Lincoln AI Computing Survey (LAICS) Update

Albert Reuther, Peter Michaleas, Michael Jones,
Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner
MIT Lincoln Laboratory Supercomputing Center
Lexington, MA, USA
{reuther,pmichaleas,michael.jones,vijayg,sid,kepner}@ll.mit.edu

Abstract—**This paper is an update of the survey of AI accelerators and processors from past four years, which is now called the Lincoln AI Computing Survey – LAICS (pronounced “lace”). As in past years, this paper collects and summarizes the current commercial accelerators that have been publicly announced with peak performance and peak power consumption numbers. The performance and power values are plotted on a scatter graph, and a number of dimensions and observations from the trends on this plot are again discussed and analyzed. Market segments are highlighted on the scatter plot, and zoomed plots of each segment are also included. Finally, a brief description of each of the new accelerators that have been added in the survey this year is included.**

Index Terms—**Machine learning, GPU, TPU, tensor, dataflow, CGRA, accelerator, embedded inference, computational performance**

I. Introduction

A number of announcements, releases, and deployments of artificial intelligence (AI) accelerators from startups and established technology companies have occurred in the past year. Perhaps most notable is the emergence of very large foundation models that are able to generate prose, poetry, images, etc. based on training using vast amounts of data usually collected via internet data crawls. Much technical press has been focused on how effective the resulting tools will be for various tasks, but also there is much discussion about the training of these models. But from an accelerator perspective, it is the very same accelerators that are aimed towards training more modestly sized models that are used for training these very large models. The very large models are just using many more accelerators simultaneously in a synchronous parallel manner, and they are interconnected with very high bandwidth networks. But beyond that news, not much has changed in the overall trends and landscape. Hence, this paper just updates what was discussed in last year’s survey.

For much of the background of this study, please refer to one of the previous IEEE-HPEC papers that our team has published [1]–[4]. This background includes an explanation of the AI ecosystem architecture, the history of the emergence of AI accelerators and accelerators in general, a more detailed explanation of the survey scatter plots, and a discussion of broader observations and trends.

II. Survey of Processors

This paper is an update to IEEE-HPEC papers from the past four years [1]–[4]. This survey continues to cast a wide net to include accelerators and processors for a variety of applications including defense and national security AI/ML edge applications. The survey collects information on all of the numerical precision types that an accelerator supports, but for most of them, their best inference performance is in int8 or fp16/bf16, so that is what usually is plotted. This survey gathers performance and power information from publicly available materials including research papers, technical trade press, company benchmarks, etc. The key metrics of this public data are plotted in Figure 1, which graphs recent processor capabilities (as of Summer 2023) mapping peak performance vs. power consumption, and Table I summarizes some of the important metadata of the accelerators, cards, and systems, including the labels used in Figure 1.

The x-axis indicates peak power, and the y-axis indicate peak giga-operations per second (GOps/s), both on a logarithmic scale. The computational precision of the processing capability is depicted by the geometric marker used. The form factor is depicted by color, which shows the package for which peak power is reported. Blue corresponds to a single chip; orange corresponds to a card; and green corresponds to entire systems (single node desktop and server systems). Finally, the hollow geometric objects are peak performance for inference-only accelerators, while the solid geometric figures are performance for accelerators that are designed to perform both training and inference.

A reasonable categorization of accelerators follows their intended application, and the five categories are shown as ellipses on the graph, which roughly correspond to performance and power consumption: Very Low Power for wake word detection, speech processing, very small sensors, etc.; Embedded for cameras, small UAVs and robots, etc.; Autonomous for driver assist services, autonomous driving, and autonomous robots; Data Center Chips and Cards; and Data Center Systems. A zoomed in scatter plot for each of these categories is shown in the subfigures of Figure 2.

For most of the accelerators, their descriptions and commentaries have not changed since last year so please refer to the papers of the last four years for descriptions and commentaries. Several new releases are included in this update.

