BUNDESPATENTGERICHT

17 W (pat) 5/21

(Aktenzeichen)

Verkündet am
15. März 2022
…

B E S C H L U S S

In der Beschwerdesache

betreffend die Teilanmeldung 10 2014 020 058.2

…

hat der 17. Senat (Technischer Beschwerdesenat) des Bundespatentgerichts auf die mündliche Verhandlung vom 15. März 2022 unter Mitwirkung des Vorsitzenden Richters Dipl.-Phys. Dr. Morawek, des Richters Dipl.-Phys. Dr. Forkel, der Richterin Akintche und des Richters Dipl.-Phys. Dr. Städele

beschlossen:

ECLI:DE:BPatG:2022:150322B17Wpat5.21.0

Die Beschwerde wird zurückgewiesen.

G r ü n d e

I.

Die vorliegende Patentanmeldung ist aufgrund einer Teilungserklärung im Beschwerdeverfahren aus der Patentanmeldung 10 2014 119 048.3 entstanden und trägt die Bezeichnung

„ Ausführung von Verarbeitungsvorgängen in einer SIMD-Verarbeitungseinheit “.

Die Stammanmeldung, die eine GB-Priorität vom 18. Dezember 2013 in Anspruch nimmt, wurde durch Beschluss der Prüfungsstelle für Klasse G06F des Deutschen Patent- und Markenamts vom 20. Januar 2020 aus Gründen des Bescheids vom 24. September 2019 zurückgewiesen. In dem in Bezug genommenen Bescheid ist sinngemäß ausgeführt, mangels einer ausreichenden Definition des unter Schutz zu stellenden Gegenstands erfülle der (damalige) Patentanspruch 1 nicht die Anforderungen nach §34 Abs. 3 Nr. 3 PatG und sei aus diesem Grund nicht gewährbar. Selbst wenn dieser Mangel keine Berücksichtigung fände, sei der Gegenstand des Patentanspruchs 1 mit Rücksicht auf den der Druckschrift D1 entnehmbaren Stand der Technik nicht neu und beruhe gegenüber der Druckschrift D2 nicht auf einer erfinderischen Tätigkeit.

Die gegen den Beschluss der Prüfungsstelle gerichtete Beschwerde vom 19 Februar 2020 war beim Senat unter dem Aktenzeichen 17 W (pat) 20/20 anhängig und wurde mit Beschluss vom 15. Juni 2021 zurückgewiesen. Im Laufe des Beschwerdeverfahrens erklärte die Anmelderin mit Schriftsatz vom

23. Dezember 2020 die Teilung der Anmeldung gegenüber dem Bundespatentgericht. Das Deutsche Patent- und Markenamt legte für die Trennanmeldung eine Akte mit dem Aktenzeichen 10 2014 020 058.2 an und stellte am 3. März 2021 den fristgerechten Eingang der erforderlichen Unterlagen und der Gebühren fest (siehe Gerichtsakte Bl. 14).

Die Anmelderin stellt den Antrag,

1. das nachgesuchte Patent mit folgenden Unterlagen zu erteilen:

   Patentansprüche 1 bis 9, eingegangen am 9. März 2022,

   Beschreibung Seiten 1 bis 34, eingegangen am 23. Dezember 2020,

   9 Blatt Zeichnungen mit Figuren 1 bis 10, eingegangen am 23 Dezember 2020.

2. hilfsweise die Anmeldung zur Fortführung des Prüfungsverfahrens an das Deutsche Patent- und Markenamt zurückzuverweisen.

Im Prüfungsverfahren der Stammanmeldung vor dem Deutschen Patent- und Markenamt sind u.a. die Druckschriften

G9 Giesen, Fabian: A trip through the Graphics Pipeline 2011, part 8, 10.07.2011. URL: https://fgiesen.wordpress.com/2011/07/10/a-trip-through-the-graphics-pipeline-2011-part-8/, [abgerufen am 10.09.2019]

und

D2 Rhu, Minsoo; Erez, Mattan: Maximizing SIMD Resource Utilization in GPGPUs with SIMD Lane Permutation. In: SIGARCH Comput. Archit. News, Volume 41, June 2013, Issue 3, S. 356-367. - ISSN 0163-5964. http://doi.acm.org/10.1145/2508148.2485953, [abgerufen am 18.04.2016]

genannt worden. Vom Senat wurde zusätzlich die Druckschrift

D11 US 2007 / 0 182 746 A1

eingeführt.

Der geltende **Patentanspruch 1** lautet mit einer möglichen Gliederung versehen:

M1 SIMD(Single Instruction Multiple Data)-Verarbeitungseinheit,

die ausgelegt ist,

M2 eine Mehrzahl von Verarbeitungsvorgängen zu verarbeiten,

M2a die jeweils Arbeitselemente bis zu einer vorbestimmten Höchstanzahl umfassen,

M2b wobei die Arbeitselemente eines Verarbeitungsvorgangs ausgelegt sind, eine gemeinsame Sequenz von Befehlen an entsprechenden Datenelementen auszuführen, wobei

M2c ein Verarbeitungsvorgang durch Blöcke von Arbeitselementen gebildet wird, die jeweils genau vier Arbeitselemente umfassen und

M2d sich auf ein jeweiliges Pixel-Quad beziehen, wobei jedes Pixel-Quad einen 2x2 Block von vier Pixeln umfasst,

M3 wobei die SIMD-Verarbeitungseinheit umfasst:

eine Gruppe von Verarbeitungsbahnen, die ausgelegt sind, Befehle von Arbeitselementen eines bestimmten Verarbeitungsvorgangs über vier aufeinander folgende Verarbeitungszyklen auf SIMD-Weise auszuführen,

M3a wobei jede der Verarbeitungsbahnen der Gruppe ausgelegt ist, Befehle aller Arbeitselemente für einen jeweiligen Block von Arbeitselementen, der sich auf ein jeweiliges Pixel-Quad bezieht, über die vier aufeinander folgenden Verarbeitungszyklen hinweg auszuführen; und

M4 Logik, die mit der Gruppe von Verarbeitungsbahnen gekoppelt ist, die ausgelegt ist, die Gruppe von Verarbeitungsbahnen zu veranlassen, die Ausführung eines gegebenen Verarbeitungszyklus zu überspringen,

wenn keine gültigen Arbeitselemente für die Ausführung in der Gruppe von Verarbeitungsbahnen in diesem Verarbeitungszyklus eingeplant sind.

