

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

NVIDIA Corporation,
Petitioner,

v.

Onesta IP, LLC,
Patent Owner.

Case No. IPR2026-00011

**PETITION FOR *INTER PARTES* REVIEW
OF CLAIMS 1-4 AND 7-10 OF
U.S. PATENT NO. 11,841,803
UNDER 35 U.S.C. §312 AND 37 C.F.R. §42.104**

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
II.	MANDATORY NOTICES.....	4
	A. Real Party-in-Interest.....	4
	B. Related Matters	4
	C. Counsel.....	5
	D. Service Information	5
III.	CERTIFICATION OF GROUNDS FOR STANDING.....	5
IV.	OVERVIEW OF THE CHALLENGE AND RELIEF REQUESTED	6
	A. Grounds	6
	B. Requested Relief.....	6
	C. Prior Art.....	6
V.	'803 PATENT OVERVIEW.....	7
	A. Technology Background	7
	B. The '803 Patent's Alleged Problem and Purported Solution.....	8
	C. Prosecution History	12
	D. Claims.....	13
VI.	PRIOR ART OVERVIEW.....	14
	A. Kelleher	14
	B. Collins	18
	C. Koker.....	21
	D. Sistla.....	23
VII.	LEVEL OF ORDINARY SKILL IN THE ART.....	24
VIII.	CLAIM CONSTRUCTION.....	25
	A. "Chiplet".....	25
	B. "Dedicated for inter-chiplet communications".....	26
IX.	INVALIDITY GROUNDS	29
	A. Ground I: Claims 1-4, 7, and 10 Are Obvious Over Kelleher and Collins.....	29
	1. Claim 1.....	29

2.	Claim 2.....	53
3.	Claim 3.....	54
4.	Claim 4.....	58
5.	Claim 7.....	61
6.	Claim 10.....	61
B.	Ground II: Claims 8 and 9 Are Obvious Over Kelleher, Collins, Koker, and Sistla	67
1.	Claim 8.....	67
2.	Claim 9.....	98
X.	CONCLUSION.....	103
	TABLE OF CLAIMS.....	I
	TABLE OF EXHIBITS FOR U.S. PATENT 11,841,803 PETITION FOR <i>INTER PARTES</i> REVIEW	V
	CERTIFICATE OF SERVICE	1
	WORD COUNT CERTIFICATE OF COMPLIANCE	1

NVIDIA requests IPR of claims 1-4 and 7-10 of USP11,841,803 (“’803 patent”)(EX1001-’803).

I. INTRODUCTION

Patent Owner, Onesta, acquired the ’803 patent from AMD in November 2024, less than a year after the patent issued and shortly after Onesta was formed. Its only business since then appears to be a campaign to monetize its acquired IP, including initiating patent litigation.

Onesta is asserting the ’803 patent against NVIDIA in an ITC investigation and in a now-stayed case in the Western District of Texas. NVIDIA produces advanced graphics processing units (“GPUs”)—the most advanced GPUs in the world—that are used by this country’s most innovative companies for developing artificial intelligence (“AI”) technology, data centers, and high-performance computing applications. In the ITC, Onesta seeks to use the ’803 patent to obtain an order excluding NVIDIA’s advanced GPUs from the U.S.

But there is good reason that, of its 13,000 patents, AMD was willing to part with the ’803 patent shortly after issuance: the alleged invention was well-known years before the claimed June 2019 priority date.

The ’803 patent claims a system with (1) GPU “chiplets”—smaller, individual processing units that are combined to form a complete processor; (2) a “passive crosslink ... dedicated for inter-chiplet communications” connecting the chiplets; and (3) a CPU coupled to one of the GPU chiplets via a bus. The patent does not

claim to have invented chiplets, CPUs, or connections between components. Indeed, the patent acknowledges that “[c]hiplets have been used successfully in CPU architectures to reduce cost of manufacture and improve yields.” EX1001-’803, 1:46-51. Instead, the ’803 patent asserts that using chiplets in GPU architectures allegedly presented challenges, because GPUs rely on highly parallel workflows—demanding synchronized operations across multiple GPU chiplets—and because CPUs are traditionally programmed to address a single GPU, rather than an array of GPU chiplets. EX1001-’803, 1:52-67. The ’803 patent purports to solve those alleged challenges by connecting the CPU over a bus to a single chiplet, thus preserving the monolithic GPU programming model for the CPU, and then connecting all the GPU chiplets via a passive crosslink. But, as discussed below, such systems were known **14 years** before the ’803 patent’s filing date.

First, Kelleher (USP7,598,958)(EX1004-Kelleher)—an NVIDIA patent filed over 14 years prior to the ’803 patent’s June 2009 filing date—discloses a system, like the ’803 patent, with (1) multiple GPU chiplets that the CPU views as a single (monolithic) GPU; (2) a crosslink for inter-chiplet communications; and (3) the CPU coupled to one GPU chiplet via a bus. **Second**, although Kelleher does not explicitly disclose whether its crosslink is “passive” or “dedicated for inter-chiplet communications” (i.e., used for transmission of only chiplet-to-chiplet signals), Collins (USP11,610,862)(EX1005-Collins)—a prior art Intel patent—discloses a

dedicated passive crosslink that, when incorporated into Kelleher's system, is free of signaling pathways linking the chiplets to other components, such as conductive pathways known as "through silicon vias" ("TSVs") that link the chiplets to power, ground, etc. **Third**, certain dependent claims of the '803 patent require some "cache" memory to be accessible by a specific GPU chiplet and other cache to be accessible by all GPU chiplets. To the extent these claims are not obvious over Kelleher and Collins, they would have been obvious based on a combination of those references with Koker (USP10,909,652)(EX1031-Koker) and Sistla (USPN2009/0037658)(EX1047-Sistla), which teach the claimed cache memory in a chiplet system.

Neither Kelleher, Collins, Koker, nor Sistla was considered by the USPTO during prosecution of the '803 patent. As explained below, a POSITA would have been motivated to combine Kelleher's, Collins's, Koker's, and Sistla's teachings to obtain the '803 patent claims and would have had a reasonable expectation of success in doing so. All four are directed to modular design of GPU or CPU systems, and Collins, Koker, and Sistla provide routine implementation details (physical layout, memory architecture, etc.) that a POSITA would have been motivated to use to maintain performance and memory coherency in Kelleher's chiplet-based system. Accordingly, the challenged claims are invalid.

IPR was intended to eliminate patents, like the '803 patent, that never should

have issued. Onesta will no doubt argue for discretionary denial based on the pending ITC case, but Onesta also asserts the '803 patent in its stayed WDTX case. Accordingly, the ITC case, which is not dispositive of the invalidity issues in the WDTX case, will not resolve all issues concerning the '803 patent. Moreover, the broad remedy Onesta seeks in the ITC—excluding NVIDIA's advanced GPUs from the U.S.—would set the U.S. back in the global race for AI leadership, disadvantage U.S. companies relative to their foreign competitors, risk U.S. national security, and fundamentally harm U.S. innovation. Indeed, NVIDIA GPUs are so critical to the U.S.'s national interest that they are subject to strict export control restrictions, yet Onesta seeks to impose that same restriction on all *U.S.* companies. Onesta should not be permitted to use a plainly invalid patent to hold the U.S. economy hostage.

II. MANDATORY NOTICES

A. Real Party-in-Interest

The real party-in-interest and Petitioner for this Petition is NVIDIA Corporation.

B. Related Matters

Pursuant to 37 C.F.R. §42.8(b)(2), NVIDIA states that the '803 patent has been asserted or challenged in the following cases:

- *Onesta IP, LLC v. NVIDIA Corp.*, No. 1:25-cv-00587 (W.D. Tex.)
(EX1024-District-Court-Complaint);

- *Certain Integrated Circuits, Electronic Devices Containing the Same, and Components Thereof*; Inv. No. 337-TA-1450 (ITC) (EX1023-ITC-Complaint; EX1025-ITC-Ex50-Infringement; EX1026-ITC-Ex.19-Domestic-Industry; EX1043-Claim-Construction-Chart).

C. Counsel

Lead Counsel: Grant K. Rowan (Reg. No. 41,278)

Backup Counsel: Michael J. Summersgill (*pro hac vice* to be filed)

Amanda L. Major (*pro hac vice* to be filed)

Harry D. Hanson (*pro hac vice* to be filed)

Sarah J. Murphy (Reg. No. 80,104)

D. Service Information

E-mail: WH-Nvidia-IPRService@wilmerhale.com

Post and hand delivery: Wilmer Cutler Pickering Hale and Dorr LLP

2100 Pennsylvania Avenue NW

Washington, DC 20037

Telephone: (202) 663-6000

Fax: (202) 663-6363

NVIDIA consents to service via email at the email addresses above.

III. CERTIFICATION OF GROUNDS FOR STANDING

The '803 patent is available for IPR and NVIDIA is not estopped from requesting IPR for the challenged claims.

IV. OVERVIEW OF THE CHALLENGE AND RELIEF REQUESTED

A. Grounds

NVIDIA challenges claims 1-4 and 7-10 under 35 U.S.C. §103 as noted below:

Ground	Claims	References
I	1-4, 7, 10	Kelleher/Collins
II	8-9	Kelleher/Collins/Koker/Sistla

B. Requested Relief

NVIDIA requests cancelling the challenged claims. This Petition, supported by Dr. Subramanian's declaration (EX1002), demonstrates a reasonable likelihood that NVIDIA will prevail on the challenged claims. 35 U.S.C. §314(a).

Dr. Subramanian is a Professor of Microtechnology at EPFL in Switzerland and leads the Laboratory for Advanced Fabrication Technologies. He has a Ph.D. in electrical engineering from Stanford University and was a Professor at U.C. Berkeley before joining EPFL. His research has focused on design, fabrication, characterization, and modeling of semiconductor devices/integrated circuits, including chiplet-based architectures. EX1002-Decl., ¶¶3-13.

C. Prior Art

NVIDIA relies on:

1. USP 7,598,958 ("Kelleher")(EX1004-Kelleher)

2. USP 11,610,862 (“Collins”)(EX1005-Collins)
3. USP 10,909,652 (“Koker”)(EX1031-Koker)
4. USPN 2009/0037658 (“Sistla”)(EX1047-Sistla)

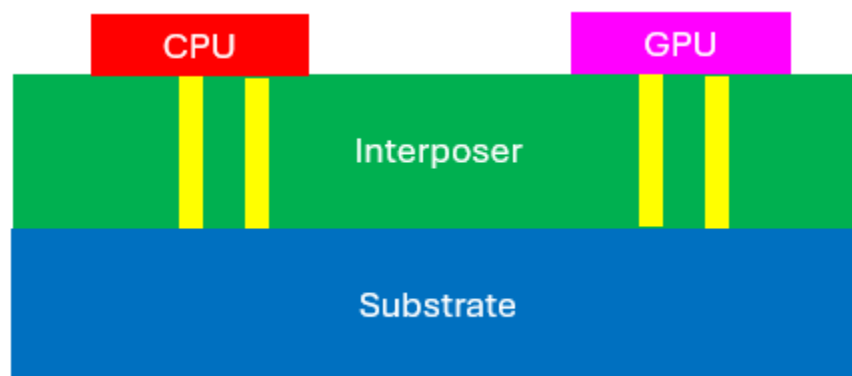
V. ’803 PATENT OVERVIEW

A. Technology Background

A central processing unit (“CPU”) is the “brains” of a computer. A CPU executes software instructions, performs calculations, and handles the overall function of a computer system. EX1002-Decl., ¶36.

A graphics processing unit (“GPU”) efficiently processes data-intensive tasks in parallel. GPUs are used for AI, parallel processing, and high-performance computing applications. GPUs were well-known since at least 1999, when NVIDIA launched the first commercial GPU. EX1002-Decl., ¶37.

Many types of connections between semiconductor components also have been well-known. For example, buses, bridges, and links were widely used long before the ’803 patent’s filing date. As the patent acknowledges, one known technique mounted multiple chips (“dies”) side-by-side on an “interposer,” as shown below. EX1001-’803, 1:6-16. EX1002-Decl., ¶38, ¶57



An “interposer” is layered between (1) a CPU and GPU and (2) a “substrate” and contains conductive pathways (not shown) to enable the CPU and GPU to exchange information. The interposer contains vertical conductors (“vias”) for communicating power and other signals between the CPU or GPU and substrate. Because the interposer is typically made of silicon and the vias extend through the interposer, these vias are “through silicon vias” (“TSVs”). EX1002-Decl., ¶¶36-40.

B. The ’803 Patent’s Alleged Problem and Purported Solution

According to the ’803 patent, “[c]omputing devices ... require[] the integration of more performance and features into increasingly smaller spaces,” but “[c]onventional monolithic die designs are becoming increasingly expensive to manufacture.” EX1001-’803, 1:6-9, 1:45-46. Before the ’803 patent, one well-known way to address this purported problem was using chiplets—smaller, processing units that are combined to form a complete processor. Rather than manufacturing one, large “monolithic” chip, manufacturers created multiple smaller chiplets that are connected together. Chiplets—also called “multi-chip modules”—

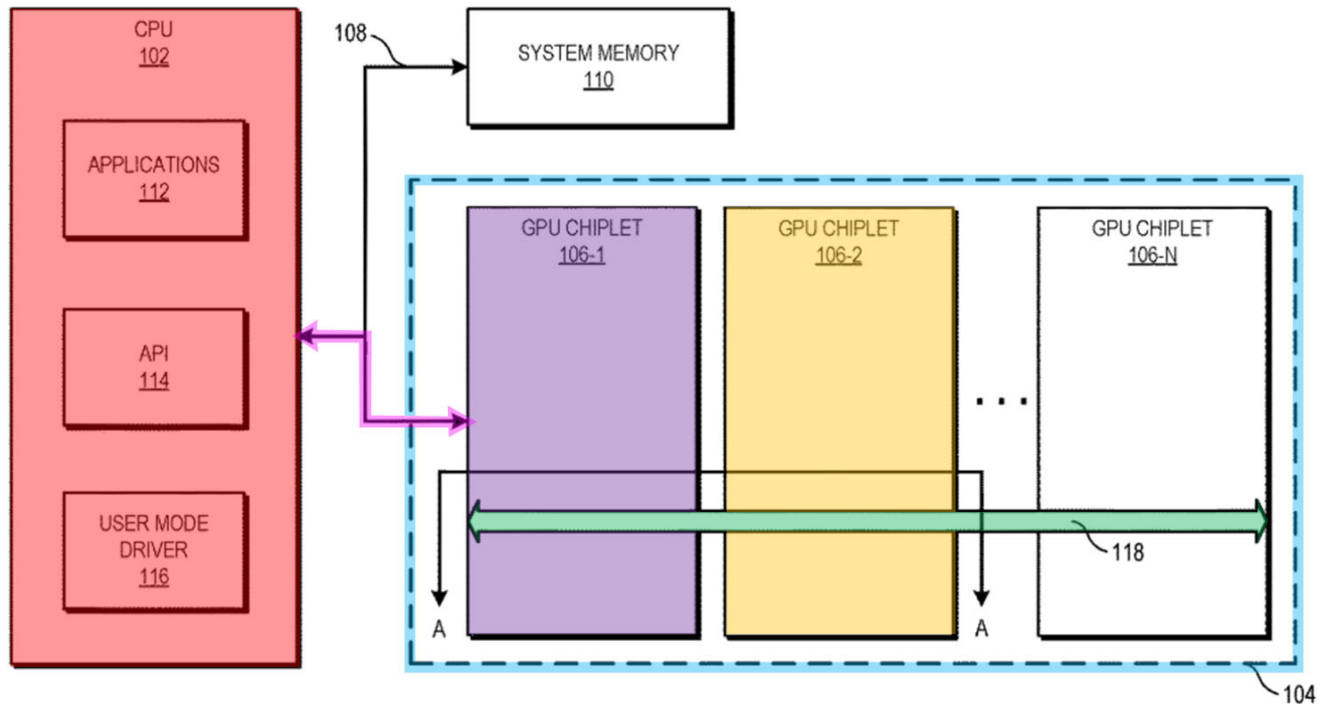
reduce manufacturing complexity and cost, because defects in one chiplet are addressed by discarding just that chiplet rather than a larger, monolithic chip. EX1002-Decl., ¶¶41-55; ¶61.

The '803 patent acknowledges that “[c]hiplets have been used successfully in *CPU* architectures to reduce cost of manufacture and improve yields.” EX1001-'803, 1:46-51. But the patent asserts that using chiplets in *GPU* architectures presented challenges. The patent states that CPUs are programmed to address a single GPU, and thus, programming CPUs to interact with multiple GPU chiplets is challenging. EX1001-'803, 1:62-67. EX1002-Decl., ¶62.

The '803 patent alleges that the system in Figure 1 (below)¹ overcomes the above-described challenges because it “improve[s] system performance by using GPU chiplets while preserving the current [CPU] programming model”:

¹ All annotations added.

100



EX1001-'803, 2:1-2, Figure 1. EX1002-Decl., ¶63.

First, the system requires multiple GPU “chiplets” (purple, gold, white) instead of a one larger, single GPU. The specification defines “chiplet” as:

- 1) a chiplet includes an active silicon die containing part of the computational logic used to solve a full problem (i.e., the computational workload is distributed across multiples of these active silicon dies);
- 2) chiplets are packaged together as a monolithic unit on the same substrate; and
- 3) the programming model preserves the concept that these separate computational dies are a single monolithic unit (i.e., each chiplet is not exposed as a separate device

to an application that uses the chiplets for processing computational workloads).

EX1001-'803, 2:41-53. EX1002-Decl., ¶64.

Second, the system uses a “passive crosslink” “dedicated for inter-chiplet communications” (i.e., transmitting signals among only GPU chiplets). EX1001-'803, Abstract, 2:11-14, 3:57-4:2, 4:16-21, Claim 1. EX1002-Decl., ¶65-68.

Third, the CPU communicates with *only one* GPU chiplet via a bus, and that chiplet relays information to other chiplets (e.g., chiplet) via the dedicated, passive crosslink. EX1001-'803, 3:42-49. Accordingly, the CPU purportedly views the separate GPU chiplets as a “single die, monolithic GPU.” EX1001-'803, 2:30-33, 3:49-56. EX1002-Decl., ¶¶69-70.

Figure 2 (below) shows a cross-section of the GPU chiplets and dedicated passive crosslink. EX1001-'803, 1:28-30, Figure 2. The patent characterizes the passive crosslink connecting the chiplet and chiplet as “dedicated” to inter-chiplet communications because it does not have pathways, such as TSVs (yellow), for transmitting signals between the chiplets and other components.

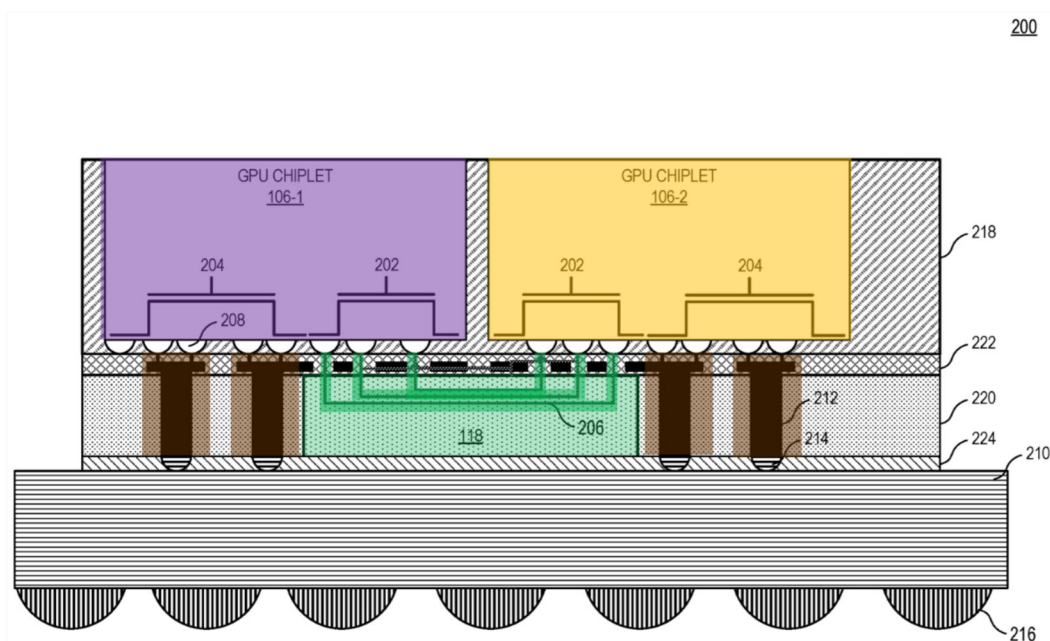


FIG. 2

EX1001-'803, Figure 2, 4:16-21 (“[T]he passive crosslink 118 does not contain any through silicon vias (TSVs)....”), 3:62-4:42 (chiplets’ physical device (“PHY”) region 202 “has ... conductor structures *dedicated to the transmission of chiplet-to-chiplet signals*, and a non-PHY region 204, which has conductor structures that are tailored more to the conveyance of power and ground and/or chiplet-to-circuit board signals”), 4:34-44 (power and I/O “routed around the interposer die (i.e., passive crosslink 118) ... thereby replacing ... conventional TSVs”). EX1002-Decl., ¶¶65-72.

C. Prosecution History

During prosecution, the Examiner relied on US 2019/0123022 (“Teig”) to reject '803 patent claim 1. EX1003-FH, 876-878, 848, 749, 643-648, 610-611, 557-

562. Teig stacked two dies (chiplets) on opposite sides of an interposer, which communicably coupled the dies to each other. One of the dies contained CPU and GPU cores, while the other contained cache memory for those cores. EX1027-Teig, [0148]. Based on this, Applicants argued that Teig does not disclose a “passive crosslink ... dedicated for inter-chiplet communication”:

[C]laim [1] requires that ‘the passive crosslink is dedicated for inter-chiplet communications’. In contrast, ... Teig merely states, ‘the first IC die is *face-to-face mounted with a passive interposer that electrically connects the die to* circuits outside of the 3D chip or *to other dies that are face-to-face mounted or back-to-face mounted on the interposer.*’ *Nowhere ... does Teig disclose that the passive interposer is dedicated for inter-chiplet communications.*

EX1003-FH, 847. The Examiner then allowed the claim. EX1003-FH, 6. EX1002-Decl., ¶¶73-84.

D. Claims

Representative Claim 1 recites:

1. A system, comprising:

a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes:

the first GPU chiplet communicably coupled to the CPU via a bus; and

a second GPU chiplet communicably coupled to the first GPU chiplet
via a passive crosslink,

wherein the passive crosslink is dedicated for inter-chiplet
communications.

EX1002-Decl., ¶71.

Dependent claims add requirements regarding the number of chiplets, the
arrangement of conductor structures, and memory hierarchy. As described herein,
these features were well-known before the '803 patent's priority date. EX1002-
Decl., ¶72

VI. PRIOR ART OVERVIEW

A. Kelleher

Kelleher was filed 14 years before the '803 patent's June 2019 priority date
and is prior art under AIA-35 U.S.C. §102(a)(1). EX1004-Kelleher, Cover.
EX1002-Decl., ¶85.

In Figure 2 (below), Kelleher discloses a system just like the '803 patent
system:

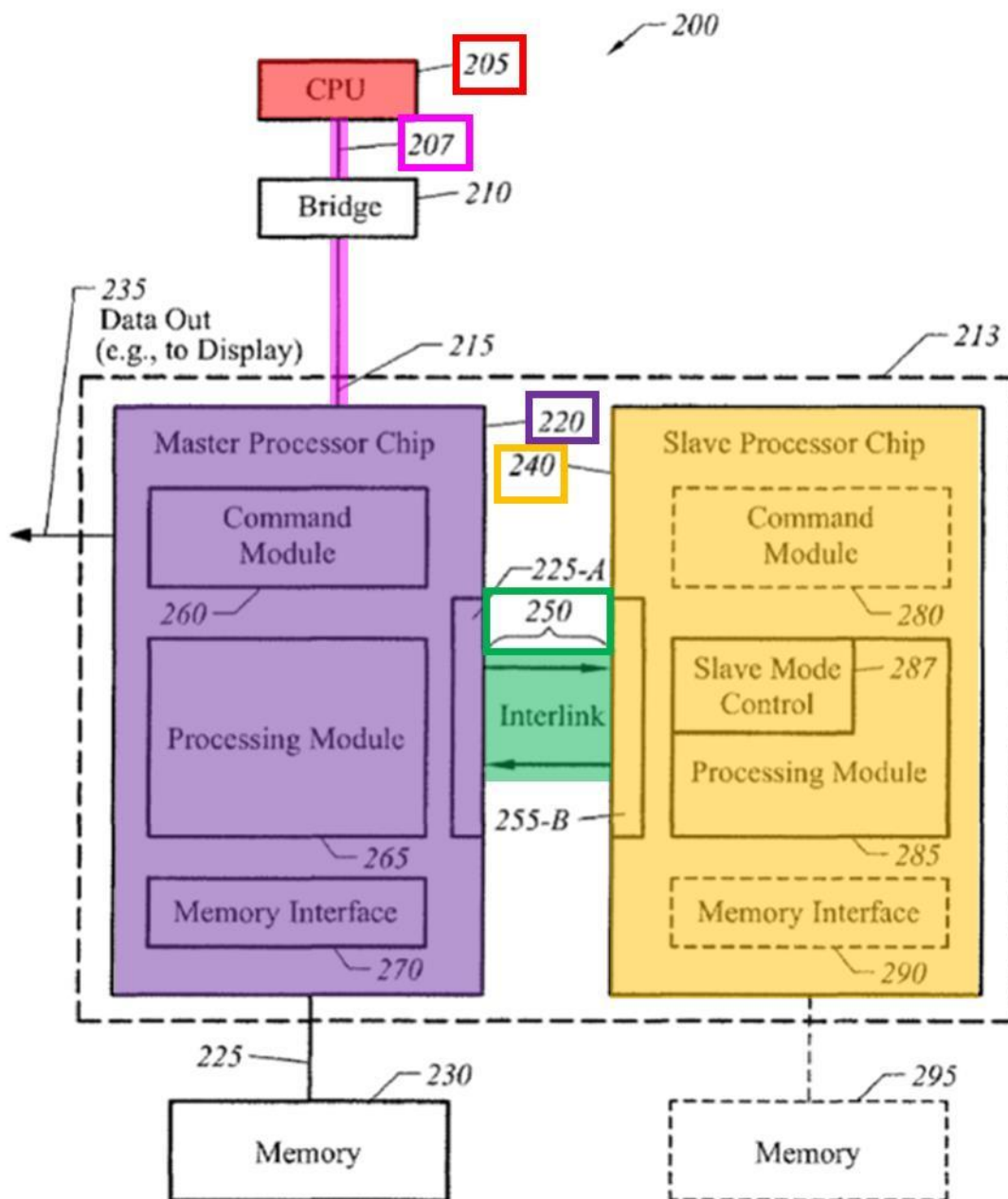


FIG. 2

EX1004-Kelleher, Figure 2. EX1002-Decl., ¶86.