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

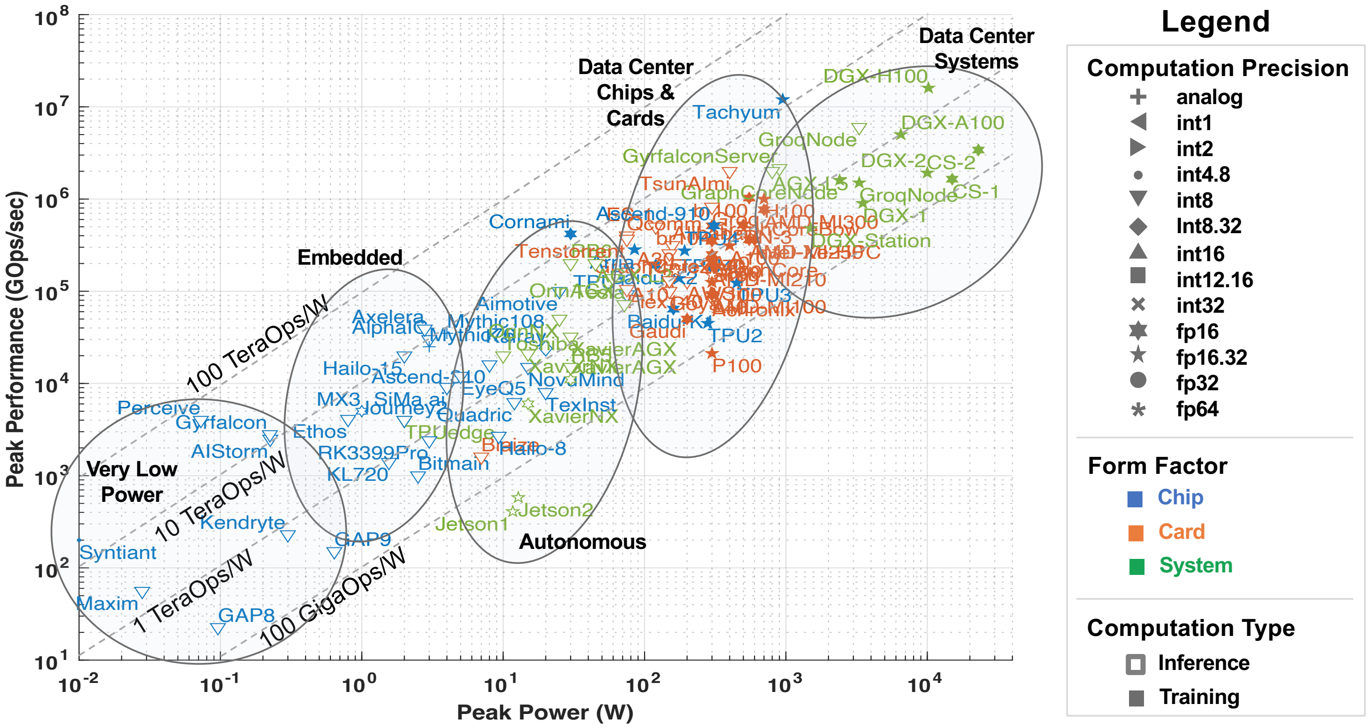
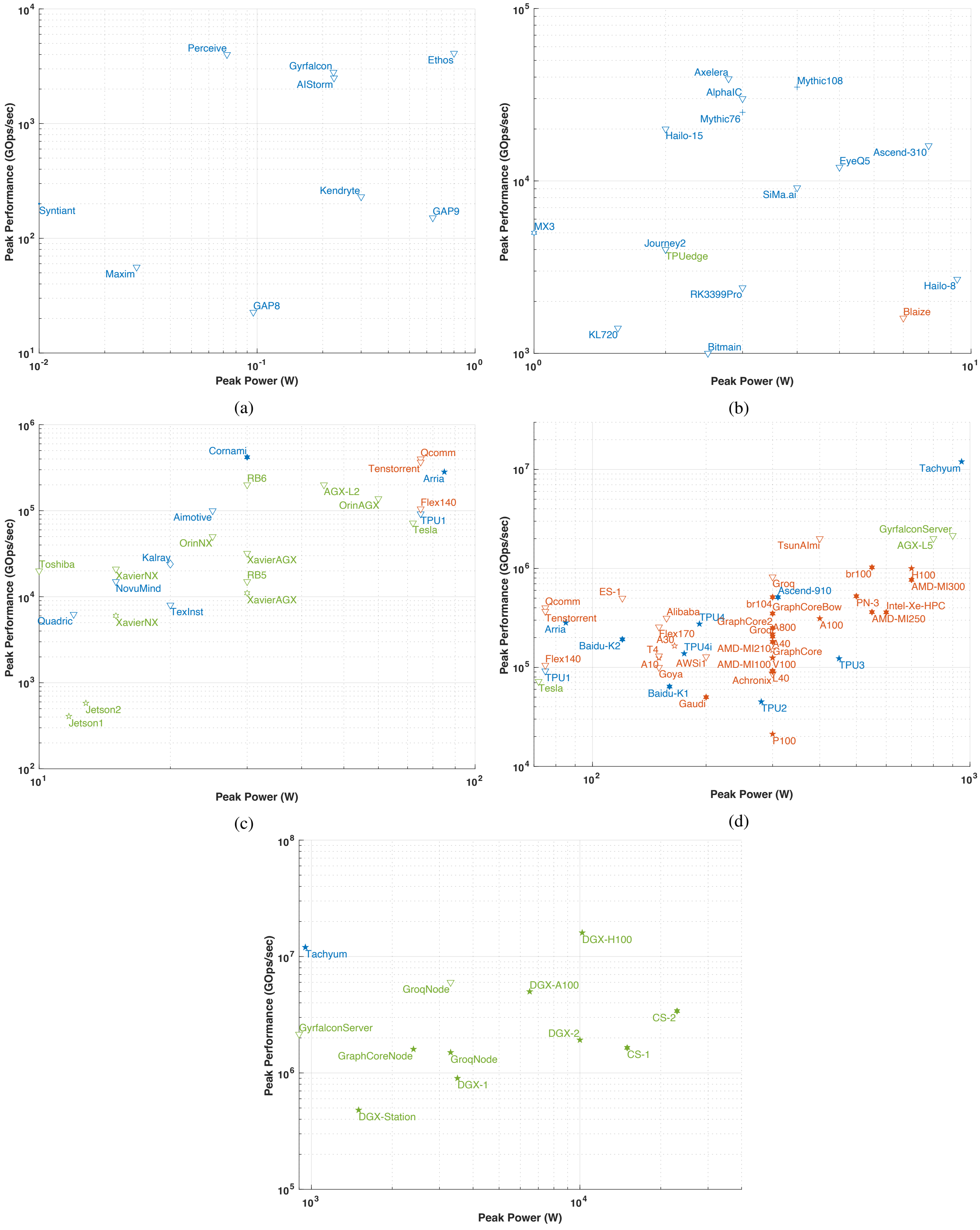


