

SPU-001 TC2

Sparse Processing Unit 1

Sparse Processing Unit 1

SPU-001 TC2 and Evaluation Board Datasheet

Version 0.8.0

FEMTOSENSE

The Future of AI is Sparse

SPU-001 TC2

Sparse Processing Unit 1

Disclaimer	3
Overview	5
Applications	5
Product Features	5
Precision Support	6
Layer and Operator Support	6
System Diagrams	7
Specifications	15
End-to-End Task Performance	17
Evaluation Board PCB Specifications	18
Contact Information	21
Notice	21

SPU-001 TC2

Sparse Processing Unit 1

Disclaimer

©Femtose Inc. (“Femtose”). All rights reserved.

No portion of this document may be reproduced or transmitted in any form without the express written permission of Femtose. Nothing contained in this document should be construed as granting any license or right to use proprietary information without express written permission of Femtose.

This version of the document supersedes all previous versions.

Notice

Femtose, to the fullest extent permitted by law, provides this document “as-is”, and disclaims all warranties, either express or implied, statutory or otherwise, including but not limited to the implied warranties of merchantability, non-infringement of third parties’ rights, and fitness for particular purpose. Femtose assumes no liability for any error in this document and for damages, whether direct, indirect, incidental, consequential or otherwise, that may result from such errors, including but not limited to loss of data or profits. The content in this document is subject to change without prior notice. Femtose reserves the right to make changes to said content without prior notification to users.

This datasheet contains preliminary information intended for design and evaluation. Information such as packaging dimensions and pinouts are subject to change in future revisions.

SPU-001 TC2

Sparse Processing Unit 1

Revision History

Version	Date	Notes
0.1.0	12/15/2022	Datasheet initialized for SPU-001 test chip 2 (TC2). The following specifications are estimates that will be validated and updated as simulation and silicon verification information become available.
0.2.0	1/6/2023	Updated pinout from TC1 reference to TC2 CSP form factor. Add projected TC2 application performance and efficiency metrics
0.3.0	2/3/2023	Update from EVK1 placeholder to preliminary EVK2 (EVB2)
0.3.1	2/23/2023	Added new operator support for sliding windowed attention and more spectral operations
0.4.0	3/24/2023	Update EVK2 (EVB2) PMOD board schematic
0.5.0	6/5/2023	Update packaging specifications
0.6.0	8/18/2023	Added QFN package information
0.7.0	9/20/2023	Updated QFN package and chip tray information and updated EVB2/EVB3 schematics
0.8.0	10/23/2023	Updated end-to-end task metrics

SPU-001 TC2

Sparse Processing Unit 1

Overview

The SPU-001 is an ultra-low-power AI co-processor designed to run sparse neural network inference in SWAP-constrained edge devices. It supports a wide variety of neural network layers and operators suitable for audio, speech, and general 1-D time series data. Native sparsity support allows weight matrices to be compressed over 90% in on-chip SRAM. Activations can be sparsified upwards of 90% as well, leading to 100x reduction in energy and 10x reduction in storage requirements when both forms of sparsity are present. While sparse neural networks will provide the best performance-energy tradeoff, SPU-001 is capable of running dense networks.

SPU-001 TC2 is a production-quality sample of the SPU-001 co-processor; the form factor and performance is what is expected from a volume run of SPU-001 co-processors.

EVB2/EVB3 is an eval board containing SPU-001 TC2, designed to slot into a PMOD SPI header. In addition to allowing the SPU to plug into a variety of microcontroller or FPGA development boards, the EVK allows the current draw of the SPU to be measured through its test points.

Applications

- Noise Reduction/Speech Enhancement
- Sound Event/Scene Classification
- Bio-signal Event Classification
- Intelligent Beamforming
- Other Speech/Audio Inference Tasks

Product Features

- Self-contained coprocessor with SPI interface to host processor
- 1 MB total SRAM for storing compressed neural network parameters
- 8 datapaths, each containing a vector processing unit
- Weight sparsity support
 - Custom compression scheme for parameter matrices
- Activation sparsity support
 - Zero skipping for node outputs
- Low leakage
- Power gating
 - Ultra-low-power sleep/retention mode for inactive cores
- Flexible 1.8-3.3 V DVDD I/O power
- 0.8V core power
- 1.56 mm x 2.24 mm x 0.4 mm pitch WLCSP15, or 4mm x 4mm x 0.5mm pitch QFN20 package
- 22nm ULL process

