

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

**DECLARATION OF DINESH MANOCHA, PH.D. REGARDING
PATENT OWNER'S PRELIMINARY RESPONSE
PURSUANT TO 37 C.F.R. § 42.107(a)**

I. Qualifications

1. I am the Paul Chrisman Iribe Chair of Computer Science and Electrical & Computer Engineering and Distinguished University Professor at the University of Maryland at College Park. I hold PH.D. and M.S. degrees in Computer Science from the University of California at Berkeley and a B.E. degree in Computer Science and Engineering from the Indian Institute of Technology. I have been involved in teaching and research in computer engineering for the past 48 years and am a named inventor on 17 U.S. patents, many of which are directly related to GPUs and graphics processing.

2. My skills and experience are in areas of computer graphics, computational geometry, high performance computing, and symbolic computation. Specifically, I have:

- developed many well-known software packages for GPU-based algorithms, collision detection, triangulation, multi-agent simulation, texture compression, solid modeling, etc., which have been downloaded by more than 500,000 users worldwide and license to more than 60 industrial organizations including Intel, Microsoft, Disney, Ford, Kawasaki, Siemens, Phillips Labs, MSC Software, Lockheed Martin, Raytheon, and others;
- published more than 850 papers in the leading conferences and journals in computer graphics, computational geometry, databases, multimedia, high performance computing, symbolic computation, and robotics;
- served as a member of the editorial board or guest editor of eleven leading journals in computer graphics, robotics, geometry processing and scientific computing;

- co-founded three companies, including Impulsonic that developed physics-based audio simulation technologies and was acquired in Valve Inc.;
- won many industry awards, including the Distinguished Alumni Aware from the Indian Institute of Technology and the Distinguished Career in Computer Science Award from the Washington Academy of Sciences. I am a fellow of AAAI, AAAS, ACM, IEEE, and NAI; and
- worked in the broad field of GPU computing since 1999 and published a large number of papers in this area.

3. Because of my background, training, and experience, I am qualified as an expert to opine on the challenged patent. A more detailed account of my work experience and other qualifications is listed in my Curriculum Vitae, which is submitted as Exhibit 2002.

II. MATERIALS CONSIDERED

4. In preparing this declaration, I have reviewed the Petition and the declaration of Petitioner's expert, as well as the materials that they cite, including the asserted prior art.

III. UNDERSTANDING OF THE LAW

5. In preparing and expressing my opinions and considering the subject matter of the challenged patent, I am relying on certain basic legal principles that counsel have explained to me. These principles are discussed below.

A. Claim Construction

6. I understand that the scope of a patent is defined by its claims. I understand that a claim can be:

- a. a method claim, which covers a process/activity, or
- b. an apparatus claim, which covers a physical machine, etc.

7. I understand that the first step in determining the validity of a claim is for the claim to be properly construed. I have been further advised that, in an inter partes review proceeding, the claims of a patent are typically given their ordinary and customary meaning as would be understood by one of ordinary skill in the art in the context of the claims and the patent disclosure (specification).

8. I further understand that an inventor can provide special definitions for specific claim term(s) within the patent. However, any special definition for a claim term must be set forth in the specification with reasonable clarity and precision.

B. Invalidity

9. I have been informed that a claim may be invalid as obvious if the subject matter described by the claim, as a whole, would have been obvious to a POSITA at the time the claimed invention was made. I understand that the standard for obviousness in an inter partes review proceeding is by a preponderance of the evidence.

10. I have also been informed that a determination of obviousness involves an analysis of the scope and content of the prior art, the similarities between the claimed invention and the prior art, and the level of ordinary skill in the art. I have been informed and understand that a prior art reference should be viewed as a whole.

11. I have been informed that, in considering whether an invention for a claimed combination would have been obvious, I may assess whether there are apparent reasons to combine known elements in the prior art in the manner claimed in view of interrelated teachings of multiple prior art references, the effects of demands known to the design community or present in the market place, and/or the background knowledge possessed by a POSITA. I also understand that other principles may be relied on in evaluating whether a claimed invention would have been obvious.

12. I have been informed that, in making a determination as to whether or not the claimed invention would have been obvious to a POSITA, one is to consider certain objective indicators of non-obviousness if they are present, such as: commercial success of product(s) practicing the claimed invention; long-felt but unsolved need; teaching away; unexpected results; copying; and praise by others in the field.

