Article

Ultra-Low Power and High-Throughput SRAM Design to Enhance AI Computing Ability in Autonomous Vehicles

Youngbae Kim *, Shreyash Patel, Heekyung Kim, Nandakishor Yadav and Kyuwon Ken Choi

DA-Lab, Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, IL 60616, USA; spatel115@hawk.iit.edu (S.P.); hkim104@hawk.iit.edu (H.K.); nkyadav.vlsi@gmail.com (N.Y.); kchoi12@iit.edu (K.K.C.)
* Correspondence: ykim102@hawk.iit.edu

Abstract: Power consumption and data processing speed of integrated circuits (ICs) is an increasing concern in many emerging Artificial Intelligence (AI) applications, such as autonomous vehicles and Internet of Things (IoT). Existing state-of-the-art SRAM architectures for AI computing are highly accurate and can provide high throughput. However, these SRAMs have problems that they consume high power and occupy a large area to accommodate complex AI models. A carbon nanotube field-effect transistors (CNFET) device has been reported as a potential candidates for AI devices requiring ultra-low power and high-throughput due to their satisfactory carrier mobility and symmetrical, good subthreshold electrical performance. Based on the CNFET and FinFET device's electrical performance, we propose novel ultra-low power and high-throughput 8T SRAMs to circumvent the power and the throughput issues in Artificial Intelligent (AI) computation for autonomous vehicles. We propose two types of novel 8T SRAMs, P-Latch N-Access (PLNA) 8T SRAM structure and single-ended (SE) 8T SRAM structure, and compare the performance with existing state-of-the-art 8T SRAM architectures in terms of power consumption and speed. In the SRAM circuits of the FinFET and CNFET, higher tube and fin numbers lead to higher operating speed. However, the large number of tubes and fins can lead to larger area and more power consumption. Therefore, we optimize the area by reducing the number of tubes and fins without compromising the memory circuit speed and power. Most importantly, the decoupled reading and writing of our new SRAMs cell offers better low-power operation due to the stacking of device in the reading part, as well as achieving better readability and writability, while offering read Static Noise Margin (SNM) free because of isolated reading path, writing path, and greater pull up ratio. In addition, the proposed 8T SRAMs show even better performance in delay and power when we combine them with the collaborated voltage sense amplifier and independent read component. The proposed PLNA 8T SRAM can save 96%, while the proposed SE 8T SRAM saves around 99% in writing power consumption compared with the existing state-of-the-art 8T SRAM in FinFET model, as well as 99% for writing operation in CNFET model.

Keywords: CNFET (Carbon Nanotube Filed Effect Transistor); FinFET (Fin Field Effect Transistor); read SNM free; low power SRAM; 8-T SRAM; CNFET SRAM; FinFET SRAM

1. Introduction

The low power and high throughput issues are becoming increasingly important in Very large-scale integration (VLSI) system, microprocessor, and Static Random Access Memory (SRAM) designs for Artificial Intelligent (AI) application, such as in autonomous vehicles [1] and Internet of Things (IoT) [2]. The causes of these low power problems can be classified as follows. (1) In modern chip designs, most system-on-chips (SoCs) for IoT application incorporate a large amount of caches built by SRAM cells. (2) As the deep learning models for autonomous vehicles become more and more complex, the area of the chip for deep learning accordingly is getting bigger, and this large areas causes the power consumption problem of the integrated circuits (ICs). For these reasons,

many recent chip designers are looking for resolutions to achieve high power efficiency without compromising the chip speed and stability. The biggest problem with low power consumption of traditional chips, Metal Oxide Semiconductor FET (MOSFET), is that gate leakage current can no longer reduce the specific point size of the MOSFET. Because of this size limitation, the supply voltage is no longer lowered, so designers have been working hard to solve this limitation. To solve these MOSFET limitations, recently, many state-of-the-art SRAMs use Fin Field Effect Transistor (FinFET)s instead of MOSFETs. In addition to the traditional FinFET device, many companies are developing an improved FinFET type. However, while most state-of-the-art SRAM devices have changed from MOSFET to FinFET, the SRAM structure still uses traditional 6T (Figures 1 and 2) and is lagging behind. Developing a device is the easiest way to increase SRAM performance, but it has the disadvantage that it increases process cost and development cost. Therefore, to maximize SRAM performance, not only device development but also SRAM structure development must be carried out.
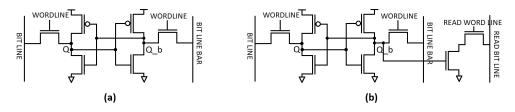

Figure 1. Conventional state-of-the-art SRAM cell structure: (**a**) 6T; and (**b**) 8T (Source [3]).

Therefore, we maximize low-power and high-speed effect by proposing a new SRAM structure as well as changing the device in this paper. Furthermore, in this paper, we maximize the performance for AI chip by using not only FinFET but also Carbon Nanotube Filed Effect Transistor (CNFET) device, a next-generation device that has not yet been used in state-of-the-art SRAM.

In using the FinFET device, because the fin number on the FinFET device limits the design space of the circuitry that requires the proper WL ratio transistor, especially the SRAM cell, the fin number must be taken into account to ensure stable read and write. Therefore, we analyze and apply the optimal number of fins for using FinFET in this paper. Since the large number of fins can make larger area and more power consumption, we optimize the area by reducing the number of fins without compromising the memory circuit speed and power. For low power and speed, this FinFET was first used by Intel Corporation and showed a significant reduction in dynamic power compared to conventional MOSFETs and is still being actively researched [4,5].

A Carbon Nanotube Filed Effect Transistor (CNFET) device has been reported as a potential candidate for AI devices requiring low power and high throughput due to their satisfactory carrier mobility and symmetrical, good subthreshold electrical performance.

	Company	Node(nm)	Area(um^2)	$V_{dd,1}$(V)	SNM_1(V)	$V_{dd,2}$(V)	SNM_2(V)
Planar Bulk MOSFET	Intel	32	0.171	~0.9	~0.2		
	TSMC	32	0.15	1.0	0.22	0.8	0.2
	IBM	32	0.157	0.9	0.213		
	UMC	28	0.124	0.9	0.179	0.7	0.144
	SAMSUNG	20		0.9	0.255		
	IBM	22	0.1	0.9	0.22	0.7	0.148
FinFETs	IBM	22	0.094 0.063	0.8 0.8	0.2 0.15	0.4 0.4	0.1 0.05
	TSMC	20	0.1	0.85 0.65		0.45	0.09
	Intel	22	0.092 0.108				

Figure 2. State-of-the-art SRAM performance.

The CNFET can be an alternative to solve these problems, which is now regarded as a next-generation transistor. Due to the power efficiency and resistance of CNFET, SRAM design using CNFET is very effective at low power and speed. This has the advantage of having excellent electrical properties in terms of high speed and low power. In addition, because the effective gate capacitance per unit width of CNTFET is more than twice that of p-MOSFET, compatibility with the high-k gate dielectric is certainly an asset to CNTFET.

