

# SPU-001

Sparse Processing Unit 1

---

## Sparse Processing Unit 1

*SPU-001*

*Version 0.2.0*

FEMTOSENSE

Sparse AI That Makes Sense

# SPU-001

## Sparse Processing Unit 1

---

Disclaimer.....	3
Overview.....	5
Applications.....	5
Product Features.....	5
Precision Support.....	5
Layer and Operator Support.....	6
System Diagrams.....	7
Specifications.....	13
End-to-End Task Performance.....	15
Contact Information.....	16
Notice.....	16

# SPU-001

## Sparse Processing Unit 1

---

### Disclaimer

©Femtosense, Inc. (“Femtosense”). All rights reserved.

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

This version of the document supersedes all previous versions.

### Notice

Femtosense, 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. Femtosense 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. Femtosense 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

## Sparse Processing Unit 1

---

### Revision History

Version	Date	Notes
0.1.0	1/25/2024	Initialized production part datasheet
0.2.0	3/18/2024	MP power measurements, T&R Packaging Info, reference to integration guide, replaced EVB4 info with typical application circuits

# SPU-001

## Sparse Processing Unit 1

---

### Overview

The SPU-001 is an ultra-low-power AI co-processor designed to run sparse neural network inference in size, weight, power, and cost-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.

### 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
- 1.59 mm x 2.27 mm x 0.4 mm pitch WLCSP15.
- 22nm ULL process

### Precision Support

# SPU-001

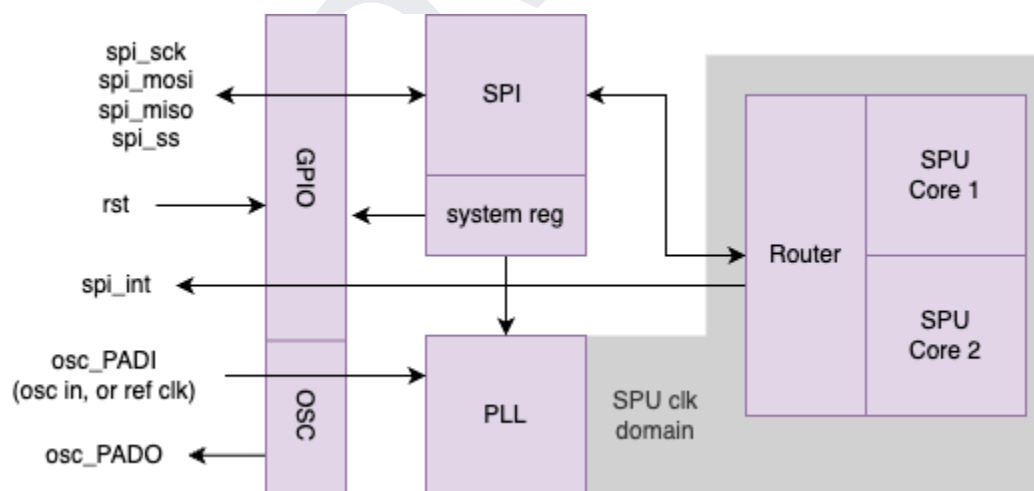
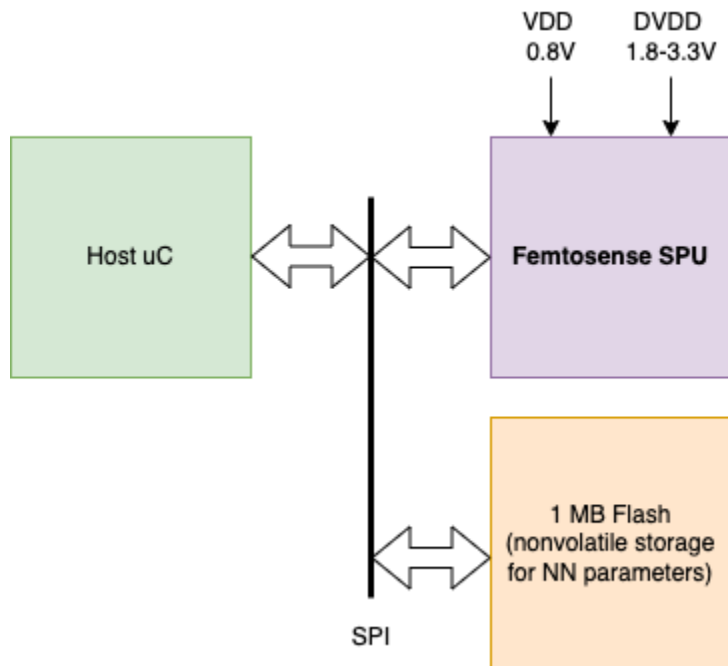
## Sparse Processing Unit 1

