



- (51) **International Patent Classification:**
G06N 3/04 (2006.01) *G06N 3/063* (2006.01)
- (21) **International Application Number:**
PCT/CA2019/050187
- (22) **International Filing Date:**
15 February 2019 (15.02.2019)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/710,488 16 February 2018 (16.02.2018) US
62/664,190 29 April 2018 (29.04.2018) US
- (71) **Applicant: THE GOVERNING COUNCIL OF THE UNIVERSITY OF TORONTO** [CA/CA]; Banting Institute, Room 413, 100 College St., Toronto, Ontario M5G 1L5 (CA).
- (72) **Inventors: MOSHOVOS, Andreas;** 24 Wellesley St. W., Suite 1613, Toronto, Ontario M4Y 2X6 (CA). **DELMAS LASCORZ, Alberto;** 567 Bathurst St., Toronto, Ontario M5S 2P8 (CA). **POULOS, Zisis;** 250 St. George St., Suite 401, Toronto, Ontario M5R 3L8 (CA). **MALONE STUART, Dylan;** 72 Arnold Ave., Toronto, Ontario M5S 3B2 (CA). **JUDD, Patrick;** 21034 Vanedie Line RR4, Merlin, Ontario N0P 1W0 (CA). **SHARIFY, Sayeh;** 2885 Bayview Ave., Unit 215, Toronto, Ontario M2K 0A3 (CA). **MAHMOUD, Mostafa;** 30 Charles St. West, Apt. 1715, Toronto, Ontario M4Y 1R5 (CA). **NIKOLIC, Milos;** 60 Southport St., Apt. 524, Toronto, Ontario M6S 3N4 (CA). **SIU, Kevin Chong Man;** 141 Carolbreen Square, Toronto, Ontario M1V 1J1 (CA).

(54) **Title:** NEURAL NETWORK ACCELERATOR

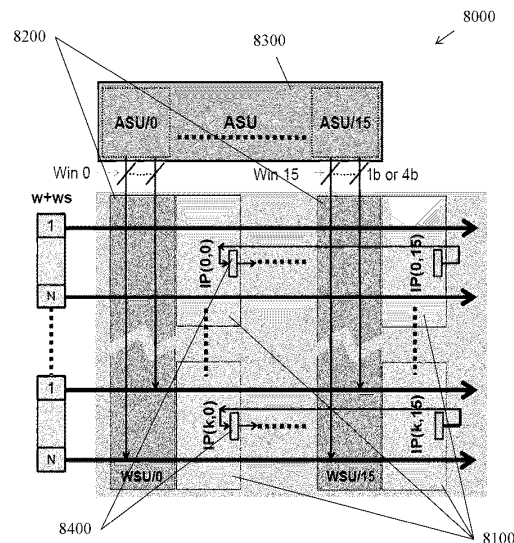


FIG. 8

(57) **Abstract:** Described is a neural network accelerator tile for exploiting input sparsity. The tile includes a weight memory to supply each weight lane with a weight and a weight selection metadata, an activation selection unit to receive a set of input activation values and rearrange the set of input activation values to supply each activation lane with a set of rearranged activation values, a set of multiplexers including at least one multiplexer per pair of activation and weight lanes, where each multiplexer is configured to select a combination activation value for the activation lane from the activation lane set of rearranged activation values based on the weight lane weight selection metadata, and a set of combination units including at least one combination unit per multiplexer, where each combination unit is configured to combine the activation lane combination value with the weight lane weight to output a weight lane product.

[Continued on next page]

(74) **Agent: HEER, Christopher**; 25 Adelaide Street East, Suite 1000, Toronto, Ontario M5C 3A1 (CA).

(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, 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.

(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))