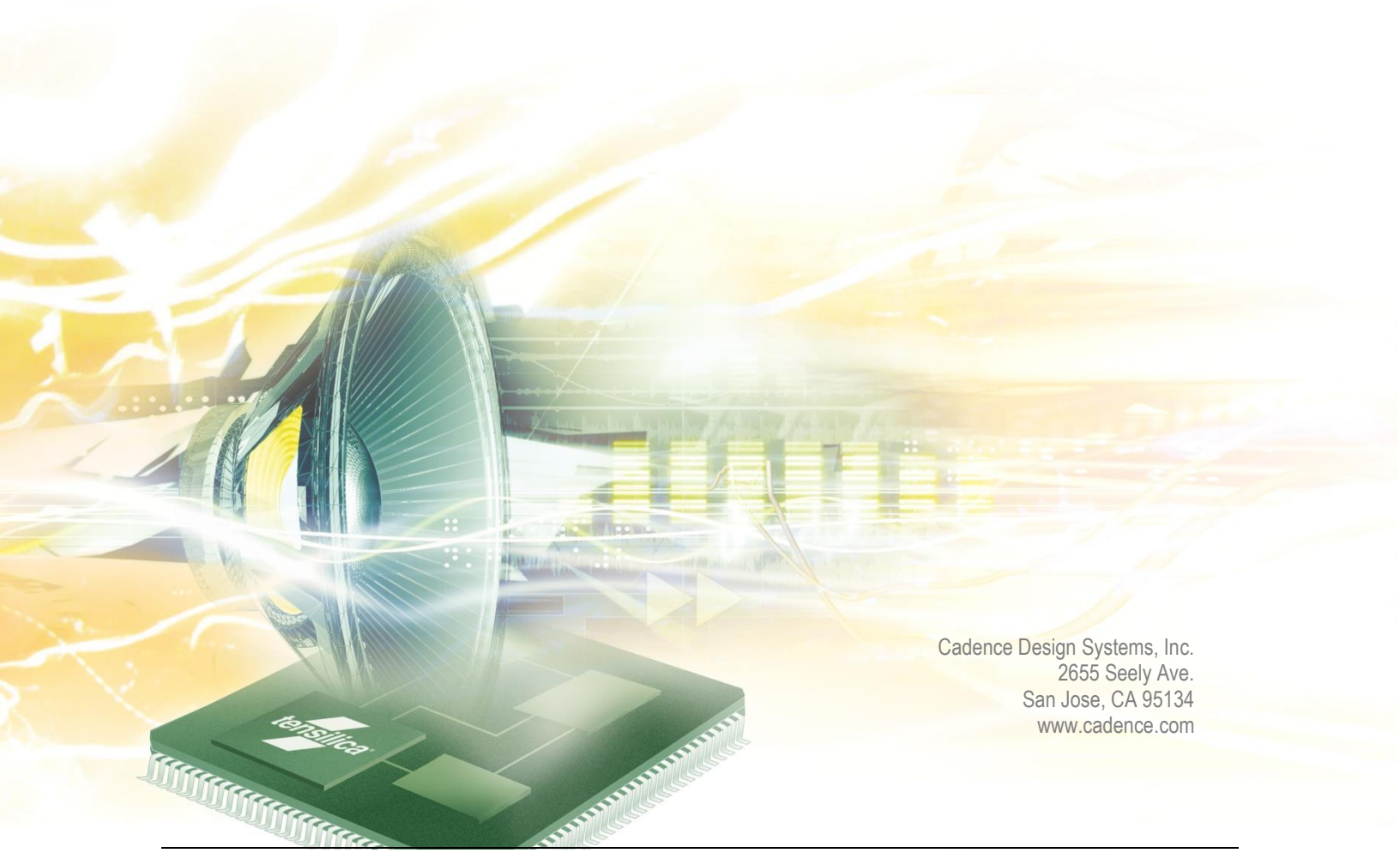
cadence®

Fusion G3 Neural Network Library

Performance report



Cadence Design Systems, Inc.
2655 Seely Ave.
San Jose, CA 95134
www.cadence.com

Version 1.0
November 2024

Contents`

Tables

Change History

Version	Changes
1.0	Initial version

1. Introduction

The Fusion G3 Neural Network (NN) Library is an optimized implementation of various low-level NN kernels. The low-level NN kernels are the basic building blocks for operators and networks in neural network frameworks with a generic and simple interface.

