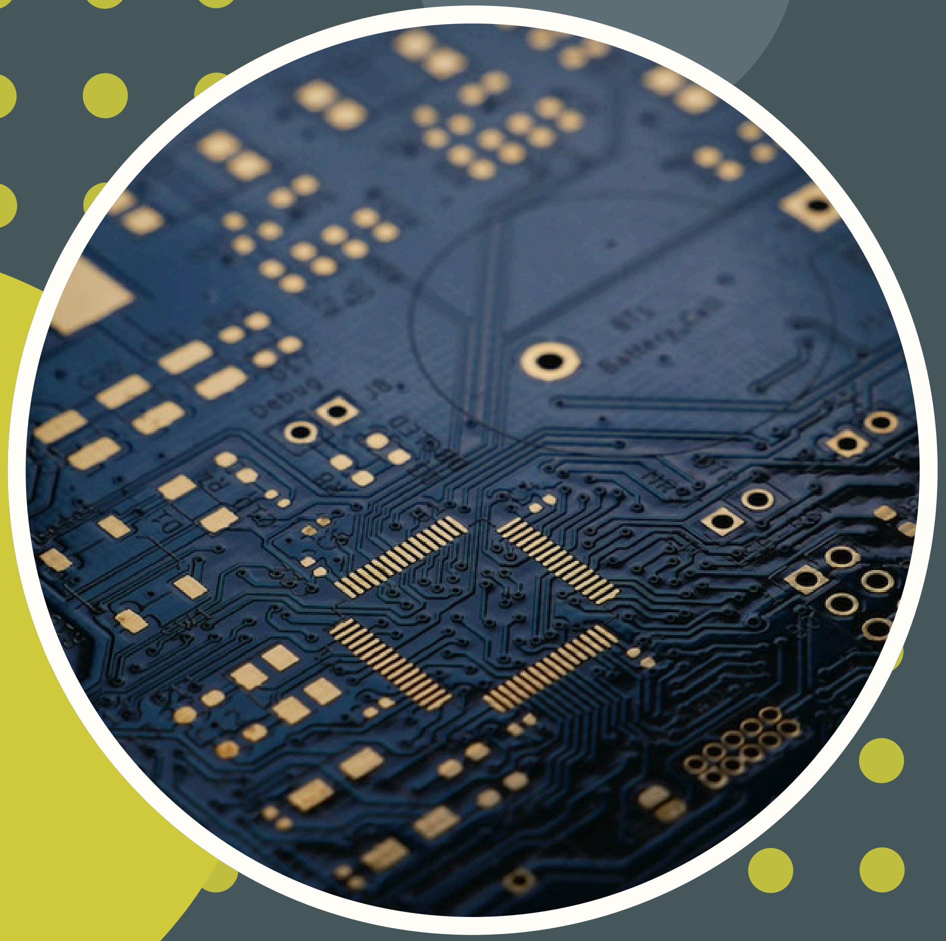
CONVOLVE

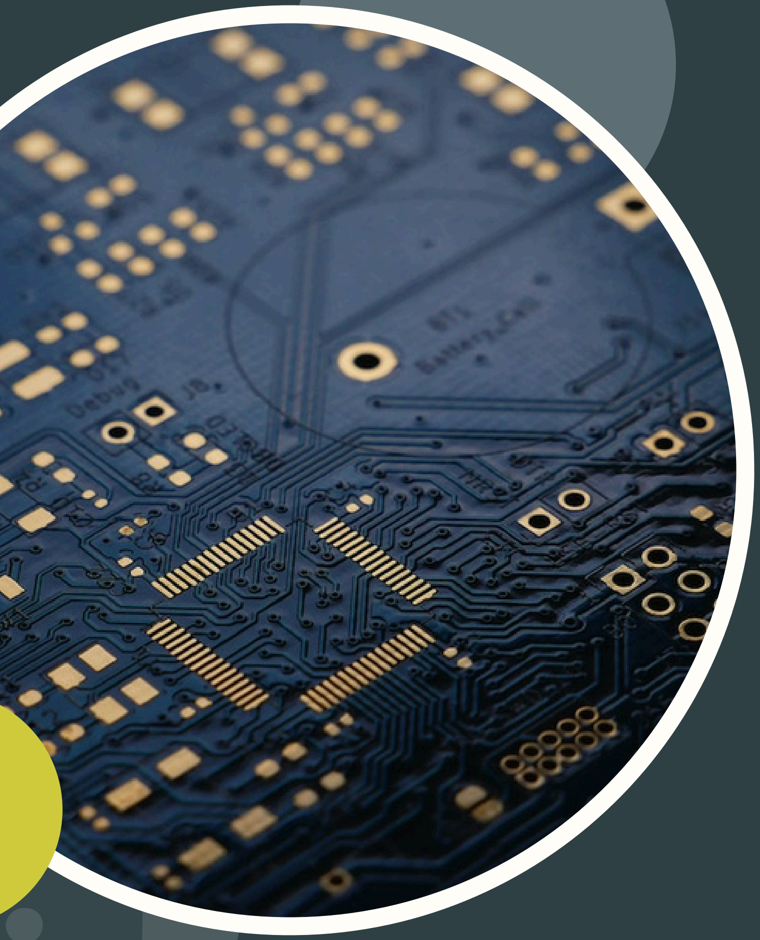


Seamless design of smart edge processors

www.convolve.eu

CONVOLVE

Table Of Contents



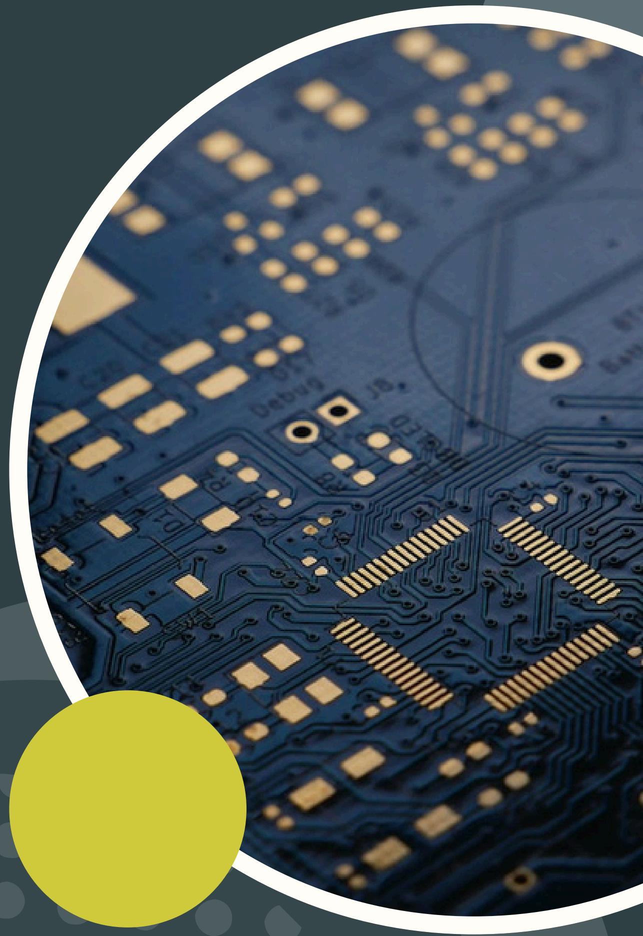
Introduction

CONVOLVE enforces the EU's position in the design and development of smart edge processors, such that it can become a dominant player in the global edge-processing market.

Emerging and rapidly growing Deep Learning (DL) has shown exceptional performance in addressing complex tasks, such as image classification or object detection, and the rapid growth in these methodologies has allowed the accuracy to improve continuously thanks to the more advances in training methods. While the accuracy of such tasks is continuously improving, thanks to the more complex network topologies that have been developed and deployed, it imposes challenges in supporting such inference efficiently on resource-constrained edge devices.

Traditional computing platforms such as CPUs and GPUs are insufficient to infer such networks in constrained systems due to limited power and area budgets, real-time and/or latency requirements, etc. Therefore, for such systems, dedicated hardware accelerators are designed and deployed.