SPU-001 TC2

Sparse Processing Unit 1

Precision Support

Configuration	Weights	Activations
STD	INT4	INT8
EIGHTS	INT8	INT8
DBL	INT8	INT16

Layer and Operator Support

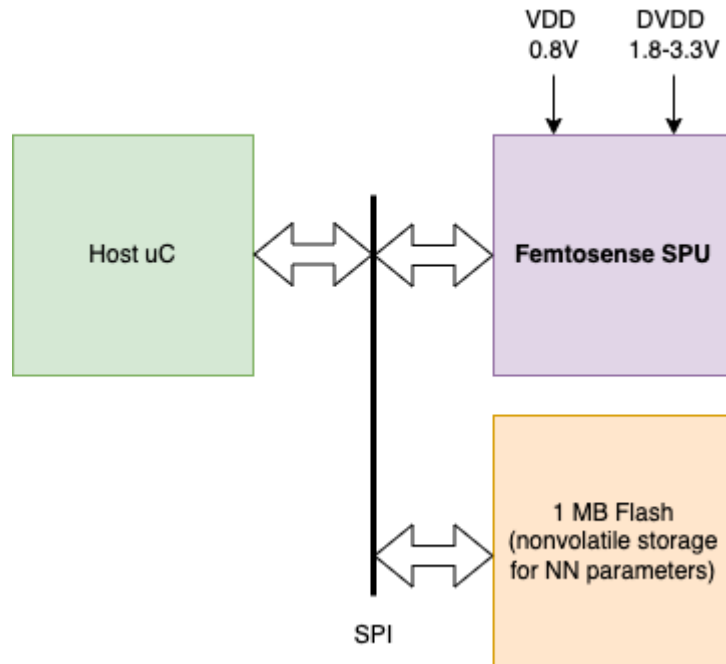
Nonlinear Functions	Unidirectional RNNs	Audio Features	Optimization Utilities
ReLU, PreLU, etc. Sigmoid Softmax Tanh Exp, Log Sin, Cos Reciprocal, Sqrt, etc.	Unidirectional RNN Unidirectional SRU Unidirectional GRU Unidirectional LSTM State Space Models (SSMs) <i>Custom Unidirectional RNN Layers</i>	FFT/IFFT RFFT/IRFFT STFT/ISTFT DCT MelFilterBank MFCC <i>arbitrary sized FFT</i>	Sparse Weights Sparse Activations Quantization-aware Training Heterogenous Precision Keras Compression Package PyTorch Compression Package
Dense Layers	Normalization Layers	Convolutional Layers	Causal Attention Layers
Fully Connected	LayerNorm BatchNorm (1D)	Temporal Conv1D (TCN) Depthwise TCN	Linear Attention Sliding Attention H3 SSM

SPU-001 TC2

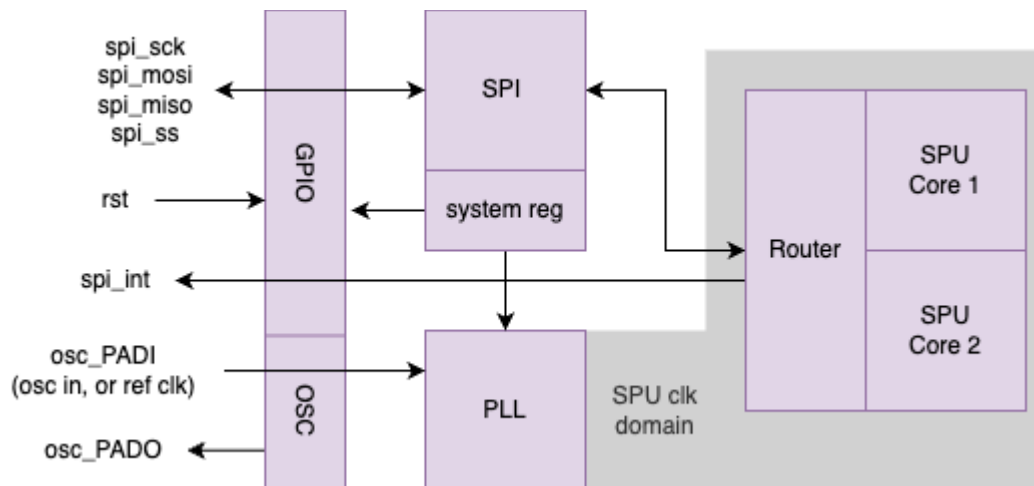
Sparse Processing Unit 1

System Diagrams

Example System Diagram



Internal Organization



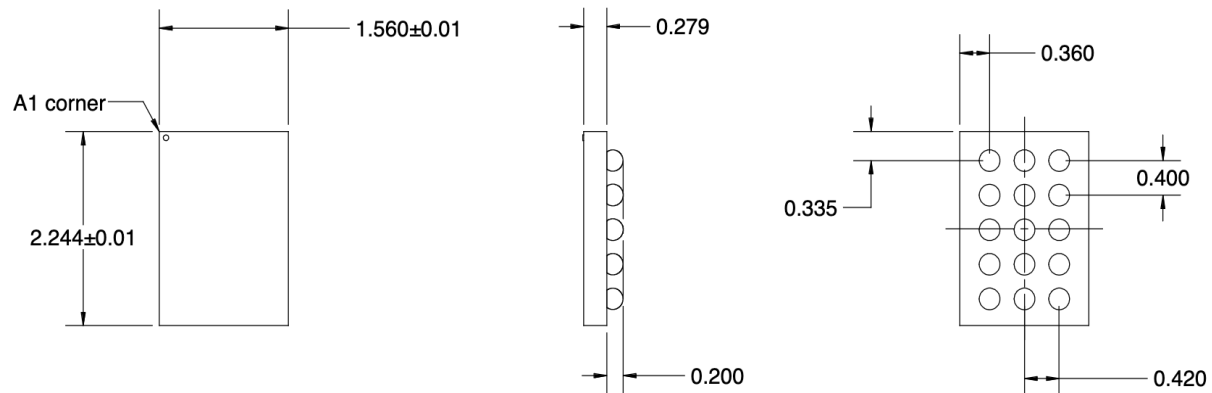
SPU-001 TC2

Sparse Processing Unit 1

POD Diagram - WLCSP

SPU-001-TC2-WLCSP

All dimensions in the drawing are mm. The view in the left panel is from the top, and the view in the right panel is from the bottom (ball array-side) of chip. The ball array is a regular grid at 420um x-pitch, 400um y-pitch. Note offset to pin 1: array is centered in width of the chip but slightly off-center in its height. Ball size is 250um. Die thickness (not including balls) is 11 mil.



