RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE INTELLIGENCE ARTIFICIELLE

HOLIGRAIL

HOLIstic approaches to GReener model Architectures for Inference and Learning

Olivier Sentieys (Université de Rennes, Inria, IRISA)
Olivier Bichler (CEA List/LIAE, Saclay)
Mohamed Tamaazousti (CEA List/LVML, Saclay)
Florent De Dinechin (INSA Lyon, CITI)
Fabrice Rastello (Inria Grenoble)
Adrien Prost-Boucle (Grenoble-INP, CNRS)

anr® agence nationale de la recherche

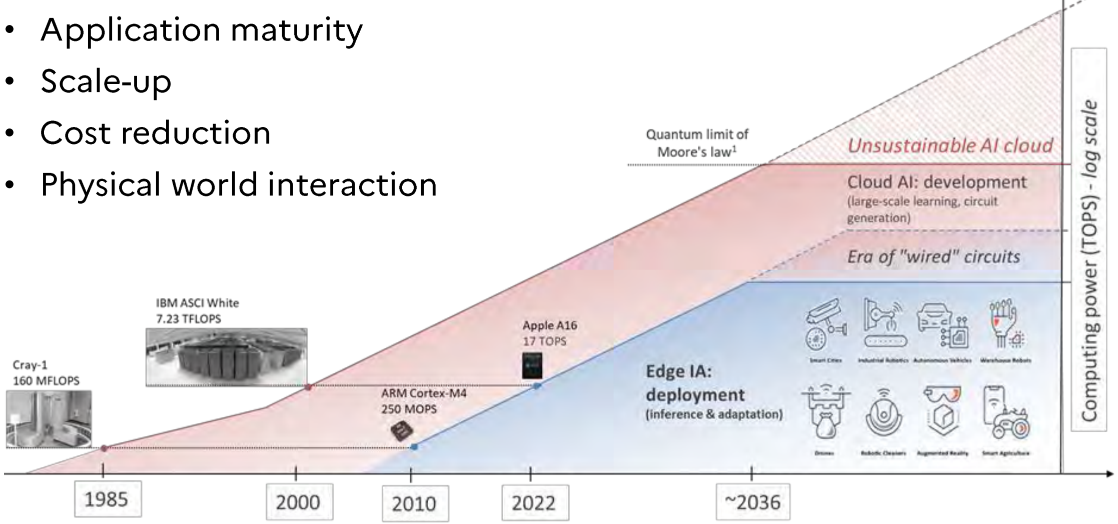
RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

FRANCE 2030 — PROGRAMME DE RECHERCHE INTELLIGENCE ARTIFICIELLE

Motivation

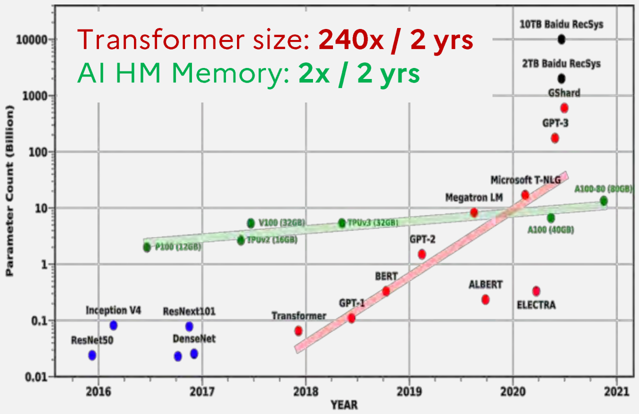
Deployment, from cloud to edge, a necessity:

- Application maturity
- Scale-up
- Cost reduction
- Physical world interaction



*AI's "Moore law" is **two order of magnitude** faster than silicon's*

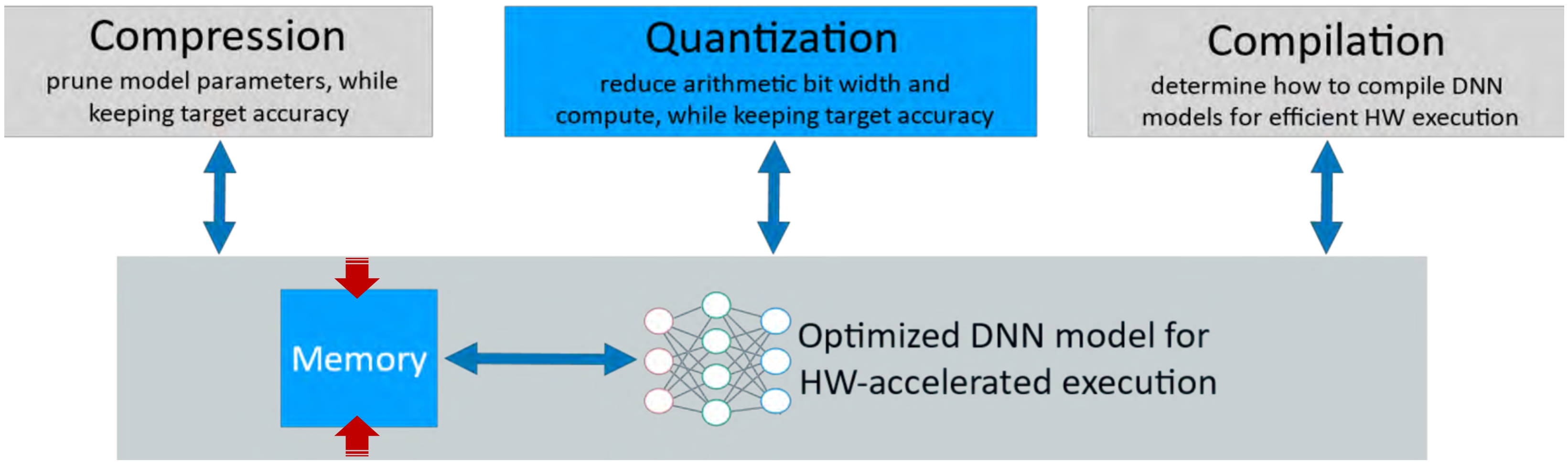
*Evolution of the number of parameters is **much higher** than available on-chip memory*



➔ Holigrail targets breakthroughs in algorithms **coding compactness**, arithmetic **operators efficiency** and related **compiler optimizations** for both training and inference