First, like the '803 patent, Kelleher discloses a “master GPU” chip and a “slave GPU” chip that meet the '803 patent’s “chiplet” definition (EX1001-'803, 2:41-53):

- 1) Kelleher’s master chip and slave chip each includes an active silicon die (with a “processing module”) containing part of the computational logic used to solve a full problem (i.e., the computational workload is distributed across these dies). EX1004-Kelleher, 3:38-41 (“[In] GPU chip 220 is a processing module 265 that performs a processing operation....”), 3:51-53 (“Slave GPU chip 240 includes a processing module 285 capable of performing a processing function....”), 4:3-19 (“[I]f processing modules 265 and 285 are substantially identical, *about half of the processing that would ordinarily be performed by processing module 265 is off-loaded to processing module 285....*”);
- 2) Kelleher’s GPU chip and GPU chip are packaged together as a monolithic unit on the same substrate. EX1004-Kelleher, 3:46-50 (“Slave GPU chip 240 is *packaged with GPU chip 220* as a multi-chip module 213 *that may be used as a replacement for an individual GPU chip....* [M]ulti-chip module 213 may include a package ... mounted onto a board.”), Figure 3 (master/slave chips on same substrate); and

3) Kelleher's programming model preserves the concept that these separate computational dies are a single monolithic unit (i.e., each chiplet is not exposed as a separate device to applications using the chiplets for processing computational workloads). EX1004-Kelleher, 5:9-15 (“[O]nly the master GPU 220 receives graphics instructions from CPU 205. As result, replacing a single GPU chip with a multi-chip module having a master/slave pair of GPUs is a seamless change to driver software within CPU 205. *[T]he driver software [in the CPU] perceives the master/slave pair as a single GPU.*”).

EX1002-Decl., ¶¶87-93.

Second, like the '803 patent, Kelleher's system uses a crosslink—an interlink—for communicating between the master and slave chiplets. EX1004-Kelleher, Figure 2 (interlink connecting only master/slave GPU chips), 4:20-35 (“Slave GPU chip 240 is coupled to GPU chip 220 via an interlink 250.... [I]nterlink 250 facilitate[s] seamlessly offloading processing tasks to slave GPU chip 240 as if processing module 285 of slave GPU chip 240 was an internal extension of processing module 265 of GPU chip 220.”), 4:3; compare EX1001-'803, 6:55-62 (“the processing system 500 utilizes a *master-slave topology*” where one chiplet “is designated as the master chiplet” and “other GPU chiplets communicate with the CPU 102 indirectly via the passive crosslink 118 and are designated as the slave chiplets....”). EX1002-Decl., ¶¶94-96.

Third, like the '803 patent, Kelleher's CPU communicates with only the master chiplet via a bus and that chiplet relays information to the slave chiplet via the interlink. EX1004-Kelleher, Figure 2, 3:8-10 ("[CPU] 205 is coupled to ... [GPU] chip 220 through a suitable bus interface."), 5:9-12 ("[O]nly the master GPU 220 receives graphics instructions from CPU 205."), 4:20-45 ("Slave GPU chip 240 is coupled to GPU chip 220 via an interlink 250.... [I]nterlink 250 facilitate[s] seamlessly offloading processing tasks to slave GPU chip 240 as if processing module 285 of slave GPU chip 240 was an internal extension of processing module 265 of GPU chip 220."), 6:55-62 (quoted above). EX1002-Decl., ¶97.

Like the '803 patent, Kelleher's CPU considers the multiple chiplets to be a single "monolithic" GPU, thereby facilitating programming of the CPU. EX1004-Kelleher, 5:9-15 ("[O]nly the master GPU 220 receives graphics instructions from CPU 205. As result, replacing a single GPU chip with a multi-chip module having a master/slave pair of GPUs is a seamless change to driver software within CPU 205. [T]he driver software *perceives the master/slave pair as a single GPU.*"); compare EX1001-'803, 3:49-56. EX1002-Decl., ¶98.

B. Collins

Collins was filed nine months before the '803 patent's June 2019 priority date and is prior art under AIA-35 U.S.C. §102(a)(2). EX1005-Collins, Cover. EX1002-Decl., ¶99.

In Figure 1 (below), Collins discloses “passive” interconnects called
“**embedded multi-die interconnect bridges**” (“**EMIBs**”):

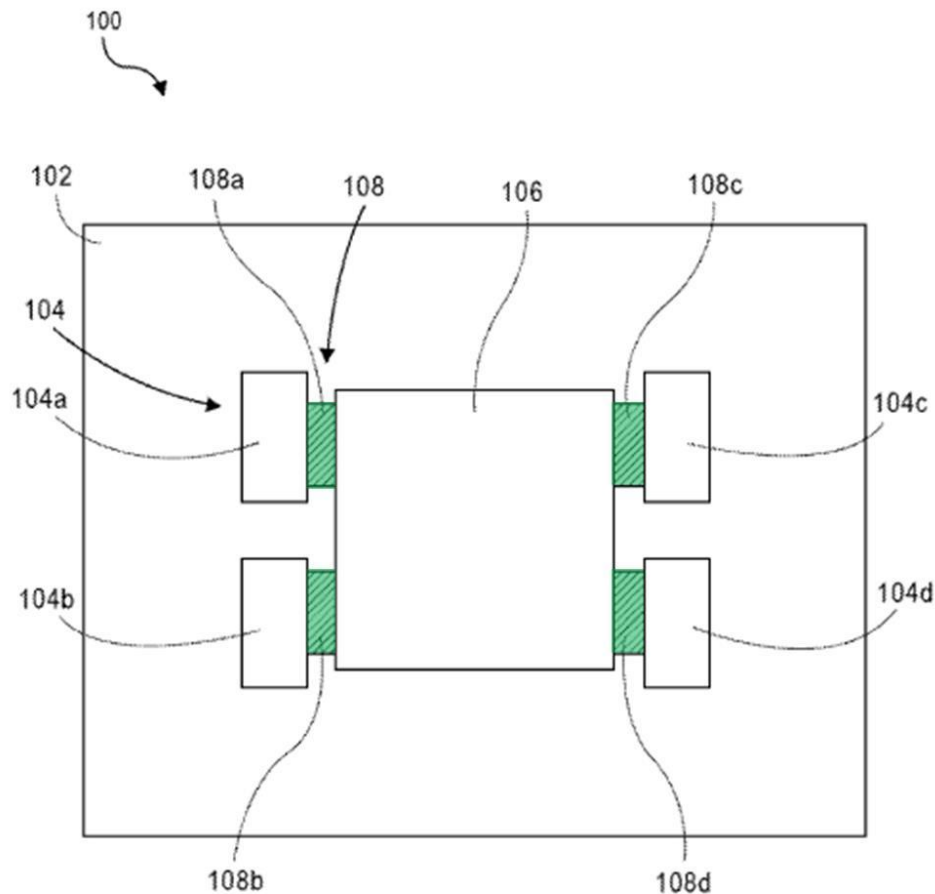


FIG. 1

EX1005-Collins, Figure 1, 3:50-57 (“[T]he interconnect structure 108 includes one or more embedded multi-die interconnect bridges (EMIBs).... Each of the EMIBs *include ... passive silicon with conductive elements ... for conducting signals.*”).

EX1002-Decl., ¶100.

While Figure 1 shows **EMIBs** connecting chiplets to memory (EX1005-Collins, 3:50-57), Collins discloses using the **passive interconnect** for chiplet-to-

chiplet communications. EX1005-Collins, 4:24-39 (“[T]he EMIBs 108 *may couple two or more of the chiplets 104 together*.... The chiplets 104 coupled ... via the EMIBs 108 may utilize the EMIBs 108 for communication between the chiplets 104....”), 16:24-27. Figure 2 shows Collins’s **interconnect** used in the chiplet-to-memory embodiment:

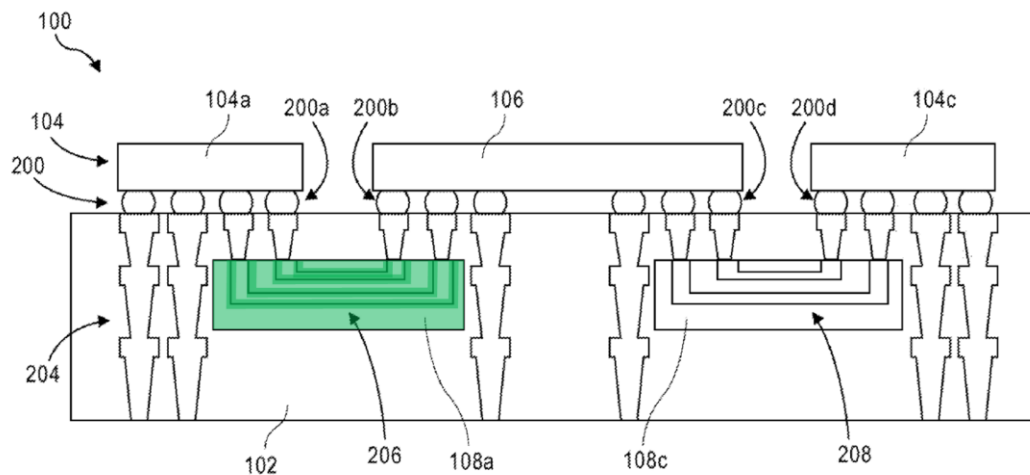
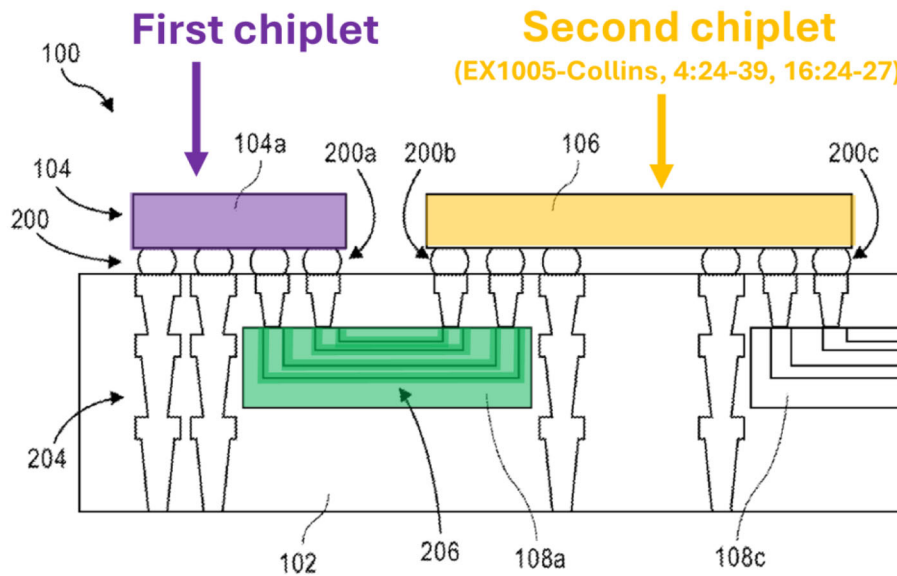


FIG. 2

EX1005-Collins, Figure 2. But as applied to the chiplet-to-chiplet embodiment (modified figure below), the **passive interconnect** connects multiple chiplets:



EX1005-Collins, Modified Figure 2 (replacing memory device with **second chiplet** (4:24-39, 16:24-27)). Here, the **EMIB** includes a **passive interposer** that is free of signaling pathways linking the chiplets to other components, and is thus dedicated for inter-chiplet communications. EX1002-Decl., ¶101-104.

C. Koker

Koker was filed three months before the '803 patent's priority date and is prior art under AIA-35 U.S.C. §102(a)(2). EX1031-Koker, Cover. EX1002-Decl., ¶105.

In Figure 3C (below), Koker teaches a "[GPU] 380 which includes ...

graphics processing resources arranged into multi-core groups [MCGs] 365A-365N” (e.g., MCG 356A and MCG 356B).

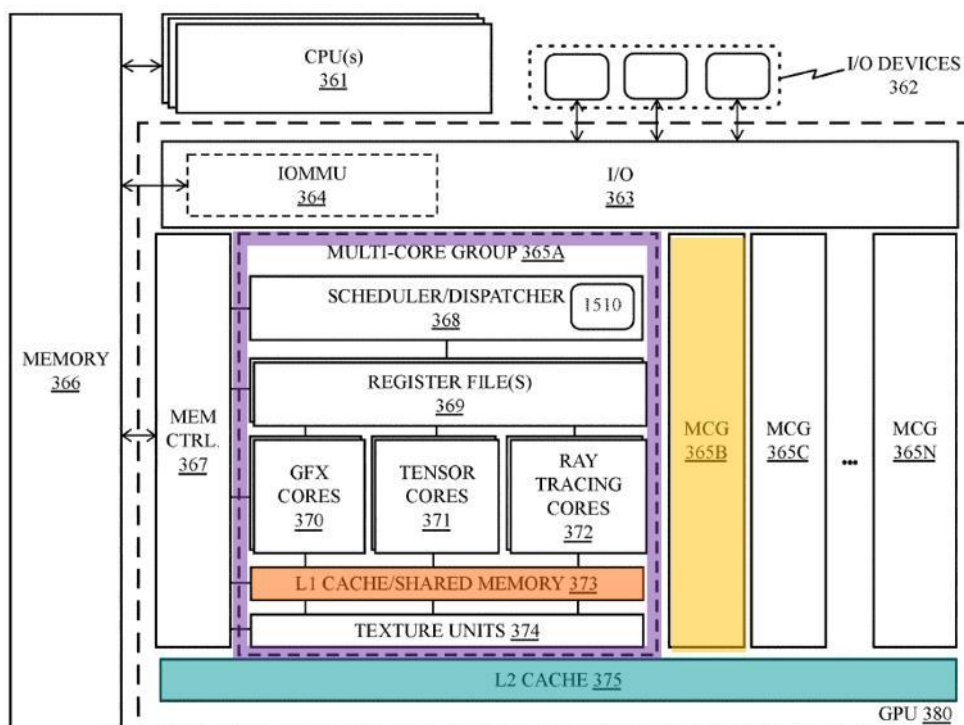


FIG. 3C

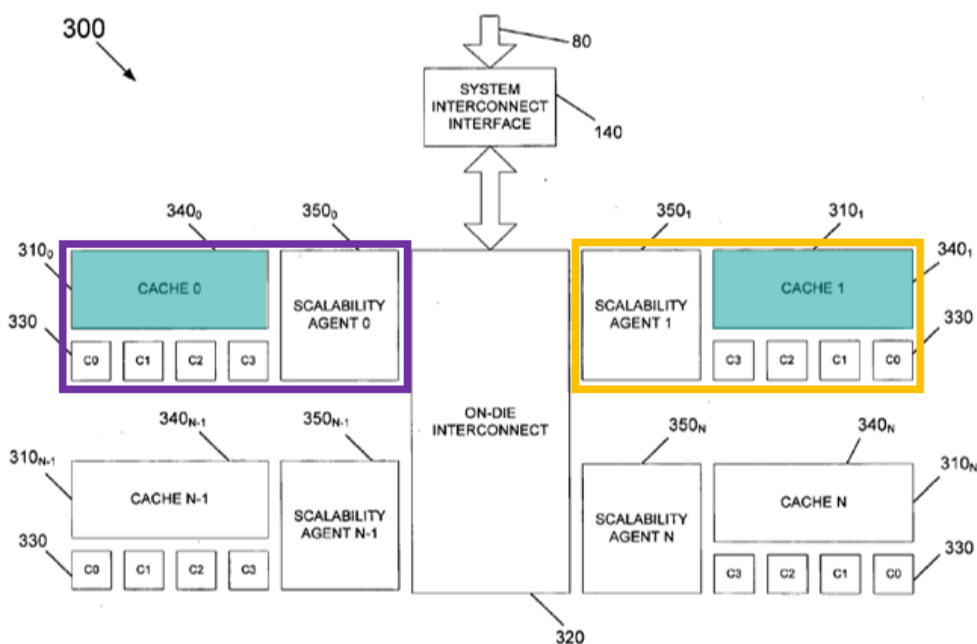
EX1031-Koker, Figure 3C, 13:36-42. EX1002-Decl., ¶106.

Koker teaches a “cache memory hierarchy” including an L1 cache 373, which is local to the MCG 365A (and MCG 365B) and an L2 cache 375, which is shared among all MCGs 365A-365N. EX1031-Koker, 13:56-66 (“[L]evel 1 (L1) caches and shared memory units 373 store graphics data ... locally within each multi-core group 365A.... A Level 2 (L2) cache 375 shared by all ... multi-core groups 365A-365N....”), 13:38-42. EX1002-Decl., ¶107.

D. Sistla

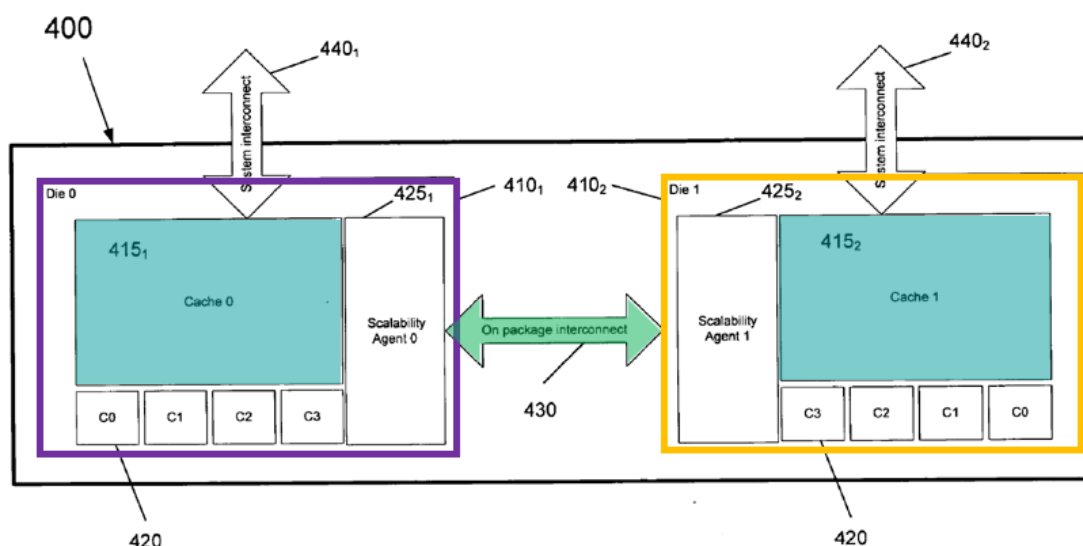
Sistla published ten years before the '803 patent's priority date and is prior art under AIA-35 U.S.C. § 102(a)(1). EX1031-Sistla, Cover. EX1002-Decl., ¶108.

Figure 1 (below) shows a processor 300 having a first “core-cache cluster” 310₀ with a first cache 340₀ and a second “core-cache cluster” 310₁ with a second cache 340₁.



EX1047-Sistla, Figure 2. The caches 340₀ and 340₁ form a “distributed shared cache architecture” that constitutes a “single” last level cache (“LLC”) shared by processing cores (C0 to C3) of both core-cache clusters 310₀ and 310₁. EX1047-Sistla, [0011] (“[C]ore-cache clusters 310 may be ... a distributed shared cache architecture.”), [0012] (“[T]he individual caches to act together as a shared last level cache (LLC).”). EX1002-Decl., ¶109.

Figure 2 (below) shows a processor 400 containing the **first core-cache cluster** in a **first die 410₁** and a **second core-cache cluster** in a **second die 410₂**. The dies 410₁ and 410₂ are coupled by a **point-to-point interconnect**.



EX1047-Sistla, Figure 2, [0012], [0013] (“[P]rocessor 400 ... includes a **first die 410₁** and a **second die 410₂** connected by ... a **point-to-point (PTP) interconnect**.”). The first and second dies 410₁ respectively contain **portions 415₁ and 415₂** of the **LLC**. EX1047-Sistla, [0012], [0013]. EX1002-Decl., ¶110.

VII. LEVEL OF ORDINARY SKILL IN THE ART

A POSITA would have had a bachelor’s degree in computer science, computer engineering, electrical engineering, or equivalent field, and three to five years of experience with semiconductor layout design, or a master’s degree or Ph.D. in the same fields, and one to two years of experience with semiconductor layout design. EX1002-Decl., ¶¶111-116.

VIII. CLAIM CONSTRUCTION

Claims are construed “in accordance with the ordinary and customary meaning of such claim[s] as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.” 37 C.F.R. §42.100(b). In this Petition, two terms require construction. EX1002-Decl., ¶¶117-123.

A. “Chiplet”

NVIDIA, Onesta, and the Staff Attorney in the ITC (“Staff”) agree to construing “chiplet” based on the ’803 patent’s definition:

a device that includes the following characteristics:

- 1) a chiplet includes an active silicon die containing part of the computational logic used to solve a full problem (i.e., the computational workload is distributed across multiples of these active silicon dies);
- 2) chiplets are packaged together as a monolithic unit on the same substrate; and
- 3) the programming model preserves the concept that these separate computational dies are a single monolithic unit (i.e., each chiplet is not exposed as a separate device to an application that uses the chiplets for processing computational workloads).

EX1043-Claim-Construction-Chart, 2; EX1001-’803, 2:41-53 (defining “chiplet”).

EX1002-Decl., ¶¶124-125.

B. “Dedicated for inter-chiplet communications”

“Dedicated for inter-chiplet communication” should be construed as “used for transmission of only chiplet-to-chiplet signals.” The ITC Staff agrees. EX1043-Claim-Construction-Chart, 4. EX1002-Decl., ¶126.

First, the plain meaning supports NVIDIA’s construction. “Dedicated” means “committed entirely to a single purpose or device.” EX1034-Dictionary-of-Computer-Science, 149; *see also* EX1035-Dictionary-of-Multimedia, 118. “Dedicated for inter-chiplet communications” thus means “used for transmission of only chiplet-to-chiplet signals.” EX1002-Decl., ¶127.

Second, the patent carefully distinguishes between “PHY regions” and “non-PHY regions” of the chiplets. EX1001-’803, Figure 2 (showing separate PHY/non-PHY regions). It explains that the “PHY regions” are dedicated for chiplet-to-chiplet signals and interface electrically with the passive crosslink. EX1001-’803, 3:63-4:2 (“[E]ach GPU chiplet 106 is constructed with a *physical device (PHY) region 202* ... *dedicated to the transmission of chiplet-to-chiplet signals,*”), 4:8-16 (“*[T]races 206 [of passive crosslink 118] interface electrically with conductor structures of the PHY regions 202* of the GPU chiplets 106.”) It then explains that the non-PHY regions handle other signaling and that such signaling is “routed around the ... passive crosslink.” EX1001-’803, 4:36-44 (“The conductive pillars 212 connect signals between the GPU chiplets 106 and the substrate in areas where the HDCL

die [i.e., the high density crosslink interconnect chip used for the passive crosslink 118 (3:38-41)] is not present.... In this manner, ***power and input/output (I/O) lines of the non-PHY regions 204 are routed around the interposer die (i.e., passive crosslink 118)*** using redistribution (RDL) technology, thereby replacing usage of conventional TSVs.”), 3:63-4:2 (“[E]ach GPU chiplet 106 is constructed with ... a non-PHY region 204 ... ***tailored more to the conveyance of power and ground and/or chiplet-to-circuit board signals.***”), Figure 2 (showing no TSVs in the passive interposer die/passive crosslink), 4:22-31 (“The non-PHY regions 204 ... interface electrically with a circuit board 210 (or any other substrate) by way ... conductive pillars 212.”). The chiplets’ separate regions for communicating different signals confirms that “dedicated for inter-chiplet communication” means “used for transmission of only chiplet-to-chiplet signals” EX1002-Decl., ¶¶128-131.

Third, the patent distinguishes the “passive crosslink,” which is “dedicated for inter-chiplet communications,” from regions containing TSVs. EX1001-’803, Figure 2 (no TSVs in passive interposer/crosslink), 4:16-21 (“***[T]he passive crosslink 118 does not contain any through silicon vias (TSVs)***). In this manner, the passive crosslink 118 is a passive interposer die ***that communicably couples and routes communications between the GPU chiplets 106....***”), 4:41-44 (“[P]ower and input/output (I/O) lines of the non-PHY regions 204 ***are routed around the ...***

passive crosslink 118... thereby replacing usage of conventional TSVs.”).

EX1002-Decl., ¶¶132-133.

Finally, as discussed in Section V.C [Prosecution-History], Applicants distinguished claim 1 from Teig’s interposer, which is not “dedicated” for inter-chiplet communications, because it was used for chiplet-to-memory communications. EX1027-Teig, [0148]; EX1003-FH, 847. EX1002-Decl., ¶134.

In the ITC investigation, Onesta proposes construing “dedicated for inter-chiplet communication” as “configured to provide communication only between GPU chiplets.” EX1043-Claim-Construction-Chart at 4. Onesta is attempting to rewrite the claims to allow the passive crosslink to convey signals *other than* inter-chiplet communications, including transmitting ground and power signals between a chiplet and other components that the specification explicitly states are instead “*routed around*” the crosslink. In other words, Onesta argues that the claimed crosslink need only be *partially dedicated* or *dedicated some of the time*. Onesta’s argument is directly contrary to the claims, the specification, and the file history, as explained above, and should be rejected. EX1002-Decl., ¶135.

As explained below, the challenged claims are invalid under NVIDIA’s,

Onesta's or any reasonable construction.² EX1002-Decl., ¶135.

IX. INVALIDITY GROUNDS

A. Ground I: Claims 1-4, 7, and 10 Are Obvious Over Kelleher and Collins

1. Claim 1

(a) [1.pre]: A system, comprising:

Kelleher discloses a “graphics system.” EX1004-Kelleher, Figure 2, 3:6-13; Elements [1.a]-[1.d]. EX1002-Decl., ¶¶137-139.

² In the ITC, the Staff proposes construing “a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communication” as “a non-active semiconductor die, which die does not include any through silicon vias and which die is used for transmission of only chiplet-to-chiplet signals.” Onesta asserts that “passive crosslink” means “a non-active connecting structure” and that “dedicated for inter-chiplet communication” means “configured to provide communication only between GPU chiplets” (as noted above). EX1043-Claim-Construction-Chart, 4. NVIDIA submits that this term does not need to be construed for the purpose of this Petition. Nonetheless, as explained below, the challenged claims are invalid under these constructions or any reasonable construction. EX1002-Decl., ¶136.

- (b) [1.a]: a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array,

Figure 2 of Kelleher (below) discloses a CPU 205 communicably coupled to a first GPU chiplet (master GPU chip) of a GPU chiplet array (multi-chip module).

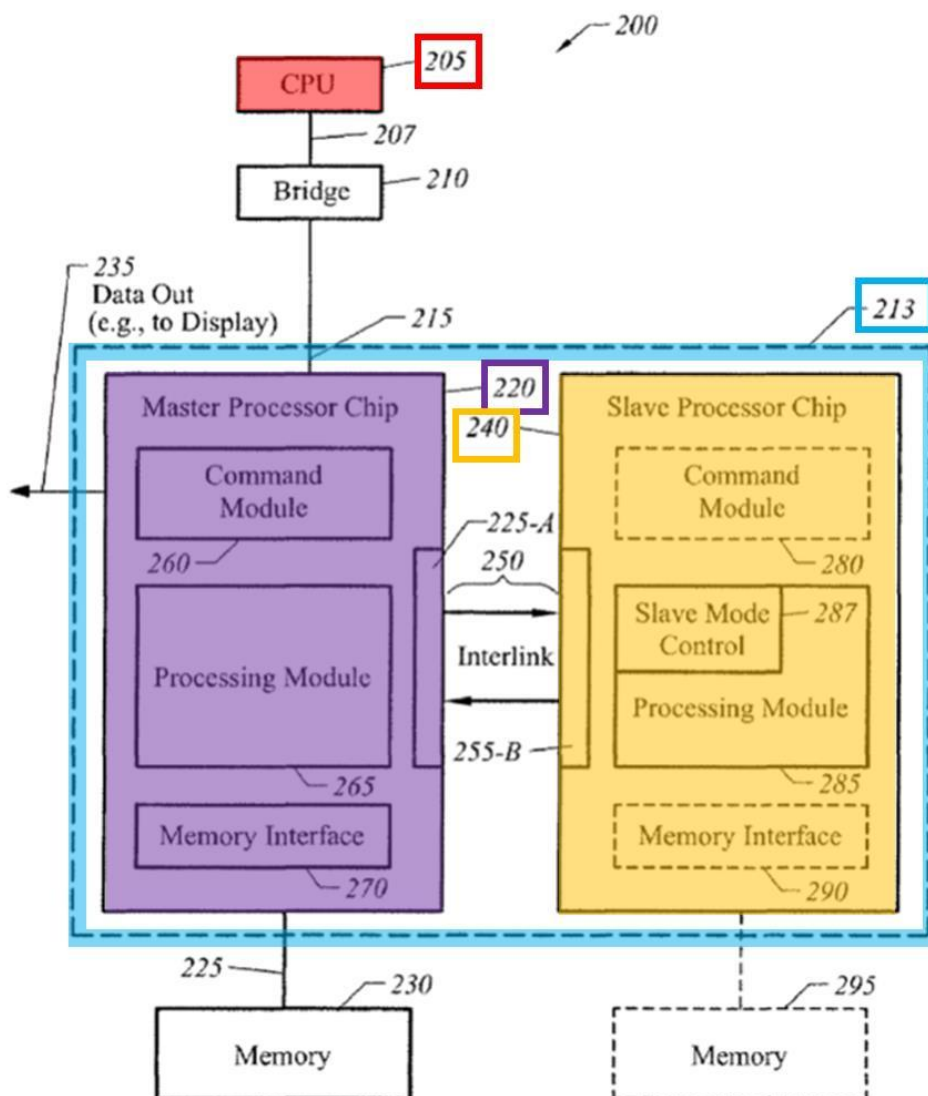


FIG. 2

EX1004-Kelleher, Figure 2. EX1002-Decl., ¶¶140-141.

