Using MRAM in Inference Engine Application

Terry Torng

***Stealth Startup**
co-founder
Gyrfalcon Technology Inc

Santa Clara, CA
August 2019

Outline

- Why NVM for AI
- Application Specific Accelerator **Architecture**
- Embedded MRAM AI Accelerator
- Case study examples

Core Challenge To AI: Energy Efficincy

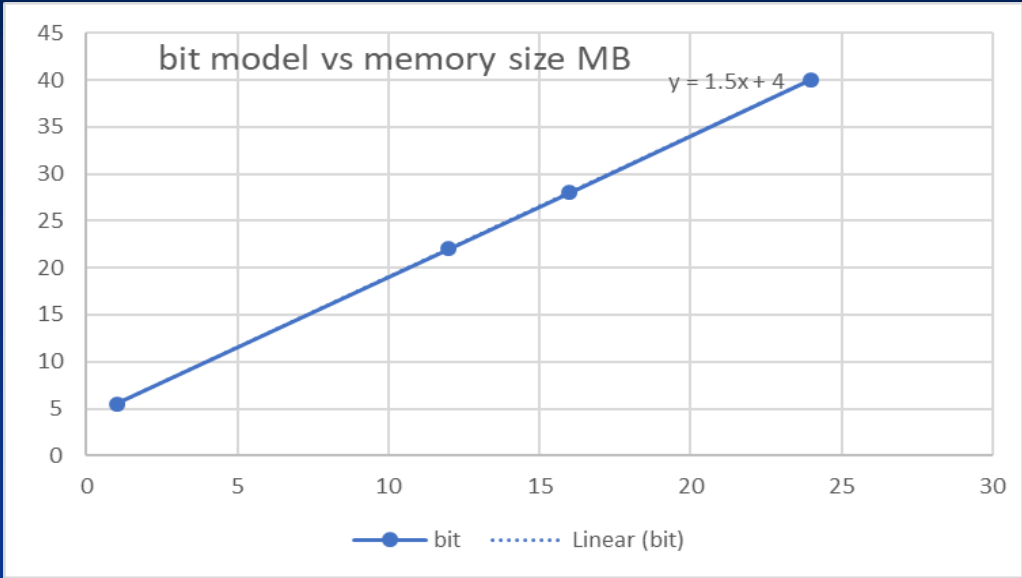
- **Data Center Energy Use is Growing….**
- "Global data centers used roughly **3%** of total electricity in 2016, and will **double** every four years"
- Radoslav Danilak, December 15, 2017

- "Global IP traffic will increase nearly threefold over the next five years, and will have increased 127-fold from 2005 to 2021."
- Bill Kleyman, Mar 09, 2018

- **Edge and IoT Devices….**
- "AI is hungry for processing power. IoT is projected to exceed 20b devices by 2020. There are currently 10b internet-connected devices, doubling to 20 billion will require massive increases to our data center infrastructure, which will massively increase our **electricity consumption**." Radoslav Danilak, December 15, 2017