In this paper, we propose a reliable low power high throughput SRAM that is useful for AI computation, using these advantages of FinFET and CNFET to secure these conventional SRAM shortcomings.

2. Memory Issue for AI Computation in Autonomous Vehicles

In recent years, AI has been extensively researched to benefit society, from voice recognition to autonomous vehicle [6] (e.g., Level 2 autonomous driving system based on AI technology of Tesla [7], Nvidia, Intel/Mobileye, Waymo, Cruise, and classical OEMs such as Toyota, VW, Daimler AG, etc.).

In addition, many challenges are underway to speed up the evolution of these AIs. In this paper, we focus on the optimization of hardware, especially memory, for AI. The most important point in researching optimized memory for these AIs is to identify the best memory type for a particular AI function and find the best way to integrate various memory solutions together. Figure 3 shows the power consumption rate of AI chip for Ford Fusion's autonomy system [8,9]. In this regard, AI faces two major memory limitations: speed and power efficiency.

Component	Structure	Maps	Added Weight	Increased Drag	Camera	Sonar	Radar	Small LiDAR	GPS_INS	DSRC	Compute
Total 100%	4%	3%	15%	10%	4%	2%	5%	5%	2%	9%	41%

Figure 3. Power consumption rate of AI chip for Ford Fusion's autonomy system.

The power demands of these AIs also make it an effort to extend AI beyond data centers (i.e., the cloud) where power is readily available, but it is not easy because it causes a variety of other issues (e.g., power consumption in network and security issues). Therefore, we propose an energy efficient memory SRAM with no speed loss to enable AI.

Figure 4 shows the memory usage of the core cluster Processing Element (PE) block during the Convolutional Neural Networks (CNN) process for AI [10,11]. Figure 5 shows the percentage of memory for each CNN model's compute. Since each PE block requires its own memory for I/O, this memory becomes a crucial part in terms of power consumption and speed. This is an area that must be addressed first and foremost in developing AI hardware architecture. Recently, Renesas announced 16-nm technology AI SRAM for autonomous vehicles and supported fast read times of 313 ps at 0.8 V [12,13].

Autonomous vehicles are undergoing a lot of research and testing for higher levels on public roads, and Level 2 has already been popularized in many vehicles. Many AI chips for autonomous vehicles not only focus on recognition accuracy and speed for fast object recognition, but power consumption is also evolving. However, solutions to optimize power consumption are still lagging behind the others.
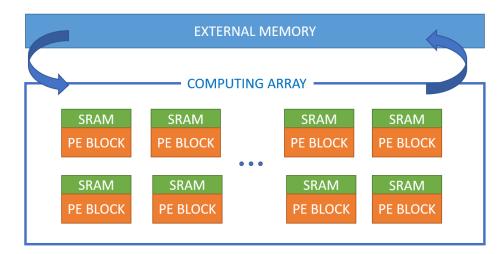
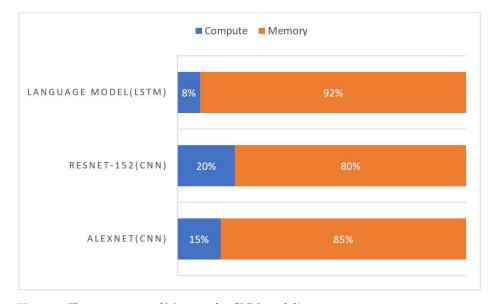**Figure 4.** AI SRAM Configuration of PE Array for CNN Computing.

Figure 5. The percentage of Memory for CNN model's compute.

Definitely, in future generations of the autonomous vehicles, especially Electric Vehicle (EV)s other than Internal Combustion Engine (ICE) vehicles, serious power consumption problems are expected because such power consumption issues can affect the range of EVs. In addition, even in an ICE vehicle, the power consumption issues can create fuel economy and battery space issues. For these reasons, many companies are investing in low-power and high-throughput AI chip development. For example, Nvidia has more than 370 automotive partners as the primary hardware supplier for autonomous driving solutions. With an 8 core CPU and 512 core graphics processing unit (GPU), the Xavier platform delivers 30 Trillion Operations Per Second (TOPS) and consumes 30 watts. The Pegasus platform, which is more suitable for autonomous driving solutions, can perform 320 TOPS with the addition of two Xavier chips and two GPUs and consumes 500 watts or 0.64 TOPS.

Intel is developing low-power chips optimized for self-driving cars, Tesla is developing its own low-power chips for Autopilot, and Qualcomm is building the necessary communication hardware with low power and efficiency in mind [14].

From this point of view, our proposed new SRAM aims to support future AI chip development by proposing a reliable low-power and high-throughput method for the memory that occupies the most power consumption in these AI chips for reducing the power consumption.

3. Background of New Devices Technology

In this section, we discuss the power consumption and performance limitations in the conventional MOSFET when it use to develop AI application. Further, the advantages of the FinFET and CNFET devices are discussed for the AI applications.

3.1. Limitations of MOSFET for AI Application

There are two types of MOSFET, the Depletion and Enhancement type MOSFETs use an electrical field created by a gate voltage to alter the flow of the charge carriers, through the semi conductive drain-source channel. The MOSFET cannot be further down scaled for the high speed AI computation due to fabrication limitations, increase in leakage current sources, and reduction in threshold voltage. This is the main reason that we do not use MOSFET and that it is being replaced by FinFET or CNFET in yjr future. In recent years, for down scaling of Complementary metal–oxide–semiconductor (CMOS) technology, many researchers have spent a lot of time reducing the thickness of the silicon dioxide gate dielectric. Despite these efforts, there are still many problems with the SiO2 thickness reduction due to the increased depletion of Poly-Si gate, the ingress of gate dopant into channel areas, and the high direct tunneling gate leakage current. These problems can cause serious problems in dielectric integrity, reliability, and power consumption. As we have well recognized, the increased gate leakage current in MOSFET is a problem to continue down scaling. The down scaling of MOSFETs to the nanometer regime leads to the short channel effects, which further degrade the system performance and reliability. At high frequencies, MOSFETs are not reliable, as switching takes place from PMOS to NMOS, which results in more power consumption, as both MOSFETs conduct briefly.

3.2. FinFET Technology for AI Application

Much research and development is being done to develop transistors with low power and high throughput for AI computation. One such development is called FinFET. FinFETs have the advantage of using a fin-like structure instead of the flattened design of the conventional CMOS, allowing engineers to build faster and more compact circuits and computer chips. The term 'FinFET' refers to a non-planar, double-gate transistor fabricated on an Silicon on Insulator (SOI) substrate based on a single-gate transistor design [15]. The gate is deposited to form the structure by surrounding the fin, as shown in Figures 6 and 7.