Zu den nebengeordneten Patentansprüchen 8 und 9 sowie zu den Unteransprüchen 2 bis 7 wird auf die Akte verwiesen.

II.

Die zulässige Beschwerde konnte keinen Erfolg haben, da der Gegenstand des Patentanspruchs 1 nicht auf einer erfinderischen Tätigkeit beruht und somit nicht patentfähig ist (§ 1 Abs. 1 i. V. m. § 4 Satz 1 PatG).

1. Der Senat ist für die Behandlung der vorliegenden Teilanmeldung zuständig. Durch die Teilungserklärung erhält das Bundespatentgericht die Entscheidungskompetenz über die neue Teilanmeldung, weil deren Gegenstand mit der Beschwerde in der Beschwerdeinstanz angefallen ist (Schulte, PatG, 11. Aufl., § 39 Rdnr. 60 m.w.N.; Busse/Keukenschriyver , PatG, 9. Aufl., §39 Rdnr. 24; BGH GRUR 2019, 766 - *Abstandsberechnungsverfahren*, Leitsatz b)).

Im vorliegenden Fall war das Beschwerdeverfahren gegen die Zurückweisung der Stammanmeldung 10 2014 119 048.3 mit Eingang der Beschwerde am 19. Februar 2020 beim Bundespatentgericht anhängig. Über diese Beschwerde wurde durch Beschluss vom 15. Juni 2021 entschieden. Die Teilungserklärung vom 23. Dezember 2020 ging somit während der Anhängigkeit des Beschwerdeverfahrens beim Bundespatentgericht ein.

2. Ausgehend von der Feststellung des Deutschen Patent- und Markenamts im Schreiben vom 3. März 2021 bestehen an der Wirksamkeit der Teilung keine Zweifel.

3. Die vorliegende Teilanmeldung betrifft die Datenverarbeitung in SIMD-Verarbeitungseinheiten („SIMD“ = „single instruction multiple data“).

In der Beschreibungseinleitung der Teilanmeldung wird ausgeführt, dass SIMD-Verarbeitungseinheiten Datenelemente parallel verarbeiten können und daher besonders nützlich sind, wenn derselbe Befehl an einer großen Anzahl von Datenelementen auszuführen ist. Beispielsweise könne eine Grafikverarbeitungseinheit (GPU) eine SIMD-Verarbeitungseinheit verwenden, um Verarbeitungsvorgänge an einer großen Anzahl von Pixeln eines computergenerierten Bilds durchzuführen (Seite 1, Zeile 6 bis 18 der Beschreibung der Teilanmeldung; alle Seiten- und Zeilenangaben in den Abschnitten **II.3** und **II.4** dieses Beschlusses beziehen sich auf diese Beschreibung).

Ein Verarbeitungsvorgang könne aus einer Mehrzahl von Arbeitselementen bestehen, bei deren Ausführung eine gemeinsame Sequenz von Befehlen an entsprechenden Datenelementen ausgeführt werde. Eine SIMD-Verarbeitungseinheit könne Verarbeitungsbahnen umfassen, die jeweils ausgelegt sind, einen Befehl eines Arbeitselements in mehreren Verarbeitungszyklen auszuführen. Falls ein Verarbeitungsvorgang nur teilweise belegt sei oder ungültige Arbeitselemente umfasse, blieben einige „Verarbeitungsschlitze“ - d.h. Zeitfenster,

in denen jeweils ein Befehl in einer bestimmten Verarbeitungsbahn abgearbeitet werden kann - ungenutzt (Seite 1, Zeile 20 bis Seite 3, Zeile 21 sowie Figuren 1 und 2; der in Figur 1 dargestellte Verarbeitungsvorgang 100 mit 32 Arbeitselementen wird gemäß Figur 2 auf 16 Verarbeitungsbahnen in zwei Verarbeitungszyklen ausgeführt).

Moderne Grafikschnittstellen zur Anwendungsprogrammierung wie OpenGL und DirectX definierten Befehle, die mit Pixeln in einem 2x2-Pixelquad arbeiten. Es sei z.B. häufig erforderlich, die Änderungsrate einer variierenden Menge von verschiedenen Pixeln mittels eines „Gradienten“-Vorgangs zu bestimmen. Die Befehle der Grafikschnittstellen ließen die Entfernung von „leeren“ Pixelverarbeitungsschlitzen (die ungültigen Arbeitselementen entsprechen) nicht zu, wenn Arbeitselemente in Verarbeitungsvorgängen verpackt werden (Seite 3, Zeile 23 bis 29).

