MIPS ARC VPX DSP IP Processor:

AI Data Compression Option

Highlights

- Enhanced data efficiency
- Improved computational performance
- Compliance and standardization
- Simplified memory management
- Resource optimization

STU conversion

- Conversion of INT4 formats to/from INT8
- Conversion of BF16 formats to/from FP32
- Conversion of OCP-FP8 formats to FP32 or FP16
- Conversion of OCP-MX formats to FP32

Target Applications

- Large Language Models (LLM) across multiple applications

Overview

The AI Data Compression Option for MIPS ARC® VPX DSP IP offers advanced data conversion and compression capabilities, enhancing the efficiency and performance of AI and machine learning (ML) applications, including large language models (LLMs). For system-on-chip (SoC) designers, the option provides more efficient data handling, improved computational performance, and simplified memory management, all while ensuring compliance with industry standards.

BF16, OCP-FP8 and OCP-MX Support in ARC VPX DSP Processors

The AI data compression option for ARC VPX, when combined with the vector floating point unit (FPU) option for VPX, adds support for INT4, BF16, OCP-FP8 and OCP-MX data compression to any ARC VPX DSP IP Processor.

BFloat16 (BF16) is a 16-bit floating point format comprised of one sign bit, eight exponent bits, and seven mantissa bits. It maintains the number range from the 32-bit IEEE 754 single-precision floating-point format (FP32) while reducing the precision from 24 bits to 8 bits.

OCP 8-bit Floating Point (OCP-FP8) provides a standardized format for floating point encodings in 8-bit.

OCP-MX is a micro-scaling format that compresses data by sharing the common part of the exponent across a set of 32 elements, which helps reduce the memory footprint and improve hardware performance and efficiency, reducing overhead and operational costs.

A typical use case for OCP-MX is in applications involving LLMs. It helps in efficiently reading dictionaries and reading/writing the key and value (KV) cache of the underlying transformer, which are fundamental components in LLM architectures.

The AI Data Compression Option for VPX supports vectorization of INT4, BF16, OCP-FP8 and OCP-MX data types in the channel dimension used for computing inner products during matrix multiplication, which is fundamental in neural network (NN) computations.

The AI Data Compression Option for ARC VPX DSP Processors is fully compliant with the OCP specifications based on:

- OCP 8-bit Floating Point Specification (OFP8), Rev. 1.0, Approved: June 20, 2023
- OCP Microscaling Formats (MX) Specification, Version 1.0, Sep 2023

	XM (DDR/CSM)	L1 (UCCM)	XM -> L1	L1 -> XM
	INT4	INT8	Sign-extend	Saturate
	BF16	FP32	Mantissa pad	Round, Saturate
OCP-FP8	E5M2	FP16 | FP32	Mantissa pad, exponent bias offset	not supported
	E4M3	FP16 | FP32		
OCP-MX	MXE5M2	FP32		not supported
	MXE4M3	FP32		
	MXE3M2	FP32		
	MXE2M2	FP32		
	MXE2M1	FP32		
	MXEINT8	FP32		

Table 1: Supported data-type conversions

All data format conversions are performed on the fly in the DMA, uncompressing data when moving from system memory into the VCCM and compressing data when moving from VCCM to system memory.

Hardware Features for OCP-MX

Scaling Factors Storage

The OCP-MX scaling factors are stored in CSM or DDR memory in line with the 32 elements.

Tensor Padding

The OCP-MX tensors are padded to a multiple of 32 elements+1 scale per 32 elements to ensure alignment and consistency. The inline storage of scaling factors and consistent tensor padding simplify memory management, reducing latency and improving access times.

Deliverables

- RTL supporting the mentioned data-type conversions.
- Confidence test library (CCTs), demonstrating DMA conversions

About MIPS:

MIPS by GlobalFoundries delivers software to silicon with RISC-V for building physical AI platforms. MIPS delivers software-hardware co-design, optimized AI, and custom ASSP design and manufacturing. Together with ARC, MIPS delivers the open, standards-based processor IP portfolio for embedded applications. Physical AI is built on MIPS.

For more information, visit **www.mips.com/arc.**

MIPS ARC by GF © MIPS 2026 All Rights Reserved | MIPS by GlobalFoundries. **www.mips.com**