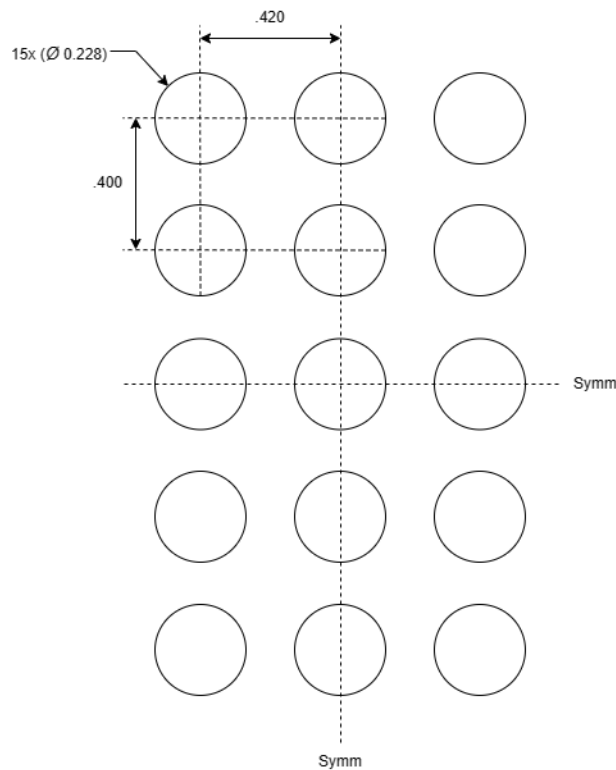
SPU-001 TC2

Sparse Processing Unit 1

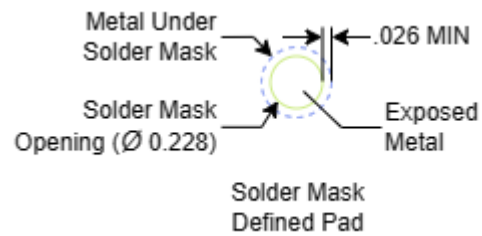
PCB Land Pattern and Fanout - WLCSP

SPU-001-TC2-WLCSP

All dimensions in the drawing are mm. Exposed pad size is recommended to be 228um. A solder mask defined (SMD) pad design is recommended on the right. Using the recommended SMD dimensions, a 280um via-in-pad can be used to fanout the 3 interior pads.



Land Pattern Example
(Exposed Metal Shown)



SPU-001 TC2

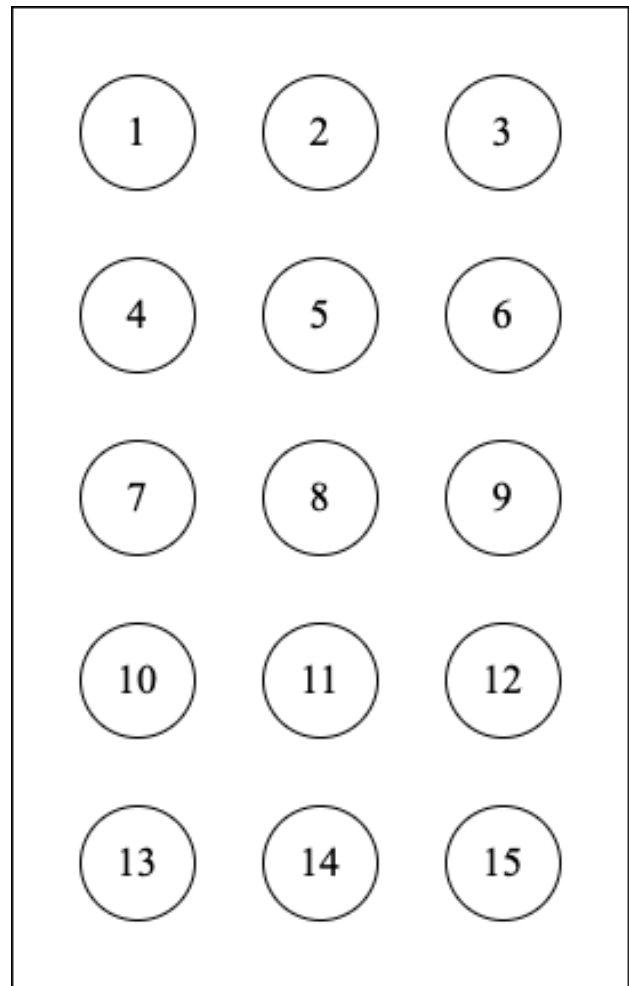
Sparse Processing Unit 1

Pinout - WLCSP

Note: view is from the top (looking “through” the chip).

