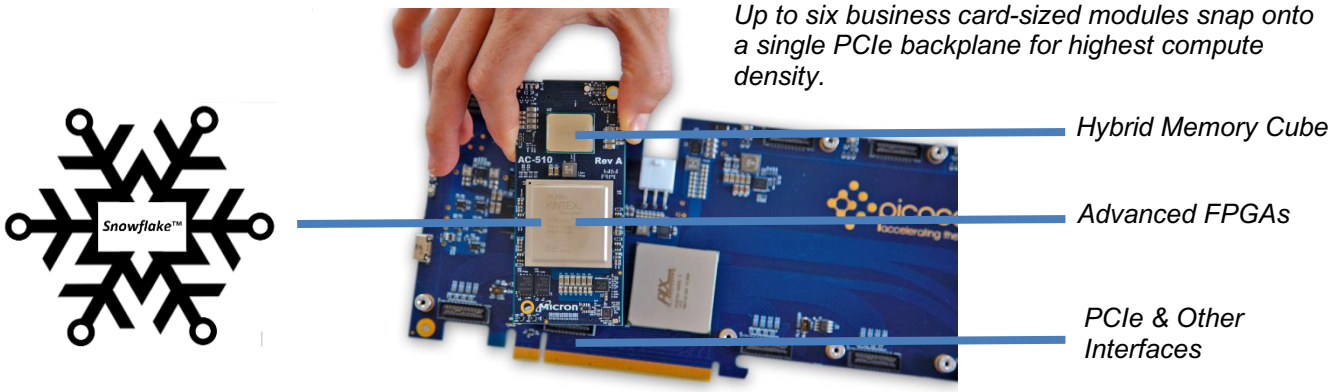
Snowflake™

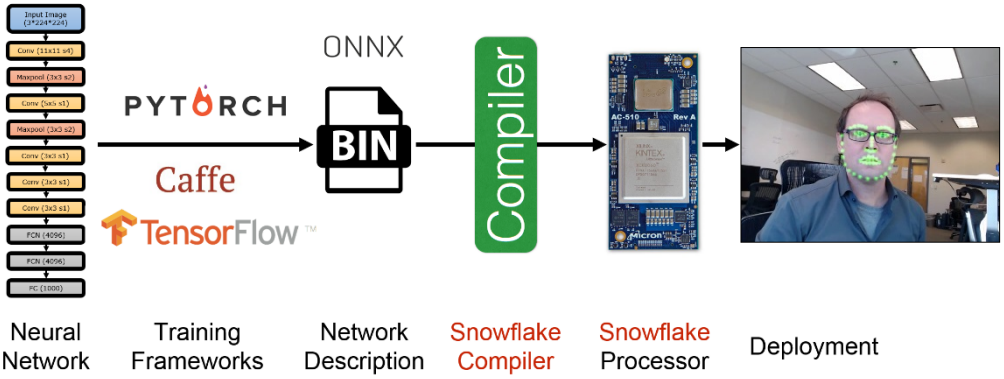
Deep Neural Network Accelerator

Machine Learning—Without the Need for Programming!

- **Best performance/Watt of any deep learning solution**
- **Framework-agnostic development pathways—PyTorch, TensorFlow, Caffe**
- **Easy-to-use compiler requires no Verilog/VHDL expertise**
- **Most efficient scalability with highest compute density**
- **Exploits the power of FPGAs with the ease of GPUs**

Our state-of-the-art deep learning solutions comprise a modular FPGA-based architecture with Micron's Hybrid Memory Cube running Forward Next's high-performance Snowflake neural network IP. Our fully integrated SDK takes trained neural network files and compiles them directly into the accelerator—*with no need for any programming*—enabling direct, rapid deployment from framework to application.

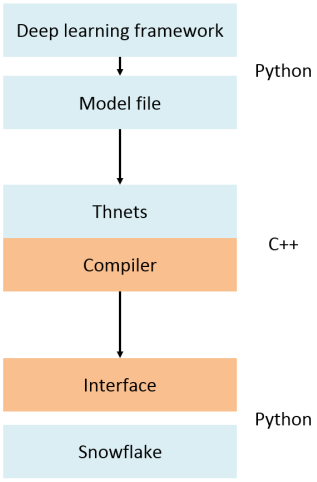




www.fwdnxt.com • info@fwdnxt.com • 1281 Win Hentschel Blvd. West, Suite E1935, Lafayette, IN 47906

OPERATION: COMPILER STEPS

- Model creation
- Parsing
- Partition + assignment
- Code generation
- Execute



SYSTEMS SUMMARIES

Snowflake model	512	1K	2K	3K
FPGA	Micron AC510	2x Micron AC510	4x Micron AC510	6x Micron AC510
Accelerator cores	512	1024	2048	3072
Clock Freq.	187 MHz	187 MHz	187 MHz	187 MHz
Peak Throughput	191 G-ops/s	383 G-ops/s	766 G-ops/s	1,148 G-ops/s
Memory	4 GB HMC	8 GB HMC	16 GB HMC	24 GB HMC
Memory B/W	60 GB/s	120 GB/s	240 GB/s	360 GB/s
Power system	24 W	48 W	96 W	144 W

Snowflake model	256	1K	2K	3K
FPGA	1x Xilinx ZC706	2x Micron AC510	4x Micron AC510	6x Micron AC510
Accelerator cores	256	1024	2048	3072
Clock Freq.	250 MHz	187 MHz	187 MHz	187 MHz
Peak Throughput	128 G-ops/s	383 G-ops/s	766 G-ops/s	1,148 G-ops/s
Memory	1GB DDR3	8GB HMC	16GB HMC	24GB HMC
Memory B/W	4.2 GB/s	18 GB/s	36 GB/s	54 GB/s
Power 1 board	12 W	24 W	24 W	24 W
Power system	12 W	48 W	96 W	144 W
Efficiency	99.3%	98.2%	98.2%	98.2%