The main distinctive of FinFET is that the conducting channel is swaddled by a thin Si "fin", as shown in Figure 6, which forms the body of the device. The fin thickness determines the effective channel length of the device. The CMOS devices have a gate only on one side, but FinFET devices have three side gates, which is called multi-gate FinFET. FinFETs can be operated in two modes, namely Short-Gate (SG) mode and Independent-Gate (IG) mode, as shown in Figure 8a,b, respectively.
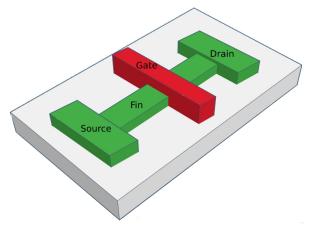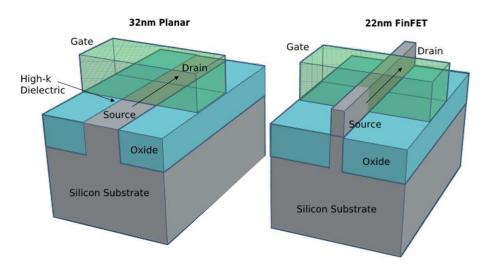

Figure 6. The Basic Structure of FinFET.

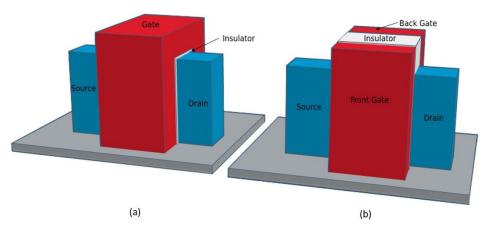**Figure 7.** Planar CMOS Structure vs. Non Planar FinFET Structure.

Figure 8. (**a**) Short-hate FinFET; and (**b**) independent-gate FinFET.

The number of gates that control the conduction channel could depend on the mode of operation, and the front and rear gate terminals are shortened in SG mode. In IG mode, the gate terminal is disconnected and two gates control the channel. The advantages of FinFET over MOSFETs are from the FinFET model structure consisting of the following regions with low doping silicon fin region: poly-silicon region, source and drain contact region, highly doped, and gate region oxide (SiO2).

The advantages of FinFET for AI application are as follows [16]:

- Maintain excellent control of channel effect in sub micron system and expandability of the transistor, thus small transistors may have greater intrinsic benefits and much lower off-state currents than conventional transistors
- Ability of matching behavior
- Low cost
- Technical maturity higher than plane Double-Gate (DG)
- Suppressed Short Channel Effect (SCE)

Based on these advantages, FinFET is reported as an excellent technology that provides better performance and reliability for AI computation.

3.3. CNFET Technology for AI Application

Carbon Nanotube Filed Effect Transistor (CNFET) devices have been reported as potential candidates for AI devices requiring low-power and high-throughput due to their satisfactory carrier mobility and symmetrical, good subthreshold electrical performance.

Single-wall carbon nanotube (CNT) refers to a tube formed by rolling a sheet of graphene. It can be a metal or semiconductor in accordance with the chiral vector (m, n) that indicates the direction in which the graphene sheet is rolled. The semiconductor type of carbon nanotubes is a promising high-performance channel material substitution because it is easier to control and has strong current density.

CNFET uses a set of parallel carbon nanotubes to construct a channel between the source and drain terminals compared to a bulk CMOS FET, as shown in Figure 9. The carbon nanotubes provide much larger driving current compared to the traditional channel inducted by the gate-body voltage. The threshold voltage of CNFET is mostly determined by the diameter of the carbon nanotubes, instead of multiple factors in bulk CMOS technology that make the circuits suffer from large indie subthreshold voltage variation [17,18], which are associated with the chiral vector defined as:

$$D_{CNT} = \frac{a}{\pi}\sqrt{m^2 + n^2 + mn} \tag{1}$$

$$V_{th} \approx \frac{E_g}{2e} = \frac{aV\pi}{\sqrt{3} \times qD_{CNT}} \tag{2}$$

q (q = energy gap) is equal to the unit electron charge, a is equal to the CNT atomic distance of 2.49, πV is equal to the carbon π to π bond energy of 3.033 eV in the rigid bonding model, e is the unit electron charge, and CNT D is a CNT diameter. The size of the CNFET can be easily adjusted by setting the number of tubes. Since n-type and p-type have the same carrier mobility, P-CNFET and N-CNFET with the same number of carbon nanotubes have the same strength characteristics.
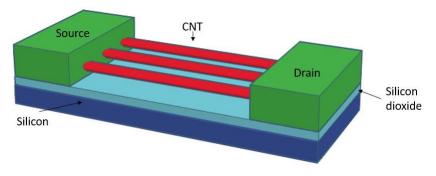

Figure 9. Structure of the CNFET.

Most single-walled nanotube (SWNT)s are close to 1 nm in diameter. This paper uses 1.5 nm and it can be billions of times longer. The structure of a SWNT can be represented by a complete cylinder wrapping a one-atom-thick graphite layer called graphene. The manner in which graphene sheets are wrapped is represented by a pair of indices (n, m), as shown in Figures 10 and 11. The integers n and m represent the number of unit vectors along the two directions of the honeycomb crystal lattice. If m = 0, as shown in Figure 11, the nanotubes are referred to as zigzag nanotubes. If m = n, the nanotubes are referred to as armchair nanotubes, as shown in Figure 10. Otherwise, they are called chirals.

Based on the electrical performance of the CNFET device, our proposed SRAM based on the satisfactory carrier mobility and good subthreshold electrical performance, has indicated a performance improvement of 99% reduction in power consumption and 97% improvement in delay compared to the state-of-the art SRAM with same condition. These power consumption and delay performance improvement results greatly affect the AI computation in the PE Block.
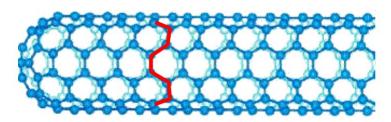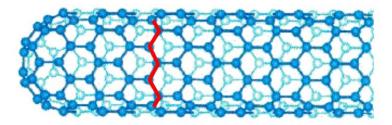
Figure 10. n = m (chiral angle = 30° tubes are called "armchair"). Adapted from Weisman, 2004, p. 24.

Figure 11. m = 0 (chiral angle = 0° tubes are called "zigzag"). Adapted from Weisman, 2004, p. 24.

4. Proposed SRAM Design for AI Application

The performance of write operation is considered very important for AI applications as it helps significantly improve the parallel write operations required by AI computations. Therefore, we propose two new structures to significantly improve these write performances while reducing power consumption and speed.

