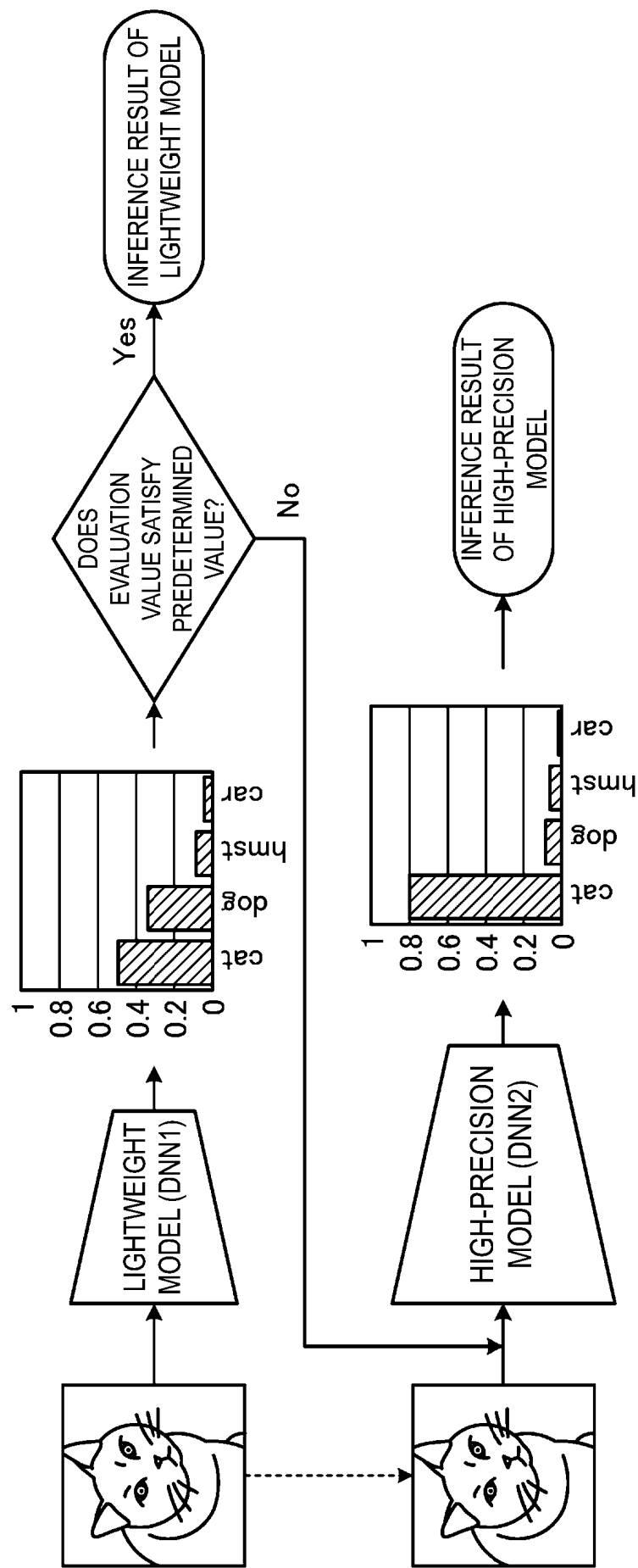


Fig. 1



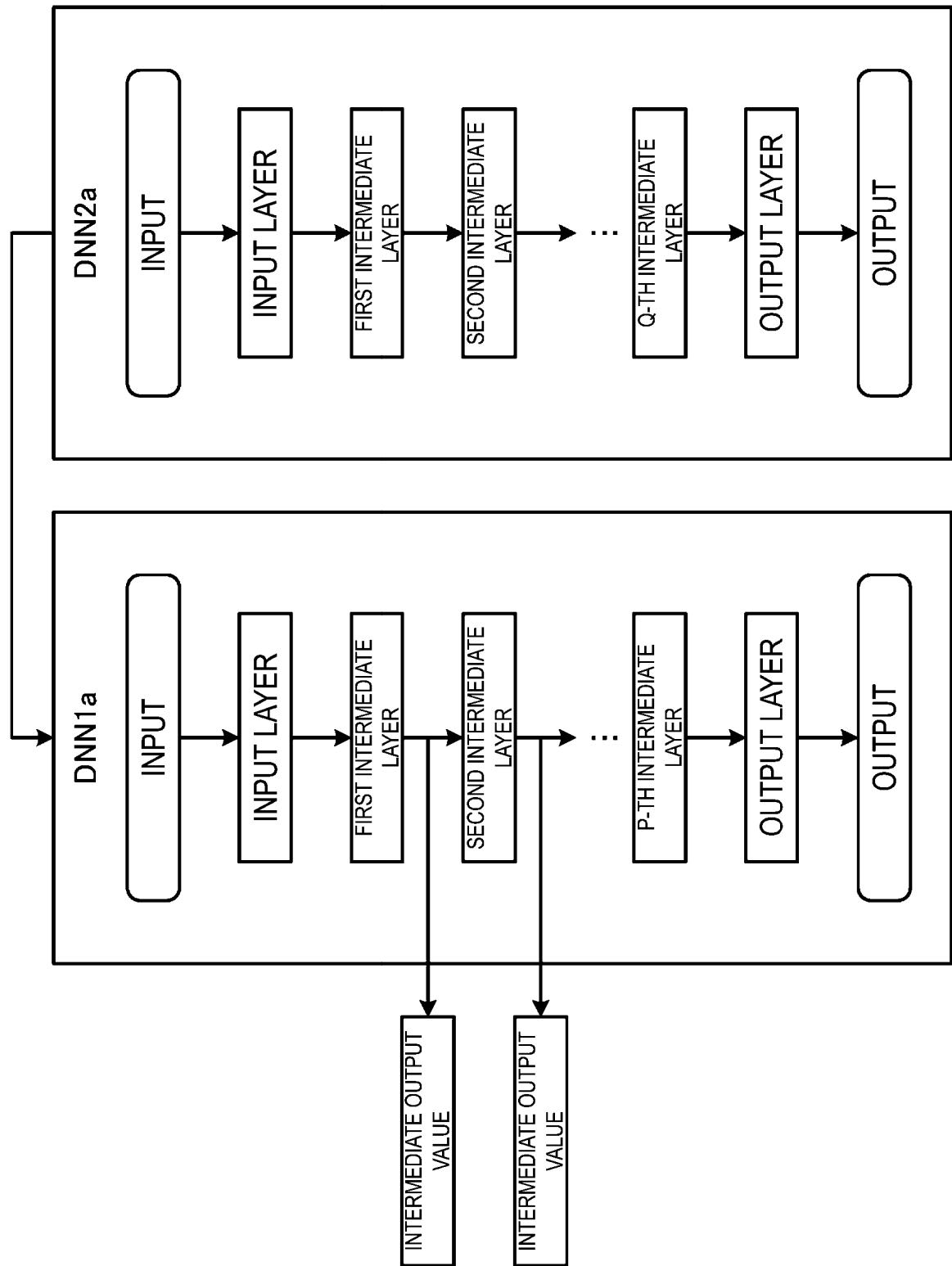


Fig. 2

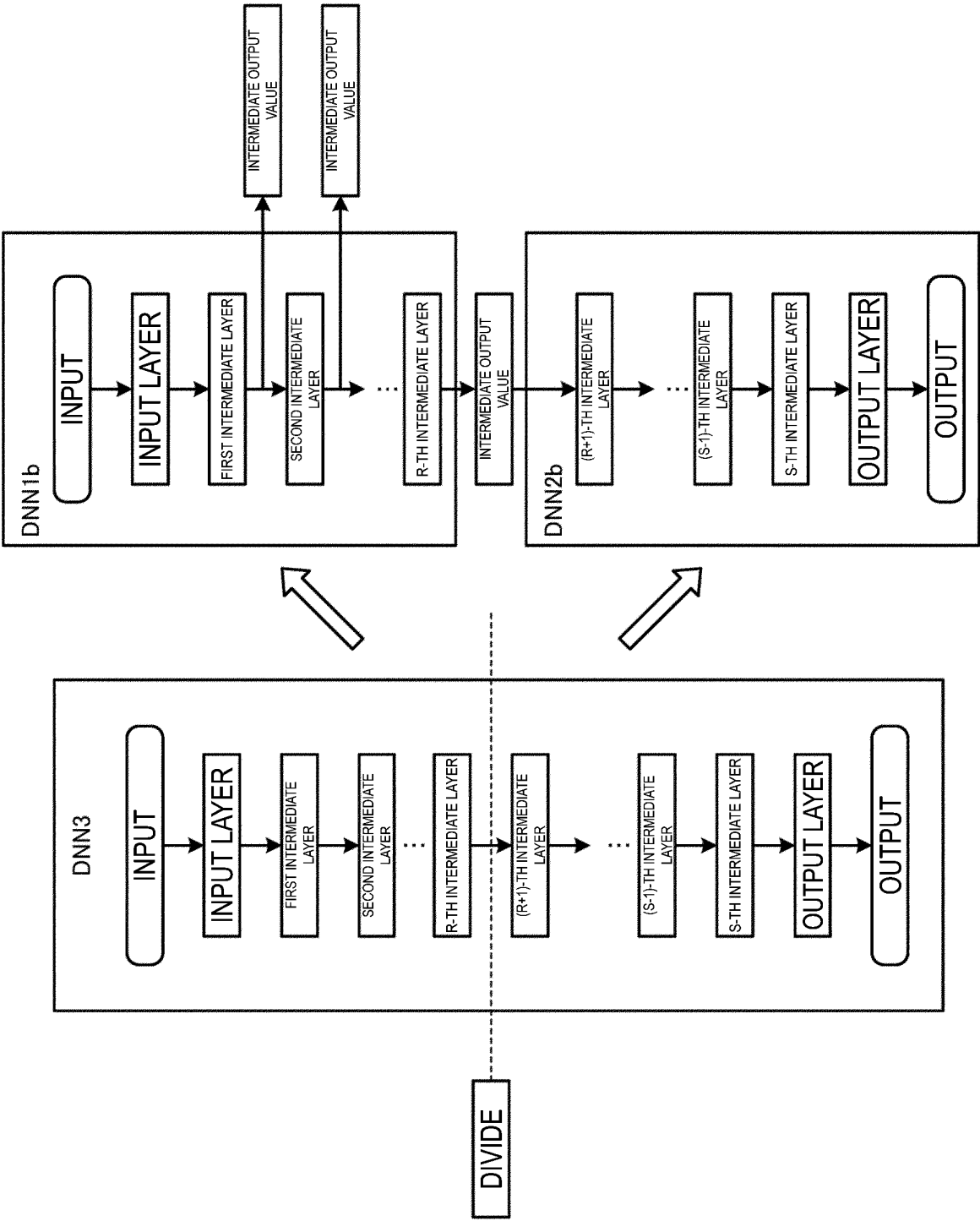