The Fusion G3 NN Library package includes the source code containing low-level kernel implementations. The current version of the library implements activation, basic operation, normalization and reorg functions as low-level kernels.

This document provides the code-size memory requirements, and timings (cycles) information for low-level NN kernels. The details of the APIs available in Fusion G3 NN Library can be found in FusionG3-NNLib-API.pdf.

Note	This version of the library supports Fusion G3 DSPs with the SP-VFPU (Single Precision Vector Floating Point Unit).
Note	This version of the Fusion G3 NN Library is tested with the xt-clang/xt-clang++ compilers using Xtensa Software Tools from RI-2022.10 release.

2. Fusion G3 NN Library Performance

The following table provides details of the library version, core information and build target used for producing the performance data.

Table 2-1 Details of Setup Used for Measurements

Library Name	FusionG3 Neural Network Library
Library Version	1.0
Library API Version	1.0
Core Name	Fusion G3
Tool Chain	RI-2022.10
Build Target	Release

The memory usage and performance figures are provided for all the kernels available in Fusion G3 NN library.

2.1 Memory Requirements

The NN library is provided as a single library archive. The Text and Read-Only Data (RODdata) sizes of the archive are shown in **Error! Reference source not found.**2.

Table 2-2 Library Text and ROData Sizes

DSP	Neural Network Library	
	Text (in Bytes)	Data (in Bytes)
Fusion G3 (with NN, SP-VFPU)	98484	300

The following table provides Kernel-level code sizes. The Kernel-level code size is defined as the code size taken by the kernel and its dependent functions.

Note The Kernel-level code sizes are independently calculated for each kernel. If two kernels share dependent functions, then the total code size required for including both kernels is less than the sum of individual sizes mentioned in the table below. This also explains why the sum of Text sizes in Table 2-3 is greater than the Text size of the entire library.

Table 2-3 Kernel Level Text Sizes

Kernel Name	Text (in Bytes))
xa_nn_vec_softmax_dim_f32_f32	7692
xa_nn_elm_add_32x32_32	504
xa_nn_elm_add_scalar_32x32_32	348
xa_nn_elm_add_broadcast_5D_32x32_32	7108
xa_nn_elm_add_f32xf32_f32	456
xa_nn_elm_add_scalar_f32xf32_f32	392
xa_nn_elm_add_broadcast_5D_f32xf32_f32	7373
xa_nn_elm_dequantize_asym4_f32	2042
xa_nn_elm_dequantize_asym4u_f32	2042
xa_nn_elm_dequantize_asym8_f32	1990
xa_nn_elm_dequantize_asym8u_f32	1990
xa_nn_elm_dequantize_asym16_f32	2016
xa_nn_elm_dequantize_asym16u_f32	1984
xa_nn_elm_dequantize_sym4_f32	1991
xa_nn_elm_dequantize_sym4u_f32	2039
xa_nn_elm_dequantize_sym8_f32	1943
xa_nn_elm_dequantize_sym8u_f32	1991
xa_nn_elm_dequantize_sym16_f32	1972
xa_nn_elm_dequantize_sym16u_f32	2020
xa_nn_elm_mul_scalar_32x32_32	392

xa_nn_elm_mul_32x32_32	332
xa_nn_elm_mul_broadcast_5D_32x32_32	6348
xa_nn_elm_mul_scalar_f32xf32_f32	380
xa_nn_elm_mul_f32xf32_f32	376
xa_nn_elm_mul_broadcast_5D_f32xf32_f32	5934
xa_nn_elm_quantize_f32_asym4	2957
xa_nn_elm_quantize_f32_asym4u	2941
xa_nn_elm_quantize_f32_asym8	2925
xa_nn_elm_quantize_f32_asym8u	2893
xa_nn_elm_quantize_f32_asym16	2925
xa_nn_elm_quantize_f32_asym16u	2877
xa_nn_elm_quantize_f32_sym4	2634
xa_nn_elm_quantize_f32_sym4u	2634
xa_nn_elm_quantize_f32_sym8	2614
xa_nn_elm_quantize_f32_sym8u	2614
xa_nn_elm_quantize_f32_sym16	2582
xa_nn_elm_quantize_f32_sym16u	2582
xa_nn_native_layer_norm_f32_f32	6982
xa_nn_cat	649

