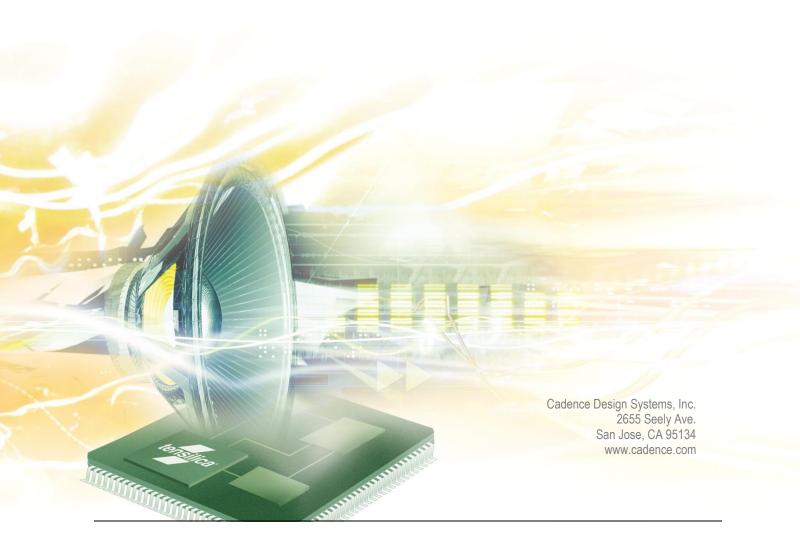


Fusion G3 Neural Network Library

Performance report





© 2024 Cadence Design Systems, Inc. All rights reserved.

Cadence Design Systems, Inc. (Cadence), 2655 Seely Ave., San Jose, CA 95134, USA.

Trademarks: Trademarks and service marks of Cadence Design Systems, Inc. (Cadence) contained in this document are attributed to Cadence with the appropriate symbol. For queries regarding Cadence's trademarks, contact the corporate legal department at the address shown above or call 1-800-862-4522.

All other trademarks are the property of their respective holders.

Patents: Licensed under U.S. Patent Nos. 7,526,739; 8,032,857; 8,209,649; 8,266,560; 8,650,516

Restricted Print Permission: This publication is protected by copyright and any unauthorized use of this publication may violate copyright, trademark, and other laws. Except as specified in this permission statement, this publication may not be copied, reproduced, modified, published, uploaded, posted, transmitted, or distributed in any way, without prior written permission from Cadence. This statement grants you permission to print one (1) hard copy of this publication subject to the following conditions:

- The publication may be used solely for personal, informational, and noncommercial purposes;
- The publication may not be modified in any way;
- Any copy of the publication or portion thereof must include all original copyright, trademark, and other proprietary notices and this permission statement,
- The information contained in this document cannot be used in the development of like products or software, whether for internal or external use, and shall not be used for the benefit of any other party, whether or not for consideration; and
- Cadence reserves the right to revoke this authorization at any time, and any such use shall be discontinued immediately upon written notice from Cadence.

Disclaimer: Information in this publication is subject to change without notice and does not represent a commitment on the part of Cadence. The information contained herein is the proprietary and confidential information of Cadence or its licensors, and is supplied subject to, and may be used only by Cadence's customer in accordance with, a written agreement between Cadence and its customer. Except as may be explicitly set forth in such agreement, Cadence does not make, and expressly disclaims, any representations or warranties as to the completeness, accuracy or usefulness of the information contained in this document. Cadence does not warrant that use of such information will not infringe any third party rights, nor does Cadence assume any liability for damages or costs of any kind that may result from use of such information.

Restricted Rights: Use, duplication, or disclosure by the Government is subject to restrictions as set forth in FAR52.227-14 and DFAR252.227-7013 et seq. or its successor.

For further assistance, contact Cadence Online Support at https://support.cadence.com/. Copyright © 2024 Cadence Design Systems, Inc. All rights reserved.