13. I understand that, for such objective evidence to be relevant to the non-obviousness of claim, there must be a causal relationship (called a “nexus”) between

the claim and the evidence. I also understand that this nexus must be based on a novel element of the claim rather than something available in the prior art.

14. I have also been informed and understand that, when considering the obviousness of a patent claim, one should consider whether a reason or motivation existed for combining the elements of the references in the manner claimed, and that the prior art must create a reasonable expectation of success in producing the claimed subject matter.

IV. TECHNOLOGY BACKGROUND AND THE CHALLENGED '803 PATENT

15. 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 two or more semiconductor chips mounted side-byside 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.

For example, as compared with CPUs, GPUs are designed for higher throughput, processing huge volumes of data in parallel to complete large tasks quickly. However, the data a GPU processes includes not only fully parallel sections but also work that requires synchronous ordering across different sections. 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.

16. 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.

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.

18. "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":

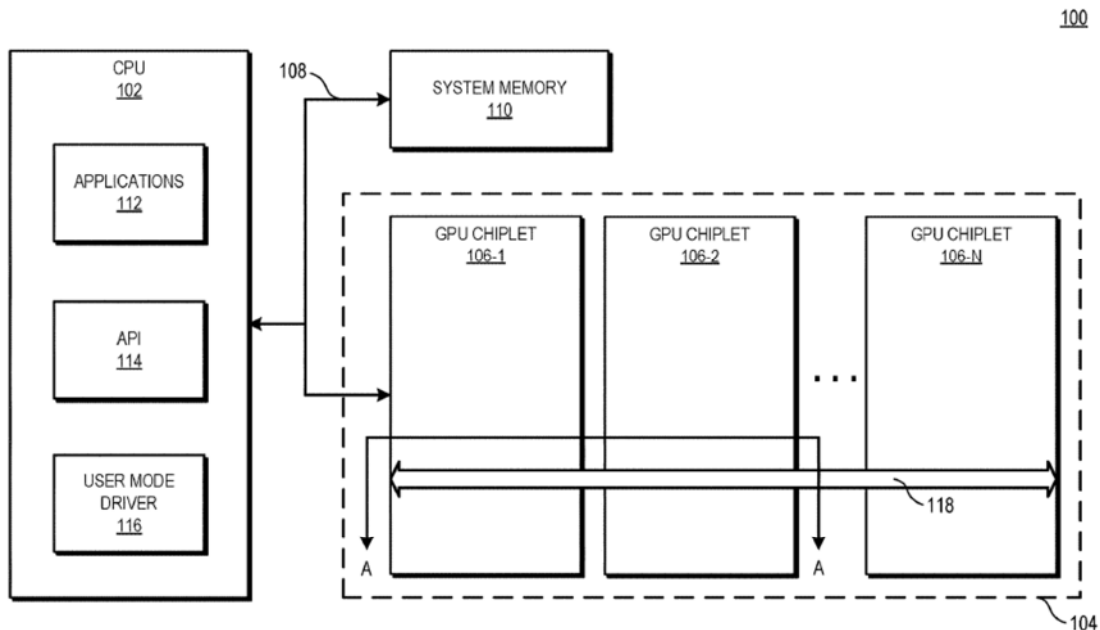


FIG. 1

Ex. 1001, 2:34-36, Fig. 1.

19. “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.