Eine **Aufgabe** wird in der Beschreibung der vorliegenden Teilanmeldung nicht ausdrücklich genannt. Jedoch lässt der Text auf Seite 7, Zeile 25 bis 28 aus Sicht des Senats auf die Aufgabenstellung schließen, eine SIMD-Verarbeitungseinheit bereitzustellen, die ausgelegt ist, die Anzahl von Verarbeitungsschlitzen zu verringern, die aufgrund ungültiger Arbeitselemente in Verarbeitungsvorgängen ungenutzt bleiben.

Als **Fachmann**, der mit der Lösung dieser Aufgabe betraut wird, ist ein Informatiker oder Ingenieur der Fachrichtung Elektrotechnik mit mehrjähriger Erfahrung in der Entwicklung paralleler Datenverarbeitungsarchitekturen - insbesondere für Grafikanwendungen - anzusehen.

4. Zur Lehre von Patentanspruch 1

Patentanspruch 1 sieht eine SIMD-Verarbeitungseinheit vor (Merkmal **M1**), d.h. eine Recheneinheit, die es gestattet, denselben Befehl gleichzeitig auf mehrere unterschiedliche Daten anzuwenden (Seite 1, Zeile 6 bis 18).

Die SIMD-Verarbeitungseinheit ist ausgelegt, eine Mehrzahl von Verarbeitungsvorgängen zu verarbeiten, die jeweils Arbeitselemente bis zu einer vorbestimmten Höchstanzahl umfassen (Merkmale **M2**, **M2a**). Ein Arbeitselement kann einen Teilvorgang eines Verarbeitungsvorgangs und insbesondere eine Sequenz von Befehlen umfassen (Seite 1, Zeile 20 bis 27).

Die Arbeitselemente eines Verarbeitungsvorgangs sollen ausgelegt sein, eine „gemeinsame“ Sequenz von Befehlen an entsprechenden Datenelementen auszuführen (Merkmal **M2b**). Dies soll bedeuten, dass die Arbeitselemente einer Gruppe von Arbeitselementen des Verarbeitungsvorgangs dieselbe Sequenz von Befehlen umfassen, die auf entsprechende Datenelemente anzuwenden sind (Seite 1, Zeile 23 bis 27).

Ferner soll ein Verarbeitungsvorgang durch Blöcke von Arbeitselementen gebildet werden, die jeweils genau vier Arbeitselemente umfassen und sich auf ein jeweiliges Pixel-Quad beziehen, wobei jedes Pixel-Quad einen 2x2-Block von vier Pixeln umfasst (Merkmale **M2c**, **M2d**). Ein Block von Arbeitselementen stellt eine zusammenhängende Menge von Arbeitselementen dar, die in einer oder mehreren Dimensionen angeordnet sind (vgl. z.B. die Figuren 1 und 2 der Teilanmeldung).

Die SIMD-Verarbeitungseinheit umfasst ferner eine Gruppe von Verarbeitungsbahnen, die ausgelegt sind, Befehle von Arbeitselementen eines bestimmten Verarbeitungsvorgangs über vier aufeinander folgende Verarbeitungszyklen auf SIMD-Weise auszuführen (Merkmal **M3**). Dass Befehle von Arbeitselementen „auf SIMD-Weise“ ausgeführt werden, bedeutet aus fachmännischer Sicht, dass einzelne Befehle an alle Verarbeitungsbahnen geschickt und dort in Abhängigkeit ihrer „Breite“ und der Anzahl der Verarbeitungsbahnen entweder in einem einzigen Verarbeitungszyklus oder in mehreren aufeinanderfolgenden Verarbeitungszyklen ausgeführt werden (Seite 1, Zeile 6 bis 18; siehe auch Druckschrift **D2**, Abschnitt 2.1, erster und zweiter Satz).

Jede der Verarbeitungsbahnen der Gruppe soll ferner ausgelegt sein, Befehle aller Arbeitselemente für einen jeweiligen Block von Arbeitselementen, der sich auf ein jeweiliges Pixel-Quad bezieht, über die vier aufeinander folgenden Verarbeitungszyklen hinweg auszuführen (Merkmal **M3a**). Damit wird insbesondere zum Ausdruck gebracht, dass die einzelnen Pixelwerte eines Pixelquads nacheinander ausschließlich in einer einzigen Verarbeitungsbahn verarbeitet werden (im Beispiel der Figur 6 sollen die Arbeitselemente 0 bis 3, mit denen offensichtlich Berechnungen an den Pixeln eines ersten Pixelquads vorgenommen werden, ausschließlich in der ganz linken Verarbeitungsbahn ausgeführt werden, die Arbeitselemente 4 bis 7 in der zweiten Verarbeitungsbahn von links usw.). Diese spaltenweise „vertikale“ Anordnung der Blöcke von Arbeitselementen wird in der Teilanmeldung auch als „Spalte-zuerst-Ordnung“ bezeichnet (Seite 14, Zeile 21 bis Seite 15, Zeile 13). Da die Befehle der Arbeitselemente gemäß Merkmal **M3** über vier aufeinanderfolgende Verarbeitungszyklen auf SIMD-Weise ausgeführt werden sollen, implizieren die Merkmale **M3** und **M3a** im Falle einer „Spalte-zuerst-Ordnung“ der Arbeitselemente, dass auf all diejenigen Pixelquads, deren zugehörige Arbeitselemente jeweils in einer von mehreren benachbarten „Spalten“ eines Verarbeitungsvorgangs liegen, dieselben Befehle angewendet werden.

