Ultra-low power features of Nema GPUs

Georgios Keramidas, MSc, Ph.D
Chief Scientific Officer

Think Silicon S.A.
g.keramidas@think-silicon.com

Think Silicon is a privately held company founded in 2007 by the core team of Atmel MMC IC group

Intellectual Property semiconductor cores in the field of **computer graphics** for **mobile/embedded** devices

KHRONOS GROUP
CONNECTING SOFTWARE TO SILICON

OpenWF OpenVG OpenGL|ES Vulkan™

Company ID

- HQ and R&D: Patras / Athens, Greece
- Sales & marketing offices: USA, Canada, Germany, Japan, Taiwan
- Target markets: wearables, mobile, home automat./appliances
- Silicon proven technology

Licensees

SEQUANS COMMUNICATIONS

dialog SEMICONDUCTOR

MICROCHIP

FARADAY

LATTICE SEMICONDUCTOR

Undisclosed Tier1

Partners

ambiq micro Intelligent Energy-Efficient ICs

Technische Universität Berlin

SAMSUNG

codeplay

LPGPU

HiPEAC

EU-funded Projects

1. GPU-WEAR: Ultra-low power heterogeneous GPUs for Wearable/IoT devices

Period: Jun 2016 – May 2018, Budget: 2M, SME Instrument – Phase 2

2. LPGPU2: Low-Power Parallel Computing on GPUs 2

Period: Jan 2016 – June 2018, Total Budget: 3M, 5 Partners

Innovation Action, Customised and low power computing

3. TETRAMAX: TEchnology TRAnsfer via Multinational Application eXperiments

Period: Sep 2017 – Aug 2021, Budget: 7M, 22 Partners

Innovation Action, Customized and Low-Energy Computing

Problem in IoT/Wearable Devices

High quality visual experience is limited to High End mobile devices

Problem:

Application Processors* provide high quality visual experience

- Power Hungry
- High BOM (cost)

Microcontrollers** power the massive cost driven products

- Poor visual experience
- Long design cycles

Solution:

Think Silicon's graphics technology delivers:

- Application Processor Quality Graphics
- On the microcontrollers power and cost envelop
- Short design cycles

Think Silicon's goal is to bring smartphone-level visual experience to any size / cost display device!

* Qualcomm Snapdragon, TI OMAP, NXP i.MX, MediaTek Helio etc.

** ST STM32, NXP Kinetis, Microchip PIC32 etc.

Solution

NEMA® | GPU-Series – The only no compromise graphics solution

Device	Application Processors	Microcontrollers	Microcontrollers with NEMA
Battery Life	Days	Weeks	Weeks
BOM	High	Mid	Low
Software Development time	Weeks	Months	Weeks
Graphics Quality*	High	Low	High

Specific GPU type *(Nema\|p)* fully optimized/customized for GUIs (Graphical User Interface) acceleration → GUI apps are the common app in IoT/Wearable devices

Mobile Graphics: Past-Present-Future

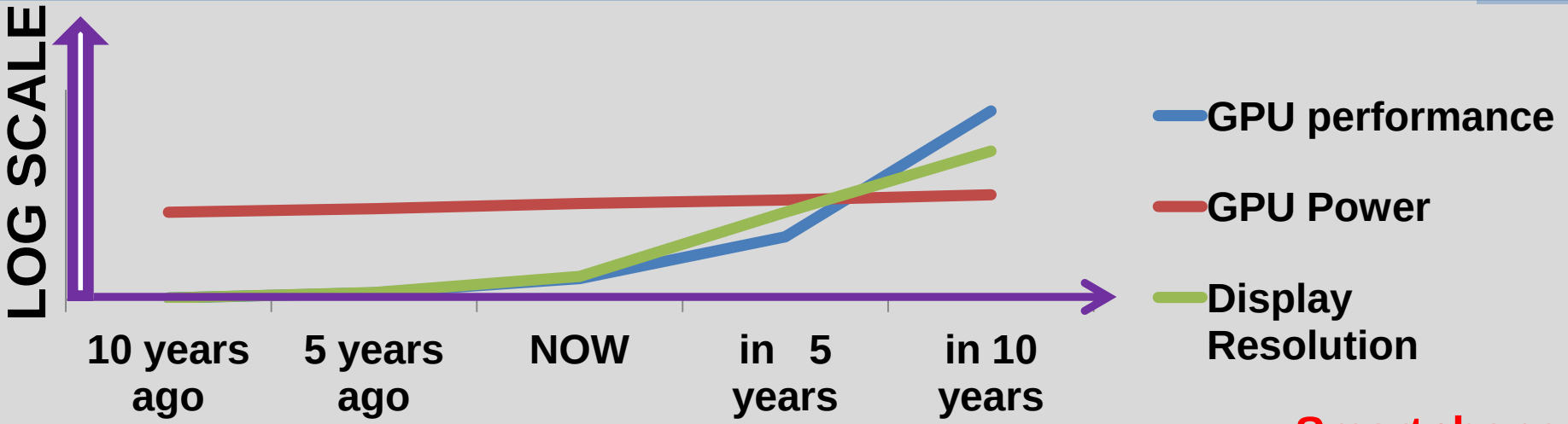
past

present

Visualize data from various sources
(Internet of Things paradigm)

future

Moore's Law in Mobile GPUs



Smartphone GPUs

Applications → in next 5 years will need 4-5x the current GPU perf.

Display resolution→ exponential increase (4K displays are here)

BUT… Power→ roughly under the same power budget (few hundreds milliwatts)

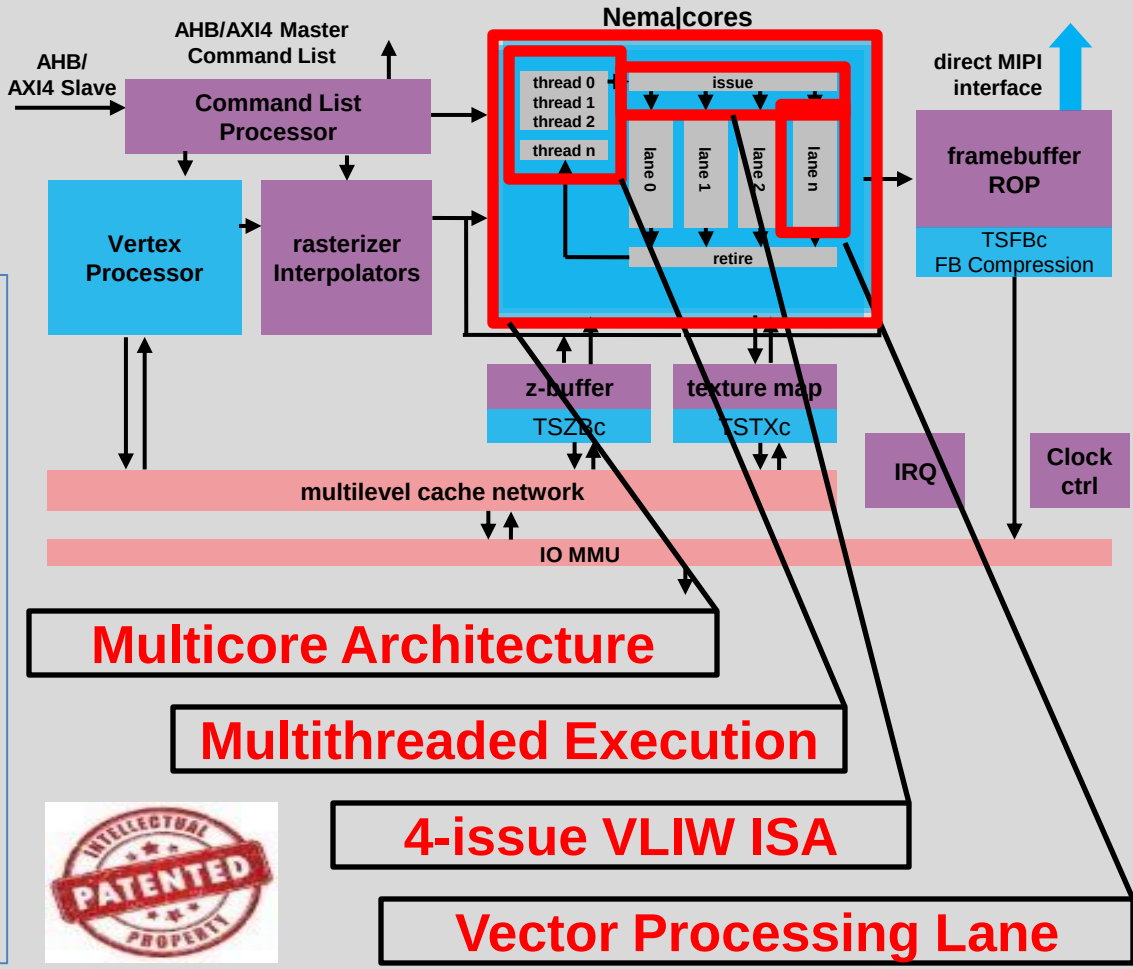
Multi-level power optimizations are required