The CPU is “communicably coupled” to the GPU chip, which receives commands/instructions from the CPU. EX1004-Kelleher, 3:6-13 (“A central processing unit (CPU) 205 is coupled to a graphics processing unit (GPU) chip 220 through a suitable bus interface.”), 3:29-34 (“GPU chip 220 receives a graphics command stream from CPU 205....”), 5:9-11, 3:14-26; Element [1.b]. EX1002-Decl., ¶¶142-145.

In Figure 2 (above), the first GPU chiplet (GPU chip 220) is part of a GPU chiplet array (multi-chip module), which includes the chip 220 and a GPU chip 240. EX1004-Kelleher, 3:45-50 (“GPU chip 240 is packaged with GPU chip 220 as a multi-chip module 213....”), 5:39-65. EX1002-Decl., ¶¶146-148.

Kelleher’s GPU chips 220 and 240 satisfy the ’803 patent’s “chiplet” definition (EX1001-’803, 2:41-53). EX1002-Decl., ¶¶149-150.

First, the ’803 patent states that “a chiplet includes an active silicon die containing part of the computational logic used to solve a full problem (*i.e.*, the computational workload is distributed across multiples of these active silicon dies).” EX1001-’803, 2:44-47. Kelleher’s GPU chips 220 and 240 each includes an active silicon die containing a “processing module.” EX1004-Kelleher, 3:10-13 (“[A] ‘chip’ is an integrated circuit that resides on a die....”), 3:38-41 (“[O]ne of the stages in the graphics pipeline of GPU chip 220 is a *processing module* 265 that performs a processing operation....”), 3:51-53 (“Slave GPU chip 240 includes a *processing*

module 285 capable of performing a processing function....”), 4:3-19. A POSITA would have understood Kelleher’s “dies” to be silicon dies, which were standard in the semiconductor industry. EX1031-Koker, 61:15-50 (“Building ... larger **silicon dies** is challenging for a variety of reasons.... [A] graphics processing unit ... is composed from **diverse silicon chiplets**....”); EX1029-Schaller, 53-54, 57; EX1008-Jerger, *1-*2; EX1011-Kannan, ii, 1-2, 6-13, 18, 44; EX1030-Deng-&-Maly, 450-452. Furthermore, Kelleher explains that “Ziptronix, Inc.” “provides ... services that permit a non-adhesive bond to be formed between a die and a corresponding interconnect on another chip.” EX1004-Kelleher, 6:41-46. A Ziptronix article describes using silicon dies in such bonding processes. EX1028-Ziptronix, 308-315. EX1002-Decl., ¶¶151-155.

Further, each of the **chip 220** and **chip 240** contains part of the computational logic for solving a full problem because the computational workload is distributed between processing modules of the chips **220** and **240**. EX1004-Kelleher, 4:3-19 (“**GPU chip 220** acts as the master processor and utilizes processing module 285 [of **GPU chip 240**] as an additional processing resource that is used in parallel with processing module 265 [of **GPU chip 220**] [A] load balancing module ... determines a distribution of work to be performed by processing modules 265 and 285.... [A]bout half of the processing that would ordinarily be performed by

processing module 265 [of chip 220] is off-loaded to processing module 285 [of chip 240].”). EX1002-Decl., ¶156.

Second, the '803 patent states that “chiplets are packaged together as a monolithic unit on the same substrate...” EX1001-'803, 2:47-48. In Figure 3 (below), Kelleher's chips 220 and 240 are packaged together as a monolithic unit on the same substrate (package 305).

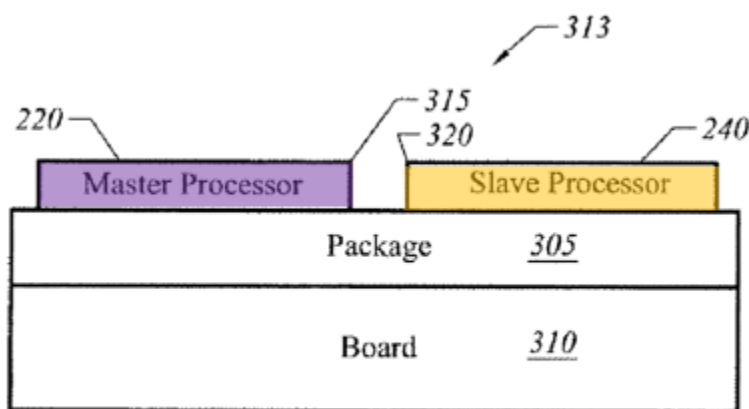


FIG. 3

EX1004-Kelleher, Figure 3, 3:46-50 (“Slave GPU chip 240 is packaged with GPU chip 220 as a multi-chip module 213.... [M]odule 213 may include a package ... mounted onto a board.”), 5:39-44 (“GPU chip 240 and GPU chip 220 are packaged on top of a ... package 305....”). EX1002-Decl., ¶¶157-158.

Third, the '803 patent states that “the programming model preserves the concept that these separate computational dies are a single monolithic unit (i.e., each chiplet is not exposed as a separate device to an application that uses the chiplets for

processing computational workloads).” EX1001-’803, 2:48-54. Applications running on Kelleher’s CPU consider the GPU chips 220 and 240 to be a single monolithic unit. EX1004-Kelleher, 4:63-5:18 (“[O]nly the master GPU 220 receives graphics instructions from CPU 205. As result, replacing a single GPU chip with a multi-chip module having a master/slave pair of GPUs is a *seamless change to driver software within CPU 205*. That is, *the driver software perceives the master/slave pair as a single GPU*.”), 5:19-37. EX1002-Decl., ¶¶159-160.

Onesta has agreed in the ITC to construe “chiplet” per the ’803 patent’s requirements above. EX1043-Claim-Construction-Chart, 2. EX1002-Decl., ¶124.

- (c) [1.b]: wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and

Figure 2 of Kelleher (below) discloses the first GPU chiplet (chip 220) communicably coupled to the CPU via a bus (bus 215 and/or bus 207). EX1002-Decl., ¶¶161-162.

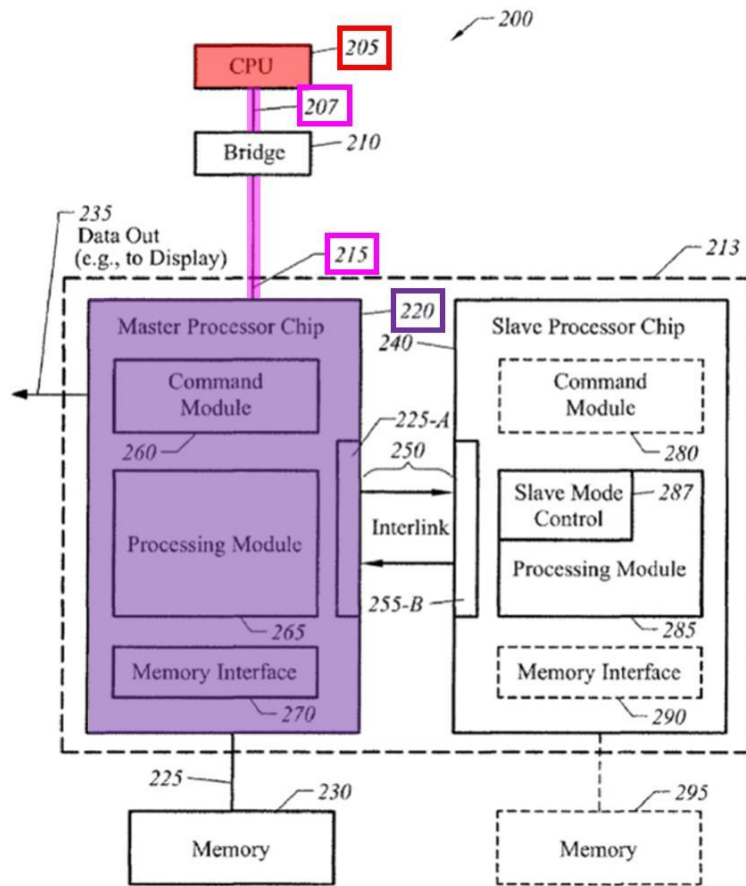


FIG. 2

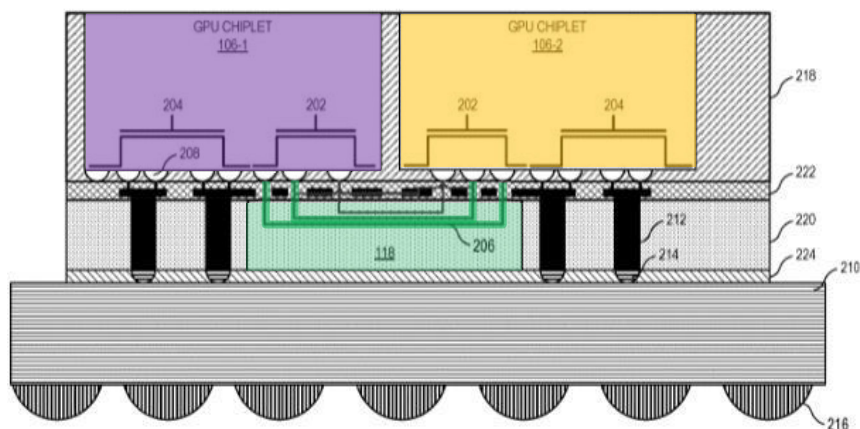
EX1004-Kelleher, Figure 2. Specifically, the CPU is connected to the GPU chip via the bus 207, bridge 210, and bus 215. EX1004-Kelleher, 3:14-24 (“[A] bridge 210 ... couple[s] CPU 205 to GPU chip 220. [A] front side bus 207 may be used to communicatively couple CPU 205 to bridge 210 whereas a ... [PCI]-Express bus 215 may be used to communicatively couple GPU chip 220 to bridge 210.”), 3:6-13 (“[CPU] 205 is coupled to a ... [GPU] chip 220 through a suitable bus interface.”). EX1002-Decl., ¶163.

The '803 patent explains that the first GPU chiplet and CPU may be connected via buses and bridges. EX1001-'803, 2:59-67 (“GPU chiplet 106-1 communicat[es] with the CPU 102 over a direct connection or *via other buses, bridges....*”). EX1002-Decl., ¶164.

(d) [1.c]: [wherein the GPU chiplet array includes:] a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink,

The Kelleher-Collins combination discloses Element [1.c]. EX1002-Decl., ¶¶165-166.

As background, Figure 2 of the '803 patent (below) discloses a **passive crosslink 118** coupling chiplets **106-1** and **106-2**:



EX1001-'803, Figure 2, 3:57-4:2. The patent does not explain in detail what a “passive crosslink” is, but notes that it includes a **passive interposer** with **conductive traces 206** communicably coupling the chiplets **106-1** and **106-2**. EX1001-'803, 4:3-21 (“GPU chiplets 106 are communicably coupled by way of the **passive crosslink**

118.... The passive crosslink 118 includes ... conductor traces [206].... The traces 206 interface electrically with ... the GPU chiplets 106 by way of conducting pathways.... [T]he passive crosslink 118 is a passive interposer die that communicably couples and routes communications between the GPU chiplets 106....”). EX1002-Decl., ¶¶167-169.

Similarly, in Figure 2 of Kelleher (below), the GPU chiplet array (multi-chip module) includes a second GPU chiplet (chip 240) communicably coupled to the first GPU chiplet (chip 220) via an interlink.

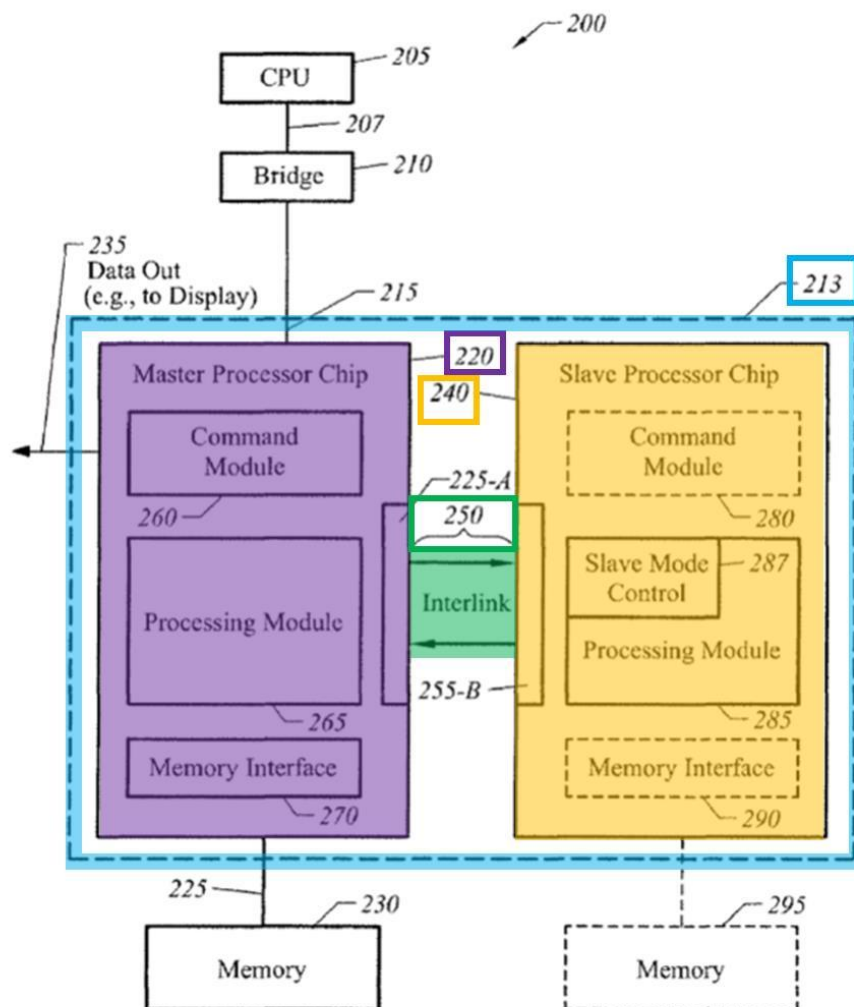
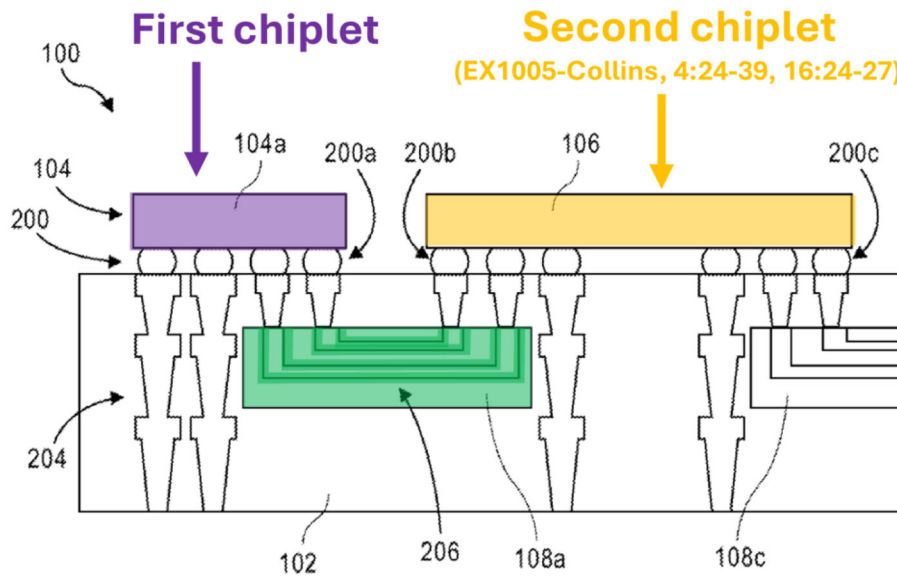


FIG. 2

EX1004-Kelleher, Figure 2, 3:46-50 (“Slave GPU chip 240 is packaged with GPU chip 220 as a multi-chip module 213....”), 4:20-35 (“Slave GPU chip 240 is coupled to GPU chip 220 via an interlink 250.... [I]nterlink 250 facilitate[s] seamlessly offloading processing tasks to slave GPU chip 240 as if processing module 285 of slave GPU chip 240 was an internal extension of processing module 265 of GPU chip 220.”), 7:1-8. EX1002-Decl., ¶¶170. 175.

Like the '803 patent, Figure 2 of Collins show a “passive” **crosslink** (**interconnect structure/embedded multi-die bridge (EMIB)** with **conductive paths**). EX1005-Collins, Figure 2, 3:49-4:23 (“[T]he **interconnect structure 108** includes ... **embedded multi-die interconnect bridges (EMIBs)**, where each of the **EMIBs** are embedded within the substrate assembly 102. The substrate assembly 102 includes conductive elements that couple the **chiplets 104** and the memory device 106 [or **second chiplet** (4:24-39, 16:24-27)] to the **EMIBs**. Each of the **EMIBs include a piece of passive silicon with **conductive elements** located within the passive silicon....**”), 5:42-47 (“[T]he **interconnect structure 304** comprises an **interposer**.... The **interposer can be passive ... with ... conductive paths located within the silicon**.”). EX1002-Decl., ¶171.

Although Figure 2 shows the **EMIB** coupling a **chiplet** and **memory device**, Collins discloses that the **EMIB** can couple **multiple chiplets**. EX1005-Collins, 4:24-39 (“[T]he **EMIBs 108** may couple **two or more of the chiplets 104** together.... The chiplets 104 coupled with each other via the EMIBs 108 may utilize the EMIBs 108 for communication between the chiplets 104”), 16:24-27 (“[T]he **interconnect structure** further **couples a physical layer (PHY) of the first chiplet to a PHY of the second chiplet**.”). A modified version of Figure 2 (below) shows the **interconnect structure** in the **chiplet-to-chiplet** embodiment:



EX1005-Collins, Modified Figure 2 (replacing memory device with **second chiplet** (4:24-39, 16:24-27)). EX1002-Decl., ¶¶172-174.

Since Collins’s **EMIB/interconnect structure/interposer** (with **conductive paths**) is “passive” (EX1005-Collins, 3:49-4:64, 5:42-47), it discloses the “passive crosslink” under the proposed constructions of NVIDIA (“passive crosslink”) and Onesta (“a non-active connecting structure”). Collins further discloses the “passive crosslink” under the Staff’s proposed construction: “a non-active [passive] semiconductor die [**embedded multi-die bridge (EMIB)/interconnect/interposer**], which die does not include any through silicon vias.” *See also* Element 1[d], Claim 2. EX1002-Decl., ¶¶172-174.

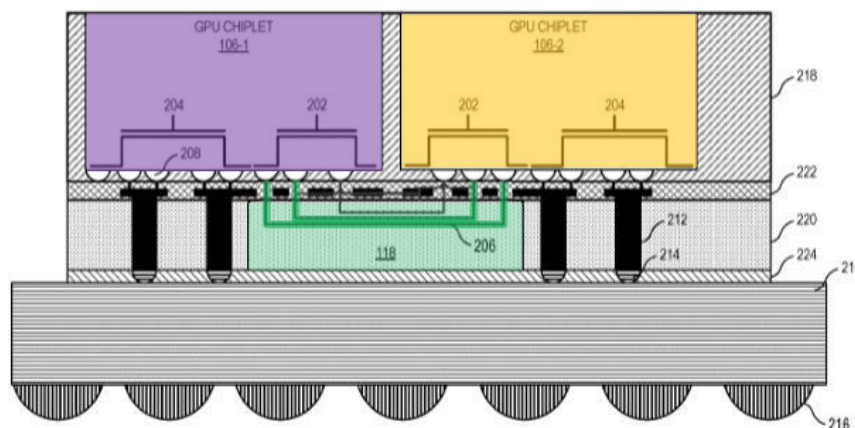
Accordingly, the Kelleher-Collins combination discloses that Kelleher's GPU chiplet array (**multi-chip module**) includes Kelleher's second GPU chiplet (**chip 240**) communicably coupled to Kelleher's first GPU chiplet (**chip 220**) via Collins's passive crosslink (**EMIB/interposer with conductive elements**). EX1002-Decl., ¶¶175-177.

(e) Motivation to Combine Kelleher and Collins

A POSITA would have been motivated to combine Kelleher's and Collins's teachings by communicably coupling Kelleher's GPU chips 220 and 240 with Collins's **passive interconnect/EMIB**. EX1002-Decl., ¶178.

First, Kelleher and Collins are analogous art to each other and the challenged claims, because they all relate to communicably coupling chiplets mounted on the same substrate. EX1002-Decl., ¶179.

- Figure 2 of the '803 patent (below) discloses **first** and **second** chiplets mounted on the same substrate, and an **interposer** communicably couples the chiplets.



EX1001-'803, Figure 2, 4:3-21 (“[T]he passive crosslink 118 is a passive interposer die that communicably couples and routes communications between the GPU chiplets....”). EX1002-Decl., ¶180.

- Figures 2-3 of Kelleher (below) disclose a **first chiplet 220** and a **second chiplet 240** mounted on the same substrate (package), and an **interlink** communicably couples the chiplets.

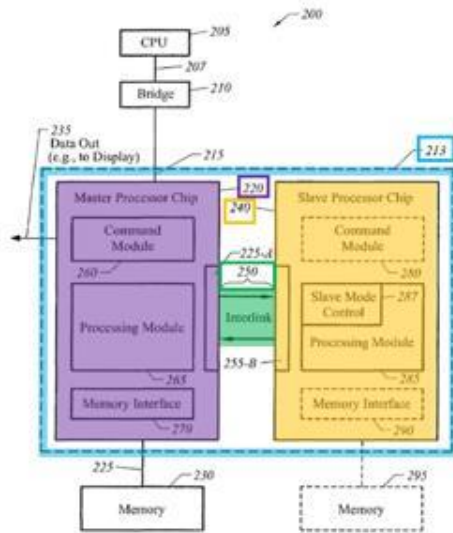


FIG. 2

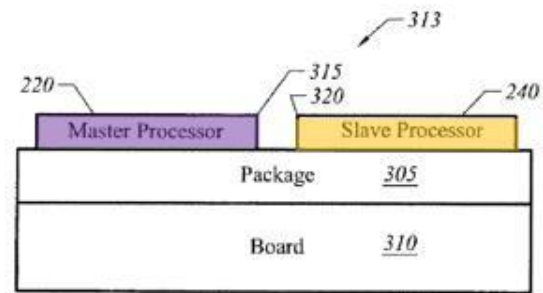


FIG. 3

Kelleher, Figs. 2-3, 4:20-35 (“GPU chip 240 is coupled to GPU chip 220 via an **interlink 250**.... [S]ending and receiving data on **interlink 250** does not exceed a timing margin for processing data.”), 5:39-57 (“GPU chip 240 and GPU chip 220 are packaged on ... package 305....”). EX1002-Decl., ¶181.

- Collins discloses an **EMIB** connecting multiple chiplets together. EX1005-Collins, 4:24-39 (“The chiplets 104 coupled with each other via the EMIBs

108 may utilize the EMIBs 108 for communication between the chiplets 104....”) EX1002-Decl., ¶182.

Second, a POSITA would have been motivated to combine Kelleher’s and Collins’s teachings by communicably coupling Kelleher’s GPU chips 220 and 240 with Collins’s **passive interconnect structure/EMIB**. As explained below, incorporating Collins’s passive interconnect structure into Kelleher’s system would provide more bandwidth for transferring data, but without needing additional components to address heat issues. EX1002-Decl., ¶183.

In Figure 3 of Kelleher (below), GPU chips 220 and 240 are mounted in an “edge-to-edge” configuration on package 305, and the interlink’s sub-links (not shown) extend along the top of package 305 between the right edge 315 of the **chip 220** and the left edge 320 of the **chip 240**.

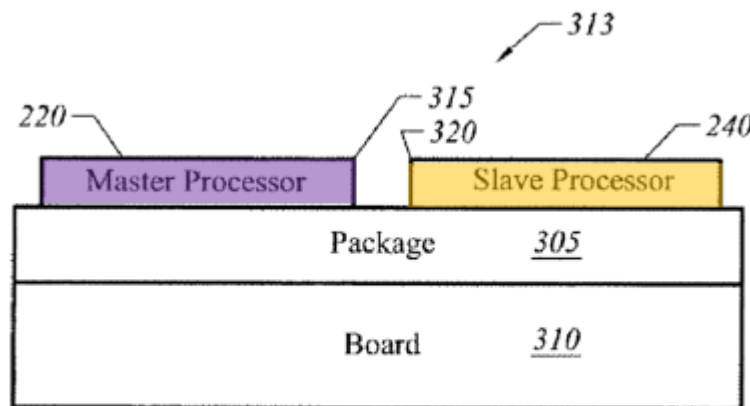


FIG. 3

EX1004-Kelleher, Figure 3, 5:39-57 (“GPU chip 240 and GPU chip 220 are packaged on top ... package 305.... An edge-to-edge configuration permits only a comparatively small number of ... sub-links to be formed between edges 315 and 320.”). Since the sub-links only extend along opposing edges 315 and 320 of the chips, the edge-to-edge configuration limits the number of sub-links that can communicably couple the chips, which may not provide enough bandwidth for transferring data. EX1004-Kelleher, 5:45-48 (quoted above), 5:51-57 (“[A] drawback of an edge-to-edge packaging configuration is that it may not permit an interlink 250 to be formed having sufficient latency and bandwidth attributes required to support parallel processing of stages in high performance GPUs requiring a large bandwidth for data transfer....”). EX1002-Decl., ¶184.

To overcome this problem, Kelleher discloses a “parallel” configuration in Figure 4 (below) where the GPU chips 220 and 240 are mounted on opposite sides of the package.

