Pathways for Electron Device Research in the AI Era

Jens Trommer[1]

NaMLab gGmbH, Nöthnitzer Str. 64a, 01187 Dresden, Germany

Blue Sky Abstract

With the recent emerge of AI applications chip designers face new challenges to combat the immense power needs arising from the new technology. To keep up with the rapid developments on the system and software level, device research should focus on two parallel pathways: first, emerging devices have to be developed together with their respective CMOS baseline technology, either by monolithic co-integration or by hetero-integration. Second, more attention has to be brought to the operation of the individual device under the constraints of the circuit or system around it. As a result, not always the most scaled, most performant individual device solution will make a market impact, but the solution that functionally-enhances the AI accelerator system most seamlessly.

The big picture

- Emerging devices enable new styles of analog computing to bring down the power needs of AI hardware.
- there are two overarching pathways that researches have to pursue towards this goal: A co-integration with classical CMOS elements, either monolithically or via hetero-integration and device-circuit-co-design, including variability and reliability concerns.
- Challenges and opportunities along both pathways are exemplary illustrated on ferroelectric FET, Ferroelectric Tunnel Junctions, and Reconfigurable FETs. The same paradigms generally apply for any other emerging non-CMOS concept.
- Classical Figures-of-Merits have to be rethought to estimate the gain under those boundary conditions.

The Power of Artificial Intelligence

Today's public perception of technological evolution is dominated by progress originating from the computer science domain: social media algorithms, self-driving autonomous transportation, or even futuristic visions like the meta-verse. More recently, the release of the *GTP4* engine has brought artificial intelligence (AI) to the spotlight of the public[1]. Within just two years, we have already seen AI innovations taking over. The application space of AI goes way beyond chatbots: it includes machine vision and speech recognition, path and timetable scheduling in aviation, cryptography, and graph analysis, as well as robotics, healthcare, and automotive[2]. In addition, we have seen Moore's law, the classical rules of scaling down transistor node sizes with each technological generation, decline. So, in some sense, one can ask: Do we need to develop new electron devices at all, or is it all software and circuit design going forward?

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

At a glance, the question seems to be answered when looking at top-notch AI accelerator hardware like the newly spoiled Blackwell series of Nvidia. However, the rapid technological revolution initiated by AI comes with a massive downside: the underlying architectures used for executing the deep learning models powering AI cannot keep up with the need for a massive amount of parallel data processing and storage capability. In fact, the energy required by processing AI with classical CPU/GPUs would outclass world energy production in the next decade, if the current trajectories are followed[3]. On the one hand, this is governed by the rapid growth of the AI chip market, which is predicted to almost double its share of the overall semiconductor business at least every five years from now. On the other hand, it is based on fundamental constraints originating from the Shannon limit and thus cannot be solved with digital architectures based on CMOS. As opposed to this, direct analog computation has tremendous potential to reduce computation power and ensure sustainability (Fig. 1). As a simple yet relevant example, vector-matrix-multiplications can be directly executed analogically by exploiting Ohm's and Kirchhoff's laws inside a memory array[4], thus improving the most energy hungry function for realizing attention mechanism in AI hardware. To pursue this digital-to-analog transition the utilization of emerging device concepts for Logic-In-Memory (LiM) or post-Shannon analog accelerators is imperative. At the same time, it opens up two challenging pathways for the research on emerging electron devices: the new device concepts have to be integrated into existing CMOS hardware, but also be tuned to operate reliable inside their respective circuit environment.

A Matter of Integration

Generally, the three competing approaches illustrated in Fig. 2 are of importance with respect to fabrication of AI accelerators with emerging devices. In the first one, the emerging element is directly co-integrated into a CMOS chip's front-end-of-line (FEOL). However, this monolithic system-on-chip (SoC) approach will create several constraints for the emerging device itself. The baseline technology's design rules and process parameters must be respected, as the co-integration process must not affect the baseline technology. Naturally, these constraints will lead to individual devices with less optimal performance values than an independently optimized device in a lab-scale environment. Two emerging device concepts that have already demonstrated monolithic SoCs capabilities are ferroelectric field effect transistors (FeFETs) and reconfigurable field effect transistors (RFETs), as FeFETs enable a non-volatile multilevel storage option[5], while RFETs provide adaptability by providing different transport modes on request[6]. Most notably, both concepts can be used as non-linearity to facilitate analog data processing[5,7]. The seamless FEOL co-integration achieved with those devices reduced the need for interfacing circuitry or data routing between memory and processing units, the main driver for power consumption in AI applications. The main drawbacks of the approach are the high complexity of the process and, thus, the considerable time to market and cost expected to yield the first product. Thus, the second hetero-integration approach relaxes those constraints by moving the emerging device away from the FEOL and into the back-end-of-line (BEOL). This way, an individual set of design rules can be deduced, easing a co-integration with the CMOS baseline. Utilizing the third dimension will increase the density of elements per area. In turn, the emerging device will have to obey the temperature budget for BEOL processing, typically a fabrication below 400°C. On the transistor side, reconfigurable Thin-Film-Transistors (TFT) are inherently suitable for such an integration, as they do