A Holistic Approach to Low Power

NEMA® GPU Low Power Consumption

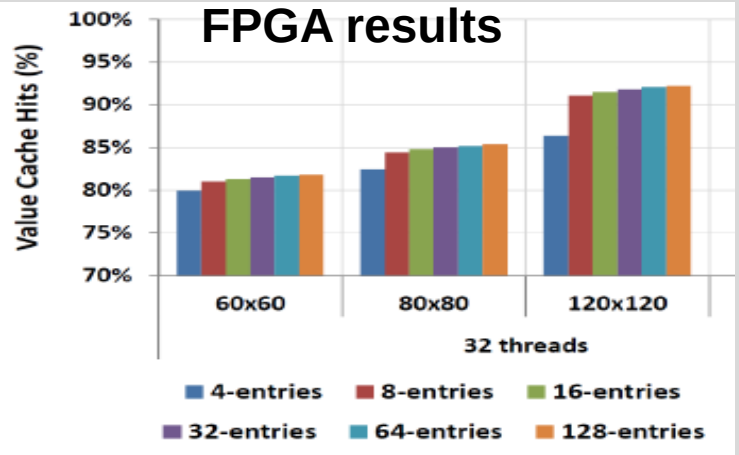
HW-level Parallelism is the key for Ultra-low Power Designs

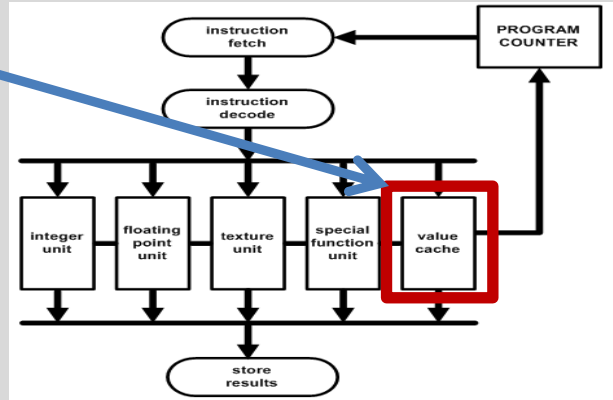
- **Very low gate-count with extensive HW folding techniques** (starting at 180K gates)
- **Very high IPC VLIW Processor core allows very low clock speeds** (Up to 12 operations per clock cycle)
- **Datapath width adjusted to Display Characteristics**
  (e.g. Targeting a 2 bpp display can be computed with less accuracy than 16 half FP, 10 or 8 bit precision with no visual impact)
- **Hierarchical Clock Throttling/skipping**
  Each submodule is throttled to match its workload



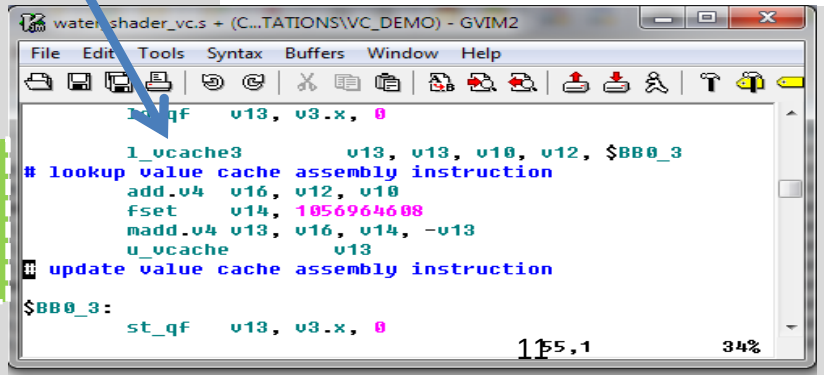
Approximate Memoization Mechanism

- **Value Cache:** Skip computations if called with identical or similar arguments
- Extra functional unit in GPU data path managed by ISA instructions visible to GPU compiler /assembler
- Extension of GPU ISA
- Automatic Insertion methodology implemented in LLVM





Assembly Produced By LLVM



Reduction in Memory Accesses

High quality Frame-Buffer Compression

- fixed-rate, 4bpp or 6bpp with or without alpha
- lossy
- real time on-the-fly (compression & decompression)

Texture Compression

- Off-line lossy (Real Time Decompression)
  - Fix-rate compression, 4bpp or 6bpp
  - Alpha support

- **Reduction of upto 6x in system bandwidth**
- **Enables use of onchip SRAM for graphics buffers**

- **Read a 320x320 FB (Smartwatch size)**
  - **Off-chip 32-bit -> 400 Kb -> 12800000 pJ**
  - **Off-chip 16-bit -> 200 Kb -> 6400000 pJ**
  - **On-chip TSBc -> 50 Kb -> 160000 pJ**

- **Energy Reduction by up to 80 times (!!!)**
- **Further savings on on-chip bus fabric**

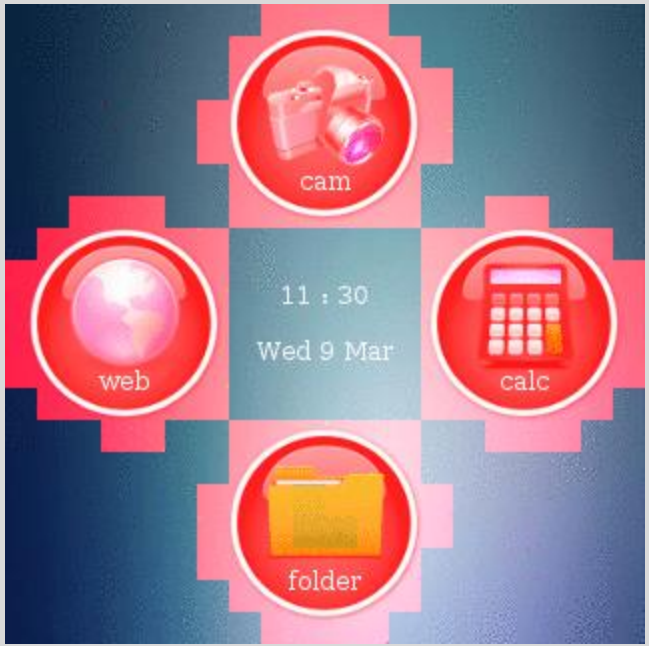
Original image
32bpp

TSFBc
4bpp

TSFBc
6bpp

Reduction on Communication Traffic with Screen

Nema GPU keeps track of affected screen regions that were modified and only updates those regions to the screen (on panels with internal framebuffer)



