New Hardware

Recommendations for New Hardware

Specialized hardware (HW)

The development of efficient hardware is essential for running services, orchestrators and SAMs efficiently at the edge and within federated networks. Europe must address memory costs (for AI), energy consumption, and ecological impact, potentially leveraging non-volatile memory for direct edge execution. Additionally, the next generation of SAMs should incorporate learning through experiences or allow to the efficient execution of digital twins to maintain Europe's competitive edge in AI (embedded AI). In the field of AI accelerators, the focus should be on inference (becoming more and more important with the approach pioneered by OpenAI o1 and o3) or on fine tuning. Reducing the transfer of data is key to reach lower levels of power consumption. This can be achieved with near- or in-memory computing (NMC or IMC), direct execution from the storage of parameters (hence eliminating the need for RAM), etc…

Beyond purely digital hardware (HW)

Investigation of new **accelerators using non digital technologies**, going from exact computations (digital computation) to more approximate computing (neural networks are universal approximators, quantum computing results are stochastics, optimization techniques using Bayesian, Ising approaches can solve complex problems) should be also investigated in the context of providing more efficient services to the next computing paradigm (NCP) ecosystem.

YOU HAVE TO SEE THIS...
MY NEW COMPUTER CONSISTS ONLY OF BIO-BASED MATERIALS!
YES, THE DOG LIKES IT TOO..
HiPEAC
ARNULF.BE

Introduction

The changes in the hardware arena since the publications of the HiPEAC Vision of 2023 and 2024 may be minor, but there have certainly been developments. Artificial intelligence (AI) accelerator hardware is dominating the profits of the top hardware manufacturers. Tensor processing units (TPUs) are increasingly augmented by general-purpose graphics processing units (GPGPUs), but manufacturers of both processor architectures are based outside Europe.

The influence of AI is being felt in all aspects of the economy, including the technology sector, where it supports design and implementation of both hardware and software. AI is also spreading towards the edge of the continuum, making smarter applications possible in the home, as well as smart equipment in the field. We expect to see this growth towards the edge increasing and finding its way into as yet underexplored applications in the coming years. This creates challenges and opportunities for European players [SemiWiki-NPU-2024].

Training AI is an important part of its application; inference is another important one. A market is opening for efficient inference engines, tailored to specific needs of the place in the computing continuum where it takes place. AI applications for specific domains do not require a general-purpose AI application, but one tuned to the needs of the domain with tuned hardware. Of course, this will increase the diversity of devices for AI, but AI itself can be used to design these application specific AI devices.

All seems quiet on the quantum computer front. But even though investments in quantum computer start-ups levelled out in 2024 [EETimes-March] the field is still progressing. An often-made remark is that the race for quantum computers is a marathon, not a sprint; it will take endurance and time to achieve quantum computers that can demonstrate their superiority over classical computers in specific practical compute challenges.

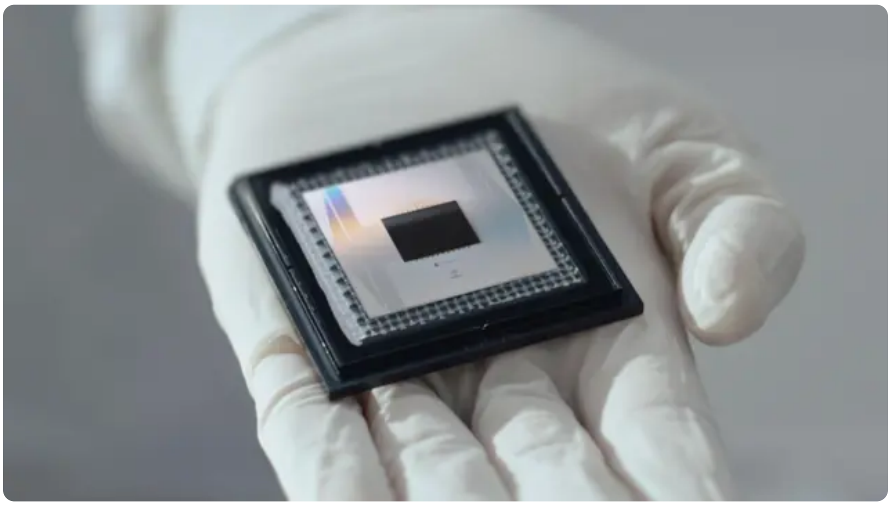

Figure 1: Google's 105 qubit Willow Chip which achieved exponential error reduction, produced in 2024. [BBC-Willow]

Towards the end of 2024, Google announced its Willow chip [BBC-Willow][Google-Willow], a chip which stands out not so much for its number of qubits (105) as for demonstrating that exponential error reduction is possible with a linearly increasing number of logical qubits to form a logical qubit. See Figure 1.