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

## Sparse Processing Unit 1

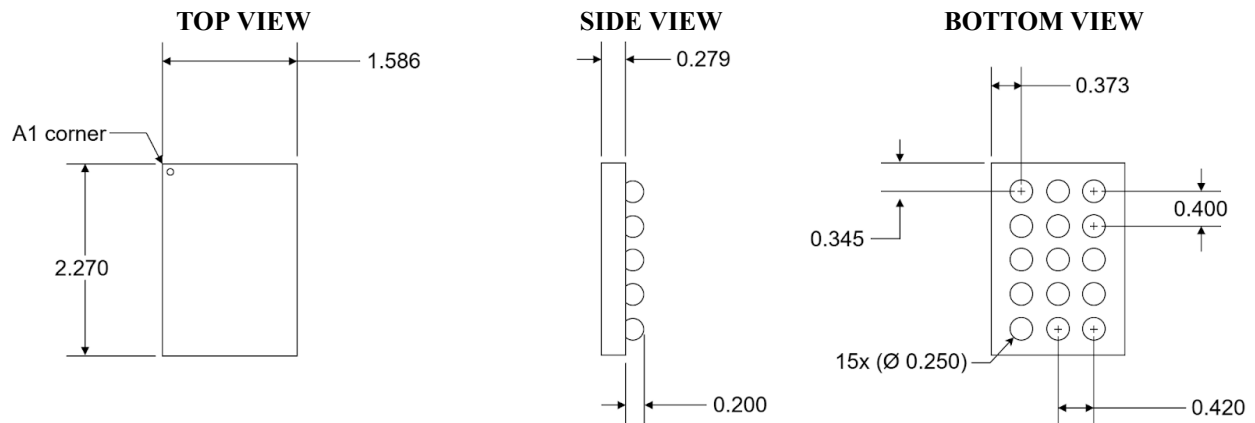


# SPU-001

## Sparse Processing Unit 1

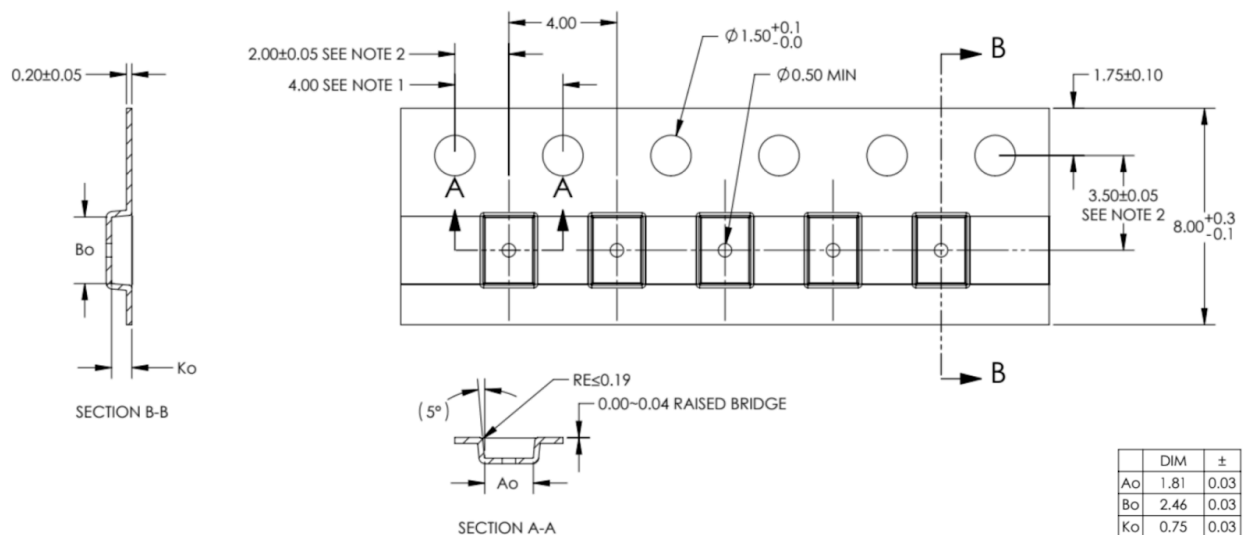
### POD Diagram

All dimensions in the drawing are mm. The ball array is a regular grid at 420μm x-pitch, 400μm y-pitch. Note that the ball array is centered in width of the chip, but slightly off-center in height by 10μm.



### Tape & Reel Packaging

All dimensions in the drawing are mm.



- NOTES:
1. 10 SPROCKET HOLE PITCH CUMULATIVE TOLERANCE ±0.2
  2. POCKET POSITION RELATIVE TO SPROCKET HOLE MEASURED AS TRUE POSITION OF POCKET, NOT POCKET HOLE.
  3. A<sub>o</sub> AND B<sub>o</sub> ARE MEASURED ON A PLANE AT A DISTANCE "R" ABOVE THE BOTTOM OF THE POCKET.