Selective Framebuffer Update

Reduction on update regions → Reduces communication to off-chip display

Example: 320x320 Image

Continuous Update: 12800 nJ
Partial Update: 5632 nJ

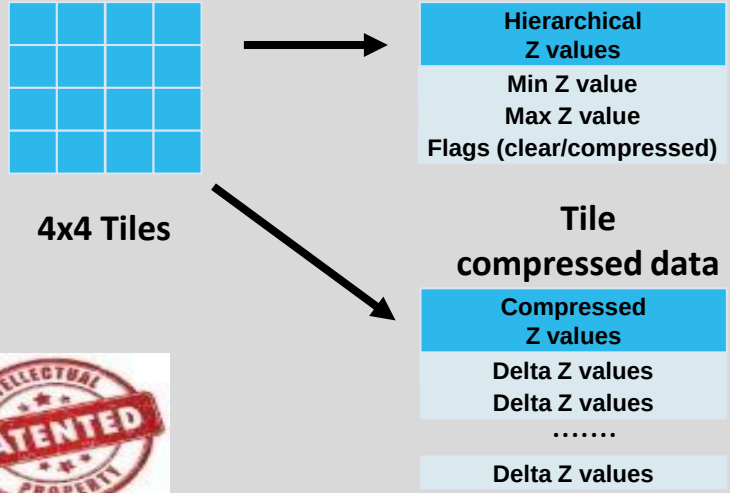
56% Reduction in screen update traffic

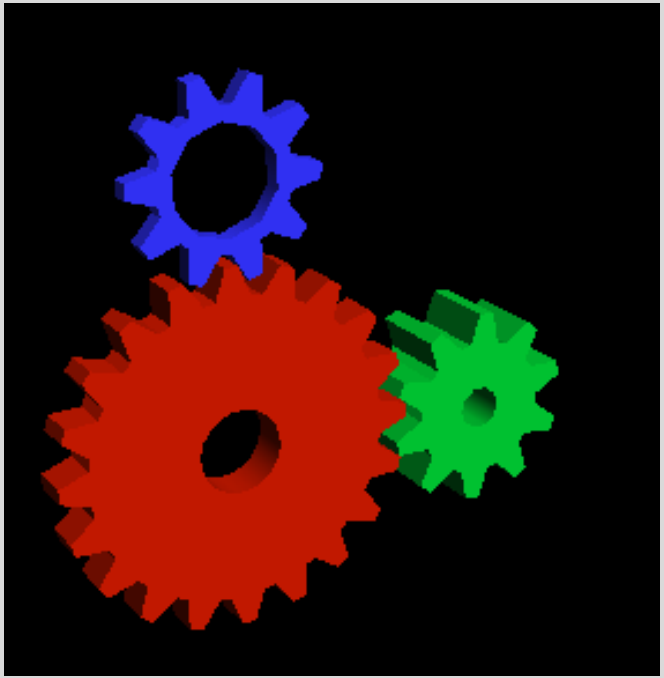
Further Reduction on Memory Accesses: Z-Buffer

Tiled Hierarchical Z-Buffer Compression

Reduction of memory transactions by:

- **Read 80%**
- **Write 70%**



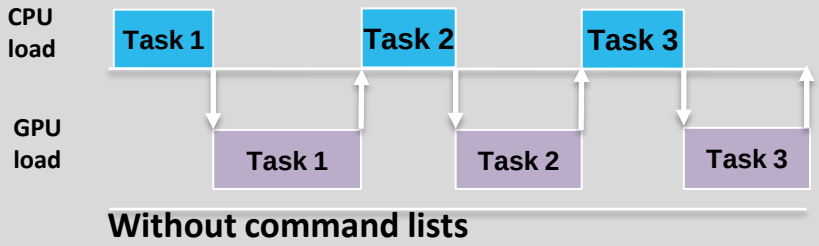


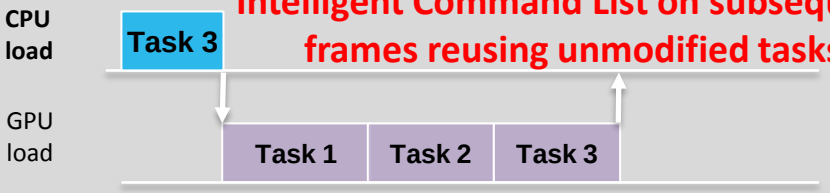
Reduced CPU load

ARM M-series
Synopsys ARC

Reduced CPU load

- **Lean Library and Intelligent command lists**
- **Graphics API customized to Display Dimensions**
- **Graphics API customized to Object Dimensions**

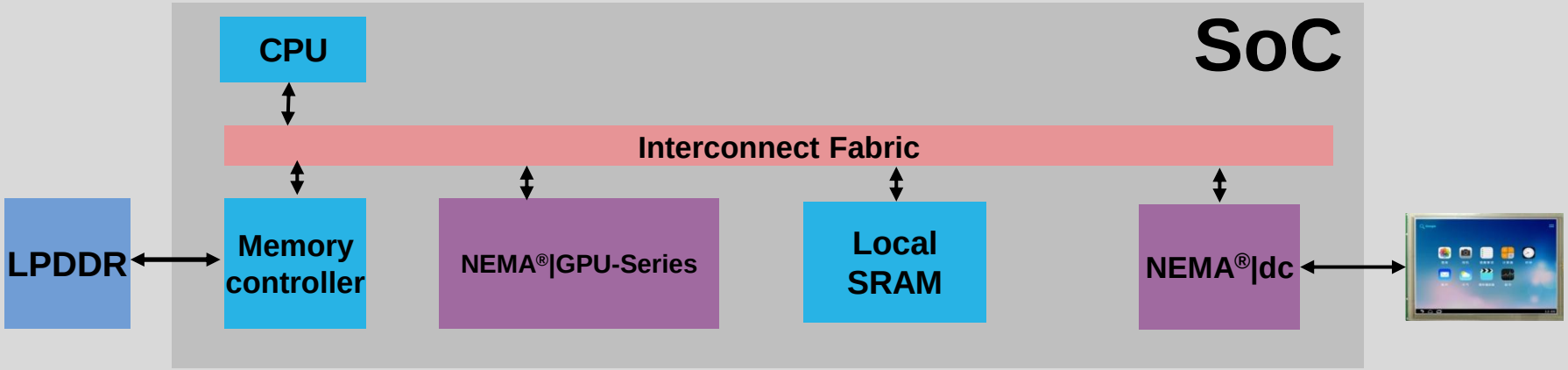




Lean Library (NemaGFX API):
RAM: 1 KBytes
ROM: 24 KBytes

Intelligent command list Reduces CPU load by 30%* in demo example workload (*scene dependent)

Putting it all together



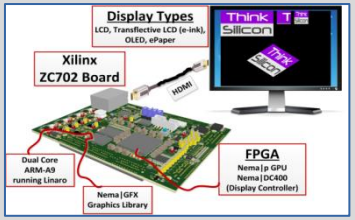
	Nema|p	Conventional	Description
CPU load	**1.2 MIPS**	**~ 2 MIPS**	Reduction using command lists
Memory reduction	**0.68 mW**	**55 mW**	40-80 x Reduction in Power * (use of compression & on-chip RAM) (*DRAMPower estimation)
GPU Active Power	**1.2 mW**	**[illegible] mW**	Low Frequency and throttling (peak)
GPU Leakage Power	**0.06 mW**	**0.30 mW**	Low Gatecount and Power Down
Screen communication	**1.4 mW**	**3mW**	55% reduction by smart partial updates
Total Power Consumption	**3.2 mW**	**74 mW**	Peak Subsystem Power

23x

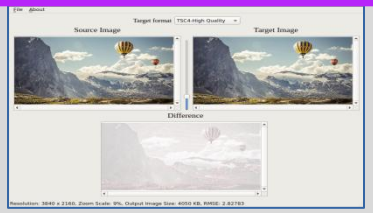
NEMA® GPU: Ecosystem

Nema|SDK Not just a GPU, but a complete Ecosystem of Tools

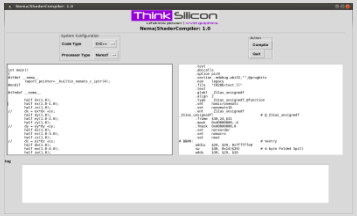
NEMA®|GUI Builder
Rapid GUI Design Toolkit that allows drag&drop creation of advanced GUI in minutes instead of months



NEMA®|Bits
An EVK Kit for technology evaluation and pre-silicon application development



NEMA®|PixPresso
Asset management and image optimization for optimal visual appearance and efficient memory utilization



NEMA®|ShaderEditor
A GUI tool that compiles GLSL/C++/OpenCL Code to Nema Assembly for Vertex and Fragment Shader Cores

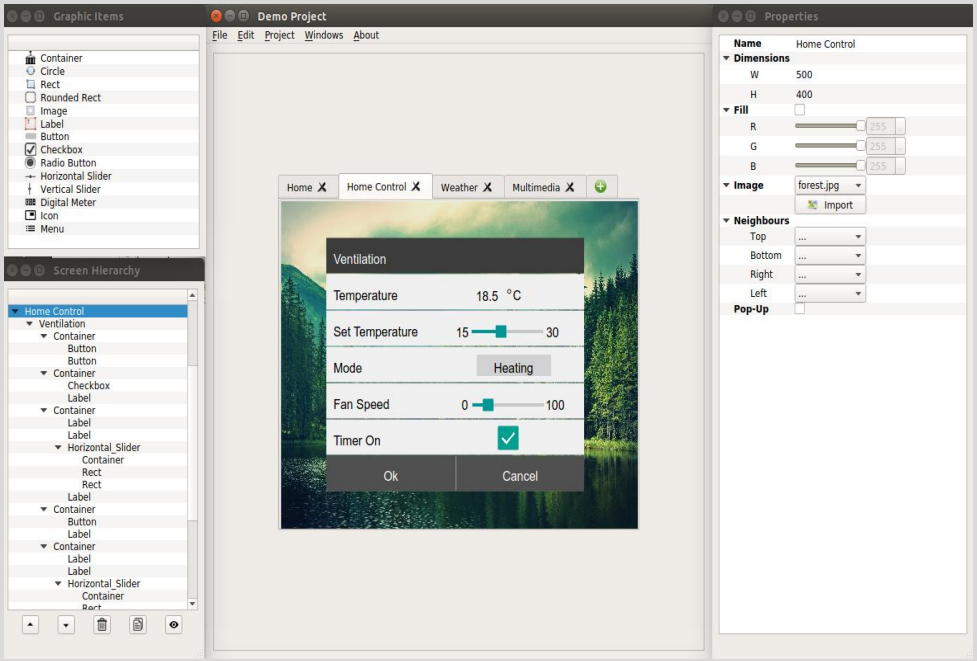
*NEMA®|Profiler
GPU Profiler based on CodeXL, monitors system activity, performance and power and identifies bottlenecks

