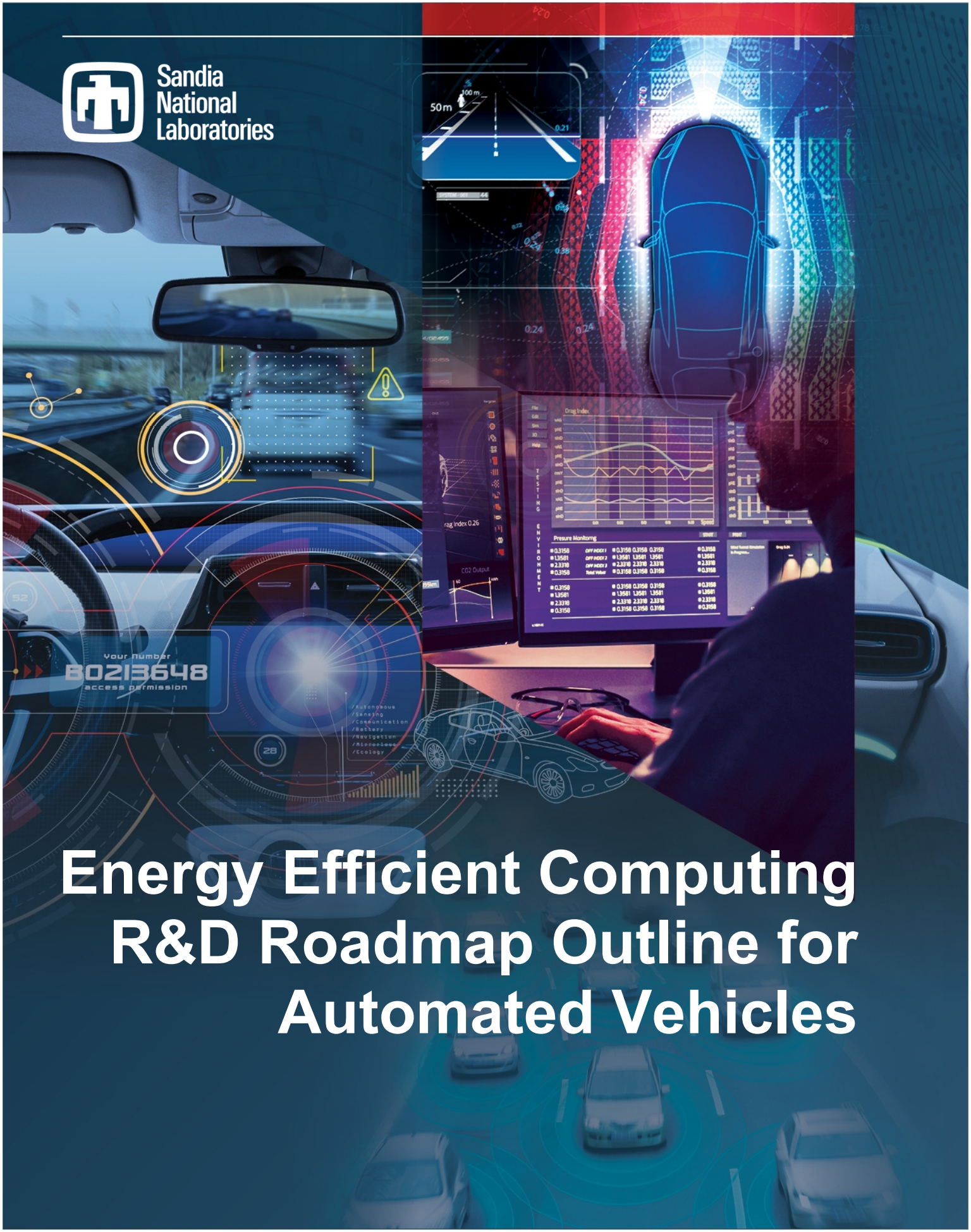

Energy Efficient Computing R&D Roadmap Outline for Automated Vehicles

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.

Prepared by Sandia National Laboratories in partnership with:

ARM

Rob Aitken

Carnegie Mellon University

Yorie Nakahira

Hewlett Packard Enterprise

John Paul Strachan
Kirk Bresniker

Intel

Ian Young

Sandia National Laboratories

Zhiyong Li
Lennie Klebanoff
Carrie Burchard
Suhas Kumar
Matt Marinella
William Severa
Alec Talin
Craig Vineyard
Christian Mailhiot

University of Michigan

Robert Dick
Wei Lu

USCAR

Jace Mogill

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

This page left blank.

CONTENTS

TABLE OF FIGURES

ACRONYMS

ABBREVIATION	DEFINITION
3D	Three-dimensional
AI	Artificial intelligence
ASIC	Application-specific integrated circuits
AutoML	Automated Machine Language
AV	Automated vehicle
CAN	Controller area network
CAV	Connected and automated vehicles
CGRA	Coarse-grained reconfigurable architecture
CMOS	Complementary metal-oxide semiconductor
CNN	Convolutional Neural Network
CPU	Central processing unit
DARPA	Defense Advanced Research Projects Agency
DAV	Decentralized Automated Vehicles
DOE	Department of Energy
EECAV	Energy Efficient Computing for Automated Vehicles
FeRAM	Ferroelectric Random Access Memory
FET	Field-effect transistors
FHE	Fully Homomorphic Encryption
FPGA	Field programmable gate array
GPU	Graphics processing unit
HBM	High Bandwidth Memory
IC	Integrated circuit
IoT	Internet-of-Things
IRDS	International Roadmap for Devices and Systems
LiDAR	Light Detection and Ranging
MAC	Multiply accumulate (operation)
ML	Machine Learning
NAS	Neural Architecture Search
NC-FET	Negative capacitance field-effect transistor
NPU	Network Processing Unit
ODD	Operation Design Domain

OEM	Original Equipment Manufacturer
OP	Operation
OPS	Operations per Second
PCIe	Peripheral component interconnect express
PCM	Phase Change Memory
R&D	Research and development
RASM	Reliability, Availability, Serviceability, Maintainability
ReRAM	Resistive Random Access Memory
SAE	Society of Automotive Engineers
SerDes	Serializer/deserializer
SNL	Sandia National Laboratories
SWaP	Size, Weight, and Power
TA	Technical Area
TFET	Tunnel field-effect transistors
TOPS	Tera Operations per Second
USCAR	United States Council for Automotive Research
Wh	Watt-hour

1. INTRODUCTION

Automated vehicles (AV) hold great promise for improving safety, as well as reducing congestion and emissions. In order to make automated vehicles commercially viable, a reliable and high-performance vehicle-based computing platform that meets ever-increasing computational demands will be key. Given the state of existing digital computing technology, designers will face significant challenges in meeting the needs of highly automated vehicles without exceeding thermal constraints or consuming a large portion of the energy available on vehicles, thus reducing range between charges or refills. The accompanying increases in energy for AV use will place increased demand on energy production and distribution infrastructure, which also motivates increasing computational energy efficiency.

To meet both energy efficiency and computational performance goals, targeted research and development (R&D) is needed in those technical areas that most impact computation and energy efficiency in the context of the size, weight, power, and thermal constraints of a vehicle. Our approach to this goal was to assemble a "Roadmap Team," with representatives from national lab, academia, and industry (microelectronics and automotive), in order to develop a "Roadmap Outline" for the advancement of energy efficient computing for automated vehicles (EECAV). The purpose of the Roadmap Outline is to identify the R&D challenges that must be overcome for the realization of highly automated driving in retail vehicles with low power consumption and high computational performance. The retail-vehicle focus ensures that the AV R&D problems are considered in the context of the most complicated ownership, service, and support scenario, identifying technical problems while being cognizant of additional economic and regulatory constraints. In addition, we focused on long-term R&D problems that would typically be the target of public or private investment, and hopefully complementary to ongoing industry investments. Above all, we aimed to provide a neutral technical assessment, without bias or favoritism to a particular technology, ensuring technically sound input for an eventual long-term EECAV roadmap that can guide R&D investment. The Roadmap Outline is purposefully high-level and intended to serve as the starting point of a more comprehensive energy efficient computing roadmap that will include broader partners in a next phase.

As the automation level of vehicles increases and the operation design domain (ODD) expands, the required computation will increase significantly on AVs in order to perform dynamic driving tasks reliably and safely. Conventional complementary metal-oxide semiconductor (CMOS) digital computing technology is approaching the end of Moore's Law scaling, as illustrated in *Figure 1*. The energy/power performance ratios of computer chips faces the difficulty of breaking the CMOS bottleneck around picojoule/(multiply-accumulate) operation and has shown an indication of hitting an inflection point. The projected computing performance as shown will not be able to meet the demand of highly automated AV functions alone. A new co-design paradigm shift is needed where the research activity into computer chips, architecture, algorithms, and sensors are highly integrated to advance the energy efficiency of AV computing.

