Design and Thermal Analysis of 2.5D and 3D Integrated System of a CMOS Image Sensor and a Sparsity-Aware Accelerator for Autonomous Driving

Janak Sharda, *Graduate Student Member, IEEE,* Madison Manley, *Graduate Student Member, IEEE*, Ankit Kaul, *Graduate Student Member, IEEE*, Wantong Li, *Graduate Student Member, IEEE*, Muhannad Bakir, *Senior Member, IEEE,* and Shimeng Yu, *Senior Member, IEEE*

***Abstract*—For the autonomous driving application, data movement has increased rapidly between a CMOS Image sensor (CIS) and the processor due to increase in image resolution. Advanced packaging techniques like 2.5D/3D integration have been proposed to reduce the data movement energy between memory and processor. In this work, we explore the use of such techniques to integrate a CIS and a backend accelerator on a silicon interposer. The data movement energy from CIS to the accelerator is thus reduced by 100× compared to using the conventional MIPI links. We perform thermal simulations to study the impact of the thermal coupling of CIS and accelerator and ensure a peak temperature increase of less than 5 °C. We also vary the distance between the CIS and the processor to study the trade-offs between energy savings and peak temperature. Next, we assume the 3D stacked CIS and accelerator to reduce the data movement further and obtain an energy efficiency of 45.81 TOPS/W. Now we observe a heat dissipation challenge with an increase in the peak temperature of more than 85 °C. Hence, we scale down the operational frequency and study the trade-off between performance degradation and reduction in peak temperature, while maintaining the accurate multi-object tracking on the BDD100k dataset for autonomous driving.**

***Index Terms*—3D-stacked CIS, advanced packaging, 2.5D/3D integration, thermal modeling, near-pixel compute, hardware accelerator, autonomous driving**

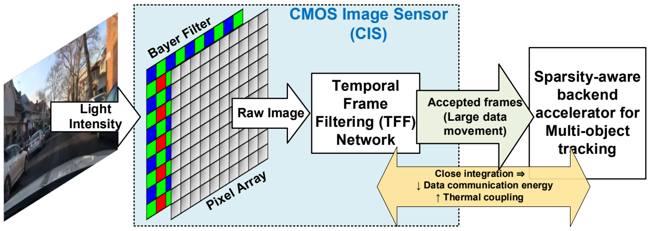
CMOS Image Sensor (CIS)
Bayer Filter
Pixel Array
Light Intensity
Raw Image
Temporal Frame Filtering (TFF) Network
Accepted frames (Large data movement)
Sparsity-aware backend accelerator for Multi-object tracking
Close integration ⇒
↓ Data communication energy
↑ Thermal coupling

Fig. 1. Schematic of the complete video processing pipeline. The CIS takes light intensity as the input which is converted to a digital image. Redundant frames are rejected by the Temporal Frame Filtering (TFF) network. The accepted frames are processed by the backend accelerator which runs the QDTrack network to perform multi-object tracking (modified from [4]).

I. INTRODUCTION

Deep learning algorithms [1] for autonomous driving have advanced rapidly in recent years. Typically, a CIS captures an image and sends it to a backend processor for multi-object tracking using state-of-the-art algorithms like the QDTrack network [2]. For autonomous driving applications, such algorithms require processing a high-resolution image (e.g., 1296×720 image in BDD100k dataset [3]) and hence are difficult to run in real-time. Due to the high resolution of the images, large amounts of data movement from the CIS to the processor is required, which increases energy consumption and reduces the time for the backend processing. Conventionally, Mobile Industry Processor Interface (MIPI) links have been used to integrate CIS with a processor, which are slow and energy inefficient. Advanced packaging techniques such as 2.5D/3D integration [5] have been used to efficiently transfer data from the memory to the processor by closely integrating them using Cu-Cu hybrid bonding (HB), μbumps, and through-silicon via (TSV). These techniques reduce energy consumption by reducing the interconnect parasitics between memory and processor. In this work, a sparsity-aware backend accelerator is designed and integrated with the CIS using techniques like 2.5D/3D integration to reduce the data movement latency and energy. Although such close integration reduces the latency and energy of data transfer, it increases the thermal coupling between the processor and CIS, hence increasing the chip temperature. An increase in chip temperature can reduce the image quality by increasing the thermal noise in the CIS as well as decrease the performance of the pixel circuit. Prior works have been proposed on using advanced packaging techniques to integrate a CIS with a processor [6][7], however, they do not study the thermal impact of such processors running such deep learning algorithms.

