chosen for this product family.

The total area consumed by multiple chiplets is typically greater than a monolithic chip with equivalent functionality. While this could theoretically cause a corresponding increase in the overall package size, the size of the SP3 processor package used by AMD EPYC™ processors is primarily determined by the large number of package pins required to support the eight DDR memory channels, 128 lanes of PCIe plus other miscellaneous I/O, and all the power and ground connections.

4) Chiplet-Package Co-design Challenges: While routing on the package substrate was the only practical option, that does not imply that it was simple to do. This section discusses some of the challenges and the solutions related to package-level routing and power delivery. These are topics that are often outside the attention of many computer architecture researchers, but we wish to highlight how high-level architecture decisions can have major downstream impacts on the rest of the overall design.

The package routing layers are already heavily utilized not only for IFOP but also for external I/O connections, escaping out the multiple DDR4 channels, and delivering power and ground across the entire package. Figure 11(a) shows a schematic view of the first-generation AMD EPYC™ processor's package routing for the DDR4 memory buses, PCIe I/O lanes, and IFOP. Figure 11(b) shows how the package routing resources have already been consumed by the first-generation AMD EPYC™ processor [3]. To enable the necessary package-level connectivity for our second-generation architecture, significant co-design was required between our packaging and silicon teams.

Figure 11(c) shows the package routing for the second-generation AMD EPYC™ processor. The overall layout and floorplan for the CCDs, the IOD, and the package routing had to be coordinated from the outset. Note that to provide our customers with the option for seamless upgrades, the first and

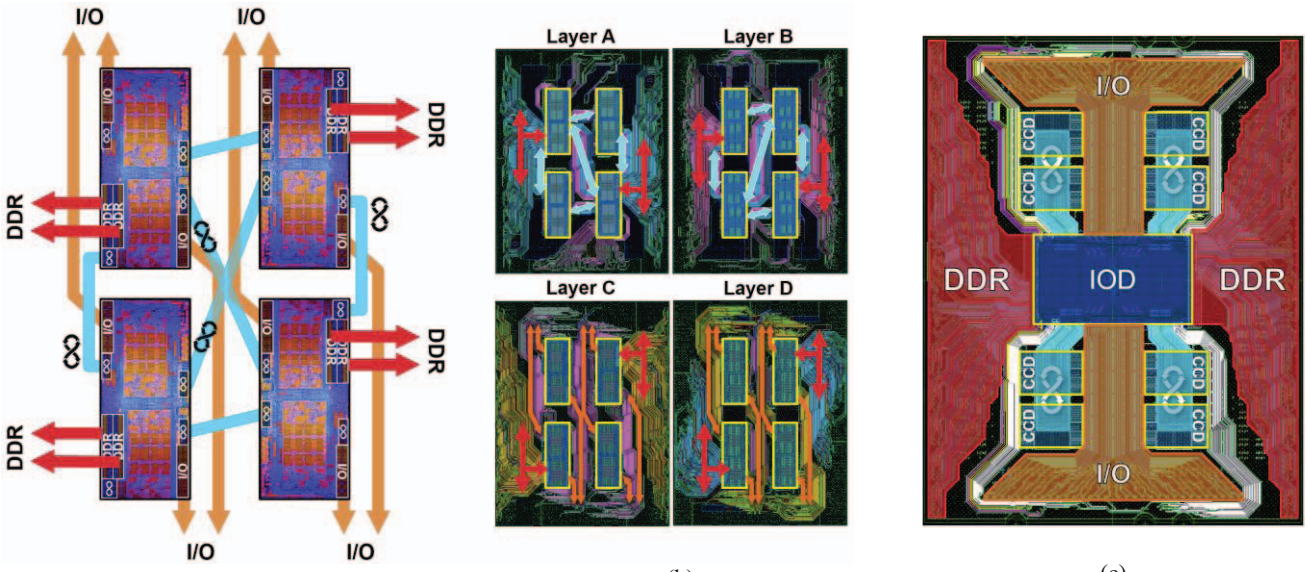


Figure 11. (a) Schematic view of first-generation AMD EPYC™ processor package routing for DDR (red), I/O (orange), and Infinity Fabric™ links (cyan), (b) multi-layer package routing layout of the first-generation AMD EPYC™ processor package, and (c) annotated package routing layout of the second-generation AMD EPYC™ processor package.

second-generation AMD EPYC™ processors are designed to be socket compatible. However, that creates a challenge to providing connectivity to all the CCDs because we cannot increase the package size to make room for more routes. There are four "inner" CCDs directly adjacent to the central IOD, and then there are four more "outer" CCDs toward the top and bottom edges of the package. The enabling co-design innovation was finding a way to route the IFOP directly underneath the CCDs. Figure 11(c) shows how this opens clear routing paths between the two columns of CCDs to allow the PCIe routes to escape to the top and bottom edges of the package. In a similar fashion, the multiple channels of DDR4 can escape directly from the IOD to the right and left edges of the package.

Unfortunately, routing underneath the CCDs only works if there are available routing resources in the package.