Fig. 3

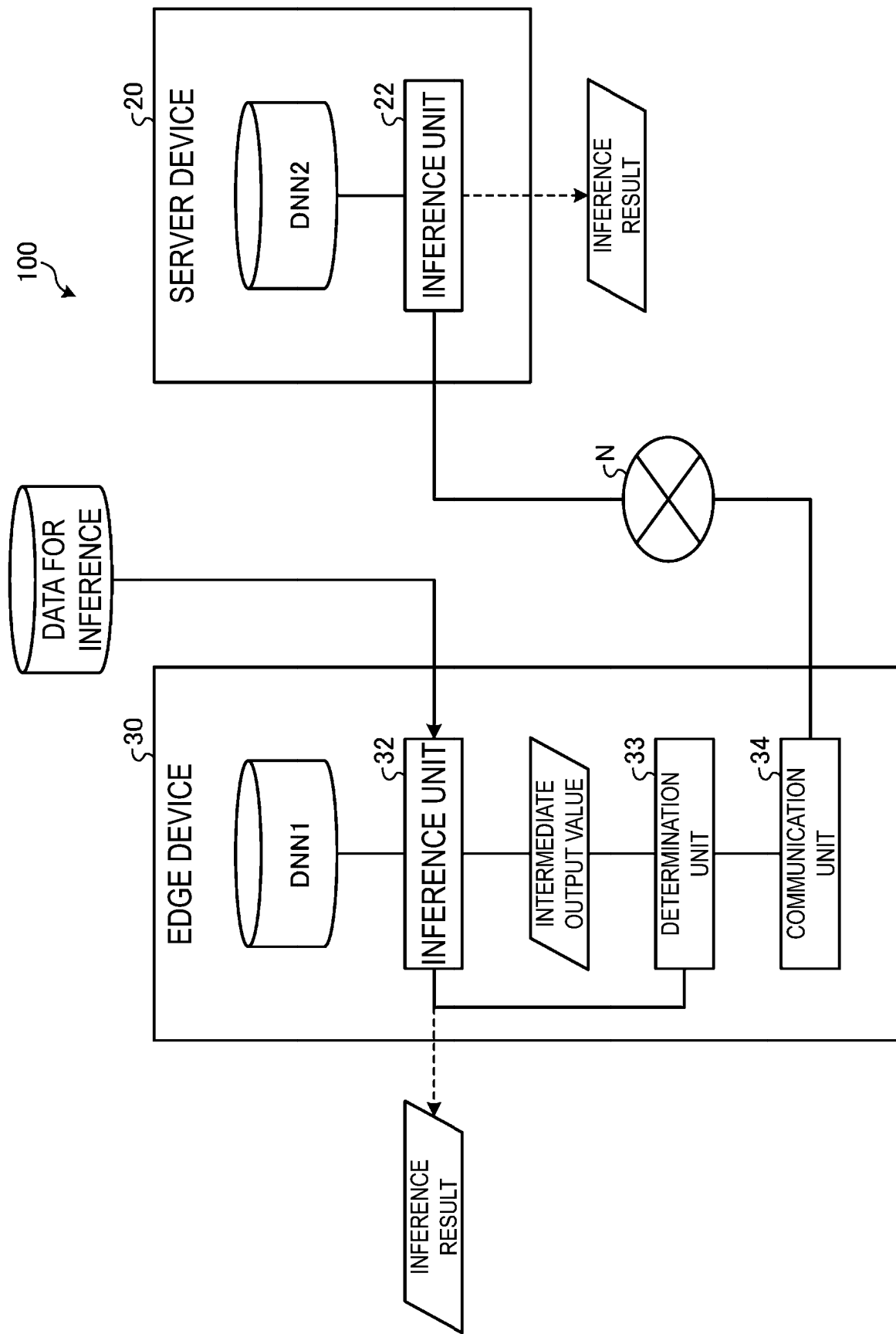


Fig. 4

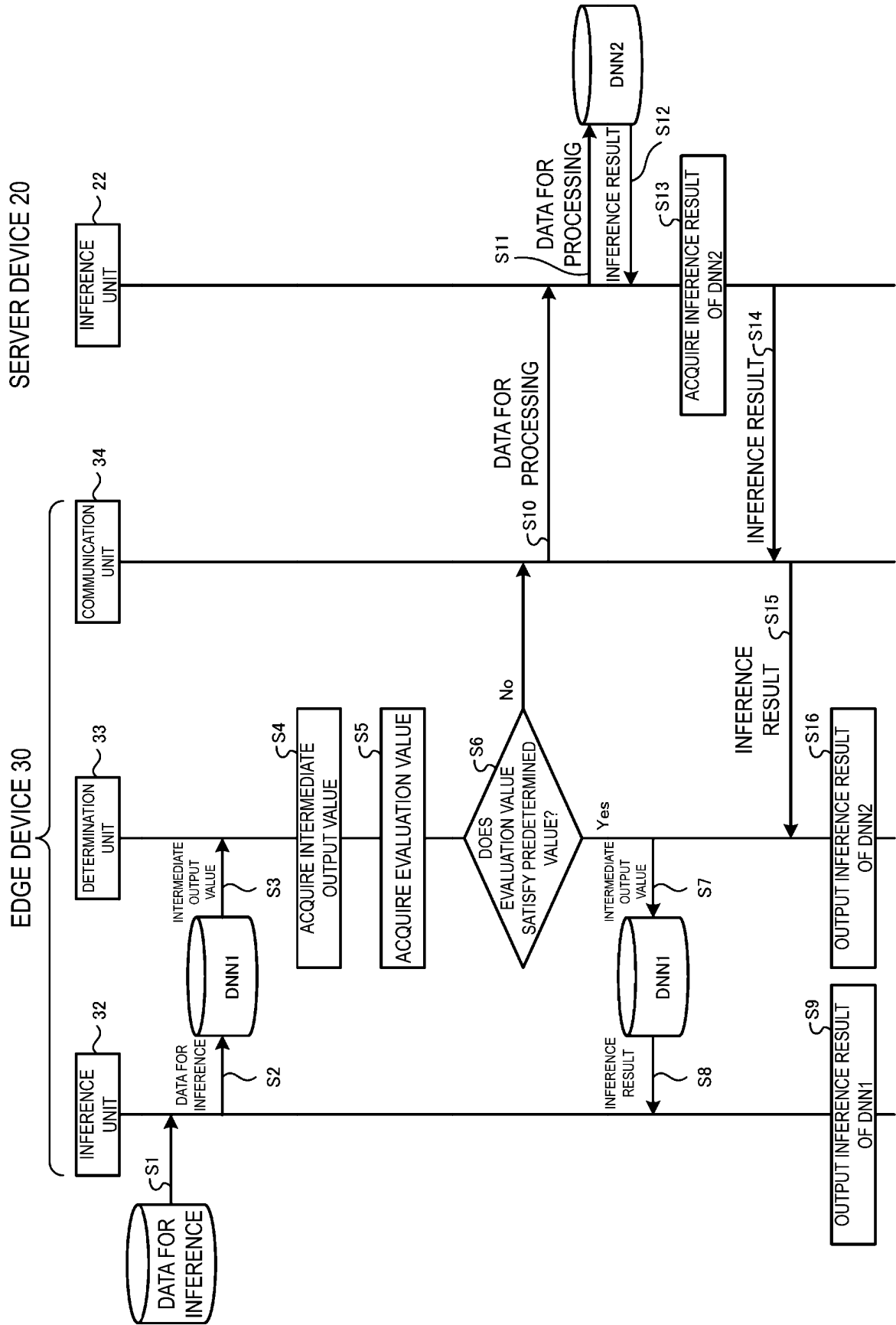


Fig. 5

Fig. 6

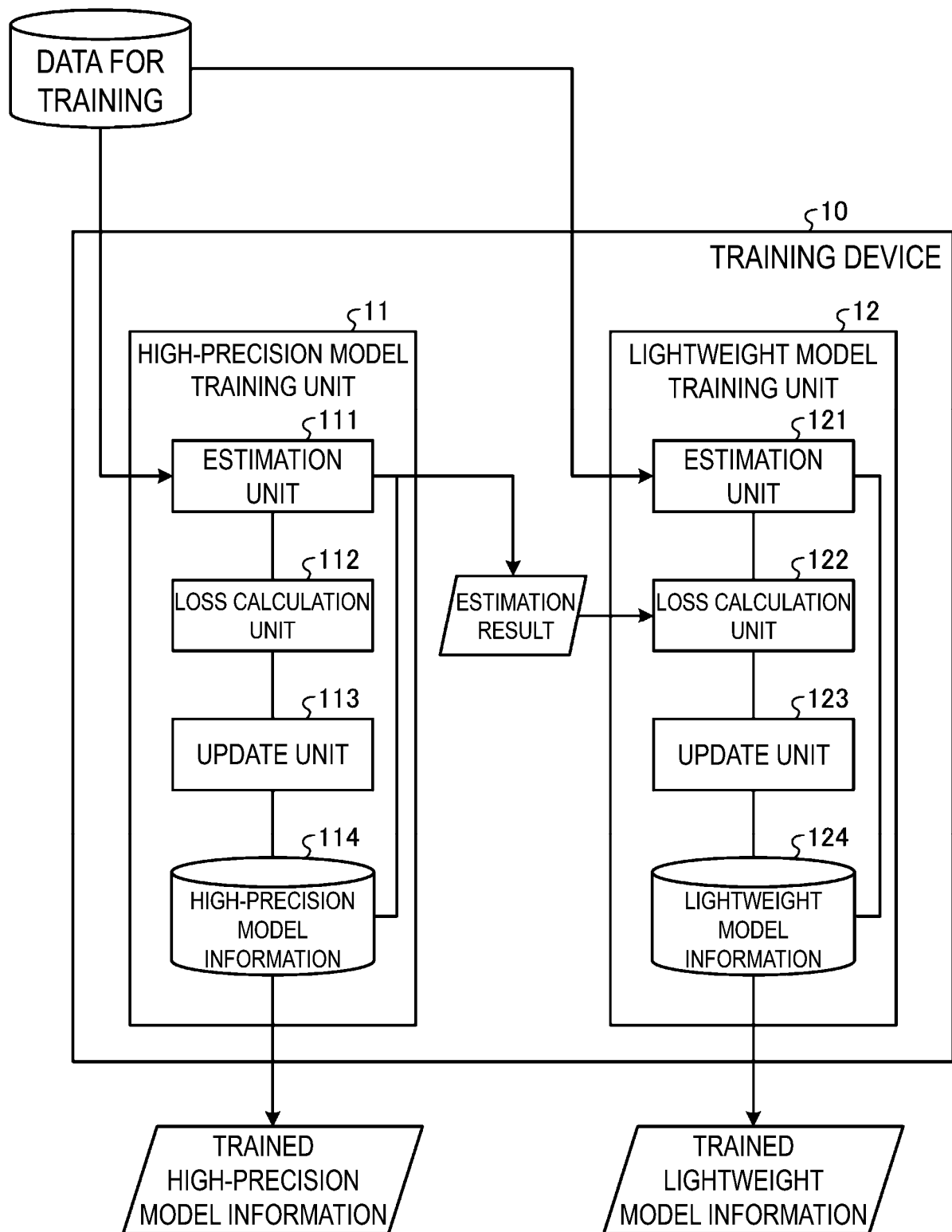