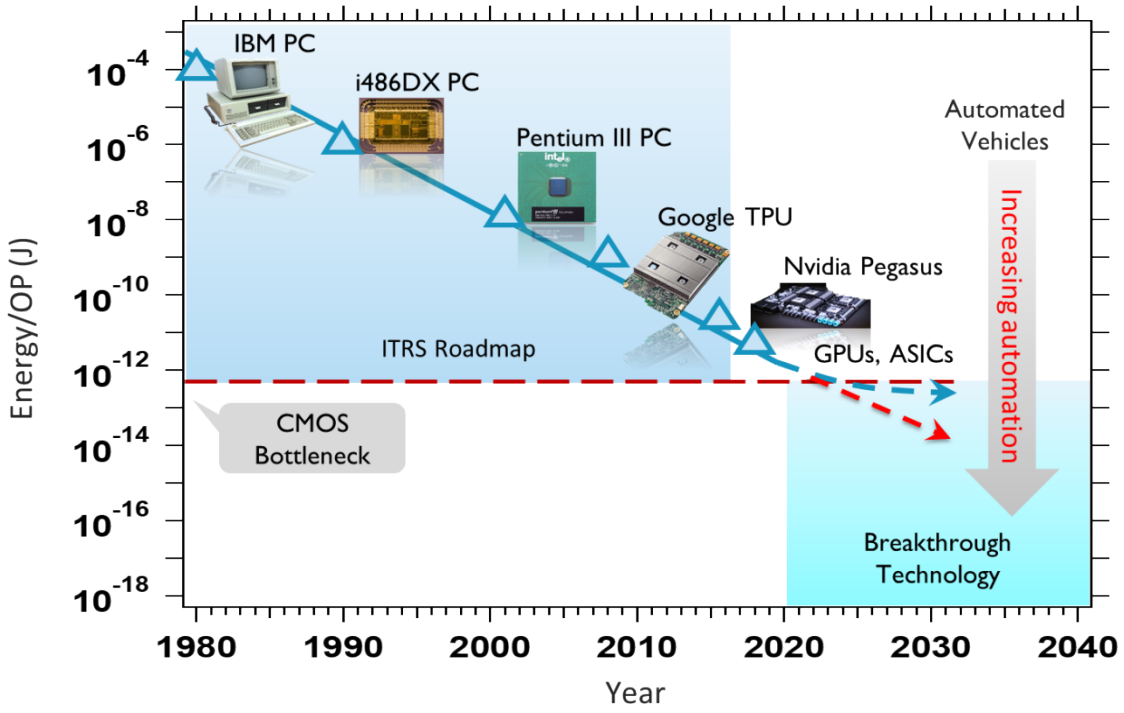

Figure 1. *Energy per Operation (OP) plotted versus Year. A new co-design paradigm will be needed to meet the energy efficient computing requirements of highly automated driving.*

AVs will rely on sophisticated inference systems that make hard decisions based on complex, multi-modal sensed data. Communication among vehicles, and with infrastructure, could provide many benefits. However, wireless communication may experience intermittent outages due to interference and obstructions. Therefore, time- and safety-critical decisions must be possible on-vehicle, not dependent on the cloud. This requires local computation. Unfortunately, existing automated commercial AV computer vision prototypes require an order of magnitude more power than will be practical to cool in future commodity vehicles.

An energy expenditure of 300 Wh/mile is typical for modern non-automated electric vehicles.[1] For a vehicle speed of 60 mph, this corresponds to a power expenditure of 18 kW for electric vehicle locomotion. Highly automated vehicles currently being tested by leading developers generally expend about 2.5 - 4 kW of power on the sensing and computation necessary for

[1] Kane, M., "Electric Car Energy Consumption (EPA) Compared," April 1, 2019, https://insideevs.com/

inference with similar numbers noted for more recent systems.[2,3,4,5] This level of power consumption for automated driving represents a burdensome fraction of the available onboard power. It is difficult to narrow this 2.5 - 4 kW range further for the current AV prototypes, as existing systems will trade off safety and functionality against energy efficiency, but we will use an estimate of 3 kW for discussion purposes. We anticipate that a 300 W "all-in" target for on-board AV computation and associated support equipment will be required to meet thermal constraints in economically viable vehicles. Even if the thermal problem could be resolved, there would likely be significant resistance to consuming much over 10% of locomotion power for onboard computing, which can otherwise be used for the travel range of the AVs.

It is instructive to compare the 3 kW of current developmental platform vehicles to the 300 W power target in order to identify potential sources for the required improvement. The possibilities for transitioning from the current 3 kW to 300 W are depicted in *Figure 2*.

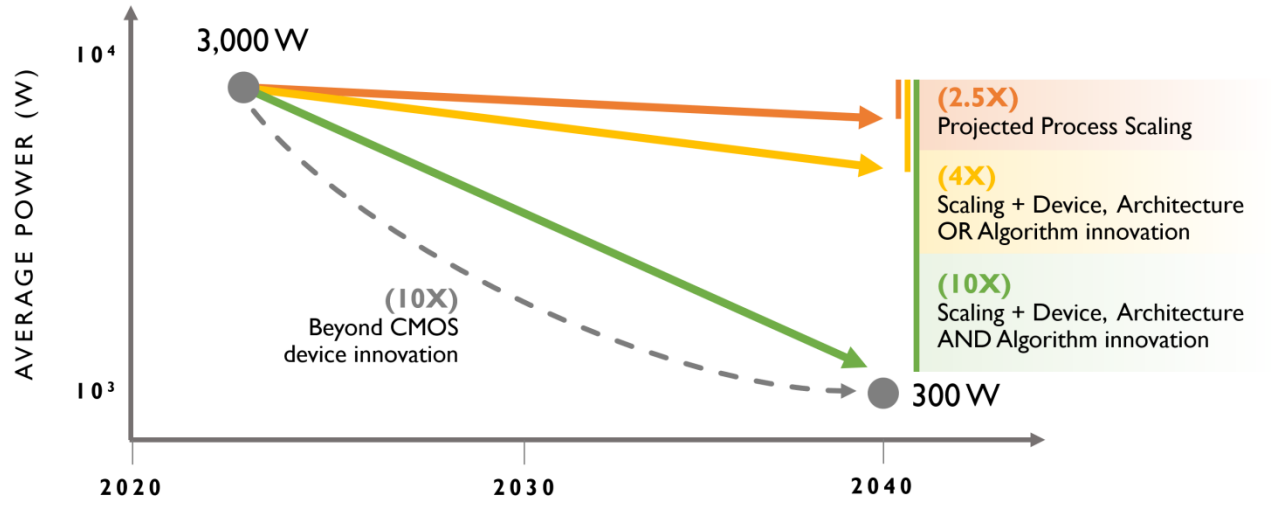

Figure 2. Different technology development paths to the energy efficiency improvements required for automated driving.

---

[2] Stewart, J., Self-Driving Cars Use Crazy Amounts of Power, and It's Becoming a Problem, Wired, Feb. 2018. https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/.

[3] Pant, Y.V., Abbas, H., Nischal, K.N., Kelkar, P., Kumar, D., Devietti, J., Mangharam, R., "Power-efficient Algorithms for Autonomous Navigation," Proc. International Conference on Complex Systems Engineering (ICCSE), Nov. 2015.

[4] Hamza, K., Willard J., Chu, K. and Laberteaux, K.P., Modeling the Effect of Power Consumption in Automated Driving Systems on Vehicle Energy Efficiency for Real-World Driving in California. Transportation Research Record 2673 (4): 339–47. https://doi.org/10.1177/0361198119835508.

[5] National Academies of Sciences, Engineering, and Medicine 2021. Assessment of Technologies for Improving Light-Duty Vehicle Fuel Economy 2025-2035. Washington, DC: The National Academies Press. https://doi.org/10.17226/26092.

Extrapolating based on projections from the International Roadmap for Devices and Systems (IRDS),[6] we estimate that the energy efficiency of CMOS/FinFET technology (Process Scaling) will improve by 2.5X between 2020 and 2040. Additionally, with a 1.5X improvement in energy efficiency due to device, architecture, *or* algorithm innovation, the estimated power reduction improves to 4X by 2040, but is still insufficient. A breakthrough beyond-CMOS device innovation could conceivably, by itself, achieve the 300 W power target. However, building on advances in Process Scaling, Devices, Architecture, *and* Algorithms to reach the required 10X power reduction, as shown in *Figure 2*, has a higher probability of success than focusing on only one level of the design process.

Figure 2's depiction of the required 10X improvement to AV computational energy efficiency is semi-quantitative because there are several factors which are not considered for simplicity. For example, there are likely to be regulatory burdens (security, safety and privacy) on the sensor array and associated computation which will affect the power budget. Also, the (3 kW) power burden varies somewhat amongst the current prototype vehicles. Improvements exceeding 10X are valuable and can be traded off for improvements in reliability and functionality.

The remainder of this Roadmap Outline describes the R&D problems that must be overcome for highly automated driving with 300 W power consumption for computing in retail vehicles.

2. BACKGROUND

In the early phases of the roadmapping activity, a survey was conducted of published technical literature to assess if there had been prior attempts to evaluate the state of R&D with regard to AV computation, in general, and AV computational energy efficiency, specifically. The purpose was to be aware of prior work in order to avoid duplication and have relevant prior literature inform our efforts. Although the AV technical field is quite broad, our review of the literature was limited to the topic of AV computational energy efficiency and the technical issues that directly affect it. The papers were discovered by searching the technical literature (i.e., using Google Scholar) using search terms such as "automated vehicle computation" and "AV computer energy efficiency."

Overall, this literature search found few previously published articles or reviews discussing the computational requirements needed for AV operation. Essentially, all prior literature assessing future AV activity assumes the computational capability will be there. A few articles provided discussion of the demand side of the problem, namely the sensor/camera data input that must be processed. No discussion is provided in the prior literature on energy efficiency associated with computation for AV, except for a 2019 Sandia National Laboratories (SNL) workshop report[7] which documents a progenitor activity to the current roadmapping activity. Thus, there has been no prior roadmapping activity on the EECAV topic.

This literature review lent confidence that our R&D AV roadmapping activity is providing new investigation of the computational needs for AV operation and the required energy efficiency

[6] International Roadmap of Devices and Systems 2020 Update: More Moore, 2020.

[7] Mailhiot, C., Severa, W.M., Moen, C.D., Jones, T.B., "Workshop on Advanced Computing for Connected and Automated Vehicles," Nov. 2019, SAND2019-14117.

with which that computation must be conducted. A summary of the literature review will be available on Sandia website or upon request.

Draft technical content for the Roadmap Outline was presented at an online EECAV workshop held with the AV and related technical communities on May 11 and May 12, 2021 to present the work of the Roadmap Team, discuss the general AV challenges from a "required R&D" perspective, and receive feedback on the draft Roadmap Outline and the R&D problem areas. The Agenda for the EECAV Workshop, copies of the workshop presentations and a summary of workshop will be available on Sandia website or upon request.