3. Timings – Low-level kernels

Following table provides cycles information for various low-level kernels for the given parameters. “Low-Level Kernel name” column specifies the name of the kernel, “Parameters” column specifies the parameters and shape of inputs used for performance measurement, “Average Cycles” column specifies the average cycles taken by the kernel when the kernel is fed with the parameters specified in “Parameters” column, “Performance Metric” column specifies the average cycles required for calculation of one output sample.

Table 3-1 Low-Level Kernels Timings

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_vec_softmax_dim_f32_f32	axis pointer = NULL input - 12,197,197	2270671	4.88
	axis pointer = NULL input - 1,3,172	2840	5.5
	axis = 0 input – 12,197,197	4424632	9.5
	axis = 0 input – 96,197,197	21928928	5.89
	axis = 2 input - 1,3,172	3409	6.61
	Axis = 3 Input - 1, 12, 197, 197	2988186	6.42
xa_nn_elm_add_32x32_32	Input1 - 1,40,14,14 Input2 - 1,40,14,14	3000	0.38
	Input1 - 1,1024,28,28 Input2 - 1,1024,28,28	301116	0.375
	Input1 - 1,3,64 Input2 - 1,3,64	132	0.69
xa_nn_elm_add_scalar_32x32_32	Input1 - 1,3,1 Input2 - 1	28	9.3
	Input1 - 1,288,14,14 Input2 - 1	14156	0.25
xa_nn_elm_add_broadcast_5D_32x32_32	Input1 - 1,8,512 Input2 - 1,1,512	1702	0.41
	Input1 - 1,8,3,3 Input2 - 3,3	627	8.7
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	9927	4.13
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	422504	0.40
xa_nn_elm_add_f32xf32_f32	Input1 - 1,40,14,14 Input2 - 1,40,14,14	2988	0.38
	Input1 - 1,1024,28,28 Input2 - 1,1024,28,28	301104	0.38
	Input1 - 1,3,64 Input2 - 1,3,64	120	0.63
xa_nn_elm_add_scalar_f32xf32_f32	Input1 - 1,3,1	32	10.7

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input2 - 1		
	Input1 - 1,288,14,14 Input2 - 1	14157	0.25
xa_nn_elm_add_broadcast_5D_f32xf32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1700	0.42
	Input1 - 1,8,3,3 Input2 - 3,3	596	8.3
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11603	4.84
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	401985	0.383
xa_nn_elm_mul_32x32_32	Input1 - 1,3,172 Input2 - 1,3,172	253	0.49
	Input1 - 1,96,14,14 Input2 - 1,96,14,14	7111	0.378
	Input1 - 1,64,224,224 Input2 - 1,64,224,224	1204535	0.375
xa_nn_elm_mul_scalar_32x32_32	Input1 - 1,64 Input2 - 1,64	65	1.01
	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37873	0.25
xa_nn_elm_mul_broadcast_5D_32x32_32	Input1 - 1,16,1,1 Input 2 - 1,16,56,56	60710	1.21
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	778	8.1
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11723	4.88
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	393768	0.376
xa_nn_elm_mul_f32xf32_f32	Input1 - 1,3,172 Input2 - 1,3,172	253	0.49
	Input1 - 1,96,14,14 Input2 - 1,96,14,14	7111	0.377
	Input1 - 1,64,224,224 Input2 - 1,64,224,224	1204535	0.375
xa_nn_elm_mul_scalar_f32xf32_f32	Input1 - 1,64 Input2 - 1,64	67	1.04
	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37873	0.25
xa_nn_elm_mul_broadcast_5D_f32xf32_f32	Input1 - 1,16,1,1 Input2 - 1,16,56,56	60710	1.2
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	778	8.1
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11723	4.88
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	418476	0.4
	Axis = 0 Input - 2, 197, 768	227413	0.75

