Advanced Computing for Automated and Connected Vehicles

for

DOE Vehicle Technologies Office
May 2019

DRAPER

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

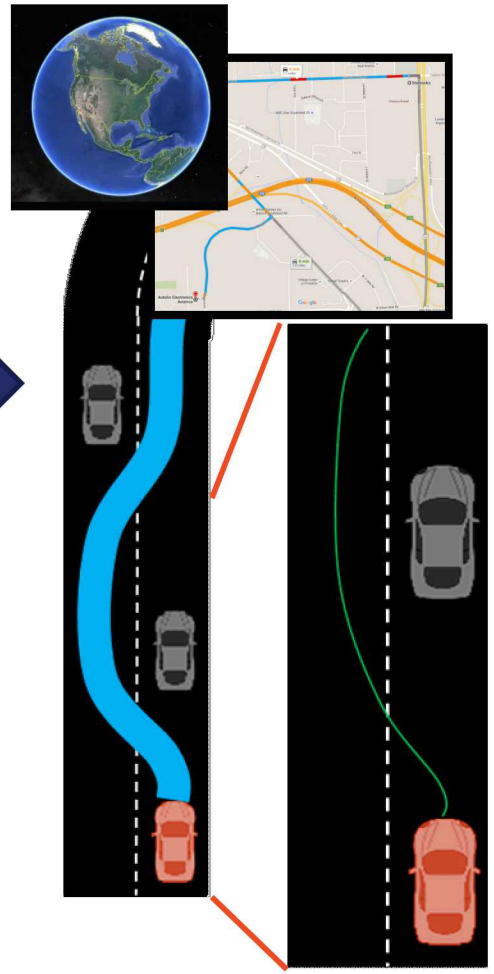
Outline

- Brief history of self-driving car development
- Computing requirements for automated driving
- Computing technology
- Taking action
- Backups
  - example of device-level microelectronic innovation
  - safety considerations

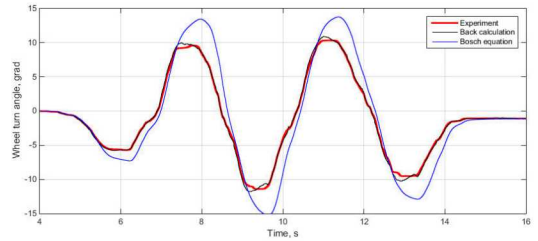
Self-Driving Cars

(Very Roughly) Approximating Human Driving

Perceive
Roads, cars, pedestrians, signals



Think
Plan routes, lanes of travel, precision maneuvers around traffic



Act
Control the steering, throttle and brake to follow plans

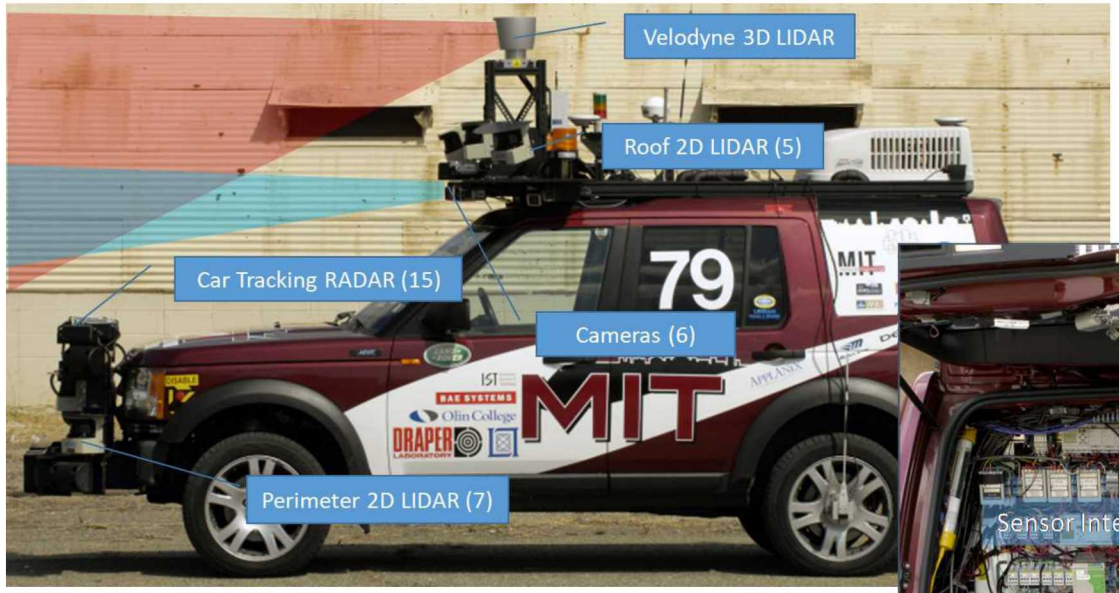
DRAPER

DARPA Self-Driving Grand Challenges

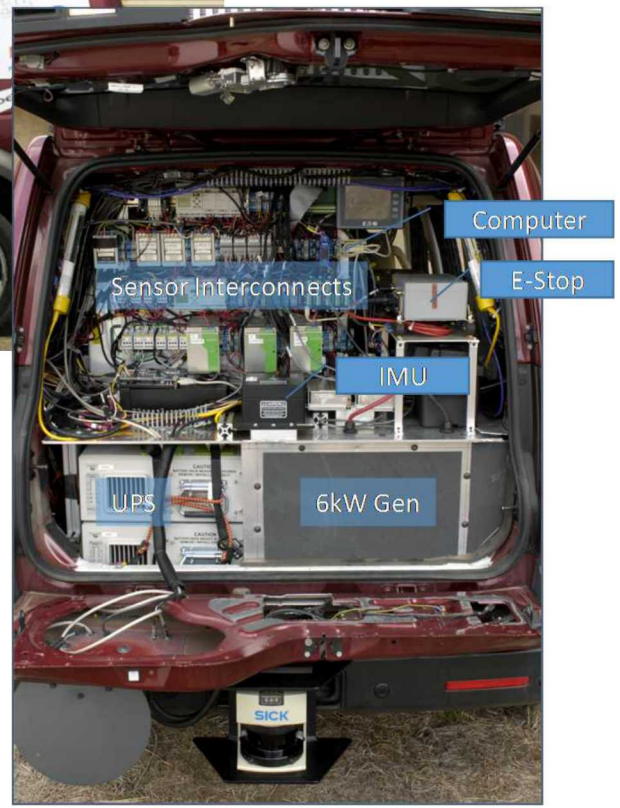
- DARPA Pushed Self-Driving Technology for Military Operations
- 2004 & 2005 Off-Road Races
  - 130 miles through the desert
  - 2004: no finishers
  - 2005: Stanford, CMU
- 2007 Urban Challenge
  - 60 mile mission < 6 hours
  - Drive on city streets, obey traffic rules
  - Robot & human traffic interaction
  - CMU, Stanford, VT, MIT were top 4 finishers (6 total)

DRAPER

Team MIT DARPA Urban Challenge Racer - Talos



13 LIDAR, 15 RADAR, 6 Cameras, and a 500lb Computer



DRAPER

Urban Challenge Was Possible Because DARPA *Ignored* …

Brake, Turning, and Traffic Signals

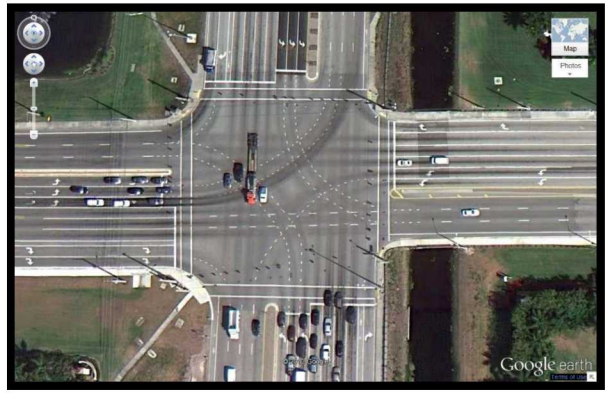
- All intersections 4-way stops and mapped
- No visual sign reading required
- Very low speeds, awkward behavior OK

Pedestrians and Complex Traffic

- No pedestrians on course
- Controlled, low density traffic

Roadway, Lane Detection & Weather

- Precision map of road & lane locations provided
- Southern CA weather



Long Range Car Tracking (km Scale)

- Low speed short range (30 m) intersections
- Controlled traffic

DRAPER

2008 | 2012 | 2014 | 2016 | 2018

CMU

Stanford

Virginia Tech

MIT/Draper

DARPA Urban Challenge (2007)

Demonstrated fully autonomous driving under strictly controlled conditions

Google Self-Driving Car R&D

Leaders from the winning DARPA teams are hired into Google, begin SDC program with lessons learned

Lexus

Delphi

Ford

Google

Moves into fleet testing & custom sensor and vehicle design

Uber

Begins small fleet operations in Pittsburgh

Tesla

First generation Autopilot Hardware in Model S and X – limited highway driving assistant, lacks strict controls on use

Waymo

Testing ~ 600 vehicle fleet in 6 States, Limited L5 autonomy taxi service starts

Second generation Autopilot Hardware in Model S,3,X – most highway driving, limited local roads, slight limits on use

Fatal accident in Arizona and concerns about performance stall efforts

OEM's & Suppliers

Start to consider automated driving as viable for future vehicles, begin internal R&D. Public road testing by Delphi.

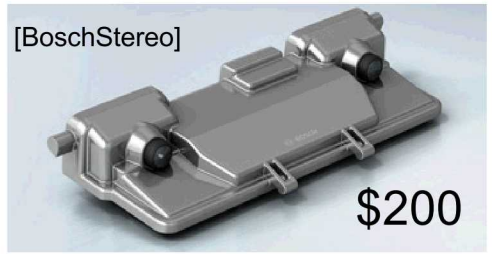
GM, Nissan, Volvo Deploy Automated Highway Driving

All limit usage and monitor driver to avoid over-trust by consumers

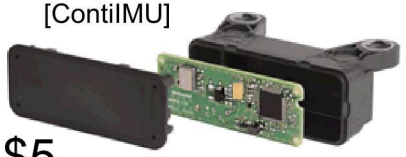
DRAPER

Costs of Sensing and Computing

Assisting a Human Driver on the Highway



Stereo / Mono HD EO, Multispectral Cameras w/ lane & object tracking/classification

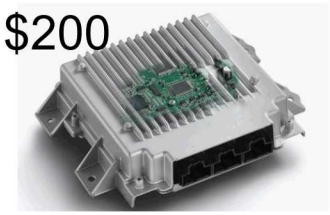


$5
IMU/GPS
(5m accurate)

3D Radar with target tracking

•$800

Air-cooled low-power (5W) computer w/ Giga-Flop Processing



Autopilot uses customized automotive cameras and radars with mid-range GPU computing

• $2000?

Self-Driving, Replacing a Human

[VeloLidar] $5,000 x 4

LIDAR (Short Range)

[IbeoScala] $30,000

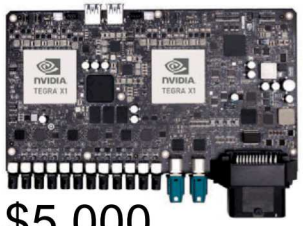
LIDAR (Long Range)

[SBGGPSINS] $40,000
IMU/GPS (10cm accurate)

[FLIREOCam]
HD+ EO/IR Cameras,

•$97,000

Liquid-cooled high power (200W) Terra-Flop Processing



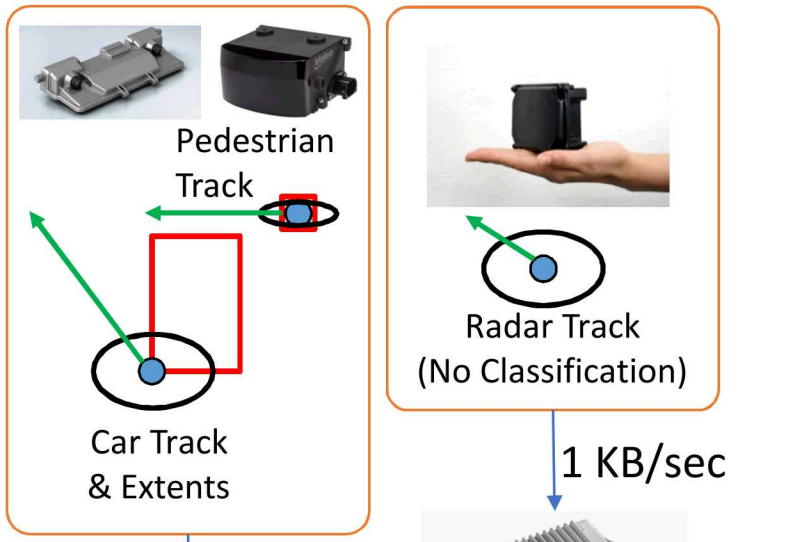
$5,000

GM NISSAN BMW VOLVO

DRAPER

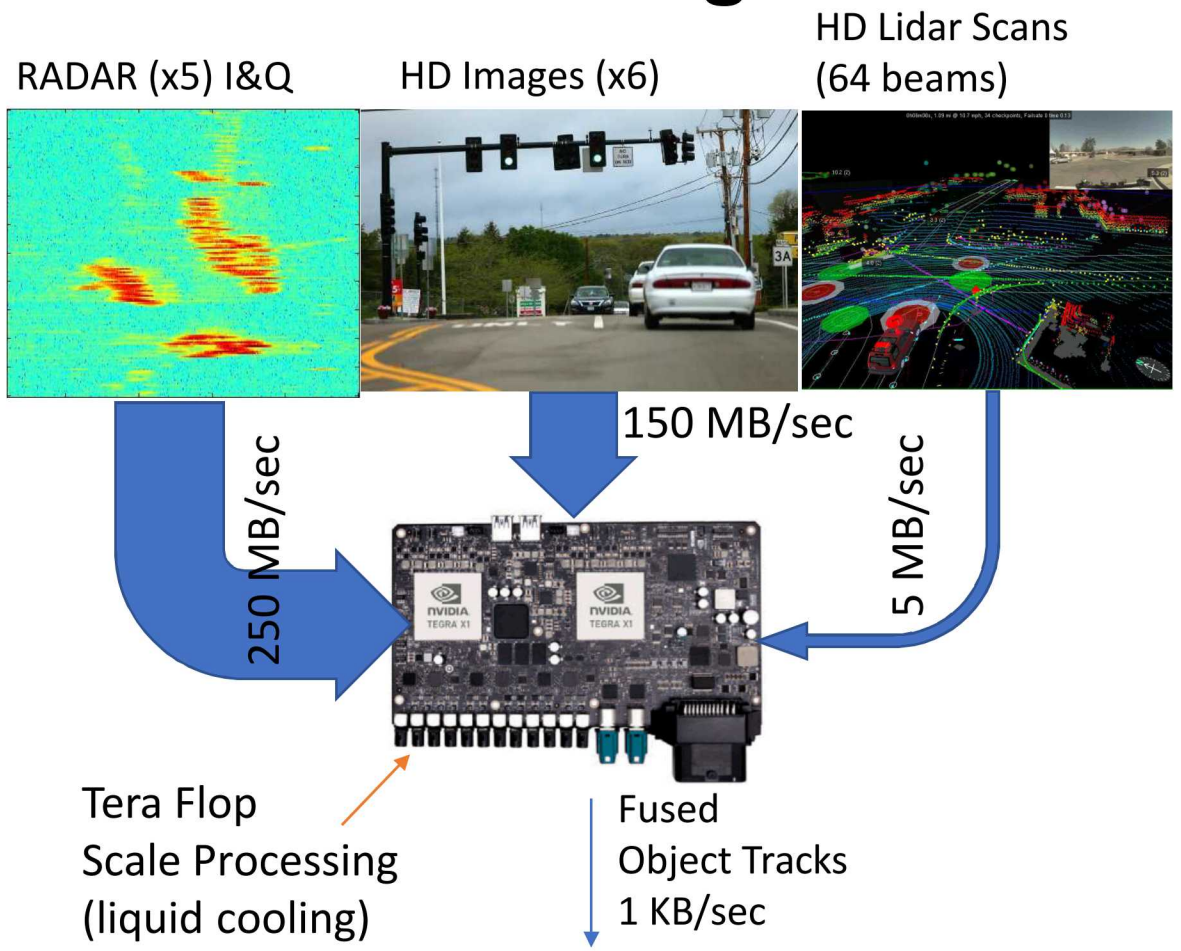
WAYMO Uber ARGO AI cruise

Computing for Driver Assistance vs. Self-Driving Research



Little to no influence on the classification and tracking algorithms in each ADAS sensor