Need to make us greener…
AI and 5G coming……

More Memory

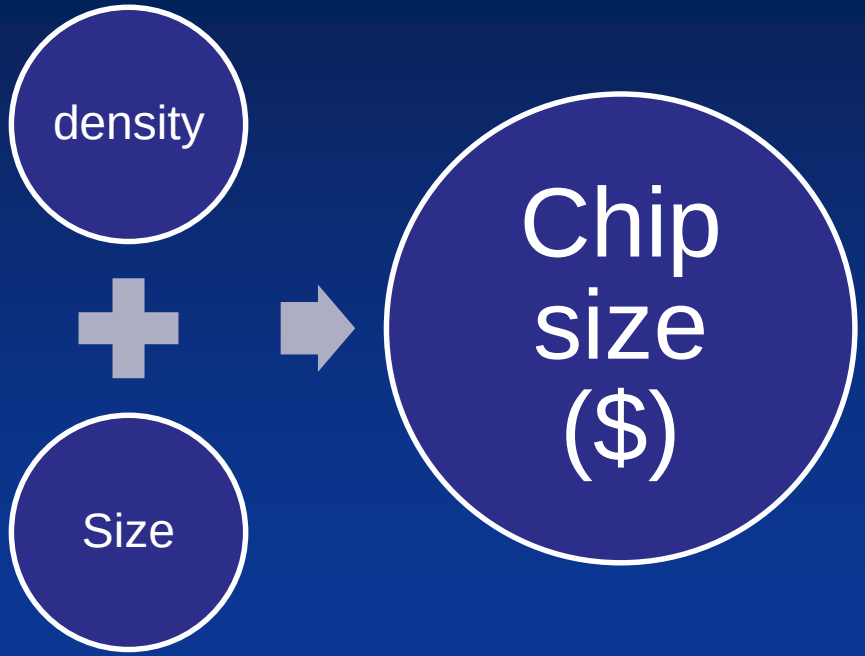


Accuracy

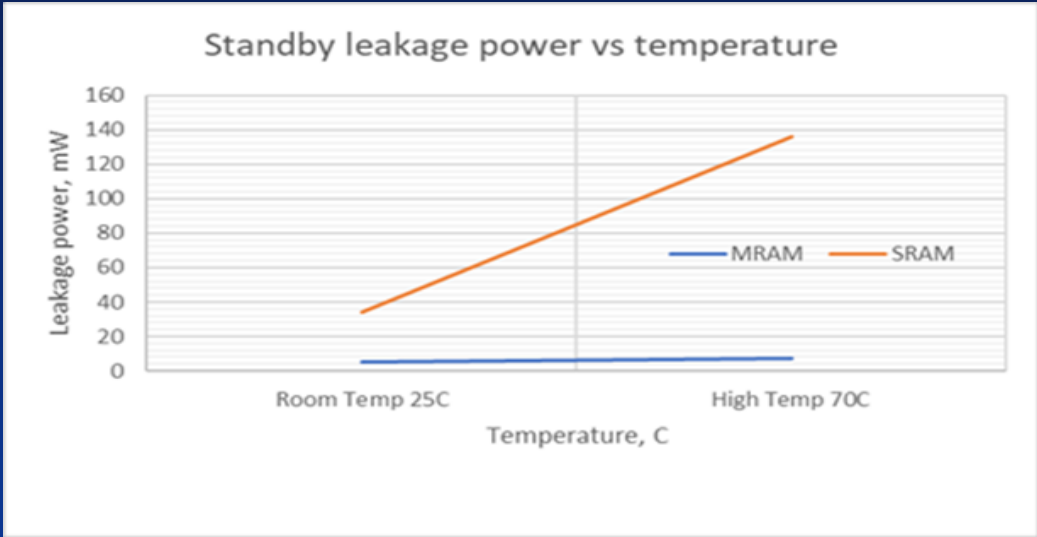
The higher the bit model,
The better the accuracy.

