

## ZenDNN 6.0 Support Matrix

	zentr	zentr
<b>Release Version</b>	<ul style="list-style-type: none"> <li>2.21.0.0</li> <li>2.20.0.0</li> </ul>	<ul style="list-style-type: none"> <li>2.12.0.2</li> <li>2.11.0.2</li> </ul>
<b>Release Date</b>	Jul, 2026	Jul, 2026
<b>manylinux Support (Based on Alma Linux GCC version 13.3)</b>	manylinux_2_28	manylinux_2_28
<b>glibc Version</b>	2.28	2.28
<b>Python Version</b>	<ul style="list-style-type: none"> <li>3.13 through 3.10 for 2.21.0.0</li> <li>3.13 through 3.9 for 2.20.0.0</li> </ul>	3.13 through 3.10
<b>Native Framework versions</b>	<ul style="list-style-type: none"> <li>TensorFlow 2.21.0 through 2.16.1</li> </ul>	<ul style="list-style-type: none"> <li>PyTorch 2.12.0, 2.11.0</li> <li>vLLM 0.23.0 through 0.20.0</li> </ul>
<b>Compiler</b>	GCC v13.3	GCC v13.3
<b>Wheel File Release</b>	Yes	Yes
<b>Wheel file built with ZenDNN 6.0.0 and AOCL-DLP 5.2+</b>	Yes	Yes
<b>Wheel File Built with FBGEMM 1.2.0</b>	Yes	Yes
<b>C++ Interface Release</b>	Yes	No
<b>Java Interface Support</b>	Yes, v1.2.0-SNAPSHOT	No
<b>Compatible with OS</b>	<ul style="list-style-type: none"> <li>Ubuntu 22.04, 24.04</li> <li>RHEL 9.2, 9.5, 10.2</li> </ul>	<ul style="list-style-type: none"> <li>Ubuntu 22.04, 24.04</li> <li>RHEL 9.2, 9.5, 10.2**</li> </ul>
<b>Not Compatible with OS</b>	N/A	N/A
<b>Compatible H/W</b>	4 <sup>th</sup> , 5 <sup>th</sup> , and 6 <sup>th</sup> Gen AMD EPYC™ processors	4 <sup>th</sup> , 5 <sup>th</sup> , and 6 <sup>th</sup> * Gen AMD EPYC™ processors
<b>Data Types</b>	<ul style="list-style-type: none"> <li>FP32</li> <li>BF16</li> </ul>	<ul style="list-style-type: none"> <li>FP32</li> <li>FP16</li> <li>BF16</li> <li>A16W4</li> <li>DA8W8</li> </ul>
<p>** Use libstdc++.so.6 library which supports GLIBCXX_3.4.30</p> <p>* This release introduces limited functional support for running LLMs using float16 precision with vLLM on 6<sup>th</sup> Gen AMD EPYC™ processors.</p> <p>A16W4 - A16 refers to activation as BF16 dtype: we support both asymmetric and symmetric Quantization for weights</p> <p>DA8W8 - DA8 is dynamic INT8 quantization of activation: we support symmetric quantization for activation and weights.</p>		
<p><b>System's GCC version should be greater than or equal to 9.4 for torch.compiler flow</b></p>		