

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
U.S. Patent No. 11,841,803

**PATENT OWNER'S PRELIMINARY RESPONSE
PURSUANT TO 37 C.F.R. § 42.107(a)**

TABLE OF CONTENTS

	Page
I. INTRODUCTION	1
II. TECHNOLOGY BACKGROUND AND THE CHALLENGED '803 PATENT	4
III. CLAIM CONSTRUCTION	11
IV. THE KELLEHER AND COLLINS REFERENCES.....	11
A. Kelleher (Ex. 1004).....	11
B. Collins (Ex. 1005)	12
V. THE PETITION FAILS TO ESTABLISH THE REQUIRED LIKELIHOOD OF SUCCESS	12
A. All Grounds Fails To Disclose Or Render Obvious “A Passive Crosslink, Wherein The Passive Crosslink Is Dedicated For Inter-Chiplet Communication”.....	13
1. Ground I Fails To Disclose Or Render Obvious “Dedicated For Inter-Chiplet Communication”.....	14
2. Ground I Fails To Disclose Or Render Obvious A “Passive” Crosslink.....	31
VI. CONCLUSION.....	34

TABLE OF AUTHORITIES

Page(s)

Cases

<i>ActiveVideo Networks, Inc. v. Verizon Comm’ns, Inc.</i> , 694 F.3d 1312 (Fed. Cir. 2012)	31
--	----

UPDATED TABLE OF EXHIBITS

Exhibit	Description
2001	Declaration Of Dinesh Manocha, Ph.D.
2002	<i>Curriculum Vitae</i> of Dinesh Manocha, Ph.D.
2003-2100	<i>Reserved</i>
2101	Amended Procedural Schedule (Order No. 15) (December 1, 2025) from 337-TA-1450
2102	Exhibit 803-A-01 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2103	Exhibit 803-A-02 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2104	Exhibit 803-A-03 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2105	Exhibit 803-A-04 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2106	Exhibit 803-A-05 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2107	Exhibit 803-A-06 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2108	Exhibit 803-A-07 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2109	Exhibit 803-A-11 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2110	Exhibit 803-A-12 to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2111	Exhibit 803-B to NVIDIA's November 20, 2025, Supplemental Invalidity Contentions from 337-TA-1450
2112	Onesta's September 29, 2025, Preliminary Rebuttal To NVIDIA's Invalidity Contentions from 337-TA-1450
2113	Defendants' Unopposed Motion To Stay Pursuant To 28 U.S.C. § 1659(a) (Dkt 20) (June 4, 2025) from Case No. 1:25-cv-586-ADA (WDTX)
2114	Text Order Granting Motion to Stay Case (Dkt 21) (July 2, 2025) from Case No. 1:25-cv-586-ADA (WDTX)
2115	ITC Motion to Stay Pending IPR Stats

2116	Guidance on USPTO’s Recission of ‘Interim Procedure for Discretionary Denials in AIA Post-Grant Proceedings with Parallel District Court Litigation’” (the “Guidance Memo”)
2117	Section 337 Statistics: Average Length of Investigations (April 11, 2025)
2118	Notice of Institution of Investigation (May 20, 2025) from 337-TA-1450
2119	Joint Comment On The Public Interest Of The United States Patent And Trademark Office And The United States Department Of Justice (November 25, 2025)

I. INTRODUCTION

The Petition should be denied because each of the two asserted grounds fails to disclose multiple limitations of the challenged claims of U.S. Patent Number 11,841,803 (“the ’803 patent”).

Claim 1, the only challenged independent claim, recites “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.” The only challenged dependent claims (2-4, 7-10) all depend from claim 1. Thus, all challenged claims require this limitation.

The Petition relies exclusively on the combination of Kelleher and Collins for the claimed “passive crosslink . . . dedicated for inter-chiplet communication,” Pet., 36-53, but neither Kelleher nor Collins discloses or renders obvious such a crosslink, nor is there any motivation to combine the references, as they are fundamentally incompatible and from technical different fields.

First, neither Kelleher nor Collins discloses or renders obvious a crosslink “dedicated for inter-chiplet communication.” The Petition admits that Kelleher does not disclose whether its “interlink 250”—which the Petition alleges is a crosslink—is dedicated for inter-chiplet communications. Pet., 2, 17. The Petition relies on Collins to fill that gap, but Collins also does not disclose a crosslink that is “used for transmission of only chiplet-to-chiplet signals,” as Petitioner’s construction of the

“dedicated for inter-chiplet communication” requires. Collins repeatedly explains that the purpose of its interconnect structure and EMIBs is to couple a chiplet and **a memory device**. *E.g.*, Ex. 1005, 1:7-9, 2:3-7, 3:49-51, 3:65-4:23, 5:18-29, 6:2-22, 6:63-7:10. Indeed, Collins’s entire focus is on coupling a chiplet and memory device, rather than providing chiplet-to-chiplet communication. *Id.*, 1:7-9 (“[T]he present disclosure relates to semiconductor packages with chiplet and memory device coupling.”).

Second, neither Kelleher nor Collins discloses or renders obvious a “passive crosslink.” The Petition admits that Kelleher does not disclose that its “interlink 250”—which the Petition alleges is a crosslink—is “passive.” Pet., 2, 17. Instead, the Petition relies solely on Collins for this limitation. Pet., 39. But Collins also fails to disclose a “passive” crosslink; rather, Collins only states that its interconnect structure and/or the EMIBs contained in that interconnect structure—both of which the Petition ambiguously alleges is the claimed crosslink—may **include some** passive elements. But just because these components **include** passive elements does not mean that these structures themselves **are** passive, because neither Collins nor the Petition’s allegations foreclose that the interconnect structure or EMIBs contain additional, active elements as well.

Third, a POSITA would not have been motivated to combine Kelleher and Collins in the manner proposed by the Petition. The Petition’s motivation-to-

combine arguments fail for numerous reasons. Kelleher and Collins are not analogous art to each other or the challenged claims. Kelleher describes a multi-chip module comprising a “master GPU chip” and a “slave GPU chip” (in the primary embodiment, the latter being a “floorswept” chip, meaning one that has defects but is usable enough to serve some purpose, Ex. 1004, 7:43-47), and Collins relates to coupling a **CPU** chiplet, not a **GPU** chiplet (as explained below) to a memory device, not another chiplet. A POSITA looking to implement a GPU chiplet array would not have looked to Collins, because Collins does not concern **GPU** chiplets, but rather only ordinary **CPU** chiplets. A POSITA would not have looked to a CPU reference to improve a GPU reference and further would have understood that CPU technologies are not applicable to GPU technologies, and certainly cannot simply be mixed and matched without any modification whatsoever, as the Petition argues. Collins’s interconnect structure and EMIBs simply would not be suitable for performing GPU-specific functions as in Kelleher.

Further, the Petition ignores that Kelleher uses pre-packaged chips and uses external wire connections to further mount the chips on an additional package. Collins, by contrast, uses unpackaged dies and connects them through a silicon “interposer,” a much more integrated manufacturing process. This would be a massive manufacturing and design change, and for no benefit. In fact, the Petition ignores the numerous drawbacks and incompatibilities that the combination would

introduce, including higher latency, a more complex and expensive manufacturing process, and thermal mechanical problems, without providing any explanation for how those could be overcome, or why a POSITA would make such a massive change in light of them.

Thus, all grounds fail to disclose or render obvious this limitation, and institution should be denied.

Accordingly, the Board should deny institution of the Petition.

