REASON: Accelerating Probabilistic Logical Reasoning for Neuro-Symbolic Intelligence






Zishen Wan, Che-Kai Liu, Jiayi Qian, Hanchen Yang,
Arijit Raychowdhury, Tushar Krishna

Georgia Institute of Technology, Atlanta, GA

Email: zishenwan@gatech.edu

Cognitive Reasoning in Real World

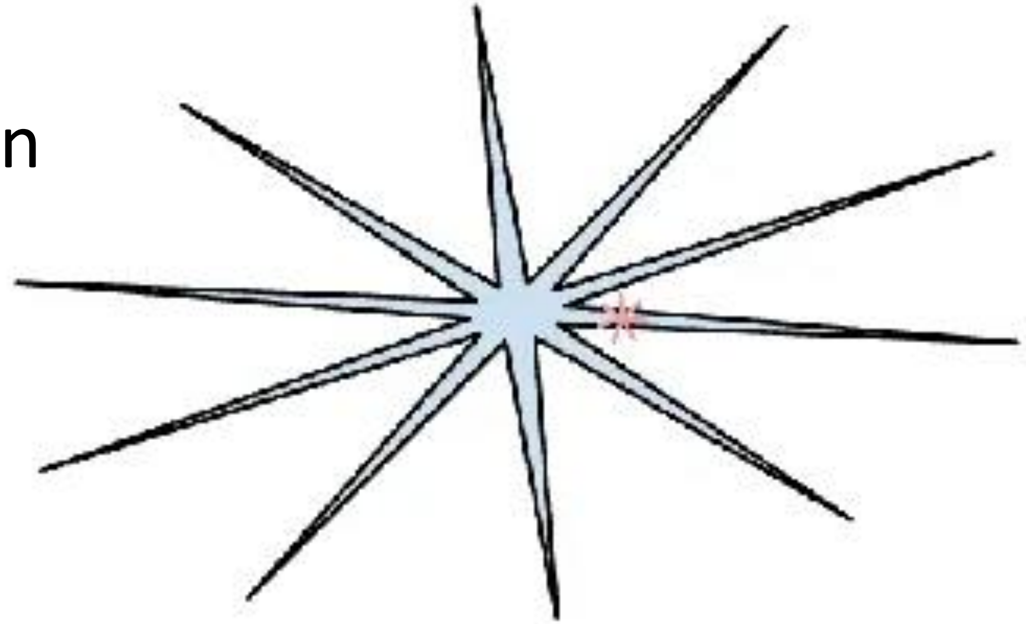
- Cognitive reasoning pioneers the next frontier of physical intelligence

(From Figure AI, Jiayuan Mao)

Achieving human-level cognition requires both intuitive thinking and deliberative reasoning

Is a Monolithic LLM Enough?

- ✅ Pattern recognition & language modeling
- ✅ Data-driven learning
- ❌ Reliable Multi-step logical deduction
- ❌ Uncertainty-aware reasoning
- ❌ Verifiability and robustness



