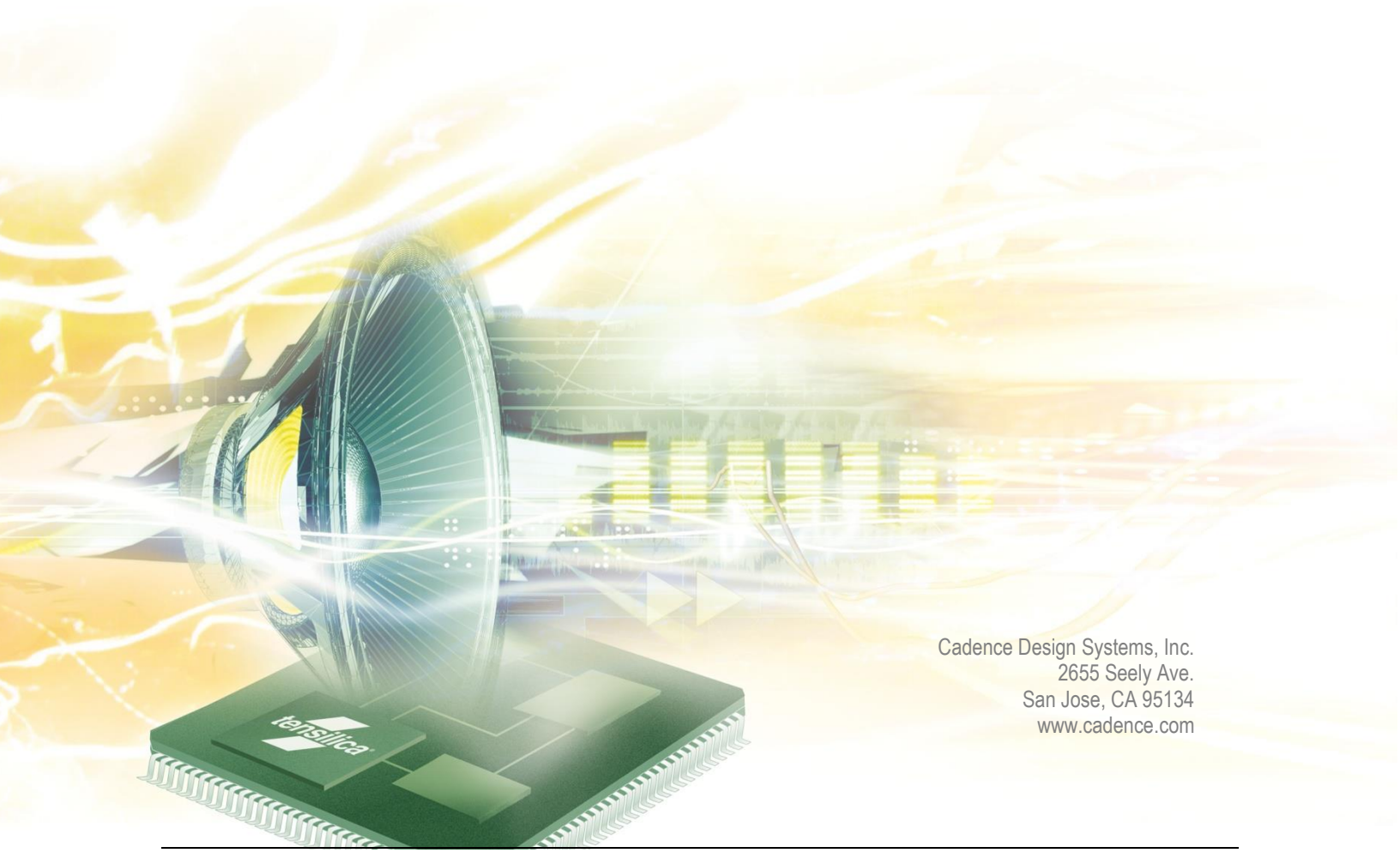
cadence®

Fusion G3 Neural Network Library

Performance report



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

Version 1.1
December 2024

Contents`

1. Introduction 1
2. Fusion G3 NN Library Performance 2
2.1 Memory Requirements 3
3. Timings – Low-level kernels 6
4. References 20

Tables

Table 2-1 Details of Setup Used for Measurements 2
Table 2-2 Library Text and RODataSizes 3
Table 2-3 Kernel Level Text Sizes 3
Table 3-1 Low-Level Kernels Timings 6

