Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Análise de Desenvolvimento de Sistemas

Marcos Vinícius Bertelli Bocamino

Multi-Chip Module em Processadores Modernos:

Uma análise das arquiteturas baseadas em chiplets

Americana, SP

2026

Marcos Vinícius Bertelli Bocamino

Multi-Chip Module em Processadores Modernos:

Uma análise das arquiteturas baseadas em chiplets

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas na área de concentração em Arquitetura e Organização de computadores.

Orientador: Prof. Me. Rossano Pablo Pinto

Este trabalho corresponde à versão final do Trabalho de Conclusão de Curso apresentado por Marcos Vinícius Bertelli Bocamino e orientado pelo Prof. Me. Rossano Pablo Pinto.

Americana, SP

2026

FICHA CATALOGRÁFICA – Biblioteca Fatec Americana Ministro Ralph Biasi- CEETEPS Dados Internacionais de Catalogação-na-fonte

BOCAMINO, Marcos Vinícius Bertelli

Multi-chip module em processadores modernos: uma análise das arquiteturas baseadas em chiplets. / Marcos Vinícius Bertelli Bocamino – Americana, 2026.

57f.

Monografia (Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - - Faculdade de Tecnologia de Americana Ministro Ralph Biasi – Centro Estadual de Educação Tecnológica Paula Souza

Orientador: Prof. Ms. Rossano Pablo Pinto

1. Computadores 2. Hardware 3. Processadores. I. BOCAMINO, Marcos Vinícius Bertelli II. PINTO, Rossano Pablo III. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana Ministro Ralph Biasi

CDU: 681.31
681.31
681.31

Elaborada pelo autor por meio de sistema automático gerador de ficha catalográfica da Fatec de Americana Ministro Ralph Biasi.

Marcos Vinícius Bertelli Bocamino

Multi-Chip Module em Processadores Modernos: Uma Análise das Arquiteturas Baseadas em Chiplets

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas pelo Centro Paula Souza – FATEC Faculdade de Tecnologia de Americana Ministro Ralph Biasi.
Área de concentração: Análise e Desenvolvimento de Sistemas.

Americana, 11 de junho de 2026.

Banca Examinadora:

Rossano Pablo Pinto
Mestre
Fatec Americana "Ministro Ralph Biasi"

José Luiz Zem
Doutor
Fatec Americana "Ministro Ralph Biasi"

Ana Lúcia Spigolon
Especialista
Fatec Americana "Ministro Ralph Biasi"

AGRADECIMENTOS

Eu gostaria de agradecer a todos os professores, educadores e a minha mãe, que me ajudaram a chegar até aqui e me serviram de inspiração e motivação para continuar aprendendo e prosseguindo.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

O presente trabalho contou com o auxílio de ferramentas de inteligência artificial generativa, especificamente o modelo ChatGPT (OpenAI, GPT 5.5 Thinking, 2026) e Claude (Anthropic, Sonnet 4.6, 2026), utilizado nas seguintes atividades:

1. Apoio à estruturação e organização dos capítulos;
2. Revisão gramatical e melhoria de clareza textual;
3. Auxílio na formulação de frases e parágrafos a partir do conteúdo técnico elaborado pelo autor;
4. Formatação do documento em relação às normas.

Ressalta-se que a pesquisa, a seleção das fontes, a análise do conteúdo técnico, a organização final das ideias e a responsabilidade pela versão final do trabalho são de autoria e responsabilidade do autor.

Marcos Vinícius Bertelli Bocamino

Americana, SP

2026

RESUMO

O avanço do desempenho computacional esteve, durante décadas, associado à miniaturização dos transistores e à continuidade da Lei de Moore. Entretanto, a crescente complexidade dos processos de fabricação, a redução da taxa de rendimento (*yield rate*) em *chips* de grande área e o aumento dos custos associados aos nós avançados tornaram o modelo monolítico progressivamente mais limitado para atender às demandas atuais de processamento. Nesse contexto, este trabalho tem como objetivo analisar a tecnologia *Multi-Chip Module* (MCM) como solução arquitetural às limitações físicas, produtivas e econômicas dos *chips* monolíticos, compreendendo suas motivações produtivas, suas principais arquiteturas e suas aplicações em cenários de alta demanda computacional. A pesquisa utiliza método bibliográfico, de natureza descritiva e exploratória, baseando-se na análise de artigos científicos, documentação técnica, *white papers* de fabricantes e publicações especializadas da indústria de semicondutores. O estudo apresenta inicialmente os fundamentos da fabricação de semicondutores, o impacto do *yield rate* e o conceito de integração *multi-die* e *chiplets*. Em seguida, analisa arquiteturas comerciais baseadas em MCM, incluindo processadores AMD Epyc, Intel Xeon Scalable, Intel Ponte Vecchio, Apple M1 Ultra e aceleradores AMD Instinct. Também são discutidas tecnologias de encapsulamento avançado, como integração 2.5D, 3D, HBM, EMIB, CoWoS e Foveros. Os resultados indicam que o MCM deixou de ser apenas uma alternativa produtiva ao *chip* monolítico e passou a atuar como estratégia arquitetural para ampliar escalabilidade, especialização funcional, largura de banda de memória e viabilidade produtiva. As aplicações analisadas em supercomputadores, inteligência artificial e simulações científicas demonstram que arquiteturas modulares já compõem a base de sistemas de alto desempenho, como El Capitan, Frontier e Aurora. Conclui-se que MCM, *chiplets* e encapsulamento avançado tendem a permanecer como elementos fundamentais na evolução de processadores, aceleradores e plataformas computacionais modernas.

Palavras-chave: Computadores; *Hardware*; Processadores.

ABSTRACT

The advancement of computational performance was, for decades, associated with transistor miniaturization and the continuity of Moore's Law. However, the increasing complexity of manufacturing processes, the reduction of yield rate in large-area chips, and the rising costs associated with advanced nodes have progressively limited the monolithic model in meeting current processing demands. In this context, this work aims to analyze Multi-Chip Module (MCM) technology as an architectural solution to the physical, productive, and economic limitations of monolithic chips, considering its manufacturing motivations, main architectures, and applications in high computational demand scenarios. The research uses a bibliographic method, with a descriptive and exploratory nature, based on the analysis of scientific papers, technical documentation, manufacturer white papers, and specialized publications from the semiconductor industry. The study initially presents the fundamentals of semiconductor manufacturing, the impact of yield rate, and the concept of multi-die and chiplet integration. It then analyzes commercial architectures based on MCM, including AMD EPYC processors, Intel Xeon Scalable, Intel Ponte Vecchio, Apple M1 Ultra, and AMD Instinct accelerators. Advanced packaging technologies are also discussed, such as 2.5D and 3D integration, HBM, EMIB, CoWoS, and Foveros. The results indicate that MCM is no longer only a productive alternative to monolithic chips, but has become an architectural strategy to improve scalability, functional specialization, memory bandwidth, and manufacturing feasibility. The applications analyzed in supercomputers, artificial intelligence, and scientific simulations demonstrate that modular architecture already forms the basis of high-performance systems, such as El Capitan, Frontier, and Aurora. It is concluded that MCM, chiplets, and advanced packaging tend to remain fundamental elements in the evolution of modern processors, accelerators, and computing platforms.

Keywords: *Computers; Hardware; Processors*.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MCM	*Multi-Chip Module*
HPC	*High Performance Computing*
IA	Inteligência Artificial
LLM	*Large Language Model*
CPU	*Central Processing Unit*
GPU	*Graphics Processing Unit*
SoC	*System-on-Chip*
UMA	*Uniform Memory Access*
NUMA	*Non-Uniform Memory Access*
CVD	*Chemical Vapor Deposition*
LPCVD	*Low Pressure Chemical Vapor Deposition*
PECVD	*Plasma-Enhanced Chemical Vapor Deposition*
DUV	*Deep Ultraviolet* (Ultravioleta Profundo)
EUV	*Extreme Ultraviolet* (Ultravioleta Extremo)
CCX	*Core Complex*
CCD	*Core Complex Die*
PCIe	*Peripheral Component Interconnect Express*
CXL	*Compute Express Link*
DRAM	*Dynamic random-access memory*
HBM	*High Bandwidth Memory*
HBM2E	*High Bandwidth Memory 2 Extended*
HBM3	*High Bandwidth Memory 3*
TSV	*Through-Silicon Via*
I/O	Input/Output
BIOS	*Basic Input/Output System*
RDL	*Redistribution Layer*
CoWoS	*Chip on Wafer on Substrate*
AMX	*Advanced Matrix Extensions*
DSA	*Data Streaming Accelerator*
IAA	*In-Memory Analytics Accelerator*
QAT	*QuickAssist Technology*
DLB	*Dynamic Load Balancer*

EMIB	*Embedded Multi-die Interconnect Bridge*
FP32	*Floating Point 32-bit*
TFLOPS	*Tera Floating Point Operations per Second*
GCD	*Graphics Compute Die*
XCD	*Accelerator Complex Die*
OAM	*Open Accelerator Modules*
APU	*Accelerated Processing Unit*
ROCm	*Radeon Open Compute*
FinFET	*Field-Effect Transistor*

SUMÁRIO

1 INTRODUÇÃO

Durante décadas, o avanço do desempenho computacional foi sustentado pela Lei de Moore e pelo escalonamento contínuo de transistores em chips monolíticos. No entanto, à medida que os processos de litografia se aproximam dos limites físicos do silício, o ritmo de ganhos por geração desacelerou significativamente, enquanto os custos de desenvolvimento e fabricação em processos (ou “nós”) avançados cresceram de forma significativa (WARD et al., 2020). Esse cenário é agravado pela demanda contínua e crescente por processamento, impulsionada por datacenters, supercomputadores, computação em nuvem, inteligência artificial e HPC (*High Performance Computing*), setores estes que exigem saltos de desempenho que o escalonamento monolítico tradicional já não é capaz de entregar com viabilidade econômica.

Nesse cenário, surgem propostas alternativas de continuidade da evolução dos semicondutores, como a chamada Lei de Expansão Tau (*Tau Scaling Law*), apresentada pela Huawei em 2026, que propõe deslocar o foco da miniaturização geométrica para a redução do atraso de propagação dos sinais. Embora ainda recente e sem validação industrial ampla, essa proposta reforça a percepção de que a evolução dos processadores passa a depender cada vez mais de estratégias arquiteturais, de integração e de encapsulamento, além da simples redução dimensional dos transistores (HUAWEI, 2026; BAPTISTA; PAN, 2026).

Diante dessas limitações, a indústria de semicondutores buscou alternativas arquiteturais capazes de contornar as restrições físicas e econômicas do modelo monolítico. O método *Multi-Chip Module* (MCM), baseado na integração de múltiplos módulos de silício, denominados *dies*, em um único encapsulamento, consolidou-se como uma das respostas mais relevantes a esse desafio, ganhando destaque comercial como uma proposta de solução a partir de 2017 com o lançamento da família de processadores EPYC de primeira geração da AMD (AMD, 2017), e expandindo-se progressivamente até o estado atual de presença crescente em segmentos estratégicos da indústria.

O tema foi escolhido devido à sua relevância crescente no cenário de desenvolvimento de hardware de alto desempenho e pela escassez de materiais acadêmicos consolidados em língua portuguesa que abordem o MCM sob uma perspectiva técnica e aplicada. A análise desta tecnologia contribui para a

compreensão dos fundamentos que orientam o projeto de processadores modernos utilizados em aplicações críticas de IA e HPC.

O objetivo geral deste trabalho é analisar a tecnologia *Multi-Chip Module* como solução arquitetural às limitações físicas e econômicas dos chips monolíticos, compreendendo suas motivações produtivas, suas principais arquiteturas e suas aplicações em contextos de alta demanda computacional.

Como objetivos específicos, o trabalho se propõe a: Descrever o processo de fabricação de semicondutores e suas limitações físicas e econômicas no contexto de chips monolíticos, com ênfase no conceito de taxa de aproveitamento, mais conhecido como *yield rate*; Identificar como a arquitetura MCM surge como resposta às restrições impostas pelo *yield rate* e pelo custo crescente de processos avançados de fabricação; Apresentar e comparar arquiteturas baseadas em MCM e *chiplets* adotadas pela indústria, analisando soluções comerciais de diferentes segmentos, como servidores, aceleradores de alto desempenho e SoCs (*System-on-a-Chip*); Analisar como o MCM é aplicado em contextos de alta demanda computacional, especialmente em HPC e inteligência artificial.

O método científico de pesquisa utilizado foi o dedutivo, de natureza descritiva e exploratória, baseando-se na análise de artigos científicos, documentação técnica, *white papers* oficiais de fabricantes e publicações especializadas da indústria de semicondutores. A abordagem exploratória justifica-se pela natureza dinâmica do tema, cujos desenvolvimentos mais relevantes são recentes e distribuídos em fontes técnicas diversas.

Para auxílio na estruturação e revisão textual deste trabalho, foram utilizados os modelos de linguagem ChatGPT (OpenAI, GPT 5.5 Thinking, 2026) e Claude (Anthropic, Sonnet 4.6, 2026). Os modelos foram utilizados como ferramenta de apoio à escrita e organização das ideias, com base em pesquisa e conteúdo técnico desenvolvidos pelo autor. As referências bibliográficas, os dados técnicos e as análises apresentadas são de responsabilidade do autor.

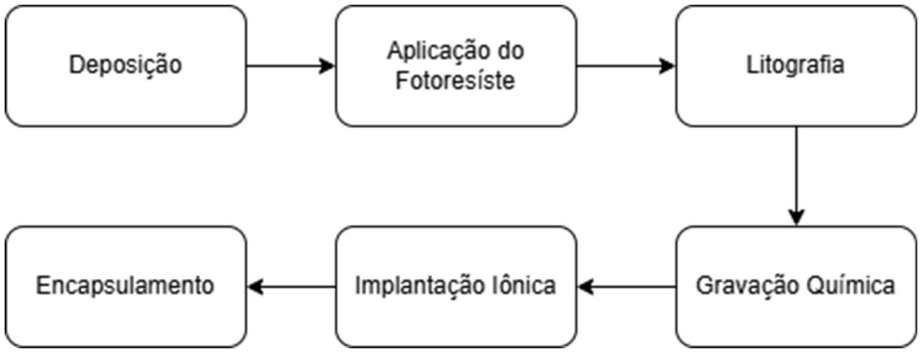
O trabalho foi estruturado em cinco capítulos. O primeiro capítulo apresenta a introdução, contextualizando o problema, os objetivos e a metodologia adotada. O segundo capítulo aborda o processo de fabricação de semicondutores, discutindo as limitações físicas e econômicas dos chips monolíticos e o impacto do *yield rate* na

viabilidade produtiva. O terceiro capítulo apresenta as arquiteturas baseadas em MCM, comparando diferentes implementações adotadas pela indústria. O quarto capítulo analisa as aplicações da tecnologia em cenários de HPC e inteligência artificial, examinando casos concretos de uso. Com base nas informações levantadas ao longo dos capítulos anteriores, o quinto capítulo reserva-se às considerações finais do estudo.

2 FABRICAÇÃO DE CHIPS

A fabricação de processadores modernos consiste em múltiplas etapas de alta precisão, executadas sobre *wafers*, discos ultrafinos de silício cortados a partir de um lingote, em ambientes altamente controlados, conhecidos como *clean rooms* (salas limpas). O processo combina técnicas de deposição, litografia, gravação química, implantação iônica e encapsulamento, como exemplificado na Figura 1, para construir chips com bilhões de transistores.

Figura 1 – Fluxo do Processo de Fabricação de Processadores.



Fonte: Elaborado pelo autor com base em Li e Timings (2023).

Conforme as estruturas internas dos microchips diminuem de tamanho, essas etapas passam a demandar maior precisão e controle, em decorrência do aumento da densidade de transistores e da redução das dimensões físicas das estruturas do chip. Considerando isso, as seções a seguir detalham os processos de fabricação, limitações físicas e econômicas que emergem com a miniaturização, juntamente com abordagens de integração que vieram na tentativa de contornar as limitações (LI; TIMINGS, 2023).