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

not require perfect crystalline channel material to be operational[8]. On the memory side, resistive switching devices, but also ferroelectric tunnel junctions (FTJs) have been proven to be monolithically integrable in the BEOL[4]. Lastly, instead of dealing with cross-interaction between different technologies on a single chip, a system can also be constructed from modular chiplets of dedicated accelerators, each optimized for a specific task exploiting a different set of technological solutions[9]. These chiplets are combined on a shared host substrate, featuring an electrical interface for routing signals through an interposer or a through-silicon-via (TSV), as already pursued for the current high-performance digital products. While this pathway yields the most freedom for designing the individual electron device technologies, it puts additional constraints on the chip packaging and interfacing circuitry, having to match different voltage levels and performances. More importantly, the need to optimize multiple SoCs as well as their assembly on the die-level is a considerable cost factor, currently limiting the applicability of chiplet-based hetero-integration to high-performance architectures. With respect to emerging devices, chiplets are mainly interesting for modern LiM accelerators, for example based on complementary FeFET-based analog content-addressable memories[10] or high-performance devices utilizing new transport regimes, such as dirac-source transistors made from two-dimensional materials[11].

Circuit Level Operation Matters

Besides the integration challenge, it is also vital not to decouple the emerging device development from the circuit or system it operates in. While higher on-currents are almost always more desirable for classical digital design, they might pose a power risk in high-parallel computing. If currents are summed over many devices, the maximal currents are somewhat limited by the current load capacity of the BEOL wiring; thus, a slower device with better leakage properties might be favoured. Factors like programming linearity, read power efficiencies, and failure resistance gain increasing importance. Again, not the most scaled device will always show the most desired property. A highly-scaled device may show higher variability and, thus, worse reliability features than a more relaxed one, putting constraints on the evaluation circuit[12]. In addition, some properties might change completely when altering the device size. For example, both FeFETs and FTJs have been shown to transition from a gradual switching behavior to a rapid statistical switching behavior when scaled to their smallest physical possible size[5]. Device-circuit interactions become even more complex once reliability constraints are considered. On the one hand, it is important to note that not all devices in an electrical system will see the same amount of stress over the product's lifetime. Thus, the circuit has to be designed in a way that it is fully operational with a combination of a fresh device at time-zero and devices that have already been stressed for nearly 10 years. The different degradation effects to consider here are endless: bias-temperature instability, hot-carrier injection, time-dependent-dielectric-breakdown, endurance cycling, retention, imprint, thermal stress, and more. While these effects can be investigated independently to gain a deeper physical understanding, they often relate or compete with each other. For example, a trade-off between switching speed and retention has been seen based on the pulse train programming in both FeFETs and FTJs[13]. For the other example case of RFETs with programmable p- and n-type operation, they have to withstand positive and negative voltage stress in a randomly alternating fashion, leading to a new stress pattern not observed in classical CMOS design[14]. To that, a comprehensive reliability assessment should be carried out for all

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

emerging technologies, identifying all worst-case conditions of the new technology followed by a circuit level stress test inducing a smart combination of those for product lifetime test[15]. Afterward, those constraints can be fed into modern design-technology-co-optimization (DTCO) tools to optimize cell layouts and physical design[12].

Conclusions

To summarize, with the transition from digital to analog computation, careful ***device-circuit-co-design*** will play a significant role in ensuring the usability of emerging elements, providing features such as non-linearities or multilevel storage for analog computation. At the same time, the influence of imperfections will play an increased importance, making the product-level assessment of variability and reliability inevitable to ensure a stable operation over the whole lifetime. Lastly, as CMOS will not go anywhere anytime soon, new technologies should always be regarded with respect to their **CMOS-*co-integrability***. Both paradigms put external constraints on the device design and individual device performance metrics. While these constraints are valid for all electron devices, FeFETs, RFETs and FTJs have been used as three specific examples to illustrate future pathways in providing added value to an SoC by multilevel storage and non-linearity, respectively. Finally, it is important to note that not always the most-scaled high-performance switch will also lead to the best system-level performance. As a result, the typical ***figures-of-merit*** (FoM) used to benchmark electron devices will have to be re-evaluated for any given scenario.

Acknowledgement.

This work was partially funded by the European Union under grant agreement no. 101135316 and no. 101016776. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the EU or the EC. Neither the EU nor the granting authority can be held responsible for them.

Biography.