Change History

Version	Changes
1.0	Initial version
1.1	Added performance cycles for div, sub, exp, slice, permute, mean kernels

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	1.1
Library API Version	1.1
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)	242467	916

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	7679
xa_nn_elm_add_32x32_32	500
xa_nn_elm_add_scalar_32x32_32	344
xa_nn_elm_add_broadcast_5D_32x32_32	6784
xa_nn_elm_add_f32xf32_f32	448
xa_nn_elm_add_scalar_f32xf32_f32	396
xa_nn_elm_add_broadcast_5D_f32xf32_f32	7146
xa_nn_elm_dequantize_asym4_f32	2779
xa_nn_elm_dequantize_asym4u_f32	2763
xa_nn_elm_dequantize_asym8_f32	2763
xa_nn_elm_dequantize_asym8u_f32	2759
xa_nn_elm_dequantize_asym16_f32	2827
xa_nn_elm_dequantize_asym16u_f32	2875

xa_nn_elm_dequantize_sym4_f32	2944
xa_nn_elm_dequantize_sym4u_f32	3074
xa_nn_elm_dequantize_sym8_f32	2944
xa_nn_elm_dequantize_sym8u_f32	3074
xa_nn_elm_dequantize_sym16_f32	2976
xa_nn_elm_dequantize_sym16u_f32	3090
xa_nn_elm_mul_scalar_32x32_32	384
xa_nn_elm_mul_32x32_32	336
xa_nn_elm_mul_broadcast_5D_32x32_32	6064
xa_nn_elm_mul_scalar_f32xf32_f32	376
xa_nn_elm_mul_f32xf32_f32	372
xa_nn_elm_mul_broadcast_5D_f32xf32_f32	5590
xa_nn_elm_quantize_f32_asym4	3682
xa_nn_elm_quantize_f32_asym4u	3666
xa_nn_elm_quantize_f32_asym8	3682
xa_nn_elm_quantize_f32_asym8u	3666
xa_nn_elm_quantize_f32_asym16	3638
xa_nn_elm_quantize_f32_asym16u	3606
xa_nn_elm_quantize_f32_sym4	3550
xa_nn_elm_quantize_f32_sym4u	3534
xa_nn_elm_quantize_f32_sym8	3550
xa_nn_elm_quantize_f32_sym8u	3534
xa_nn_elm_quantize_f32_sym16	3468
xa_nn_elm_quantize_f32_sym16u	3468
xa_nn_native_layer_norm_f32_f32	7126
xa_nn_cat	1005

xa_nn_elm_sub_32x32_32	504
xa_nn_elm_sub_scalar_32x32_32	348
xa_nn_elm_sub_broadcast_5D_32x32_32	6707
xa_nn_elm_sub_f32xf32_f32	456
xa_nn_elm_sub_scalar_f32xf32_f32	392
xa_nn_elm_sub_broadcast_5D_f32xf32_f32	7150
xa_nn_elm_div_f32xf32_f32	4964
xa_nn_elm_div_32x32_32	4192
xa_nn_elm_div_32x32_f32	1720
xa_nn_elm_div_scalar_f32xf32_f32	1572
xa_nn_elm_div_scalar_32x32_32	1168
xa_nn_elm_div_scalar_32x32_f32	840
xa_nn_elm_div_broadcast_5D_f32xf32_f32	39317
xa_nn_elm_div_broadcast_5D_32x32_32	34108
xa_nn_elm_div_broadcast_5D_32x32_f32	17497
xa_nn_elm_exp_f32_f32	1217
xa_nn_slice	1390
xa_nn_permute	5247