This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE. This work was supported in part by the DARPA IP2 program under contract number HR00112190116, and the CHIMES center of the SRC/DARPA JUMP 2.0 program. *(Corresponding author: S. Yu).*

The authors are affiliated with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta 30332, USA (email: shimeng.yu@ece.gatech.edu).

Moreover, to reduce the energy consumption and latency of the system further, designing the accelerator in an advanced technology node is desired. This scaling increases the chip power density which results in an increase in the chip temperature. To prevent obstruction of light, the photodiode needs to be on top tier without any heat spreader. As a result, the 3D integration of such a near-pixel compute (NPC) [8] engine with CIS makes it difficult to dissipate the excess heat caused by running the accelerator. This further increases the chip temperature which results in a decrease in the operating frequency of the chip. Thus, a complete thermal-aware study is required to design such an NPC-based 3D stacked CIS.

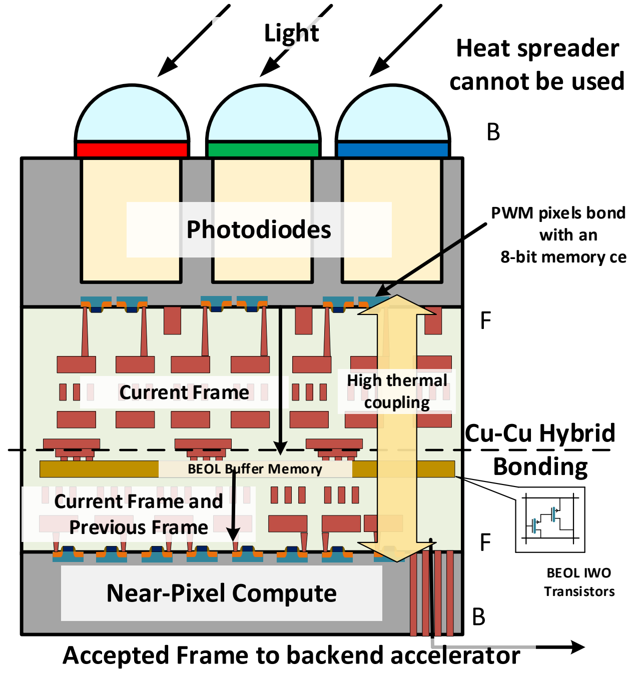
Light
Heat spreader cannot be used
B
Photodiodes
PWM pixels bonded with an 8-bit memory cell
F
Current Frame
High thermal coupling
Cu-Cu Hybrid Bonding
BEOL Buffer Memory
Current Frame and Previous Frame
F
BEOL IWO Transistors
Near-Pixel Compute
B
Accepted Frame to backend accelerator

Fig. 2. 3D stacked CIS. The top tier comprises photodiodes and the bottom tier comprises DCIM based near-pixel compute (NPC) engine and IWO-FETs based buffer memory to run the TFF network. The two tiers are integrated using Cu-Cu hybrid bonding. Note that, such close integration and absence of heat spreader results in an increase in peak temperature (modified from [9]).

This work extends our prior work done in the EDTM 2023 conference [10] by redesigning the custom backend accelerator instead of using an off-the-shelf FPGA, and reevaluating the thermal impact of its 3D integration with the CIS instead of 2.5D integration. The main contributions of this work are as follows:

1. PE array-based backend accelerator with sparsity-aware multiply-and-accumulate (MAC) units is designed.
2. The accelerator is integrated with the 3D stacked CIS described in [9] using 2.5D integration on a silicon interposer. Evaluations and thermal simulations are performed for the complete system design.
3. The accelerator is scaled to an advanced technology node (e.g., 7 nm) and integrated using more aggressive 3D integration, and its thermal impact on performance is evaluated.
4. Cut-off temperature of various components is evaluated, and the circuits are modified slightly to ensure a sufficient thermal budget for the 3D integrated design.

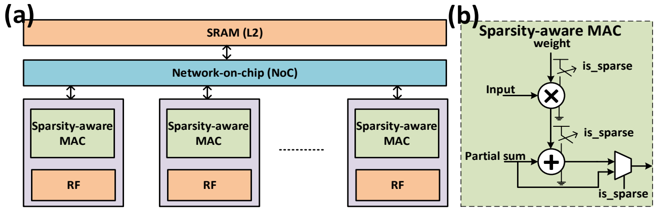

Fig. 3. (a) PE array-based backend accelerator schematic. The NoC is decided by the dataflow, which in our case is kc-p dataflow from [11]. (b) Sparsity-aware 8b×8b MAC unit, with power gated multiplier and adder. The signal '*is_sparse*' = 1 when weight = 0. ((a) modified from [11]).