Ferner soll die SIMD-Verarbeitungseinheit eine mit der Gruppe von Verarbeitungsbahnen gekoppelte „Logik“ - also logische Schaltungen - umfassen, die ausgelegt sind, die Gruppe von Verarbeitungsbahnen zu veranlassen, die Ausführung eines gegebenen Verarbeitungszyklus zu überspringen, wenn keine gültigen Arbeitselemente für die Ausführung in der Gruppe von Verarbeitungsbahnen in diesem Verarbeitungszyklus eingeplant sind (Merkmal **M4**). Damit wird zum Ausdruck gebracht, dass ein Verarbeitungszyklus übergangen werden kann, in dem ausschließlich ungültige Arbeitselemente ausgeführt werden sollen. Dadurch verringert sich die Anzahl ungenutzter „Verarbeitungsschlitze“ (Seite 20, Zeile 4 bis 14), d.h. die Anzahl von Zeitfenstern, in denen keine Befehle zur Ausführung vorgesehen sind.

Ein Arbeitselement ist dann ungültig, wenn es sich auf ein ungültiges Datenelement bzw. einen ungültigen Pixelwert bezieht (Seite 4, Zeile 14 bis 15). Ein Pixelwert eines Pixelquads wiederum kann ungültig sein, wenn an dem Pixelwert keine oder nur sehr wenige Befehle auszuführen sind (Seite 10, Zeile 10 bis 16; Seite 24, Zeile 12 bis 23). Aus Sicht des Fachmanns stellen auch maskierte oder ausgeblendete Arbeitselemente, deren Ausführungsergebnisse nicht weiterverwendet werden, ungültige Arbeitselemente im Sinne von Merkmal **M4** dar (Seite 28, Zeile 4 bis 15).

5. Der Patentanspruch 1 ist so klar und deutlich gefasst, dass sein Schutzbereich hinreichend sicher vorhersehbar ist. Außerdem ist die damit beanspruchte Lehre in der Anmeldung so deutlich und vollständig offenbart, dass ein Fachmann sie ausführen kann.

6. Der Patentanspruch 1 ist nicht gewährbar, da sein Gegenstand nicht auf einer erfinderischen Tätigkeit beruht.

6.1 Für die Beurteilung der beanspruchten Lehre sind die Druckschriften **G9** und **D2** von besonderer Bedeutung.

6.1.1 Die Druckschrift **G9** ist ein Teil einer einführenden Textreihe zur parallelen Datenverarbeitung in einer Grafikpipeline (Seite 1, erster Absatz sowie zweiter Absatz, erster Satz), in dem die Parallelisierung von Rasterungs- und Shadingoperationen (vgl. Abschnitte „Going wide during rasterization“, „You need to go wider!“), die Interpolation von Pixelattributen (vgl. Abschnitte „Attribute interpolation“, „“Centroid“ interpolation is tricky“) und weitere Teilaspekte des Shadings (vgl. Abschnitt „The actual shader body“) diskutiert werden.

Dieser Druckschrift ist insbesondere zu entnehmen, dass bei der Grafikverarbeitungsarchitektur „Fermi“ der Firma NVidia mehrere Warps, die aus maximal 32 Threads bestehen (und somit im Sinne der Merkmale **M2** und **M2a** als

eine Mehrzahl von Verarbeitungsvorgängen angesehen werden können, die jeweils eine Höchstanzahl von 32 Arbeitselementen umfassen), an die Shader-Einheiten einer GPU geschickt werden (Abschnitt „You need to go wider!“ - „for NVidia, the unit of dispatch to shader units is 32 threads, which they call a “Warp“.“). Für den Fachmann ist selbstverständlich, dass die Anzahl der Threads pro Warp ein bei der Definition einer Grafikverarbeitungsarchitektur frei wählbarer Designparameter ist, so dass die 32 Threads z.B. auch auf 2, 4 oder 8 einzelne Warps verteilt werden können.

Die Pixelwerte, die bei der Ausführung der Warps verarbeitet werden, sind in Pixelquads angeordnet, d.h. in 2x2-Blöcken (vgl. Figur auf Seite 3), wobei jedes Pixel als ein Thread behandelt werden kann (Abschnitt „You need to go wider!“, erster Absatz - „Each quad has 4 pixels (each of which in turn can be handled as one thread)“). Alle 32 Threads eines Warps führen auf den Pixelwerten von jeweils acht aufeinanderfolgenden Pixelquads gleichzeitig dieselben Befehle aus (vgl. Seite 5, letzter Absatz - „work on all elements of each batch usually proceeds in lockstep. All “threads“ run the same code, at the same time“ i. V. m. Seite 5, vorletzter Absatz- „multiple batches (or „Warps“ on NVidia hardware […])“ sowie Abschnitt „You need to go wider!“, erster Absatz - „we need to grab 8 incoming quads […] before we can send off a batch“; vgl. Merkmal **M2b**). Daher stellt die in Druckschrift **G9** beschriebene GPU eine SIMD-Verarbeitungseinheit dar (Merkmal **M1**), die ausgelegt ist, eine Mehrzahl von Verarbeitungsvorgängen entsprechend den Merkmalen **M2** bis **M2b** zu verarbeiten. Da ein Warp aus einer durch 4 teilbaren Anzahl von Threads besteht und sich somit in Viererblöcke von Arbeitselementen einteilen lässt, liegt auch Merkmal **M2c** vor.

Die Druckschrift **G9** zeigt jedoch nicht ausdrücklich, dass sich die Arbeitselemente eines solchen Viererblocks gemäß Merkmal **M2d** auf ein jeweiliges Pixelquad beziehen, und offenbart auch keines der Merkmale **M3**, **M3a** und **M4**, die Einzelheiten der Verarbeitung der Viererblöcke in einer Gruppe von Verarbeitungsbahnen betreffen.