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 ROData Sizes	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 (ROData) sizes of the archive are shown in **Error! Reference source not found.**2.

Table 2-2 Library Text and ROData Sizes

DCD	Neural Network Library	
DSP	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)
	axis pointer = NULL		
	input - 12,197,197		
	axis pointer = NULL		
	input - 1,3,172		
	axis = 0	4366471	9.37598452
xa_nn_vec_softmax_dim_f32_f32	input – 12,197,197		
	axis = 0	21890173	5.87550917
	input – 96,197,197		
	axis = 2	3455	6.69573643
	input - 1,3,172		
	Axis = 3	2974070	6.38612607
	Input - 1, 12, 197, 197		
	Input1 - 1,40,14,14	3060	0.390306
	Input2 - 1,40,14,14	204470	0.075440
xa_nn_elm_add_32x32_32	Input1 - 1,1024,28,28	301176	0.375149
	Input2 - 1,1024,28,28 Input1 - 1,3,64	192	1
	Input2 - 1,3,64	192	Į.
	Input1 - 1,3,1	84	28
	Input2 - 1	04	20
xa_nn_elm_add_scalar_32x32_32	Input1 - 1,288,14,14	14211	0.251754
	Input2 - 1		
	Input1 - 1,8,512	1811	0.442139
	Input2 - 1,1,512		
	Input1 - 1,8,3,3	943	13.09722
xa_nn_elm_add_broadcast_5D_32x32_32	Input2 - 3,3		
	Input1 - 4,6,10,10	10295	4.289583
	Input2 - 4,6,10,1		
	Input1 - 1,16,1,1	578148	0.551365
	Input2 - 1,16,256,256	2017	0.000010
	Input1 - 1,40,14,14	3047	0.388648
	Input2 - 1,40,14,14	201162	0.275422
xa_nn_elm_add_f32xf32_f32	Input1 - 1,1024,28,28	301163	0.375133
	Input2 - 1,1024,28,28 Input1 - 1,3,64	179	0.932292
	Input2 - 1,3,64	113	0.332232
vo nn alm add asslar f20vf20 f20	Input1 - 1,3,1	85	28.33333
xa_nn_elm_add_scalar_f32xf32_f32	IIIputi - 1,5,1	03	20.0000



Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input2 - 1		. ,
	Input1 - 1,288,14,14 Input2 - 1	14213	0.251789
	Input2 - 1 Input1 - 1,8,512 Input2 - 1,1,512	1791	0.437256
	Input1 - 1,8,3,3 Input2 - 3,3	887	12.31944
xa_nn_elm_add_broadcast_5D_f32xf32_f32	Input2 - 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
	Input1 - 1,3,172 Input2 - 1,3,172	318	0.616279
xa_nn_elm_mul_32x32_32	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
	Input1 - 1,64 Input2 - 1,64	119	1.859375
xa_nn_elm_mul_scalar_32x32_32	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37927	0.250681
	Input1 - 1,16,1,1 Input 2 - 1,16,56,56	62546	1.246532
xa_nn_elm_mul_broadcast_5D_32x32_32	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
	Input1 - 1,3,172 Input2 - 1,3,172	313	0.606589
xa_nn_elm_mul_f32xf32_f32	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
	Input1 - 1,64 Input2 - 1,64	121	1.890625
xa_nn_elm_mul_scalar_f32xf32_f32	Input1 - 1,12,64,197 Input2 - 1,12,64,197	37923	0.250654
	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
xa_nn_elm_mul_broadcast_5D_f32xf32_f32	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
	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	9950	0.066100659
xa_nn_cat -8-bit	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
	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	19358	0.128600659
xa_nn_cat – 16-bit	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
	Axis = 0 Input1 - 1, 1, 224, 224 Input2 - 1, 1, 224, 224 Input3 - 1, 1, 224, 224	38174	0.25360066
xa_nn_cat – 32-bit	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
	Input - 1,40,14,14	6261	0.798597
xa_nn_elm_quantize _f32_asym4u	Input - 1,3,1	319	106.3333



Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Channel = 3 1,3,8,4	1201	12.51042
xa_nn_elm_quantize _f32_asym4 – channel based	Channel = 1 Input - 1,12,64,197	116132	0.767581
	Channel = 3 Input - 1,3,8,4	1189	12.38542
xa_nn_elm_quantize _f32_asym4u – channel based	Channel = 1 Input - 1,12,64,197	116098	0.767357
va no alm quantina (22 agum)	Input - 1,40,14,14	6266	0.799235
xa_nn_elm_quantize _f32_asym8	Input - 1,3,1	323	107.6667
	Input - 1,40,14,14	6261	0.798597
xa_nn_elm_quantize _f32_asym8u	Input - 1,3,1	321	107
xa_nn_elm_quantize _f32_asym8	Channel = 3 1,3,8,4	1202	12.52083
- channel based	Channel = 1 Input - 1,12,64,197	116136	0.767608
xa_nn_elm_quantize _f32_asym8u	Channel = 3 Input - 1,3,8,4	1200	12.5
- channel based	Channel = 1 Input - 1,12,64,197	116121	0.767509
	Input - 1,40,14,14	6264	0.79898
xa_nn_elm_quantize _f32_asym16	Input - 1,3,1	330	110
	Input - 1,40,14,14	6261	0.798597
xa_nn_elm_quantize _f32_asym16u	Input - 1,3,1	324	108
xa_nn_elm_quantize _f32_asym16	Channel = 3 1,3,8,4	1215	12.65625
- channel based	Channel = 1 Input - 1,12,64,197	116108	0.767423
xa_nn_elm_quantize _f32_asym16u	Channel = 3 Input - 1,3,8,4	1204	12.54167
- channel based	Channel = 1 Input - 1,12,64,197	116078	0.767225
	Input - 1,40,14,14	5271	0.672321
xa_nn_elm_quantize _f32_sym4	Input - 1,3,1	310	103.3333
	Input - 1,40,14,14	5265	0.671556
xa_nn_elm_quantize _f32_sym4u	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 = 3 1,3,8,4	1183	12.32292
- channel based	Channel = 1 Input - 1,12,64,197	97132	0.642
xa_nn_elm_quantize _f32_sym4u	Channel = 3 Input - 1,3,8,4	1174	12.22917
- channel based	Channel = 1 Input - 1,12,64,197	97110	0.641854
(00 - 0	Input - 1,40,14,14	5272	0.672449
xa_nn_elm_quantize _f32_sym8	Input - 1,3,1	309	103
	Input - 1,40,14,14	5265	0.671556
xa_nn_elm_quantize _f32_sym8u	Input - 1,3,1	307	102.3333
xa_nn_elm_quantize_f32_sym8	Channel = 3 1,3,8,4	1184	12.33333
– channel based	Channel = 1 Input - 1,12,64,197	97125	0.641954
