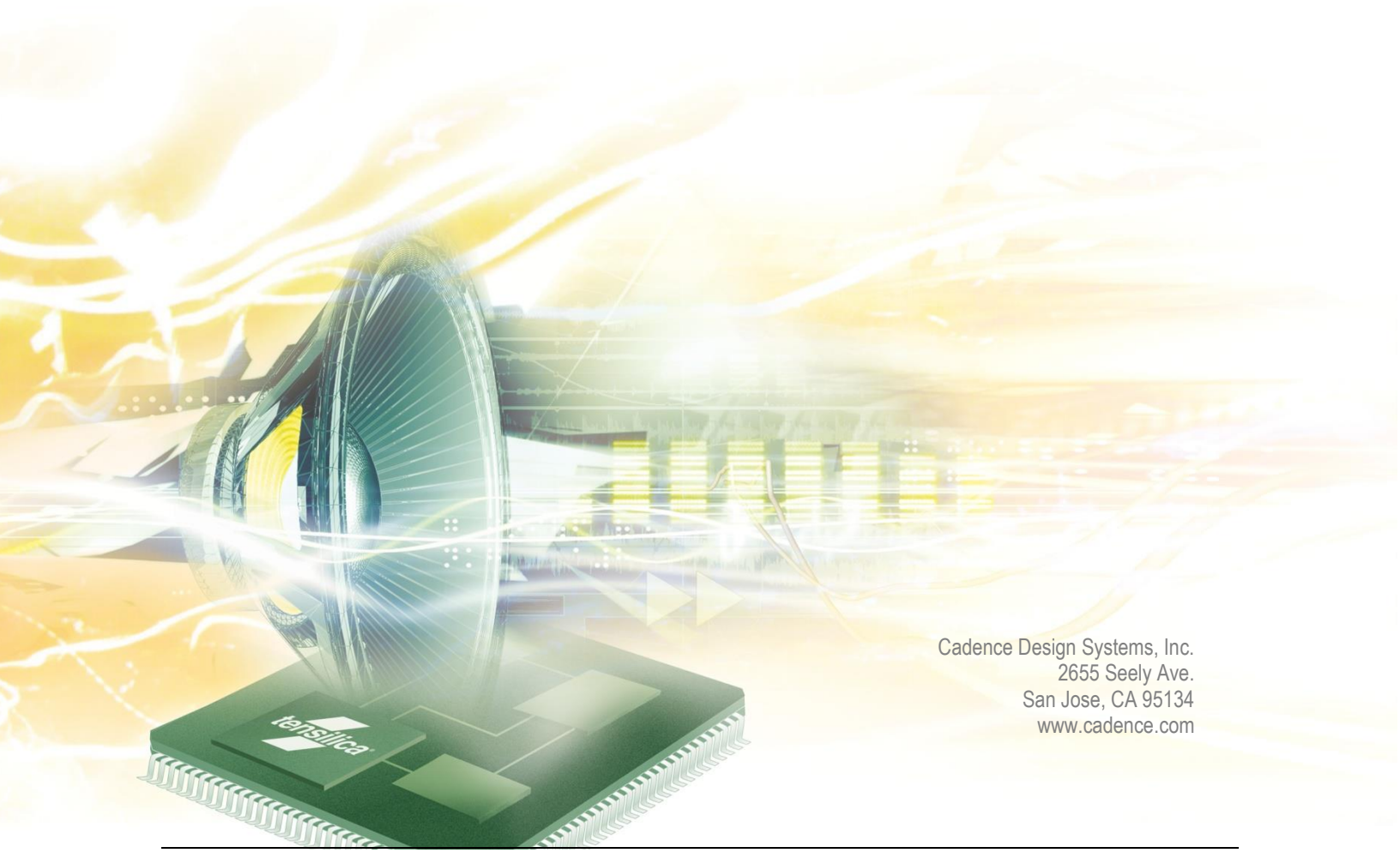
cadence®

Fusion G3 Neural Network Library

Performance report



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

Version 1.2
February 2025

Contents

Tables

Change History

Version	Changes
1.0	Initial version
1.1	Added performance cycles for div, sub, exp, slice, permute, mean kernels
1.2	Added performance cycles for clamp, sigmoid, sqrt, rsqrt, tanh, lt, transpose, where, sub kernels. Updated the report with cycles for quantize and dequantize kernels with packed implementaion (4-bit).

