FVLLMONTI

Call: **H2020-FETPROACT-2020-01**

Grant Agreement no. **101016776**

Deliverable D6.7 – Technology impact and exploitation innovation – Year 1

Start date of the project: 1st January 2021
Duration: 50 months
Project Coordinator: Cristell MANEUX - University of Bordeaux
Contact: Cristell MANEUX - cristell.maneux@ims-bordeaux.fr

DOCUMENT CLASSIFICATION

Title	Technology impact and exploitation innovation - Y1
Deliverable	D6.7
Estimated Delivery	28/02/2022 (M12+2)
Date of Delivery Foreseen	28/02/2022 (M12+2)
Actual Date of Delivery	28/02/2022 (M12+2)
Authors	Jens Trommer – P6 – NLB Guilhem Larrieu – P2 – CNRS-LAAS Ian O'Connor – P3 – ECL-INL Jens Trommer – P6 – NLB Giovanni Ansaloni – P4 – EPFL Chhandak Mukherjee – P1 – UBx
Approver	Cristell Maneux – P1 – UBx
Work package	WP6
Dissemination	PU
Version	V1.0
Doc ID Code	D6.7_FVLLMONTI_P6-NLB-220228
Keywords	Technology impact, exploitation, assessment

DOCUMENT HISTORY

VERSION	PUBLICATION DATE	CHANGE
1.0	28.02.2022	Year 1 Version

DOCUMENT ABSTRACT

This document describes the initial technology impact and exploitation action assessment associated with FVLLMONTI. The aims and visions of the project are set into the bigger overall socio-economic context. It is described why the disruptive N2C2 concept based on emerging nanowire technologies can change the neuromorphic circuit market. Current market size and segmentation, as well as competitor technologies, are referenced.
FVLLMONTI is a multi-partner project spanning from emerging vertical nanowire technologies via design-enablement, design-technology-co-optimization (DTCO), and circuit design up to hardware-software-co-optimization. Key innovations across all layers of technology development are identified and assessed regarding their future technological impact. The major risks for the potential application of those innovations are analyzed. Last but not least, potentially exploitable project assets are identified, and plans and actions towards their exploitation are given. The document will be updated every year, covering the new findings within the project and the changes in the competitive landscape and market segmentation.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101016776.

TABLE OF CONTENT

LIST OF FIGURES AND TABLES

LIST OF ACRONYMS / GLOSSARY

ANN: Artificial Neuronal Network
AP: Ambipolar
CNN: Convolutional Neuronal Network
CAGR: Compound Annual Growth Rate
CMC: Compact Model Consortium
CMOS: Complementary Metal Oxide FET Technology
DRAM: Dynamic Random Access Memory
DTCO: Design-Technology-Co-Optimization
EDA: Electronic Design Automation
EDP: Energy-Delay-Product
Fe: Ferroelectric
FET: Field Effect Transistor
FVLLMONTI: Acronym for this project
GPU: Graphic Processing Unit
IOT: Internet of Things
JL: Junctionless
NN: Neural Network
ML: Machine Learning
NLP: Natural Language Processing
N2C2: Neuronal Network Compute Cube
PC: Polarity-Controllable
PDP: Power-Delay-Product
P&R: Place and Route
RRAM: Resistive Random Access Memory
SNN: Spiking Neuronal Networks
SRAM: Static Random Access Memory
SME: Small or Medium Enterprise
TSMC: Taiwan Semiconductor Manufacturing Company

1. EXECUTIVE SUMMARY

Neuromorphic computing combines specific hardware and software that implements a behavior mimicking the biological function of a Neural Network. This enables a highly parallel computation with reduced power consumption compared to classic circuit architecture. Neuromorphic circuits are predicted to yield high benefits in many application areas, such as consumer electronics, industrial electronics, automotive, financial services, cyber security, and other more fringe markets like medical, aerospace, and defense. Thus, high market growth has been forecast for the period of the project and beyond. While the main market share right now is governed by the application of image recognition, natural language processing is expected to overtake this within the coming years. With the increased complexity of the underlying algorithms, increased performance of the executing hardware is targeted. Simultaneously, low-power and low-volume hardware are needed to enable mobile applications, which do not rely on cloud storage access on a server.

In FVLLMONTI, a candidate for such a hardware accelerator is being developed based on vertical nanowire technologies. By stacking multiple vertical nanowires on top of each other, a higher density of operations/area can be achieved. Hardware demonstrators for junctionless, polarity controllable, and ferroelectric nanowire devices will be fabricated. Either all of those devices or a subset will be combined to yield different technology platforms for circuit design. Logic blocks are developed in a design-technology-co-optimization (DTCO) fashion, including parasitic modeling of unique test structures. 3D place & route algorithms will be demonstrated to pave the way for EDA for 3D stacked technologies. Based on these 3D circuit cells, neuronal network compute cube (N2C2) circuit blocks will be designed. The concept of the N2C2 block as a highly flexible 3D building block to enable dense hardware accelerators that are precisely tuned to the needs of a given machine learning application is highly disruptive. In FVLLMONTI, we target two specific and timely application scenarios on these accelerators: automated speech recognition and machine translation.