Need for DNN Compression

From Sensors to Clouds, for both Inference and Training

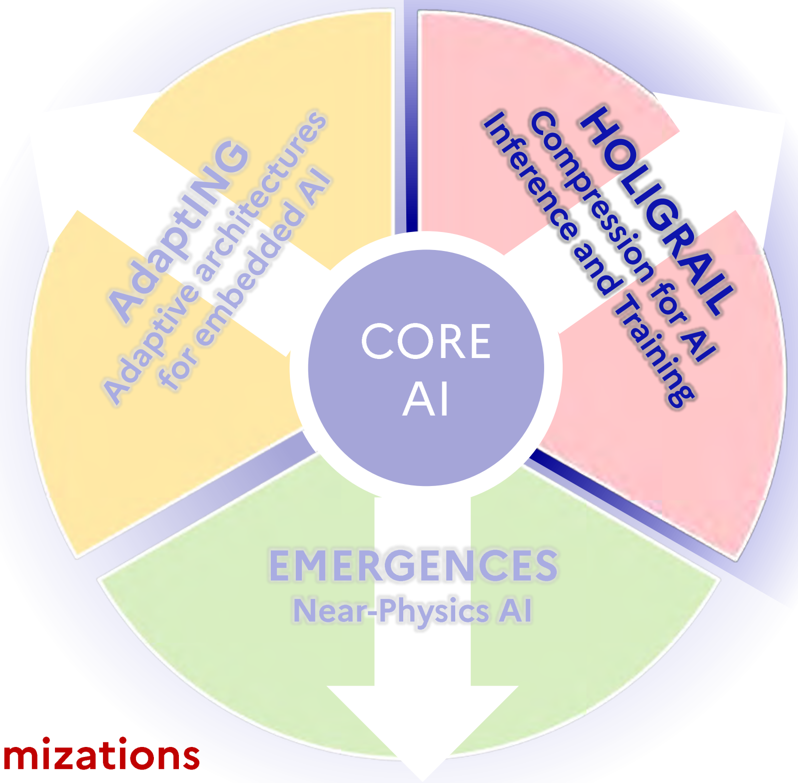


*Don't spread **memory**, compact data!*

From Cloud to Edge

Research Challenges

- Extreme quantization
- Structured sparsity
- Maximum entropy coding
- Tensor methods
- Low-precision training
- Compiler and architectural support



All are related to Hardware-Aware Optimizations

Challenges: Quantization

Quantization effects: the good

- Reduced memory usage, reduced energy, faster execution
- Less silicon area, more parallelism and performance

Quantization effects: the bad

- Less precision results in lower accuracy

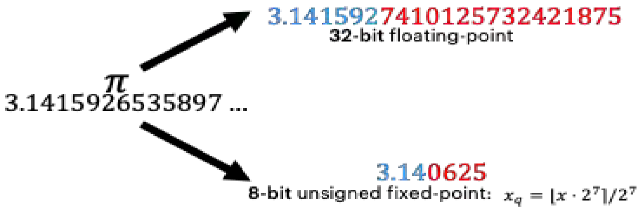
ADD energy (pJ)			
INT8	INT32	FP16	FP32
0.03	0.1	0.4	0.9
30x energy reduction			

MULT energy (pJ)			
INT8	INT32	FP16	FP32
0.2	3.1	1.1	3.7
18.5x energy reduction			

Memory access energy (pJ)	
Cache (64-bit)	
8KB	10
32KB	20
1MB	100
External Mem.	
DRAM	1300-2600

Need for new methods for extreme quantization

- Below 8 bits, ternary/binary
- Quantize weights and activations
- Non-standard number representation formats
- Complex DNN models (e.g., transformers, LLM)



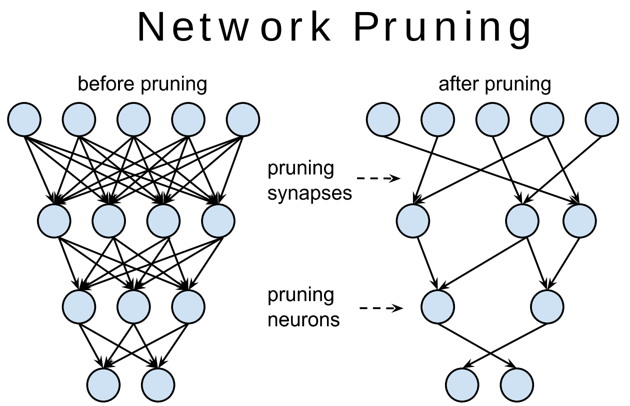
Challenges: Sparsity

Sparsity is intrinsically present in DL models

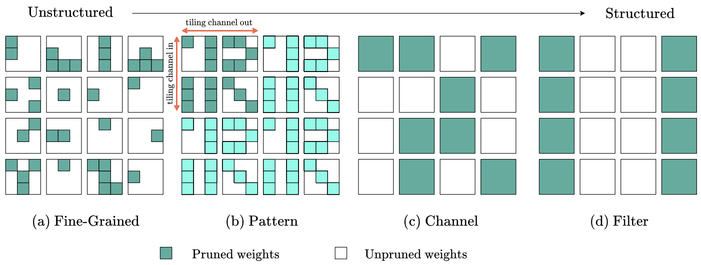
But naturally unstructured and thus difficult to exploit by data-parallel hardware

Propose methods for structured sparsity

- More direct and efficient parallel implementations
- Combined with extreme quantization
- Explore learned data-dependent structured sparsity
- Explore automaton schemes