PIN #	ID
1	SPI_MOSI
2	VDD
3	VSS
4	SPI_SCK
5	SPI_MISO
6	RST
7	SPI_SS
8	DVDD
9	VSSA
10	SPI_INT
11	VDDA
12	OSC_PADI
13	VSS
14	VDD
15	OSC_PAD

SPU-001-TC2-WLCSP



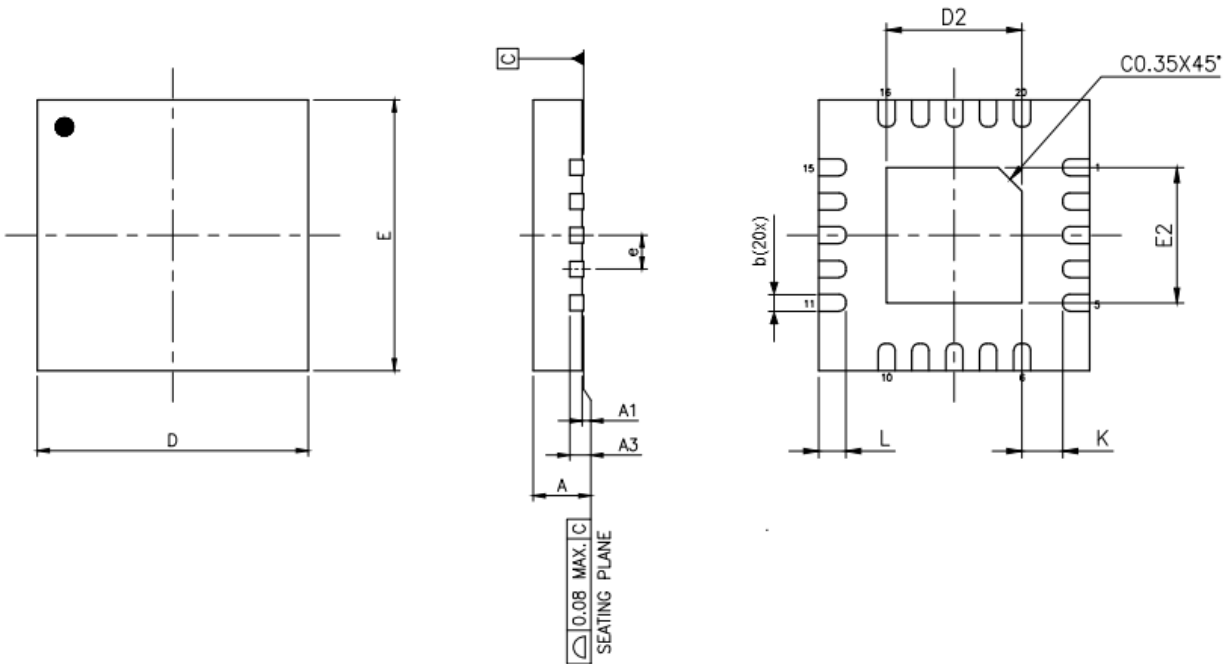
SPU-001 TC2

Sparse Processing Unit 1

POD Diagram - QFN

SPU-001-TC2-QFN

All dimensions in the drawing are mm. The view in the left panel is from the top, and the view in the right panel is from the bottom (pad-side) of the chip. The pads are 500um pitch, and the outline is 4x4mm. Package thickness is 33 mil.



SYMBOL	MIN	NOM.	MAX
A	0.80	0.85	0.90
A1	0.00	0.02	0.05
A3	0.203 REF.		
D	4.00 BSC		
E	4.00 BSC		
e	0.50 BSC		
K	0.20	-	-
b	0.20	0.25	0.30
D2	1.95	2.00	2.05
E2	1.95	2.00	2.05
L	0.30	0.40	0.50