The current early development status in each of these areas is briefly summarized. The potential impact of the key innovations is discussed, and potential risks for future exploitation are assessed. Also, specific plans and actions regarding the exploitation strategy for the individual assets are given.

The developed devices, models, characterization techniques, software tools, circuits, and algorithms all show potential to be exploited beyond the project's runtime. However, the full potential will be unlocked by combining all developments towards the FVLLMONTI vision.

2. ANALYSIS OF THE OVERALL SOCIO-ECONOMIC CONTEXT

I. BACKGROUND

Neuromorphic computing combines specific brain-inspired integrated hard- and software that implements a behavior mimicking the biological function of a Neural Network (NN). There are several types of these artificial neuronal networks (ANN); the most important to name are convolutional neuronal networks (CNN) and spiking neuronal networks (SNN). Neural Networks aim to reach the massively parallel processing ability of the human brain along with low power consumption. Key features envisioned are stochastic operations, pattern recognition, fault tolerance, faster computation, and scalability. Data storage (memory) and the processing units are integrated seamlessly (non-volatile logic or logic-in-memory), different from classic computing architectures suffering from the infamous von-Neumann bottleneck.

Neuromorphic chips can be designed in digital, analog, or mixed-signal fashion. Artificial neurons from analog designs are typically used to resemble the characteristic spiking behavior of a biological neuron. Artificial synapses are used to generate patterns and learning algorithms. In theory, analog architectures need fewer transistors to emulate a specific function and thus consume less energy than digital neuromorphic chips. However, the analog architecture leads to higher noise, lowering the precision. Digital designs, on the other hand, are more precise compared to analog chips. Their digital structure eases on-chip programming and data evaluation. This flexibility allows users to implement various algorithms with low-energy consumption accurately.

Digital neuromorphic hardware can be implemented on a large variety of chips as long as they provide embedded memory functionality. Non-volatile memories are envisioned to provide a high gain in power efficiency compared to volatile memories such as SRAM or DRAM. Ongoing developments cover both classical memories, such as Charge-Trapping-based devices, and emerging memories, such as resistive random-access memory (R-RAM) or ferroelectric FETs (Fe-FET).

The main competitor technology for neuromorphic hardware is graphic processing units (GPU).

Neuromorphic computing relies on the skillful combination of software, e.g., a machine learning (ML) algorithm and the underlying hardware accelerator. Given the advanced state of current CMOS manufacturing sites, this poses a high hurdle for small and medium enterprises (SME) or start-ups without a network to compete in the market.

II. USER NEEDS, MARKET TRENDS, AND SEGMENTATION

The general market of neuromorphic circuits is expected to proliferate over the next years, independent of the specific underlying hardware which will be used. Several major market analysis companies[1,2,3,4] predict a Compound Annual Growth Rate (CAGR) in the range of 47.4% to 89.1%

[1] https://www.databridgemarketresearch.com
[2] https://www.alliedmarketresearch.com/
[3] https://www.researchandmarkets.com/
[4] https://www.mordorintelligence.com/

during the runtime of this research project and it is not to be assumed that this trend stops anytime soon. This is mainly issued in the widespread range of possible applications of neuromorphic technology. Applicable end-user industries for neuromorphic circuitry include the all-encompassing areas of consumer electronics, industrial electronics, automotive, financial services, cybersecurity, as well as other more fringe markets like medical, aerospace and defense.

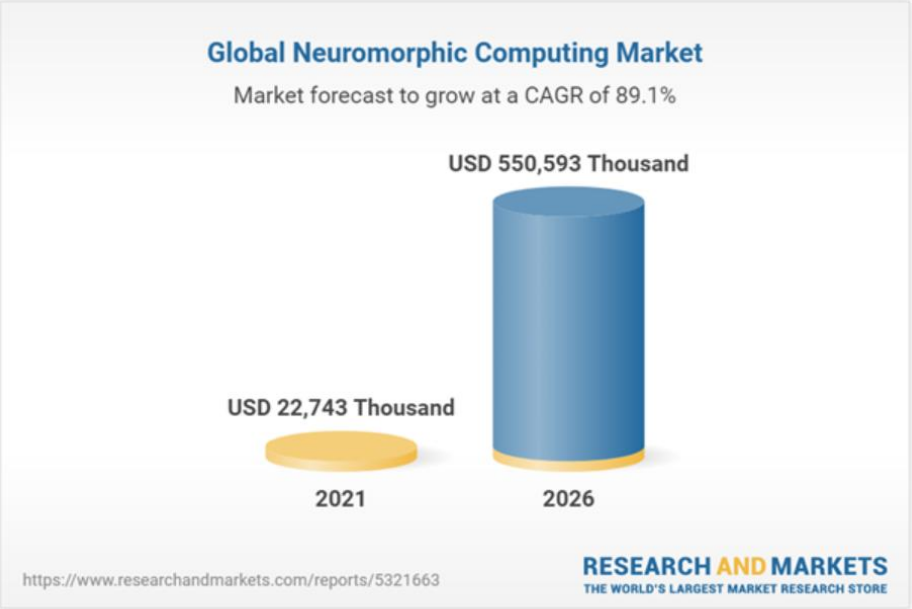

Figure 1 Global neuromorphic computing market forecast for the project duration; publicly available at given link.