Jens Trommer received the Dipl.-Ing. Degree in electronic and sensor materials from TU Bergakademie Freiberg, Germany in 2011 and the Dr.-Ing. Degree in electrical engineering from TU Dresden, Germany in 2017. Currently, he is holding the position of a Senior Scientist at NaMLab gGmbH, Dresden, Germany, leading the "Emerging Devices" research group. His research interest comprises reconfigurable and ferroelectric devices and their transfer into industrial circuit applications. He is author or co-author of more than 90 peer-reviewed publications and 1 patent. In 2023 he has been elected to the rank Senior Member at the Institute of Electrical and Electronics Engineers.

References

1. Sanderson, K. (2023). GPT-4 is here: what scientists think. Nature *615*, 773–773.

2. Christensen, D.V., Dittmann, R., Linares-Barranco, B., Sebastian, A., Gallo, M.L., Redaelli, A., Slesazeck, S., Mikolajick, T., Spiga, S., Menzel, S., et al. (2022). 2022 roadmap on neuromorphic

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

computing and engineering. Neuromorphic Comput. Eng. *2*, 022501. https://doi.org/10.1088/2634-4386/ac4a83.

3. Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. Nature *561*, 163–166. https://doi.org/10.1038/d41586-018-06610-y.

4. Amirsoleimani, A., Alibart, F., Yon, V., Xu, J., Pazhouhandeh, M.R., Ecoffey, S., Beilliard, Y., Genov, R., and Drouin, D. (2020). In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives. Adv. Intell. Syst. *2*, 2000115. https://doi.org/10.1002/aisy.202000115.

5. Mulaosmanovic, H., Breyer, E.T., Dünkel, S., Beyer, S., Mikolajick, T., and Slesazeck, S. (2021). Ferroelectric field-effect transistors based on HfO2: a review. Nanotechnology *32*, 502002. https://doi.org/10.1088/1361-6528/ac189f.

6. Mikolajick, T., Galderisi, G., Rai, S., Simon, M., Böckle, R., Sistani, M., Cakirlar, C., Bhattacharjee, N., Mauersberger, T., Heinzig, A., et al. (2022). Reconfigurable field effect transistors: A technology enablers perspective. Solid-State Electron. *194*, 108381. https://doi.org/10.1016/j.sse.2022.108381.

7. Simon, M., Mulaosmanovic, H., Sessi, V., Drescher, M., Bhattacharjee, N., Slesazeck, S., Wiatr, M., Mikolajick, T., and Trommer, J. (2022). Three-to-one analog signal modulation with a single back-bias-controlled reconfigurable transistor. Nat. Commun. *13*, 7042. https://doi.org/10.1038/s41467-022-34533-w.

8. Park, J.-M., Bae, J.-H., Eum, J.-H., Jin, S.H., Park, B.-G., and Lee, J.-H. (2017). High-Density Reconfigurable Devices With Programmable Bottom-Gate Array. IEEE Electron Device Lett. *38*, 564–567. https://doi.org/10.1109/LED.2017.2679343.

9. Lau, J.H. (2023). Recent Advances and Trends in Chiplet Design and Heterogeneous Integration Packaging. J. Electron. Packag. *146*. https://doi.org/10.1115/1.4062529.

10. Liu, X., Katti, K., He, Y., Jacob, P., Richter, C., Schroeder, U., Kurinec, S., Chaudhari, P., and Jariwala, D. (2024). Analog content-addressable memory from complementary FeFETs. Device *2*, 100218. https://doi.org/10.1016/j.device.2023.100218.

11. Qin, L., Tian, H., Li, C., Xie, Z., Wei, Y., Li, Y., He, J., Yue, Y., and Ren, T.-L. (2024). Steep Slope Field Effect Transistors Based on 2D Materials. Adv. Electron. Mater. *10*, 2300625. https://doi.org/10.1002/aelm.202300625.

12. Karner, M., Rzepa, G., Schleich, C., Schanovsky, F., Kernstock, C., Karner, H.-W., Baumgartner, O., and Stanojevic, Z. (2024). An Efficient and Accurate DTCO Simulation Framework for Reliability and Variability-Aware Explorations of FinFETs, Nanosheets, and Beyond. In 2024 8th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), pp. 1–3. https://doi.org/10.1109/EDTM58488.2024.10511441.

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

13. Lancaster, S., Lomenzo, P.D., Engl, M., Xu, B., Mikolajick, T., Schroeder, U., and Slesazeck, S. (2022). Investigating charge trapping in ferroelectric thin films through transient measurements. Front. Nanotechnol. *4*. https://doi.org/10.3389/fnano.2022.939822.

14. Galderisi, G., Mikolajick, T., and Trommer, J. (2024). Reliability of Reconfigurable Field Effect Transistors: Early Analysis of Bias Temperature Instability. In 2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), pp. 1–7. https://doi.org/10.1109/IPFA61654.2024.10691045.