6.1.2 Die Druckschrift **D2** befasst sich mit Methoden zur Maximierung der Ressourcenauslastung in Grafikverarbeitungseinheiten, die nach dem SIMD-Prinzip arbeiten und aus mehreren Streaming-Multiprozessoren mit parallelen Verarbeitungsbahnen („execution lanes“, „SIMD lanes“) bestehen. In diesen Bahnen werden Threads (d.h. parallelisierbare Befehlsfolgen) ausgeführt, die ein Programmierer zu sog. kooperativen Threadfeldern („CTAs“ = „cooperative thread arrays“) gruppiert hat, die ihrerseits aus mehreren „Warps“ bestehen - d.h. aus kleineren Threadgruppen, die denselben Code umfassen (vgl. Druckschriftentitel i. V. m. Abschnitt 2.1). Die Anzahl gleichzeitig ausführbarer Threads (die „$SIMD_{width}$“, vgl. Abschnitt 3.3, Gleichung (1) mit darunterstehendem Text - „$SIMD_{width}$ designates the width of the SIMD pipeline“) hängt dabei von der verwendeten Grafikarchitektur ab; in den Figuren 1, 2, 5 und 7 sind beispielsweise vier Verarbeitungsbahnen gezeigt.

Damit sind in der Lehre der Druckschrift **D2** bereits die Merkmale **M1, M2, M2a** und **M2b** verwirklicht. Denn ein nach dem SIMD-Prinzip arbeitender Streaming-Multiprozessor ist eine SIMD-Verarbeitungseinheit im Sinne von Merkmal **M1**. Die Verarbeitungsbahnen eines solchen Multiprozessors bilden eine Gruppe und sind ausgelegt, die Befehle der Threads mehrerer CTAs - d.h. die Befehle der Arbeitselemente mehrerer Verarbeitungsvorgänge - auszuführen (Merkmal **M2**). Die Anzahl von Threads pro CTA ist ein Anwendungsparameter und stellt ebenso wie die Anzahl von Threads pro Warp eine Höchstanzahl von Arbeitselementen dar (vgl. Abschnitt 2.1, vorletzter Satz; Merkmal **M2a**). Da ein Streaming-Multiprozessor eine SIMD-Verarbeitungseinheit ist, wird ein CTA über mehrere Verarbeitungsbahnen hinweg in mehreren Verarbeitungszyklen ausgeführt, so dass jede Verarbeitungsbahn eine Sequenz von Befehlen (selbstverständlich an entsprechenden Datenelementen) ausführt, die sie mit den anderen Bahnen gemeinsam hat (Merkmal **M2b**).

Ferner ist den einleitenden Abschnitten der Druckschrift **D2** zu entnehmen, dass ein sogenanntes SIMT-Ausführungsmodell („SIMT“ = „single-instruction multiple-

thread“) es ermöglicht, in Grafikprozessoren („GPUs“) effiziente SIMD-Pipelines zu verwenden und gleichzeitig in den ausgeführten Threads beliebige Kontrollflüsse vorzusehen. Die GPU-Hardware würde auch bedingte Verzweigungen unterstützen, so dass jede SIMD-Verarbeitungsbahn ihren eigenen logischen Thread ausführen könne (vgl. Abschnitt 1, erster Absatz sowie Abschnitt 2.1 und Abschnitt 2.2, erster Satz).

Die unabhängige Verarbeitung von Verzweigungen würde durch hardwaregenerierte Bitmasken ermöglicht, welche kennzeichneten, ob ein Thread aktiv sei oder nicht. Gemäß dem SIMT-Ausführungsmodell würde die Ausführung einer divergenten Verzweigung teilweise serialisiert, da der „wahre“ und der „falsche“ Pfad nacheinander ausgeführt und die Threads auf dem jeweils nicht aktiven Pfad maskiert werden müssten. Bei jeder Divergenz würden Threads maskiert und nicht ausgeführt, so dass sich die Anzahl aktiver Threads und damit die Auslastung der SIMD-Verarbeitungsbahnen weiter reduziere und die Anzahl „vergeudeter Verarbeitungsschlitze“ („wasted execution slots“) erhöhe (vgl. Abschnitt 2.2 i. V. m. Figur 1; Abschnitt 2.3, erster Satz; s. auch Abstract und Abschnitt 1, erster Absatz).

Um die Auslastung der SIMD-Verarbeitungsbahnen zu verbessern, beschreibt die Druckschrift **D2** u.a. die Methode der Threadblock-Verdichtung („Thread block compaction“ = „TBC“; vgl. Abschnitt 2.3 sowie Figuren 1 und 2). Dabei werden Threads innerhalb eines CTA unter Beibehaltung ihrer Verarbeitungsbahn (ihrer „home lane“) umgeordnet (vgl. Abschnitt 2.3 i. V. m. Figuren 1 und 2 - die Warps „- - - 7“ und „- - A - “ des CTA B und die Warps „4 5 6 -“ und „8 9 - B“ des CTA C werden umgeordnet, so dass die Warps „- - A 7“ und „- - - -“ bzw. „4 5 6 B“ und „8 9 - -“ gebildet werden). Falls nach einer solchen Verdichtung ein Warp entsteht, der ausschließlich maskierte Threads - d.h. ungültige Arbeitselemente im Sinne von Merkmal **M4** - enthält, erhöht sich die Auslastung der SIMD-Verarbeitungsbahnen, da der zur Ausführung dieses Warps vorgesehene Verarbeitungszyklus übergangen wird (vgl. Figur 1 - ohne Verdichtung werden die Warps „- - - 7“ und „-

- A -“ des CTA B in zwei Verarbeitungszyklen ausgeführt, nach der Verdichtung genügt stattdessen die Ausführung des verdichteten Warps „- - A 7“ in einem einzigen Verarbeitungszyklus). Der Druckschrift **D2** ist ebenfalls zu entnehmen, dass Warps übergangen werden, die bereits vor einer Threadblock-Verdichtung ausschließlich aus ungültigen Arbeitselementen bestehen; so ist der in Figur 1 (a) gezeigte Warp „- - - -“ des CTA B bereits in dem in Figur 1 (b) gezeigten Ausführungsablauf nicht mehr vorhanden.

