

EXHIBIT 803-A-02

Akhil Arunkumar et al., *MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability*, Procs. ISCA '17 (2017) (“Arunkumar”)

Akhil Arunkumar et al., *MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability*, Procs. ISCA '17 (2017) (“Arunkumar”) anticipates and/or renders obvious (alone or in combination with other references), including at least under Complainant’s apparent infringement theory:

- claims 1-3, 7-10 of U.S. Patent No. 11,841,803 (“the ’803 patent”) under 35 U.S.C. §§ 102 and/or 103.

NVIDIA has charted the portions of the reference that correspond to each element, including at least under Complainant’s apparent interpretation of the claims as reflected in Complainant’s infringement contentions. Where the chart below states that the reference “discloses,” “teaches,” or “suggests” a claim limitation, such disclosure, teaching, or suggestion may be express, implied, or inherent. The cited evidence is merely illustrative, and NVIDIA reserves the right to cite alternative or additional evidence.

No part of this exemplary chart construes, or is intended to construe, the specification, file history, or claims of the asserted patents. In addition, nothing stated in this chart shall be treated as an admission or suggestion that NVIDIA agrees with Complainant regarding either the scope of any of the asserted claims or Complainant’s apparent interpretation of the scope of the asserted claims, or that NVIDIA’s accused products meet any limitations of the claims.

To the extent that Complainant contends that Arunkumar does not disclose one or more limitations of the asserted claims, it would have been obvious to combine the teachings of Arunkumar with (1) the knowledge of one of ordinary skill in the art, and/or (2) the teachings of any of the prior art set forth in NVIDIA’s other invalidity charts and cover pleading, to show all the limitations of the asserted claims, including for the reasons set forth in NVIDIA’s Invalidity Contention Cover Pleading, which are incorporated herein.

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
1[Pre]: A system, comprising:	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Arunkumar at Fig. 1. Figure 1: MCM-GPU: Aggregating GPU modules and DRAM on a single package. Arunkumar at Fig. 3.

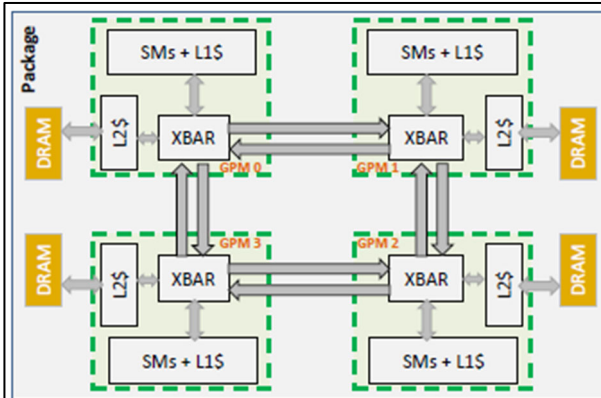


Figure 3: Basic MCM-GPU architecture comprising four GPU modules (GPMs).

Arunkumar at Fig. 5.

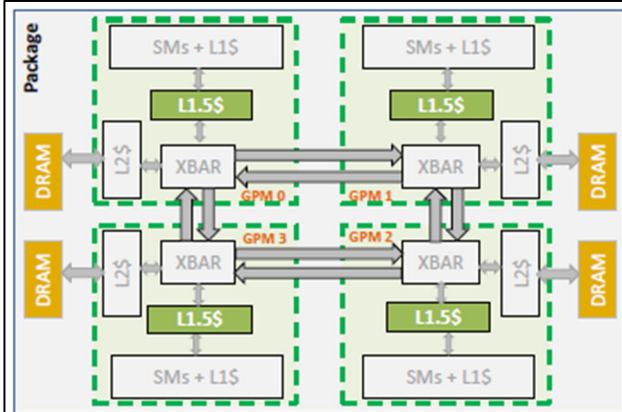


Figure 5: MCM-GPU architecture equipped with L1.5 GPM-side cache to capture remote data and effectively reduce inter-GPM bandwidth and data access latency.

Arunkumar at Table 3.

Number of GPMs	4
Total number of SMs.	256
GPU frequency	1GHz
Max number of warps	64 per SM
Warp scheduler	Greedy then Round Robin
L1 data cache	128 KB per SM, 128B lines, 4 ways
Total L2 cache	16MB, 128B lines, 16 ways
Inter-GPM interconnect	768GB/s per link, Ring, 32 cycles/hop
Total DRAM bandwidth	3 TB/s
DRAM latency	100ns

Table 3: Baseline MCM-GPU configuration.