Intelligence level

Higher Density



Low or No Leakage Power



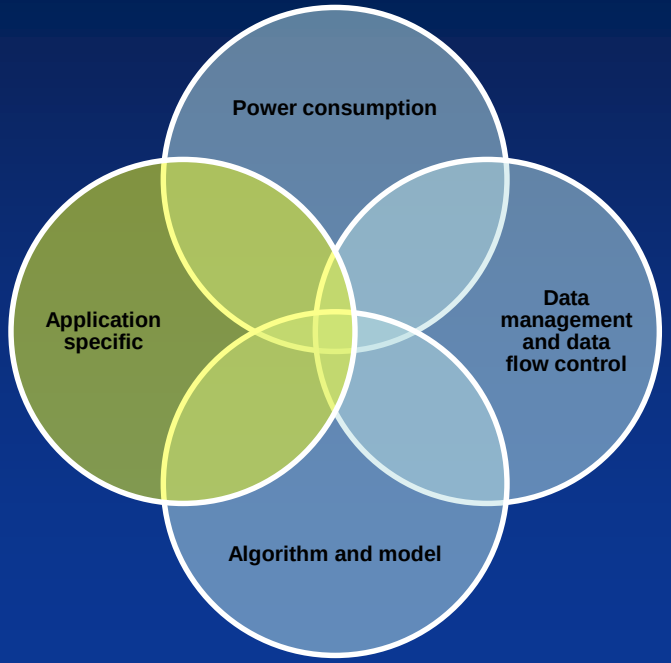
Active power

Memory intensive applications

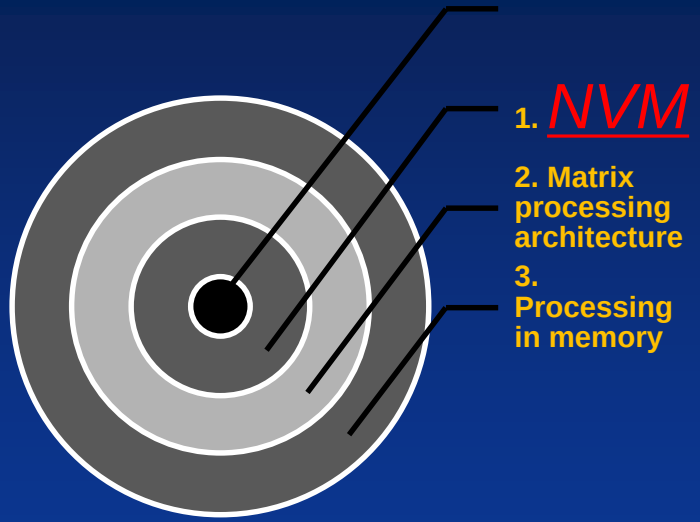
Low duty cycle applications

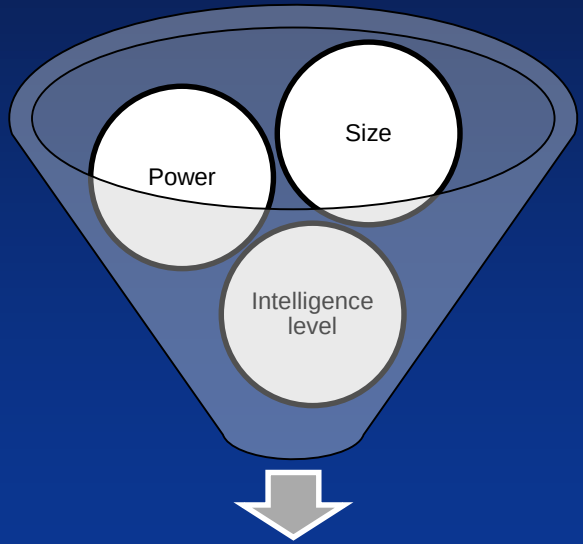
Architecture

Challenges for the Edge AI

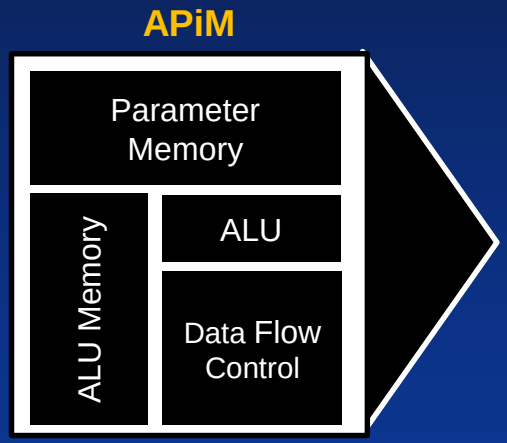


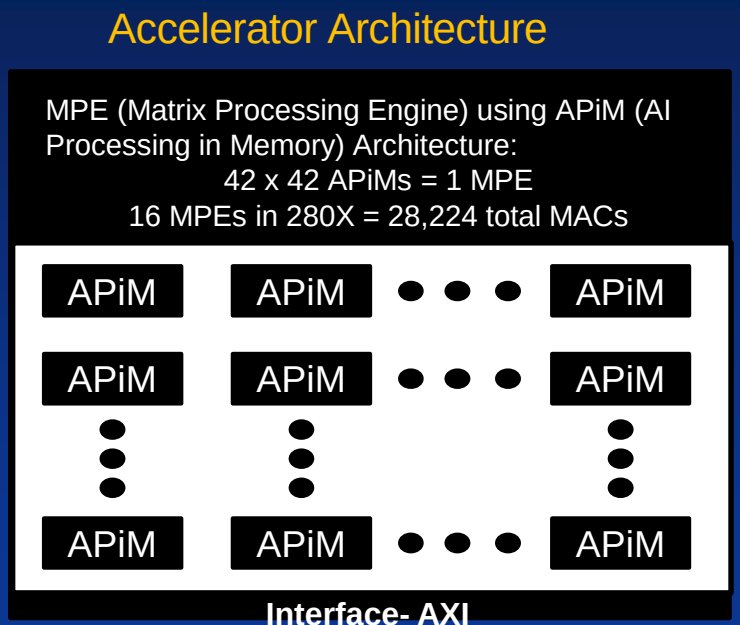
Hardware, Software and Memory co-design



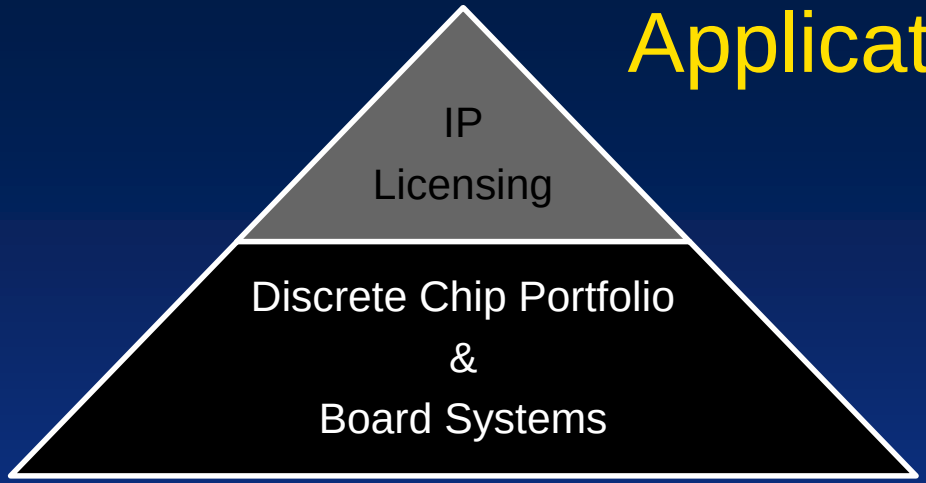


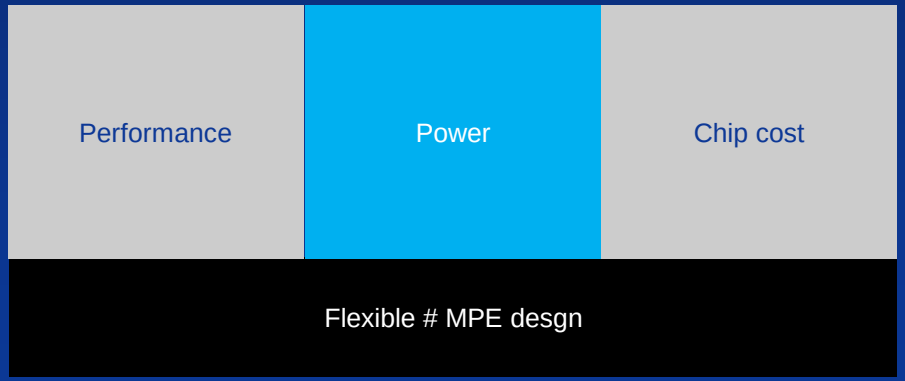