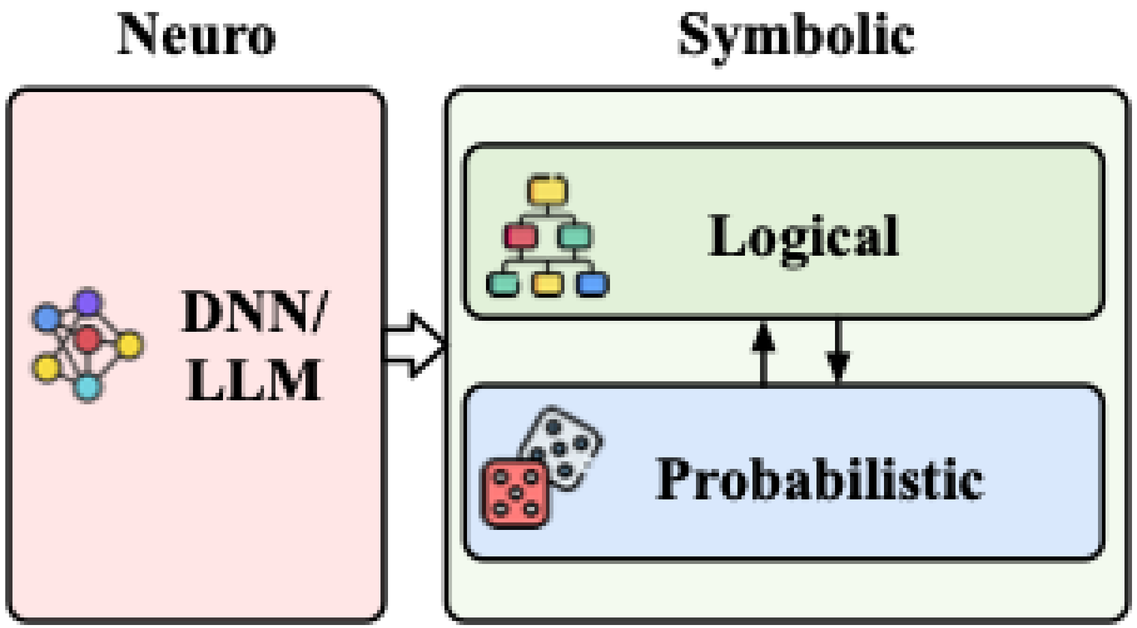
Jagged performance of monolithic LLM

Andrej Karpathy, Eureka Labs & OpenAI, 2025 LLM Year in Review

Key issue: only scaling parameters substitutes compute for structure

AI Systems Are Becoming Compositional

Perception & representation learning



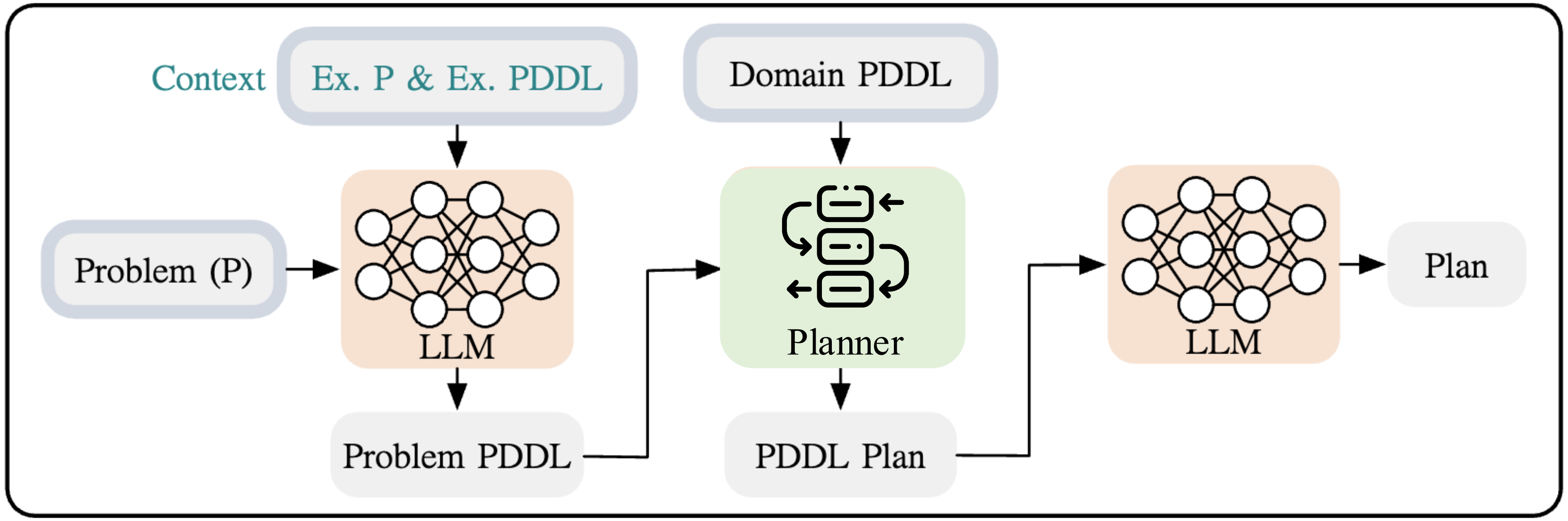
Explicit logic, rules, and constraints that are interpretable and verifiable

Model uncertainty and belief updates, reason under ambiguity

LLMs are excellent at intuition — but verification and uncertainty require explicit components

Neuro-Symbolic Example: Trustworthy Planning

Grasp bottle | Free gripper | Grasp can | Place can | Re-grasp bottle | Place bottle



- **Task**: long-horizon multi-step planning
- **Modules**:
  - LLM: natural language interpreter, task decomposition, generate program
  - PDDL & Symbolic solver: planning domain definition language-based planner to ensure feasibility & correctness

Liu et al, “LLM+P: Empowering Large Language Models with Optimal Planning Proficiency”, 2025

Neuro-Symbolic Example: Scientific Reasoning

Google DeepMind

AlphaGeometry: An Olympiad-level AI system for geometry

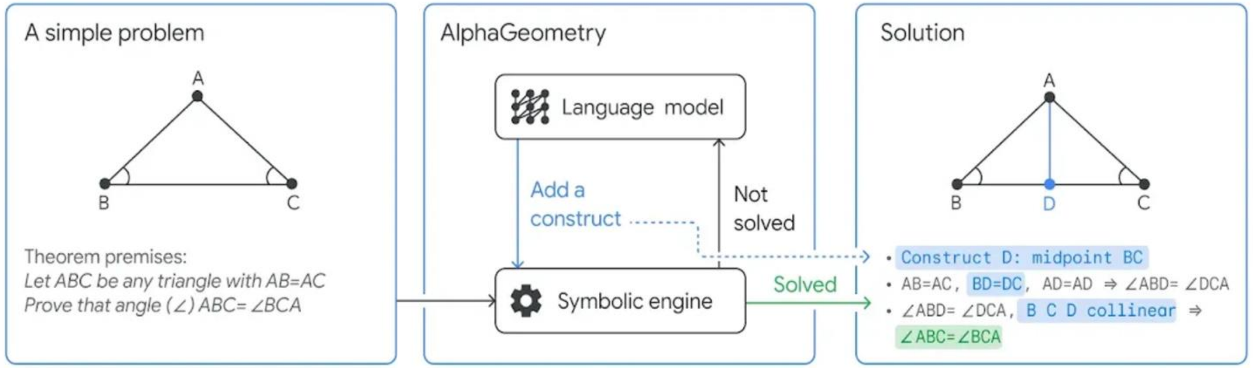
17 JANUARY 2024

Trieu Trinh and Thang Luong

Share

AlphaGeometry adopts a neuro-symbolic approach

AlphaGeometry is a neuro-symbolic system made up of a neural language model and a symbolic deduction engine, which work together to find proofs for complex geometry theorems. Akin to the idea of "thinking, fast and slow", one system provides fast, "intuitive" ideas, and the other, more deliberate, rational decision-making.