4.1. Decoupling the Read and Write Operation

The proposed new 8-T SRAMs are shown in Figures 12 and 13. It can be seen that the writing part and reading part are completely separate and have an independent structure. In the read part, you can see the advantage of providing lower power consumption because the device is stacked. However, there is an area penalty due to the fact that read and write are separated. However, as we can share the read part with several write parts in actual SRAM design in future fabrication, we can easily solve this area problem. This alternative SRAM cell design has the advantage of successfully coping with the difficulty of maintaining proper operation at high yield constraints in the low threshold operating region and providing characteristics for stable cache realization in Negative Temperature Coefficient (NTC)-based systems.
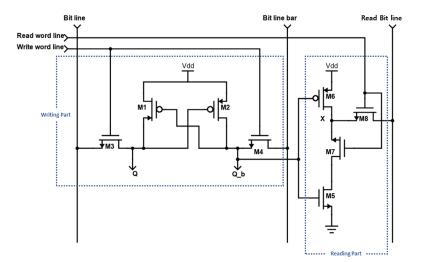

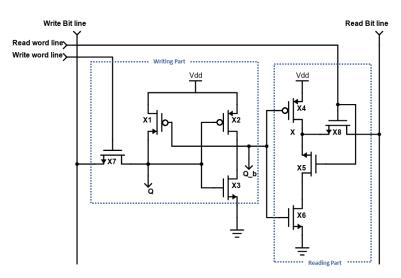**Figure 12.** Proposed PLNA 8T SRAM cell.

Figure 13. Proposed SE 8T SRAM cell.

4.2. P-Latch N-Access (PLNA) 8T SRAM Design

The biggest difference from the traditional SRAM cell is that we have completely separated the read and write parts to reduce switching power consumption, as shown in Figure 12.

The proposed 8T SRAM cell uses only two driver transistors M1 and M2. Thus, during a write operation, the circuit has the advantage of avoiding a strong or rapid current flowing between the latch and the access transistor. This has the effect of improving write speed and saving a lot of power consumption [19]. Although, in the reading operation, our suggestion provides low power by stacking the device on the reading part, there is a corresponding area penalty. However, this area penalty is not large because we can share this single reading part with multiple writing parts in a SRAM design to real field [20]. This new SRAM cell design successfully addresses the difficulty of maintaining proper operation at high yield constraints in the sub threshold operating region. The sizes of FinFET and CNFET of the SRAM are described in Table 1.

Table 1. Size of FinFET and CNFET in PLNA 8T SRAM.

No	TR	No. of Fins (FinFET)	No. of Tubes (CNFET)
1	M1	1	1
2	M2	1	1
3	M3	2	2
4	M4	2	2
5	M5	4	4
6	M6	1	1
7	M7	4	4
8	M8	2	2

4.3. Single-Ended (SE) 8T SRAM Design

To decrease the switching power consumption, which occupies the largest portion of power consumption, we eliminate the switching of the weak inverter. In our case, we have different read and write parts, as shown in Figure 13. In our design, we use two access transistors, X7 and X8. The pull-up transistor is made weaker than the access transistor, whereas the driver transistor is made stronger than the access transistor. This type of constraint is called read stability and write stability.

FinFET has fin-like structure, while CNFET has tubes. Therefore, the sizes of FinFET and CNFET are determined by number of fins and tubes they have, respectively. The large number of fins and tubes can make larger area and more power consumption. Therefore,

we optimize the number of fins and tubes without compromising the memory circuit speed and power. The comparison of sizes in FinFET and CNFET is shown in Table 2.

Table 2. Size of FinFET and CNFET in SE 8T SRAM.

No	TR	No. of Fins (FinFET)	No. of Tubes (CNFET)
1	X1	1	1
2	X2	1	1
3	X3	4	4
4	X4	1	1
5	X5	4	4
6	X6	4	4
7	X7	2	2
8	X8	2	2

5. Performance Verification

5.1. SRAM Operation

The proposed novel FinFET and CNFET SRAMs are compared with the conventional FinFET 8T SRAM [21,22]. All FinFETs and CNFETs used in the experiment were simulated at room temperature (27 °C) and VDD = 0.9 v in 32-nm technology. All transistors in SRAM cells use the minimum size adjustment shown in Tables 1 and 2 considering area, power, and speed. In addition, the size of the read access transistor is adjusted to $2\times$ minimum sizing in order to balance speed and power consumption. The sizes of the transistors used in conventional 8T SRAM cells are $1\times$, $2\times$, and $4\times$ minimum for the pull-up, access, and pull-down transistors, respectively. In the whole SRAM circuit for the experiment, M9 and X9 are scaled up to reduce the ground bounce effect. In a SRAM whole circuit, as shown in Figures 14 and 21, designed for an environment similar to an actual SRAM circuit in the real field, the produced SRAM whole circuit contains SRAM cells as well as peripherals that are as close to real as possible. The bit line and bit line bar are pre-discharged to the ground before the write operation, and the write enable signal is asserted at the same time as the start of the write operation. For the analysis of writing and reading, we used a pseudo-random sequence of 0110100.
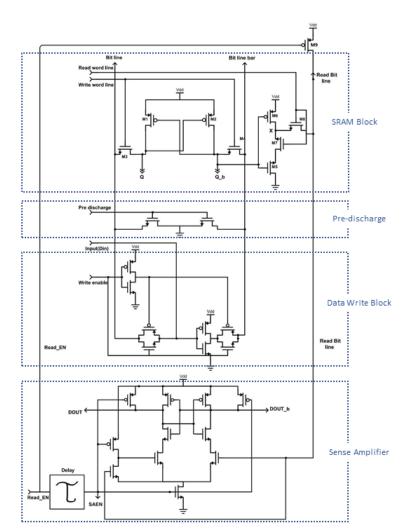
Figure 14. Proposed PLNA 8T SRAM cell operation.

5.2. PLNA 8T SRAM

The write operation starts with asserting write word line by data write block (Figure 14), and all control signals for the operation are shown in Table 3. Q and Qb are loaded to the states on Bit Line and Bit Line Bar, respectively. By adding the separated reading access path, we can solve some issues of the reading operation. Because of the stacking technology applied to the reading part, the proposed SRAM cell can operate on even lower voltage and observe lower leakage. In addition, the reading speed only depends on the discharging rate through the transistors of the reading part and the sensitivity of sense amplifier.

Table 3. Status of control signals for proposed PLNA 8T SRAM.