Arunkumar at *1 (Abstract) (“Historically, improvements in GPU-based high performance computing have been tightly coupled to transistor scaling. As Moore’s law slows down, and

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	the number of transistors per die no longer grows at historical rates, the performance curve of single monolithic GPUs will ultimately plateau. However, the need for higher performing GPUs continues to exist in many domains. To address this need, in this paper we demonstrate that package-level integration of multiple GPU modules to build larger logical GPUs can enable continuous performance scaling beyond Moore's law. Specifically, we propose partitioning GPUs into easily manufacturable basic GPU Modules (GPMs), and integrating them on package using high bandwidth and power efficient signaling technologies. We lay out the details and evaluate the feasibility of a basic Multi-Chip-Module GPU (MCMGPU) design. We then propose three architectural optimizations that significantly improve GPM data locality and minimize the sensitivity on inter-GPM bandwidth. Our evaluation shows that the optimized MCM-GPU achieves 22.8% speedup and 5x inter-GPM bandwidth reduction when compared to the basic MCM-GPU architecture. Most importantly, the optimized MCM-GPU design is 45.5% faster than the largest implementable monolithic GPU, and performs within 10% of a hypothetical (and unbuildable) monolithic GPU. Lastly we show that our optimized MCM-GPU is 26.8% faster than an equally equipped Multi-GPU system with the same total number of SMs and DRAM bandwidth.”). Arunkumar at *1 (§ 1) (“GPU-based compute acceleration is the main vehicle propelling the performance of high performance computing (HPC) systems, machine learning and data analytics applications in large-scale cloud installations, and personal computing devices. In such devices, each computing node or computing device typically consists of a CPU with one or more GPU accelerators. The path forward in any of these domains, either to exascale performance in HPC, or to human-level artificial intelligence using deep convolutional neural networks, relies on the ability to continuously scale GPU performance. As a result, in such systems, each GPU has the maximum possible transistor count at the most advanced technology node, and uses state-of-the-art memory technology. Until recently, transistor scaling improved single GPU performance by increasing the Streaming Multiprocessor (SM) count between GPU generations. However, transistor scaling has dramatically slowed down and is expected to eventually come to an end. Furthermore, optic and manufacturing limitations constrain the reticle size which in turn constrains the maximum die size (e.g. $\approx$

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	800mm²). Moreover, very large dies have extremely low yield due to large numbers of irreparable manufacturing faults. This increases the cost of large monolithic GPUs to undesirable levels. Consequently, these trends limit future scaling of single GPU performance and potentially bring it to a halt.” (citations omitted)). Arunkumar at *2 (§ 1) (“Leveraging this new integration tier, we propose a novel Multi-Chip Module GPU (MCM-GPU) architecture that enables continued GPU performance scaling despite the slowdown of transistor scaling and photoretic limitations. Our proposal aggregates multiple GPU Modules (GPMs) within a single package as illustrated in Figure 1. First, we detail the basic MCM-GPU architecture that leverages NVIDIA’s state-of-the-art Ground Reference Signaling (GRS). We then optimize our proposed MCM-GPU design using three architectural innovations targeted at improving locality and minimizing inter-GPM communication: (i) hardware caches to capture remote traffic in the local GPM, (ii) distributed and batched co-operative thread array (CTA) scheduling to better leverage inter-CTA locality within a GPM, and (iii) first touch page allocation policy to minimize inter-GPM traffic. Overall, this paper makes the following contributions:” (citation omitted)). Arunkumar at *2 (§ 1) (“We present modular MCM-GPU with 256 SMs and discuss its memory system, on-package integration, and signaling technology. We show its performance sensitivity to inter-GPM bandwidth both analytically and via simulations. Our evaluation shows that since inter-GPM bandwidth is lower than a monolithic GPU’s on-chip bandwidth, an onpackage non-uniform memory access (NUMA) architecture is exposed in the MCM-GPU.”). Arunkumar at *2 (§ 1) (“We propose a locality-aware MCM-GPU architecture, better suited to its NUMA nature. We use architectural enhancements to mitigate the penalty introduced by nonuniform memory accesses. Our evaluations show that these optimizations provide an impressive 5x inter-GPM bandwidth reduction, and result in a 22.8% performance speedup compared to the baseline MCM-GPU. Our optimized MCMGPU architecture achieves a

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	44.5% speedup over the largest possible monolithic GPU (assumed as a 128 SMs GPU), and comes within 10% of the performance of an unbuildable similarly sized monolithic GPU.”). Arunkumar at *3 (§ 3.1) (“In this paper we propose the MCM-GPU as a collection of GPMs that share resources and are presented to software and programmers as a single monolithic GPU. Pooled hardware resources, and shared I/O are concentrated in a shared on-package module (the SYS + I/O module shown in Figure 1). The goal for this MCM-GPU is to provide the same performance characteristics as a single (unmanufacturable) monolithic die. By doing so, the operating system and programmers are isolated from the fact that a single logical GPU may now be several GPMs working in conjunction. There are two key advantages to this organization. First, it enables resource sharing of underutilized structures within a single GPU and eliminates hardware replication among GPMs. Second, applications will be able to transparently leverage bigger and more capable GPUs, without any additional programming effort.”).
1[A]: a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array,	Including at least under Complainant’s apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Arunkumar at Fig. 1.