The article is divided as follows: Section II describes the prior work for the autonomous driving algorithm and the 3D stacked CIS design. Section III describes the design of the backend accelerator and its integration with the CIS. Section IV describes the simulation methods and the results.

II. Background

A. Algorithm for autonomous driving

In this work, the BDD100k dataset [2] is used as the dataset for the autonomous driving problem, on which multi-object tracking is performed using the QDTrack network [3]. Fig. 1 describes the complete image processing pipeline. The image is first filtered using a temporal frame filtering (TFF) network [4], and we call it frontend pre-processing to reduce the data volume to be transmitted to the backend processing (e.g., for multi-object tracking). The TFF network is a 3-layer convolutional neural network that takes the current and the previous frames as the input and gives an output score. It rejects frames with scores less than a certain threshold to reject redundant frames without much degradation in accuracy. The accepted frames are then passed through the QDTrack network with a ResNet-50 backbone, which segments the image and performs multi-object tracking on it. For performing accurate multi-object tracking, BDD100k dataset has been generated in [3] at 5 frames per second (FPS). The TFF network reject 40% of the frames, therefore, the backend processor needs to process data at 3 FPS or with a total latency of less than 333 ms. Prior work [4] shows that the TFF network performs frame dropping with a minimal drop in the multi-object tracking accuracy. This reduces the amount of data transfer from the CIS to the backend processor as well as reduces the operating frequency of the backend processor, hence decreasing the overall energy consumption of the system.

B. Hardware for frontend and backend processing

Fig. 2 shows the schematic of 3D stacked CIS [9] for frontend processing including the image capturing and the TFF network execution. The CIS takes light intensity as the input and converts it to a digital signal which is the input to the TFF network. The near-pixel compute (NPC) based engine runs the TFF network. The 3D stacked CIS comprises of 2-tier design, where the top tier comprises photodiodes and its peripheral circuit, and the bottom tier comprises buffer memory and the NPC engine to run the TFF network. The top tier is designed in the 40 nm technology node considering the CIS availability in today's foundry offerings (e.g., from TSMC [12]), and the bottom tier is designed in the 22 nm technology node for better performance and power efficiency. The two tiers are integrated using HB with a 3 µm pitch for global shutter operation, as described in [9]. The buffer memory comprises of back-end-of-line (BEOL) Tungsten-doped Indium Oxide Transistors (IWO-FET) based 2T-eDRAM cell [13] to store the previous frame, in the 22 nm technology node. Since, the TFF network needs to run at 5 FPS, the 2T-eDRAM cell needs to have a retention time of at least 200 ms, to avoid any refresh penalty. The wide bandgap of IWO-FET offers ultra-low leakage current and could satisfy such retention requirements in 2T-eDRAM cells. The NPC circuit comprises a digital compute-in-memory (DCIM) based design, similar to the prototype chip by TSMC [14].

This 3D stacked CIS is first assumed to be integrated with a backend FPGA using 2.5D integration [10] on a silicon interposer to reduce energy consumption due to data movement. The RC parasitics of the 2.5D links are modeled with a pitch of 8 µm using models described in [15]. Thermal modeling is performed and a peak temperature of 84.8 °C of CIS is found (shown in Fig. 6(a) as baseline design), even with a heat spreader and active air cooling that covers the entire FPGA die. This high temperature is due to the high energy consumption of the FPGA die, which inhibits its potential for 3D integration with the CIS. As a result, an energy-efficient sparsity-aware backend accelerator is designed and integrated with the CIS that paves the way for a fully 3D integrated system, as described in the following section.

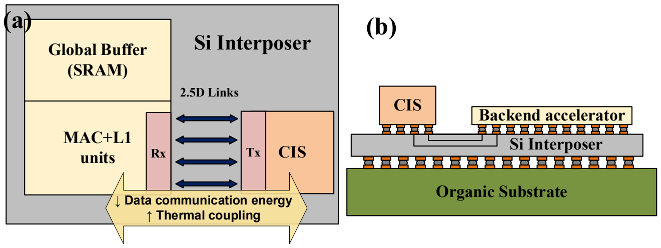

Fig. 4. (a) The backend accelerator comprises MAC+L1(SRAM) units and an L2 global buffer. Tx and Rx are simple inverter chain-based driver circuits. (b) The two chips are integrated on Si interposer using 2.5D links (modified from [10]).