	Hold	Read	Write 1/0
Bit line	0	0	1/0
Bit line bar	0	0	1/0
Read Bit line	0	1	0
Read word line	0	1	0
Write word line	0	0	1
Pre discharge	0	0	0/1
Write enable	0	0	1
Read enable	0	1	0
Input(Din)	0	0	1/0

5.2.1. Writing Operation

For writing operation, the data are first loaded to Bit Line and Bit Line Bar, and then the word line is asserted to 1. The process starts with turning on M2, and then the Qb is charged or discharged to the state on Bit Line Bar. Note that M4 is designed to be stronger than M2. Thus, the state on the Qb is forced to be the state on Bit Line Bar even if it is different with the latter. The same condition happens on the other side, and finally the state be saved in the cell, as shown in Figures 15 and 16.
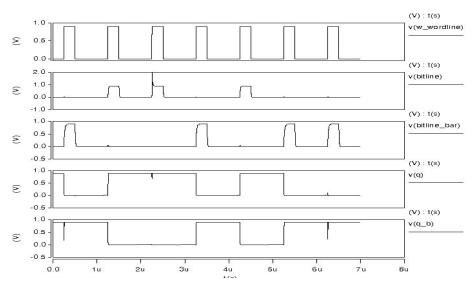

Figure 15. Waveform of write operations in FinFET (sequence of 0110100).
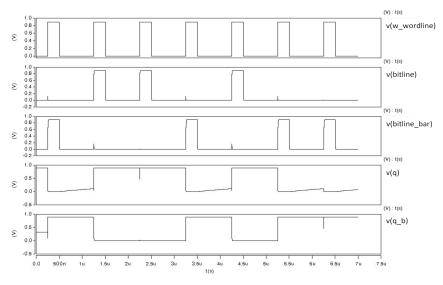

Figure 16. Waveform of write operations in CNFET (sequence of 0110100).

5.2.2. Reading Operation

In the case of reading operation, as mentioned above, we use an independent reading part. During the reading operation, we switch on the M7 and M8 through the read word line. At the same time, the writing part is switched off through the write word line to reduce the power consumption. As a result, our proposed SRAM has the principle of operating only four transistors in all operations of write and read, even though it has eight transistors. The results of all reading operations are shown in Figures 17–20.
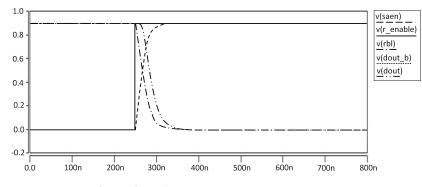
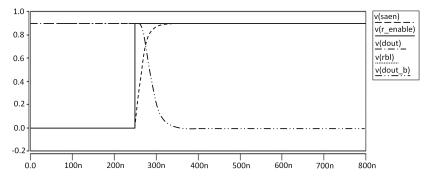**Figure 17.** Waveform of Reading 0 operation in FinFET.

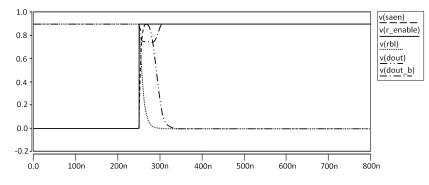**Figure 18.** Waveform of Reading 1 operation in FinFET.

Figure 19. Waveform of Reading 0 operation in CNFET.

Figure 20. Waveform of Reading 1 operation in CNFET.

Depending on the value of Qb, node X is 0 or 1. For Reading 0, the read bit line is pre-charged to vdd. Since Qb is 1, M5 is activated and the read bit line discharges though M8, M7 and M5. Because M7 and M5 are stacked, the power consumption can be reduced. For Reading 1, Q remains 1 and QB remains 0. This is because the node X is 1 and the read bit line can be read without discharging.

The proposed PLNA 8T SRAM is suitable for our target low power and high speed, but it has some loss in stability. Even though this structure has less stability, this low stability is consistent with the criteria for using SRAM, so it is be a big problem in the actual use of the SRAM output, but we compensate for the low stability with the following new SE structure.

5.3. SE 8T SRAM

The write operation using single-ended structure can be started by asserting write word line by writing block (write enable and input). Then, Q and Qb are loaded to the states on bit line, as shown in Figure 21. And all control signals for the operation are shown in Table 4. As the same approach as PLNA 8T SRAM, by adding the separated reading access path, we can solve some issues of the reading operation. Due to the stacking of devices on the read port, the proposed SRAM can operate on even lower voltage and observe lower leakage [23]. In addition, the reading speed only depends on the discharging rate through the transistors of the reading part and the sensitivity of sense amplifier.

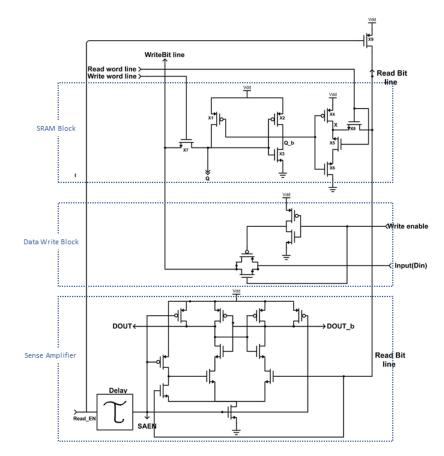**Figure 21.** Proposed SE 8T SRAM cell operation.

Table 4. Status of control signals for proposed SE 8T SRAM.

	Hold	Read	Write 1/0
Write Bit line	0	0	1/0
Write Word line	0	0	1
Read word line	0	1	0
Read Bit line	0	1	0
Write enable	0	0	1
Read enable	0	1	0
Input(Din)	0	0	1/0

5.3.1. Writing Operation

To perform writing operation, if the data are to be written as 1, then write enable signal is turned on, the data signal is to changed to 1, the word line is activated, and data are written as 1. If the data to be written are 0, then the data signal is changed to 0. When the data signal is changed to 0, the write bit line is connected to ground and we can write 0. Since we have different read and write operations, the switching of the inverter is eliminated and the speed of both read and write operation is increased. And the results of all writing operations are shown in Figures 22 and 23.

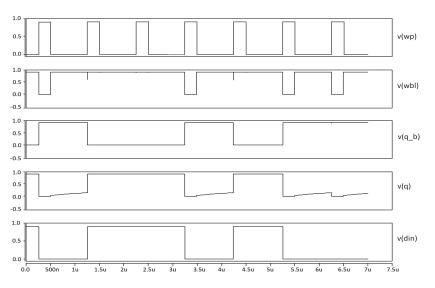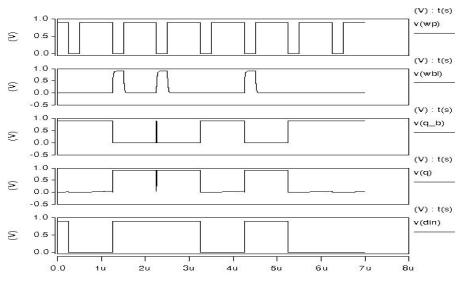**Figure 22.** Waveform of write operations in FinFET (Sequence of 0110100).

Figure 23. Waveform of write operations in CNFET (Sequence of 0110100).

5.3.2. Reading Operation

For the reading operation, in our design, we have separate read access blocks. In the reading mode, X8 and X7 are turned on by asserting read lines. All results for the reading operation are shown in Figures 24–27.

