



(51) International Patent Classification:

G06F 9/50 (2006.01) G06F 3/045 (2006.01)
G06N 3/063 (2023.01)

(21) International Application Number:

PCT/US2024/053102

(22) International Filing Date:

25 October 2024 (25.10.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/593,413 26 October 2023 (26.10.2023) US

(71) Applicant: **GOOGLE LLC** [US/US]; 1600 Amphitheatre
Parkway, Mountain View, California 94043 (US).

(72) Inventors: **PIQUERAS, Enrique**; c/o Google LLC, 1600
Amphitheatre Parkway, Mountain View, California 94043
(US). **BRADBURY, James Edward Kahn**; c/o Google
LLC, 1600 Amphitheatre Parkway, Mountain View, Cali-
fornia 94043 (US). **SAETA, Brennan**; c/o Google LLC,
1600 Amphitheatre Parkway, Mountain View, California
94043 (US). **DOUGLAS, Sholto Francis Alexandre**; c/
o Google LLC, 1600 Amphitheatre Parkway, Mountain
View, CA 94043 (US). **LEVSKAYA, Anselm**; c/o Google
LLC, 1600 Amphitheatre Parkway, Mountain View, Cali-
fornia 94043 (US). **SCHUH, Parker Edward**; c/o Google
LLC, 1600 Amphitheatre Parkway, Mountain View, Cali-
fornia 94043 (US). **POPE, Reiner**; c/o Google LLC, 1600
Amphitheatre Parkway, Mountain View, California 94043
(US).

(54) Title: EFFICIENTLY SERVING MACHINE-LEARNED MODEL COMPUTATIONS WITH HIGH THROUGHPUT AND
LOW LATENCY

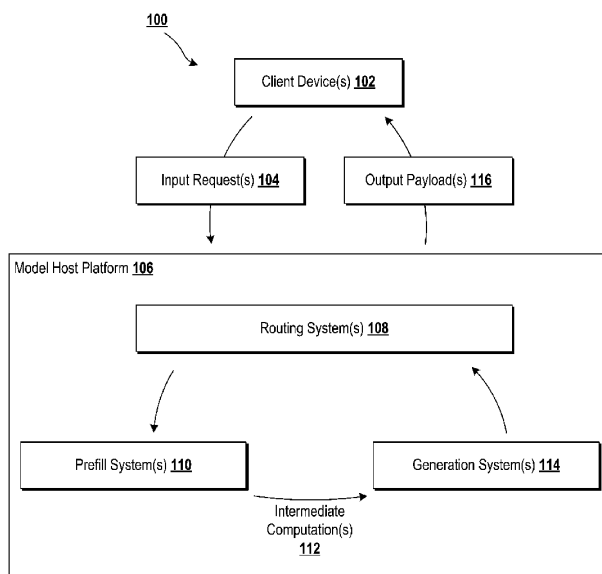


FIG. 1

(57) **Abstract:** An example method includes receiving input requests to process a plurality of input sequences using the machine-learned sequence processing model to generate a plurality of output sequences respectively corresponding to the plurality of input sequences; generating a plurality of initial attention tensors respectively for the plurality of input sequences, wherein: one or more respective initial attention tensors are generated for each respective input sequence in parallel over input elements of the respective input sequence; and the one or more respective initial attention tensors are generated in one or more batches having a first batch size using a prefill system that comprises one or more prefill computing devices and executes one or more layers of the machine-learned sequence processing model; and autoregressively generating, using the plurality of initial attention tensors, a plurality of output elements for each of the plurality of output sequences in one or more batches having a second batch size, wherein: the plurality of output elements are generated

[Continued on next page]

(74) **Agent: NOEL, James H.** et al.; Dority & Manning, P.A.,
P.O. BOX 1449, Greenville, South Carolina 29602-1449
(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).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

using a generation system that comprises one or more generation computing devices.