xa_nn_elm_quantize _f32_sym8u	Channel = 3 Input - 1,3,8,4	1181	12.30208
- channel based	Channel = 1 Input - 1,12,64,197	97123	0.64194
	Input - 1,40,14,14	5269	0.672066
xa_nn_elm_quantize _f32_sym16	Input - 1,3,1	315	105
	Input - 1,40,14,14	5273	0.672577
xa_nn_elm_quantize _f32_sym16u	Input - 1,3,1	317	105.6667
xa_nn_elm_quantize _f32_sym16	Channel = 3 1,3,8,4	1193	12.42708
- channel based	Channel = 1 Input - 1,12,64,197	97119	0.641914
xa_nn_elm_quantize _f32_sym16u	Channel = 3 Input - 1,3,8,4	1191	12.40625
- channel based	Channel = 1 Input - 1,12,64,197	97100	0.641788
	Input - 1,40,14,14	3246	0.414031
xa_nn_elm_dequantize_asym4_f32	Input - 1,3,1	259	86.33333
	Input - 1,40,14,14	3243	0.413648
xa_nn_elm_dequantize_asym4u_f32	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	Channel = 3 1,3,8,4	1011	10.53125
based	Channel = 1 Input - 1,12,64,197	58717	0.388094
xa_nn_elm_dequantize_asym4u_f32 - channel	Channel = 3 Input - 1,3,8,4	1013	10.55208
based	Channel = 1 Input - 1,12,64,197	58730	0.388179
	Input - 1,40,14,14	3244	0.413776
xa_nn_elm_dequantize_asym8_f32	Input - 1,3,1	260	86.66667
	Input - 1,40,14,14	3244	0.413776
xa_nn_elm_dequantize_asym8u_f32	Input - 1,3,1	261	87
xa_nn_elm_dequantize_asym8_f32	Channel = 3 1.3.8.4	1010	10.52083
- channel based	·	58743	0.388265
xa_nn_elm_dequantize_asym8u_f32		1012	10.54167
- channel based		58780	0.38851
	Input - 1,40,14,14	3251	0.414668
xa_nn_elm_dequantize_asym16_f32	Input - 1,3,1	267	89
	Input - 1,40,14,14	3252	0.414796
xa_nn_elm_dequantize_asym16u_f32	Input - 1,3,1	268	89.33333
xa nn elm dequantize asym16 f32	Channel = 3 1,3,8,4	1006	10.47917
- channel based	Channel = 1 Input - 1,12,64,197	58744	0.388272
xa_nn_elm_dequantize_asym16u_f32	Channel = 3 Input - 1,3,8,4	1021	10.63542
- channel based	Channel = 1 Input - 1,12,64,197	58757	0.388358
	Input - 1,40,14,14	2255	0.287628
xa_nn_elm_dequantize_sym4_f32	Input - 1,3,1	247	82.33333
	Input - 1,40,14,14	2260	0.288265
xa_nn_elm_dequantize_sym4u_f32	Input - 1,3,1	250	83.33333



Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
4 700	Channel = 3	957	9.96875
xa_nn_elm_dequantize_sym4_f32 – channel based	1,3,8,4 Channel = 1 Input - 1,12,64,197	39763	0.262816
xa_nn_elm_dequantize_sym4u_f32	Channel = 3 Input - 1,3,8,4	963	10.03125
- channel based	Channel = 1 Input - 1,12,64,197	39820	0.263193
	Input - 1,40,14,14	2261	0.288393
xa_nn_elm_dequantize_sym8_f32	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 = 3 1,3,8,4	963	10.03125
- channel based	Channel = 1 Input - 1,12,64,197	39763	0.262816
xa_nn_elm_dequantize_sym8u_f32	Channel = 3 Input - 1,3,8,4	962	10.02083
- channel based	Channel = 1 Input - 1,12,64,197	39796	0.263034
	Input - 1,40,14,14	2257	0.287883
xa_nn_elm_dequantize_sym16_f32	Input - 1,3,1	249	83
	Input - 1,40,14,14	2267	0.289158
xa_nn_elm_dequantize_sym16u_f32	Input - 1,3,1	251	83.66667
xa_nn_elm_dequantize_sym16_f32	Channel = 3 1,3,8,4	971	10.11458
- channel based	Channel = 1 Input - 1,12,64,197	39768	0.262849
xa_nn_elm_dequantize_sym16u_f32	Channel = 3 Input - 1,3,8,4	982	10.22917
- channel based	Channel = 1 Input - 1,12,64,197	39830	0.263259
	Input1 - 1,3,8,4 Intput2 - 1,3,8,4	158	1.645833
xa_nn_elm_sub_32x32_32	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	14216	0.251842
	Input2 - 1	400	0.711000
	Input1 - 224,1 Input2 - 1	160	0.714286
	Input1 - 1,8,512	1766	0.431152
	Input2 - 1,1,512	1700	0.101102
xa_nn_elm_sub_broadcast_5D_32x32_32	Input1 - 1,8,3,3	885	12.29167
Xd_III_GIII_3db_bloddcd3t_3b_32X32_32	Input2 - 3,3	10011	4.055000
	Input1 - 4,6,10,10	10214	4.255833
	Input2 - 4,6,10,1 Input1 - 1,3,8,4		
	Intput2 - 1,3,8,4	148	1.541667
xa_nn_elm_sub_f32xf32_f32	Input1 - 1,256,28,28		
xa_IIII_eIIII_Sub_I32xI32_I32	Input2 - 1,256,28,28	75376	0.375558
	Input1 - 39,39	740	0.400774
	Input2 - 39,39 Input1 - 1,3,1	713	0.468771
	Input2 - 1,3,1	86	28.66667
	Input1 - 1,288,14,14		20.00001
xa_nn_elm_sub_scalar_f32xf32_f32	Input2 - 1	14216	0.251842
	Input1 - 224,1		
	Input2 - 1	160	0.714286
	Input1 - 1,8,512 Input2 - 1,1,512	1749	0.427002
	Input1 - 1,8,3,3	1743	0.427002
xa_nn_elm_sub_broadcast_5D_f32xf32_f32	Input2 - 3,3	856	11.88889
	Input1 - 4,6,10,10		
	Input2 - 4,6,10,1	11914	4.964167
vo no alm div 20v20 f20 made 0	Input1 - 1,4,8,8 Input2 - 1,4,8,8	896	3.5
xa_nn_elm_div_32x32_f32 – mode 0 – high performance	Input1 - 1,16,112,112	090	3.5
riigii porioriilarios	Input2 - 1,16,112,112	552128	2.750956633
	Input1 - 1,4,8,8		
xa_nn_elm_div_scalar_32x32_f32 – mode 0	Input2 - 1	215	0.83984375
high performance	Input1 - 1,16,112,112	50327	0.250752352
	Input2 - 1 Input1 - 1,16,1,1	30327	0.250752352
	Input2 - 1,16,56,56	266695	5.315190529
xa_nn_elm_div_broadcast_5D_32x32_f32	Input1 - 1,3,8,4		
– mode 0 – high performance	Input2 - 1,3,1,4	1701	17.71875
	Input1 - 4,6,10,10	04004	0.4005
	Input2 - 4,6,10,1 Input1 - 1,4,8,8	21894	9.1225
xa_nn_elm_div_32x32_32 - mode 1	Input2 - 1,4,8,8	970	3.7890625
- high performance	Input1 - 1,16,112,112		
V r	Input2 - 1,16,112,112	602314	3.001006457
	Input1 - 1,4,8,8	000	4 4474075
xa_nn_elm_div_scalar_32x32_32 – mode 1 - high performance	Input2 - 1 Input1 - 1,16,112,112	286	1.1171875
- nign penomiance	Input1 - 1,10,112,112	100510	0.500787229
	I II Pate I	100010	0.000101220



