UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MERCEDES-BENZ USA, LLC,

Petitioner,

v.

DAEDALUS PRIME LLC,

Patent Owner.

Case No.: IPR2023-01344

U.S. Patent No. 9,575,895

PATENT OWNER'S SUR-REPLY

TABLE OF CONTENTS

I. INTRODUCTION 1

II. CLAIM CONSTRUCTION 2

III. THE CHALLENGED CLAIMS ARE NOT UNPATENTABLE IN VIEW OF SINHAROY 4

IV. THE REPLY FAILS TO CONVINCINGLY DEMONSTRATE THAT XU IS PRIOR ART 11

V. THE CHALLENGED CLAIMS ARE NOT UNPATENTABLE IN VIEW OF TANG 14

VI. CONCLUSION 15

TABLE OF AUTHORITIES

CASES

ServiceNow Inc., v. Hewlett-Packard Co.,
IPR2015-00716 (PTAB Aug. 26, 2015) 12

OTHER AUTHORITIES

PTAB Consolidated Trial Practice Guide (Nov. 2019)..... 12

EXHIBIT LIST

EXHIBIT NO.	DESCRIPTION
2001	Transcript of Deposition of Dr. Horst
2002	Declaration of Michael C. Brogioli, Ph.D.
2003	Transcript of Deposition of Dr. Horst re Reply Declaration

I. INTRODUCTION

MB USA's Reply fails to demonstrate why either Sinharoy or Tang teach or suggest a "single caching agent" that includes "a plurality of distributed portions each associated with a corresponding one of [a] plurality of cores," as required by the challenged claims. Accordingly, the Board should reject MB USA's challenges and find in favor of Patent Owner, Daedalus Prime.

As to Sinharoy, MB USA pegs its challenge on an alleged disclosure of a "logical single cache controller," yet MB USA's own expert admits the challenged claims do not recite a "logical single cache controller." Moreover, when questioned during deposition, that expert was forced to concede that the alleged support for this reading of Sinharoy (which relies on a controller knowing certain cache states of data) is, in fact, absent from the reference. Dr. Horst admitted that on a cache miss, the controller would not have any knowledge of the state of the cache line in question, hence, just a Sinharoy states, the remaining cache controllers would be forced to snoop the cache miss to determine whether or not that cache line was present in any of their respective caches. This demonstrates that Dr. Brogioli was correct in his assessment that there is more than one cache controller in Sinharoy—not surprising since Sinharoy itself explicitly says this is the

case—and the challenged claims' requirement of a single cache controller is not met. The fact that the different cache controllers may "work together" as MB USA earnestly alleges is immaterial. What matters is that there are separate controllers.

As to Tang, nothing in the Reply should cause the Board to rethink its original assessment that, "Tang's Figure 10 shows 'all of the ways of this last level cache are shared across all of the cores of the processor,' [and a]s a result of the sharing, [distributed portions of] the 'caching agent' [are] not [each] 'associated with a corresponding one of the plurality of cores.'" Indeed, MB USA's expert, Dr. Horst, confirmed once again that in Tang, the last level cache and all of its associated CPU data ways and controllers communicates with and is shared by each of the cores of the processor. As the challenged claims require something markedly different—specifically, distributed portions of a single caching agent, that are *each associated with a corresponding one of a plurality of cores*—Tang cannot be read as teaching or suggesting the subject matter recited in those claims.

II. CLAIM CONSTRUCTION

MB USA insists that notwithstanding the claims already recite a single caching agent that includes a plurality of distributed portions, the

"single caching agent" should be understood as "one caching agent for the processor, which may be formed from distributed portions." Reply at 4-5. Not only is this redundancy unnecessary, it introduces potential uncertainty where there should be none.