III. Design and Modeling Methodologies

A. Backend accelerator design

The backend accelerator is a PE array-based accelerator designed using the MAESTRO tool [11]. Fig. 3(a) describes the PE array with a network-on-chip (NoC) and a shared L2 SRAM buffer. Since ResNet-50 takes nearly 70% of the computation for the QDTrack we use the kc-p dataflow where the parallelism is at the kernel and channel level and is observed to be efficient for ResNet-50. Next, a sparsity-aware PE is designed, where each PE can perform eight 8-bit MAC operations. For a sparsity-aware design, the PE is power gated with a '*is_sparse*' signal, when the weight is 0, the PE is turned OFF. Since the '*is_sparse*' signal is transmitted along with the weight, each weight is transmitted as a 9-bit signal. Along with that, read/write for each weight only happens if the '*is_sparse*' signal is '1'. Considering 90% weight sparsity of the QDTrack network [16], the total model size is just 0.34 MB. Each PE has 2.3 kb of Register files for local storage, and the complete processor has a shared 3.66 MB L2 SRAM as the global buffer to store the weights of the network as well as the intermediate computation results. The PEs are connected through a NoC decided by the kc-p dataflow and the energy estimation is done by finding the capacitance of the NoC wires.

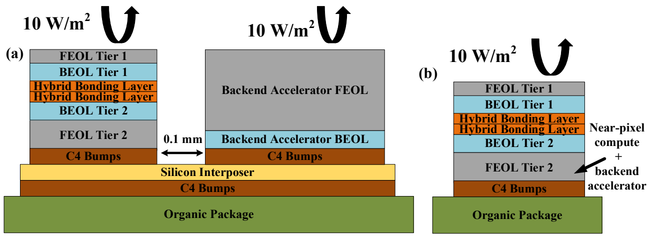

Fig. 5. Simulations are performed for the above schematics in Ansys mechanical APDL software, using the parameters described in the table I. (a) Schematic for 2.5D integrated design with 22 nm node backend processor ("Design 1"). (b) Schematic for 3D integrated design with 7 nm node backend processor ("Design 2").

B. 2.5D integration

Next, 2.5D integration of CIS and backend processor (either off-the-shelf FPGA or our custom-designed accelerator) on a silicon interposer is modeled, as shown in Fig. 4, where the ASIC is the backend accelerator. RC parasitics of the 2.5D links are modeled as described in [15]. Additionally, parasitics due to µbump and bump pad are modeled. To find out the peak temperature, the complete system is modeled and simulated in Ansys mechanical APDL, as shown in Fig. 5(a). The air cooling is assumed at the top surfaces for each die stack. Table I provides the list of parameters used for thermal modeling. The power density and the area of each component are obtained from Cadence simulations. For C4 bumps, the area is kept same as other tiers in the stack, with the thickness of 50 µm. Although closer integration of CIS and backend processor reduces the 2.5D link length and hence the parasitics, it might result in higher thermal coupling and higher peak temperature.

Table I: Modeling Specifications

Tier	Thickness (x, y, z) (cm×cm×µm)	Thermal Conductivity (x, y, z) (W/mK)
Tier 1 Bulk	0.388×0.315×4	149 (25 °C)
Tier 1 BEOL	0.388×0.315×2	200, 200, 3
HB Layer	0.388×0.315×3	14.163
Tier 2 BEOL	0.388×0.315×2	200, 200, 3
Tier 2 Bulk	0.388×0.315×120	149 (25 °C)
Accelerator BEOL	0.67×0.67×2	200, 200, 3
Accelerator Bulk	0.67×0.67×128	149 (25 °C)
C4 Bumps	50 µm	0.3, 0.3, 2.5
Interposer	2×2×1000	149 (25 °C)
Package	2.5×2.5×1000	70, 70, 3
TIM	1×1×30	2.9

Therefore, we vary the distance between the two dies and calculate the peak temperature and the energy per bit (EPB) for transmitting the data from CIS to the backend processor. We restrict the maximum distance to 5 mm, as beyond 5 mm the energy consumption due to the interposer links is quite large. Due to the relatively high thermal conductivity of the silicon interposer, the variation of peak temperature within the above-mentioned distance range is negligible (< 0.1 °C), we select 0.1 mm as the distance between the two dies to minimize the EPB.