Figure 24. Waveform of Reading 0 operations in FinFET.

Figure 25. Waveform of Reading 1 operations in FinFET.

Figure 26. Waveform of Reading 0 operations in CNFET.

Figure 27. Waveform of Reading 1 operations in CNFET.

Based on the value of Qb, the value at node X is 1 or 0, because we use and inverter that inverts the value at node Qb. To Read 0, the bit line on which is to be performed is pre-charged to vdd. Depending on the value of Qb and node, Read 1 or Read 0 operations are performed. In our case, when Qb is 1, node X is 0. When read word line is asserted and the value at node x is 0, 0 is read by sense amplifier. To Read 1, Q is holding the value 1

and Qb has value 0. Thus, the value at node x is 1 and read bit line is not discharged. Thus, the sense amplifier is Read 1. Since M7 and M5 are in series combination, we have a stack effect here and in this way we can reduce static power.

5.4. A Latch Type Voltage Sense Amplifier

Sense amplifier (SA) is designed as shown in Figure 28, and the virtual read bit line bar (VRBLB) in the SA for output is produced by M10 and M11. In the off-state, even though the current flows, it is extremely small, so the power consumption is not significant, and, because nodes 2 and 3 are pulled at M7 and M8, both inputs, as well as RBL, can be seen pulling to Vdd.The SA starts working when asset SAEN is 1, which turns on M9 and turns off M7 and M8. Node 1 is pulled down immediately after M9 turns on. The drain currents of M5 and M6 start discharging nodes 2 and 3. With a difference between RBL and VRBLB, the drain current through M5 and M6 is also different. During the Read 1 operation, the delay is much shorter because there is no change in value in the RBL. During the Read 0 operation, the RBL is released as 0 through the read approach path. M10 is turned off, held until the value of VRBLB is high, and gives the signal a reversal of RBL.The latch circuit is controlled by the current through M5 and M6, and the output voltage of D-OUT and D-OUT-b is determined by the current of M5 and M6. The output voltage of DOUT and DOUT-b is amplified through the control of the latch.
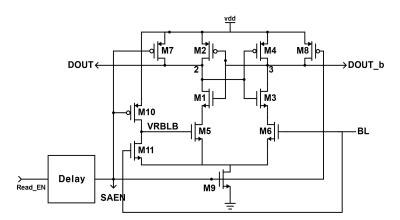

Figure 28. Schematic of sense amplifier.

6. Stability Analysis

Figure 29 shows the Static Noise Margin (SNM) for each proposed SRAM structure. SNM shows the tolerance for noise before the SRAM cell risks losing 'memorized' bits. Therefore, it is necessary to secure a minimum SNM when designing the SRAM structure, which can be confirmed as a square, as shown in Figure 29. To find the SNM, we considered two consecutive inverters, as shown in Figure 30. Drawing both Vout (Vin) DC characteristics, but swapping the X/Y axis for the second, we found the SNM graphically with the largest square diagonal that fits the continuous DC characteristic.

The two new structures we proposed show lower SNM results than conventional 8T SRAM. We have put more emphasis on high-throughput and ultra-low power for AI computations, even though the stability is slightly reduced compared to the conventional 8T SRAM. However, since the stability was reduced under the conditions that satisfy the performance required for AI computation, the performance of our SRAM shows the same results as the existing 8T SRAM, and it is reliable because it was analyzed using an actual model. In addition, the proposed SRAM offers read SNM free because of a separated reading section from an internal node of latch. There is no feedback from the read access circuit, which makes it read SNM free.

While reading and writing operations show the same results as conventional state-of-the-art SRAM, high-throughput and ultra-low power performances are what we aim at in the SRAM design for AI computation.
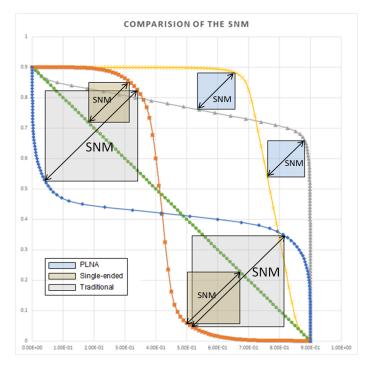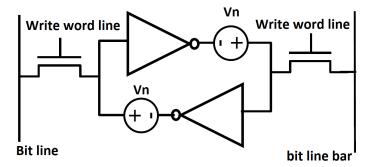
Figure 29. Comparisons of the SNM.

Figure 30. Back-to-back inverter circuit for checking the SNM.

7. Performance Analysis

For reliable data comparison, we applied the same transistor size ratio, capacitor size, voltage source, temperature, and test circuit as the proposed SRAM to the comparable tradition 8T FinFET SRAM, and all conditions were equally applied.

The traditional 8T FinFET data are a result of re-simulating the traditional 8T SRAM structure [21,22] under the same conditions as our proposed 8T SRAM. The total memory power consumption and delay may vary depending on type of the applied AI chip and CNN model. Therefore, we measured power consumption and delay with each sequence separately (e.g., Reading 0, Reading 1, Writing 0, and Writing 1) for reliable measurement results. For this reason, we are showing the power consumption for Readings 0 and 1 and Writings 0 and 1 separately in the table, not just total power consumption of the memory. This is more accurate and reliable information when we do not know which AI chip the memory is applied to. For reliable comparison, all conditions (input data, size, temperature, etc.) of the traditional SRAM and the proposed SRAM are set the same.

The purpose of this simulation is to show the possibility of achieving improved results in terms of power consumption and speed even by simply changing the SRAM cell structure, and the improved results for our new structure compared to the traditional SRAM structure were analyzed through the following session.

7.1. PLNA 8T SRAM

We used a BSIM-Independent Multi Gate (IMG) HSPICE FinFET model and Stanford CNFET HSPICE model [24] in 32-nm technology. We chose chirality vector as (19,0), diameter 1.5 nm for CNFET. The Stanford University CNFET model includes enhanced modes of channeling single-walled carbon nanotubes; each device can have at least 1–10 carbon nanotubes with user-specified chirality. In addition, this model is based on a quasi-ballistic transport picture and is open for use by the researcher, which includes an accurate description of the capacitor network in the CNFET.