As explained in Patent Owner's Response, the claims of the '895 patent use the term "single" consistent with its "plain and ordinary meaning" to indicate one, and only one, caching agent. POR at 8. Furthermore, the claims already explicitly require this one caching agent to include distributed portions, each associated with a corresponding core. Ex. 1001 at 7:58-61; 8:37-39; 9:6-10. The Board's, and now MB USA's, proposed construction, however, would have the claims understood as reciting one caching agent for a processor, multicore processor, or first multicore processor, as applicable, which may be formed from distributed portions and that includes a plurality of distributed portions each associated with a corresponding core. Thus, it becomes unclear whether it is the processor, multicore processor, or first multicore processor, which may be formed from the distributed portions or whether it is the caching agent which may be so formed. And, if it is the caching agent that *may* be so formed, why then do the claims go on to demand the caching agent *includes* distributed portions? Is there a difference between optionally being formed from distributed portions and including

distributed portions? If not, what purpose is served by introducing the option of including distributed portions ("which *may be* formed from distributed portions") only to remove that option in place of an insistence that the caching agent includes the distributed portions? And, if there is a difference, then certainly it is not one that was addressed in the Petition or the Reply.

The purpose of claim construction is not to introduce such unnecessary confusion into the meaning of the claims. It is the Board's preliminary construction, not Patent Owner's that introduces ambiguity, to use MB USA's phrasing, Reply at 4, and leaves open the questions of just what might be formed of distributed portions and that also might include distributed portions. All of this confusion can be avoided simply by adopting Patent Owner's proposed construction of the "single caching agent" as meaning "one" caching agent, just as the specification of the '895 patent describes. *See, e.g.*, Ex. 1001 at 1:62-64; 2:24-26, 64-66; 3:16-19; 6:30-39; 7:13-17.

III. THE CHALLENGED CLAIMS ARE NOT UNPATENTABLE IN VIEW OF SINHAROY

MB USA's Reply spells out its challenge based on Sinharoy as follows: "'Petitioner does not contend that the L2 caches and L3 cache regions in Sinharoy are structurally one single caching agent. Instead, the

Petition states that 'the caching agent, comprising the centralized cache controllers for the L2 and L3 caches including RC and write machines, is formed as a single caching agent for the processor.'"" Reply at 6. Sinharoy itself dispels this notion. In the cache hierarchy of Sinharoy, each L2 cache, if which there are eight, "is comprised of a single centralized controller, with two address-hashed cache data arrays." Ex. 1003 at 18. And, each L3 cache region, of which there are also eight, is also comprised of "a single centralized controller, with four address hashed eDRAM cache data banks and four address hashed SRAM directory banks." *Id*. at 19. ***Nothing*** in Sinharoy states or suggests that all of these individual single centralized controllers, each a component of their own respective cache, are somehow "formed as a single caching agent for the processor," as MB USA alleges.

First, MB USA admits the individual cache controllers are ***not*** part of the same physical unit. Reply at 6; and see Ex. 1018 at 69:17 – 70:17, 75:6-18 (Dr. Brogioli explaining the separate physical nature of Sinharoy's cache controllers); Ex. 2003 at 15:15-20 (Dr. Horst admitting the separate physical nature of the cache controllers in Sinharoy).

Second, MB USA's mischaracterization of snooping activities described in Sinharoy does not support its position. According to the Reply, Sinharoy only indicates that cache accesses that miss a given L2 and L3

region are *possibly* snooped by other L3 regions. "That is, not all of the other L3 caches need to be snooped." In MB USA's view, this must mean there is a single, logical caching agent. Reply at 7-8. This argument ignores the explicit text of Sinharoy. In fact, Sinharoy states,

> Storage accesses that miss both the L2 cache and the 4-MB local L3 region are broadcast to the coherence fabric and snooped by the memory controller, ***other L2 caches***, possibly other L3 caches, ***and by the seven remote 4-MB L3 regions that comprise the remaining 28 MB of the on-chip L3 cache***.

Ex. 1003 at 19 (emphasis added). In other words, on a cache miss by one core's L2 and L3 caches, ***all*** of the remaining controllers for ***each*** of the other cores' L2 and L3 cache regions snoop the storage access so as to determine whether the requested data is present in their respective caches. Ex. 2002 at ¶¶ 36-38. MB USA is simply wrong when alleging that these cache controllers do not act independently. As Dr. Brogioli explained, "A person of ordinary skill in the art would understand and appreciate that the need to snoop the cache miss by all of the individual cache controllers of the other cores' respective L2 and L3 cache controllers indicates that no one single cache controller is or is acting as a single—that is, a solitary—caching agent for the processor." *Id.* at ¶ 37.