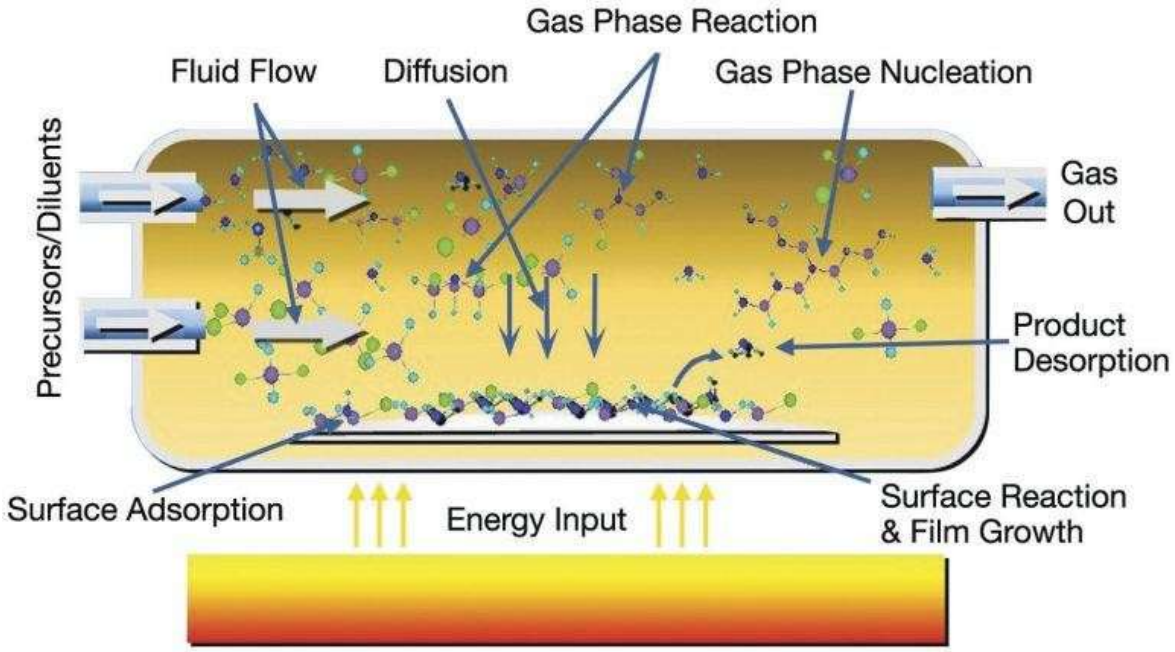
2.1 O processo de fabricação

Uma das primeiras etapas é a deposição, que ocorre sobre um *wafer* de silício monocristalino (KALPAKJIAN; SCHMID, 2014) produzido com níveis muito elevados de pureza e polido até atingir uma superfície extremamente lisa. Nesta etapa, camadas finas de materiais semicondutores, isolantes e condutores são depositadas sobre o *wafer* conforme o tipo de estrutura a ser formada (LI; TIMINGS, 2023).

A deposição pode ser realizada por diferentes técnicas, variando conforme as condições de pressão, temperatura e uso de sistemas a vácuo. Entre esses métodos estão a deposição a vácuo, o *sputtering* e a deposição química em fase vapor (*Chemical Vapor Deposition*, CVD), representada na Figura 2. Nesse método, o filme

é formado por meio da reação ou decomposição de compostos gasosos. Variações como LPCVD e PECVD permitem alterar condições de pressão, uniformidade e temperatura do processo (KALPAKJIAN; SCHMID, 2014).

Figura 2 – Ilustração do processo de deposição química em fase vapor (CVD).

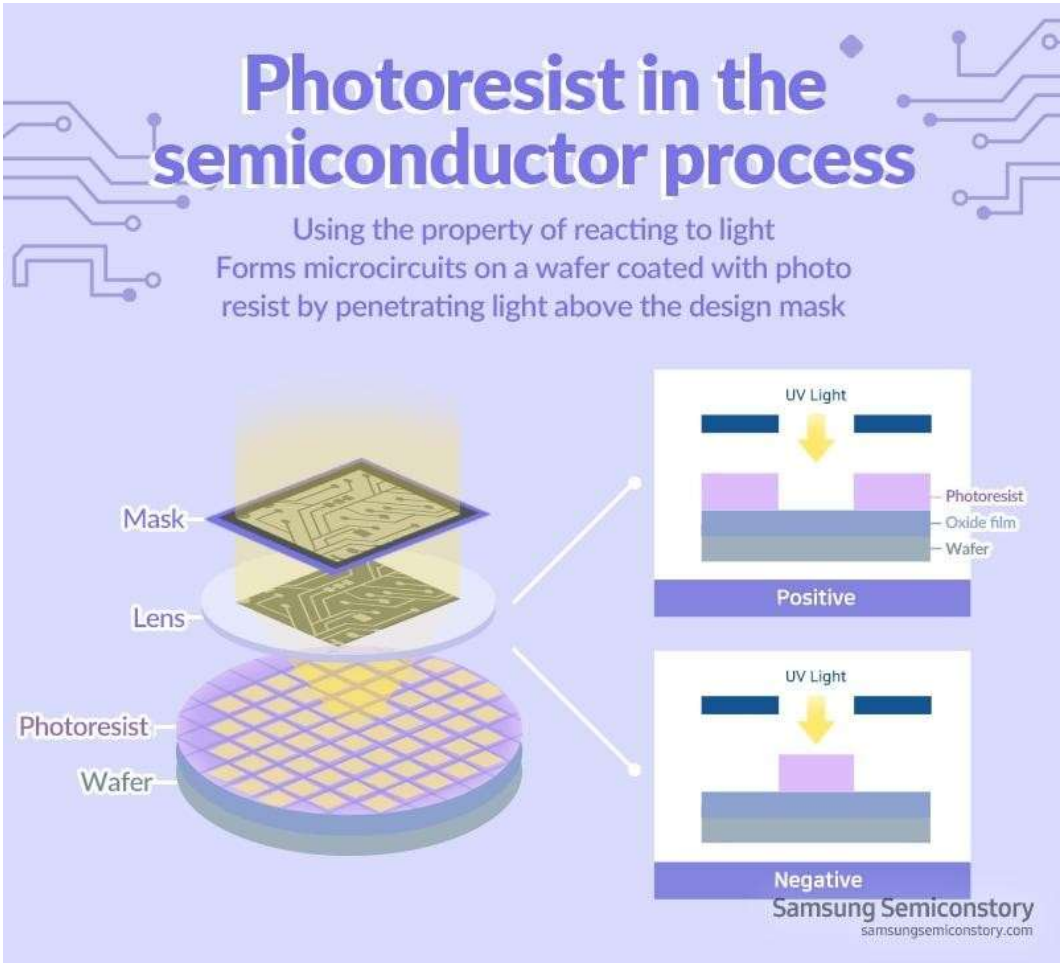


Fonte: MKS Instruments ([s.d.]).

Na sequência, o *wafer* recebe uma camada fotossensível denominada fotorresiste (*photoresist coating*), com a finalidade de viabilizar a transferência dos padrões do circuito durante a litografia. Existem fotorresistes positivos e negativos, que diferem quanto à forma como reagem à luz ultravioleta. No fotorresiste positivo, as áreas expostas à luz ultravioleta alteram sua estrutura e tornam-se mais solúveis. Por outro lado, no fotorresiste negativo, as áreas expostas se polimerizam, tornando-se mais resistentes e difíceis de dissolver. O fotorresiste positivo é mais comum na produção de semicondutores, pois sua maior capacidade de resolução o torna uma escolha mais adequada para a etapa de litografia (LI; TIMINGS, 2023).

Na Figura 3, a Samsung Semiconductor ilustra o uso dos fotorresistes positivo e negativo e suas respectivas reações à luz ultravioleta, incluindo elementos associados à etapa de litografia, como lente e máscara.

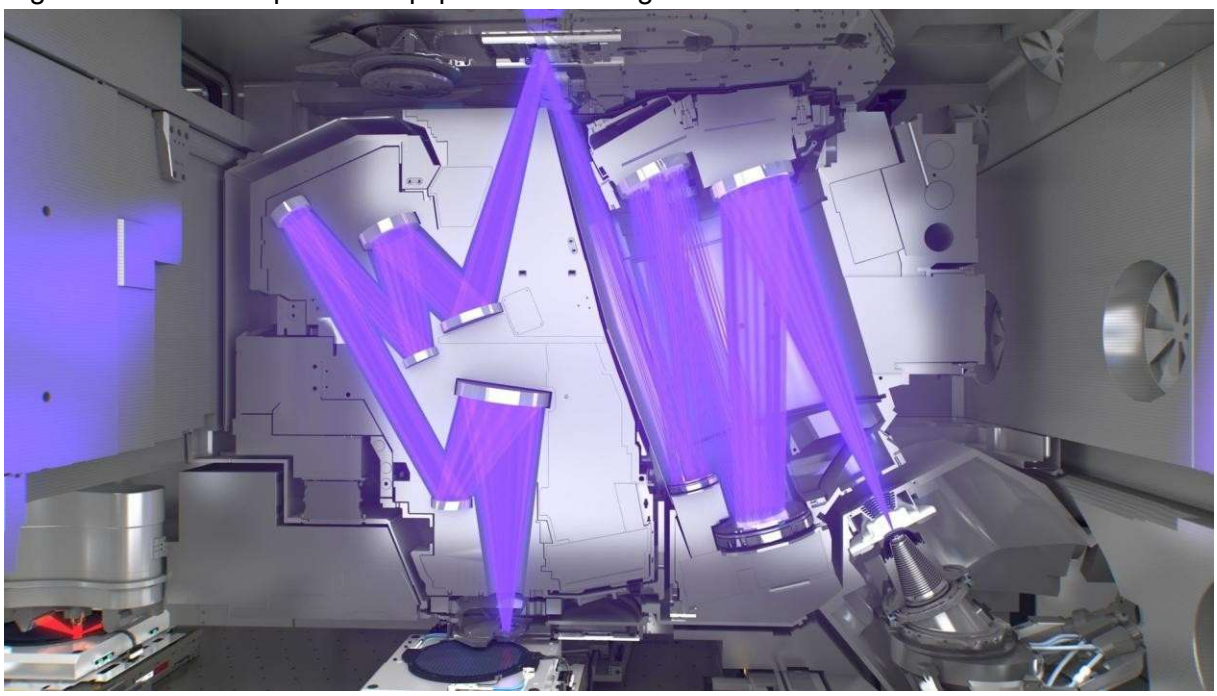
Figura 3 – Aplicação do fotorresiste no processo de fotolitografia.



Fonte: Samsung Semiconductor ([s.d.]a).

A etapa seguinte é a litografia, processo responsável por transferir os padrões do circuito para a superfície do *wafer* através de uma máscara. Para isso, a luz é projetada sobre o *wafer* por meio de um retículo (*reticle*), que contém uma máscara com o padrão do circuito a ser transferido. O sistema óptico, composto por lentes nos sistemas de ultravioleta profundo (DUV) ou por espelhos nos sistemas de ultravioleta extremo (EUV), reduz e foca o padrão sobre a camada de fotorresiste. No caso da litografia EUV, a maioria dos materiais absorve a luz ultravioleta extrema, por isso, os espelhos tornam-se necessários ao processo. Esses espelhos possuem mais de 100 camadas de materiais específicos, selecionados para maximizar a reflexão da luz EUV. A organização desses espelhos dentro de uma máquina de litografia é ilustrada na Figura 4 (ASML, [s.d.]; LI; TIMINGS, 2023; IMEC, [s.d.]b).

Figura 4 – Sistema óptico de equipamento de litografia EUV.



Fonte: ASML ([s.d.]).

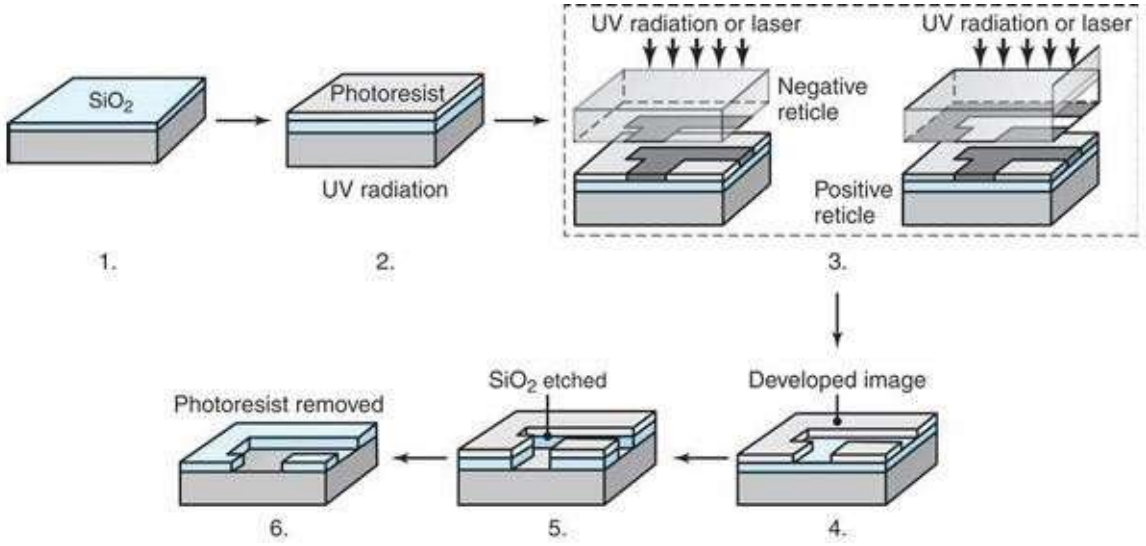
A luz altera quimicamente o fotorresiste nas áreas expostas e, após a revelação, o padrão passa a estar definido nessa camada, orientando os processos posteriores de gravação química e formação das estruturas microscópicas do circuito integrado. Essa técnica utilizada na produção de semicondutores é denominada fotolitografia (IMEC, [s.d.]b). Em processos avançados de fabricação, utiliza-se a litografia por ultravioleta extremo, o EUV (*Extreme Ultraviolet*), que emprega luz com comprimento de onda de 13,5 nanômetros, significativamente menor que o ultravioleta profundo utilizado anteriormente, permitindo a criação de estruturas em escala nanométrica com maior precisão e densidade (IMEC, [s.d.]b).

Após a litografia, inicia-se a etapa de gravação química, denominada *etch*. Nesse processo, o *wafer* é submetido a tratamento térmico e revelação química, removendo seletivamente partes do fotorresiste e expondo um padrão tridimensional de canais abertos sobre a superfície. Existem dois tipos de gravação: o método úmido (*wet etching*), que utiliza banhos químicos, dissolvendo as áreas expostas, e o método seco (*dry etching*), que emprega gases para definir o padrão sobre o *wafer* (LI; TIMINGS, 2023). Os processos de gravação devem atuar com precisão crescente sobre as estruturas condutoras sem que comprometam a integridade estrutural do chip. Em projetos de chips modernos, como memórias 3D NAND, o controle preciso

da profundidade de gravação é crítico e cada vez mais desafiador, já que essas memórias podem chegar a até 175 camadas sobrepostas (LI; TIMINGS, 2023).

A Figura 5 ilustra como a superfície de um *wafer* é manipulada até o processo de gravação química e como ele altera a superfície do *wafer*.

Figura 5 – Processo de transferência de padrão de circuito para *wafer*.

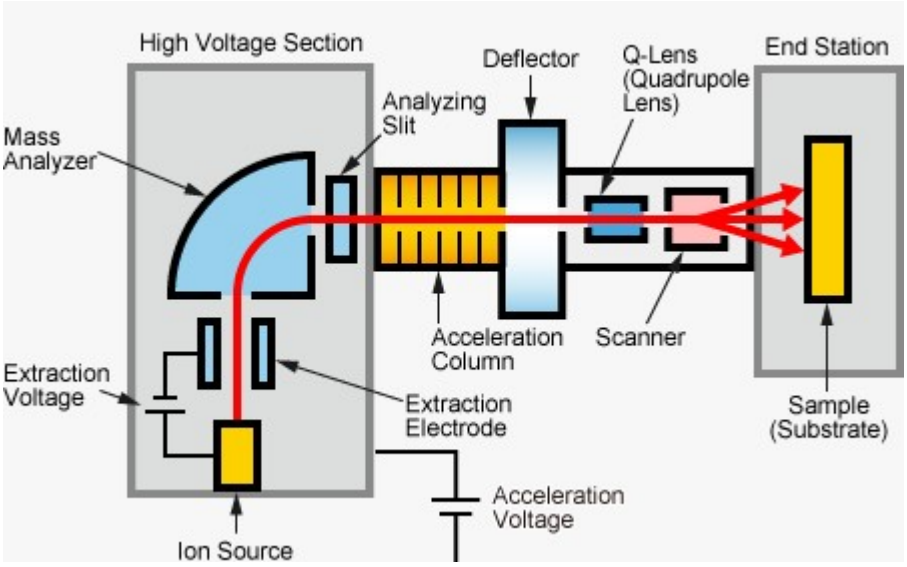


Fonte: Kalpakjian e Schmid (2014, p. 812, fig. 28.13).