- **Task**: Olympiad-level geometry solving
- **Modules**:
  - LLM: construct auxiliary points and lines
  - Symbolic: algebraic deductive reasoning

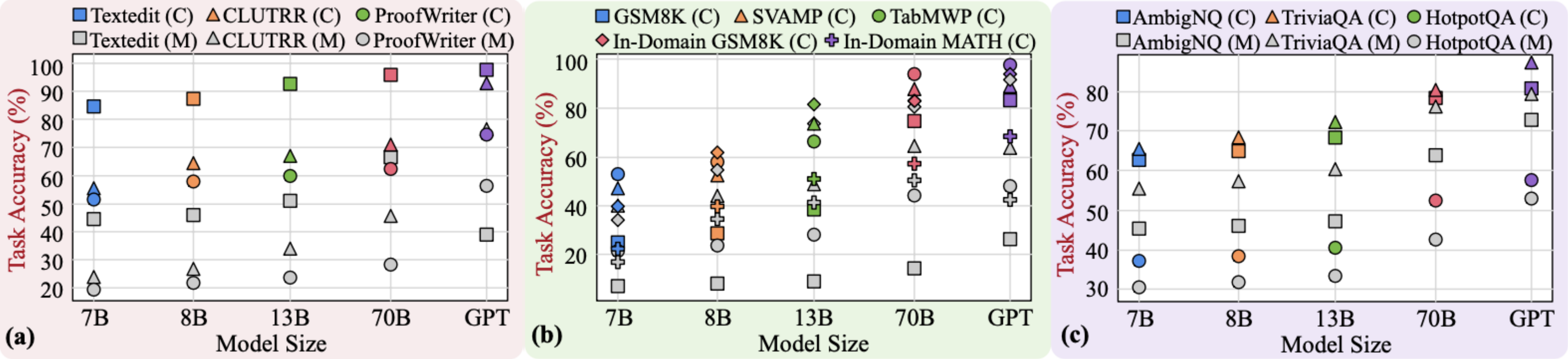
Trinh et al, "Solving Olympiad Geometry without Human Demonstrations", Nature 2024

Compositional AI Scales Better

Complex Reasoning Tasks

Math Reasoning Tasks

Question-Answering Tasks

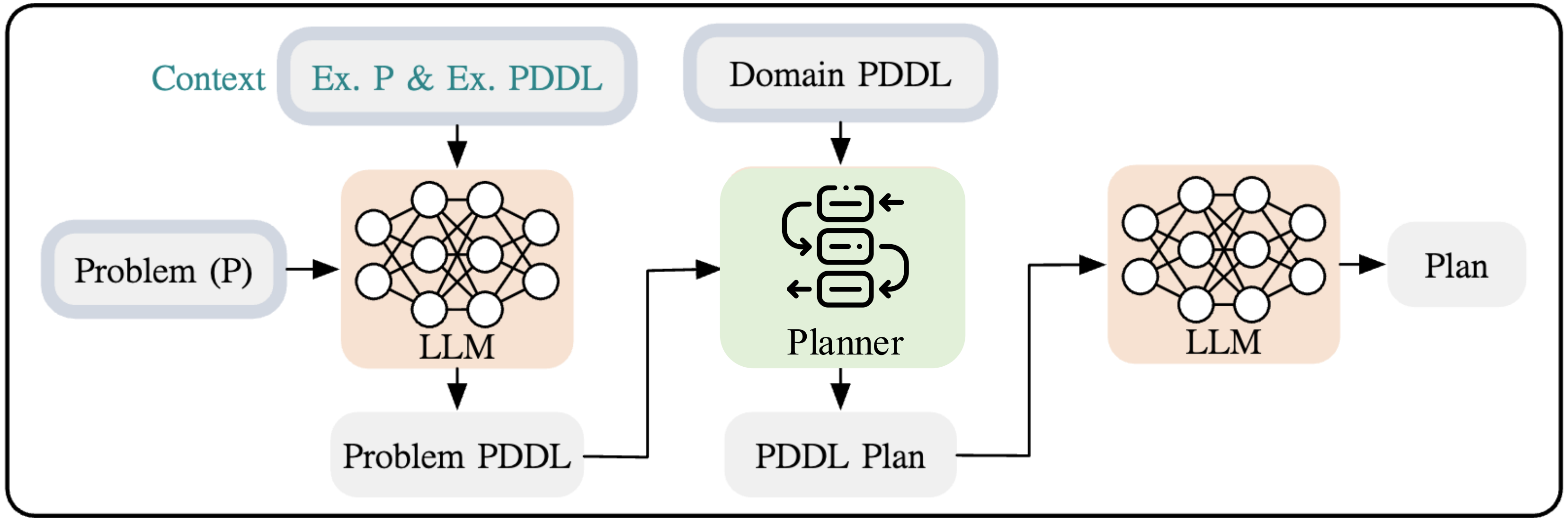


- Neuro-symbolic systems outperform monolithic LLMs
- Smaller models achieve comparable or higher accuracy