Fig. 7

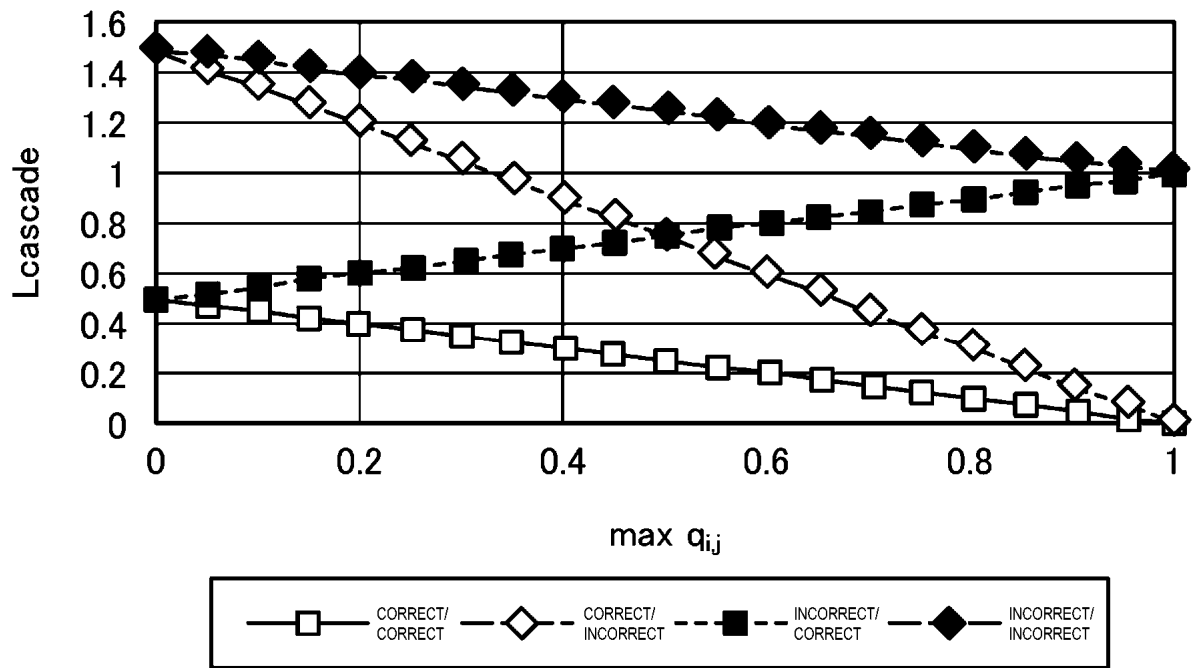


Fig. 8

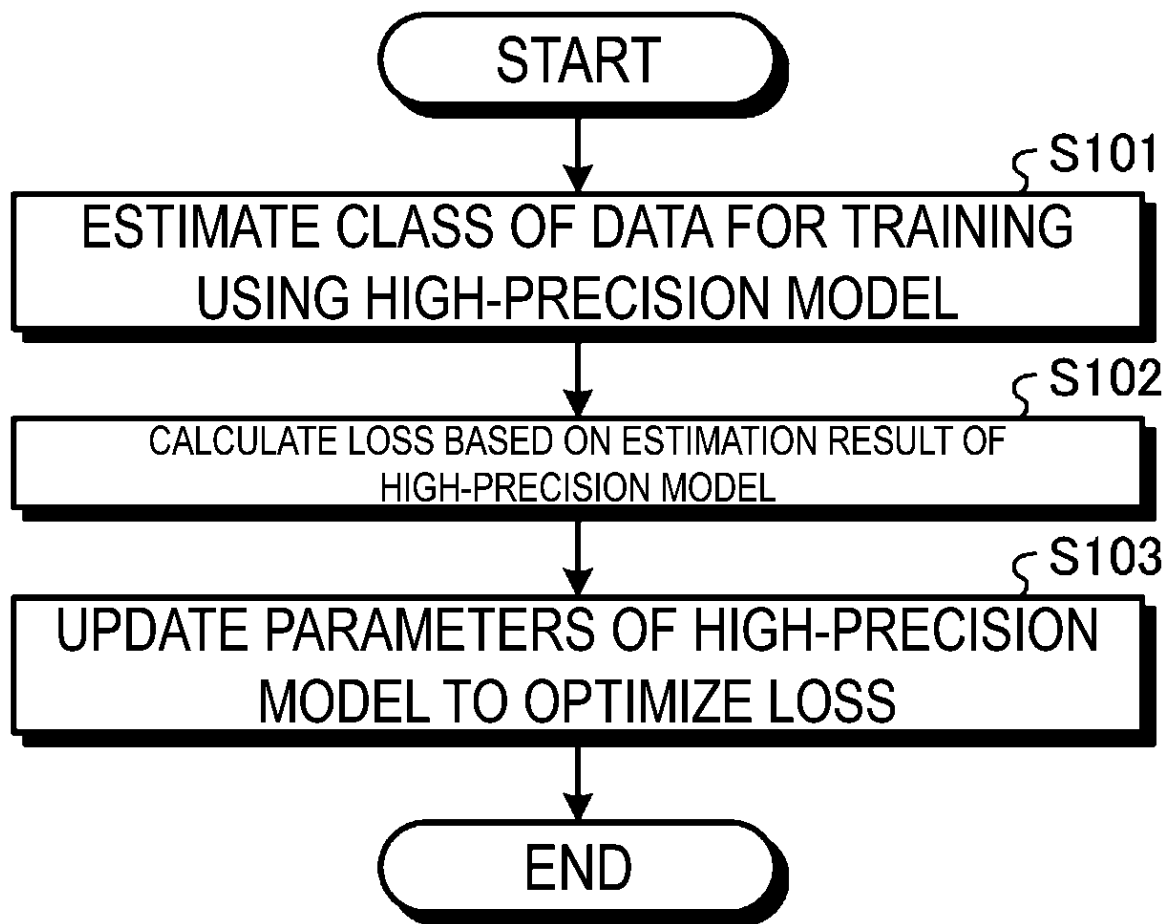
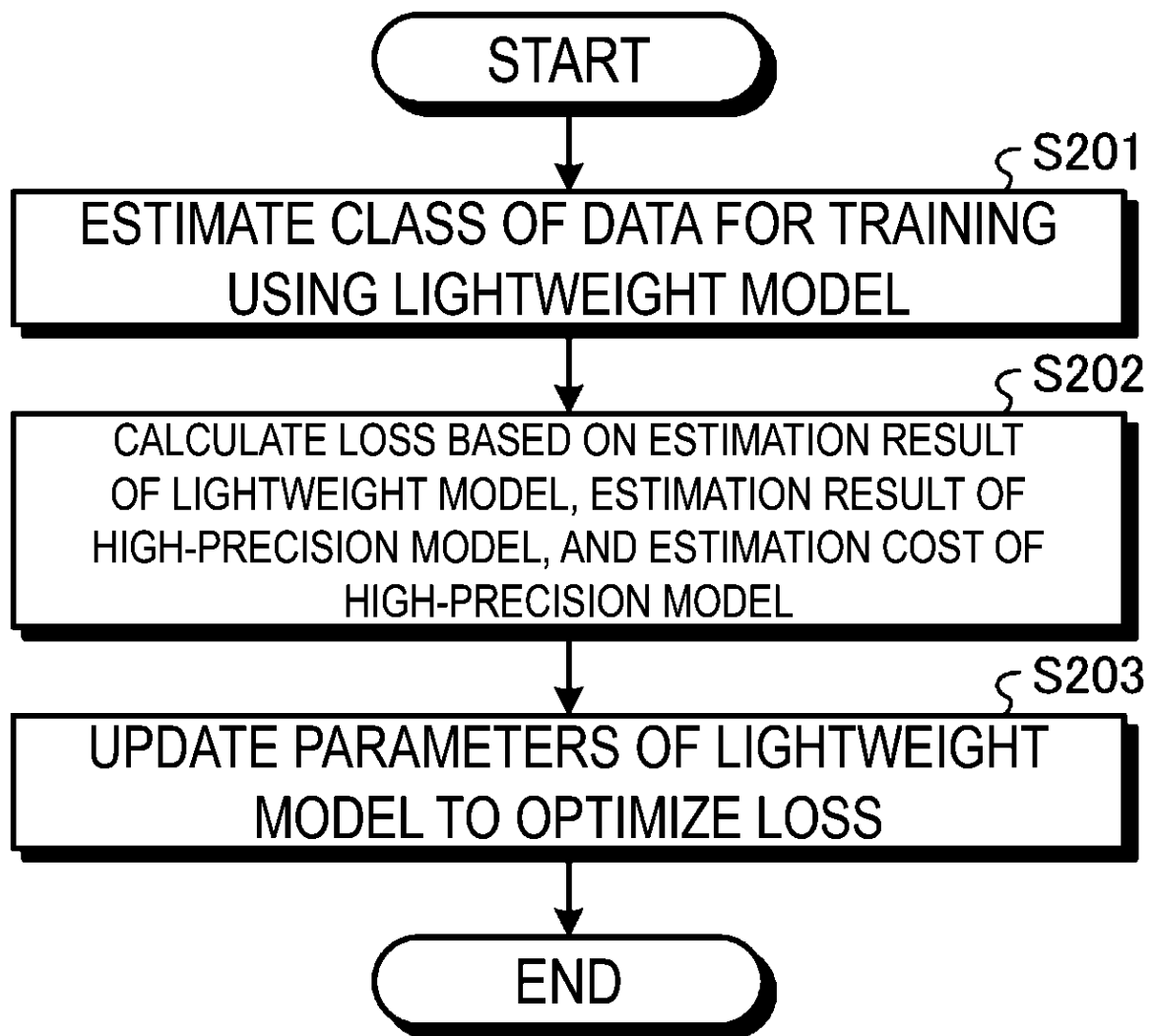