II. TECHNOLOGY BACKGROUND AND THE CHALLENGED '803 PATENT

The challenged '803 patent claims generally relate to an improved computing system comprising GPU chiplets and a particular type of crosslink between them. The '803 patent explains that “[c]onventional monolithic die designs are becoming increasingly expensive to manufacture.” Ex. 1001, 1:45-48. Prior to the '803 patent, “[s]ome conventional multi-chip modules include[d] two or more semiconductor chips mounted side-by-side on a carrier substrate or in some cases on an interposer (so-called ‘2.5D’) that is, in-turn, mounted on a carrier substrate.” Ex. 1001, 1:12-16. Additionally, chiplets—small, modular integrated circuits that can be combined with other chiplets to create more complex systems—had been “used successfully in CPU architecture to reduce cost of manufacture and improve yields.” *Id.* However, several barriers exist when applying a chiplet architecture to GPUs. As

compared with CPUs, GPUs are designed for high throughput, processing huge volumes of data in parallel to complete large tasks quickly while also maintaining synchronous ordering across different sections. Ex. 2001, ¶15. The '803 patent acknowledges these challenges in its Background section:

GPU work by its nature includes parallel work. However, the geometry that a GPU processes includes not only sections of fully parallel work but also work that requires synchronous ordering between different sections. Accordingly, a GPU programming model that spreads sections of work on different threads is often inefficient because the parallelism is difficult to distribute across multiple different working groups and chiplets as it is difficult and expensive computational to synchronize the memory contents of shared resources throughout the entire system to provide a coherent view of the memory to applications. Additionally, from a logical point of view, applications are written with the view that the system only has a single GPU. That is, even though a conventional GPU includes many GPU cores, applications are programmed as addressing a single device. So, it has been historically challenging to bring chiplet design methodology to GPU architectures.

Id., 1:51-67. To overcome these challenges, the '803 patent provides “systems and methods that utilize high bandwidth passive crosslinks for coupling GPU chiplets” designed to “improve system performance by using GPU chiplets while preserving the current programming model.” *Id.*, 2:1-4. “The GPU chiplet breaks up systems-

on-a-chip (SoC) into smaller functional groupings referred to as ‘chiplets’ that perform functions of the various cores of the SoC (e.g., a GPU).” *Id.*, 2:14-17.

Another critical aspect of the ’803 patent’s system is its memory architecture. “Currently, various architectures already have at least one level of cache (e.g., L3 or other last level cache (LLC)) that is coherent across the entire conventional GPU die.” *Id.*, 2:18-20. In the ’803 patent’s system, “the 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.” *Id.*, 2:20-24. Accordingly, “[d]uring operations, a memory address request from the CPU to the GPU is transmitted only to a single GPU chiplet, which then communicates with a high bandwidth passive crosslink to locate the requested data. From the CPU’s point of view, it appears to be addressing a single die, monolithic GPU. This allows for use of a large-capacity, multi-chiplet GPU that appears as a single device to an application.” *Id.*, 2:26-33.

“FIG. 1 is a block diagram illustrating a processing system 100 employing high bandwidth passive crosslinks for coupling GPU chiplets in accordance with some embodiments”:

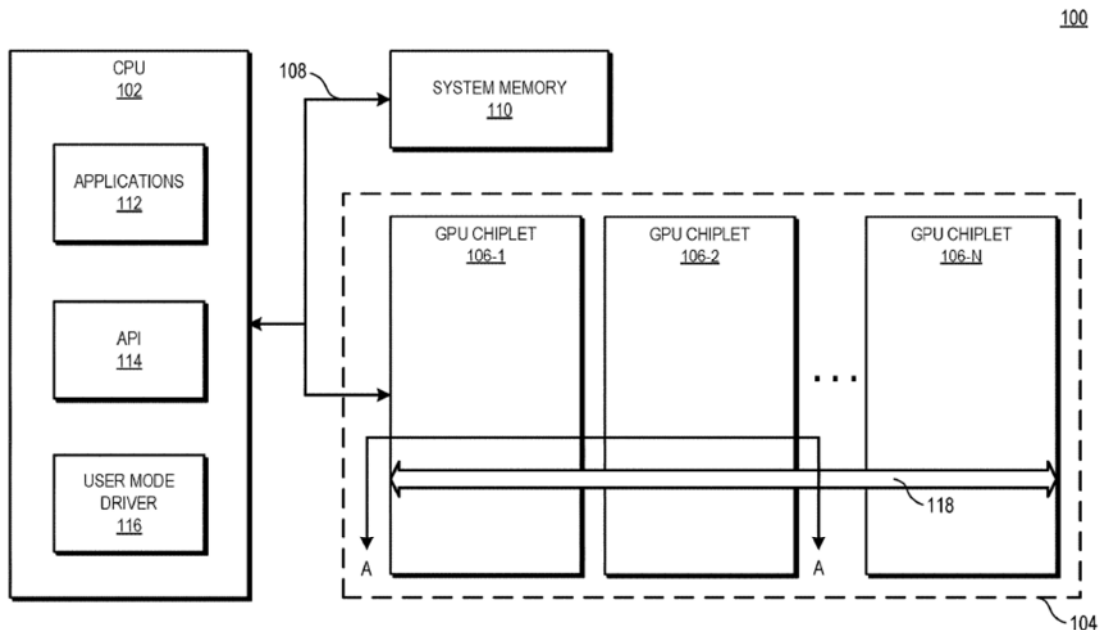


FIG. 1

Ex. 1001, 2:34-36, Fig. 1. “In the depicted embodiment of FIG. 1, a passive crosslink 118 communicably couples the GPU chiplets 106... to each other.” *Id.*, 3:32-34. “In various embodiments, the passive crosslink 118 includes an interconnect chip such as a high-density crosslink (HDCL) die interposer or other similar technology for inter-chiplet communications.” *Id.*, 3:38-41. “Subsequently, any inter-chiplet communications is routed through the passive crosslink 118 as appropriate to access memory channels on other GPU chiplets 106.” *Id.*, 3:46-49. “In this manner, the GPU chiplet-based system 100 includes GPU chiplets 106 that are addressable as a single, monolithic GPU from a software developer’s

perspective... and therefore avoids requiring any chiplet-specific considerations on the part of a programmer or developer.” *Id.*, 3:49-56.

“FIG. 4 is a block diagram illustrating a floor plan view of a GPU chiplet in accordance with some embodiments”:

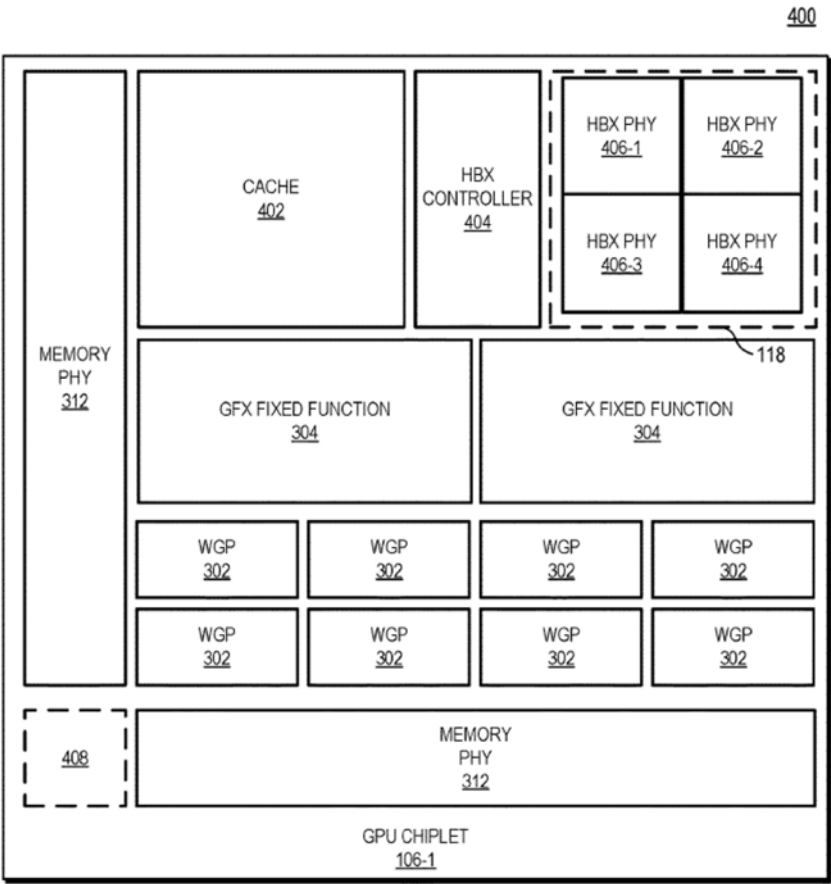


FIG. 4

Ex. 1001, Fig. 4. “The GPU chiplet 106-1 [] includes a hierarchical cache memory 402... and memory PHY 312. The GPU chiplet 106-1 further includes a passive crosslink controller 404 and one or more passive crosslink PHYs 406-1, 406-2, 106-3, and 406-4.” *Id.*, 5:49-58. “The passive crosslink controller 404 connects to the last level cache (LLC) of the GPU chiplet 106-1... and handles routing between the LLC and electrically-active portions of the logic of data fabric crossbars.” *Id.*, 5:59-63.

Figure 5 illustrates how individual GPU chiplets may be combined in a GPU chiplet array in accordance with some embodiments:

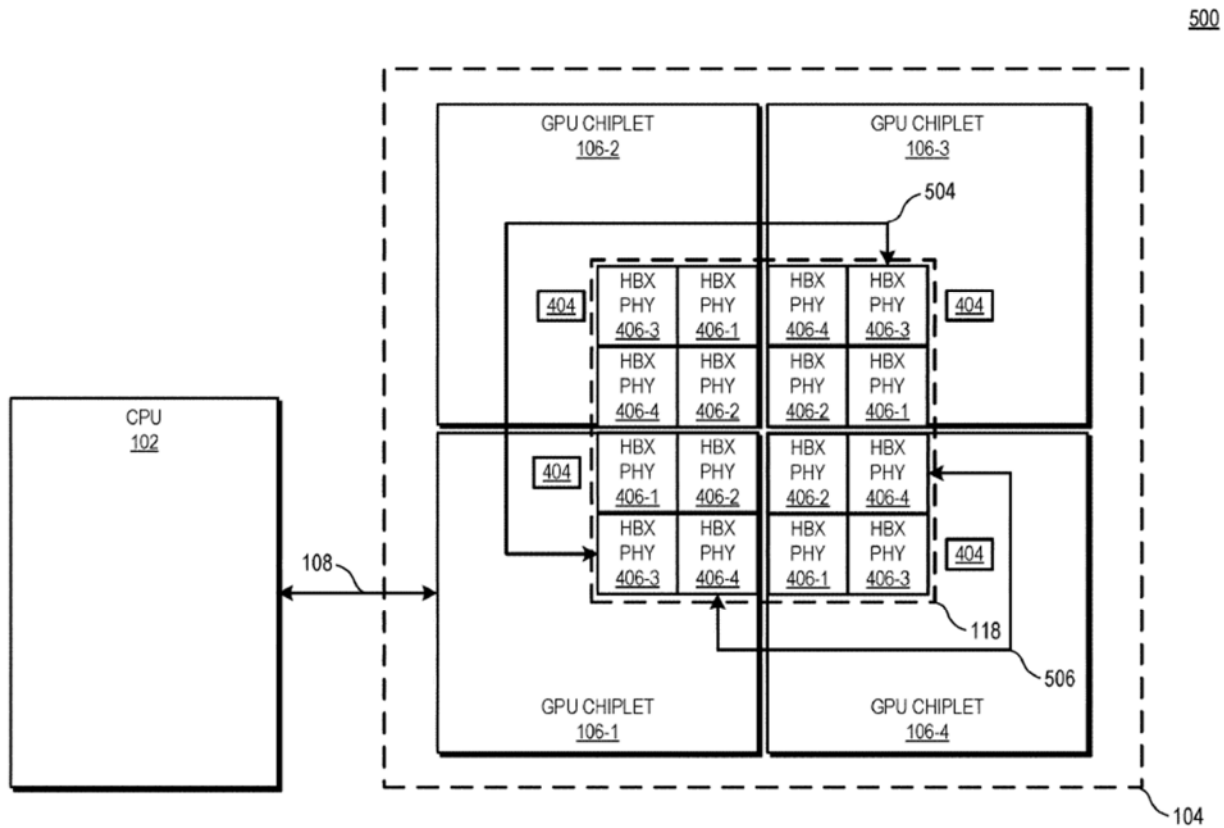


FIG. 5

Ex. 1001, Fig. 5. As shown, the passive crosslink PHYs 406-1, 406-2, 106-3, and 406-4 are of each GPU chiplet are arranged such that passive crosslink 118 communicably couples all GPU chiplets in the array. Each passive crosslink PHY “corresponds to a dedicated communication channel that routes to the L3 cache memory 310 of” a particular GPU chiplet. *Id.*, 6:41-54.

The challenged claims of the ’803 patent capture the above-described features.

Independent 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.

III. CLAIM CONSTRUCTION

The Petition argues that the terms “chiplet” and “dedicated for inter-chiplet communications” require construction. Pet., 25-29. The Petition fails under Petitioner’s constructions, which the Patent Owner applies here without agreeing they are correct.

IV. THE KELLEHER AND COLLINS REFERENCES

A. Kelleher (Ex. 1004)

Kelleher describes a master/slave architecture for combining two GPUs as a multi-chip module, much like the “conventional” prior art described in the ’803 patent as being distinct from its approach. Ex. 1001, 1:12-16; Ex. 1004, 2:7-10. Kelleher’s stated benefit is that the master GPU can offload some of its work to the slave GPU to process in parallel. *Id.*, 2:15-17.

Unlike the ’803 patent, which employs a passive crosslink dedicated for inter-chiplet communication, Kelleher describes only a generic interlink that the Petition

concedes is not described as passive or dedicated to inter-chiplet communication. Pet. at 2. In comparison to Kelleher, the '803 patent's passive crosslink is dedicated for inter-chiplet communication and provides a much higher density and bandwidth than Kelleher's interlink. Ex. 2001, ¶27.

B. Collins (Ex. 1005)

Collins “relates to semiconductor packages with chiplet and memory device packaging.” Ex. 1005, 1:7-9. “The semiconductor package may further include an interconnect structure that couples the first chiplet to a first memory channel of the memory device and the second chiplet to a second memory channel of the memory device.” *Id.*, Abstract. The purpose of the interconnect structure is to couple a chiplet and a memory device, and the reference describes a number of different architectures, all of which only describe chiplet-to-memory connections, not dedicated chiplet connections. *E.g.*, Ex. 1005, 1:7-9, 2:3-7, 3:49-51, 3:65-4:23, 5:18-29, 6:2-22, 6:63-7:10. As such, unlike the '803 patent's passive crosslink, Collins's interconnect structure is not dedicated for inter-chiplet communication, and also is not described as being passive.

V. THE PETITION FAILS TO ESTABLISH THE REQUIRED LIKELIHOOD OF SUCCESS

The Petition asserts only one ground (Ground I) against independent claims, which is the sole independent claim challenged in the Petition.

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

Pet., 6. This preliminary response explains why Ground 1 fails to render obvious claim 1. Ground II merely adds two additional references (Koker and Sistla) to the combination of Ground I to attempt to address dependent claims 8 and 9, which depend from claim 1. Ground II does not change Ground I’s analysis for claim 1 and thus Ground II fails for the same reasons Ground I does. Thus, all grounds fail to establish the required likelihood of success and institution should be denied.