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	1.2
Library API Version	1.2
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) default implementation	366927	1367
Fusion G3 (with NN, SP-VFPU and ENABLE_4BIT_PACK)	431324	1367

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 (Default implementation)	4137
xa_nn_elm_dequantize_asym4u_f32 (Default implementation)	4121
xa_nn_elm_dequantize_asym4_f32 (ENABLE_4BIT_PACK macro enabled)	9766

xa_nn_elm_dequantize_asym4u_f32 (ENABLE_4BIT_PACK macro enabled)	9766
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 (Default implementation)	4175
xa_nn_elm_dequantize_sym4u_f32 (Default implementation)	4175
xa_nn_elm_dequantize_sym4_f32 (ENABLE_4BIT_PACK macro enabled)	8503
xa_nn_elm_dequantize_sym4u_f32 (ENABLE_4BIT_PACK macro enabled)	8503
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 (Default implementation)	6146
xa_nn_elm_quantize_f32_asym4u (Default implementation)	6146
xa_nn_elm_quantize_f32_asym4 (ENABLE_4BIT_PACK macro enabled)	18738

xa_nn_elm_quantize_f32_asym4u (ENABLE_4BIT_PACK macro enabled)	16450
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 (Default implementation)	5389
xa_nn_elm_quantize_f32_sym4u (Default implementation)	5389
xa_nn_elm_quantize_f32_sym4 (ENABLE_4BIT_PACK macro enabled)	16268
xa_nn_elm_quantize_f32_sym4u (ENABLE_4BIT_PACK macro enabled)	16097
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
xa_nn_mean_f32_f32	9053
xa_nn_tanh_f32_f32	4062
xa_nn_sigmoid_f32_f32	3010
xa_nn_elm_rsqrt_f32_f32	924
xa_nn_elm_sqrt_f32_f32	1532
xa_nn_elm_clamp_8_8	352
xa_nn_elm_clamp_scalar_8_8	468
xa_nn_elm_clamp_broadcast_5D_8_8	12719
xa_nn_elm_clamp_8u_8u	352
xa_nn_elm_clamp_scalar_8u_8u	468
xa_nn_elm_clamp_broadcast_5D_8u_8u	12719
xa_nn_elm_clamp_16_16	384
xa_nn_elm_clamp_scalar_16_16	488
xa_nn_elm_clamp_broadcast_5D_16_16	12895
xa_nn_elm_clamp_f32_f32	416
xa_nn_elm_clamp_scalar_f32_f32	492
xa_nn_elm_clamp_broadcast_5D_f32_f32	13151
xa_nn_elm_less_f32xf32_bool	436
xa_nn_elm_less_scalar_f32xf32_bool	500

xa_nn_elm_less_broadcast_5D_f32xf32_bool	6885
xa_nn_elm_where_f32xf32_f32	432
xa_nn_elm_where_broadcast_5D_f32xf32_f32	12401
xa_nn_elm_sub_32xf32xf32_f32	572
xa_nn_elm_sub_scalar_32xf32xf32_f32	732
xa_nn_elm_sub_broadcast_5D_32xf32xf32_f32	7632
xa_nn_elm_sub_32xf32x32_f32	576
xa_nn_elm_sub_scalar_32xf32x32_f32	748
xa_nn_elm_sub_broadcast_5D_32xf32x32_f32	7667
xa_nn_elm_sub_f32x32xf32_f32	528
xa_nn_elm_sub_scalar_f32x32xf32_f32	388
xa_nn_elm_sub_broadcast_5D_f32x32xf32_f32	7240
xa_nn_elm_sub_f32x32x32_f32	532
xa_nn_elm_sub_scalar_f32x32x32_f32	388
xa_nn_elm_sub_broadcast_5D_f32x32x32_f32	7257