Figure 12(a) shows a simplified view of the VDDM power delivery on the "Zeppelin" chiplet (VDDM is a regulated power supply rail supporting the L3 cache SRAM arrays) [25]. The real core supply (RVDD) is not shown for clarity, but it is also distributed in the package layer as input to the on-chip regulators. VDDM is delivered by a low-drop out (LDO) linear regulator, which drives the power down to the thick copper layers in the package substrate to help distribute VDDM across the entire area of the L3 cache. The power can then be delivered back up to power the L3 cache. The low-resistance package layers are very effective for power distribution, but unfortunately this is where we also want to route IFOP for the second-generation architecture.

Figure 12(b) shows how the package and CCD were jointly co-designed to simultaneously address power delivery requirements while freeing up the necessary package routing resources to enable IFOP to tunnel underneath the CCDs. The VDDM distribution was brought on to the CCD and it instead utilizes redistribution layer (RDL) metal. The challenge was that the RDL metal is more resistive than the thick copper layers in the package, and so the span (power distribution distance) of the LDOs had to be reduced.

Figure 12(b) shows how the LDOs have been reoriented and distributed along the sides of the L3 cache. The LDOs drive power down to the RDL, and then only need to fan out a relatively shorter distance through the resistive RDL routes. This carefully choreographed chip-package layout kept the VDDM IR drop to within 10mV while freeing up the package routing layers (labeled "Signal Routes" at the bottom of Figure 12(b)) for IFOP to pass through.

Another packaging-related challenge was that the chiplets manufactured in different process nodes normally would have utilized different bump pitches to connect the chiplets to the package substrate. In particular, the 12nm IOD bump pitch is 150µm, while the 7nm CCD bump is 130µm. The result would have been that the different types of chiplets would be at different heights after assembly, which could potentially be problematic for providing a level and uniform interface to the cooling solution. To address this, we engineered a copper pillar-based bump for the IOD, which was also compatible with the 7nm CCDs, to ensure uniform post-assembly chiplet heights. The transition to a copper-pillar package interface also provided denser bump pitches and enabled higher maximum current (electromigration) limits.

5) Improved Memory Performance: The first-generation AMD EPYC™ processor's organization with memory and

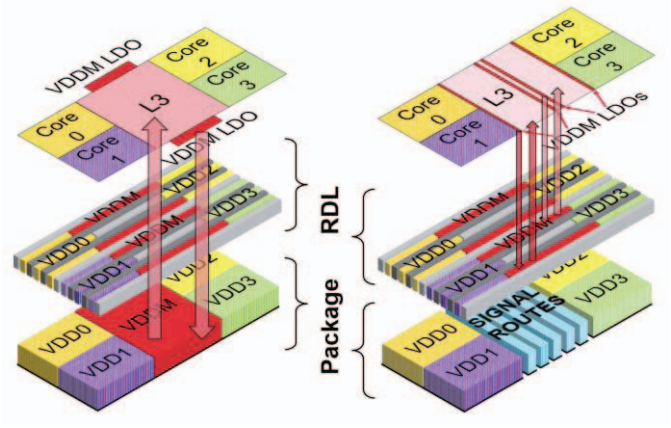


Figure 12. (a) "Zen" CPU VDDM distribution via the package plane, (b) "Zen 2" CPU VDDM distribution via RDL only.

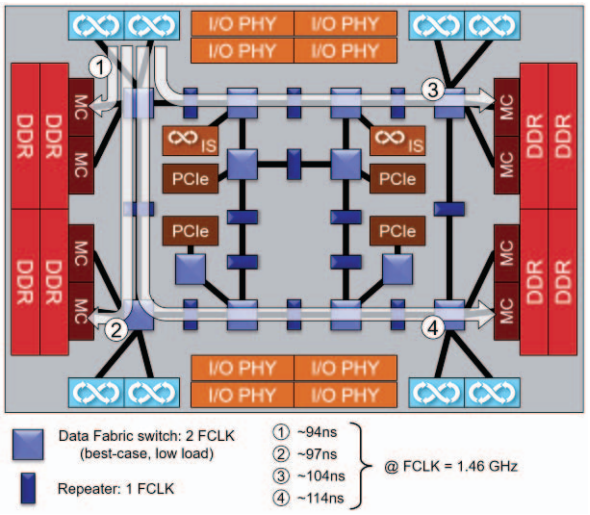


Figure 13. IOD data fabric topology and measured memory latencies to the four quadrants.

I/O distributed across the four chiplets provided an architecture that only required a single chiplet design, but it also introduced intra-package NUMA effects as discussed in Section IV-A. Furthermore, IFOP must support demand requests from the chiplet to the remote memory channels, requests from the remote cores to the local memory channels, and I/O in both directions as well.