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		
	axis pointer = NULL input - 1,3,172		
	axis = 0 input – 12,197,197	4366471	9.37598452
	axis = 0 input – 96,197,197	21890173	5.87550917
	axis = 2 input - 1,3,172	3455	6.69573643
	Axis = 3 Input - 1, 12, 197, 197	2974070	6.38612607
xa_nn_elm_add_32x32_32	Input1 - 1,40,14,14 Input2 - 1,40,14,14	3060	0.390306
	Input1 - 1,1024,28,28 Input2 - 1,1024,28,28	301176	0.375149
	Input1 - 1,3,64 Input2 - 1,3,64	192	1
xa_nn_elm_add_scalar_32x32_32	Input1 - 1,3,1 Input2 - 1	84	28
	Input1 - 1,288,14,14 Input2 - 1	14211	0.251754
xa_nn_elm_add_broadcast_5D_32x32_32	Input1 - 1,8,512 Input2 - 1,1,512	1811	0.442139
	Input1 - 1,8,3,3 Input2 - 3,3	943	13.09722
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	10295	4.289583
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	578148	0.551365
xa_nn_elm_add_f32xf32_f32	Input1 - 1,40,14,14 Input2 - 1,40,14,14	3047	0.388648
	Input1 - 1,1024,28,28 Input2 - 1,1024,28,28	301163	0.375133
	Input1 - 1,3,64 Input2 - 1,3,64	179	0.932292
xa_nn_elm_add_scalar_f32xf32_f32	Input1 - 1,3,1	85	28.33333

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input2 - 1		
	Input1 - 1,288,14,14 Input2 - 1	14213	0.251789
xa_nn_elm_add_broadcast_5D_f32xf32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1791	0.437256
	Input1 - 1,8,3,3 Input2 - 3,3	887	12.31944
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11736	4.89
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	410214	0.391211
xa_nn_elm_mul_32x32_32	Input1 - 1,3,172 Input2 - 1,3,172	318	0.616279
	Input1 - 1,96,14,14 Input2 - 1,96,14,14	7176	0.381378
	Input1 - 1,64,224,224 Input2 - 1,64,224,224	1204344	0.375037
xa_nn_elm_mul_scalar_32x32_32	Input1 - 1,64 Input2 - 1,64	119	1.859375
	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37927	0.250681
xa_nn_elm_mul_broadcast_5D_32x32_32	Input1 - 1,16,1,1 Input 2 - 1,16,56,56	62546	1.246532
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	801	8.34375
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11764	4.901667
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	422610	0.403032
xa_nn_elm_mul_f32xf32_f32	Input1 - 1,3,172 Input2 - 1,3,172	313	0.606589
	Input1 - 1,96,14,14 Input2 - 1,96,14,14	7168	0.380952
	Input1 - 1,64,224,224 Input2 - 1,64,224,224	1204336	0.375035
xa_nn_elm_mul_scalar_f32xf32_f32	Input1 - 1,64 Input2 - 1,64	121	1.890625
	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37923	0.250654
xa_nn_elm_mul_broadcast_5D_f32xf32_f32	Input1 - 1,16,1,1 Input2 - 1,16,56,56	66075	1.316865
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	809	8.427083
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	12642	5.2675
	Input1 - 1,16,1,1 Input2 - 1,16,256,256	389723	0.371669
	Axis = 0 Input - 2, 197, 768	227364	0.751388

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	126164	0.833889
xa_nn_native_layer_norm_f32_f32 – High precision	Axis = 0 Input - 2, 197, 768	756931	2.501490456
	Axis = 2 Input - 1, 197, 768	621881	4.110359824
xa_nn_cat –8-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	9950	0.066100659
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	957	4.984375
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	9848	0.065090948
xa_nn_cat – 16-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	19358	0.128600659
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	957	4.984375
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	19304	0.127590948
xa_nn_cat – 32-bit	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	38174	0.25360066
	Axis = 4 Input1 - 1, 3, 8, 4, 1 Input2 - 1, 3, 8, 4, 1	957	4.984375
	Axis = 1 Input1 - 1, 1, 768 Input2 - 1, 196, 768,	38216	0.25259095
xa_nn_elm_quantize _f32_asym4	Input - 1,40,14,14	6264	0.79898
	Input - 1,3,1	320	106.6667
xa_nn_elm_quantize _f32_asym4u	Input - 1,40,14,14	6261	0.798597
	Input - 1,3,1	319	106.3333