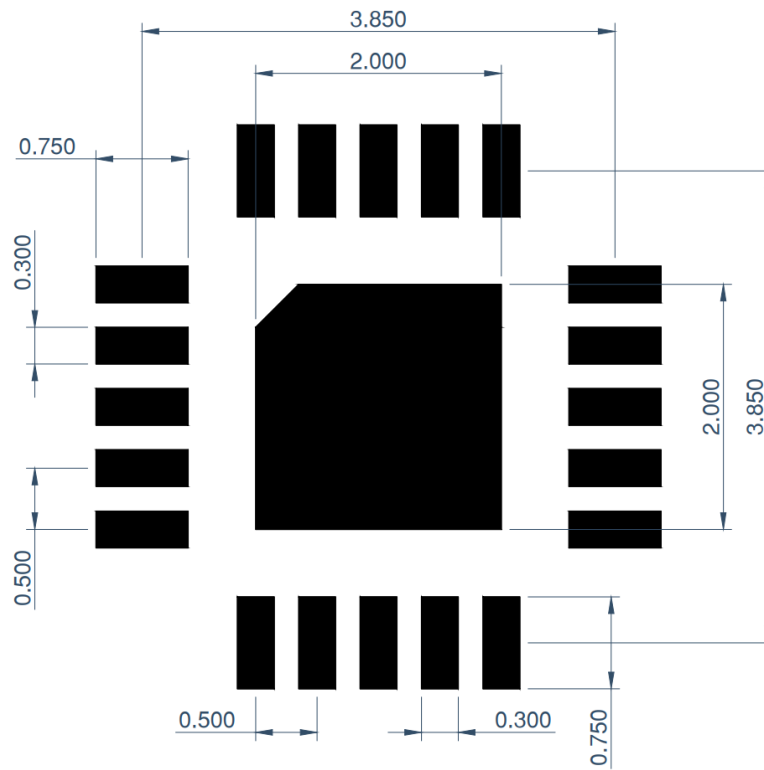
SPU-001 TC2

Sparse Processing Unit 1

PCB Land Pattern - QFN

All dimensions in the drawing are mm.

SPU-001-TC2-QFN



SPU-001 TC2

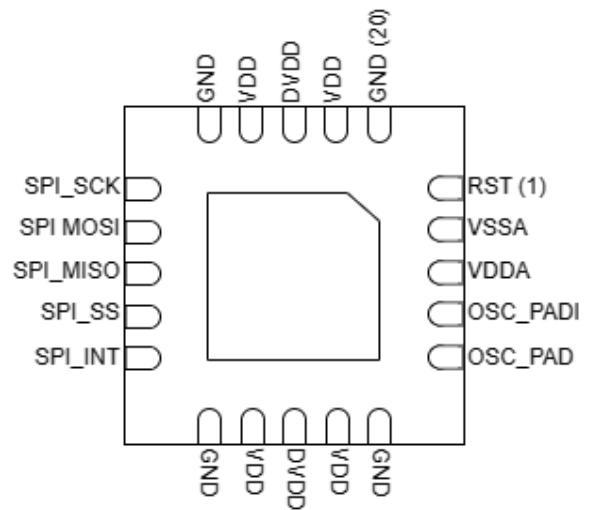
Sparse Processing Unit 1

Pinout - QFN

Note: view is from the bottom.

SPU-001-TC2-QFN

PIN #	ID
1	RST
2	VSSA
3	VDDA
4	OSC_PADI
5	OSC_PAD
6	GND
7	VDD
8	DVDD
9	VDD
10	GND
11	SPI_INT
12	SPI_SS
13	SPI_MISO
14	SPI_MOSI
15	SPI_SCK
16	GND
17	VDD
18	DVDD
19	VDD
20	GND
PAD	GND