Fig. 1: Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

- Based on similar technology of its Cloud AI 100 accelerator, Qualcomm has released two versions of its robotics AI system platform, the RB5 and RB6, in the past few years. Both are competing in the same low power system-on-a-chip market as the NVIDIA Jetson product line, and are aimed at integration in applications including robotics, driver assist, modest UAVs, etc. [99], [100].
- The Memryx MX3 AI accelerator chip is a startup that was spun out of the University of Michigan. It is designed to be deployed with a host CPU to greatly speed up AI inference, consuming about 1W of power. It computes activations with bf16 numerical precision, and store model parameter weight at 4-bit, 8-bit, and 16-bit integer precisions, which can be set on a layer-by-layer basis [68], [69].
- On the heels of it's Hailo-8 AI accelerator, Hailo has released a lower power variant, the Hailo-15. The Hailo-15 targets the Internet Protocol (IP) camera market, and it is a SoC that includes a CPU, a digital signal processor (DSP) accelerator, and a neural accelerator, which all draw less than 2W [52].
- Startup Esperanto Technologies has released their first processor accelerator called the ET-SoC-1. Each chip is comprised of 1,088 64-bit ET-Minion RISC-V cores, each of which have scalar, vector, and tensor units along with L1 cache/scratchpad memory. Their key application is training and inference for recommender systems, which have a balanced mix of scalar, vector, and tensor operations [30], [31].
- Baidu has started deploying its second-generation Kunlun accelerator, Kunlun II. Baidu touted that the Kunlun II is 2-3 times faster than the original Kunlun [20].
- The Chinese GPU startup Biren emerged from stealth mode to announce and release two high performance GPUs: the BR100 and BR104. The BR104 is a single die GPU, while the BR100 combines two dies/chiplets in the same package [21], [22].
- AMD has announced the followup to their Instinct MI250 GPU called the Instinct MI300A, which will be a multi-chiplet CPU-GPU Accelerated Processing Unit (APU) integrated package. The announcement showed package photos of two CPU dies integrated with six GPU dies. [13], [14]
- While Intel announced their high-end AI GPU a few years ago, details continued to be scarce until this past year. Enough performance numbers were announced for the Intel Xe-HPC (codename Ponte Vecchio) to include it in this year's survey [58]–[60]. Along with the Xe-HPC, Intel also announced and started shipping two inference-oriented GPU cards, the Flex 140 and Flex 170 [61].
- After announcing and shipping their Hopper H100 GPUs in systems at the end of 2022, NVIDIA has started shipping DGX servers, which integrate eight H100 GPUs [81]. NVIDIA has also released a high-performance Ampere GPU, the A800, that is aimed at the Chinese market, which reportedly performs at approximately 70% peak performance of the A100 [77]. Finally, NVIDIA has released a new Ada Lovelace GPU family which is aimed at the data center inference and graphics rendering (gaming) farm markets. The first specifications were released for the L40 GPU, which are included in this survey [88].



(a)

(b)

(c)

(d)

(e)

Fig. 2: Zoomed regions of peak performance vs. peak power scatter plot: **(a)** very low power, **(b)** embedded, **(c)** autonomous, **(d)** data center chips and cards, **(e)** data center systems.

TABLE I: List of accelerator metadata and labels for plots.