*NEMA®|PowerModel
Power Model based on performance counters can estimate the system power consumption

* Under development

NEMA®|GUI-Builder

NEMA®|GUI-Builder allows the rapid creation of GUIs for Bare/RTOS or Linux embedded systems within minutes instead of weeks



Optimized Graphics Assets (images, icons, fonts)

Automatic Generation of optimized code

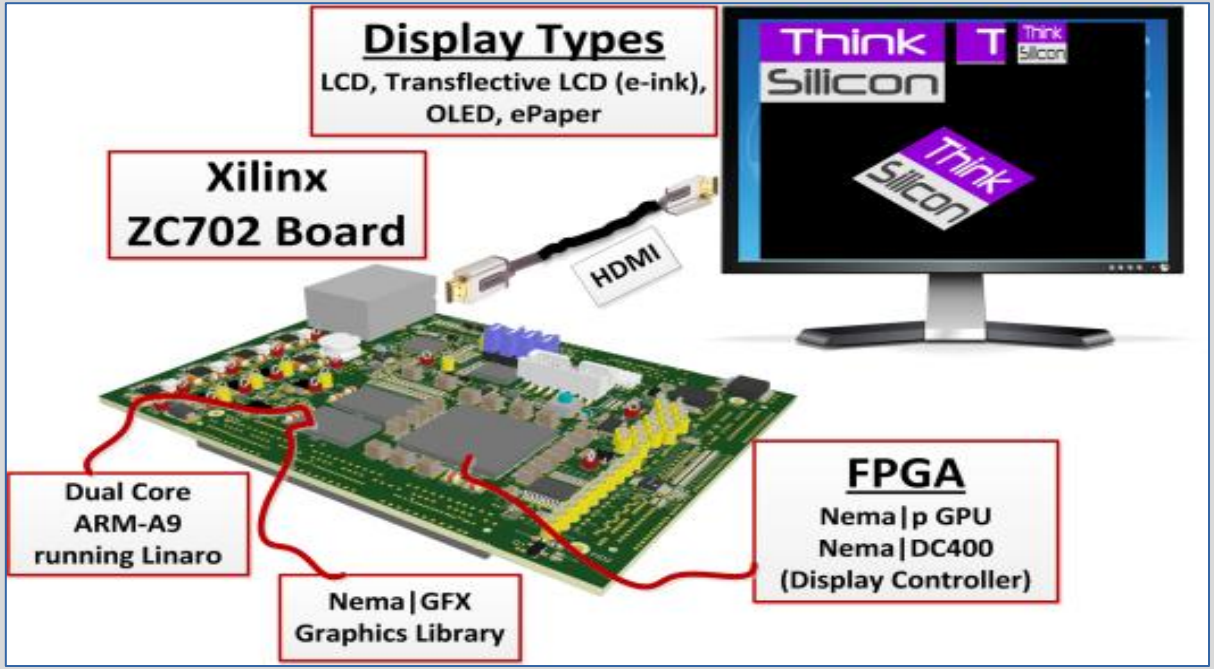
Application Code

NemaGFX Library

HAL OS

NEMA®|Bits

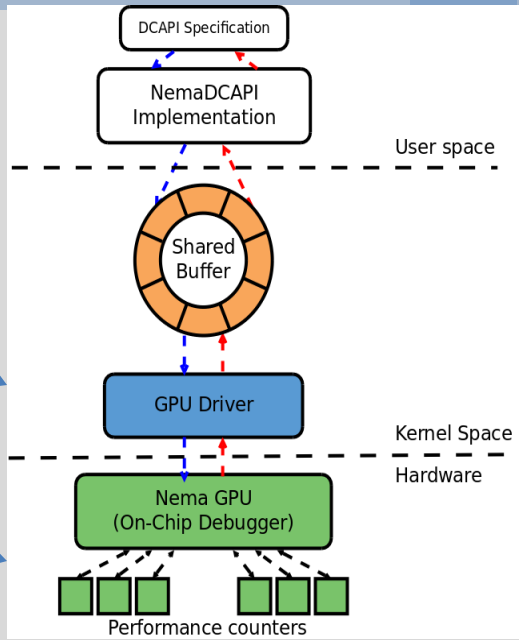
NEMA®|Bits is an Evaluation Kit based on Xilinx Zynq 702 board. NEMA®|Bits allows technology evaluation and pre-silicon application development



NEMA® | Profiler

- NEMA® | Profiler is being developed as part of LPGPU2 project → LPGPU2 tool customized for Nema GPUs

- Think Silicon developed an **on-chip Debugger Architecture (150 performance counters), w/ 3 different adaptive sampling modes** for more accurate performance statistics



Performance strikes in Think Silicon GPUs executing Image processing algorithms

Indicate the problematic and power consuming areas of the execution

NEMA®|PowerModel

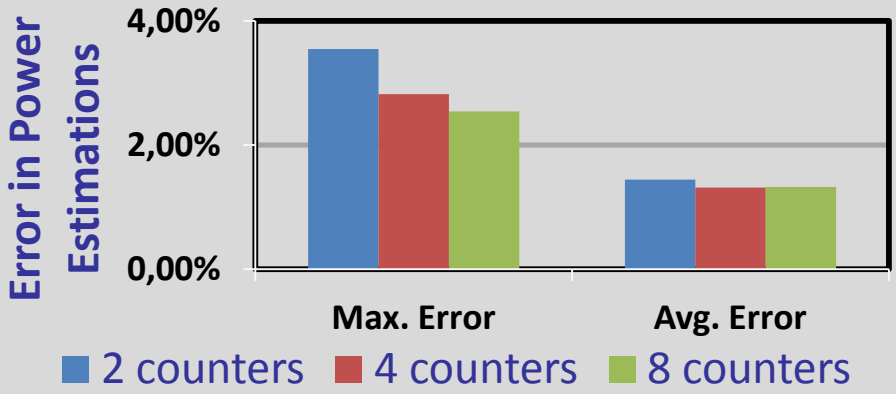
NEMA®|PowerModel allows highly accurate power estimations based on performance counter data. Not only a power model but complete methodology to:

GPU Configuration
Requested Accuracy
Training Suite

→ **NEMA PowerModel** →

Power Model 1, Set of Counters 1 → Instr. Mix 1
Power Model 2, Set of Counters 2 → Instr. Mix 2
.....
.....
Power Model N, Set of Counters N → Instr. Mix N

Fully automated methodology



Multiple Power Models; One of each piece of code (instruction mix) being executed → Highly-accurate power estimation with limited perf. Counters

Results for TSMC LP @ 40nm; Similar results for FDSOI @28nm

Nema GPU Technology beyond Graphics

Think Silicon evaluates a number of applications running on NEMA GPUs and explores Hardware Enhancements