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_native_layer_norm_f32_f32 – High performance	Axis = 2 Input - 1, 197, 768	126238	0.834
xa_nn_native_layer_norm_f32_f32 – High precision	Axis = 0 Input - 2, 197, 768	756674	2.5
	Axis = 2 Input - 1, 197, 768	617092	4.07
xa_nn_cat –8-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	9649	0.064
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	638	3.33
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	9587	0.063
xa_nn_cat – 16-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	19059	0.13
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	638	3.33
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	19104	0.13
xa_nn_cat – 32-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	37875	0.25
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	638	3.33
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	37811	0.25
xa_nn_elm_quantize _f32_asym4	Input - 1,40,14,14	6072	0.77
	Input - 1,3,1	137	45.67
xa_nn_elm_quantize _f32_asym4u	Input - 1,40,14,14	6072	0.77
	Input - 1,3,1	136	45.67
xa_nn_elm_quantize _f32_asym4 – channel based	Channel = 3 1,3,8,4	862	8.98

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Channel = 1 Input - 1,12,64,197	115091	0.76
xa_nn_elm_quantize _f32_asym4u – channel based	Channel = 3 Input - 1,3,8,4	860	8.98
	Channel = 1 Input - 1,12,64,197	115079	0.76
xa_nn_elm_quantize _f32_asym8	Input - 1,40,14,14	6070	0.77
	Input - 1,3,1	135	45
xa_nn_elm_quantize _f32_asym8u	Input - 1,40,14,14	6068	0.77
	Input - 1,3,1	133	45.67
xa_nn_elm_quantize _f32_asym8 – channel based	Channel = 3 1,3,8,4	881	9.2
	Channel = 1 Input - 1,12,64,197	115122	0.76
xa_nn_elm_quantize _f32_asym8u – channel based	Channel = 3 Input - 1,3,8,4	873	9.1
	Channel = 1 Input - 1,12,64,197	115116	0.76
xa_nn_elm_quantize _f32_asym16	Input - 1,40,14,14	6068	0.77
	Input - 1,3,1	137	45.67
xa_nn_elm_quantize _f32_asym16u	Input - 1,40,14,14	6066	0.77
	Input - 1,3,1	137	45.67
xa_nn_elm_quantize _f32_asym16 – channel based	Channel = 3 1,3,8,4	865	9.01
	Channel = 1 Input - 1,12,64,197	115056	0.76
xa_nn_elm_quantize _f32_asym16u – channel based	Channel = 3 Input - 1,3,8,4	883	9.2
	Channel = 1 Input - 1,12,64,197	115061	0.76
xa_nn_elm_quantize _f32_sym4	Input - 1,40,14,14	5071	0.65
	Input - 1,3,1	130	43.33
xa_nn_elm_quantize _f32_sym4u	Input - 1,40,14,14	5071	0.65
	Input - 1,3,1	130	43.33
xa_nn_elm_quantize _f32_sym4 – channel based	Channel = 3 1,3,8,4	742	7.73
	Channel = 1 Input - 1,12,64,197	96000	0.64
xa_nn_elm_quantize _f32_sym4u – channel based	Channel = 3 Input - 1,3,8,4	741	7.72
	Channel = 1 Input - 1,12,64,197	95997	0.64
xa_nn_elm_quantize _f32_sym8	Input - 1,40,14,14	5070	0.65
	Input - 1,3,1	129	43