Very low input data bandwidth and processing requirements for fusion



Very high input data bandwidth and processing performance requirements

Optimized fusion across all sensor inputs within one core set of algorithms

DRAPER

Many (familiar) Challenges Remain

Brake, Turning, and Traffic Signals

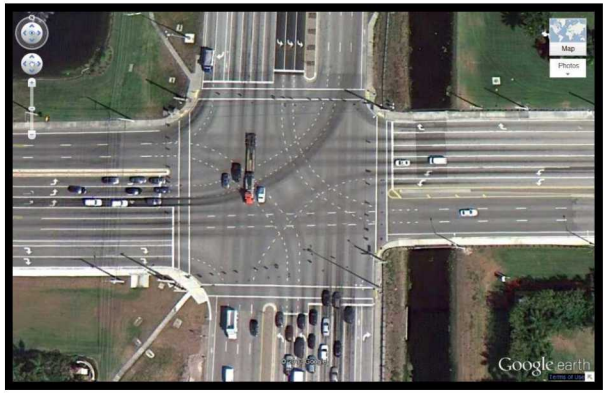
- Recognizing car signals limited to nearest neighbors
- Traffic lights must be 3D mapped

Pedestrians and Complex Traffic

- Recognizing pedestrians is 80% (in good visibility)
- Prediction of intent for pedestrians and cars is still rudimentary

Roadway, Lane Detection & Weather

- Most (including Waymo) still rely on HD maps of all roads/lanes before driving
- Growing, but still very limited exposure to inclement weather



Long Range Car & Pedestrian Tracking (km scale)

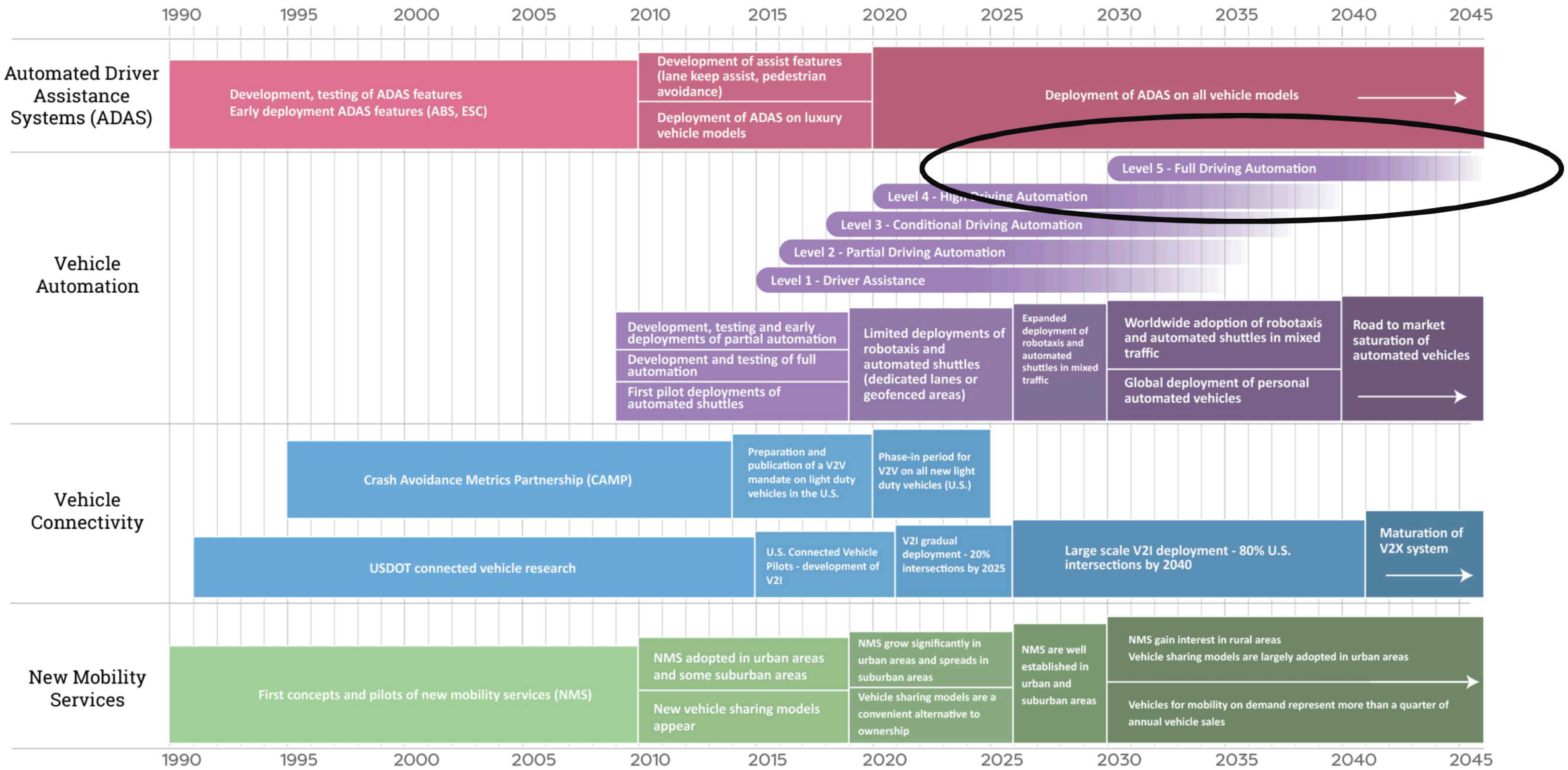
- Reliable car tracking is limited to 200 m ranges
- Pedestrian tracking 100 m in best visibility

Self-driving efforts today are carefully selecting and mapping their environments to maximize reliability

DRAPER

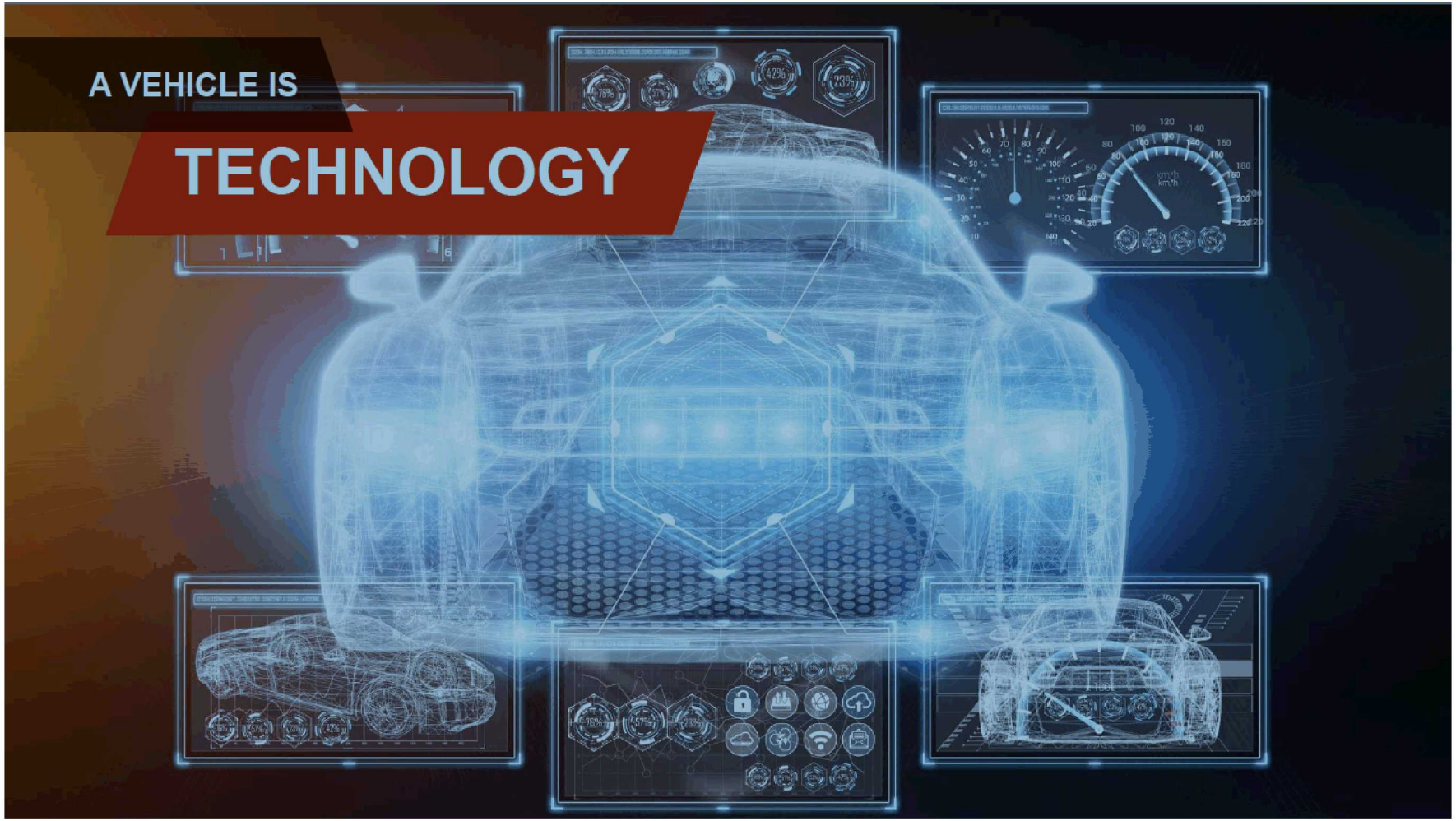
Advanced Computing Requirements for Automated Driving

Ten-year gap predicted between Level 4 and Level 5 consumer deployment



Center for Automotive Research, Ann Arbor, MI

Advanced computing compromises more than just perception and logic

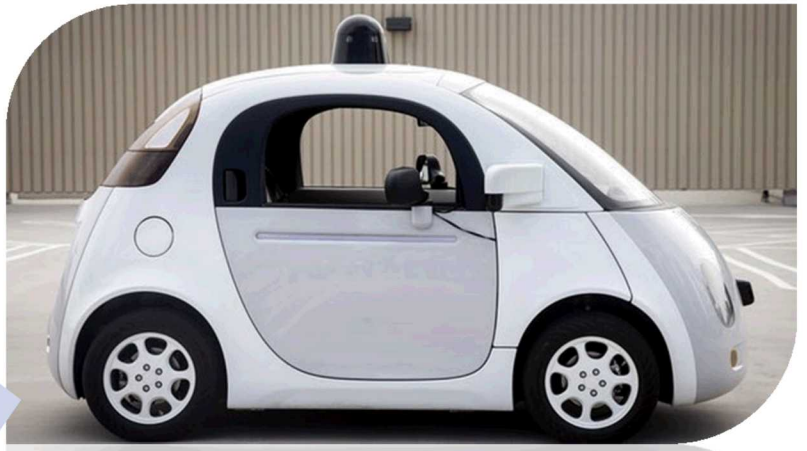


Horsepower	→	Data-power
Performance	→	Connectivity

Projected computing performance and power

HPC must meet CAV size, weight, and power constraints

~1 petaflops
~100 W or less
~10 TOPS/watt



Full level 5 automated driving

>10x less power

>10x compute

>100x power performance

Early prototype self-driving

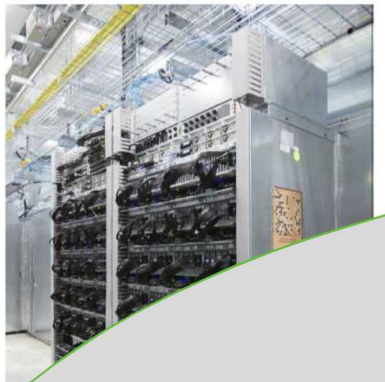
https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