The overall layout of the second-generation AMD EPYC™ processor's chiplets resembles a star topology, which provides much more uniform memory access latencies. Each memory request from a CCD takes a direct hop to the IOD, and then from there the high-performance data fabric routes the request to one of the eight targeted memory channels. Note that some memory channels are still closer or farther from each of the CCDs and so some NUMA effects remain, but they are greatly reduced compared to the prior generation approach. Figure 13 shows the overall IOD data fabric topology. The data fabric utilizes a hybrid ring-ladder topology, where the demand memory requests are routed along the external ring, and the I/O traffic moves along the interior "ladder" of the data fabric (IS in the figure is for inter-socket (IS) traffic for symmetric multiprocessing (SMP) platforms). Figure 13 also shows the latencies from one of the CCDs to memory channels in each of the four corners of the IOD (measured on an AMD EPYC™ 7002 Series processor with DDR2933 memory with DRAM page-missing traffic at low load). In contrast to the first-generation organization where some memory requests could be locally serviced on a given chiplet without any IFOP hops, in the second-generation approach all requests always require an IFOP hop to get from the CCD to the IOD. Despite this mandatory IFOP hop, the "local" memory access latency (labeled ① in Figure 13) is only 4ns slower than the same-chiplet memory access in the first-generation architecture (i.e., 94ns versus 90ns) due to intense engineering efforts to squeeze out cycles along the memory paths. The path to the next-closest pair of memory channels ② incurs approximately 3ns of additional latency roundtrip, or approximately 4 FCLKs (fabric clocks). In the worst case to access the memory channels on the opposite corner of the IOD ④, the data fabric adds approximately 7 FCLKs in each direction, or approximately 20ns roundtrip assuming a 1.46GHz FCLK. Note that this is a ~30ns improvement over the inter-chiplet memory latency of the first-generation AMD EPYC™ processor. For a memory access pattern uniformly distributed across all eight channels, the average memory latency improves by ~24ns (~19% reduction), and the difference between nearest and farthest memory channels improves from ~51ns down to ~20ns, which represents a ~61% reduction in the variance/range of in-socket NUMA effects of the overall design.

6) Overall Cost and Performance: An objective of the second-generation AMD EPYC™ processor was to enable an architecture that delivered scalable performance with a cost structure that also scaled linearly with the capability of the system (in contrast to monolithic chips where the cost scales super-linearly with SoC function/die size). Figure 14 shows the relative cost of five different possible configurations ranging from 64 cores down to 16 cores. For comparison, we also show the projected cost for hypothetical monolithic SoCs with equivalent core counts. Note that no cost is shown for a monolithic 64-core SoC because the total chip size would be greater than 1000mm^2, which greatly exceeds the reticle limit.

The chart illustrates a few important trends that really demonstrate the power and value of our chiplet methodology. The first is that across all configurations, the final silicon cost is significantly lower than any of the monolithic equivalents. The second is that the cost scales linearly with a gentle slope

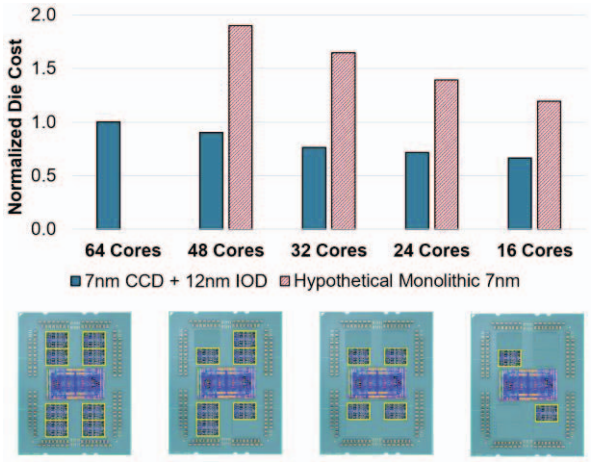


Figure 14. Normalized processor costs as core counts vary compared to hypothetical monolithic die, and visualization of core-count configuration by varying the number of CCDs.

	EPYC™ 7601 CPU (180W TDP)	EPYC™ 7742 CPU (225W TDP)	Improvement
Cores	32	64	100%
Total Transistors (B)	19.2	38.74	102%
Total Silicon Area (mm2)	852	1008	18%
SPECrate® 2017_int_base (2P)	272	663	144%
SPECrate® 2017_fp_base (2P)	259	511	97%

Figure 15. Comparison of first- and second-generation AMD EPYC™ processors.

as the core count is varied. The bottom portion of Figure 14 also illustrates how the different core counts can be achieved by simply depopulating CCDs from the package. This visually shows how with only two tapeouts (and only one of those being in the leading-edge 7nm node), we are able to flexibly enable an entire server product stack including the 64-core option that would otherwise be both technologically and economically impractical to manufacture.

Figure 15 shows a comparison between the first-generation and second-generation AMD EPYC™ processors. The chiplet approach enabled a doubling in total core count per socket. The total number of transistors more than doubled, with a total count of over 38 billion devices, although total silicon area in the package only increased a modest 18%. This metric simply counts total mm^2 of silicon and does not treat the denser 7nm silicon any differently than 14nm, and so the relatively small increase in total silicon highlights the density advantage of the 7nm process node. Finally, Figure 15 shows the overall performance as measured on SPECrate® for double-socket (2P) server platforms[1]. The performance uplift is a combination of the doubled core count, higher clock speeds, higher IPC of the newer "Zen 2" microarchitecture, and a higher supported power limit (TDP) for the second-generation products.