Nonetheless, dedicated hardware accelerators have their limitations to cope with different network topologies and are typically optimized towards specific tasks with specific networks. Accordingly, we need to evaluate the network topologies in detail to improve the flexibility of hardware accelerators and utilize this information in devising hardware architecture for such accelerators.



CONVOLVE

Objectives

1. Improve energy efficiency by 100x
2. Reduce design time by 10x
3. Achieve long-term hardware security
4. Enable smart AI models

CONVOLVE

Objectives

Objective 1 Improve energy efficiency by 100x

Achieve 100x improvement in energy efficiency compared to state-of-the-art commercial off-the-shelf (COTS) solutions by developing near-threshold self-healing dynamically reconfigurable accelerators. This involves the development of an Ultra-Low-Power (ULP) library with novel architectural and micro-architectural accelerator building blocks, in short ULP blocks, having common or standard interfaces, and optimized at micro-architecture, circuit, and device levels. Different architectural paradigms will be evaluated, such as Compute-in-Memory (CIM), Compute Near Memory (CNM), and Coarse-Grained Reconfigurable Arrays (CGRA), all keeping processing very close to the memory to reduce energy consumption.

The accelerator blocks are optimized to execute the computation patterns of both Artificial Neural Networks (ANN) and Spiking Neural Networks (SNN) efficiently. To further reduce energy consumption, support for application dynamism will be provided in ULP blocks to dynamically adapt computational precision, data path width, early-termination, skipping layers/neurons, etc. Leakage will be reduced by advanced power management, and by using non-volatile ReRAM-based crossbar units. Novel self-healing mechanisms will be introduced to deal with hardware variability at near-threshold regions.

CONVOLVE

Objectives

Objective 2 Reduce design time by 10x

Reduce design time by 10x to be able to quickly implement a ULP edge AI processor combining innovations from the different levels of the stack for a given application using a compositional ULP design approach. CONVOLVE researches efficient design-space exploration (DSE) techniques combining different levels of hierarchy in a compositional way, i.e., hardware and software components can be seamlessly glued together while guaranteeing overall behaviour and reliability; this deals with the SoC heterogeneity and supports the efficient mapping of applications to hardware architectures. Designing ULP accelerator blocks with common interfaces will allow these to be plugged into a modular architecture template. We will then generate an SoC architecture using these modular architecture templates after performing automated DSE; this allows the evaluation of all architecture possibilities.

CONVOLVE

Objectives

Objective 3 Achieve long-term hardware security

Provide hardware security against known attacks and real-time guarantees by compositional Post Quantum Cryptography (PQC) and real-time Trusted Execution Environment (TEE). We will design PQC accelerator blocks with standard interfaces that can be plugged into a modular architecture template to make hardware secure, even in the long term (over a decade).

Furthermore, CONVOLVE develops design-for-security shames and makes sure that all security features can be added in a compositional manner while providing real-time guarantees. We will explore design for robustness, to deal with in-field failures and non-ideal real-world environments.

CONVOLVE

Objectives

Objective 4 Enable smart AI models

CONVOLVE will develop smarter AI models to be combined with ULP accelerators. The project will explore AI models which dynamically adapt to the data input, such that the 'common input case' can be executed much more efficiently. This will dramatically reduce energy consumption. Furthermore, inspired by the redundancy and self-healing properties of biological brains, we enhance reliability by online (re-)learning, and adapting parameters and weights on the fly. This requires new and cheap learning algorithms.

Finally, CONVOLVE investigates whether spiking neural networks (SNNs) could have an edge w.r.t. ANNs for certain application domains, especially for streaming input and always-power-on attention blocks.