Deep Learning

Artificial Intelligence
Neural Networks

GPGPU

General Purpose Computing

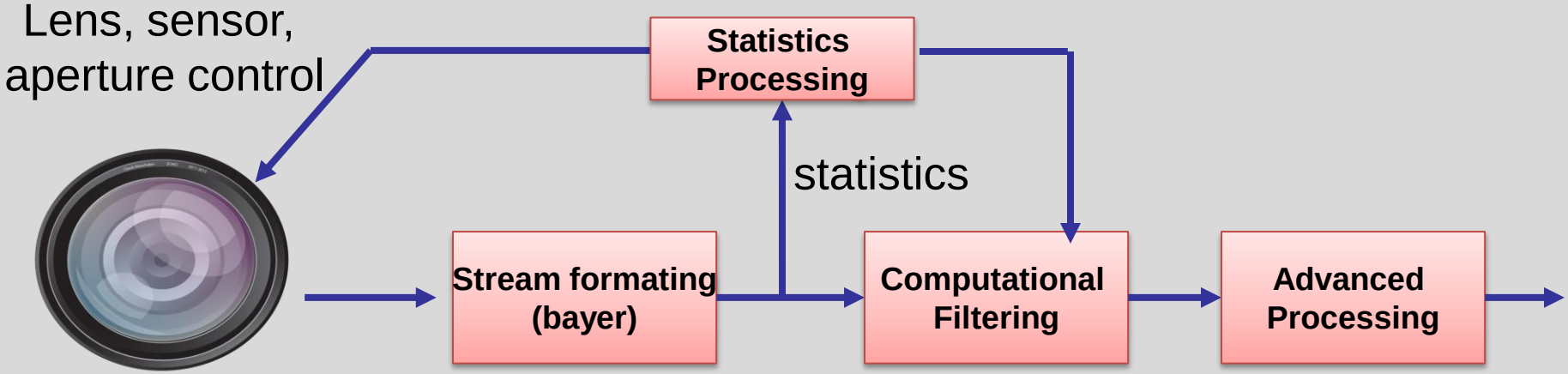
VR/AR

Virtual Reality
Augmented Reality

Image Processing

Computation Photography
Computer Vision

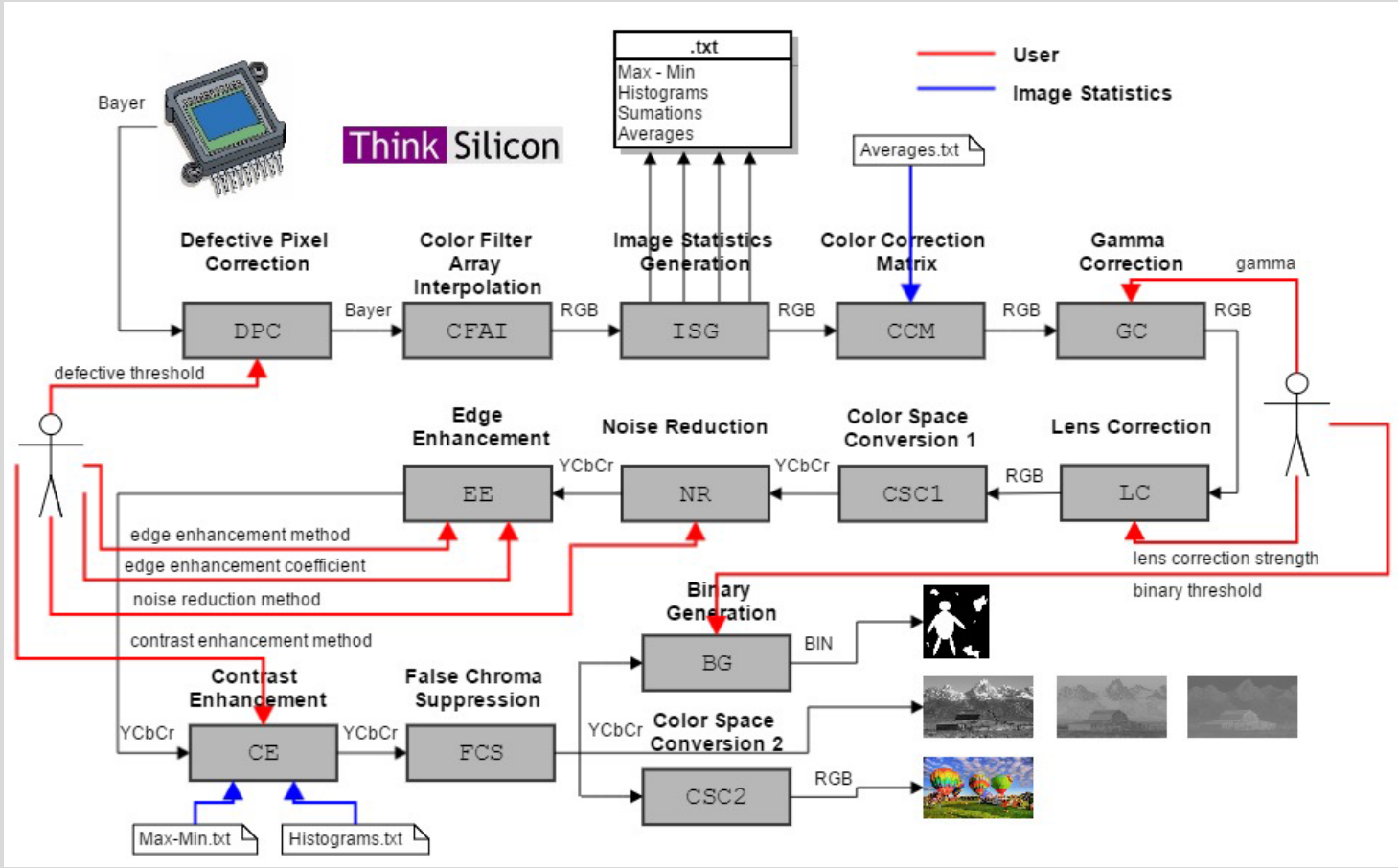
Versatile Programmable Image Signal Pipeline (ISP). It brings flexible camera processing capabilities to systems with limited resources (silicon area, memory)



Standard Sensor Filters

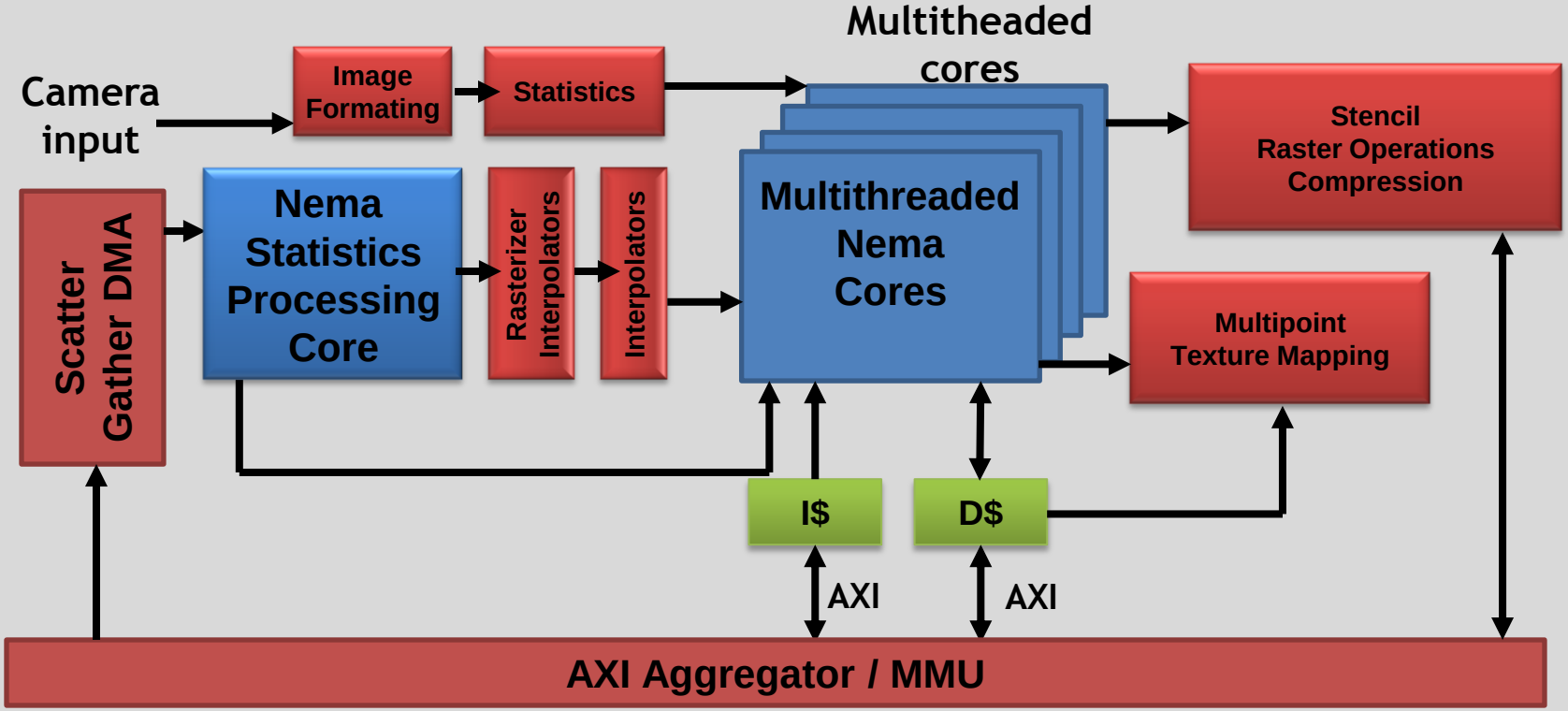
Bayer Filter - CFA(Color Filter Array) - Defective Pixel Correction - Image Filters (3x3 and above) - HDR Processing - Color Correction 3x3 Matrix - Noise Reduction - Edge Enhancement - Gamma Correction - Support re-sampling filter - Image Statistics- Cropping/Windowing