Third, MB USA contends that "a POSITA would have understood that the cache controller for a L3 region has knowledge about whether the lines in its local L3 region are held in any other L2 and L3 caches as well." Reply at 8. But this is irrelevant. Any knowledge that an individual cache controller has about data in its cache comes from associated cache line state information maintained by that controller. Ex. 2003 at 26:18 – 27:22 (Dr. Horst discussing the cache states reported in Table 2 of Sinharoy). And, in the circumstance of a cache miss, the controller associated with the miss has no information about the requested data because that data is not included in its cache. *Id*. at 33:19 – 35:2. This is precisely why a cache miss needs to be snooped by the controllers of the remaining L2 and L3 caches because there is no single controller for the processor maintaining some kind of global cache state information that can be used by all of the individual L2 and L3 regions. Instead, each cache region has its own associated controller, just a Sinharoy states. Ex. 2002 at ¶¶ 36-38; Ex. 1003 at 18, 19. The fact that some data may be shared, that is accessible by two different cache controllers, Ex. 2003 at 28:17-20, does not mean there is just a single cache controller for the processor and the need for cache misses to be snooped by all of the other individual cache controllers demonstrates this point.

Similarly, MB USA's reference to the load and capacity balancing among the L3 cache regions afforded by various heuristics does not make its case for a single cache controller. Sinharoy explains that various cache management policies are employed to route data and instructions to where they are best stored, for example, based on accesses. Ex. 1003 at 17. The fact that this routing is needed and employed again reinforces the idea that the cache controllers are independent. A single cache controller would be able to access any cache data wherever it is located. No routing of data would be needed. Instead, since data routing among caches is employed, this points to the independent nature of the cache controllers to allow access by those controllers.

MB USA goes on to argue that since L3 regions can use space in other L3 areas associated with cores that are disabled, this must mean there is only a single cache controller. Reply at 9-10. But this too is wrong. MB USA's argument is premised on consideration of a circumstance where one L3 region makes use of another L3 region associated with a core that is disabled. Sinharoy explains that in those circumstances the otherwise unused L3 region is treated as a victim cache. Ex. 1003 at 19. But this does not imply there is only a single controller. Instead, Sinharoy explains that a replacement policy included in each of the individual L3 cache controllers

categorizes an L3 cache's cache lines so as to include those installed as victims. *Id*. at 20. In other words, when a particular L3 region evicts a cache line to another L3 cache region, the controller associated with the evicted line maintains a record of that cache line. There is no single cache controller maintaining a global cache line state, rather each individual controller is maintaining cache line state for its data. This is entirely consistent with the circumstance addressed by Dr. Brogioli where one L3 region is attempting to evict a cache line to another L3 region that is being used by another active core. In such a situation, "the decision falls to the targeted local L3 region as to whether to accept or reject the cache line." Ex. 2002 at ¶ 38 (citing Sinharoy at 20). This need for acceptance is evidence of the separate nature of the cache controllers, *id*., and that nature does not change when the targeted L3 region is one associated with an inactive core.

MB USA's attempt to somehow rely on Sinharoy's table of cache state information (Table 2) is equally unavailing. According to MB USA, "Table 2 shows that, in certain cases, a local L3 region does not need to snoop all the other L2 or L3 regions to obtain a shared cache line." Reply at 10. But this is immaterial. A snoop may not be needed where, for example, a cache line is shared and a particular cache controller has all of the cache line state information it needs. *See,* Ex. 2003 at 26:18 – 27:22. But the telling

point is that cache misses must be snooped by all other cache controllers. If there were just a single controller, this would not be required. That it is demonstrates the separate nature of the controllers in Sinharoy. Ex. 2002 at ¶¶ 36-38. The Reply offers no rebuttal to this simple and conclusive point.