In 2020 the image recognition segment accounted for more than 40% of the global neuromorphic computing market share[2]. This single application alone puts consumer electronics in the lead position in terms of market segmentation. However, it is expected that the market share of this application will soon be overtaken by real-time speech recognition and translation applications, which are the focus of the FVLLMONTI project. In order to enable this, a higher performance of the underlying hardware is needed.

While consumer electronics have the highest market share at the project start, automotive is the fastest growing industry to adopt neuromorphic chips. All the premium car manufacturers are investing heavily to achieve Level 5 of vehicle autonomy, which in turn, is anticipated to generate huge demand for AI-powered neuromorphic chips[2]. One and a half hour of driving, the typical time an American person spends a car each day, may lead to up to 4 Terabytes of data as estimated by an Intel analysis[5]. Processing this amount of data the classical way will require well over Teraflops-scale compute power and will have a noticeable impact on the driving range of electric vehicles. Thus, the autonomous driving market requires constant improvement in AI algorithms for high throughput with low-power requirements. Improvements are needed on software as well as hardware level.

[5] https://newsroom.intel.de/news-releases/self-driving-cars-theres-big-meaning-behind-one-big-number-4-terabytes/#gs.muc9dl

Beyond automotive, neuromorphic chips are expected to play a big role in Industrial Internet of Things (IoT), robotics, and healthcare, which need device for edge computing. Edge computing is a distributed computing paradigm that brings computation and data storage closer to the sources of data. In edge computing, a task which would be usually done by a central server processing unit is executed decentralized at the 'edge' of the network. This approach is expected to improve response times and save bandwidth. Edge computing can be seen as the adversary concept to cloud computing. The edge computing segment is expected to account for a 95% share of the overall neuromorphic computing market by 2026[6], putting a dense and lightweight kind of technology as targeted by FVLLMONTI into focus. Industrial IoT combines real-time processing, hardware optimization capabilities, and ubiquitous connectivity for IoT systems to maximize the efficiency of machines and the throughput of the entire process. Edge computing is one of the core components of accelerating the journey towards an industry 4.0. Also, applications from other areas, such as automotive sensors, gesture recognition, natural language processing (NLP), and virtual assistants, are expected to work with the edge computing principle, providing benefits to low-power and low-latency systems.

Finally, multiple features provided by neuromorphic circuits, such as self-learning capability, highly-parallel operation, and pattern recognition, are also beneficial for the adoption in financial services and cyber security. In the financial services landscape, neuromorphic chips are suitable options for predicting unconventional and high-frequency trading patterns. In the cyber security landscape, solutions tend to sequentially match small chunks of data against a library of suspicious patterns. The nature of current cybersecurity protocols follows a proactive and counter-response approach. Neuromorphic chips are envisioned to identify patterns in encrypted packets that could point to malicious payloads inside the traffic, where the aim is to work towards predictive cyber security protocols.

The overall neuromorphic computing market is segmented in hardware and software, where a close entanglement is needed from both sides to be competitive. The market is dominated by US and Asian Pacific (Japan, China, South Kora, India) companies, followed by Europe (Great Britain, Germany, France, Italy). In Europe, the share of edge computing applications as compared to cloud computing applications is higher due to its relevance for the automotive industry. In 2020 the major shareholder among European countries was Great Britain.

III. STAKEHOLDERS AND COMPETITORS

The market for neuromorphic computing is in an intermediate development phase. On the one hand, the need for skillfully combining ultra-scaled hardware and complex software algorithms poses a high hurdle for start-ups and small-medium-enterprises (SMEs) to compete successfully. On the other hand, development is driven by many players both in research and academia, thus there is no single company holding a monopoly. Hereafter a list of key market players is given, divided by region. These key-market players are somewhat competitors but can also serve as potential stakeholders once interested in the FVLLMONTI technology.

