

Vivienne Sze

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Computing challenge for self-driving cars

JACK STEWART TRANSPORTATION 02.06.18 08:00 AM

SELF-DRIVING CARS USE CRAZY AMOUNTS OF POWER, AND IT'S BECOMING A PROBLEM



Shelley, a self-driving Audi TT developed by Stanford University, uses the brains in the trunk to speed around a racetrack autonomously.

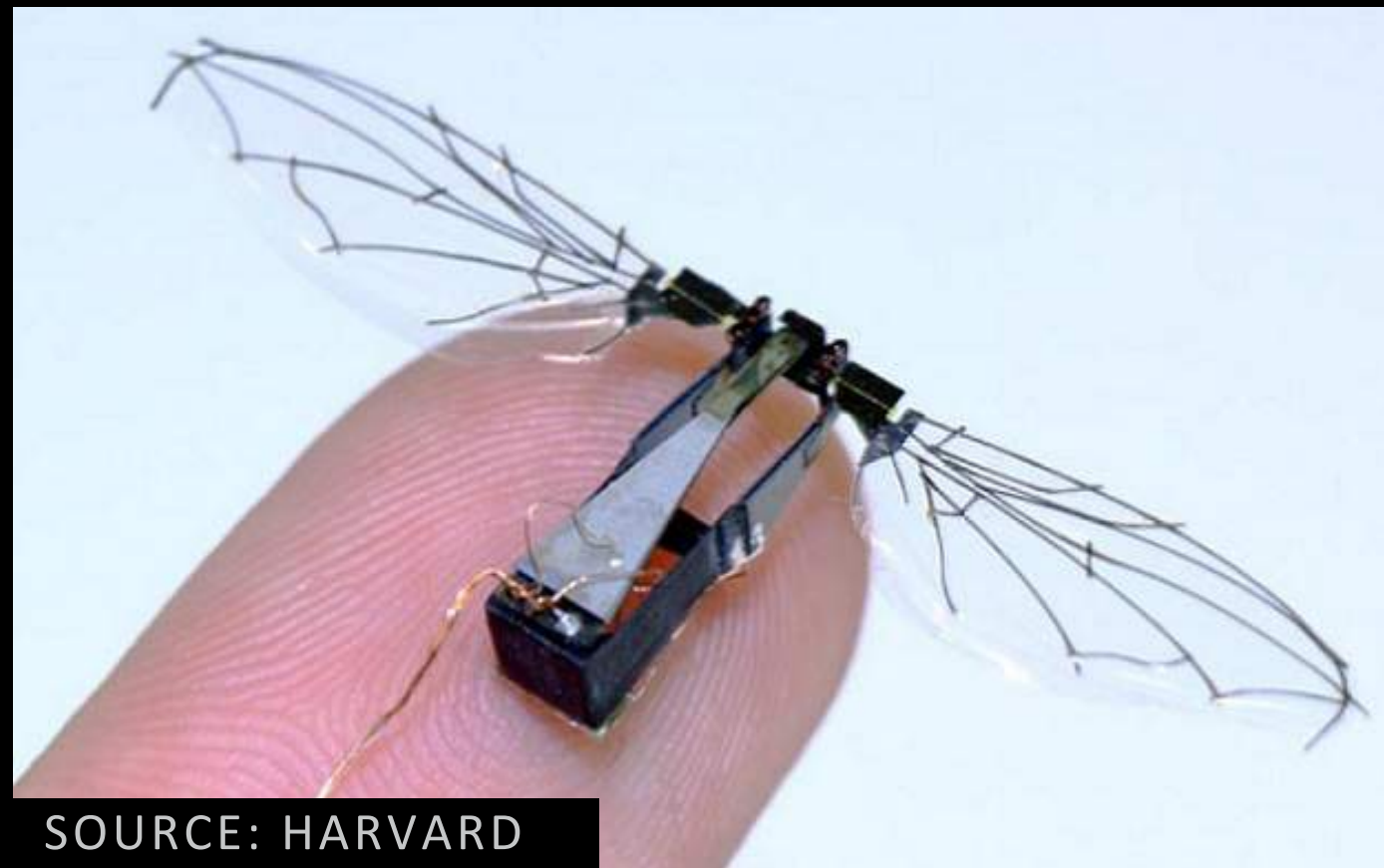
NIKKI KAHN/THE WASHINGTON POST/GETTY IMAGES

Cameras and radar generate ~6 gigabytes of data every 30 seconds

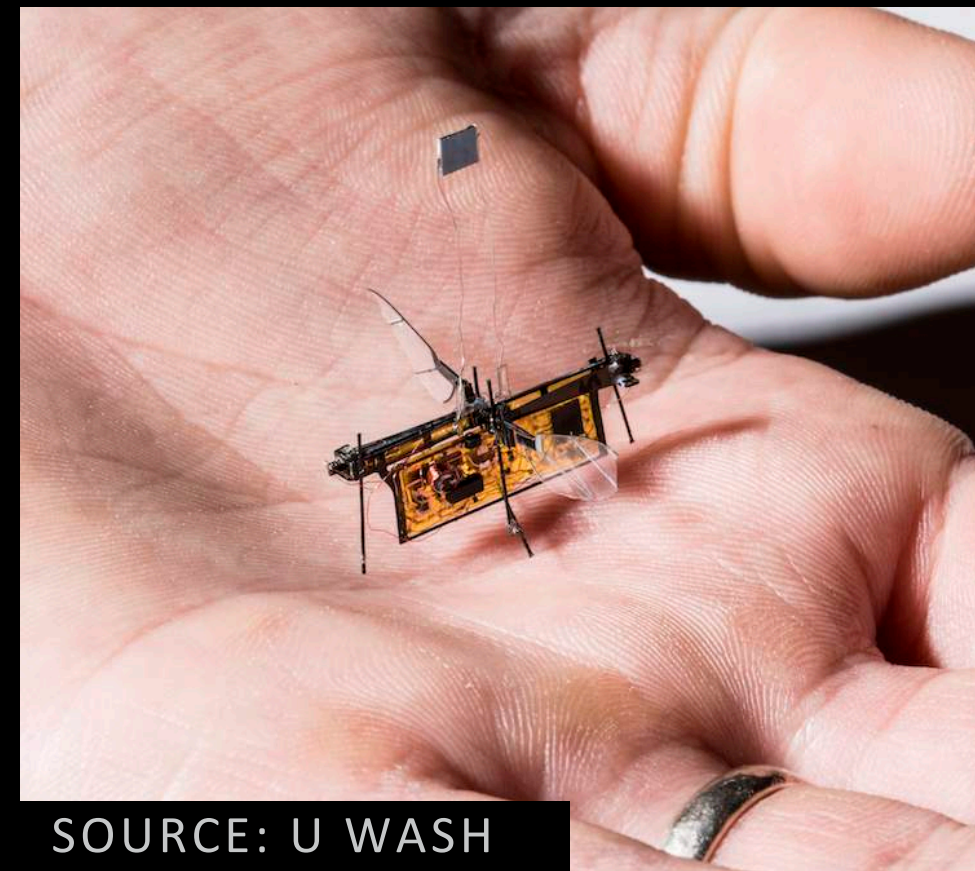
Self-driving car prototypes use approximately 2,500 Watts of computing power

Generates wasted heat and some prototypes need water-cooling

Robots consuming < 1 Watt for actuation



SOURCE: HARVARD



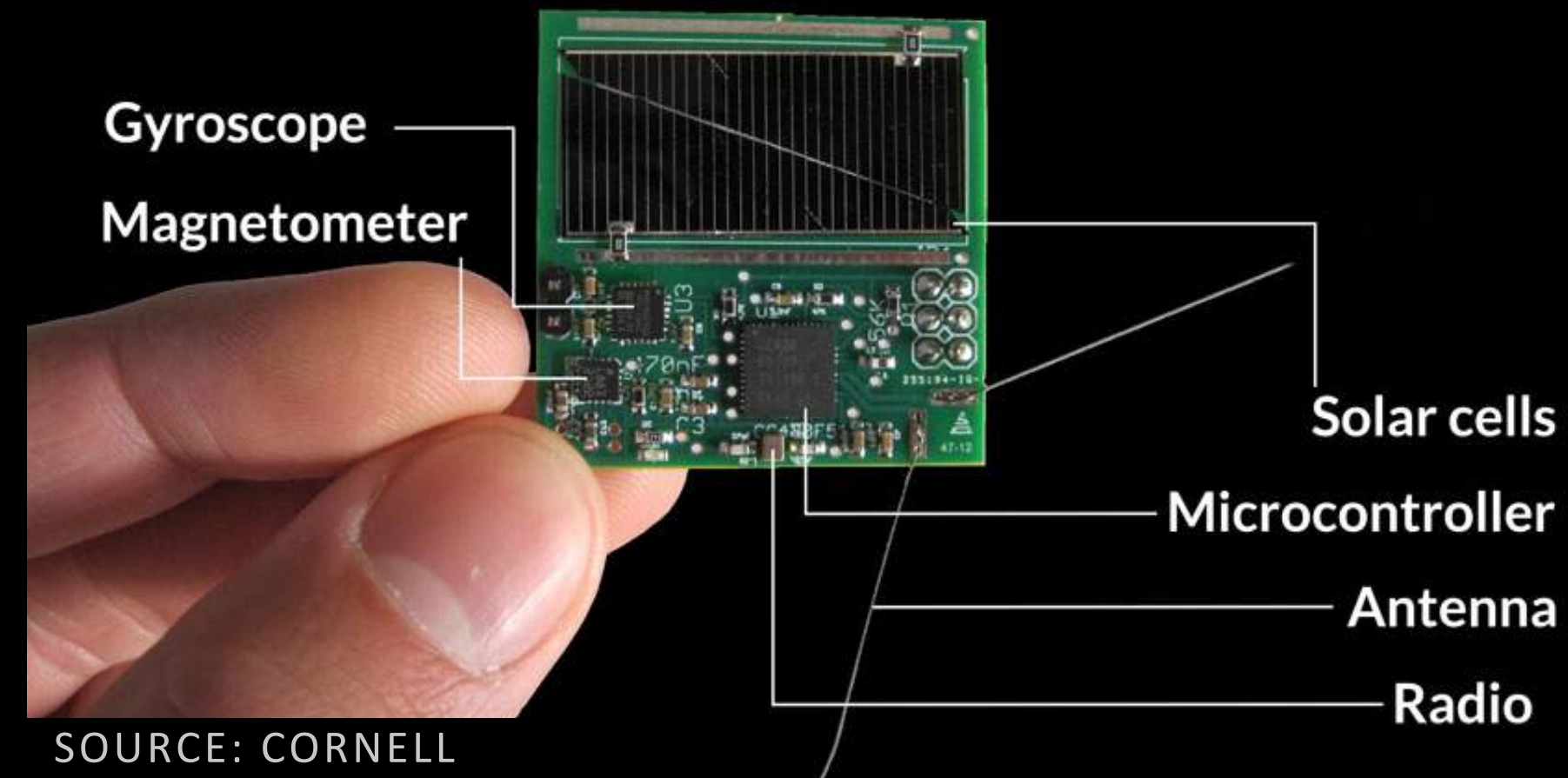
SOURCE: U WASH

Low energy robotics

- Miniature aerial vehicles
- Lighter than air vehicles
- Miniature satellites
- Micro unmanned gliders