3. ROADMAP SCOPE AND TIMELINES

3.1. Scope

The Roadmap Team developed a couple of "boundary conditions" for the roadmapping activity. The first of these is the location of the computational capacity. There are several technical scenarios within which AVs may operate. At one end of the spectrum, the guidance and control of AV operation relies heavily on intelligent infrastructure through the "vehicle to X" or "V2X" communication. A majority of the computation task on the vehicle in this case can be shifted to off-board through low latency and high bandwidth communication network as part of the infrastructure. At the other end, all computation and sensing hardware reside on-board the AV and are self-sufficient to support the dynamic driving tasks within the operation design domain.

The team decided that the latter scenario, in which all of the computational capacity resides on the vehicle, established a boundary condition for this work. This choice was made because input from the Original Equipment Manufacturers (OEMs) indicated commercial retail AVs would have to be "self-sufficient" for entry into the automotive market, particularly since any intelligent infrastructure, if developed, would initially be sparse and therefore often unavailable. Moreover, communication with external infrastructure would rely on wireless communication that is often subject to intermittent outages due to interference, obstruction, and distance. Identification and adoption of the "compute on vehicle" boundary condition places extreme importance on the electrical energy efficiency with which computation is performed. It also opens up for R&D consideration all technical issues associated with hosting computational capacity on the vehicle, including the energy, space, and weight it takes to do so. Having said that, it should be noted that in a mature AV technology scenario, there may be a compromise between the two scenarios, where a substantial amount of "compute on vehicle" is supplemented with opportunistic "V2X" to save local power, compute more efficiently, or enable multi-vehicle cooperative driving. In other words, a mature AV technology could involve a composite of vehicular and infrastructure sensing and computation. For simplicity and to conceptually bound our R&D scope, we assume almost all the computation resides on the vehicle.

The second boundary condition considered is computational power. Given that the computational capacity resides on the vehicle, the team examined the likely constraint on electrical power that will be consumed for computation on commercially viable AV systems. The group settled on a constraint of 300 W for the electrical power consumed by the on-board computers, sensors, and any supporting peripherals to enable the automated driving functions.

3.2. Timelines

With the boundary conditions of self-sufficiency and power worked out, the Roadmap Team then discussed the timeline associated with the AV enterprise. It became clear in the initial discussions that each team member had their own timeline in mind. The representatives from the chip makers typically had the "chip commercialization" timeline in mind, which is the timeline for establishing commercial-ready (for mass consumption) chips needed for widespread AV application. The automotive OEMs had their "OEM Implementation" timeline, which specifies the timing of the OEM implementation of commercial-ready chips in vehicles. The researchers on the team were focused on when R&D needed to be performed (i.e., the R&D Timeline). Thus, in specifying the Roadmap Outline, it was important to develop three separate but mutually consistent timelines for R&D, chip commercialization, and OEM implementation. As shown in *Figure 3*, the roadmap was developed using these three conceptually distinct timelines:

1. The timeline for R&D that enables AV systems.
2. The timeline for establishing a commercially ready compute system for AV computation.
3. The timeline for OEM implementation of those systems in future mass-produced AV.

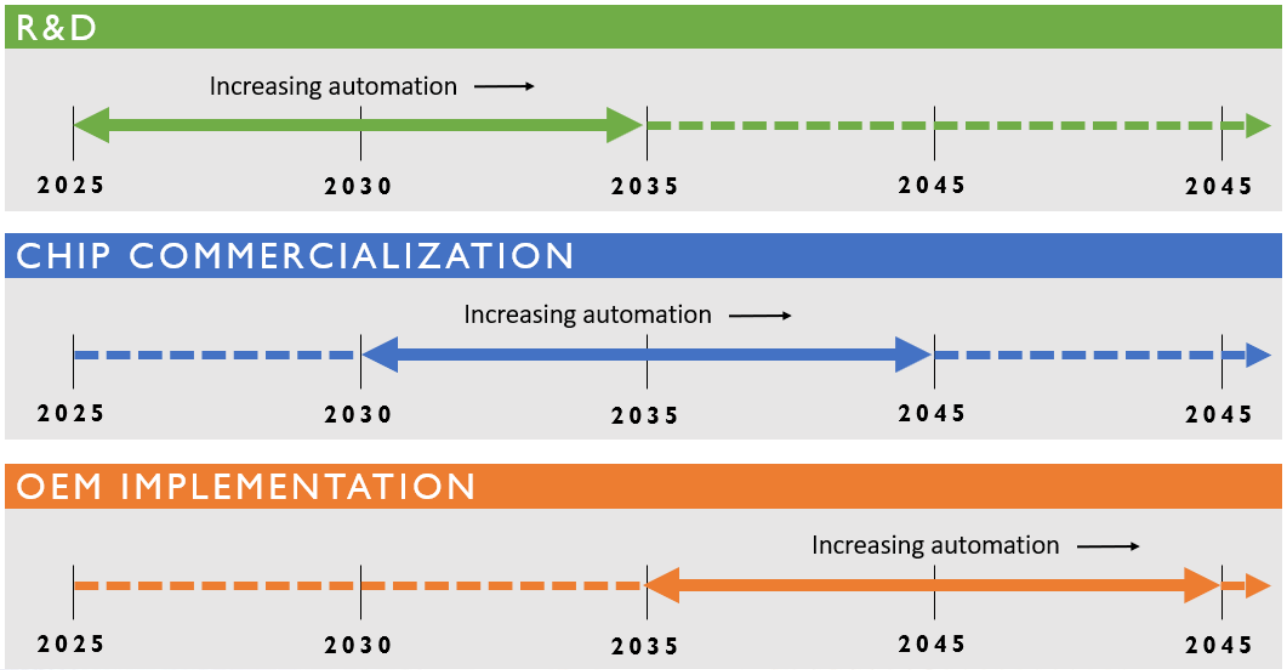

Figure 3. *Specification of the R&D Timeline, Chip Commercialization Timeline, and OEM Implementation Timeline needed for widespread adoption of mass-produced retail AV. The solid double-headed arrows indicate the time considered for each activity in the EECAV Roadmap Outline.*

The three timelines are related as follows: each timeline considers a 10-year period of activity with the end goal being widespread implementation of AV in the retail vehicle market. The team considered the typical lead time by which R&D must precede commercial implementation of technology. While a longer period could be argued for, it was decided five years was the lead time for R&D to be commercially implemented. Thus, if we assume the R&D Timeline started in 2025, the R&D for EECAV would extend from 2025 to 2035, targeting an ever-increasing level of

vehicle automation. With commercialization pushed out in time five years from R&D initiation, *Figure 3* shows that the Chip Commercialization Timeline would therefore extend from 2030 to 2040, again with increasing computational capacity supporting increasing automation in the vehicle. It would be safe to assume that even if chips were being sold in quantity (i.e., not a prototype technology), it would take five years for the OEMs to incorporate them in mass-produced retail AVs. Thus, the OEM Implementation Timeline extends from 2035 to 2045, pushed out in time from the Chip Commercialization Timeline by five years, with increasing levels of vehicle automation during that time.

3.3. Technical Areas for Research and Development

Energy efficient computing is required not only for AV, but for nearly all industries, making the identified R&D problems for AV of interest to many industries such as the Internet-of-Things (IoT), mobile systems, edge computing and data center computers. However, unlike these applications, the AV challenge presents a unique nexus of characteristics, namely, safety critical, small footprint, wide temperature range, low weight, high impact on society, and long-term deployment (~15 years) in a car. The long-term nature of AV technology is an especially distinguishing characteristic, placing high importance on upgradeability.

AVs will need sophisticated inference systems that make hard decisions based on complex, multi-model sensed data. Time- and safety-critical decisions are needed on-vehicle, which require high-performance local computation on the vehicle. As shown in *Figure 2,* current partially autonomous (but supervised) commercial AV prototypes require about 3 kW, more power than will be practical to thermally manage in future retail AVs. Moreover, this accounts for ~20 % of the total energy consumption of the current prototype passenger vehicles, including the energy for locomotion. To make thermal management in mass-market AVs practical and support long-range highly automated driving, it will be necessary to reduce the average "all-in" AV power consumption to 300 W, 10X lower than currently required. The exact improvement required to reach a level where cooling is economical and requires no major advances depends on the particular existing AV used as a baseline, but we can say with some confidence that an order of magnitude improvement in energy efficiency will be needed. This requires improving full-system energy efficiency, including improving the efficiency of hardware so it can do the same calculations using less energy as well as improving the efficiency of algorithms so they can produce the same results with less computation.

The Roadmap Team has considered both approaches and identified R&D problems in four general technical areas (TAs). These TAs encompass essential technical disciplines that impact computational energy efficiency, along with the technical problems in these areas that motivate targeted and impactful R&D investment for realizing EECAV.

The four technical areas are:

I. Chips: Materials, Devices, and Circuits.
II. Chips: Architecture, Safety, and Security.
III. Algorithms and Data Management.
IV. Sensors Data Interface.

Four teams were formed – one for each TA, and each team worked to identify R&D gaps within their TAs. The team leaders were cognizant of the partial overlaps among the TAs (e.g., the interactions between chips and architectures, and between architectures and algorithms) and considered them when formulating R&D recommendations. These topical R&D areas are listed in each of the TA sections to follow.

4. TECHNICAL AREAS FOR RESEARCH AND DEVELOPMENT

4.1. Technical Area I. Chips: Materials, Devices, and Circuits

Technical Area I (TA-I) focuses on the materials, devices, and circuits that are the building blocks of computer chips operating in the demanding automotive environment, and within the power budget allotted for computation. The ultimate computational "supply" such a chip must provide is determined by the computational "demand" of the perception, planning, routing, navigation, compliance and other essential functions of AV. These "demand" aspects are considered in the other technical areas, while TA-I strives to provide more capable and more energy efficient technology at the fundamental chip level.