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	4366475	9.37599311
	axis = 0 input – 96,197,197	21890177	5.875510245
	axis = 2 input - 1,3,172	3459	6.703488372
	Axis = 3 Input - 1, 12, 197, 197	2974074	6.386134659
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 (Default implementation)	Input - 1,40,14,14	6300	0.803571429
	Input - 1,3,1	359	119.6666667
xa_nn_elm_quantize _f32_asym4u (Default implementation)	Input - 1,40,14,14	6300	0.803571429
	Input - 1,3,1	360	120

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_quantize _f32_asym4 – channel based (Default implementation)	Channel = 3 1,3,8,4	743	7.739583333
	Channel = 1 Input - 1,12,64,197	116246	0.76833492
xa_nn_elm_quantize _f32_asym4u – channel based (Default implementation)	Channel = 3 Input - 1,3,8,4	736	7.666666667
	Channel = 1 Input - 1,12,64,197	116223	0.7681829
xa_nn_elm_quantize _f32_asym4 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	11236	1.433163
	Input - 1,3,1	378	126
xa_nn_elm_quantize _f32_asym4u (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	10265	1.309311
	Input - 1,3,1	376	125.3333
xa_nn_elm_quantize _f32_asym4 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 1,3,8,4	839	8.739583
	Channel = 1 Input - 1,12,64,197	211195	1.395906
xa_nn_elm_quantize _f32_asym4u – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 Input - 1,3,8,4	842	8.770833
	Channel = 1 Input - 1,12,64,197	192342	1.271296
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 = 3	1215	12.65625

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
– channel based	1,3,8,4		
	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 (Default implementation)	Input - 1,40,14,14	5295	0.675382653
	Input - 1,3,1	337	112.3333333
xa_nn_elm_quantize _f32_sym4u (Default implementation)	Input - 1,40,14,14	5295	0.675382653
	Input - 1,3,1	337	112.3333333
xa_nn_elm_quantize _f32_sym4 – channel based (Default implementation)	Channel = 3 1,3,8,4	701	7.302083333
	Channel = 1 Input - 1,12,64,197	97186	0.642356705
xa_nn_elm_quantize _f32_sym4u – channel based (Default implementation)	Channel = 3 Input - 1,3,8,4	701	7.302083333
	Channel = 1 Input - 1,12,64,197	97186	0.642356705
xa_nn_elm_quantize _f32_sym4 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	10250	0.675382653
	Input - 1,3,1	358	112.3333333
xa_nn_elm_quantize _f32_sym4u (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	9244	0.675382653
	Input - 1,3,1	354	112.3333333
xa_nn_elm_quantize _f32_sym4 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 1,3,8,4	824	7.302083333
	Channel = 1 Input - 1,12,64,197	192305	0.642356705
xa_nn_elm_quantize _f32_sym4u – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 Input - 1,3,8,4	783	7.302083333
	Channel = 1 Input - 1,12,64,197	173136	0.642356705
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

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
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 (Default implementation)	Input - 1,40,14,14	3268	0.416837
	Input - 1,3,1	283	94.33333
xa_nn_elm_dequantize_asym4u_f32 (Default implementation)	Input - 1,40,14,14	3267	0.416709
	Input - 1,3,1	282	94
xa_nn_elm_dequantize_asym4_f32 – channel based (Default implementation)	Channel = 3 1,3,8,4	702	7.3125
	Channel = 1 Input - 1,12,64,197	58833	0.38886
xa_nn_elm_dequantize_asym4u_f32 – channel based (Default implementation)	Channel = 3 Input - 1,3,8,4	698	7.270833
	Channel = 1 Input - 1,12,64,197	58821	0.388781
xa_nn_elm_dequantize_asym4_f32 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	4306	0.549235
	Input - 1,3,1	314	104.6667
xa_nn_elm_dequantize_asym4u_f32 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	4304	0.54898
	Input - 1,3,1	312	104