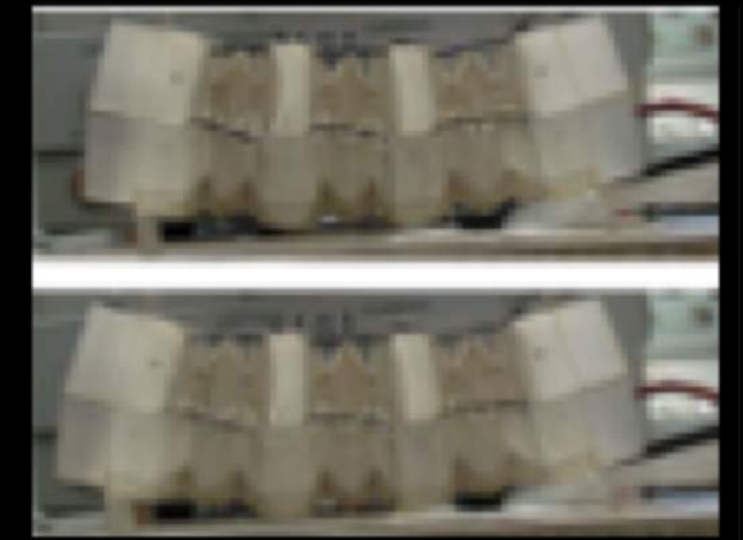
SOURCE: GEORGIA TECH



SOURCE: CORNELL



SOURCE: CMU



SOURCE: MIT, HARVARD



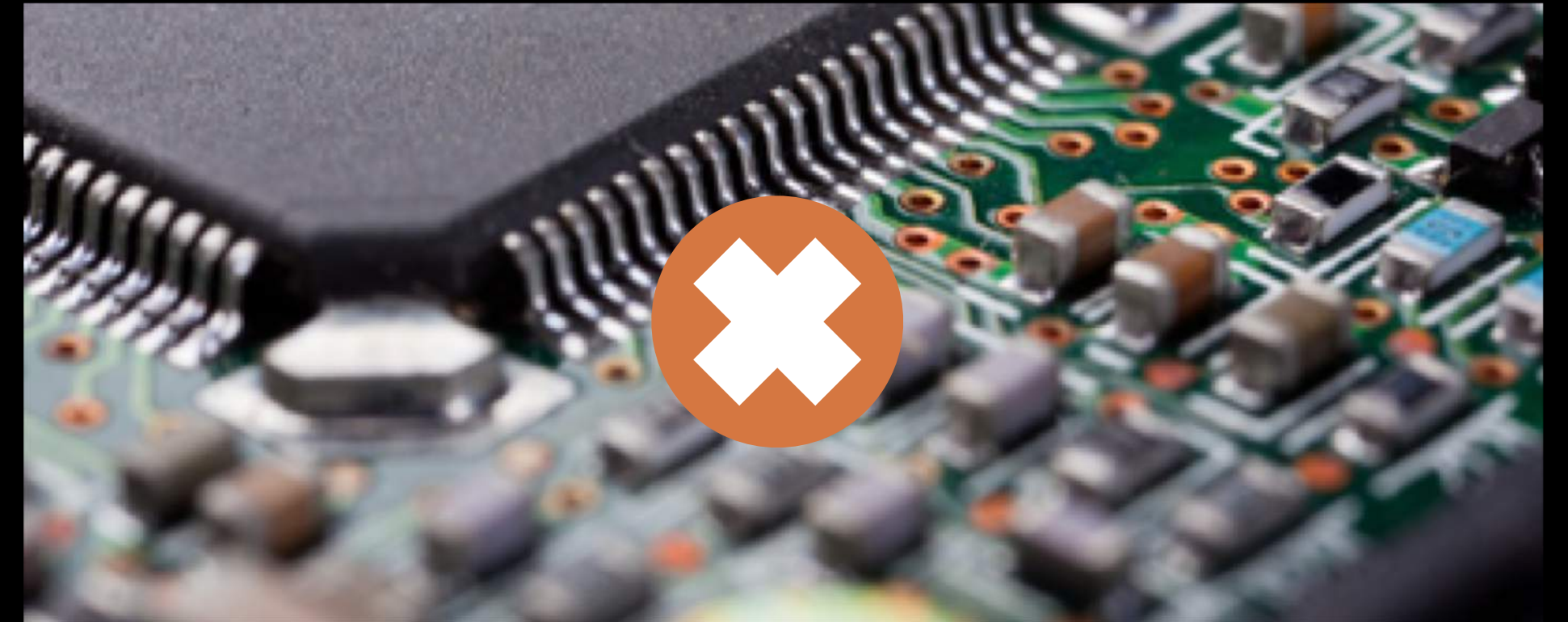
SOURCE: MIT, HARVARD

Existing processors consume too much power

< 1 Watt

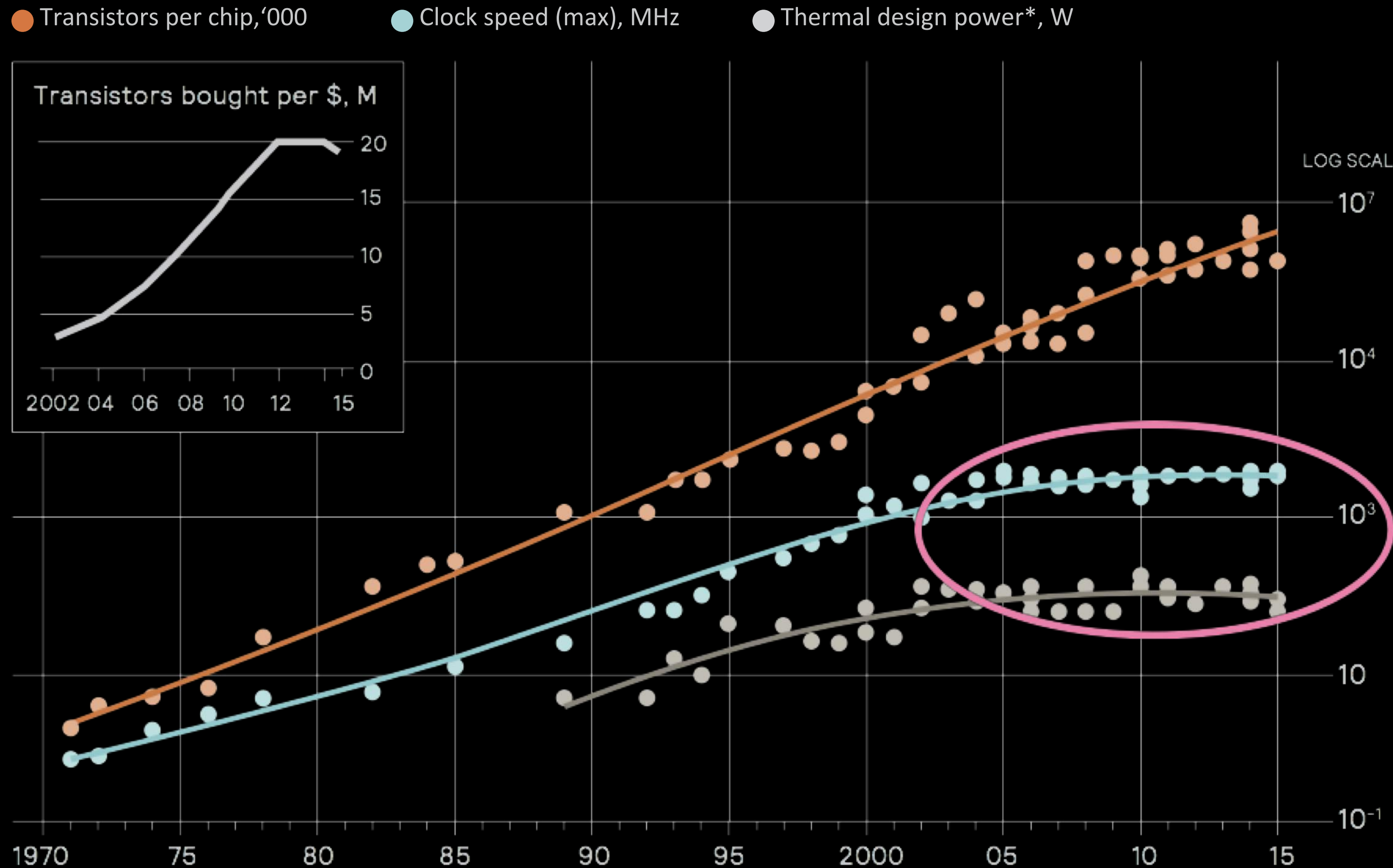


> 10 Watt



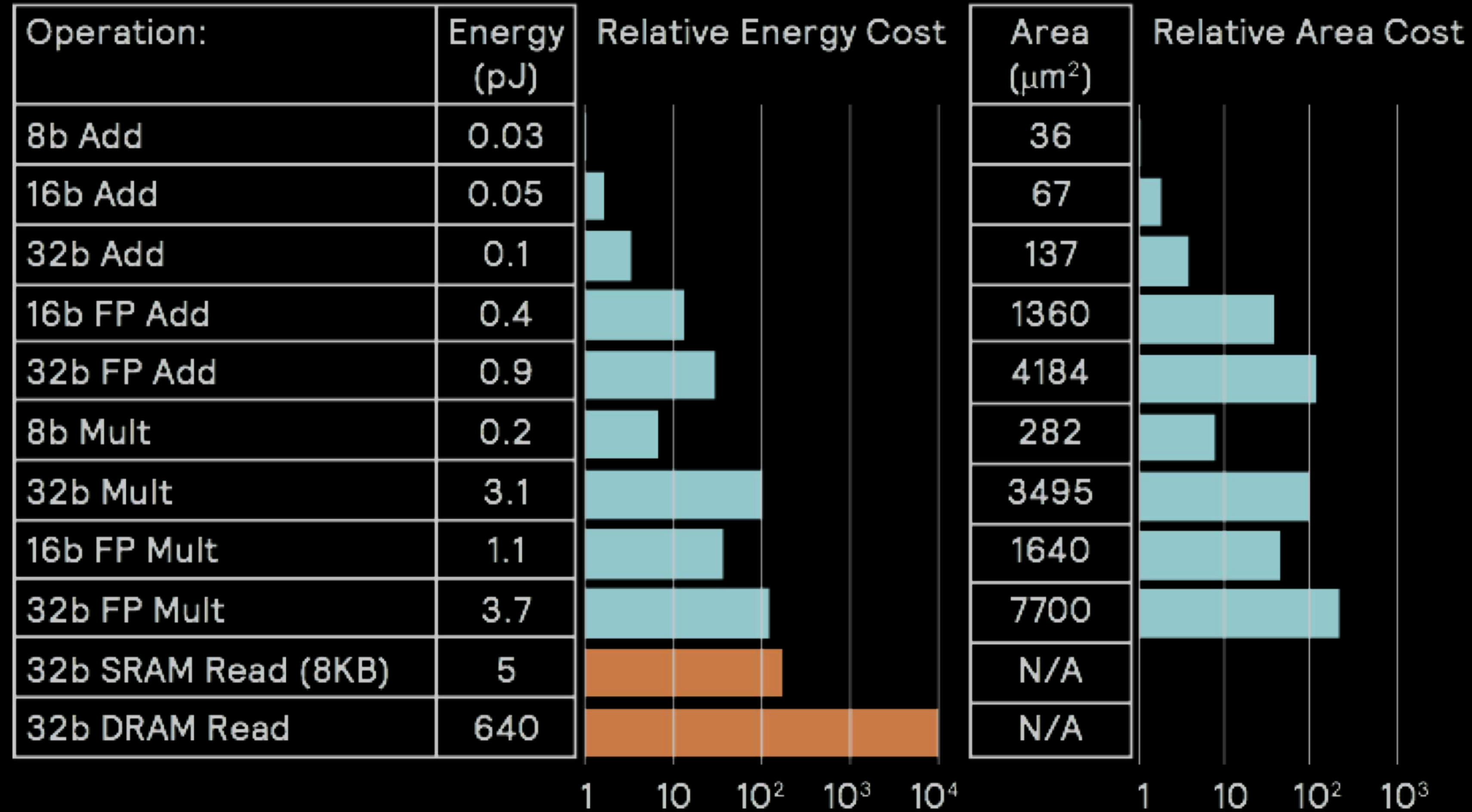
Transistors are NOT getting more efficient

Slowdown of Moore's Law and Dennard Scaling
General purpose microprocessors are not getting faster or more efficient



- Need specialized hardware for significant improvement in speed and energy efficiency
- Redesign computing hardware from the ground up!

Power dominated by data movement



Memory access is **orders of magnitude** higher energy than compute

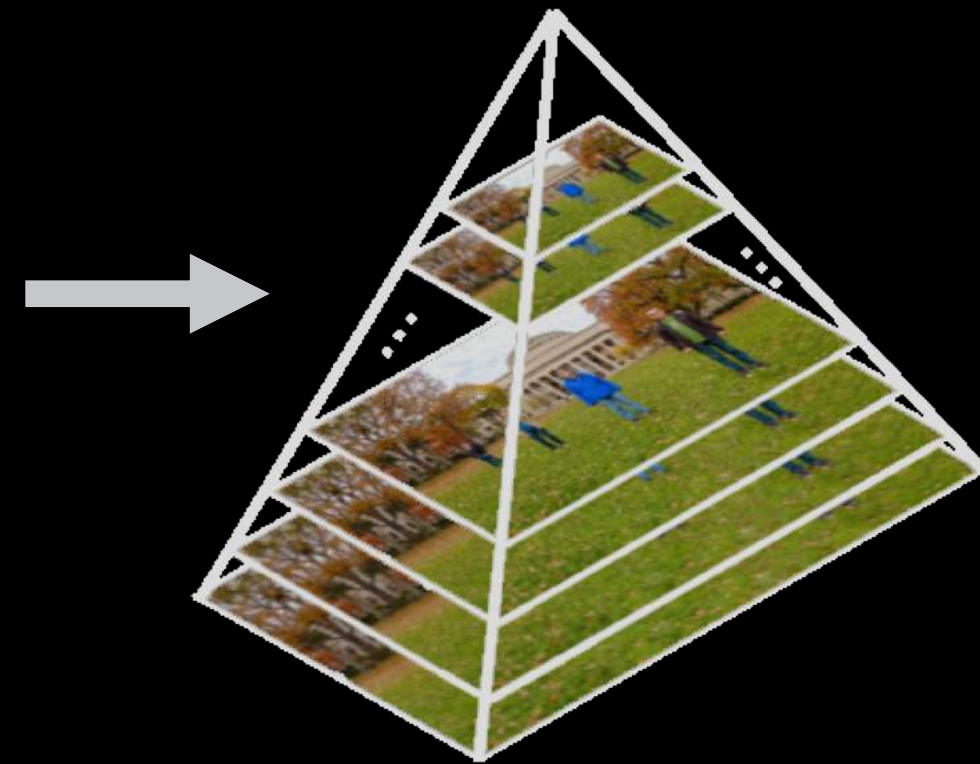
Autonomous navigation uses a lot of data