Company	Product	Label	Technology	Form Factor	References
Achronix	VectorPath S7t-VG6	Achronix	FPGA	Card	[5]
Aimotive	aiWare3	Aimotive	dataflow	Chip	[6]
AIStorm	AIStorm	AIStorm	dataflow	Chip	[7]
Alibaba	HanGuang 800	Alibaba	dataflow	Card	[8]
AlphaIC	RAP-E	AlphaIC	dataflow	Chip	[9]
Amazon	Inferentia	AWSi1	dataflow	Card	[10], [11]
AMD	MI100	AMD-MI100	GPU	Card	[12]
AMD	MI210	AMD-MI210	GPU	Card	[12]
AMD	MI250	AMD-MI250	GPU	Card	[12]
AMD	MI300	AMD-MI300	GPU	Card	[13], [14]
ARM	Ethos N77	Ethos	dataflow	Chip	[15]
Axelera	Axelera Test Core	Axelera	dataflow	Chip	[16]
Baidu	Baidu Kunlun 200	Baidu-K1	dataflow	Chip	[17]–[19]
Baidu	Baidu Kunlun II	Baidu-K2	dataflow	Chip	[20]
Biren Technology	br100	br100	GPU	Card	[21], [22]
Biren Technology	br104	br104	GPU	Card	[21], [22]
Bitmain	BM1880	Bitmain	dataflow	Chip	[23]
Blaize	El Cano	Blaize	dataflow	Card	[24]
Canaan	Kendrite K210	Kendryte	CPU	Chip	[25]
Cerebras	CS-1	CS-1	dataflow	System	[26]
Cerebras	CS-2	CS-2	dataflow	System	[27]
Cornami	Cornami	Cornami	dataflow	Chip	[28]
Enflame	Cloudblazer T10	Enflame	CPU	Card	[29]
Esperanto	ET-SoC-1	ES-1	CPU	Card	[30], [31]
Google	TPU Edge	TPUedge	tensor	System	[32]
Google	TPU1	TPU1	tensor	Chip	[33], [34]
Google	TPU2	TPU2	tensor	Chip	[33], [34]
Google	TPU3	TPU3	tensor	Chip	[33]–[35]
Google	TPU4i	TPU4i	tensor	Chip	[35]
Google	TPU4	TPU4	tensor	Chip	[36]
GraphCore	C2	GraphCore	dataflow	Card	[37], [38]
GraphCore	C2	GraphCoreNode	dataflow	System	[39]
GraphCore	Colossus Mk2	GraphCore2	dataflow	Card	[40]
GraphCore	Bow-2000	GraphCoreBow	dataflow	Card	[41]
GreenWaves	GAP8	GAP8	dataflow	Chip	[42], [43]
GreenWaves	GAP9	GAP9	dataflow	Chip	[42], [43]
Groq	Groq Node	GroqNode	dataflow	System	[44]
Groq	Groq Node	GroqNode	dataflow	System	[44]
Groq	Tensor Streaming Processor	Groq	dataflow	Card	[37], [45]
Groq	Tensor Streaming Processor	Groq	dataflow	Card	[37], [45]
Gyrfalcon	Gyrfalcon	Gyrfalcon	PIM	Chip	[46]
Gyrfalcon	Gyrfalcon	GyrfalconServer	PIM	System	[47]
Habana	Gaudi	Gaudi	dataflow	Card	[48], [49]
Habana	Goya HL-1000	Goya	dataflow	Card	[49], [50], [50]
Hailo	Hailo-8	Hailo-8	dataflow	Chip	[51]
Hailo	Hailo-15H	Hailo-15	dataflow	Chip	[52]
Horizon Robotics	Journey2	Journey2	dataflow	Chip	[53]
Huawei HiSilicon	Ascend 310	Ascend-310	dataflow	Chip	[54]
Huawei HiSilicon	Ascend 910	Ascend-910	dataflow	Chip	[55]
Intel	Arria 10 1150	Arria	FPGA	Chip	[56], [57]
Intel	Mobileye EyeQ5	EyeQ5	dataflow	Chip	[24]
Intel	Xe-HPC	Intel-Xe-HPC	GPU	Card	[58]–[60]
Intel	Flex140	Flex140	GPU	Card	[61]
Intel	Flex170	Flex170	GPU	Card	[61]
Kalray	Coolidge	Kalray	manycore	Chip	[62], [63]
Kneron	KL720	KL720	dataflow	Chip	[64]
Maxim	Max 78000	Maxim	dataflow	Chip	[65]–[67]
MemryX	MX3	MX3	dataflow	Chip	[68], [69]
Mythic	M1076	Mythic76	PIM	Chip	[70]–[72]
Mythic	M1108	Mythic108	PIM	Chip	[70]–[72]
NovuMind	NovuTensor	NovuMind	dataflow	Chip	[73], [74]
NVIDIA	Ampere A10	A10	GPU	Card	[75]
NVIDIA	Ampere A100	A100	GPU	Card	[76]
NVIDIA	Ampere A800	A800	GPU	Card	[77]
NVIDIA	Ampere A30	A30	GPU	Card	[75]
NVIDIA	Ampere A40	A40	GPU	Card	[75]
NVIDIA	DGX Station	DGX-Station	GPU	System	[78]
NVIDIA	DGX-1	DGX-1	GPU	System	[78], [79]
NVIDIA	DGX-2	DGX-2	GPU	System	[79]
NVIDIA	DGX-A100	DGX-A100	GPU	System	[80]
NVIDIA	DGX-H100	DGX-H100	GPU	System	[81]
NVIDIA	H100	H100	GPU	Card	[82]
NVIDIA	Jetson AGX Xavier	XavierAGX	GPU	System	[83]
NVIDIA	Jetson NX Orin	OrinNX	GPU	System	[84], [85]
NVIDIA	Jetson AGX Orin	OrinAGX	GPU	System	[84], [85]
NVIDIA	Jetson TX1	Jetson1	GPU	System	[86]
NVIDIA	Jetson TX2	Jetson2	GPU	System	[86]
NVIDIA	Jetson Xavier NX	XavierNX	GPU	System	[83]
NVIDIA	DRIVE AGX L2	AGX-L2	GPU	System	[87]
NVIDIA	DRIVE AGX L5	AGX-L5	GPU	System	[87]
NVIDIA	L40	L40	GPU	Card	[88]
NVIDIA	Pascal P100	P100	GPU	Card	[89], [90]
NVIDIA	T4	T4	GPU	Card	[91]
NVIDIA	Volta V100	V100	GPU	Card	[90], [92]
Perceive	Ergo	Perceive	dataflow	Chip	[93]
Preferred Networks	MN-3	PN-3	multicore	Card	[94], [95]
Quadric	q1-64	Quadric	dataflow	Chip	[96]
Qualcomm	Cloud AI 100	Qcomm	dataflow	Card	[97], [98]
Qualcomm	QRB5165	RB5	GPU	System	[99]
Qualcomm	QRB5165N	RB6	GPU	System	[100]
Rockchip	RK3399Pro	RK3399Pro	dataflow	Chip	[101]
SiMa.ai	SiMa.ai	SiMa.ai	dataflow	Chip	[102]
Syntiant	NDP101	Syntiant	PIM	Chip	[103], [104]
Tachyum	Prodigy	Tachyum	CPU	Chip	[105]
Tenstorrent	Tenstorrent	Tenstorrent	multicore	Card	[106]
Tesla	Tesla FSC	Tesla	dataflow	System	[107], [108]
Texas Instruments	TDA4VM	TexInst	dataflow	Chip	[109]–[111]
Toshiba	2015	Toshiba	multicore	System	[112]
Untether	TsunAImi	TsunAImi	PIM	Card	[113]

III. Summary

This paper updates the Lincoln AI Computing Survey (LAICS) of deep neural network accelerators that span from extremely low power through embedded and autonomous applications to data center class accelerators for inference and training. We presented the new full scatter plot along with zoomed in scatter plots for each of the major deployment/market segments, and we discussed some new additions for the year. The rate of announcements and releases has continued to be consistent as companies compete for various embedded, data center, cloud, and on-premises HPC deployments.

IV. Data Availability

The data spreadsheets and references that have been collected for this study and its papers will be posted at https://github.com/areuther/ai-accelerators after they have cleared the release review process.

Acknowledgement

We express our gratitude to Masahiro Arakawa, Bill Arcand, Bill Bergeron, David Bestor, Bob Bond, Chansup Byun, Nathan Frey, Vitaliy Gleyzer, Jeff Gottschalk, Michael Houle, Matthew Hubbell, Hayden Jananthan, Anna Klein, David Martinez, Joseph McDonald, Lauren Milechin, Sanjeev Mohindra, Paul Monticciolo, Julie Mullen, Andrew Prout, Stephan Rejto, Antonio Rosa, Charles Yee, and Marc Zissman for their support of this work.

References

[1] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Ai and ml accelerator survey and trends." IEEE, 9 2022, pp. 1–10.

[2] ——, "Ai accelerator survey and trends," 9 2021, pp. 1–9.

[3] ——, "Survey of machine learning accelerators," 2020, pp. 1–12.