We set the optimal Vdd to 0.9v considering all the characteristics, low power, and speed of the transistor model. CNFET transistor in 4T PLNA cell used the minimum size for both area and low power, as shown in Table 1. For the CNFET design, we chose the chirality vector as (19,0) using a common method (3) of making carbon nanotubes into semiconductors as below. The delay measurement value was measured at 50% of the Vdd according to each transition. All measurements were conducted through Netlist and Hspice, and circuit and waveform were checked through Cadence Virtuoso and Synopsys Cosmos Scope for function verification.

$$|m - n| = 3k \pm 1; k : integer \tag{3}$$

By using the CNT diameter in Equation (4) as below, we set the CNT diameter as 1.5 nm.

$$D_{CNT} = \frac{a}{\pi} \sqrt{m^2 + n^2 + mn} \tag{4}$$

In addition to the SRAM simulation, we constructed the 8-T CNFET and FinFET SRAM whole circuits, as shown in Figures 14 and 21, for simulations as similar to actual circuits in real field as possible. Both the bit line and bit line bar were pre-discharged with ground before writing operation, and writing and reading operations were analyzed and simulated using pseudo-random sequence input as 0110100.

As shown in Table 5, in the design of the FinFETs and CNFETs, due to the elimination of a strong or rapid current flowing between the latch and the access transistor and the optimization of the CNFET model, the proposed SRAM reduces the static power consumption by up to 73% in Hold 0 and Hold 1 modes. The results of dynamic power consumption are shown in Table 6.

Table 5. Static power consumption comparison.

Power (W)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Hold 0	1.228×10^{-9}	6.852×10^{-10}	5.591×10^{-10}	44.2%/54.5%
Hold 1	1.228×10^{-9}	6.562×10^{-10}	3.315×10^{-10}	46.6%/73%

Table 6. Dynamic power consumption comparison.

Power (W)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Write 0	1.906×10^{-5}	7.330×10^{-7}	3.599×10^{-8}	96%/99%
Write 1	1.882×10^{-5}	5.001×10^{-7}	7.074×10^{-8}	97%/99%
Read 0	1.081×10^{-9}	1.560×10^{-10}	4.575×10^{-11}	85.6%/95.8%
Read 1	1.094×10^{-9}	1.532×10^{-10}	6.068×10^{-11}	86%/94.5%

In the write operation '1', the dynamic power of the proposed PLNA 8T SRAM can be seen as a reduction of approximately 99% in CNFET compared to the conventional 8T FinFET SRAM. For Writing 0, the dynamic power of the our proposed CNFET 8T SRAM is reduced by 99%. In addition, the reading part of the design in the actual SRAM design can be shared for all writing part 4T PLNAs in the SRAM column; thus, the proposed SRAM

design is able to improve the area compared to the conventional SRAM design. In other words, one reading part and several writing parts can be composed together.

In Table 7, because of the optimized SRAM cell size and characteristics of the PLNA and SE structure in the writing part, we can see a 90% improvement in writing operation speed. In addition, the independently separated read operation and the speed of the sense amplifier (Figure 28) had a great effect on improving read delay time.

Table 7. Delay comparison.

Delay (s)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Write 0	4.255×10^{-10}	4.047×10^{-11}	1.419×10^{-11}	90.5%/96.7%
Write 1	5.643×10^{-10}	7.551×10^{-11}	6.452×10^{-11}	86.6%/88.6%
Read 0	1.511×10^{-9}	1.418×10^{-9}	2.659×10^{-10}	6.2%/82.4%
Read 1	1.640×10^{-9}	1.637×10^{-9}	3.097×10^{-10}	0.2%/81.1%

Since our design has separate read and write operations, it makes our design novel compared to conventional SRAM cell.

7.2. SE 8T SRAM

In Table 8, the static power is calculated and compared with traditional 8T FinFET SRAM. The improvement of our new SE 8T SRAM is that CNFET has better performance compared to FinFET.

Table 8. Static power consumption comparison.

Power (W)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Hold 0	1.228×10^{-9}	3.552×10^{-10}	2.679×10^{-10}	71%/78.2%
Hold 1	1.228×10^{-9}	2.561×10^{-10}	2.299×10^{-10}	79.1%/81.3%

After simulation, we achieved 71% improvement for Hold 0 and 79.1% for Hold 1 in FinFET and 78.2% for Hold 0 and 81.3% for Hold 1 in CNFET.

Table 9 shows the performance results of dynamic power for traditional and proposed 8T SRAM using FinFET and CNFET. For Write 0 and 1, we have around 99% improvement compared to traditional 8T SRAM for FinFET and CNFET. For Read 0 and 1, we have 86% for FinFET and 95% for CNFET. As shown in Table 10, since we have discrete read and write operations, which have different control paths, we can show that the proposed 8T SRAM cell is faster than the traditional 8T SRAM cell.

Table 9. Dynamic power consumption comparison.

Power (W)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Write 0	1.906×10^{-5}	3.533×10^{-8}	2.395×10^{-8}	99%/99%
Write 1	1.882×10^{-5}	1.624×10^{-7}	7.269×10^{-8}	99%/99%
Read 0	1.081×10^{-9}	1.560×10^{-10}	4.575×10^{-11}	85.6%/95.8%
Read 1	1.094×10^{-9}	1.532×10^{-10}	6.068×10^{-11}	86%/94.5%

Table 10. Delay Comparison.

Delay (s)	Traditional 8TFinFET	Proposed 8TFinFET	Proposed 8TCNFET	Improvement (FinFET/CNFET)
Write 0	4.255×10^{-10}	3.736×10^{-11}	1.039×10^{-11}	91.2%/97.6%
Write 1	5.643×10^{-10}	1.586×10^{-10}	3.503×10^{-11}	71.9%/93.8%
Read 0	1.511×10^{-9}	1.418×10^{-9}	2.659×10^{-10}	6.2%/82.4%
Read 1	1.640×10^{-9}	1.637×10^{-9}	3.097×10^{-10}	0.2%/81.1%

8. Conclusions

To accommodate complex CNN models, existing state-of-the-art SRAMs have reached physical limitations to increasing power efficiency due to scaling and density. This has been a major limiting factor in improving AI computing performance and the need for new SRAMs. Various SRAMs for AI have been proposed by many researchers, but a lot of power consumption is being used to provide high-throughput. However, in various AI applications, such as autonomous vehicles, security systems, and Internet of Things (IoT), power consumption is an important factor that cannot be compromised.

Therefore, in this paper, we propose two types of novel 8-T SRAMs, PLNA 8T SRAM and SE 8T SRAM, based on CNFET and FinFET models for high-throughput, ultra-low power, better readability, and writ-ability for AI application. Our proposed new SRAM can be used as memory for I/O in PE blocks for AI computations. Since the memory for I/O of each PE block occupies most of the power consumption in CNN, our new memory will bring a lot of power consumption reduction in AI chip. In addition, our proposed memory maximizes the effect by improving the throughput as well as the power consumption in the AI chips.

We compared the existing state-of-the-art 8-T FinFET SRAM with our proposed SRAMs. The proposed SRAM achieved low dynamic and static power dissipation due to single-ended scheme and PLNA scheme. Experimental results show that our proposed design reduced static power consumption up to 81.3% for Hold 1 in CNFET. Furthermore, due to the combination of independent reading access component and voltage sense amplifier, the reading power consumption of our 8T SRAM was improved up to 95.8% for Reading 0 and 86% for Reading 1. Moreover, because of the separated reading part from an internal node of latch, there is no feedback from the read access circuit, which makes it read SNM free.

