



(51) International Patent Classification:

G06F 15/173 (2006.01) G06N 3/063 (2023.01)
G06F 15/80 (2006.01)

(21) International Application Number:

PCT/EP2024/054036

(22) International Filing Date:

16 February 2024 (16.02.2024)

(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: HAGER, Pascal; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). MOONS, Bert; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL).

AE Eindhoven (NL). KHOV, Gua Hao; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). GEURSEN, Alexander; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). SCHMUCK, Manuel; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). GOETSCHALCKX, Koen; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). UYTTERHOEVEN, Roel; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). KOUMOUI, Spyridoula; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). STANISAVLJEVIC, Milos; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). MACH, Stefan; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). ZARUBA, Florian; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). LU, Yi; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). BUSSOLINO, Beatrice; c/o AXELERA AI

(54) Title: DATAFLOW ARCHITECTURES TO ACCELERATE EXECUTIONS OF INTERDEPENDENT OPERATIONS SUCH AS INVOLVED IN AI INFERENCES

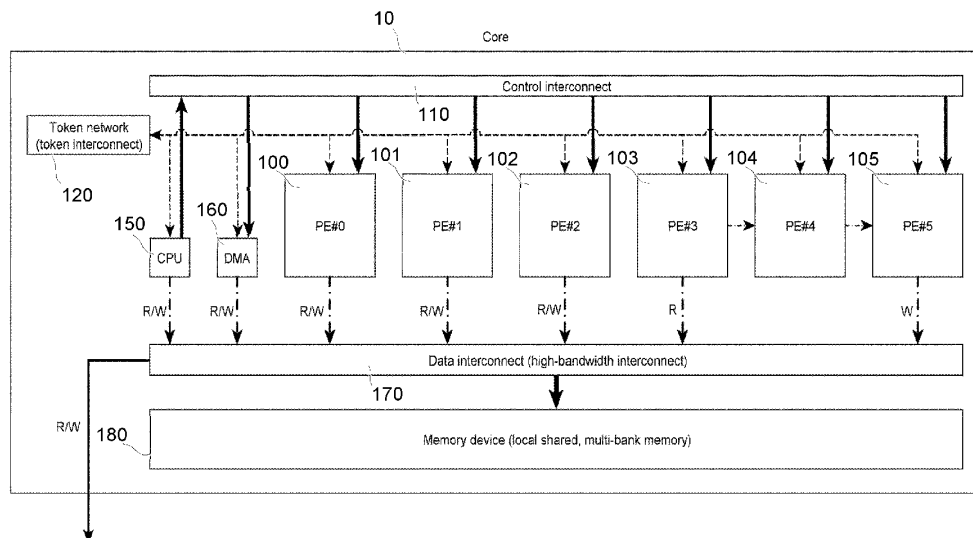


FIG. 1

(57) Abstract: The invention is notably directed to a data processing device (10) for executing interdependent operations. The device comprises a control unit (150), e.g., a CPU, a memory device (180), several processing elements (100 - 105), and distinct interconnects (110, 120, 170). The processing elements are configured to execute operations, e.g., matrix-vector multiplication (MVM) operations as involved in artificial neural network (ANN) executions. The interconnects include: a data interconnect (170), which interfaces the memory device (180) with at least some of the processing elements; a control interconnect (110), which interfaces the control unit (150) with each of the processing elements; and a token interconnect (120), which interfaces each of the processing elements with at least another one of these processing elements. The control unit is configured to dispatch commands to the processing elements through the control interconnect. The commands specify conditions for executing the operations in accordance with a predetermined

[Continued on next page]

BV, High Tech Campus 5, 5656 AE Eindhoven (NL). **HÜT-TER, Noah**; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL). **ELEFThERIOU, Evangelos**; c/o AXELERA AI BV, High Tech Campus 5, 5656 AE Eindhoven (NL).

(74) **Agent: E. BLUM & CO. AG**; Hofwiesenstrasse 349, 8050 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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

interdependence of the operations. In addition, the processing elements are configured to send and/or to receive tokens via the token interconnect, in accordance with said conditions, to orchestrate the executions of the operations. In operation, executing these operations cause at least some of the processing elements to read from and/or write to the memory device through the data interconnect. The above hardware architecture is scalable. The invention is further directed to related systems (comprising clusters of such data processing devices) and related methods.