C. Scaling to 7 nm and 3D integration

To improve performance and reduce energy consumption, the NPC engine on the bottom tier of the 3D stacked CIS and the backend accelerator are both scaled to the 7 nm technology node. This also reduces the area of the NPC die to 2.1 mm^2, creating an area imbalance with a 40 nm CIS die (10 mm^2). This makes it difficult to integrate the two dies using HB. After performing scaling, the combined area of the NPC engine and the backend accelerator reduces to nearly 10 mm^2. Since this is equal to the area of the top CIS tier, we integrate both the NPC engine and the backend accelerator on a single monolithic die at the 7 nm technology node and place them at the bottom tier to make a true 3D stack. The top tier comprising photodiodes is kept unchanged at 40 nm technology nodes as the photodiodes do not follow similar scaling laws as logic with advanced technology nodes. This reduces the overall form factor, removes the use of silicon interposer, removes the need to use a specialized process for HB with area imbalance, as described in [17], and reduces the energy consumption due to data movement even further. Now, the bottom tier comprises both, the NPC engine and the backend accelerator and is integrated with the top tier using HB. This further reduces the EPB for data transfer since the images only need to travel over the BEOL metal layers from the DCIM-based engine to the PE array-based accelerator. Further, we still require the TFF network since it helps to reduce the number of frames to be processed by the larger backend QDTrack network. We maintain the heterogeneous design for the DCIM-based NPC engine and PE array-based backend accelerator, as DCIM is efficient for small neural networks like the TFF network while PE arrays are scalable for large networks like the QDTrack network. The 2.5D integrated design is referred to as "Design 1" and the 3D integrated design is referred to as "Design 2".

However, since the photodiode tier needs to be on the top without a heat spreader, the system cannot have a heat sink, making it difficult to dissipate heat in such 3D integration. To reduce the temperature, we reduce the frequency of operation of the PE-based backend processor and evaluate its effect on the peak temperature and latency of the system.

D. Temperature cut-off

TABLE II: CUT-OFF TEMPERATURES OF VARIOUS COMPONENTS

	Cut-off temperature (°C)
Photodiodes	125
IWO-FETs based buffer memory	95.8
Digital components	85

Next, we check the temperature cut-offs of each of the components for both 2.5D/3D designs. We simulate each circuit component in Cadence Virtuoso at a high temperature and check the functionality to obtain the peak temperature. The cut-off for the digital components is assumed to be 85 °C. The cut-off temperatures of each of the components are defined as follows: photodiodes at higher temperatures increase the thermal noise and decrease the retention time at the storage node capacitance. This reduces the dynamic range and degrades the quality of an 8-bit input image, we consider the temperature corresponding to the dynamic range of 50 dB as the cut-off temperature. For IWO-FETs-based buffer memory, an increase in temperature results in an increase in the leakage current and hence reduces the retention time. For IWO buffer memory, we require a retention time of 200 ms. Table III shows the cut-off temperatures for various components. Due to the small form factor of the overall system and high thermal coupling, the temperature is rather homogeneous throughout the system. As a result, we keep the overall temperature cut-off to be the lowest of all the above values which is 85 °C.

IV. Results and Discussion

The sparsity-aware MAC is designed in Cadence Virtuoso using 40 nm foundry PDK and energy and latency are extracted. The power-performance-area (PPA) is scaled to 22 nm and 7 nm technology nodes using NeuroSim projection [18]. Power consumption and latency of individual components are obtained from SPICE simulations and used along with the cycle-level data obtained from MAESTRO simulations [11]. The backend accelerator comprises 10k PEs along with the local L1 SRAM storage per PE, NoC, and L2 SRAM as a global buffer. Table III summarizes the design parameters and the output metrics for designs 1 and 2. The data transmission energy over 2.5D links reduces by over 117× as compared to conventional MIPI links, as shown in Table IV.

TABLE III: PPA FOR THE BACKEND ACCELERATOR AT DIFFERENT TECHNOLOGY NODES

	Design 1	Design 2
Technology node for backend processor and DCIM	22	7
Integration type	2.5D	3D
Operating Frequency (MHz)	200	41
Power consumption (mW)	206.3	21.3
Latency (ms)	152.7	332.9
Area (mm^2)	41.72	8.11
Frame rate (FPS)	6.5	3
Energy efficiency (TOPS/W)	10.30	45.81

TABLE IV: 2.5D LINK ENERGY CONSUMPTION

	MIPI Link	2.5D Links
Energy/Bit (pJ/bit)	12.5	0.11
Power Consumption (mW)	0.84	0.007