CONVOLVE

Preliminary Work-Package Results

WP1
Use case and application requirements, metrics and baseline

WP2
Self-configurable modular ULP accelerator blocks

WP3
Composable Real-Time and Hardware Security

WP4
Algorithmic principles for ultra-low power neural network processing

WP5
Transparent and compositional programming flow

WP6
Compositional architecture DSE and SoC generation

WP7
Application mapping, benchmarking and integrated demos

CONVOLVE

Preliminary Work-Package Results

WP1 Use case and application requirements, metrics and baseline

WP1 lays the groundwork for edge application use-case by defining requirements and benchmarks for smart edge processors. Initial point demos will guide target adjustments, supporting technical work package topics with application-focused approach. Key achievements include delivering official documentations D1.1 and D1.2, establishing code bases for work in WP2, WP4, WP5 and WP6 at TUE, and publishing state-of-the-art research on quantization, efficient deployment of neural networks for resource efficient speech quality prediction, acoustic scene analysis and image processing.

Additionally, student supervision and dissemination efforts such as TinyML and Danish Digitalization, Data Science and AI 2.0 amplify the project's societal impact, ensuring a robust foundation for advancing Europe's edge computing capabilities.

CONVOLVE

Preliminary Work-Package Results

WP2 Self-configurable modular ULP accelerator blocks

WP2 focuses on the development of the key components and blocks of the targeted accelerators. Focus will be done not only on low power design aspects to contribute to 100x+ energy efficiency improvement, but also on dynamic configurability (to adapt the accelerators depending on the needs), modularity (to speed-up the design process), and self-healing (to deal with the non-idealities of hardware). The key achievements of WP2 includes the following points:

- successful simulation-based demonstration of energy-efficient analog and digital CIM accelerator design.
- Timely delivery of results for deliverables where deliverables D2.1, D2.2, D2.3 and D2.6.
- Establishing cross-workpackage synchronisation where use case applications from WP1 are exposed to benchmark the ULP blocks developed in WP2. In this case, an object detecion application from the Vinotion usecase has been used to benchmark the ULP accelerators.
- Publication of CONVOLE research output.

CONVOLVE

Preliminary Work-Package Results

WP3 Composable Real-Time and Hardware Security

WP3 aims at simultaneously providing long-term, software and hardware security as well as real-time guarantees of the device. In D3.1 the relevant attack scenarios were identified and include side-channel attacks based on physical access to the device, mutually distrustful applications on a single device and even powerful attackers with access to large-scale quantum computers. To account for this wide range of scenarios, WP3 is working on four interrelated frontiers: Trusted Execution Environments (TEE), Post-quantum cryptgraphy (PQC), security of compute-in-memory (CIM) and real-time guarantees. TEEs provide a means of isolating software and its (potential) vulnerabilities as well as offering secure boot and remote attestation to ensure that the running software has not been tampered with.

We have a working prototype based on Keystone and the Rocket core. PQC is the vital component to equip the chip with security measures that will resist even future attackers, i.e., if large-scale quantum computers become available. We have conducted a careful exploration of schemes to determine the best fit for the specific needs of Convolve. On the CIM frontier, we are able to extract neural network weights from the crossbars via power side-channels and are working on effective countermeasures. Real-time guarantees pose a particular problem in combination with TEEs and the current focus lies on developing a strategy to consolidate both paradigms harmoniously. For the coming months we work on further advances on each of the four frontiers individually as well as the integration within WP3 and the entirety of the Convolve project.

Preliminary Work-Package Results

WP4 Algorithmic principles for ultra-low power neural network processing

The freedom to explore radically new ideas in WP4 provides a high-risk, high-reward environment to contribute to the low-power goals of CONVOLVE, focusing on the target applications supplied by the industrial partners on the project.

Work to date on WP4 includes:

- Optimisations of the compute efficiency in traditional deep neural networks by limiting operand precision (quantization) and limiting connectivity between neurons ('sparsification' and pruning), all with minimal loss of performance and classification accuracy
- Control flow techniques (called Dynamic Neural Networks) that can detect when sufficient processing has been carried out on an input, avoiding unneeded computation
- New, more efficient learning rules to replace the costly backdrop algorithm

CONVOLVE

Preliminary Work-Package Results

WP5 Transparent and compositional programming flow

WP5 targets transparent and compositional programming flow to realise an effective heterogeneous ULP accelerator SOC. This is done by developing a modular compiler to generate high-performance and secure code for the rapidly evolving matrix of applications and many-accelerator hardware. Our xDSL framework integrates all parts of compilation, featuring neural network ingestion, low-level hardware abstractions, micro-kernel code generation, workload-specific optimizations, and secure, interactive compilation.