Gemäß Druckschrift **D2** können die Threads vor der Verdichtung im Rahmen einer „SIMD lane permutation (SLP)“ über die Verarbeitungsbahnen hinweg permutiert werden (vgl. Abschnitte 4.1 und 4.3 i. V. mit den Figuren 7 bis 9 sowie Abschnitt 4.4 - „Implementation of SLP […] Enabling SLP on top of TBC […]“). Durch diese zusätzliche Maßnahme können in vielen Fällen noch mehr Verarbeitungszyklen übergangen werden, so dass eine noch höhere Auslastung der Verarbeitungsbahnen erzielt wird (vgl. Figur 14).

Es ist selbstverständlich, dass das Übergehen der Warps durch logische Schaltungen der Streaming-Multiprozessoren gesteuert werden muss, die mit den einzelnen Verarbeitungsbahnen gekoppelt sind. Daher verwirklicht die Lehre der Druckschrift **D2** auch das Merkmal **M4**.

Die in Figur 7 (a) im Zusammenhang mit einer Threadblock-Verdichtung („TBC“) gezeigten quadratischen Blöcke mit 16 Threads entsprechen jeweils einem CTA, dessen Befehle in vier Verarbeitungsbahnen in vier Verarbeitungszyklen abgearbeitet werden. Ein solcher CTA besteht aus vier „Spalten“, die jeweils genau vier Threads enthalten und als „Blöcke von Arbeitselementen“ im Sinne von Patentanspruch 1 angesehen werden können (z.B. umfasst die ganz linke Spalte des CTA, der dem links in Figur 7 (a) gezeigten quadratischen Block entspricht, die Threads „0 4 8 C“; die zweite Spalte von links die Threads „- - - -“ usw.; Merkmal **2c**).

Die vier Verarbeitungsbahnen, die gemäß dem in Figur 7 (a) gezeigten Beispiel verwendet werden, sind dazu ausgelegt, die Befehle der 16 Threads über vier aufeinanderfolgende Verarbeitungszyklen auf SIMD-Weise auszuführen, indem einzelne Befehle an alle vier Verarbeitungsbahnen geschickt und dort ausgeführt werden (vgl. Abschnitt 2.1 - „streaming multiprocessors (SM) […] each SM contains a number of parallel execution lanes […] that operate in SIMD fashion"; Merkmal **M3**).

Jede einzelne dieser vier Verarbeitungsbahnen ist selbstverständlich ausgelegt, die Befehle aller 16 Threads - und damit die Befehle der vier Thread-„Spalten" - über die vier Verarbeitungszyklen hinweg auszuführen (**Teilmerkmal** von Merkmal **M3a**).

Die Druckschrift **D2** befasst sich allerdings nicht speziell mit der Verarbeitung von Pixelquads und offenbart daher nicht ausdrücklich, dass sich vier Arbeitselemente einer „Spalte" von Arbeitselementen jeweils auf ein Pixel-Quad beziehen, das einen 2x2-Block von vier Pixeln umfasst (Merkmal **M2d** sowie **verbleibendes Teilmerkmal** „[…] Block von Arbeitselementen, der sich auf ein jeweiliges Pixel-Quad bezieht" von Merkmal **M3a**).

6.2 Der Gegenstand von Patentanspruch 1 ist durch den aufgezeigten Stand der Technik nahegelegt.

In Druckschrift **G9** ist beschrieben, dass einige der Pixel eines Pixelquads, die an die Shader-Einheiten der GPU geschickt werden, unsichtbar sind, so dass an ihnen kein Shading durchgeführt werden muss (Seite 2, letzter Absatz und Seite 3, erster Absatz i. V. m. der Figur auf Seite 3 - die Werte der hellgrau hinterlegten unsichtbaren Pixel werden weder für ein Shading noch für Gradientenberechnungen verwendet, die Werte der mittelgrau hinterlegten, ebenfalls unsichtbaren Helferpixel („helper pixels") nur für Gradientenberechnungen, und die Werte der dunkelgrau hinterlegten sichtbaren Pixel sowohl für ein Shading als auch für Gradientenberechnungen). Die

unsichtbaren Pixel sind zwar maskiert, jedoch wird auch an ihnen zwangsläufig ein Shading durchgeführt (vgl. Seite 2, zweiter vollständiger Absatz - „all pixels in a quad, even the masked ones, get shaded“).

Die Werte der unsichtbaren Pixel stellen „ungültige“ Pixelwerte im Sinne des Patentanspruchs 1 dar, da an ihnen selbstverständlich weniger Befehle als an den Werten der sichtbaren Pixel auszuführen sind. Dementsprechend sind die maskierten Threads, die der Verarbeitung der Werte der unsichtbaren Pixel dienen, keine gültigen Arbeitselemente im Sinne von Patentanspruch 1.