Structured pruning provides higher efficiency



Challenges: Maximum entropy coding

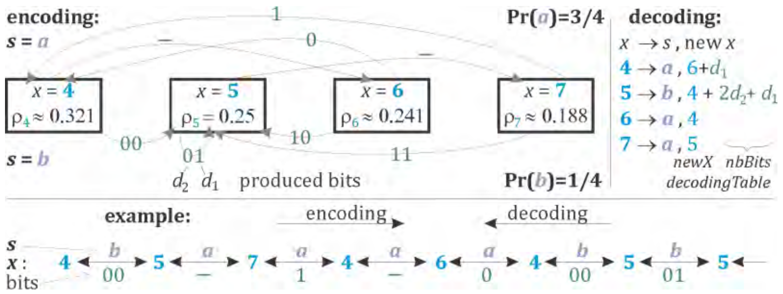
In information theory, entropy measure the quantity of information, in fractional bits

A coding that maximizes entropy per bit should be efficient (no useless bit)

Inversely, minimizing entropy regardless of coding allows efficient compression!

Propose training objectives that enable entropy-based compression:

- Minimize the entropy, not the bit-width
- Exploit arithmetic compression
- Trade memory for (little) computation

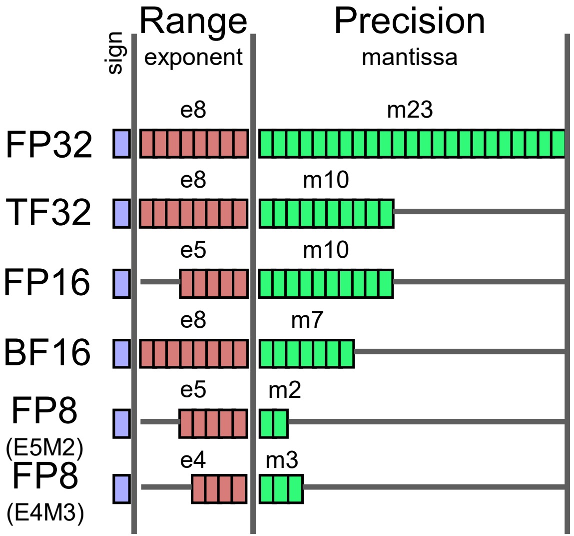


Challenges: Low-precision training

Carbon footprint of DNN training

> *Analyzing the carbon footprint of current natural-language processing models shows an alarming trend: **training one huge model for machine translation emits the same amount of CO2 as five cars in their lifetimes (fuel included)*** [Strubell *et al.*, ACL 2019]

- Many more operations than inference, more pressure on memory access, much more difficult to accelerate



FP8 (8-bit floating-point) currently under standardization by IEEE P3109 working group

Need for significant reduction of training carbon footprint

- Mixed-precision, run-time adaptation, analytical models
- Low-precision floating-point and variable-precision variants
- New models or training algorithms

Challenges: Compiler and architectural support

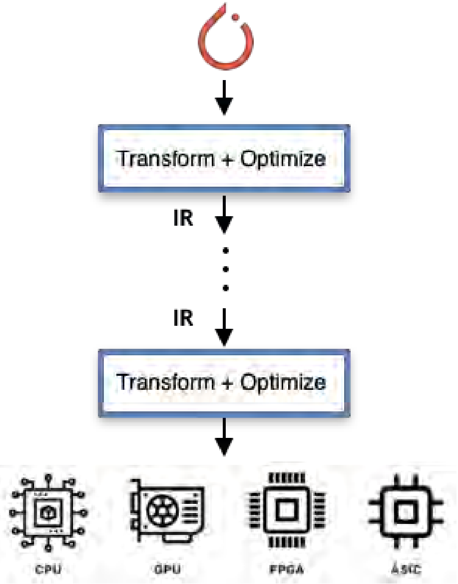
Compiling and generating code from a DNN framework is still an unsolved problem

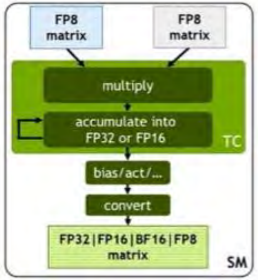
Current runtime and compiler rely on:

- Optimized DNN specific libraries
- Pattern specific (for tensors) compiler optimization

Need for new compiler techniques and frameworks

- Reduce amount of computation and memory footprint
- Exploit current context (sparsity, extreme quantization)
- Revisiting pattern-specific optimization strategies
- Analytical and statistical performance models (high-level)
- Co-design compiler / optimization / hardware





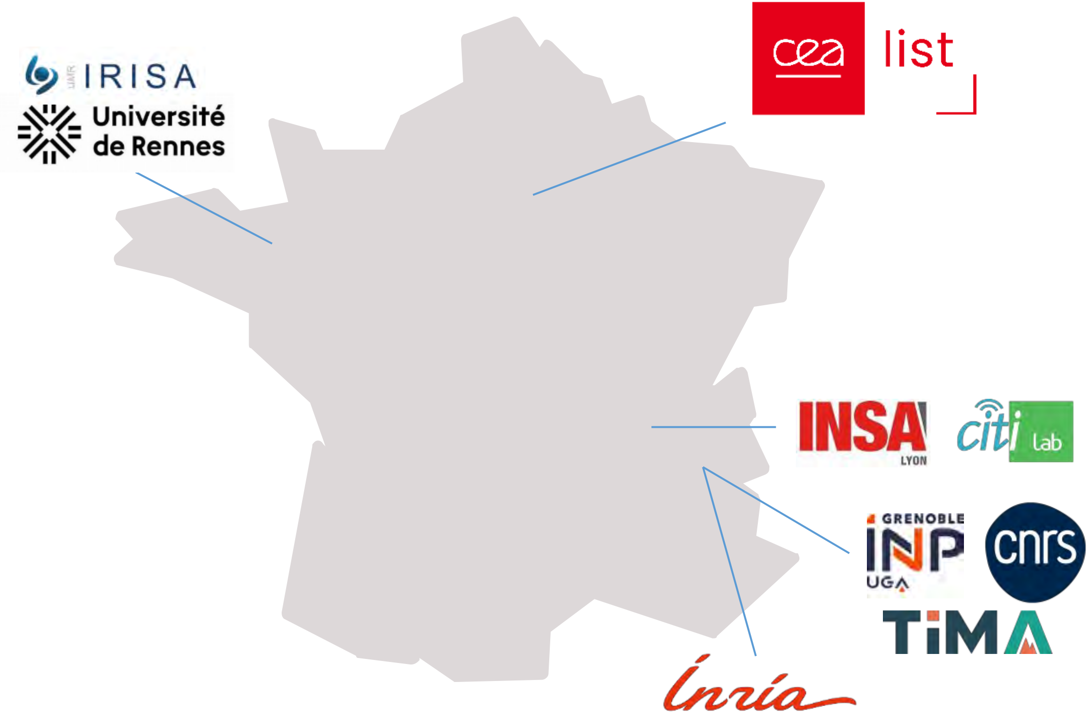
Nvidia Hopper GH100 GPU

FP8 support in tensor cores provides up to 4x speedup

HOLIGRAIL Partners

Complementary skills in

- machine learning, optimization,
- computer arithmetic,
- hardware architecture,
- hardware acceleration,
- compiler optimizations,
- coding theory



Consortium members

Université de Rennes, Inria, IRISA

Project coordinator

Olivier Sentieys, Silviu Filip

Previous works in the field

- Computer arithmetic and architecture for embedded systems
- Energy-efficient hardware accelerators, for machine learning and data mining
- Approximate computing: reduced-precision arithmetic, numerical accuracy analysis
- Low-precision training
- Quantization-aware training

Commissariat à l'énergie atomique et aux énergies alternatives (CEA)

Olivier Bichler, Mohamed Tamaazousti

Previous works in the field

- Architectures based on resistive memory technology and neuromorphic computing
- Development of the **N2D2** deep learning quantization and deployment framework
- Computer vision applications
- Deep neural networks representation and compression
- Random tensor theories for machine learning models