During the last few years, simulating quantum systems, e.g. for drug discovery, was touted as one of the most important practical applications of quantum computing for the near future. However, AI-based applications are now challenging that role as they are already showing great acceleration of that task on classical computers (see [Labiotech] and [Acellera]). However, this does not mean that quantum computing is becoming obsolete before it reaches practical maturity, because drug discovery is just one application. It should not lead to quantum computing disillusion, as we are still in the process of finding out for which problems quantum computing is a winner over classical computing (cracking encryption keys being one).

Europe is strong in quantum compute expertise, but when the technology moves from the lab to the market Europe must be able to commercialize this expertise. Geopolitical changes over the last decade also make it necessary for Europe to strive for technical sovereignty now more than ten years ago, and quantum computing is definitely one of the topics relevant for sovereignty. Europe should intensify its intra-European cooperation, guarding against techno-nationalism. Setting up European technology centres that stimulate European independence from other global factions might function as a glue between countries.

Although outside HiPEAC's field, it should be noted that Europe's presence in semiconductor manufacturing equipment is quite strong. ASML stands out, but other critical equipment, such as that used for deposition and measurement, or metrology, also comes predominantly from European vendors. In addition, the percentage of capital investment for manufacturing equipment by chip manufacturers has risen in recent years from around 10% to approximately 20% [EETimes-November].

Efficient hardware: Continue the quest for lower power and improved performance

The growth of AI towards the edge of the continuum creates opportunities for Europe to become a player in AI. AI requires a lot of processing, and that requires energy. As energy is scarcer at the edge, low-power architectures with just the right amount of processing power are required. This is an emerging development, and Europe has an opportunity to build a strong presence in this technology. It can do that also by initiating standards initiatives, leading the way to fast integration and introduction of edge AI technology. Europe has a strong position in embedded chips, which it should strive to retain. Designers of embedded systems are used to dealing with constraints, a skillset that European companies can leverage in designing edge AI processing components.

Strengthen European sovereignty in hardware technology and manufacturing

With the European Chips Joint Undertaking as a major and very important step to increase independence, Europe must also ensure that it covers all the important steps in the manufacturing chain, not only chip manufacturing. Europe's presence in the supply of manufacturing equipment is already strong. But it still depends on non-European suppliers for critical raw material for chip production. It is very probably not possible to fully eliminate this dependency, e.g. because of the geological location of such supplies, but Europe should strive to minimise such dependencies.

Another aspect is sovereignty in key technologies. Europe's presence in e.g. photonics production is not very high, even though the level of research in Europe of this technology is high. Photonics is key for data communication. If Europe misses out on such key technologies, its sovereignty in ICT and its applications such as AI is threatened. (The Chips Joint Undertaking has a paragraph on strengthening integrated photonics production.)

New accelerators: Prepare for the integration and hetero integration of new hardware technology

Hetero integration refers to the integration of different types of materials, devices, or technologies into a single system or chip. By leveraging the strengths of diverse materials and technologies, hetero integration is a key technology paving the way for more advanced, efficient, and versatile computing systems that would be critical in the context of the NCP.

Although practical applications of quantum hardware seem to be ten years away, this is a field that Europe should not leave to others. The technology has the potential to become key in society and in the economy. Private investments in this technology are levelling out, possibly because the technology is growing out of the startup environment. It requires scaling up to an industrial application level as the next step, which is beyond the capabilities of private investors and requires existing (European) companies to step in. In this respect it has also been noted that more (European) company research should be directed to this technology to prepare for commercial application `[EETimes-March]`.

Quantum technology expertise is spread over Europe, a situation which is currently supported by the way projects are financed. It should be investigated whether a stronger concentration of efforts in the development of this technology would be beneficial for its progress. Strategic alliances with European non-EU-based research groups and companies should be investigated.

Encourage modularity and standardization at the hardware level

For better design and optimization of hardware/software, support of complex architectures including modular AI to make scalability relatively easy is key.

Modular design of digital systems should be encouraged. With modularity comes the need for standardization, not just at the hardware level (connectors, metrics, etc.), but above all at the level of software stacks and their interactions with the various hardware levels. This need, traditionally expressed in the rather late stages of digital systems integration, must be considered right from the design stage of the hardware and software components that make them up. This is the case, for example, with the hetero integration of technologies mentioned in the previous paragraph. Not only must the various chips be designed to interconnect physically with each other, but also to operate logically as a functional whole.

The integration of AI at all levels of the NCP concept also illustrates this need for modularity and standardization. Ensuring that an AI model integrated into the system, with its requirements on data and hardware resources, can perform the expected function depends on its ability to seamlessly interface with it.

Support a European ecosystem that encourages a strong link between information sciences and basic research on emerging hardware, including quantum computing

European universities, research centres and companies must be at the forefront of basic research in information sciences: information encoding, programming paradigms and computing complexity to cite a few. It is important that research on new hardware for computing devices – such as spintronics and photonics – are linked to progress in information science in order for cross-fertilization between those disciplines to occur. For example, a computing concept like the “Ising spins machine” might appear as a good idea from a hardware elaboration point of view but might require complete new knowledge from an information or programming perspective. Research into new hardware devices for computing must therefore be carried out in close collaboration with advances in information science and algorithms.

