Internal distribution code:

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

Datasheet for the decision of 5 August 2024

Case Number: T 1710/23 - 3.5.06

Application Number: 18163807.3

Publication Number: 3396546

IPC: G06F9/50, G06F8/41, G06T1/20

Language of the proceedings: EN

Title of invention:
COMPUTE OPTIMIZATION MECHANISM FOR DEEP NEURAL NETWORKS

Applicant:
INTEL Corporation

Headword:
Deep learning GPU/INTEL

Relevant legal provisions:
EPC Art. 84

Keyword:
Claims - clarity (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 1710/23 - 3.5.06

D E C I S I O N
of Technical Board of Appeal 3.5.06
of 5 August 2024

Appellant: (Applicant)	INTEL Corporation 2200 Mission College Blvd. 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 18 April 2023 refusing European patent application No. 18163807.3 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 18 April 2023, to refuse European patent application 18163807.3, on grounds of added subject-matter, lack of inventive step, and lack of clarity. The following document was cited as closest prior art in the appealed decision:

D3: US 2009/265528 A1.

II. A notice of appeal was received on 14 June 2023, the appeal fee being paid on the same day. A statement of grounds of appeal was received on 28 August 2023.

III. The appellant requested that the decision under appeal be set aside and a patent granted on the basis of the claims of a main or one of three auxiliary requests. The appellant made a conditional request for oral proceedings.

IV. The board issued a summons to oral proceedings. In an annex to the summons, the board set out its preliminary opinion that the decision under appeal should be upheld.

V. On 4 July 2024, the appellant filed a response, maintaining its requests.

VI. The appellant requests that the decision under appeal be set aside and a patent granted on the basis of claims 1 to 8 of the main request filed with the statement of grounds of appeal, or claims 1 to 8 of one of auxiliary requests 1 to 3 that were the object of

the refusal, filed during the oral proceedings on 28 March 2023.

The further text on file is:

description pages
2 to 84 as originally filed,
1, 1a and 85 to 87 received on 27 December 2021;

drawing sheets
1 to 39 as originally filed.

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

"An apparatus to facilitate compute optimization, comprising:

a compute mechanism (610);

a central processing unit, CPU, (612) including one or more processors;

a graphics processing unit, GPU, (614) including

a plurality of processing units (700(a)-700(n)) each comprising a plurality of execution units, EUs (705, 706, 707), wherein the plurality of EUs (705, 706, 707) comprise a first EU type (705(a)-705(n)) and a second EU type (706(a)-706(n));

wherein the compute mechanism (610) transmits software hints to the graphics processing unit (614);

wherein the hints indicate that the graphics processing unit (614) is to power down, or bypass, higher power EUs (705, 706, 707) if processing of instructions requires less processing intensive EUs (705, 706, 707); wherein

the GPU (614) is implemented to process matrix operations in deep learning applications, wherein the processing units (700(a)-700(n)) being included within

memory to eliminate data transfers related to the deep learning matrix operations."

VIII. Compared to the main request, claim 1 of auxiliary request 1 comprises the additional feature that "the memory comprises a high bandwidth memory, HBM, (750), wherein the HBM comprises a first memory channel (752(0)) and a first processing unit (700(a)) included in the first memory channel (752(0)), and a second memory channel (752(1)) and a second processing unit (700(b)) included in the second processing channel (752(1))".

IX. Compared to the main request, the last paragraph of claim 1 is in auxiliary request 2 replaced by the following two paragraphs:

" wherein the compute mechanism (610) is to implement a register file (358, 334) within the GPU (614) to perform matrix-to-vector transformations and vector-to-matrix transformations;
wherein the compute mechanism (610) is to modify data within the register file (358, 334) without being required to move the data during operations to perform the transformations".

X. Compared to auxiliary request 2, the last paragraph of claim 1 is in auxiliary request 3 replaced by the following paragraph:

" wherein the compute mechanism (610) is to modify register content of the register file (358, 334) without being required to move matrix/vector data during operations to perform the transformations, wherein the matrix/vector data is stored in contiguous register blocks, the operations include a source

register and a destination register, the source register includes register address start limit, stride of an array, number of elements and element size, and once the operations are performed the results are stored in the destination register".

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

Reasons for the Decision

1. *The invention*

From claim 1 it can be inferred that the application aims to increase the efficiency of graphics processing units (GPUs) for deep learning applications.