Consortium members

Université Grenoble Alpes, Inria Corse

Fabrice Rastello, *Christophe Guillon*

Previous works in the field

- Automatic parallelization and compiler back-end optimization
- Pattern-specific compiler optimization for hardware accelerators
- Performance debugging based on binary instrumentation
- Compiler design

Université Grenoble Alpes, Grenoble-INP, CNRS, TIMA

Adrien Prost-Boucle, *Olivier Muller, Frédéric Pétrot*

Previous works in the field

- High-level synthesis
- Optimized implementations for FPGA
- Digital circuit accelerators for AI inference
- Focus on resources and energy usage

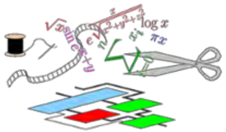
Consortium members

INSA Lyon, CITI Lab, Inria

<u>Florent De Dinechin</u>, Anastasia Volkova

Previous works in the field

- Hardware and software computer arithmetic
- FPGA arithmetic and computing
- Formal proof for arithmetic algorithms
- Low-level implementations of artificial neural networks
- **FloPoCo** arithmetic cores generator software project

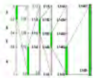


WP0: Management, Dissemination, and Valorization

WP1: Model Compression and Optimization

Compression framework at the algorithm level: unified formulation including pruning, quantization and decomposition:

- Tensor decomposition
- Quantization-aware training
- Mixed-precision quantization
- Maximum entropy coding



WP2: Optimizing Sparse Computation and Quantization

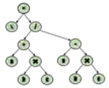
Enhance efficiency and speed of sparse computations in deep neural networks, taking into account quantization:

- Generation of optimized codelets
- Fast heuristics for compressibility
- Quantization-aware sparsity

WP3: Arithmetic Operators and Code Generation

Optimize neural networks at the micro-architecture level, down to compiler back-end techniques:

- Small formats arithmetic units
- Memory access optimizations
- Resource-constrained execution



From algorithm optimization to hardware mapping

WP4: Tool Development and Validation

Integration and interoperability with existing reference frameworks. Develop hardware librairies, prototypes or demonstrators to showcase HOLIGRAIL results, fully synchronized with the outcomes of ADAPTING and DEEPGREEN.

WP0: Management, Dissemination, and Valorization

WP1: Model Compression and Optimization

PhD	cea list	(+ IJCLab)
PhD	cea list	
PostDoc	cea list	
PostDoc	INSA	
PhD	Université de Rennes	
PhD	cea list	+ Université de Rennes
PhD	cea list	+ inria

WP2: Optimizing Sparse Computation and Quantization

PostDoc	Université de Rennes	+ inria
PhD	Université de Rennes	
PhD	inria	+ Université de Rennes
PhD	inria	+ TIMA
PostDoc	inria	· Université de Rennes
PhD	TIMA	+ Université de Rennes

WP3: Arithmetic Operators and Code Generation

PhD	TIMA	+ inria
PhD	TIMA	
PhD	Université de Rennes	+ INSA
PhD	INSA	+ Université de Rennes
PhD	INSA	+ TIMA
PostDoc	cea list	+ inria
PostDoc	INSA	+ TIMA

From algorithm optimization to hardware mapping

WP4: Tool Development and Validation

5 research engineers and 3 PostDoc, partially shared with WP1, WP2 and WP3

Project expected outcomes

- Breakthroughs in efficiency for inference and training algorithms on specialized hardware:
  - Algorithms coding compactness
  - Arithmetic operators efficiency
  - Compiler optimizations
- Dissemination in high-quality publications in journals and conferences with high impact
- Open-source software (e.g., MPTorch, FloPoCo, MLIR/LLVM) and hardware specifications
- Integration in embedded system solutions, in particular the DeepGreen framework (aidge)

Manpower

- 14 PhD students, 7 post-doctoral fellows (16.5 p.y), 4 research eng. (10.5 p.y)

p.y: person year

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030

PROGRAMME DE RECHERCHE

INTELLIGENCE ARTIFICIELLE

Retrouvez toutes nos actualités

anr agence nationale de la recherche

Relation with other projects from the PEPR