~100 teraflops
~1000 W
~0.1 TOPS/watt

TOPS == Trillion (tera) Operations

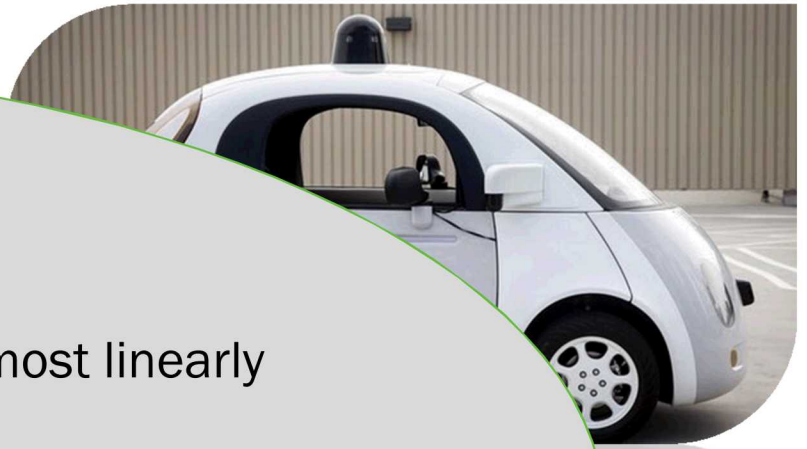
CAV == Connected & Automated Vehicle

Projected computing performance and power



HPC must meet CAV size, weight, and power

~1 petaflops

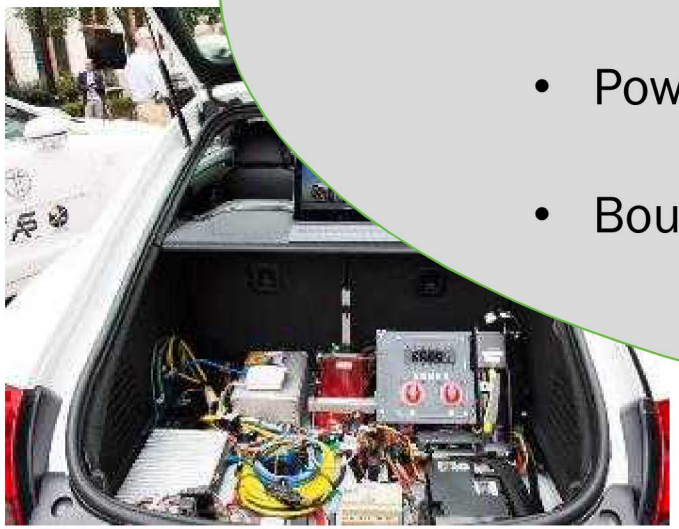


Power requirements

- Assume battery electric storage increases at most linearly
- Reduce energy consumption from ~10% of battery to ~1%
- Power includes communication, cooling, and redundancy
- Bounding target -- the human brain is a 30 W system

5

d driving



Early prototype self-driving

~100 te

~1000 W

~0.1 TOPS/watt

https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

CAV == Connected & Automated Vehicle

Projected computing performance and power

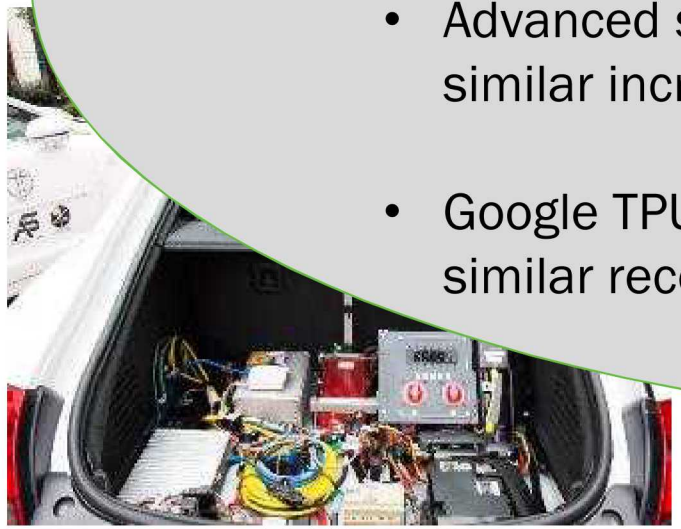
HPC must meet CAV size, weight, and power

~1 petaflops

Computing requirements

- Estimating computing operations per second to grow by at least an order of magnitude in moving from Level 4 to Level 5 driving
- Advanced sensing results in > 10X increase in data, requiring similar increase in computing capability
- Google TPU and Apple A11 are estimated ~ 1 TOPS/W at die level; similar recent claims by Nvidia and Tesla

driving



Early prototype self-driving

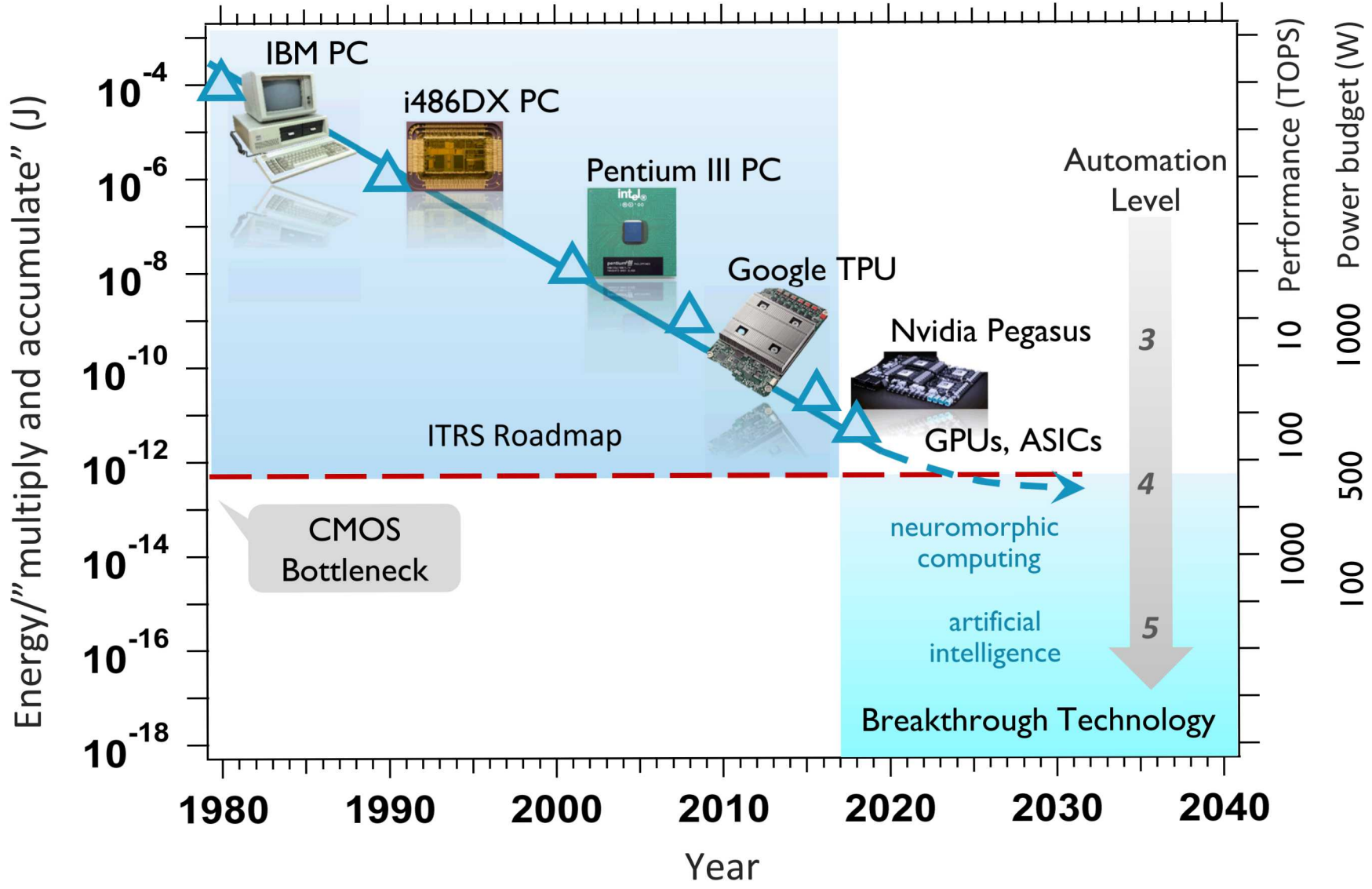
~1000 W

~0.1 TOPS/watt

https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

CAV == Connected & Automated Vehicle

Energy efficiency of computing and the need of CAV's – a challenging problem



Computing Technology

Why now for computing industry? — convergence of four critical trends

Technology:

- End of Dennard scaling: power becomes the key constraint
- Slowdown in Moore's Law evidenced by flattening of transistor cost takedown

Architectural:

- Limitation and inefficiencies in exploiting instructional level parallelism and in the prevailing von-Neumann architecture

Applications:

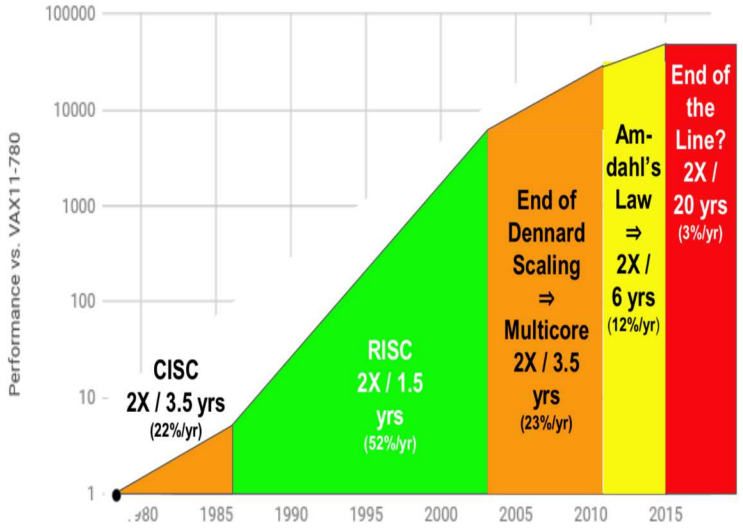
- Shift from desktop of mobile and IoT
- Ultra-scale cloud computing and artificial intelligence/machine learning workloads

Industry collaborations:

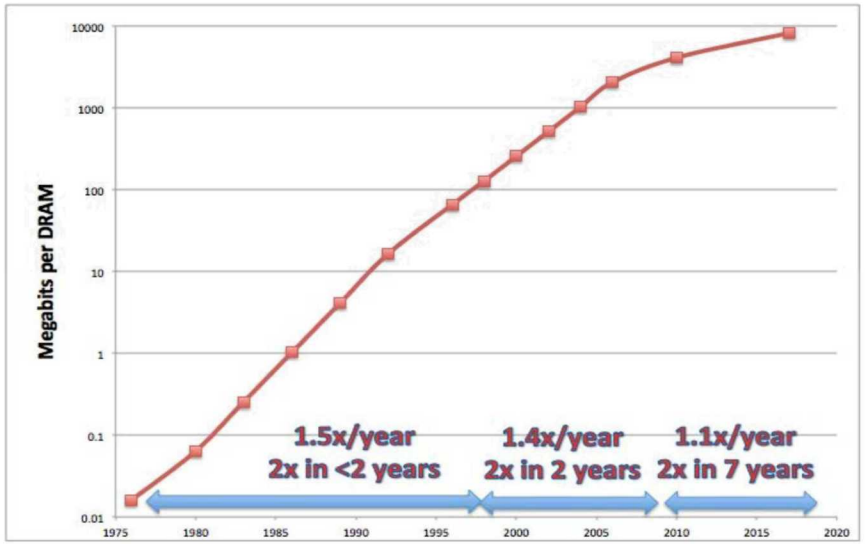
- End of International Technology Roadmap for Semiconductors (ITRS) roadmap
- Decline in SRC participation and the end of SEMATECH (absorbed in SUNY)