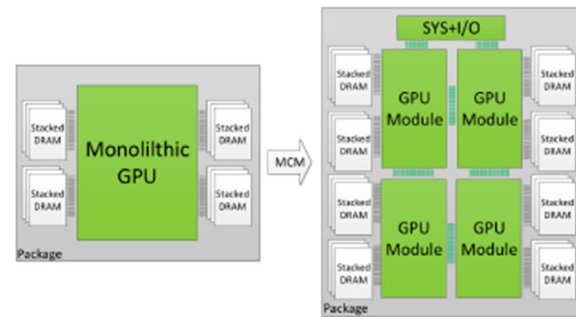


Figure 1: MCM-GPU: Aggregating GPU modules and DRAM on a single package.

Arunkumar at Fig. 3.

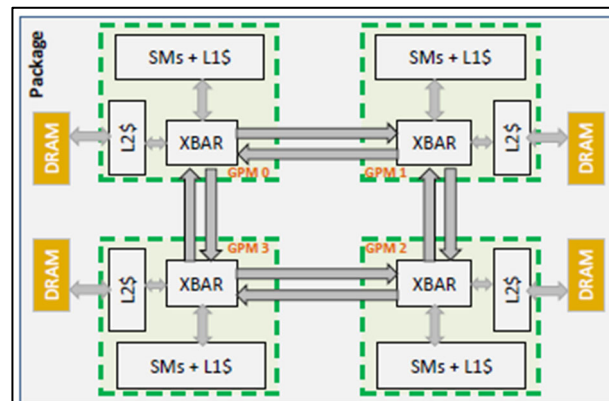


Figure 3: Basic MCM-GPU architecture comprising four GPU modules (GPMs).

Arunkumar at Fig. 5.

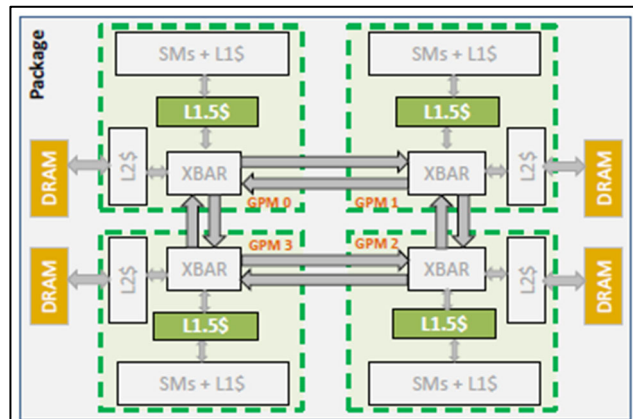


Figure 5: MCM-GPU architecture equipped with L1.5 GPM-side cache to capture remote data and effectively reduce inter-GPM bandwidth and data access latency.

Arunkumar at *1-2 (§ 1) (“An alternate approach to scaling performance without exceeding the maximum chip size relies on multiple GPUs connected on a PCB, such as the Tesla K10 and K80. However, as we show in this paper, it is hard to scale GPU workloads on such “multi-GPU” systems, even if they scale very well on a single GPU. This is due to multiple unsolved challenges related to work partitioning, load balancing, and data sharing across the slow on-board interconnection network. However, due to recent advances in packaging and signaling technologies, package-level integration provides a promising integration tier that lies between the existing on-chip and on-board integration technologies.” (citations omitted)).

Arunkumar at *3 (§ 3) (“The proposed Multi-Chip Module GPU (MCM-GPU) architecture is based on aggregating multiple GPU modules (GPMs) within a single package, as opposed to today’s GPU architecture based on a single monolithic die. This enables scaling single GPU