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_quantize _f32_asym4 – channel based	Channel = 3 1,3,8,4	1201	12.51042
	Channel = 1 Input - 1,12,64,197	116132	0.767581
xa_nn_elm_quantize _f32_asym4u – channel based	Channel = 3 Input - 1,3,8,4	1189	12.38542
	Channel = 1 Input - 1,12,64,197	116098	0.767357
xa_nn_elm_quantize _f32_asym8	Input - 1,40,14,14	6266	0.799235
	Input - 1,3,1	323	107.6667
xa_nn_elm_quantize _f32_asym8u	Input - 1,40,14,14	6261	0.798597
	Input - 1,3,1	321	107
xa_nn_elm_quantize _f32_asym8 – channel based	Channel = 3 1,3,8,4	1202	12.52083
	Channel = 1 Input - 1,12,64,197	116136	0.767608
xa_nn_elm_quantize _f32_asym8u – channel based	Channel = 3 Input - 1,3,8,4	1200	12.5
	Channel = 1 Input - 1,12,64,197	116121	0.767509
xa_nn_elm_quantize _f32_asym16	Input - 1,40,14,14	6264	0.79898
	Input - 1,3,1	330	110
xa_nn_elm_quantize _f32_asym16u	Input - 1,40,14,14	6261	0.798597
	Input - 1,3,1	324	108
xa_nn_elm_quantize _f32_asym16 – channel based	Channel = 3 1,3,8,4	1215	12.65625
	Channel = 1 Input - 1,12,64,197	116108	0.767423
xa_nn_elm_quantize _f32_asym16u – channel based	Channel = 3 Input - 1,3,8,4	1204	12.54167
	Channel = 1 Input - 1,12,64,197	116078	0.767225
xa_nn_elm_quantize _f32_sym4	Input - 1,40,14,14	5271	0.672321
	Input - 1,3,1	310	103.3333
xa_nn_elm_quantize _f32_sym4u	Input - 1,40,14,14	5265	0.671556
	Input - 1,3,1	307	102.3333

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_quantize _f32_sym4 – channel based	Channel = 3 1,3,8,4	1183	12.32292
	Channel = 1 Input - 1,12,64,197	97132	0.642
xa_nn_elm_quantize _f32_sym4u – channel based	Channel = 3 Input - 1,3,8,4	1174	12.22917
	Channel = 1 Input - 1,12,64,197	97110	0.641854
xa_nn_elm_quantize _f32_sym8	Input - 1,40,14,14	5272	0.672449
	Input - 1,3,1	309	103
xa_nn_elm_quantize _f32_sym8u	Input - 1,40,14,14	5265	0.671556
	Input - 1,3,1	307	102.3333
xa_nn_elm_quantize _f32_sym8 – channel based	Channel = 3 1,3,8,4	1184	12.33333
	Channel = 1 Input - 1,12,64,197	97125	0.641954
xa_nn_elm_quantize _f32_sym8u – channel based	Channel = 3 Input - 1,3,8,4	1181	12.30208
	Channel = 1 Input - 1,12,64,197	97123	0.64194
xa_nn_elm_quantize _f32_sym16	Input - 1,40,14,14	5269	0.672066
	Input - 1,3,1	315	105
xa_nn_elm_quantize _f32_sym16u	Input - 1,40,14,14	5273	0.672577
	Input - 1,3,1	317	105.6667
xa_nn_elm_quantize _f32_sym16 – channel based	Channel = 3 1,3,8,4	1193	12.42708
	Channel = 1 Input - 1,12,64,197	97119	0.641914
xa_nn_elm_quantize _f32_sym16u – channel based	Channel = 3 Input - 1,3,8,4	1191	12.40625
	Channel = 1 Input - 1,12,64,197	97100	0.641788
xa_nn_elm_dequantize_asym4_f32	Input - 1,40,14,14	3246	0.414031
	Input - 1,3,1	259	86.33333
xa_nn_elm_dequantize_asym4u_f32	Input - 1,40,14,14	3243	0.413648
	Input - 1,3,1	260	86.66667