[4] ——, "Survey and benchmarking of machine learning accelerators." Institute of Electrical and Electronics Engineers Inc., 9 2019. [Online]. Available: https://doi.org/10.1109/HPEC.2019.8916327

[5] G. Roos, "Fpga acceleration card delivers on bandwidth, speed, and flexibility," 11 2019. [Online]. Available: https://www.eetimes.com/fpga-acceleration-card-delivers-on-bandwidth-speed-and-flexibility/

[6] "aiware3 hardware ip helps drive autonomous vehicles to production," 10 2018. [Online]. Available: https://aimotive.com/news/content/1223

[7] R. Merritt, "Startup accelerates ai at the sensor," 2 2019. [Online]. Available: https://www.eetimes.com/startup-accelerates-ai-at-the-sensor/

[8] T. Peng, "Alibabas new ai chip can process nearly 80k images per second," 2019. [Online]. Available: https://medium.com/syncedreview/alibabas-new-ai-chip-can-process-nearly-80k-images-per-second-63412dec22a3

[9] P. Clarke, "Indo-us startup preps agent-based ai processor," 8 2018. [Online]. Available: https://www.eenewsanalog.com/news/indo-us-startup-preps-agent-based-ai-processor/page/0/1

[10] J. Hamilton, "Aws inferentia machine learning processor," 11 2018. [Online]. Available: https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor/

[11] C. Evangelist, "Deep dive into amazon inferentia: A custom-built chip to enhance ml and ai," 1 2020. [Online]. Available: https://www.cloudmanagementinsider.com/amazon-inferentia-for-machine-learning-and-artificial-intelligence/

[12] R. Smith, "Amd announces instinct mi200 accelerator family: Taking servers to exascale and beyond," 11 2021. [Online]. Available: https://www.anandtech.com/show/17054/amd-announces-instinct-mi200-accelerator-family-cdna2-exacale-servers

[13] T. P. Morgan, "Amd teases details on future mi300 hybrid compute engines," 1 2023.

[14] ——, "The third time charm of amds instinct gpu," 6 2023. [Online]. Available: https://www.nextplatform.com/2023/06/14/the-third-time-charm-of-amds-instinct-gpu/

[15] D. Schor, "Arm ethos is for ubiquitous ai at the edge," 2 2020. [Online]. Available: https://fuse.wikichip.org/news/3282/arm-ethos-is-for-ubiquitous-ai-at-the-edge/

[16] S. Ward-Foxton, "Axelera demos ai test chip after taping out in four months," 5 2022. [Online]. Available: https://www.eetimes.com/axelera-demos-ai-test-chip-after-taping-out-in-four-months/

[17] J. Ouyang, X. Du, Y. Ma, and J. Liu, "Kunlun: A 14nm high-performance ai processor for diversified workloads," vol. 64, 2 2021, pp. 50–51.

[18] R. Merritt, "Baidu accelerator rises in ai," 7 2018. [Online]. Available: https://www.eetimes.com/baidu-accelerator-rises-in-ai/

[19] C. Duckett, "Baidu creates kunlun silicon for ai," 7 2018. [Online]. Available: https://www.zdnet.com/article/baidu-creates-kunlun-silicon-for-ai/

[20] A. Shilov, "Baidu unveils kunlun ii ai chip: Rival for nvidia a100," 8 2021. [Online]. Available: https://www.tomshardware.com/news/baidu-unveils-kunlun-ii-processor-for-ai

[21] O. Peckham, "Chinese startup biren details br100 gpu," 8 2022. [Online]. Available: https://www.hpcwire.com/2022/08/22/chinese-startup-biren-details-br100-gpu/

[22] A. Shilov, "Chinese gpu firm biren plans ipo to better compete against nvidia," 7 2023. [Online]. Available: https://www.tomshardware.com/news/biren-mulls-ipo

[23] B. Wheeler, "Bitmain soc brings ai to the edge," 2 2019. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php%3Fnum=5975%26year=2019%26tag=3

[24] M. Demler, "Blaize ignites edge-ai performance," pp. 1–5, 9 2020. [Online]. Available: https://www.blaize.com/wp-content/uploads/2020/09/Blaize-Ignites-Edge-AI-Performance.pdf

[25] L. Gwennap, "Kendryte embeds ai for surveillance," 3 2019. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=5992

[26] A. Hock, "Introducing the cerebras cs-1, the industrys fastest artificial intelligence computer," 11 2019. [Online]. Available: https://www.cerebras.net/introducing-the-cerebras-cs-1-the-industrys-fastest-artificial-intelligence-computer/

[27] T. Trader, "Cerebras doubles ai performance with second-gen 7nm wafer scale engine," 4 2021. [Online]. Available: https://www.hpcwire.com/2021/04/20/cerebras-doubles-ai-performance-with-second-gen-7nm-wafer-scale-engine/

[28] "Cornami achieves unprecedented performance at lowest power dissipation for deep neural networks," 10 2019. [Online]. Available: https://cornami.com/1416-2/

[29] P. Clarke, "Globalfoundries aids launch of chinese ai startup," 12 2019. [Online]. Available: https://www.eenewsanalog.com/news/globalfoundries-aids-launch-chinese-ai-startup

[30] D. R. Ditzel and the Esperanto team, "Accelerating ml recommendation with over 1,000 risc-v/tensor processors on esperanto's et-soc-1 chip," *IEEE Micro*, vol. 42, pp. 31–38, 5 2022.

[31] D. Schor, "A look at the et-soc-1, esperantos massively multi-core risc-v approach to ai," 7 2021. [Online]. Available: https://fuse.wikichip.org/news/4911/a-look-at-the-et-soc-1-esperantos-massively-multi-core-risc-v-approach-to-ai/

[32] "Edge tpu," 2019. [Online]. Available: https://cloud.google.com/edge-tpu/

[33] N. P. Jouppi, D. H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson, "A domain-specific supercomputer for training deep neural networks," *Commun. ACM*, vol. 63, p. 6778, 6 2020. [Online]. Available: https://doi.org/10.1145/3360307

[34] P. Teich, "Tearing apart google's tpu 3.0 ai coprocessor," 5 2018. [Online]. Available: https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/

[35] N. P. Jouppi, D. H. Yoon, M. Ashcraft, M. Gottscho, T. B. Jablin, G. Kurian, J. Laudon, S. Li, P. Ma, X. Ma, T. Norrie, N. Patil, S. Prasad, C. Young, Z. Zhou, D. Patterson, and G. Llc, "Ten lessons from three generations shaped google's tpuv4i." IEEE Computer Society, 6 2021, pp. 1–14.