Semantic Understanding

- High frame rate
- Large resolutions
- Data expansion



2 MILLION PIXELS



10x-100x MORE PIXELS

Geometric Understanding

- Growing map size



Visual-inertial localization

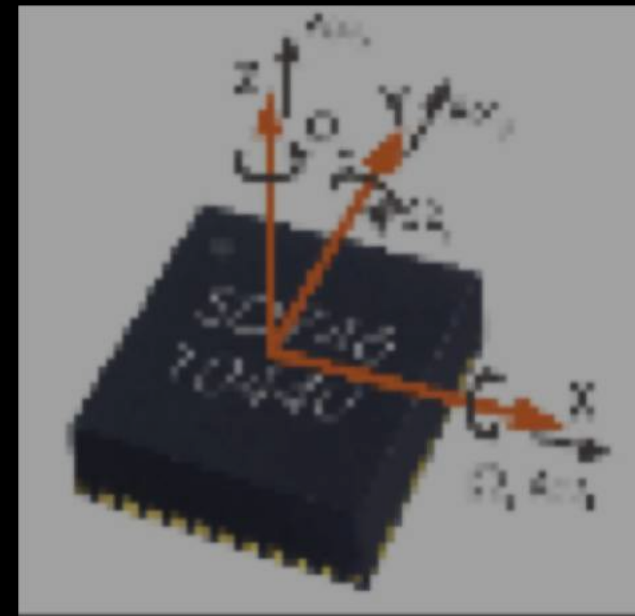
Determines location/orientation of robot from images and IMU

Image sequence



IMU

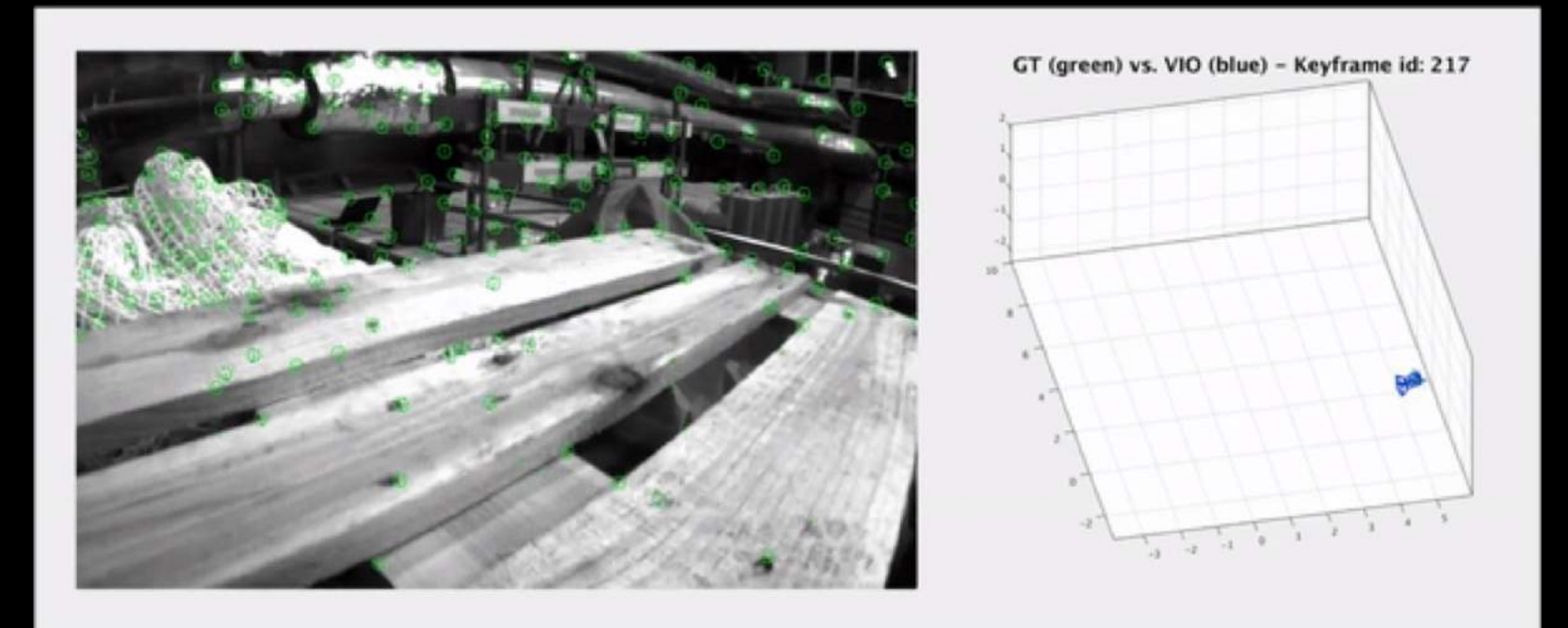
INERTIAL MEASUREMENT UNIT



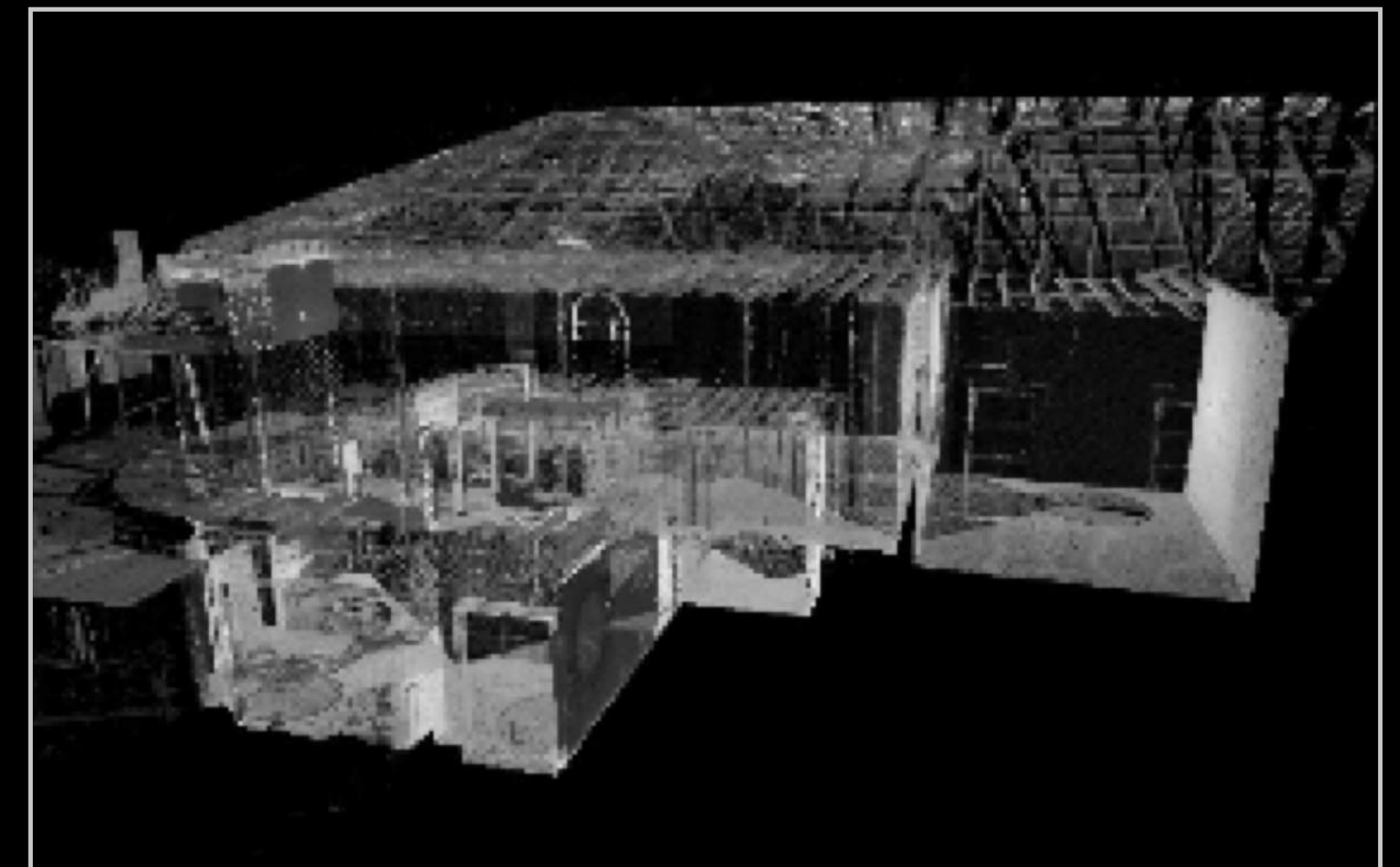
*SUBSET OF SLAM ALGORITHM
(SIMULTANEOUS LOCALIZATION
AND MAPPING)

Visual-Inertial
Odometry (VIO)

