(12) International Application Status Report

Received at International Bureau: 19 December 2019 (19.12.2019)
Information valid as of: 13 May 2020 (13.05.2020)
Report generated on: 19 September 2020 (19.09.2020)

(10) Publication number:
WO2020/117926

(43) Publication date:
11 June 2020 (11.06.2020)

(26) Publication language:
English (EN)

(21) Application Number:
PCT/US2019/064454

(22) Filing Date:
04 December 2019 (04.12.2019)

(25) Filing language:
English (EN)

(31) Priority number(s):
16/211,954 (US)

(31) Priority date(s):
06 December 2018 (06.12.2018)

(31) Priority status:
Priority document received (in compliance with PCT Rule 17.1)

(51) International Patent Classification:
G06F 17/16 (2006.01); **G06F 9/38** (2006.01); ***G01T 1/20*** (2006.01); ***G06N 3/02*** (2006.01)

(71) Applicant(s):
ADVANCED MICRO DEVICES, INC. [US/US]; 2485 Augustine Drive Santa Clara, California 95054 (US) *(for all designated states)*

(72) Inventor(s):
NEMLEKAR, Milind N.; 2485 Augustine Drive Santa Clara, California 95054 (US)

(74) Agent(s):
SHEEHAN, Adam D.; Davidson Sheehan LLP 6836 Austin Center Blvd., Suite 320 Austin, Texas 78731 (US)

(54) Title (EN): PIPELINED MATRIX MULTIPLICATION AT A GRAPHICS PROCESSING UNIT

(54) Title (FR): MULTIPLICATION MATRICIELLE EN PIPELINE AU NIVEAU D'UN PROCESSEUR GRAPHIQUE

(57) Abstract:
(EN): A graphics processing unit (GPU) [100] schedules recurrent matrix multiplication operations at different subsets of CUs [110, 111, 112, 113] of the GPU. The GPU includes a scheduler [104] that receives sets of recurrent matrix multiplication operations [103, 114], such as multiplication operations associated with a recurrent neural network (RNN). The multiple operations associated with, for example, an RNN layer are fused into a single kernel, which is scheduled by the scheduler such that one work group is assigned per compute unit, thus assigning different ones of the recurrent matrix multiplication operations to different subsets of the CUs of the GPU. In addition, via software synchronization of the different workgroups, the GPU pipelines the assigned matrix multiplication operations so that each subset of CUs provides corresponding multiplication results to a different subset, and so that each subset of CUs executes at least a portion of the multiplication operations concurrently.

(FR): Un processeur graphique (GPU) [100] planifie des opérations de multiplication matricielle récurrente au niveau de différents sous-ensembles de CU [110, 111, 112, 113] du GPU. Le GPU comprend un planificateur [104] qui reçoit des ensembles d'opérations de multiplication matricielle récurrente [103, 114], telles que des opérations de multiplication associées à un réseau de neurones bouclé (RNN). Les multiples opérations associées, par exemple, à une couche RNN sont fusionnées en un noyau unique, qui est planifié par le programmateur de telle sorte qu'un groupe de travail est attribué par unité de calcul, ce qui permet d'attribuer différentes opérations de multiplication matricielle récurrente à différents sous-ensembles de CU du GPU. De plus, par l'intermédiaire d'une synchronisation logicielle des différents groupes de travail, le GPU effectue en pipeline les opérations de multiplication matricielle attribuées de sorte que chaque sous-ensemble de CU fournit des résultats de multiplication correspondants à un sous-ensemble différent, et de sorte que chaque sous-ensemble de CU exécute au moins une partie des opérations de multiplication simultanément.

International search report:
Received at International Bureau: 30 March 2020 (30.03.2020) [KR]

International Report on Patentability (IPRP) Chapter II of the PCT:

Not available

(81) Designated States:

AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW

European Patent Office (EPO) : AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR

African Intellectual Property Organization (OAPI) : BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG

African Regional Intellectual Property Organization (ARIPO) : BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW

Eurasian Patent Organization (EAPO) : AM, AZ, BY, KG, KZ, RU, TJ, TM