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_dequantize_asym4_f32 – channel based	Channel = 3 1,3,8,4	1011	10.53125
	Channel = 1 Input - 1,12,64,197	58717	0.388094
xa_nn_elm_dequantize_asym4u_f32 – channel based	Channel = 3 Input - 1,3,8,4	1013	10.55208
	Channel = 1 Input - 1,12,64,197	58730	0.388179
xa_nn_elm_dequantize_asym8_f32	Input - 1,40,14,14	3244	0.413776
	Input - 1,3,1	260	86.66667
xa_nn_elm_dequantize_asym8u_f32	Input - 1,40,14,14	3244	0.413776
	Input - 1,3,1	261	87
xa_nn_elm_dequantize_asym8_f32 – channel based	Channel = 3 1,3,8,4	1010	10.52083
	Channel = 1 Input - 1,12,64,197	58743	0.388265
xa_nn_elm_dequantize_asym8u_f32 – channel based	Channel = 3 Input - 1,3,8,4	1012	10.54167
	Channel = 1 Input - 1,12,64,197	58780	0.38851
xa_nn_elm_dequantize_asym16_f32	Input - 1,40,14,14	3251	0.414668
	Input - 1,3,1	267	89
xa_nn_elm_dequantize_asym16u_f32	Input - 1,40,14,14	3252	0.414796
	Input - 1,3,1	268	89.33333
xa_nn_elm_dequantize_asym16_f32 – channel based	Channel = 3 1,3,8,4	1006	10.47917
	Channel = 1 Input - 1,12,64,197	58744	0.388272
xa_nn_elm_dequantize_asym16u_f32 – channel based	Channel = 3 Input - 1,3,8,4	1021	10.63542
	Channel = 1 Input - 1,12,64,197	58757	0.388358
xa_nn_elm_dequantize_sym4_f32	Input - 1,40,14,14	2255	0.287628
	Input - 1,3,1	247	82.33333
xa_nn_elm_dequantize_sym4u_f32	Input - 1,40,14,14	2260	0.288265
	Input - 1,3,1	250	83.33333

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_dequantize_sym4_f32 – channel based	Channel = 3 1,3,8,4	957	9.96875
	Channel = 1 Input - 1,12,64,197	39763	0.262816
xa_nn_elm_dequantize_sym4u_f32 – channel based	Channel = 3 Input - 1,3,8,4	963	10.03125
	Channel = 1 Input - 1,12,64,197	39820	0.263193
xa_nn_elm_dequantize_sym8_f32	Input - 1,40,14,14	2261	0.288393
	Input - 1,3,1	248	82.66667
xa_nn_elm_dequantize_sym8u_f32	Input - 1,40,14,14	2258	0.28801
	Input - 1,3,1	250	83.33333
xa_nn_elm_dequantize_sym8_f32 – channel based	Channel = 3 1,3,8,4	963	10.03125
	Channel = 1 Input - 1,12,64,197	39763	0.262816
xa_nn_elm_dequantize_sym8u_f32 – channel based	Channel = 3 Input - 1,3,8,4	962	10.02083
	Channel = 1 Input - 1,12,64,197	39796	0.263034
xa_nn_elm_dequantize_sym16_f32	Input - 1,40,14,14	2257	0.287883
	Input - 1,3,1	249	83
xa_nn_elm_dequantize_sym16u_f32	Input - 1,40,14,14	2267	0.289158
	Input - 1,3,1	251	83.66667
xa_nn_elm_dequantize_sym16_f32 – channel based	Channel = 3 1,3,8,4	971	10.11458
	Channel = 1 Input - 1,12,64,197	39768	0.262849
xa_nn_elm_dequantize_sym16u_f32 – channel based	Channel = 3 Input - 1,3,8,4	982	10.22917
	Channel = 1 Input - 1,12,64,197	39830	0.263259
xa_nn_elm_sub_32x32_32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	158	1.645833
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75386	0.375608
	Input1 - 39,39 Input2 - 39,39	692	0.454964
xa_nn_elm_sub_scalar_32x32_32	Input1 - 1,3,1 Input2 - 1	85	28.33333

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,288,14,14 Input2 - 1	14216	0.251842
	Input1 - 224,1 Input2 - 1	160	0.714286
xa_nn_elm_sub_broadcast_5D_32x32_32	Input1 - 1,8,512 Input2 - 1,1,512	1766	0.431152
	Input1 - 1,8,3,3 Input2 - 3,3	885	12.29167
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	10214	4.255833
xa_nn_elm_sub_f32xf32_f32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	148	1.541667
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75376	0.375558
	Input1 - 39,39 Input2 - 39,39	713	0.468771
xa_nn_elm_sub_scalar_f32xf32_f32	Input1 - 1,3,1 Input2 - 1	86	28.66667
	Input1 - 1,288,14,14 Input2 - 1	14216	0.251842
	Input1 - 224,1 Input2 - 1	160	0.714286
xa_nn_elm_sub_broadcast_5D_f32xf32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1749	0.427002
	Input1 - 1,8,3,3 Input2 - 3,3	856	11.88889
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	11914	4.964167
xa_nn_elm_div_32x32_f32 – mode 0 – high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	896	3.5
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552128	2.750956633
xa_nn_elm_div_scalar_32x32_f32 – mode 0 – high performance	Input1 - 1,4,8,8 Input2 - 1	215	0.83984375
	Input1 - 1,16,112,112 Input2 - 1	50327	0.250752352
xa_nn_elm_div_broadcast_5D_32x32_f32 – mode 0 – high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	266695	5.315190529
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1701	17.71875
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	21894	9.1225
xa_nn_elm_div_32x32_32 – mode 1 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	970	3.7890625
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602314	3.001006457
xa_nn_elm_div_scalar_32x32_32 – mode 1 - high performance	Input1 - 1,4,8,8 Input2 - 1	286	1.1171875
	Input1 - 1,16,112,112 Input2 - 1	100510	0.500787229