A. All Grounds Fails To Disclose Or Render Obvious “A Passive Crosslink, Wherein The Passive Crosslink Is Dedicated For Inter-Chiplet Communication”

Claim 1, the only challenged independent claim, recites “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.” The only challenged dependent claims (2-4, 7-10) all depend from claim 1. Thus, all challenged claims require this limitation.

The Petition relies exclusively on the combination of Kelleher and Collins for the claimed “passive crosslink.” Pet., 36-53. However, neither Kelleher nor Collins

discloses or renders obvious a “passive crosslink [] dedicated for inter-chiplet communication.” Thus, all grounds fail to disclose or render obvious this limitation, and institution should be denied.

1. Ground I Fails To Disclose Or Render Obvious “Dedicated For Inter-Chiplet Communication”

Ground I fails to disclose or render obvious a crosslink that is “dedicated for inter-chiplet communication.” Neither Kelleher nor Collins discloses or renders obvious this limitation. Further, a POSITA would not have been motivated to combine Kelleher and Collins, especially in the manner proposed by the Petition.

a. Neither Kelleher nor Collins Discloses Or Renders Obvious “Dedicated For Inter-Chiplet Communication”

The Petition admits that Kelleher does not disclose that its “interlink 250”—which the Petition alleges is a crosslink—is dedicated for inter-chiplet communications. Pet., 2 (“Kelleher does not explicitly disclose whether its crosslink is ‘passive’ or ‘dedicated for inter-chiplet communications’”), 17. The Petition relies on Collins to fill that gap, but Collins also does not disclose a crosslink that is “used for transmission of only chiplet-to-chiplet signals,” as Petitioner’s construction of the “dedicated for inter-chiplet communication” requires.

Collins repeatedly explains that the purpose of its interconnect structure and its embedded multi-die interconnect bridges (“EMIBs”) is to couple a chiplet and *a*

memory device. *E.g.*, Ex. 1005, 1:7-9, 2:3-7, 3:49-51, 3:65-4:23, 5:18-29, 6:2-22, 6:63-7:10. Indeed, Collins’s entire focus is on coupling a chiplet and memory device, rather than providing chiplet-to-chiplet communication. *Id.*, 1:7-9 (“[T]he present disclosure relates to semiconductor packages with chiplet and memory device coupling.”); Ex. 2001, ¶32. Indeed, as the Petition acknowledges, during prosecution, the ’803 patent was allowed over the Teig reference, which disclosed an interposer coupling between devices (such as memory) other than chiplets. Pet., 12-13 (citing Ex. 1003, 6, 847). Thus, Collins’s interconnect structures and EMIBs are not “dedicated for inter-chiplet communication,” as required by all challenged claims.

The Petition relies on a misread of Collins’s brief statement that “[t]he chiplets 104 *coupled* with each other via the EMIB s 108 may utilize the EMIB s 108 for communication between the chiplets 104,” and, similarly, “Example 5 may include the semiconductor package of 25 example 1 or any other example herein, wherein the interconnect structure further *couples* a physical layer (PHY) of the first chiplet to a PHY of the second chiplet.” See Pet., 39 (citing Ex. 1005, 4:24-39, 16:24-27) (emphasis added). Looking at the statement in context makes clear that Collins is not describing a dedicated connection, but rather an indirect “coupling” of one chiplet to another ***through memory***, which is not a dedicated chiplet-to-chiplet

connection. The quotation the Petition relies on regards Figure 1, in which chiplets 104 connect to memory device 106 via EMIBs 108.