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	performance by increasing the number of transistors, DRAM, and I/O bandwidth per GPU. Figure 1 shows an example of an MCM-GPU architecture with four GPMs on a single package that potentially enables up to 4× the number of SMs (chip area) and 2× the memory bandwidth (edge size) compared to the largest GPU in production today.”). Arunkumar at *3 (§ 3.1) (“In this paper we propose the MCM-GPU as a collection of GPMs that share resources and are presented to software and programmers as a single monolithic GPU. Pooled hardware resources, and shared I/O are concentrated in a shared on-package module (the SYS + I/O module shown in Figure 1). The goal for this MCM-GPU is to provide the same performance characteristics as a single (unmanufacturable) monolithic die. By doing so, the operating system and programmers are isolated from the fact that a single logical GPU may now be several GPMs working in conjunction. There are two key advantages to this organization. First, it enables resource sharing of underutilized structures within a single GPU and eliminates hardware replication among GPMs. Second, applications will be able to transparently leverage bigger and more capable GPUs, without any additional programming effort.”). Arunkumar at *5 (§ 4) (“We use an NVIDIA in-house simulator to conduct our performance studies. We model the GPU to be similar to, but extrapolated in size compared to the recently released NVIDIA Pascal GPU [17]. Our SMs are modeled as in-order execution processors that accurately model warp-level parallelism. We model a multi-level cache hierarchy with a private L1 cache per SM and a shared L2 cache. Caches are banked such that they can provide the necessary parallelism to saturate DRAM bandwidth. We model software based cache coherence in the private caches, similar to state-of-the-art GPUs. Table 3 summarizes baseline simulation parameters.”). Arunkumar at *5 (§ 4) (“We study a diverse set of 48 benchmarks that are taken from four benchmark suites. Our evaluation includes a set of production class HPC benchmarks from the CORAL benchmarks [6], graph applications from Lonestar suite [43], compute applications from Rodinia [24], and a set of NVIDIA in-house CUDA benchmarks. Our application set

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	covers a wide range of GPU application domains including machine learning, deep neural networks, fluid dynamics, medical imaging, graph search, etc. We classify our applications into two categories based on the available parallelism—high parallelism applications (parallel efficiency $\geq 25\%$) and limited parallelism applications (parallel efficiency $< 25\%$). We further categorize the high parallelism applications based on whether they are memory-intensive (M-Intensive) or compute-intensive (C-Intensive). We classify an application as memory-intensive if it suffers from more than 20% performance degradation if the system memory bandwidth is halved. In the interest of space, we present the detailed per-application results for the M-Intensive category workloads and present only the average numbers for the C-Intensive and limited-parallelism workloads. The set of M-Intensive benchmarks, and their memory footprints are detailed in Table 4. We simulate all our benchmarks for one billion warp instructions, or to completion, whichever occurs first.”). Arunkumar at *11 (§ 7) (“Finally, we propose to expose the MCM-GPU as a single logical GPU via hardware innovations and extensions to the driver software to provide programmer- and OS-transparent execution. While there have been studies that propose techniques to efficiently utilize multi-GPU systems [20, 23, 33, 36], none of the proposals provide a fully transparent approach suitable for MCM- GPUs.”). This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
1[B]: wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and	Including at least under Complainant’s apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. See evidence cited for ’803 patent limitation 1[A].

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
1[C]: a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink,	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Arunkumar at Fig. 1. Figure 1: MCM-GPU: Aggregating GPU modules and DRAM on a single package. Arunkumar at Fig. 3.

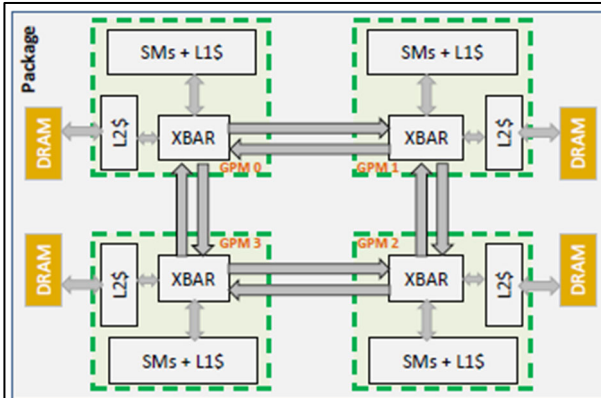


Figure 3: Basic MCM-GPU architecture comprising four GPU modules (GPMs).

Arunkumar at Fig. 5.

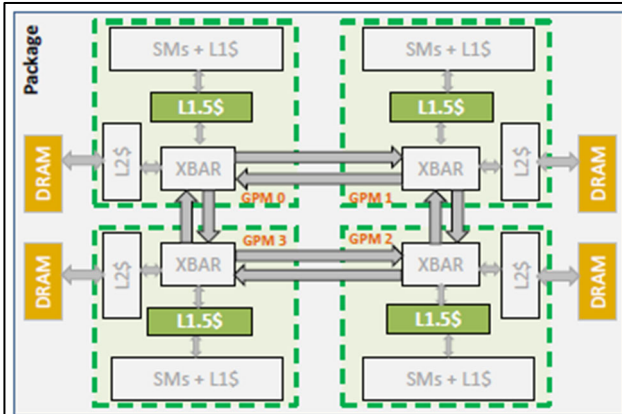


Figure 5: MCM-GPU architecture equipped with L1.5 GPM-side cache to capture remote data and effectively reduce inter-GPM bandwidth and data access latency.

Arunkumar at Table 3.

