Internal distribution code:

(A) [-] Publication in OJ
(B) [-] To Chairmen and Members
(C) [-] To Chairmen
(D) [X] No distribution

Datasheet for the decision of 7 August 2024

Case Number: T 0046/22 - 3.5.06

Application Number: 19218493.5

Publication Number: 3654185

IPC: G06F9/50, G06F8/41, G06T1/20, G06F9/455, G06N3/04, G06N3/08, G06N3/063

Language of the proceedings: EN

Title of invention:
COMPUTE OPTIMIZATION MECHANISM FOR DEEP NEURAL NETWORKS

Applicant:
Intel Corporation

Headword:
EU types/INTEL

Relevant legal provisions:
EPC Art. 56

Keyword:
Inventive step - (no)

Decisions cited:

Catchword:

Europäisches Patentamt
European Patent Office
Office européen des brevets

Beschwerdekammern

Boards of Appeal

Chambres de recours

Boards of Appeal of the
European Patent Office
Richard-Reitzner-Allee 8
85540 Haar
GERMANY
Tel. +49 (0)89 2399-0

Case Number: T 0046/22 - 3.5.06

D E C I S I O N
of Technical Board of Appeal 3.5.06
of 7 August 2024

Appellant: (Applicant)	Intel Corporation 2200 Mission College Boulevard Santa Clara, CA 95054 (US)
Representative:	Samson & Partner Patentanwälte mbB Widenmayerstraße 6 80538 München (DE)
Decision under appeal:	**Decision of the Examining Division of the European Patent Office posted on 5 August 2021 refusing European patent application No. 19218493.5 pursuant to Article 97(2) EPC.**

Composition of the Board:

Chairman	M. Müller
Members:	G. Zucka
	B. Müller

Summary of Facts and Submissions

I. The appeal is against the decision by the examining division, dispatched with reasons on 5 August 2021, to refuse European patent application 19218493.5, on the basis that the subject-matter of claim 1 of both requests was not novel, Article 54 EPC, in view of the following document:

D1: US 2011/072243 A1.

II. The following documents are introduced by the board:

D3: S. Chetlur *et al.*: "cuDNN: Efficient Primitives for Deep Learning", 18 December 2014 (3rd version), available on the Internet at https://arxiv.org/pdf/1410.0759v3.pdf, retrieved by the board on 29 February 2024;

D4: M. Harris: "Maxwell: The Most Advanced CUDA GPU Ever Made", 18 September 2014, available on the Internet at https://developer.nvidia.com/blog/maxwell-most-advanced-cuda-gpu-ever-made/, retrieved by the board on 5 March 2024;

D7: Whitepaper "NVIDIA Tesla P100", 2016, available on the Internet at https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf, retrieved by the board on 19 March 2024.

III. A notice of appeal was received on 11 August 2021, the appeal fee being paid on the same day. A statement of grounds of appeal was received on 15 December 2021.

IV. The appellant requested that the decision under appeal be set aside and a patent granted on the basis of either request that was the object of the appealed decision. The appellant made a conditional request for oral proceedings.

V. The board issued a summons to oral proceedings. In an annex to the summons, the board set out its preliminary opinion according to which the appealed decision should be upheld.

VI. On 5 July 2024, the appellant filed a response to the summons and maintained its requests.

VII. The appellant requests that the decision under appeal be set aside and a patent be granted on the basis of claims 1 to 4 of the main request or claims 1 to 3 of the auxiliary request that were the object of the refusal, both filed on 12 November 2020.

The further text on file is:

description pages
2 to 100 as originally filed,
1 and 1a received on 17 June 2020;

drawing sheets
1 to 39 as originally filed.

VIII. Claim 1 of the main request reads as follows:

"A graphics processing unit (214, 410-413) comprising one or more multiprocessors (234, 325),
characterized in that at least one of the one or more multiprocessors (234, 325) including:

a register file (258, 334A, 334B, 1606) to store operands;

a first processing unit (700(a)) including a plurality of execution units (705(a)-705(n)) of a first type,

the first type of execution units being floating point units to process, using the register file, deep learning matrix math operations on a first set of operands;

and

a second processing unit (700(b)), being different from the first processing unit (700(a)), the second processing unit (700(b)) including a plurality of execution units of a second type;

the second type of execution units (706(a)-706(n)) comprising integer arithmetic logic units to perform general-purpose graphics processing unit operations on a second set of operands."

IX. Claim 1 of the auxiliary request distinguishes itself from that of the main request in that the second type of execution units comprises not only integer arithmetic logic units but also floating point units to perform general-purpose graphics processing unit operations on a second set of operands.

X. At the end of the oral proceedings, the chairman announced the board's decision.

Reasons for the Decision

1. *The invention*

The application is in the area of graphics processing units comprising at least one multiprocessor (first line of claim 1).

According to the statement of grounds of appeal (page 16, second and third last paragraphs), the aim of the application is to increase processing efficiency, more specifically to process matrix and graphics operations more efficiently.