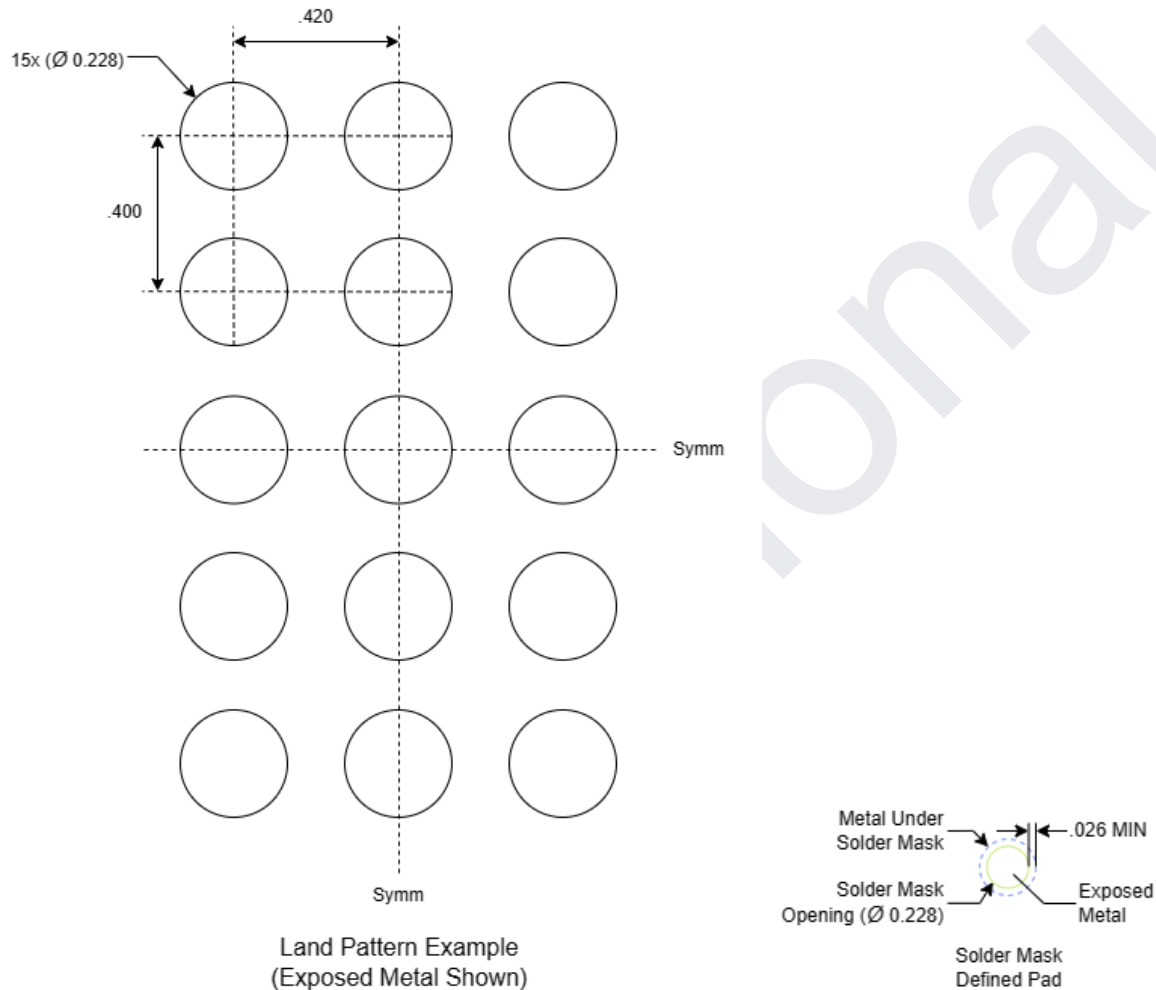


# SPU-001

## Sparse Processing Unit 1

### PCB Land Pattern and Fanout

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



### RoHS assembly

The solder ball material is lead-free SAC405 and should be