R&D within TA-I is critically needed today, as CMOS scaling is approaching its physical limits. Thus, there is interest in new materials, devices, and circuits as the industry simultaneously pursues three "thrusts" for chip development:

1. **Chip Thrust 1:** Improved CMOS approaches, with continued transistor scaling, while addressing the associated problems such as increasing dynamic power and sub-threshold leakage currents.
2. **Chip Thrust 2**: CMOS-like approaches to explore alternative device/materials technologies such as field-effect transistors (FETs), tunnel field-effect transistors (TFETs), negative capacitance FETs (NC-FETs), 2-dimensional transistors, and 2-terminal memories, etc., which offer transistor-like functions with enhancements and better efficiencies.
3. **Chip Thrust 3:** Physical computing devices, which seek not merely enhanced transistor-like functions, but physics-driven complex behaviors that cannot be expressed by CMOS-like technologies. Examples include devices and computing based on ionic, quantum, and photonic processes.

These three chip technology thrusts need to be pursued simultaneously, although they will probably commercially mature and be deployed at different times. All these technical approaches need to be pursued by the AV sector to address a common set of key issues facing AVs.

Harsh environment: AVs will find themselves in almost every environment conceivable with extremes of cold, hot, dry, and wet, combined with the shock/vibration challenge of the automotive environment. Given the harsh environments faced by AV computers compared to stationary data centers or consumer electronics, computing in AV has specific demands. Availability is clearly one of the highest priorities, where availability means the percentage of time the AV functionality is operating to ensure safety. Availability includes serviceability, system tolerance to computational failures, the stability/reliability of the hardware to extremes of temperature in AV, mechanical stress, resilience to moisture and corroding chemicals, and even cosmic ray (or radiation) impact, particularly in high altitude regions. Some of these issues are

present in data-center-level computing systems, laptops, wearable electronics, remote sensing electronics, etc., but the different environmental issues are critical to different extents in these other applications.

Size, Weight, and Power (SWaP): An AV will have SWaP physical constraints within which large amounts of data and computing have to be packed, which imposes design constraints on the circuits and supports the use of three-dimensional (3D) circuit technologies (e.g., 3D NAND Flash memories). These constraints could possibly encourage the strategic use of off-vehicle storage (e.g., within data centers) for non-critical or latency-tolerant tasks. For example, object recognition libraries must be stored on-vehicle since they are of critical and urgent safety importance (e.g., recognizing a human on the road), whereas route mapping libraries might be stored off-vehicle since real-time rerouting is less time-sensitive.

Latency: Lives are on the line with AV technology. Rapid response is required to navigate safely and confidently in complicated traffic and situational environments. Data centers and consumer electronics generally do not require critical low-latency hardware for the computation tasks. However, in AVs, the entire data lifecycle, from acquisition/aggregation to analysis to action, needs to be less than the ~150 milliseconds typical of human reaction times. Nearly all the sensor data must be processed with extremely low latency because there is a possibility that any sensor input could correspond to a critical situation (e.g., a human on the road).

Reconfigurability: Automobiles are typically on the road for ~15 years, but cell phones and laptops typically have much shorter lifespans. This introduces an unusual requirement for reconfigurability: the ability of technology to be upgraded while still resident on the vehicle. Artificial intelligence (AI) and machine learning (ML) algorithms are the key enablers for vision, classification, learning, and prediction tasks. However, the field of algorithms research is highly dynamic: algorithms change frequently. This poses a significant challenge to hardware and chip designs. High performance and energy efficiency are typically achieved with special-purpose application-specific integrated circuits (ASICs). However, a new algorithm may not run efficiently on an ASIC that was optimized for a previous algorithm; a new ASIC design may be needed, which in turn leads to a high cost of ownership. Since vehicles are owned by consumers for many years, the technology must be easily upgradable with the existing computing system in-vehicle. At the same time, reconfigurability can preserve data easily in-vehicle and mitigate other safety and security risks associated with computing hardware service during the lifecycle of a vehicle.

Low Volumes: The relatively modest volume for early AV market introduction may pose another challenge for the cost-effective chip manufacturing process. A high priority for any R&D in materials, devices, and circuits is to remain compatible with current semiconductor foundry processes to leverage the existing semiconductor manufacturing infrastructure. This nonetheless leaves room for innovation. Technology development to enable compatibility with prevailing manufacturing technologies will be especially important for Thrusts 2 and 3.

Safety and Predictable Degradation: Safety is the number one filter any AV technology must pass through to find its way onto an AV. Safety encompasses functional safety, availability, reliability, performance, and predictable degradation, and thus overlaps with some of the other issues outlined above. While R&D has to focus on improving reliability in harsh environments, it is equally important to understand chip lifetime, and when performance and reliability will begin to

falter. Especially in AV applications, such a prediction becomes important because a potentially failing hardware being deployed on the roads could pose a public danger. In addition, software quality engineering and the system resilience to software faults will be an important R&D area.

Privacy: AVs are likely to eventually be immersed in a matrix of highly interconnected computing nodes, which may include neighboring AVs, data centers, live traffic monitoring stations, weather stations, logistics and fleet control, etc. In such a highly interconnected environment, there will be significant sharing of data, which makes privacy of prime importance. Computing chips designed for AV applications will need to incorporate hardware or circuits that are dedicated to maintaining data integrity, for instance, via encryption.

Taking into consideration the above issues, some of the identified TA-I (Chips: Materials, Device, and Circuits) R&D challenges are:

1. **Discovering new, or improved, materials and processing techniques for increased thermal/mechanical/radiation robustness for automotive environments for the life of a vehicle (~15 years).**
2. **Developing low latency and low power (< 300 W) on-board computing circuits, such as "in-memory" computing hardware, where memory and logic/computing are integrated.**
3. **Creating computing devices and circuits that offer reconfigurability (e.g., in response to a new algorithm or learning from road conditions).**
4. **Integration of novel materials, devices, and circuits into existing manufacturing technologies and tools.**

4.1.1. *New or Improved Materials and Processing for Increased Thermal/ Mechanical/Radiation Robustness*

With computation needing to be placed on the AV itself, there arise significant challenges in the robustness needed for the chip technology. This includes chip survivability in thermally cycling conditions (constantly changing hot and cold environments), shock resistance, and stability under cosmic ray bombardment during high-altitude driving. These are all challenging environmental concerns needing to be solved in the near-term if AVs are to enjoy widespread deployment. Although there are already many semiconductor components in current vehicles, what is novel in the AV application is having to meet more severe SWaP constraints while also operating safely in the harshest terrestrial conditions in mass-produced retail vehicles.

Potential research in this area includes failure points and interconnects. Failure points typically are found at interfaces, so increased robustness is required to thermal/mechanical/radiation stress. This could be addressed through new treatments and materials, particularly those compatible with existing processes. For interconnects, new and/or augmented interconnects are needed for the automotive environment (e.g., integrated circuit (IC) wiring, packaging interposers, and heterogeneous integration technologies).

4.1.2. *Low Latency and Low Power Devices and Circuits*

Potential research areas within this R&D Challenge area include enabling non-von Neumann architectures (removing the separation between data storage and compute), which can lead to

significant energy and latency savings by minimizing on-chip data movement that dominates these metrics. It is also important to develop memory technologies (Flash, ferroelectric-based, resistive-based etc.) that are especially automotive-compatible (e.g., ability to withstand harsh temperature swings and weather). Development of low power/energy circuits must also consider the harsh environmental changes mentioned above, which lead to issues with drift, instabilities, variabilities and drops in precision. Analog circuits and asynchronous event-driven circuits deserve special attention owing to their promise of low power and latency.

4.1.3. *New Computing Circuits and Devices for Reconfigurability and High Performance*

In the immediate future, development of low-cost, high-speed Field Programmable Gate Arrays (FPGAs) is necessary to enable reconfigurability. Material-, device-, and circuit-level research to achieve such high-performance FPGAs are worth the investment. Furthermore, devices and circuits for Coarse-Grained Reconfigurable Architectures (CGRA) deserve attention since CGRAs hold promise for supporting a variety of autonomous driving applications. Hardware reconfigurability is important, as the hardware needs to be compatible with advanced and developing algorithms (e.g., computer vision) and efficient computing models (e.g., cellular networks) that can enable software-level generality while maintaining the best possible performance of the underlying hardware. Hardware compatibility with brain-inspired or neuromorphic computing approaches will help promote reconfigurability, quick learning and low computational energy.

4.1.4. *Integration of Novel Materials, Devices, and Circuits into Existing Manufacturing Technologies and Tools.*

Improving energy-efficiency in computing is of interest across all industries, and longer-term R&D is needed as efficiency improvements through scaling down transistors are decreasingly effective. Thus, there is interest in new materials, devices, and circuits for “beyond Moore” and “beyond CMOS” technologies. Such approaches explore alternative technologies such as Tunnel FETs (TFETs), negative capacitance FETs (NC-FETs) or carbon-based alternatives, for example. Also promising are new materials, devices, and circuits that enable more efficient architectures beyond the von Neumann layout. This includes the development of non-volatile memory technologies that can increase memory densities. Examples include Resistive Random Access Memory (ReRAM), Ferroelectric RAM (FeRAM), Phase-change memory (PCM) and related technologies. Importantly, such novel devices allow the coupling of stored data with compute operations, enabling in-memory or near-memory computing circuits and architectures to be developed. These architectures can address many of the specific machine vision, signal processing, and real-time decision-making compute challenges in autonomous vehicles.

Figure 4 lists the TA-I R&D problems in a simplified 2-axis format to convey the timing (near-term, long-term) associated with solving these R&D problems, and likely impact (high impact, very high impact) of the solutions on realizing widespread highly automated AVs. This format is admittedly simplistic in that is does not capture all the dimensions of the AV problem, such as investment costs, impact on other industries and risk, just to name several. For these 2-axis plots, “Near Term” investment means in the timeframe 2025 – 2030; “Longer Term” investment means in the timeframe 2030 – 2035. The impact scale is somewhat subjective. “High Impact” conveys very

important R&D that is needed for AV development. "Very High Impact" conveys R&D that can substantially change the direction of AV technical development.