Author Contributions: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Y.K. and S.P.; Writing—original draft, Writing—review and editing, Y.K., S.P., H.K., N.Y. and K.K.C.; Supervision, Funding acquisition, Project administration, K.K.C. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Power Consumption Vital for AI Self-Driving Cars. Available online: https://www.aitrends.com/ai-insider/power-consumption-vital-for-ai-self-driving-cars (accessed on 10 January 2021).
2. Optimizing IoT Power Consumption. Available online: https://www.iotworldtoday.com/2020/05/08/optimizing-iot-power-consumption (accessed on 10 January 2021).
3. Alorda, B.; Torrens, G.; Bota, S.; Segura, J. 8T vs. 6T SRAM cell radiation robustness: A comparative analysis. *Microelectron. Reliab.* **2011**, *51*, 350–359. [CrossRef]

4. Siers, S.; Damaraju, S.; George, V.; Jahagirdar, S.; Khondker, T.; Milstrey, R.; Sarkar, S.; Stolero, I.; Subbiah, A. The first 22 nm IA multi-CPU and GPU system-on-chip using tri-gate transistors. In Proceedings of the 2012 IEEE Asian Solid State Circuits Conference (A-SSCC), Kobe, Japan, 12–14 November 2012; pp. 9–12. [CrossRef]

5. Rahman, A.; Bai, P.; Curello, G.; Hicks, J.; Jan, C.; Jamil, M.; Park, J.; Phoa, K.; Rahman, M.S.; Tsai, C.; et al. Reliability studies of a 22 nm SoC platform technology featuring 3-D tri-gate, optimized for ultra low power, high performance and high density application. In Proceedings of the 2013 IEEE International Reliability Physics Symposium (IRPS), Anaheim, CA, USA, 14–18 April 2013; pp. PI.2.1–PI.2.6. [CrossRef]

6. Brown, B. The Social Life of Autonomous Cars. *Computer* **2017**, *50*, 92–96. [CrossRef]

7. SAE J3016: Level of Driving Automation. Available online: https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic (accessed on 10 January 2021).

8. Self-Driving Cars and Power Consumption. Available online: https://nitinvaish.medium.com/self-driving-cars-and-power-consumption-new-chip-designs-4c723659f8cd (accessed on 10 January 2021).

9. Self-Driving Cars Have Power Consumption Problems. Available online: https://www.therobotreport.com/self-driving-cars-power-consumption (accessed on 10 January 2021).

10. Choi, S.; Bong, K.; Han, D.; Yoo, H. CNNP-v2: An Energy Efficient Memory-Centric Convolutional Neural Network Processor Architecture. In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hsinchu, Taiwan, 18–20 March 2019; pp. 38–41. [CrossRef]

11. Bong, K.; Choi, S.; Kim, C.; Kang, S.; Kim, Y.; Yoo, H. 14.6 A 0.62mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on haar-like face detector. In Proceedings of the 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 5–9 February 2017; pp. 248–249. [CrossRef]

12. Okagaki, T.; Shibutani, K.; Matsushita, H.; Ojiro, H.; Morimoto, M.; Tsukamoto, Y.; Nii, K.; Onozawa, K. Area and performance study of FinFET with detailed parasitic capacitance analysis in 16 nm process node. In Proceedings of the 2015 International Conference on Microelectronic Test Structures, Tempe, AZ, USA, 23–26 March 2015; pp. 141–144. [CrossRef]

13. Takahashi, C.; Shibahara, S.; Fukuoka, K.; Matsushima, J.; Kitaji, Y.; Shimazaki, Y.; Hara, H.; Irita, T. 4.5 A 16 nm FinFET heterogeneous nona-core SoC complying with ISO26262 ASIL-B: Achieving 10-7 random hardware failures per hour reliability. In Proceedings of the 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 31 January–4 February 2016; pp. 80–81. [CrossRef]

14. Self-Driving Cars Use Crazy Amounts of Power. Available online: https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip (accessed on 10 January 2021).

15. Gaynor, B.D.; Hassoun, S. Fin Shape Impact on FinFET Leakage with Application to Multithreshold and Ultralow-Leakage FinFET Design. *IEEE Trans. Electron Devices* **2014**, *61*, 2738–2744. [CrossRef]

16. Banu, R.; Shubham, P. Design and Performance Analysis of 6T SRAM Cell in 22 nm CMOS and FINFET Technology Nodes. In Proceedings of the 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), Bangalore, India, 16–17 March 2017; pp. 38–42. [CrossRef]

17. Wang, W.; Yu, Z.; Choi, K. High SNM 6T CNFET SRAM cell design considering nanotube diameter and transistor ratio. In Proceedings of the 2011 IEEE International Conference on Electro/Information Technology, Mankato, MN, USA, 15–17 May 2011; pp. 1–4. [CrossRef]

18. Deng, J.; Wong, H.P. A Compact SPICE Model for Carbon-Nanotube Field-Effect Transistors Including Nonidealities and Its Application—Part I: Model of the Intrinsic Channel Region. *IEEE Trans. Electron Devices* **2007**, *54*, 3186–3194. [CrossRef]

19. Wang, C.; Lee, C.; Lin, W. A 4-kb Low-Power SRAM Design with Negative Word-Line Scheme. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2007**, *54*, 1069–1076. [CrossRef]

20. Kim, Y.; Tong, Q.; Choi, K.; Lee, Y. Novel 8-T CNFET SRAM cell design for the future ultra-low power microelectronics. In Proceedings of the 2016 International SoC Design Conference (ISOCC), Jeju, Korea, 23–26 October 2016; pp. 243–244. [CrossRef]

21. Monica, M.; Chandramohan, P. Characterization of 8T SRAM cells using 16 nm FinFET technology. In Proceedings of the 2016 International Conference on Signal Processing and Communication (ICSC), Noida, India, 26–28 December 2016; pp. 403–406. [CrossRef]

22. Kutila, M.; Paasio, A.; Lehtonen, T. Comparison of 130 nm technology 6T and 8T SRAM cell designs for Near-Threshold operation. In Proceedings of the 2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS), College Station, TX, USA, 3–6 August 2014; pp. 925–928. [CrossRef]

23. Patel, S.; Kim, Y.; Choi, K. Novel Low Power FinFET SRAM Cell Design with Better Read and Writabilty For Cache Memory. In Proceedings of the 2018 International SoC Design Conference (ISOCC), Daegu, Korea, 12–15 November 2018; pp. 44–45. [CrossRef]

24. Stanford CNFET Model. Available online: https://nano.stanford.edu/stanford-cnfet-model (accessed on 21 January 2021).