Localization

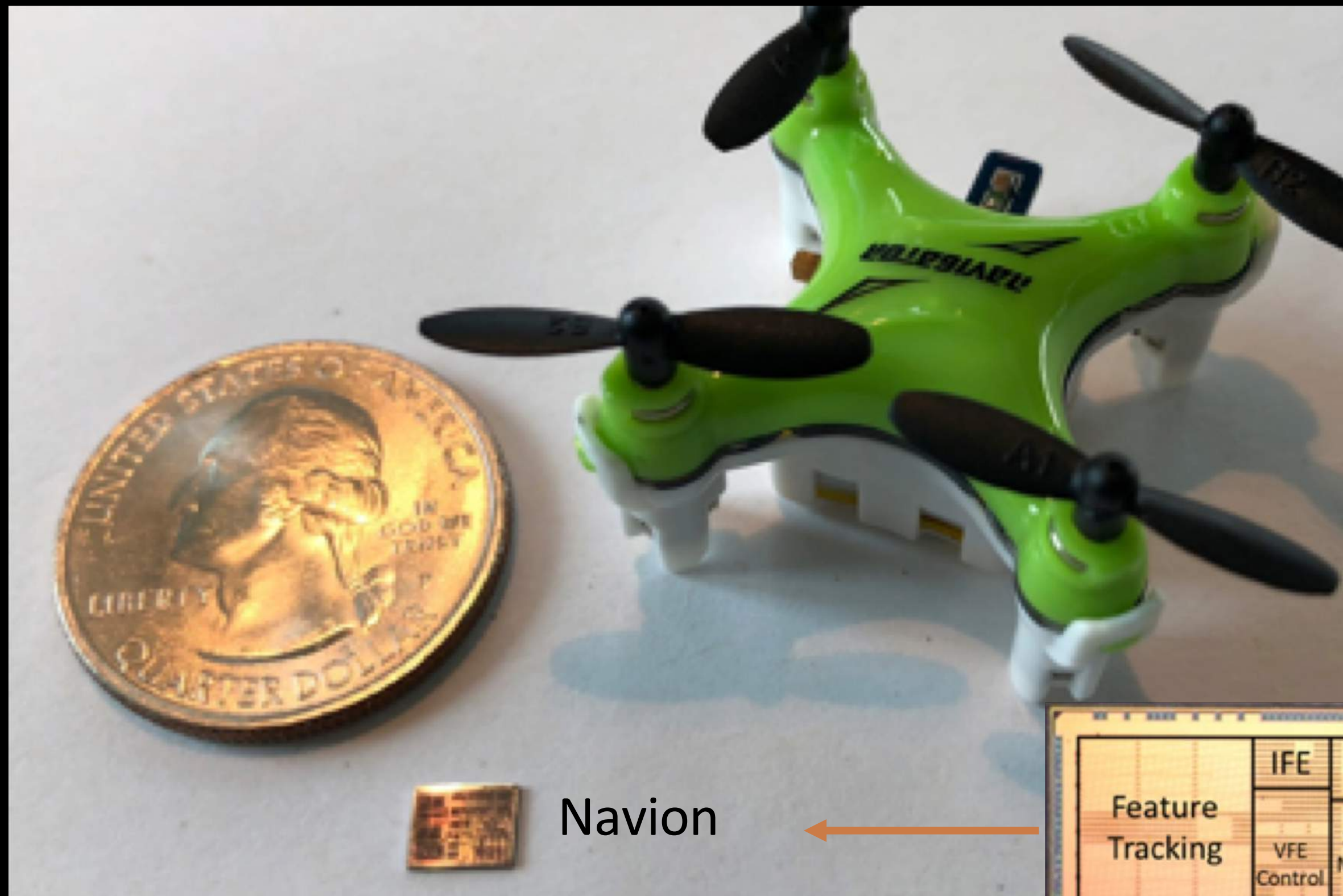


Mapping



Localization at under 25 mW

First chip that performs complete Visual-Inertial Odometry

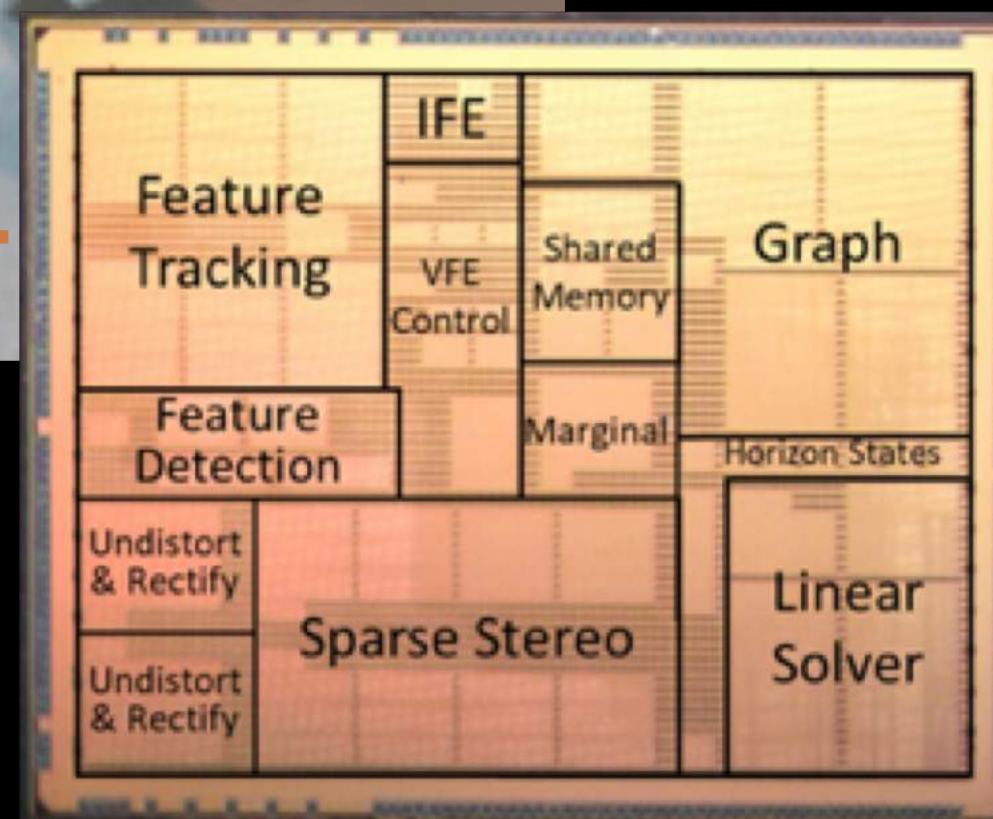


Front-End for Camera
(Feature detection, tracking, and outlier elimination)

Front-End for IMU
(Pre-integration of accelerometer and gyroscope data)

Back-End Optimization of Pose Graph

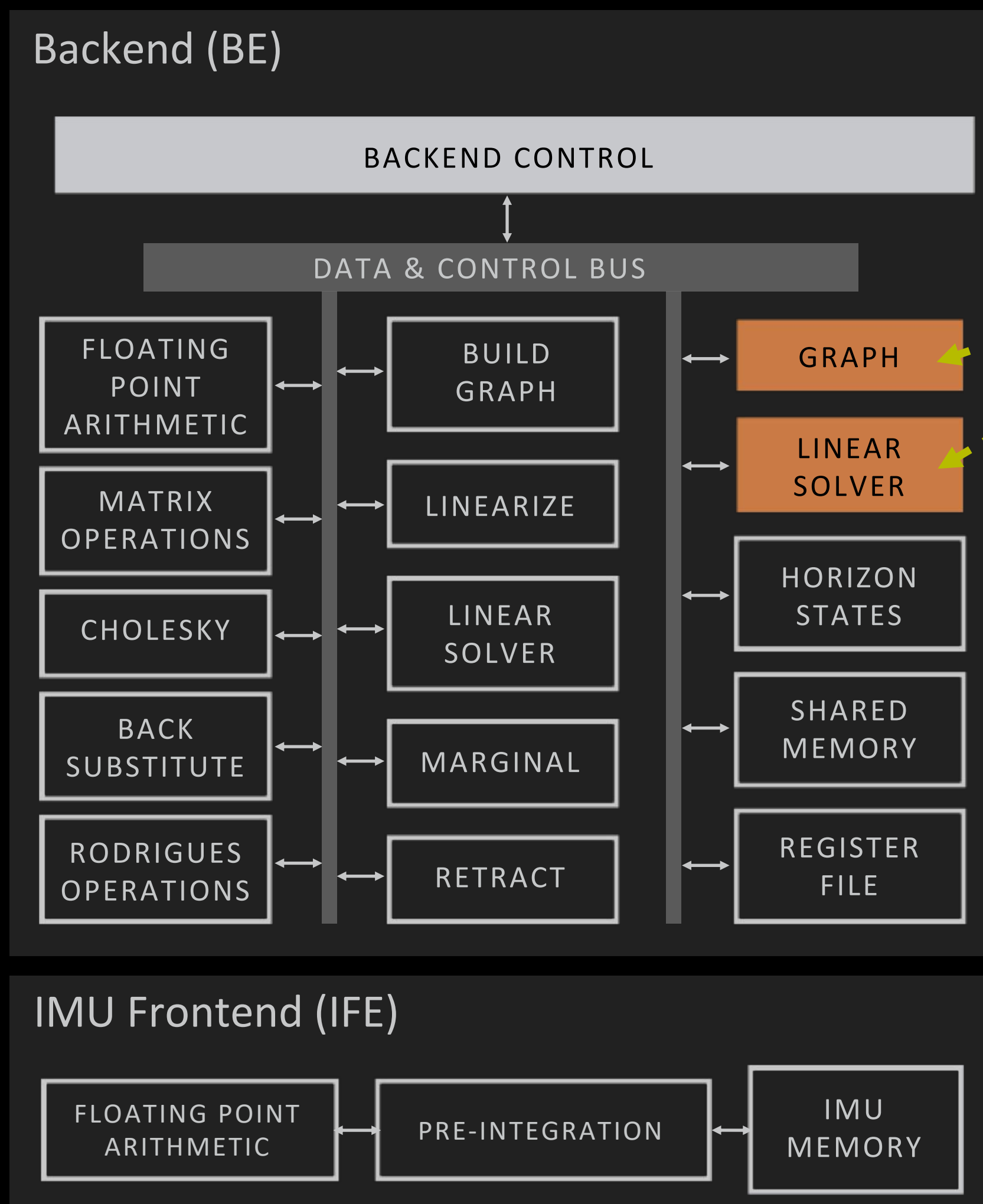
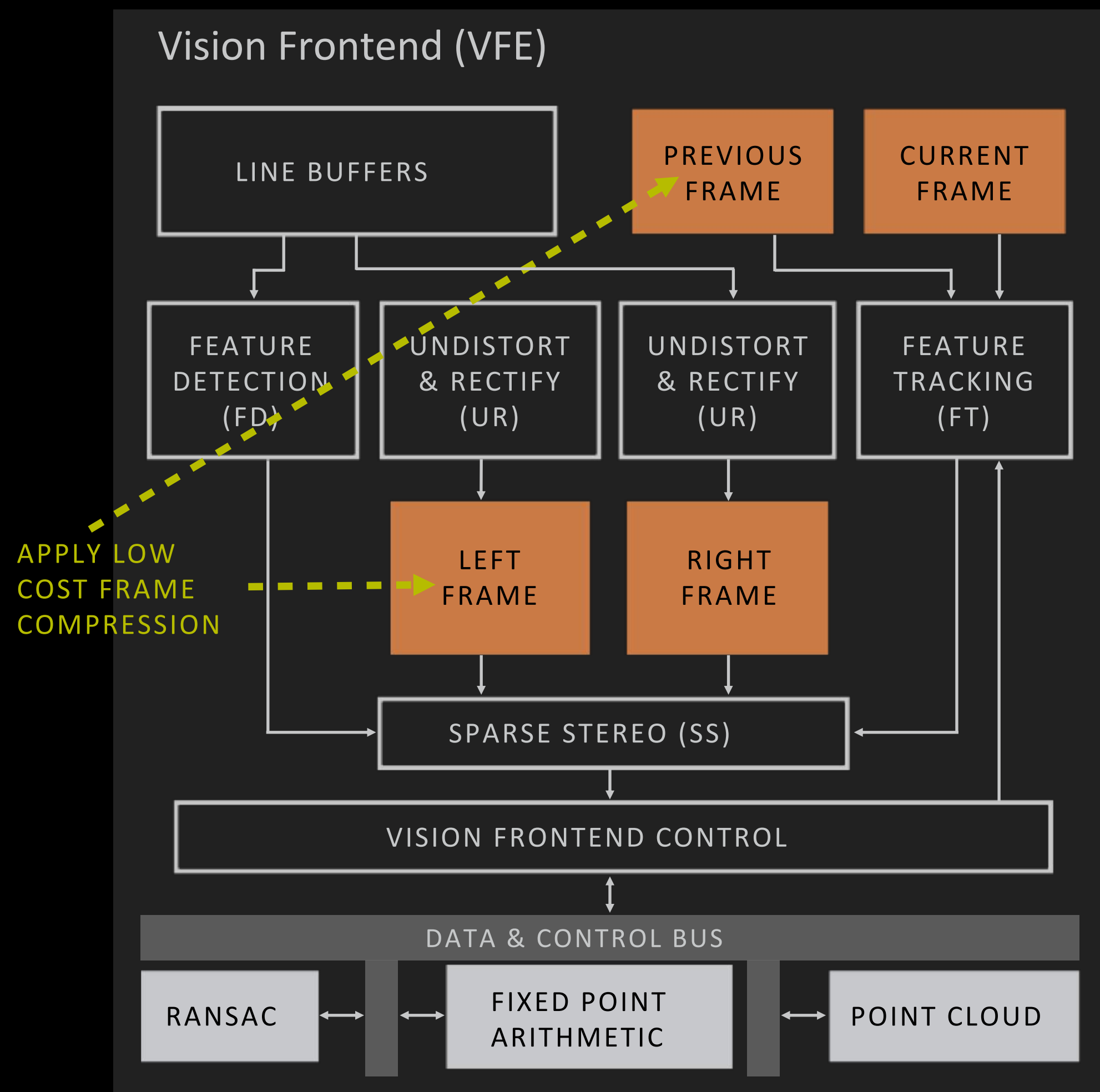
Consumes **684×** and **1582×** less energy than mobile and desktop CPUs, respectively



Technology	65nm CMOS	Supply	1 V
Chip area (mm ²)	4.0 × 5.0	Resolution	752 × 480
Core area (mm ²)	3.54 × 4.54	Camera Rate	28 – 171 fps
Logic Gates	2,043 kgates	Keyframe Rate	16 – 90 fps
SRAM	854KB	Average Power	24 mW
VFE Frequency	62.5 MHz	GOPS	10.5 – 59.1
BE Frequency	83.3 MHz	GFLOPS	1 – 5.7