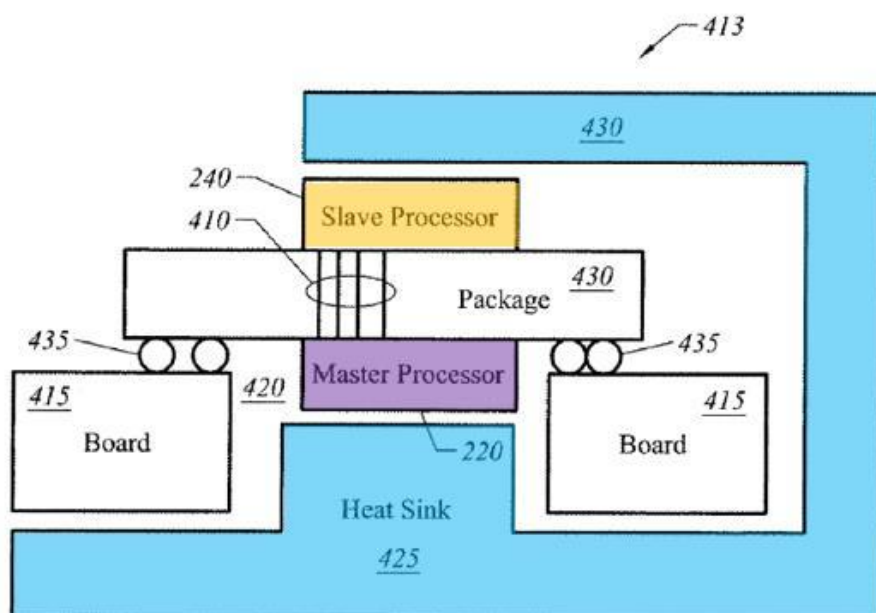


FIG. 4

EX1004-Kelleher, Figure 4. As shown, sub-links (“vias”) 410 of the interlink are created across the bottom area of the GPU chip, through the package, to corresponding portions across the top area of the GPU chip. More sub-links can be created across the top and bottom areas of the chips than along respective side edges of the chips (as in Figure 3). Thus, the “parallel” configuration provides more bandwidth for transferring data. EX1004-Kelleher, 5:58-6:23 (“FIG. 4 shows ... a multi-chip module 413 for implementing a low latency, high bandwidth interlink 250.... GPU chip 220 and slave GPU 240 are flip-chip mounted to opposite sides of a package 405.... Conductive vias 410 within package 405 are used to form interlink 250. The vias 410 may be distributed across the area of GPU chip 220 and

slave GPU 240 to provide a combined (total) bandwidth that scales with the number of vias.”). EX1002-Decl., ¶¶185-189.

However, this “parallel” configuration has disadvantages. Since GPU chips generate substantial heat, when the top and bottom surfaces of the GPU chip and GPU chip face each other in close proximity, additional measures—like providing heat sinks (Figure 4 above)—are required to cool the chips so they operate properly. EX1004-Kelleher, 6:53-55 (“GPU chips dissipate a substantial amount of heat.... Consequently, packaging two GPU chips in a multi-chip module 213 will increase heat dissipation. Thermal modeling is desirable to optimize the heat sinking and cooling for multi-chip module 213.”), 6:16-21 (“[T]he board 415 onto which the package is mounted has a hole 420 to permit both GPU chip 220 and slave processor 240 to be cooled to a heat sink 425.”). Adding heat sinks increases the size and cost of the system. EX1002-Decl., ¶190.

In light these disadvantages, a POSITA would have been motivated to use Collins’s interconnect structure/EMIB shown in Modified Figure 2 (below) (see Element [1.c]) to communicably couple Kelleher’s GPU chip and GPU chip.

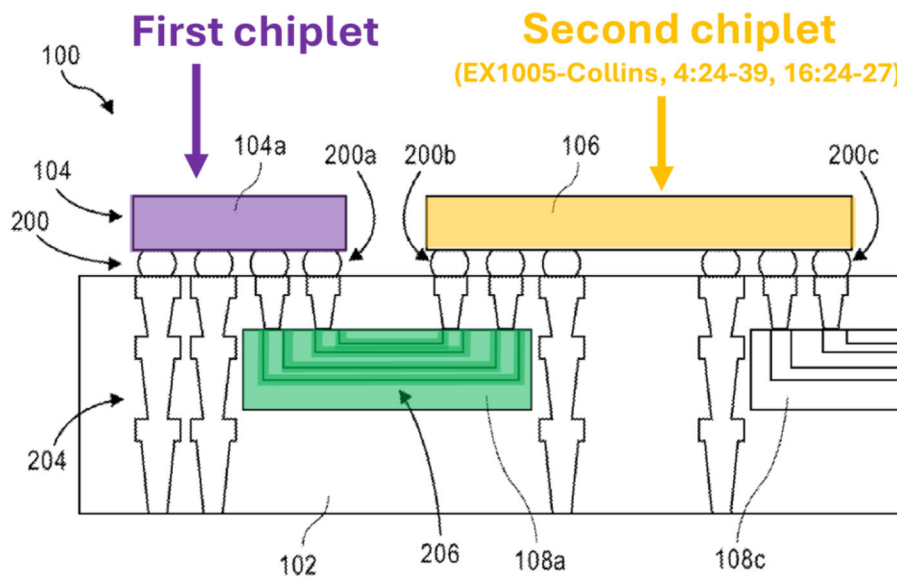


FIG. 2

EX1005-Collins, Modified Figure 2; Element [1.c]. **First**, by using Collins's **EMIB**, the number of **sub-links** coupling Kelleher's **chip** and **chip** is not limited to the number of sub-links that can be formed along the opposing edges of the chips (like Kelleher's Figure 3 embodiment). Instead, more sub-links can be formed on the bottom surfaces of the chips, which increases the bandwidth for transferring data between the chips. **Second**, by using Collins's **EMIB**, only the small opposing edges of the **first** and **second** chips face each other, rather than the chips' large top and bottom surfaces (like Kelleher's Figure 4 embodiment). Accordingly, the need for heat sinks is reduced, which reduces the size and cost of a system containing the chips. EX1002-Decl., ¶¶191-193.

Third, the advantages of using an EMIB as a “dedicated ultra-high-density interconnection between dies within [a] package” were well-known. EX1037-Liu, at [0028]. Because EMIBs are close to the chips, the conductors used are significantly smaller than those required to traverse a full interposer. EX1037-Liu, [0028]. This reduced conductor length “help[s] to significantly reduce loading and directly boost[s] performance.” EX1037-Liu, [0028]. Additionally, EMIBs are less susceptible to “warping”—unintended distortion of the interposer. EX1037-Liu, [0029]; *see also* EX1038-Yazdani, 13-16 (describing challenges associated with large silicon interposers, including cost, reliability, warpage, cracks, and thermal-stresses). Finally, EMIBs require fewer microbumps and TSVs, making them less complex and more cost-effective to manufacture than large interposers. EX1037-Liu, at [0029]._EX1002-Decl., ¶¶194-196.

Fourth, a POSITA would have had a reasonable expectation of success of connecting Kelleher’s GPU chip and GPU chip with Collins’s passive interconnect/EMIB. While Figure 2 of Collins (above) shows the EMIB connecting a GPU chip and a *memory device*, Collins teaches that the EMIB “may couple *two or more of the chiplets 104....*” EX1005-Collins, 4:32-39. Moreover, Collins describes implementation details for coupling two chiplets: “the interconnect structure ... couples a physical layer (PHY) of the first chiplet to a PHY of the second chiplet.” EX1005-Collins, 16:24-27. Thus, a POSITA would have readily

understood how Collins's EMIB would couple Kelleher's two chiplets to facilitate inter-chiplet communication: coupling Kelleher's GPU chip's PHY to its GPU chip's PHY. This implementation would have been well within the skill of a POSITA, who would have had a reasonable expectation of success achieving chiplet-to-chiplet communication without undue experimentation. *See* EX1010-Lau, 19-21; EX1021-ODSA, 6; EX1014-Elsherbini, [0096]. EX1002-Decl., ¶¶47, 197-198.

Accordingly, a POSITA would have been motivated to use Collins's teachings of a passive interconnect structure/EMIB as the interlink in Kelleher's system, because it would have yielded predictable results and is aligned with Kelleher's performance objectives. EX1006-Coskun, *3 §3.3; EX1020-Stow, 728, 731-732. The combination involves applying Collins's well-known passive interconnect structure technique to Kelleher's well-known chiplet system that was ready for improvement, and it would have been a matter of routine design choice within a POSITA's skill. EX1002-Decl., ¶¶51-53, 199-200.

(f) [1.d]: wherein the passive crosslink is dedicated for inter-chiplet communications.

The Kelleher-Collins combination teaches the passive crosslink (Collins's passive interconnect/EMIB) is dedicated for inter-chiplet communications (only between Kelleher's GPU chips 220 and 240). EX1002-Decl., ¶¶201-202.

As noted in Section VIII [Claim-Construction], “dedicated for inter-chiplet communications” should be construed as “used for transmission of only chiplet-to-chiplet signals.” In Figure 2 of Kelleher (below), the **interlink** transmits signals between GPU chips 220 and 240.

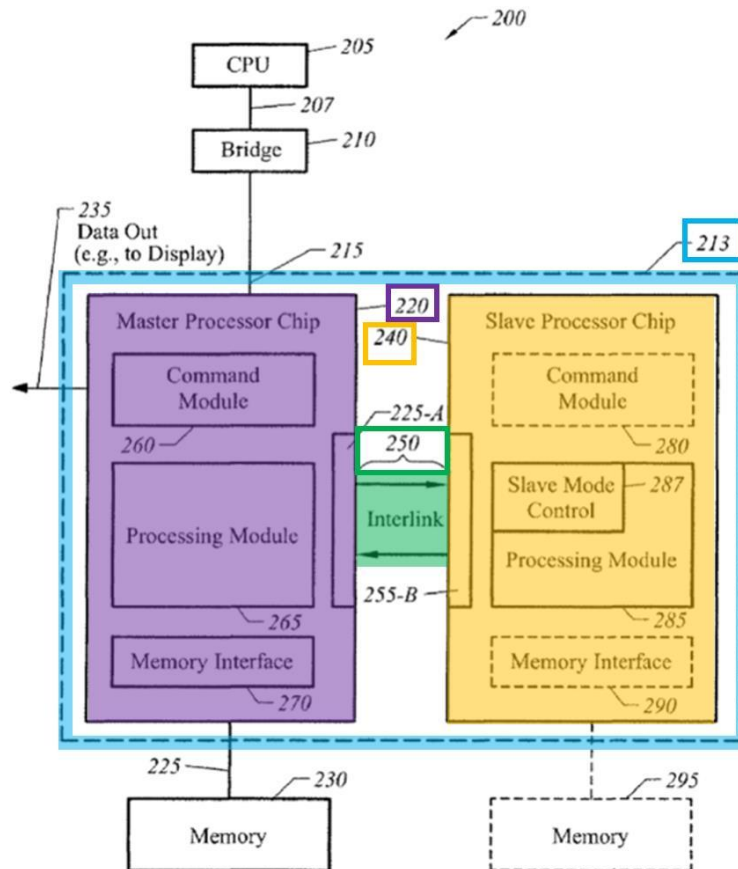


FIG. 2

EX1004-Kelleher, Figure 2, 4:20-35 (“GPU chip 240 is coupled to GPU chip 220 via an **interlink** 250.... [A]ttributes of **interlink** 250 facilitate seamlessly offloading processing tasks to **slave GPU chip 240 as if processing module 285 of slave GPU chip 240 was an internal extension of processing module 265 of GPU chip 220.**”),

Figures 4-11 (interlink transmitting signals between GPU chips). EX1002-Decl., ¶¶203-205.

Further, the Kelleher-Collins combination teaches that the **passive crosslink** (Collins's **interconnect/EMIB**) is dedicated for inter-chiplet communications (*i.e.*, used for transmission of only chiplet-to-chiplet signals between Kelleher's GPU chips 220 and 240). Element [1.c]; Section IX.A.1.(e) [MTC-Kelleher-Collins]. EX1002-Decl., ¶¶205-206.

Namely, in Modified Figure 2 of Collins (below), the **passive EMIB** is used for transmitting signals only between the **first** and **second** chiplets. EX1005-Collins, 4:24-39 ("The chiplets 104 coupled with each other via the EMIBs 108 may utilize the EMIBs 108 for communication between the chiplets 104"), 16:24-27. The EMIB is also free of signaling pathways linking the **first** and **second** chiplets to other components.

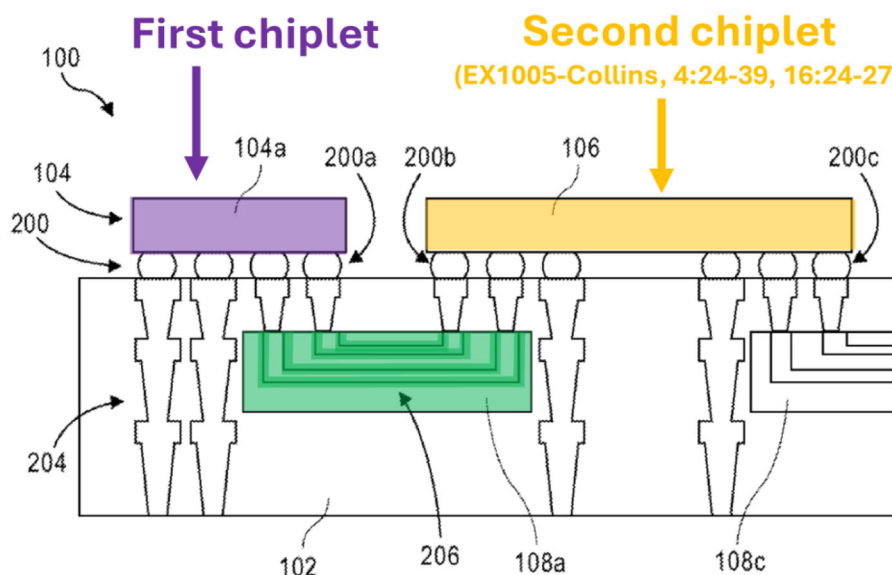


FIG. 2

EX1005-Collins, Modified Figure 2 (replacing memory device with **second chiplet** (4:24-39, 16:24-27)). EX1002-Decl., ¶¶207-208.

Therefore, when Collins’s **passive EMIB** communicably couples Kelleher’s GPU chips **220** and **240**, the Kelleher-Collins combination teaches that the passive crosslink (Collins’s **passive EMIB**) transmits only chiplet-to-chiplet signals between Kelleher’s GPU chip **220** and **240** and is free of signaling pathways linking the chips **220** and **240** to other components (e.g., through silicon vias (TSVs)). EX1005-Collins, 4:24-39 (“The chiplets 104 coupled with each other via the EMIBs 108 may utilize the EMIBs 108 for communication between the chiplets 104 . . .”); Element [1.c]; Section IX.A.1.(e) [MTC-Kelleher-Collins]; *compare* EX1001-’803, 4:12-21

(“[T]he passive crosslink 118 does not contain any through silicon vias (TSVs).”). EX1002-Decl., ¶209.

The Kelleher-Collins combination also discloses a passive crosslink that is “dedicated for inter-chiplet communications” under the Staff’s and Onesta’s proposed constructions. As noted above, the combination discloses that Collins’s *interconnect/EMIB/interposer* provides communication only between Kelleher’s GPU chiplets (per Onesta’s construction) and the *interconnect/EMIB/interposer* is a “die...used for transmission of only chiplet-to-chiplet signals” (per the Staff’s construction). Element [1.c], Claim 2. EX1002-Decl., ¶210.

Accordingly, the Kelleher-Collins combination teaches the passive crosslink (Collins’s *passive EMIB*) dedicated for inter-chiplet communications (only between Kelleher’s GPU chips 220 and 240). EX1002-Decl., ¶211.

2. Claim 2

(a) The system of claim 1, wherein the passive crosslink comprises a passive interposer die.

Collins teaches that its passive crosslink (*interconnect structure/EMIB*) comprises a *passive interposer die*. EX1005-Collins, 3:49-64 (“[T]he *interconnect structure 108* includes one or more *embedded multi-die interconnect bridges (EMIBs)*.... [T]he *interconnect structure 108* may include active EMIBs with repeaters designed into the bridge, *interposers*....”), 5:46-47 (“The *interposer can be passive* ... with ... conductive paths....”). EX1002-Decl., ¶¶212-216.

3. Claim 3

- (a) [3.a]: The system of claim 1, wherein the first GPU chiplet comprises a first PHY region including conductor structures for chiplet-to-chiplet communications,**

The Kelleher-Collins combination teaches that the first GPU chiplet (Kelleher's **chip 220**) comprises a first PHY region (**PHY I/O region**) including conductor structures (**solder balls**) for chiplet-to-chiplet communications (between Kelleher's GPU chips **220** and **240**). EX1002-Decl., ¶¶217-220.

As background, in Figure 2 of the '803 patent (below), the **first GPU chiplet** comprises a **first PHY region 202** having **conductor structures** for inputting/outputting signals from/to the **second chiplet** via the **passive crosslink**.³

³ Although symbols 202 and 204 look like transistors, a POSITA would have understood that they delineate PHY and non-PHY regions, respectively. EX1001-'803, 3:57-4:2; EX1004-Kelleher, 3:64-4:2. EX1002-Decl., ¶222 n.4.

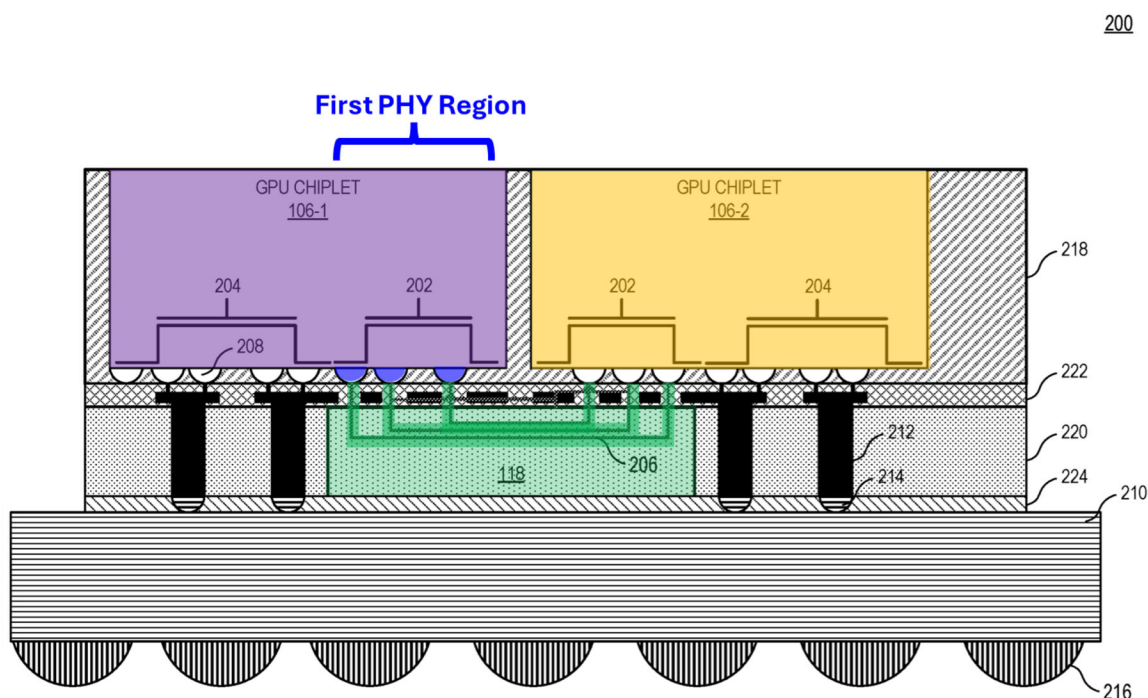


FIG. 2

EX1001-'803, Figure 2, 3:57-4:21 (“[E]ach GPU chiplet 106 is constructed with a *physical device (PHY) region 202*, which has various ... *conductor structures* dedicated to the transmission of chiplet-to-chiplet signals.... The *passive crosslink 118* includes ... *[conductor] traces 206*.... The *traces 206* interface electrically with *conductor structures of the PHY regions 202* of the GPU chiplets 106....”) The '803 patent explains that the conductive structures include “solder bump[s], micro bump[s], and the like.” EX1001-'803, 4:22-35. EX1002-Decl., ¶¶221-223.

In Modified Figure 2 of Collins (below), the *first chiplet* comprises a *first PHY region (PHY I/O)* for inputting/outputting signals from/to a *second PHY region*

(PHY I/O) of the **second chiplet** (EX1005-Collins, 4:24-39, 16:24-27) via the **passive crosslink**.

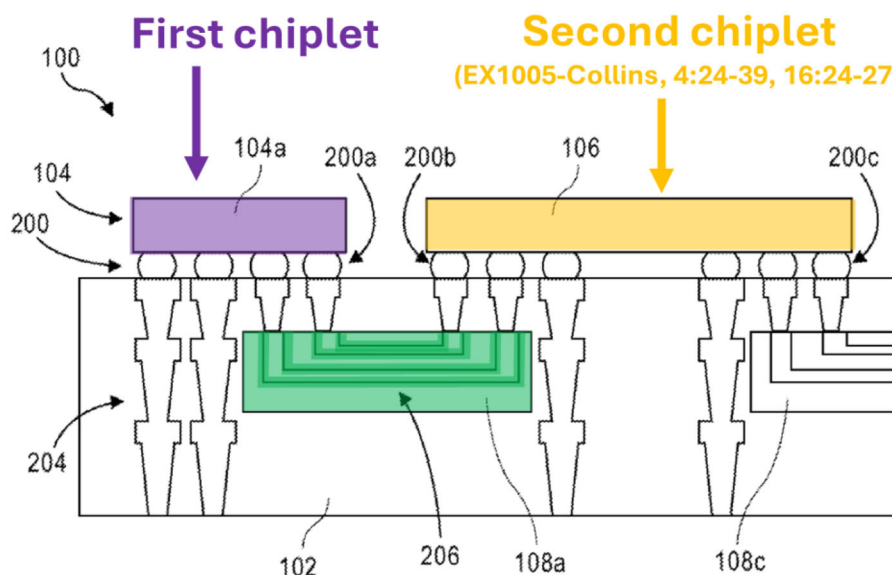


FIG. 2

EX1005-Collins, Modified Figure 2 (replacing memory device with **second chiplet** (4:24-39, 16:24-27)), 1:20-24, 16:24-27 (“[T]he **interconnect structure** further couples a **physical layer (PHY)** of the **first chiplet** to a **PHY** of the **second chiplet**.”), 3:34-48 (“The memory device 106 [or **second chiplet** (4:24-39, 16:24-27)] **includes a PHY I/O**.... **For distributed PHY I/O**, the memory device 106 [or **second chiplet** (4:24-39, 16:24-27)] **includes a plurality of PHY I/Os**....”), 3:65-4:23 (“[E]ach of the **EMIBs** can couple each of the chiplets 104 to a ... different **distributed PHY I/Os** [of the **second chiplet**]. For example, the first **EMIB 108a** can couple the **first chiplet 104a** to a **first distributed PHY I/O** [of the **second chiplet**], the second EMIB

108b can couple the second chiplet 104b to a second distributed PHY I/O *[of the second chiplet]/....*”), 8:6-11, 11:4-18, 12:32-46. EX1002-Decl., ¶¶224-226.

Figure 2 of Collins (above) shows that the PHY I/Os of the first and second chiplets (EX1005-Collins, 4:24-39, 16:24-27) have external conductor structures (elements 200a and 200b comprising a ball grid array). EX1005-Collins, 4:46-67 (“The interconnect elements 200 couple the first chiplet 104a ... and the memory device 106 [or second chiplet (4:24-39, 16:24-27)] to conductive elements 204 of the substrate assembly 102.... [T]he interconnect elements 200 are ... a ball grid array.”) A POSITA would have understood that balls in a ball grid array are the same as the solder bumps in Figure 2 of the ’803 patent. EX1001-’803, 4:25-27. EX1036- Mearig-&-Goers, 434. EX1002-Decl., ¶¶227-229.

Accordingly, the Kelleher-Collins combination teaches that the first GPU chiplet (Kelleher’s chip 220) comprises a first PHY region (Collin’s PHY I/O region) including conductor structures (solder balls) for chiplet-to-chiplet communications (between Kelleher’s chips 220 and 240). EX1002-Decl., ¶230.

(b) [3.b]: and further wherein the second GPU chiplet comprises a second PHY region including conductor structures for chiplet-to-chiplet communications.

As explained for Element [3.a], the Kelleher-Collins combination teaches that the second GPU chiplet (Kelleher’s chip 240) comprises a second PHY region (Collins’s PHY I/O) including conductor structures (solder balls) for chiplet-to-

chiplet communications (between Kelleher's chips 220 and 240). Element [3.a].
EX1002-Decl., ¶¶231-232.

4. Claim 4

- (a) **[4.a]: The system of claim 3, further comprising: a third GPU chiplet communicably coupled to the first GPU chiplet via the passive crosslink,**

The Kelleher-Collins combination teaches a third GPU chiplet (Kelleher's second slave GPU chip) communicably coupled to the first GPU chiplet (Kelleher's master chip 220) via the passive crosslink (Collins's interconnect structure/EMIB).
EX1002-Decl., ¶¶233-236.

Although Kelleher's Figure 2 embodiment (below) shows a master GPU chip and one slave GPU chip connected via an interlink, Kelleher discloses that the master GPU chip may have a second slave GPU chip.

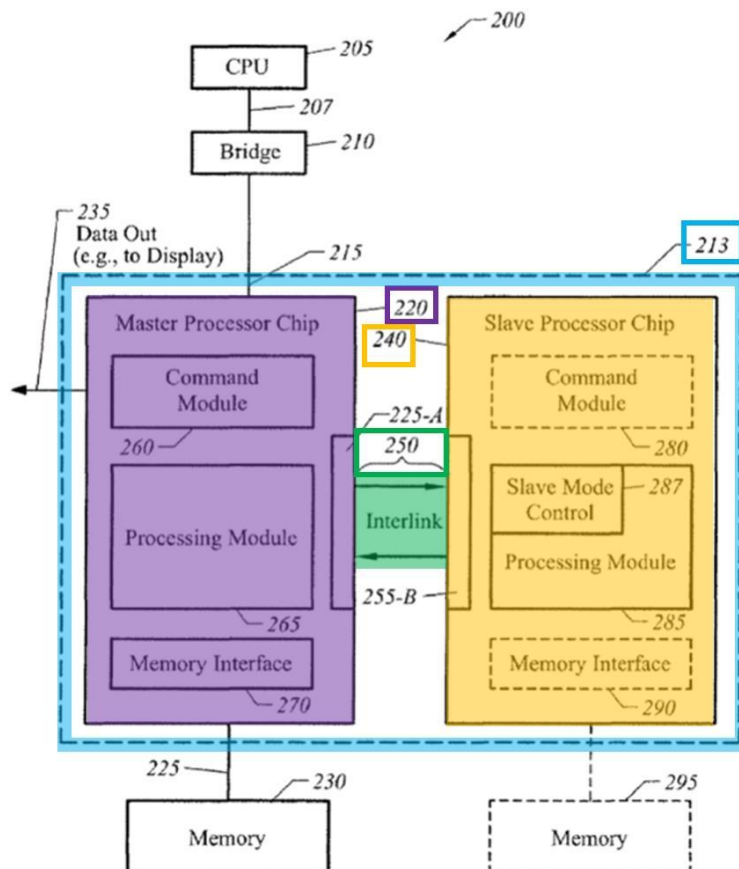


FIG. 2

EX1004-Kelleher, Figure 2, 8:8-13 (“[I]t may be desirable to utilize *two or more slave GPUs for each master GPU.*”), 10:45-59. A POSITA would have understood that the second slave GPU chip (and the *first slave GPU chip*) would be communicably coupled to the *master GPU chip* via the *interlink*. See Element [1.a]. EX1002-Decl., ¶¶237-239.

Similarly, Collins teaches that its *interconnect structure/EMIB* may connect three chiplets. EX1005-Collins, 4:32-41 (“*EMIBs 108 may couple two or more of the chiplets 104 together....* The chiplets 104 coupled ... via the EMIBs 108 may

utilize the EMIBs 108 for communication between the chiplets 104”). EX1002-Decl., ¶240.