Assim que concluída a gravação química, o *wafer* pode ser submetido à implantação iônica (*ion implantation*), etapa em que sua superfície é bombardeada com íons positivos ou negativos com o objetivo de ajustar as propriedades elétricas do silício. Por natureza, o silício puro não é um condutor nem um isolante perfeito, pois suas propriedades elétricas ficam entre os dois extremos. A introdução controlada de íons eletricamente carregados no cristal de silício permite modular o fluxo de eletricidade e, assim, formar os transistores, que são os componentes eletrônicos fundamentais dos microchips (LI; TIMINGS, 2023). Concluída a implantação, as seções restantes do fotorresiste que protegiam as áreas que não deveriam ser modificadas são removidas, encerrando o ciclo de uma camada de circuito (LI; TIMINGS, 2023).

Conforme ilustrado na Figura 6, o processo de implantação iônica envolve a geração, seleção, aceleração e direcionamento de íons até o substrato semicondutor (MATSUSADA PRECISION, 2026).

Figura 6 – Esquemática de um equipamento de implantação iônica.

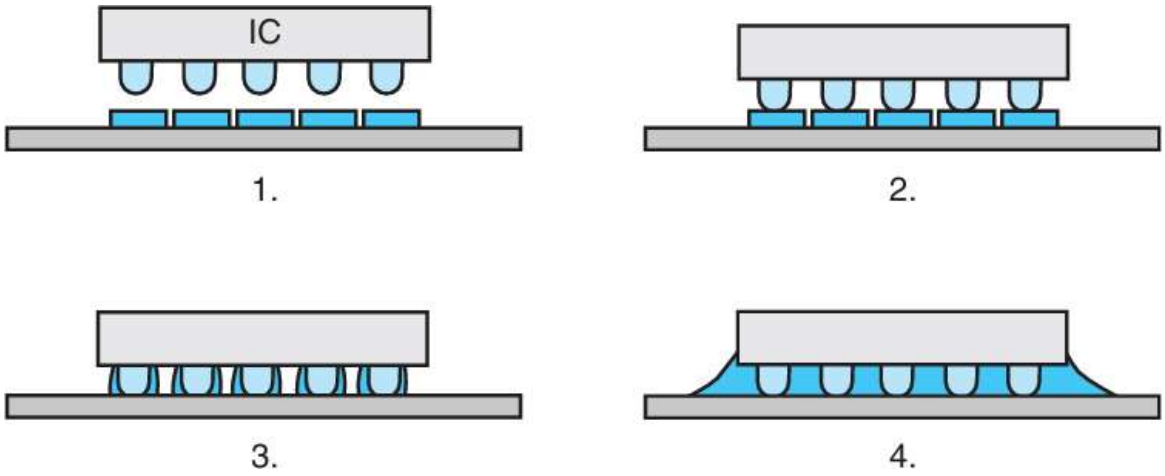


Fonte: Matsusada Precision (2026).

Por fim, a etapa final é o encapsulamento (*packaging*). Para extrair os chips do *wafer*, este é cortado com uma serra de diamante, resultando em unidades individuais denominadas *dies*. Um *wafer* de 300 mm, o formato mais utilizado na indústria, pode conter desde alguns poucos até milhares de *dies*, a depender do tamanho do chip produzido. Cada *die* é então fixado sobre um substrato (*substrate*), uma base que utiliza folhas metálicas para direcionar os sinais de entrada e saída do chip para os demais componentes do sistema. Por fim, um dissipador de calor (*heat spreader*) é posicionado sobre o conjunto, atuando como um componente metálico de proteção e dissipação térmica, geralmente de alumínio com componentes de cobre, que auxilia na manutenção de temperatura adequada do microchip (LI; TIMINGS, 2023).

Uma das tecnologias utilizadas nessa etapa é o *flip-chip*, em que o *die* é posicionado de forma invertida e conectado ao substrato por meio de esferas de solda. Essa abordagem permite estabelecer conexões elétricas diretamente entre a superfície ativa do *die* e o substrato, reduzindo a necessidade de fios de ligação e favorecendo maior densidade de interconexões. A Figura 7 apresenta um exemplo desse tipo de encapsulamento, ilustrando uma das possibilidades de integração física e elétrica entre o chip e o substrato (KALPAKJIAN; SCHMID, 2014, p. 834).

Figura 7 – Exemplo de encapsulamento por tecnologia *flip-chip*.

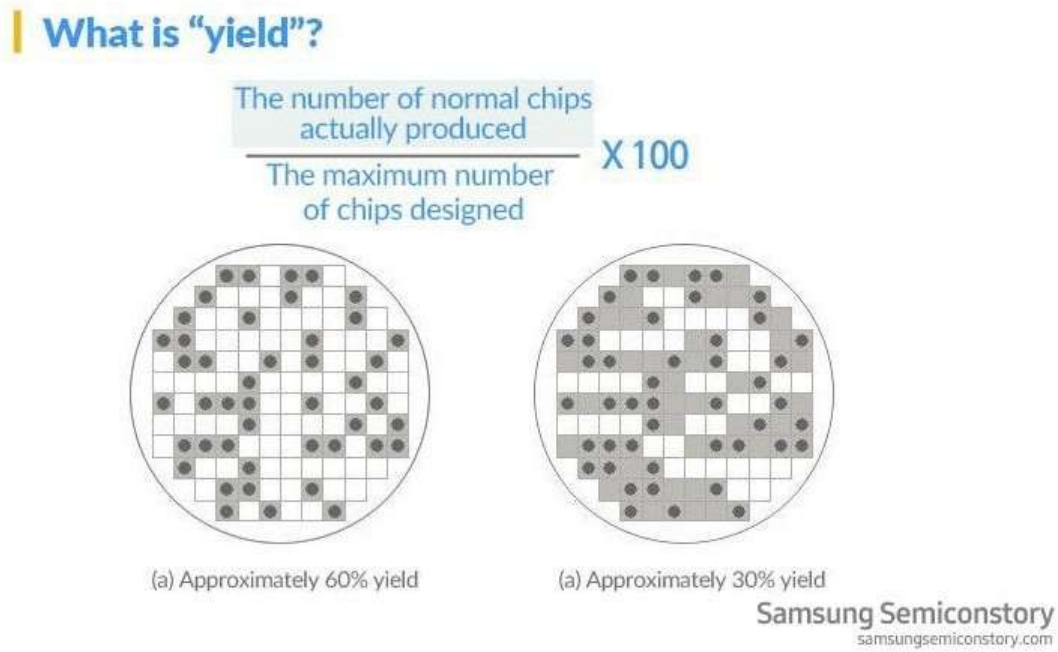


Fonte: Kalpakjian e Schmid (2014, p. 834, fig. 28.33).

2.2 Taxa de aproveitamento – *Yield Rate*

A taxa de aproveitamento, denominada *yield rate*, corresponde à proporção de chips funcionais obtidos em relação ao total de unidades que se pode produzir em um *wafer*. De maneira prática, representa a eficiência do processo de fabricação: quanto maior o *yield rate*, menor o desperdício e o custo por unidade produzida (KHOPE; KSHIRSAGAR, 2025). A Figura 8 apresenta uma representação visual desse conceito, demonstrando como defeitos distribuídos pelo *wafer* reduzem a quantidade de *dies* funcionais.

Figura 8 – Representação do *yield rate*.



Fonte: SAMSUNG SEMICONDUCTOR ([s.d.]b).

Embora a Figura 8 apresente valores ilustrativos de aproveitamento, o *yield rate* não possui um percentual fixo de aceitabilidade. Segundo Kalpakjian e Schmid (2014, p. 835), essa taxa pode variar de poucos pontos percentuais em processos novos até mais de 90% em linhas de fabricação maduras, pois depende de fatores como o processamento do *wafer*, ligação, encapsulamento e testes.

As perdas de aproveitamento na fabricação de *wafers* ocorrem sob duas maneiras principais: perdas de linha (*line yield*) e perdas de *die* (*die yield*). As primeiras resultam de danos físicos ao *wafer* ou de erros no processo de fabricação. As segundas decorrem predominantemente de defeitos microscópicos distribuídos pela superfície do *wafer*, como partículas, curtos-circuitos ou falhas em camadas isolantes, que comprometem o funcionamento individual de cada *die* (LEACHMAN, 2014).

Segundo Malloy (2025), um *die* de silício corresponde a uma pequena peça quadrada ou retangular de material semicondutor, que serve como base física para a fabricação de circuitos integrados.

A área do *die* exerce influência direta sobre o *yield rate*: quanto maior o *die*, maior a probabilidade de que ao menos um defeito esteja presente em sua superfície, reduzindo a taxa de aproveitamento. Esse fenômeno torna a produção de chips de grande área progressivamente menos viável do ponto de vista econômico. Além do

impacto dos defeitos, a área máxima de exposição do *die* é fisicamente limitada pelo tamanho do retículo utilizado nos sistemas de litografia, que na indústria atual restringe a área máxima de um *die* monolítico a 858 mm² (HAN et al., 2024). Perante este limite, técnicas alternativas para a criação de chips avançados ressurgiram, como o *Multi-Chip Module*.

2.3 *Multi-Die & Chiplet*

Anteriormente, com as arquiteturas monolíticas tradicionais, um único *die* de silício concentrava toda a funcionalidade de um chip. Nos dias de hoje, com as soluções *multi-die*, os fabricantes podem distribuir funcionalidades complexas entre vários *dies*, cada um fabricado com o processo mais adequado para sua função específica (MALLOY, 2025).

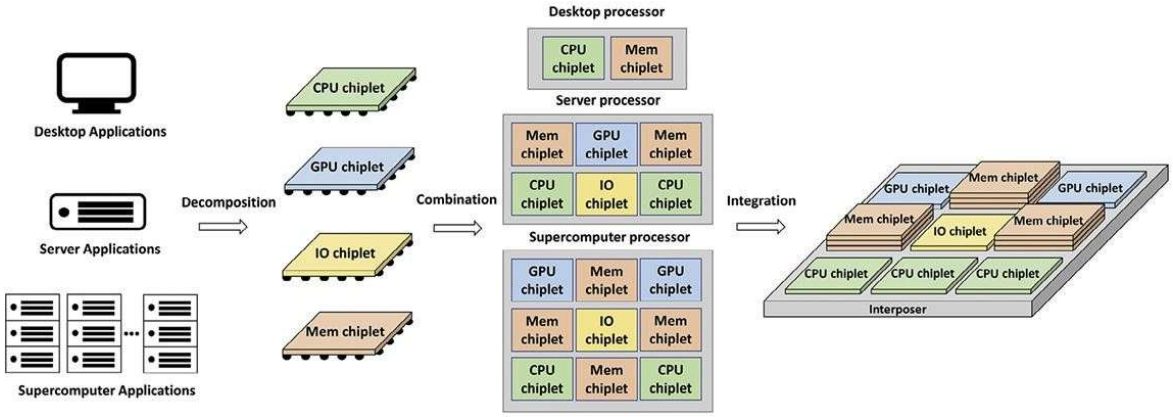
A ideia por trás da integração é poder alocar tudo em um único *die*, que, em tese, resulta em um desempenho maior do que a separação em múltiplos *dies*. Entretanto, atualmente, alguns componentes que estariam no chip não se beneficiam tanto de novos processos de fabricação, o que significa que produzir um chip completamente integrado resulta em áreas que poderiam ser implementadas com um processo mais barato, uma vez que a adoção integral de um processo mais avançado eleva os custos sem trazer benefícios além da integração em um único *die* (MOYER, 2026, p. 6-7). Na produção de chips, muitos chips derivam de um mesmo projeto de *die*, sendo comum que os modelos sejam organizados em séries, com versões com diferentes configurações de recursos e funcionalidades. A criação de um *die* diferente para cada modelo significaria criar um conjunto de máscaras para cada versão. Processos avançados de fabricação são caros por si só, e os custos das máscaras vêm crescendo progressivamente. Por exemplo, para o nó de 5 nm, um conjunto pode custar entre US$ 10 milhões e US$ 20 milhões (MOYER, 2026, p. 7).

Devido a esse custo, os engenheiros de arquitetura precisam mapear como criar vários modelos de uma série a partir de um único projeto de *die*. Por exemplo, com um *die* de oito núcleos, usá-lo também para vender unidades de quatro e dois núcleos. Isso é possível desabilitando núcleos que não devem estar presentes de acordo com a versão. Dessa forma, um chip de dois núcleos será na verdade uma unidade de oito núcleos com seis deles desabilitados, o que gera um desperdício significativo de silício. Para que essa abordagem funcione, é necessário vender

modelos mais modestos com lucro, mesmo que eles venham de um *die* com capacidade superior à comercializada (MOYER, 2026, p. 8).

De acordo com Moyer (2026), a indústria está lidando com esses desafios com alternativas ao método monolítico tradicional, reduzindo a tendência de integração, separando o *die* em múltiplos *dies* menores e montando em um único encapsulamento. O resultado ainda é o mesmo circuito, porém cada *die* será menor, melhorando a taxa de aproveitamento, reduzindo o custo por *die*. Cada *die* pode ser fabricado no processo mais econômico que atenda ao desempenho necessário, reduzindo o custo ao reservar processos mais avançados para módulos que realmente precisam ou que tiram maior vantagem disso.

Figura 9 – Esquema de reuso de IP por *chiplets*.



Fonte: HAN et al. (2024, p. 1435).

A lógica de reutilização de blocos funcionais denominados *chiplets* pode ser observada na Figura 9, com exemplificação de diferentes aplicações. *Chiplets* podem ser integrados por meio do método Multi-Chip Module, que é definido como uma unidade de encapsulamento única que possui dois ou mais chips e um substrato de interconexão, funcionando em conjunto, formando um sistema maior. O MCM também é responsável pelo gerenciamento de entradas e saídas, pelo controle térmico, pelo suporte mecânico e pela proteção dos componentes internos contra o ambiente externo (MADPCB, [s.d.]). O conceito foi desenvolvido no início da década de 1980, representando uma evolução do encapsulamento híbrido tradicional no contexto dos circuitos integrados (MADPCB, [s.d.]).

2.4 Arquiteturas baseadas em chiplets

A adoção de *chiplets* não determina uma única forma de organização arquitetural. Embora o conceito esteja associado à divisão de um sistema em múltiplos módulos menores, a forma como esses módulos são conectados e organizados pode variar de acordo com o padrão de acesso à memória, a função de cada *chiplet*, a interconexão utilizada e o nível de escalabilidade desejado. Nesse contexto, Han et al. (2024, p. 1437-1439) classificam arquiteturas baseadas em *chiplets* em diferentes categorias, sendo elas *symmetric-chiplet, NUMA-chiplet, cluster-chiplet, heterogeneous-chiplet e hierarchical-chiplet.*

A arquitetura *symmetric-chiplet* é composta por múltiplos *chiplets* de computação idênticos, que acessam a memória de maneira compartilhada e uniforme ou por meio de recursos de I/O (*Input/Output*), como uma rede de roteadores ou de recursos de interconexão entre os *chiplets*, como um *interposer*. Segundo Han et al. (2024, p. 1437), essa organização apresenta um efeito semelhante ao modelo UMA (*Uniform Memory Access*), pois a memória uniforme pode ser acessada de forma equivalente por todos os *chiplets*. Entre suas vantagens estão a possibilidade de dividir cargas de trabalho entre *chiplets*, manter equilíbrio na distribuição das tarefas, reduzir movimentações desnecessárias de dados e fornecer certo nível de redundância, uma vez que outros *chiplets* podem assumir funções caso um módulo falhe.

A arquitetura *NUMA-chiplet* também utiliza múltiplos *chiplets* interconectados, porém diferencia-se pelo fato de a memória ser compartilhada e, ao mesmo tempo, distribuída na arquitetura. Nesse modelo, cada *chiplet* pode possuir sua própria memória local, como DRAM (*Dynamic Random-Access Memory*) ou HBM (*High Bandwidth Memory*), tornando o acesso à memória dependente da posição do *chiplet* e da localização dos dados. Essa característica aproxima a arquitetura do modelo NUMA (*Non-Uniform Memory Access*), no qual acessos locais tendem a apresentar menor latência do que acessos remotos a memórias associadas a outros *chiplets* (HAN et al., 2024, p. 1437-1438).