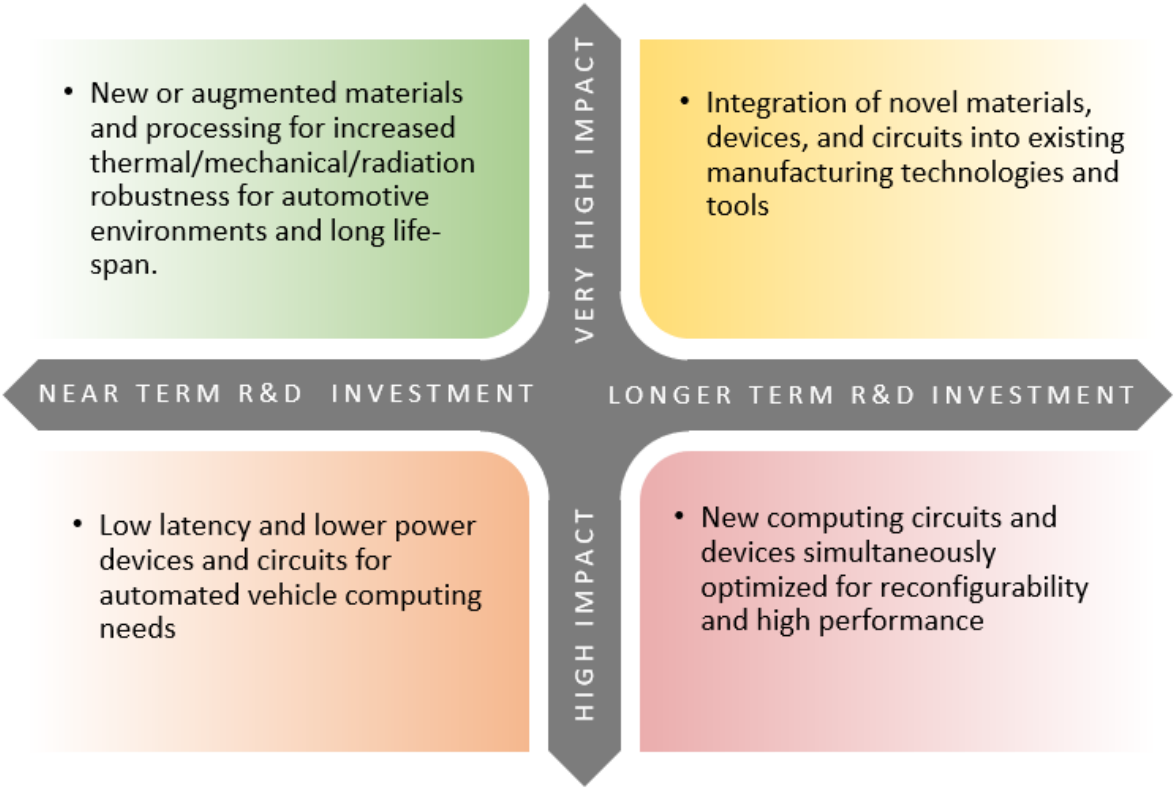

Figure 4. *R&D Challenges for TA-I - Chips: Materials, Devices, and Circuits.*

4.2. Technical Area II: Chips: Architecture, Safety, and Security

TA-II focuses on the computer architecture for energy-efficient AV computation, with emphasis on safety and security while limiting total power to 300 W. The computational throughput needed for highly automated driving and the technology needed to compute with high energy efficiency begins with the chip. However, the chips that enable AV are the physical instantiation of a system based around an architecture that provides not only a structure but also an anchor point for software and for communication within and outside the system itself. In addition, this architecture needs to meet energy consumption, performance, and in-vehicle footprint/weight constraints while also supporting safety and security.

We assume that the vast majority (and all safety-critical) of the required computational technology resides on the vehicle and consider in TA-II the many aspects of the computer architecture that touch on AV operation, including data throughput, memory and the security of the computational activity during vehicle operation. A key computing problem is handling the tasks of "computer vision" and "machine learning," which despite the terminology is a different way of "seeing" from human vision.

Passenger and pedestrian safety are paramount, and computation for AV requires a high degree of certainty on calculational results, an area known as functional safety for digital logic. Societal safety is also involved in a broad way, and questions of system security naturally arise.

The R&D challenges identified in TA-II follow.

1. **Exploring the optimized use of distributed, heterogeneous multiprocessor systems (such as CPUs, GPUs, and neural accelerators) to support the algorithms needed for EECAV.**
2. **Determining the type of on-board network/interconnect strategies that optimizes computational energy efficiency.**
3. **Developing improved memory (addressable and storage) and bandwidth in support of AV and determining where these are located within the system.**
4. **Identifying when (and if) computational "demand" starts to require consideration of "off-vehicle" computation.**

4.2.1. *Distributed, Heterogenous Multiprocessor System Architectures to Support EECAV Algorithms*

Future AVs will require distributed computing systems consisting of smart sensors that process incoming data locally as well as centralized vehicle computers which perform data fusion, perception and navigation. These systems will be needed to support advanced autonomy algorithms for AVs, and will connect a range of different computing processors, including CPUs, GPUs, NPUs, and other special purpose accelerators. Research is needed to better define, develop, and optimize these heterogeneous system architectures, while meeting all of the AV system constraints. The key constraints that must be considered follow.

1. Chip-level functional safety cannot be compromised or downgraded.
2. High performance and throughput must be obtained while meeting the computing system power ceiling of about 300W, which creates a requirement for high energy efficiency.
3. The computing system must meet automotive reliability, robustness, resilience, and fault tolerance requirements. Furthermore, due to the complexity of these computing systems, the security implications and possible new vulnerabilities must be understood and, where possible, mitigated.
4. Reprogrammability, upgradeability, reusability across multiple vehicles is desired.

These constraints are important not just for the processing system, but also for communication networks and memory systems, as discussed below.

4.2.2. *Network and Interconnectivity Architecture for Energy-Efficient AVs*

As described in *Section 4.2.1*, new architectures will be defined to connect heterogenous processors. As such, it will be necessary to define the networks that will connect these chips. Research is needed to understand the bandwidth and latency requirements both at the chip-to-chip level and from the car to the infrastructure, as well as protocols that might meet those

requirements. We must understand both where existing protocols can be leveraged, and where there are gaps that will require new communication protocols to be implemented.

The communication network must satisfy automotive constraints. First, functional safety of the system cannot be compromised by the interconnection network. In defining new connectivity, we must understand the possible failure modes, and consider when the system must fail safe versus fail operationally. Furthermore, the communications cannot weaken the reliability or fault tolerance of the entire system. Finally, new communication protocols and network architectures may introduce new security concerns. These must be identified and paths to mitigation should be found.

4.2.3. *Defining Memory and Storage Needs for AVs*

With the new heterogeneous computing multiprocessor architectures discussed above, memory needs will change and bandwidth and latency both need to improve to support new processing requirements. More memory and storage will certainly be needed to support new algorithms such as deep neural networks, but exactly where in the system the memory will physically reside needs to be investigated. Furthermore, questions such as the requirement of coherence across parts of the system should be understood, as these decisions directly affect computing and network architectures, and require these three elements to be defined holistically.

New, higher performance memory technologies may be required to meet these new memory and latency requirements, such as High Bandwidth Memory and emerging Storage Class Memory solutions. Approaches such as compute in- or near-memory and non-Von Neumann approaches are also promising methods to achieve high efficiency computing by greatly reducing data movement. These techniques should be considered and their fit in the overall heterogeneous architecture should be explored. Again, as with the previous two research areas, functional safety, reliability, and security implications of these new memory technologies must be considered.

4.2.4. *Identifying When (and if) Computational "Demand" Starts to Require Consideration of "Off-vehicle" Computation*

As described previously in Scope (*Section 3.1*), the Roadmap Team decided that having all of the computational capacity residing on the vehicle established a boundary condition for this work. However, it would be important to understand when significant "off-vehicle" computation would be necessary. Other transportation systems (e.g., air and rail) operate with centralized control points rather than as a collection of independent autonomous objects. These centralized systems require significantly less compute than decentralized systems but come with a corresponding loss of autonomy – airplanes only take off when air traffic control says they can, while cars are free to travel unless told they can't, by a stoplight for example. In addition to the existing cultural and infrastructure expectations of individual vehicle autonomy, an autonomous vehicle that requires connection to the outside world to operate has functional and security failure points that a vehicle with optional connectivity does not have, and likely would not be compatible with human-driven cars. As a result, mandated connectivity and centralized control are unlikely in the first generations of autonomous vehicles. However, autonomous driving is known to have "long tail" properties, which may make offloading some rare and challenging operations better than

attempting to perform them on-vehicle, especially if other vehicles can supply their data about the same situation ("is that a tire on a dark road, or a bump in the pavement?"). Some complex situations (e.g., intersections of high-volume high-speed routes) or safety challenges (dense fog on a freeway) may have more efficient solutions where local infrastructure and information from other vehicles contributes to the decision-making process. Identifying such situations and determining the likelihood that they will result in a significant amount of off-vehicle computation, along with an expected timeline, is an important research consideration.

Figure 5 lists these TA-II R&D problems to convey the required timing and likely impact of solving these technical problems.

