AIX v2: Flexible High Performance AI Inference Accelerator for Datacenters

SK telecom is Korea’s largest telecommunications company with more than 29 million mobile subscribers, which accounts for around 50 percent of the total market.

AIX v2 - AI Inference Accelerator for Datacenters

**Building a New ICT Ecosystem through Openness and Collaboration**

**AIX v2 Architecture Highlights**

**Dynamic Command Scheduler and Parallel Execution Units**

**Matrix Multiplication and Convolution Unit**

**Dynamic Compute Optimization**

- Computation Graph Optimization
- Quantization, Pruning
- Model Parallel Execution

**Static Compute Optimization**

- Compiled Dataflow Graph
- HW Scheduled Parallel Execution

**Dynamic Dataflow Graph Scheduling**

- 2xINT8 MACs per DSP
- Dynamic Precision with FP Scaler

- 775 MHz MAC Frequency
- 84% of DSPs

**Ideal TOPS Effective Performance (YOLO v2)**

- **Batch 128**
  - **TOPS**: 29
  - **FPS**: 1,587

**Parallel Computing w/ Dynamic Resource Clustering**

**Effective Per-Frame Tensor Cache Capacity**

**Maximized Computing Efficiency**

**Computation Unit Utilization**