Na arquitetura *cluster-chiplet*, os *chiplets* são organizados em grupos, ou *clusters*, podendo formar sistemas de grande escala com grande número de núcleos. Nessa organização, cada cluster é formado por *chiplets* interconectados e pode possuir memória individual, podendo também executar seu próprio sistema

operacional. A comunicação entre *clusters* ocorre por métodos de interconexão de alto desempenho e pode utilizar passagem de mensagens. Segundo Han et al. (2024, p. 1438), esse modelo favorece escalabilidade, pois permite estruturar sistemas maiores a partir de agrupamentos menores de *chiplets*.

A arquitetura *heterogeneous-chiplet* é caracterizada pela integração de *chiplets* com funções distintas dentro de um mesmo sistema. Em vez de utilizar apenas módulos idênticos ou organizados de forma uniforme, esse modelo combina blocos especializados, como unidades de processamento, memória, I/O ou aceleradores específicos. Essa abordagem permite que cada *chiplet* seja projetado de acordo com sua função e, quando necessário, fabricado em processos tecnológicos diferentes, explorando a flexibilidade produtiva e funcional da integração baseada em *chiplets*.

Por fim, a arquitetura *hierarchical-chiplet* organiza os componentes em diferentes níveis de hierarquia. Nesse modelo, núcleos podem ser agrupados em estruturas menores, que se conectam internamente por redes de menor nível, enquanto a comunicação entre *chiplets* ocorre por uma rede de interconexão superior. Essa organização cria uma hierarquia de comunicação e memória, na qual largura de banda, latência, consumo de energia e custo variam conforme o nível de acesso. Por isso, o projeto precisa equilibrar a distribuição das cargas de trabalho e o uso dos recursos de interconexão para manter a eficiência do sistema (HAN et al., 2024, p. 1438-1439).

Tabela 1 – Comparação de tipos de arquiteturas baseadas em chiplets.

Arquitetura	**Organização**	**Acesso à memória**	**Vantagens principais**
Symmetric-chiplet	Múltiplos *chiplets* idênticos	Memória compartilhada e uniforme - UMA	Equilíbrio de tarefas; redução de movimentação de dados; redundância entre *chiplets*
NUMA-chiplet	*Chiplets* interconectados com memória distribuída	Memória compartilhada e distribuída - NUMA	Menor latência nos acessos locais; memória dedicada por *chiplet*
Cluster-chiplet	*Chiplets* em grupos (*Clusters*)	Pode ser individual por *cluster*	Alta escalabilidade; possibilidade de sistema operacional independente por *cluster*
Heterogeneous-chiplet	*Chiplets* com diferentes funções integrando um mesmo sistema	Varia de acordo com a função de cada *chiplet*	Especialização funcional; flexibilidade de integração e fabricação
Hierarchical-chiplet	*Chiplets* organizados em diferentes níveis hierárquicos	Acesso hierárquico, variando por nível	Hierarquia de comunicação e memória; adaptabilidade por nível

Fonte: Elaborado pelo autor com base em Han et al. (2024).

Conforme sintetizado na Tabela 1, as arquiteturas baseadas em *chiplets* descritas por Han et al. (2024) diferem quanto à organização, à forma de acesso à memória e às vantagens principais. Essas diferentes classificações demonstram que arquiteturas baseadas em *chiplets* não devem ser compreendidas apenas como a divisão física de um chip monolítico em múltiplos *dies*. Cada categoria envolve decisões arquiteturais específicas relacionadas ao acesso à memória, à comunicação entre módulos, à especialização funcional e ao nível de escalabilidade pretendido. Dessa forma, essa classificação fornece uma base conceitual para a análise de arquiteturas comerciais baseadas em *chiplets* nos capítulos seguintes.

2.5 2.5/3D *packaging* e interconexões

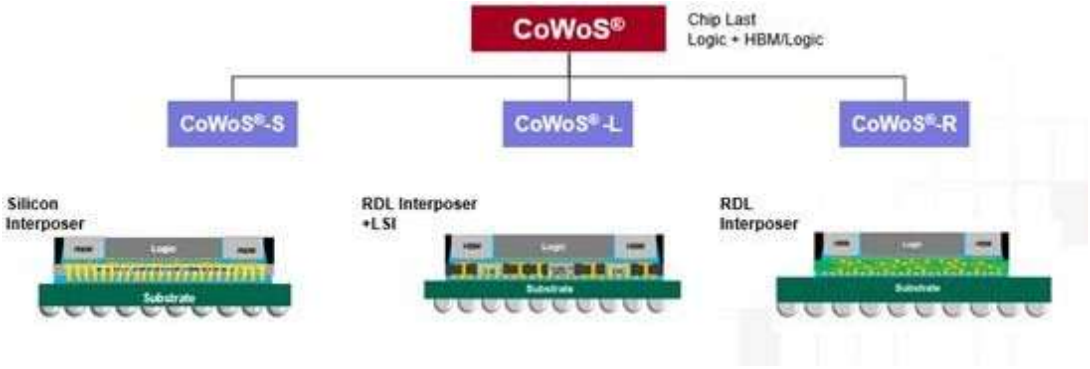
Conforme a evolução das arquiteturas baseadas em *chiplets*, o encapsulamento deixou de exercer apenas a função de proteção física e conexão externa do chip, passando a desempenhar papel central na integração elétrica, térmica e funcional dos componentes. Em chips compostos por múltiplos *dies*, a forma como esses módulos são posicionados e interligados influencia diretamente a largura de banda, latência, consumo de energia, escalabilidade e o custo do sistema. Nesse contexto, as integrações 2D, 2.5D e 3D oferecem diferentes formas de conectar *chiplets* e memórias de alta largura de banda dentro de um mesmo encapsulamento.

Na integração 2D, os *dies* são posicionados lado a lado sobre um substrato orgânico ou outro meio de interconexão. A integração 2.5D mantém essa disposição lateral, mas adiciona um elemento intermediário de alta densidade, como um *interposer*, permitindo maior densidade de sinais e menor distância elétrica entre os componentes. Segundo Han et al. (2024, p. 1434), tecnologias 2D, 2.5D e 3D permitem diferentes formas de integração de *chiplets*, variando conforme o uso de substratos, *interposer*, *microbumps*, TSVs (*Through-Silicon Vias*) e empilhamento vertical.

Um método relevante de integração 2.5D é o CoWoS (*Chip-on-Wafer-on-Substrate*), tecnologia da TSMC voltada a aplicações de alto desempenho. A TSMC define o CoWoS como um processo de encapsulamento avançado baseado em *interposer*, projetado para integração homogênea e heterogênea, sendo especialmente adequado para aplicações de HPC.

Conforme demonstrado na Figura 10, essa tecnologia possui três variações, voltadas a diferentes necessidades de integração. O CoWoS-S utiliza *interposer* de silício, oferecendo alta densidade de interconexão entre lógica e memória. O CoWoS-L combina *interposer* RDL (*Redistribution Layer*) com ponte de silício local, para equilibrar densidade de conexão e flexibilidade de integração. Já o CoWoS-R utiliza apenas o *interposer* RDL, priorizando flexibilidade e custo do substrato. Essas variações mostram que o encapsulamento 2.5D pode ser ajustado conforme os requisitos de desempenho, integração e custo do sistema (TSMC, [s.d.]).

Figura 10 – Comparação entre CoWoS-S, CoWoS-L e CoWoS-R.



Fonte: TSMC ([s.d.]).

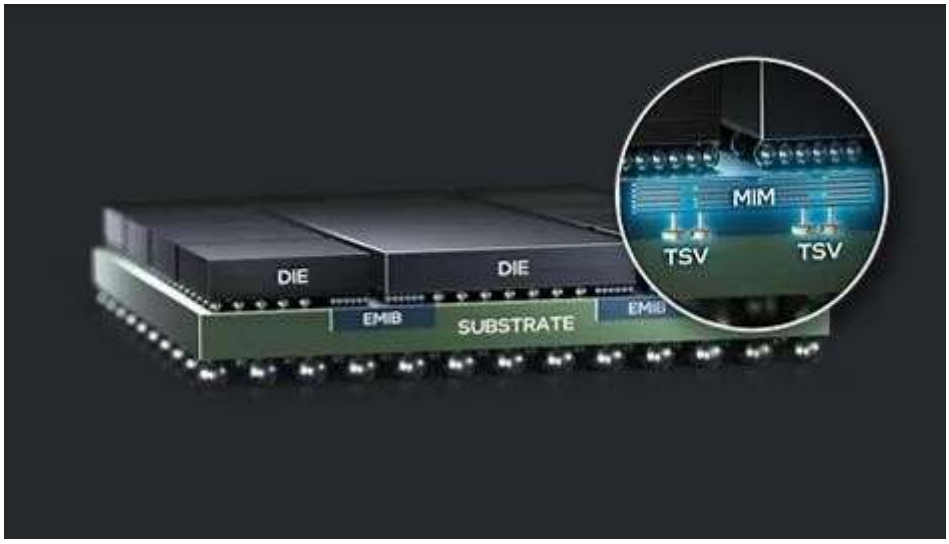
Essa tecnologia permite integrar múltiplos *dies* e memórias HBM, ampliando a comunicação entre lógica e memória no mesmo encapsulamento (TSMC, [s.d.]).

Já a integração 3D envolve o empilhamento vertical de *dies*, reduzindo a distância entre circuitos e ampliando a densidade de integração. A Imec destaca que essa abordagem pode aproximar processador e memória, além de combinar subsistemas heterogêneos fabricados em tecnologias diferentes, favorecendo sistemas grandes e complexos (IMEC, [s.d.]a).

Nesse contexto, a TSMC apresenta o SoIC (*System on Integrated Chip*) como uma tecnologia de empilhamento 3D voltada à integração de *chiplets* no nível do chip. Esse método permite empilhar dies similares ou diferentes, aumentando a densidade de interconexão e reduzindo o tamanho do sistema. Em aplicações de HPC, a TSMC associa suas tecnologias 3DFabric à combinação de SoIC, CoWoS e outras soluções para atender demandas de desempenho, largura de banda e eficiência em centros de dados e servidores de alto desempenho (TSMC, [s.d.]).

Além da TSMC, a Intel também desenvolve tecnologias próprias de encapsulamento avançado. O Intel EMIB (*Embedded Multi-die Interconnect Bridge*) foi apresentado como uma solução 2.5D eficiente para conectar múltiplos *dies* complexos.

Figura 11 – Tecnologia EMIB 2.5D da Intel.



Fonte: Recorte de Intel ([s.d.]b).

Diferente de um *interposer* completo, o EMIB utiliza uma ponte de silício embutida no substrato do encapsulamento, como exibido na Figura 11, permitindo conexões densas entre *dies* em regiões específicas. A Intel descreve o EMIB como aplicável tanto à integração lógica-lógica quanto à integração entre lógica e HBM, sendo uma alternativa para sistemas que exigem alta largura de banda entre componentes próximos (INTEL, [s.d.]b).

A Intel também apresenta tecnologias de empilhamento 3D próprias, como o *Foveros Direct*, que utiliza interconexão híbrida cobre-cobre para empilhar *chiplets* sobre um *die* base ativo. A combinação de EMIB e *Foveros* permite a criação de sistemas heterogêneos complexos, nos quais múltiplos *dies* e pilhas 3D podem ser integrados em um mesmo encapsulamento (INTEL, [s.d.]b).

Assim, os diferentes métodos de integração e encapsulamento permitem ajustar a arquitetura conforme a prioridade do projeto, como densidade de interconexão, largura de banda, custo, eficiência ou integração heterogênea. Por isso, tecnologias 2.5D e 3D são fundamentais para viabilizar arquiteturas baseadas em *chiplets* em aplicações de HPC, inteligência artificial e centros de dados.

3 ARQUITETURAS BASEADAS EM MCM

Após a apresentação dos fundamentos de fabricação, *yield rate* e integração *multi-die*, este capítulo analisa arquiteturas comerciais baseadas em *Multi-Chip Module* com o objetivo de observar como os conceitos discutidos anteriormente são aplicados em processadores reais, com foco inicial na família AMD EPYC 7001, que utilizou do método *Multi-Chip Module*, na tentativa comercial da AMD de retornar ao mercado de servidores (TIRIAS RESEARCH, 2018).

3.1 AMD EPYC 7001: “Zeppelin”

A primeira geração dos processadores AMD EPYC, lançada em 2017, representou uma das principais adoções comerciais modernas do MCM em CPUs voltadas a servidores. A série EPYC 7001 foi apresentada pela AMD como uma plataforma para *datacenters*, com modelos de até 32 núcleos Zen, suporte a grande largura de banda de memória e ampla disponibilidade de linhas PCIe (AMD, 2017).

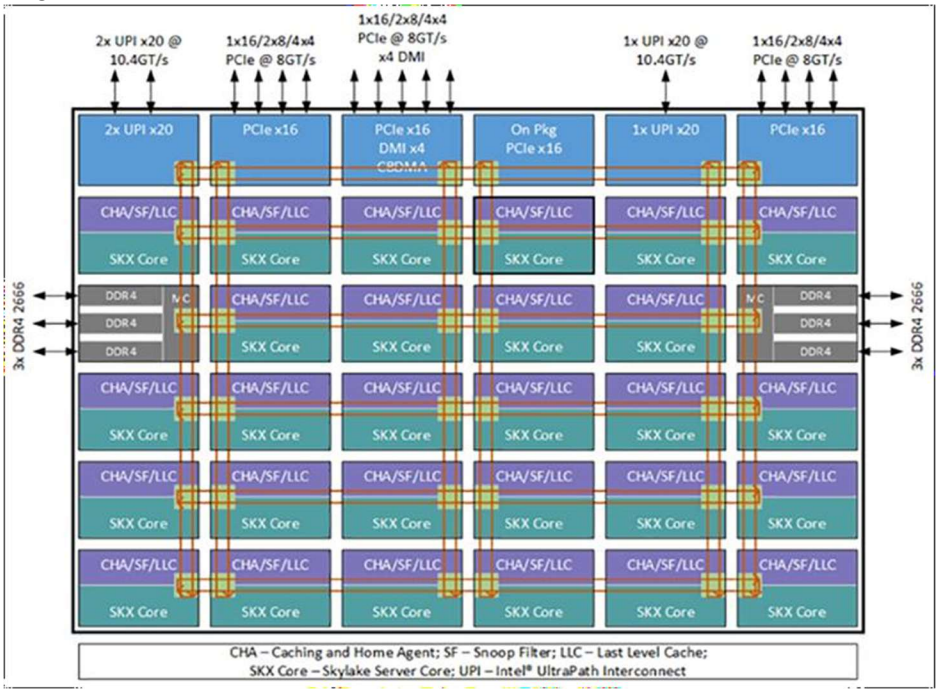
Segundo a Tirias Research (2018), os EPYC 7001 foram construídos a partir de quatro *dies* “Zeppelin” integrados em um único encapsulamento por meio de MCM. Cada *die* possuía até oito núcleos, permitindo que a AMD alcançasse o marco de 32 núcleos em um único processador. Essa configuração possibilitou a criação de uma CPU para servidores com alta contagem de núcleos, ampla largura de banda de memória de até aproximadamente 145 GB/s e grande capacidade de I/O, com até 128 trilhas PCIe, sem depender da fabricação de um único *die* monolítico de grande área.

No mesmo período, a Intel lançava a família Xeon Scalable, anunciada como sua nova plataforma para *datacenters*, com modelos de até 28 núcleos de alto desempenho (SPELMAN, 2017). Esse contraste evidencia duas abordagens distintas para o mercado de servidores: de um lado, a estratégia tradicional da Intel; de outro, a abordagem *multi-die* adotada pela AMD para atingir maior contagem de núcleos e escalabilidade.

Os Xeon Scalable Skylake-SP utilizavam uma arquitetura interna baseada em *mesh*, demonstrada na Figura 12, na qual núcleos, controladores de memória, *cache* e interfaces de I/O eram conectados por caminhos horizontais e verticais dentro do

processador. Esse método buscava permitir que um núcleo do processador se comunicasse com outro usando a rota mais curta possível (MULNIX, 2022a).