Accordingly, the Kelleher-Collins combination teaches a third GPU chiplet (Kelleher’s second slave chip) communicably coupled to the first GPU chiplet (Kelleher’s **chip 220**) via the passive crosslink (Collins’s **interconnect structure/EMIB**). Element [1.c]. EX1002-Decl., ¶241.

(a) [4.b]: wherein the third GPU chiplet comprises a third PHY region including conductor structures for chiplet-to-chiplet communications.

The Kelleher-Collins combination teaches that the third GPU chiplet (Kelleher’s second slave GPU chip) comprises a third PHY region (Collins’s **PHY I/O**) including conductor structures (**solder balls**) for chiplet-to-chiplet communications (between Kelleher’s **master chip** and second slave GPU chip). EX1002-Decl., ¶¶242-243.

As explained for Element [4.a], the Kelleher-Collins combination teaches that Kelleher’s second slave GPU chip performs chiplet-to-chiplet communications with Kelleher’s **master chip** via the passive crosslink (Collins’s **interconnect structure/EMIB**). Element [4.a]; EX1005-Collins, 4:32-41. As explained for Elements [3.a]-[3.b], the Kelleher-Collins combination teaches that Kelleher’s **master chip** and **first slave chip** comprise first and second PHY regions (Collins’s **PHY I/Os**) including conductor structures (**solder balls**) for chiplet-to-chiplet

communications. For similar reasons, the Kelleher-Collins combination teaches that Kelleher's second slave GPU chip comprises a third PHY region (Collins's **PHY I/Os**) including conductor structures (**solder balls**) for chiplet-to-chiplet communications. Elements [3.a]-[3.b]. EX1002-Decl., ¶¶244-246.

5. Claim 7

- (a) The system of claim 1, wherein the passive crosslink communicably couples all GPU chiplets in the GPU chiplet array.**

As explained for Elements [1.a] and [1.c], in the Kelleher-Collins combination, the passive crosslink (Collins's **interconnect structure/EMIB**) communicably couples Kelleher's GPU chiplets **220** and **240** in Kelleher's GPU chiplet array (**multi-chip module 213**). Elements [1.a], [1.c]. As noted in Element [1.a] and in Figure 2 of Kelleher, the chips **220** and **240** are all GPU chiplets in the **module**. Element [1.a]; EX1004-Kelleher, Figure 2. EX1002-Decl., ¶¶247-250.

6. Claim 10

- (a) The system of claim 1, further comprising: a plurality of conductive pillars coupling a circuit board substrate to a first non-PHY region of the first GPU chiplet and a second non-PHY region of the second GPU chiplet.**

The Kelleher-Collins combination teaches conductive pillars (Collins's **conductive elements**) coupling a circuit board substrate (Collins's **printed circuit board (PCB)**/Kelleher's **circuit board**) to a **first non-PHY region** of Kelleher's **first**

GPU chiplet 220 and a second non-PHY region of Kelleher's second GPU chiplet 240. EX1002-Decl., ¶¶251-253.

As background, Figure 2 of the '803 patent (below) discloses conductive pillars coupling a circuit board substrate to non-PHY regions 204 of the first and second GPU chiplets.

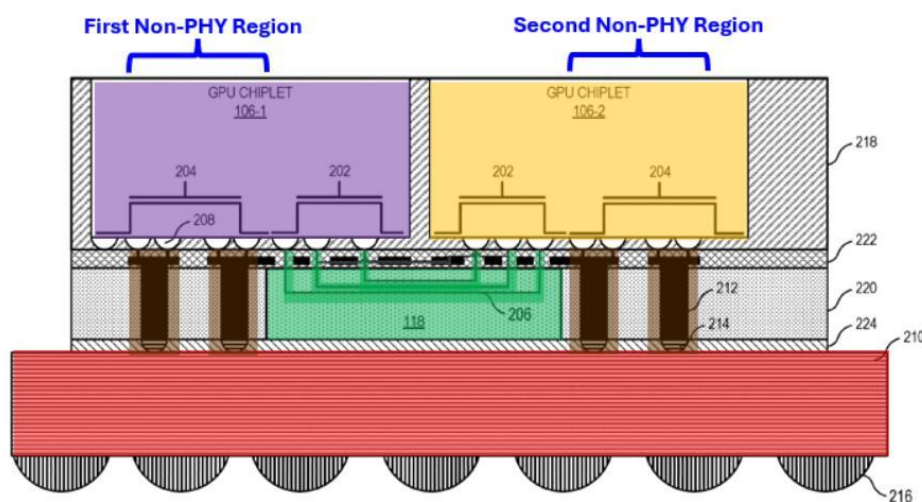


FIG. 2

EX101-'803, 3:63-4:2, 4:22-24 (“The non-PHY regions 204 of the GPU chiplets 106 interface electrically with a circuit board 210 ... by way of ... conductive pillars 212.”). EX1002-Decl., ¶¶254-256.

As explained for Elements [3.a]-[3.b], in Collins, the first and second chiplets have a PHY regions that communicate via the interconnect structure/EMIB. See Elements [3.a]-[3.b]. As shown in Modified Figure 2 of Collins (below), the first and second chiplets (EX1005-Collins, 4:24-39, 16:24-27) have non-PHY regions for

coupling to a **printed circuit board (PCB)** (not shown) via conductive pillars (conductive elements 204).

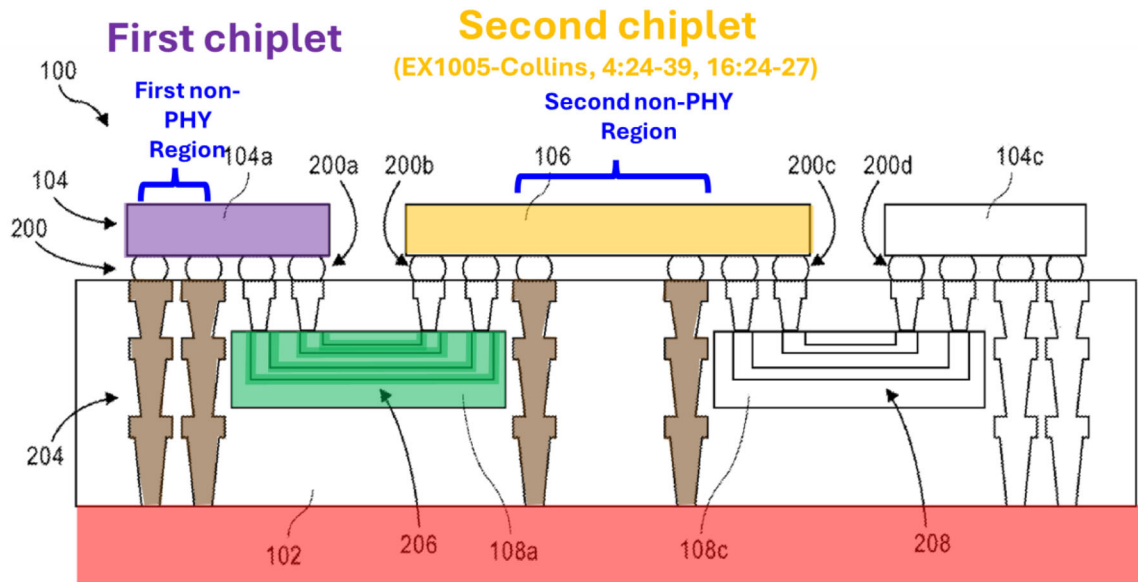


FIG. 2

EX1005-Collins, Modified Figure 2 (replacing memory device with **second chiplet** (4:24-39, 16:24-27)), 4:40-5:29. EX1002-Decl., ¶257.

Figure 10 of Collins (below) shows a semiconductor package 1004 containing a **chiplet 1010** and memory device 1012 [or **second chiplet** (EX1005-Collins, 4:24-39, 16:24-27)] connected via an **EMIB 1008**.

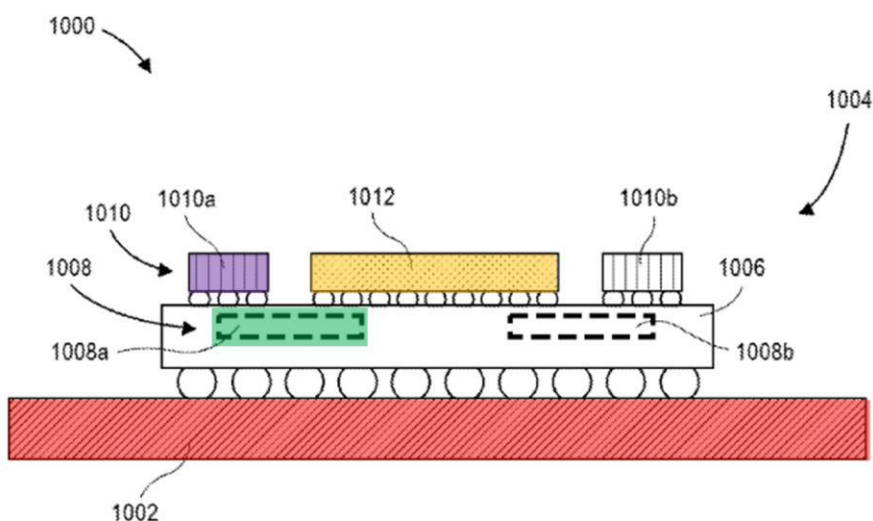


FIG. 10

EX1005-Collins, Figure 10, 13:26-35 (“[P]ackage 1004 includes a substrate assembly 1006... [and] ... an **interconnect structure 1008**. [T]he **interconnect structure 1008** is illustrated as a **first EMIB 1008a**...”), 14:8-15 (“The **interconnect structure 1008** provides the coupling among the **chiplets 1010** and the memory device 1012 [or **second chiplet** (EX1005-Collins, 4:24-39, 16:24-27)]...”). EX1002-Decl., ¶¶258-260.

Although Figure 10 shows the **EMIB 1008a** connecting the **chiplet 1010a** and memory device 1012, like the Figure 2 embodiment, the **EMIB 1008a** may connect multiple chiplets together. EX1005-Collins, 14:12-16 (“[T]he **interconnect structure 1008** may couple ... the chiplets 1010 with each other...”), 4:32-39. EX1002-Decl., ¶261.

Figure 10 of Collins also shows the substrate assembly 1006 (similar to substrate assembly 204 (in Figure 2 (above)) mounted on a **PCB 1002**. EX1005-Collins, 13:11-35 (“The substrate assembly 1006 is mounted and coupled to the **PCB 1002**.... [T]he **interconnect structure 1008** is illustrated as a first **EMIB 1008a** ... embedded within the substrate assembly 1006.”). EX1002-Decl., ¶261.

A POSITA would have understood that, in Figure 10, the **non-PHY regions** of the **first chiplet 1010a** (and second chiplet) and memory device 1012 would couple to the **PCB 1002** (not shown) via conductive pillars (**conductive elements 204**) (as taught by the Figure 2 embodiment). EX1005-Collins, 13:15-25 (“[In Figure 10,] [t]he semiconductor package 1004 is mounted to ... the **PCB 1002**. The *semiconductor package 1004 may implement ... the features [of]* the semiconductor packages (such as *the semiconductor package 100 (FIG. 1)*....”), 1:35-37 (“**FIG. 2** illustrates a cross-sectional, side view of ... *semiconductor package of FIG. 1*.”). EX1002-Decl., ¶262.

Figure 3 of Kelleher (below) also shows GPU chiplets **220** and **240** coupled to a **circuit board** via a substrate (package).

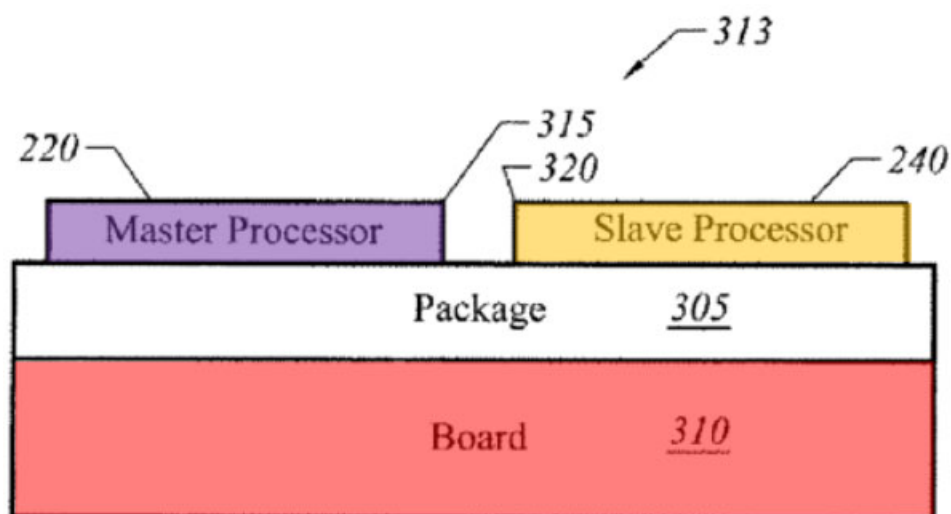


FIG. 3

E1004-Kelleher, Figure 3, 5:39-44 (“GPU chip 240 and GPU chip 220 are packaged on ... package 305 that is then placed on a board 310.”). EX1002-Decl., ¶¶263-264.

A POSITA would have known that Kelleher’s GPU chips 220 and 240 receive power from the circuit board via the package to operate, and thus, there must be some connection between (1) the chips 220 and 240 and (2) the board 310 via the package. EX1004-Kelleher, 3:35-50, 5:39-57, 5:66-6:52. Collins teaches a typical way to provide electrical connections to chiplets via a package/substrate: forming conductive elements through the package/substrate as shown in Figure 2. EX1002-Decl., ¶265.

Accordingly, as shown in the annotated figure below, the Kelleher-Collins combination teaches using conductive pillars (Collins’s conductive elements) to couple a circuit board substrate (Kelleher’s board) to a first non-PHY region of the

Kelleher's first GPU chiplet and a second non-PHY region of Kelleher's second GPU chiplet.

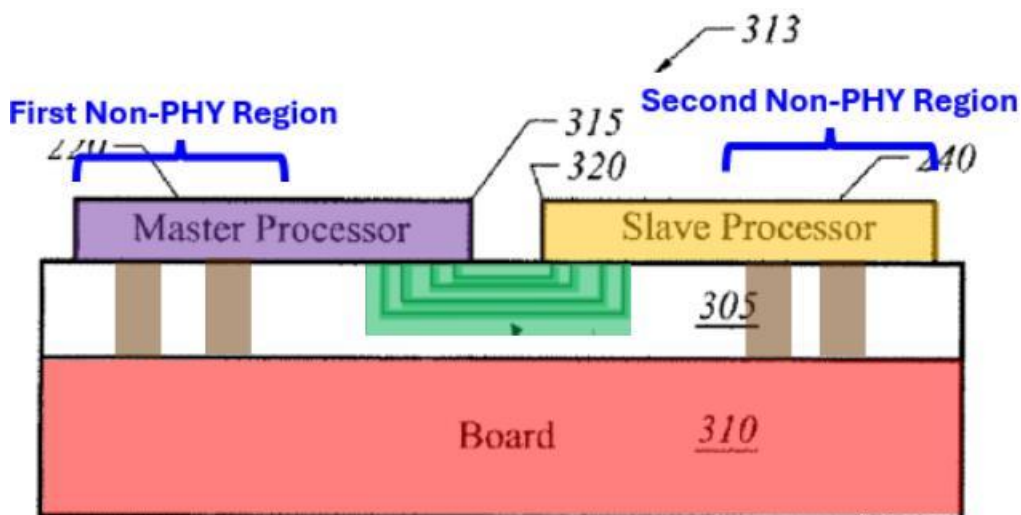


FIG. 3

EX1004-Kelleher, Modified Figure 3. EX1002-Decl., ¶266.

B. Ground II: Claims 8 and 9 Are Obvious Over Kelleher, Collins, Koker, and Sistla

1. Claim 8

- (a) [8.a]: The system of claim 1, further comprising: a first cache memory hierarchy at the first GPU chiplet,

The Kelleher-Collins-Koker-Sistla combination teaches a first cache memory hierarchy (Koker's L1 cache/shared memory 373/L2 cache 375 (modified per Sistla)) at Kelleher's first GPU chiplet 220 (corresponding to Koker's MCG 365A).

EX1002-Decl., ¶¶267-270.

As background, Figure 3 of the '803 patent (below) shows first and second chiplets 106-1 and 106-2 having a “cache memory hierarchy”—a tiered system of memories having different levels (e.g., Level 1 (L1), Level 2 (L2), and Level 3 (L3) caches).

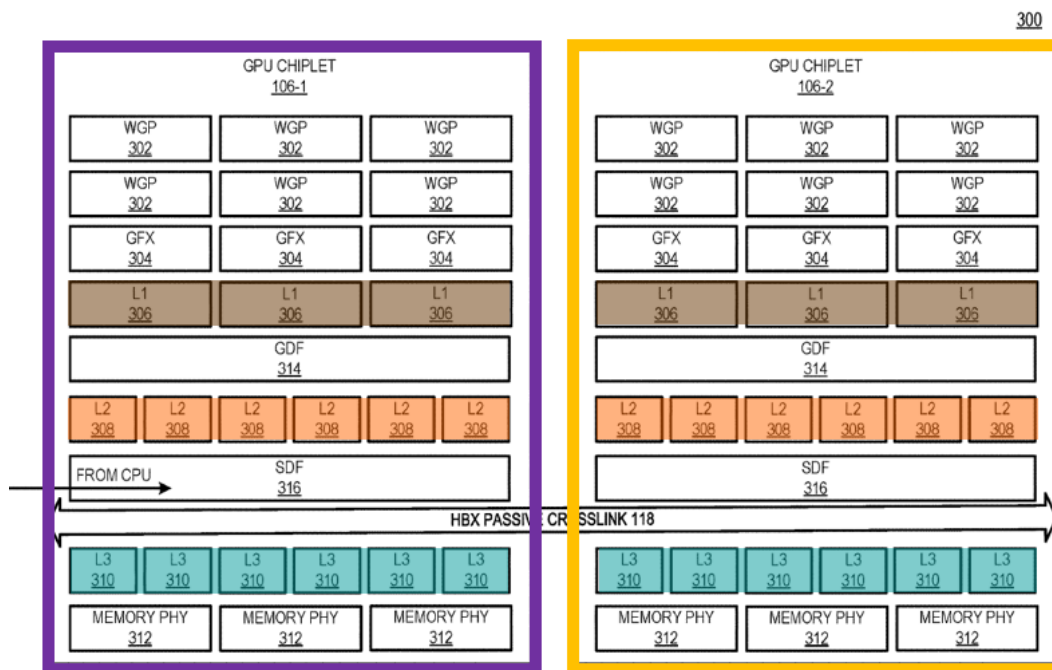


FIG. 3

EX1001-'803, Figure 3, 5:1-18 (“FIG. 3 ... illustrat[es] a *cache hierarchy of GPU chiplets*.... Each of the GPU chiplets 106-1, 106-2 include[s] ... workgroup processors 302 (WGP) and ... fixed function blocks 304 (GFX) that communicate with a given channel’s L1 cache memory 306. Each GPU chiplet 106 also includes ... L2 cache memory 308 banks ... [and] L3 cache memory 310 channels....”).
EX1002-Decl., ¶¶270-272.

Like the '803 patent, Figure 3C of Koker (below) discloses that each of the first and second MCGs 365A and 365B is associated with a cache memory hierarchy with different levels of cache: **Level 1 (L1)** and **Level 2 (L2)** cache.

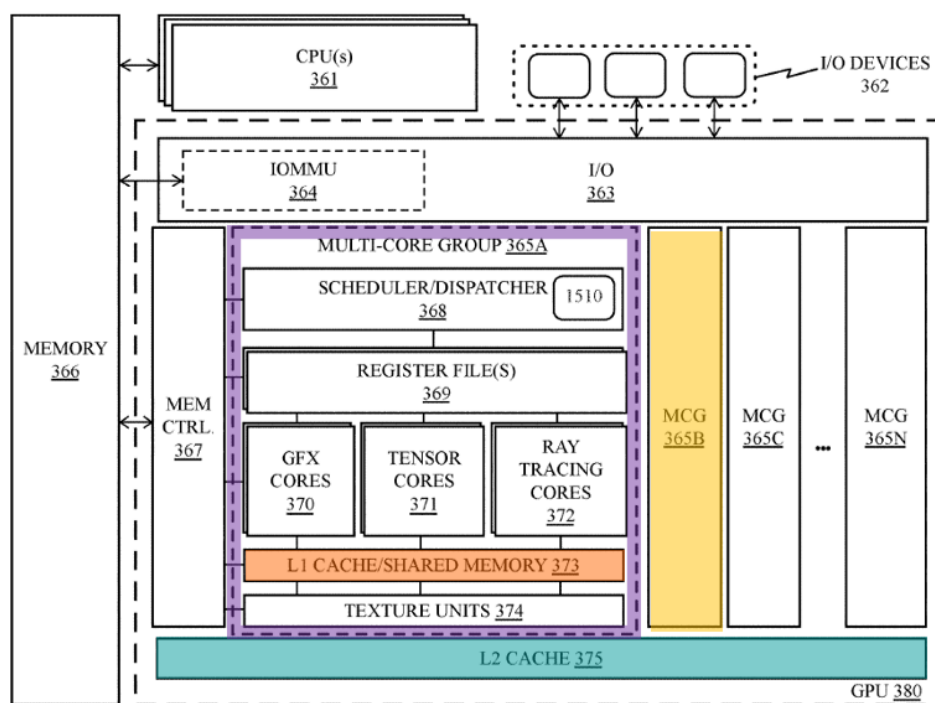


FIG. 3C

EX1031-Koker, Figure 3C, 13:36-42 (“FIG. 3C illustrates a ... [GPU] 380 which includes ... multi-core groups 365A-N. While the details of only a single multi-core group 365A are provided, ... the other multi-core groups 365B-365N may be equipped with the same ... graphics processing resources.”), 13:55-66 (“[L]evel 1 (L[1]) caches and shared memory units 373 store graphics data ... locally.... A Level 2 (L2) cache 375 shared by all or a subset of the multi-core groups 365A-365N stores graphics data....”); *see also* EX1031-Koker, 17:37-40, 40:29-38 (“*The cache memory hierarchy may include at least one level of instruction and data*

cache within each processor core and one or more levels of shared mid-level cache, such as a Level 2 (L2), Level 3 (L3), Level 4 (L4), or other levels of cache....”). EX1002-Decl., ¶¶273-276.

Figure 3 of the ’803 patent (below) discloses that the L3 cache is a “last level cache” (“LLC”) that is included within the first and second chiplets 106-1 and 106-2.

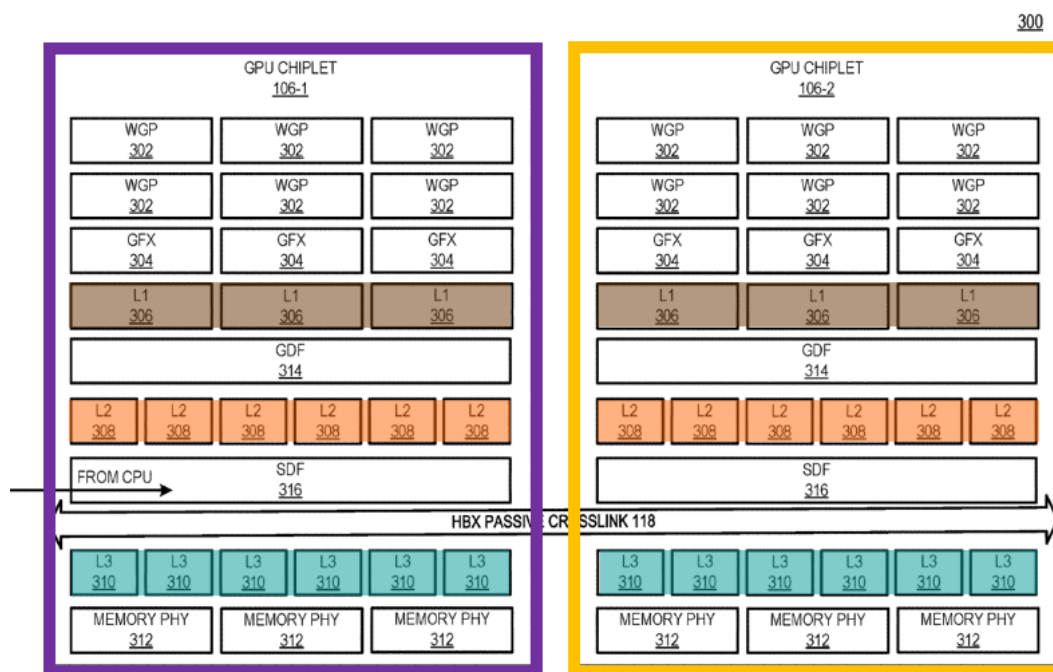


FIG. 3

EX1001-’803, Figure 3, 5:14-18 (“[T]he L3 level (or other last level) of cache is unified and coherent across all of the GPU chiplets 106.”). EX1002-Decl., ¶277.

In Koker, the L2 cache is an LLC because, as shown in Figure 3C (below), it is the highest level of cache before external memory 366.