A full-blown Programmable ISP



Programmable ISP in Nema GPUs (internal project)

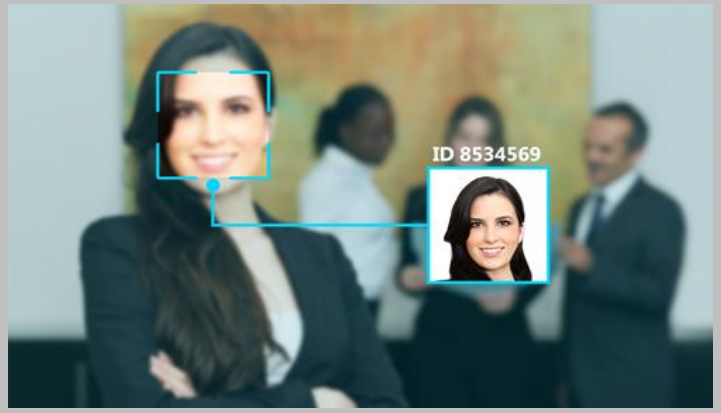
- Based on the Nema GPU multicore architecture family
- Additional Instructions compared to graphics-only architecture
- Multithreaded processing unit with compact VLIW instruction set



Deep Learning Acceleration: Nema GPUs vs DSP

Automotive Applications

Security Applications



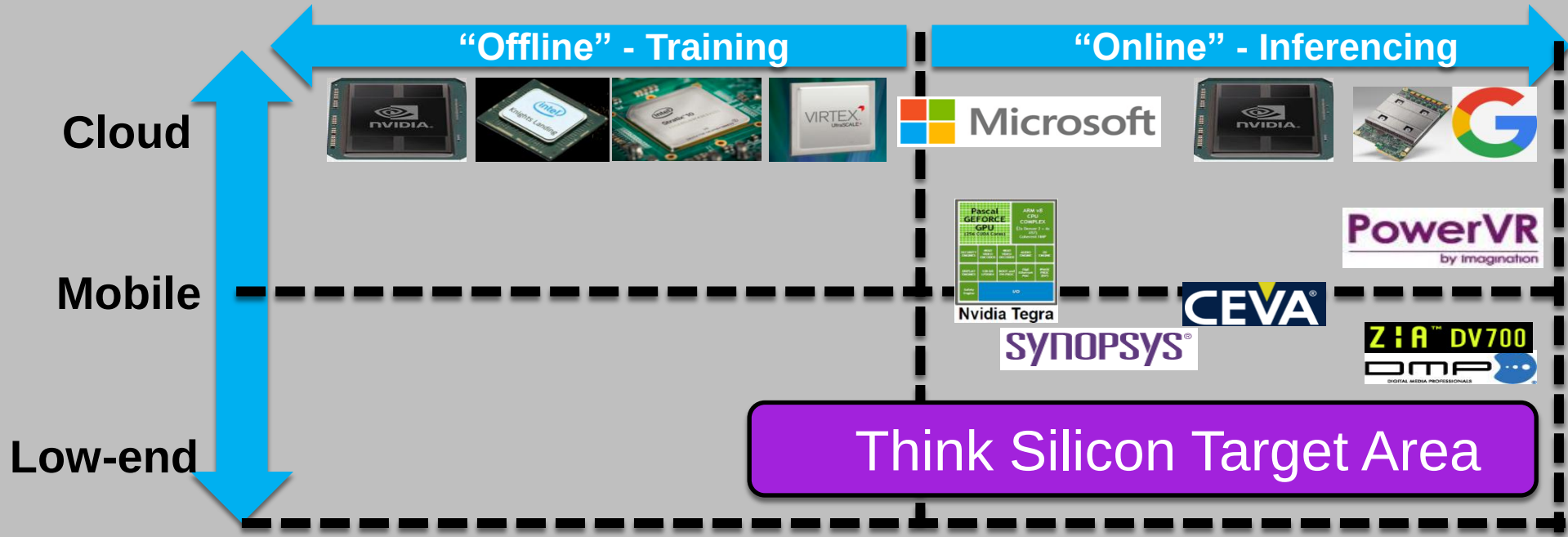
Required Features	NEMA	DSP
Multi-cores (scalability)	✓	✓
Fast MAC operations (8-bits)	✓	✓
Memory-latency-hiding-capabilities	✓	✗
Data Reuse Optimizations	✓	✗
Fast Memory (Image) Copies/DMAs	✓	✗
Approximate Calculations (for low power)	✓	✗
Real-time Image compression/decompression	✓	✗

Gartner Cool Vendor 2016

‘Novel Semiconductors for Neural Networks, 2016”

Customized DNN Accelerators

Customized NN Accelerators are in their INFANCY



Target Applications

Automotive Applications

Security Applications

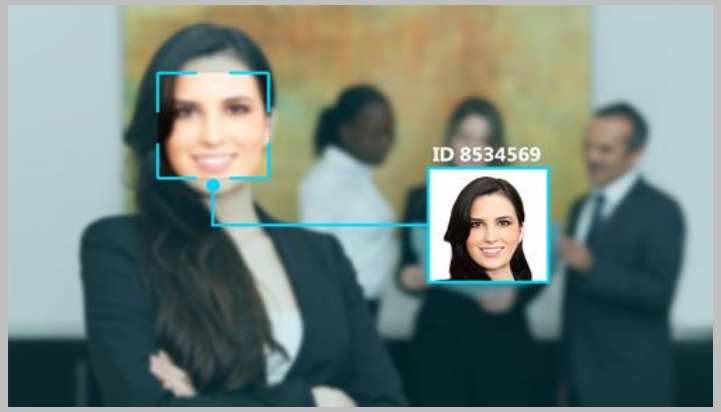
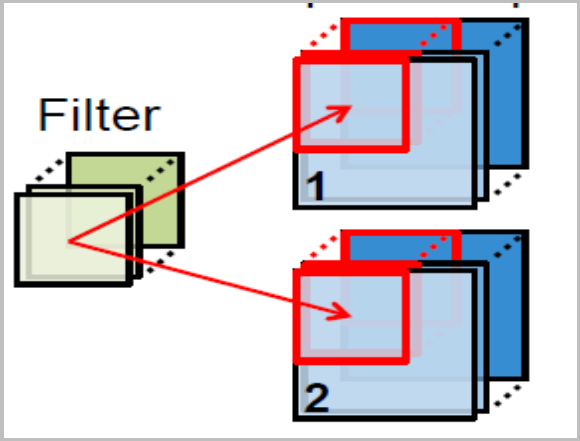


Image Processing / Classification Applications

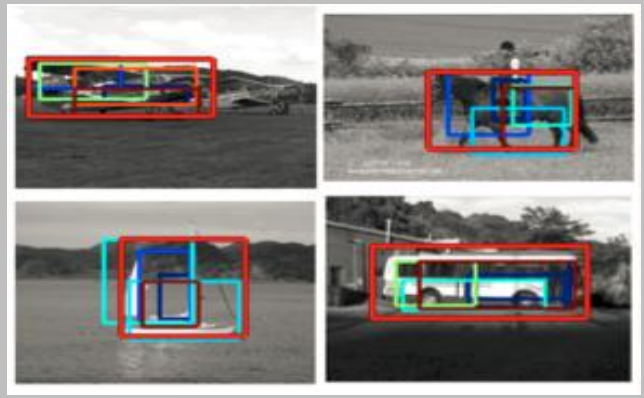
Relatively simple applications that cannot be executed in low-end devices using classical libraries/approaches (e.g., OpenCV)