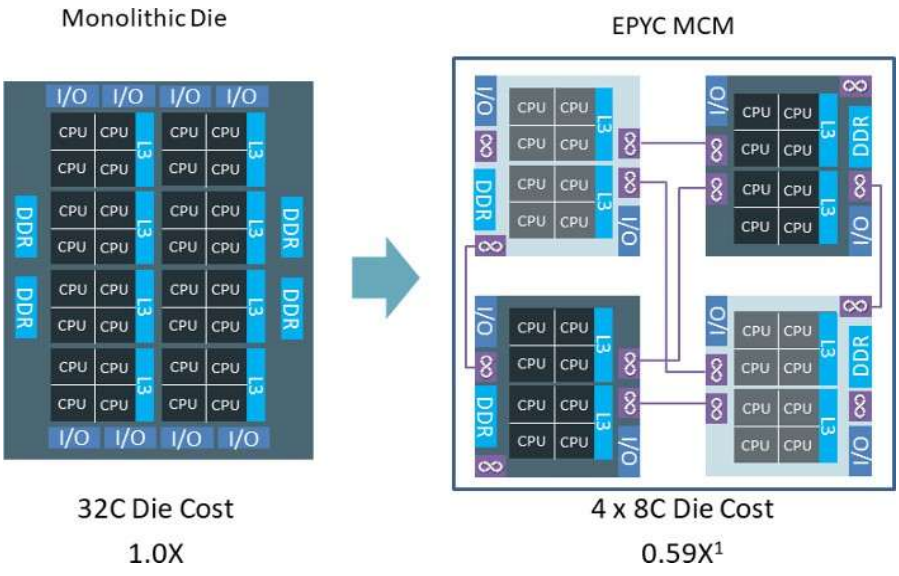
Figura 12 – Arquitetura Mesh da família Intel Xeon Scalable.



Fonte: Mulnix (2022a).

A comparação entre a abordagem monolítica e a solução MCM adotada pela AMD pode ser observada na Figura 13, juntamente à sua estimativa de custo de produção dos *dies* e a interconexão entre eles.

Figura 13 – Comparação entre *die* monolítico e arquitetura EPYC MCM.



Fonte: AMD apud Tirias Research (2018, p. 2).

A principal vantagem produtiva dessa abordagem estava no uso de *dies* menores e mais viáveis para fabricação. A Tirias Research (2018) aponta que o encapsulamento dos EPYC era composto por quatro *dies* com área de 213 mm²,

totalizando 852 mm² de silício. Caso essa lógica fosse implementada de maneira monolítica em um único *die*, mesmo com economia de área pela remoção de parte das interconexões entre *dies*, o chip hipotético ainda teria cerca de 777 mm², o que resultaria em um *yield rate* menor e custo de fabricação maior. Dessa forma, o uso de múltiplos *dies* permitiu à AMD reduzir custos e melhorar o aproveitamento produtivo.

Essa abordagem, entretanto, também introduziu desafios arquiteturais. Como os quatro *dies* possuíam seus próprios controladores de memória e eram interligados pelo *Infinity Fabric*, tecnologia proprietária de interconexão da AMD, o acesso à memória passou a seguir um modelo NUMA, no qual a latência varia conforme a localização dos dados em relação ao núcleo que os acessa. Assim, o ganho produtivo e econômico do MCM veio acompanhado da necessidade de otimização do acesso à memória pelo sistema operacional e pelas aplicações.

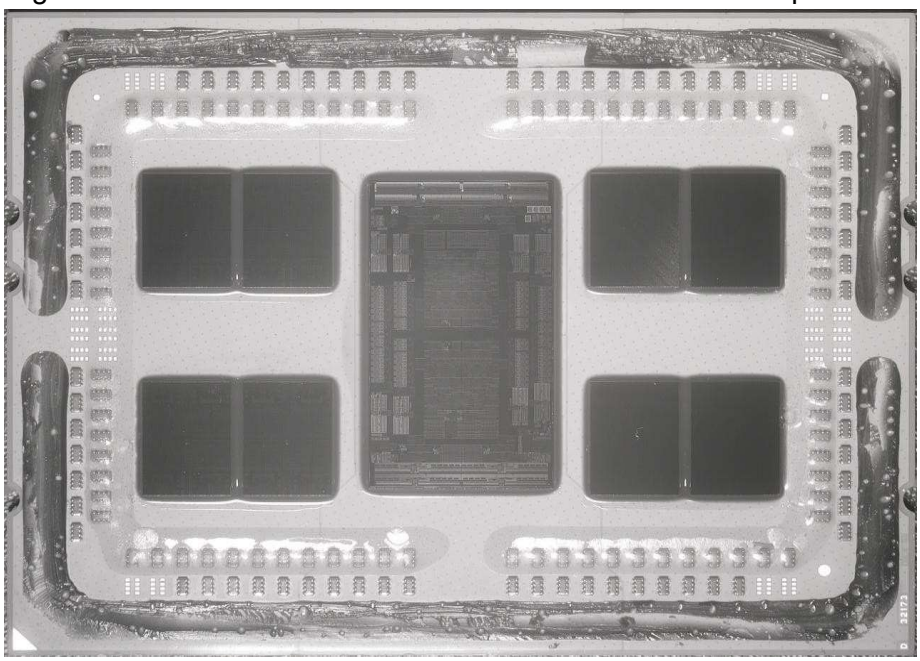
3.2 AMD EPYC 7002: maturidade da arquitetura *multi-die*

Os EPYC 7002 foram a segunda geração dos processadores AMD EPYC, trazendo uma evolução muito significativa na forma como aplicavam o método MCM. Enquanto os EPYC 7001 utilizavam múltiplos *dies* Zeppelin, cada um contendo núcleos, controladores de memória e interfaces de I/O, os EPYC 7002 reorganizaram a arquitetura ao separar os módulos de computação dos módulos de entrada e saída.

Nessa geração, a AMD adotou uma arquitetura híbrida *multi-die*, composta por *dies* de computação que contêm os núcleos e o *cache*, conectados a um *I/O die* central. Essa separação permitiu que os módulos de computação, denominados pela AMD de CCDs (*Core Complex Dies*), fossem fabricados em processo de 7 nm, enquanto o módulo responsável pelas funções de I/O, acesso à memória e segurança fosse implementado em processo de 14 nm, reduzindo a necessidade de aplicar o nó mais avançado a todos os componentes do processador (AMD, 2020).

A Figura 14 permite observar a organização física do encapsulamento de um EPYC 7702, da família EPYC 7002, com os módulos de computação distribuídos ao redor do *I/O die* central.

Figura 14 – Interior de um EPYC 7702 ES em infravermelho próximo.



Fonte: Fritzchens Fritz (2019).

A família EPYC 7002 alcançou até 64 núcleos e 128 *threads* por soquete, com suporte a oito canais de memória DDR4 e 128 linhas PCIe Gen4. Além disso, ampliou o *cache* L3 para até 32 MB por CCD*,* totalizando até 256 MB nos modelos mais avançados da família, o que aumentou significativamente a escalabilidade em relação à geração anterior (AMD, 2020).

Com esses avanços, os EPYC 7002 não apenas dobraram a contagem máxima de núcleos em relação à geração anterior, mas também demonstraram uma aplicação mais madura do MCM. A arquitetura deixou de apenas replicar *dies* completos no encapsulamento e passou a distribuir funções de forma mais especializada entre *chiplets* de computação e um *die* central de *I/O*.

De acordo com a AMD (2020), o projeto da série como um SoC permitiu eliminar a necessidade de múltiplos *chips* externos de suporte, ajudando a reduzir os custos dos servidores.

Outro fator relevante da série 7002 foi a manutenção de compatibilidade com parte da infraestrutura da geração anterior. Segundo a AMD (2020), algumas funcionalidades dos processadores EPYC de segunda geração exigem atualização de BIOS (*Basic Input/Output System*) quando utilizadas em placas-mãe projetadas para a primeira geração EPYC, enquanto o uso de todos os recursos disponíveis requer uma placa-mãe projetada especificamente para a segunda geração. Dessa

forma, a AMD preservou certo grau de retrocompatibilidade, permitindo adoção gradual em servidores já existentes, ainda que nem todos os recursos fiquem disponíveis em plataformas anteriores.

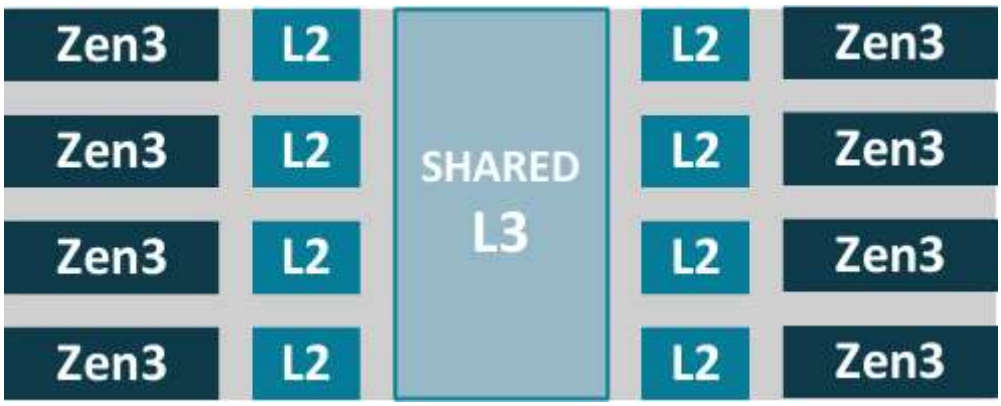
Além das melhorias arquiteturais e de escalabilidade, os EPYC 7002 trouxeram avanços notórios em eficiência energética. Em testes da Phoronix comparando *Naples*, geração dos EPYC 7001, e *Rome*, geração dos EPYC 7002, a nova geração apresentou avanço expressivo em eficiência energética e desempenho por watt, mantendo temperaturas próximas às da geração anterior, mesmo com ganhos significativos de desempenho (LARABEL, 2019).

3.3 AMD EPYC 7003: maturação dos módulos e 3D V-Cache

A terceira geração dos processadores AMD EPYC, conhecida como EPYC 7003, deu continuidade ao esquema de organização baseada em *chiplets* adotado na geração anterior. A arquitetura manteve a separação entre módulos de computação e o *I/O die* central, preservando a lógica de especialização dos blocos funcionais dentro do encapsulamento. Dessa forma, os EPYC 7003 representam menos uma mudança estrutural completa e mais um refinamento da abordagem *multi-die* consolidada nos EPYC 7002.

A principal evolução da família está associada à arquitetura Zen 3, que reorganizou internamente os blocos de núcleos e *cache*, melhorando a comunicação entre os núcleos e o acesso ao *cache* L3, que pode ser observado na Figura 15.

Figura 15 – Estrutura de um CCD Zen 3.



Fonte: AMD (2022, p. 4).

O conjunto interno ao CCD pode ser dividido em múltiplos CCX (*Core Complex*), como a AMD denomina. Esses CCX funcionam como módulos internos ao

CCD; nas arquiteturas Zen anteriores ao Zen 3, cada CCD possuía dois CCX de até quatro núcleos. Essa mudança reforça a maturidade da abordagem baseada em *chiplets*, pois mantém a lógica modular da geração anterior enquanto aprimora o desempenho interno dos módulos de computação.

Um ponto relevante da geração 7003 dos EPYC é a introdução da tecnologia 3D V-Cache da AMD em modelos específicos, na qual um *chiplet* adicional de *cache* é empilhado verticalmente sobre o *die* de computação. Essa abordagem amplia a capacidade de *cache* sem depender apenas do aumento da área horizontal do *die*, demonstrando uma nova etapa na evolução das técnicas de integração: além de distribuir funções em múltiplos *dies* no plano horizontal, torna-se possível empilhar componentes verticalmente para aumentar densidade e desempenho.

Assim, os EPYC 7003 demonstram de maneira prática que a evolução do MCM não ocorre apenas pela adição de mais *dies* no encapsulamento, mas também pelo refinamento da organização interna dos *chiplets* e pela adoção de técnicas de integração tridimensional. Essa geração reforça a tendência de que processadores modernos passem a combinar modularidade, especialização funcional e empilhamento vertical como caminhos para contornar limitações de área, custo e escalabilidade.

3.4 Xeon Scalable – Sapphire Rapids

A quarta geração dos processadores Intel Xeon Scalable, conhecida pelo codinome Sapphire Rapids, foi lançada oficialmente em 10 de janeiro de 2023 (INTEL, 2023) e representa uma mudança importante na organização das CPUs de *datacenter* da Intel. Enquanto a família Xeon Scalable baseada em Skylake-SP e Ice Lake utilizava uma arquitetura interna em malha dentro de uma organização monolítica, os Sapphire Rapids passaram a adotar uma estrutura baseada em múltiplos *tiles*, aproximando-se das estratégias de integração *multi-die* discutidas anteriormente, como ilustrado na Figura 16.

Figura 16 – Comparação visual de arquitetura entre Ice Lake e Sapphire Rapids.

Fonte: Biswas (2021, p. 6).

A geração introduziu suporte a tecnologias como DDR5, PCIe 5.0 e CXL (*Compute Express Link*), além de aceleradores integrados voltados a tarefas específicas, recurso destacado pela Intel como um dos diferenciais da 4ª geração Xeon Scalable (MULNIX, 2022b; INTEL, [s.d.]).

Entre eles, podemos destacar o Intel AMX (*Advanced Matrix Extensions*), voltado à aceleração de cargas de inteligência artificial e operações de matriz; o Intel DSA (*Data Streaming Accelerator*), direcionado à movimentação e cópia eficiente de dados; o Intel IAA (*In-Memory Analytics Accelerator*) voltado à análise e compressão de dados em memória; o Intel QAT (*QuickAssist Technology*), aplicado à compressão e criptografia; e o Intel DLB (*Dynamic Load Balancer*), para balanceamento de cargas em redes e sistemas distribuídos (BISWAS, 2021).

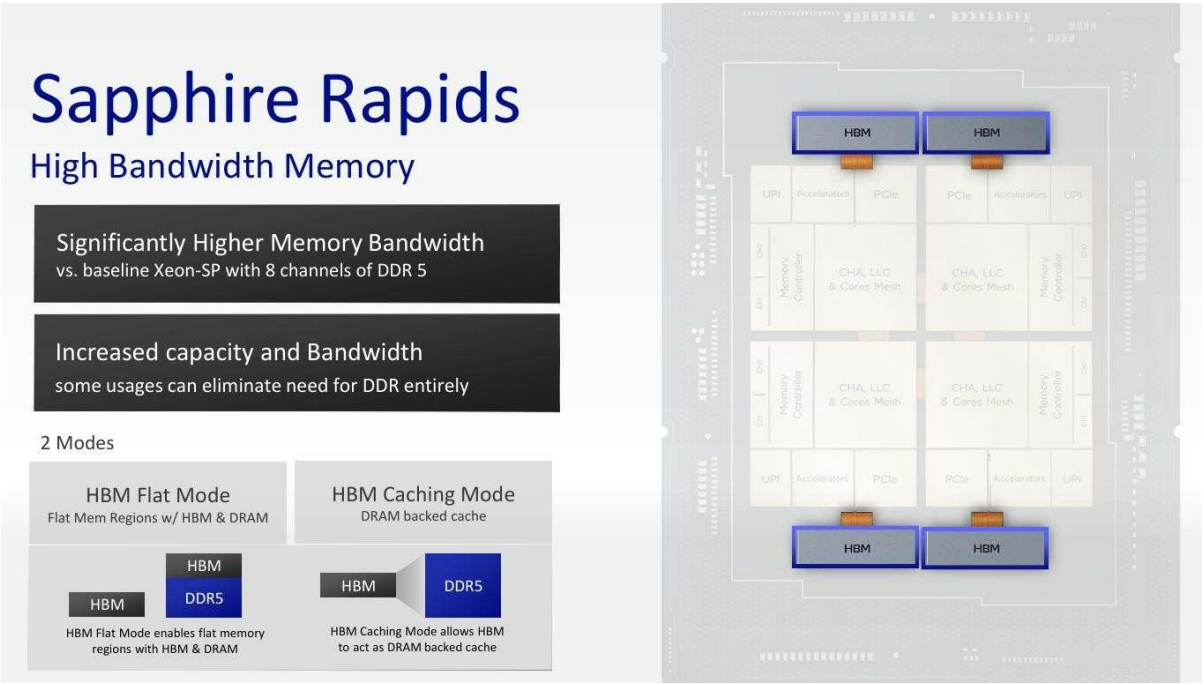
No aspecto físico e arquitetural, os Sapphire Rapids adotaram uma organização composta por múltiplos *tiles* interligados pela tecnologia EMIB de interconexão 2.5D da Intel, que possibilitou a integração de diferentes blocos dentro de um mesmo encapsulamento, mantendo a comunicação de alta densidade entre os componentes. Dessa forma, a geração representa uma transição em relação às arquiteturas monolíticas anteriores, aplicando a lógica de modularidade também em CPUs Xeon voltadas a *datacenters*.