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_dequantize_asym4_f32 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 1,3,8,4	1049	10.92708
	Channel = 1 Input - 1,12,64,197	77469	0.512036
xa_nn_elm_dequantize_asym4u_f32 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 Input - 1,3,8,4	1045	10.88542
	Channel = 1 Input - 1,12,64,197	77457	0.511957
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 (Default implementation)	Input - 1,40,14,14	2273	0.289923
	Input - 1,3,1	267	89
xa_nn_elm_dequantize_sym4u_f32 (Default implementation)	Input - 1,40,14,14	2273	0.289923
	Input - 1,3,1	267	89

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_dequantize_sym4_f32 – channel based (Default implementation)	Channel = 3 1,3,8,4	654	6.8125
	Channel = 1 Input - 1,12,64,197	39893	0.263675
xa_nn_elm_dequantize_sym4u_f32 – channel based (Default implementation)	Channel = 3 Input - 1,3,8,4	654	6.8125
	Channel = 1 Input - 1,12,64,197	39893	0.263675
xa_nn_elm_dequantize_sym4_f32 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	3294	0.420153
	Input - 1,3,1	285	95
xa_nn_elm_dequantize_sym4u_f32 (ENABLE_4BIT_PACK macro enabled)	Input - 1,40,14,14	3294	0.420153
	Input - 1,3,1	285	95
xa_nn_elm_dequantize_sym4_f32 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 1,3,8,4	944	9.833333
	Channel = 1 Input - 1,12,64,197	58446	0.386302
xa_nn_elm_dequantize_sym4u_f32 – channel based (ENABLE_4BIT_PACK macro enabled)	Channel = 3 Input - 1,3,8,4	944	9.833333
	Channel = 1 Input - 1,12,64,197	58446	0.386302
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 = 3	971	10.11458

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
– channel based	1,3,8,4		
	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
	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_sub_32xf32xf32_f32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	155	1.614583333
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75383	0.375592913
	Input1 - 39,39 Input2 - 39,39	732	0.481262327

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_sub_scalar_32xf32xf32_f32	Input1 - 1,3,1 Input2 - 1	94	31.33333333
	Input1 - 1,288,14,14 Input2 - 1	19526	0.345911281
	Input1 - 224,1 Input2 - 1	199	0.888392857
xa_nn_elm_sub_broadcast_5D_32xf32xf32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1824	0.4453125
	Input1 - 1,8,3,3 Input2 - 3,3	990	13.75
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	14137	5.890416667
xa_nn_elm_sub_32xf32x32_f32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	152	1.583333333
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75380	0.375577966
	Input1 - 39,39 Input2 - 39,39	729	0.479289941
xa_nn_elm_sub_scalar_32xf32x32_f32	Input1 - 1,3,1 Input2 - 1	94	31.33333333
	Input1 - 1,288,14,14 Input2 - 1	19526	0.345911281
	Input1 - 224,1 Input2 - 1	199	0.888392857
xa_nn_elm_sub_broadcast_5D_32xf32x32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1825	0.445556641
	Input1 - 1,8,3,3 Input2 - 3,3	977	13.56944444
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	14136	5.89
xa_nn_elm_sub_f32x32xf32_f32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	156	1.625
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75384	0.375597895
	Input1 - 39,39 Input2 - 39,39	733	0.48191979
xa_nn_elm_sub_scalar_f32x32xf32_f32	Input1 - 1,3,1 Input2 - 1	90	30
	Input1 - 1,288,14,14 Input2 - 1	14217	0.251860119
	Input1 - 224,1 Input2 - 1	161	0.71875
xa_nn_elm_sub_broadcast_5D_f32x32xf32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1823	0.445068359
	Input1 - 1,8,3,3 Input2 - 3,3	991	13.76388889
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	12216	5.09
xa_nn_elm_sub_f32x32x32_f32	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	153	1.59375