Evidence of Moore's Law slowdown – scaling is in a crisis

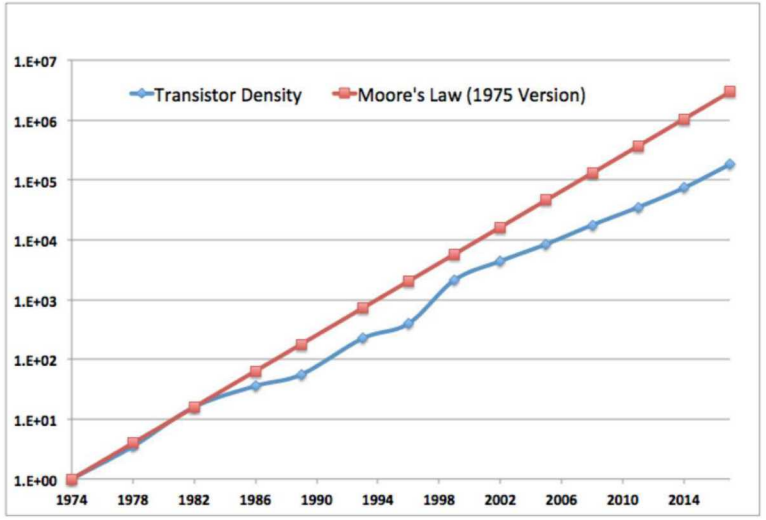
CPU single core performance



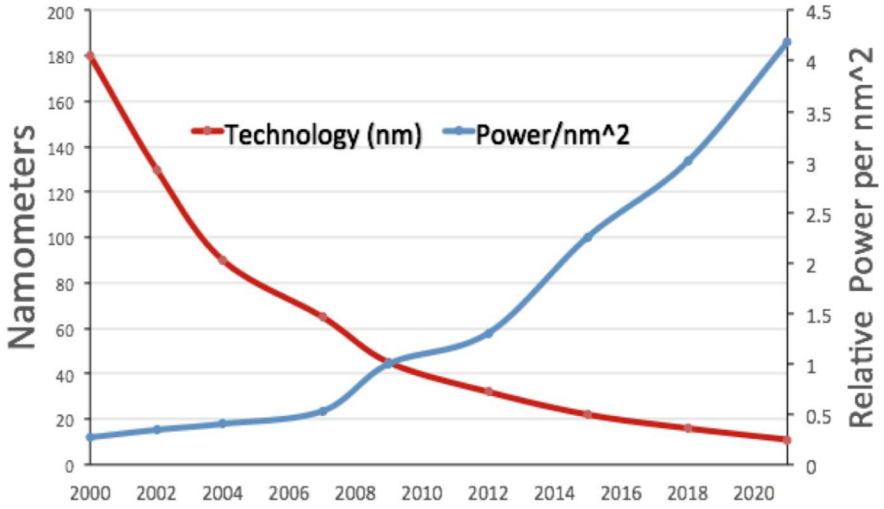
DRAM density roll-off



Intel CPU density roll-off

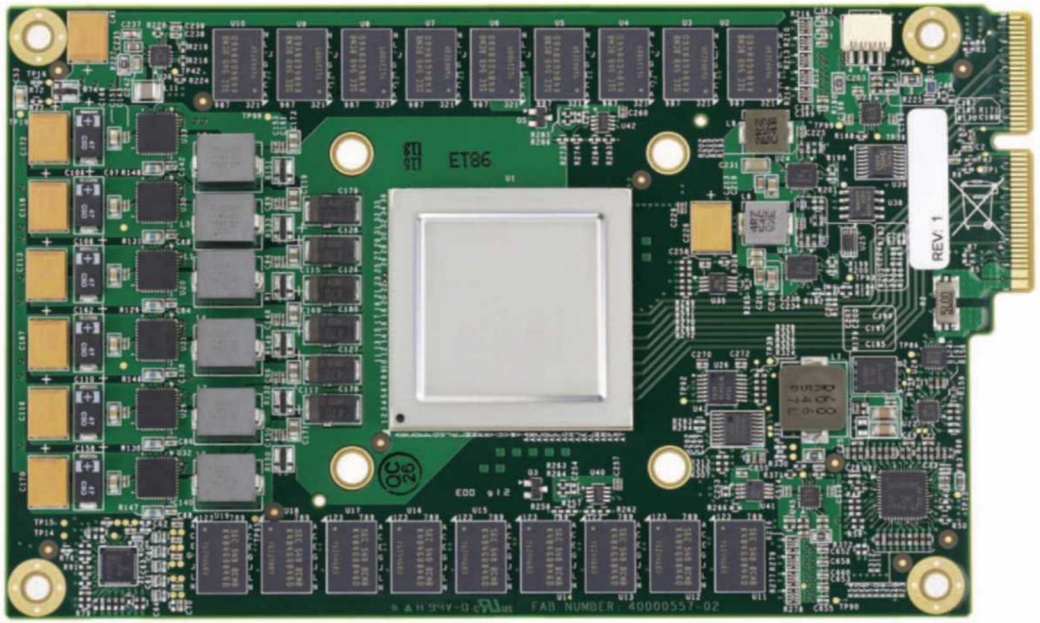


Transistor power takes off

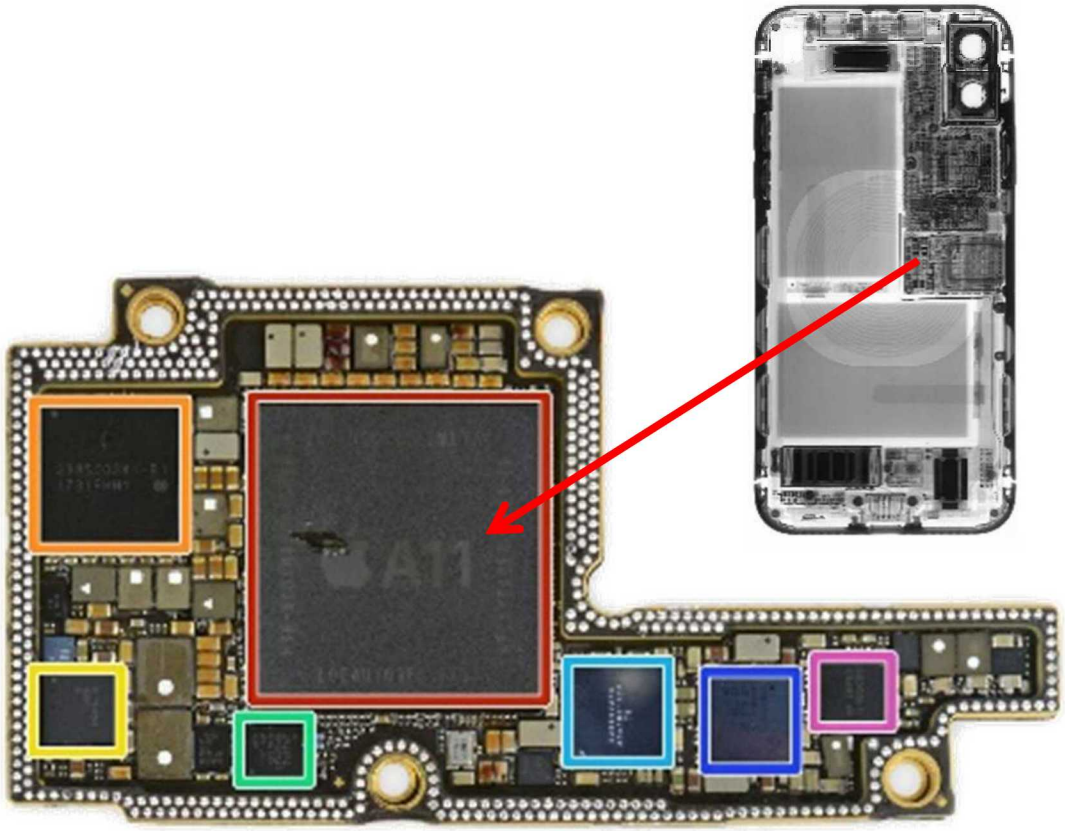


Digital accelerators will get us to ~ 1 TOPS/watt

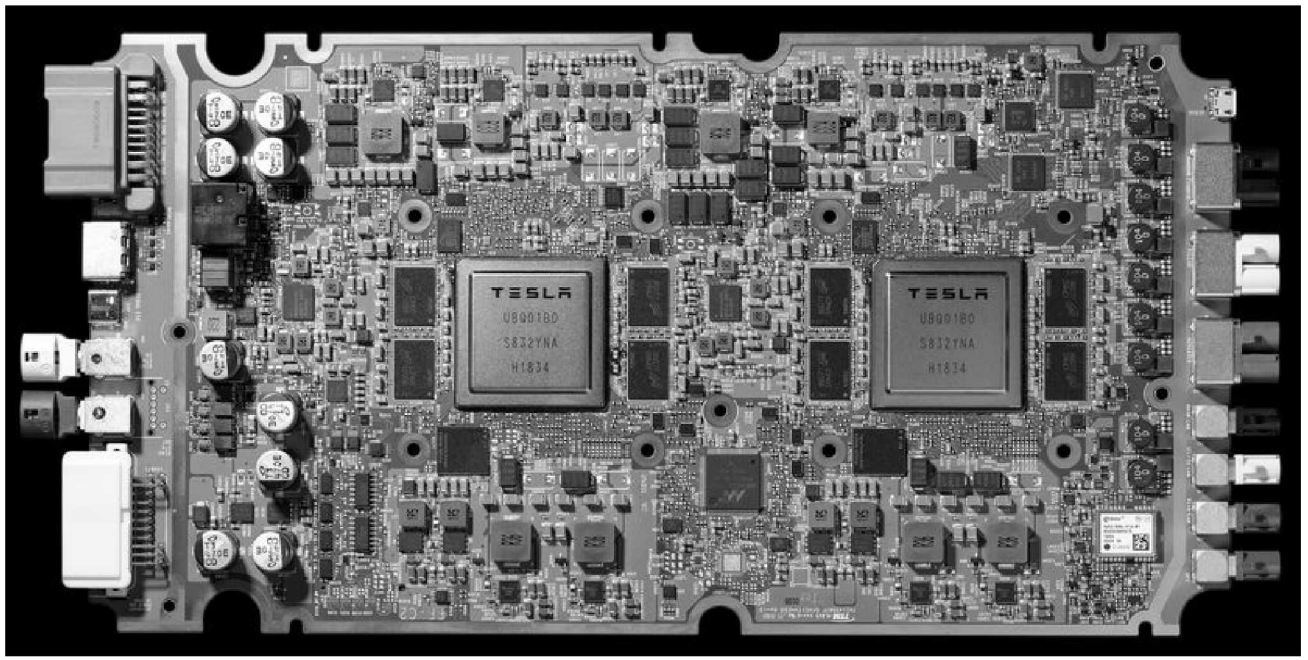
- Tensor Processing Unit (TPU): machine learning accelerator in use by Google since 2015
- Core: matrix multiply unit
- Die level performance of 2.3 TeraOps/W



- Apple A11: iPhone 8 and X main system on a chip (SoC) processor
- Estimated > 1 TeraOp/W, or < 1 pJ/op



In the news...

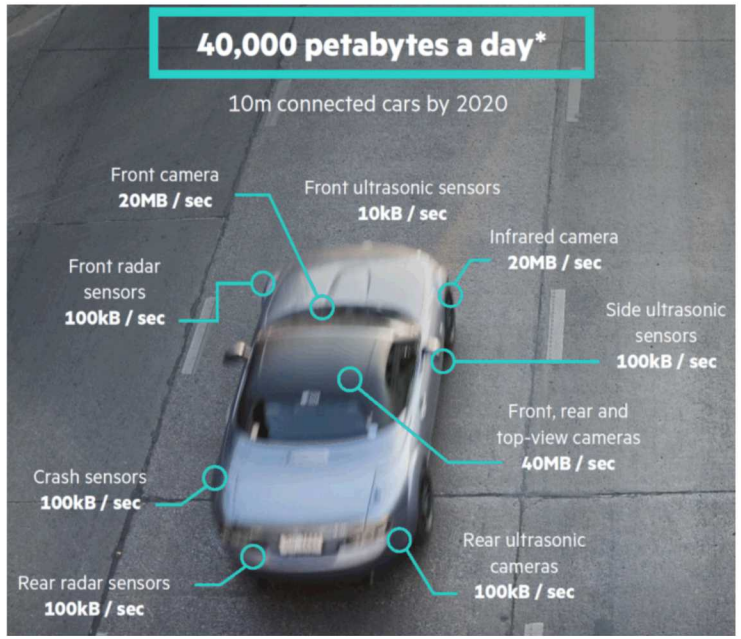


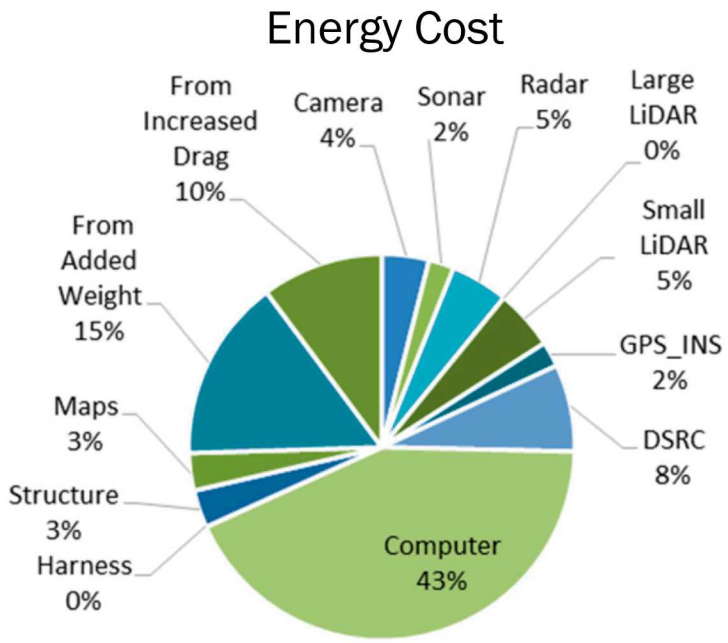
Tesla's new self-driving computer contains the "best chip in the world," according to Elon Musk. Credit: Tesla, April 2019

Nvidia agrees with Tesla's take on self-driving cars, but corrects specifics – Digital Trends, April 2019

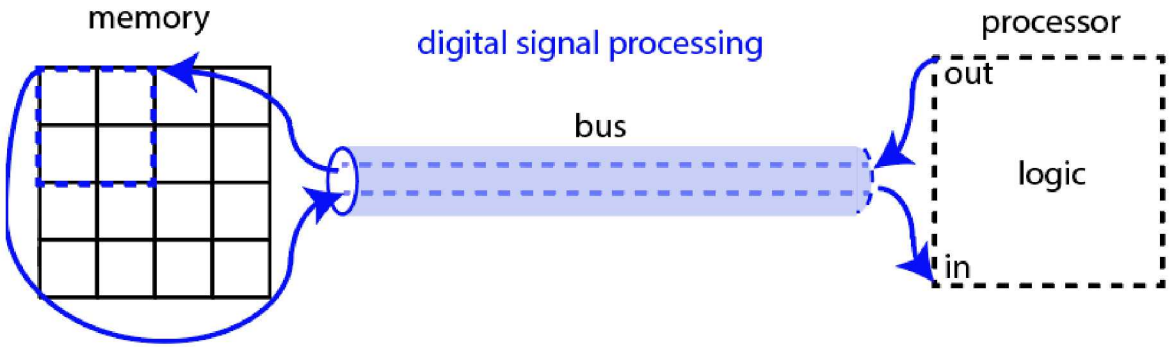
"The Xavier processor features a programmable CPU, GPU, and deep learning accelerators, delivering 30 TOPs. We built a computer called DRIVE AGX Pegasus based on a two-chip solution, pairing Xavier with a powerful GPU to deliver 160 TOPS, and then put two sets of them on the computer, to deliver a total of 320 TOPS"

New approach is needed to meet low-energy computing demand

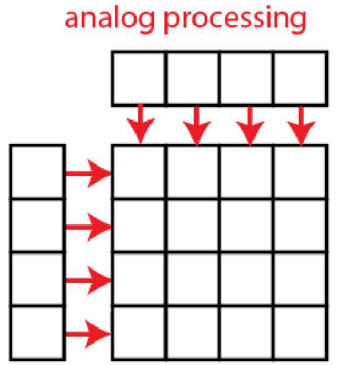




Gawron et. al. *Environ. Sci. Technol.* 52, 3249-3256 (2018)

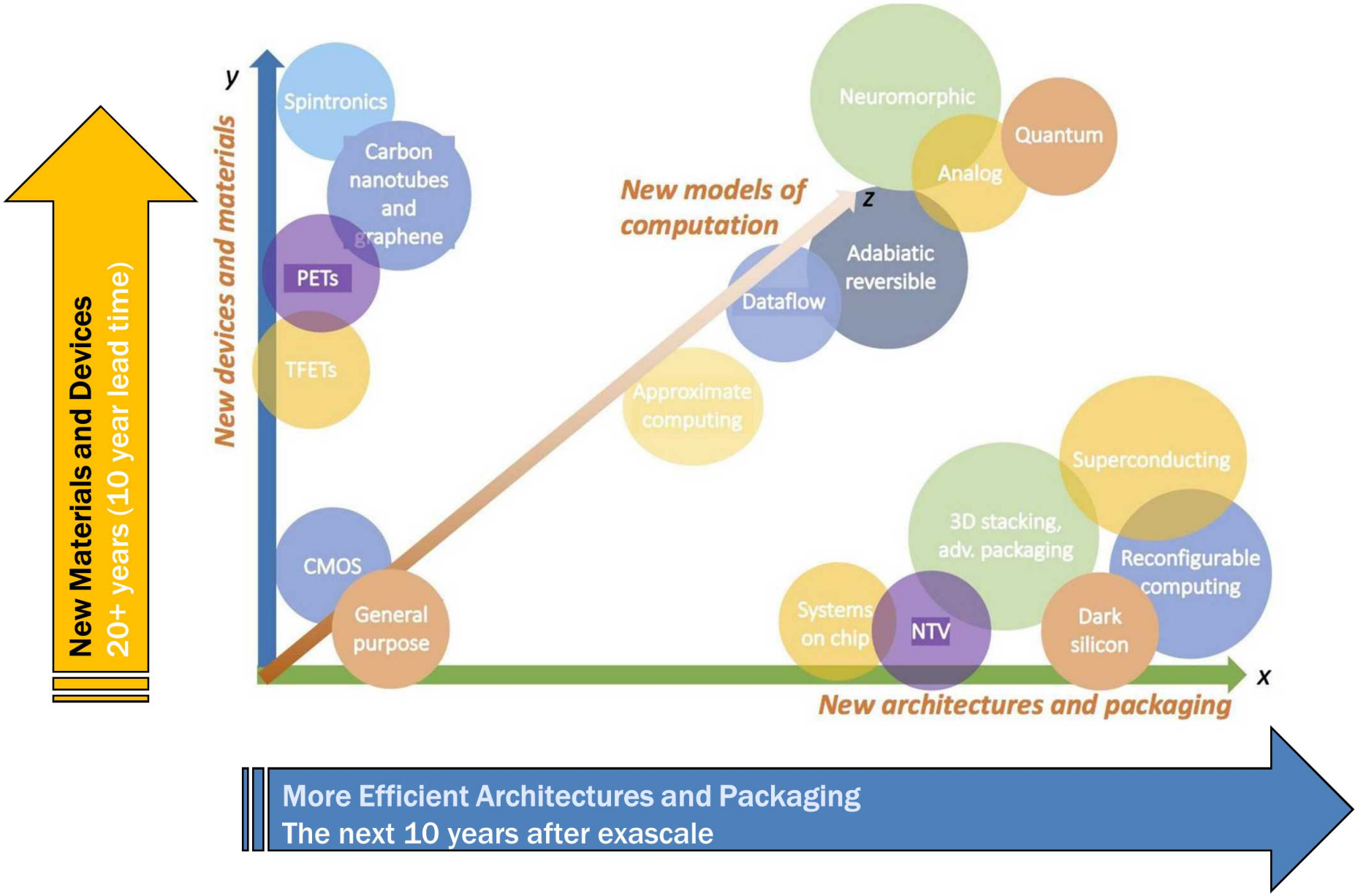


fundamental flaw: processing separate from memory creates efficiency bottleneck



logic and memory combined to circumvent digital bottleneck

Many unproven candidates yet to be investigated at scale



Taking Action

National priorities, July 2018

EXECUTIVE OFFICE OF THE PRESIDENT
WASHINGTON, D.C.