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

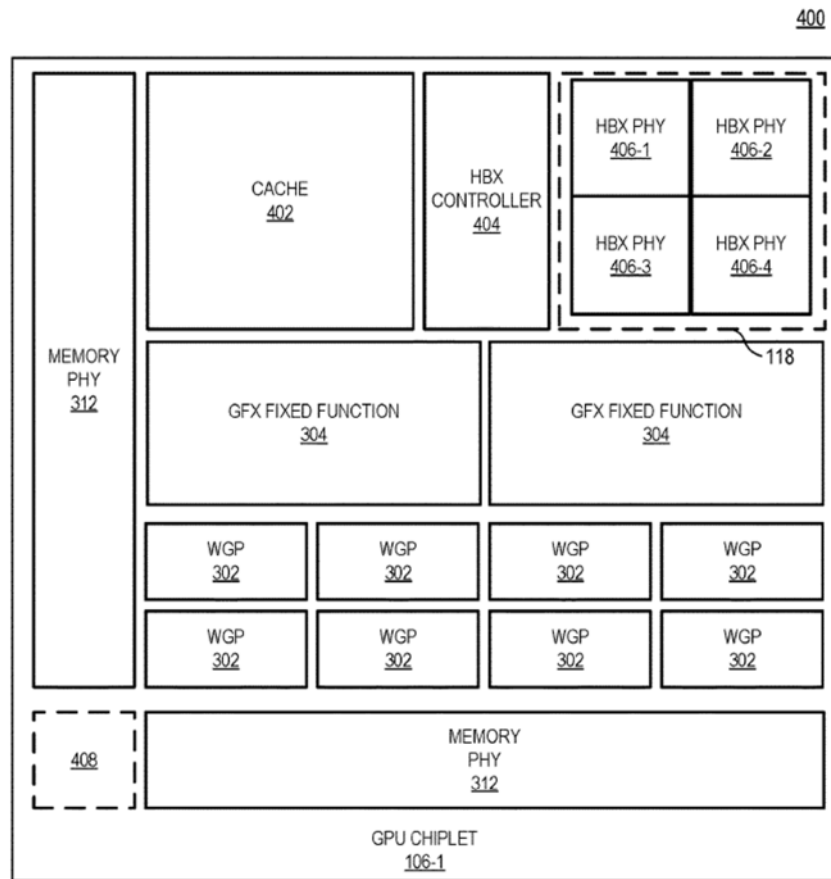


FIG. 4

Ex. 1001, Fig. 4.

21. “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.