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,256,28,28 Input2 - 1,256,28,28	75381	0.375582948
	Input1 - 39,39 Input2 - 39,39	730	0.479947403
xa_nn_elm_sub_scalar_f32x32x32_f32	Input1 - 1,3,1 Input2 - 1	91	30.33333333
	Input1 - 1,288,14,14 Input2 - 1	14215	0.251824688
	Input1 - 224,1 Input2 - 1	159	0.709821429
xa_nn_elm_sub_broadcast_5D_f32x32x32_f32	Input1 - 1,8,512 Input2 - 1,1,512	1820	0.444335938
	Input1 - 1,8,3,3 Input2 - 3,3	977	13.56944444
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	12216	5.09
xa_nn_elm_div_32x32_f32 – mode 0 – high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	898	3.5078125
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552130	2.750966598
xa_nn_elm_div_scalar_32x32_f32 – mode 0 – high performance	Input1 - 1,4,8,8 Input2 - 1	219	0.85546875
	Input1 - 1,16,112,112 Input2 - 1	50331	0.250772282
xa_nn_elm_div_broadcast_5D_32x32_f32 – mode 0 – high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	267611	5.333446269
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1702	17.72916667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	21893	9.122083333
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	287	1.12109375
	Input1 - 1,16,112,112 Input2 - 1	100511	0.500792211
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 1 - High performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	358080	7.136479592
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1855	19.32291667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	30006	12.5025
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	361	1.41015625

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,16,112,112 Input2 - 1	150697	0.750842036
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 2 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	362560	7.225765306
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1951	20.32291667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	30966	12.9025
xa_nn_elm_div_f32xf32_f32 – mode 0 - high performance	Input1 - 1,4,8,8 Input2 - 1,4,8,8	897	3.50390625
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552129	2.750961615
xa_nn_elm_div_scalar_f32xf32_f32 – mode 0 - high performance	Input1 - 1,4,8,8 Input2 - 1	205	0.80078125
	Input1 - 1,16,112,112 Input2 - 1	50317	0.250702527
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	1685	17.55208333
	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	216	0.83984375
	Input1 - 1,16,112,112 Input2 - 1	50328	0.250757334
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 1 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	287327	5.726383131
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1724	17.83333333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	28852	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
	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	218	0.84765625
	Input1 - 1,16,112,112 Input2 - 1	50330	0.250767299
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high performance	Input1 - 1,16,1,1 Input2 - 1,16,56,56	285535	5.690668846
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1700	17.58333333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	28372	11.82041667

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_div_32x32_f32 – mode 0 – high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	898	3.5078125
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552130	2.750966598
xa_nn_elm_div_scalar_32x32_f32 – mode 0 – high precision	Input1 - 1,4,8,8 Input2 - 1	560	2.1875
	Input1 - 1,16,112,112 Input2 - 1	276176	1.376036352
xa_nn_elm_div_broadcast_5D_32x32_f32 – mode 0 – high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	266694	5.315170599
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1639	17.07291667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	24453	10.18875
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	574	2.2421875
	Input1 - 1,16,112,112 Input2 - 1	301246	1.500946668
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 1 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	355392	7.082908163
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1777	18.51041667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	46326	19.3025
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	611	2.38671875
	Input1 - 1,16,112,112 Input2 - 1	326339	1.625672632
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 2 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	361664	7.207908163
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1885	19.63541667
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	48726	20.3025
xa_nn_elm_div_f32xf32_f32 – mode 0 - high precision	Input1 - 1,4,8,8 Input2 - 1,4,8,8	897	3.50390625
	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552129	2.750961615
xa_nn_elm_div_scalar_f32xf32_f32 – mode 0 - high precision	Input1 - 1,4,8,8 Input2 - 1	539	2.10546875
	Input1 - 1,16,112,112 Input2 - 1	263627	1.31351144

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
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	542	2.1171875
	Input1 - 1,16,112,112 Input2 - 1	276158	1.375946668
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 1 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	289119	5.762097417
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1679	17.48958333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	44452	18.52166667
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	535	2.08984375
	Input1 - 1,16,112,112 Input2 - 1	276151	1.37591179
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high precision	Input1 - 1,16,1,1 Input2 - 1,16,56,56	286431	5.708525989
	Input1 - 1,3,8,4 Input2 - 1,3,1,4	1655	17.23958333
	Input1 - 4,6,10,10 Input2 - 4,6,10,1	44932	18.72166667
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
	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	684	0.1484375

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Start – 4 End – 21 Step – 2 Slice dim - 0		
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 Slice dim - 1	580	0.647321
	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	93101	0.618496227