July 31, 2018

M-18-22

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: MICK MULVANEY
DIRECTOR, OFFICE OF MANAGEMENT AND BUDGET

MICHAEL KRATSIOS
DEPUTY ASSISTANT TO THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

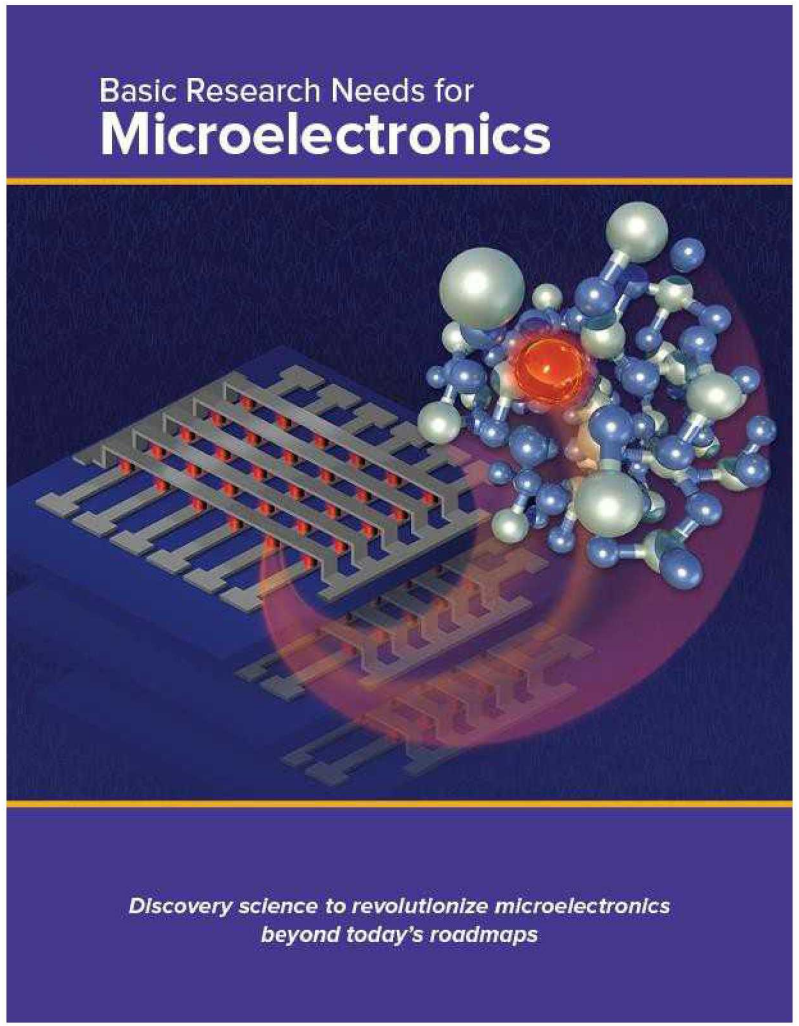
SUBJECT: FY 2020 Administration Research and Development Budget Priorities

"Agencies should prioritize investment in research and infrastructure to maintain U.S. leadership in strategic computing, *from edge devices to high-performance computing, that accelerates delivery of low power, high performance devices*; supports a national high-performance computing ecosystem; and explores novel pathways to advance computing in a post-Moore's Law era".

Semiconductor Industry Association policy recommendations, April 2019

"Semiconductors — the tiny chips that enable modern technologies — are critical to America's economy, job creation, technology leadership, and national security. For 50 years, America has led the world in semiconductor innovation, driving transformative advances in nearly every modern technology, from computers to mobile phones to the Internet itself. Today, semiconductors underpin the most exciting "must-win" technologies of the future, including artificial intelligence to power self-driving cars and other autonomous systems, quantum computing to analyze huge volumes of data and enhance digital encryption, and advanced wireless networks to seamlessly connect people at unprecedented speeds and security.

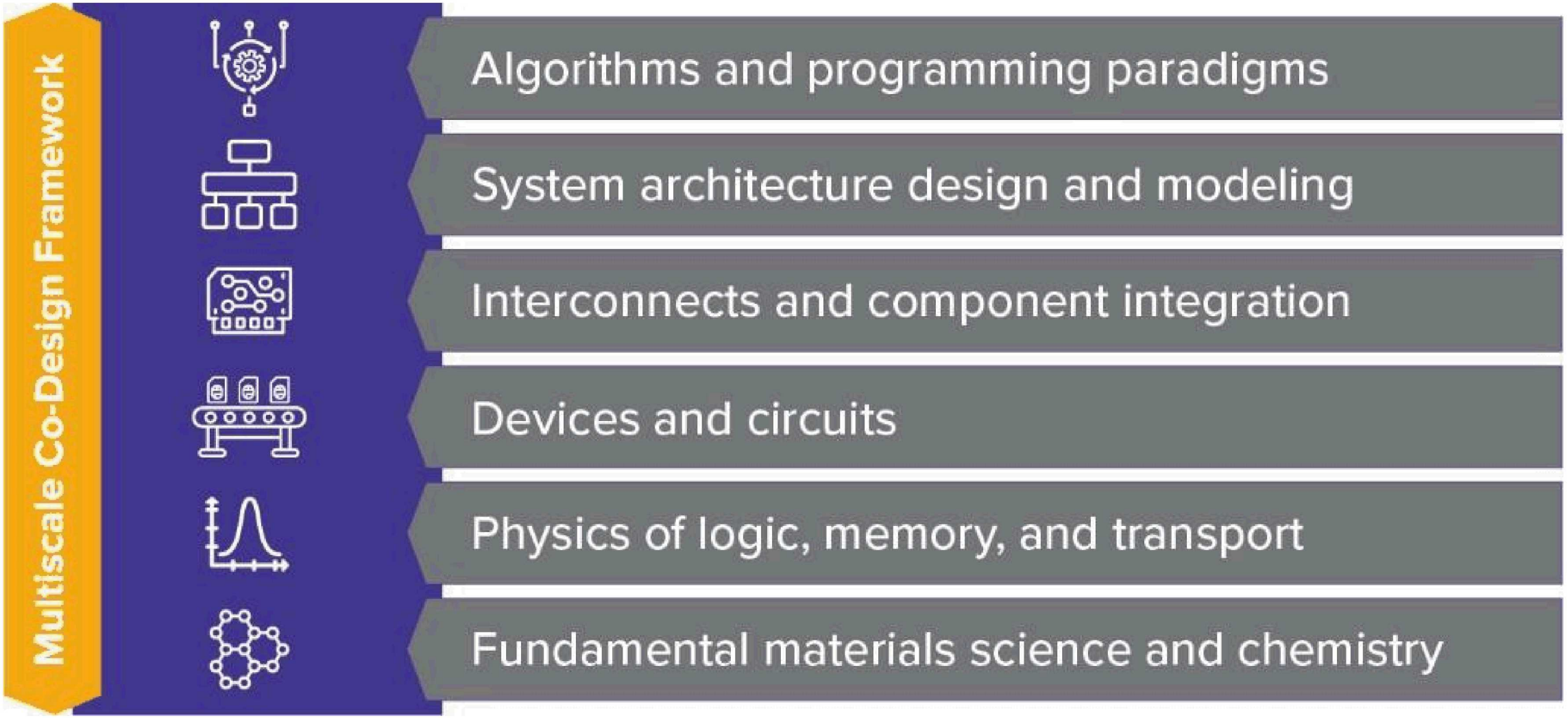
To secure America's leadership in these future technologies for the next 50 years, the United States must continue to lead the world in semiconductor research, design, and manufacturing."



Call To Action

- Significant challenges as CMOS extends below 5nm
- The end to Moore's Law will impact U.S. industry and competitiveness
- The importance of this issue and its technical complication will require *innovative approaches* to keep the U.S. in a leadership position
- Solving a problem of this scale will require *"whole of government"* approach and a robust *public/private partnership* to apply the best research from *industry, academia and government research facilities* to allow the U.S. to successfully make this technology transition
- **DOE, and particularly the Office of Science, will play a significant role in this effort**
- **DOE-SC was charged with organizing a *Basic Research Needs Workshop* to define the highest priority research directions**

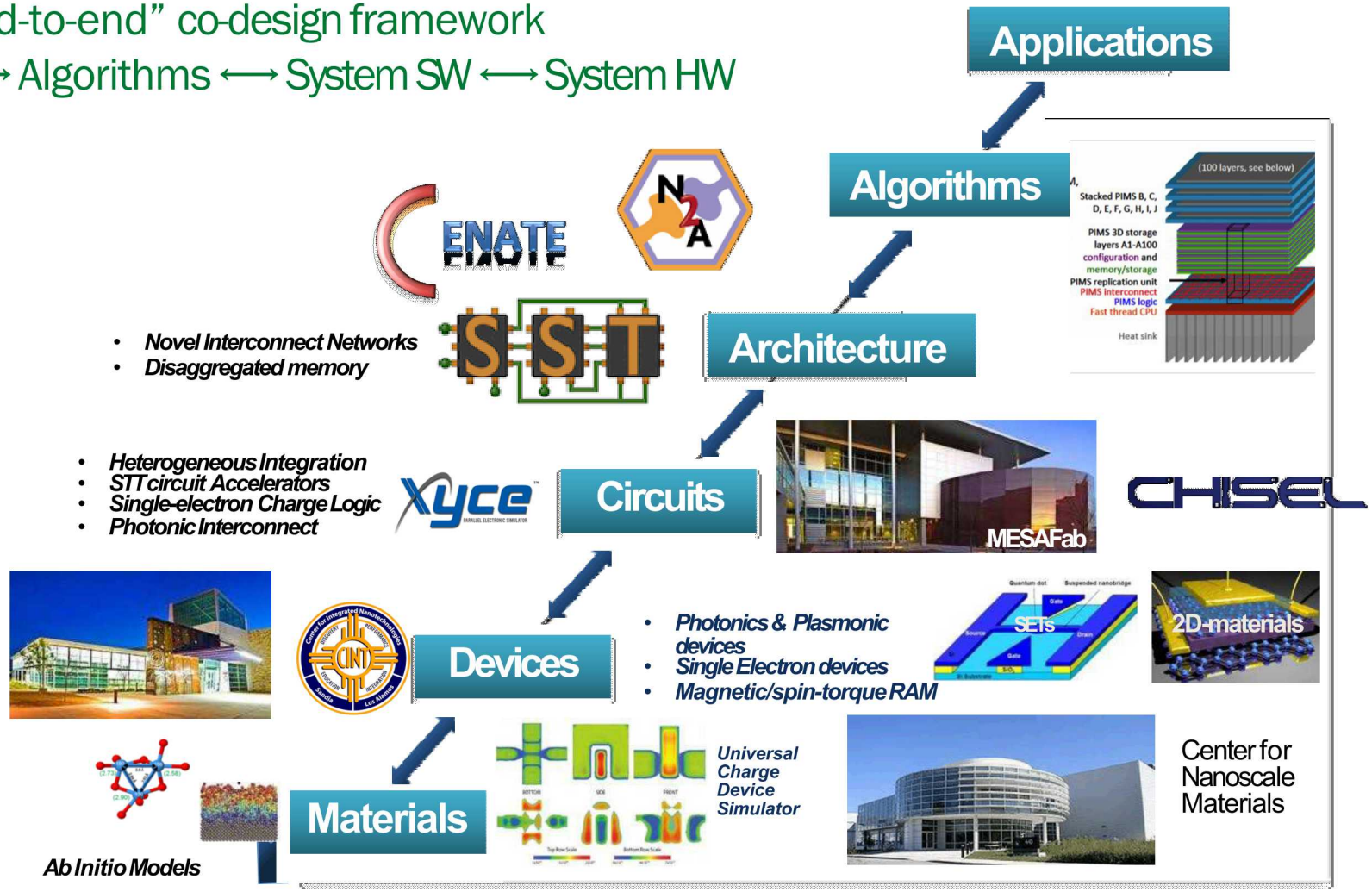
Principles of co-design underpin five priority research directions



BES/ASCR/HEP Basic Research Needs for Microelectronics Workshop, October 2018

Multiscale co-design approach – a collaborative approach to meet the urgent need

- Develop an "end-to-end" co-design framework
- Applications ⟷ Algorithms ⟷ System SW ⟷ System HW



Validating the need with industry

Will fully automated vehicles be viable with conventional computing approaches, or will they require a step-change in computing?

What are the energy requirements to support on-board sensing and computing for fully automated vehicles?

What advanced computing approaches could reduce the energy requirements for fully automated vehicles while meeting their computational requirements?