In more detail, we have built the front-end compilation chain, targeting both MLIR and LLVM IR. We leverage MLIR and LLVM ecosystems for RISC-V-specific optimizations – such as instruction reordering to enable instruction fusion – and work on integrating design space exploration (ZigZag/Stream) tools from partners to optimise and match the code to CONVOLVE accelerators.

At the code generation front, our domain-specific compiler lowers operations to custom RISC-V ISA accelerator extensions. By leveraging linear algebra operations, we target Snitch's custom ISA extensions, transforming nested loops, at the heart of popular ML kernels, from improved data streaming and accelerated computation. We have also contributed to the IREE OSS ML compiler framework, having developed an ONNX importer that converts ONNX modules to Torch MLIR for IREE compilation. Along with other ecosystem updates, we work on adapting IREE for the Snitch architecture (via runtime plugins and code generation). We have introduced a scalable cache model for ML (affine-heavy) programs, significantly reducing analysis time and enabling efficient updates after program modifications. Lastly, in terms of secure compilation, we have facilitated peephole rewrite validation at the intermediate representation level by automating integration with interactive theorem provers.

CONVOLVE

Preliminary Work-Package Results

WP6 Compositional architecture DSE and SoC generation

WP6 deals with automated compositional system architecture design space exploration (DSE) and SoC generation. This is, on the one hand, done by providing a multi-accelerator architecture simulator, Stream, to analytically study optimal SoC architectures. On the other hand, a modular architecture template can easily hook up one or multiple ML and security accelerators to a RISC-V host and tightly coupled memory system. The resulting design subsequently enters an automated design time instantiation flow for run-time flexible SoC generation.

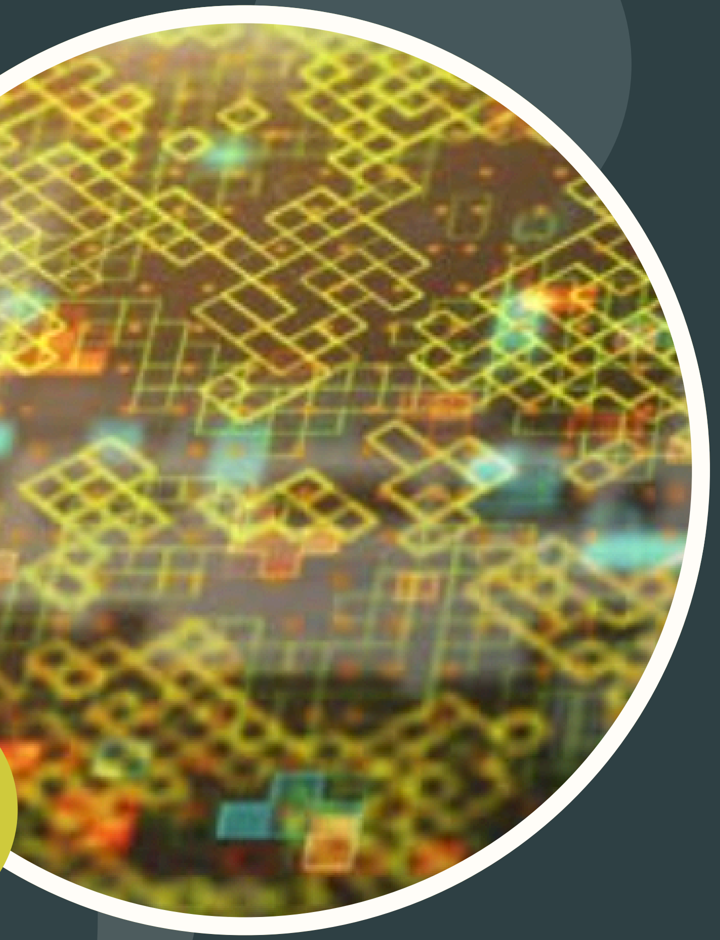
CONVOLVE

Preliminary Work-Package Results

WP7 Application mapping, benchmarking and integrated demos

WP7 aims to show in a unified manner the work and developments throughout the different working packages in CONVOLVE. By mapping the use cases' applications into the different hardware and software developments, the requirements are confirmed and validated. The analysis of the performance improvement in the point demonstrators of each of the work packages with respect to a benchmark is conducted throughout the project.