assembled at higher temperatures appropriate for RoHS compliance.

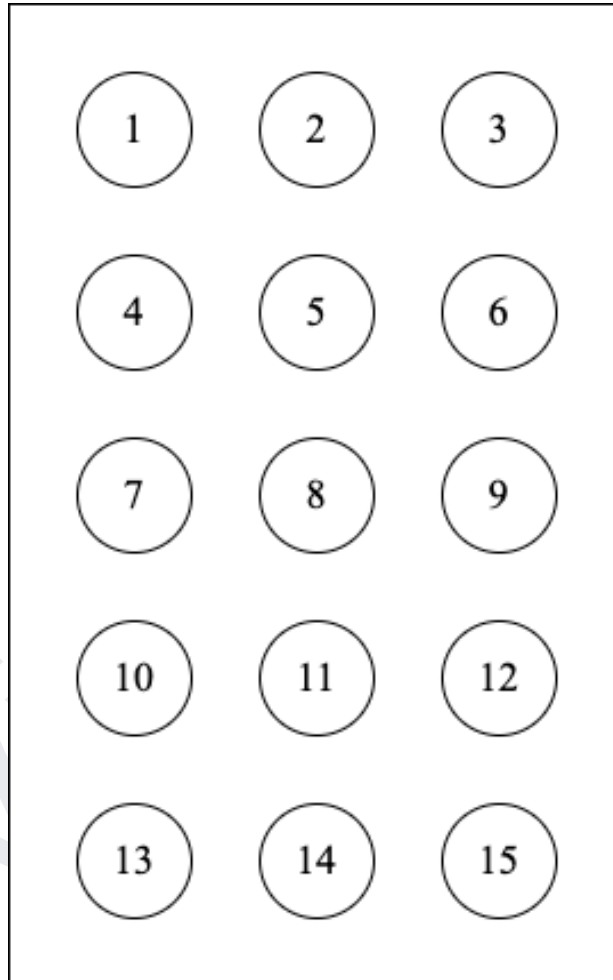
# SPU-001

## Sparse Processing Unit 1

### Pinout

Note: view is from the **top** (looking down “through” the chip), equivalent to PCB pad layout.