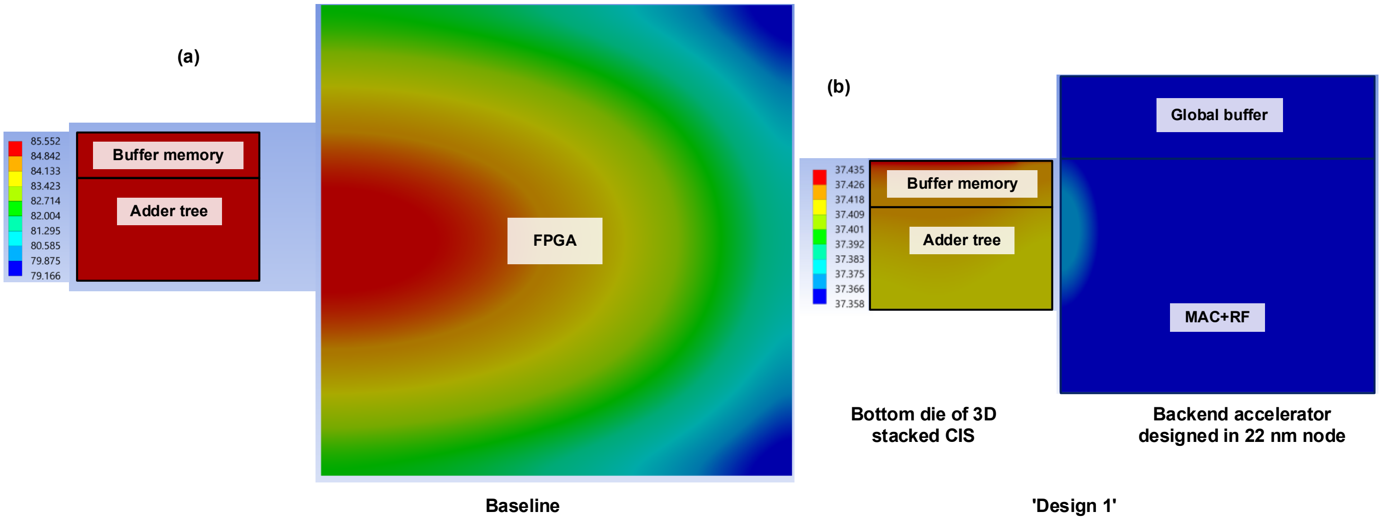

Fig. 6. (a) Thermal map for 2.5D integration of FPGA and CIS (from [10]). (b) Thermal map for the 2.5D integration of CIS and custom accelerator ("Design 1") described in Fig. 5 (a). The backend accelerator is designed in 22 nm node.

Next, we perform thermal simulations in Ansys Mechanical APDL. Table I summarizes the parameters used to perform the simulations. An ambient temperature of 35 °C is assumed, considering the elevated temperature during driving conditions. For design 1, the schematic described in Fig. 5(a) is used. Design 2 only requires the 3D stacked CIS on an organic substrate and the schematic is described in Fig. 5(b). Fig. 6 shows the thermal map obtained for different configurations. Although the heat generation is maximum in the backend accelerator, similar temperature throughout the package is due to high thermal coupling in such closely integrated systems. For design 1, we obtain a peak temperature of 37.4 °C, with the temperature being nearly the same throughout the package, as shown in Fig. 6(b). In case of design 1, the temperature is higher at the edge closer to the CIS due to higher coupling. In case of design 2, the temperature is higher in the MAC+RF part due to its higher power dissipation. If running at the same frequency of 200 MHz as the 22nm, Fig. 7 (a) shows the thermal map of design 2 as shown in the schematic described in Fig. 5 (b), indicating an excessive heat dissipation problem in the 3D integration. Use of a heat spreader in design 2 will prevent light from entering the photodiodes. Hence, lowering the operating frequency is necessary to lower the peak temperature. Fig. 7 (b) describes the thermal map of design 2 operated at a lowered 41 MHz to keep the peak temperature within 85 °C. The lower operating frequency results in an increase in the inference latency of the QDTrack network. Fig. 8 describes the obtained peak temperature for design 2 as a function of latency. From this curve, we can choose the operating frequency based on the thermal budget of the chip. For accurate multi-object tracking at 3 FPS, the latency should be less than 333 ms. This corresponds to an operating frequency of 41 MHz with a peak temperature of 85 °C. Table V summarizes the comparison of the designed accelerators in this work with other state-of-the-art accelerators for CIS. The proposed 2.5D design can run at a high frequency of 175 MHz to achieve a frame rate of 14.7 FPS while the peak temperature only increases by 2.75 °C. The 3D design can perform accurate multi-object tracking by achieving frame rate of 3 FPS and keep the temperature under 85 °C. As compared to other state-of-the-art CIS, our 3D stacked CIS is