[6] https://www.globenewswire.com/

- USA : Intel Corporation, BrainChip Holdings Ltd., Qualcomm Incorporated, Hewlett Packard Enterprise Development, IBM Corporation, Numenta, General Vision Inc., Vicarious Robotics, Cisco Systems
- Asia : Huawei (China), SAMSUNG (South Korea), Nepes Corp (South Korea)
- Europe : SixSq (Switzerland), SynSense AG (Switzerland), Infineon (Germany), Nokia (Finland), Axelera AI (Belgium), STMicroelectronics (France)
- Rest of World : Applied Brain Research Inc. (Canada), Saguna (Israel)

The primary competitor technology outside the neuromorphic computing world are GPU-based systems. Initially developed for video processing, GPUs use parallel processing to perform mathematical operations. Due to this parallel architecture, they also achieve the high-throughput targeted with neuromorphic technologies. They are more accessible and easier to program than typical neuromorphic chips. GPUs already have high support in terms of drivers and software for deep-learning algorithms. However, they are very power-intensive and thus unsuitable for a lightweight mobile device. Also, covering the expected market share for neuromorphic circuits fully by GPUs would lead to conflicts with worldwide sustainability aims regarding carbon reduction. There are only two prominent vendors for GPUs worldwide: Nvidia and AMD.

3. ANALYSIS AND RISK ASSESSMENT OF KEY TECHNOLOGICAL INNOVATION OF THE PROJECT

I. TECHNOLOGY IMPACT OF VERTICAL NANOWIRE PLATFORMS FOR NEURONAL NETWORKS

In FVLLMONTI two basic types of disruptive nanowire technologies are researched. Junctionless (JL) transistors on the one hand, and ambipolar transistors with active polarity control (PC) on the other hand. Both serve different roles within the technology platform. While the JL transistor is expected to replace classical CMOS transistors at ultra-scaled nodes, the polarity-controllable ambipolar transistors pose an add-on function, which can be co-integrated into either JL or classic CMOS, where the added benefit is needed. Thus, different performance levels have to be achieved for both technologies to make an impact. For the JL gate-all-around nanowire technology, we have analyzed the energy-delay-product (EDP) in a technology-agnostic approach already at the beginning of the project[7]. Comparisons with the EDP of a baseline 7 nm FinFET technology have been carried out. The analysis motivates a new 3D neural network compute cube (N2C2) concept on the system level. Our results show that a 10x gain in EDP can be achieved for a physical vertical nanowire FET gate length of 14 nm. Even for the actual fabricated devices at the project start an 4.3x EDP gain has been predicted[8]. The actual value that can be achieved depends on the 'compactness' of the final designs, as given by an arbitrary factor A_c (representing the circuit-level footprint gain of vertical over lateral

[7] I. O'Connor *et al*., "Analysis of Energy-Delay-Product of a 3D Vertical Nanowire FET Technology," doi: 10.1109/EuroSOI-ULIS53016.2021.9560180

[8] Y. Guerfi and G. Larrieu "Vertical Silicon Nanowire Field Effect Transistors with Nanoscale Gate-All-Around" doi: 10.1016/j.sse.2016.12.008

technologies) in Fig. 2, which has to be optimized by a design-technology-co-optimization (DTCO) procedure. An even higher gain can be expected after several gate layers of transistors have been stacked on-top of each other. The potential impact is harder to predict precisely for polarity-controllable transistors, which are in a much earlier development stage. On the one hand, the benefit of adding these devices to the system level must first be demonstrated. On the other hand, the fact that an add-on functionality to CMOS is targeted here, instead of a replacement, lowers the entry-level for potential applications once these applications are found.

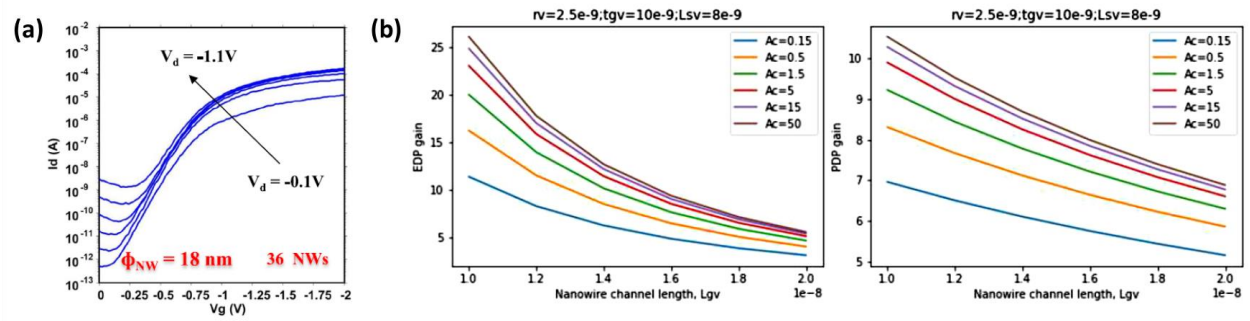

Figure 2 Evaluation of the technology impact of the vertical junctionless nanowire transistor platform[7]. (a) p-type transfer characteristics of nanowire FETs with a nanowire diameter of 18 nm. (b) Early technology impact estimation based on EDP and PDP gain calculations as compared to a 7 nm FinFET technology. rv is the nanowire diameter, tgv the dielectric thickness and Lsv the space length. Ac is the gain factor for the 'compactness' of the design at the circuit level[8].