NN Processing Requires …. Low Power GPUs

Optimized for Data Reuse



Optimized for fast DMAs

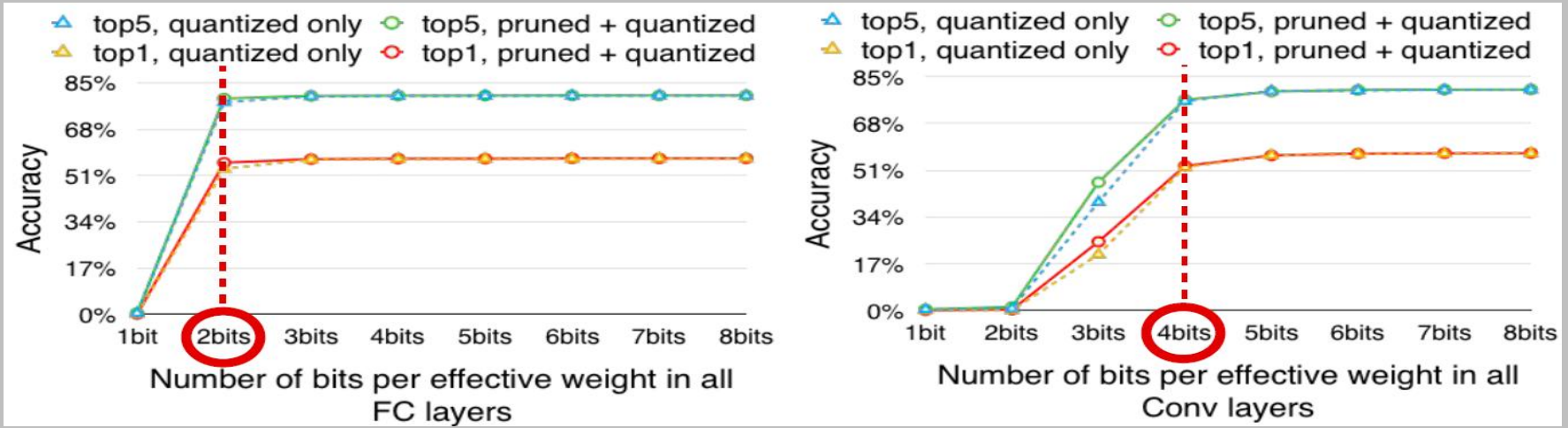


- **Value Cache (Approximate Value Re-uses):** Data re-use can be done at various levels (Filter weights, Activations, Features memory, Parameters memory)

- **Scratchpad and Cache coordinated Memory Systems optimized for fast DMAs of "misaligned" data**

NN Processing Requires Flexible Datapath Precision

Bits Per Weight



- **Datapath width Adjustable at Compile Time & flexible Datapath consisting of both Fixed Point and Floating Point Units of Variable Ranges:** Different layers require different precision in weight operations & quantized tensors

NN Processing Requires Multi-Layer Compression Support

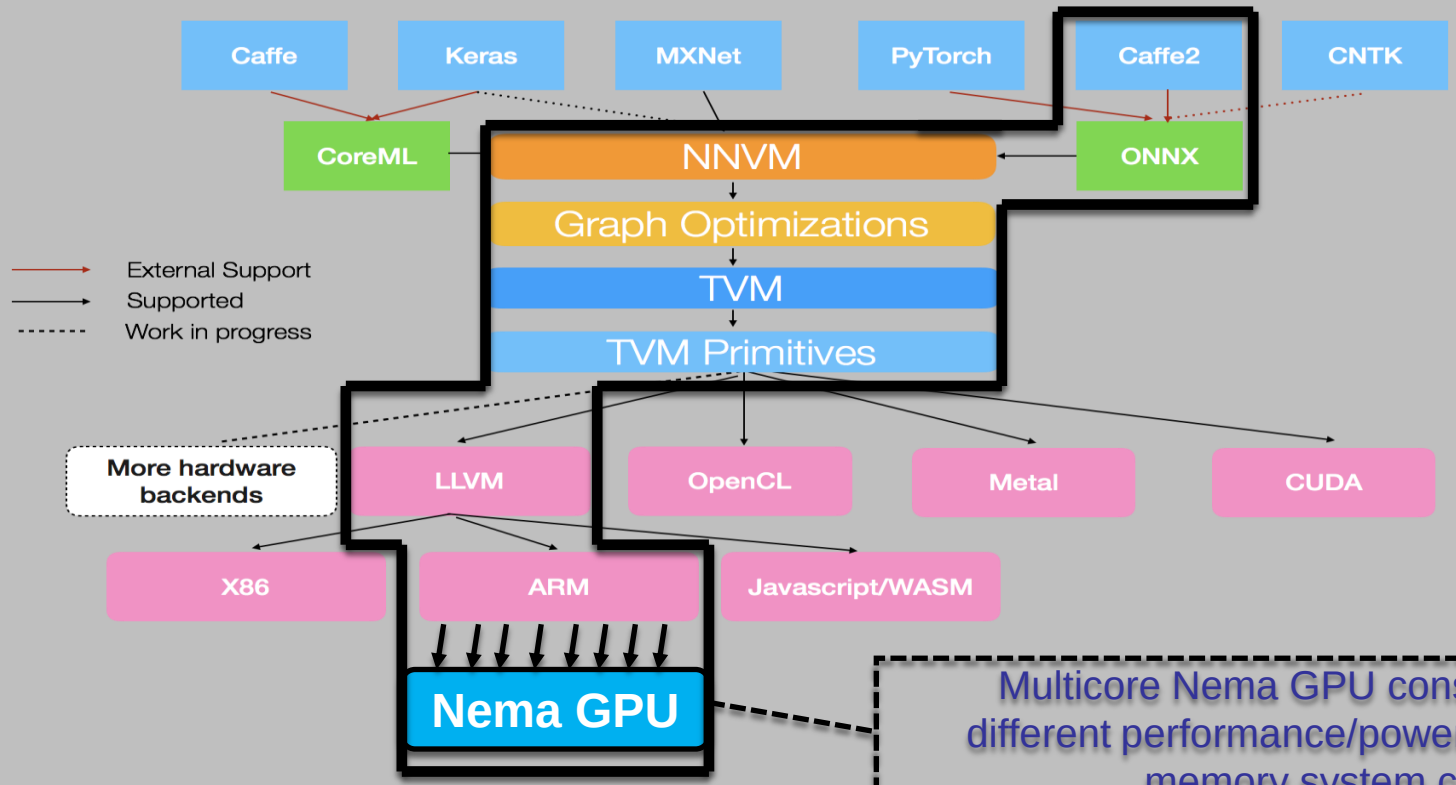
Real-time Lossy (De)Compression of Tensor Data

Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
LeNet-300	1070KB	27KB	**40x**	98.36%	98.42%
LeNet-5	1720KB	44KB	**39x**	99.20%	99.26%
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
SqueezeNet	4.8MB	0.47MB	**10x**	80.32%	80.35%

- **Nema Proprietary Texture and FrameBuffer Compression Techniques:** Real-time Compression/De-compression of tensor data with minimal accuracy loss

NemaNN Software Ecosystem

NNVM Framework (Amazon) for Evaluation/Benchmarking

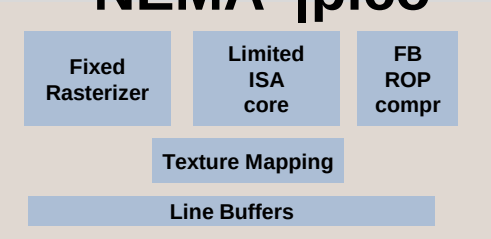


Illustrated path represents Think Silicon benchmarking/ characterization methodology

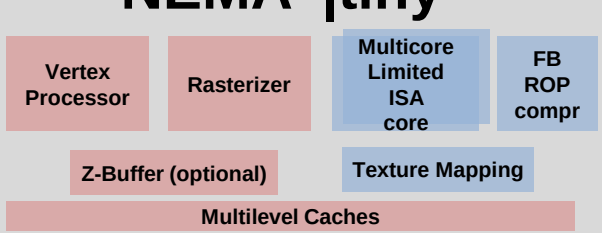
Multicore Nema GPU consisting of 3 cores with different performance/power tradeoffs and different memory system configurations