[1]. Results obtained from the SPEC® website as of January 3, 2020. (http://www.spec.org/cpu2017)

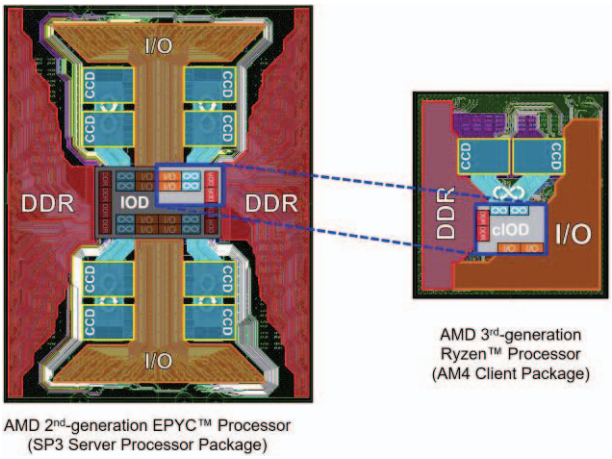


Figure 16. Construction of the third-generation AMD Ryzen™ processor by reusing CCDs and building a client IOD with extensive IP leverage from the server IOD.

Beyond enabling new levels of performance, another benefit of our combined CCD and IOD approach is that it enables more flexible inventory management. With conventional monolithic SoCs, the relatively long lead time of modern silicon manufacturing forces a company to forecast how many of each part should be ordered. If too few of a popular part gets ordered, then a company could potentially face shortages which could potentially lead to missed revenue opportunities as well as customers potentially buying products from competitors. If too many parts are ordered, then a company may be faced with excess inventory that might become challenging to move. The second-generation AMD EPYC™ processor enables a later-binding approach, where we can wait until after the chips have been returned from manufacturing to assemble the chiplets into 16, 24, …, 64-core parts. If market demands shift and we need to produce more of one part or another, we have the potential to be agile and assemble more or fewer chiplets to increase the supply of the desired products.

V. Case Study: AMD Ryzen™ Processors

The chiplet-based architecture of the second-generation AMD EPYC™ processor was a significant feat of silicon-package co-design and proved to be highly effective for addressing the needs of our server, enterprise, and high-performance computing customers. However, AMD also produces many products to serve a wide range of other target markets.

A. AMD Ryzen™ Processor Organization

The "Zeppelin" chiplet is a complete SoC with cores, memory, I/O, and all system functionality for standalone operation. For the first-generation AMD Ryzen™ processors, we were able to take a single "Zeppelin" chiplet and put it in a client AM4 package to provide a desktop processor with eight cores, two DDR4 memory channels, and 24 lanes of I/O.

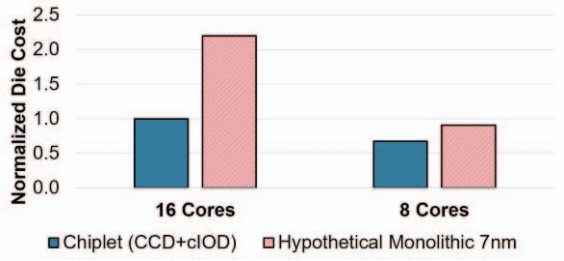


Figure 17. Normalized AMD Ryzen™ processor costs as core counts vary compared to hypothetical monolithic die.

	Ryzen™ 2700X CPU	Ryzen™ 3950X CPU	Improvement
Cores	8	16	100%
1T Fmax (GHz)	4.3	4.7	9%
NT Base Freq. (GHz)	3.9	3.95	1%
Total Transistors (B)	4.8	9.69	102%
Total Silicon Area (mm2)	213	273	28%
Cinebench 1T	434	527	21%
Cinebench NT	4020	8862	120%

Figure 18. Comparison of first- and second-generation AMD Ryzen™ processors.

We later used the silicon from the second-generation AMD EPYC™ processor to create the third-generation AMD Ryzen™ processor with two CCDs and a "client IOD" (cIOD), shown in Figure 16. The CCDs directly utilize the same silicon design as that used by the server products. The cIOD is a new die, but it heavily leverages the server IOD design. Figure 16 shows how the 125mm^2, 2.09 billion transistor cIOD is effectively a quarter-sized version of the server IOD, which features two DDR channels, 32 lanes of PCIe I/O, and two IFOP ports to the CCDs. The result is an industry-leading 16-core high-performance desktop processor without requiring a new 7nm tapeout and a cost-effective and highly-leveraged cIOD. Analogous to the impracticality of a monolithic 64-core server processor, a 16-core desktop processor would likely not be economically practical if implemented as a monolithic die, but chiplets make it possible. Similar to the server products, the overall architecture of the third-generation AMD Ryzen™ processor enables product definition flexibility by simply reducing the number of CCDs to one.