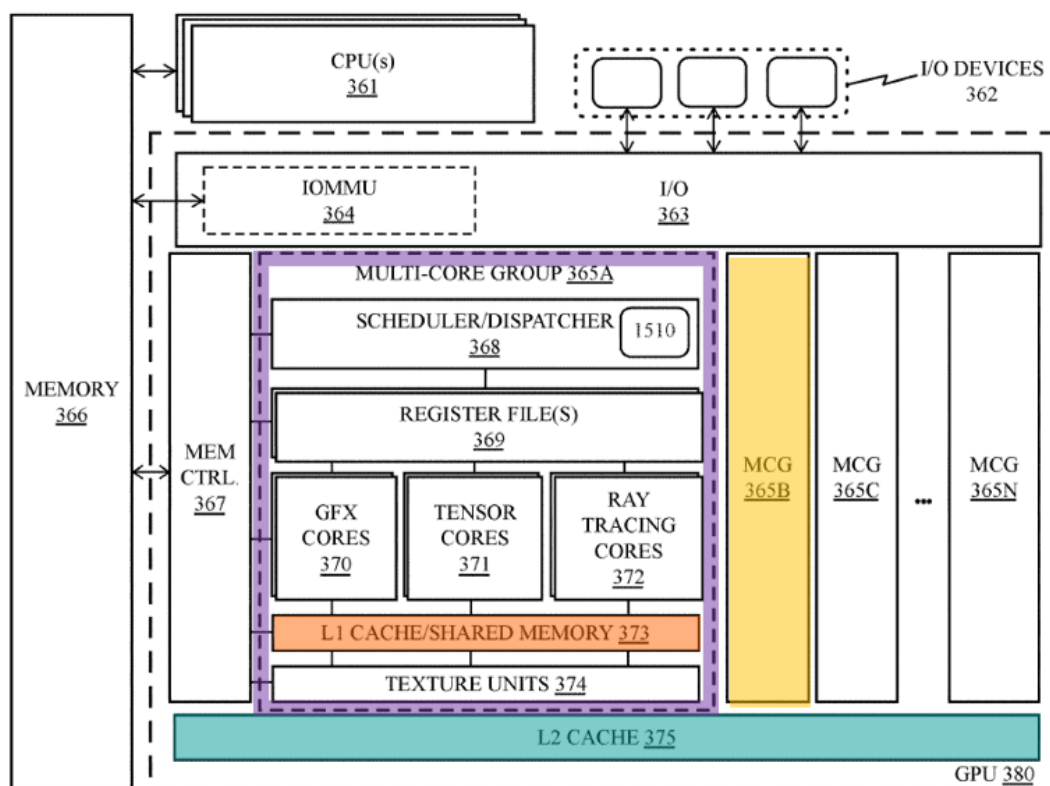
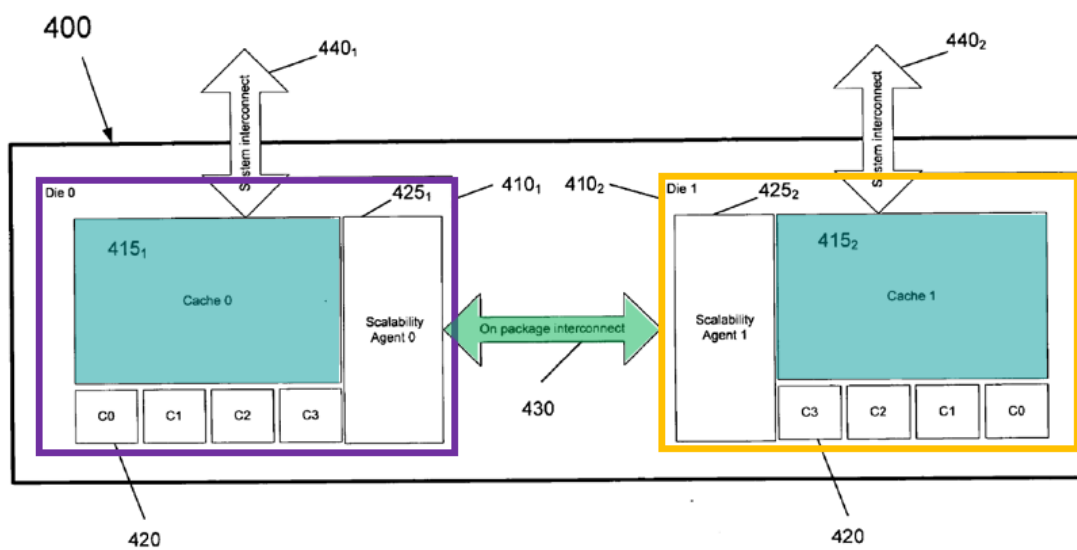


FIG. 3C

EX1031-Koker, Figure 3C, 40:29-42 (“[T]he **highest level of cache before external memory** is classified as the **LLC**.”), 13:64-14:2 (“[T]he **L2 cache 375** may be shared across ... multi-core groups 365A-365N.... [M]emory controllers 367 couple the GPU 380 to **a memory 366 which may be a system memory**....”). EX1002-Decl., ¶¶278-279.

While Koker’s **L2 cache** is outside the MCGs 365A-365N, Sistla teaches including the **LLC** within dies/chiplets. In Figure 2 (below), Sistla shows a processor comprising first and second dies 410₁ and 410₂ (corresponding to **first** and **second** chiplets, respectively) connected via a **point-to-point (PTP) interconnect**.

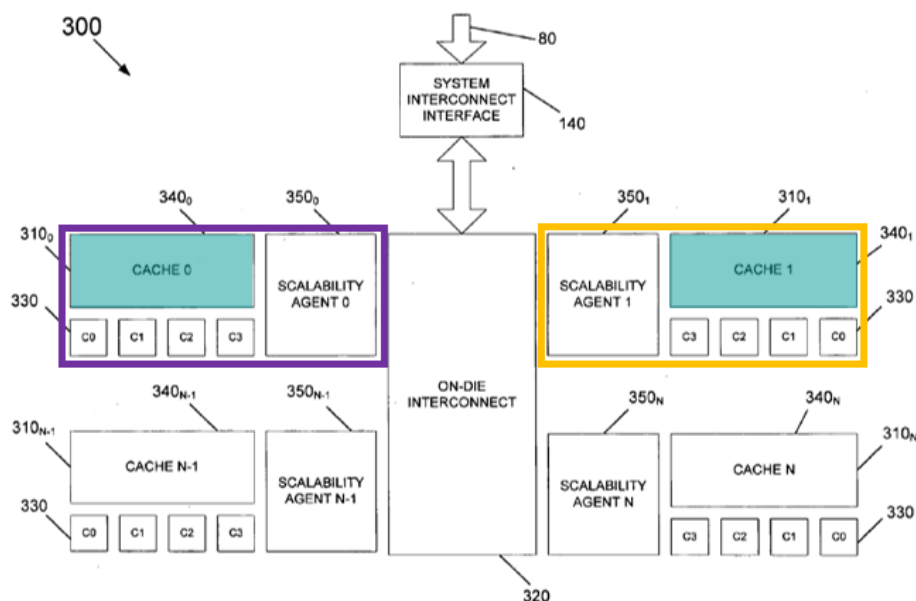


EX1047-Sistla, Figure 2, [0013] (“[P]rocessor 400 ... includes a first die 410₁ and a second die 410₂ connected by ... a point-to-point (PTP) interconnect.”) The first die/chiplet contains a first cache 415₁, and the second die/chiplet contains a second cache 415₂. EX1047-Sistla, [0013] (“Each die 410 (generically) may include ... a cache 415₁ and 415₂....”). EX1002-Decl., ¶¶280-281, 284.

Sistla further explains that the caches 415₁ and 415₂ constitute a “distributed shared cache architecture” and form an LLC, which is shared among cores 420 of the first and second die/chiplets 410₁ and 410₂. EX1047-Sistla, [0011]-[0013]. EX1002-Decl., ¶¶282, 284.

For example, in Figure 1 (below), Sistla discloses a first “core-cache cluster” 310₀ (corresponding to first die/chiplet 410₁) having a first cache 340₀ (corresponding to cache 415₁) and a second “core-cache cluster” 310₁

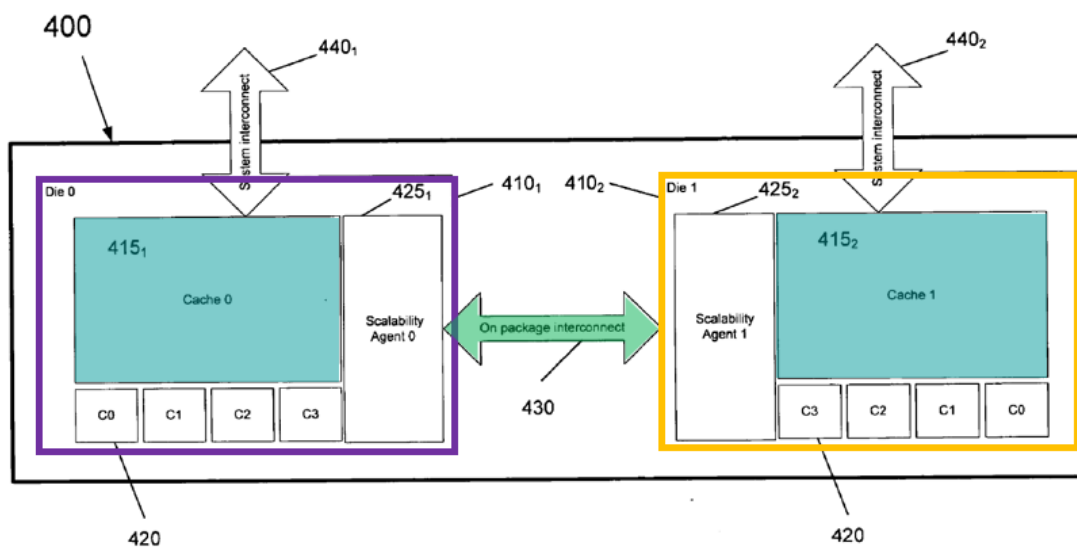
(corresponding to **second die/chiplet 410₂**) having a **second cache 340₁** (corresponding to **cache 415₂**).



EX1047-Sistla, Figure 1, [0011] (“[P]rocessor 300 includes ... core-cache clusters 310₁-310_n.... Each core-cache cluster 310 includes ... cores 330 that share one of a plurality of **caches 340₁-340_N**....”). The **caches 340₀-340₁** form “a distributed shared cache architecture” constituting a “**single**” **LLC** that the cores (C0-C3) 300 of the “core-cache clusters” 310₀ and 310₁ share. EX1047-Sistla, [0011] (“The architecture of core-cache clusters 310 may be ... a *distributed shared cache architecture*.”), [0011]-[0012] (“**[C]aches 340₀-340₁** ... *act together as a shared last level cache (LLC)*....”). EX1002-Decl., ¶¶283-284.

Sistla’s Figure 2 embodiment (below) is similar to the Figure 1 embodiment (above) but includes the “core-cache clusters” within the **first die/chiplet 410₁**

(having the cache 415₁) and the second die/chiplet 410₂ (having the cache 415₂).



EX1047-Sistla, Figure 2, [0013] (“FIG. 1 ... include[es] a processor formed on a single die.... [I]n FIG. 2, processor 400 ... includes a first die 410₁ and a second die 410₂.... Each die 410 (generically) may include at least one core-cache cluster including a cache 415₁ and 415₂ and a plurality of cores 420.”). Like the Figure 1 embodiment, the caches 415₁ and 415₂ form a shared LLC for the dies 410₁ and 410₂. EX1047-Sistla, [0013] (“[T]he dies [410₁ and 410₂] can be joined at the interconnect, thus doubling the number of cores.... [T]he caches 415₁ and 415₂ will now behave like a shared cache.... [I]f each die were a 4-core die with 8 megabytes (M) cache, then the combined 8-core processor may behave as a 16M cache.”), [0011]-[0012]. EX1002-Decl., ¶285.

As explained below in Section IX.B.1.(d) [MTC-Kelleher-Collins-Koker-Sistla], a POSITA would have been motivated by Sistla’s teachings to include

different portions of Koker's **last level L2 cache** within Koker's MCGs 365A-365N. Accordingly, the Koker-Sistla combination teaches a first cache memory hierarchy (Koker's **L1 cache/shared memory 373/L2 cache 375 (modified by Sistla)**) at Koker's **first MCG 365A**. When Koker's teachings are incorporated into Kelleher (as explained below in Section IX.B.1.(c) [MTC-Kelleher-Collins-Koker]) and modified by Sistla's teachings (as discussed in Section IX.B.1.(d) [MTC-Kelleher-Collins-Koker-Sistla]), the combined teachings disclose a first cache memory hierarchy (Koker's **L1 cache/shared memory 373/L2 cache 375 (modified by Sistla)**) at Kelleher's **first GPU chiplet 220** (corresponding to Koker's **MCG 365A**). EX1002-Decl., ¶¶286-288.

(b) [8.b]: wherein a first level of the first cache memory hierarchy is coherent within the first GPU chiplet; and

The Kelleher-Collins-Koker-Sistla combination teaches that a first level (Koker's **L1 cache/shared memory 373**) of the first cache memory hierarchy (Koker's **L1 cache/shared memory 373/L2 cache 375 (modified by Sistla)**) is coherent within Kelleher's **first GPU chiplet 220** (corresponding to Koker's **MCG 365A**). EX1002-Decl., ¶¶289-290.

As background, in Figure 3 of the '803 patent (below), multiple components (WGP 302/GFX 304) of the **first GPU chiplet** access a local first level (**L2 cache**)

of the hierarchy (L1 cache/L2 cache/L3 cache), and thus, the L2 cache is coherent within the first chiplet.

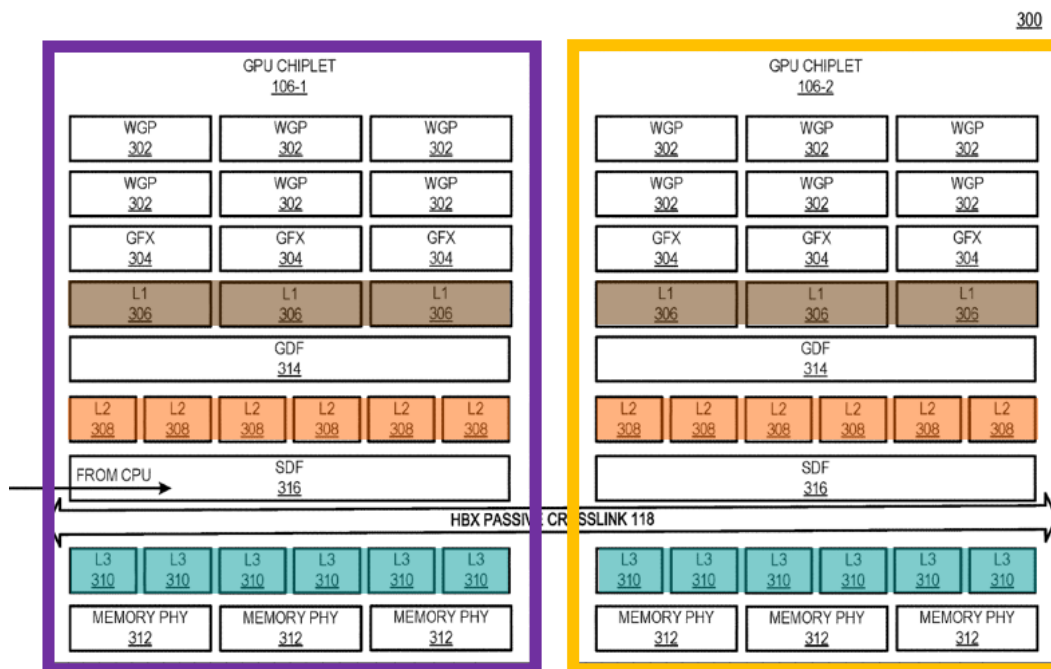


FIG. 3

EX1001-'803, Figure 3, 5:1-25 (“*Each GPU chiplet 106 also includes ... L2 cache memory 308 banks that are individually accessible.... The L2 level of cache is coherent within a single chiplet.... [E]ach of the workgroup processors [WGP] 302 and fixed function blocks [GFX] 304 [within the first chiplet 106-1] ... access data stored in any bank of the L2 cache memory 308.*”); Element [8.a]. EX1002-Decl., ¶¶291-293.

Since multiple WGPs and GFXs of the first chiplet access the L2 cache, a POSITA would have understood that the L2 cache is “coherent” to ensure that the data stored in the cache remains valid and up-to-date, even when different

components read/write data from/to the same L2 cache memory locations. EX1048-Nayyar, 20 (“Cache coherence schemes (protocols) ensure that each request from a processor gets the most up-to-date copy of the block.”). EX1002-Decl., ¶¶294-295.

Similarly, in Figure 3C of Koker (below), multiple components (GFX, tensor, and ray tracing cores 370-372) of the first MCG 365A access a first level (L1 cache/shared memory) of the first cache memory hierarchy (L1 cache/shared memory/L2 cache(modified by Sistla)), and thus, a POSITA would have understood that the L1 cache is coherent within the first MCG.

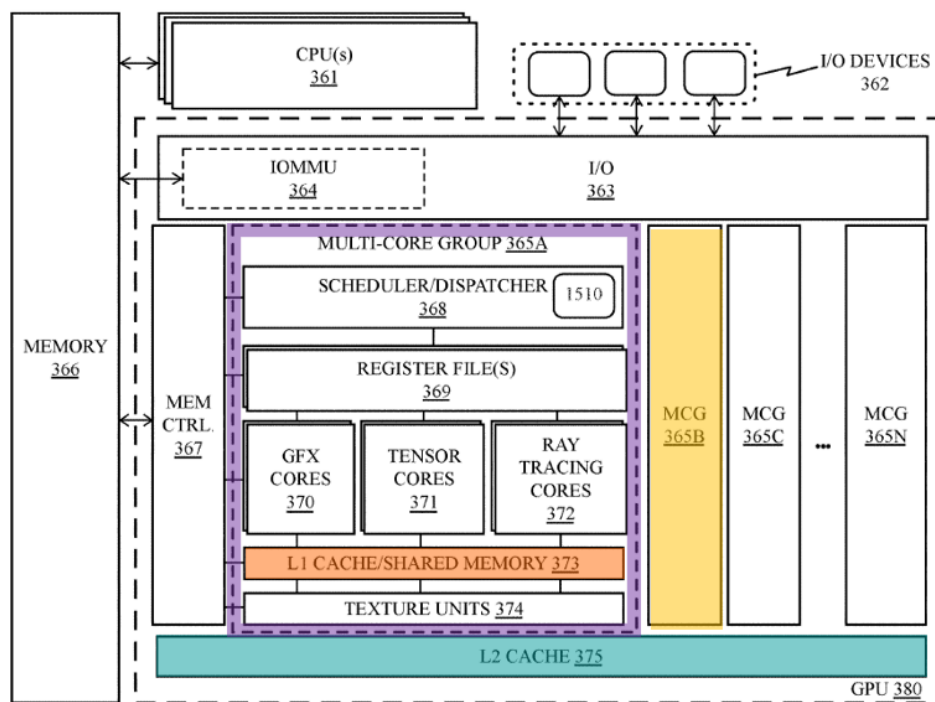


FIG. 3C

EX1031-Koker, Figure 3C. EX1002-Decl., ¶¶296-297.

Namely, the cores 370-372 execute graphics “threads”—sets of graphics computer instructions. EX1031-Koker, 13:43-47 (“[A] multi-core group 365A may

include a set of graphics cores [GFX] 370, a set of tensor cores 371, and a set of ray tracing cores 372. A scheduler/dispatcher 368 schedules and dispatches the graphics threads for execution on the various cores 370, 371, 372.”). Moreover, the L1 cache/shared memory 373 is a local to the MCG 365A and stores graphics data used by the MCG 365A’s cores 370-372 while they execute threads. EX1031-Koker, 13:56-59 (“[L]evel 1 (L1) caches and shared memory units 373 store graphics data ... *locally within each multi-core group 365A.*”) Since the multiple cores 370-372 share the local L1 cache/shared memory 373 and read/write data from/to the cache/shared memory, a POSITA would have understood that the L1 cache/shared memory is coherent within MPG 365A. EX1002-Decl., ¶298.

Indeed, Koker states that “[c]ache snooping/coherency techniques are well understood by those of skill in the art and will not be described in detail” (EX1031-Koker, 17:55-58), confirming that a POSITA would have understood standard coherency mechanisms, including that the local L1 cache is coherent within MPG 365A. EX1031-Koker at 17:47-49 (“Coherency is maintained ... in the various caches 462A-462D.”), 40:29-38 (“The cache memory hierarchy may include ... levels of shared mid-level cache, such as a Level 2 (L2), Level 3 (L3), Level 4 (L4), or other levels of cache.... *[C]ache coherency logic maintains coherency between the various cache units....*”). EX1002-Decl., ¶¶299-300.

Accordingly, Koker teaches that a first level (L1 cache/shared memory) of the first cache memory hierarchy (L1 cache/shared memory/L2 cache (modified by Sistla)) is coherent within the MCG 365A. When Koker's teachings are incorporated into Kelleher (as explained below), the Kelleher-Collins-Koker combination teaches that a first level (Koker's L1 cache/shared memory) of the first cache memory hierarchy (Koker's L1 cache/shared memory/L2 cache (modified by Sistla)) is coherent within Kelleher's first GPU chiplet 220 (corresponding to Koker's MCG 365A). EX1002-Decl., ¶¶301-303.

(c) Motivation to Combine Kelleher, Collins, and Koker

A POSITA would have been motivated, with a reasonable expectation of success, to incorporate Koker's cache memory hierarchy teachings into the Kelleher-Collins system. EX1002-Decl., ¶¶304-305.

First, Kelleher, Collins, and Koker are analogous art to each other and the challenged claims because they all relate to communicably coupling discrete/dedicated chiplets/multi-core groups within a GPU. Section IX.A.1.(e) [MTC-Kelleher-Collins] (explaining Kelleher, Collins, and challenged claims relate to communicably coupling discrete chiplets).

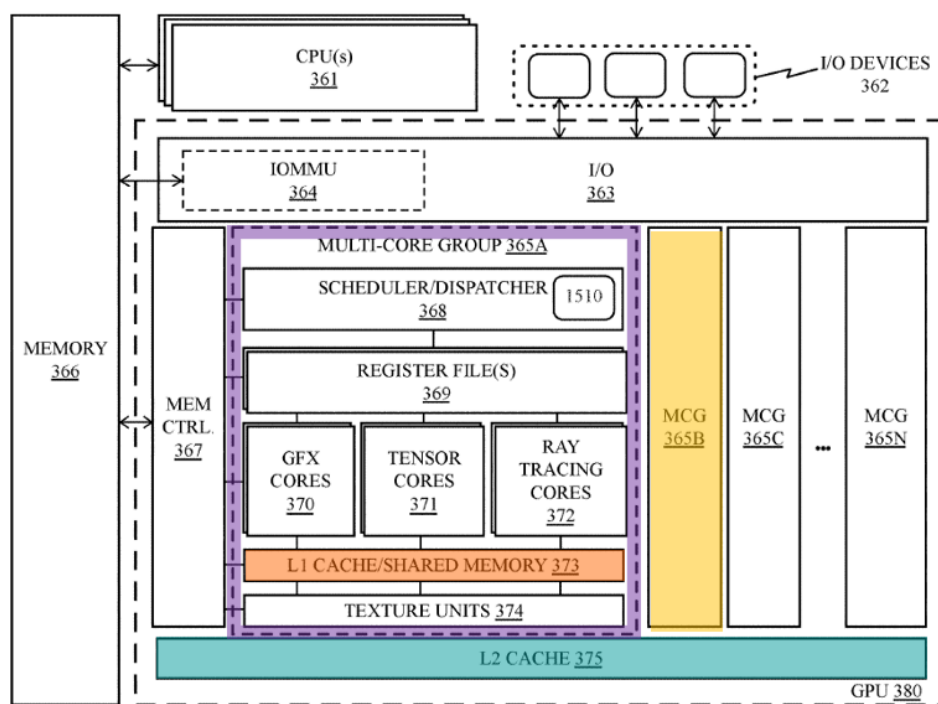


FIG. 3C

EX1031-Koker, Figure 3C (above), 12:21-24 (“FIG. 3C illustrates a ... [GPU] 380 which includes ... multi-core groups 365A-365N.”), 13:36-42, 13:61-66. EX1002-Decl., ¶¶306-308.

Second, a POSITA would have recognized that Kelleher’s and Koker’s systems divide processing resources into chiplets/multi-core groups to perform GPU operations in parallel and improve overall performance. EX1002-Decl., ¶309.

For example, in Figure 2 (below), Kelleher’s system includes a **chiplet 220** (with processing module 265) and a **chiplet 240** (with processing module 285) that execute tasks in parallel to improve GPU performance.

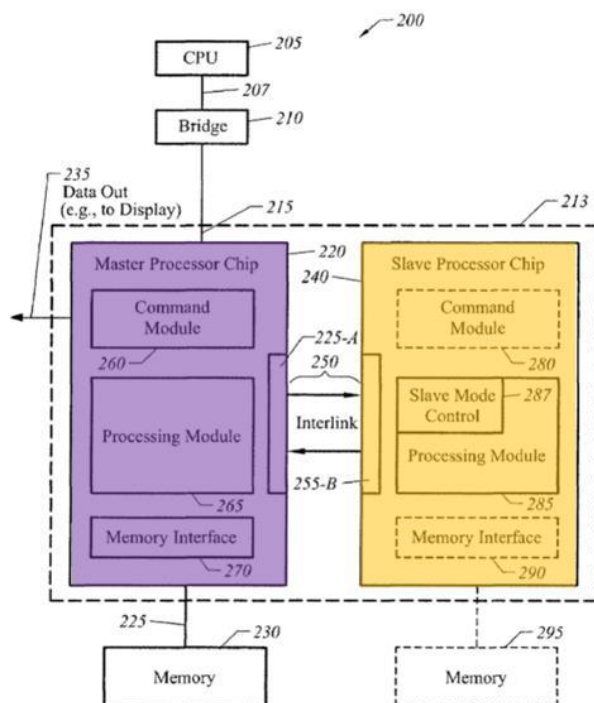


FIG. 2

EX1004-Kelleher, Figure 2, 4:3-19 (“GPU chip 220 ... *utilizes processing module 285 [of GPU chip 240] as an additional processing resource that is used in parallel with processing module 265 [of GPU chip 220]*.... [A]bout half of the processing that would ordinarily be performed by processing module 265 is off-loaded to processing module 285, *permitting an approximately two-fold improvement in performance....*”). EX1002-Decl., ¶¶310-311.

Kelleher’s Figure 8 (below) shows additional details of the processing resources divided between the GPU chip 720 and GPU chip 740.

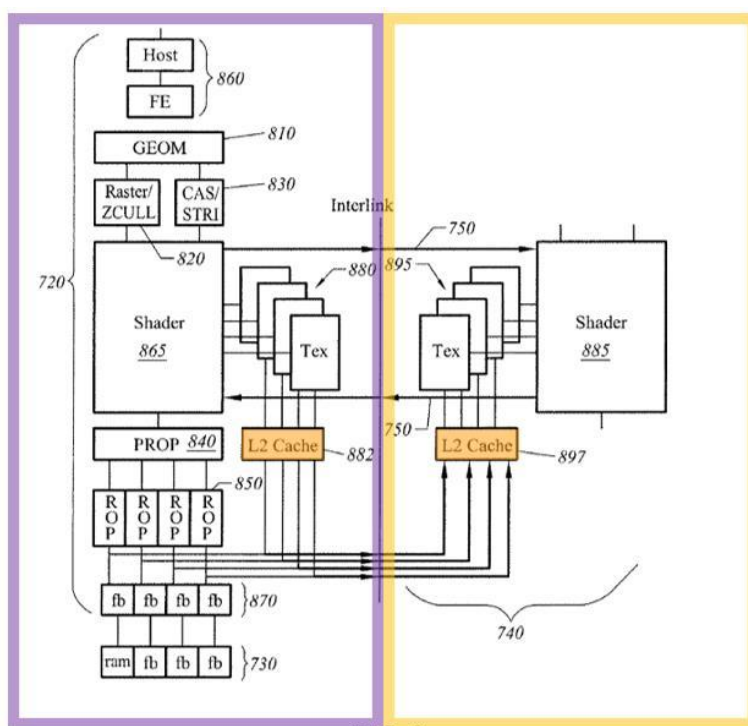


FIG. 8

EX1004-Kelleher, Figure 8. Each of the GPU chips 720 and 740 has processing resources including texture modules 880/895, a shader stage 865/885, and L2 cache 882/897. EX1004-Kelleher, 7:9-47. EX1002-Decl., ¶¶312-313.

Similarly, in Figure 3C (below), Koker's system includes first and second MCGs 365A and 365B, each having dedicated processing resources for executing tasks in parallel to improve GPU performance.