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

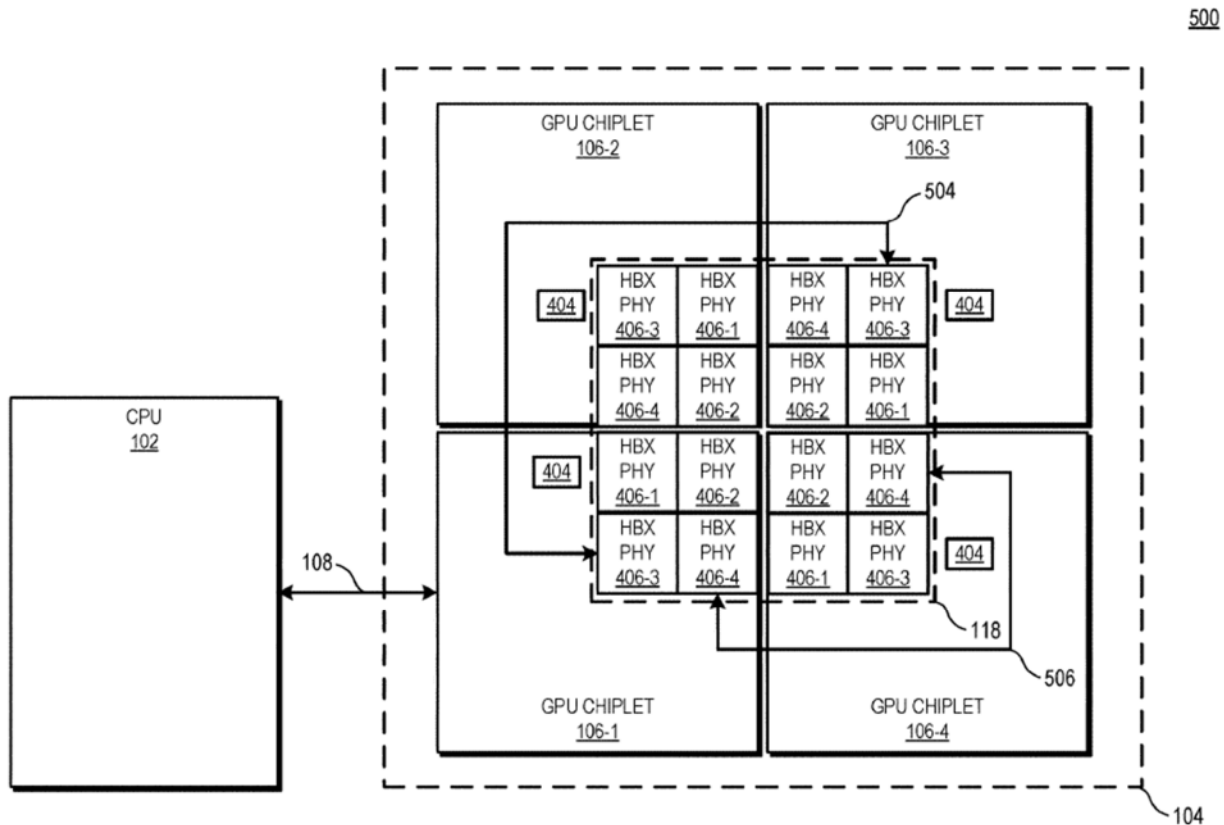


FIG. 5

Ex. 1001, Fig. 5.

23. As shown above in Figure 5, 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.

24. 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.

V. CLAIM CONSTRUCTION

25. The Petition argues that the terms “chiplet” and “dedicated for inter-chiplet communications” require construction. Pet., 25-29. In my opinion, the Petition fails under Petitioner’s constructions, which I have applied here, although I do not necessarily agree that Petitioner’s constructions are correct..

VI. THE KELLEHER AND COLLINS REFERENCES

A. Kelleher (Ex. 1004)

26. Kelleher describes a master/slave architecture for combining two GPUs as a multi-chip module. Ex. 1004, 2:7-10. The chips are coupled using a high-bandwidth interlink. *Id.*, 2:10-13. 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.

27. Unlike the '803 patent, which employs a passive crosslink dedicated for inter-chiplet communication, Kelleher describes only a generic interlink. 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.

B. Collins (Ex. 1005)

28. 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 primary purpose of the interconnect structure 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. As such, unlike the '803 patent's

passive crosslink, Collins's interconnect structure is not dedicated for inter-chiplet communication.

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

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

29. 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.”

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

30. 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”

31. 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.

32. 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.”). 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.”.

33. The Petition relies on a misread of Collins’s brief statements 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 these statements in context makes clear that Collins is not describing a dedicated connection, but rather 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.