B. Overall Cost and Performance

The modular chiplet design approach of the second-generation AMD Ryzen™ processor provides the same types of cost savings and scalability benefits as we were able to achieve with the second-generation AMD EPYC™ processors. Figure 17 shows the relative die cost of both 16-core (two CCDs) and 8-core (one CCD) chiplet implementations compared to hypothetical monolithic 7nm designs. These results only show the silicon costs and do not reflect the additional engineering benefits from reusing the CCDs and server IOD IP in terms of amortization of both design and verification efforts, product configurability, and time-to-market.

Figure 18 shows a comparison between first- and second-generation AMD Ryzen™ processors. Similar to the comparison of the AMD EPYC™ processors, the chiplet approach and 7nm technology enabled a doubling of core count. Another similar trend is the doubling in overall transistor count while total silicon area increases by only 28%. Performance is reported using the Cinebench R20 benchmark. The single-threaded (1T) performance improvements come primarily from enhancements to the "Zen 2" core microarchitecture and increased clock speeds (both parts were measured with the same 105W TDP). Maximum 1T clock frequency is very difficult to increase, and the "Zen 2" microarchitecture combined with 7nm enabled a critical 400MHz improvement. The all-cores (NT) performance results highlight the benefits of our chiplet approach. The maximum all-cores clock speed is similar between the two generations of processors[2], but the doubling in core count is responsible for the majority of the NT performance gains, with the rest coming from the core IPC uplift, memory system improvements, power-performance efficiency optimizations, and other enhancements.

C. Additional Chiplet Optimizations

In Section III we discussed how prior to assembly, individual chiplets can be tested to match up parts with similar performance. However, parametric variations still exist within each individual chiplet, leading to some potential differences between cores. This effect can be magnified by the higher core counts that our chiplet design approach enables, and we have observed up to 200MHz variations in the maximum clock frequency (F_{max}) across cores in the same CCD. While legacy boost techniques did not take advantage of the faster cores, we now utilize an algorithm that characterizes each CCD's cores at boot time to generate a list of cores in order of F_{max} capabilities. As shown in Figure 19, this list is made

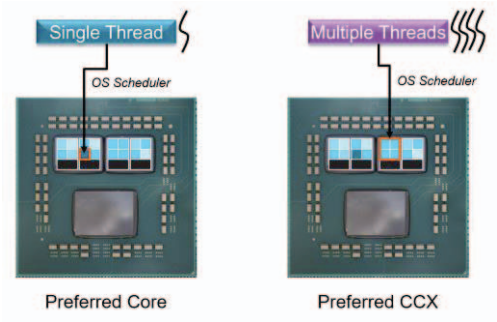


Figure 19. Boot-time characterization of each CCD's cores and core complexes (CCX) enable better utilization of the many cores in the system.

[2] Performance results including measured clock speeds were tested by AMD Performance Labs as of December 13, 2019. Results may vary.

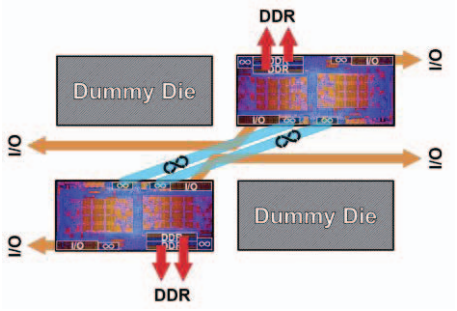


Figure 20. First-generation AMD Ryzen™ Threadripper™ processor created with two "Zeppelin" chiplets, two dummy die, and new package routes.

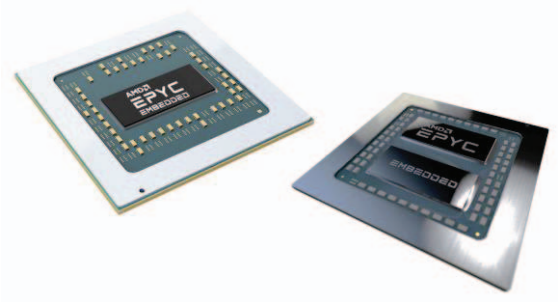


Figure 21. AMD EPYC™ 3000 Series embedded processors utilizing one or two chiplets.

available to the operating system (OS), so that when the OS runs a single high-performance thread it can schedule the program to the highest performing core. Each eight-core CCD consists of a pair of core complexes (CCX) each with an L3 cache and four associated cores. Like the core-level characterization, we also determine the best-performing CCXs, which allows the OS to schedule threads from a multi-threaded workload to the fastest CCX.

The core and CCX characterization processes occur every time a system is booted. This enables the characterization processes to adapt to the specific system that the processor is seated in; for example, systems may be equipped with different cooling solutions or board-level components such as voltage regulators themselves may vary across instances. An additional benefit of boot-time characterization is that as processors age, the preferred core may change over time and so we can continue to deliver the highest performance possible for each part by selecting the fastest core under the current circumstances.

VI. Expanding the Chiplet Approach

The chiplet-based architectures of the second-generation AMD EPYC™ processor and AMD Ryzen™ processor were highly effective for addressing the needs of our server and client offerings. However, chiplets also provided opportunities in a range of other target markets.