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



**TOP VIEW**

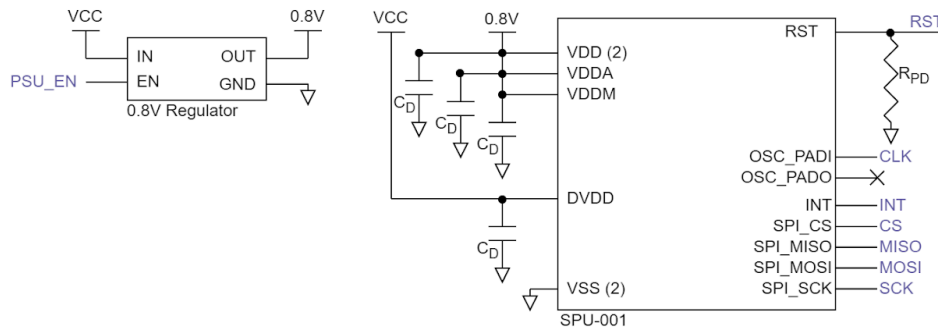
# SPU-001

## Sparse Processing Unit 1

### Typical Application Circuit

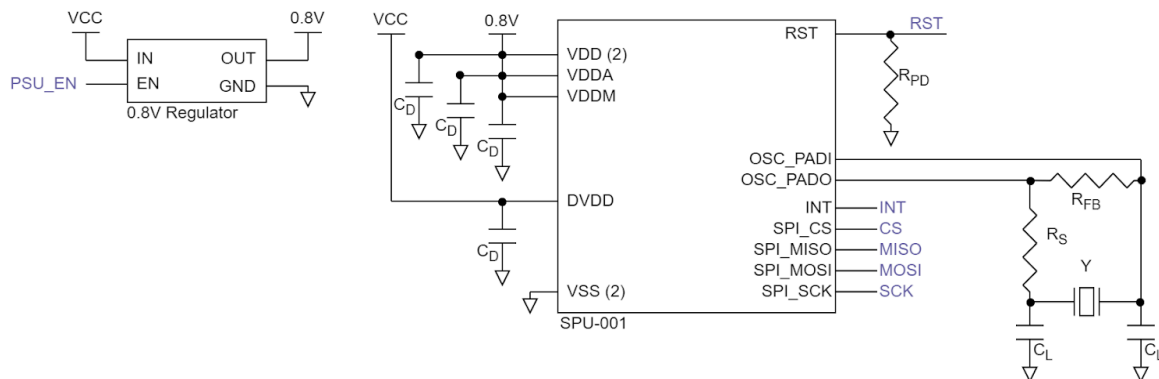
A full integration guide covering hardware and firmware integration is available in the separate document “SPU Integration Guide.” A summary of the typical application circuit is also given below.

The following schematic shows a typical application circuit when the SPU is clocked by an external reference clock (CLK). Purple signals represent SPU control IO from adjacent systems (e.g. host MCU). **PSU\_EN** is optional, and only required if your system is sensitive to boot-up power consumption so that the SPU boot sequence can be precisely controlled. Ideally, the SPU’s power rails are brought up with the reset held. This could help prevent an indeterminate state (with indeterminate power consumption) if the SPU is powered up while not under reset.



If the interior VDDA, VDDM, and INT pads cannot be practically routed out with your PCB technology, INT should be left disconnected, and VDDA, VDDM should be connected to VDD. In this case, only one  $C_D$  is needed for the combined VDD/VDDA/VDDM connection (in addition the  $C_D$  on DVDD), and the INT can be accessed via a SPI register read instead. More information about this layout is available in the “SPU Integration Guide” document.

For a crystal oscillator reference, the schematic below shows a typical application circuit:



The following values are used in the schematics:

# SPU-001

## Sparse Processing Unit 1

---

Symbol	Value	Comment
C <sub>D</sub>	100nF	+/-20%
C <sub>L</sub>	22pF	+/-1%
Y	32.7680KHZ	12.5pF load, 70KΩ ESR
R <sub>FB</sub>	4.7MΩ	+/-1%
R <sub>S</sub>	1Ω	+/-1%
R <sub>PD</sub>	100KΩ	Optional
VCC	IO Voltage (host voltage)	0.8V - 3.3V

Additional details about alternative crystals or reference clocking architectures are included in the “SPU Integration Guide” document.

# SPU-001

## Sparse Processing Unit 1

### Specifications

**NOTE:** As noted elsewhere, many figures are provisional, subject to complete characterization. Figures are measured from silicon unless otherwise noted.

#### Absolute Maximum Ratings

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

#### Recommended Operating Conditions

Parameter	Min	Typical	Max	Unit
$T_{case}$	-20	25	85	°C

#### Electrical Specifications

Parameter	Description	Conditions	Min	Typical	Max	Unit
$V_{dd}$	Core voltage <sup>1</sup>		0.72	0.80	0.88	V
$V_{ddIO}$	IO Pad voltage		1.62	1.8-3.3	3.63	V
$I_{core}$	Peak VDD current <sup>2</sup>	$V_{dd} = 0.8$ V, $F_{VCO} = 300$			60	mA
$I_{IO}$	Peak DVDD current <sup>3</sup>				4	mA
$P_{leak\_min}$	Always-on Leakage <sup>4</sup>	25C, chip powered, PLL off, osc off, memories off		80		μW
$P_{leak\_max}$	Leakage with all domains powered <sup>5</sup>	25C, chip powered, PLL off, osc off, memories on		220		μW

<sup>1</sup>Max frequency (PLL VCO frequency) will only be guaranteed within a smaller range. Pending characterization.