Key methods to reduce data size

Navion: Fully integrated system — no off-chip processing or storage

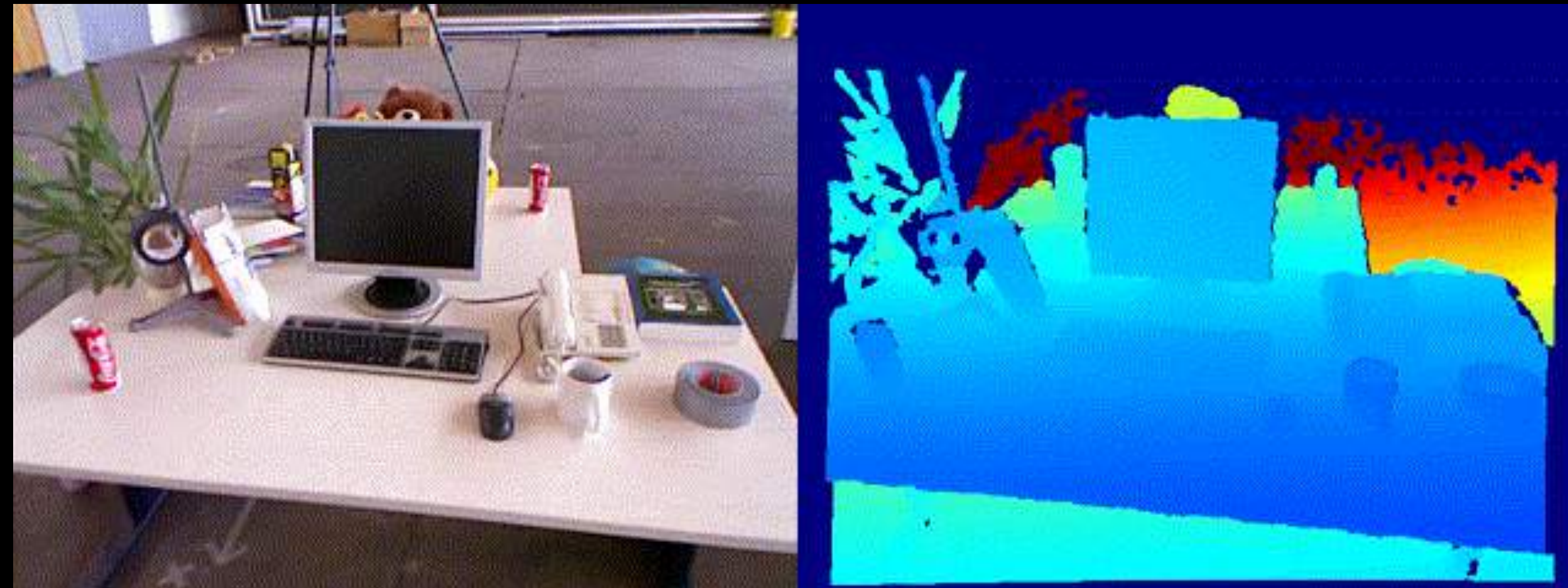


EXPLOIT SPARSITY IN GRAPH AND LINEAR SOLVER

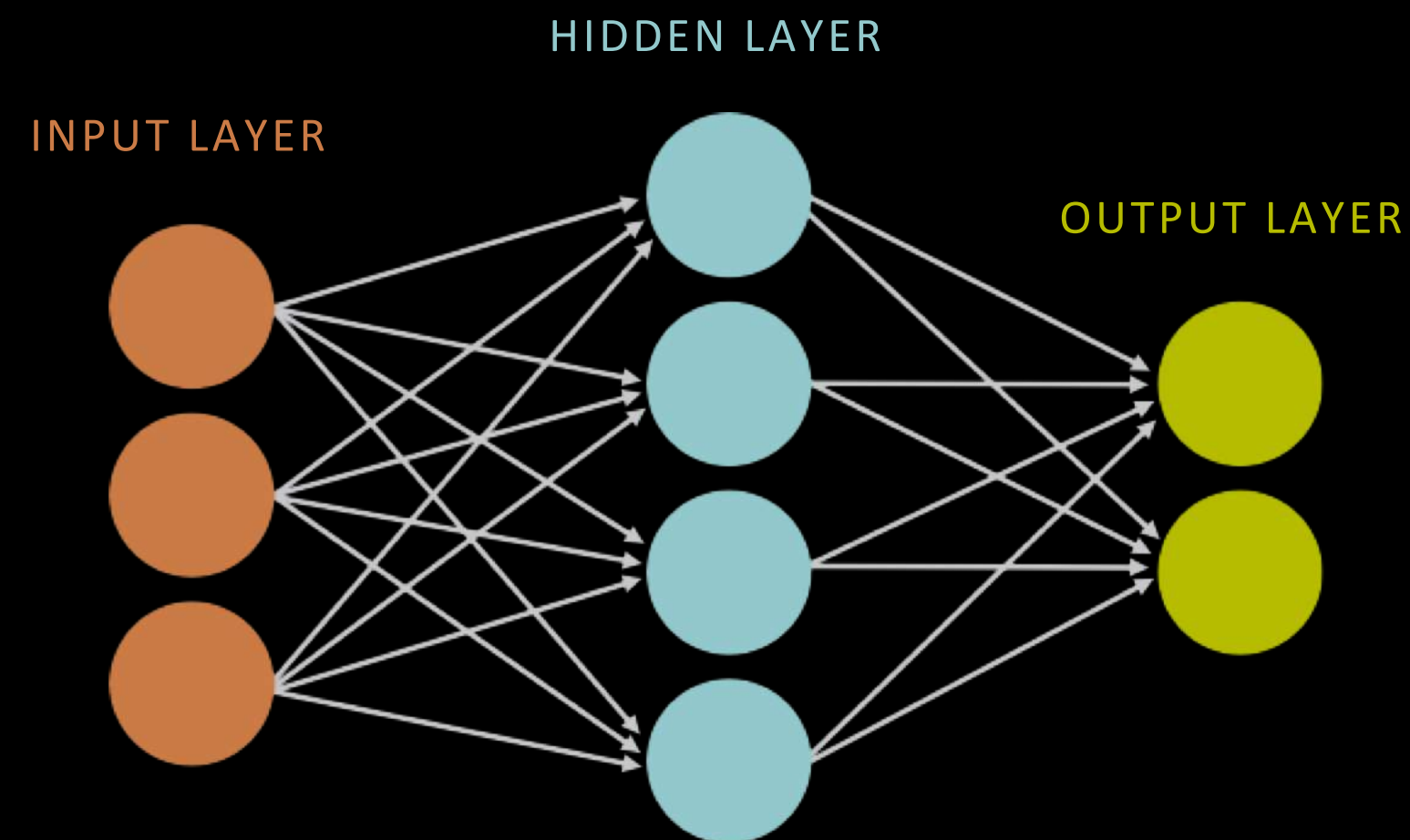
Use **compression** and **exploit sparsity** to reduce memory down to **854kB**

Understanding the environment

Depth Estimation



Semantic Segmentation



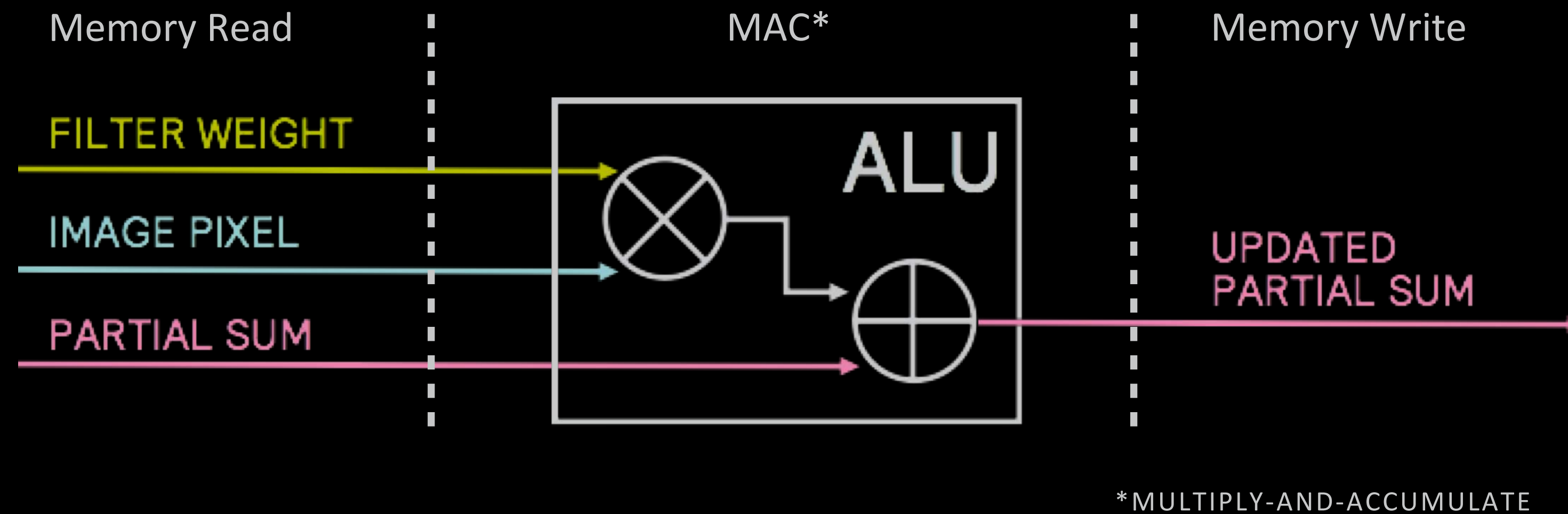
State-of-the-art approaches use **Deep Neural Networks** which require up to several hundred millions of operations and weights to compute!

> 100× more complex than video compression

Properties we can leverage

Operations exhibit **high parallelism**
→ high throughput possible

Memory Access is the Bottleneck



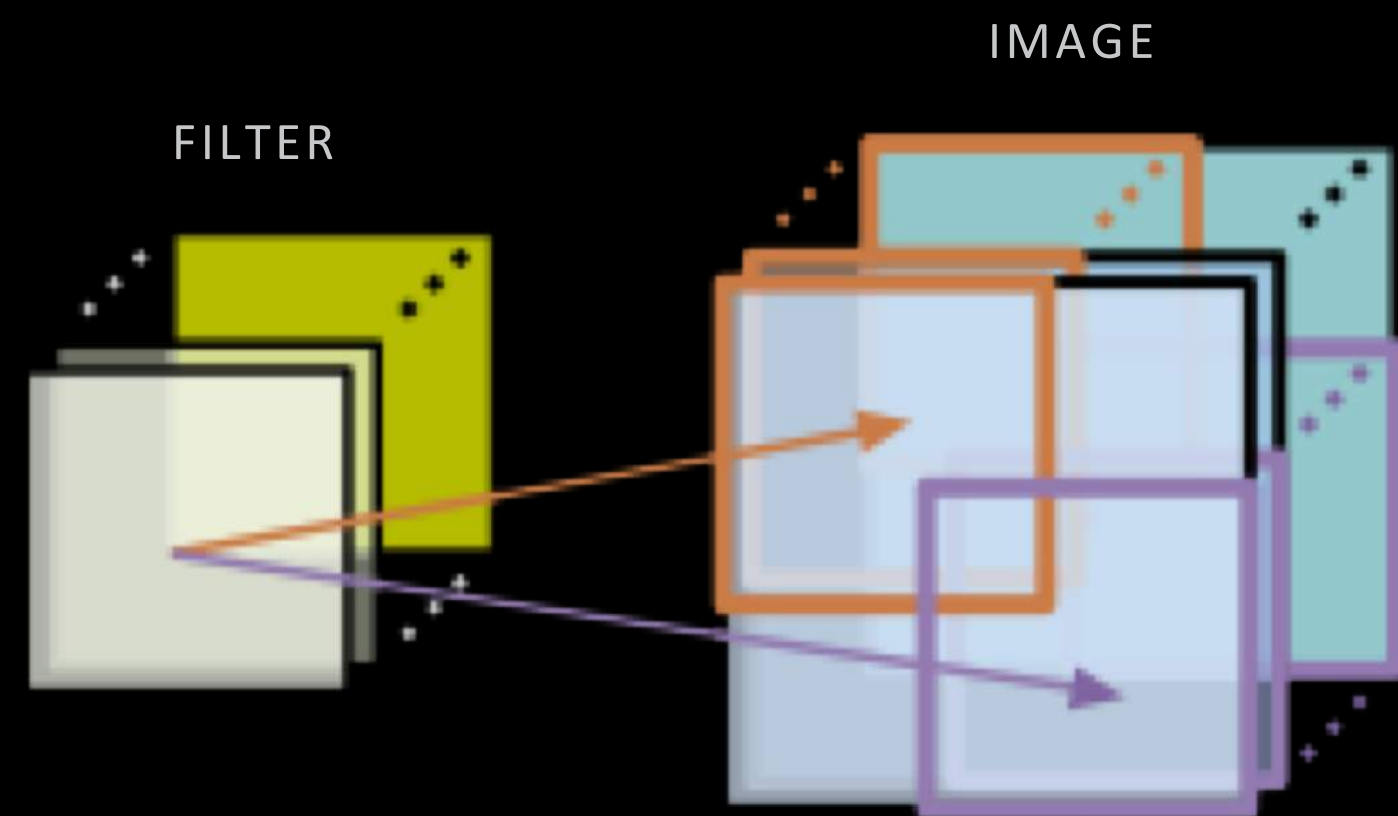
Worst Case: all memory R/W are **DRAM** accesses

Example: AlexNet has 724M MACs
→ 2896M DRAM accesses required

Properties we can leverage

Operations exhibit **high parallelism**
→ high throughput possible

Input data reuse opportunities (up to 500x)



CONVOLUTIONAL REUSE
(PIXELS, WEIGHTS)