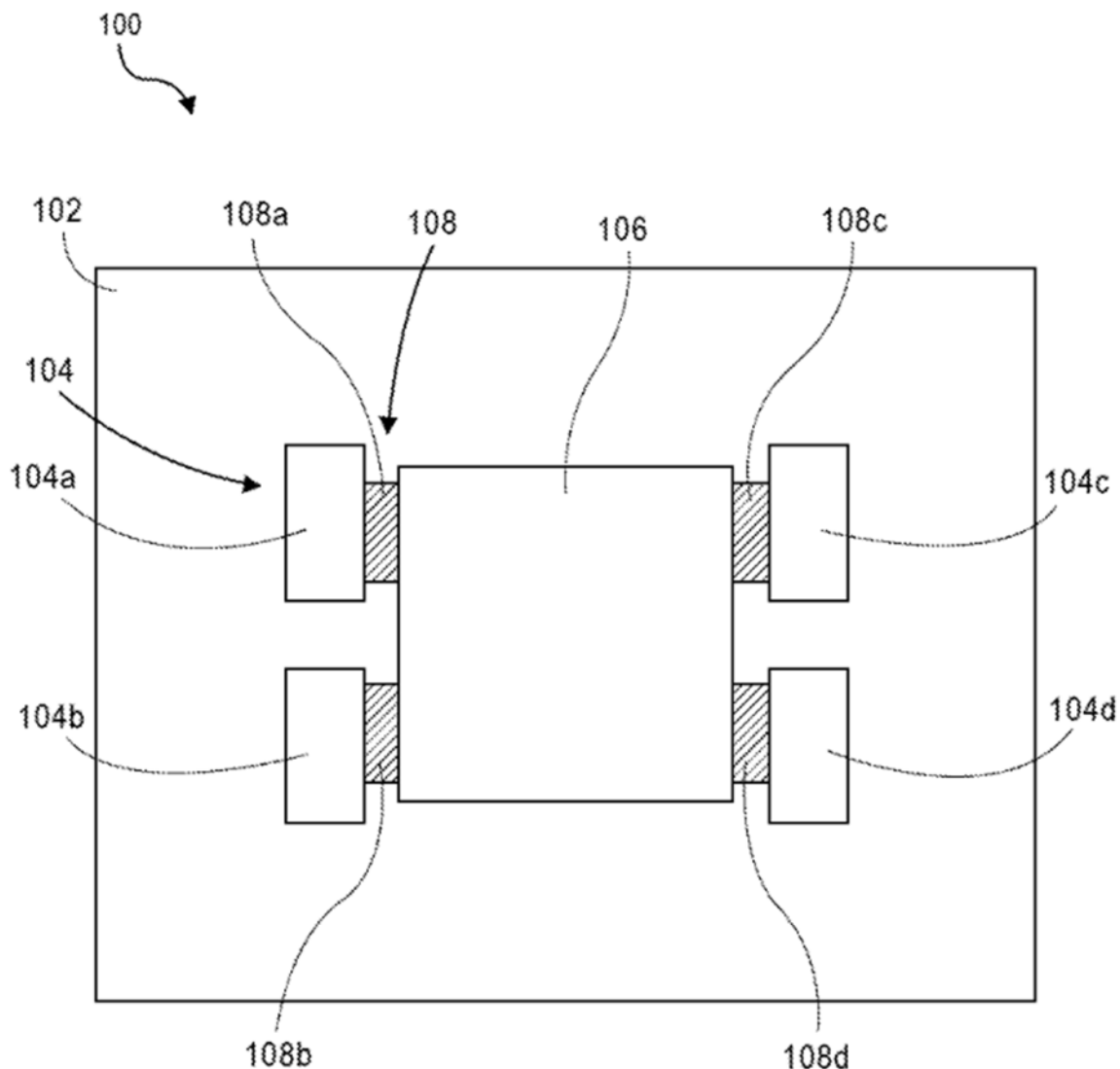


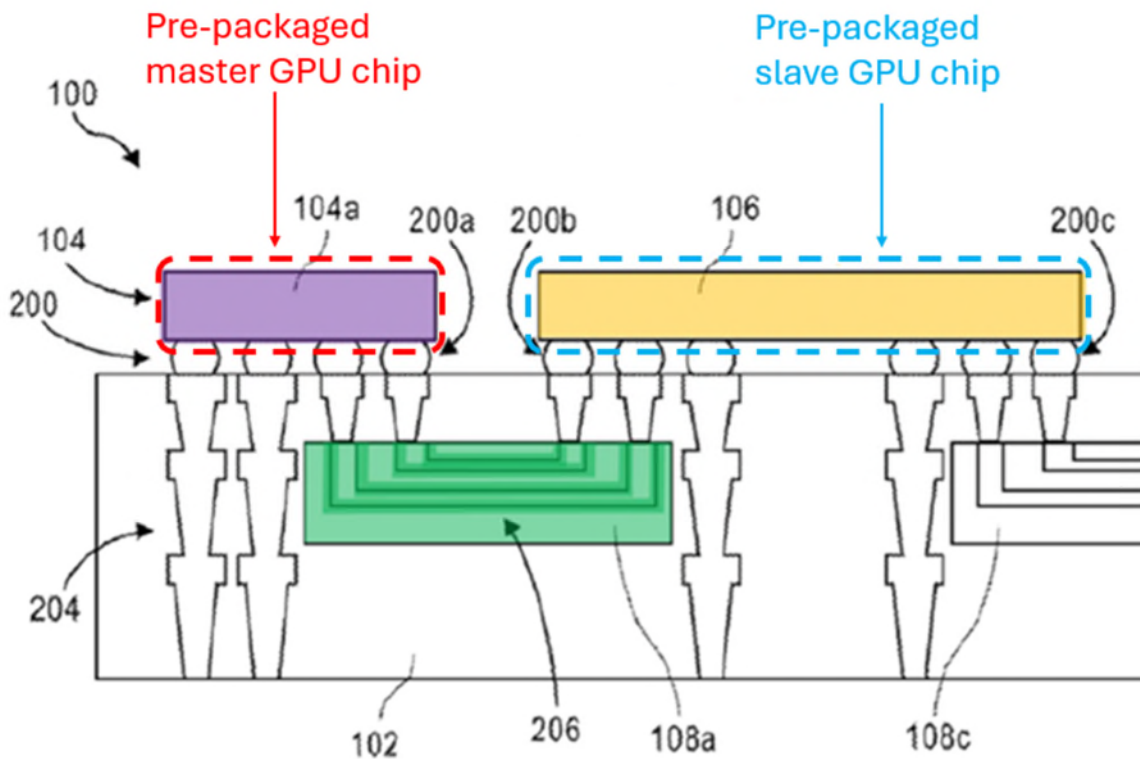
FIG. 1

Ex. 1005, Fig. 1. The specification then describes variations in the “interconnect structure” in the form of changing the numbers of chiplets, EMIBs, and how they connect to memory. Ex. 1005, 3:49-3:39. There is no discussion of removing the

memory or changing this fundamental architecture. Viewed in that context, the statements that “the EMIBs 108 may couple two or more of the chiplets 104 together in some embodiments” and “[t]he chiplets 104 coupled with each other via the EMIBs 108 may utilize the EMIBs 108 for communication between the chiplets 104” simply mean that whereas the architecture above appears to only allow data flow between memory and a chiplet, which is its main purpose, it can also allow one chiplet to communicate to another, but a POSITA would know this would be *indirect* communication, through memory, as no architecture change is described or depicted. That is, it is apparent that even in these embodiments, this chiplet-to-chiplet coupling is ***in addition to*** coupling chiplets to a memory device. Ex. 2001, ¶34. Even in these embodiments, the interconnect structure couples a chiplet to a memory device ***and*** another chiplet, and is thus still not “***dedicated*** for inter-chiplet communication.” Ex. 2001, ¶34. As a POSITA would know, the device of Collins would not function without the chiplet-to-memory connection, as that is its main purpose, and thus, Collins does not disclose any embodiment in which its interconnect structure and/or EMIBs are dedicated for inter-chiplet communication, as required by all challenged claims.

b. The Petition Fails To Set Forth A Motivation To Combine Kelleher And Collins In The Proposed Manner

Despite neither Kelleher nor Collins disclosing a crosslink dedicated for inter-chiplet communication,” the Petition alleges that “[t]he Kelleher-Collins combination teaches the passive crosslink (Collins’s passive interconnect/EMIB) is dedicated for inter-chiplet communications.” Pet., 49. As an initial matter, the Petition does not explain how the Kelleher-Collins combination could teach a crosslink “dedicated for inter-chiplet communications” when neither reference in the combination individually teaches such a crosslink. The Petition’s proposed combination is “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 element).” Pet., 41. The Petition includes a modified version of Collins’s Figure 2 that purports to illustrate the proposed combination whereby Kelleher’s pre-packaged GPU chips would be connected with Collins’s interconnect structure and/or EMIBs:



Pet., 47 (annotated in Petition and further annotated here). Each of Kelleher’s chips “is an integrated circuit that resides on a die, with the die packaged to permit electrical power and signal connections to be made to the integrated circuit.” Ex. 1004 at 3:10-13. The Petition’s combination fails for multiple reasons.

First, the Petition is incorrect that “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.” Pet., 41. Neither Kelleher nor Collins relate to communicably coupling GPU chiplets, and they refer to different technology from one another as well. Kelleher describes a multi-chip module comprising a master GPU chip and a slave GPU chip (a “floorswept” chip, meaning

one that has defects but is usable enough to serve some purpose, Ex. 1004, 7:43-47), and Collins relates to coupling a CPU chiplet—not a GPU chiplet (as explained below)—to a memory device, not another chiplet. Ex. 2001, ¶37. Thus, a POSITA working to implement inter-chiplet communication in a GPU chiplet array would not have looked to either Kelleher or Collins. Moreover, a POSITA looking to modify Kelleher’s GPU chip architecture would not have looked to Collins, because Collins does not concern **GPU** chiplets, but rather only ordinary **CPU** chiplets. A POSITA would have understood that implementing a chiplet architecture for GPUs presents unique challenges as opposed to CPU chiplets. Ex. 2001, ¶38. Indeed, the only mention of “graphics” or “GPU” in Collins is the “graphics processor 1130” in Figure 11. Ex. 1005, 14:52, Fig. 11. Figure 11 and its accompanying text are boilerplate descriptions of a generic computer system and Collins makes little to no connection between the chiplet structure and Figure 11, much less the graphics processor 1130. *Id.*, 14:17-15:62. Nor does Collins imply its chiplets are GPU chiplets. Collins discusses only a generic chiplet and does not go into any detail on the purpose or benefit of utilizing chiplets, other than “[u]tilizing chiplets may provide for more good dies per wafer, greater flexibility to bin based on performance/leakage, and greater flexibility to scale the die complex size (logic elements or cores) for different product applications than larger dies.” Ex. 1005, 3:6-10. These benefits are generic to any application of chiplets, and not specific to

GPU chiplets. Ex. 2001, ¶38. Given this lack of detail and the timeframe of Collins, in addition to the above, a POSITA would have understood that Collins's architecture is designed for CPU chiplets. *Id.*