Low-Level Kernel name	Parameters	Average Cycles	Performance Metric (cycles / output sample)
	Input1 - 1,16,1,1	054400	7.004004004
vo no alm div broadcost ED 20v20 20	Input2 - 1,16,56,56 Input1 - 1,3,8,4	354493	7.064991231
xa_nn_elm_div_broadcast_5D_32x32_32 - mode 1 - High performance	Input2 - 1,3,6,4	1866	19.4375
g portoniac	Input1 - 4,6,10,10		1011010
	Input2 - 4,6,10,1	29762	12.40083333
	Input1 - 1,4,8,8		0.040-
xa_nn_elm_div_32x32_32 – mode 2	Input2 - 1,4,8,8 Input1 - 1,16,112,112	976	3.8125
- high performance	Input2 - 1,16,112,112	602320	3.001036352
	Input1 - 1,4,8,8	002020	0.00100002
xa_nn_elm_ div_scalar_32x32_32 - mode 2	Input2 - 1	360	1.40625
- high performance	Input1 - 1,16,112,112	4	
	Input2 - 1	150696	0.750837054
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	360765	7.189991231
xa_nn_elm_div_broadcast_5D_32x32_32	Input1 - 1,3,8,4	300703	7.103331231
- mode 2 - high performance	Input2 - 1,3,1,4	1923	20.03125
	Input1 - 4,6,10,10		
	Input2 - 4,6,10,1	30722	12.80083333
vo nn alm div f20vf20 f20 mada 0	Input1 - 1,4,8,8	894	3.4921875
xa_nn_elm_div_f32xf32_f32 – mode 0 - high performance	Input2 - 1,4,8,8 Input1 - 1,16,112,112	094	3.4921073
mgn penemunoe	Input2 - 1,16,112,112	552126	2.750946668
	Input1 - 1,4,8,8		
xa_nn_elm_div_scalar_f32xf32_f32	Input2 - 1	201	0.78515625
– mode 0 - high performance	Input1 - 1,16,112,112	50242	0.050600507
	Input2 - 1 Input1 - 1,16,1,1	50313	0.250682597
	Input2 - 1,16,56,56	281948	5.619180485
xa_nn_elm_div_broadcast_5D_f32xf32_f32	Input1 - 1,3,8,4		
– mode 0 - high performance	Input2 - 1,3,1,4	1682	17.52083333
	Input1 - 4,6,10,10	07400	44.00044007
	Input2 - 4,6,10,1 Input1 - 1,4,8,8	27169	11.32041667
xa_nn_elm_div_f32xf32_f32 – mode 1	Input2 - 1,4,8,8	902	3.5234375
- high performance	Input1 - 1,16,112,112		0.020.0.0
	Input2 - 1,16,112,112	552134	2.750986527
	Input1 - 1,4,8,8	0.45	0.00004075
xa_nn_elm_div_scalar_f32xf32_f32 - mode 1 - high performance	Input2 - 1	215	0.83984375
- mode - mgn penormance	Input1 - 1,16,112,112 Input2 - 1	50327	0.250752352
	Input1 - 1,16,1,1	00021	0.200102002
	Input2 - 1,16,56,56	289116	5.762037628
xa_nn_elm_div_broadcast_5D_f32xf32_f32	Input1 - 1,3,8,4	,=:-	4= 000
– mode 1 - high performance	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	Input1 - 1,4,8,8	20040	12.02041001
- high performance	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	Input1 - 1,4,8,8 Input2 - 1	217	0.84765625
- mode 2 - high performance	Input1 - 1,16,112,112 Input2 - 1	50329	0.250762317
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	286428	5.708466199
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high performance	Input2 - 1,10,30,00 Input1 - 1,3,8,4 Input2 - 1,3,1,4	1688	17.58333333
- mode 2 - mgn penormanec	Input2 - 1,0,1,1 Input1 - 4,6,10,10 Input2 - 4,6,10,1	28369	11.82041667
xa_nn_elm_div_32x32_f32 - mode 0	Input2 - 4,8,8 Input2 - 1,4,8,8	896	3.5
- high precision	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552128	2.750956633
xa_nn_elm_div_scalar_32x32_f32 – mode 0	Input1 - 1,4,8,8 Input2 - 1	556	2.171875
- high precision	Input1 - 1,16,112,112 Input2 - 1	276172	1.376016422
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	266695	5.315190529
xa_nn_elm_div_broadcast_5D_32x32_f32 – mode 0 – high precision	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	Input1 - 1,4,8,8 Input2 - 1,4,8,8	970	3.7890625
- high precision	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602314	3.001006457
xa_nn_elm_div_scalar_32x32_32 - mode 1	Input1 - 1,4,8,8 Input2 - 1	573	2.23828125