Risk assessment: Probability: low, medium, high / impact: low, medium, high

Outperformance by planar technologies: high high The most significant risk to the industrial exploitation of the underlying nanowire technologies is the rapid progress of conventional planar technologies driven by industry. Major semiconductor manufacturers, such as Intel, Samsung and TSMC, have announced to move towards 'stacked-nanosheets' in production roughly by the end of the project. While the devices are branded under various names, such as *RibbonFET*[9] or *Multibridge-FET*[10] all of them share the basic design principle shown in Fig. 3(c). They are still a planar technology, where multiple thin channels are stacked on top of each other, thus increasing the device's width and performance without increasing the area[11]. On the other hand, having multiple nanowires in parallel in a vertical technology would increase the footprint again. In a simple approximation (Fig.3(d)), vertical technologies lose their advantage of having a lower single device area once more than three layers can be sufficiently stacked. Therefore, the main competitive advantage of vertical technologies remains the stacking of multiple layers of devices on-top of each other, which goes hand in hand with the reduced routing overhead. A key task here is to determine how many stacked vertical devices are needed to make this advantage outweigh the individual device performance of planar technologies.

[9] https://www.guru3d.com/news-story/intel-introduces-ribbonfet-transistor-architecture.html
[10] https://samsungatfirst.com/mbcfet/
[11] https://spectrum.ieee.org/the-nanosheet-transistor-is-the-next-and-maybe-last-step-in-moores-law

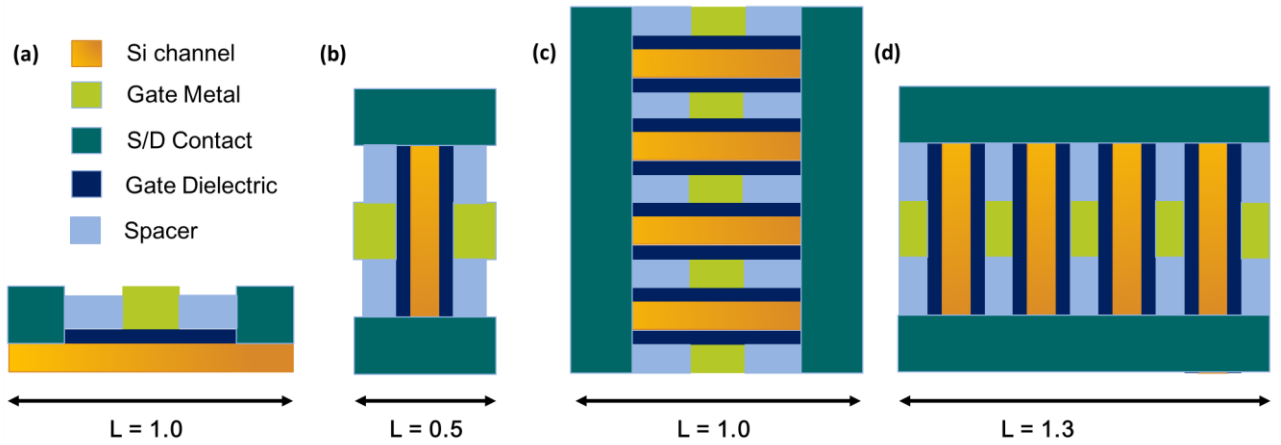

Figure 3 Simple area comparison of lateral and vertical transistor geometries assuming equal gate length, channel length, and spacer lengths across all technologies. (a) single layer planar MOSFET, (b) single nanowire vertical transistor, (c) stacked-nanosheet transistor with four overall layers, (d) parallel vertical nanowire array device with four channels. While a 50% area gain can be predicted for the single nanowire variant, the competitive advantage is lost if four or more layers are assumed to drive a single device in parallel. This simple assessment motivates the need for stacking multiple vertical devices on-top of each other as targeted in FVLLMONTI.

Equal p- and n-type performance: **Medium** ● **medium** To provide a fully complementary platform, p- and n-type devices with relatively equal performance have to be demonstrated. In classical bulk technology, n-type devices are favored over p-type devices due to the higher intrinsic mobility in silicon. However, as dimensions approach the nanoscale, other effects such as achieving good ohmic contacts for each carrier become more prominent. While platinum silicide has been shown to be a good candidate material for p-type contacts, the performance in n-type devices is expected to be lower. Good n-type performance remains demonstrated with the vertical junctionless technology. Polarity-controllable transistors might help mitigate this risk, providing p- or n-type functionality on request using an external electrical programming voltage (electrostatic doping).

Industrial property rights on vertical technologies: **low** ● **high** As of writing this report, no company on the market fabricates vertical nanowire transistors on an industrial scale. However, strong competition is underway at the research and development level. For example, IBM and Samsung are working on a vertical nanowire tunnel-FET technology, which they have announced to provide as much as 85% power reduction as compared to scaled FinFET[12]. In addition to the big players, fabless companies, like Unisantis from Singapore, have acquired a considerable amount of industrial property rights for vertical nanowire technologies[13]. These activities pose the risk of limiting the freedom of action of the FVLLMONTI technology. In order to compete, the consortium partners hold essential key patents to ensure the exploitation of the FVLLMONTI results. It is vital that the consortium place and advertise these in the proper spotlight on the market. Also, new patents should be filed, wherever possible, to ensure the continued freedom to operate of the partners.

