Facial Recognition with the Gemini® Associative Processing Unit

Introduction

Facial recognition systems are increasingly standard in a wide range of settings: from smart airports to automated retail experiences and beyond. Effective facial recognition systems must extract accurate results with minimum latency between data capture and positive identification and be able to handle multiple queries almost simultaneously. However, current solutions require very large numbers of costly GPUs and CPUs to deliver performance at scale. The market seeks a new solution that can meet performance expectations while keeping infrastructure and TCO (Total Cost of Ownership) low.

Introducing a new facial recognition search engine powered by GSI's Gemini® Associative Processing Unit (APU) on the GSI APU board. Gemini® delivers accurate search results while reducing search times in large databases from many minutes to fractions of a second, all with very low power consumption.

GSI's face recognition search engine can process 32-bit and 64-bit floating-point vectors with 128 features and above (no upper limit). A scalable solution, large datasets and multiple queries can be processed using one or multiple APU boards, with no loss of performance and with very low power consumption—approximately 3.5 times lower power consumption when compared with CPU only systems.

GSI APU Board with Gemini® APU
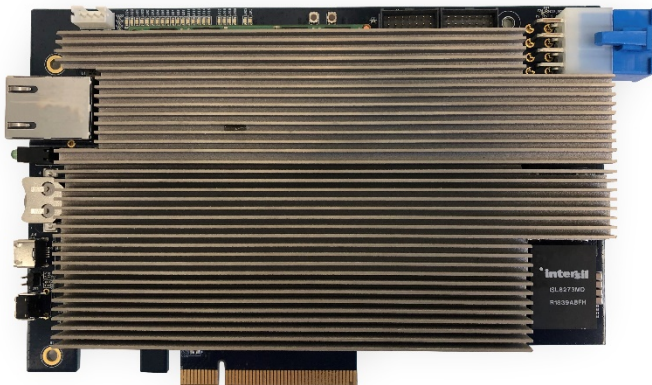
Board Specifications	
Thermal Design Power (TDP)	60W*
PCI IDs	Serial controller: GSI Technology Gemini® (Leda-G®)
APU clocks	400 MHz
PCI express interface	Gen 3 x 8
Board dimensions	254 x 111.15 mm

*40W power consumption at 400 MHz tested for Facial Recognition search

Spec.	One Board	Multiple Boards
Maximum memory clock	2666 MHz	2666 MHz
Memory size	16 GB	N x 16 GB
Memory bus width	64 bits	N x 64 bits
Peak memory bandwidth	21.3 GB/s	N x 21.3 GB/s

Search Solution
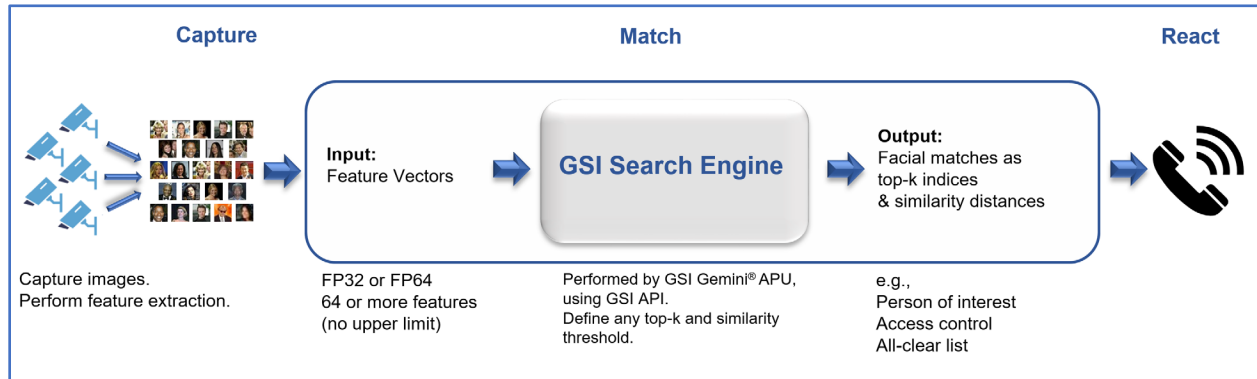
Performance and Power Consumption

The following tables show Gemini® performance and power consumption for face recognition, on databases of 128 thousand to 8 million records. Records and queries are represented by 32-bit floating point vectors with 256 features. k = 25. Operating frequency is 400 MHz. A single APU can process databases of up to 200 million faces.

1 APU Board

Number of Queries	Processing time (milliseconds)				
	128K	256K	512K	1M	8M
1	0.06	21.3	29.0	42.0	218
10	0.40	36.8	50.0	65.0	247
50	1.99	46.9	138	172	412
100	5.03	101	159	213	922
1000	47.4	545	741	883	2080

2 APU Boards

Number of Queries	Processing time (milliseconds)				
	128K	256K	512K	1M	8M
1	0.06	0.06	26.0	35.0	156
10	0.40	0.43	39.0	52.0	197
50	1.99	2.00	86.0	109	366
100	5.03	5.30	138	171	451
1000	47.4	47.5	485	583	1640

Typical Power Consumption, APU vs. CPU
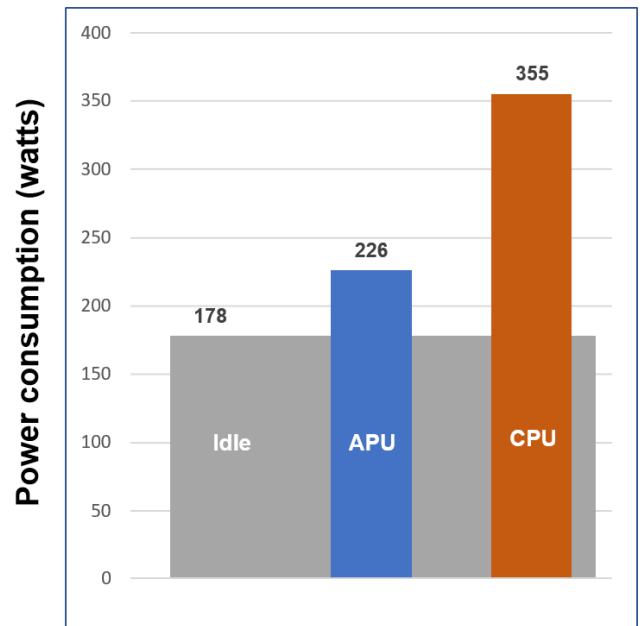
The APU reduces search power consumption by a factor of 3.5.

For more information on the Gemini® APU, please visit **www.gsitechnology.com/APU**.