15. Ettisserry, D., Visconti, A., Bonanomi, M., Pazzocco, R., Locatelli, A., Sebastiani, A., Chavan, A., Hollander, M., Servalli, G., Calderoni, A., et al. (2024). Comprehensive Reliability Assessment of 32Gb (Hf,Zr)O2-Based Ferroelectric NVDRAM. In 2024 IEEE International Reliability Physics Symposium (IRPS), pp. 1–8. https://doi.org/10.1109/IRPS48228.2024.10529336.

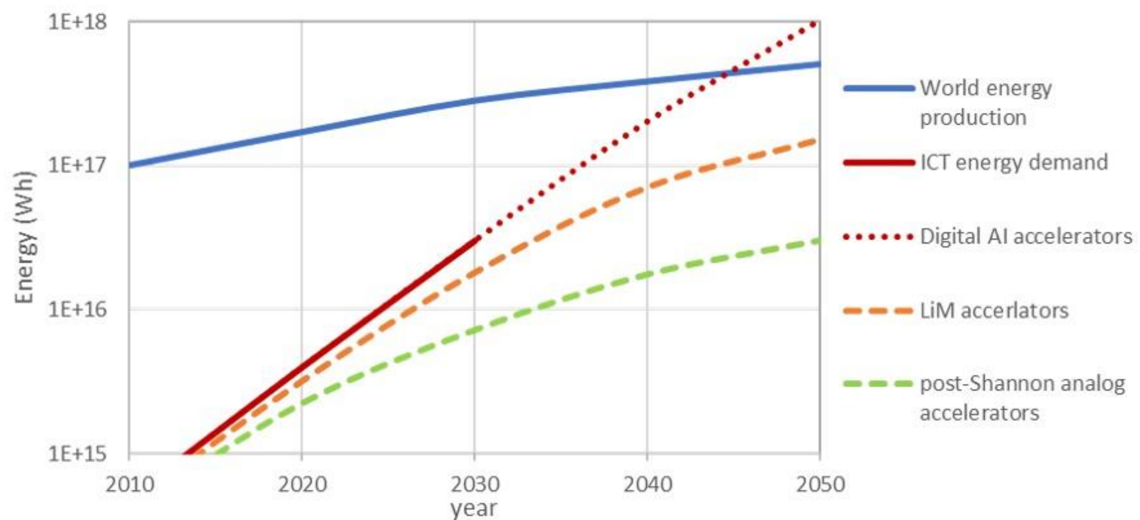

Figure 1: Estimating how the future worldwide total information and communication technology (ICT) demand will outpace the world's electrical energy production, if the present trajectory is followed [3]. In conclusion, a paradigm shift away from digital accelerators towards logic-in-memory (LiM) or analog accelerators must be drawn to continue the development of AI-aided systems successfully.

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656

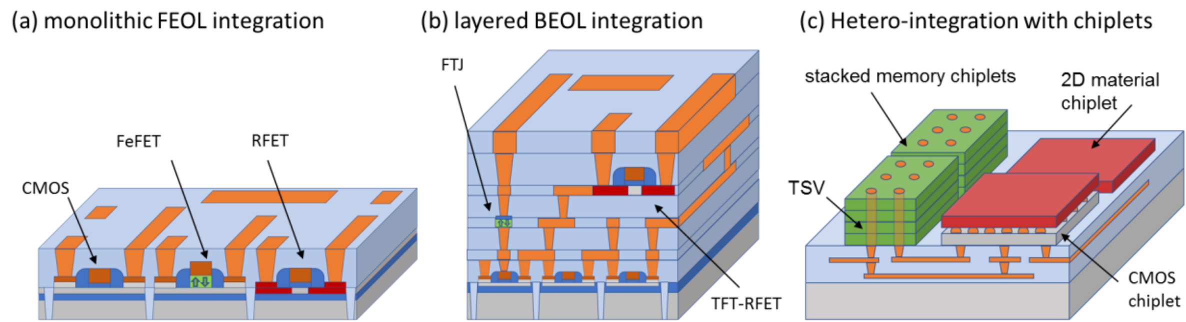

Figure 2: Pathways for co-integration of emerging device concepts with CMOS technology. (a) monolithic integration directly into the front-end-of-line (FEOL) of the CMOS process, (b) layered co-integration approach expanding to the back-end-of-line (BEOL), (c) hetero-integration using chiplets connected via interposer and through-silicon vias (TSV) or by direct chip-on-chip bonding. Examples are illustrated for ferroelectric field effect transistors (FeFETs), ferroelectric funnel junctions (FTJs), and reconfigurable field effect transistors (RFETs).

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

Cite as: Trommer, Jens Device, Volume 3, Issue 1, 100656

https://doi.org/10.1016/j.device.2024.100656