Neuro-symbolic scale *algorithmically* better than monolithic LLMs

But Systems Hit a Wall

Grasp bottle | Free gripper | Grasp can | Place can | Re-grasp bottle | Place bottle



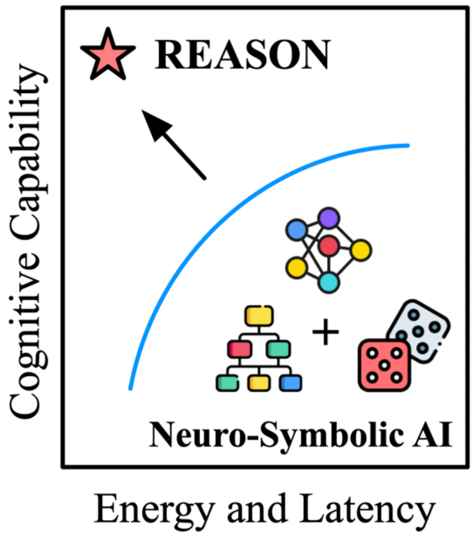
- ❑ **Task**: long-horizon multi-step planning
- ❑ **Modules**:
  - ❑ LLM: natural language interpreter, task decomposition, generate program

Symbolic components take **>100 seconds** on CPU-GPU per task

Symbolic reasoning is slow, latency explodes with task complexity

REASON Enables Efficient Neuro-Symbolic Cognition

Goal:

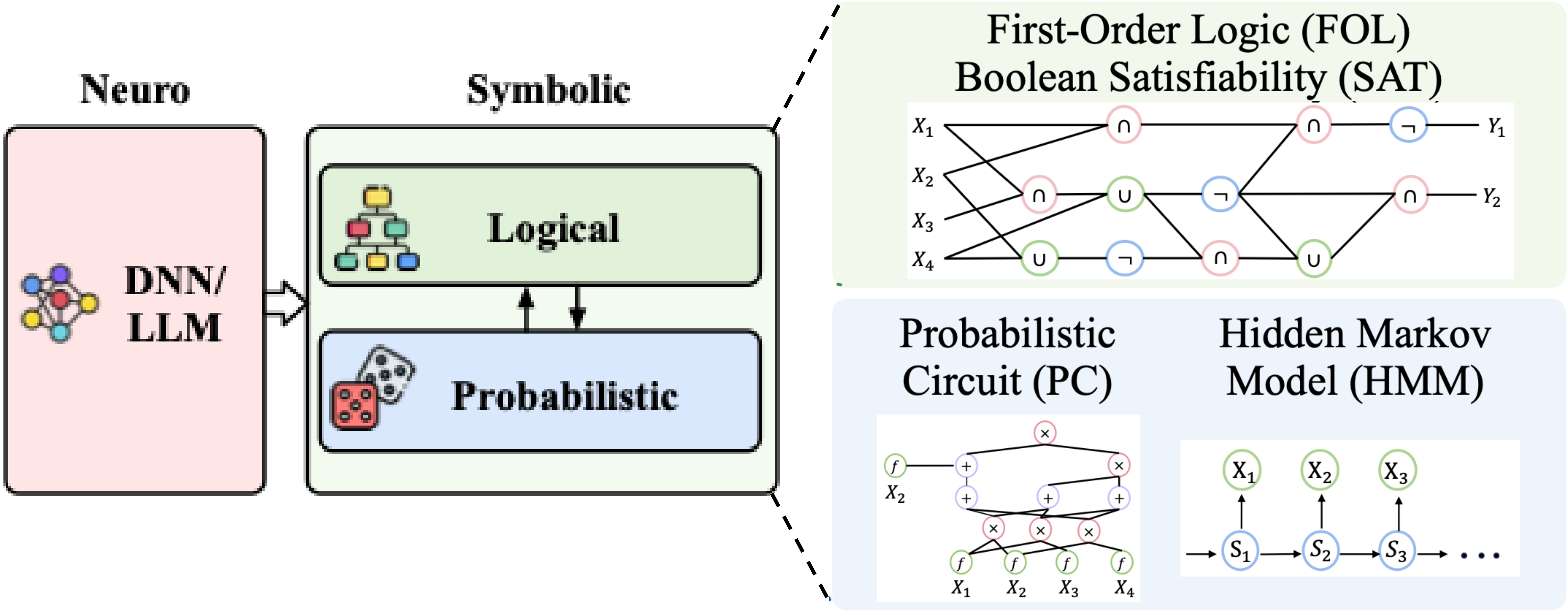


Challenge 1: Heterogeneous & diverse operators

Challenge 2: Inefficient processing, Low hardware utilization

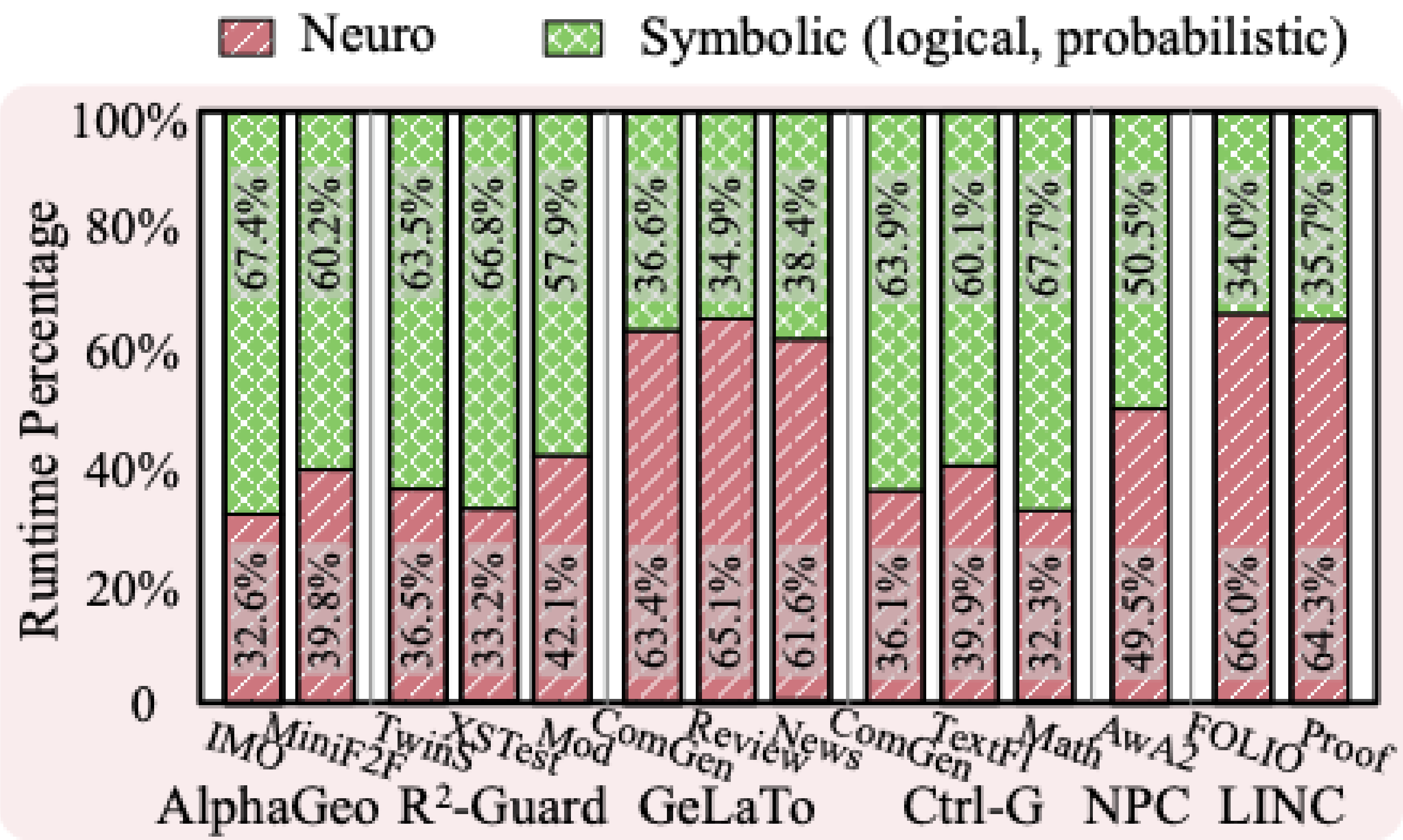
Challenge 3: Complex control flow

Challenge 1: Diverse Symbolic Operators



Core computational primitives: logical reasoning, probabilistic reasoning, sequential reasoning

Challenge 2: Long Runtime Latency



Neuro:
- LLM

Symbolic:
- First-order Logic
- Boolean Satisfiability (SAT) solver
- Probabilistic Circuit
- Hidden Markov Model

Symbolic (Logical and probabilistic) reasoning accounts for large portion of end-to-end runtime