To this end, the GPU of the present application includes a plurality of processing units, each comprising a plurality of execution units (EUs), wherein the plurality of EUs comprise a first EU type and a second EU type, the compute mechanism transmits software hints to the graphics processing unit indicating that the GPU is to power down, or bypass, higher power EUs if processing of instructions requires less processing intensive EUs, and the GPU is implemented to process matrix operations in deep learning applications, wherein the processing units are included within memory to eliminate data transfers related to the deep learning matrix operations (claim 1).

2. *Clarity; Article 84 EPC*

2.1 In the annex to its summons (point 6), the board had raised a number of clarity issues. Only the issues raised in points 6.3 to 6.7 were discussed at length during the oral proceedings. These form the basis for the board's conclusion regarding clarity of the main and auxiliary requests as formulated below. The other issues are not relevant to the board's conclusion and are not dealt with in the present decision.

2.2 It is not clear from claim 1 of the main request what an execution (EU) "type" is meant to refer to, or what is the impact on the claimed apparatus of the existence of two EU types, especially since no reference to those types is made later in the claim.

The appellant pointed out in its response to the summons (points 17 to 20) that the skilled person will read in the remainder of claim 1 that there are "higher power EUs" and "less processing intensive EUs", in claim 5 that the "first type" is selected to process a first type of application workload and the "second type" to process a second type of application workload, and in the description [00154] that the EU types may differ in the number of threads that may be processed, the number of registers per thread, or any other processing characteristic. For instance, 3D applications may require a larger number of threads and smaller thread register space, while media applications may require a smaller number of threads with larger register space.

During the oral proceedings, the appellant additionally observed that according to the description [00155], EU

types or configurations may be designed for specific deep learning use models.

According to the board, the above examples merely give a general idea to the skilled person what said EU types could be. They are not a substitute for a definition of those types in the claim itself.

In particular, the claim leaves open whether the "EU types" correspond to the "higher power" and "less processing intensive" EUs mentioned later in the claim. The description par. [00155] cited by the appellant would rather tend to indicate that the types are not necessarily linked to power consumption or processing intensity. Indeed, no connection exists between such values and particular deep learning use models.

2.3 It is not clear according to which criteria EUs are considered to be "higher power" or "less processing intensive". This is in particular true if EU types are merely designed for specific deep learning use models (description [00155]; see the above reasoning).

2.4 The feature according to which the compute mechanism transmits software "hints" to the graphics processing unit, wherein the "hints" indicate that the graphics processing unit is to power down, or bypass, higher power EUs if processing of instructions requires less processing intensive EUs, is not clear.

Firstly, the context to be considered when deciding which kind of EUs the processing of instructions requires is not clear. Is the context one of individual instructions, tasks/functions, threads, or something else?

Secondly, it is not clear whether the compute mechanism or the graphics processing unit takes such decision. When taking the claim literally, the hints merely are an indication for the graphics processing unit to power down or bypass higher power instructions if processing of instructions requires less processing intensive EUs, which would seem to leave it to the GPU to make the decision whether less processing intensive EUs are required.

Thirdly, the process which would allow to make such a decision is not clear. Is the decision already made in advance and entered in some table, or is some kind of monitoring performed to allow making the decision, and if so which kind of monitoring?

Fourthly, the nature of the "hints" is not clear. During the oral proceedings, the appellant declared that the term should be understood as "instructions".

Fifthly, the consequence of the transmission of the "hint" is not clear. The claim leaves it for instance open whether or under which conditions the GPU actually powers down or bypasses higher power EUs after having received a corresponding indication in the form of a "hint".

2.5 It is not clear what it means for processing units to be included "within" memory.

The appellant explained during the oral proceedings that the processing units are included in the channels as illustrated in figure 7B of the application. The application however provides no detail about a structure of the channels which would make such

inclusion possible. In its simplest form, a channel would actually be nothing more than a conducting wire.

The appellant stressed that the "channels" in case of for instance HBM memory would be a substantially more complex structure than a simple wire. However, even when considering the particular case of HBM memory, which is not part of claim 1 in the main request, it is still not clear what a positioning of the processing units within the channels would mean in practice.

2.6 It is furthermore apparent that data transfers related to deep learning matrix operations will not be "eliminated" because of an inclusion of processing units within memory. Data transfers would still be necessary, even if the path for such transfers were shortened.

2.7 The main request therefore does not satisfy the requirements of Article 84 EPC (clarity).

2.8 The features discussed under points 2.2 to 2.4 above appear in claim 1 of all auxiliary requests. None of the auxiliary requests 1 to 3 therefore satisfy the requirements of Article 84 EPC, for the reasons given under those points for claim 1 of the main request.

Order

For these reasons it is decided that:

The appeal is dismissed.

The Registrar: The Chairman:

L. Stridde Martin Müller

Decision electronically authenticated