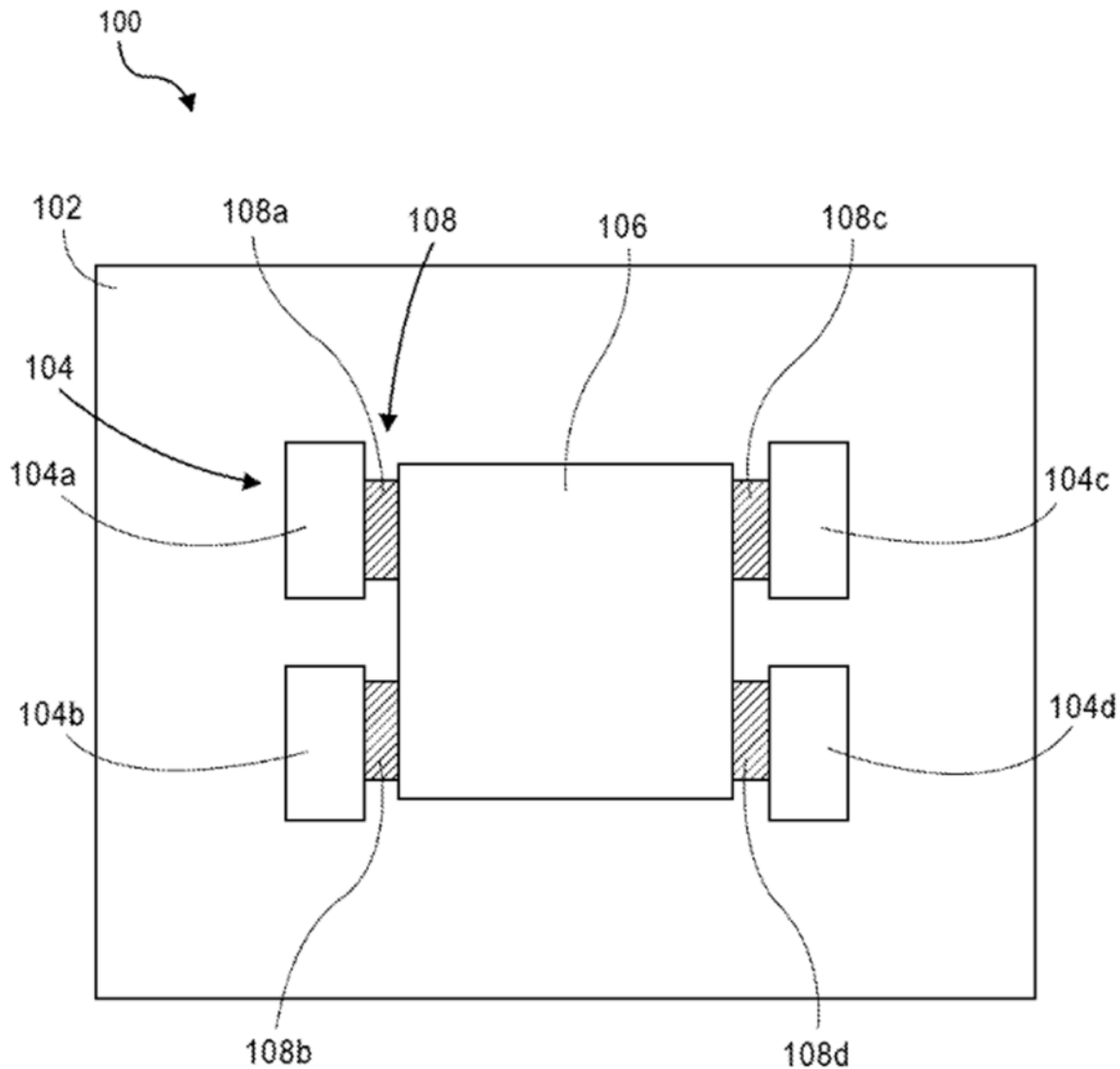


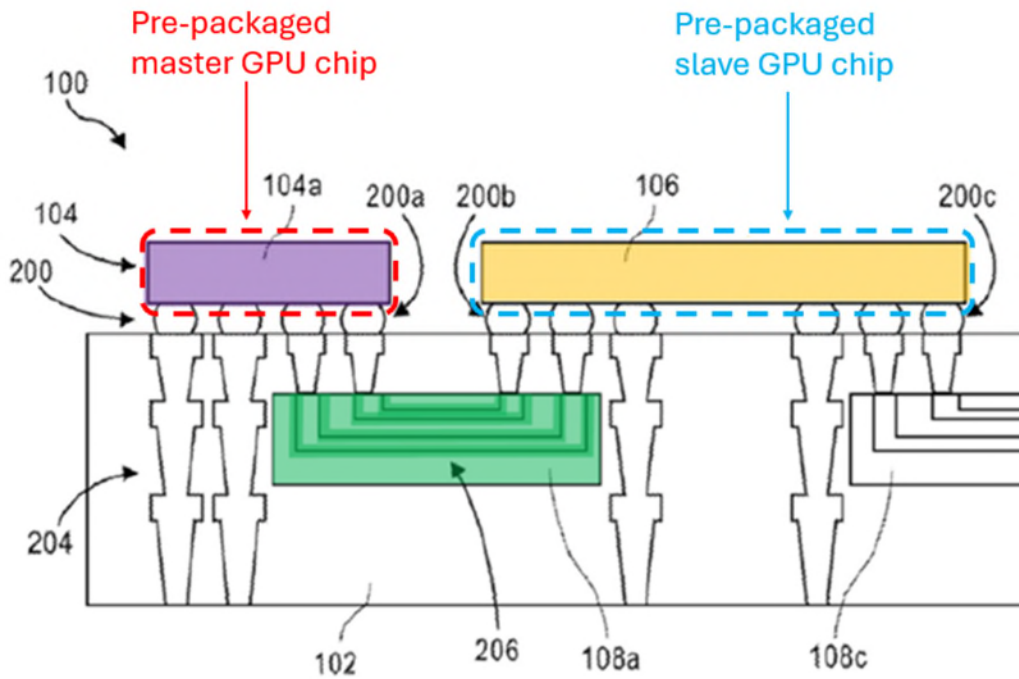
FIG. 1

Ex. 1005, Fig. 1.