A. AMD Ryzen™ Threadripper™ Processors

AMD identified opportunities for extremely high-performance, high-core-count processors for high-end desktops (HEDT) and workstations. In particular, many professional content creation tasks such as computer-aided design and high-resolution video rendering (e.g., for Hollywood feature movies) require massive CPU resources to, for example, reduce movie rendering times for higher end-user productivity. Figure 20 shows how AMD leveraged the first-generation AMD EPYC™ processor by depopulating two of the four chiplets and replacing them with dummy silicon die to preserve the mechanical integrity of the overall package, combined with some additional customizations to the package, firmware, BIOS, etc., to create first-generation AMD Ryzen™ Threadripper™ processors with up to 16 cores. In subsequent generations, additional chiplets were enabled, and later we also utilized the chiplet designs from the second-generation AMD EPYC™ processor for 64-core HEDT offerings. This is another interesting example of how creative co-design of both silicon chiplets and packaging enabled AMD to quickly react to emerging market data and new opportunities.

B. AMD EPYC™ Embedded Processors

The high-performance embedded processor market has a range of applications with requirements for substantial I/O needs coupled with a range of desired core counts. Example applications include embedded networking, storage, medical imaging, and industrial control solutions. Figure 21 shows how our "Zeppelin" chiplet designs combined with packaging optimized for high-performance embedded use cases can produce additional products appropriate for this market. The first generation of AMD EPYC™ Embedded 3000 series processors include single- and dual-chiplet configurations with up to 16 cores in a package designed to be appropriate for the target embedded market form factors. Again, a single chiplet design can enable multiple product options. In a similar fashion, chiplet designs from the second-generation AMD EPYC™ server processors have been leveraged for the AMD EPYC™ 7000 series embedded processors.

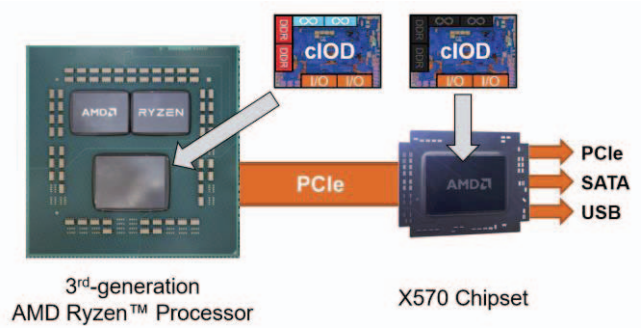


Figure 22. Reuse of the cIOD chiplet enables a fully-featured PCIe 4.0 chipset for AMD Ryzen™ processor motherboards.

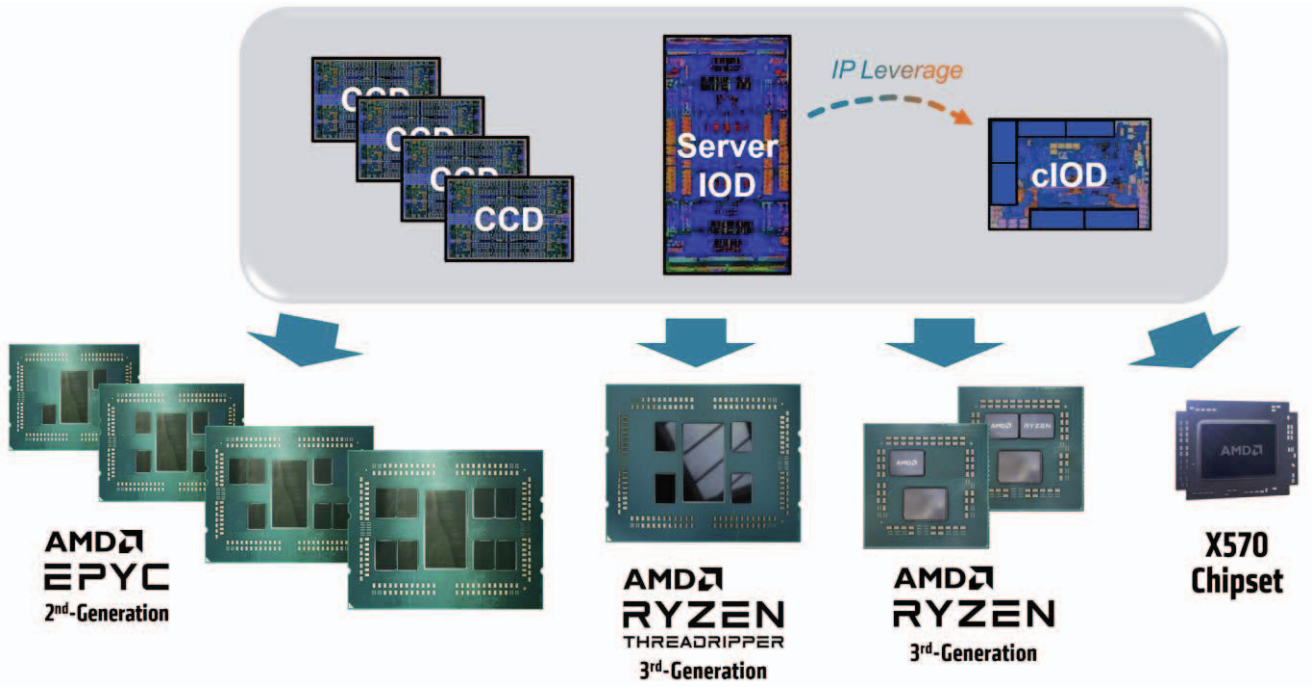


Figure 23. Mapping of AMD chiplets to multiple families of high-performance products.

