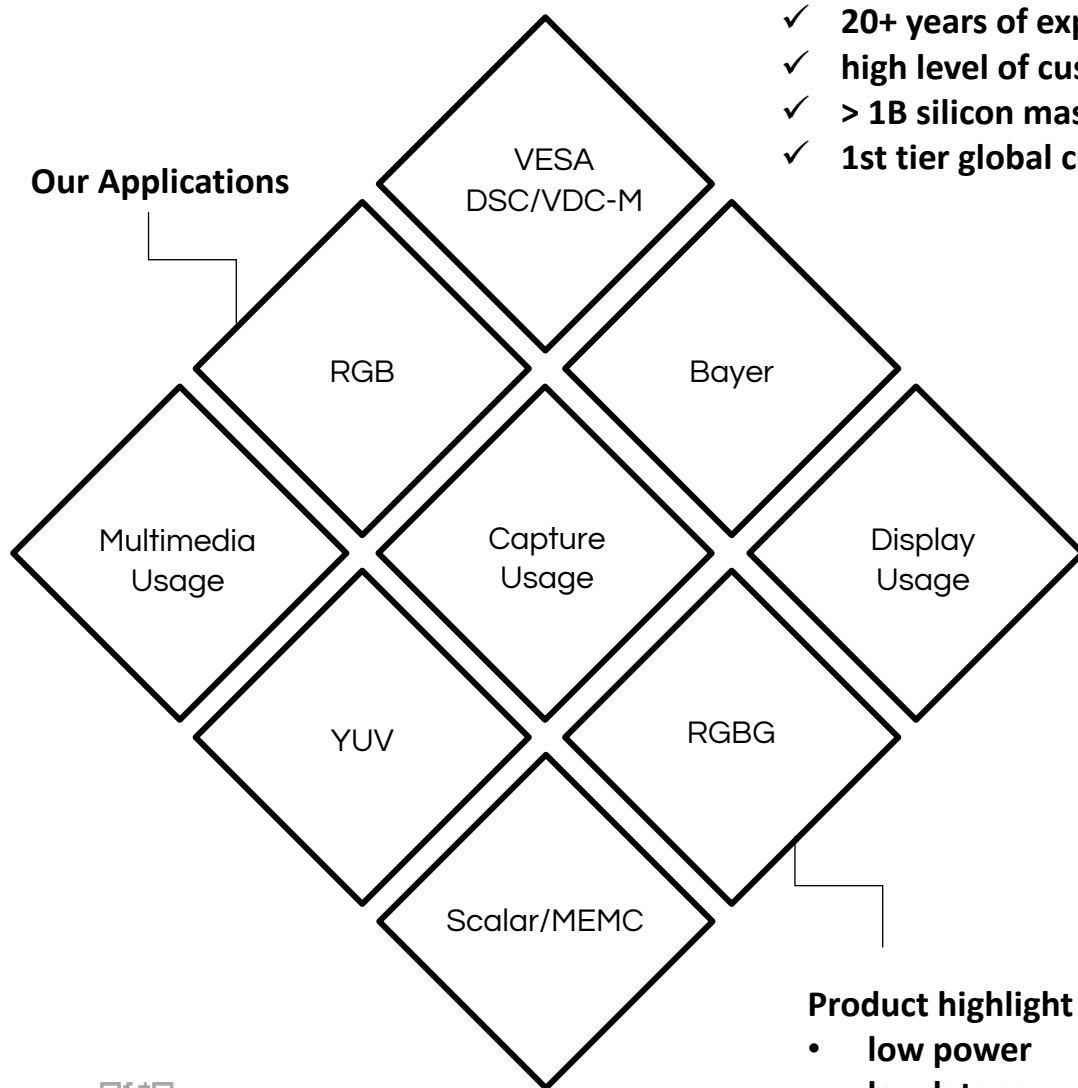


# TITC

## Image Compression IP specialist

- ✓ 20+ years of experience
- ✓ high level of customization
- ✓ > 1B silicon mass produced
- ✓ 1st tier global customers

### Our Applications



### Product highlight features:

- low power
- low latency
- small area

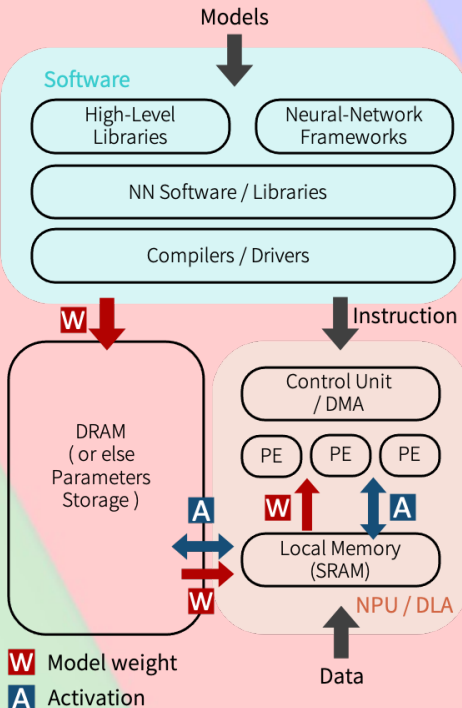


Tel: +886-3-5829011  
☺ [www.titc-usa.com](http://www.titc-usa.com)

# TITC N-Series IP

## Model weight/Activation for AI

N-series IPs offer an efficient, lossless solution for reducing the storage and bandwidth demands of AI models. By compressing both model weights and activations, it significantly lowers data traffic power consumption, cache SRAM cost, and DRAM space usage. The algorithm achieves near-theoretical compression ratios and maintains consistent performance across different models. With minimal hardware cost, ultra-low latency, and high throughput, the solution features an adaptive, entropy-aligned design and a parallel hardware architecture that scales to meet mainstream DRAM bandwidth requirements.



### ➤ TITC AI Inference Device IP

Usage / Series		capture / N-series
IP Name		TITC_N1
Data	Type	Weight/feature map
	Bit-Depth	int8
Compression	Type	Lossless
	Unit	16 data / T (= int8 * 16)
Performance	Throughput	16 data / T (= 128bit / T)
Note		* Ultra high throughput with ultra low latency * Tiny gate count with no SRAM in need

Note  
If specifically for CNN, Activation also can be described as 'Feature Map' .

Model Weight	Model	Size (byte)	Compression Ratio	
			zip	TITC_N1
CNN	mobilenet_v1	4,210,112	57.00%	65.53%
	yolo_v2	15,855,536	61.56%	67.25%
	private_a	9,009,472	83.08%	91.39%
	private_b	14,782,144	53.16%	57.81%
Transformer	bert	108,310,272	60.55%	67.65%
	gpt2	354,823,168	58.56%	64.64%
	llama3	8,030,261,248	52.69%	59.64%

Feature Map	Model	Size (byte)	Compression Ratio	
			zip	TITC_N1
CNN	private_a	78,151,680	57.17%	58.88%
	private_b1	4,516,762	40.16%	53.38%
	private_b2	30,870,800	56.48%	65.53%