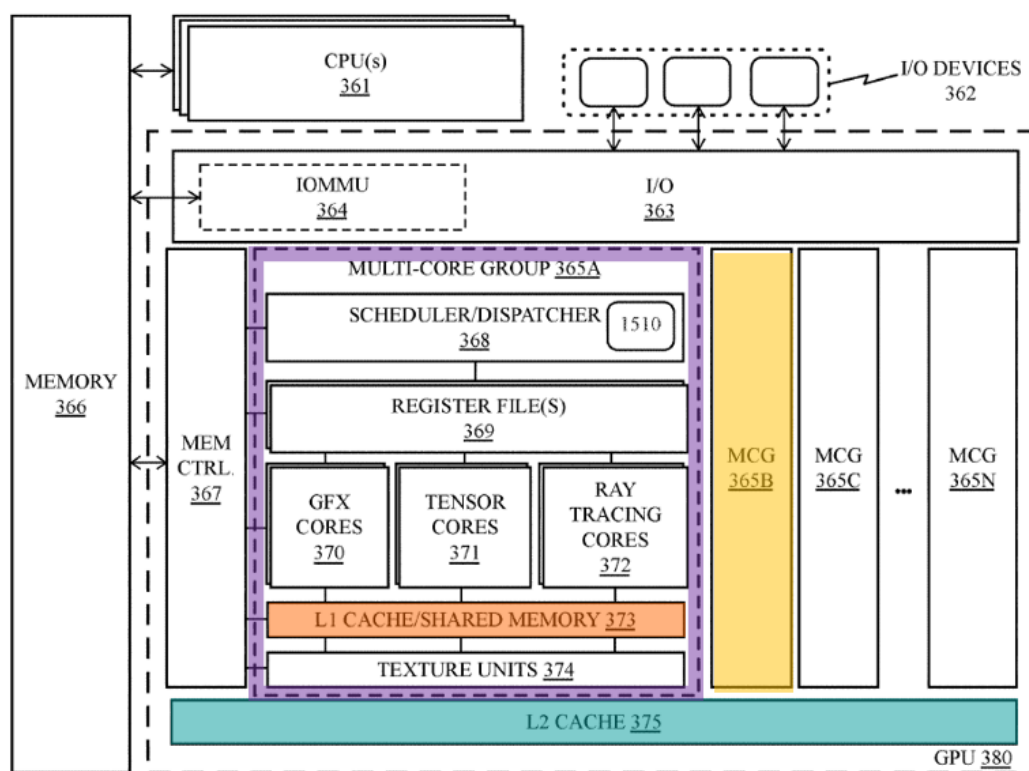


FIG. 3C

EX1031-Koker, Figure 3C, 12:21-27 (“FIG. 3C illustrates a ... [GPU] 380 which includes *dedicated sets of graphics processing resources arranged into multi-core groups 365A-365N*.... [T]he multi-core groups 365A-365N can be ... *capable of simultaneous execution of a large number of execution threads.*”). Specifically, as explained for Element [8.b], each of the first and second MCGs 365A and 365B has processing resources including cores 370-372 and an *L1 cache/shared memory 373*. Element [8.b]; EX1031-Koker, 13:43-61. EX1002-Decl., ¶¶314-315.

Third, a POSITA would have been motivated to incorporate Koker’s “cache memory hierarchy” teachings into Kelleher’s system to improve Kelleher’s system’s

parallel execution of GPU tasks and its overall performance. EX1002-Decl., ¶¶312-325.

Although Figure 8 of Kelleher (below) shows that the GPU chip and GPU chip each includes a local L2 cache, Kelleher provides no detail regarding how these caches operate, interact with other caches, integrate into a cache hierarchy system, and maintain coherency.

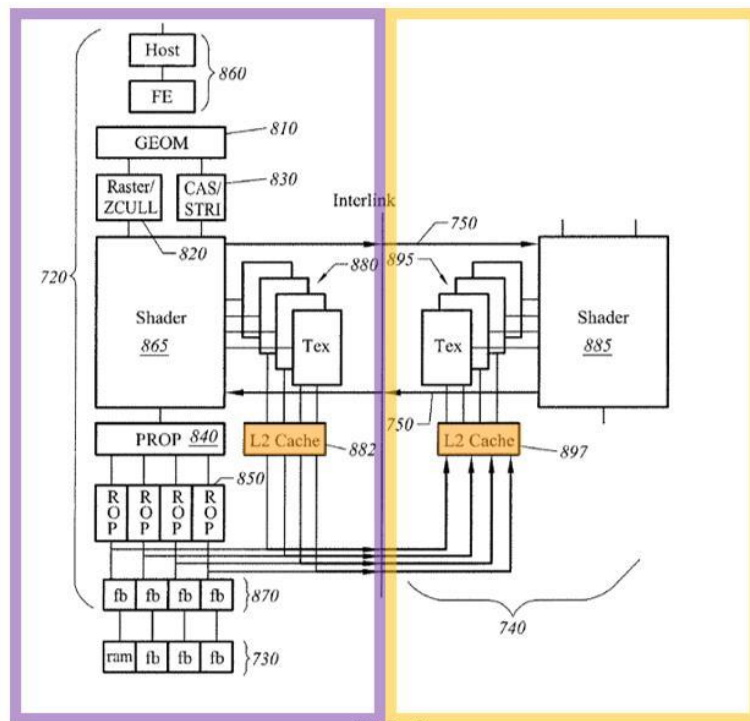
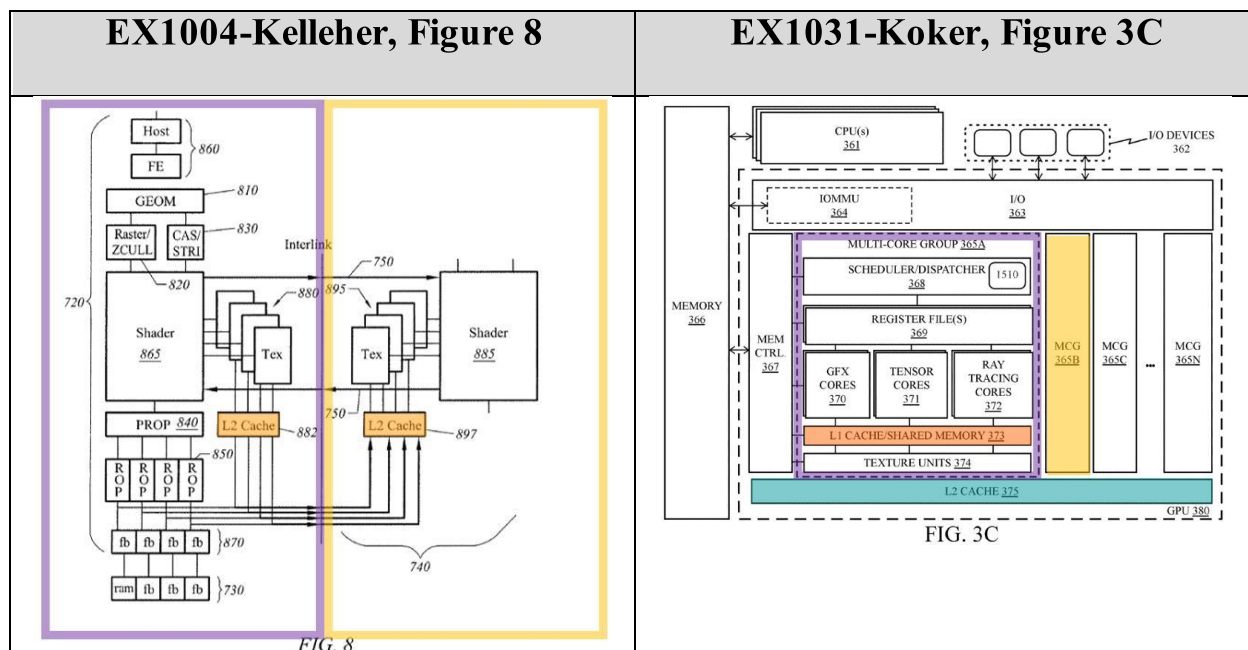


FIG. 8

EX1004-Kelleher, Figure 8, 7:9-47 (briefly mentioning “cache” four times.), 1:6-7:8, 7:48-11:9 (not mentioning “cache.”). EX1002-Decl., ¶¶315-316.

Accordingly, a POSITA would have looked to other sources—such as Koker—for guidance on implementing an effective cache hierarchy system. EX1002-Decl., ¶317.

Like Kelleher's GPU chip and GPU chip, each with local cache (Figure 8 below), Figure 3C of Koker (below) discloses that its first MCG and second MCG each has local cache.



EX1004-Kelleher, Figure 8, 7:9-34; EX1031-Koker, Figure 3C, 13:56-59 (“[L]evel 1 (L[1]) caches and shared memory units 373 store graphics data ... *locally* within each multi-core group 365A.... “), 13:38-42. EX1002-Decl., ¶¶314-315.

While Kelleher does not disclose how its local cache integrates with other caches in a cache memory hierarchy, Koker does. Koker teaches that each MCG (e.g., first MCG) uses its local cache to store data used *locally within the MCG* and that a higher-level cache—the L2 cache—is *shared among all MCGs* (e.g. first and second MCGs) so that the MCGs can execute multiple concurrent software instruction threads. EX1031-Koker, 13:56-66 (“[L]evel 1 (L1) caches and shared

memory units 373 store graphics data ... *locally within each multi-core group 365A* [e.g. MCG 365A or 365B].... A Level 2 (L2) cache 375 *shared by all ... multi-core groups 365A-365N* stores graphics data and/or instructions for multiple concurrent graphics threads....”), 17:37-40, 40:29-38 (“The *cache memory hierarchy* may include ... *levels of shared mid-level cache*, such as a Level 2 (L2), Level 3 (L3), Level 4 (L4), or other levels of cache....”). EX1002-Decl., ¶316.

A POSITA would have understood that, by sharing a higher-level L2 cache among Kelleher’s chiplets (as Koker teaches), Kelleher’s chiplets can efficiently share the latest, up-to-date data among the chiplets, which enables the chiplets to simultaneously process software instruction threads more quickly. EX1002-Decl., ¶¶58-60, 317-321.

Koker further teaches how to keep caches—e.g., its local and shared caches—“coherent” to ensure that data stored in the caches remains valid and up-to-date, even when different components (*i.e.*, Kelleher’s processing resources within one chiplet or Kelleher’s multiple chiplets) read/write data from/to the same cache memory locations. EX1031-Koker, 17:47-58 (“Coherency is maintained for data and instructions stored in the various caches 462A-462D, 456.... [E]ach cache may have cache coherency logic/circuitry associated therewith to communicate to over the coherence bus 464 in response to detected reads or writes to particular cache lines....

[A] cache snooping protocol is implemented over the coherence bus 464 to snoop cache accesses.”). EX1002-Decl., ¶322.

Koker’s teachings are directly applicable to Kelleher’s chiplets, which like Koker’s MCGs, are dedicated, modular processing units that would benefit from both local and shared caches to coordinate data across Kelleher’s entire system. By incorporating Koker’s cache memory hierarchy into Kelleher’s chiplet-based architecture, a POSITA would have achieved several benefits: higher cache capacity, reduced latency for local workloads, and reliable data sharing across processing units. *See generally* EX1048-Nayyar; EX1049-Balasubramonian. EX1002-Decl., ¶323.

Fourth, a POSITA would have had a reasonable expectation of success of implementing Koker’s cache memory hierarchy within Kelleher’s system. Both Kelleher’s master and slave GPU chips and Koker’s first and second MCGs contain dedicated processing resources for simultaneously executing graphics tasks in parallel. Koker explains how to implement a cache hierarchy using a local cache within each chiplet/MCG and a shared cache for all chiplets/MCGs. EX1031-Koker, Figure 3C, 13:56-66. Further, as Koker notes, implementing cache coherency techniques into Kelleher’s system was routine and well-within a POSITA’s skill. EX1031-Koker 17:55-58. EX1002-Decl., ¶324.

Accordingly, a POSITA would have been motivated to incorporate Koker's teachings into the Kelleher-Collins combination, because applying Koker's cache memory hierarchy teachings to the Kelleher-Collins combination would provide cache coherency and improved performance across chiplets, ensuring that information stored in memory remains accurate and can be shared across chiplets despite the chiplets' physical separation. The combination would have been a predictable and advantageous design choice, with a reasonable expectation of success, yielding a more coherent, higher-performing multi-chiplet GPU. EX1002-Decl., ¶325.

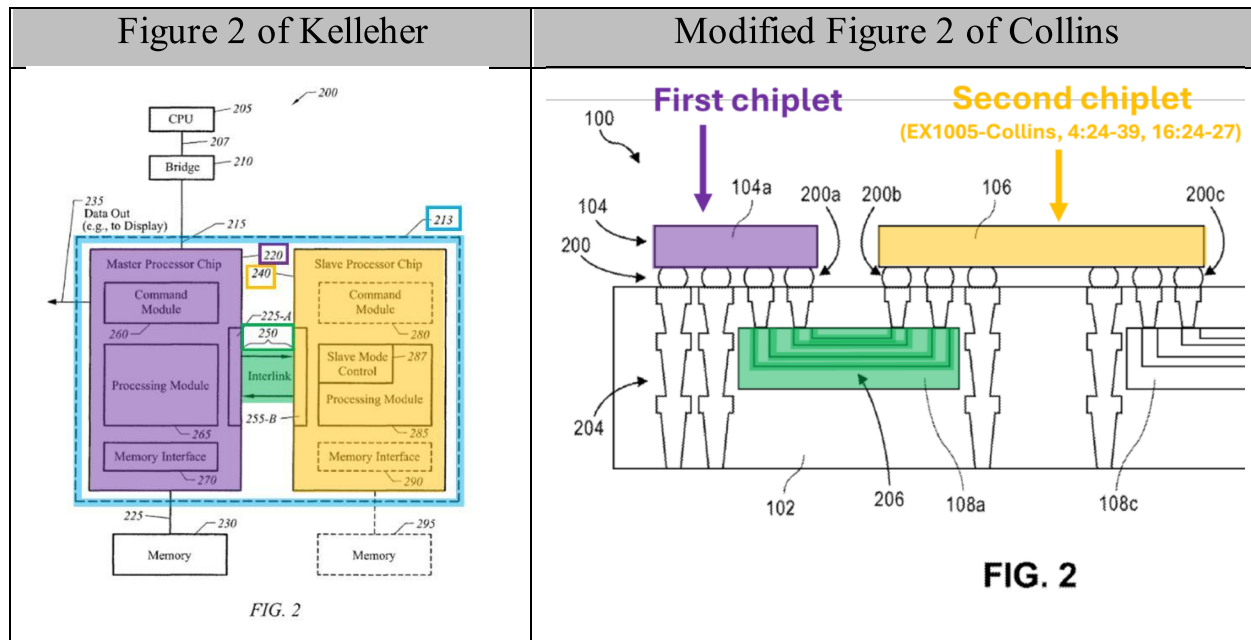
(d) Motivation to Combine Kelleher, Collins, Koker, and Sistla

A POSITA would have been motivated, with a reasonable expectation of success, to incorporate Sistla's teachings of a "distributed shared cache architecture" into the Kelleher-Collins-Koker system. EX1002-Decl., ¶326.

A POSITA would have been motivated to combine Kelleher's and Collins's teachings by communicably coupling Kelleher's GPU chips 220 and 240 (first and second GPU chiplets) with Collins's interconnect structure/EMIB (passive crosslink). Section IX.A.1.(e) [MTC-Kelleher-Collins]. A POSITA also would have been motivated to incorporate Koker's cache memory hierarchy teachings into the Kelleher-Collins system. Section IX.B.1.(c) [MTC-Kelleher-Collins-Koker]. EX1002-Decl., ¶327.

A POSITA would have further been motivated to combine the Kelleher's, Collins's, Koker's, and Sistla's teachings, because the references are analogous art to each other and the challenged claims since each of them relates to communicably coupling chiplets/dies/MCGs within a GPU/CPU. EX1002-Decl., ¶¶328-329.

Kelleher and Collins teach communicably coupling chiplets as explained in Section IX.A.1.(e) [MTC-Kelleher-Collins] and for Elements 1[c]-1[d]. Moreover, as shown in Kelleher's Figure 2 and Collins's Modified Figure 2 below, the Kelleher-Collins combination discloses communicably coupling **first** and **second** chiplets via a **passive crosslink** that is dedicated for inter-chiplet communications. See Elements 1[c]-1[d]; Section IX.A.1.(e) [MTC-Kelleher-Collins].



EX1004-Kelleher, Figure 2; EX1005-Collins, Modified Figure 2 (replacing memory device 106 with chiplet (4:24-39 and 16:24-27)). EX1002-Decl., ¶330.

As shown below, Koker's Figure 3C discloses communicably coupling **first** and **second** MCGs 365A-365B (corresponding to **first** and **second** chiplets).

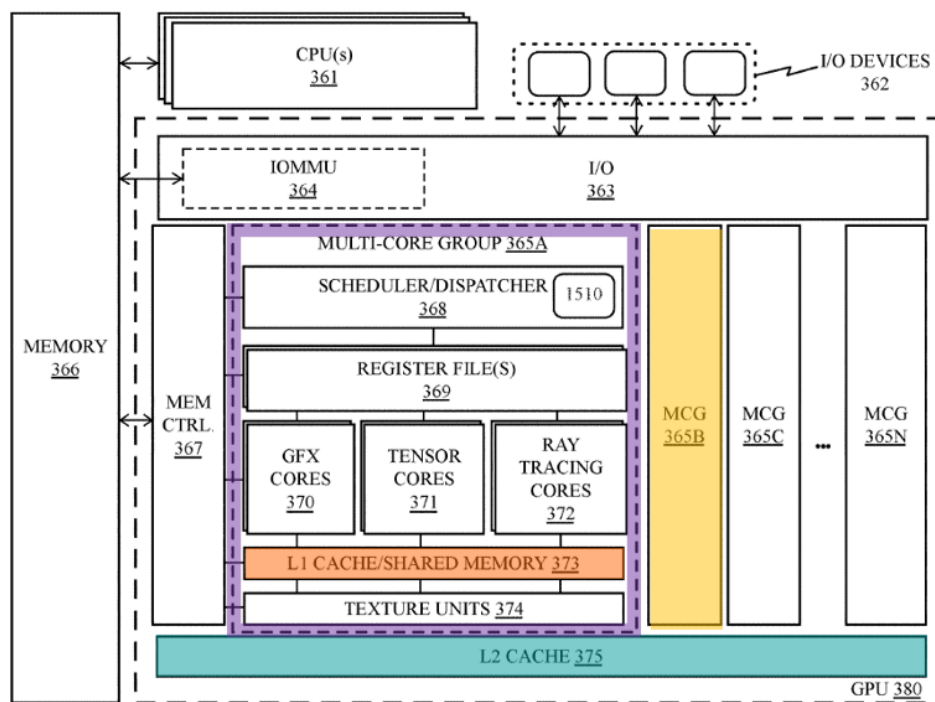


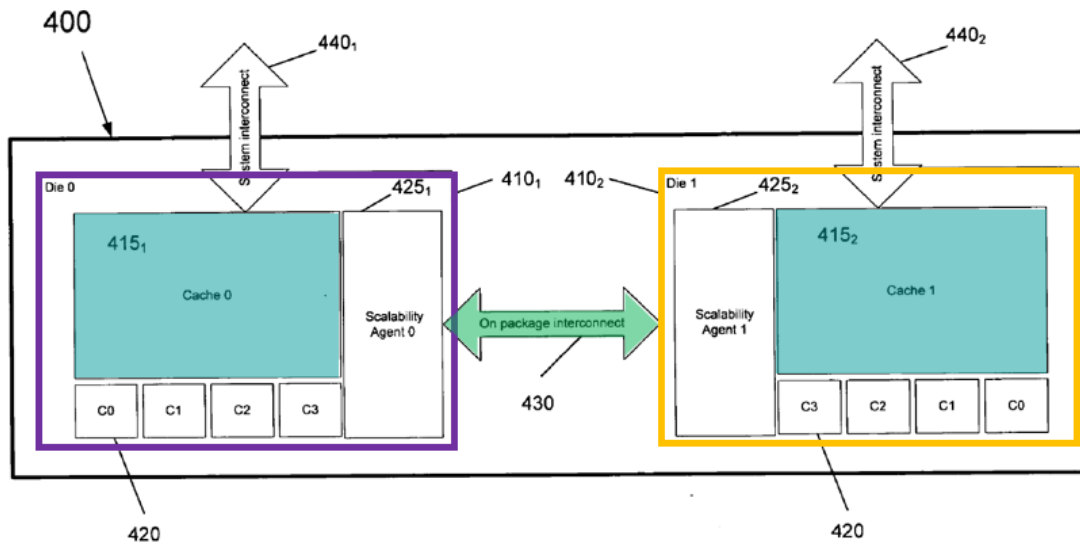
FIG. 3C

EX1031-Koker, Figure 3C. The **first** and **second** MCGs 365A-365B share the same **L2 cache 375** and thus, are at least indirectly communicably coupled. EX1031-Koker, 13:56-14:2 (“[T]he L2 cache 375 may be shared across ... multi-core groups 365A-365N.”). EX1002-Decl., ¶331.

As explained for Elements 8[a]-8[b] and in Section IX.B.1.(c) [MTC-Kelleher-Collins-Koker], Koker teaches implementation details for a cache memory hierarchy. Moreover, in light of Sistla's teachings, a POSITA would have motivated to more efficiently incorporate Koker's **last level L2 cache** (which is outside Koker's

first and second MCGs 365A-365B,) into the first and second chiplets of the Kelleher-Collins combination. EX1002-Decl., ¶¶331-333.

For example, Sistla's Figure 2 below teaches a first die/chiplet 410₁ and a second die/chiplet 410₂ communicably coupled via a point-to-point (PTP) interconnect.



EX1047-Sistla, Figure 2, [0013] (“[P]rocessor 400 ... includes a first die 410₁ and a second die 410₂ connected by ... point-to-point (PTP) interconnect.”). As explained above for Elements 8[a]-8[b], Sistla teaches incorporating first and second portions 415₁ and 415₂ of the shared LLC in the first and second die/chiplets, respectively. EX1047-Sistla, [0011]-[0013]; Elements 8[a]-8[b]. EX1002-Decl., ¶¶334-335.

Per Sistla's teachings, a POSITA would have been motivated, with a reasonable expectation of success, to modify Koker's cache memory hierarchy by

including portions of Koker's last level L2 cache in the first and second chiplets of the Kelleher-Collins combination. EX1002-Decl., ¶¶58-60, 336.

Incorporating Sistla's "aggregated core-cache cluster" teachings into the Kelleher-Collins-Koker system provides many advantages. Sistla's configuration increases the system's scalability because core-cache clusters (chiplets) can be added to the system without degrading the system's operation. EX1047-Sistla, [0009] ("[T]he 'scalability agent protocol' ... provides "graceful" scalability where the responsibility for maintaining memory coherency is approximately shared equally among the core-cache clusters despite any increase or decrease in number.... The latencies to the cache in the core-cache cluster substantially remain unchanged as the number of cores increases."), [0010] ("[T]he scalability agent (i) aggregates core-cache clusters to appear as one caching agent, (ii) handles local cache coherence between core-cache clusters on the same IC, and (iii) support scalability so that the operations of a core-cache cluster are not significantly effected if other core-cache clusters are added."), [0011] ("The architecture of core-cache clusters 310 may be ... a *distributed shared cache architecture*.... According to this architecture, *multi-core processor 300 enables the latency of a first shared cache 340₁ to remain substantially constant despite increases in the number of cores in processor 300....*"). EX1022-Milic, 129-130; EX1033-Arunkumar, *4-*5. EX1002-Decl., ¶¶58-60, 339.

Sistla teaches that, by including portions of an LLC within different dies/chiplets to create an “aggregated core-cache cluster” configuration, a device (e.g., a CPU) coupled to the aggregated core-cache clusters views them as a single caching agent. EX1047-Sistla, [0014] (“[T]he aggregation [of “core-cache clusters”] *appears as a single caching agent to devices* communicatively coupled to a system interconnect.”), Figure 2 (system interconnect 440), [0002] (“Performance of a multi-core processor may be optimized by ... *ensuring that aggregated processor cores appear as one caching agent in order to avoid scalability issues and interconnect saturation.*”), [0010] (“[T]he scalability agent [] ... *aggregates core-cache clusters to appear as one caching agent....*”). This further helps the CPU view the aggregated system as a single, monolithic device, which as explained above, facilitates programming the CPU. EX1002-Decl., ¶337.

Kelleher discloses a similar objective: configuring its **first** and **second** GPU chips **220** and **240** so that a **CPU’s** driver software perceives the two chiplets as a single GPU.

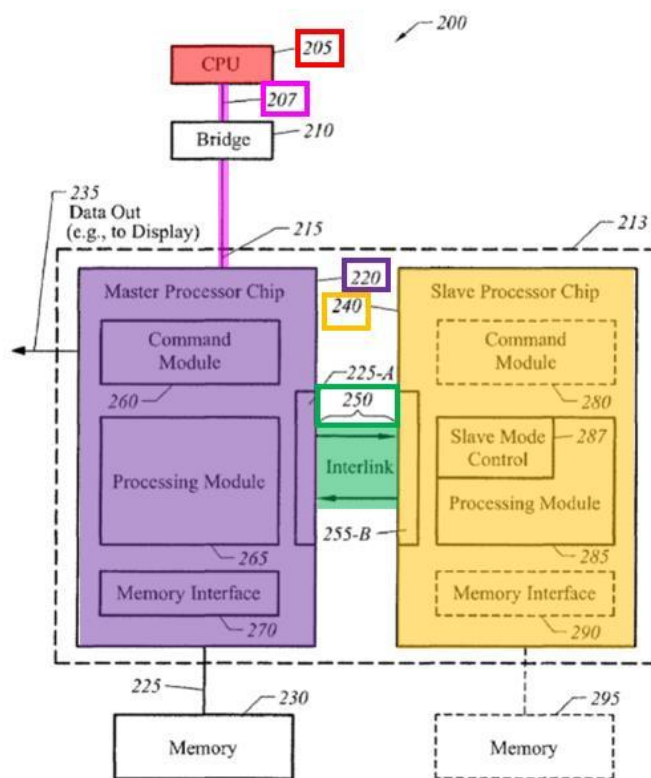


FIG. 2

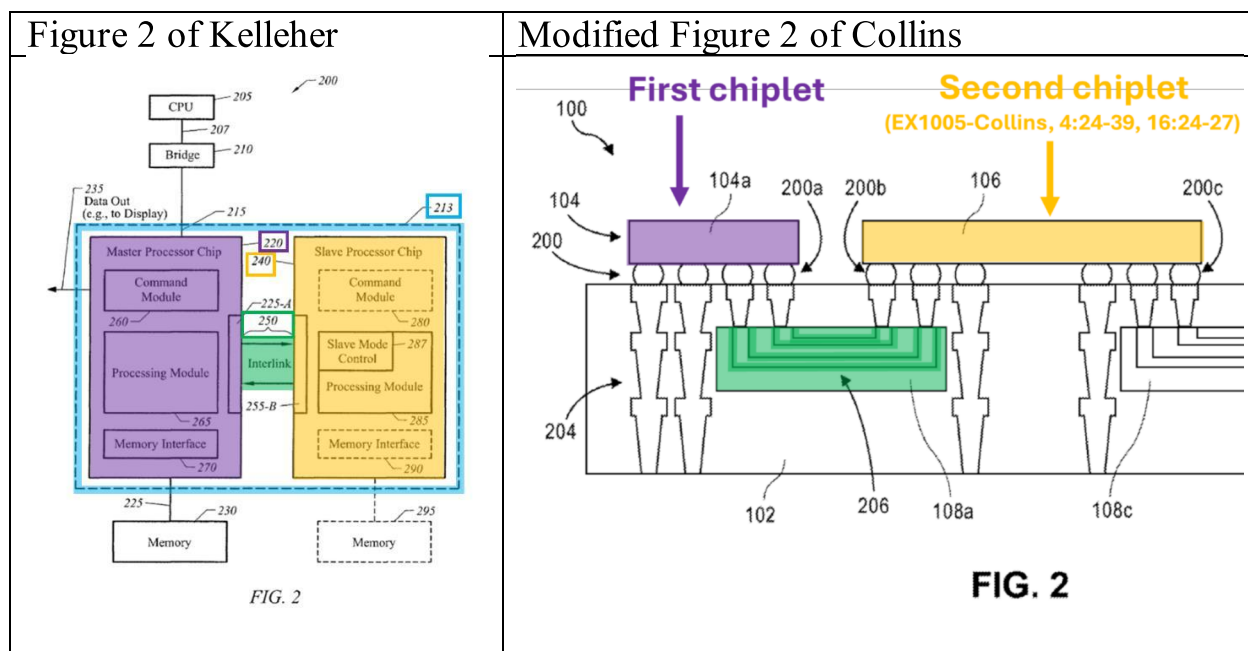
EX1031-Kelleher, Figure 2, 4:63-5:19 (“[O]nly the master GPU 220 receives graphics instructions from CPU 205. As result, replacing a single GPU chip with ... a master/slave pair of GPUs is a seamless change to driver software within CPU 205. *That is, the driver software perceives the master/slave pair as a single GPU.*”).
EX1002-Decl., ¶338.

