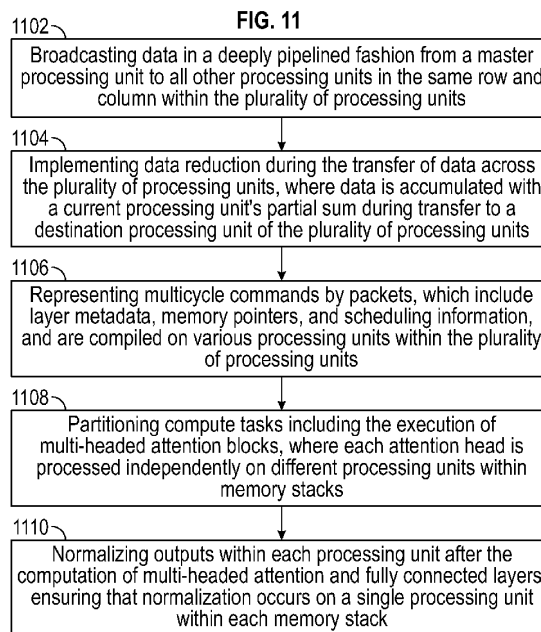




- (51) **International Patent Classification:**
G06N 3/02 (2006.01) *G06N 3/08* (2023.01)
- (21) **International Application Number:**
PCT/US2024/040380
- (22) **International Filing Date:**
31 July 2024 (31.07.2024)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
63/530,849 04 August 2023 (04.08.2023) US
18/789,431 30 July 2024 (30.07.2024) US
- (71) **Applicant:** **EXPEDERA, INC.** [US/US]; 3211 Scott Boulevard, Suite #204, Santa Clara, California 95054 (US).
- (72) **Inventors:** **CHUANG, Shang-Tse**; 3211 Scott Boulevard, Suite #204, Santa Clara, California 95054 (US). **MA, Siyad Chih-Hua**; 3211 Scott Boulevard, Suite #204, Santa Clara, California 95054 (US). **CHOLE, Sharad Vasantrao**; 3211 Scott Boulevard, Suite #204, Santa Clara, California 95054 (US). **CALAMVOKIS, Costas**; 3211 Scott Boulevard, Suite #204, Santa Clara, California 95054 (US).
- (74) **Agent:** **BATHURST, K. Brian** et al.; Carr & Ferrell LLP, 411 Borel Ave, Suite 603, San Mateo, California 94402 (US).
- (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,

(54) **Title:** MACHINE LEARNING MODEL SCALABILITY WITH DISTRIBUTED MULTI-LAYER PROCESSING



(57) **Abstract:** Machine learning model scalability with distributed multi-layer processing is disclosed herein. A method for processing and deploying machine learning models that enhances scalability and efficiency by executing a subset of a neural network on each of a plurality of interconnected processing units. The method involves partitioning compute tasks across these processing units to reduce latency, including broadcast and reduction processes for inputs and outputs. It also includes managing the allocation of samples in a batch to specific master processing units within the distributed arrangement and synchronizing computation between fully connected layers within each processing unit. Additionally, the method implements data reduction during the transfer of data across the processing units, wherein data is accumulated with a current processing unit's partial sum as it is transferred to the destination processing unit.

[Continued on next page]

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:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*