Accelerator Architecture





Application specific

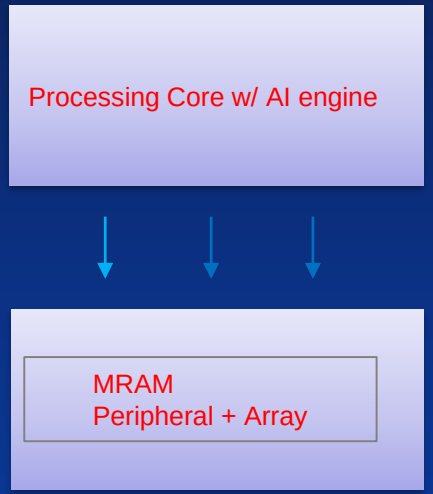




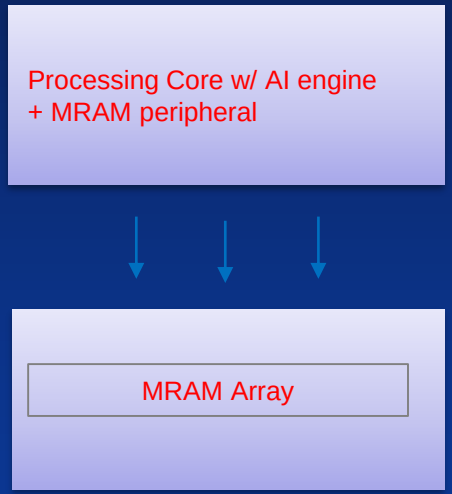
Embedded MRAM AI Accelerator

Hardware/Software/MRAM co-design

Typical Embedded MRAM AI Design

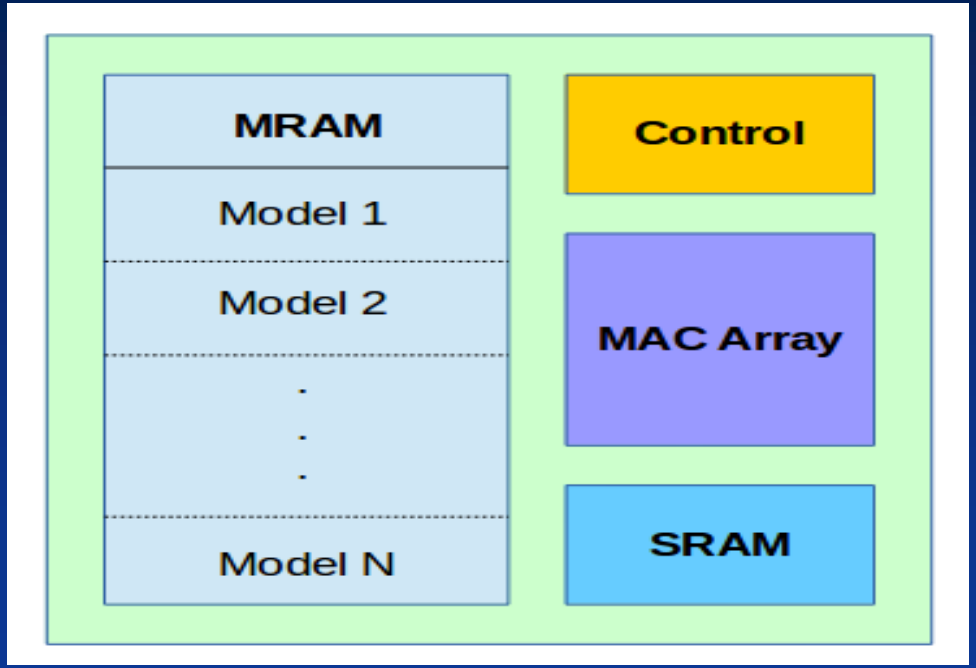


GTI's GME Engine Design



	Other eMRAM	RRAM (R&D)	GTI'S eMRAM	SRAM
No Leakage power	Yes	Yes	Yes	Very High
Cell Size	Small	Small	Small	Very Big
No Size penalty for distributed memory	No	No	Yes	No
Endurance	E6 to E12	E5	E9 to E15	E15
NVM	Yes	Yes	Yes	No