Additionally, providing portions of Koker’s last level L2 cache within the Kelleher-Collins combination’s first and second chiplets, instead of using an external LLC, reduces the components that are coupled to the chiplets with relatively

large conductors/connectors. Thus, the overall size of the system would be reduced.

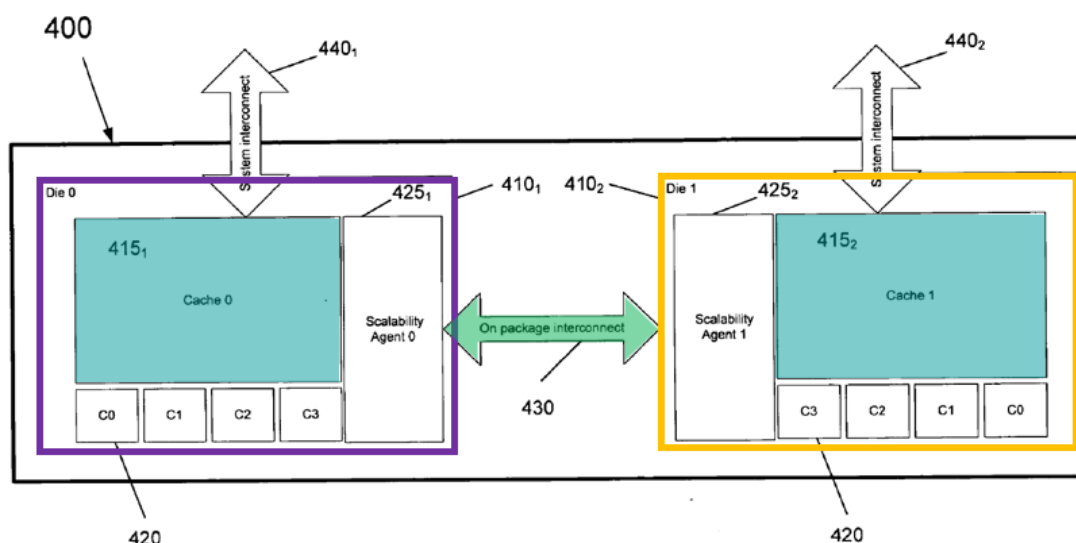
EX1002-Decl., ¶340.

A POSITA would have had a reasonable expectation of success to modify the Kelleher-Collins-Koker system by providing **first and second portions** of Koker's LLC (L2 cache) within the system's **first** and **second** chiplets, respectively, per Sistla's teachings. As illustrated below, the Kelleher-Collins combination teaches a **first** and **second** chiplet communicably coupled via a dedicated **passive crosslink**.



EX1004-Kelleher, Figure 2; EX1005-Collins, Modified Figure 2; Elements 1[c]-1[d]; Section IX.A.1.(e) [MTC-Kelleher-Collins]. EX1002-Decl., ¶341.

Figure 2 of Sistla (below) similarly shows “a **point-to-point interconnect** communicably” coupling **first** and **second** dies/chiplets.



EX1047-Sistla, Figure 2, [0013]. Moreover, Sistla further provides implementation details an aggregated core-cluster system design where **first** and **second** dies/chiplets contain portions of shared LLC and discloses advantages of this configuration, as explained above. EX1002-Decl., ¶342.

A POSITA would have understood how to include portions of the shared LLC in the **first** and **second** dies/chiplets without undue experimentation and doing so would have been well within the skill of a POSITA. EX1002-Decl., ¶343.

Thus, a POSITA would have been motivated, per Sista's teachings, to incorporate Koker's LLC in Kelleher-Collins system's **first** and **second** chiplets, because doing so would have yielded predictable results that achieve Kelleher's performance objectives, and such combination would have been a routine design choice within a POSITA's skill. EX1002-Decl., ¶¶344-345.

(e) [8.c]: a second cache memory hierarchy at the second GPU chiplet,

The Kelleher-Collins-Koker-Sistla combination teaches a second cache memory hierarchy (Koker's L1 cache/shared memory and L2 cache (modified by Sistla)) at Kelleher's second GPU chiplet 240 (corresponding to Koker's MCG 365B).

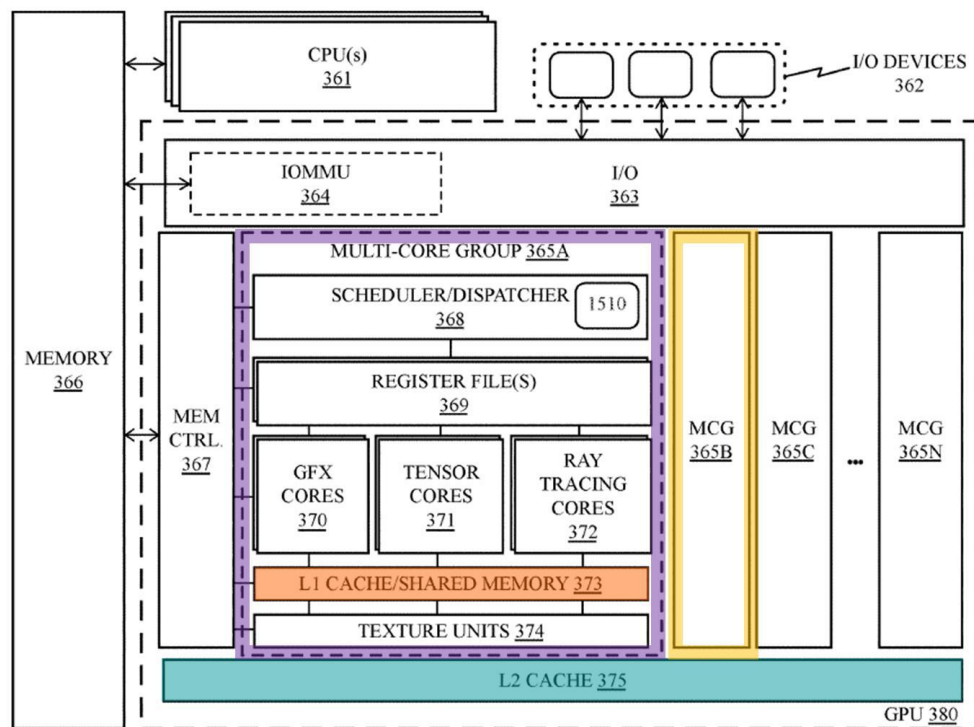


FIG. 3C

EX1031-Koker, Figure 3C, 13:39-51 (“While the details of only a single multi-core group 365A are provided, ... the other multi-core groups 365B-365N may be equipped with the same ... resources.”); Element [8.a], Section IX.B.1.(d) [MTC-Kelleher-Collins-Koker-Sistla]. EX1002-Decl., ¶¶346-350.

- (f) **[8.d]: wherein a first level of the second cache memory hierarchy is coherent within the second GPU chiplet.**

The Kelleher-Collins-Koker-Sistla combination teaches that a first level (Koker's L1 cache/shared memory) of the second cache memory hierarchy (Koker's L1 cache/shared memory/L2 cache (modified by Sistla)) is coherent within Kelleher's second GPU chiplet 240 (corresponding to Koker's MCG 365B). EX1031-Koker, Figure 3C, 13:39-51 (quoted above); Element [8.b]. EX1002-Decl., ¶¶351-354.

2. Claim 9

- (a) **[9.a]: The system of claim 8, further comprising: a unified cache memory including both a last level of the first cache memory hierarchy and a last level of the second cache memory hierarchy,**

The Kelleher-Collins-Koker-Sistla combination teaches a unified cache memory (Koker's L2 cache (modified by Sistla)) including both:

- a last level (Koker's L2 cache (modified by Sistla)) of the first cache memory hierarchy (Koker's L1 cache/shared memory (in MCG 365A)/L2 cache (modified by Sistla)) used by Kelleher's chip 220 (corresponding to Koker's MCG 365A), and
- a last level (Koker's L2 cache (modified by Sistla)) of the second cache memory hierarchy (Koker's L1 cache/shared memory (in MCG 365B)/L2

cache (modified by Sistla)) used by Kelleher's chip 240 (corresponding to Koker's MCG 365B).

EX1002-Decl., ¶¶355-359.

As background, in Figure 3 of the '803 patent, the L3 cache is the last level of the first and second cache memory hierarchies (L1 cache/L2 cache/L3 cache) at the first and second chiplets 106-1 and 106-2, respectively.

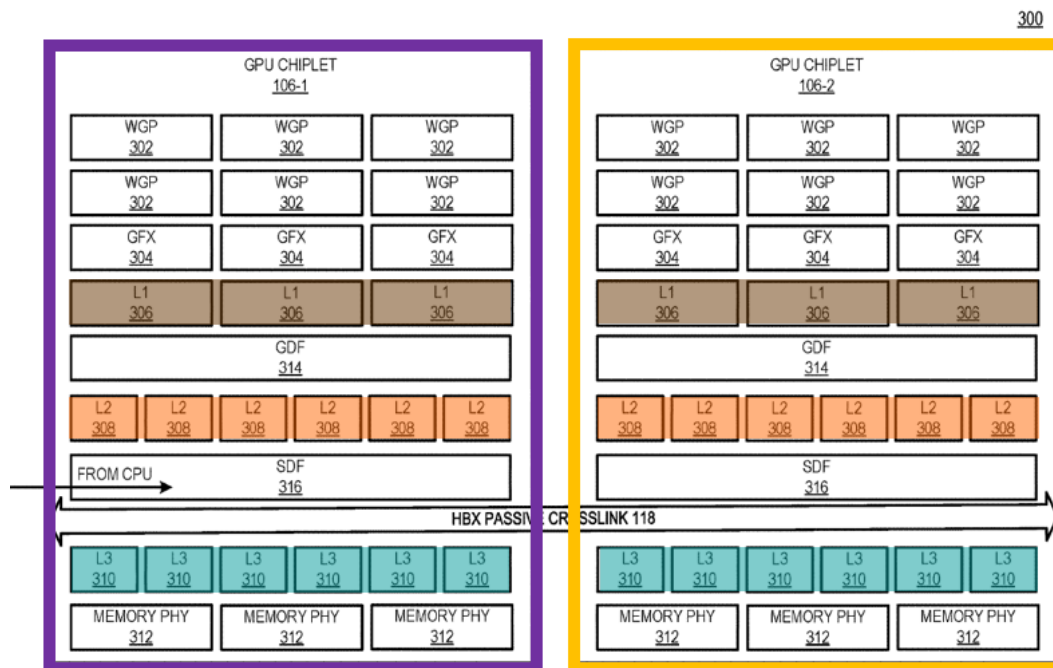


FIG. 3

EX1001-'803, Figure 3, 5:1-18 ("FIG. 3 ... illustrat[es] a *cache hierarchy of GPU chiplets*.... Each of the GPU chiplets 106-1, 106-2 include ... L1 cache memory 306. Each GPU chiplet 106 also includes ... L2 cache memory 308 ... [and] ... L3 cache memory 310 channels.... [T]he L3 level (or other last level) of cache is

unified and coherent across all of the GPU chiplets 106.”). EX1002-Decl., ¶¶360-362.

Since the L3 cache is shared by and coherent across the first and second chiplets 106-1 and 106-2, it is a “unified” cache memory. EX1001-’803, 5:16-18 (“[T]he L3 level (or other last level) of cache is unified and coherent across all of the GPU chiplets 106.”), 2:18-26 (“[T]he chiplet-based GPU architecture positions those physical resources (e.g., LLC) on different dies and *communicably couples those physical resources such that the LLC level is unified and remains cache coherent across all GPU chiplets.*”). EX1002-Decl., ¶363.

Like the ’803 patent, Figure 3C of Koker shows that the L2 cache is the LLC of the first memory hierarchy (L1 cache/shared memory 373/L2 cache) at the MCG 365A and is the LLC of the second memory hierarchy (L1 cache/shared memory 373/L3 cache) at MCG 365B.

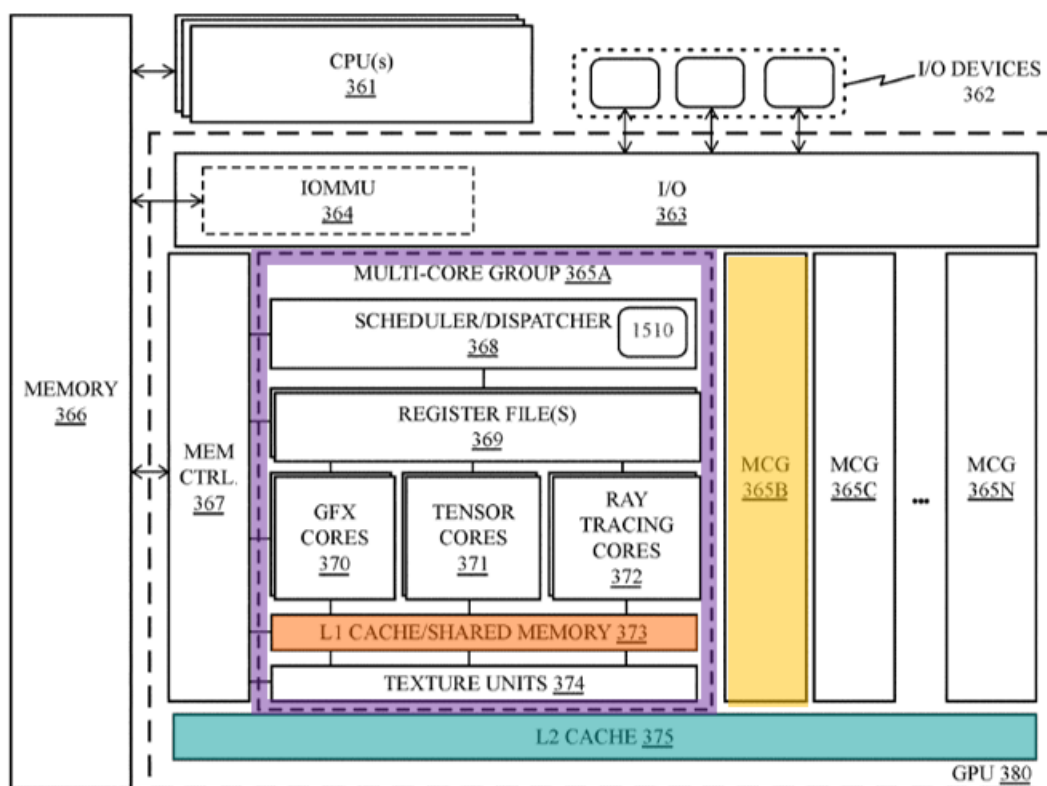


FIG. 3C

EX1031-Koker, Figure 3C (showing L2 cache as LLC for MCG 365A, MCG 365B), 13:38-54, 13:64-14:2, 40:29-42; Element [8.a] (explaining Koker’s L2 cache is LLC). EX1002-Decl., ¶¶364-365.

Since Koker’s L2 cache is shared by and coherent across the MCG 365A and MCG 365B, a POSITA would have known that it is a “unified” cache memory. EX1031-Koker, 13:55-66 (“A Level 2 (L2) cache 375 shared by all ... of the multi-core groups 365A-365N....”), 17:43-58 (“Cache snooping/coherency techniques are well understood by those of skill in the art....”); Elements [8a]-[8b], [8d]. EX1002-Decl., ¶366.

A POSITA would have further understood that Koker's L2 cache (modified by Sistla) is a "unified" cache memory because it is located on different MCGs 365A and 365B, shared by the MCGs 365A and 365B, and communicably couples resources of the MCGs 365A and 365B. Elements 8[a], 8[c]; Section IX.B.1.(d) [MTC-Kelleher-Collins-Koker-Sistla]; *compare* EX1001-'803, 2:18-24 ("[T]he chiplet-based GPU architecture positions those physical resources (e.g., LLC) on different dies [chiplets 106-1 and 106-2] and communicably couples those physical resources such that the LLC level is unified and remains cache coherent across all GPU chiplets."). EX1002-Decl., ¶¶366-367.

Accordingly, the Kelleher-Collins-Koker-Sistla combination teaches a unified cache memory (Koker's L2 cache (modified by Sistla)) including both a last level (Koker's L2 cache (as modified by Sistla)) of the first cache memory hierarchy (Koker's L1 cache/shared memory/L2 cache (modified by Sistla)) used by Kelleher's chip 220 (corresponding to Koker's MCG 365A) and a last level (Koker's L2 cache (modified by Sistla)) of the second cache memory hierarchy (Koker's L1 cache/shared memory/L2 cache (modified by Sistla)) used by Kelleher's chip 240 (corresponding to Koker's MCG 365B). EX1002-Decl., ¶¶368-370.

- (a) [9.b] wherein the unified cache memory is coherent across all GPU chiplets of the GPU chiplet array.

The Kelleher-Collins-Koker-Sistla combination teaches that the unified cache memory (Koker's L2 cache (modified by Sistla)) is coherent across all GPU chiplets

(Kelleher's chips 220 and 240 (corresponding to Koker's MCGs 365A and 365B)) of the GPU chiplet array (Kelleher's module 213). EX1031-Koker, 13:55-66, 17:43-58; EX1047-Sistla, [0021]-[0042]; Elements [9.a], [8.a]-[8.d]. EX1002-Decl., ¶¶371-375.

X. CONCLUSION

Petitioner respectfully requests institution of IPR.

Respectfully submitted,

/Grant K. Rowan/
Grant K. Rowan
Registration No. 41,278

Table of Claims

Claim 1	
Limitation	Claim Language
[1.pre]	A system, comprising:
[1.a]	a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array,
[1.b]	wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and
[1.c]	[wherein the GPU chiplet array includes:] a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink,
[1.d]	wherein the passive crosslink is dedicated for inter-chiplet communications.

Claim 2
The system of claim 1, wherein the passive crosslink comprises a passive interposer die.

Claim 3	
Limitation	Claim Language
[3.a]	The system of claim 1, wherein the first GPU chiplet comprises a first PHY region including conductor structures for chiplet-to-chiplet communications,
[3.b]	and further wherein the second GPU chiplet comprises a second PHY region including conductor structures for chiplet-to-chiplet communications.

Claim 4	
Limitation	Claim Language
[4.a]	The system of claim 3, further comprising: a third GPU chiplet communicably coupled to the first GPU chiplet via the passive crosslink,
[4.b]	wherein the third GPU chiplet comprises a third PHY region including conductor structures for chiplet-to-chiplet communications.

Claim 7
The system of claim 1, wherein the passive crosslink communicably couples all GPU chiplets in the GPU chiplet array.

Claim 8	
Limitation	Claim Language
[8.a]	The system of claim 1, further comprising: a first cache memory hierarchy at the first GPU chiplet,
[8.b]	wherein a first level of the first cache memory hierarchy is coherent within the first GPU chiplet; and
[8.c]	a second cache memory hierarchy at the second GPU chiplet,
[8.d]	wherein a first level of the second cache memory hierarchy is coherent within the second GPU chiplet.

Claim 9	
Limitation	Claim Language
[9.a]	The system of claim 8, further comprising: a unified cache memory including both a last level of the first cache memory hierarchy and a last level of the second cache memory hierarchy,
[9.b]	wherein the unified cache memory is coherent across all GPU chiplets of the GPU chiplet array.

Claim 10
The system of claim 1, further comprising: a plurality of conductive pillars coupling a circuit board substrate to a first non-PHY region of the first GPU chiplet and a second non-PHY region of the second GPU chiplet.

Table of Exhibits for U.S. Patent 11,841,803 Petition for *Inter Partes* Review

Exhibit	Description
1001	U.S. Patent No. 11,841,803 (“’803 patent”)
1002	Declaration of Dr. Vivek Subramanian Regarding U.S. Patent No. 11,841,803 (“Decl.”)
1003	File History for U.S. Patent No. 11,841,803 (“FH”)
1004	U.S. Patent No. 7,598,958 (“Kelleher”)
1005	U.S. Patent No. 11,610,862 (“Collins”)
1006	Coskun et. al, A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems, PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN (2018) (“Coskun”)
1007	RESERVED
1008	Jerger et al., NoC Architectures for Silicon Interposer Systems, IEEE/ACM INTERNATIONAL SYMPOSIUM ON MICROARCHITECTURE (2015) (“Jerger”)
1009	RESERVED
1010	John Lau, Heterogeneous Integrations, Springer (2019) (“Lau”)
1011	Ajaykumar Kannan, Enabling Interposer-Based Disintegration of Multi-Core Processors, University of Toronto Department of Electrical & Computer Engineering (2015) (“Kannan”)
1012-1013	RESERVED
1014	International Patent Publication No. WO2019132971 (“Elsherbini”)
1015-1019	RESERVED
1020	Dylan Stow, et. al., Cost-Effective Design of Scalable High-Performance Systems Using Active and Passive Interposers, IEEE/ACM INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN (ICCAD) (2017) (“Stow”)
1021	The Open Domain-Specific Architecture WorkGroup, A Chiplet-Based Open Architecture, Draft v0.9 (2018) (“ODSA”)

Exhibit	Description
1022	Ugljesa Milic et. al., BEYOND THE SOCKET: NUMA-AWARE GPUS, <i>in</i> PROCEEDINGS OF MICRO-50 (2017) (“Milic”)
1023	Complaint, <i>In the Matter of</i> CERTAIN INTEGRATED CIRCUITS, ELECTRONIC DEVICES CONTAINING THE SAME, AND COMPONENTS THEREOF, Inv. No. 337-TA-1450 (“ITC-Complaint”)
1024	Complaint, <i>Onesta IP, LLC v. NVIDIA Corp</i> , No. 1:25-cv-00586 (W.D. Tex.) (“District-Court-Complaint”)
1025	ITC Exhibit 50, <i>In the Matter of</i> CERTAIN INTEGRATED CIRCUITS, ELECTRONIC DEVICES CONTAINING THE SAME, AND COMPONENTS THEREOF, Inv. No. 337-TA-1450 (“ITC-Ex50-Infringement”)
1026	ITC Exhibit 19, <i>In the Matter of</i> CERTAIN INTEGRATED CIRCUITS, ELECTRONIC DEVICES CONTAINING THE SAME, AND COMPONENTS THEREOF, Inv. No. 337-TA-1450 (“ITC-Ex19-Domestic-Industry”)
1027	U.S. Patent Publication No. 2019/0123022 (“Teig”)
1028	Paul Enquist, Room Temperature Direct Wafer Bonding for Three Dimensional Integrated Sensors, SENSORS AND MATERIALS, Vol. 17, No. 6 (2005) (“Ziptronix”)
1029	Robert R. Schaller, Moore’s Law: Past, Present, and Future, IEEE SPECTRUM (June 1997) (“Schaller”)
1030	Yangdong (Steven) Deng & Wojciech Maly, 2.5D System Integration: A Design Driven System Implementation Schema, IEEE (2004) (“Deng-&-Maly”)
1031	U.S. Patent No. 10,909,652 (“Koker”)
1032	RESERVED
1033	Akhil Arunkumar et. al, MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability <i>in Proceedings of ISCA ’17</i> (June 2017) (“Arunkumar”)
1034	Dictionary of Computer Science (7th ed.), Oxford University Press (2016) (“Dictionary-of-Computer-Science”)

Exhibit	Description
1035	Dictionary of Multimedia (4th ed.), Franklin, Beedle & Associates Inc. (2005) (“Dictionary-of-Multimedia”)
1036	Joel Mearig & Bill Goers, An Overview of Manufacturing BGA Technology, IEEE/CPMT Int’l Electronics Manufacturing Technology Symposium (1995) (“Mearig-&-Goers”)
1037	U.S. Patent Publication No. 2018/0240778 (“Liu”)
1038	Farhang Yazdani, Design and Direct Assembly of 2.5D/3D Rigid Silicon Interposer on PCB, IMAPS Proceedings (2014) (“Yazdani”)
1039	RESERVED
1040	U.S. Patent Publication No. 2001/0005873 (“Yasuda”)
1041	U.S. Patent No. 10,475,147 (“Price”)
1042	U.S. Patent Publication No. 2007/0273699 (“Sasaki”)
1043	Amended Joint Proposed Claim Construction Chart, <i>In the Matter of</i> CERTAIN INTEGRATED CIRCUITS, ELECTRONIC DEVICES CONTAINING THE SAME, AND COMPONENTS THEREOF, Inv. No. 337-TA-1450 (“Claim-Construction-Chart”)
1044	Chris McClanahan, History and Evolution of GPU Architecture, Georgia Tech. (2010) (“McClanahan”)
1045	Thiruvengadam Vijayaraghavan et.al, Design and Analysis of an APU for Exascale Computing, IEEE International Symposium on High Performance Computer Architecture (2017) (“Vijayaraghavan”)
1046	NVIDIA Tesla P100, Whitepaper, The Most Advanced Datacenter Accelerator Ever Built Featuring Pascal GP100, the World’s Fastest GPU (2016) (“NVIDIA-P100”)
1047	U.S. Patent Publication No. 2009/0037658 (“Sistla”)
1048	Raman Nayyar, “Performance Analysis of a Hierarchical, Cache-Coherent, Shared Memory Based, Multi-processor System” (1993) (“Nayyar”)
1049	Rajeev Balasubramonian et. al, “Multi-Core Cache Hierarchies,” Morgan & Claypool Publishers (2011) (“Balasubramonian”)

Exhibit	Description
1050	Reza Ghaffarian, Book of Knowledge (BOK) for NASA Electronic Packaging Roadmap (2015) (“Ghaffarian”)

CERTIFICATE OF SERVICE

Pursuant to an agreement between Petitioner and Patent Owner, I hereby certify that, on October 15, 2025, I caused a true and correct copy of the foregoing materials:

- Petition for *Inter Partes* Review of U.S. Patent No. 11,841,803 Under 35 U.S.C. § 312 and 37 C.F.R. § 42.104
- Appendix of Claims
- Table of Exhibits for Petition for *Inter Partes* Review of U.S. Patent No. 11,841,803
- Exhibits 1001–1006, 1008, 1010–1011, 1014, 1020–1031, 1033–1038, 1040–1050
- Word Count Certificate of Compliance
- Fee Authorization
- Power of Attorney

to be served on Patent Owner (Davidson Sheehan LLP, 6836 Austin Center Blvd. Suite 320, Austin, TX 78731, United States) via electronic mail to:

- | | |
|--------------------------|-------------------------|
| • ARizk@mintz.com | • MTRenaud@mintz.com |
| • AMetell@mintz.com | • SFDavenport@mintz.com |
| • ALoBue@mintz.com | • SMOBrien@mintz.com |
| • CXu@mintz.com | • WAMeunier@mintz.com |
| • CHerndon@mintz.com | • YLiu@mintz.com |
| • HMEdge@mintz.com | • Onesta-itc@mintz.com |
| • MAKarambelas@mintz.com | |

Respectfully Submitted,

/Sarah J. Murphy/
Sarah J. Murphy
Registration No. 80,104

WORD COUNT CERTIFICATE OF COMPLIANCE

I hereby certify that the foregoing Petition for *Inter Partes* Review of U.S. Patent No. 11,841,803 contains 13,917 words as measured by the word processing software used to prepare the document, in compliance with 37 C.F.R. § 42.24(d).

Respectfully submitted,

Dated: October 15, 2025

/Sarah J. Murphy/
Sarah J. Murphy
Registration No. 80,104