Fig. 9



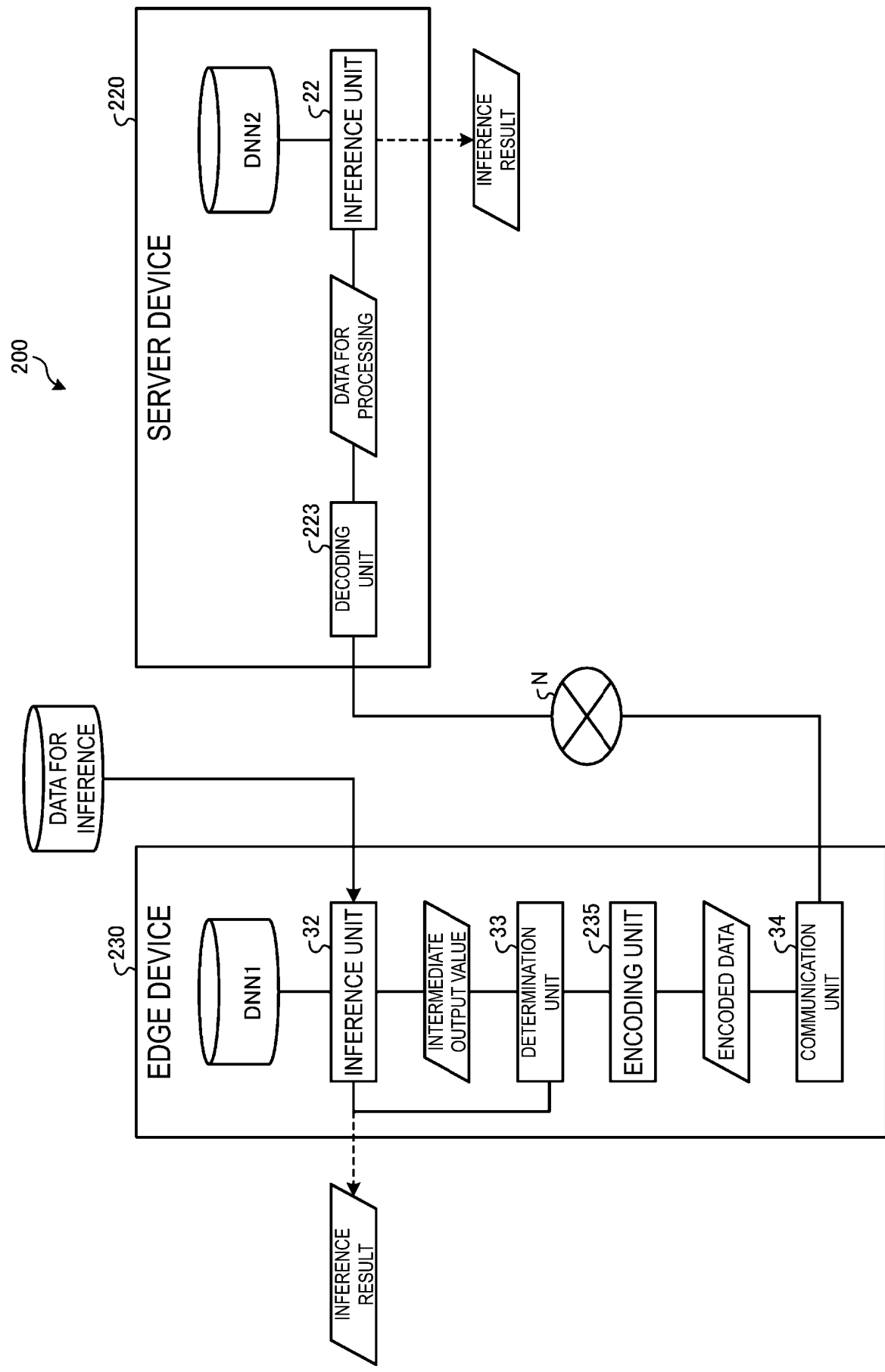


Fig. 10

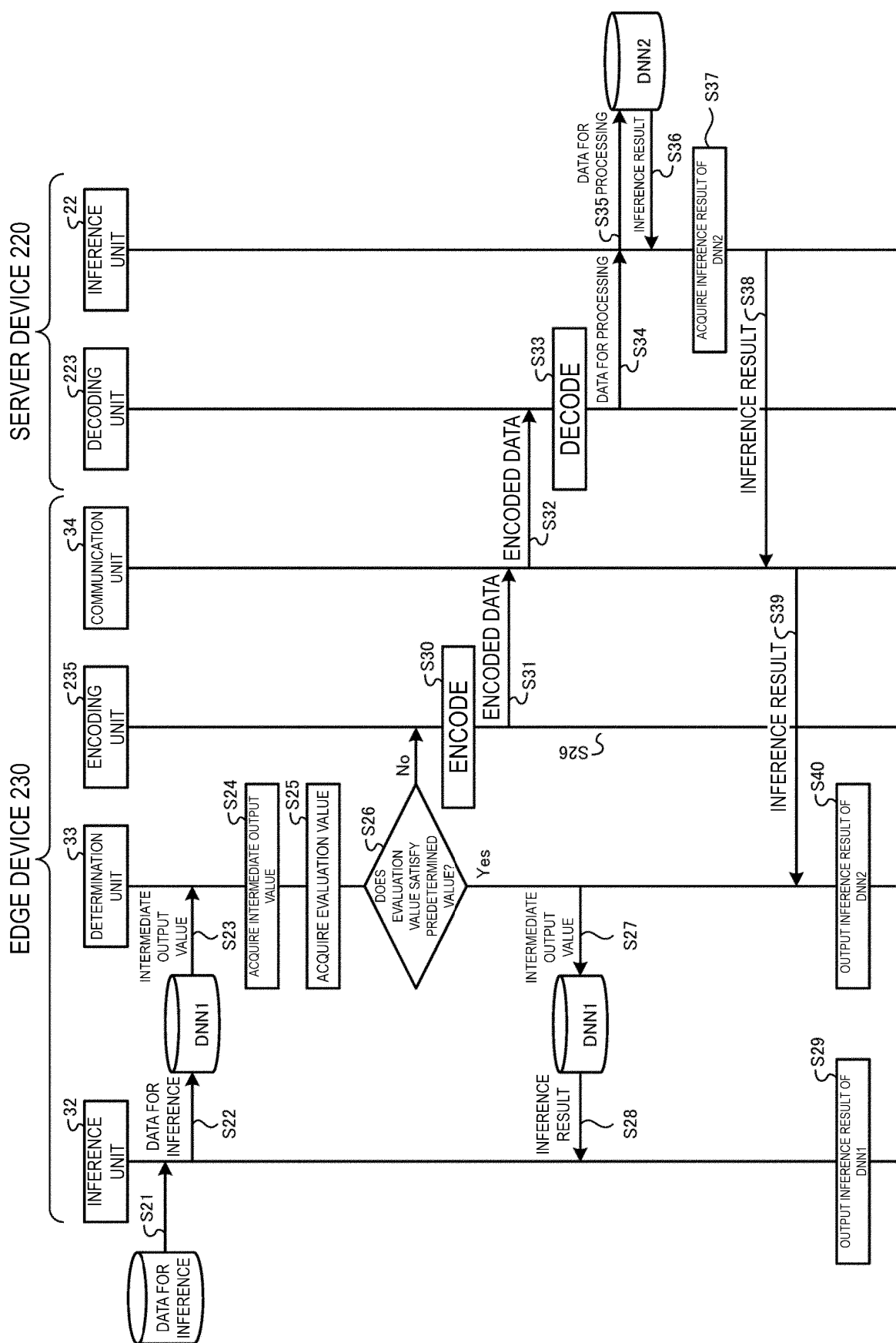


Fig. 11

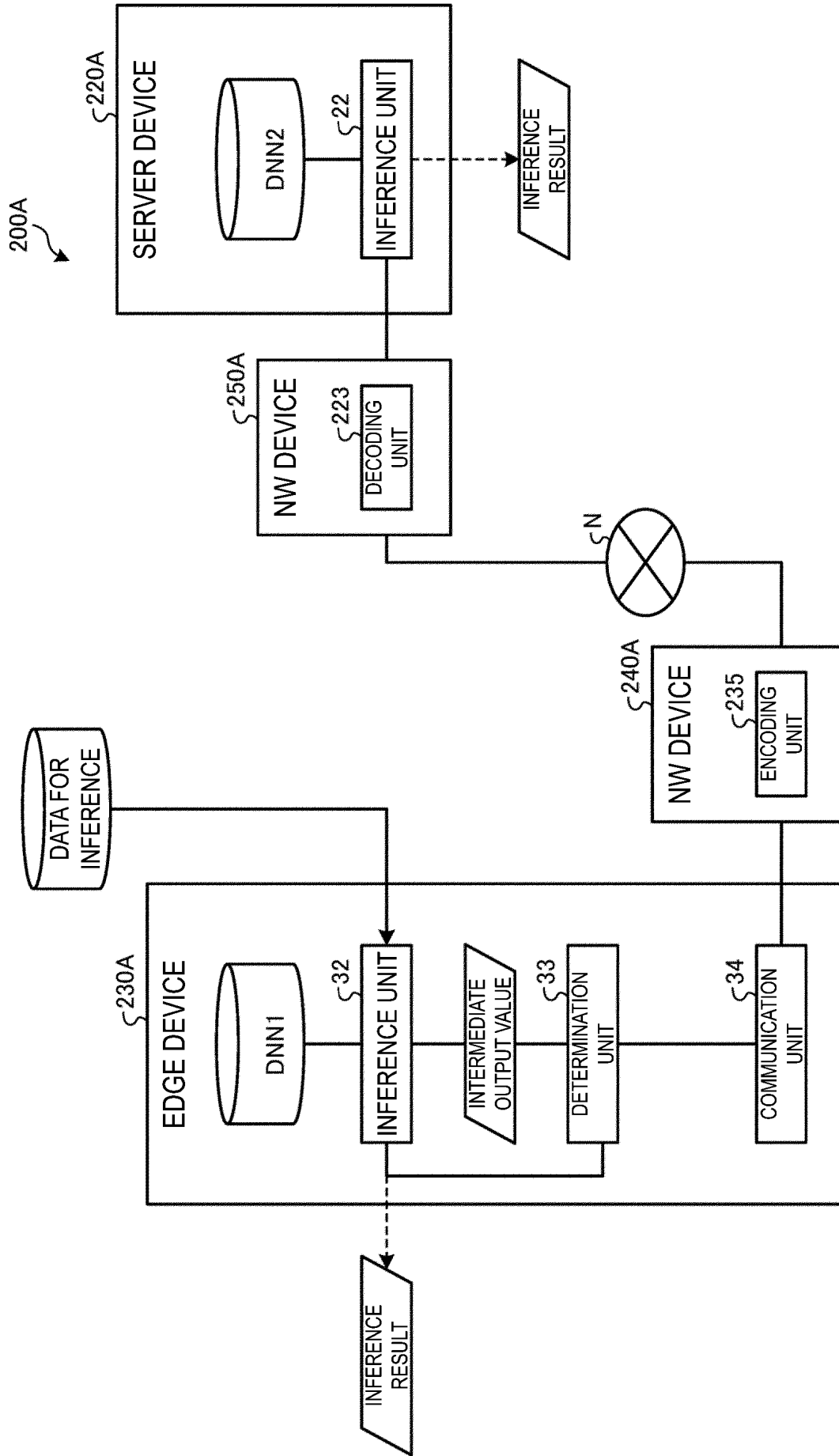