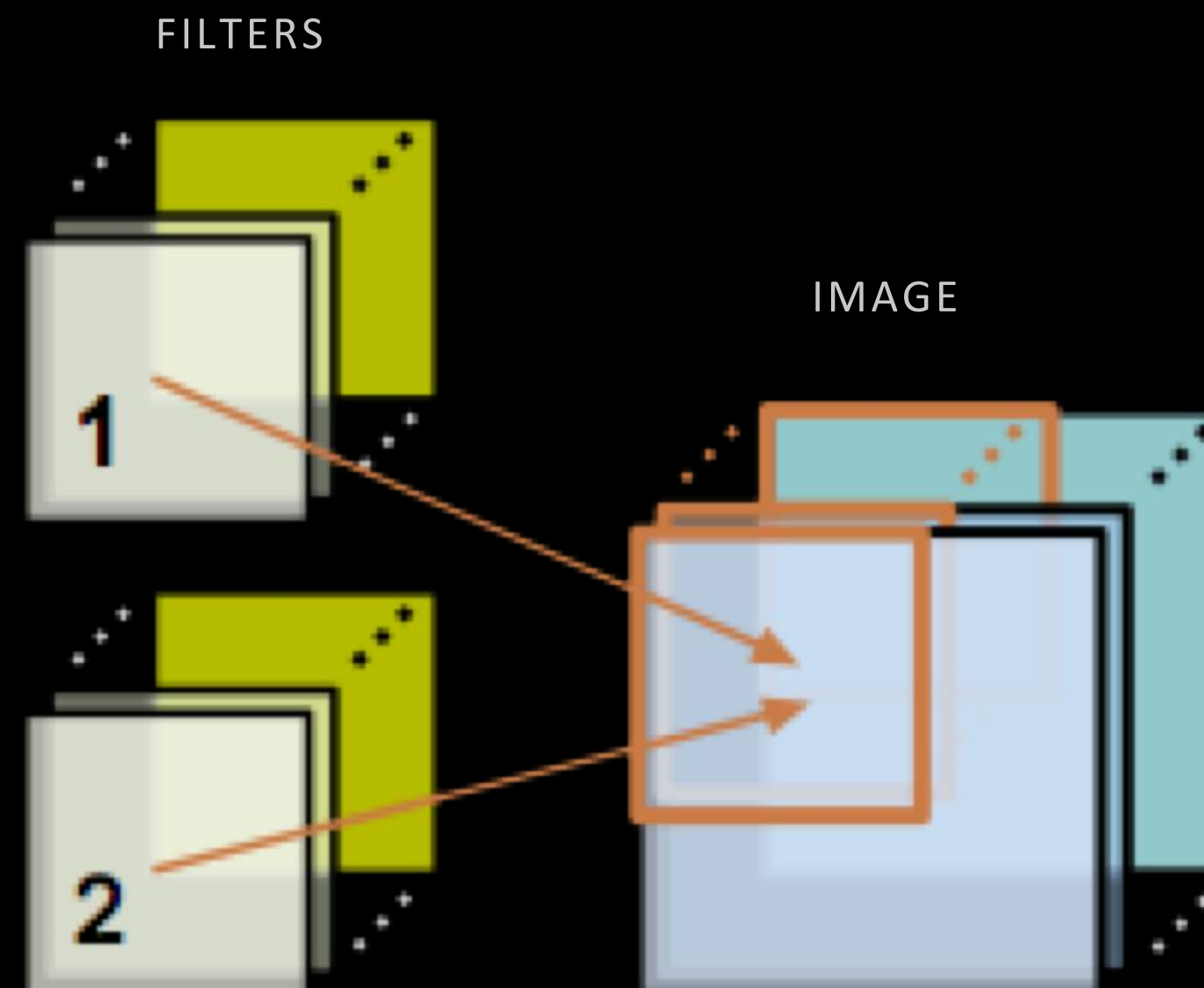
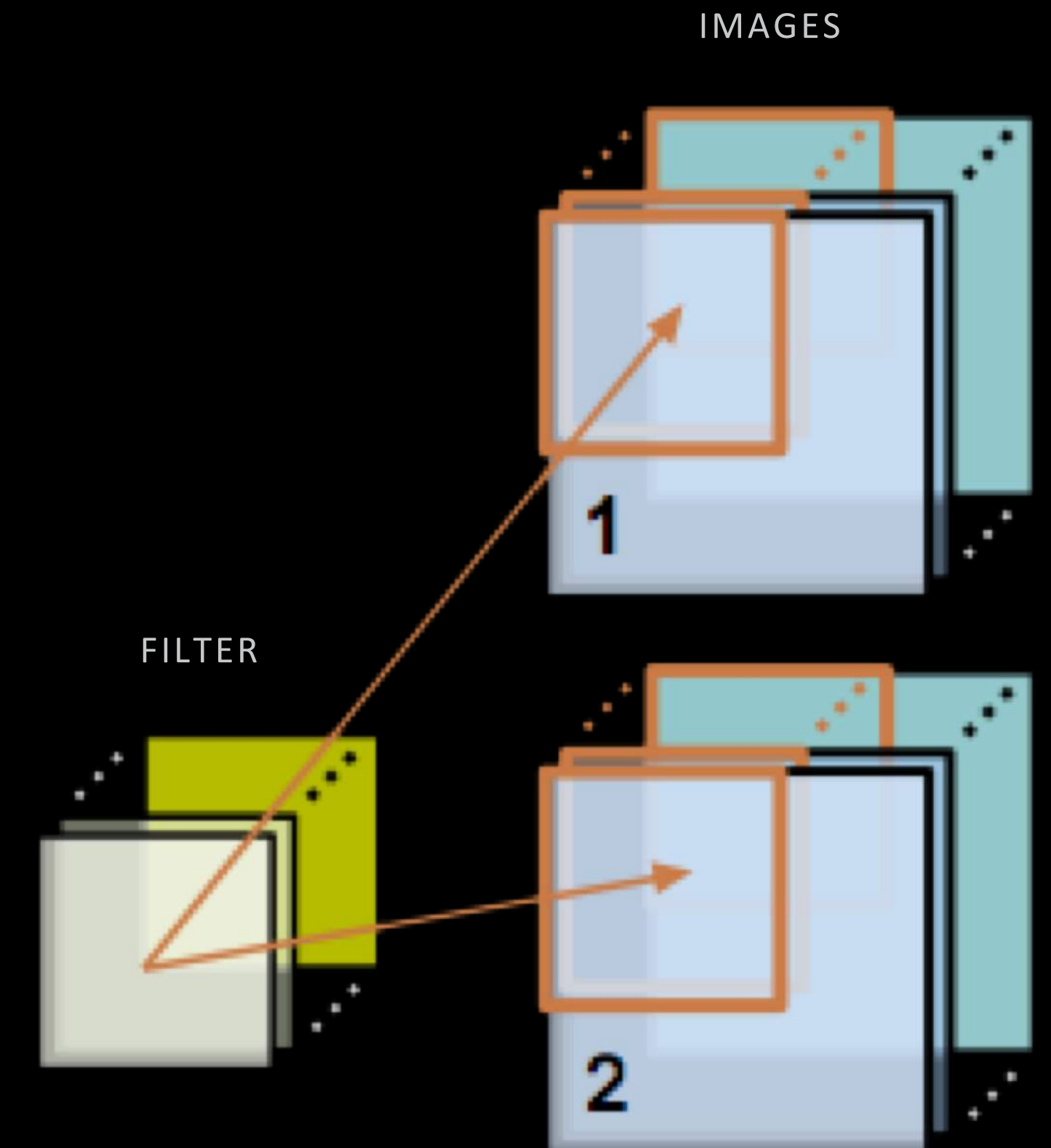
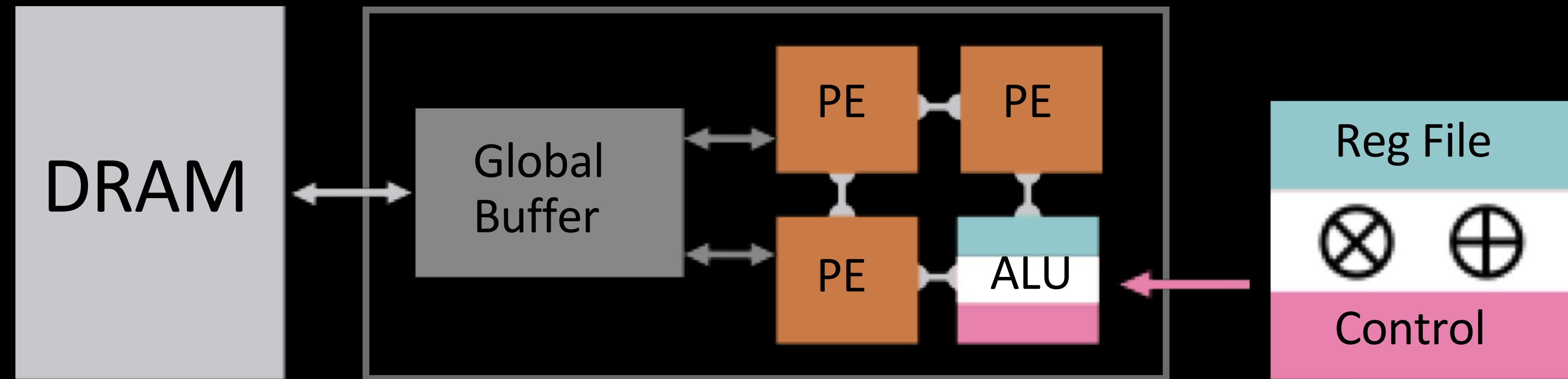


IMAGE REUSE
(PIXELS)



FILTER REUSE
(WEIGHTS)

Exploit data reuse at low-cost memories



Specialized hardware with small (< 1kB) low cost memory near compute



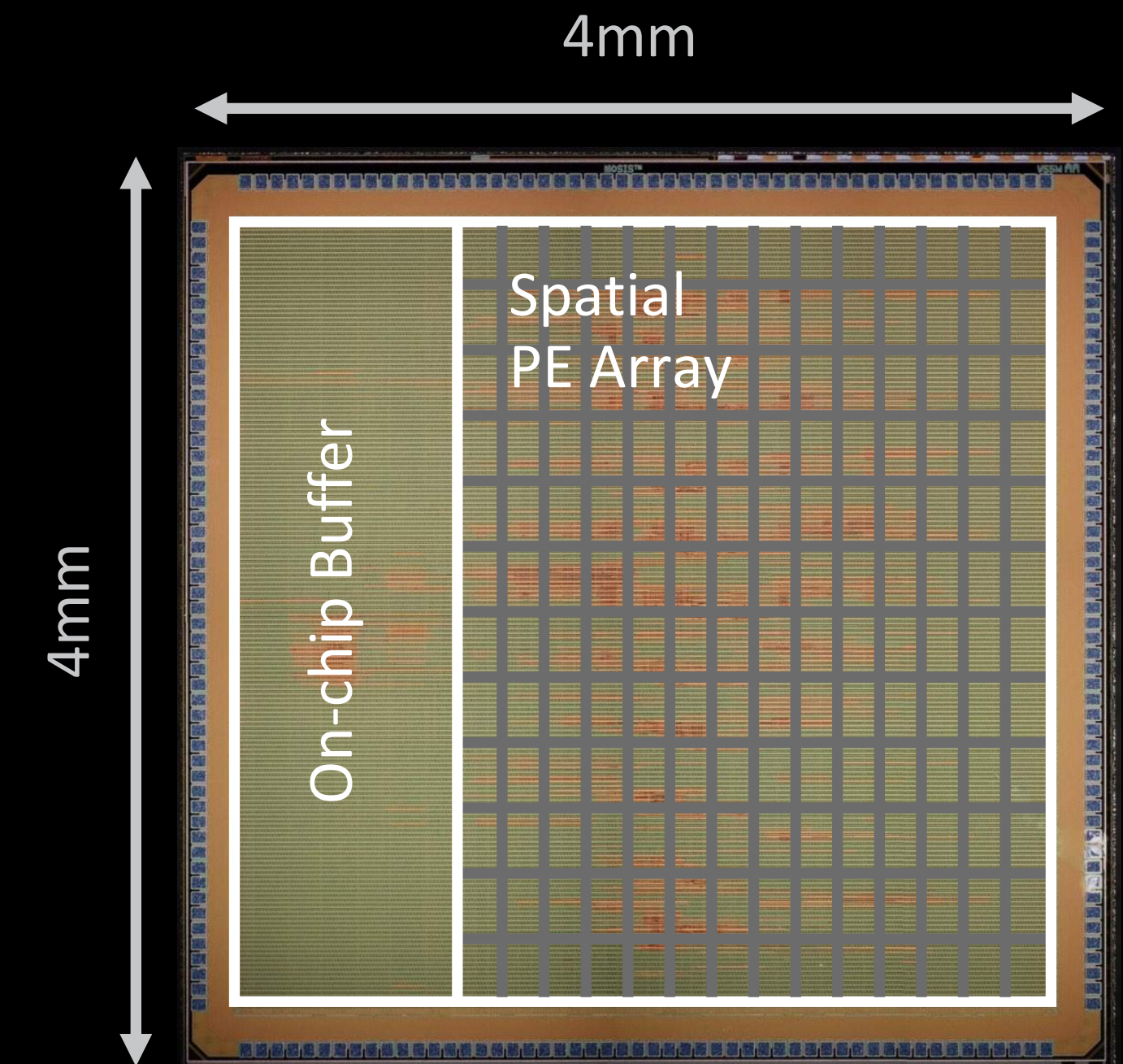
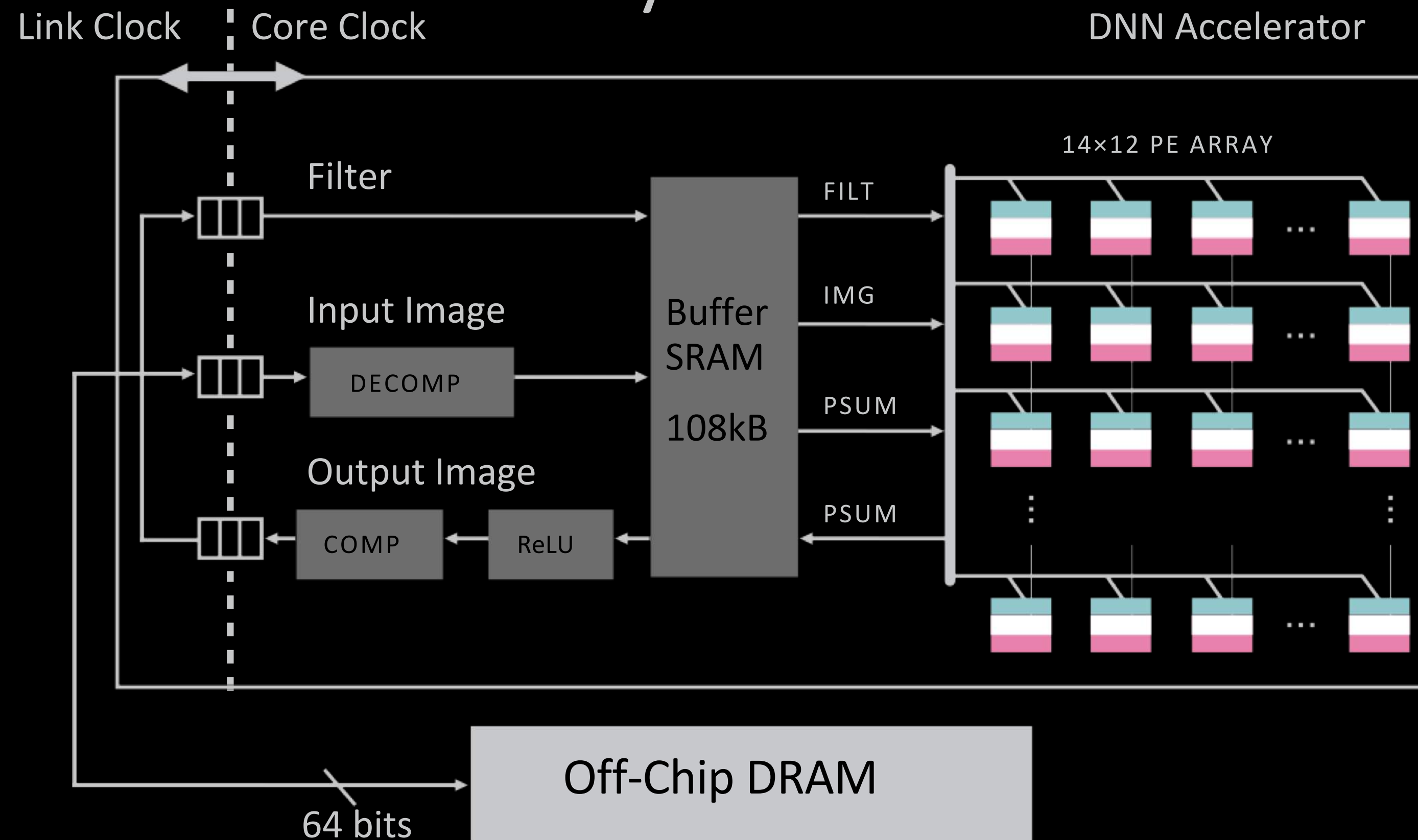
*MEASURED FROM A COMMERCIAL 65nm PROCESS