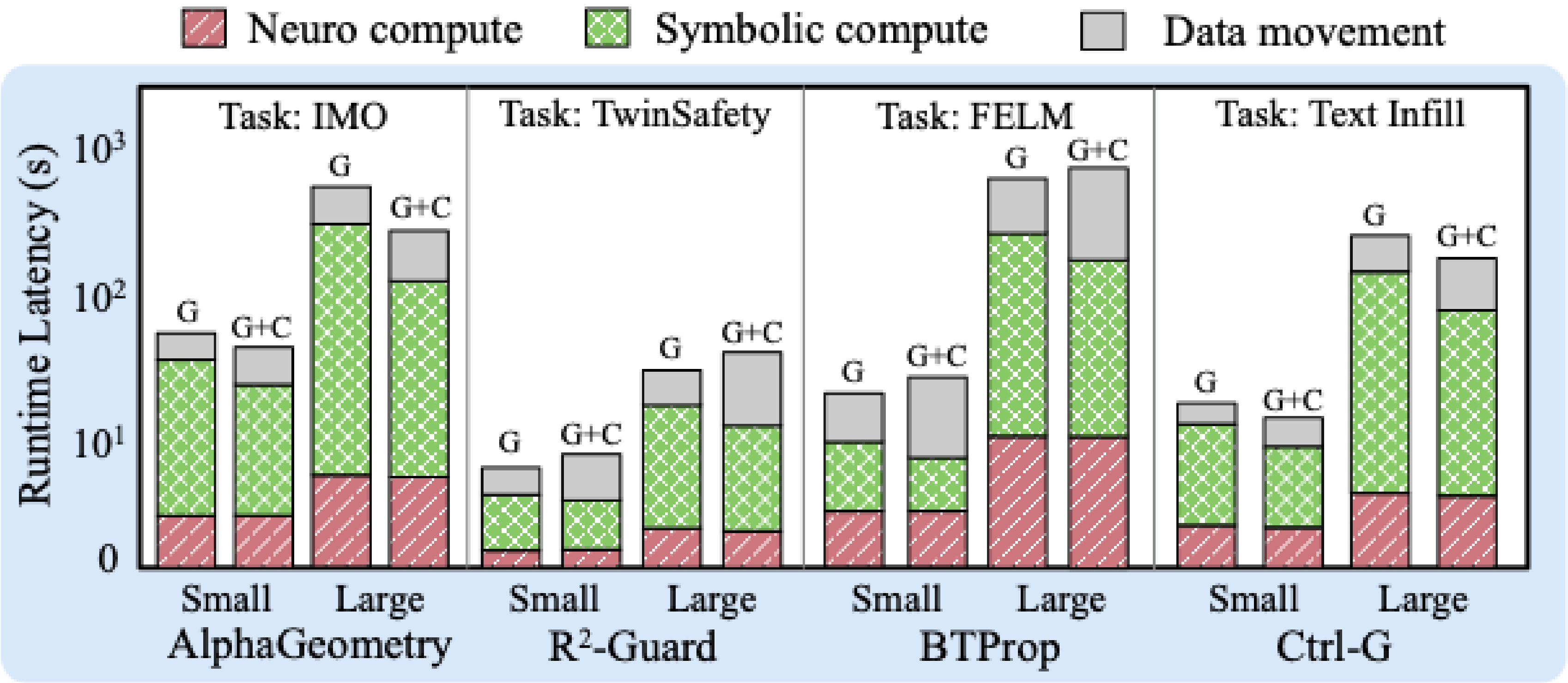
Why GPUs Inefficient for Symbolic Kernels?

	Neural Kernel		Logical Kernel		Probabilistic Kernel	
	MatMul	Softmax	Sparse MatVec	FOL	Marginal	Bayesian
Compute Efficiency						
Compute Throughput (%)	96.8	62.2	32.5	14.7	35.0	31.1
ALU Utilization (%)	98.4	72.0	43.9	29.3	48.5	52.8
Memory Behavior						
L1 Cache Throughput (%)	82.4	58.0	27.1	20.6	32.4	37.1
L2 Cache Throughput (%)	41.7	27.6	18.3	12.4	24.2	27.5
L1 Cache Hit Rate (%)	88.5	85.0	53.6	37.0	42.4	40.7
L2 Cache Hit Rate (%)	73.4	66.7	43.9	32.7	50.2	47.6
DRAM BW Utilization (%)	39.8	28.6	57.4	70.3	60.8	68.0
Control Divergence and Scheduling						
Warp Execution Efficiency (%)	96.3	94.1	48.8	54.0	59.3	50.6
Branch Efficiency (%)	98.0	98.7	60.0	58.1	63.4	66.9
Eligible Warps/Cycle (%)	7.2	7.0	2.4	2.1	2.8	2.5

Symbolic kernels suffer from low ALU utilization, low cache hit rate, high data movement, complex control flow, and low warp and branch efficiency

Even optimized GPU kernels suffer low utilization due to irregular symbolic operations

How About CPU for Symbolic, GPU for Neuro?