Fig. 12

Fig. 13

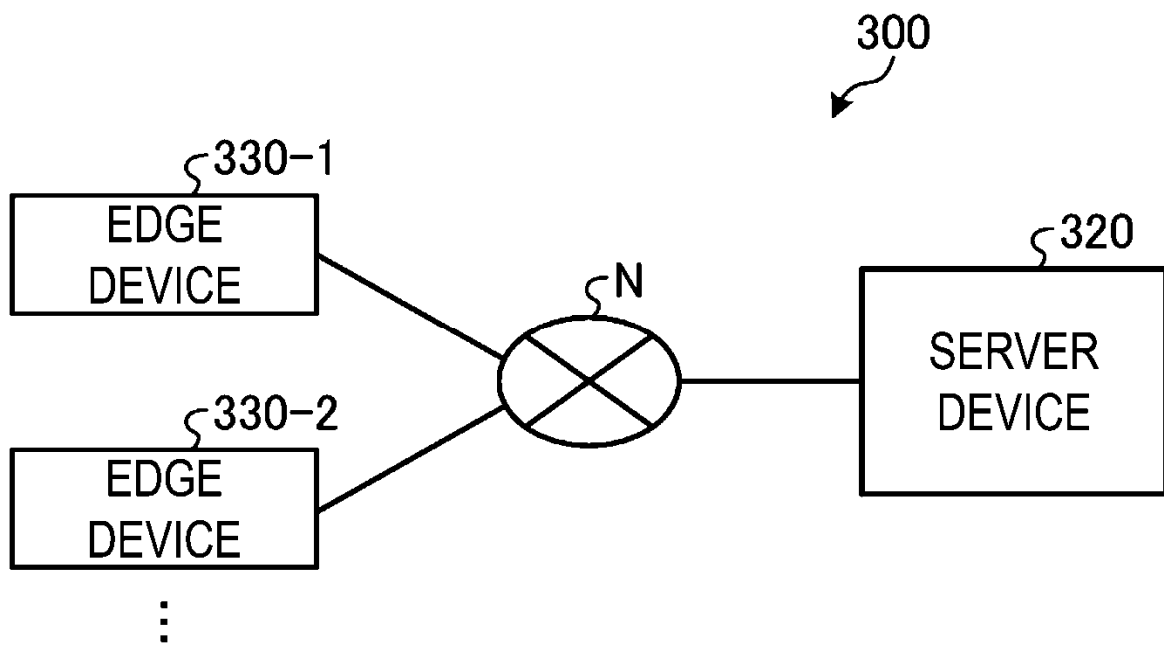


Fig. 14

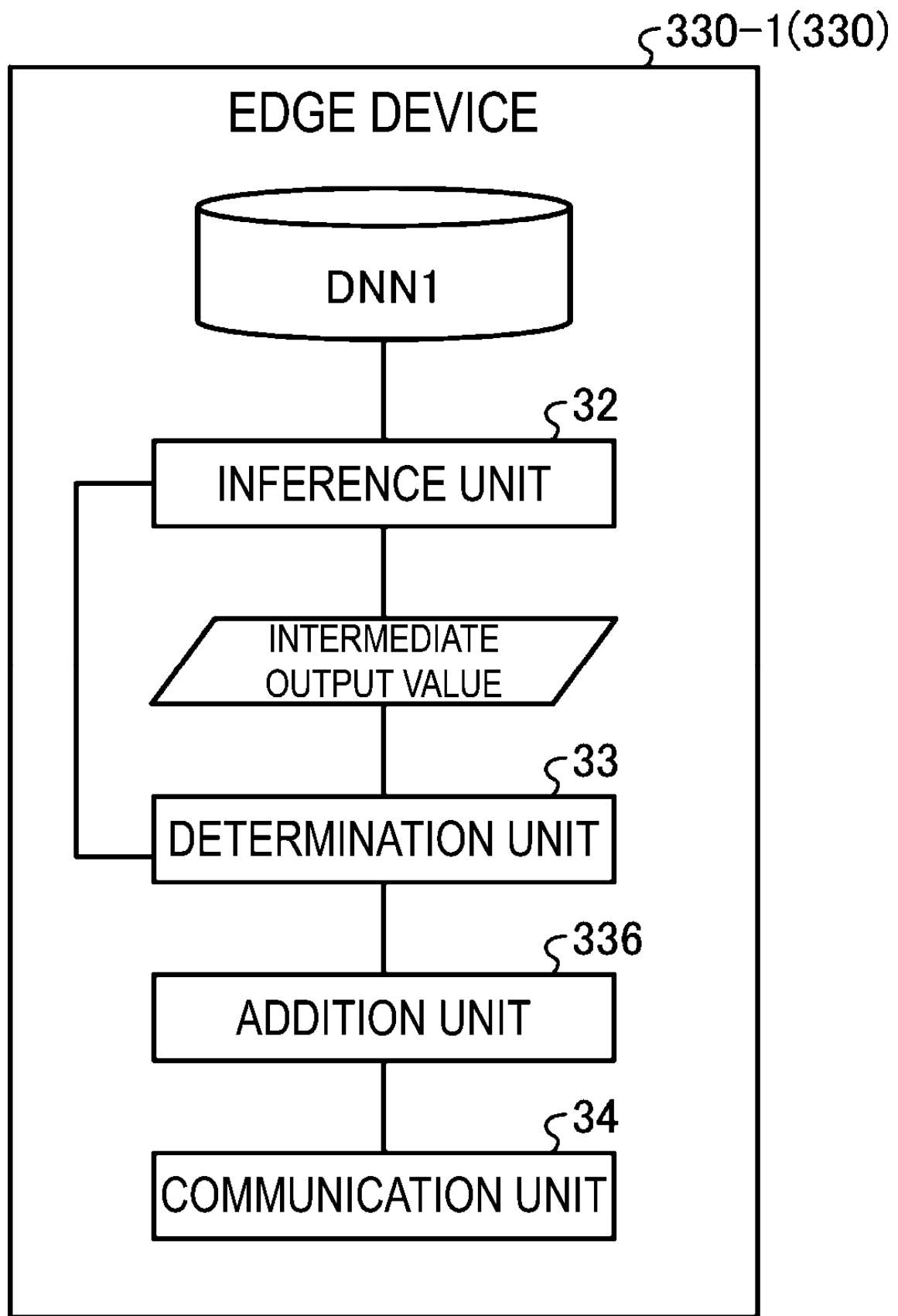
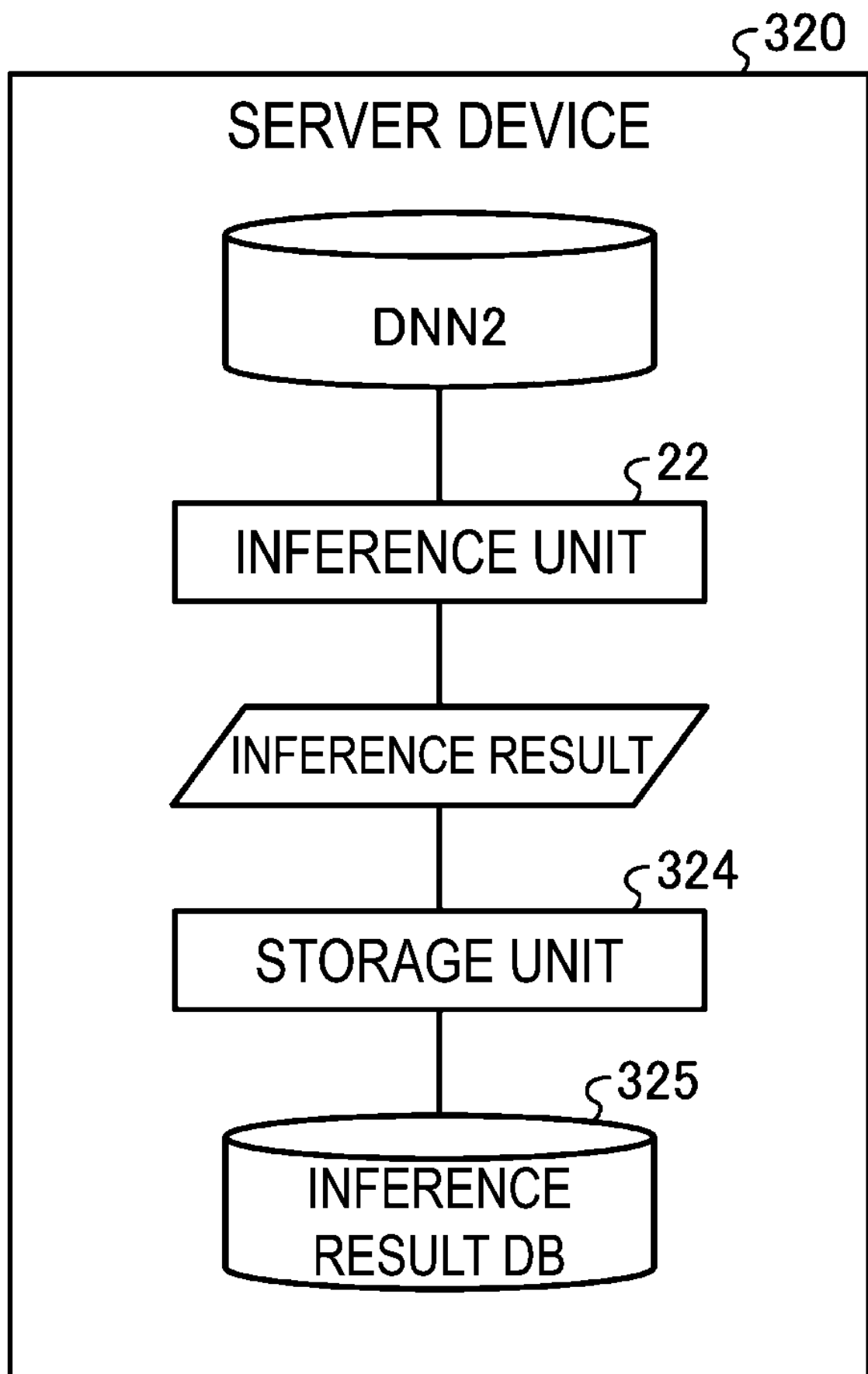