The architecture of GPUs and their interconnects in terms of parallelism and memory organization is very different from that of a CPU, and they cannot simply be mixed and matched. Ex. 2001, ¶39. While CPUs are designed for general purpose computing, GPUs are designed to provide high throughput for certain types of applications that can exploit the inherent parallelism. *Id.* This is seen in the references themselves, as Kelleher is concerned with parallelism and load balancing from a “master” to a “slave,” Ex. 1004, 4:3-19, whereas that is not a concern or an aspect of the architecture in Collins, in which memory sits in the middle of various CPU chiplets that access it. Ex. 1005, 3:65-4:23, Fig. 1. The techniques for memory hierarchy and cache coherence for GPU chiplets are also very different from CPU chiplets. Ex. 2001, ¶40. For example, because GPUs are designed for graphics rendering operations, they require access to video memory which stores image data such as textures and shaders. *Id.* Moreover, GPUs generally have smaller cache sizes than CPUs because GPUs have fewer pipeline stages; due to the small cache sizes, GPU memory must be organized differently to account for fewer memory blocks fitting in the cache at any given time. *Id.* Given these differences, directly using CPU chiplet methods and architectures would result in poor performance from

GPU chiplets, which would defeat the purpose of using GPUs. Ex. 2001, ¶40. Moreover, the software or programming interface for systems with multiple GPU chiplets is very different from that of a system with CPU chiplets. Ex. 2001, ¶40. Indeed, the '803 patent addresses some of these unique challenges. Ex. 1001, 1:45-67. The Petition is deficient as it does not consider or address any of these incompatibilities or provide any solution for addressing them. Thus, a POSITA looking to modify Kelleher's GPU chip architecture, or to implement a GPU chiplet array, would not have looked to Collins, because its teachings are inapplicable to and incompatible with the GPU context. Ex. 2001, ¶41.

More specifically, Kelleher's interlink transmits GPU data and performs GPU-specific functions that could not be performed by Collins's interconnect structure or EMIB, including routing "[t]exture cache requests and return texture data," and other implementation-specific functions such as "exchang[ing] messages to set up a master-slave relationship. Ex. 1004, 7:33-35, 5:28-29; Ex. 2001, ¶42. In contrast, even setting aside whether Collins's interconnect is dedicated for inter-chiplet communication, which it is not, as explained above, as to the discussion that the Petition argues discloses such a structure, Collins's EMIBs do not transmit any data, not CPU data and certainly not GPU data, but only are used for "clock distribution communications, register control communications, interface testing function communications, *or some combination thereof*"—rather, as discussed

throughout Collins, data is shared through memory reads and writes between chiplet and memory, whereas any communication between chiplets, which is not through a dedicated connection, is for control information only. Ex. 1005, 4:34-39, 2:53-5:29; Ex. 2001, ¶42. Thus, Collins's interconnect structure/EMIBs are not suitable for GPU data traffic, and cannot perform any of the other GPU-specific functions of Kelleher's interlink. Ex. 2001, ¶42. The Petition is deficient because it fails to address these issues and proposes no solution to make Collins's interconnect work in Kelleher. Thus, a POSITA would have understood that Collins's interconnect structure and EMIBs are incompatible with Kelleher's multi-chip GPU architecture. *Id.*

Second, even if Kelleher and Collins were analogous or from the same field, which they are not, the Petition's rationales for why a POSITA would have been motivated to combine Collins and Kelleher make no sense. The Petition alleges that "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." Pet., 43. For one thing, a POSITA would not undertake the massive redesign required to use an intra-substrate link as in Collins in Kelleher. Each of Kelleher's chips "is an integrated circuit that resides on a die, with the die packaged to permit electrical power and signal connections to be made to the integrated circuit," which are then further packaged side by side. Ex. 1004 at

3:10-13, Fig. 3, 5:39-65. That is, Kelleher is concerned with taking pre-packaged, individually functional chips and further connecting them on an additional package, which is why it uses traditional packaging, connecting, and cooling methods. *See* Ex. 1004, 4:63-5:18. Specifically, Kelleher is concerned with “economies of scale associated with the design, manufacture, and marketing of GPU chips.” *Id.*, 1:45-50. “As a result [of economies of scale], the cost of a GPU chip also increases if the production volume decreases below some minimum amount.” *Id.* Moreover, “GPU manufacturers have difficulty simultaneously serving the different market segments.... Therefore, what is desired in an improved apparatus, system, and method for designing GPU chips to service different market segments.” *Id.*, 1:51-2:3 Kelleher’s proposed solution is using “floorswept unit[s] having a defect that renders it unusable as a stand-alone or master GPU” as a slave GPU chip, and pairing it with a non-defective GPU chip as the master GPU chip. *Id.*, 7:36-47 (“[S]lave GPU chip 740 does not require a fully functional command module [], geometry stage [], or raster stage []. As a result, slave GPU chip 740 may be a GPU similar in design to master GPU chip 720 but that is identified as a ‘floorswept’ unit having a defect that renders it unusable as a stand-alone or master GPU.”).

In contrast, Collins’s interconnect is substantially different; the chiplets are not individually packaged, with existing external connections, and then re-packaged onto an additional substrate. They are simply on one substrate from the outset. Ex.

1005, 2:53-4:40. A POSITA would not look to a reference like Collins to modify Kelleher, and any modification requires far too many redesigns to be considered a simple substitution of a connection as the Petition argues. Ex. 2001, ¶44

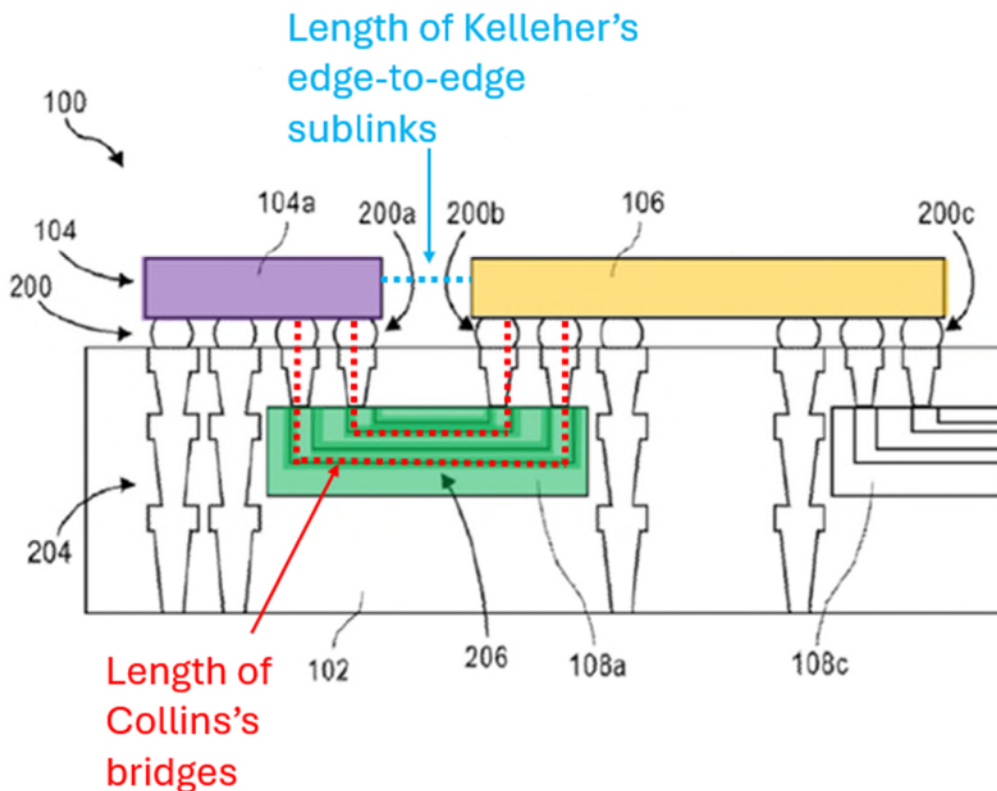
For example, a POSITA would have understood that Collins's EMIBs could not be physically connected to Kelleher's pre-packaged chips, because the pitch and diameter of the solder bumps in EMIBs are much smaller and would not function within Kelleher. Ex. 2001, ¶45. Typically, connections to standalone GPU chips employ C4 bump solder connections. Ex. 2001, ¶45; Ex. 1010 at, *e.g.*, 15, 16 (showing numerous chips employing C4 bumps on a package substrate, including AMD and NVIDIA GPUs). In contrast, EMIBs employ microbumps, which have a much smaller pitch and diameter as compared to conventional C4 bumps, as stated in Liu (Petitioner's Exhibit 1037, ¶[0039]), which is cited by the Petition in its motivation-to-combine analysis. Pet., 48. Liu confirms that microbumps of an EMIB are "generally smaller than the diameter of C4 bumps." Ex. 1037, ¶[0039]; Pet., 48. And regardless whether C4 specifically would be expected in Kelleher, which it would, the Petition concedes that the size of the connections on Kelleher's packaging would not be compatible with Collins's interconnect, stating "[b]ecause EMIBs are close to the chips, the conductors used are significantly smaller than those required to traverse a full interposer," as purportedly is the case in Kelleher. Pet., 48. But this means that, among other reasons, one cannot simply swap Collins's