Latency	High	High	Low	Low
Dynamic power	High	High	Low	Low

Block Diagram for MRAM AI chip to load multiple models



Industry's 1st Production AI Chip (Lightspeeur®2802M) With Embedded MRAM

Additional Specifications:

- 9.9 TOPS/W
- 22nm ASIC
- 20-50% power savings (SRAM, "other MRAM")

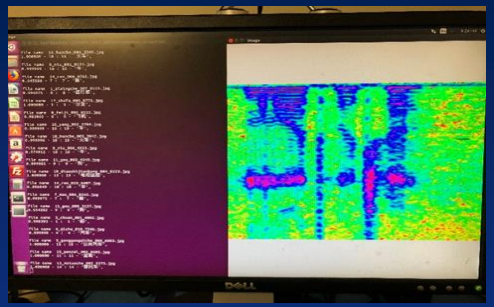
Customization Options for Large Scale Customers:

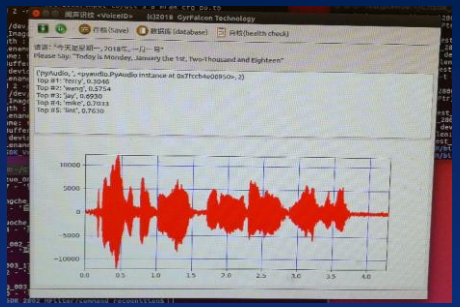
- Up to 5 ns Read Speed
- Supports multiple models on single chip
- **Flexible intelligence level**