Finally, the Reply's discussion of the read/claim (RC) machines mentioned by Sinharoy provides no basis for concluding that only a single cache controller is present. MB USA contends that there are 24 RC machines shared across all of the L3 caches and that this must mean there is one logical cache controller. Reply at 11-12. Not so. Sinharoy makes clear that each L3 region employs its own pool of 24 RC machines:

> Storage accesses that miss the L2 cache access the 4-MB local L3 region via the core/L2 dispatch port. Those that hit in the 4-MB local L3 region are managed by a pool of eight read machines. Prefetches and some L2 store misses also access the L3 region via the core/L2 dispatch port and are managed by a pool of 24 lightweight RC machines.

Ex. 1003 at 19. Notice that Sinharoy is describing cache management within a local L3 region. That local L3 region is "managed by a pool of 24 lightweight RC machines." The RC machines are not shared across all of the

L3 regions of the cache (imagine the latencies that would exist if that were the case!). Instead, each L3 region employs its own set of RC machines.

Perhaps most damaging to MB USA's case is that the challenge is based on an attempt to demonstrate a logical single cache controller in Sinharoy, yet MB USA's expert admits the challenged claims do not recite a "logical" single cache controller. *See, e.g.*, Ex. 2003 at 57:12-19. Instead, the claims recite one cache controller that includes distributed portions. Sinharoy, on the other hand, describes separate cache controller for each L2 region and separate cache controllers for each L3 region. Ex. 1003 at 18, 19. Consequently, MB USA's challenge should be spurned.

IV. THE REPLY FAILS TO CONVINCINGLY DEMONSTRATE THAT XU IS PRIOR ART

MB USA maintains that Xu should be considered prior art because Mr. MacPherson offered testimony concerning its publication at a conference he did not state he attended, its later dissemination, although he did not state he had personal knowledge of that distribution, and a purported copyright record of the U.S. Copyright Office (Ex. 1019) which has been offered for the first time with Petitioner's Reply and itself has not been authenticated. Reply at 13-14. None of these rationales is convincing.

As to Mr. MacPherson's testimony, his declaration offers no indication that he had personal knowledge of any publication of Xu at the 2011 NORCHIP conference and merely states in conclusory fashion that Xu was so published. Ex. 1013 at ¶ 13. Conclusory testimony of this sort is not something the Board should consider. *See, e.g.*, PTAB Consolidated Trial Practice Guide at p. 35 (Nov. 2019). For example, Mr. MacPherson offers no indication he attended the conference in order to have personal knowledge of any materials that were distributed at the conference. And, he offers no testimony indicating how he came to believe that Xu was so distributed. Instead, Mr. MacPherson says only that he obtained a copy of Xu through the IEEE Xplore website, on or about Aug. 1, 2023, Ex. 1013 at ¶ 9, which is also the regular means by which such articles are made available to the public. *Id*. at ¶ 7. Whether or not Xu was "publicly available" on or about Aug. 1, 2023 says nothing about whether it is properly considered prior art to the '895 patent. To the extent Mr. MacPherson is relying on statements on the face of the Xu paper itself, *id*. at ¶ 10, those statements are hearsay and presumptively inadmissible. *See* Fed. R. Evid. 801(c), 802. In short, there is nothing in Mr. MacPherson's declaration that rises to the level of proving that Xu is prior art. *See ServiceNow Inc., v. Hewlett-Packard Co.,* IPR2015-00716, Paper No. 13, slip op. at 8, 15-17 (PTAB Aug. 26, 2015) (Board

denying institution where Petitioner failed to meet its burden to prove references asserted in its petition were prior art despite relying on an affidavit from an internet archive, expert declarations, two press releases, and the copyright dates from the face of the documents.).