<sup>2</sup>Peak current scales roughly linearly with PLL VCO frequency

<sup>3</sup>Assuming medium pad drive strength. Only a single output pin should only ever be simultaneously switching

<sup>4</sup>Typical part. VDD + VDDM + VDDA currents. Pending further characterization

<sup>5</sup>Typical part. VDD + VDDM currents. Pending further characterization

# SPU-001

## Sparse Processing Unit 1

### Clocking Specifications

Parameter	Conditions	Min	Typical	Max	Unit
PLL VCO Frequency (core frequency)	$V_{dd} > 0.76V$ , -20C			200 <sup>6</sup>	MHz
SPI_sck Frequency (IO frequency)				50	MHz
PLL lock time <sup>7</sup>	1 MHz ref clk.			< 500	$\mu s$
PLL max multiplier				8192	

### Maximum Performance

Metric	Conditions	Value	Unit
Raw Computational Efficiency	200 MHz <sup>8</sup> VCO, int4 weights, int8 activations	250	GOPS/W
Effective Computational Efficiency	90% weight and activation sparsity, 200 MHz <sup>9</sup> , int4 weights, int8 activations	25	ETOPS/W
Raw Throughput	200MHz <sup>10</sup> , int 4 weights, int8 activations	12.8	GOPS
Effective Throughput	90% weight and activation sparsity, 200 MHz <sup>11</sup> , int4 weights, int8 activations	1.28	ETOPS

<sup>6</sup> Final binning strategy is TBD, final characterization pending. There will be a bin at least this fast, given a -20C, -5% VDD worst-case operating point.

<sup>7</sup> PLL lock time is linearly related to ref clock frequency. Capped at  $T_{ref} \times 500$ , but should be lower in practice

<sup>8</sup> Final binning strategy TBD. Faster bins that exceed this performance are planned

<sup>9</sup> Final binning strategy TBD. Faster bins that exceed this performance are planned

<sup>10</sup> Final binning strategy TBD. Faster bins that exceed this performance are planned

<sup>11</sup> Final binning strategy TBD. Faster bins that exceed this performance are planned

# SPU-001

## Sparse Processing Unit 1

### End-to-End Task Performance

Task (Dataset)	Model Version	Use Case	Model Architecture	Performance (Metric)	Power (VDD+VDDM)	Latency (algorithm)	Model Size
Ultra-low Power Speech Enhancement (Custom)	AINRGP_16khz_4hop_8algo_v4	Intelligent transparency mode, speech enhancement (TWS earbuds, hearing aids, headsets, etc.)	FemtoseNSE proprietary DNN	6.4 dB (SISDRi, Café Env.) 11.1 dB (SISDRi, Car Env.)	809 $\mu$ W	8 ms	615 kB
Ultra-low Latency Speech Enhancement (Custom)	AINRGP_16khz_1hop_2algo_v2	Intelligent transparency mode, speech enhancement (TWS earbuds, hearing aids, headsets, etc.)	FemtoseNSE proprietary DNN	7.8 dB (SISDRi, Café Env.) 14.1 dB (SISDRi, Car Env.)	3.4 $\mu$ W	2 ms	738 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)	403 $\mu$ W	32 ms	525 kB
Wakeword Detection (Alexa)	WWD ALEXA_8khz_16ms_v0	Streaming wake word detection (end to end)	FemtoseNSE Pruned LSTM with spectral frontend	98.8% (F1 Score)	165 $\mu$ W	32 ms	212 kB
Spoken Language Understanding (SLU) - English	pre-release	Streaming local intent detection (end to end)	FemtoseNSE Pruned LSTM with spectral frontend	pre-release	404 $\mu$ W	32 ms	pre-release

# SPU-001

## Sparse Processing Unit 1

---

### Contact Information

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

**Web:** [www.femtosome.ai](http://www.femtosome.ai)

**Email:** [info@femtosome.ai](mailto:info@femtosome.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.