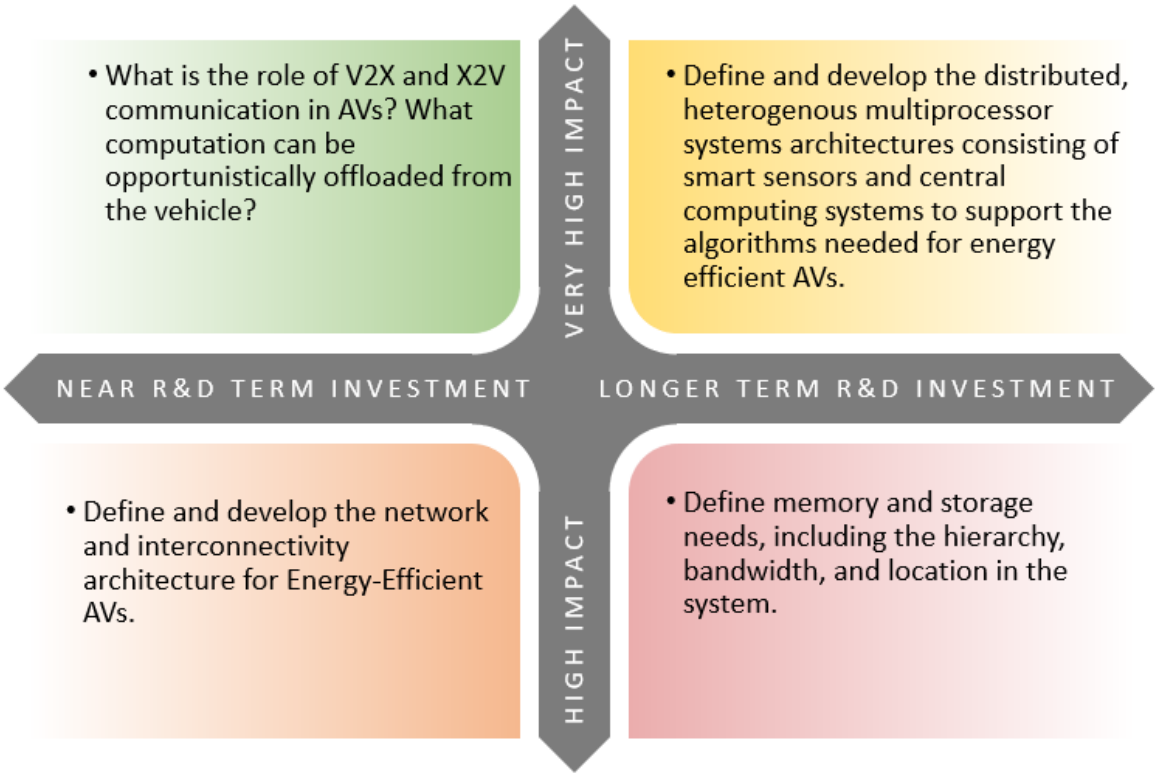

Figure 5. *R&D challenges for TA-II - Architecture, Safety, and Security.*

4.3. Technical Area III: Algorithms and Data Management

Technical Area III (TA-III) identifies algorithmic and software challenges that, if solved, would lead to reduced computational demand and greater functionality via more optimal analysis of data from cameras, light-detection and ranging (LiDAR) systems, as well as other sensor data sources. Energy-efficient computation is supported by appropriate chips and architectures. However, improved algorithms and data management can reduce computational demand and provide new functionality, often without reducing decision quality, thus improving energy consumption, inference accuracy, throughput, reliability, and security.

From the perspective of algorithms, highly automated vehicles operate by iterative processes in which data are sensed, transformed using signal processing, possibly compressed, and ultimately used for inference to support navigational and other decisions, enacted via actuation and logged

for compliance. The types of algorithms and data management approaches used impact all stages of these processes. Moreover, algorithms impose requirements on other technical areas, e.g., computer architecture and the types of sensors used. They also determine the functionality and performance levels possible.

We seek to determine which algorithm and data management R&D areas are most important to enable safe, secure, and economically successful deployment of highly automated vehicles. The interplay of software and hardware requires co-design because estimating chip performance (e.g., multiply-accumulate (MAC) throughput and energy consumption) requires knowledge of algorithms and, vice versa, estimating algorithm performance (e.g., latency, inference throughput and accuracy) requires knowledge of chip hardware.

The R&D challenges for TA-III follow.

1. **Efficiency Optimization: for example, sparse sampling/processing of only important data, which may benefit from sensor fusion and near-sensor signal processing or by improving the algorithms for tasks on latency and energy consumption critical paths, such as dynamic object detection.**

2. **Co-Optimizing Algorithms and Implementation Platforms: this includes developing fault-tolerant machine learning techniques.**

3. **Data and Training: for example, methods to determine accuracy, such as testing with carefully selected validation data, potentially during on-line learning.**

4. **Managing Data Retention and Locations: design decisions will be subject to performance, memory, communication, privacy, and legal constraints.**

4.3.1. *Efficiency Optimization*

Many of the successes of recent deep learning approaches result from using grossly over-parameterized systems to support generalization by learning the properties of large datasets. This approach is generally computationally intensive, with the training and inference costs of state-of-the-art networks increasing over time. There has been some attention paid to efficiency as a first-order design concern in recent years, with work generally focusing on reducing the number of parameters or MAC operations. However, even efficient ML techniques often remain computationally intensive or sacrifice accuracy. We believe that substantial efficiency improvements remain to be realized via algorithmic improvements including carefully determined spatial and temporal sampling distributions, improvements in network design going beyond adjustments in layer and neuron counts and better understanding of the relationships between specific training samples and their influence on learned parameters.

Efficiency enhancements can be applied at several stages of the design process, including initial design, post-hoc modifications, and deployment mappings. Automated Machine Learning (AutoML) and Neural Architecture Search (NAS) use ML techniques to define new algorithms. These methods represent a new and growing sub-field with the promise of finding new, efficient neural network approaches, but models will need to be optimized with awareness of architectural constraints. Distillation (teaching an efficient algorithm to mimic a more resource-

intensive algorithm) and pruning/compression (removing or reducing the sizes of parameters or activations in a network) work to reduce computational demand without undermining accuracy. Lastly, as algorithms are mapped onto a particular hardware target, appropriate bit-precision and effective floating-point-to-fixed-point mappings can help to achieve compact, high-performance designs.

The ML techniques for perception, data-fusion, planning and control have shown great promise in achieving intelligent behavior. However, due to their "black-box" nature, they often have to go through a safety verification process. This process can be computationally expensive (as it needs to search over possible risks) but must be performed under strict latency constraints.

In a V2X framework, there are additional, unique challenges. Decentralized learning and federated learning will require improvements both in algorithmic performance and energy efficiency. Current methods for maintaining privacy, like Fully Homomorphic Encryption (FHE), are computationally expensive and would require considerable improvement for widespread adoption in AV applications.

We believe that the above methods warrant further R&D and are likely candidates for producing lean AI algorithms. However, other potentially promising directions have been omitted for brevity. The field is moving quickly; many of the efficiency enhancing ideas used for AV in the coming decade have yet to be discovered.

4.3.2. *Co-Optimizing Algorithms and Implementation Platforms*

AVs will require advances in both algorithms and the computational substrates supporting them. The majority of the computational load may be handled by Deep Learning AI algorithms, including convolutional neural networks, recurrent neural networks, multi-layer perceptrons, and their derivatives. These algorithms are well-suited for video, RADAR, and LiDAR processing necessary for vehicle perception. Advancements in convolutional neural networks such as capsule networks and attention-based approaches (e.g., transformers and squeeze-and-excitation networks) will likely foster necessary developments in capability, and recent advances in multi-layer perceptrons render them competitive, especially when efficiency is of primary importance.

Demands on computational substrates and advances in devices and circuits may require architectural changes. Both classical and convolutional neural networks and many other classes of ML techniques require numerous matrix-vector multiplies and thus MAC operations. However, newer methods are benefiting from other operations as well, such as matrix-matrix multiplication and table lookups. Additionally, Hidden Markov Models will also be important as these are often used for downstream processing, such as dynamic object detection and modeling.

Given the diversity of workloads and the need to accommodate future, currently unknown, workloads, counting MAC operations does not adequately determine architectural requirements. Memory access and communication properties are important, and caching may not solve memory latency problems. A key problem for many ML techniques is the need to rapidly access a very wide range of often unpredictable memory locations. AV systems should be evaluated based on their first-order optimization objectives: inference or learning latency, not best-case MAC throughput; average system power consumption for realistic workloads considering the use of power management states, not peak processor power consumption (except for power delivery

network design); and accuracy properly defined to account for the consequences of errors. The co-design of algorithms with hardware will be critical for success.

4.3.3. *Data and Training*

The character of training data is critical for AI systems. However, AV training faces special requirements for safety and trust. It is impractical (likely impossible) to capture all real-world traffic dynamics in a static dataset or simulator. Furthermore, environmental effects, wear, and age may affect deployed sensors and processors. Understanding the role and implications of training approaches and training data selection and augmentation are critical for successful AVs.

We will need methods for quantifying uncertainty in vehicle perception and estimating the completeness of training data. Bayesian methods, including stochastic or Bayesian neural networks, may help estimate epistemic uncertainty. Additionally, advances in computational learning theory may help inform when algorithms need to be retrained or updated, e.g., due to a training dataset distribution that no longer matches the current data distribution. Data will often not be independent and identically distributed. Online learning methods may help alleviate an initial data requirement by being able to learn "on the fly." However, online learning methods introduce greater challenges for safety, assuredness, and validation. It is currently unknown how to measure lost performance due to neglecting on-line/continuous learning. Computational learning theory may also be able to address this concern.

AVs may be able to capture data useful for updating AI models, particularly if the AV encounters unlikely "edge cases." However, in such scenarios, data privacy must be considered, and both technical and legal challenges exist. Given the challenges of obtaining representative datasets, simulation will play a role in training AV AI systems. Additionally, methods such as surrogate tasks and contrastive learning may help improve generalization, both from dataset to deployment and from simulation to the real-world.

4.3.4. *Managing Data Retention and Locations Distributed Algorithms and Data*

Near-sensor signal processing to transform and compress data may be needed and will influence algorithm design. It may be possible to eliminate standard signal processing stages and hardware components, feeding raw data into ML algorithms. Covariance shift (mismatches between ML testing and training data) resulting from changes to a signal processing pipeline may reduce accuracy in these cases. While some methods exist to address this challenge, it is not fully solved, particularly for the assured or trusted operation needed in CAVs. Sensor fusion will help to identify unimportant data. Data compression and compressive sensing (and variants in which inference accuracy, not reconstruction accuracy is optimized) will also be valuable. However, near-sensor signal processing and inference must avoid undermining system-level sensor fusion techniques.