Fig. 15



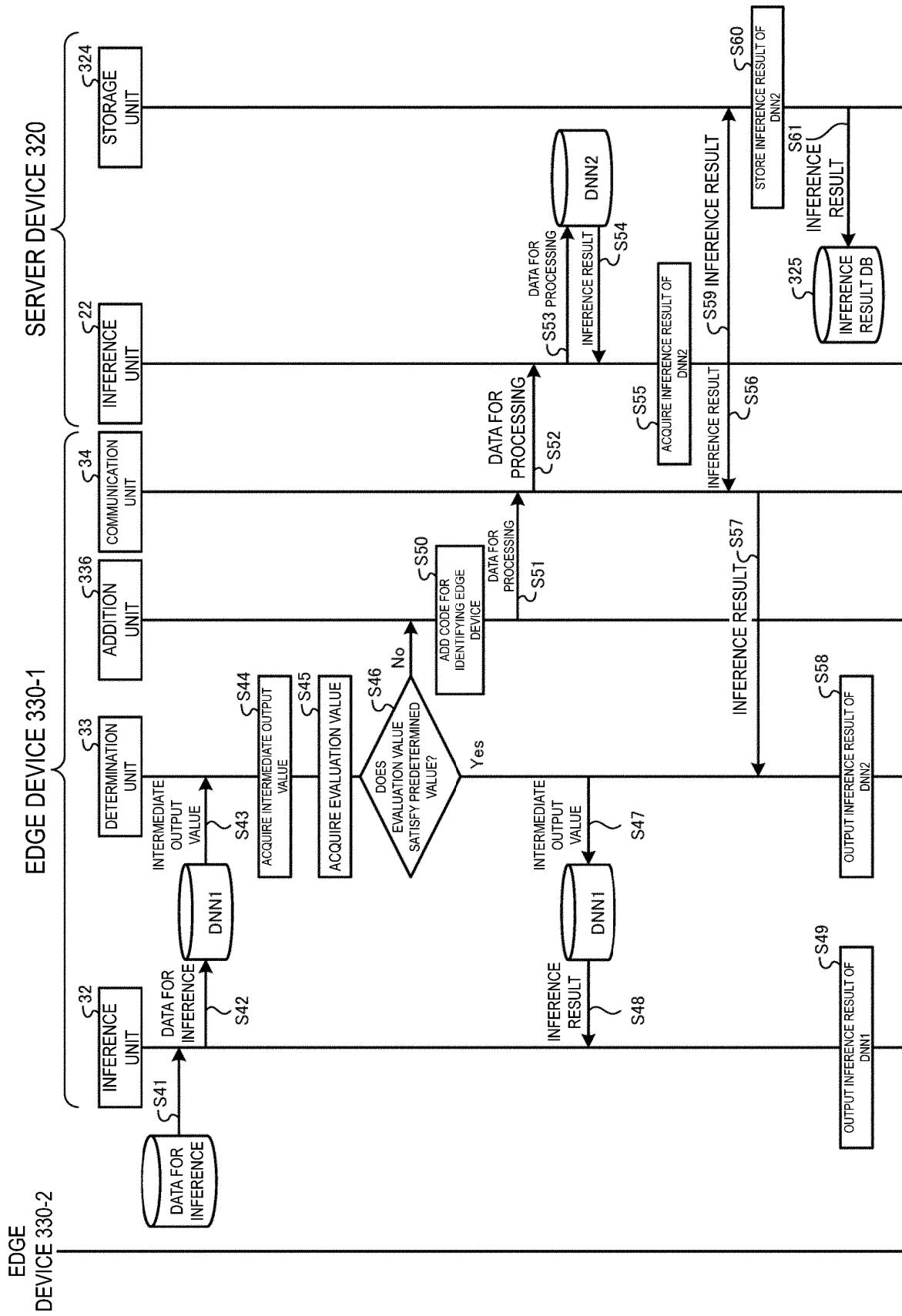


Fig. 16

Fig. 17

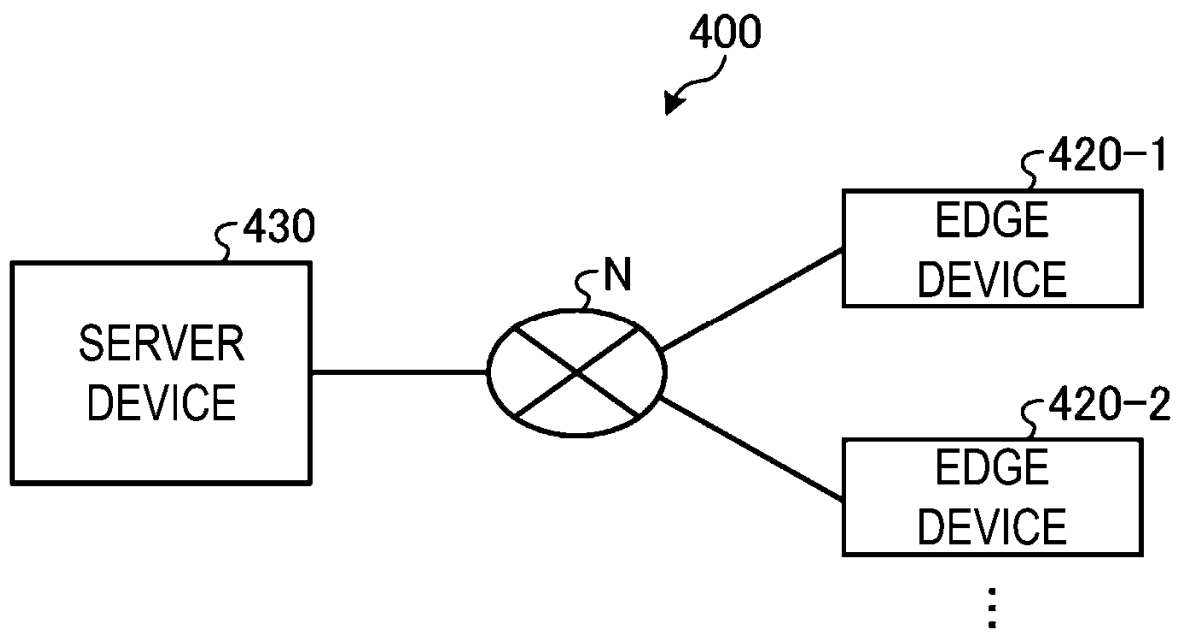


Fig. 18

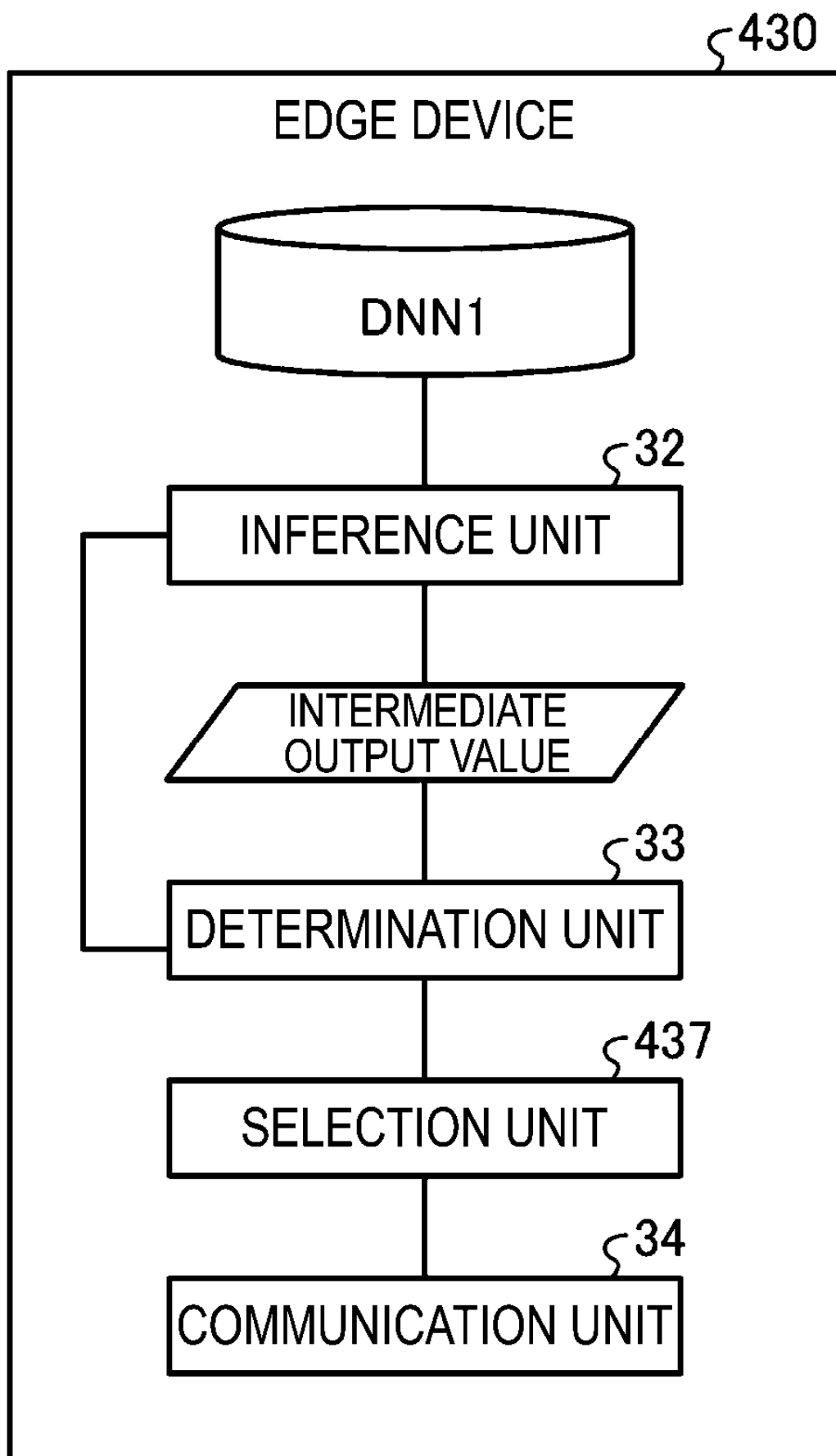


Fig. 19

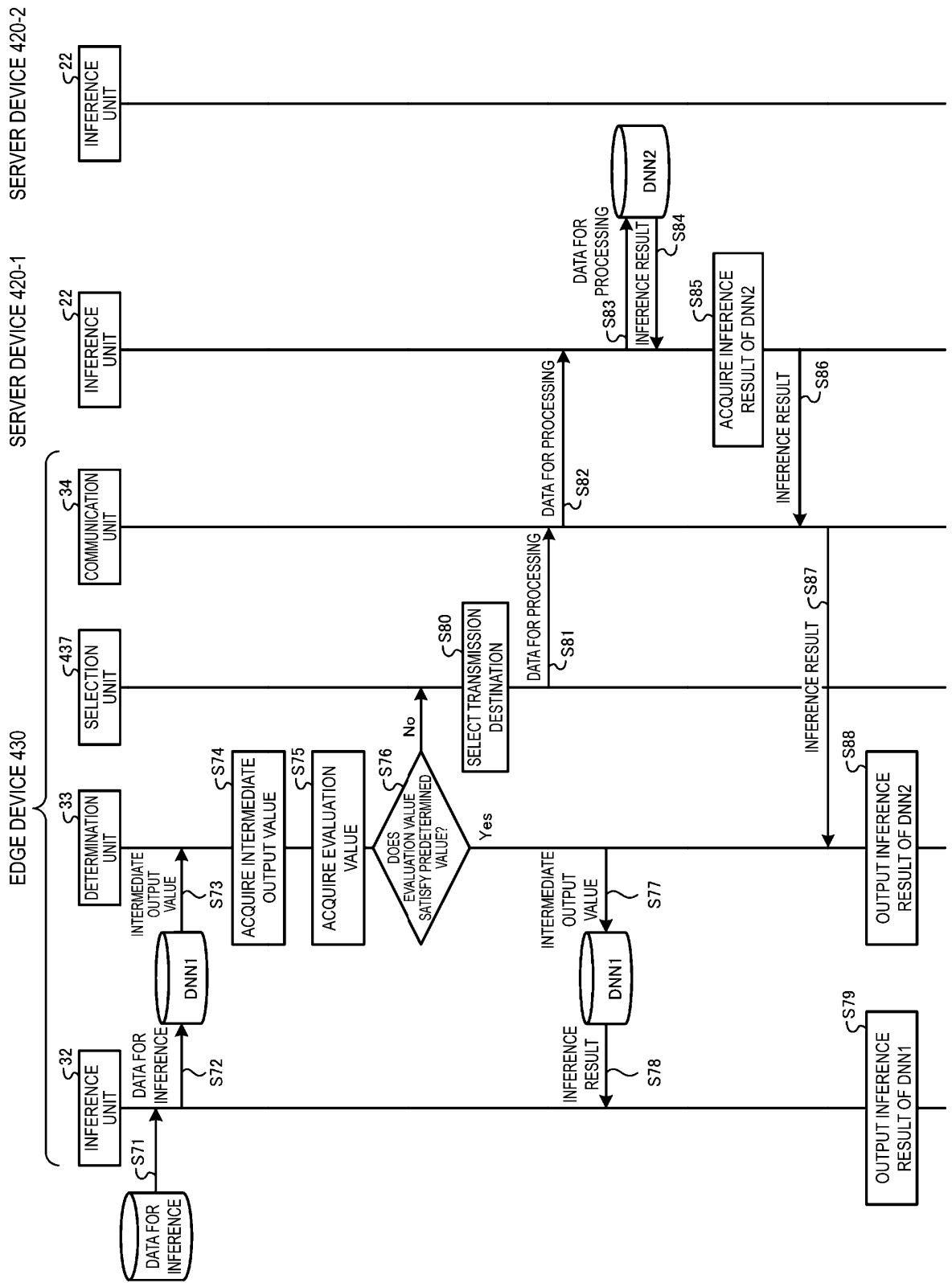


Fig. 20

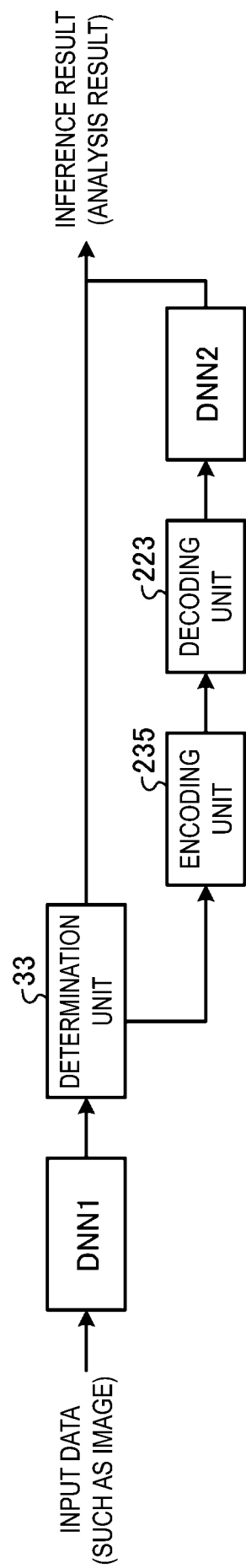


Fig. 21

DNN1, DNN2	DETERMINATION UNIT	ENCODING UNIT, DECODING UNIT, COMMUNICATION CONTENT
(1-A) WHEN LOAD OF DNN1 IS LOW, INFERENCE IS PERFORMED TO END IN DNN1 AS MUCH AS POSSIBLE (WHEN HW RESOURCES ARE INSUFFICIENT, SENDING TO DNN2 IS PERFORMED AND REST IS EXECUTED IN DNN2)	(2-A) DATA IS SENT TO DNN2 IN PARALLEL WITH EXECUTION OF DNN1, AND WHEN LOAD IS NOT HEAVY, PARALLEL PROCESSING IS ALSO PERFORMED IN DNN2	(3-A) ORIGINAL IMAGE OR PROCESSING RESULT OF DNN1 IN ADDITION TO ORIGINAL IMAGE IS TRANSMITTED AS IT IS
(1-B) ONE DNN IS DIVIDED AT SPECIFIC LAYER ACCORDING TO HW RESOURCES AND AMOUNT OF PROCESSING (INTO DNN1 AND DNN2). PROCESSING IS PERFORMED HALF-WAY IN DNN1, AND REST IS EXECUTED IN DNN2 EXAMPLE: DIVISION POINT IS CHANGED AT ANY TIME ACCORDING TO LOAD, SUCH AS FIFTH LAYER AT LOAD PEAK AND TWELFTH LAYER AT NORMAL TIME	(2-B) DATA IS SENT TO DNN2 IN PARALLEL WITH EXECUTION OF DNN1, AND DNN2 PERFORMS PREFERENTIAL CONTROL AS LOW-PRIORITY REQUEST	(3-B) ONLY NECESSARY INFORMATION IS TRANSMITTED TO DNN2 ACCORDING TO RESULT OF DNN1