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_quantize _f32_sym8u	Input - 1,40,14,14	5070	0.65
	Input - 1,3,1	127	42.33
xa_nn_elm_quantize _f32_sym8 – channel based	Channel = 3 1,3,8,4	735	7.66
	Channel = 1 Input - 1,12,64,197	95999	0.64
xa_nn_elm_quantize _f32_sym8u – channel based	Channel = 3 Input - 1,3,8,4	734	7.66
	Channel = 1 Input - 1,12,64,197	95996	0.64
xa_nn_elm_quantize _f32_sym16	Input - 1,40,14,14	5070	0.65
	Input - 1,3,1	127	42.33
xa_nn_elm_quantize _f32_sym16u	Input - 1,40,14,14	5071	0.65
	Input - 1,3,1	129	43
xa_nn_elm_quantize _f32_sym16 – channel based	Channel = 3 1,3,8,4	737	7.66
	Channel = 1 Input - 1,12,64,197	96005	0.64
xa_nn_elm_quantize _f32_sym16u – channel based	Channel = 3 Input - 1,3,8,4	739	7.66
	Channel = 1 Input - 1,12,64,197	95995	0.64
xa_nn_elm_dequantize_asym4_f32	Input - 1,40,14,14	3067	0.39
	Input - 1,3,1	98	32.67
xa_nn_elm_dequantize_asym4u_f32	Input - 1,40,14,14	3065	0.39
	Input - 1,3,1	97	32.22
xa_nn_elm_dequantize_asym4_f32 – channel based	Channel = 3 1,3,8,4	586	6.11
	Channel = 1 Input - 1,12,64,197	57716	0.38
xa_nn_elm_dequantize_asym4u_f32 – channel based	Channel = 3 Input - 1,3,8,4	585	6.11
	Channel = 1 Input - 1,12,64,197	57695	0.38
xa_nn_elm_dequantize_asym8_f32	Input - 1,40,14,14	3063	0.39
	Input - 1,3,1	92	30.67
xa_nn_elm_dequantize_asym8u_f32	Input - 1,40,14,14	3061	0.39
	Input - 1,3,1	91	30.33
xa_nn_elm_dequantize_asym8_f32 – channel based	Channel = 3 1,3,8,4	569	6.11
	Channel = 1 Input - 1,12,64,197	57679	0.38

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_dequantize_asym8u_f32 – channel based	Channel = 3 Input - 1,3,8,4	568	6.11
	Channel = 1 Input - 1,12,64,197	57658	0.38
xa_nn_elm_dequantize_asym16_f32	Input - 1,40,14,14	3062	0.39
	Input - 1,3,1	92	30.67
xa_nn_elm_dequantize_asym16u_f32	Input - 1,40,14,14	3059	0.39
	Input - 1,3,1	91	30.33
xa_nn_elm_dequantize_asym16_f32 – channel based	Channel = 3 1,3,8,4	568	6.11
	Channel = 1 Input - 1,12,64,197	57666	0.38
xa_nn_elm_dequantize_asym16u_f32 – channel based	Channel = 3 Input - 1,3,8,4	567	6.11
	Channel = 1 Input - 1,12,64,197	57633	0.38
xa_nn_elm_dequantize_sym4_f32	Input - 1,40,14,14	2074	0.27
	Input - 1,3,1	90	30
xa_nn_elm_dequantize_sym4u_f32	Input - 1,40,14,14	2077	0.27
	Input - 1,3,1	90	30
xa_nn_elm_dequantize_sym4_f32 – channel based	Channel = 3 1,3,8,4	523	5.45
	Channel = 1 Input - 1,12,64,197	38665	0.26
xa_nn_elm_dequantize_sym4u_f32 – channel based	Channel = 3 Input - 1,3,8,4	523	5.45
	Channel = 1 Input - 1,12,64,197	38701	0.26
xa_nn_elm_dequantize_sym8_f32	Input - 1,40,14,14	2069	0.27
	Input - 1,3,1	84	28
xa_nn_elm_dequantize_sym8u_f32	Input - 1,40,14,14	2072	0.27
	Input - 1,3,1	84	28
xa_nn_elm_dequantize_sym8_f32 – channel based	Channel = 3 1,3,8,4	510	5.3
	Channel = 1 Input - 1,12,64,197	38616	0.26
xa_nn_elm_dequantize_sym8u_f32 – channel based	Channel = 3 Input - 1,3,8,4	506	5.3
	Channel = 1 Input - 1,12,64,197	38652	0.26
xa_nn_elm_dequantize_sym16_f32	Input - 1,40,14,14	2072	0.27
	Input - 1,3,1	83	27.67
xa_nn_elm_dequantize_sym16u_f32	Input - 1,40,14,14	2075	0.27

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input - 1,3,1	83	27.67
xa_nn_elm_dequantize_sym16_f32 – channel based	Channel = 3 1,3,8,4	512	5.33
	Channel = 1 Input - 1,12,64,197	38651	0.26
xa_nn_elm_dequantize_sym16u_f32 – channel based	Channel = 3 Input - 1,3,8,4	512	5.33
	Channel = 1 Input - 1,12,64,197	38687	0.26

4. References

[1] FusionG3-NNLib-API.pdf