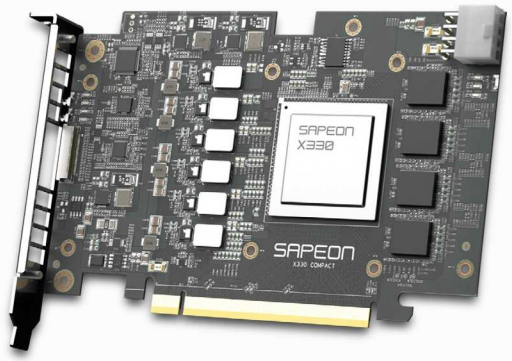
SAPEON
X330 Product brief

X330 can accelerate inference workloads for a variety of AI models, including large language models, with very low power consumption. With a design based on the 7nm process, X330 increases computing performance by up to four times compared to previous products.
Nevertheless, thanks to its power-efficient design, the increase in peak power usage is minimized, and power consumption is effectively controlled for every workload. Accordingly, X330 is the most powerful and efficient AI processor to choose from servers that require AI inference performance.

X330 supports not only integer operations but also floating-point operations, allowing it to respond to a variety of AI inference workloads.
In particular, it supports a variety of 8-bit and 16-bit formats, including FP8 format, providing a range of options for more efficient AI inference. Additionally, X330 is equipped with a video codec that can process 4K 60fps multi-stream to efficiently support multi-modal AI. As a result, the diverse spectrum of AI inference workloads that our customers demand can be supported by X330.

X330 is provided as a PCIe card for easy use in servers and data centers. X330 Compact is a more efficient card with the PCIe FHHL standard, while X330 Prime provides more powerful performance with PCIe FHFL standard. Both X330 Compact and X330 Prime connect quickly to the host CPU with the PCIe Gen 5 Interface. Depending on the customer's usage environment, two types of PCIe cards can be selected to build the most efficient AI infrastructure.



X330 provides the most efficient AI inference performance for servers and data centers

X330 Compact | Prime Specifications


		X330 Compact	X330 Prime
Precision		FP 16/8 bit, INT 8 bit	
8 bit performance		367 TFLOPS	734 TFLOPS
Memory	Type	GDDR6 X 8 (ECC)	GDDR6 X 16 (ECC)
	Capacity	16 GB	32 GB
	Bandwidth	256 GB/s	512 GB/s
Host interface		PCIe Gen5 16 Lane	
TDP		75~120 W	250 W
Form factor		FHHL Single	FHFL Single
CODEC		Encoder: H264 / VP8 / MPEG-4 Decoder: HEVC / H264 / VP8 / MPEG-4	

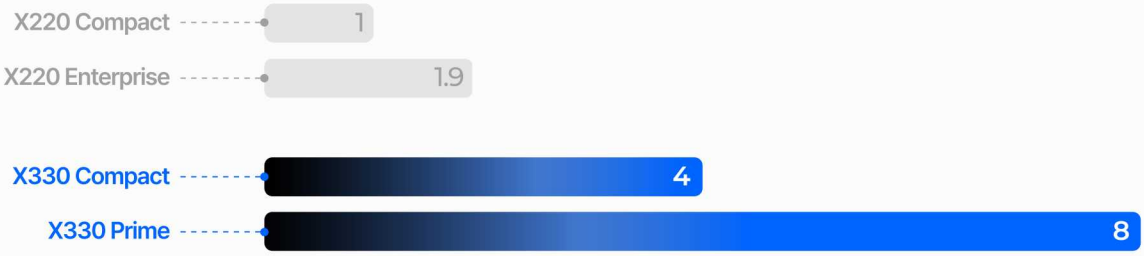
All products are subject to change without prior notice

X330 Architecture: The most efficient and comprehensive architecture design for AI inference

X330's Architecture is designed to provide hardware support for the functions required for AI inference operations in data centers. X330 integrates four powerful AI cores with 64K MAC matrix blocks (MXC) and 16 neural vector processors (NVP). To reduce the burden on the host for artificial intelligence application processing and to manage X330's resources efficiently, 16 RISC-V-based CPUs are included while a high-performance video codec cluster are equipped to support video streams of 4K 60fps. The chip also includes a layout and format converter (LFCVT) for versatile input and output processing. This architectural design makes X330 a comprehensive single-chip solution for AI computation in server environments.

AI inference performance comparison

Resnet-50 v1.5 performance



Up to four times more powerful performance than previous products

According to the results of testing models used by MLPerf™, which provides a standard benchmark for measuring machine learning system performance, X330 Compact and X330 Prime each provide up to four times faster performance than their predecessors. This proves that X330 Compact and X330 Prime provide powerful performance not only in theoretical computing performance but also in real-world use cases.

Convenient, developer-friendly Zero-touch™ AI full stack

SAPEON's goal is to minimize the time it takes for customers to deploy AI model networks on X330. X330's SDK supports AI models based on industry standard ONNX. Framework formats such as TensorFlow or Pytorch can be converted to ONNX using a variety of open-source converters. Our compiler automatically converts our customers' AI models into a format suitable for running on X330. After this, the SDK performs a series of processes for AI inference, including optimization.

In addition to SDK for convenient use of X330, SAPEON separately provides Artiference™ and Trainer™, SW platforms for AI full stack, to help customers provide more efficient inference services. Artiference™ is a cloud serving platform that allows customers to easily deploy SAPEON-based inference services. Trainer™ helps developers easily create an efficient neural network for X330 by supporting retraining of existing models when developing new applications.