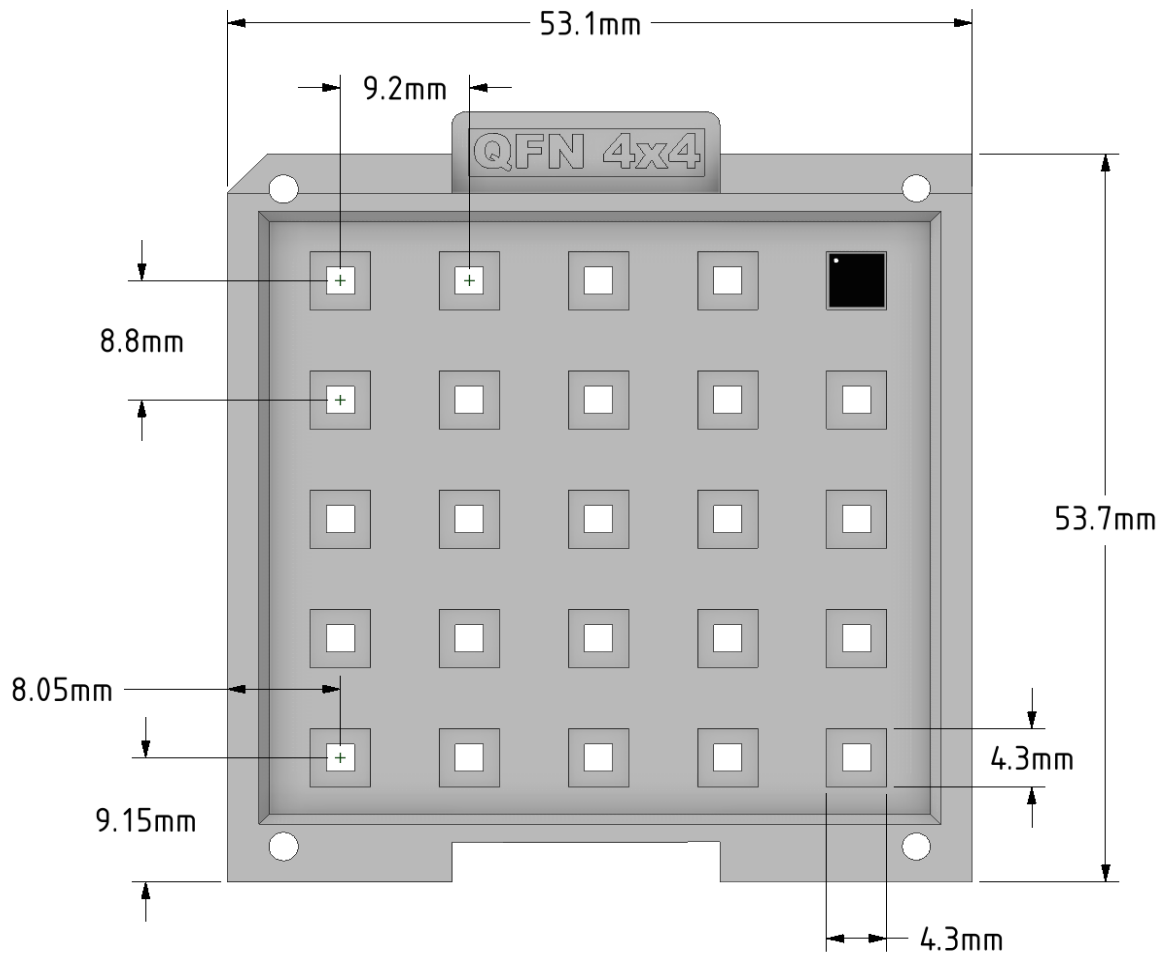
SPU-001 TC2

Sparse Processing Unit 1

Packaging Tray- QFN

Note: view is from the top.

SPU-001-TC2-QFN



SPU-001 TC2

Sparse Processing Unit 1

Specifications

NOTE: 12/15/2022: Numbers reflect pre-silicon projections for SPU-001 test chip 2 (“TC2”), but grounded by post-silicon measurements from test chip 1 (“TC1”).

Absolute Maximum Ratings

Parameter	Rating
Storage Temperature	TBD
Device Voltage, V_{dd}	+0.88 V

Recommended Operating Conditions

Parameter	Min	Typical	Max	Unit
T_{case}	-40	25	85	°C

Electrical Specifications

Parameter	Description	Conditions	Min	Typical	Max	Unit
V_{dd}	Core voltage ¹		0.72	0.80	0.88	V
V_{ddIO}	IO Pad voltage		1.8		3.3	V
I_{core}	Peak core current ²	$V_{dd} = 0.8 \text{ V}$, $F_{VCO} = 300$			60	mA
I_{IO}	Peak IO pad current ³				4	mA
P_{leak_min}	Minimum leakage power	25C, chip powered, PLL off, osc off, memories off		61		μW
P_{leak_max}	Maximum leakage power	25C, chip powered, PLL off, osc off, memories on		170		μW
P_{active_dp}	Active power per datapath			21		μW/MHz

¹ Part should work for V_{dd} within range, but max operating frequency is not guaranteed away from typical operating point. Qualification pending.

² Estimated; scales roughly linearly with PLL VCO frequency

³ Estimated; assuming medium pad drive strength. Only a single output pin should only ever be simultaneously switching

SPU-001 TC2

Sparse Processing Unit 1

Clocking Specifications

Parameter	Conditions	Min	Typical	Max	Unit
PLL VCO Frequency (core frequency)				300	MHz
SPI_sck Frequency (IO frequency)				50	MHz
PLL lock time ⁴	1 MHz ref clk.			< 500	μs
PLL max multiplier				8192	

Maximum Performance

Metric	Conditions	Value	Unit
Raw Computational Efficiency	300 MHz VCO, int4 weights, int8 activations	380	GOPS/W
Effective Computational Efficiency	90% weight and activation sparsity, 300 MHz, int4 weights, int8 activations	38	ETOPS/W
Raw Throughput	300MHz, int 4 weights, int8 activations	19.2	GOPS
Effective Throughput	90% weight and activation sparsity, 300 MHz, int4 weights, int8 activations	1.9	ETOPS

⁴ PLL lock time is linearly related to ref clock frequency. Capped at $T_{ref} \times 500$, but should be lower in practice