connection into Kelleher. As such, absent a significant post-manufacturing reconstruction of the chip packaging, Kelleher's pre-packaged GPU chips would have been incompatible with Collins's EMIBs. Ex. 2001, ¶45. The Petition ignores this concern and provides no analysis or explanation of how it could work and is therefore deficient.

Further, according to the Petition, Collins's interconnect structure would address Kelleher's stated drawback that "[a]n edge-to-edge configuration permits only a comparatively small number of low latency sub-links to be formed between edges 315 and 320" and thus "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." Pet., 44 (citing Ex. 1005, 5:39-57). But the Petition does not explain how Collins's interconnect would meet Kelleher's latency and bandwidth requirements.

Nothing in Collins suggests that its interconnects offer enough bandwidth to perform intensive graphics data or GPU data processing, as Collins does not disclose that its interconnects can couple **GPU** chiplets. As a POSITA would know, graphics processing is highly demanding and highly resource-intensive and bandwidth-intensive, much more so than the generic CPU processing of Collins. Ex. 2001, ¶47. And instead of providing improved latency, Collins would worsen the latency in

Kelleher, cutting against a motivation to combine, because the path through which data travels is longer in the Collins link than in the Kelleher link:



Pet., 47 (annotated in Petition and further annotated here). Thus, the Petition fails to set forth an adequate justification for why a POSITA would have been motivated to use Collins's interconnect structure/EMIB to couple Kelleher's two GPU chips.

Further, the Petition points to alleged benefits of not requiring a heat sink as in Kelleher's Figure 4 embodiment. Pet., 46. A POSITA would have understood that a heat sink is a commonplace component across electronics and removal of such

a common component does not justify such a massive redesign as that proposed by the Petition, which fundamentally changes the manufacturing of the devices, increasing cost and complexity, which cuts against the goal of Kelleher, which are to be able to use off-the-shelf and/or floorswept, packaged GPUs connected in a simple way, rather than carving and burying an EMIB as in Collins, which adds costs and complexity that far exceed that of an ordinary heat sink. Ex. 2001, ¶50. Collins also never mentions heat issues or that it provides a solution to any heat issues; thus, the Petition's argument that adding Collins would solve or mitigate Kelleher's heat issues is speculative at best. In fact, Collins's design would likely **aggravate** heat issues by creating a Coefficient of Thermal Expansion (CTE) mismatch between the substrate and the silicon in EMIBs. Ex. 2001, ¶51. Moreover, Kelleher's inventors already designed alternatives to Figure 4. Figures 5 and 6 are alternative designs that do not require a separate heat-sink structure, and thus the Kelleher inventors were satisfied with the design as-is, while being fully aware of a potential heat dissipation concern. Collins does not contain any advantage that is not already contemplated by Kelleher. Ex. 2001, ¶52.

Third, even if a POSITA would have been motivated to use Collins's interconnect structure/EMIB in Kelleher—which it would not, as explained above—the combination would not be in the way proposed by the Petition, as it is illogical and contradicts the purposes of both references. The Petition presumes that a

POSITA would have been motivated to: (1) use Collins's interconnect structure to couple Kelleher's two GPU chips, but (2) not use Collins's interconnect structure as a means to connect Kelleher's GPU chips to Kelleher's memory device (only for an alleged chip-to-chip connection), which is needed in Kelleher and Collins and which is the primary purpose of Collins's interconnect structure. Ex. 2001, ¶53. Alternatively, to the extent that the Petition's combination proposes, as illustrated in its "Modified Figure 2" to simply replace the memory device with a GPU chiplet, a POSITA would have understood that this would result in a non-functioning system, as the chiplets require memory to perform processing, and neither Kelleher nor Collins provide for internal memory within the chips.¹ Ex. 2001, ¶53. Either way, a POSITA would not have been motivated to combine Collins and Kelleher in the manner proposed by the Petition. Ex. 2001, ¶53. Any combination of Collins and Kelleher that could actually function (which would not be motivated as explained above) would use Collins's interconnect structure similar to as illustrated in

¹ Kelleher discloses cache memory within its GPU chips (Ex. 1004, 7:24-25), but a POSITA would have understood that a GPU chiplet array would require non-cache memory, because cache memory is too small to hold all of the data necessary for intensive graphics processing. Ex. 2001, ¶53.

Collins's Figure 1, with Kelleher's GPU chips replacing "chiplets 104") and interconnect structure 108 coupling the GPU chips to a memory device:

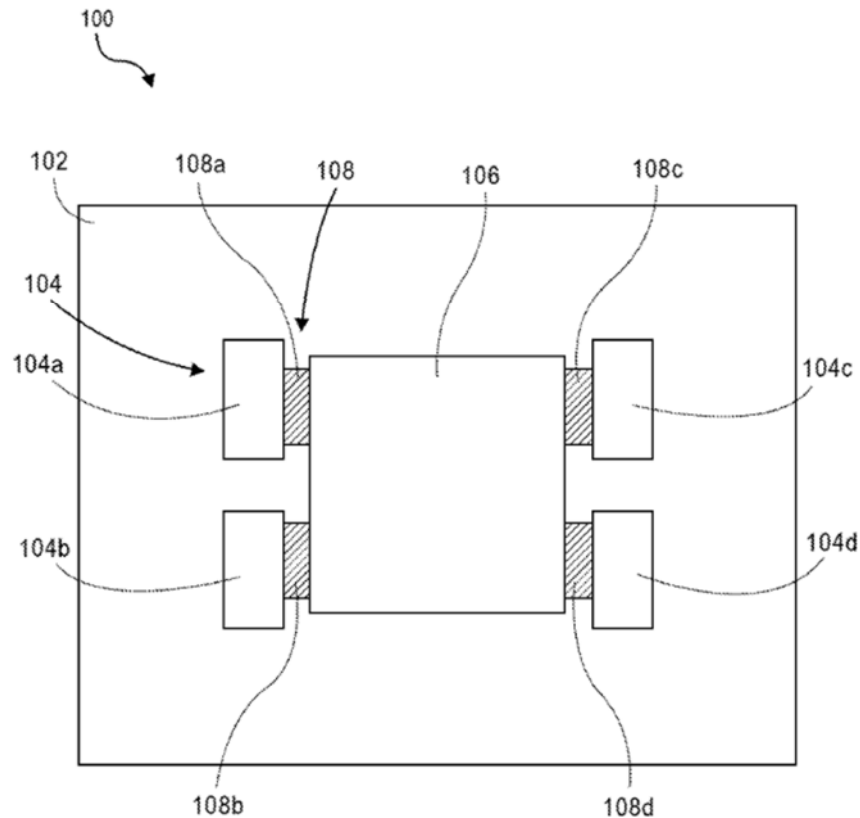


FIG. 1

Ex. 1005, Fig. 1. But of course, this would not meet the challenged claim language because such an interconnect structure would not be “dedicated for inter-chiplet communication,” as required by the challenged claims. Ex. 2001, 54. For this reason, the Petition instead proposes a contrived, hindsight-driven “Kelleher-Collins combination” in an attempt to shoehorn the combination to meet the claim language. Thus, the Petition’s motivation-to-combine arguments “fail[] to explain why a

person of ordinary skill in the art would have combined elements from specific references *in the way the claimed invention does.*” *ActiveVideo Networks, Inc. v. Verizon Comm’ns, Inc.*, 694 F.3d 1312, 1328 (Fed. Cir. 2012) (emphasis in original).