When data and computational resources are distributed (as in the case of distributed algorithms), this introduces research challenges, especially in the case of AV tasks with hard real-time deadlines. It is very likely that algorithms earlier in the processing chain will be deep learning based (such as vehicle perception) whereas later algorithms may be rule-based, or at least amenable to analyzing and understanding the decision-making logic. This delineation is partly due to computational cost and task complexity and partly due to explainability requirements.

This chain of processes complicates meeting real-time requirements. For example, dynamic object detection is computationally intensive and generally starts after the perception stage is finished. Many of today's tools for estimating worst-case latencies are pessimistic and would benefit from optimization.

Determining which data should be retained, in what form, and where will be a research challenge, as there will be conflicting objectives including efficiency, latency, accuracy, privacy, storage cost, and security.

The R&D challenges for TA-III are shown in *Figure 6* with their impact and timing indicated.

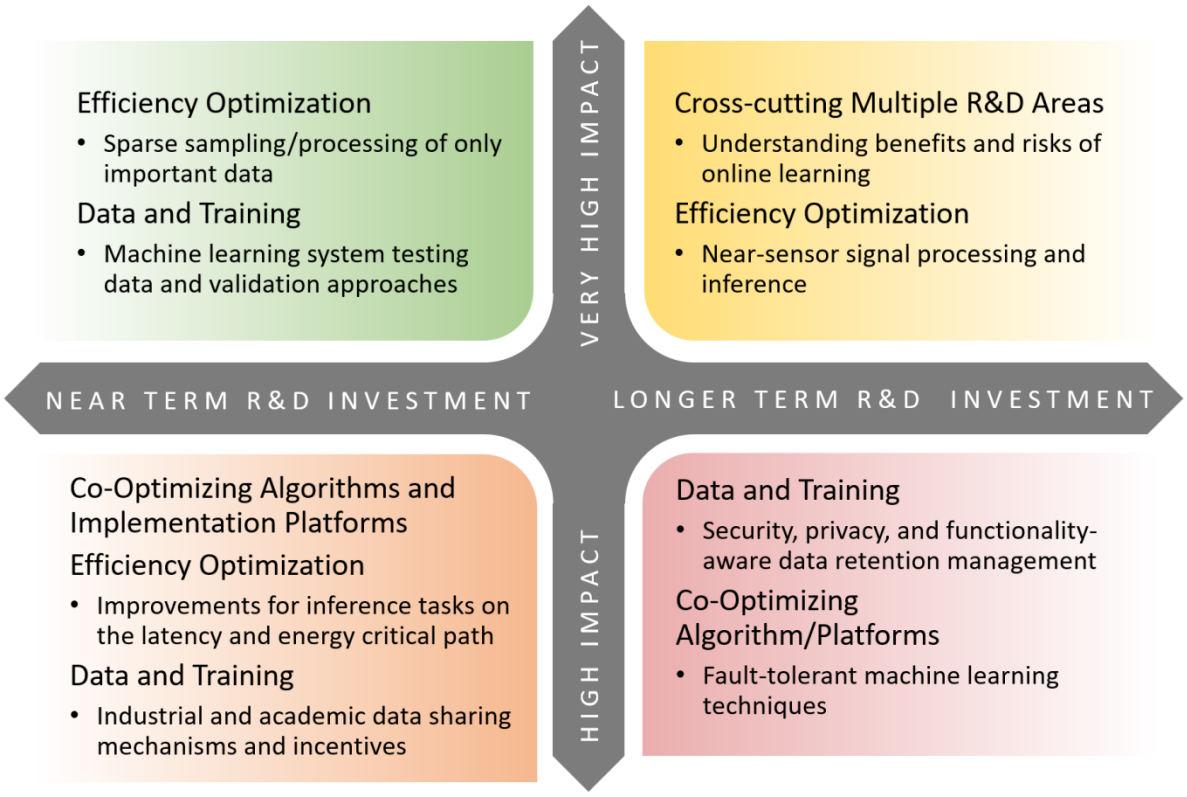

Figure 6. *R&D Challenges for TA-III - Algorithm and Data Management*

4.4. Technical Area IV: Sensors Data Interface

Technical Area IV (TA-IV) is concerned with the links from the external physical world to AV computers, and among physically distributed computers in a vehicle. The computational requirements are intimately connected to the nature of the sensors and their interfaces.

The data networks and computers used for automated driving are generally adapted from data centers where best-effort performance and uniform bandwidth support an ever-changing workload. However, AV workloads bear little resemblance to a data center workload due to fail-operational requirements and real-time requirements. The 15+ year service life of AVs means diagnostic tools and replacement parts must have a high degree of backward compatibility.

Sensor data dominates the workload in an AV system, and the data may be processed locally inside the sensor as well as duplicated to several destinations to distribute the processing work.

Sensors themselves receive very little data, so those interconnects can be asymmetric. Indeed, a case can be made for several kinds of interconnects in-vehicle: controller area network (CAN) for low bandwidth latency sensitive sensors and actuators (e.g., wheel rotation sensor or brake actuator), Ethernet or Peripheral Component Interconnect express (PCIe) for mixed criticality systems, and point-to-point serializer/deserializer (SerDes) for asymmetric high bandwidth sensors (e.g., cameras).

Topics that were examined included how "smart" sensors influence interconnects, fragmentation of the programming model, and generally how sensors and computers may be connected in-vehicle.

The R&D challenges for TA-IV follow.

1. **Evaluating tradeoffs between smart sensors and central computing.**
2. **Data versus task migration for dynamic power management.**
3. **Exploiting asymmetric bandwidth utilization of networks to improve energy efficiency.**
4. **Determining the R&D needed to anticipate advances in sensors and computers over a long (~15 year) vehicle lifespan to maintain forward and backward capability.**

4.4.1. *Tradeoffs Between Smart Sensors and Central Computing: How Smart Should a Sensor Be?*

The decreasing cost of computational capabilities means sensors that historically included only a detector now often include some amount of computing capability as well. The tradeoffs and long-term implications of in-sensor processing are multi-dimensional and not constrained to the sensor. For example, data bottlenecks arising from insufficient network bandwidth can be remedied by performing data reduction in the sensor, but the in-sensor data reduction might limit what algorithms can be applied. Distributed processing may add overhead, component cost, or complexity not present with central processing, but it may facilitate functional safety by replicating data and computations for fail-operational resiliency. Thus, the additional costs are offset by the savings of not implementing functional safety separately.

This is a high priority for research funding support. Making sensors "smarter" implies moving computation close to the sensor, with many implications for the required computational and energy efficiency. For example, if data processing at the sensor produces a reduction of the data (e.g., converting an 8 MB image to a 400 B object list), significant energy and cost benefits may be had by replacing a high bandwidth interconnect between the sensor and processor with reduced-bandwidth, inexpensive, energy-efficient interconnect. Examining how to do this could have a big impact on energy demands and the needed energy efficiency.

In addition, current approaches to functional safety require redundancy. The distributed nature of smart sensors may help support functional safety requirements more efficiently than fully redundant modules. The extent to which this is possible needs research.

4.4.2. *Data versus Task Migration for Dynamic Power Management*

Distributed and heterogeneous systems offer choices about where computation is performed and data are stored: data can be moved from storage to where a process is executing, or the

process can move to execute near to where the data are stored. Tradeoffs include the processor types, energy and time costs for moving data and tasks, bandwidth consumed, types and amount of memory, etc.

Dynamic task migration capabilities can also be used to implement functional safety, such as moving work and/or data from a failed unit to another device. Potential synergies may blur the line between load balancing and fail-operational safety without adding complexity.

4.4.3. *Exploiting Asymmetric Bandwidth Utilization of Networks to Improve Energy Efficiency*

Network technologies typically provide symmetric transmit and receive bandwidth, a reasonable design when the use case for the network is not known in advance or the topology changes over time. However, in-vehicle computing systems have the benefit of changing slowly, if ever, and have well-defined use cases that change little over time. Specifically, high resolution cameras primarily send data and only need to receive control commands – bandwidth utilization may differ by six orders of magnitude or more between transmit and receive, creating the opportunity for significant savings by making the network asymmetric. Likewise, a wheel rotation sensor produces data but does not consume it, meaning it is over-provisioned with respect to received bandwidth.

Possible approaches may include physically distinct transmit and receive interfaces operating at different speeds or building upon existing low-power mode when there is no activity, dynamically shutting down an idle link as opposed to provisioning a smaller network.

4.4.4. *Advancements in Sensors and Computers over a Long (~15 year) Vehicle Lifespan to Maintain Forward and Backward Compatibility*

Despite best efforts to maintain compatibility of software and hardware over time, the rapid rate of innovation means compatibility is often broken before the device itself breaks. Over the 15+ year lifespan of a vehicle, technology will evolve so that replacement components may be much more capable and different in nature than the devices they replace, changing the economics of dynamic load balancing. Research is needed to develop best practices for future-proofing sensors, computers, and their interconnects.

Security concerns which are difficult to predict may also limit backward compatibility or prevent upgrades.

The R&D challenges for TA-IV are shown in *Figure 7* with their impact and timing indicated.

