



- (51) International Patent Classification:
G06F 7/544 (2006.01)
- (21) International Application Number:
PCT/EP2021/087303
- (22) International Filing Date:
22 December 2021 (22.12.2021)
- (25) Filing Language: English
- (26) Publication Language: English
- (71) Applicant: **AXELERA AI BV** [NL/NL]; High Tech Campus 5 (HTC5), 5656 AE Eindhoven (NL).
- (72) Inventors: **KHADDAM-ALJAMEH, Riduan**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **ELEFThERIOU, Evangelos**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL).

(NL). **PAPISTAS, Ioannis**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **KATSELAS, Leonidas**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **HAGER, Pascal**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **ROOSELEER, Bram**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **MOONS, Bert**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **COSEMANS, Stefan**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **UYTTERHOEVEN, Roel**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **GARCEA, Giuseppe**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **POLIAKOV, Dmitri**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **YI, Lu**; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL). **VAN LOON, Jeroen**; c/o AXELERA AI

(54) Title: IN-MEMORY PROCESSING BASED ON MULTIPLE WEIGHT SETS

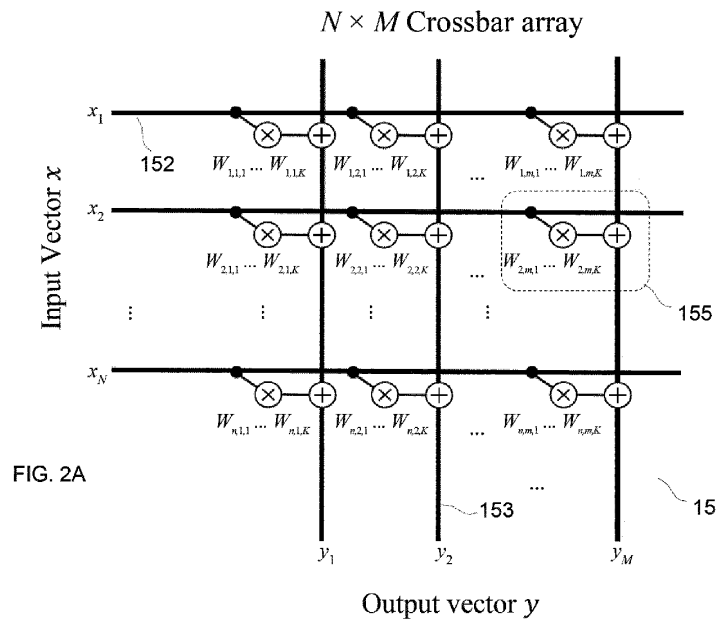


FIG. 2A

(57) **Abstract:** The invention is notably directed to a method of in-memory processing, the aim of which is to perform matrix-vector calculations. The method relies on a device having a crossbar array structure (15). The latter includes N input lines (152) and M output lines (153), which are interconnected at cross-points defining $N \times M$ cells (155), where $N \geq 2$ and $M \geq 2$. The cells include respective memory systems, each designed to store K weights $W_{i,j,k}$, where $K \geq 2$. Thus, the crossbar array structure includes $N \times M$ memory systems, which are capable of storing K sets of $N \times M$ weights. In order to perform multiply-accumulate (MAC) operations, the method first enables $N \times M$ active weights for the $N \times M$ cells by selecting, for each of the memory systems, a weight from its K weights and setting the selected weight as an active weight. Next, signals encoding a vector of N components are

[Continued on next page]

BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL).
MACHIELS, Brecht; c/o AXELERA AI BV, HTC5, High Tech Campus, 5656 AE Eindhoven (NL).

(74) **Agent: E. BLUM & CO. AG**; Vorderberg 11, 8044 Zürich (CH).

(81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): 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, IT, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE

applied to the N input lines of the crossbar array structure. This causes the latter to perform MAC operations based on the vector and the $N \times M$ active weights. Eventually, output signals obtained in output of the M output lines are read out to obtain corresponding values. This allows distinct sets of weights to be locally enabled at the crossbar array, which makes it possible to locally perform rotations of the weights and accordingly reduce the frequency of data exchanges with a memory unit. This, in turn, reduces idle times of the crossbar array structure. Thus, the proposed approach makes it possible to substantially reduce the frequency of data transfers, which results in speeding up computations. Plus, the weights may possibly be prefetched, while performing MAC operations in accordance with the currently active weights. Of particular advantage is that the prefetching steps are at least partly hidden through pipelining. The invention is further directed to related devices, systems, and computer program products.