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input – 1000,768, Permute_vector - 1, 0,	466636	0.607598958
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2478	12.90625
	Input – 1, 768, 196, Permute_vector - 1,0,2	9916	0.065874787
xa_nn_permute – unsigned 8-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	93101	0.618496227
	Input – 1000,768, Permute_vector - 1, 0,	466636	0.607598958
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2478	12.90625
	Input – 1, 768, 196, Permute_vector - 1,0,2	9916	0.065874787
xa_nn_permute – 16-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	93716	0.622581845
	Input – 1000,768, Permute_vector - 1, 0,	470503	0.612634115
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	2211	11.515625
	Input – 1, 768, 196, Permute_vector - 1,0,2	19319	0.128341571
xa_nn_permute – 32-bit	Input – 1, 768, 196, Permute_vector - 2,1,0	101931	0.677156409
	Input – 1000,768, Permute_vector - 1, 0,	508886	0.662611979
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	1805	9.401041667
	Input – 1, 768, 196, Permute_vector - 1,0,2	38134	0.253334928
xa_nn_mean_f32_f32	Input – 1,16,56,56 Dims – 3	55310	61.72991071
	Input - 1,3,64 Dims – 2	560	186.6666667
	Input – 1,240,14,14 Dims - 2	49306	14.67440476
	Input – 1,16,56,56 Dims - 1,3	39398	703.5357143
	Input – 3,1,240,14,14 Dims - 0,2,3	46498	3321.285714
	Input – 1,240,14,14 Dims – 1	13875	70.79081633
	Input – 3,1,240,14,14 Dims - 3,0,2	46494	3321
xa_nn_tanh_f32_f32	Input – 1,512	3446	6.73046875
	Input1 – 1,8,3,3 Input2 – 1,8,3,3 Cond – 1,8,3,3	147	2.041666667

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_where_f32xf32_f32	Input1 – 1,12,197,197 Input2 – 1,12,197,197 Cond – 1,12,197,197	232968	0.500244789
xa_nn_elm_where_broadcast_5D_f32xf32_f32	Cond – 1,12,197,1 Input1 – 1,12,197,197 Input2 – 1,12,197,197	289197	0.620983535
	Cond – 3,3 Input1 – 1,8,3,3 Input2 – 1,8,3,3	1040	14.44444444
	Cond – 1,12,197,197 Input1 – 1,12,197,1 Input2 – 1,12,197,197	317563	0.681892946
xa_nn_elm_less_f32xf32_bool	Input1 – 1,40,14,14 Input2 – 1,40,14,14	3059	0.390178571
	Input1 – 1,1024,28,28 Input2 – 1,1024,28,28	301175	0.375148228
	Input1 – 1,3,64 Input2 – 1,3,64	191	0.994791667
xa_nn_elm_less_scalar_f32xf32_bool	Input1 – 1,3,1 Input2 – 1	83	27.66666667
	Input1 – 1,288,14,14 Input2 – 1	14217	0.251860119
xa_nn_elm_less_broadcast_5D_f32xf32_bool	Input1 – 1,8,512 Input2 – 1,1,512	1787	0.436279297
	Input1 – 1,8,3,3 Input2 – 3,3	793	11.01388889
	Input1 – 4,6,10,10 Input2 – 4,6,10,1	10801	4.500416667
xa_nn_sigmoid_f32_f32	Input – 1,512	2672	5.21875
	Input – 1,256,28,28	953584	4.751196
	Input – 197,768	718896	4.751586
xa_nn_elm_rsqrt_f32_f32	Input – 1,512	915	1.787109375
	Input – 1,3,1	112	37.33333333
	Input – 1,256,28,28	301203	1.500732
	Input – 197,768	227091	1.500972
xa_nn_elm_sqrt_f32_f32	Input – 1,512	1754	3.42578125
	Input – 1,256,28,28	627353	3.125762
	Input – 197,768	472953	3.126011