[36] O. Peckham, "Google cloud's new tpu v4 ml hub packs 9 exaflops of ai," 5 2022. [Online]. Available: https://www.hpcwire.com/2022/05/16/google-clouds-new-tpu-v4-ml-hub-packs-9-exaflops-of-ai/

[37] L. Gwennap, "Groq rocks neural networks," 1 2020. [Online]. Available: http://groq.com/wp-content/uploads/2020/04/Groq-Rocks-NNs-Linley-Group-MPR-2020Jan06.pdf

[38] D. Lacey, "Preliminary ipu benchmarks," 10 2017. [Online]. Available: https://www.graphcore.ai/posts/preliminary-ipu-benchmarks-providing-previously-unseen-performance-for-a-range-of-machine-learning-applications

[39] "Dell dss8440 graphcore ipu server," 2 2020. [Online]. Available: https://www.graphcore.ai/hubfs/Leadgenassets/DSS8440IPUServerWhitePaper_2020.pdf

[40] S. Ward-Foxton, "Graphcore takes on nvidia with second-gen ai accelerator," 7 2020. [Online]. Available: https://www.eetimes.com/graphcore-takes-on-nvidia-with-second-gen-ai-accelerator/

[41] M. Tyson, "Graphcore bow ipu introduces tsmc 3d wafer-on-wafer processor," 3 2022. [Online]. Available: https://www.tomshardware.com/news/graphcore-tsmc-bow-ipu-3d-wafer-on-wafer-processor

[42] "Gap application processors," 2020. [Online]. Available: https://greenwaves-technologies.com/gap8_gap9/

[43] J. Turley, "Gap9 for ml at the edge," 6 2020. [Online]. Available: https://www.eejournal.com/article/gap9-for-ml-at-the-edge/

[44] N. Hemsoth, "Groq shares recipe for tsp nodes, systems," 9 2020. [Online]. Available: https://www.nextplatform.com/2020/09/29/groq-shares-recipe-for-tsp-nodes-systems/

[45] D. Abts, J. Ross, J. Sparling, M. Wong-VanHaren, M. Baker, T. Hawkins, A. Bell, J. Thompson, T. Kahsai, G. Kimmell, J. Hwang, R. Leslie-Hurd, M. Bye, E. R. Creswick, M. Boyd, M. Venigalla, E. Laforge, J. Purdy, P. Kamath, D. Maheshwari, M. Beidler, G. Rosseel, O. Ahmad, G. Gagarin, R. Czekalski, A. Rane, S. Parmar, J. Werner, J. Sproch, A. Macias, and B. Kurtz, "Think fast: A tensor streaming processor (tsp) for accelerating deep learning workloads," 5 2020, pp. 145–158. [Online]. Available: https://doi.org/10.1109/ISCA45697.2020.00023

[46] S. Ward-Foxton, "Gyrfalcon unveils fourth ai accelerator chip — ee times," 11 2019. [Online]. Available: https://www.eetimes.com/gyrfalcon-unveils-fourth-ai-accelerator-chip/

[47] "Solidrun, gyrfalcon develop arm-based edge optimized ai inference server," 2 2020. [Online]. Available: https://www.hpcwire.com/off-the-wire/solidrun-gyrfalcon-develop-edge-optimized-ai-inference-server/

[48] L. Gwennap, "Habana offers gaudi for ai training," 6 2019. [Online]. Available: https://habana.ai/wp-content/uploads/2019/06/Habana-Offers-Gaudi-for-AI-Training.pdf

[49] E. Medina and E. Dagan, "Habana labs purpose-built ai inference and training processor architectures: Scaling ai training systems using standard ethernet with gaudi processor," *IEEE Micro*, vol. 40, pp. 17–24, 3 2020. [Online]. Available: https://doi.org/10.1109/MM.2020.2975185

[50] L. Gwennap, "Habana wins cigar for ai inference," 2 2019. [Online]. Available: https://www.linleygroup.com/mpr/article.php?id=12103

[51] S. Ward-Foxton, "Details of hailo ai edge accelerator emerge," 8 2019. [Online]. Available: https://www.eetimes.com/details-of-hailo-ai-edge-accelerator-emerge/

[52] ——, "Hailo adds vision processor socs for smart cameras," 3 2023. [Online]. Available: https://www.eetimes.com/hailo-adds-vision-processor-socs-for-smart-cameras/

[53] "Horizon robotics journey2 automotive ai processor series," 2020. [Online]. Available: https://en.horizon.ai/product/journey

[54] Huawei, "Ascend 310 ai processor," 2020. [Online]. Available: https://e.huawei.com/us/products/cloud-computing-dc/atlas/ascend-310

[55] ——, "Ascend 910 ai processor," 2020. [Online]. Available: https://e.huawei.com/us/products/cloud-computing-dc/atlas/ascend-910

[56] M. S. Abdelfattah, D. Han, A. Bitar, R. DiCecco, S. O'Connell, N. Shanker, J. Chu, I. Prins, J. Fender, A. C. Ling, and G. R. Chiu, "Dla: Compiler and fpga overlay for neural network inference acceleration," 8 2018, pp. 411–4117. [Online]. Available: https://doi.org/10.1109/FPL.2018.00077

[57] N. Hemsoth, "Intel fpga architecture focuses on deep learning inference," 7 2018. [Online]. Available: https://www.nextplatform.com/2018/07/31/intel-fpga-architecture-focuses-on-deep-learning-inference/

[58] A. Shilov, "Intel's ponte vecchio xe-hpc gpu boasts 100 billion transistors," 3 2021. [Online]. Available: https://www.tomshardware.com/news/intel-xe-hpc-ponte-vecchio-examined