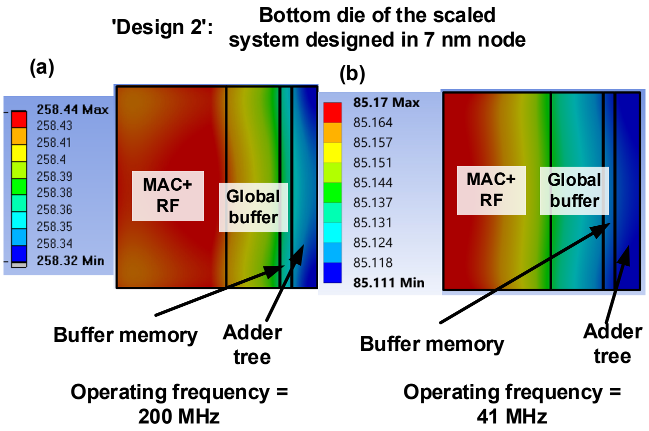

Fig. 7. (a) Thermal map for 3D integration of CIS and custom accelerator ("Design 2") described in Fig. 5 (b). The backend accelerator and the NPC circuit is scaled to 7 nm technology node and integrated together at the bottom tier using Cu-Cu hybrid bonding. The operating frequency is 200 MHz and the peak temperature is exceeding 250 °C. (b) Thermal map of "Design 2" when run at a lower frequency of 41 MHz to obtain peak temperature of 85 °C.

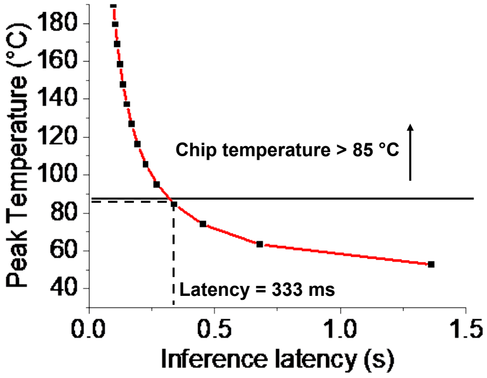

Fig. 8. The operating frequency is varied and a plot between the peak temperature vs. inference latency is obtained. Based on the required operating conditions with latency < 333 ms and peak temperature < 85 °C, we chose the operating frequency to be 41 MHz.

TABLE V: COMPARISON WITH OTHER STATE-OF-THE-ART NEAR-PIXEL COMPUTE BASED SENSORS

	ISSCC'17 [19]	**ISSCC'21 [6]**	**ISSCC'21 [20]**	**ISSCC'22 [21]**	**TCASI'22 [7]**	**This Work**	**This Work**
Technology	65 nm CMOS	22 nm CMOS	65 nm CMOS	180 nm CMOS	180 nm	22 nm CMOS	7 nm CMOS
Application	Face Detection and Recognition	Object Detection	Face Detection	Image Classification	Digit Clasification	Autonomous Driving	Autonomous Driving
Workload*	Approx. Conv, FC, Haar, MaxPool, Activation	ISP, Mobilenet_v1	Conv, Haar	Conv, FC, ReLU, MaxPool	FC, Activation	Conv, ReLU, MaxPool	Conv, ReLU, MaxPool
Thermal-aware design	No	No	No	No	No	Yes	Yes
Activation/weight precisions	-/1.5b	8b/8b	8b/1.5b	3b/1.5b	1b/1b	8b/8b	8b/8b
Area (mm^2)	16	62	4	5.37	4.7	8.7	8.7
Supply voltage (V)	0.8/2.5	0.8	0.8	0.8 V	0.8	0.9	0.75
Operating frequency (MHz)	100	262.5	-	5	-	200	41
Array Size	320×240	4056×3040	160×128	126×126	32×32	1296×720	1296×720
Processing type	Mixed-Signal	Digital	Mixed-Signal	Mixed-Signal	Analog	Digital	Digital
Compute Paradigm	Near-Pixel	Near-Pixel	In-Pixel	Near-Pixel	In-Pixel	Near-Pixel	Near-Pixel
Power consumption	620 µW	278.8 mW	42-106 µW	80.4 µW	147 nW-537 nW	206.3	21.3
Energy efficiency (TOPS/W)	1.24	4.97	0.15-3.64	-	4.7-17.3	10.3	45.81
Compute density (TOPS/mm^2)	0.03	0.02	3.7×10^{-6}-63×10^{-6}	-	5×10^{-7}	0.05	0.12

*: FC: Fully-Connected layer, Conv: Convolution operation, Approx. Conv: Separable convolution, MaxPool: Maxpooling operation, ReLU: ReLU activation function, ISP: Image Signal Processor, Activation: Activation function, Haar: Haar-like filter

the only thermal-aware design capable of performing a complex task like multi-object tracking for autonomous driving, while achieving a high energy efficiency of 45.81 TOPS/W.