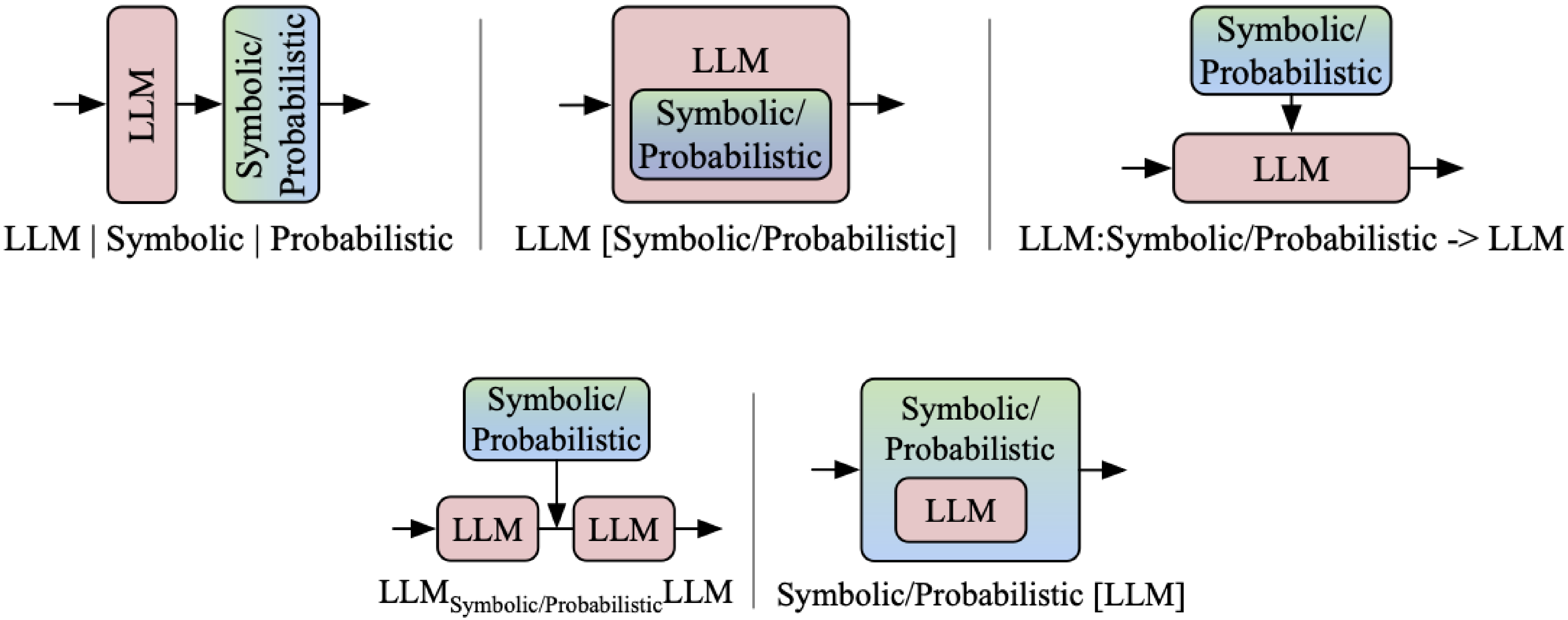


(G: GPU | G+C: GPU+CPU)

- Symbolic remain slow on CPUs: low arithmetic intensity, irregular control flow, poor locality
- Tight neuro-symbolic coupling introduces CPU-GPU communication overhead

CPU+GPU suffer from inefficient symbolic execution and frequent neuro-symbolic interaction

Challenge 3: Complex Control Flow



Tight and diverse coupling between neuro and symbolic components complicates control flow

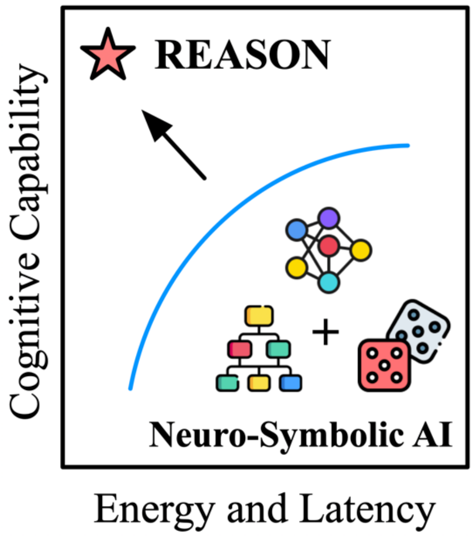
Symbolic Reasoning Is a New Workload Class

	Neuro Inference (LLM/DNN)	Neuro-Symbolic (LLM-Logic/Probabilistic)
Runtime	[Neuro] < [Neuro-Symbolic]	
Compute Kernels	Dense/sparse tensor operations	Heterogeneous kernels (tensor, logic, graph traversal, vector ops)
Arithmetic intensity	High	Low
Data access pattern	Regular, contiguous	Irregular
Control flow	Mostly static	Branch-heavy, dependent, tightly-coupled
Parallelism	Massive data parallelism	Limited, dependency-driven
Data reuse	High	Low
Performance	Throughput-oriented	Latency-critical

Supporting probabilistic logical reasoning efficiently requires new abstraction and architecture

REASON Enables Efficient Neuro-Symbolic Cognition

Goal:



