

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
20 March 2025 (20.03.2025)



(10) International Publication Number
WO 2025/058740 A1

(51) International Patent Classification:

G06F 8/41 (2018.01) G06N 20/00 (2019.01)
G06F 9/44 (2018.01) G06F 9/00 (2018.01)

(21) International Application Number:

PCT/US2024/041345

(22) International Filing Date:

07 August 2024 (07.08.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

18/465,492 12 September 2023 (12.09.2023) US

(71) Applicant: **GROQ, INC.** [US/US]; 301 Castro Street, Suite 200, Mountain View, California 94041 (US).

(72) Inventor: **JINDEL, Samir**; c/o 301 Castro Street, Suite 200, Mountain View, California 94041 (US).

(74) Agent: **WORKMAN, Parks J.** et al.; Doritiy & Manning, PA, PO Box 1449, Greenville, SC 29602 (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, 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))

(54) Title: OPTIMIZING INSTRUCTION SCHEDULING AND MEMORY ALLOCATION FOR TENSOR AND GRAPHICAL PROCESSORS USING LATTICE IMAGE DATA STRUCTURE OPTIMIZATIONS

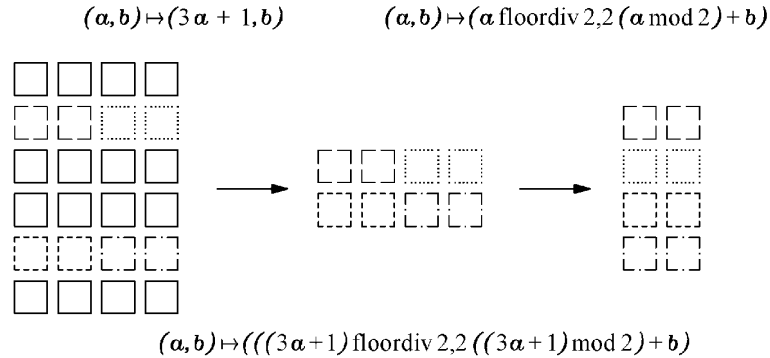


FIG. 1

(57) Abstract: Methods are provided using lattices (representing graph data structures that encode a machine learning application program) and lattice images and algorithms to accelerate optimization problems that arise during the mapping of tensor processing algorithms to specialized processors, such as for machine learning (ML) or artificial intelligence (AI), optimizations such as the scheduling of instructions and the allocation of memory. These optimizations are obtained with improved compiler methods that use lattice image and transformation algorithms to accelerate, among other tasks, memory partitioning, operation splitting, loop fusion, and optimal allocation of intermediate results of matrix mathematical operations.



WO 2025/058740 A1