As to the newly submitted record purporting to be from the U.S. Copyright Office, Ex. 1019, that record has not been authenticated. Despite Patent Owner's objections to this newly presented exhibit, see Patent Owner's Objections filed Sep. 16, 2024, MB USA has offered no supplemental evidence to authenticate the exhibit. Accordingly, Patent Owner has filed concurrently with this sur-reply a motion to exclude the exhibit. MB USA has not provided the testimony of any witness with personal knowledge of the website depicted in the printout that constitutes Ex. 1019, nor is there any other basis for concluding that the contents of the website are authentic. Hence, Ex. 1019 lacks authentication and should be deemed inadmissible. As such, it cannot be used to support the prior art status of Xu.

V. THE CHALLENGED CLAIMS ARE NOT UNPATENTABLE IN VIEW OF TANG

MB USA's Reply offers no compelling reason for the Board to second guess its original determination that Tang fails to teach or suggest a single caching agent that includes a plurality of distributed portions each associated with a corresponding one of a plurality of cores, as required by the challenged claims. The Reply admits, "in [Tang's] PBDC design, the LLC is partitioned into 'several ways' as the 'dedicated DMA Cache.'" Reply at 18. The LLC is a last level cache. Ex. 2003 at 44:6-9. Dr. Horst confirmed that this last level cache is shared by all of the cores in the processor. *Id*. at 44:10-12. And so too are the CPU data controller, prefetcher, global control logic, and cache coherence controller of the last level cache shared by all of the cores in the processor. *Id*. at 47:5-22. It is these components that control the CPU data ways (ways m+1 through n in Figure 10) of the last level cache. *Id*. at 47:5-9. And, to the extent the cache coherence controller is part of the cache controller, there is only one such cache coherence controller per processor, that is one for all of the cores. *Id*. at 47:10-16. In other words, when it comes to the cache controller in Tang, all of the cores of the processor share the same cache controller.

The challenged claims require something very different. They require a single caching agent that includes a plurality of distributed portions *each*

associated with a corresponding one of a plurality of cores. Tang cannot be read as teaching or suggesting such a caching agent. Ex. 2002 at ¶¶ 45-46. Hence, the challenged independent claims are not unpatentable in view of Tang.

VI. CONCLUSION

For at least the foregoing reasons and those in the Patent Owner's Response, the Board should find that Petitioner has failed to prove the unpatentability of any of the challenged claims.

Respectfully submitted,

Dated: October 22, 2024

/Tarek N. Fahmi/
Tarek N. Fahmi
Reg. No. 41,402

Ascenda Law Group, PC
2150 N First St., Suite 420
San Jose, CA 95131
Tel: 866-877-4883
Email: tarek.fahmi@ascendalaw.com

WORD-COUNT CERTIFICATE

I certify that the foregoing document complies with the type-volume limitation of 37 C.F.R. § 42.24(a) and (b) and contains 3,292 words (including words in drawings created by the Patent Owner or as annotations added by the Patent Owner to images reproduced in this document) based on the word count indicated by Microsoft® Word for Mac v. 16.85, excluding those portions exempted by the rule.

/Tarek N. Fahmi/
Tarek N. Fahmi
Reg. No. 41,402

CERTIFICATE OF SERVICE

The undersigned hereby certifies that a copy of the foregoing:

PATENT OWNER'S SUR-REPLY

and its accompanying exhibits (if any) were served on October 22, 2024, by filing this document though the PTAB P-TACTS System as well as by delivering a copy via email directed to the attorneys of record for the Petitioner at the following addresses:

Celine J. Crowson
Joseph J. Raffetto
Scott Hughes
Helen Y. Trac
Ryan Stephenson
Nicholas Rotz
Hogan Lovells US LLP
555 13th Street N.W.
Washington, D.C. 20004

Email:
celine.crowson@hogan.com
joseph.raffetto@hoganlovells.com
scott.hughes@hoganlovells.com
helen.trac@hoganlovells.com
ryan.stephenson@hoganlovells.com
nicholas.rotz@hoganlovells.com

The parties have agreed to electronic service in this proceeding.

Respectfully submitted,

Dated: October 22, 2024

/Tarek N. Fahmi/
Tarek N. Fahmi
Reg. No. 41,402

Ascenda Law Group, PC
2150 N First St., Suite 420
San Jose, CA 95131
Tel: 866-877-4883
Email: patents@ascendalaw.com