V. CONCLUSION

To enable the intelligence of the camera, integration of machine learning hardware with the CIS is preferred. A sparsity-aware PE array-based accelerator is designed to perform inference on the BDD100k dataset using the QDTrack algorithm for autonomous driving. The designed backend accelerator is 2.5D integrated with a CIS with NPC for efficient data transfer. The accelerator is further scaled to an advanced technology node and 3D integrated with the CIS design with lower power consumption. The thermal impact of integration is performed for both designs. Finally, we scale down the frequency to accurately perform multi-object tracking within the thermal budget. We see a trade-off in terms of the overall latency, power consumption, and peak temperature among the two designs.

ACKNOWLEDGMENT

The authors would like to thank Yandong Luo, previously at Georgia Institute of Technology, for evaluation codes and Qiucheng Wu at the University of California, Santa Barbara for inputs on the algorithm.

REFERENCES

[1] S. Grigorescu *et al.*, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics (37)*, 362– 386, 2020.

[2] J. Pang *et al.*, "Quasi-dense similarity learning for multiple object tracking," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] W. Li *et al.*, "Temporal frame filtering with near-pixel compute for autonomous driving," *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS),* 2022.

[5] F. Sheikh *et al.*, "2.5D and 3D Heterogeneous Integration: Emerging applications," in *IEEE Solid-State Circuits Magazine*, 2021.

[6] R. Eki *et al.*, "A 1/2.3inch 12.3Mpixel with on-chip 4.97TOPS/W CNN processor back-illuminated stacked CMOS image sensor," *IEEE International Solid-State Circuits Conference (ISSCC),* 2021.

[7] H. Xu *et al.*, "Senputing: An ultra-low-power always-on vision perception chip featuring the deep fusion of sensing and computing," in *IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)*, vol. 69, no. 1, pp. 232-243, 2022.

[8] F. Zhou *et al.*, "Near-sensor and in-sensor computing," *Nature Electronics*, 3.11, 664-671, 2020.

[9] J. Sharda *et al.*, "Temporal frame filtering for autonomous driving using 3D-stacked global shutter CIS with IWO buffer memory and near-pixel compute." *IEEE Transactions on Circuits and Systems I: Regular Papers,* 2023.

[10] J. Sharda *et al.*, "Thermal Modeling of 2.5D Integrated Package of CMOS Image Sensor and FPGA for Autonomous Driving," *IEEE Electron Devices Technology & Manufacturing Conference (EDTM),* 2023.

[11] H. Kwon *et al.*, "MAESTRO: A Data-Centric Approach to Understand Reuse, Performance, and Hardware Cost of DNN Mappings," *IEEE Micro*, 2020.

[12] H. Sugo *et al.,* "A dead-time free global shutter CMOS image sensor with in-pixel LOFIC and ADC using pixel-wis e connections," *IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, 2016.

[13] H. Ye *et al.*, "Double-gate W-doped amorphous indium oxide transistors for monolithic 3D capacitorless gain cell eDRAM," *IEEE International Electron Devices Meeting (IEDM)*, 2020.

[14] H. Fujiwara *et al.,* "A 5-nm 254-TOPS/W 221-TOPS/mm2 Fully-Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations," *IEEE International Solid- State Circuits Conference* (ISSCC), 2022.

[15] S. Jangam *et al.*, "Electrical Characterization of High Performance Fine Pitch Interconnects in Silicon-Interconnect Fabric," *IEEE Electronic Components and Technology Conference (ECTC)*, 2018.

[16] Y. Zhang *et al.*, "Data-Model-Circuit Tri-Design for Ultra-Light Video Intelligence on Edge Devices," *ACM Asia and South Pacific Design Automation Conference* (ASPDAC), 2023.

[17] A. Elsherbini *et al.*, "Enabling Next Generation 3D Heterogeneous Integration Architectures on Intel Process," *International Electron Devices Meeting* (IEDM), 2022

[18] X. Peng *et. al.*, "DNN+NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," *IEEE International Electron Devices Meeting (IEDM)*, 2019.

[19] K. Bong *et al.*, "A 0.62mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on Haar-like face detector," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017

[20] M. Lefebvre *et al.*, "A 0.2-to-3.6TOPS/W programmable convolutional imager SoC with in-sensor current-domain ternary-weighted MAC operations for feature extraction and region-of-interest detection," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2021.

[21] T.-H. Hsu *et al.*, "A 0.8V intelligent vision sensor with tiny convolutional neural network and programmable weights using mixed-mode processing-in-sensor technique for image classification," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2022