[12] https://newsroom.ibm.com/2021-12-14-IBM-and-Samsung-Unveil-Semiconductor-Breakthrough-That-Defies-Conventional-Design
[13] https://unisantis.com/technology/

II. TECHNOLOGY IMPACT OF TOOLS AND TECHNOLOGIES FOR ANALYSIS AND DEVELOPMENT OF EMERGING ELECTRONIC SYSTEMS

Today, the accumulated annual turnover of semiconductor manufacturers all over the world is estimated at 400 billion Euro with an average growth rate of about six to eight percent. The majority of revenue is generated with integrated circuits, with a smaller part being contributed by discrete devices. The semiconductor industry is characterized by very high investments in Research and Development. One major part of R&D spending comes from Electronic Design Automation (EDA) tools. The total market size for EDA tools in 2021 is estimated at 10 billion Euro with growth rates of roughly ten percent, even surpassing that of the semiconductor industry as a whole. While the EDA market has a diverse set of specialized competitors, it is dominated by the Big Three, namely Synopsys, Cadence, and Siemens EDA (former Mentor Graphics).

Technology Computer-Aided-Design (TCAD) systems were introduced in the 1980s, and their usage has been steadily increasing due to the increasing complexity in device structures and manufacturing processes. The TCAD share of accumulated R&D expenses amounts to an estimated world market of 200 million Euro. It represents a much smaller portion of the EDA market and is dominated by Synopsys.

The systematic and hierarchical inclusion of the FVLLMONTI innovations into GTS' DTCO tool flow will strengthen GTS position to cover all aspects of the TCAD market. It will also be of great value to address the much larger EDA market. The demonstration of a full vertically integrated DTCO flow from the technology up to the system level, combined with the predictive accuracy of GTS Nano Device Simulator, will provide a complete toolset for design-enablement and path-finding to fabless companies. Some of the most relevant companies in the field are already GTS customers. It is expected that the increased feature set in the follow-up of the project will enable GTS to 1) increase the volume with existing customers; 2) successfully engage with new customers. The innovations on the DTCO flow can be considered as low risk with medium impact.

One particular helpful activity in this regard is the model device and parasitics compact model building at UBx. In the FVLLMONTI project, the device parameter extraction strategy will play a major role in feeding into the DTCO and associated logic cell design flow. This strategy will leverage new RF test structures specifically designed on the vertical nanowire transistor technology platform as well as a fully functional SPICE compact model in Verilog-A for vertical junctionless nanowire transistors with and without a ferroelectric gate[14]. In this innovative approach, through technology layout design optimization combined with electromagnetic simulation of the test structures, intra-, and inter-cell parasitic interconnects can be modeled and minimized for the DTCO flow. Along with 3D P&R for cell layout optimization, this should further improve key performance metrics such as the energy-delay-product (EDP) of the demonstrator targeted in the FVLLMONTI project. The vertical Fe-gate nanowire transistor SPICE compact model will then be combined with the extracted parasitic interconnect network for realistic 3D logic cell simulation and extrapolation for performance prediction at projected ultimate technological dimensions.

[14] C. Maneux et al. “Modelling of vertical and ferroelectric junctionless technology for efficient 3D neural network compute cube dedicated to embedded artificial intelligence” (Invited). *67th Annual IEEE International Electron Devices Meeting (IEDM 2021)*, Dec 2021, San Fransisco, United States.

Risk assessment: Probability: low, medium, high / impact: low, medium, high

Delay in test structure delivery: low medium: Device parameter extraction for the DTCO flow relies on the delivery of new test structures for characterization. Any potential delay in test structure delivery will thus be of critical risk for subsequent design cycles. However, thorough planning of the upcoming technology runs and alignment with tasks downstream will mitigate this risk. As an alternative measure, TCAD simulation data from preceding runs could also substitute for cell performance prediction and analyses.

Technology dispersion and variability in model calibration: medium low: Anticipating dispersion and variability in the vertical nanowire transistor technology of FVLLMONTI, a comprehensive set of characterizations will be set up to determine representative devices that can be used for model calibration. Eventually, the calibrated TCAD models will be used for performance prediction and extrapolation, rendering technology variability less of a critical concern for the objectives of the FVLLMONTI project.

A completely new research aspect covered in FVLLMONTI will be the 3D place & route algorithms. These will be of increased interest in the mid-term, as 3D-integrated IC technology becomes more and more industry-relevant. Hence, the work of FVLLMONTI could put GTS in a unique starting position in this new EDA tool market, as the FVLLMONTI partners explore the field of 3D designs with a hands-on application for neuromorphic computing. Such a demonstrator could successfully highlight the real-world benefits of the 3D P&R prototype.

