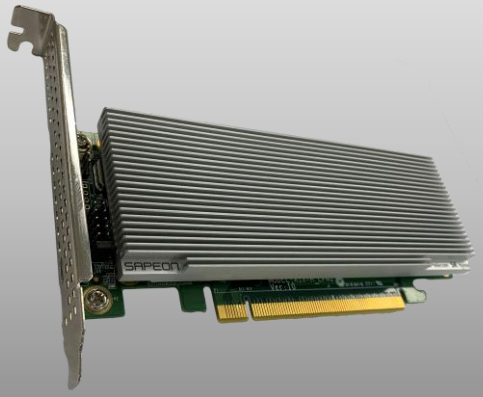
SAPEON

The Fastest Hyper-Cloud AI



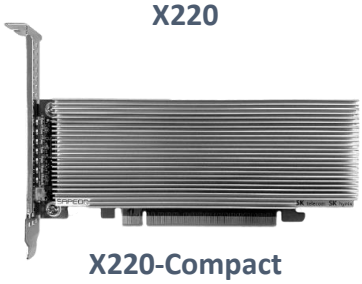
AI Inference Accelerator
SAPEON X220

SAPEON is the brand name of the AI accelerator chip for datacenter developed by SKT for the first time in Korea, and X220 is the model name of the first AI chip.

SAPEON X220 provides the optimal solution for cloud-based AI services by accelerating deep learning inference computation with low latency and high throughput.

SAPEON features high level of flexibility to run most neural networks and supports various deep learning frameworks. Once a neural network is developed and trained with a deep learning framework, SAPEON is able to run the model seamlessly without modification of the model.

X220



X220-Compact
(PCIe Low-Profile)

Computation Capability (INT8)	**87 TOPS** Boost 100 TOPS	Effective Performance (Resnet-50 Inference)	**6.7K FPS**
Precision	**INT16/8/4**	Host Interface	**PCIe Gen3 x16**
Memory Capacity	**8 GB**	Form Factor	**PCIe Low-Profile**
Memory Bandwidth	**42 GB/s**	TDP	**65 W**

KEY FEATURES

Since AI computing is too expensive due to the algorithm complexity, large-scale AI services are impracticable. Based on distinguished deep learning acceleration technologies, SAPEON enables high quality AI services to serve massive number of users.

- Ultimately optimized NPU(Neural Processing Unit) architecture only for deep learning inference
- Low-latency and high-throughput for real-time services serving a large number of users
- No additional effort needed to port your algorithm to AIX (Supporting various deep learning frameworks)
- Supporting various standard neural networks with Model-Zoo