Gemäß Druckschrift **G9** werden 25%-75% der für das Shading eines Pixelquads aufgewendeten Arbeit vergeudet, wenn viele kleine Dreiecke zu rendern sind, da auch für die unsichtbaren Helferpixel ein Shading durchgeführt wird (vgl. Seite 2 - „between 25-75% of the shading work for quads generated for triangle edges is wasted“; Seite 3 - „wasted shading work on quads“; je nachdem ob ein Pixelquad ein, zwei oder drei Helferpixel besitzt, sind 25%, 50% oder 75% der Shading-Arbeit überflüssig). Zum Zeitpunkt der Abfassung der Druckschrift **G9** erlaubten weder Programmierschnittstellen noch die Hardware eine Vereinigung von Pixelquads (Seite 3, erster Absatz, vorletzter Satz).

Der Fachmann hatte somit Veranlassung, im Stand der Technik nach Verfahren zu suchen, die es erlauben, ein Pixel-Shading zu beschleunigen, wenn eine große Zahl von Pixeln - und damit auch eine große Zahl der Threads, die zu diesen Pixeln gehören - ungültig ist.

Hierbei konnte er auf die Druckschrift **D2** stoßen, die aufzeigt, wie Warps beschleunigt verarbeitet werden können, die ungültige Threads aufweisen (s.o., Abschnitt **II.6.1.2**).

Es bot sich daher für den Fachmann an, die in Druckschrift **G9** beschriebenen Pixelquads gemäß der in der Druckschrift **D2** beschriebenen Lehre zu verarbeiten,

d.h. Streaming-Multiprozessoren mit insbesondere vier parallelen Verarbeitungsbahnen vorzusehen, jeweils 16 Threads auf einen CTA mit vier Warps zu verteilen und gemäß der Methode der Threadblock-Verdichtung mit der Zielsetzung zu komprimieren, dass für einen Verarbeitungszyklus ausschließlich ungültige Arbeitselemente vorgesehen sind, deren Ausführung übergangen werden kann.

Von dieser kombinierten Lehre unterscheidet sich der Gegenstand von Patentanspruch 1 nur durch das Merkmal **M2d** und das **verbleibende Teilmerkmal** von Merkmal **M3a**.

Die Druckschriften **D2** und **G9** überlassen es dem Fachmann, die erforderliche Anfangsverteilung der einzelnen Threads auf den CTA - und damit auch die Anfangsverteilung der Pixelwerte der Pixelquads auf die Verarbeitungsbahnen - festzulegen.

In diesem Zusammenhang war dem Fachmann hinlänglich bekannt, dass sich die Leistungsfähigkeit von Grafikprogrammen, die ein Pixel-Shading durchführen, erhöhen lässt, indem Grafikdaten mit „vertikalen“ Methoden verarbeitet werden. Diese Methoden sind dadurch charakterisiert, dass Teilabschnitte der Grafikdaten - also gerade auch Pixelquads - jeweils in einem einzigen von mehreren parallelen „vertikalen“ Kanälen verarbeitet werden (vgl. **D11**, Absatz [0044] i. V. m. Absatz [0022] und [0023], jeweils vorletzter Satz).

Daher lag es für den Fachmann auf der Hand, die Threads so auf den CTA zu verteilen, dass jeweils vier auf dasselbe Pixelquad bezogene Threads „vertikal“ in einer einzigen Spalte des CTA angeordnet sind, so dass sie einen Viererblock von Arbeitselementen bilden, bei dessen Ausführung die Pixelwerte des Pixelquads nach einer Threadblock-Verdichtung in einer einzigen Verarbeitungsbahn - je nach Anzahl ungültiger Arbeitselemente in dem CTA in bis zu vier aufeinanderfolgenden Verarbeitungszyklen - verarbeitet werden (vgl. Abschnitt **II.6.1.2**).

Auf diese Weise konnte der Fachmann somit auch zum Merkmal **M2d** sowie zum **verbleibenden Teilmerkmal** von Merkmal **M3a** gelangen, ohne erfinderisch tätig zu werden.

6.3 Die Anmelderin wendet sinngemäß ein, der Satz “This gives you a very cheap way to get derivatives at the cost of always having to shade groups of 2x2 pixels at once” in Druckschrift **G9** lenke den Fachmann gerade von einer vertikalen Anordnung der auf die Pixel eines 2x2-Pixelquads bezogenen Arbeitselemente weg in Richtung einer horizontalen Anordnung. Zudem würden in Druckschrift **D2** die Threads über die Verarbeitungsbahnen hinweg permutiert und damit gerade nicht in derselben Verarbeitungsbahn ausgeführt, wie es hingegen gemäß Patentanspruch 1 der Fall sei. Außerdem beträfen die in der Druckschrift **D11** beschriebenen „vertikalen“ Verfahren im Gegensatz zu Merkmal **M3** keine Verarbeitung „auf SIMD-Weise“, was insbesondere aus den Absätzen [0021] und [0028] der Druckschrift **D11** hervorgehe.

Diese Argumente überzeugen allerdings nicht.

So differenziert der von der Anmelderin zitierte Satz in Druckschrift **G9** nicht danach, ob Gruppen von jeweils vier Threads, die sich auf die Pixel eines 2x2-Quads beziehen, vertikal oder horizontal verarbeitet werden. Denn auch wenn diese Gruppen jeweils in einzelnen vertikalen Spalten eines Verarbeitungsvorgangs liegen, werden sie zeitgleich (“at once”) über insgesamt vier Verarbeitungszyklen hinweg verarbeitet.

Weiterhin trifft es zwar zu, dass Threads, die zunächst in einer Spalte vertikal angeordnet sind, nicht von derselben Verarbeitungsbahn abgearbeitet werden, wenn sie über die Verarbeitungsbahnen hinweg “horizontal” permutiert worden sind. Die aus der Druckschrift **D2** bekannte Threadblock-Verdichtung erfordert jedoch keine solchen Permutationen, so dass Threads, die gemäß einer bestimmten

Anfangsverteilung in einer Spalte angeordnet worden sind, aufgrund der Verdichtung nur innerhalb dieser Spalte - also nur “vertikal” - umgeordnet werden.