GTS is currently 100% export-oriented, with the main focus on the Asian market. The outcome of FVLLMONTI will address topics in the focus of European companies and institutions as highlighted by the advisory board members. The expected innovations enabled by this project will be of great interest to them and could enable GTS to win them as customers in the mid to long term.

Risk assessment: Probability: low, medium, high / impact: low, medium, high

Customer needs for 3D place & route: high high: As there is no company manufacturing vertical nanowire technologies yet, the impact of a 3D place & route algorithm is directly coupled with the emergence of an industrial-scale technology in need of this tool. We can consider this as a high risk but high reward innovation.

III. TECHNOLOGY IMPACT OF TRANSFORMER ARCHITECTURES

In FVLLMONTI, two basic types of disruptive nanowire technologies are researched. Junctionless (JL) transistors on the one hand, and ambipolar transistors with active polarity control (PC) on the other hand. Both transistor types can be combined with a ferroelectric (FE) gate layer to yield a non-volatile storage functionality, leading to four possible different device types. In the overall FVLLMONTI vision, all four transistors are combined to yield the most compact and high-performance transformer architectures. However, as intermediate steps, architectures will be researched based only on a subset of these combinations, as summarized in Table 1. Stacking

multiple transistors on-top of each other is probably crucial to yield a benefit over classical horizontal technologies. A technology variant with 2 stacked gate layers will be demonstrated in hardware within the project. The possibility of up to 3 stacked layers is considered from the circuit development point of view. As a result, transformer architectures will be developed based on various technological variants of the FVLLMONTI vision. The most promising of these variants should be identified within the project runtime and then evaluated for further exploitation.

The concept of the N2C2 block as a highly flexible 3D building block to enable dense hardware accelerators that are exactly tuned to the needs of a given machine learning application is highly disruptive. Barriers to its acceptability in mainstream architectures are therefore high. To lower these barriers, the consortium is well aware that a) small-scale hardware demonstrators must demonstrate its concrete operation, b) large-scale projections must demonstrate significantly (at least 10x) improvement over other competing approaches, and c) a solid roadmap with well-defined timescales and metrics to achieve large-scale hardware must be in place.

Risk assessment: Probability low, medium, high / impact: ● low, ● medium, ● high

N2C2 concept performance vs. cost trade-off: low ● low The highest risk for the transformer architectures arises from the fact that the N2C2 concept is highly disruptive and may not prove the necessary high-performance gain needed to justify the transition to a new technology from an economic point of view. This risk is minimized by analyzing a multitude of different technology variants. This way, it is ensured to yield a subset of devices and features, which provides the highest benefit for the envisioned N2C2 application, keeping the need for new technologies to the minimum. If the N2C2 concept by itself demonstrates less potential than expected, then the developed transistor technologies and tools can be easily transferred to other applications, minimizing the impact of this risk. Conversely, if the N2C2 concept demonstrates potential on other technologies, then the developed architecture can also be targeted to these other technologies.

Table 1: Architectural Variants to be analyzed in FVLLMONTI composed of junctionless (JL), ambipolar (AP) and ferroelectric (FE) parts in different combinations and numbers of gate stacks.

Variant	Gate Stack	No. of devivce types	Junctionless	Reconfigurable	Ferroelectric	Hardware
JL1	1	1	Yes	No	No	Yes
JL2	2	1	Yes	No	No	Yes
JL3	3	1	Yes	No	No	No
JL1FE	1	1	Yes	No	Yes	Yes
JLFE1	1	2	Yes	No	Yes	No
JLFE2	2	2	Yes	No	Yes	No
1AP	1	1	No	Yes (U)	No	Yes
JL1AP	1	2	Yes	Yes (U)	No	No
2AP	2	1	No	Yes	No	No
JL2AP	2	2	Yes	Yes	No	No
1APFE	1	2	No	Yes (U)	Yes	No
JLFE1AP	1	3	Yes	Yes (U)	Yes	No
JLFE1APFE	1	4	Yes	Yes (U)	Yes	No
FVLLMONTI	2	4	Yes	Yes	Yes	No

IV. TECHNOLOGY IMPACT OF AUTOMATIC SPEECH RECOGNITION AND MACHINE TRANSLATION

In FVLLMONTI, we plan to explore the performance of N2C2 accelerated systems. To this end, we will target two specific and timely application scenarios: automated speech recognition and machine translation. These applications will be embodied as Transformer networks[15], a novel neural network topology able to process input data streams of variable lengths to produce corresponding output data streams. In the context of the project, we plan to devise novel hardware-friendly Transformer optimizations, enabling a hardware/software co-design loop between the software architecture of the network being executed and the hardware architecture of the system hosting it.

To this end, we plan to extend the gem5-X system simulator[16] with new and dedicated components devoted to the behavioral modeling of N2C2 accelerators. The resulting environment will allow an architectural exploration of the trade-offs implied in different topologies in terms of area, runtime, and energy. It will also allow a variety of integration strategies to be studied, ranging from the tightly-coupled integration of N2C2s as per-core functional units up to loosely-coupled strategies based on arbitration and memory mapping.