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 1 - High performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	354493	7.064991231
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1866	19.4375
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	29762	12.40083333
xa_nn_elm_div_32x32_32 – mode 2 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	976	3.8125
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602320	3.001036352
xa_nn_elm_ div_scalar_32x32_32 – mode 2 - high performance	Input1 - 1,4,8,8 Input2 - 1	360	1.40625
	Input1 - 1,16,112,112 Input2 - 1	150696	0.750837054
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 2 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	360765	7.189991231
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1923	20.03125
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	30722	12.80083333
xa_nn_elm_div_f32xf32_f32 – mode 0 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	894	3.4921875
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552126	2.750946668
xa_nn_elm_div_scalar_f32xf32_f32 – mode 0 - high performance	Input1 - 1,4,8,8 Input2 - 1	201	0.78515625
	Input1 - 1,16,112,112 Input2 - 1	50313	0.250682597
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 0 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	281948	5.619180485
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1682	17.52083333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	27169	11.32041667
xa_nn_elm_div_f32xf32_f32 – mode 1 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	902	3.5234375
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552134	2.750986527
xa_nn_elm_div_scalar_f32xf32_f32 – mode 1 - high performance	Input1 - 1,4,8,8 Input2 - 1	215	0.83984375
	Input1 - 1,16,112,112 Input2 - 1	50327	0.250752352
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 1 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	289116	5.762037628
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1712	17.83333333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	28849	12.02041667
xa_nn_elm_div_f32xf32_f32 – mode 2 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	901	3.51953125

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552133	2.750981545
xa_nn_elm_div_scalar_f32xf32_f32 – mode 2 - high performance	Input1 - 1,4,8,8 Input2 - 1	217	0.84765625
	Input1 - 1,16,112,112 Input2 - 1	50329	0.250762317
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	286428	5.708466199
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1688	17.58333333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	28369	11.82041667
xa_nn_elm_div_32x32_f32 – mode 0 – high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	896	3.5
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552128	2.750956633
xa_nn_elm_div_scalar_32x32_f32 – mode 0 – high precision	Input1 - 1,4,8,8 Input2 - 1	556	2.171875
	Input1 - 1,16,112,112 Input2 - 1	276172	1.376016422
xa_nn_elm_div_broadcast_5D_32x32_f32 – mode 0 – high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	266695	5.315190529
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1640	17.08333333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	24454	10.18916667
xa_nn_elm_div_32x32_32 – mode 1 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	970	3.7890625
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602314	3.001006457
xa_nn_elm_div_scalar_32x32_32 – mode 1 - high precision	Input1 - 1,4,8,8 Input2 - 1	573	2.23828125
	Input1 - 1,16,112,112 Input2 - 1	301245	1.500941685
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 1 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	354493	7.064991231
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1776	18.5
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	46083	19.20125
xa_nn_elm_div_32x32_32 – mode 2 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	976	3.8125
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602320	3.001036352
xa_nn_elm_ div_scalar_32x32_32 – mode 2 - high precision	Input1 - 1,4,8,8 Input2 - 1	610	2.3828125
	Input1 - 1,16,112,112 Input2 - 1	326338	1.625672632
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 2 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	360765	7.189991231