(1-C) ONE DNN IS DIVIDED AT LAYER WITH LESS ENTROPY (INTO DNN1 AND DNN2). PROCESSING IS PERFORMED HALF-WAY IN DNN1, AND REST IS EXECUTED IN DNN2	(2-C) NECESSITY OF DNN2 PROCESSING IS DETERMINED BASED ON PROCESSING RESULT OF ENTIRE DNN1 OR INTERMEDIATE EXIT EXAMPLE: "RELIABILITY" IS USED AS PROCESSING RESULT	(3-C) ONE DNN IS DIVIDED AT SPECIFIC LAYER (INTO DNN1 AND DNN2), AND ONLY FEATURE OF DIVIDED LAYER IS SENT
(1-D) ONLY LOW-LOAD INFERENCE IS PERFORMED IN DNN1, AND HIGH-LOAD AND HIGH-PRECISION DNN IS PERFORMED IN DNN2	(2-D) INTERMEDIATE STATE OF DNN1 IS SENT TO DNN2	(3-D-1) DIVIDED LAYER IS QUANTIZED AND SENT TO DNN2, AND PROCESSING CONTINUES WITH RETURNING TO ORIGINAL ACCURACY AFTER RECEPTION IN DNN2
(1-E) NEXT REQUIRED DNN IS DETERMINED ACCORDING TO RESULT OF DNN1, AND SELECTED DNN2 IS EXECUTED	(2-E) DNN2 AS TRANSMISSION DESTINATION IS SELECTED BASED ON RESULT OF DNN1 EXAMPLE: OBJECT DETECTION AND PERSON DETECTION IN DNN1, SEX DETERMINATION AND VEHICLE DETECTION IN DNN2, AND NUMBER DETECTION IN DNN2	(3-D-2) DIVIDED LAYER IS QUANTIZED AND SENT TO DNN2, AND PROCESSING CONTINUES WITHOUT RETURNING TO ORIGINAL ACCURACY