- high precision	Input1 - 1,16,112,112 Input2 - 1	301245	1.500941685
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	354493	7.064991231
xa_nn_elm_div_broadcast_5D_32x32_32 – mode 1 - high precision	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	Input1 - 1,4,8,8 Input2 - 1,4,8,8	976	3.8125
- high precision	Input1 - 1,16,112,112 Input2 - 1,16,112,112	602320	3.001036352
xa_nn_elmdiv_scalar_32x32_32 - mode 2	Input1 - 1,4,8,8 Input2 - 1	610	2.3828125
- high precision	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
	Input1 - 1,4,8,8		
xa_nn_elm_div_f32xf32_f32 - mode 0 - high precision	Input2 - 1,4,8,8 Input1 - 1,16,112,112	894	3.4921875
	Input2 - 1,16,112,112	552126	2.750946668
xa_nn_elm_div_scalar_f32xf32_f32	Input1 - 1,4,8,8 Input2 - 1	535	2.08984375
- mode 0 - high precision	Input1 - 1,16,112,112 Input2 - 1	263623	1.31349151
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	281948	5.619180485
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 0 - high precision	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	Input1 - 1,4,8,8 Input2 - 1,4,8,8	902	3.5234375
- high precision	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552134	2.750986527
xa_nn_elm_div_scalar_f32xf32_f32	Input1 - 1,4,8,8 Input2 - 1	541	2.11328125
– mode 1 - high precision	Input1 - 1,16,112,112 Input2 - 1	276157	1.375941685
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	289116	5.762037628
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 1 - high precision	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	Input1 - 1,4,8,8 Input2 - 1,4,8,8	901	3.51953125
- high precision	Input1 - 1,16,112,112 Input2 - 1,16,112,112	552133	2.750981545
xa_nn_elm_div_scalar_f32xf32_f32	Input1 - 1,4,8,8 Input2 - 1	534	2.0859375
– mode 2 - high precision	Input1 - 1,16,112,112 Input2 - 1	276150	1.375906808
	Input1 - 1,16,1,1 Input2 - 1,16,56,56	286428	5.708466199
xa_nn_elm_div_broadcast_5D_f32xf32_f32 – mode 2 - high precision	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
	Input - 1,3,8,4	589	6.135417
xa_nn_elm_exp_f32_f32	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
	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	454	0.50669643
xa_nn_slice – signed 8-bit	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
	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	454	0.506696
xa_nn_slice – unsigned 8-bit	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
	Input – 1,8,128 Start – 1 End – (-1) Step – 1 Slice dim - 1	468	0.52232143
xa_nn_slice – 16-bit	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
	Input – 1, 768, 196, Permute_vector - 2,1,0	93102	0.618503
xa_nn_permute – signed 8-bit	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
	Input – 1, 768, 196, Permute_vector - 2,1,0	93102	0.618503
xa_nn_permute – unsigned 8-bit	Input – 1000,768, Permute_vector - 1, 0, Input – 1,3,8,8	466635	0.607598
	Permute_vector - 3, 2, 1, 0	2473	12.88021
	Input – 1, 768, 196, Permute_vector - 1,0,2	9969	0.066227
	Input – 1, 768, 196, Permute_vector - 2,1,0	93722	0.622622
xa_nn_permute – 16-bit	Input – 1000,768, Permute_vector - 1, 0,	470507	0.612639
	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0 Input – 1, 768, 196,	2211	11.51563
	Permute_vector - 1,0,2	19377	0.128727
	Input – 1, 768, 196, Permute_vector - 2,1,0	101938	0.677203
xa_nn_permute – 32-bit	Input – 1000,768, Permute_vector - 1, 0,	508891	0.662618
•	Input – 1,3,8,8 Permute_vector - 3, 2, 1, 0	1806	9.40625
mo mo	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