For the above reasons, the Petition fails to provide a motivation to combine Kelleher and Collins, and Ground I fails.

2. Ground I Fails To Disclose Or Render Obvious A “Passive” Crosslink

Ground I also fails to disclose or render obvious a *passive* crosslink, as required by all challenged claims. The Petition admits that Kelleher does not disclose that its “interlink 250”—which the Petition alleges is a crosslink—is “passive.” Pet., 2, 17. Instead, the Petition relies solely on Collins for this limitation. Pet., 39. But Collins also fails to disclose a “passive” crosslink.

a. Collins Fails To Disclose Or Render Obvious A “Passive” Crosslink

The Petition points to Collins’s “interconnect structure” and/or “EMIB” as the passive crosslink. Pet., 39 (“Like the ’803 patent, Figure 2 of Collins show a ‘passive’ crosslink (*interconnect structure/embedded multi-die bridge (EMIB)* with conductive paths)”), citing Ex. 1005, 3:49-4:23, 5:42-47. But Collins does not disclose that either its “interconnect structure” or its “EMIB” is passive—only that they may include passive components, allegedly the “conductive paths,” which are not the alleged passive crosslink. Collins states that the interconnect structure

includes: (1) “one or more embedded multi-die interconnect bridges (EMIBs)”;
and (2) “an interposer,” which “can be passive or active silicon with one or more
conductive paths located within the silicon.” Ex. 1005, 3:51-55, 5:42-47. Regarding
the EMIBs, Collins states that “[e]ach of the EMIBs *include* a piece of passive silicon
with conductive elements located within the passive silicon for conducting signals.”
Id., 3:57-59. Thus, even if Petitioner were correct about all of its allegations about
Collins, Petitioner only alleges that the interconnect structure and EMIBs (either or
both being the alleged claimed crosslink) may ***include*** passive elements. But just
because the interconnect structure and EMIBs ***include*** passive elements does not
mean that they ***are*** passive, because neither Collins nor the Petition’s allegations
foreclose that the interconnect structure or EMIBs contain additional, active
elements as well. Ex. 2001, ¶58. If it did not, the EMIB would be merely a
conductive path 206 itself, but that is not how the reference describes the EMIB or
the interconnect structure, which are said to be distinct from the conductive path 206.

Thus, neither Kelleher nor Collins discloses a “passive” crosslink and Ground
I fails for this additional reason.

**b. A POSITA Would Not Have Been Motivated To
Modify Kelleher To Include a Passive Crosslink**

Even if Collins did include a passive crosslink—which it does not, as
explained above—Ground 1 nevertheless fails to render obvious a passive crosslink

because a POSITA would not have been motivated to modify Kelleher to include a passive crosslink. Ex. 2001, ¶60. The Petition describes using the Collins EMIB as the interlink in Kelleher's system. Pet., 49 ("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."). But Kelleher's interlink includes circuitry to perform various functionalities (described below) which cannot be performed by an EMIB since it does not contain that circuitry and Petition did not allege that it should contain that circuitry:

FIG. 9 illustrates an exemplary physical interface for forming part of the interlink 750 between the master GPU chip and the slave GPU chip. In one embodiment the interlink is formed from many dedicated point-to-point unidirectional signals running at trans GHz frequencies. On output side 905 inputs from a plurality of shader pipeline buses are staged in a staging phase and are then serialized. A sync bit is generated to identify to the other side that a load is present. Additionally, a clock signal (iclk) is transmitted as an interlink clock to permit clock recovery/synchronization. At input side 910 received data is deserialized, staged, and sent to the graphics pipeline. Note that at input side 910 the local interlink clock, sclk, may be generated from iclk but run at a different rate than iclk.

Ex. 1004, 7:48-61; Ex. 2001, ¶60. Further, Kelleher's interlink performs functions that could not be performed by Collins's interconnect or EMIB (assuming that they are passive, as the Petition argues), including: (1) routing "[t]exture cache requests and return texture data." (Ex. 1004, 7:33-35); (2) transmitting "clock signals... to facilitate synchronization of slave GPU chip 240" (*Id.*, 4:5-8); and (3) "exchang[ing] messages to set up a master-slave relationship" (*Id.*, 5:28-29). Ex. 2001, ¶60. The Petition does not explain how a passive interconnect would be able to replace these functions in Kelleher. Any attempt to modify Kelleher by replacing its interlink with Collins's allegedly-passive interconnect structure/EMIBs would require a significant redesign of Kelleher's master-slave topology, a central feature of Kelleher. Ex. 2001, ¶61. A POSITA would not have been motivated to replace Kelleher's interlink with a passive interconnect, as the Petition alleges, given the substantial incompatibility and redesign required. The Petition does not address these issues at all, and does not explain how a combination could function, and is therefore deficient.

VI. CONCLUSION

For the above reasons, the Petition should be denied.

Dated: January 22, 2026

/William A. Meunier/

William A. Meunier (Reg. No. 41,193)

Adam S. Rizk (Reg. No. 66,867)

Michael T. Renaud (Reg. No. 44,299)

MINTZ, LEVIN, COHN, FERRIS,

GLOVSKY AND POPEO, P.C.

One Financial Center

Boston, MA 02111

Telephone: 617-348-1793

Facsimile: 617-542-2241

E-mail: WAMeunier@mintz.com

ARizk@mintz.com

MTRenaud@mintz.com

Onesta-IPR@mintz.com

Reza Dokhanchy (Reg. No. 62,795)

MINTZ, LEVIN, COHN, FERRIS,

GLOVSKY AND POPEO, P.C.

3580 Carmel Mountain Road, Suite 300

San Diego, CA 92130

Telephone: 858-314-1500

RDokhanchy@mintz.com

Counsel to Patent Owner,

Onesta IP, LLC

CERTIFICATE OF WORD COUNT

Pursuant to 37 C.F.R. §42.24(d), Patent Owner hereby certifies, in reliance on the word count of the word-processing system (Microsoft Office Word 2010) used to prepare this preliminary response, that the number of words in this paper is 6,395. This word count excludes the tables of contents, tables of authorities, certificate of word count, certificate of service, and table of exhibits.

Dated: January 22, 2026

/William A. Meunier/

William A. Meunier (Reg. No. 41,193)

CERTIFICATE OF SERVICE

I certify that a copy of Patent Owner's Preliminary Response was served by
electronic mail on the following counsel of record:

Grant K. Rowan
Wilmer Cutler Pickering Hale and Door
LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000
Fax: (202) 663-6363
WH-Nvidia-
IPRService@wilmerhale.com

Michael J. Summersgill
Amanda L. Major
Harry D. Hanson
Sara J. Murphy (Reg. No. 80,104)
Wilmer Cutler Pickering Hale and Door
LLP
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 663-6000
Fax: (202) 663-6363
WH-Nvidia-
IPRService@wilmerhale.com

Dated: January 22, 2026

/William A. Meunier/
William A. Meunier (Reg. No. 41,193)