A moonshot demonstrator will be achieved, which is ambitious and intends to prove the fulfilment of CONVOLVE's objectives through the integration of the different innovations. The final tape-out silicon will have different accelerators, an ULP architecture, and a compilation flow that enables different applications to be executed efficiently. Finally, an analysis of the market gaps and the actions that need to be taken to bring these developments to the market will be carried out.



Relevant Documents, Deliverables & References

The following will be a list to relevant documents, deliverables and references:

- D1.1 Initial requirements and use cases
- D2.1 Report on the roadmap
- D2.2 Intermediate report on the design of the targeted accelerator blocks
- D3.1 Requirements, Threats, and Vulnerabilities Analysis
- D4.1 Roadmap document for low power, high performance neural networks
- D4.2 Technical requirements for hardware accelerators for neural networks
- D5.1 Constraints and opportunities definition
- D6.1 Modular architecture template definition
- D5.2 Compiler prototype
- D6.2 Description of the gen1 performance analysis and DSE framework
- D6.3 Description SoC architecture, and the rapid design & prototyping environment

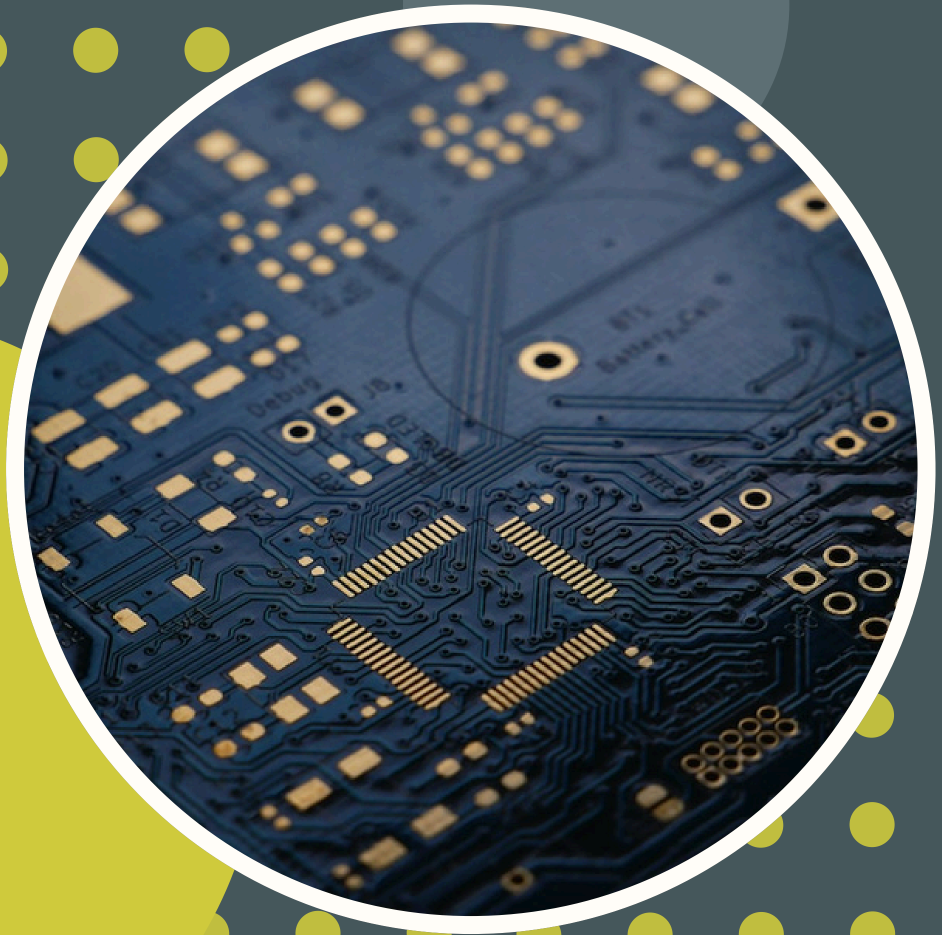
CONVOLVE

Who we are

The CONVOLVE consortium consists of 18 partners from academia and industry with strong complementary competencies in different levels of design hierarchy.