[59] "Intel provides details about sapphire rapids cpu and ponte vecchio gpu," 8 2021. [Online]. Available: https://www.hpcwire.com/off-the-wire/intel-unveils-details-about-sapphire-rapids-cpu-ponte-vecchio-gpu-ipu/

[60] A. Shilov, “Intel’s ponte vecchio is finally in the wild,” 6 2022. [Online]. Available: https://www.tomshardware.com/news/intels-ponte-vecchio-smiles-for-the-camera

[61] T. P. Morgan, “Different gpu horses for different datacenter courses,” 10 2022. [Online]. Available: https://www.nextplatform.com/2022/10/04/different-gpu-horses-for-different-datacenter-courses/

[62] B. D. de Dinechin, “Kalrays mppa manycore processor: At the heart of intelligent systems.” IEEE, 6 2019. [Online]. Available: https://www.european-processor-initiative.eu/dissemination-material/1259/

[63] P. Clarke, “Nxp, kalray demo coolidge parallel processor in ’bluebox’,” 1 2020. [Online]. Available: https://www.eenewsanalog.com/news/nxp-kalray-demo-coolidge-parallel-processor-bluebox

[64] S. Ward-Foxton, “Kneron attracts strategic investors,” 1 2021. [Online]. Available: https://www.eetimes.com/kneron-attracts-strategic-investors/

[65] ——, “Maxim debuts homegrown ai accelerator in latest ulp soc,” 11 2020. [Online]. Available: https://www.eetimes.com/maxim-debuts-homegrown-ai-accelerator-in-latest-ulp-soc/

[66] A. Jani, “Maxim showcases efficient custom ai,” 2 2021. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=6274&year=2021&tag=3

[67] M. Clay, C. Grecos, M. Shirvaikar, and B. Richey, “Benchmarking the max78000 artificial intelligence microcontroller for deep learning applications,” N. Kehtarnavaz and M. F. Carlsohn, Eds., vol. 12102. SPIE, 2022, pp. 47–52. [Online]. Available: https://doi.org/10.1117/12.2622390

[68] S. Leibson, “Adding low-power ai/ml inference to edge devices,” 4 2023. [Online]. Available: https://www.eetimes.com/adding-low-power-ai-ml-interference-to-edge-devices/

[69] S. Vicinanza, “Start-up memryx makes mx3 an ai accelerator that rivals established offerings,” 12 2022. [Online]. Available: https://circuitcellar.com/newsletter/start-up-memryx-makes-mx3-an-ai-accelerator-that-rivals-established-offerings/

[70] S. Ward-Foxton, “Mythic resizes its ai chip,” 6 2021. [Online]. Available: https://www.eetimes.com/mythic-resizes-its-analog-ai-chip/

[71] N. Hemsoth, “A mythic approach to deep learning inference,” 8 2018. [Online]. Available: https://www.nextplatform.com/2018/08/23/a-mythic-approach-to-deep-learning-inference/

[72] D. Fick, “Mythic @ hot chips 2018,” 8 2018. [Online]. Available: https://medium.com/mythic-ai/mythic-hot-chips-2018-637dfb9e38b7

[73] K. Freund, “Novumind: An early entrant in ai silicon,” 5 2019. [Online]. Available: https://moorinsightsstrategy.com/wp-content/uploads/2019/05/NovuMind-An-Early-Entrant-in-AI-Silicon-By-Moor-Insights-And-Strategy.pdf

[74] J. Yoshida, “Novuminds ai chip sparks controversy,” 10 2018. [Online]. Available: https://www.eetimes.com/novuminds-ai-chip-sparks-controversy/

[75] T. P. Morgan, “Nvidia rounds out ”ampere” lineup with two new accelerators,” 4 2021. [Online]. Available: https://www.nextplatform.com/2021/04/15/nvidia-rounds-out-ampere-lineup-with-two-new-accelerators/

[76] R. Krashinsky, O. Giroux, S. Jones, N. Stam, and S. Ramaswamy, “Nvidia ampere architecture in-depth,” 5 2020. [Online]. Available: https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/

[77] A. Shilov, “Nvidia’s chinese a800 gpu’s performance revealed,” 5 2023. [Online]. Available: https://www.tomshardware.com/news/nvidia-a800-performance-revealed

[78] P. Alcorn, “Nvidia infuses dgx-1 with volta, eight v100s in a single chassis,” 5 2017. [Online]. Available: https://www.tomshardware.com/news/nvidia-volta-v100-dgx-1-hgx-1,34380.html

[79] I. Cutress, “Nvidia’s dgx-2: Sixteen tesla v100s, 30tb of nvme, only $400k,” 3 2018. [Online]. Available: https://www.anandtech.com/show/12587/nvidias-dgx2-sixteen-v100-gpus-30-tb-of-nvme-only-400k

[80] C. Campa, C. Kawalek, H. Vo, and J. Bessoudo, “Defining ai innovation with nvidia dgx a100,” 5 2020. [Online]. Available: https://devblogs.nvidia.com/defining-ai-innovation-with-dgx-a100/

[81] H. Mujtaba, “Nvidia unveils hopper gh100 powered dgx h100, dgx pod h100, h100 pcie accelerators,” 3 2022. [Online]. Available: https://wccftech.com/nvidia-unveils-hopper-gh100-powered-dgx-h100-dgx-pod-h100-h100-pcie-accelerators/

[82] R. Smith, “Nvidia hopper gpu architecture and h100 accelerator announced: Working smarter and harder,” 3 2022. [Online]. Available: https://www.anandtech.com/show/17327/nvidia-hopper-gpu-architecture-and-h100-accelerator-announced

[83] ——, “Nvidia gives jetson agx xavier a trim, announces nano-sized jetson xavier nx,” 11 2019. [Online]. Available: https://www.anandtech.com/show/15070/nvidia-gives-jetson-xavier-a-trim-announces-nanosized-jetson-xavier-nx