34. 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. 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.” 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***

35. 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. 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 dedicated 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 using Collins’s interconnect structure and/or EMIBs:



Pet., 47 (annotated in Petition and further annotated here).

36. 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.

37. ***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. Moreover, Kelleher and Collins relate to different technologies from one another. 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. Thus, a POSITA working to implement inter-chiplet communication in a GPU chiplet array would not have looked to either Kelleher or Collins.

38. Moreover, 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 have understood that implementing a chiplet architecture for GPUs presents unique challenges as opposed to CPU chiplets. Indeed, the only mention of “graphics” or “GPU” in Collins is the “graphics processor 1130” (or “graphics CPU”) 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

chipselets, and not specific to GPU chipsets. 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 chipsets.

39. The architecture of GPUs and their interconnects in terms of parallelism and memory organization is very different from that of a CPU. 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. 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 chipsets that access it. Ex. 1005, 3:65-4:23, Fig. 1.

40. Accordingly, developing a GPU chipset array requires designing new communication architectures for GPU chipsets, as existing communication architectures designed for CPUs would not work. Additionally, the techniques for memory hierarchy and cache coherence for GPU chipsets are very different from CPU chipsets. For example, because GPUs are designed for graphics rendering or GPU computing 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. Moreover, the software or programming interface for systems with multiple GPU chiplets is very different from that of a system with CPU chiplets. Indeed, the '803 patent addresses many of these unique challenges. Ex. 1001, 1:45-67.

41. Thus, the Petition is deficient as it does not consider or address any of these incompatibilities or provide any solution for addressing them. 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.

42. 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 efficiently, 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. In contrast, Collins's EMIBs are used for "clock distribution communications, register control communications, interface testing function communications, *or some combination thereof*," rather than transmitting data, much less GPU data. Ex. 1005, 4:34-39. 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. *Id.*, 2:53-5:29. Thus, Collins's interconnect structure/EMIBs are not suitable for GPU data traffic, and cannot perform any of the GPU-specific functions of Kelleher's interlink efficiently. 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.

43. **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 makes 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 proposed by Collins, in Kelleher's architecture. 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.”).

44. In contrast, Collins’s interconnect technology 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.

45. 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 is much smaller and would not function within Kelleher. Typically, connections to standalone GPU chips employ C4 bump solder connections. 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. The Petition states, "[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. Liu confirms that microbumps of an EMIB are "generally smaller than the diameter of C4 bumps." Ex. 1037, ¶[0039]. This means that, among other reasons, one cannot simply swap Collins's connection into Kelleher. Absent a significant post-manufacturing reconstruction of the chip packaging, Kelleher's pre-packaged GPU chips would

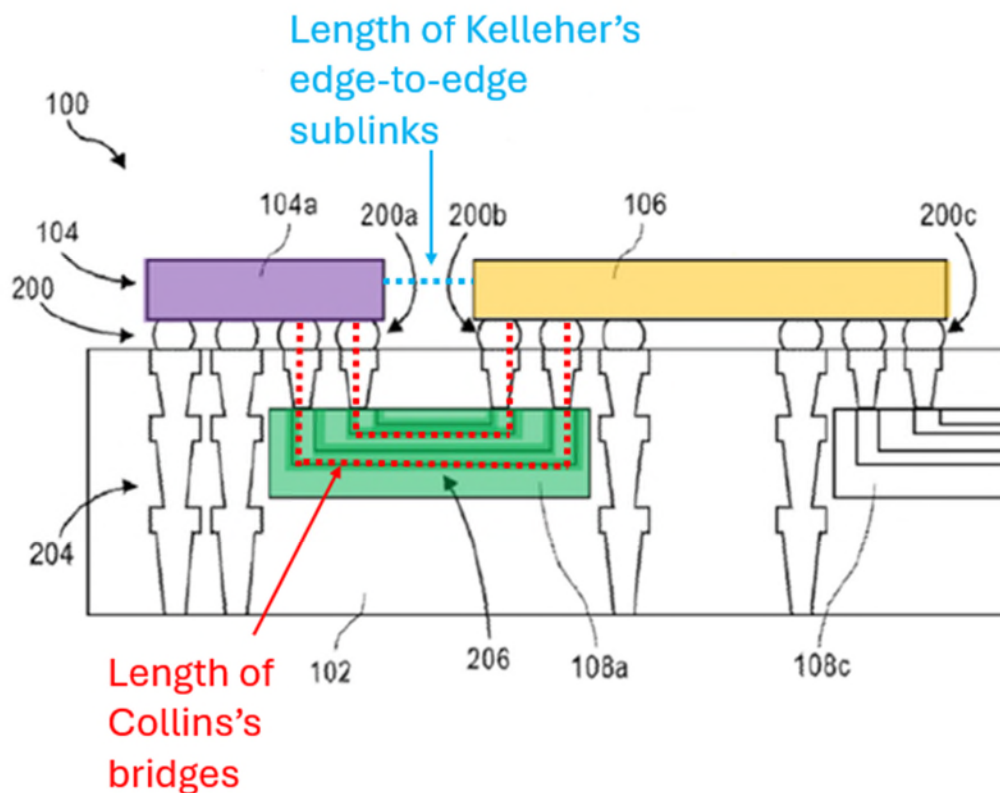
have been incompatible with Collins's EMIBs. The Petition ignores this concern and provides no analysis or explanation of how it could work.