C. The AMD X570 Chipset

Motherboards for high-performance processors such as those demanded by gamers, content creators, and other enthusiasts typically utilize chipsets to provide a full suite of I/O capabilities. Features often include additional PCIe lanes, USB ports, NVMe interfaces, and support for more SATA drives. To provide chipset support for the third-generation AMD Ryzen™ processor, we were able to directly leverage the cIOD without any CCDs in a standalone package. This is enabled by robust harvesting techniques for the cIOD chiplets, secure firmware, and other non-silicon support. Figure 22 illustrates the silicon reuse, which provides the platform with 16 PCIe gen4 I/O lanes, up to 12 SATA ports, and twelve USB ports all in addition to the PCIe, USB, and storage I/O from the AMD Ryzen™ processor itself.

VII. Conclusions

Chiplet-based design has transformed architecture at AMD. Figure 23 shows an illustration for how our chiplet-enabled approach enables a wide variety of products using a few individual chiplet designs. We have been able to creatively combine chiplets with a range of packaging solutions to produce a rich portfolio of products across a wide range of markets, create solutions that would be infeasible with monolithic designs (e.g., 64-core server), and do so while delivering great performance and value.

This paper aimed to demonstrate to the community how modular architectures can address a range of challenges in a post-Moore's Law world. At the same time, we wanted to provide a glimpse into how technology and architecture decisions not only consider traditional metrics like performance, power, and cost, but also how the decisions can help and support the construction of a diverse portfolio of products. Examples of engineering challenges such as the required silicon-package co-design for under-CCD routing are reminders that one does not simply take disparate pieces of silicon and "glue" them into a complete system. Significant thought, planning, collaboration, engineering, and creativity are needed to successfully bring all the pieces together.

In addition to the technical challenges, implementing such a widespread chiplet approach across so many market segments requires an incredible amount of partnership and trust across technology teams, business units, and our external partners. The product roadmaps across markets must be carefully coordinated and mutually scheduled to ensure that the right silicon is available at the right time for the launch of each product. Unexpected challenges and obstacles can arise, and world-class and highly passionate AMD engineering teams across the globe have risen to each occasion. The success of the AMD chiplet approach is as much a feat of engineering as it is a testament to the power of teams with diverse skills and expertise working together toward a shared set of goals and a common vision.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback and suggestions for this paper.

The authors also deeply thank the multiple worldwide AMD teams that have worked tirelessly to bring all of these products to life and whose hard work is represented in this paper.

REFERENCES

[1] Advanced Micro Devices, Inc., "Shared Resilience: AMD Response to COVID-19," https://www.amd.com/en/corporate/amd-covid-19-response, 2020.

[2] Advanced Micro Devices, Inc. "AMD Instinct™ MI100 Accelerator," https://www.amd.com/en/products/server-accelerators/instinct-mi100.

[3] Noah Beck, Sean White, Milam Paraschou, Samuel Naffziger, ""Zeppelin": An SoC for Multichip Architectures," IEEE International Solid-State Circuits Conference, February 2018.

[4] Bryan Black, Donald Nelson, Clair Webb, Nick Samra, "3D Processing Technology and its Impact on IA32 Microprocessors," International Conference on Computer Design, October 2004.

[5] Bryan Black, Murali Annavaram, Ned Brekelbaum, John DeVale, Lei Jiang, Gabriel H. Loh, Don McCauley, Patrick Morrow, Donald W. Nelson, Daniel Pantuso, Paul Reed, Jeff Rupley, Sadasivan Shankar, John Paul Shen, Clair Webb, "Die Stacking (3D) Microarchitecture," International Symposium on Microarchitecture, December 2006.

[6] Thomas Burd, Noah Beck, Sean White, Milam Paraschou, Nathan Kalyanasundharam, Gregg Donley, Alan Smith, Larry Hewitt, Samuel Naffziger, ""Zeppelin": An SoC for Multichip Architectures," IEEE Journal of Solid-state Circuits, Volume 54, No. 1, January 2019.

[7] Pat Conway, Nathan Kalyanasundharam, Gregg Donley, Kevin Lepak, Bill Hughes, "Blade Computing with the AMD Opteron™ Processor ("Magny-cours")," IEEE Hot Chips 21 Symposium, August 2009.

[8] Ronald G. Dreslinksi, David Fick, Bharan Giridhar, Gyouhou Kim, Sangwon Seo, Matthew Fojtik, Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim, Nurrachman Liu, Michael Wieckowski, Gregory Chen, Dennis Sylvester, David Blaauw, Trevor Mudge, "Centip3De: a Many-core Prototype Exploring 3D Integration and Near-threshold Computing," Communications of the ACM, Volume 56, No. 11, November 2013.