Summary

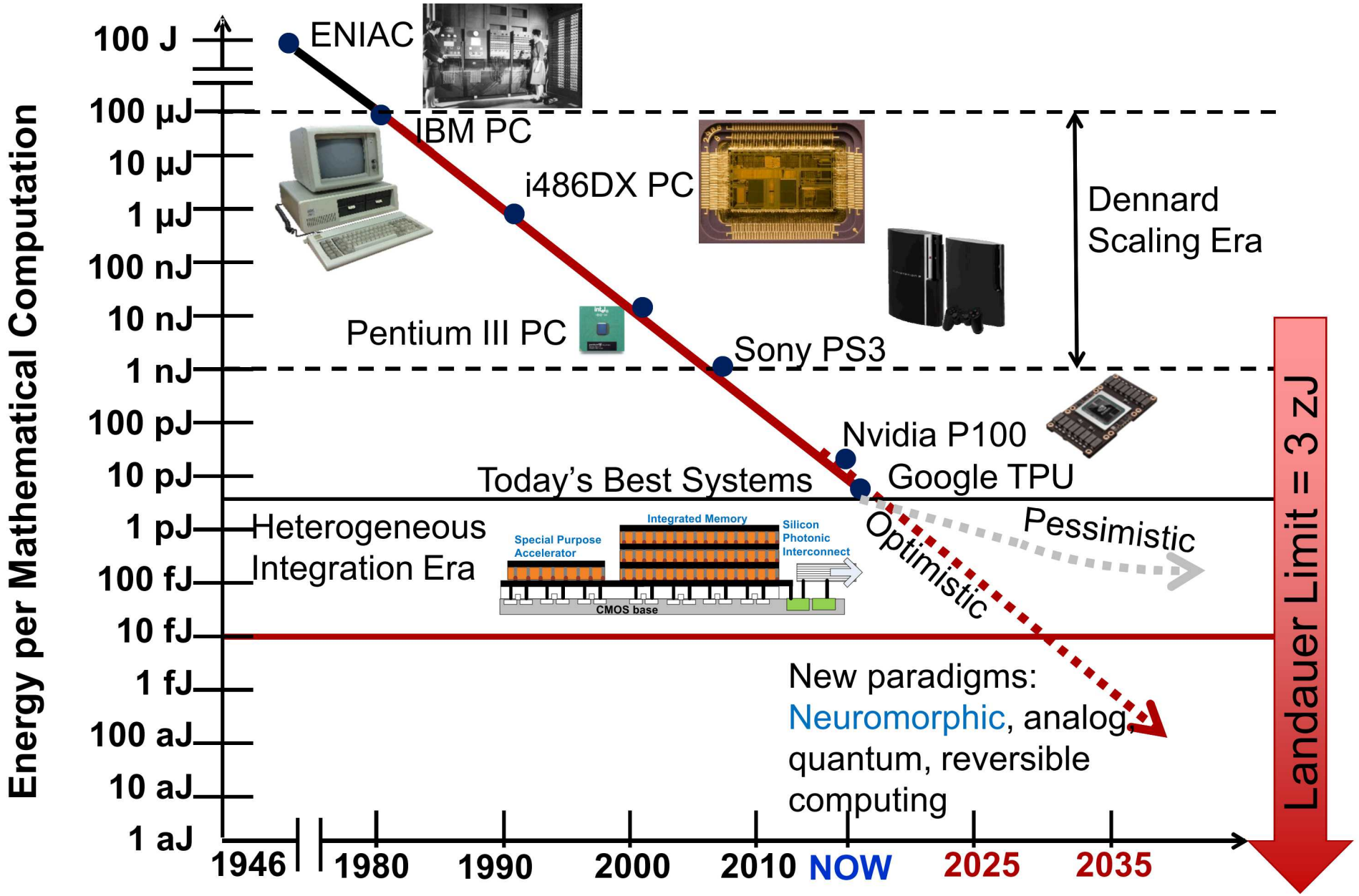
- Technology is available to get to Level 4 driving, but there appears to be a significant gap to achieve full Level 5
- We estimate that 10 TOPS/watt (minimum) should be a target computing performance goal for Level 5 driving
- There appears to be a credible path to ~ 1 TOPS/watt
- Beyond 1 TOPS/watt requires new computing architectures and microelectronic devices
- There is a role for EERE offices to advance fundamental developments in materials, device physics, and algorithm from TRL 1 up through TRL 4
  - manufacturing
  - devices and prototypes

Backups

Examples of microelectronics devices

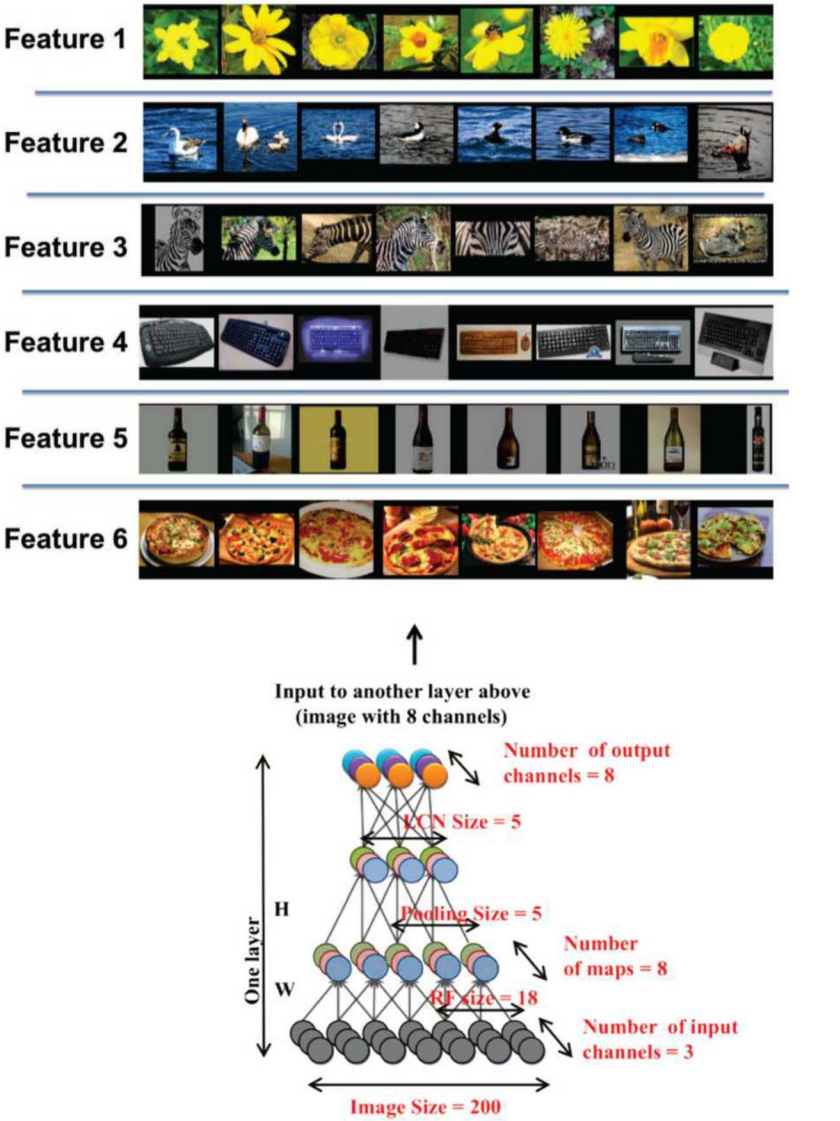
Automated driving safety

Evolution of Computing Machinery



Why should we continue these gains?

- Google Deep Learning Study
  - 16000 core, 1000 machine GPU cluster
  - Trained on 10 million 200 x 200 pixel images
  - Training required 3 days
  - Training dataset size: no larger than what can be trained in 1 week
- What would they like to do?
  - ~2 billion photos uploaded to internet per day (2014)
  - Can we train a deep net on one day of image data?
  - Assume 1000 x 1000 nominal image size, linear scaling (both assumptions are unrealistically optimistic)
  - *Requires 5 ZettaIPS to train in 3 days*
    (ZettaIPS = 10^{21} IPS; ~5 billion modern GPU cores)
  - World doesn't produce enough power for this!
  - Data is increasing exponentially with time
- Need >10^{16} - 10^{18} instruction-per-second on one IC
  - Less than 10 fJ per instruction energy budget



Q. Le, IEEE ICASSP 2013

Artificial neural networks and deep learning (example)

Computers are fast and efficient at implementing task-specific instructions

$$\frac{dx}{dt} = x^2$$

```
x=1;
dt = 0.01;
for i = 1:1000 {
    dxdt = x.^2;
    xnew = dt*dxdt;
    x = xnew; }
end
```

Computers struggle when there are no clear instructions for the task

Which one of these images is a cat?


Image recognition
Autonomous driving
Natural language processing

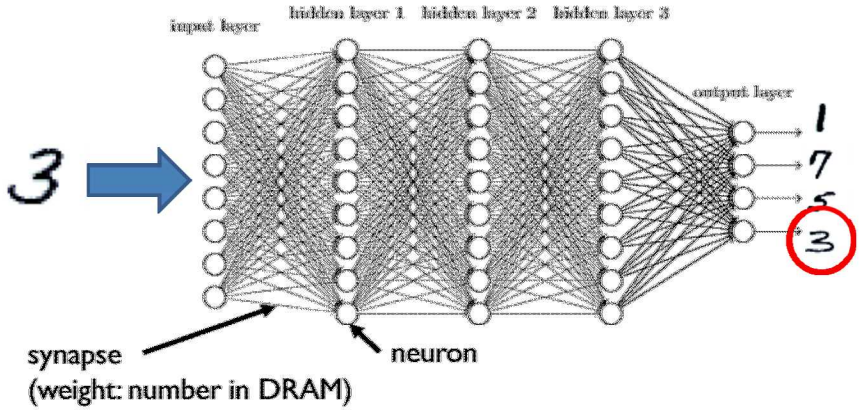
Artificial neural networks: use training examples and error backpropagation to find the matrix weights that correctly maps the input **x** onto the desired output **y**

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Artificial neural networks are power intensive

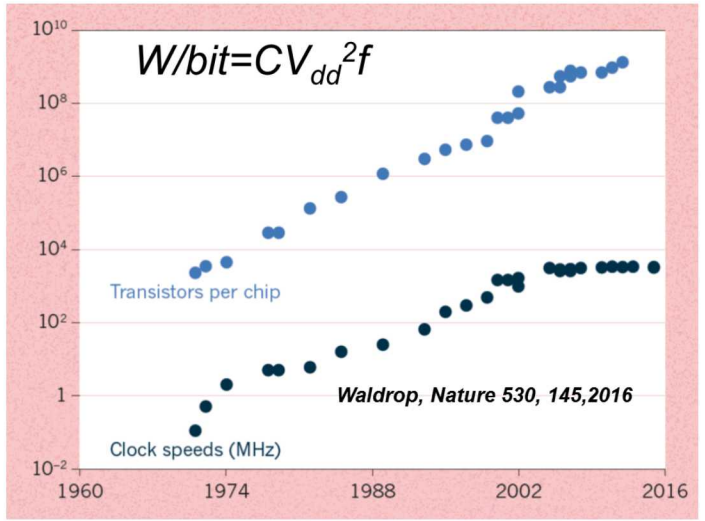
Andrew Ng, Coursera
Nawrocki et al. *IEEE Elec. Dev.* 2016

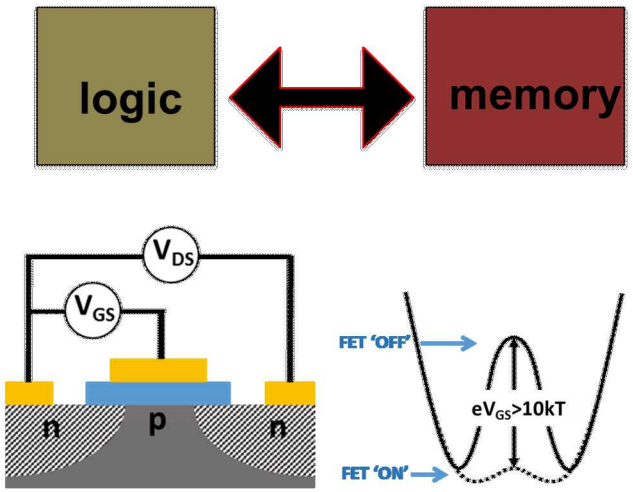
n, m > 1000

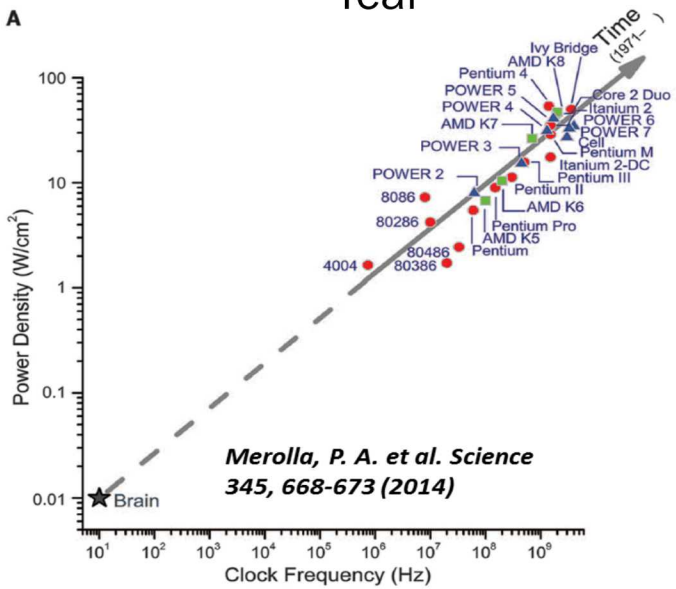


Excess power limits IC performance (example)

http://www.phys.ncku.edu.tw/~htsu/humor/fry_egg.html

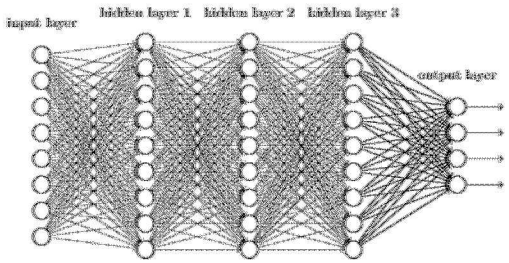






Digital vs analog implementations of neural networks (example)

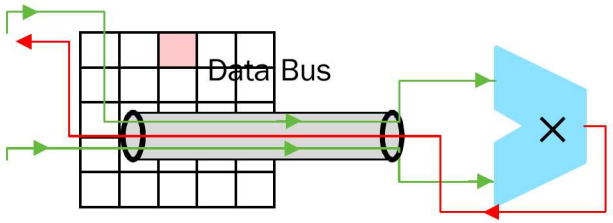
$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$