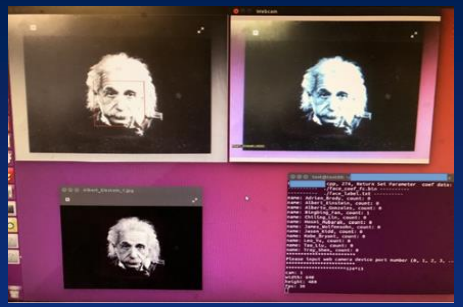
The GME™ (Gyrfalcon MRAM Engine)

	GTI's eMRAM	Other eMRAM	SRAM	RRAM
Non-volatile Memory	✓	✓	✗	✓
No Power Leakage	✓	✓	✗	✓
Small Cell Size	✓	✓	✗	✓
No Size Penalty for Distributed Memory	✓	✗	✗	✗
Low Latency	✓	✗	✓	✗
Low Dynamic Power	✓	✗	✓	✗
Endurance	E9 - E15	E6 - E12	E15	E5

Multi-filters model demo







Voice command | Voice ID | Facial Recognition

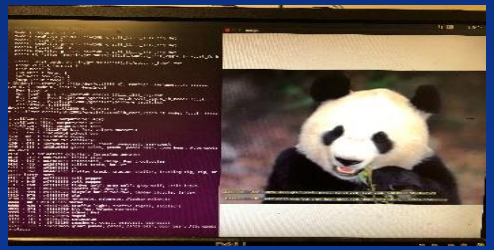


Image Classification (slide show)

Hardware/software/MRAM proprietary co-design

- Simplify circuit design
- Manufacturing friendly for foundries
  - high DR/R materials and different thermal budget processes…
- OST, SOT, voltage-controlled and MLC MRAM compatible
- Chip/wafer yield friendly

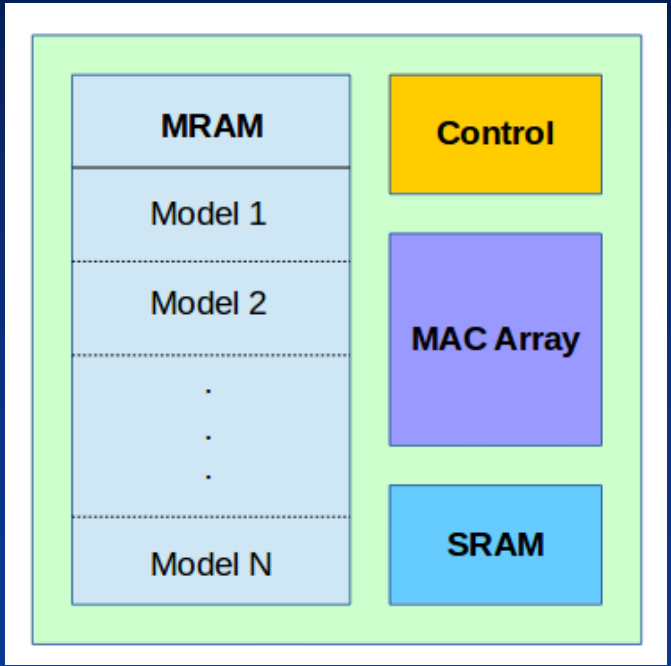
Case I – Car or Robot

- Open the door (Voice ID, voice commands or facial recognition);
- "It's me", "Wally, open the door"…….
- Start the engine;
- "Wally, let's go", "fire up" ……
- Turn on/off the radio, GPS, air condition, make phone call……**Local**
- More safe and secure for you as driver and your passengers!

Case II -- Remote areas or smart city/smart home

- **Power down due to nature disasters, human error or machine failure ---- no reload needed**