SPU-001 TC2

Sparse Processing Unit 1

End-to-End Task Performance

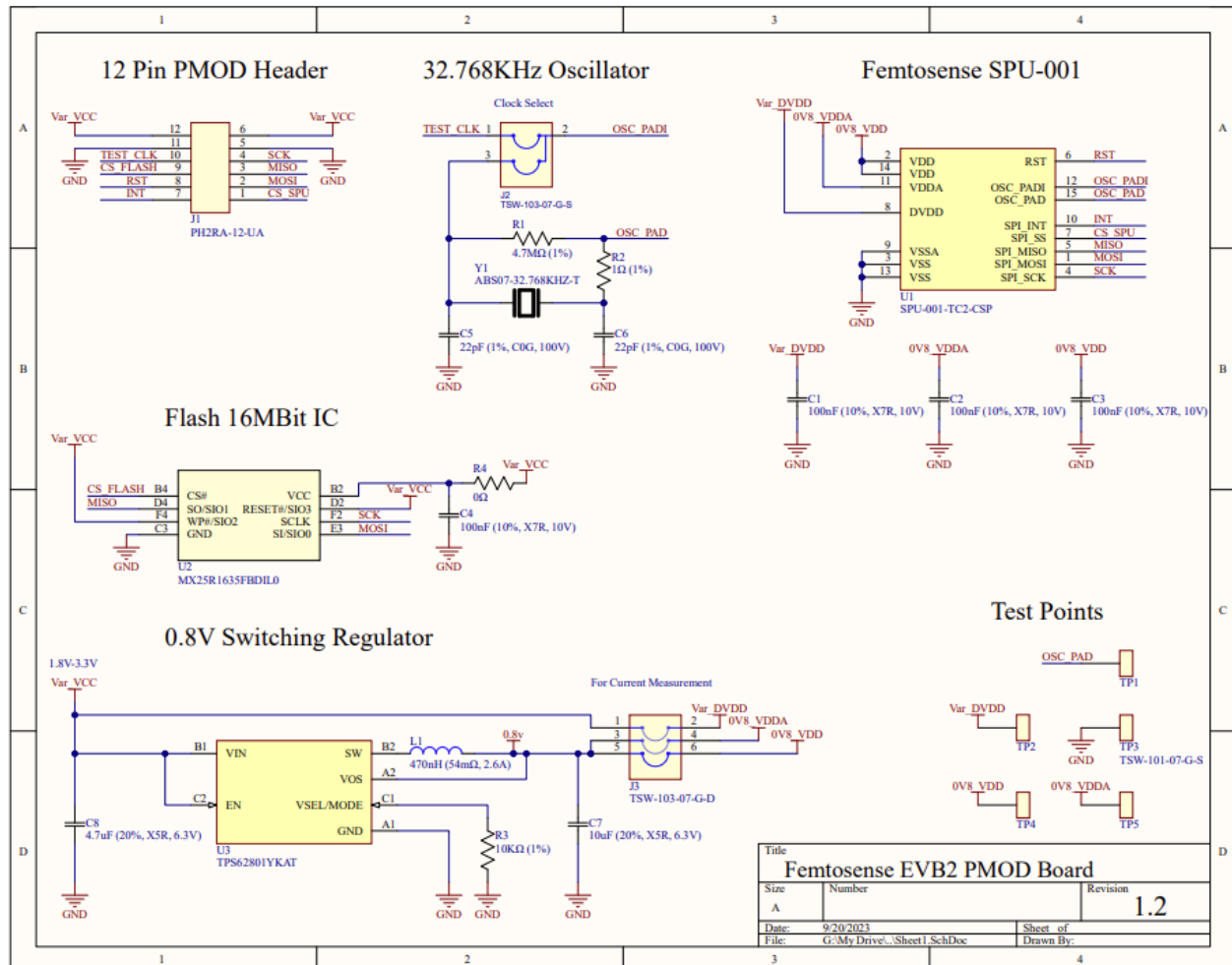
Task (Dataset)	Model Version	Use Case	Model Architecture	Performance (Metric)	Power (VDD)	Latency (algorithm)	Model Size
Wakeword Detection (Hey Snips)	WWDSNIPS_8khz_16ms_v2	Always-on voice wakeup	FemtoseNSE sparse GRU with spectral frontend	99.5% (F1-score)	193 μ W	16 ms	76.6 kB
Low Latency Speech Enhancement (Custom)	AINRGP_16khz_4hop_8algo_v3	Intelligent transparency mode, speech enhancement for hearing aids	FemtoseNSE proprietary DNN	7.2 dB (SISDRi, Caf�� Env.)	1.41 mW	8 ms	635 kB
Ultra-low Latency Speech Enhancement (Custom)	AINRGP_16khz_1hop_2algo_v0	Intelligent transparency mode, speech enhancement for hearing aids	FemtoseNSE proprietary DNN	5.0 dB (SISDRi, Caf�� Env.)	3.42 mW	2 ms	322 kB
Keyword Spotting (Google Speech Commands)	GSC_8khz_16ms_v0	Streaming local voice commands (end to end)	FemtoseNSE Dense LSTM with spectral frontend	88.88% (F1 Score)	461 μ W	16 ms	525 kB