46. 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.

47. 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 have known, graphics processing is highly demanding and highly resource-intensive and bandwidth-intensive, much more so than the generic CPU processing of Collins. GPU architectures are rather different from CPU architectures. In particular, GPUs use a very large number (e.g., thousands) of simpler cores for parallel tasks (to obtain high throughput), and utilize data parallelism by processing massive datasets simultaneously. They have been used for computer graphics and machine learning applications. While CPUs are designed for complex logic and

diverse instructions, GPUs perform simple, repetitive operations across huge data volumes, leading to massive speedups. Therefore, POSITA would not consider using the Collins's interconnect structure into Kelleher's system.

48. Moreover, Collins's interconnect structures would worsen the latency in Kelleher because the path data travels is longer in the Collins link than in Kelleher's edge-to-edge interlink:



Pet., 47 (annotated in Petition and further annotated here).

49. 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.

50. 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 in semiconductor designs and thus that the removal of a heat sink does not justify such a massive redesign. This is especially so here because implementing Collins's interconnection structure/EMIBs into Kelleher would require significant changes to the manufacturing process of the device. For example, forming EMIBs are embedded in the substrate, which requires carving a cavity into the substrate and burying the bridge in the cavity (**Embedded** Multi-die Interconnect Bridge). This process of carving and burying is chemically and mechanically difficult, because if the EMIB is not perfectly flush (coplanar) with the substrate surface, the connection to the chip will not be made. This complexity would result in decreased yield and increased manufacturing costs, as compared to Kelleher's simpler planar solution. Thus, using EMIBs in Kelleher would contradict Kelleher's central purpose of maximizing yield and decreasing manufacturing costs by using defective "floorswept" GPU chips as slave GPU chips to maximize production volume. Ex. 1004, 1:45-60 (discussing "economies of scale associated with the design, manufacture, and marketing of GPU

chips”); 7:43-47 (“[S]lave 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.”).

51. Collins also never mentions any 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. CTE is a material property that quantifies how much a material expands or contracts with temperature changes. Collins states that its interconnect structure/EMIBs are made of silicon. Ex. 1005, 3:57-59. As Kelleher acknowledges, “[h]igh performance GPU chips dissipate a substantial amount of heat.” Ex. 1004, 6:53-59. If a silicon bridge is a different material than the substrate, as the GPU heats up, the substrate (a flexible structure) will expand more than the silicon bridge (which is rigid), which results in a CTE mismatch, which is a critical reliability concern that can lead to interface failure or cracks.

52. 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 provide any advantage that is not already contemplated by Kelleher.

53. **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. The Petition's combination is not practical because it 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 to couple Kelleher's GPU chips to its memory device (only for an alleged chip-to-chip connection), which is need in Kelleher and Collins and which is the primary purpose of Collins's interconnect structure. 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 an non-functioning system, as the chiplets require memory to perform processing, and neither Kelleher nor Collins provide for internal memory within the chips.¹

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.