[9] Gary Dudeck, John Dudeck, "Design Considerations and Packaging of a Pentium® Pro Processor Based MultiChip Module for High Performance Workstation and Servers," IEEE Symposium on IC/Package Design Integration, February 1998.

[10] William Fedus, Barret Zoph, Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," arXiv 2101.03961, January 2021.

[11] Warren Flack, Robert Hsieh, Gareth Kenyon, Manish Ranjan, John Slabbekoorn, Andy Miller, Eric Beyne, Medhat Toukhy, PingHung Lu, Yi Cao, Chunwei Chen, "Large Area Interposer Lithography," IEEE 64th Electronic Components and Technology Conference, May 2014.

[12] David Greenhill, Ron Ho, David Lewis, Herman Schmit, Kok Hong Chan, Andy Tong, Sean Atsatt, Dana How, Peter McElheny, Keith Duwel, Jeffrey Schulz, Darren Faulkner, Gopal Iyer, George Chen, Hee Kong Phoon, Han Wooi Lim, Wei-Yee Koay, Ty Garibay, "A 14nm 1GHz FPGA with 2.5D Transceiver Integration," IEEE International Solid-State Circuits Conference, February 2017.

[13] Hewlett Packard Enterprise, "Cray to deliver record-setting Frontier supercomputer at ORNL," https://www.hpe.com/us/en/newsroom/press-release/2019/05/cray-to-deliver-record-setting-frontier-supercomputer-at-ornl.html, May 2019.

[14] Ajaykumar Kannan, Natalie Enright Jerger, Gabriel H. Loh, "Enabling Interposer-based Disintegration of Multi-core Processors," International Symposium on Microarchitecture, December 2015.

[15] Lawrence Livermore National Laboratory, "LLNL and HPE to partner with AMD on El Capitan, projected as world's fastest supercomputer," https://www.llnl.gov/news/llnl-and-hpe-partner-amd-el-capitan-projected-worlds-fastest-supercomputer, March 2020.

[16] Joe Macri, "AMD's Next Generation GPU and High Bandwidth Memory Architecture: FURY," IEEE Hot Chips 27 Symposium, August 2015.

[17] Michael Mantor, Ben Sander, "AMD's Radeon Next Generation GPU", IEEE Hot Chips 29 Symposium, August 2017.

[18] Gordon Moore, "Cramming More Components onto Integrated Circuits," Electronics, Volume 38, No. 8, April 1965.

[19] Shashidhar Mysore, Banit Agrawal, Navin Srivastava, Sheng-Chih Lin, Kaustav Banerjee, Timothy Sherwood, "Introspective 3D Chips," Symposium on Architectural Support for Programming Languages and Operating Systems, October 2006.

[20] Samuel Naffziger, Kevin Lepak, Milam Paraschou, Mahesh Subramony, "AMD Chiplet Architecture for High-performance Server and Desktop Products," IEEE International Solid-State Circuits Conference, February 2020.

[21] OpenAI, "AI and Compute," https://openai.com/blog/ai-and-compute/, May 2018.

[22] David Papworth, "Tuning the Pentium Pro Microarchitecture," IEEE Micro, Vol. 16, April 1996.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.

[24] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, Yuxiong He, "ZeRO: Memory Optimization Towards Training A Trillion Parameter Models," ArXiv 1910.02054, October 2019.

[25] Teja Singh, Alex Schaefer, Sundar Rangarajan, Deepesh John, Carson Henrion, Russell Schreiber, Miguel Rodriguez, Stephen Kosonocky, Samuel Naffziger, Amy Novak, "Zen: An Energy-efficient High-performance x86 Core," IEEE Journal of Solid-State Circuits, Volume 53, No. 1, January 2018.

[26] Balaram Sinharoy, Ronald N. Kalla, Joel M. Tendler, Richard J. Eickemeyer, Jody B. Joyner, "POWER5 System Microarchitecture," IBM Journal of Research and Development, Volume 49, No. 4/5, July/September 2005.

[27] Lisa T. Su, Samuel Naffziger, Mark Papermaster, "Multi-chip Technologies to Unleash Computing Performance Gains Over the Next Decade," IEEE International Electron Devices Meeting, December 2017.

[28] Lisa T. Su, "Delivering the Future of High-performance Computing," *Plenary Talk*, DARPA Electronics Resurgence Initiative Summit, July 2019.

[29] Top500 List, http://www.top500.org.

[30] Balaji Vaidyanathan, W-L. Hung, Feng Wang, Yuan Xie, N. Vijaykrishnan, M. J. Irwin, "Architecting Microprocessor Components in 3D Design Space," IEEE International Conference on VLSI Design, January 2007.

[31] Steven W. White, Sudhir Dhawan, "POWER2: Next Generation of the RISC System/6000 Family," IBM Journal of Research and Development, Volume 38, No. 5, September 1994.

ATTRIBUTION