Number of GPMs	4
Total number of SMs.	256
GPU frequency	1GHz
Max number of warps	64 per SM
Warp scheduler	Greedy then Round Robin
L1 data cache	128 KB per SM, 128B lines, 4 ways
Total L2 cache	16MB, 128B lines, 16 ways
Inter-GPM interconnect	768GB/s per link, Ring, 32 cycles/hop
Total DRAM bandwidth	3 TB/s
DRAM latency	100ns

Table 3: Baseline MCM-GPU configuration.

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	Arunkumar at *1-2 (§ 1) (“An alternate approach to scaling performance without exceeding the maximum chip size relies on multiple GPUs connected on a PCB, such as the Tesla K10 and K80. However, as we show in this paper, it is hard to scale GPU workloads on such “multi-GPU” systems, even if they scale very well on a single GPU. This is due to multiple unsolved challenges related to work partitioning, load balancing, and data sharing across the slow on-board interconnection network. However, due to recent advances in packaging and signaling technologies, package-level integration provides a promising integration tier that lies between the existing on-chip and on-board integration technologies.” (citations omitted)). Arunkumar at *2 (§ 1) (“Leveraging this new integration tier, we propose a novel Multi-Chip Module GPU (MCM-GPU) architecture that enables continued GPU performance scaling despite the slowdown of transistor scaling and photoretic limitations. Our proposal aggregates multiple GPU Modules (GPMs) within a single package as illustrated in Figure 1. First, we detail the basic MCM-GPU architecture that leverages NVIDIA’s state-of-the-art Ground Reference Signaling (GRS). We then optimize our proposed MCM-GPU design using three architectural innovations targeted at improving locality and minimizing inter-GPM communication: (i) hardware caches to capture remote traffic in the local GPM, (ii) distributed and batched co-operative thread array (CTA) scheduling to better leverage inter-CTA locality within a GPM, and (iii) first touch page allocation policy to minimize inter-GPM traffic. Overall, this paper makes the following contributions:” (citation omitted)). Arunkumar at *2 (§ 1) (“We present modular MCM-GPU with 256 SMs and discuss its memory system, on-package integration, and signaling technology. We show its performance sensitivity to inter-GPM bandwidth both analytically and via simulations. Our evaluation shows that since inter-GPM bandwidth is lower than a monolithic GPU’s on-chip bandwidth, an onpackage non-uniform memory access (NUMA) architecture is exposed in the MCM-GPU.”). Arunkumar at *3 (§ 2.3) (“Recent advances in organic package technology are expected to address today’s challenges and enable on-package integration of active components. For

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	example, next generation packages are expected to support a 77mm substrate dimension, providing enough room to integrate the MCM-GPU architecture described in this paper. Furthermore, advances in package level signaling technologies such as NVIDIA's Ground-Referenced Signaling (GRS), offer the necessary high-speed, high-bandwidth signaling for organic package substrates. GRS signaling can operate at 20 Gb/s while consuming just 0.54 pJ/bit in a standard 28nm process. As this technology evolves, we can expect it to support up to multiple TB/s of on-package bandwidth. This makes the on-package signaling bandwidth eight times larger than that of on-board signaling." (citations omitted)). Arunkumar at *3 (§ 2.3) ("The aforementioned factors make package level integration a promising integration tier, that qualitatively falls in between chip- and board-level integration tiers (See Table 2). In this paper, we aim to take advantage of this integration tier and set the ambitious goal of exploring how to manufacture a 2× more capable GPU, comprising 256 or more SMs within a single GPU package."). Arunkumar at *3 (§ 3) ("The proposed Multi-Chip Module GPU (MCM-GPU) architecture is based on aggregating multiple GPU modules (GPMs) within a single package, as opposed to today's GPU architecture based on a single monolithic die. This enables scaling single GPU performance by increasing the number of transistors, DRAM, and I/O bandwidth per GPU. Figure 1 shows an example of an MCM-GPU architecture with four GPMs on a single package that potentially enables up to 4× the number of SMs (chip area) and 2× the memory bandwidth (edge size) compared to the largest GPU in production today."). Arunkumar at *3 (§ 3.1) ("In this paper we propose the MCM-GPU as a collection of GPMs that share resources and are presented to software and programmers as a single monolithic GPU. Pooled hardware resources, and shared I/O are concentrated in a shared on-package module (the SYS + I/O module shown in Figure 1). The goal for this MCM-GPU is to provide the same performance characteristics as a single (unmanufacturable) monolithic die. By doing so, the operating system and programmers are isolated from the fact that a single logical GPU may now be several GPMs working in conjunction. There are two key advantages to this