[84] B. Funk, “Nvidia jetson agx orin: The next-gen platform that will power our ai robot overlords unveiled,” 3 2022. [Online]. Available: https://hothardware.com/news/nvidia-jetson-agx-orin

[85] “Jetson agx orin for next-gen robotics,” 2022. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/

[86] D. Franklin, “Nvidia jetson tx2 delivers twice the intelligence to the edge,” 3 2017. [Online]. Available: https://developer.nvidia.com/blog/jetson-tx2-delivers-twice-intelligence-edge/

[87] B. Hill, “Nvidia unveils ampere-infused drive agx for autonomous cars, isaac robotics platform with bmw partnership,” 5 2022. [Online]. Available: https://hothardware.com/news/nvidia-drive-agx-pegasus-orin-ampere-next-gen-autonomous-cars

[88] “Nvidia l40,” 2023. [Online]. Available: https://www.techpowerup.com/gpu-specs/l40.c3959

[89] “Nvidia tesla p100.” [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-p100/

[90] R. Smith, “16gb nvidia tesla v100 gets reprieve; remains in production,” 5 2018. [Online]. Available: https://www.anandtech.com/show/12809/16gb-nvidia-tesla-v100-gets-reprieve-remains-in-production

[91] E. Kilgariff, H. Moreton, N. Stam, and B. Bell, “Nvidia turing architecture in-depth,” 9 2018. [Online]. Available: https://developer.nvidia.com/blog/nvidia-turing-architecture-in-depth/

[92] “Nvidia tesla v100 tensor core gpu,” 2019. [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-v100/

[93] J. McGregor, “Perceive exits stealth with super efficient machine learning chip for smarter devices,” 4 2020. [Online]. Available: https://www.forbes.com/sites/tiriasresearch/2020/04/06/perceive-exits-stealth-with-super-efficient-machine-learning-chip-for-smarter-devices/

[94] “Mn-core,” 2020. [Online]. Available: https://projects.preferred.jp/mn-core/en/

[95] I. Cutress, “Preferred networks: A 500 w custom pcie card using 3000 mm2 silicon,” 12 2019. [Online]. Available: https://www.anandtech.com/show/15177/preferred-networks-a-500-w-custom-pcie-card-using-3000-mm2-silicon

[96] D. Firu, “Quadric edge supercomputer,” 4 2019. [Online]. Available: https://quadric.io/supercomputing.pdf

[97] S. Ward-Foxton, “Qualcomm cloud ai 100 promises impressive performance per watt for near-edge ai,” 9 2020. [Online]. Available: https://www.eetimes.com/qualcomm-cloud-ai-100-promises-impressive-performance-per-watt-for-near-edge-ai/

[98] D. McGrath, “Qualcomm targets ai inferencing in the cloud,” 4 2019. [Online]. Available: https://www.eetimes.com/qualcomm-targets-ai-inferencing-in-the-cloud/#

[99] S. Crowe, “Qualcomm robotics rb5 platform puts 5g, ai in developers hands,” 6 2020. [Online]. Available: https://www.therobotreport.com/qualcomm-robotics-rb5-platform-puts-5g-ai-in-developers-hands/

[100] “Robotics rb6 platform,” 2023. [Online]. Available: https://www.qualcomm.com/products/internet-of-things/industrial/industrial-automation/robotics-rb6-platform

[101] “Rockchip released its first ai processor rk3399pro npu performance up to 2.4tops,” 1 2018. [Online]. Available: https://www.rock-chips.com/a/en/News/Press_Releases/2018/0108/869.html

[102] L. Gwennap, “Machine learning moves to the edge,” pp. 1–7, 4 2020. [Online]. Available: https://www.linleygroup.com/uploads/sima-machine-learning-moves-to-the-edge-wp.pdf

[103] D. McGrath, “Tech heavyweights back ai chip startup,” 10 2018. [Online]. Available: https://www.eetimes.com/tech-heavyweights-back-ai-chip-startup/

[104] R. Merritt, “Startup rolls ai chips for audio,” 2 2018. [Online]. Available: https://www.eetimes.com/startup-rolls-ai-chips-for-audio/

[105] A. Shilov, “Tachyum teases 128-core cpu: 5.7 ghz, 950w, 16 ddr5 channels,” 6 2022. [Online]. Available: https://www.tomshardware.com/news/tachyum-teases-128-core-cpu-57-ghz-950w-16-ddr5-channels

[106] L. Gwennap, “Tenstorrent scales ai performance: Architecture leads in data-center power efficiency,” pp. 1–4, 4 2020. [Online]. Available: https://www.tenstorrent.com/wp-content/uploads/2020/04/Tenstorrent-Scales-AI-Performance.pdf

[107] E. Talpes, D. D. Sarma, G. Venkataramanan, P. Bannon, B. McGee, B. Floering, A. Jalote, C. Hsiong, S. Arora, A. Gorti, and G. S. Sachdev, “Compute solution for tesla’s full self-driving computer,” *IEEE Micro*, vol. 40, pp. 25–35, 3 2020. [Online]. Available: https://doi.org/10.1109/MM.2020.2975764

[108] “Fsd chip - tesla,” 2020. [Online]. Available: https://en.wikichip.org/wiki/tesla_(car_company)/fsd_chip

[109] S. Ward-Foxton, “Tis first automotive soc with an ai accelerator launches,” 2 2021. [Online]. Available: https://www.eetimes.com/tis-first-automotive-soc-with-an-ai-accelerator-launches/

[110] “Tda4vm jacinto processors for adas and autonomous vehicles,” 3 2021. [Online]. Available: https://www.ti.com/lit/gpn/tda4vm

[111] M. Demler, “Ti jacinto accelerates level 3 adas,” 3 2020. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=6130&year=2020&tag=3

[112] R. Merritt, “Samsung, toshiba detail ai chips,” 2 2019. [Online]. Available: https://www.eetimes.com/samsung-toshiba-detail-ai-chips/

[113] L. Gwennap, “Untether delivers at-memory ai,” 11 2020. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=6230