SPU-001 TC2

Sparse Processing Unit 1

Evaluation Board PCB Specifications

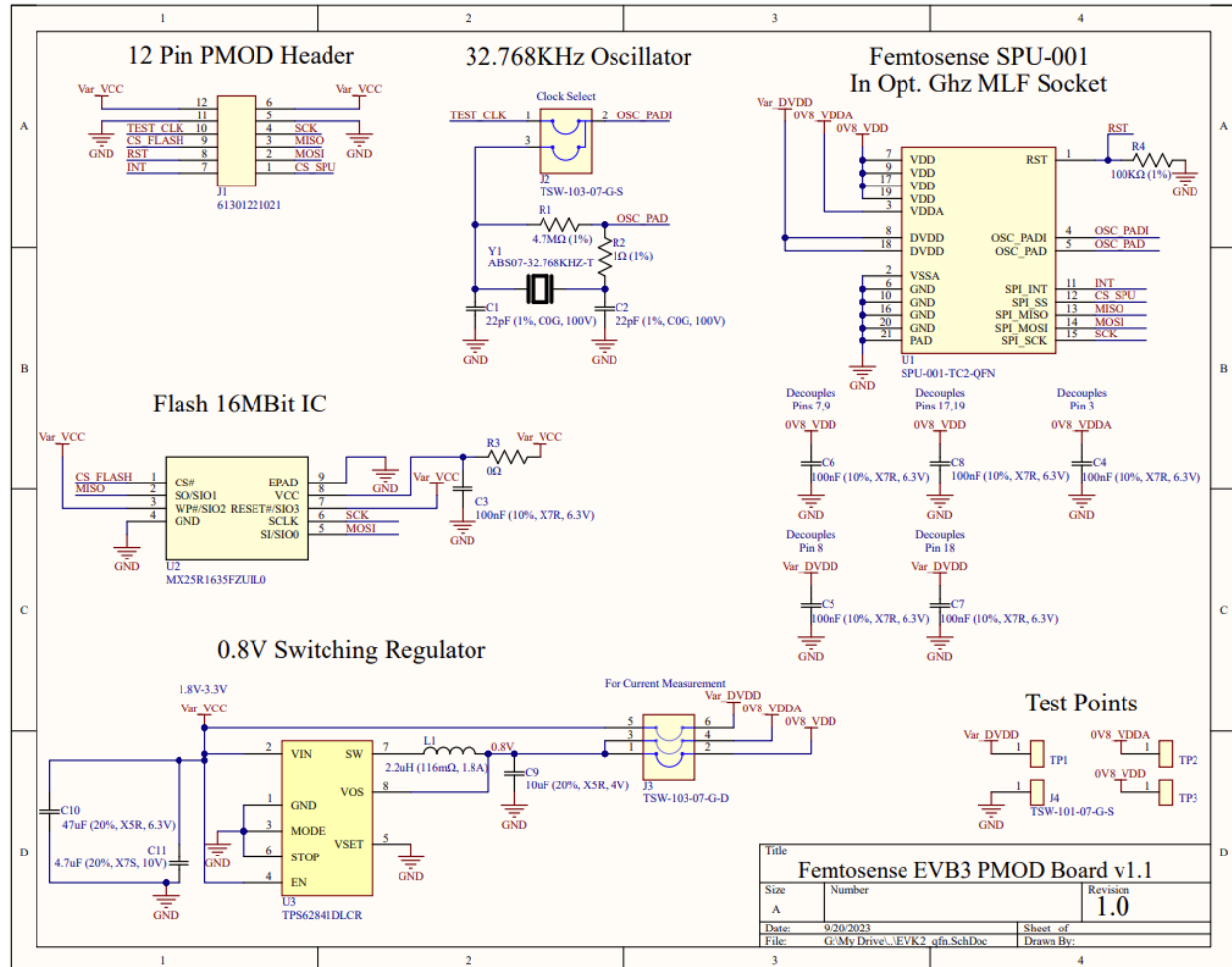
EVB2 Schematic (WLCSP)



The crystal circuit containing R1, R2, Y1, C5, C6 is optional, as the SPU can alternatively be clocked with the external signal TEST_CLK. In most applications, decoupling capacitor C1 is optional, and C2 can be combined with C3 if J3 is omitted.

Sparse Processing Unit 1

EVB3 Schematic (QFN)



The crystal circuit containing R1, R2, Y1, C1, C2 are optional, as the SPU can alternatively be clocked with the external signal TEST_CLK.

SPU-001 TC2

Sparse Processing Unit 1

EVB2/EVB3 PMOD Header Pinout Key

Digilent specification: [PMOD Pinout Key](#)

EVB2/EVB3 uses PMOD Interface Type 2A (expanded SPI)

Pin #	Signal	Direction	Alt. Signal	EVB2/EVB3 Signal
1	CS	Out		CS_SPU
2	MOSI	Out		MOSI
3	MISO	In		MISO
4	SCK	Out		SCK
5	GND			GND
6	VCC			VCC
7	GPIO	In/Out	INT	INT
8	GPIO	In/Out	RESET	RST
9	GPIO	In/Out	CS2	CS_FLASH
10	GPIO	In/Out	CS3	TEST_CLK
11	GND			GND
12	VCC			VCC

SPU-001 TC2

Sparse Processing Unit 1

CS_SPU = Chip select. Active low to enable slave device
MOSI = Master Out Slave In. Data from master to slave.
MISO = Master In Slave Out. Data from slave to master.
SCK = Serial Clock. Master provides the clock to shift data.
INT = Interrupt signal from slave to master.
RST = SPU reset
TEST_CLK = reference clock from host board. Jumper J3 must be set accordingly.

All Dimensions are in millimeters unless stated otherwise.

Contact Information

For the latest specifications, additional product information, clarification, worldwide sales and distribution locations, and information about Femtosense:

Web: www.femtosen.se.ai **Email:** info@femtosen.se.ai

Notice

The information contained herein is believed to be reliable. Femtosense makes no warranties regarding the information contained herein. Femtosense assumes no responsibility or liability whatsoever for any of the information contained herein. Femtosense assumes no responsibility or liability whatsoever for the use of the information contained herein. The information contained herein is provided "AS IS, WHERE IS" and with all faults, and the entire risk associated with such information is entirely with the user. All information contained herein is subject to change without notice. Customers should obtain and verify the latest relevant information before placing orders for Femtosense products. The information contained herein or any use of such information does not grant, explicitly or implicitly, to any party any patent rights, licenses, or any other intellectual property rights, whether with regard to such information itself or anything described by such information. Femtosense products are not warranted or authorized for use as critical components in medical, life-saving, or life-sustaining applications, or other applications where a failure would reasonably be expected to cause severe personal injury or death.