Von Neumann Digital
Separate logic and memory structures

SRAM to store the weights

Arithmetic logic unit for multiplication

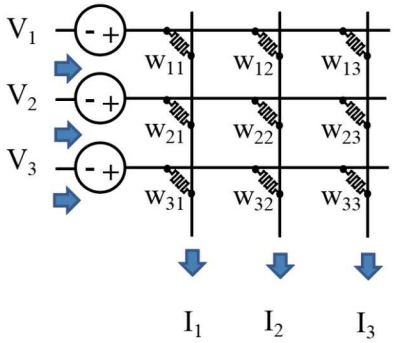


Uses established CMOS technology
Data bus results in latency and power

In-memory Parallel Analog
Use non-volatile memory

Crossbar for matrix multiplication

Conductance of each element can be changed in a predictable manner

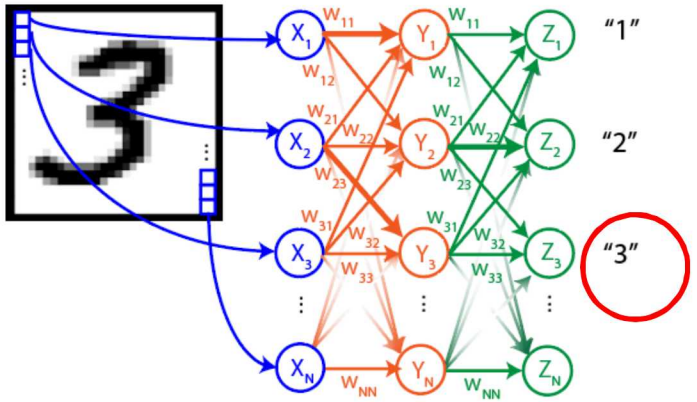


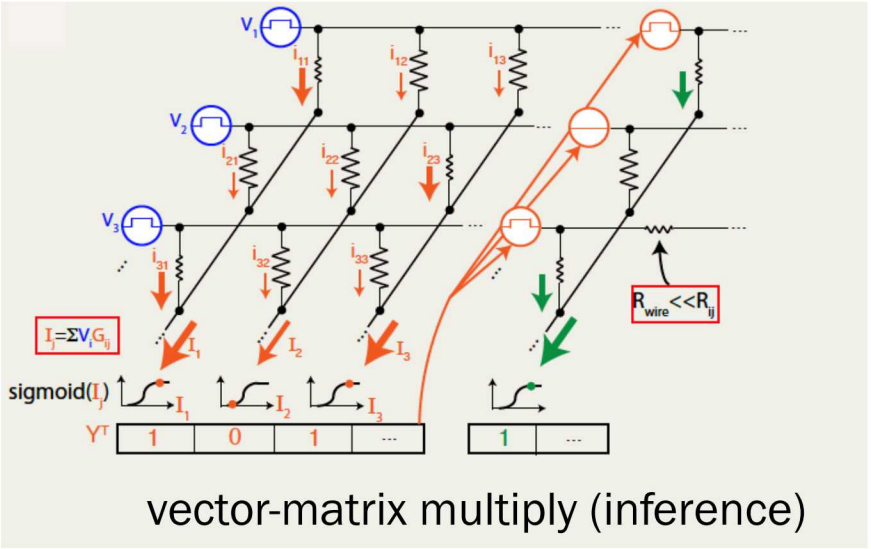
$I_1 = V_1W_{11} + V_2W_{21} + V_3W_{31}$

Simultaneous logic and memory
3 orders of magnitude less power

M. Marinella, *IEEE Circuits and Systems*, 8, 86-101, 2018
Zidan, Strachan, & Lu, *Nat. Elec.* 1, 22, 2018

Overcome CMOS bottleneck → in-memory computation (example)

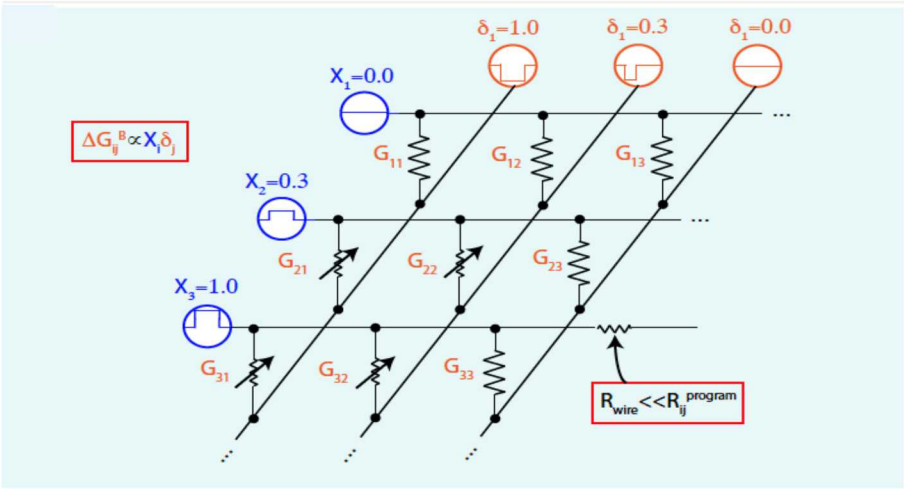




vector-matrix multiply (inference)

To beat ASICs, GPUs:

- scalable: analog signals must reach edge of 1,000 x 1,000 matrix (R>100 MΩ)
- "blind writes" with linear and symmetric programmability
- accurate: matrix operations must have CMOS equivalent accuracy
- low variation, degradation: must cycle > billion times without changes

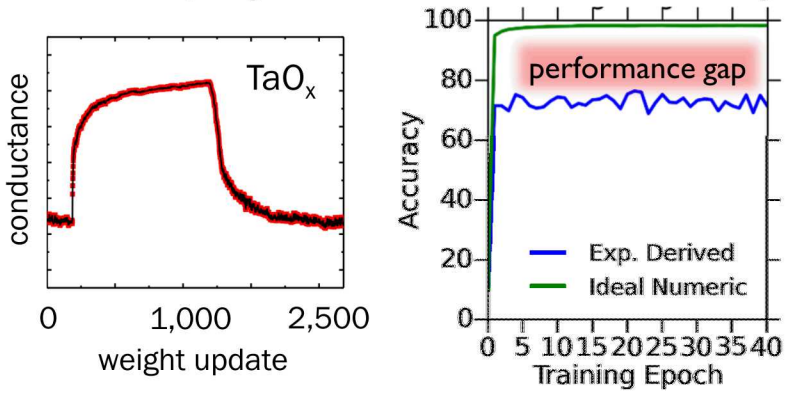


outer product (learning)

A novel material approach to neuromorphic computing

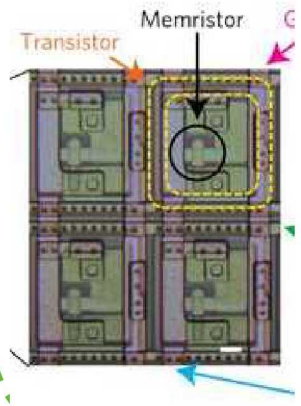
Two-terminal memristive approach

non linear, asymmetric => limited accuracy



Jacobs-Gedrim et al, 2018

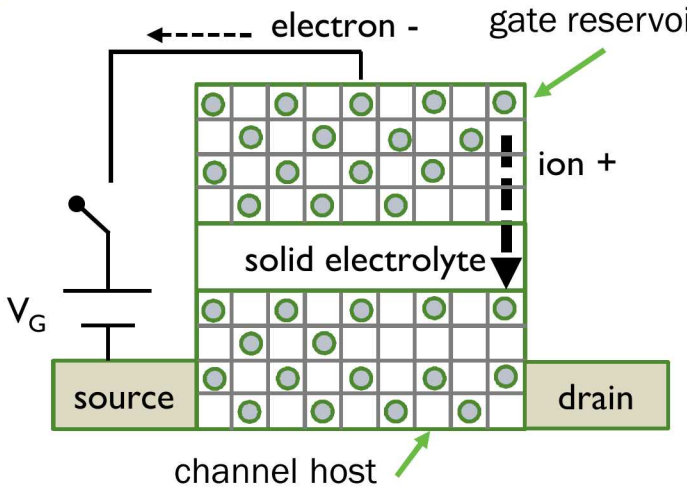
circuit parasitics => limit array sizes & efficient



- impedance ~10k requires CMOS transistor, large wires (>10μm!!!)
- Unable to support >100x100 arrays needed to beat ASICs
- Large programming currents inhibit training parallelism, devices are programmed element-by-element

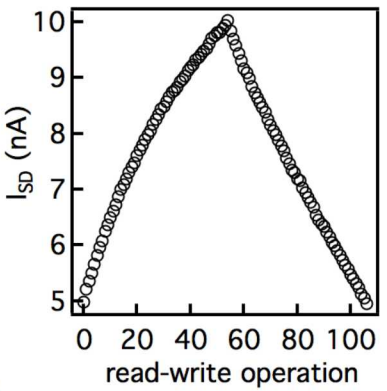
Li et. al Nature Electronics, 2018

Novel approach: ion-insertion memory



ion insertion → modulate resistance

linear, symmetric



low current/voltage

- <10 nA read/write current
- equivalent CMOS accuracy
- fully-parallel training

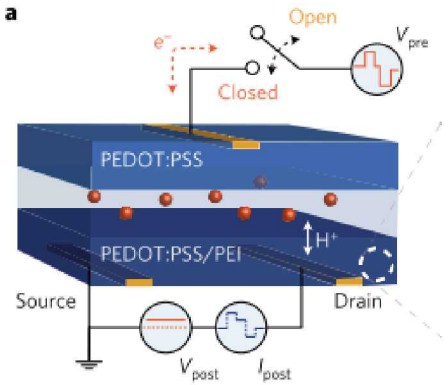
Fuller et. al. Advanced Materials, 2018

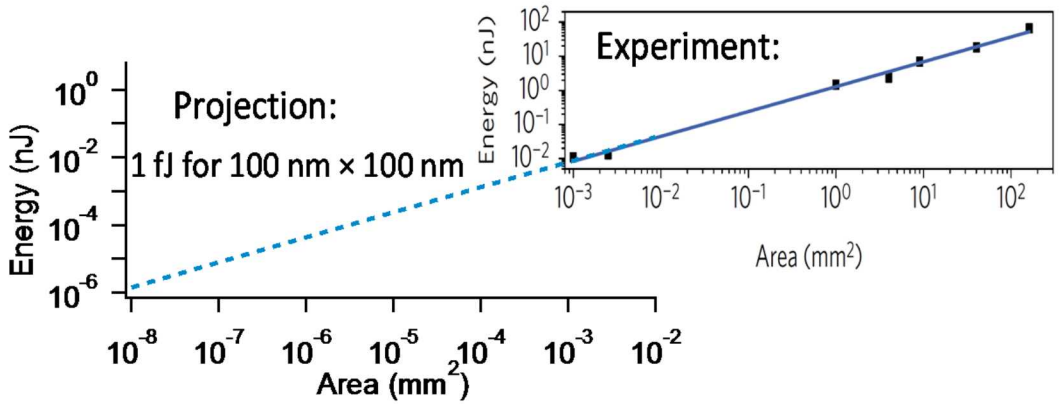
Ion-insertion devices can reach < 1 fJ / MAC**

nature materials
LETTERS
PUBLISHED ONLINE: 20 FEBRUARY 2017 | DOI: 10.1038/NMAT4856

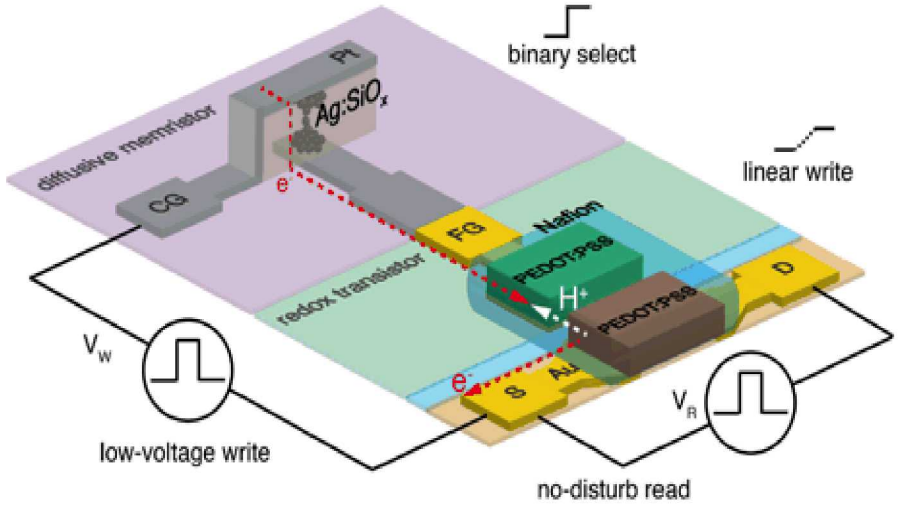
A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing

Yoeri van de Burgt[1†‡], Ewout Lubberman[1,2‡], Elliot J. Fuller[3], Scott T. Keene[1], Grégorio C. Faria[1,4], Sapan Agarwal[3], Matthew J. Marinella[5], A. Alec Talin[3]* and Alberto Salleo[1]*





binary selector enables parallel programming



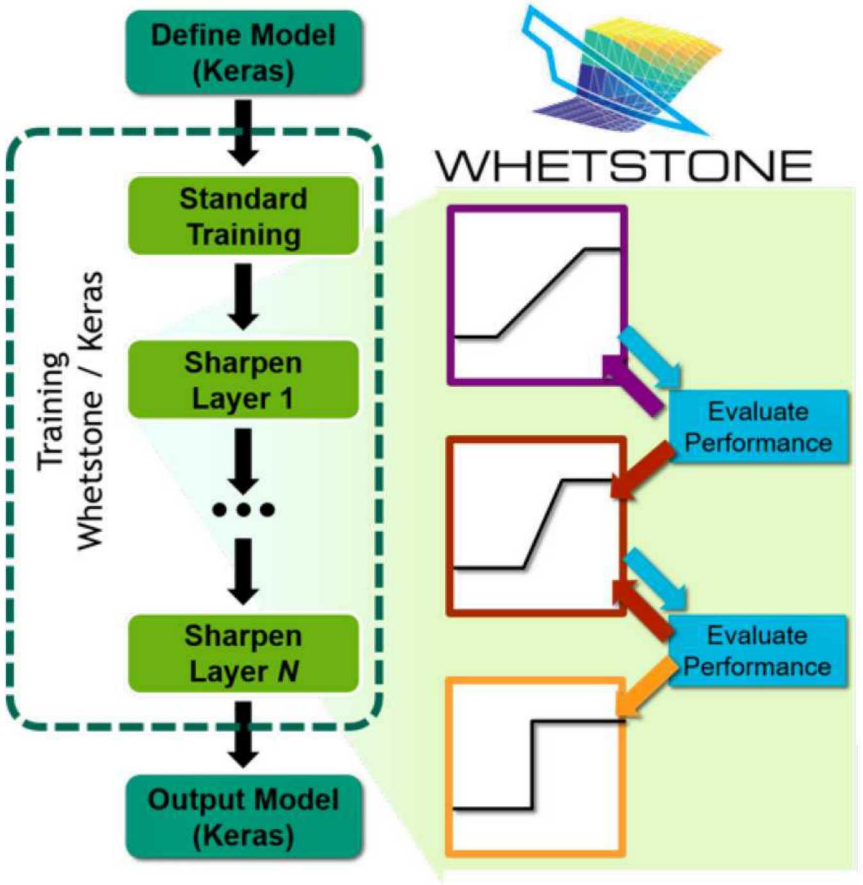
Performance demonstrated

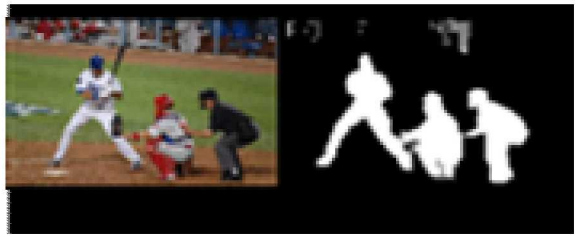
- < 1 fJ / MAC
- > 10 year retention
- $V_{program}$<10kT/q
- 1 GHz switching rate
- 10^{12} endurance
- 100 MΩ impedance