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1893	19.71875
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	48483	20.20125
xa_nn_elm_div_f32xf32_f32 – mode 0 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	894	3.4921875
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552126	2.750946668
xa_nn_elm_div_scalar_f32xf32_f32 – mode 0 - high precision	Input1 - 1,4,8,8 Input2 - 1	535	2.08984375
	Input1 - 1,16,112,112 Input2 - 1	263623	1.31349151
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 0 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	281948	5.619180485
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1607	16.73958333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	42049	17.52041667
xa_nn_elm_div_f32xf32_f32 – mode 1 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	902	3.5234375
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552134	2.750986527
xa_nn_elm_div_scalar_f32xf32_f32 – mode 1 - high precision	Input1 - 1,4,8,8 Input2 - 1	541	2.11328125
	Input1 - 1,16,112,112 Input2 - 1	276157	1.375941685
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 1 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	289116	5.762037628
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1676	17.45833333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	44449	18.52041667
xa_nn_elm_div_f32xf32_f32 – mode 2 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	901	3.51953125
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552133	2.750981545
xa_nn_elm_div_scalar_f32xf32_f32 – mode 2 - high precision	Input1 - 1,4,8,8 Input2 - 1	534	2.0859375
	Input1 - 1,16,112,112 Input2 - 1	276150	1.375906808
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	286428	5.708466199
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1652	17.20833333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	44929	18.72041667
xa_nn_elm_exp_f32_f32	Input - 1,3,8,4	589	6.135417
	Input - 1,256,28,28	853169	4.250882
	Input - 197,768	643186	4.251177

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input - 1,8,128	4526	4.419922
xa_nn_slice – signed 8-bit	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	454	0.50669643
	Input – 1,3,8,4,2 Start – 1 End – 2 Step – 1 Slice dim – 4	433	4.51041667
	Input – 39,1,512 Start – 4 End – 21 Step – 2 Slice dim - 0	684	0.1484375
xa_nn_slice – unsigned 8-bit	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	454	0.506696
	Input – 1,3,8,4,2 Start – 1 End – 2 Step – 1 Slice dim – 4	433	4.510417
	Input – 39,1,512 Start – 4 End – 21 Step – 2 Slice dim - 0	684	0.148438
xa_nn_slice – 16-bit	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	468	0.52232143
	Input – 1,3,8,4,2 Start – 1 End – 2 Step – 1 Slice dim – 4	433	4.51041667
	Input – 39,1,512 Start – 4 End – 21 Step – 2 Slice dim - 0	972	0.2109375
xa_nn_slice – 32-bit	Input – 1,8,128 Start – 1 End – (-1) Step – 1	580	0.647321

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Slice dim - 1		
	Input – 1,3,8,4,2 Start – 1 End – 2 Step – 1 Slice dim – 4	433	4.510417
	Input – 39,1,512 Start – 4 End – 21 Step – 2 Slice dim - 0	1548	0.335938
xa_nn_permute – signed 8-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	93102	0.618503
	Input – 1000,768, Permute_vector - 1, 0,	466635	0.607598
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2473	12.88021
	Input – 1, 768, 196, Permute_vector - 1,0,2	9969	0.066227
xa_nn_permute – unsigned 8-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	93102	0.618503
	Input – 1000,768, Permute_vector - 1, 0,	466635	0.607598
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2473	12.88021
	Input – 1, 768, 196, Permute_vector - 1,0,2	9969	0.066227
xa_nn_permute – 16-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	93722	0.622622
	Input – 1000,768, Permute_vector - 1, 0,	470507	0.612639
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2211	11.51563
	Input – 1, 768, 196, Permute_vector - 1,0,2	19377	0.128727
xa_nn_permute – 32-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	101938	0.677203
	Input – 1000,768, Permute_vector - 1, 0,	508891	0.662618
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	1806	9.40625
	Input – 1, 768, 196, Permute_vector - 1,0,2	38193	0.253727
xa_nn_mean_f32_f32	Input – 1,16,56,56 Dims – 3	55532	1.106744
	Input - 1,3,64 Dims – 2	557	2.901042
	Input – 1,240,14,14 Dims - 2	135850	2.887968

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input – 1,16,56,56 Dims - 1,3	56303	1.12211
	Input – 3,1,240,14,14 Dims - 0,2,3	60025	0.425347
	Input – 1,240,14,14 Dims – 1	20034	0.425893
	Input – 3,1,240,14,14 Dims - 3,0,2	60025	0.425347

4. References

[1] FusionG3-NNLib-API.pdf