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_clamp_f32_f32	Input – 1,16,112,112 Min – 1,16,112,112 Max – 1,16,112,112	100461	0.500543088
	Input – 28 Min – 28 Max – 28	124	4.428571429
xa_nn_elm_clamp_scalar_f32_f32	Input – 1,576,1,1 Min – 1 Max – 1	254	0.440972222
	Input – 1,96,14,14 Min – 1 Max – 1	4814	0.255846088
xa_nn_elm_clamp_broadcast_5D_f32_f32	Input – 1,8,1 Min – 1,8,512 Max – 1,1,512	2517	0.614501953
	Input – 1,16,1,1 Min – 1,16,56,1 Max – 1,16,56,56	67192	1.339126276
xa_nn_elm_clamp_16_16	Input – 1,16,112,112 Min – 1,16,112,112 Max – 1,16,112,112	50278	0.250508211
	Input – 28 Min – 28 Max – 28	110	3.928571429
xa_nn_elm_clamp_scalar_16_16	Input – 1,576,1,1 Min – 1 Max – 1	178	0.309027778
	Input – 1,96,14,14 Min – 1 Max – 1	2458	0.130633503
xa_nn_elm_clamp_broadcast_5D_16_16	Input – 1,8,1 Min – 1,8,512 Max – 1,1,512	1732	0.422851563
	Input – 1,16,1,1 Min – 1,16,56,1 Max – 1,16,56,56	60912	1.213966837
xa_nn_elm_clamp_8_8	Input – 1,16,112,112 Min – 1,16,112,112 Max – 1,16,112,112	25185	0.125483299
	Input – 28 Min – 28 Max – 28	95	3.392857143
xa_nn_elm_clamp_scalar_8_8	Input – 1,576,1,1 Min – 1 Max – 1	140	0.243055556
	Input – 1,96,14,14 Min – 1 Max – 1	1281	0.068080357
	Input – 1,8,1 Min – 1,8,512	1326	0.32373

Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
xa_nn_elm_clamp_broadcast_5D_8_8	Max – 1,1,512		
	Input – 1,16,1,1 Min – 1,16,56,1 Max – 1,16,56,56	51963	1.035614636
xa_nn_elm_clamp_8u_8u	Input – 1,16,112,112 Min – 1,16,112,112 Max – 1,16,112,112	25185	0.125483299
	Input – 28 Min – 28 Max – 28	95	3.392857143
xa_nn_elm_clamp_scalar_8u_8u	Input – 1,576,1,1 Min – 1 Max – 1	141	0.244791667
	Input – 1,96,14,14 Min – 1 Max – 1	1281	0.068080357
xa_nn_elm_clamp_broadcast_5D_8u_8u	Input – 1,8,1 Min – 1,8,512 Max – 1,1,512	1326	0.323730469
	Input – 1,16,1,1 Min – 1,16,56,1 Max – 1,16,56,56	51963	1.035614636
xa_nn_permute – signed 8-bit (transpose)	Input – 1, 768, 196, Permute_vector – 0, 2, 1	93031	0.618031197
	Input – 1000,768, Permute_vector - 1, 0,	466636	0.607598958
	Input – 1,3,8,8 Permute_vector – 0, 2, 1, 3	763	3.973958333
xa_nn_permute – unsigned 8-bit (transpose)	Input – 1, 768, 196, Permute_vector – 0, 2, 1	93031	0.618031197
	Input – 1000,768, Permute_vector - 1, 0,	466636	0.607598958
	Input – 1,3,8,8 Permute_vector – 0, 2, 1, 3	763	3.973958333
xa_nn_permute – 16-bit (transpose)	Input – 1, 768, 196, Permute_vector – 0, 2, 1	93646	0.622116815
	Input – 1000,768, Permute_vector - 1, 0,	470503	0.612634115
	Input – 1,3,8,8 Permute_vector – 0, 2, 1, 3	1176	6.125
xa_nn_permute – 32-bit (transpose)	Input – 1, 768, 196, Permute_vector – 0, 2, 1	101861	0.67669138
	Input – 1000,768, Permute_vector - 1, 0,	508886	0.662611979
	Input – 1,3,8,8 Permute_vector – 0, 2, 1, 3	1288	6.708333333

4. References

[1] FusionG3-NNLib-API.pdf