Farther and larger memories consume more power

Deep neural networks at under 0.3 W

Exploits data reuse for **100×** reduction in memory accesses from global buffer and **1400×** reduction in memory accesses from off-chip DRAM

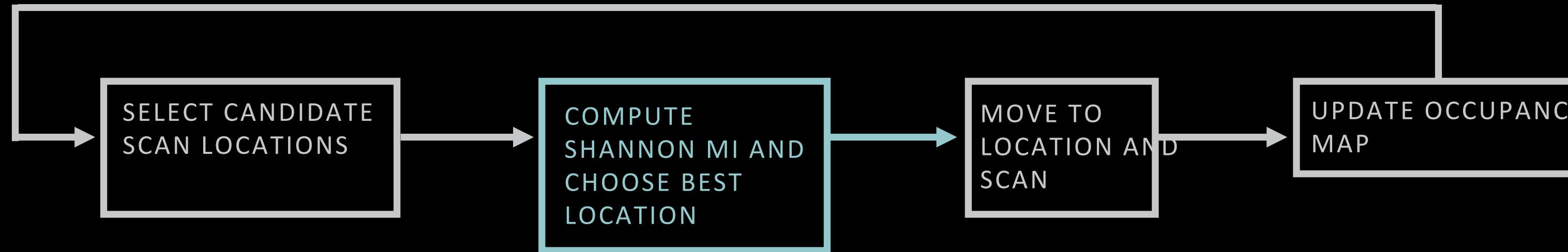
Eyeriss



Overall **> 10×** energy reduction compared to a mobile GPU

Where to go next: planning and mapping

Robot Exploration: decide where to go by computing Shannon Mutual Information



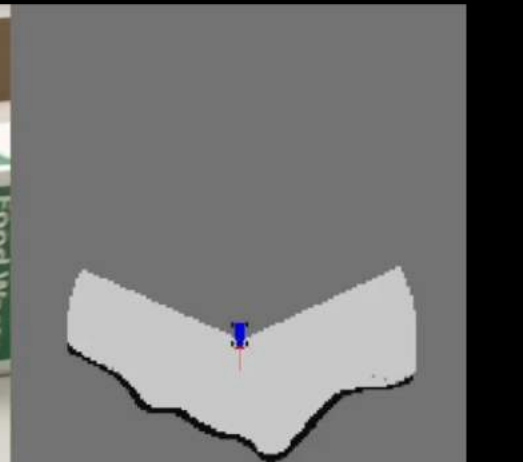
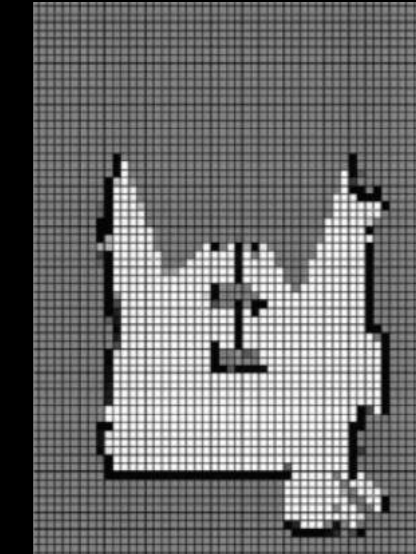
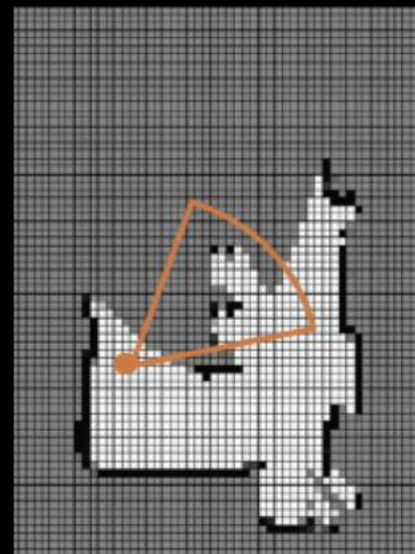
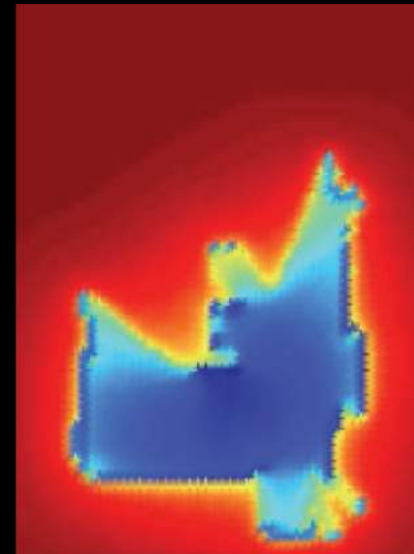
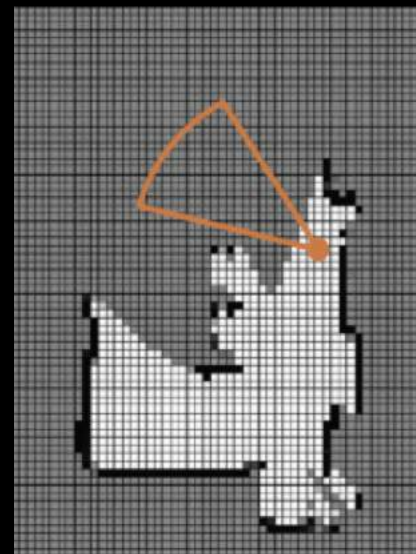
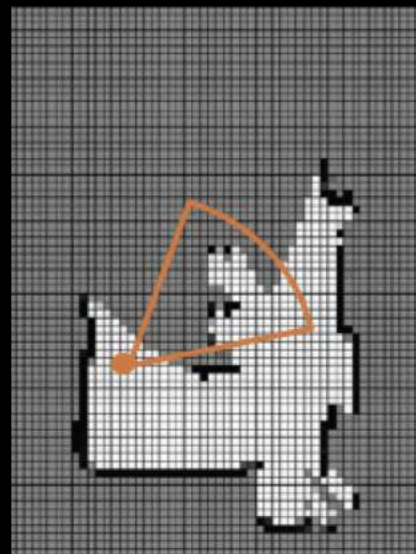
WHERE TO SCAN?

MUTUAL INFORMATION

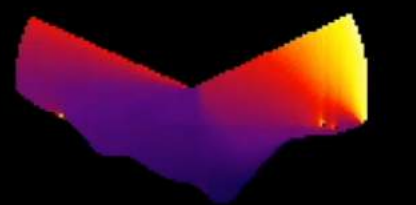
UPDATED MAP

EXPLORATION WITH A MINI RACE CAR USING MOTION CAPTURE FOR LOCALIZATION

OCCUPANCY MAP WITH PLANNED PATH

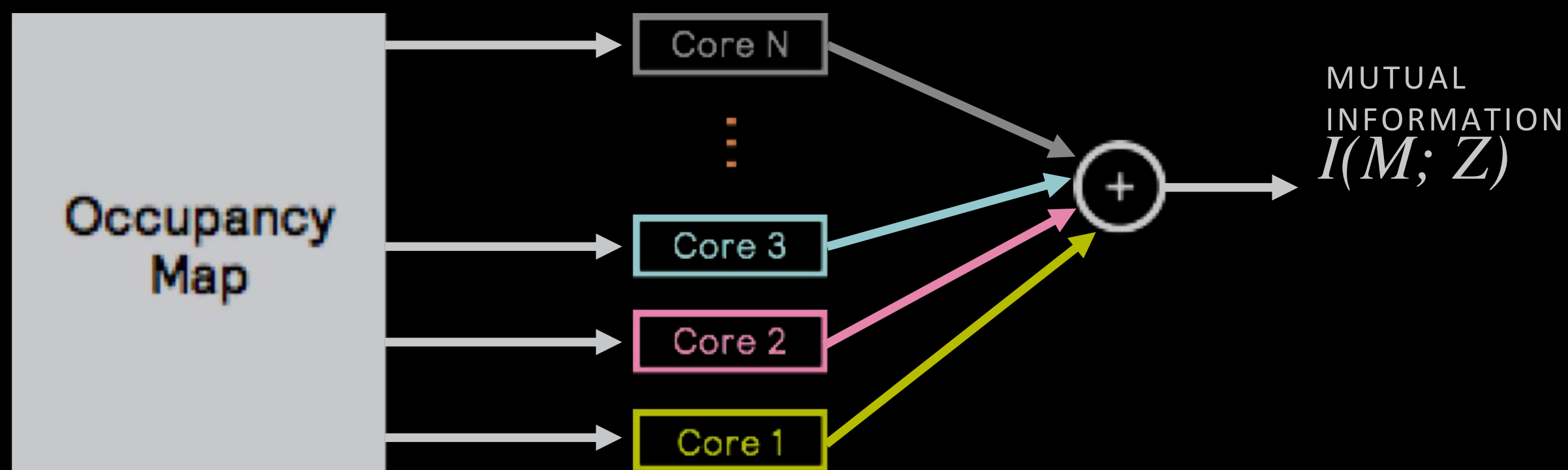
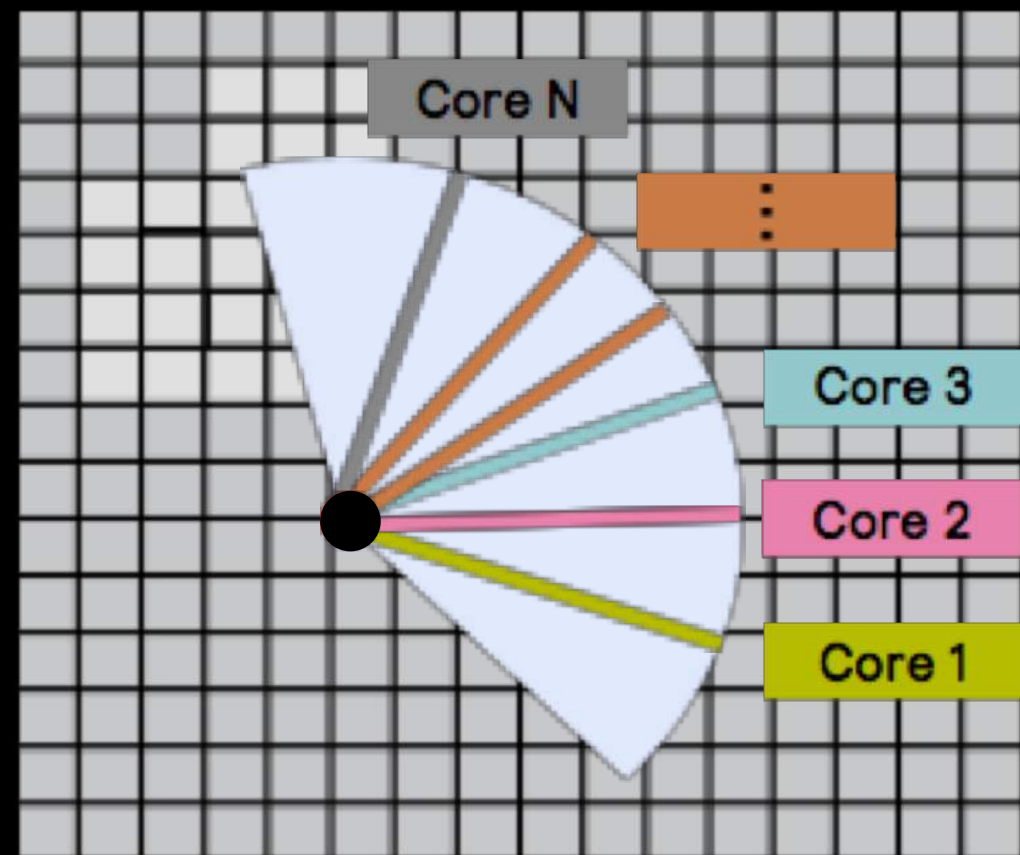