Case III –
1mw or less AI accelerator for AIoT



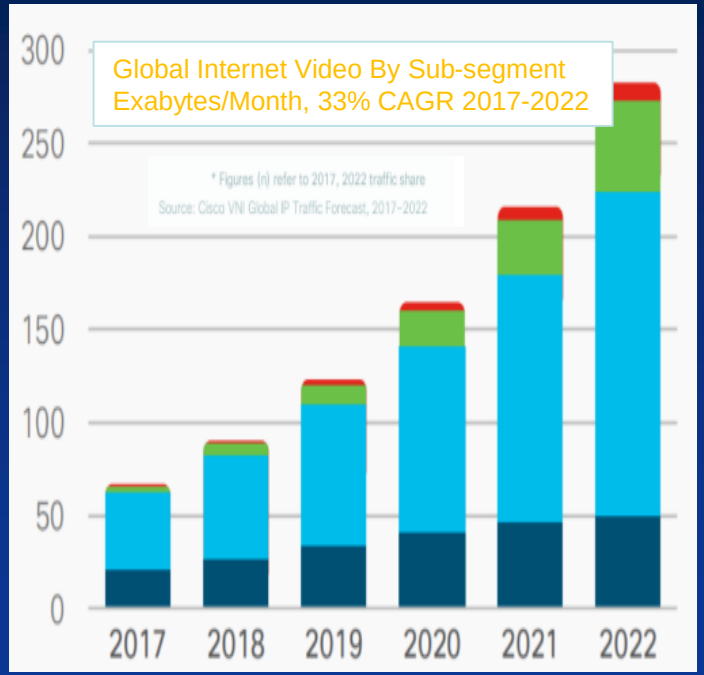
** Replace SRAM with SRAM like MRAM (ns read/write, e16 endurance besides memory density…)

**memory occupy 75 to 80+ % area

**Replace SRAM with MRAM, chip size will be half?
Besides power saving……

**** less than 1mw AI accelerator**

Case IV
Video Content Taking Over Global Data Traffic Creates Opportunities AND Challenges……



By 2021 Almost 70% of All Data Traffic Will Be Video*

Industry "Pain Points" From Uncaptured Full Value of Video Content:

1. **Can't search images within video**
2. **Existing video lacks content labels**

Video “Pain Points” Create New Industry Opportunities

Visual “Search” for Specific Content

1. Search based on picture not descriptive text
2. Images used to search existing video files for matching content

Video “Data Mining/Retrieval”

1. Converts existing video to video with labeled content
2. Labeled content becomes searchable within each frame

- Superior user experiences
- New business models

Until Now, No Solutions..

No Current Industry Standards

- For Video Data Mining/Retrieval

Tremendous Computing Efforts

- Expensive computing resources
- Manual labor intensive

- 10^3 descriptors per frame
- 10^6 comparisons to match two frames!
- 300k hours archive ~ 10^{10} frames
- $10^6 \times 10^{10} = 10^{16}$ comparisons
- 3×10^{18} FLOPS Needed!

Performance Comparison

	CDVA Standard	GTI CDVA
Input Size	640x480	640x480
CNN Model Format	Floating-point VGG-16	Fixed-point VGG-16
CNN Model Size	59M byte	4M byte
Output Vector (Descriptor) Size	512 byte	512 byte
Mean Avg Precision Score	86.81%	88.95%

PRECISE & SMALLER MODEL SIZE

UP TO **1.9x** FASTER EXTRACTION TIME THAN GPU

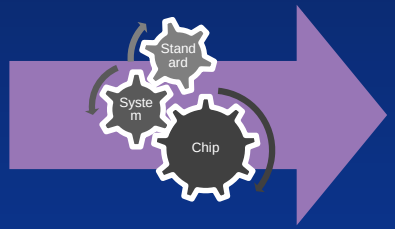
UP TO **16x** FASTER EXTRACTION TIME THAN CPU

	Pre-processing (ms)	CNN (ms)	Extract vector (ms)	Total (ms)
GTI CDVA	80 ms	88 ms	15 ms	190 ms
CDVA Standard (GPU)	260 ms	78 ms	15 ms	353 ms
CDVA Standard (CPU)	260 ms	2800 ms	15 ms	3075 ms