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	organization. First, it enables resource sharing of underutilized structures within a single GPU and eliminates hardware replication among GPMs. Second, applications will be able to transparently leverage bigger and more capable GPUs, without any additional programming effort.”). Arunkumar at *4 (§ 3.2) (“As discussed in Sections 1 and 2, moving forward beyond 128 SM counts will almost certainly require at least two GPMs in a GPU. Since smaller GPMs are significantly more cost-effective, in this paper we evaluate building a 256 SM GPU out of four GPMs of 64 SMs each. This way each GPM is configured very similarly to today’s biggest GPUs. Area-wise each GPM is expected to be 40% - 60% smaller than today’s biggest GPU assuming the process node shrinks to 10nm or 7nm. Each GPM consists of multiple SMs along with their private L1 caches. SMs are connected through the GPM-Xbar to a GPM memory subsystem comprising a local memory-side L2 cache and DRAM partition. The GPM-Xbar also provides connectivity to adjacent GPMs via on-package GRS inter-GPM links.” (citations omitted)). Arunkumar at *4 (§ 3.2) (“Figure 3 shows the high-level diagram of this 4-GPM MCMGPU. Such an MCM-GPU is expected to be equipped with 3TB/s of total DRAM bandwidth and 16MB of total L2 cache. All DRAM partitions provide a globally shared memory address space across all GPMs. Addresses are fine-grain interleaved across all physical DRAM partitions for maximum resource utilization. GPM-Xbars route memory accesses to the proper location (either the local or a remote L2 cache bank) based on the physical address. They also collectively provide a modular on-package ring or mesh interconnect network. Such organization provides spatial traffic locality among local SMs and memory partitions, and reduces on-package bandwidth requirements. Other network topologies are also possible especially with growing number of GPMs, but a full exploration of inter-GPM network topologies is outside the scope of this paper. The L2 cache is a memory-side cache, caching data only from its local DRAM partition. As such, there is only one location for each cache line, and no cache coherency is required across the L2 cache banks. In the baseline MCM-GPU architecture we employ a centralized CTA scheduler that schedules CTAs to MCM-GPU

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	SMs globally in a round-robin manner as SMs become available for execution, as in the case of a typical monolithic GPU.”). Arunkumar at *4 (§ 3.2) (“The MCM-GPU memory system is a Non Uniform Memory Access (NUMA) architecture, as its inter-GPM links are not expected to provide full aggregated DRAM bandwidth to each GPM. Moreover, an additional latency penalty is expected when accessing memory on remote GPMs. This latency includes data movement time within the local GPM to the edge of the die, serialization and deserialization latency over the inter-GPM link, and the wire latency to the next GPM. We estimate each additional inter-GPM hop latency, for a potentially multi-hop path in the on-package interconnect as 32 cycles. Each additional hop also adds an energy cost compared to a local DRAM access. Even though we expect the MCM-GPU architecture to incur these bandwidth, latency, and energy penalties, we expect them to be much lower compared to off-package interconnects in a multi-GPU system (see Table 2).”). Arunkumar at *5 (§ 3.3.3) (“In this paper we assume a low-effort, low-cost, and low-energy link design point of 768GB/s and make an attempt to bridge the performance gap due to relatively lower bandwidth settings via architectural innovations that improve communication locality and essentially eliminate the need for more costly and less energy efficient links. The rest of the paper proposes architectural mechanisms to capture data-locality within GPM modules, which eliminate the need for costly inter-GPM bandwidth solutions.”). Arunkumar at *11 (§ 6.2) (“Besides enabling performance scalability, MCM-GPUs are energy and cost efficient. MCM-GPUs are energy efficient as they enable denser integration of GPU modules on a package that alternatively would have to be connected at a PCB level as in a multi-GPU case. In doing so, MCM-GPUs require significantly smaller system footprint and utilize more efficient interconnect technologies, e.g., 0.5 pJ/b on-package vs 10 pJ/b on-board interconnect. Moreover, if we assume almost constant GPU and system power dissipation, the performance advantages of the MCM-GPU translate to additional energy savings. In addition, superior transistor density achieved by the MCM-GPU approach allows to lower GPU

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	operating voltage and frequency. This moves the GPU to a more power-efficient operating point on the transistor voltage-frequency curve. Consequently, it allows trading off ample performance (achieved via abundant parallelism and number of transistors in package) for better power efficiency.”). Arunkumar at *11 (§ 7) (“Finally, we propose to expose the MCM-GPU as a single logical GPU via hardware innovations and extensions to the driver software to provide programmer- and OS-transparent execution. While there have been studies that propose techniques to efficiently utilize multi-GPU systems [20, 23, 33, 36], none of the proposals provide a fully transparent approach suitable for MCM- GPUs.”). This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
1[D]: wherein the passive crosslink is dedicated for inter-chiplet communications.	Including at least under Complainant’s apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See</i> evidence cited for ’803 patent limitations 1[A] - 1[C]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
2. The system of claim 1, wherein the passive crosslink comprises a passive interposer die.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this claim. <i>See evidence cited for '803 patent limitation 1[D].</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
3[A]: The system of claim 1, wherein the first GPU chiplet comprises a first PHY region including conductor structures for chiplet-to-chiplet communications,	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See evidence cited for '803 patent claims 1-2.</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
3[B]: and further wherein the second GPU chiplet comprises a second PHY region including conductor structures for chiplet-to-chiplet communications.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See evidence cited for '803 patent limitation 3[A].</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
7. The system of claim 1, wherein the passive crosslink communicably couples all GPU chiplets in the GPU chiplet array.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this claim. <i>See</i> evidence cited for '803 patent limitation 1[D]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
8[A]: The system of claim 1, further comprising: a first cache memory hierarchy at the first GPU chiplet,	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See, e.g.,</i> Arunkumar at Fig. 3.