Ultimately, the goal of the FVLLMONTI project in the system exploration space is to provide a novel framework for the development and early assessment of heterogeneous, accelerator-rich systems, with a particular focus on supporting upcoming neuromorphic compute strategies. Our strategy allows the impact of low-level design choices to be assessed, down to the design of each hardware macro, on high-level metrics such as the accuracy of speech recognition and translation.

Risk assessment: Probability: low, medium, high / impact: low, medium, high

N^2C^2 acceleration does not lead to substantial speedups: low high. The runtime of Transformers is dominated by matrix operations (in particular, matrix-matrix and matrix-vector multiplications). Their regular computation and memory access patterns make them particularly well-suited for N2C2 accelerators. Moreover, the flexibility of the N2C2 design can be leveraged also to address second-order bottlenecks, e.g. the computation of activation functions. Finally, we plan to provide different integration strategies to overcome the impact of data transfer overheads in-between processors and accelerators.

Transformer architectures are not amenable to runtime optimization: low medium. The employed data representation plays a major role in achieving low energy and area. In particular, the use of small bit-width integer representation can lead to substantial efficiency gains. Our preliminary studies suggest that Transformers, similarly to other neural networks, are robust towards this optimization. Such observations are also paving the way for more structured strategies to harness robustness, such as per-tile pruning, which we are investigating as an avenue to reduce computational effort requirements.

[15] Vaswani, Ashish, et al. *"Attention is all you need."* Advances in neural information processing systems 30 (2017).
[16] Qureshi, Yasir Mahmood, et al. *"Gem5-X: A Gem5-based system level simulation framework to optimize many-core platforms."* 2019 Spring Simulation Conference (SpringSim). IEEE, 2019.

4. ANALYSIS OF POTENTIALLY EXPLOITABLE PROJECT ASSETS

I. KEY INNOVATION EXPLOITATION ACTIONS AND PLANS

The key innovations developed in FVLLMONTI and the current plans and actions on exploitation are summarized in Table 2 below.

Table 2: Key Innovations and Exploitation Plans

Key Innovation	Exploitation Plans and Actions
Vertical Nanowire Platforms (JL and PC)	▪ Key features are protected by patents / patent applications prior to project start ▪ Further patents are to be filed wherever possible to ensure freedom to operate of the project partners ▪ Promotion of patents to foundries outside the advisory board ('roadshow')
RF Test structures for Vertical FE-FET	▪ Distribution at conferences (EuroSOI-ULIS 2022 planned) ▪ Open-access publication as white paper at the end of the project
Vertical FET Compact Models (JL, FE, PC)	▪ Distribution at conferences (IEDM, VLSI-SOC, EuroSOI-ULIS), Solid-state Electronics, Distribution to users to be planned, via cloud storage options like Nanohub ▪ Contact to be established with Compact Model Consortium (CMC) for distribution ▪ Establish direct contact to EDA vendors via Advisory Board and outside of that (Synopsis, CEA-Leti, TSMC, STMicroelectronics, IBM, Globalfoundries) ▪ Promote results in business networks like Silicon Saxony
Physical Design Tools (3 D wizard integration, Interconnect 3D P&R tool)	▪ Potential exploitation development within the consortium by the SME partner GTS
Application-Aware Architecture (Gem5-X environment supporting N2C2 accelerators; 3D based systolic array architecture; efficient programming and mapping methodology)	▪ Open-source software, documentation ▪ Foster adoption in the academic and industrial community

II. ASSESS THE USABILITY OF THE RESULTS BEYOND THE CONTEXT OF THE PROJECT

In its current early state, it can be assumed that all of the key results of the projects can be used beyond the context of the project. The developed underlying technology, models, characterization techniques, and software tools can be re-used for other research topics. The development circuits and algorithms can be adapted to other applications beyond NLP. It is planned to make parts of the research results available for open access as described in the data management plan (DMP). Key innovations will be patented to secure know-how for further exploitations, where applicable.

5. CONCLUSION

This document describes the initial technology impact and exploitation action assessment associated with FVLLMONTI. Market segmentation and the socio-economical importance of neuromorphic circuits have been discussed. FVLLMONTI activities have been ranked among stakeholders and competitors. Key innovations of the project have been identified, which can be attributed to four different application areas:

- Emerging Vertical Nanowire Hardware
- Tools for Electronic Design Automation
- Transformer Architectures and Circuit Designs
- Software for Automatic Speech Recognition and Machine Translation

The research in each area can have a positive impact on their own; however, the full potential will be just unfolded by combining all developments towards the FVLLMONTI vision. The developed underlying devices, models, characterization techniques, software tools, circuits, and algorithms all show their potential to be exploited beyond the project's runtime. A first exploitation plan has been set up.

As the FVLLMONTI project progresses, the content of this deliverable will be updated to reflect changes in terms of market development or new technological insights.

FVLLMONTI

=== End of the document ===