(not achievable with pure electronic memory, i.e. FLASH)

** see slide # 18 for perspective on energy performance (1 fJ = 10^{-15} J)

Deep learning networks can efficiently adapt to existing neuromorphic platforms



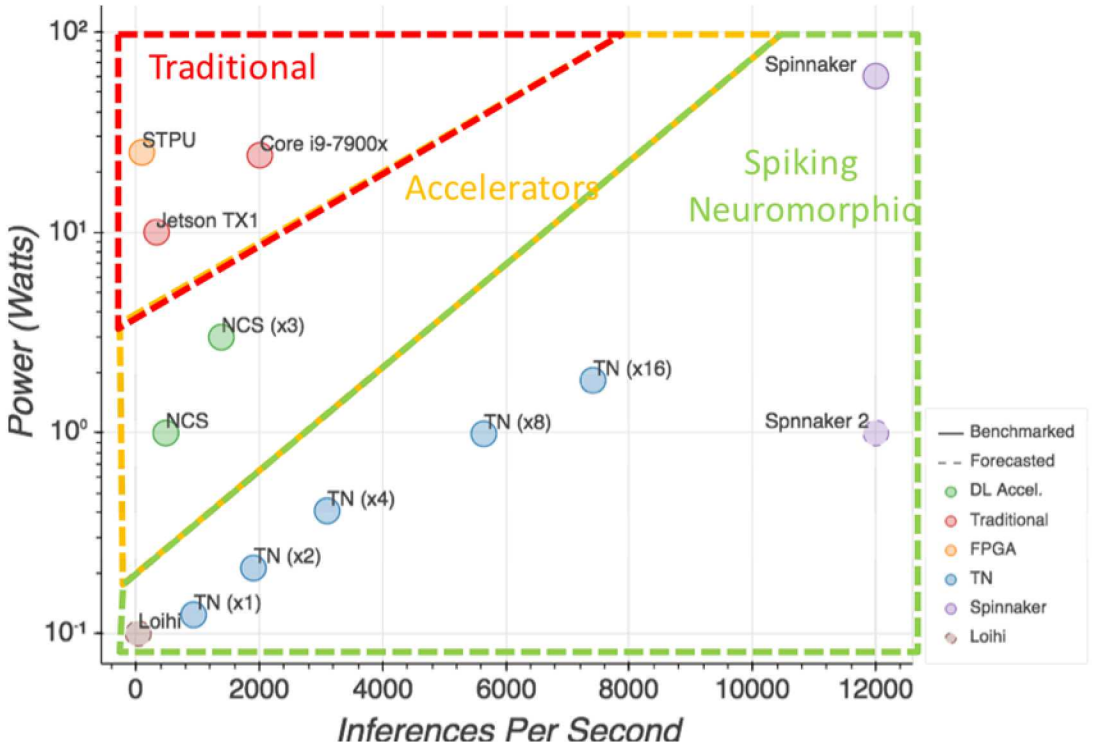


ARTICLES
https://doi.org/10.1038/s42256-018-0015-y
nature machine intelligence

Training deep neural networks for binary communication with the Whetstone method

William Severa*, Craig M. Vineyard, Ryan Dellana, Stephen J. Verzi and James B. Aimone*

The computational cost of deep neural networks presents challenges to broadly deploying these algorithms. Low-power and embedded neuromorphic processors offer potentially dramatic performance-per-watt improvements over traditional processors. However, programming these brain-inspired platforms generally requires platform-specific expertise. It is therefore difficult to achieve state-of-the-art performance on these platforms, limiting their applicability. Here we present Whetstone, a method to bridge this gap by converting deep neural networks to have discrete, binary communication. During the training process, the activation function at each layer is progressively sharpened towards a threshold activation, with limited loss in performance. Whetstone sharpened networks do not require a rate code or other spike-based coding scheme, thus producing networks comparable in timing and size to conventional artificial neural networks. We demonstrate Whetstone on a number of architectures and tasks such as image classification, autoencoders and semantic segmentation. Whetstone is currently implemented within the Keras wrapper for TensorFlow and is widely extendable.



Automakers Are Betting Big on Consumer Demand

"… although **consumers agree safety**-related technology is the **most important feature** in cars, people don't want to pay more for it." – *Business Insider, 2017*

"… we find that the **average household** is willing to pay a significant amount for automation: about $3500 for partial automation and **$4900 for full automation**." – *Daziano, 2017*

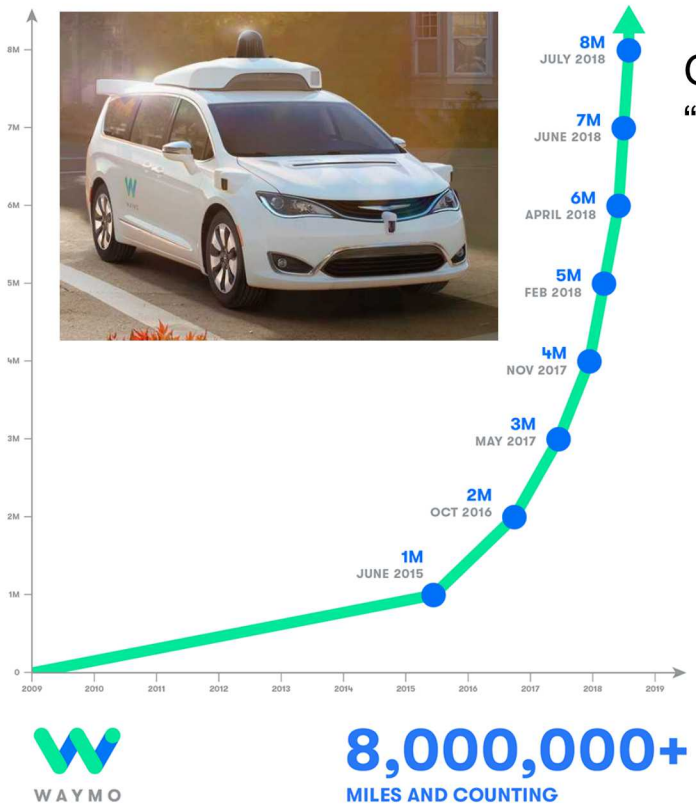
"Driver Assistance" features for emergency braking (frontal only), and traffic jam cruise control are very popular with consumers, but can only add ~ $1K to the price

Consumers are excited about "hands off" commutes to work, but only if the self-driving is "safe", and only for a limited price increase vs. normal cars

DRAPER

Safety of Self-Driving Today is Measured in Miles …

But should it be?



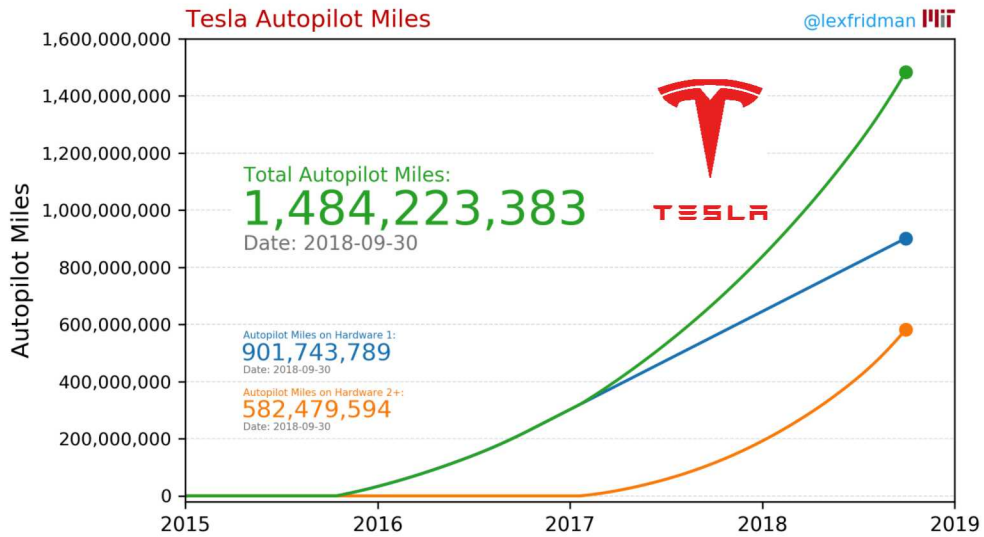
https://waymo.com/ontheroad/

Google/Waymo has – by a factor 10+ – the most "experienced" self-driving system with over $1 Billion* spent

Cruise reported **141K miles** in CA for 2016-2017

https://jalopnik.com/gm-cruise-prepping-launch-of-driverless-car-pilot-in-sa-1826571157

Only a Highway Self-Driving System (so far)…

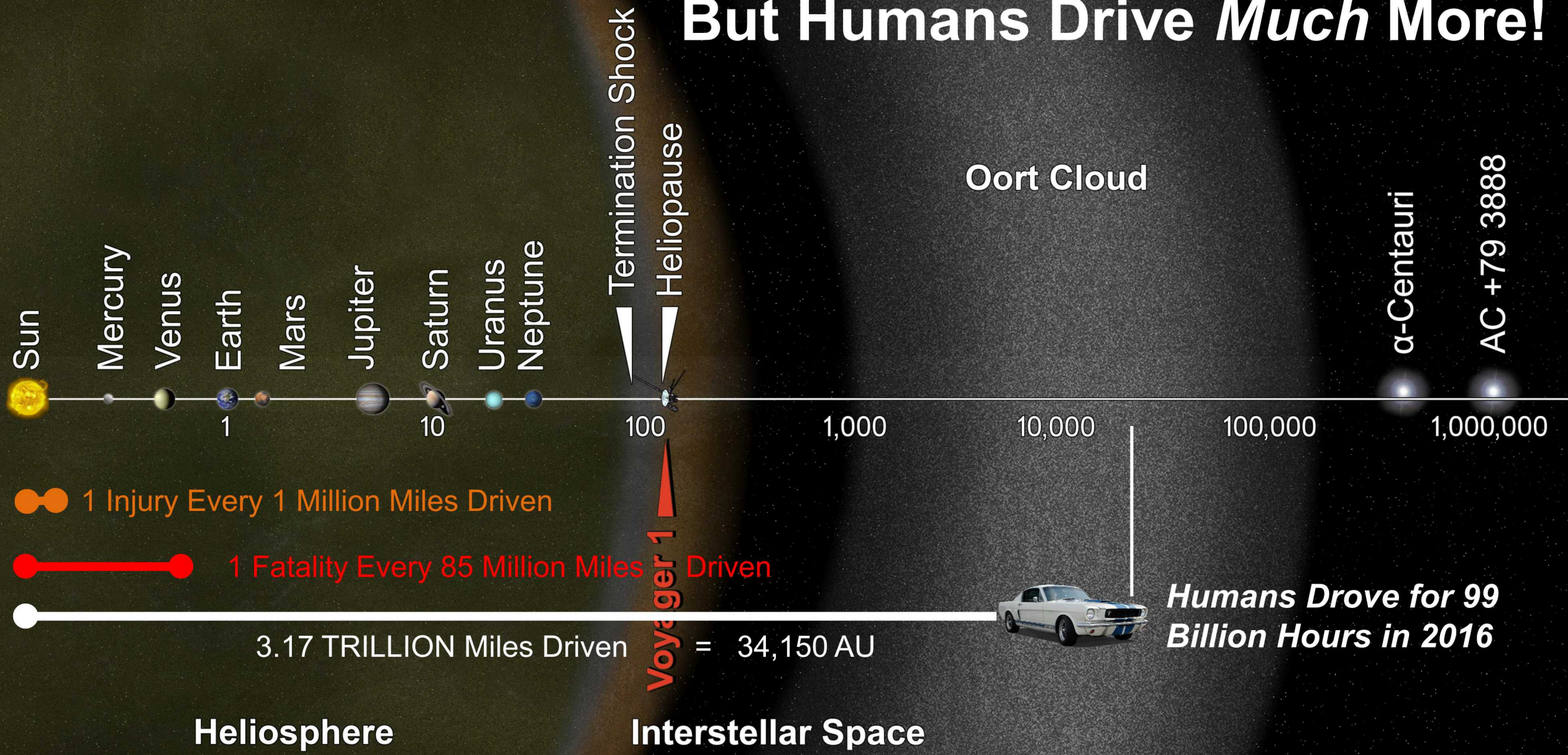


All groups tout their systems as "safe", and (someday) will reduce injuries and deaths on the road

*https://spectrum.ieee.org/cars-that-think/transportation/self-driving/google-has-spent-over-11-billion-on-selfdriving-tech

DRAPER

But Humans Drive *Much* More!



2016 United States Department of Transportation Data

All miles driven by 288 Million Humans!

https://photojournal.jpl.nasa.gov/catalog/PIA17046

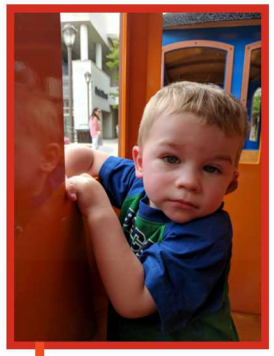
How Should We Measure Safety?

What is the safety goal for self-driving cars?

10X Better Than Human?

100X Better Than Human?

And who sets that Requirement?



My nephew has "accidents" 1 or more times an hour, they are mitigated by diapers and cuteness

12 (Estimated) Injured

WAYMO

874 Hrs

3,144,000 Injured

32,000 Hrs

37,461 Killed

2.6M Hrs

138 Injured*

1.3M Hrs

10 Killed*

*2006-2015

18M Hrs

1 | 10 | 100 | 1000 | 10,000 | 100,000 | 1 Million | 10 Million | 100 Million | 1 Billion

Mean Time Between Failure, MTBF (Hours)

Frequent | *Reasonably Probable* | *Remote* | *Extremely Remote*