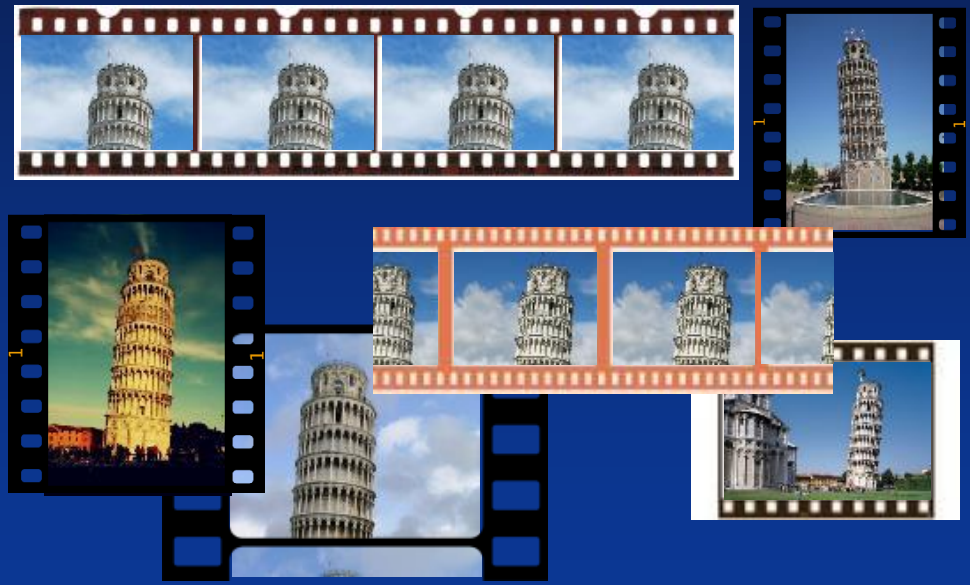
ENHANCE YOUR SEARCH EXPERIENCES

Search Query

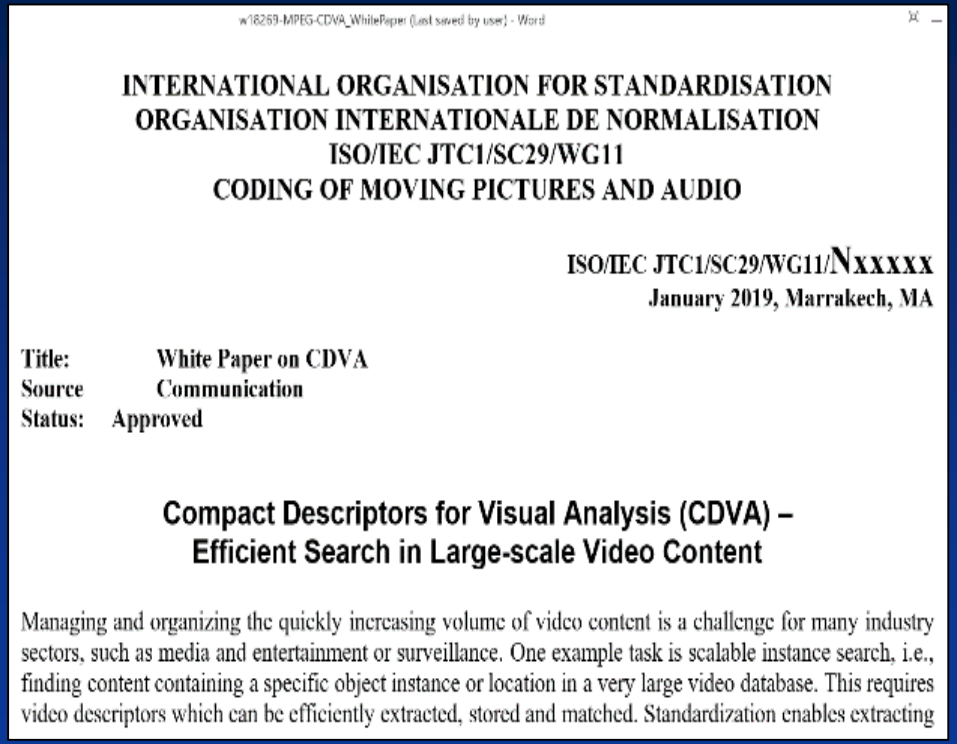
Visual Search



Search Results
(from video archives, service providers, database…)



CDVA Provides the Industry Standard…..



w18269-MPEG-CDVA_WhitePaper (Last saved by user) - Word

INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11/NXXXXX
January 2019, Marrakech, MA

Title: White Paper on CDVA
Source Communication
Status: Approved

Compact Descriptors for Visual Analysis (CDVA) – Efficient Search in Large-scale Video Content

Managing and organizing the quickly increasing volume of video content is a challenge for many industry sectors, such as media and entertainment or surveillance. One example task is scalable instance search, i.e., finding content containing a specific object instance or location in a very large video database. This requires video descriptors which can be efficiently extracted, stored and matched. Standardization enables extracting

Compact descriptors for video analysis for search and retrieval applications:

1.Enable design of interoperable object instance search applications

2.Ensure high matching performance of objects

Approved Neural Network:

VGG16(Trained By ImageNet ILSVRC) – No IP Issues

Thank you