NEMA®|GPU series

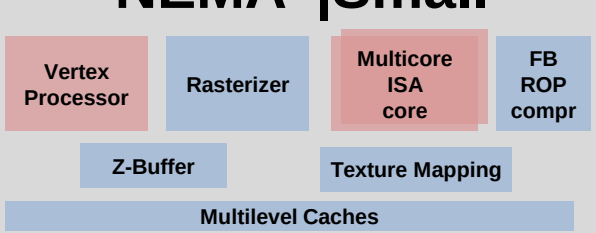
NEMA®|pico



NEMA®|tiny



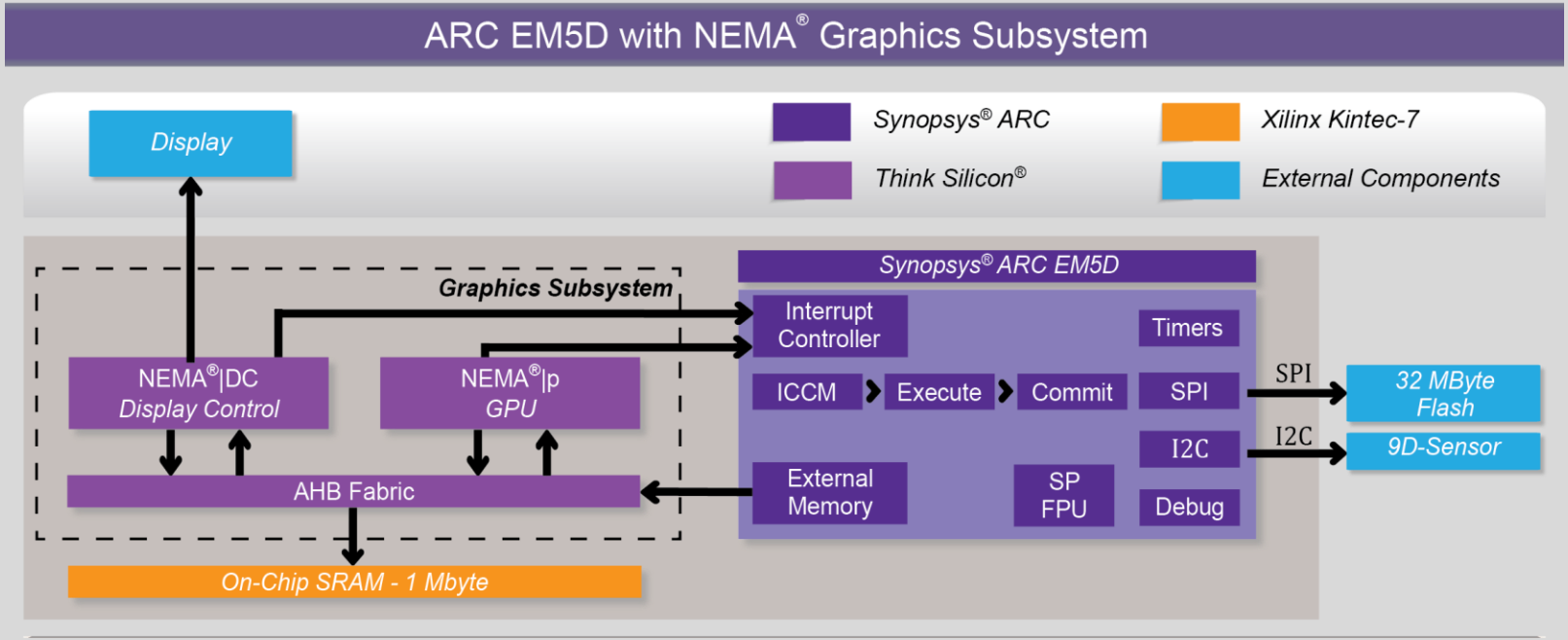
NEMA®|Small



	NEMA® | pico	NEMA® | tiny 100/200/400	NEMA® | Small 100/…/1600 *
Characteristic	2D / 2.5D GPU	Lite 3D GPU	Full 3D GPU
GPU core	1	1-4	1-16
Fragment Processor	Limited Programmable (fixed-point)	Limited Programmable (fixed-point)	Fully Programmable FP (GLSL/C/C++) LLVM
Vertex Processor	Fixed Function (Floating Point)	Fully Programmable FP (GLSL/C/C++) LLVM	Fully Programmable FP (GLSL/C/C++) LLVM
Z-Buffer Unit	no	Optional	yes
FB Compression	yes	yes	yes
Memory System (AHB/AXI4)	Line buffers	Texture/Vertex L1/L2 Caches	Texture/Vertex L1/L2 Caches
OS/ Graphics API	RTOS/Linux NemaGFX, uGFX, DirectFB	RTOS/Linux/Android NemaGFX, uGFX, DirectFB, OpenGL ES	NemaGFX, uGFX, DirectFB, OpenGL ES 3.x, Vulkan, OpenCL
Applications	Designed to build 2D display applications starting from 1.5” to 6.0” screen size with resolutions up to XGA (1024x768)	**smallest 3D (tiny) GPU in the industry**, designed to build 3D display applications starting from 1.5” to 6.0” screen size	**Smallest FULL 3D (small) GPU in the industry**, designed to build 3D display applications and compute applications

NemaNN will include a combination of Nema GPUs

ARC with NEMA® Graphics Subsystem



Kintex KC7-160T

	NEMA®|dc	NEMA®|p	Memory	ARC
LUTs	5434	30607	790	28496
FF	1828	12584	13	12318
BRAMs	3	8	218	96
DSP	6	35	0	2
MHz	33	33	33	33

ASIC 40nm LP process

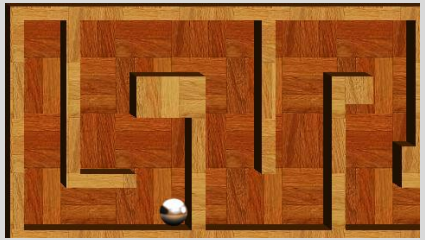
	NEMA®|dc	NEMA®|p	ARC®
Area(µm2)*	20275	177282	139822
Gates	16420	143572	113235

*memories not included

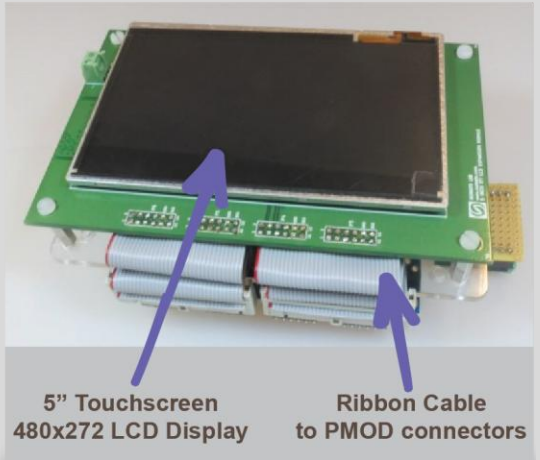
Demos: Running @33MHZ (CPU + GPU)

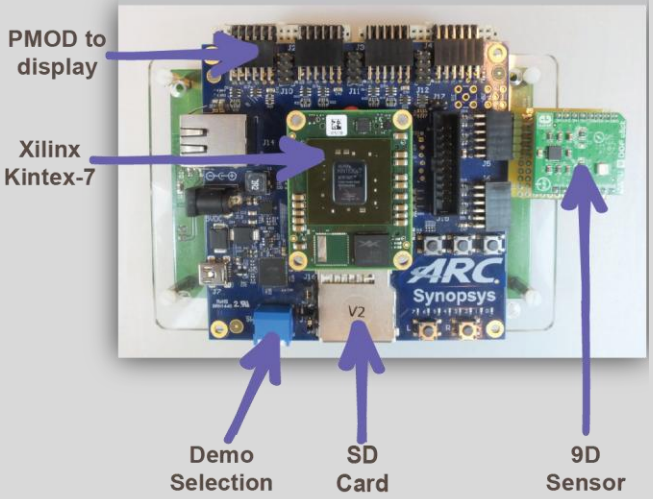
Demo 1: “Cover Flow”, 480x272, 29fps

Demo 2: “Rotating Globe”, 256x256, 36fps



Demo 3: “Maze”, 480x272, 58fps





Please visit our Booth;
You can play with the platform ☺

Think Silicon

ultra-low power | vivid graphics

HQ & DC
Patras Science Park
Rion Achaias, 26504, Greece
T. + 30 2610 911543
info@think-silicon.com

HQ North America
Ulli Mueller
Toronto / Canada
T. +1 647.824.2006
u.mueller@think-silicon.com

Sales
Christos Makiyama
Japan / Tokyo
T. +81 90.9854.1132
c.makiyama@think-silicon.com

Sales
Grace Lin
Taiwan / Taipei
T. +886. 9630.31076
g.lin@think-silicon.com

Sales
Roger Milton
North America / San Jose, CA
T. +1 408.677.6070
r.milton@think-silicon.com

Sales
Stefan Buechmann
EMEA / Germany
T. +49 170.636.5370
s.buechmann@think-silicon.com

HAMAC HELLENIC ASSOCIATION OF MOBILE APPLICATION COMPANIES

HellenicSIA

KHRONOS GROUP

mipi alliance

corallia