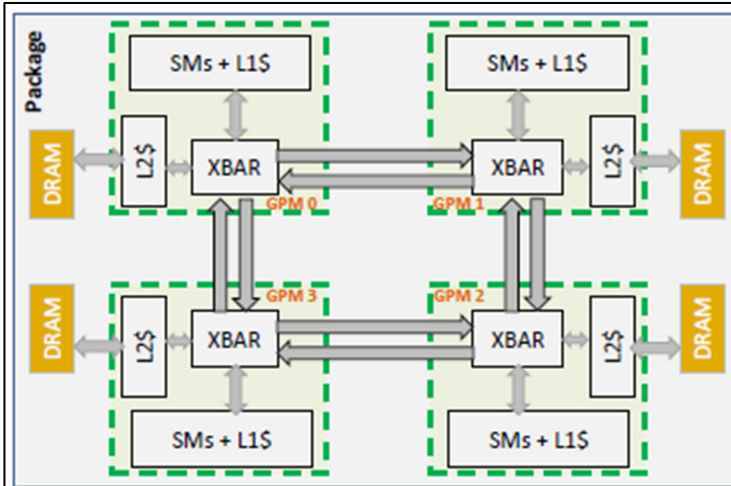


Figure 3: Basic MCM-GPU architecture comprising four GPU modules (GPMs).

Arunkumar at Fig. 5.

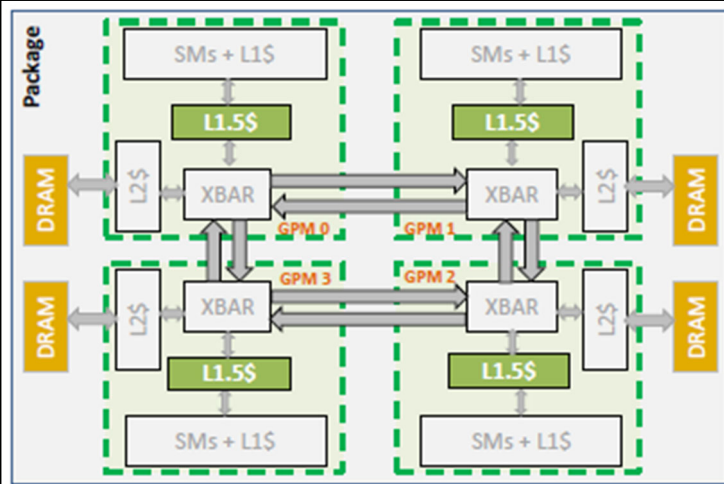


Figure 5: MCM-GPU architecture equipped with L1.5 GPM-side cache to capture remote data and effectively reduce inter-GPM bandwidth and data access latency.

Arunkumar at Table 3.

Number of GPMs	4
Total number of SMs.	256
GPU frequency	1GHz
Max number of warps	64 per SM
Warp scheduler	Greedy then Round Robin
L1 data cache	128 KB per SM, 128B lines, 4 ways
Total L2 cache	16MB, 128B lines, 16 ways
Inter-GPM interconnect	768GB/s per link, Ring, 32 cycles/hop
Total DRAM bandwidth	3 TB/s
DRAM latency	100ns

Table 3: Baseline MCM-GPU configuration.

Arunkumar at *4 (§ 3.2) (“As discussed in Sections 1 and 2, moving forward beyond 128 SM counts will almost certainly require at least two GPMs in a GPU. Since smaller GPMs are significantly more cost-effective, in this paper we evaluate building a 256 SM GPU out of four GPMs of 64 SMs each. This way each GPM is configured very similarly to today’s biggest GPUs. Area-wise each GPM is expected to be 40% - 60% smaller than today’s biggest GPU assuming the process node shrinks to 10nm or 7nm. Each GPM consists of multiple SMs along with their private L1 caches. SMs are connected through the GPM-Xbar to a GPM memory subsystem comprising a local memory-side L2 cache and DRAM partition. The GPM-Xbar also provides connectivity to adjacent GPMs via on-package GRS inter-GPM links.” (citations omitted)).