Um fator relevante dessa geração é que há modelos com memória HBM, conhecidos como Xeon Max Series. Essas versões integram memória de alta largura de banda próxima ao processador, com o objetivo de atender aplicações sensíveis à largura de banda de memória. Segundo Shipman et al. (2022), a presença de HBM

nos processadores Sapphire Rapids tem como objetivo melhorar o desempenho em cargas de trabalho HPC, especialmente quando limitadas pelo acesso à memória.

A inclusão de HBM, ilustrada na Figura 17, reforça a importância do encapsulamento avançado na evolução dos processadores modernos. Em vez de depender apenas de canais de memória externos tradicionais, a integração de memória HBM no mesmo encapsulamento reduz a distância entre processamento e memória, ampliando a taxa de transferência disponível para aplicações intensivas em dados.

Figura 17 – Organização de memória em Sapphire Rapids com HBM.



Fonte: Biswas (2021, p. 18).

Essa característica aproxima as versões com HBM dos Sapphire *Rapids* das soluções discutidas anteriormente no contexto da integração 2.5D, nas quais lógica e memória podem ser posicionadas de forma mais próxima para reduzir gargalos de comunicação. No caso dessa geração, a integração de múltiplos *tiles* por meio de EMIB e a presença de versões com HBM demonstram como o encapsulamento avançado pode ser utilizado para ampliar a largura de banda e a escalabilidade em CPUs para servidores.

A introdução dos Sapphire Rapids demonstra a ampliação da adoção de arquiteturas modulares em produtos comerciais de alto desempenho. Ao incorporar uma organização *multi-tile* em sua linha Xeon Scalable, a Intel reforça que o uso de MCM e encapsulamento avançado passou a ser uma estratégia relevante não apenas

para aumentar a contagem de núcleos, mas também para integrar recursos especializados, memória de alta largura de banda e interconexões de maior densidade em processadores voltados a *datacenters.*

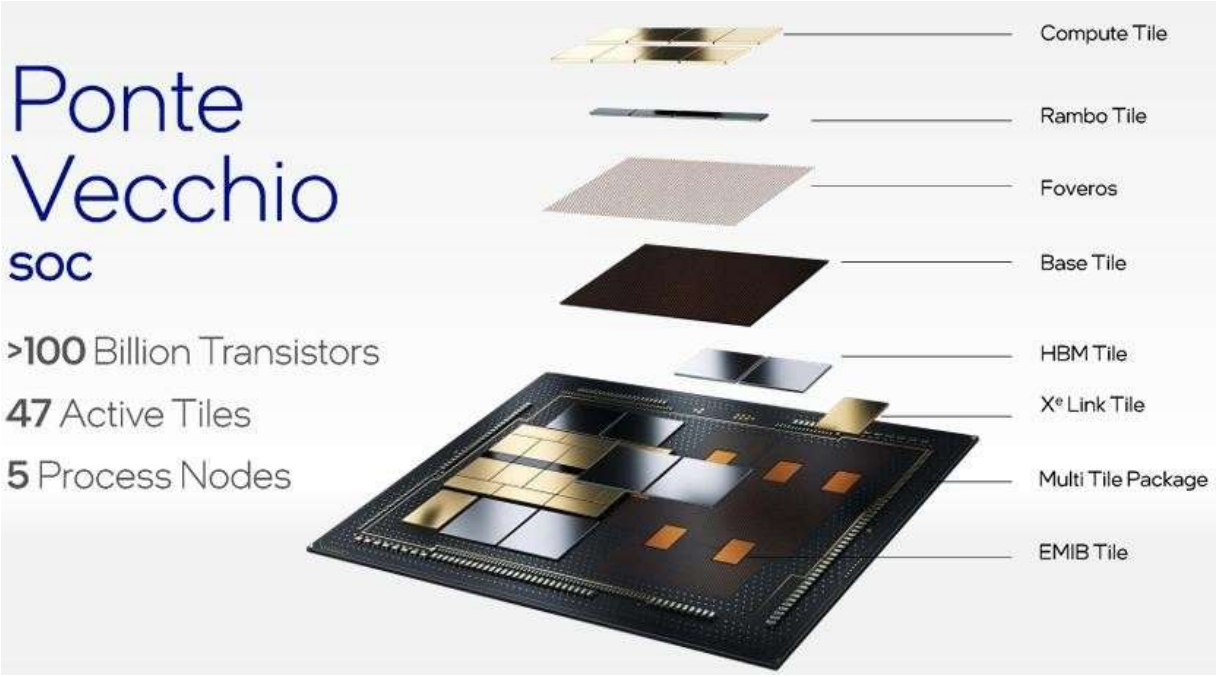
3.5 Intel Ponte Vecchio

Os aceleradores Ponte Vecchio, da Intel, representam um dos casos mais complexos de implementação MCM já produzidos comercialmente. A Intel os apresenta como GPUs (*Graphics Processing Units*), SoCs ou aceleradores voltados à computação de exaescala, especialmente em aplicações de HPC e inteligência artificial. A arquitetura adota uma organização *multi-tile* hierárquica estruturada em três níveis: *Core*, *Slice* e *Stack* (BLYTHE, 2021).

Essa organização hierárquica permite que o acelerador seja construído a partir de blocos menores e especializados, em vez de concentrar toda a lógica em um único *die* monolítico. Dessa forma, o Ponte Vecchio exemplifica uma aplicação avançada da modularidade, na qual desempenho, escalabilidade e especialização funcional dependem não apenas da quantidade de unidades de computação, mas também da forma como os tiles são organizados e interconectados.

No nível mais básico, o *Xe-core* é a unidade fundamental de computação, contendo oito motores vetoriais de 512 bits e oito motores matriciais de 4096 bits, além de *cache* L1 e unidade de carga e armazenamento (BLYTHE, 2021). Um conjunto de 16 *Xe-cores* forma um *Xe Slice*, que inclui ainda 16 unidades de *ray tracing* e 8 MB de *cache* L1. Cada *Xe Stack* agrupa até quatro *Slices*, totalizando 64 *Xe-cores*, quatro controladores HBM2E e oito *Xe Links* por *Stack* (BLYTHE, 2021). O produto resultante, a GPU Ponte Vecchio, combina dois *Stacks*, chegando a 128 *Xe-cores*, 128 unidades de *ray tracing* e oito controladores HBM2E em um único encapsulamento (BLYTHE, 2021).

Figura 18 – Organização *multi-tile* do SoC Ponte Vecchio.



Fonte: Recorte de Blythe (2021, p. 18).

No total, conforme apresentado na Figura 18, o acelerador integra 47 *tiles* ativos e cinco nós de processo distintos, com mais de 100 bilhões de transistores (BLYTHE, 2021). Embora a Intel não indique quais são todos os cinco processos utilizados, é informado que os *Compute Tiles* são fabricados no processo TSMC N5; os *RAMBO Tiles* e o *Base Tile* no Intel 7; e os *Xe Link Tiles* no TSMC N7, com cada componente alocado no nó mais adequado à sua função (GOMES et al., 2022). Essa organização evidencia uma característica central das arquiteturas MCM modernas: a possibilidade de combinar blocos especializados, fabricados em tecnologias distintas, dentro de um mesmo encapsulamento. A integração entre *tiles* é viabilizada pela combinação de *Foveros* e EMIB, resultando em um processador com envelope de potência de 600 W, desempenho superior a 45 TFLOPS em FP32 e largura de banda de memória acima de 5 TB/s (GOMES et al., 2022). Dessa forma, o Ponte Vecchio demonstra como o encapsulamento avançado deixa de ser apenas um recurso físico de montagem e passa a atuar como elemento essencial para viabilizar arquiteturas heterogêneas de alta complexidade.

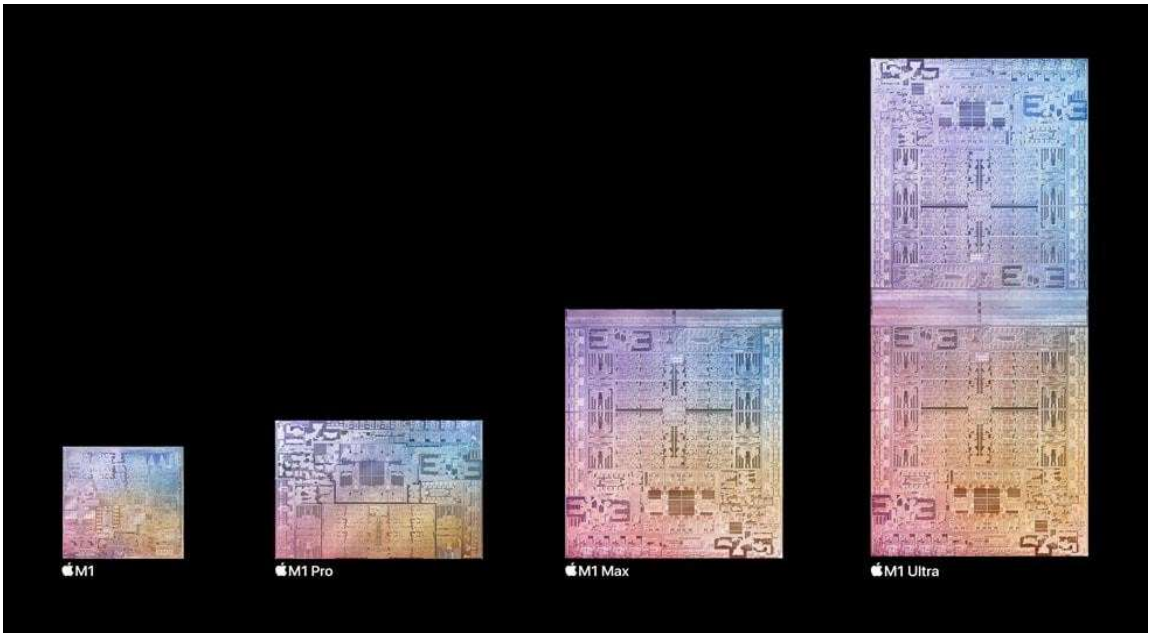
3.6 Apple M1 Ultra: integração *symmetric-chiplet* com UltraFusion

O Apple M1 Ultra é um exemplo de aplicação comercial da integração de *chiplets* fora do segmento tradicional de servidores. Lançado em 2022, este SoC da Apple foi desenvolvido a partir de uma interconexão entre dois *dies* M1 Max, por meio da tecnologia UltraFusion, uma arquitetura de encapsulamento da Apple baseada em *interposer* de silício. Essa interconexão conecta mais de 10.000 sinais entre os dois *dies* e fornece até 2,5 TB/s de largura de banda, permitindo que o sistema operacional reconheça o conjunto como um único SoC (APPLE, 2022).

Diferente de arquiteturas *NUMA-chiplet*, em que a localização da memória pode impactar a latência de acesso, o M1 Ultra, ilustrado na Figura 19, foi apresentado como uma solução capaz de operar de forma unificada para o *software*. Essa característica aproxima o projeto da categoria *symmetric-chiplet*, pois a integração combina os módulos em uma solução reconhecida como um único *chip* pelo sistema.

Assim, o M1 Ultra demonstra que a integração baseada em *chiplets* não se limita a ampliar a contagem de núcleos em processadores de servidor. Ela também pode ser empregada para escalar SoCs de alto desempenho, preservando uma experiência de software semelhante à de um *chip* monolítico, ao mesmo tempo em que evita a necessidade da fabricação de um único *die* de área maior.

Figura 19 – Comparação visual dos *dies* da família Apple M1 até o M1 Ultra.



Fonte: Apple (2022).

3.7 AMD Instinct MI250X e MI300A: aceleradores MCM para HPC e IA

A linha AMD Instinct demonstra a aplicação de arquiteturas modulares em aceleradores voltados a HPC e inteligência artificial. Enquanto os subcapítulos anteriores abordaram principalmente CPUs e SoCs, os aceleradores Instinct evidenciam como a integração *multi-die* também se tornou essencial em dispositivos dedicados a paralelismo, grande largura de banda de memória e processamento massivo de dados.

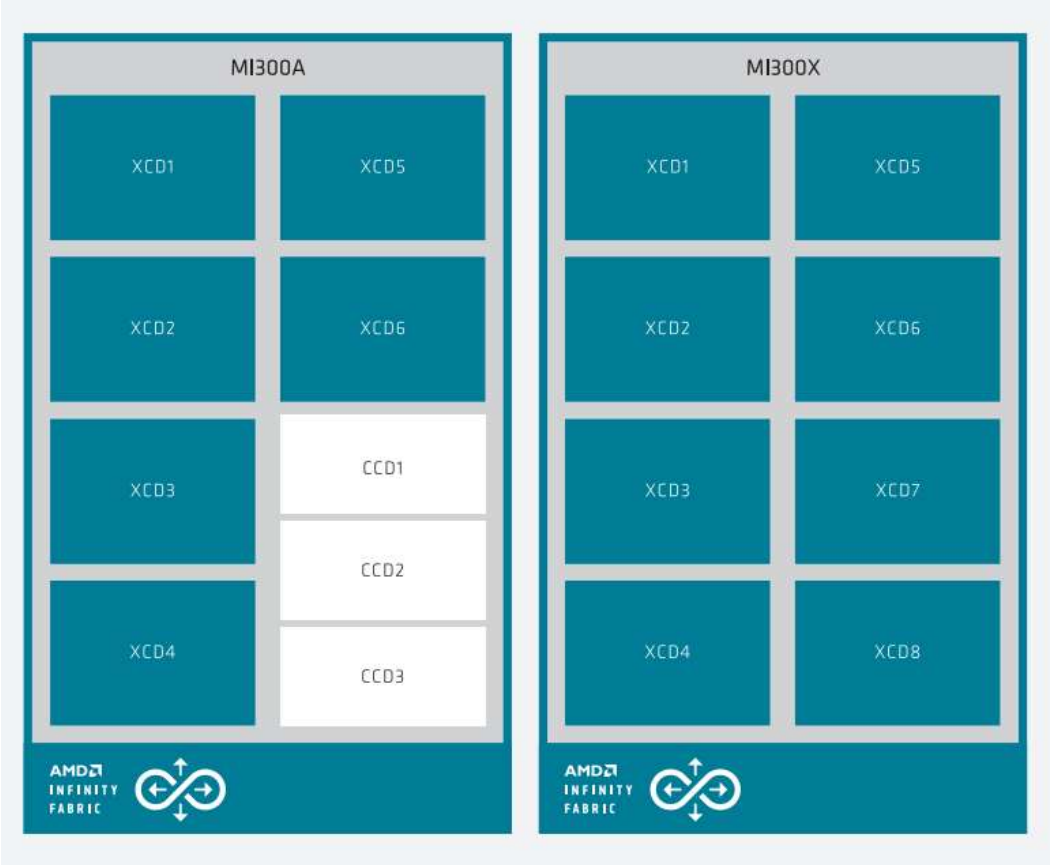
A AMD Instinct MI250X, baseada na arquitetura CDNA 2, representa um exemplo dessa abordagem em aceleradores para HPC. Segundo a AMD (2021), a arquitetura CDNA 2 utiliza o *Infinity Fabric* para viabilizar um MCM no qual a MI250 e a MI250X integram dois GCDs (*Graphics Compute Dies*) em um único acelerador no formato OAM (*Open Accelerator Module*). Em comparação, a MI210 utiliza apenas um GCD em formato PCIe. Dessa forma, a MI250X pode ser compreendida como uma solução modular que amplia recursos computacionais por meio da integração de dois *dies* gráficos dentro de um mesmo encapsulamento.

A geração MI300 representa um avanço mais amplo nessa estratégia. A AMD Instinct MI300A é uma APU (*Accelerated Processing Unit*) voltada a HPC e inteligência artificial que integra, em uma única solução, 24 núcleos AMD Zen 4, 228 unidades de computação CDNA 3 e 128 GB de memória HBM3 unificada. Essa memória apresenta um espaço de endereçamento compartilhado entre CPU e GPU, reduzindo a necessidade de movimentação explícita de dados entre processador e acelerador. A solução é interconectada pela 4ª geração da AMD *Infinity Architecture*, com largura de banda de memória de até 5,3 TB/s (AMD, 2025).

No aspecto de fabricação e encapsulamento, a MI300A reforça a lógica de especialização funcional dos *chiplets*. O *datasheet* da AMD informa o uso de processo de 5 nm FinFET para os módulos principais de computação e 6 nm FinFET para os *I/O dies*, demonstrando a possibilidade de combinar diferentes nós de fabricação em um mesmo produto (AMD, 2025). A documentação ROCm (*Radeon Open Compute*) da AMD também indica que a MI300A utiliza uma organização composta por 6 XCDs (*Accelerator Complex Die*) e 3 CCDs, enquanto a MI300X utiliza 8 XCDs, evidenciando a diferença entre uma APU híbrida CPU-GPU e um acelerador dedicado apenas à computação por GPU (AMD, 2026).