Da der Druckschrift **D11** die allgemeine Lehre zu entnehmen ist, dass bei einer „vertikalen“ Verarbeitung beliebige Befehle auf beliebige Daten angewendet werden können (vgl. die oben in Abschnitt **6.2** zitierten Textstellen der Druckschrift **D11** und ferner auch die von der Anmelderin genannten Absätze [0021] („SIMD or other instructions“) und [0028] („data (e.g., SIMD data) is pre-organized […] when instructions and associated data are such that demand processing in the vertical mode of operation […]“)), können Befehle, die in demselben Kanal „vertikal“ verarbeitet werden, auch gemäß Druckschrift **D11** zu verschiedenen SIMD-Befehlen gehören, die nacheinander ausgeführt werden. Die Druckschrift **D11** führt den Fachmann also nicht notwendigerweise auf eine Art der SIMD-Instruktionsverarbeitung, die sich von der in der Druckschrift **D2** beschriebenen Herangehensweise unterscheidet.

Da auch nach einer Threadblock-Verdichtung über die vier Spalten des CTA hinweg in jedem Verarbeitungszyklus auf den Pixelwerten vier aufeinanderfolgender „vertikal“ verarbeiteter Pixelquads dieselben Befehle ausgeführt werden (vgl. Druckschrift **G9**, Seite 5, letzter Absatz), können die Streaming-Multiprozessoren diese Befehle auch wie in Druckschrift **D2** beschrieben (vgl. Abschnitt **II.6.1.2**, drittletzter Absatz) „auf SIMD-Weise“ ausführen.

Im Übrigen geht es dem Fachmann im vorliegenden Fall auch nicht darum, die SIMD-Arbeitsweise der Streaming-Multiprozessoren der Druckschrift **D2** in technischer Hinsicht abzuändern, sondern es ist ihm vielmehr nur an einer zweckmäßigen Anfangsverteilung der Threads innerhalb des CTA gelegen. Die Druckschrift **D11** belegt in diesem Zusammenhang lediglich das fachmännische Wissen, dass es bei einem Pixel-Shading vorteilhaft sein kann, zusammenhängende Datenabschnitte (hier: die Pixelwerte eines Pixelquads) in einer einzigen Verarbeitungsbahn zu verarbeiten, da sich dadurch die

Leistungsfähigkeit einer Grafikpipeline erhöht und bei bestimmten Rechenoperationen (etwa bei Gradientenberechnungen, vgl. Absatz [0043] der Druckschrift **D11**) die Pixelwerte eines Pixelquads nicht zwischen verschiedenen Verarbeitungsbahnen ausgetauscht werden müssen.

7. Da über einen Antrag jeweils nur einheitlich entschieden werden kann, fallen mit dem Patentanspruch 1 auch die übrigen Patentansprüche 2 bis 9 (BGH GRUR 1997, 120 - Elektrisches Speicherheizgerät).

8. Eine Zurückverweisung der Sache an das Deutsche Patent- und Markenamt zur Prüfung der Teilanmeldung, wie von der Anmelderin hilfsweise beantragt, kam aus verfahrensökonomischen Gründen nicht in Betracht.

Das Bundespatentgericht hat zwar nach § 79 Abs. 3 PatG die Möglichkeit, die Sache nach seinem pflichtgemäßen Ermessen an Deutsche Patent- und Markenamt zurückzuverweisen. Dem Senat war allerdings eine eigene Entscheidung über die Teilanmeldung möglich, weil die insoweit relevanten Fragen überwiegend bereits im Verfahren über die Beschwerde gegen die Zurückweisung der Stammanmeldung hinreichend aufbereitet waren und im Übrigen mit vertretbarem Aufwand geklärt werden konnten (vgl. BGH, *Abstandsberechnungsverfahren*, a.a.O., Rdnr. 11 und Rdnr. 13). Die Sache war somit entscheidungsreif und eine Zurückverweisung nicht veranlasst.

Rechtsmittelbelehrung

Gegen diesen Beschluss steht den am Beschwerdeverfahren Beteiligten das Rechtsmittel der Rechtsbeschwerde zu. Da der Senat die Rechtsbeschwerde nicht zugelassen hat, ist sie nur statthaft, wenn gerügt wird, dass

das beschließende Gericht nicht vorschriftsmäßig besetzt war,

1. bei dem Beschluss ein Richter mitgewirkt hat, der von der Ausübung des Richteramtes kraft Gesetzes ausgeschlossen oder wegen Besorgnis der Befangenheit mit Erfolg abgelehnt war,
2. einem Beteiligten das rechtliche Gehör versagt war,
3. ein Beteiligter im Verfahren nicht nach Vorschrift des Gesetzes vertreten war, sofern er nicht der Führung des Verfahrens ausdrücklich oder stillschweigend zugestimmt hat,
4. der Beschluss aufgrund einer mündlichen Verhandlung ergangen ist, bei der die Vorschriften über die Öffentlichkeit des Verfahrens verletzt worden sind, oder
5. der Beschluss nicht mit Gründen versehen ist.

Die Rechtsbeschwerde ist innerhalb eines Monats nach Zustellung des Beschlusses beim Bundesgerichtshof, Herrenstr. 45 a, 76133 Karlsruhe, durch einen beim Bundesgerichtshof zugelassenen Rechtsanwalt als Bevollmächtigten einzulegen.

Dr. Morawek Dr. Forkel Akintche Dr. Städele

Fi