(1-F) PROCESSING IS PERFORMED IN DNN1 WHEN QUICK RESPONSE IS REQUIRED ACCORDING TO REQUIRED LATENCY, AND OTHERWISE, PROCESSING IS PERFORMED IN DNN2 BY BATCH WITH OTHER EDGES EXAMPLE: EDGE DEVICE + SERVING	(2-F) DNN2 TO BE USED IS SELECTED ACCORDING TO SPATIAL DISTANCE TO DNN2 OR OPERATION/SYSTEM STATUS (FOR EXAMPLE, AVAILABILITY OF DNN2 AND AMOUNT OF WAITING TASKS) EXAMPLE: SERVING	(3-D-3) DIVIDED LAYER IS QUANTIZED AND SENT TO DNN2, AND PROCESSING CONTINUES WITH ACCURACY RETURNED IN INTERMEDIATE LAYER IN DNN2
(1-G) INTERMEDIATE EXIT IS INSERTED INTO DNN1 AND DNN2 SO THAT VARIOUS TASK DIFFICULTY LEVELS CAN BE HANDLED	(2-G) HIGHLY SPECIFIC (CONFIDENTIAL, PRIVATE, OR LIKE) INFORMATION IS PROCESSED BY DNN1, AND OTHER PROCESSED INFORMATION IS PROCESSED BY DNN2	(3-E) ENCODING MEANS IS DESIGNED, WHICH IS OPTIMIZED FOR OUTPUT TENDENCY (ACQUIRED DURING TRAINING OF DNN1) OF EACH COMPONENT (NODE) OF FEATURE MAP TO BE TAKEN OVER TO DNN2
(1-H) DNN1 PERFORMS TRAINING ENABLING ADVANCED ADAPTIVE SELECTION IN ADDITION TO NORMAL TRAINING		(3-F) ENTIRE FEATURE MAP TAKEN OVER TO DNN2 IS SET AS ONE SPACE, AND ENCODING MEANS IS DESIGNED USING RESULT OF CLUSTERING EACH MAP IN THIS SPACE

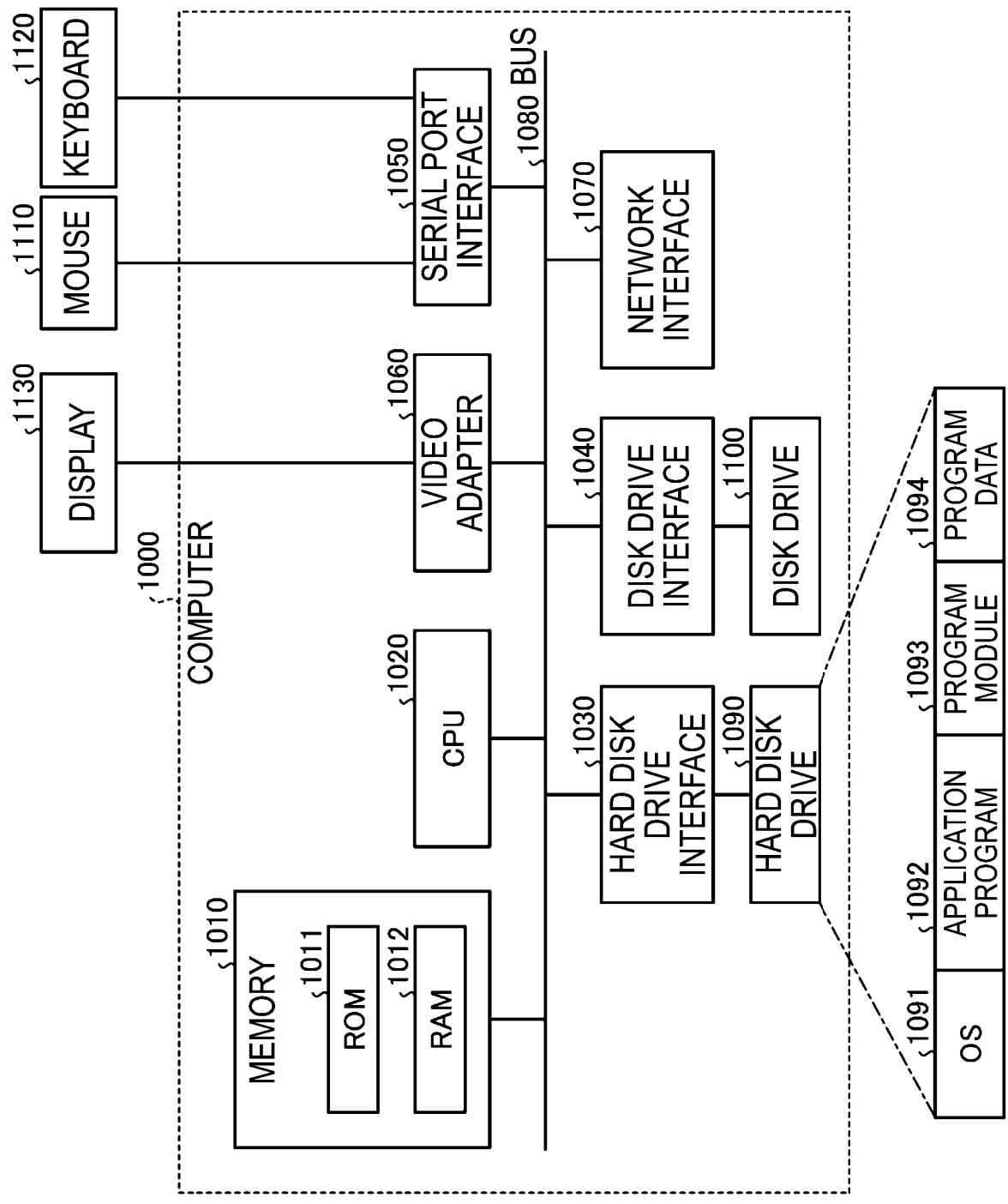


Fig. 22