Either way, a POSITA would not have been motivated to combine Collins and Kelleher in the manner proposed in the Petition. 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 Collins's Figure 1, with Kelleher's GPU chips replacing "chipllets 104" and interconnect structure 108 coupling the GPU chips to a memory device:

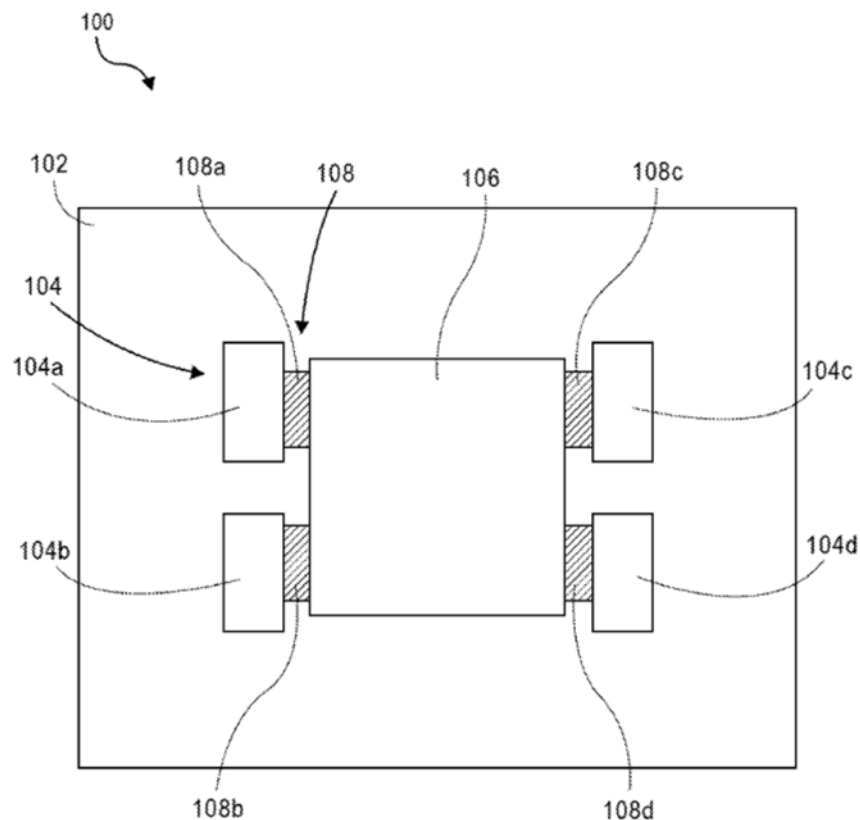


FIG. 1

Ex. 1005, Fig. 1.

54. 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. Thus, the Petition’s motivation-to-combine arguments fail to explain why a POSITA would have combined elements from Kelleher and Collins in the way the claimed invention does.

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

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

56. 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.

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

57. But Collins also fails to disclose 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 they may have passive components, allegedly the “conductive paths.” Regarding the interconnect structure, Collins states that the interconnect structure **includes**: (1) “one or more embedded multi-die interconnect bridges (EMIBs)”;²⁰⁶ 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.

58. 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 these components **include** passive elements does not mean that these components themselves **are** passive, because neither Collins nor the Petition’s allegations foreclose that the interconnect structure or EMIBs contain additional, active elements as well. 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.

59. Thus, neither Kelleher nor Collins discloses a “passive” crosslink.

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

60. 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. 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:

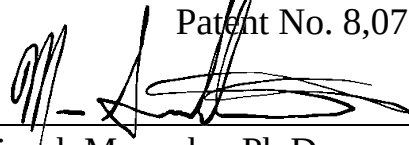
FIG. 9 illustrates an exemplary physical interface for forming part of the interlink 750 between the master GPU chip and the slave GPU chip. In one embodiment the interlink is formed from many dedicated point-to-point unidirectional signals running at trans GHz frequencies. On output side 905 inputs 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 (iclck) 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, sclck, may be generated from iclck but run at a different rate than iclck.

Ex. 1004, 7:48-61.

61. Further, Kelleher's interlink performs functions that could not be performed by Collins's interconnect or EMIB (assuming that they are passive), including: (1) routing "[t]exture cache requests and return texture data." (Ex. 1004, 7:33-35); (2) managing "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). 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. The Petition does not address these issues at all and does not explain how the combination could function.

Case No. IPR2026-00061

Patent No. 8,076,194 B2



Dinesh Manocha, Ph.D.

Dated: January 22, 2026