A MI300X pode ser mencionada como contraponto dentro da mesma família, conforme ilustrado na Figura 20. Diferente da MI300A, que combina núcleos Zen 4, GPU CDNA 3 e HBM3 em uma única APU, a MI300X é um acelerador baseado apenas em GPU, com 304 unidades de computação, 192 GB de HBM3 e largura de banda de memória de até 5,3 TB/s. A comparação entre os dois modelos mostra a flexibilidade da arquitetura MI300: a mesma família pode ser adaptada tanto para integração heterogênea CPU-GPU quanto para aceleradores com maior capacidade de memória e computação voltados a IA e HPC (AMD, 2023; AMD, 2026).

Figura 20 – Comparação entre organizações de módulos MI300A e MI300X.



Fonte: AMD (2026).

Dessa forma, a linha AMD Instinct evidencia a evolução do MCM para além de CPUs pela AMD. A MI250X demonstra a ampliação de recursos por meio da integração de múltiplos *dies* gráficos, enquanto a MI300A mostra uma etapa mais avançada de integração heterogênea, combinando CPU, GPU, I/O e HBM3 em um mesmo encapsulamento. Essas arquiteturas reforçam a importância atual do MCM, dos *chiplets* e do encapsulamento avançado para aplicações de HPC e inteligência artificial, servindo também como base para os supercomputadores a serem analisados no capítulo seguinte.

4 APLICAÇÕES PRÁTICAS ENVOLVENDO MCM

Os conceitos de fabricação e integração modular discutidos nos capítulos anteriores encontram sua expressão mais concreta nos sistemas de alto desempenho em operação atualmente. A arquitetura MCM deixou de ser uma solução experimental para se tornar uma base técnica relevante em supercomputadores de ponta, aceleradores utilizados no treinamento de modelos de inteligência artificial e sistemas de computação em nuvem, que sustentam parte significativa da infraestrutura digital global.

Em ambientes de HPC, a demanda por grande capacidade de processamento, alta largura de banda de memória e comunicação eficiente entre componentes torna o uso de arquiteturas modulares especialmente relevante. Sistemas recentes presentes na lista TOP500, como El Capitan, Frontier e Aurora, demonstram essa tendência ao utilizar processadores e aceleradores baseados em soluções *multi-die*, HBM e interconexões avançadas.

Dessa forma, os subcapítulos seguintes analisam como essas arquiteturas são empregadas em aplicações concretas, conectando os exemplos técnicos apresentados nos capítulos anteriores com seu uso em plataformas computacionais voltadas a cargas críticas e de larga escala.

4.1 MCM em HPC e supercomputadores

A presença de arquiteturas baseadas em MCM nos sistemas de maior desempenho computacional do mundo é uma realidade consolidada. Na 66ª edição da lista TOP500, publicada em novembro de 2025, quatro sistemas ultrapassaram a marca de um exaflop, isto é, a capacidade de executar aproximadamente um quintilhão de operações de ponto flutuante por segundo no *benchmark* HPL. Entre esses sistemas, El Capitan, Frontier e Aurora demonstram de forma direta a presença de processadores e aceleradores baseados em soluções *multi-die*, HBM e interconexões avançadas (TOP500, 2025a).

O El Capitan, ilustrado na Figura 21, foi implantado no *Lawrence Livermore National Laboratory* (LLNL), ocupando a primeira posição da lista com 1,809 exaflop/s no *benchmark* HPL. O sistema utiliza APUs AMD Instinct MI300A, unidades híbridas que integram núcleos de CPU Zen 4, unidades de computação CDNA 3 e memória HBM em uma solução voltada a HPC e inteligência artificial (TOP500, 2025a). Essa

arquitetura, baseada em memória compartilhada entre CPU e GPU, reduz a necessidade de replicação de dados e de transferências explícitas entre espaços de memória separados, o que representa uma vantagem para aplicações que combinam simulação científica e aceleração por GPU (TANDON et al., 2024).

Figura 21 – Supercomputador El Capitan, equipado por AMD Instinct MI300A.

Fonte: AMD (2024).

O sistema é utilizado pela *National Nuclear Security Administration* em simulações relacionadas à segurança nuclear, modelagem física de alta energia e fluxos de trabalho baseados em inteligência artificial (LLNL, 2024).

O *Frontier*, instalado no *Oak Ridge National Laboratory*, ocupa a segunda posição da lista com 1,353 exaflops/s. O sistema utiliza processadores AMD EPYC de 3ª geração e aceleradores AMD Instinct MI250X, baseados na arquitetura CDNA 2 e em uma organização modular com múltiplos *dies* gráficos. Dessa forma, o *Frontier* representa uma aplicação direta da integração MCM em aceleradores voltados a HPC, servindo como antecessor relevante da geração MI300A utilizada no El Capitan (TOP500, 2025a).

O Aurora, instalado no *Argonne Leadership Computing Facility*, ocupa a terceira posição da lista TOP500 com 1,012 exaflop/s. Desenvolvido pela HPE em colaboração com a Intel, o sistema utiliza processadores Intel Xeon CPU Max Series e aceleradores Intel Data Center GPU Max Series, baseados na arquitetura Ponte Vecchio analisada anteriormente. Sua configuração inclui 21.248 processadores Intel

Xeon Max 9470 e 63.744 aceleradores *Intel Max Series*, interligados por uma infraestrutura HPE Cray com rede Slingshot, evidenciando a aplicação de arquiteturas *multi-tile* e encapsulamento avançado em sistemas de HPC de escala extrema (HPE, 2024; TOP500, 2025a).

4.2 Aplicações em inteligência artificial e simulações científicas

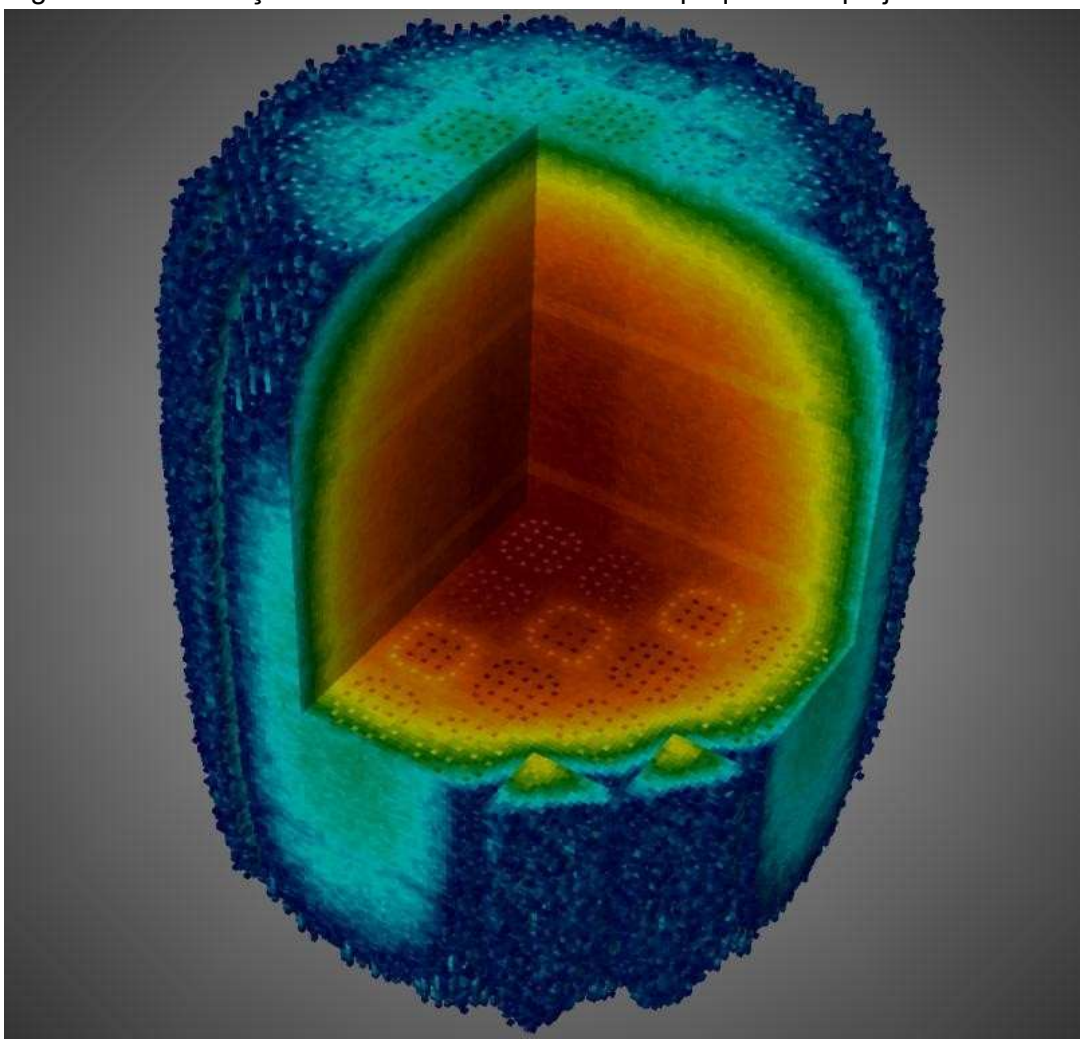
Para além dos supercomputadores, os aceleradores MCM encontraram aplicação direta no mercado de inteligência artificial, especialmente no treinamento e na inferência de LLMs (*Large Language Models*). O AMD Instinct MI300X, que conta com 192 GB de memória HBM3 e largura de banda de até 5,3 TB/s, foi projetado especificamente para modelos que exigem grande capacidade de memória, como o GPT-3 (175 bilhões de parâmetros), BLOOM (176 bilhões) e PaLM (540 bilhões) (AMD, 2024).

A arquitetura unificada do MI300A, por sua vez, trouxe benefícios específicos para aplicações científicas que alternam intensamente entre processamento de CPU e GPU. Ao compartilhar o mesmo espaço de memória física entre os núcleos Zen 4 e as unidades CDNA 3, a arquitetura elimina o gargalo de transferência de dados, simplifica o desenvolvimento de aplicações e permite o processamento incremental de códigos científicos de grande escala como o OpenFOAM, biblioteca utilizada em dinâmica de fluidos computacional (TANDON et al., 2024).

O método *Monte Carlo* é um exemplo relevante de aplicação científica que demanda grande capacidade de processamento. Esse método utiliza amostragens probabilísticas para estimar o comportamento de sistemas complexos, sendo aplicado em áreas como física de plasmas, transporte de partículas e simulações nucleares. No projeto ExaSMR, do *Oak Ridge National Laboratory*, a modelagem de pequenos reatores modulares combina transporte de nêutrons por *Monte Carlo* com dinâmica dos fluidos computacional, formando simulações multifísicas de alta fidelidade voltadas à execução eficiente em sistemas *exascale* (ORNL, [s.d.]).

A Figura 22 apresenta uma simulação multifísica desenvolvida no contexto do ExaSMR, voltada ao estudo de um reator modular pequeno, exemplificando o tipo de aplicação científica que exige grande capacidade de processamento em sistemas HPC.

Figura 22 – Simulação multifísica de reator modular pequeno no projeto ExaSMR.



Fonte: ORNL ([s.d.]).

De forma semelhante, Hirvijoki et al. (2015) apontam que simulações de partículas rápidas em plasmas de tokamak, mesmo utilizando simplificações como a formulação por centro-guia, ainda podem ser muito intensivas em CPU, justificando o uso de computação de alto desempenho. Esse cenário evidencia que aplicações científicas avançadas frequentemente exigem modelos numéricos complexos, grande quantidade de iterações e processamento paralelo em larga escala. Assim como no projeto ExaSMR, métodos baseados em Monte Carlo demonstram a necessidade de arquiteturas capazes de lidar com alto volume de cálculos e movimentação eficiente de dados. Dessa forma, essas aplicações demonstram o motivo pelo qual sistemas HPC modernos dependem de grande paralelismo, alta largura de banda de memória e aceleradores capazes de processar grandes volumes de dados científicos.

4.3 Impacto da modularidade na escalabilidade

A análise dos sistemas, tecnologias e aplicações apresentados anteriormente demonstra que a modularidade se tornou um fator central para a escalabilidade das arquiteturas de processamento modernas. Com as limitações associadas à produção de *chips* monolíticos cada vez maiores, as arquiteturas baseadas em MCM passaram a oferecer uma alternativa ao distribuir funções entre múltiplos *dies* e integrar diferentes blocos dentro de um mesmo encapsulamento. Essa abordagem possibilita a criação de soluções mais amplas e especializadas, capazes de combinar diferentes funções, ampliar a capacidade computacional, aumentar a largura de banda de memória e melhorar a viabilidade produtiva de forma mais flexível.

A modularidade também contribui para a especialização dos componentes. Como observado nas arquiteturas estudadas, diferentes blocos podem ser projetados para funções específicas, como computação geral, processamento vetorial, aceleração matricial, entrada e saída, *cache* ou memória de alta largura de banda. Essa separação permite que cada parte do sistema seja otimizada conforme sua função, inclusive utilizando processos de fabricação distintos. Essa lógica aparece em processadores como os EPYC 7002 e 7003, nos quais os módulos de computação são separados do *I/O die*, e em soluções mais complexas, como a Ponte Vecchio e a MI300A, que integram múltiplos blocos especializados em um mesmo encapsulamento.

Outro impacto relevante está na relação entre processamento e memória. Aplicações de HPC, inteligência artificial e simulações científicas frequentemente lidam com grandes volumes de dados e podem ser limitadas pela largura de banda de memória. Nesse contexto, o uso de HBM e a aproximação física entre lógica e memória reduzem gargalos de comunicação e ampliam a taxa de transferência disponível para as aplicações. Essa característica é particularmente importante em cargas de trabalho como simulações multifísicas, métodos baseados em *Monte Carlo* e modelos de inteligência artificial, que exigem grande paralelismo e movimentação intensiva de dados.

Nos supercomputadores analisados, essa tendência aparece de forma evidente, já que El Capitan, Frontier e Aurora ocupam as três primeiras posições da edição de novembro de 2025 da lista TOP500, como ilustrado na Figura 23. Esses sistemas utilizam arquiteturas modulares, memórias HBM, interconexões avançadas

e processadores ou aceleradores compostos por múltiplos *dies*. Esses sistemas demonstram como a modularidade permitiu ampliar a capacidade de processamento ao combinar diferentes componentes especializados dentro de plataformas de altíssimo desempenho.

Figura 23 – Primeiras posições da lista TOP500, Nov. 2025.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,340,000	1,809.00	2,821.10	29,685
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	1,000.00	1,226.28	15,794
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	

Fonte: TOP500 (2025b).

Dessa forma, a modularidade não deve ser compreendida apenas como uma alternativa produtiva ao modelo monolítico, mas como uma estratégia arquitetural para ampliar a escalabilidade dos sistemas. Ao combinar múltiplos *dies*, diferentes nós de fabricação, memória de alta largura de banda e interconexões avançadas, o MCM permite construir processadores e aceleradores mais adequados às demandas atuais de HPC, inteligência artificial e *datacenters*. Assim, a evolução dos sistemas analisados indica que a modularidade tende a permanecer como um dos principais caminhos para o desenvolvimento de plataformas computacionais de alto desempenho.

5 CONSIDERAÇÕES FINAIS

Este trabalho analisou a tecnologia *Multi-Chip Module* (MCM) e a sua aplicação como uma alternativa às limitações encontradas no desenvolvimento de *chips* monolíticos cada vez maiores e mais complexos. A partir da pesquisa realizada, foi possível observar que a evolução dos processadores modernos não depende apenas da redução do tamanho dos transistores, mas também de novas formas de organizar, integrar e conectar diferentes componentes dentro de um mesmo encapsulamento.

Ao longo do estudo, ficou evidente que o MCM, os *chiplets* e as tecnologias de encapsulamento avançado oferecem maior flexibilidade para o desenvolvimento de processadores e aceleradores. A possibilidade de distribuir funções entre múltiplos *dies*, utilizar diferentes processos de fabricação e integrar blocos especializados permite ampliar a escalabilidade, melhorar a viabilidade produtiva e atender melhor às demandas de desempenho dos sistemas atuais.