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
ETH zürich
ThalesAlenia Space a Thales / Leonardo company
FMI Friedrich Miescher Institute for Biomedical Research
TU Delft
KU LEUVEN
BOSCH
NXP
Jabra GN
THE UNIVERSITY of EDINBURGH
RUHR UNIVERSITÄT BOCHUM RUB
MANCHESTER 1824
UNIVERSIDAD DE MURCIA
UIR Université Internationale de Rabat THE INNOVATIVE UNIVERSITY
ViNotion Making surveillance smart
Cognitive IC Feel • Think • Act
AXELERA ARTIFICIAL INTELLIGENCE
ΕΠΙΣΕΥ ICCS
CLAIRE

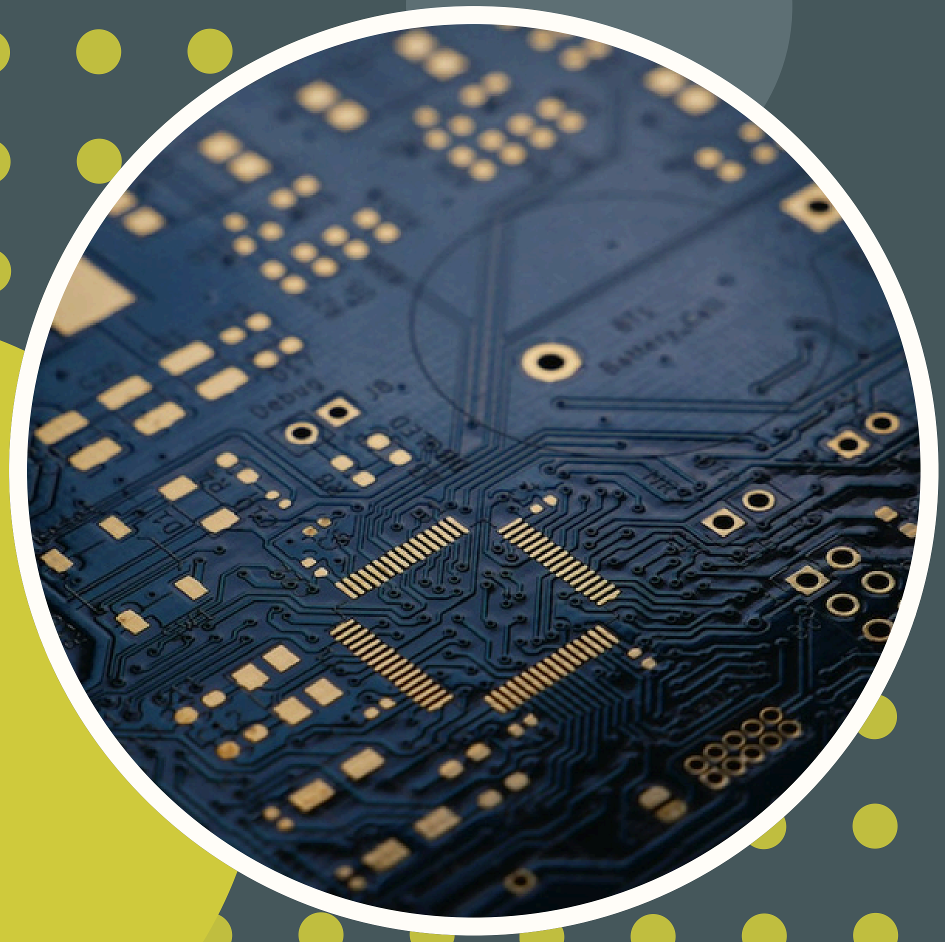
CONVOLVE



For more information »

www.convolve.eu

CONVOLVE



For more information »

www.convolve.eu