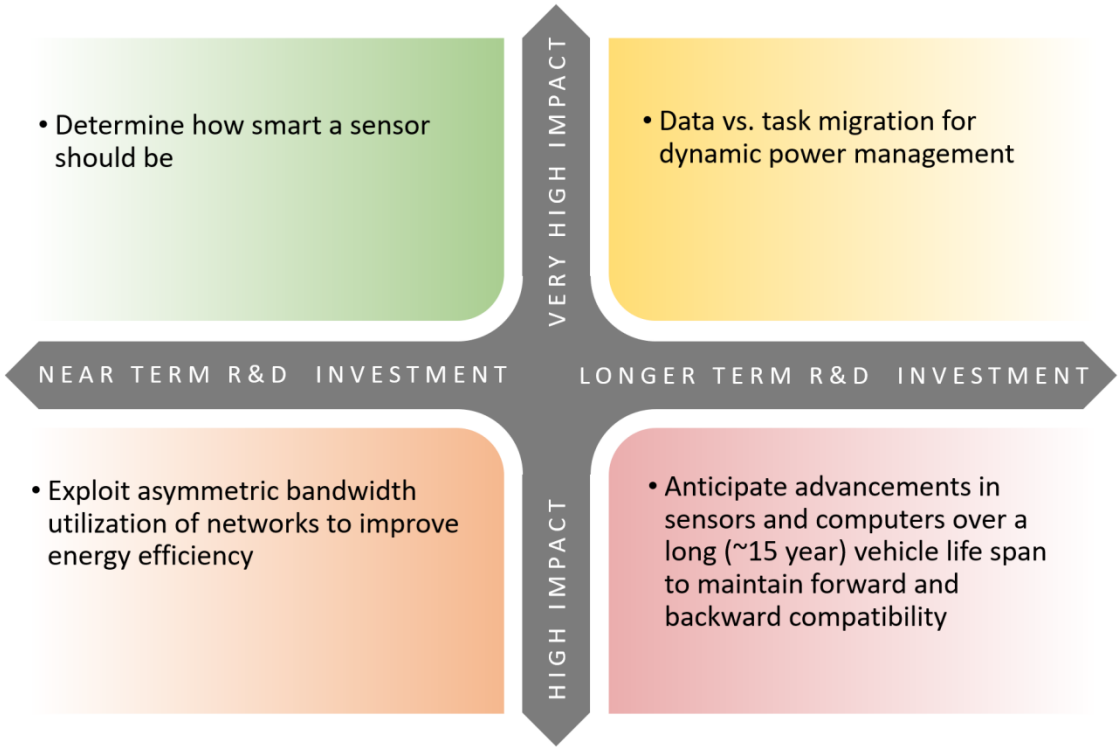

Figure 7. *R&D Challenges for TA-IV - Sensors Data Interface*

5. DIFFICULTIES IN QUANTIFYING COMPUTING PERFORMANCE IMPROVEMENT FOR FUTURE AV

While R&D problems are identified in the four technical areas, we have not specified quantitatively how much improvement is needed in any given technical area. Although we have set a limit for the total power budget of 300 W for onboard computing, we have not stated what the computational capacity target would be within this power budget to enable AV operation in 2035-2045 timeline, nor have we stated the needed computational energy efficiency improvement over the current commercial silicon CMOS technology. The Roadmap Team spent considerable time pondering this difficulty of specifying even in a semi-quantitative sense the needed levels of improvement across the four Technical Areas. The sources of the difficulty are at least two-fold. First, it is hard to define "what the AV is doing", and hence difficult to specify the amount of computation ultimately needed to support the AV capability. Second, there is uncertainty about how to best characterize (or measure) AV computational capacity and energy efficiency, and as a result, it is difficult to quantify the performance target.

Regarding the first problem, "what the AV is doing", meaningful definition of AV operational capability presents a significant challenge. The Society of Automotive Engineers (SAE) has defined Levels of Automation on a scale of 0 to 5, where Level 0 is a traditional, unautomated vehicle, and Level 5 represents ideal (full) automation, operating without human driver intervention in all situations.[8] While these Levels provide a useful high-level description of the types of AVs that

[8] https://www.sae.org/blog/sae-j3016-update

could exist, it fails to define the specific tasks or what exactly the AVs will be doing at different levels, which clearly will drive the needed computational capacity and power requirements. There needs to be developed a method for stating in a technically sound way what the targeted AV operational level is and what the assumptions are for the prevailing ODD in order to connect computational capacity and efficiency to "what the AV is doing."

The second problem involves the key performance metrics with regard to on board computing performance and energy efficiency. In order to be able to quantify the computing performance for future AVs that is associated with a (hopefully understood) AV capability, we will need to define commonly acceptable metrics to measure energy efficiency for different computing techniques. The computing hardware community typically benchmarks the energy efficiency of a processor, such as a CPU, GPU, FGPA, or special purpose accelerator using operations per second per watt (OPS/W, often in tera operation per second per watt, or TOPS/W). The numerator, operations per second, is a computing performance metric and OPS/W specifies the performance possible with a given power. While this is a useful benchmark, the result varies significantly for different tasks that the processor is being used for. Hence, in order for TOPS/W to be useful, a comparison must be made across hardware running the same algorithm, at the same fidelity, accomplishing the same task. Factors which affect TOPS/W include the algorithm, architecture, and precision of the operation.

In the case of comparing more diverse hardware, which can use different fundamental operations, it is possible to use a higher-level metric, such as inferences per second per watt. Rather than the lower level mathematical operation, we are now comparing the energy to complete an entire task (such as recognizing an image), regardless of the algorithm and architecture. This enables us to evaluate the efficiency of different approaches such as spiking and non-spiking neural systems that are designed to accomplish the same task.

Performance per watt is another possible method to benchmark various individual processors of an AV computing sub-system. However, the overall power required by the AV's computer is going to be the sum from each processor performing its own disparate tasks. It is perhaps possible that an efficiency can be given as an average performance per watt for this collective. However, given the range of processor hardware systems within an AV, it is not clear that the performance per watt average is a truly meaningful metric. Rather, it perhaps is most relevant to benchmark the power required for the AV to achieve a certain level of automation.

Perhaps a more specific, relevant metric is the number of miles per "disengagement" or collision of an AV. A disengagement is a situation where the human driver must take control of the vehicle for various reasons which range in severity, up to the point of avoiding an accident. SAE Level 5 implies that the AV never disengages or needs human intervention (i.e., there is no steering wheel in the AV) in any ODD, which may be difficult to ever fully achieve or verify. Perhaps a more achievable scenario might be that the AV can navigate surface streets, freeways, driveways, and parking in a metro area in a range of weathers, with 10-100 million miles between required disengagements. Miles per reportable disengagement have been improving for major manufacturers each year.[9] A reportable disengagement meeting specified conditions requires

[9] https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/

reporting to the California Department of Motor Vehicles, which maintains a record of these occurrences. While miles per disengagement may be a reasonable metric to track progress in the AV capabilities, it is challenging to model the miles per disengagement as a function of computing capability or efficiency. In order to better understand the relationship between miles per disengagement and computing power performance, both components will need to be broken down into fine granularity to create detailed models. Such models would improve understanding how the computing resources, algorithms, and environmental factors influence the disengagements.

In summary, more work is needed in order to specify the needed R&D progress in the four technical areas on more a quantitative scale.

6. SUMMARY

We report here a Roadmap Outline that identifies in a technically unbiased way the R&D challenges that must be overcome for the realization of highly automated driving in retail vehicles with low power consumption and high computational performance. The purposes of the high-level Roadmap Outline are twofold: to provide guidance on future public and private funding in this arena, and to stimulate a more detailed R&D Roadmap as an important next step in developing safe and reliable AV systems.

The identified R&D problems were developed under the assumption that the computational capacity for all latency- and safety-critical tasks resides on the vehicle, and that the "all-in" total electrical power devoted to computation would be 300 W for a commercially viable vehicle. We developed and considered three inter-related 10-year timelines for the AV problem: an R&D Timeline extending from 2025-2035, a Chip Commercialization Timeline from 2030-2040, and an OEM Implementation Timeline from 2035-2045. Within this AV technology development landscape, four technical areas (TAs) were identified for R&D investment: I. Chips: Materials, Device and Circuits; II. Chips: Architecture, Safety and Security; III. Algorithms and Data Management and IV. Sensors Data Interface.

Specific R&D problems were identified within each technical area, with assessments given for their timing, impact and how the problems relate to each other. In TA-I (Chips: Materials, Device and Circuits), it was found that discovery in new chip materials is needed to meet the very demanding thermal and mechanical environment of vehicle, with the materials processes capable of being integrated into the existing manufacturing technologies. The AV application demands significant reductions in chip latency and power consumption, with the long vehicle life demanding reconfigurability of the computing circuits. In TA-II (Chips: Architecture, Safety and Security), R&D is needed to explore the optimized use of distributed, heterogeneous multiprocessor systems to support AV algorithms, and to determine the types of on-board network/interconnect strategies that optimize computational energy efficiency. Furthermore, improving on-board memory and bandwidth will be critical for AV, and identifying when (if) computational "demand" will require moving some of the computation "off-vehicle." Safety needs to be developed in all systems early in the R&D development. For TA-III (Algorithms and Data Management), the R&D opportunities centered on improving the algorithmic efficiency "writ-large," for example, determining compact subsets of important data enabling efficient and accurate decision making, co-optimizing algorithms and implementation platforms to improve

fault tolerance performance, understanding how the algorithms can be trained to improve inference accuracy, and producing algorithms that optimize data motion and retention. Finally, in TA-IV R&D topics were identified to answer key questions such as how computation needs to be distributed among smart sensors and a central computer, along with R&D that improves the lifespan, composability and functional safety of the computer-sensors system. It is important that R&D be conducted on how the sensor-computer system can be made forward and backward compatible for the long (~15 years) life of the vehicle.

Unlike the computing requirements for personal computers, data centers, and high performance computer applications, computing for the AV application poses the following unique challenges: a demanding operating environment, safety and security criticality, and a large impact on society and long-term deployment (~15 years) in a retail vehicle. Since many of these AV issues are inter-related, the R&D activities within these four technical areas need to be co-designed and conducted in a holistic way to successfully meet the stiff technical challenge of developing energy efficiency computing that enables highly automated vehicles. To put computing performance improvement on a quantitative scale, more work is needed to better define “what the AV is doing”, understand quantitatively how improvements in all four technical areas affect one another, and converge on a commonly acceptable metrics to benchmark the AV computational capacity and energy efficiency. Overall, there is urgency to develop the full roadmap of advanced computing for automated vehicles if the timelines set forth in this Roadmap Outline are to be realized.

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.