Arunkumar at *4 (§ 3.2) (“Figure 3 shows the high-level diagram of this 4-GPM MCM-GPU. Such an MCM-GPU is expected to be equipped with 3TB/s of total DRAM bandwidth and 16MB of total L2 cache. All DRAM partitions provide a globally shared memory address space across all GPMs. Addresses are fine-grain interleaved across all physical DRAM partitions for maximum resource utilization. GPM-Xbars route memory accesses to the proper location (either the local or a remote L2 cache bank) based on the physical address. They also collectively provide a modular on-package ring or mesh interconnect network. Such organization provides spatial traffic locality among local SMs and memory partitions, and

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	reduces on-package bandwidth requirements. Other network topologies are also possible especially with growing number of GPMs, but a full exploration of inter-GPM network topologies is outside the scope of this paper. The L2 cache is a memory-side cache, caching data only from its local DRAM partition. As such, there is only one location for each cache line, and no cache coherency is required across the L2 cache banks. In the baseline MCM-GPU architecture we employ a centralized CTA scheduler that schedules CTAs to MCM-GPU SMs globally in a round-robin manner as SMs become available for execution, as in the case of a typical monolithic GPU.”). Arunkumar at *5 (§ 4) (“We use an NVIDIA in-house simulator to conduct our performance studies. We model the GPU to be similar to, but extrapolated in size compared to the recently released NVIDIA Pascal GPU. Our SMs are modeled as in-order execution processors that accurately model warp-level parallelism. We model a multi-level cache hierarchy with a private L1 cache per SM and a shared L2 cache. Caches are banked such that they can provide the necessary parallelism to saturate DRAM bandwidth. We model software based cache coherence in the private caches, similar to state-of-the-art GPUs. Table 3 summarizes baseline simulation parameters.” (citation omitted)). Arunkumar at *6 (§ 5.1.1) (“The first mechanism we propose to reduce on-package link bandwidth is to enhance the MCM-GPU cache hierarchy. We propose to augment our baseline GPM architecture in Figure 3 with a GPM-side cache that resides between the L1 and L2 caches. We call this new cache level the L1.5 cache as shown in Figure 5. Architecturally, the L1.5 cache can be viewed as an extension of the L1 cache and is shared by all SMs inside a GPM. We propose that the L1.5 cache stores remote data accesses made by a GPM partition. In other words, all local memory accesses will bypass the L1.5 cache. Doing so reduces both remote data access latency and inter-GPM bandwidth. Both these properties improve performance and reduce energy consumption by avoiding inter-GPM communication.”). Arunkumar at *6 (§ 5.1.1) (“To avoid increasing on-die transistor overhead for the L1.5 cache, we add it by rebalancing the cache capacity between the L2 and L1.5 caches in an iso-

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	transistor manner. We extend the GPU L1 cache coherence mechanism to the GPM-side L1.5 caches as well. This way, whenever an L1 cache is flushed on a synchronization event such as reaching a kernel execution boundary, the L1.5 cache is flushed as well. Since the L1.5 cache can receive multiple invalidation commands from GPM SMs, we make sure that the L1.5 cache is invalidated only once for each synchronization event.”). This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
8[B]: wherein a first level of the first cache memory hierarchy is coherent within the first GPU chiplet; and	Including at least under Complainant’s apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See</i> evidence cited for ’803 patent limitation 8[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA’s other invalidity charts and cover pleading.
8[C]: a second cache memory hierarchy at the second GPU chiplet,	Including at least under Complainant’s apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See</i> evidence cited for ’803 patent limitation 8[A]. This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
8[D]: wherein a first level of the second cache memory hierarchy is coherent within the second GPU chiplet.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See evidence cited for '803 patent limitation 8[B].</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
9[A]: The system of claim 8, further comprising: a unified cache memory including both a last level of the first cache memory hierarchy and a last level of the second cache memory hierarchy,	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See evidence cited for '803 patent claim 8.</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
9[B]: wherein the unified cache memory is coherent across all GPU chiplets of the GPU chiplet array.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this limitation. <i>See evidence cited for '803 patent claim 8.</i>

U.S. Patent No. 11,841,803	Akhil Arunkumar et al., <i>MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability</i> , Procs. ISCA '17 (2017)
	This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.
10. The system of claim 1, further comprising: a plurality of conductive pillars coupling a circuit board substrate to a first non-PHY region of the first GPU chiplet and a second non-PHY region of the second GPU chiplet.	Including at least under Complainant's apparent infringement theory, Arunkumar discloses and/or renders obvious this claim. <i>See evidence cited for '803 patent limitation 1[C].</i> This limitation also would have been obvious to a person of ordinary skill in the art based on (1) Arunkumar alone; (2) Arunkumar in combination with the general knowledge of a person of ordinary skill in the art; and/or (3) Arunkumar in combination with the teachings with respect to this claim element as detailed in NVIDIA's other invalidity charts and cover pleading.