Os exemplos analisados demonstram que essa abordagem já está presente em diferentes segmentos da indústria, como CPUs para servidores, aceleradores para HPC e inteligência artificial, SoCs de alto desempenho e supercomputadores. Dessa forma, o MCM não deve ser entendido apenas como uma solução de fabricação, mas também como uma estratégia arquitetural importante para lidar com cargas de trabalho que exigem grande paralelismo, alta largura de banda de memória e comunicação eficiente entre componentes.

A Tabela 2 sintetiza os principais exemplos discutidos ao longo do trabalho, destacando a forma como cada arquitetura aplica a modularidade e quais contribuições apresenta para a evolução dos processadores modernos. Essa comparação também permite observar que, embora cada solução possua objetivos e implementações distintas, todas compartilham a mesma tendência de superar as limitações do modelo monolítico por meio da integração de múltiplos blocos especializados.

Tabela 2 – Síntese comparativa das arquiteturas analisadas.

Abordagem	Organização Arquitetural	Principal contribuição	Ponto de atenção
Chip monolítico tradicional	Integra todas as funções em um único *die*	Pode oferecer comunicação interna direta e projeto mais integrado	Torna-se menos viável em chips grandes devido ao custo, área do *die*, *yield rate* e limites de fabricação
AMD EPYC 7001	Quatro *dies Zeppelin* integrados em MCM	Demonstra o uso de *dies* menores para amplicar a contagem de núcleos e reduzir riscos produtivos	Introduz desafios de latência e acesso à memória em modelo NUMA
AMD EPYC 7002 e 7003	Separação entre CCDs de computação e *I/O die* central	Representa maior maturidade de integração do MCM, com especialização funcional e melhor escalabilidade e encapsulamento 3D	Depende de interconexões eficientes entre os módulos e de organização interna mais complexa
Intel *Sapphire Rapids* e *Ponte Vecchio*	Uso de múltiplos *tiles*, EMIB, HBM e integração avançada	Mostra a adoção da modularidade também em CPUs e aceleradores Intel para *datacenters* e HPC	Aumenta a complexidade de projeto, encapsulamento e comunicação entre componentes
Apple *M1 Ultra*	Integração de dois *dies M1 Max* por interconexão de alta largura de banda	Demonstra que *chiplets* também podem escalar SoCs mantendo uma experiência próxima à de um chip único	Exige interconexão muito eficiente para preservar a percepção de unidade do sistema
AMD *Instinct* MI250X, MI300X e MI300A	Integração modular de GPU, CPU, I/O e HBM e maceleradores para HPC e IA	Evidencia o MCM como estratégia para grande paralelismo, alta largura de banda e integração heterogênea	Depende de tecnologias avançadas de encapsulamento, memória e software para explorar todo o potencial

Fonte: Elaborado pelo autor.

Com isso, os objetivos propostos foram atendidos, pois o trabalho apresentou as motivações produtivas do MCM, descreveu diferentes formas de organização baseadas em *chiplets* e analisou aplicações práticas em sistemas modernos de alto desempenho. A pesquisa também mostrou que a modularidade passou a ter papel relevante na evolução das arquiteturas de processamento, especialmente diante dos limites de custo, área e rendimento produtivo associados ao modelo monolítico.

Como limitação, destaca-se que o tema ainda está em constante evolução e depende bastante de documentação técnica, *white papers*, apresentações de fabricantes e publicações especializadas. Além disso, nem todas as empresas divulgam o mesmo nível de detalhe sobre suas arquiteturas, o que dificulta comparações totalmente equilibradas entre diferentes soluções comerciais.

Conclui-se, portanto, que o MCM se tornou uma das principais direções adotadas pela indústria de semicondutores para continuar ampliando desempenho e escalabilidade. A tendência observada indica que processadores, aceleradores e plataformas computacionais futuras devem continuar explorando arquiteturas modulares, integração heterogênea, memórias de alta largura de banda e encapsulamento avançado como caminhos para atender às demandas crescentes de HPC, inteligência artificial, *datacenters* e aplicações científicas.

REFERÊNCIAS

ADVANCED MICRO DEVICES – AMD. **AMD Accelerates Exascale Computing to New Heights Powering the Fastest Supercomputer Ever, El Capitan.** 2024. Disponível em: https://www.amd.com/en/newsroom/press-releases/2024-11-18-amd-accelerates-exascale-computing-to-new-heights-.html. Acesso em: 24 jan. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD CDNA™ 2 Architecture.** [S.l.]: AMD, 2021. Disponível em: https://www.amd.com/content/dam/amd/en/documents/instinct-business-docs/white-papers/amd-cdna2-white-paper.pdf. Acesso em: 12 dez. 2025.

ADVANCED MICRO DEVICES – AMD. **AMD Delivers Leadership Portfolio of Data Center AI Solutions with AMD Instinct MI300 Series.** 6 dez. 2023. Disponível em: https://www.amd.com/en/newsroom/press-releases/2023-12-6-amd-delivers-leadership-portfolio-of-data-center-a.html. Acesso em: 26 jan. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD EPYC™ 7002 Series Processors: A New Standard for the Modern Data Center.** abr. 2020. Disponível em: https://www.amd.com/content/dam/amd/en/documents/products/epyc/amd-epyc-7002-series-datasheet.pdf. Acesso em: 15 mar. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD EPYC 7003 Series Processors Microarchitecture Overview.** Rev. 3.0. 21 mar. 2022. Disponível em: https://docs.amd.com/v/u/en-US/overview-amd-epyc7003-series-processors-microarchitecture. Acesso em: 18 dez. 2025.

ADVANCED MICRO DEVICES – AMD. **AMD EPYC™ Datacenter Processor Launches with Record-Setting Performance, Optimized Platforms, and Global Server Ecosystem Support.** 20 jun. 2017. Disponível em: https://ir.amd.com/news-events/press-releases/detail/773/amd-epyc-datacenter-processor-launches-with-record-setting-performance-optimized-platforms-and-global-server-ecosystem-support. Acesso em: 21 mar. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD Instinct MI300 Series Accelerators.** [s.d.]. Disponível em: https://www.amd.com/en/products/accelerators/instinct/mi300.html. Acesso em: 2 fev. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD Instinct™ MI300 Series microarchitecture.** ROCm Documentation, 2026. Disponível em: https://rocm.docs.amd.com/en/latest/conceptual/gpu-arch/mi300.html. Acesso em: 28 mar. 2026.

ADVANCED MICRO DEVICES – AMD. **AMD Instinct™ MI300A APU: data sheet.** [S.l.]: AMD, 2025. Disponível em: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300a-data-sheet.pdf. Acesso em: 24 jan. 2026.

APPLE. **Apple unveils M1 Ultra, the world's most powerful chip for a personal computer.** 8 mar. 2022. Disponível em: https://www.apple.com/newsroom/2022/03/apple-unveils-m1-ultra-the-worlds-most-powerful-chip-for-a-personal-computer/. Acesso em: 14 fev. 2026.

ASML. **Lenses & mirrors.** [s.d.]. Disponível em: https://www.asml.com/en/technology/lithography-principles/lenses-and-mirrors. Acesso em: 2 maio 2026.

BAPTISTA, Eduardo; PAN, Che. **Huawei bets on speed over shrinking transistors to sidestep US chip sanctions**. *Reuters*, 29 maio 2026. Disponível em: https://www.reuters.com/world/china/huawei-bets-speed-over-shrinking-transistors-sidestep-us-chip-sanctions-2026-05-29/. Acesso em: 31 maio 2026.

BISWAS, Arijit. **Sapphire Rapids.** Hot Chips 33, 2021. Disponível em: https://hc33.hotchips.org/assets/program/conference/day1/HC2021.C1.4%20Intel%20Arijit.pdf. Acesso em: 12 fev. 2026.

BLYTHE, David. **Ponte Vecchio.** Hot Chips 33, 2021. Disponível em: https://hc33.hotchips.org/assets/program/conference/day2/hc2021_pvc_final.pdf. Acesso em: 12 fev. 2026.

GOMES, Wilfred et al. **Ponte Vecchio: A Multi-Tile 3D Stacked Processor for Exascale Computing.** HPC User Forum, mar. 2022. Disponível em: https://www.hpcuserforum.com/wp-content/uploads/2021/05/Gomes_Intel_Ponte-Vecchio_Mar2022-HPC-UF.pdf. Acesso em: 16 fev. 2026.

HAN, Yinhe *et al*. The Big Chip: Challenge, model and architecture. **Fundamental Research**, v. 4, p. 1431-1441, 2024. DOI: 10.1016/j.fmre.2023.10.020. Disponível em: https://doi.org/10.1016/j.fmre.2023.10.020. Acesso em: 25 maio 2026.

HIRVIJOKI, Eero et al. **Monte Carlo method and High Performance Computing for solving Fokker-Planck equation of minority plasma particles.** arXiv:1510.06221, 2015. Disponível em: https://arxiv.org/abs/1510.06221. Acesso em: 18 maio 2026.

HUAWEI. **Huawei présente la loi d'échelle Tau (τ), qui ouvre la voie à des avancées majeures en matière de densité de transistors et de performances des systèmes**. 25 maio 2026. Disponível em: https://www.huawei.com/fr/news/2026/5/ieee-iscas-tau-scaling. Acesso em: 31 maio 2026.

IMEC. **3D integration.** Leuven: IMEC, [s.d.]a. Disponível em: https://www.imec-int.com/en/expertise/cmos-advanced/connect/3d-integration. Acesso em: 4 maio 2026.

IMEC. **Lithography.** Leuven: IMEC, [s.d.]b. Disponível em: https://www.imec-int.com/en/semiconductor-education-and-workforce-development/microchips/how-are-microchips-made/lithography. Acesso em: 2 maio 2026.

INTEL. **4th Gen Intel® Xeon® Scalable Processors.** [s.d.]a. Disponível em: https://www.intel.com/content/www/us/en/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors-product-brief.html. Acesso em: 13 abr. 2026.

INTEL. **4th Gen Intel Xeon Sprints into the Market.** Intel Newsroom, 14 mar. 2023. Disponível em: https://newsroom.intel.com/data-center/4th-gen-intel-xeon-sprints-into-market. Acesso em: 13 abr. 2026.

INTEL. **Inovações de encapsulamento avançadas: pacotes de chips.** [s.d.]b. Disponível em: https://www.intel.com.br/content/www/br/pt/foundry/packaging.html. Acesso em: 15 mar. 2026.

KALPAKJIAN, Serope; SCHMID, Steven R. **Manufacturing Engineering and Technology.** 7. ed. 2014. Cap. 28. Disponível em: https://www.pearsoneducacion.net/docs/librariesprovider5/files_recursosmcc/kalpakjian_manufactura_ingenieria_tecnologia_7e_cap28.pdf. Acesso em: 22 maio 2026.

KHOPE, Namrata P.; KSHIRSAGAR, Ujwala A. **Yield Analysis in Semiconductor Manufacturing: Techniques, Case Studies, and Future Direction.** Journal of Information Systems Engineering and Management, v.10, n. 11s, 2025. Disponível em: https://www.jisem-journal.com/index.php/journal/issue/view/13. Acesso em: 17 maio 2026.

LARABEL, Michael. **An Extensive Look At The AMD Naples vs. Rome Power Efficiency / Performance-Per-Watt.** Phoronix, 5 dez. 2019. Disponível em: https://www.phoronix.com/review/epyc-rome-power. Acesso em: 15 mar. 2026.

LEACHMAN, Robert C. **Yield Modeling and Analysis.** Berkeley: University of California, 2014. Notas de aula da disciplina IEOR 130 – Methods of Manufacturing Improvement. Disponível em: https://fog.misty.com/perry/cod/references/yield_models.pdf. Acesso em: 12 maio 2026.

LI, Alison; TIMINGS, Jessica. **6 crucial steps in semiconductor manufacturing.** ASML, 2021. Atualizado em: out. 2023. Disponível em: https://www.asml.com/en/news/stories/2021/semiconductor-manufacturing-process-steps. Acesso em: 29 abr. 2026.

MADPCB. **MCM: Multichip Module (MCM) or Package.** [s.d.]. Disponível em: https://madpcb.com/glossary/mcm/. Acesso em: 26 maio 2026.

MALLOY, Frank. **What is a Silicon Die?** Synopsys, 22 ago. 2025. Disponível em: https://www.synopsys.com/glossary/what-is-silicon-die.html. Acesso em: 24 maio 2026.

MATSUSADA PRECISION. **Fundamentals of Ion Implantation: Processes and Equipment**. Tech Tips, 28 jul. 2025. Atualizado em: 28 abr. 2026. Disponível em: https://www.matsusada.com/column/ion-implantation.html. Acesso em: 27 maio 2026.

MKS INSTRUMENTS. **CVD Physics.** [s.d.]. Disponível em: https://www.mks.com/n/cvd-physics. Acesso em: 30 maio 2026.

MOYER, Bryon. **Chiplet Fundamentals For Engineers: eBook.** [S.l.]: Semiconductor Engineering, 2026. Disponível em: https://semiengineering.com/chiplet-fundamentals-for-engineers-ebook/. Acesso em: 22 maio 2026.

MULNIX, David L. **Intel® Xeon® Processor Scalable Family Technical Overview.** Intel, 1 dez. 2022a. Disponível em: https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html. Acesso em: 14 abr. 2026.

MULNIX, David L. **Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family.** Intel, 25 jul. 2022b. Disponível em: https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html. Acesso em: 13 abr. 2026.

OAK RIDGE NATIONAL LABORATORY – ORNL. **ExaSMR: Coupled Monte Carlo Neutronics and Fluid Flow Simulation of Small Modular Reactors.** [s.d.]. Disponível em: https://www.ornl.gov/project/exasmr-coupled-monte-carlo-neutronics-and-fluid-flow-simulation-small-modular-reactors. Acesso em: 28 maio 2026.

SAMSUNG SEMICONDUCTOR. **Photoresist.** [s.d.]a. Disponível em: https://semiconductor.samsung.com/support/tools-resources/dictionary/photoresist-printing-the-circuit-onto-the-wafer/. Acesso em: 26 maio 2026.

SAMSUNG SEMICONDUCTOR. **Yield.** [s.d.]b. Disponível em: https://semiconductor.samsung.com/support/tools-resources/dictionary/semiconductor-glossary-yield/. Acesso em: 25 maio 2026.

SHIPMAN, Galen M. et al. **Early Performance Results on 4th Gen Intel® Xeon® Scalable Processors with DDR and Intel® Xeon® processors, codenamed Sapphire Rapids with HBM.** arXiv:2211.05712, 2022. DOI: 10.48550/arXiv.2211.05712. Disponível em: https://arxiv.org/abs/2211.05712. Acesso em: 29 maio 2026.

TANDON, Suyash et al. **Porting HPC Applications to AMD Instinct MI300A Using Unified Memory and OpenMP.** arXiv, 2024. DOI: 10.48550/arXiv.2405.00436. Disponível em: https://arxiv.org/abs/2405.00436. Acesso em: 24 jan. 2026.

TIRIAS RESEARCH. **AMD Optimizes EPYC Memory with NUMA.** [S.l.]: AMD; Tirias Research, mar. 2018. Disponível em: https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/AMD-Optimizes-EPYC-Memory-With-NUMA.pdf. Acesso em: 15 mar. 2026.

TIRIAS RESEARCH. **The 2nd Generation AMD EPYC Processor Redefines Data Center Economics.** [S.l.]: AMD; Tirias Research, ago. 2019. Disponível em: https://tiriasresearch.com/wp-content/uploads/2020/04/TIRIAS_Research-The_2nd_Gerneration_AMD_EPYC_Redefines-_Data_Center_Economics.pdf. Acesso em: 15 mar. 2026.

TOP500. **El Capitan Retains #1 as JUPITER Becomes Europe's First Exascale System in the 66th TOP500 List.** 2025a. Disponível em: https://top500.org/news/el-capitan-retains-1-as-jupiter-becomes-europes-first-exascale-system-in-the-66th-top500-list/. Acesso em: 31 maio 2026.

TOP500. **November 2025.** nov. 2025b. Disponível em: https://top500.org/lists/top500/2025/11/. Acesso em: 31 maio 2026.

TSMC. **TSMC 3DFabric® for High-Performance Computing.** [s.d.]. Disponível em: https://www.tsmc.com/english/dedicatedFoundry/technology/platform_HPC_tech_WLSI. Acesso em: 16 mar. 2026.

WARD, Daniel R. et al. Atomic Precision Advanced Manufacturing for Digital Electronics. **Electronic Device Failure Analysis**, v. 22, n. 1, p. 4-10, 2020. DOI: 10.31399/asm.edfa.2020-1.p004. Disponível em: https://arxiv.org/abs/2002.11003. Acesso em: 9 mar. 2026.