Multichip

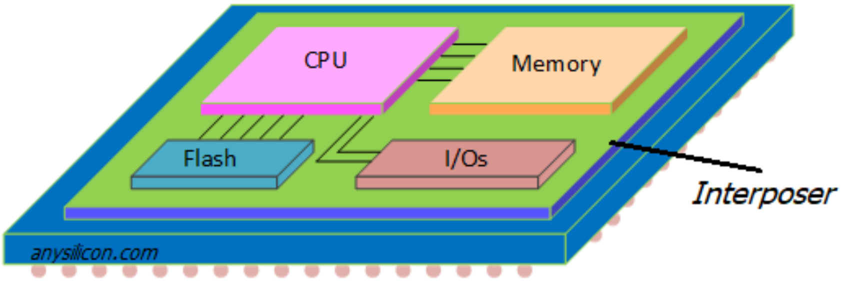

Figure 2: Schematic diagram of 3D-interposer technology (https://anysilicon.com/wp-content/uploads/2019/08/Interposer.png)

As the complexity of digital systems keeps growing, it becomes harder and harder to integrate the whole system on one chip. Moreover, some accelerators may be fabricated using fabrication technologies that are incompatible with standard digital ones.

Integrating multichip systems in one package is increasingly done using 2.5D and 3D interposers. Europe has a strong research capability in this technology. CEA-List in France is researching active interposers, meaning that the interposers themselves perform part of the active functions of the integrated system, e.g. through a network-on-chip (or better: network-on-interposer), or implementing analogue functions.

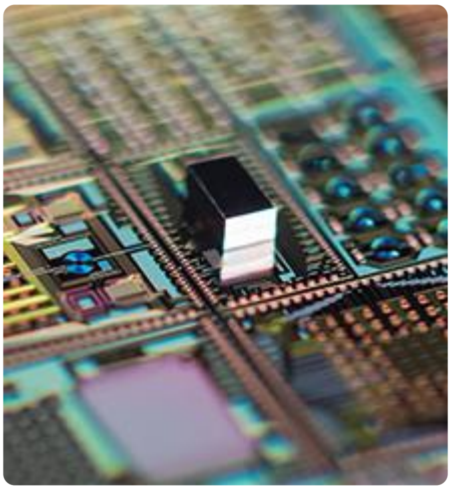

Figure 3: 3D active interposer technology from CEA [CEA-OpticalInterposers]

The Fraunhofer Gesellschaft has founded a Chiplet Center of Excellence in Dresden in 2019, researching this technology with an eye on applications in the automotive industry [Fraunhofer-IZM]. Imec in Leuven launched (in Ann Arbor in the USA) its automotive chiplet programme in October 2024. Imec is also active in standardizing chip interposer technology, paving the way to wider use of this technology. European companies active in chiplet technology include Bosch, and Quintarius [Quintarius], founded by a number of, mostly European companies. Non-European companies such as Singapore-based Silicon Box have plans to invest in chip factories in Europe [SiliconBox].

These examples show Europe's strong presence in this technology, which it should strive to keep.

Conclusion

The future of digital hardware technologies goes beyond the development of new physical components, new chip technologies or new materials. In a context of increasingly complex systems, with the arrival of AI at all levels and the need for sustainable/eco-responsible digital technology, it is a finer coupling between hardware and software technologies that is the key to new digital technologies. In this sense, the case of NCP is eloquent and illustrates our vision of future hardware technologies.

References

Acellera: https://www.acellera.com/

BBC-Willow: https://www.bbc.com/news/articles/c791ng0zvl3o

CEA-OpticalInterposers: https://www.leti-cea.com/cea-tech/leti/english/Pages/Industrial-Innovation/Demos/3D-Integration-HPC-AI.aspx

EETimes-March: https://www.eetimes.eu/ee-times-europe-magazine-march-2024/

EETimes-November: https://www.eetimes.eu/ee-times-europe-magazine-november-2024/

EETimes-September: https://www.eetimes.eu/ee-times-europe-magazine-september-2024/

Fraunhofer-IZM: https://blog.izm.fraunhofer.de/the-chiplet-center-of-excellence/

Google-Willow: https://blog.google/technology/research/google-willow-quantum-chip/

IMEC: https://www.imec.be/nl/press/internationale-auto-industrie-klopt-aan-bij-imec-voor-nieuw-type-microchips

Labiotech: https://www.labiotech.eu/best-biotech/ai-drug-discovery-europe/

Quintarius: https://en.wikipedia.org/wiki/Quintauris

SemiWiki-NPU-2024: https://semiwiki.com/artificial-intelligence/349906-get-ready-for-a-shakeout-in-edge-npus/

SiliconBox: https://www.reuters.com/technology/silicon-box-picks-piedmont-region-its-italian-34-bln-chip-plant-2024-06-28