MI SURFACE

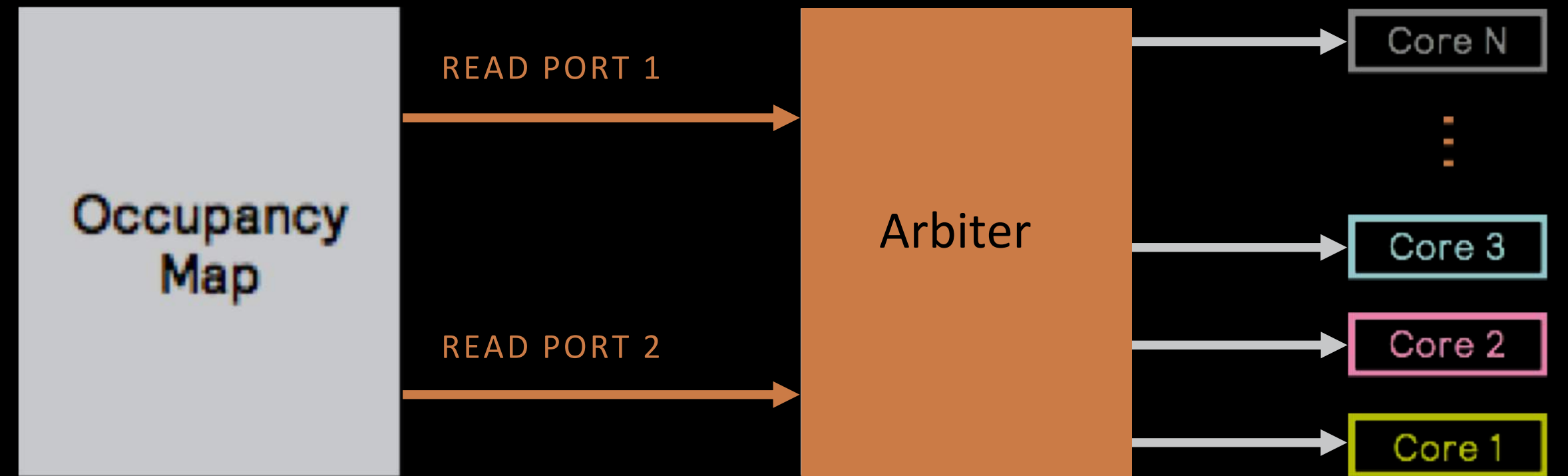


Challenge is data delivery to all cores

Process multiple beams in parallel



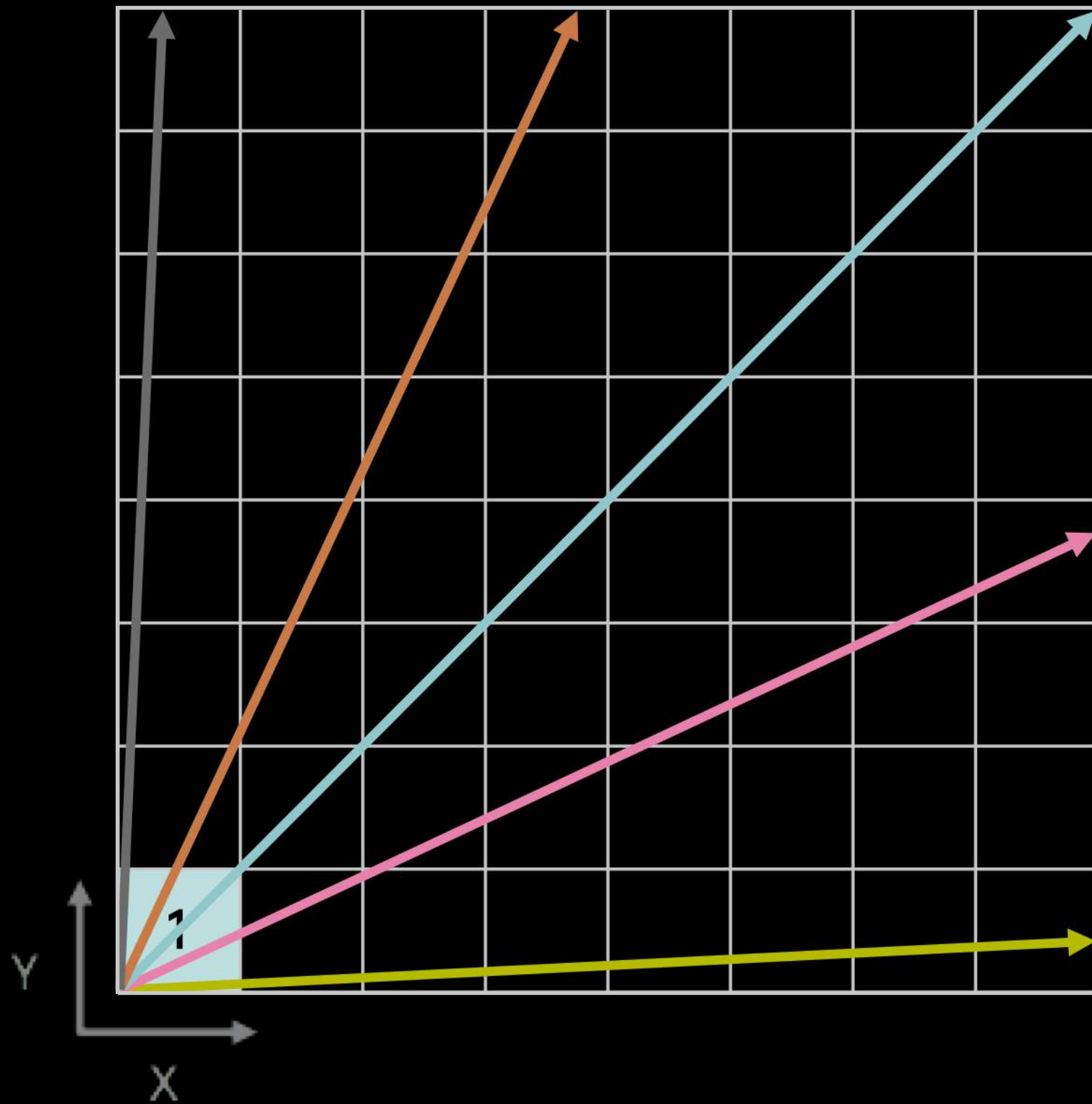
Data delivery from memory is limited



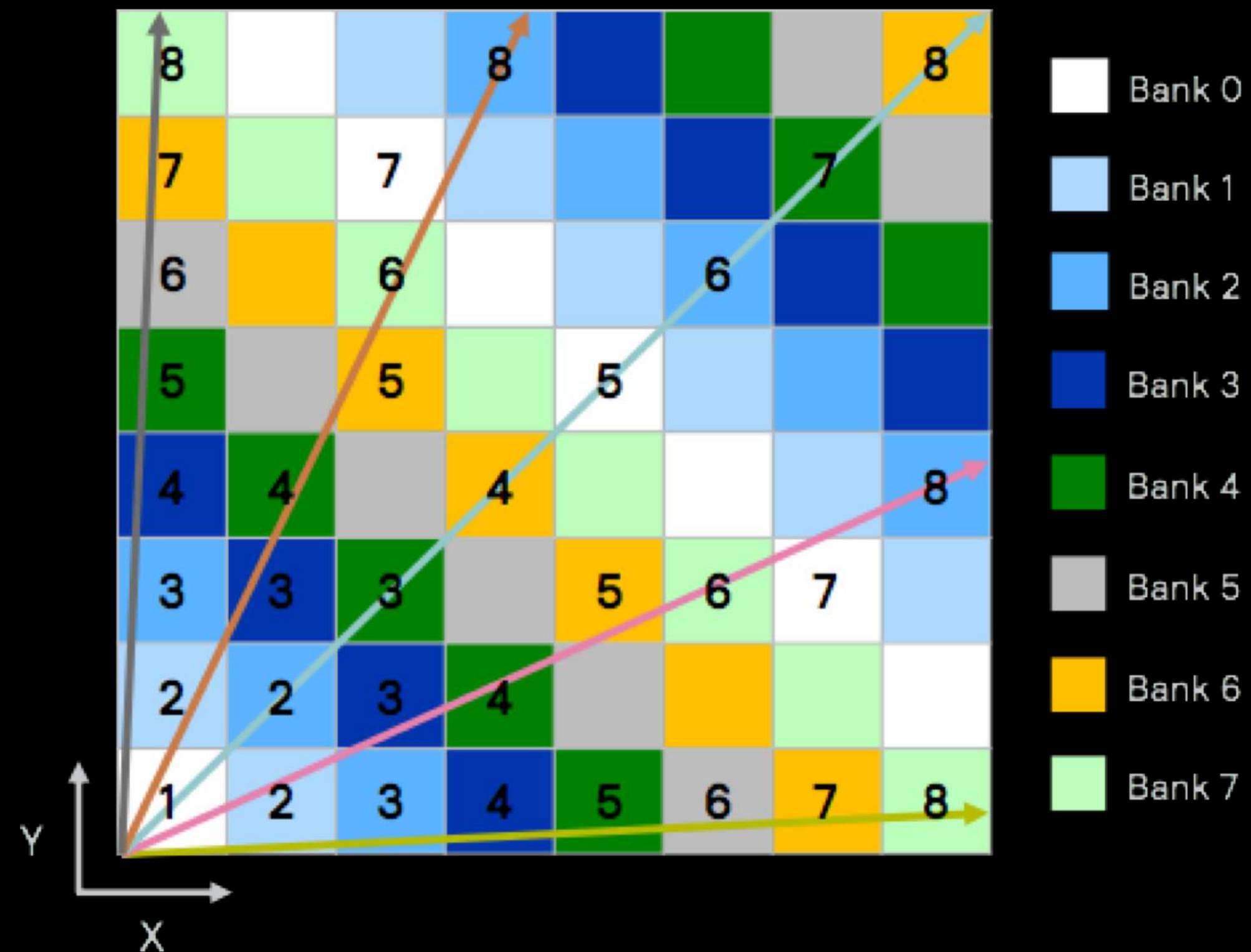
Specialized memory architecture

Break up map into **separate memory banks** and use a novel storage pattern to minimize read conflicts when processing different beams in parallel

MEMORY ACCESS PATTERN



DIAGONAL BANKING PATTERN



Compute the mutual information for an **entire map** of 20m × 20m at 0.1m resolution **in under a second** → a 100× speed up versus CPU at 1/10th of the power

Summary

Efficient computing is critical for advancing the progress of autonomous robots, particularly at the smaller scales → **Critical step to making autonomy ubiquitous!**

In order to meet computing demands in terms of power and speed, need to redesign computing hardware from the ground up → **Focus on data movement!**

Specialized hardware opens up new opportunities for the co-design of algorithms and hardware → **Innovation opportunities for the future of robotics!**