This is said to be achieved (*ibid.*, second and fourth paragraph) by having execution units (EUs) of a first type, viz. floating point units, to process deep learning matrix operations and EUs of a second type, viz. integer arithmetic logic units, to perform general-purpose graphics processing unit operations. Each EU may thus be assigned a task for which it is most suitable.

2. *Inventive step; Article 56 EPC*

2.1 To begin with, the board points out that a claim to a "graphics processing unit" (GPU) refers to an electronic circuit that is, at least initially, designed to process (computer) graphics. It cannot be denied that GPUs are currently used to perform a variety of other tasks, including deep learning applications. However, the skilled person understands that such applications use the capabilities of GPUs for performing graphics-related tasks to implement that other application. The skilled person will consequently

understand that the relevant parts of the GPU should be *suitable* for performing those other, non-graphics related tasks.

In the present case, claim 1 is directed to a GPU, not to some other kind of specialised circuit and not to a non-graphics related arrangement *comprising* a GPU. A skilled person will therefore not read the claim as comprising elements designed specifically to execute non-graphics related tasks such as deep-learning matrix math operations, given that the claim does not explicitly say so, but merely as a device comprising elements which should be able to carry out such task, according to the intention expressed in the description. The board acknowledges that some GPUs may be designed from the start with the intention to carry out non-graphics related tasks and may even be optimised for this purpose. The present claim is however not limited to such a specialised processor, but is generally directed to a "graphics processing unit".

The board considers that it is in that sense that the word "to" in the claim should be understood. In fact, no GPU known from the prior art or disclosed in the application is specially adapted for "deep learning" operations. Instead, as implied by par. [00181] of the description, the GPU should be part of a deep neural network and it is this neural network that will perform the deep learning.

2.2 The document D3 discloses a deep neural network (see section 1 "Introduction", first paragraph) using a GPU, more specifically the GPU used in an NVIDIA Tesla K40 or a Geforce GTX 980 (see section 3.2 "Performance").

2.3 As is visible in document D4 (under the heading "SMM: The Maxwell Multiprocessor"), the Geforce GTX 980 GPU, which is a GM204 Maxwell architecture GPU, comprises 8 multiprocessors (SMs).

2.4 Figure 1 of D4 shows that each multiprocessor includes a register file to store operands and a first processing unit including execution units of a first type, this first type being floating point units (FP32 CUDA cores).

The appellant submitted in its response (point 46) that D4 does not state that the CUDA cores are floating point units. However, the table in D4 does refer to "GFLOPs" (billions of floating-point operations per second), which necessarily implies that the cores carry out floating point operations. This fact is corroborated by D7, page 11, Table 1, which shows that each SM of a Maxwell GPU has 128 FP32 CUDA Cores, i.e. 32 FP32 CUDA Cores per processing block, and 4 FP64 CUDA Cores.

2.5 In D3, said floating point units are used to process deep learning (given the context of D3) matrix math operations on a first set of operands (see section 3.1 "Our approach").

2.6 Although the CUDA Cores of the Geforce GTX 980 GPU provide integer arithmetic (as they contain both an Integer ALU and a Floating-Point ALU), this GPU arguably does not contain separate first and second processing units containing respectively floating point units to process deep learning matrix math operations and integer arithmetic logic units to perform general-purpose GPU operations.

2.7 The advantage of this feature distinguishing the subject-matter of claim 1 from the device of D3 is that it renders the GPU more effective in its role of graphics processing.

2.8 It is however considered natural that a skilled person would at some moment envisage rendering a GPU more effective for the purpose for which it was originally designed, viz. the processing of graphics.

2.9 One obvious way of achieving this would be to have part of the GPU made up of a second type of execution units specialised in the performance of general-purpose graphics processing unit operations. Since those operations are frequently based on integer arithmetic (in particular colour and pixel values and geometric representations are typically represented as integers), the skilled person would envisage providing integer arithmetic logic units for this purpose.

2.10 The skilled person would thus arrive at the subject-matter of claim 1 of the main request without the need for an inventive step.

2.11 The appellant pointed out during the oral proceedings before the board that the use of a single register file by both types of execution units would not be obvious. The claim, however, only mentions a use of the register file by the first type.

2.12 The board therefore concludes that the main request does not satisfy the requirements of Article 56 EPC.

2.13 Regarding claim 1 of the auxiliary request, the board holds it to be a logical further step for the skilled person to wish to deal with graphics operations in

general, i.e also those operations which are based on floating point arithmetic. For that purpose, he or she would envisage the provision of both integer and floating point arithmetic logic units. (It is noted *obiter* that this would be a straightforward process, given that for instance the CUDA Cores of an NVIDIA Maxwell GPU already contain both units.)

2.14 The skilled person would thus arrive at the subject-matter of claim 1 of the auxiliary request without the need for an inventive step. The board concludes that the auxiliary request also does not satisfy the requirements of Article 56 EPC.

2.15 The other issues raised in the summons were not discussed during the oral proceedings, as they are not relevant to the present decision.

Order

For these reasons it is decided that:

The appeal is dismissed.

The Registrar: The Chairman:

L. Stridde Martin Müller

Decision electronically authenticated