Challenge 1: Heterogeneous & diverse operators

Challenge 2: Inefficient processing, Low hardware utilization

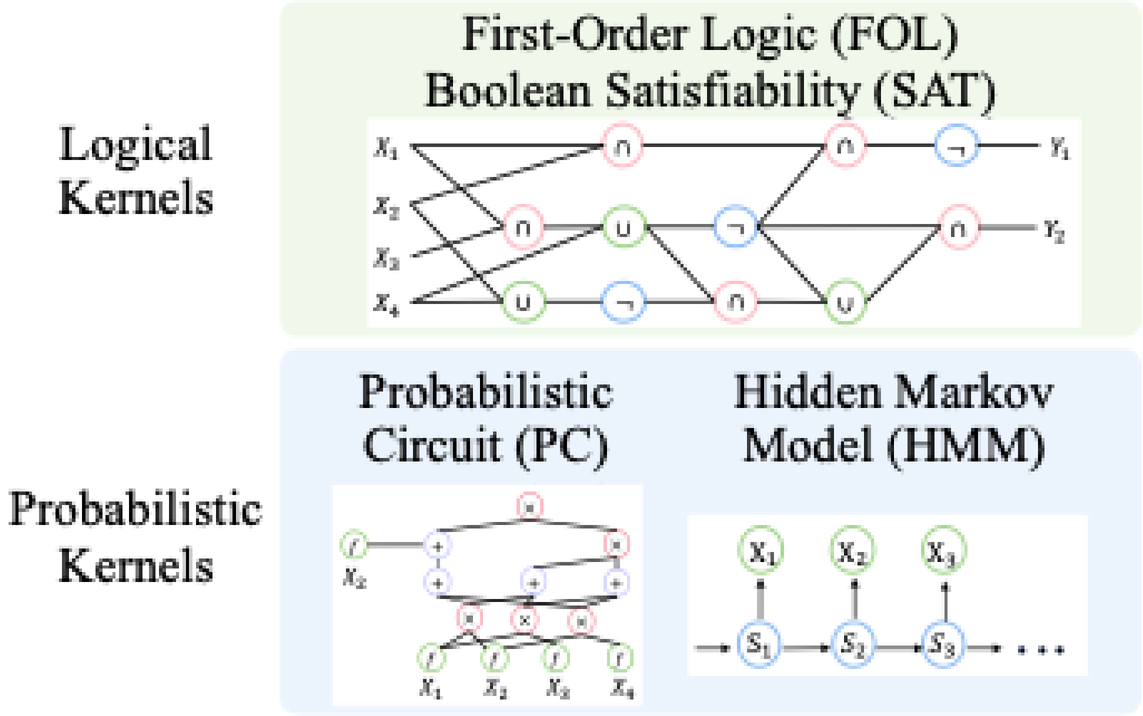
Challenge 3: Complex control flow

Technique 1: **Unified intermediate representation** of diverse symbolic kernels

Technique 2: **Flexible architecture** & **compilation flow** to support diverse symbolic kernels

Technique 3: **Programming model** with GPU **co-processor integration**

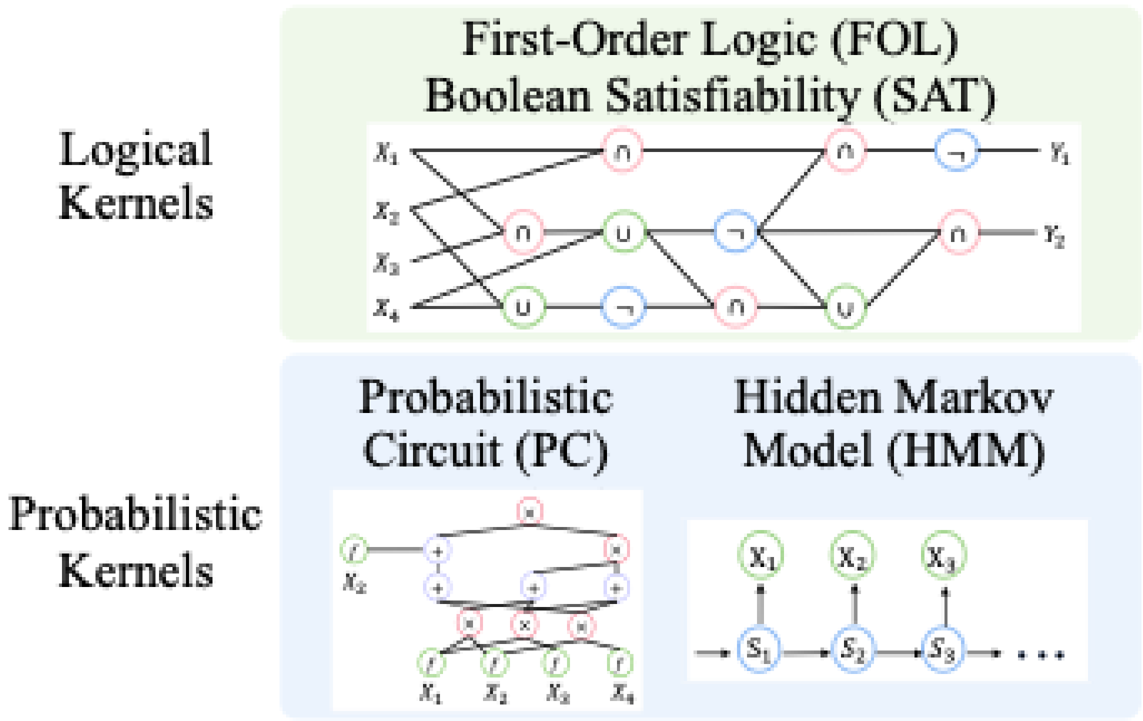
Technique 1: Unified Intermediate Representation



Kernel	DAG Nodes	DAG Edges	DAG Execution
SAT/FOL	Literals and logical ops	Logic dependency between literals, clauses, formulas	Search and deduction via traversal
PC	Primitive distributions, sum and product nodes	Weighted dependency encoding probabilistic factorization	Probability aggregation and flow propagation
HMM	Hidden state variables at each step	State transition and emission dependencies	Sequential message passing

Design a unified DAG structured-based IR for symbolic and probabilistic reasoning

Technique 1: Unified Intermediate Representation



Step 1: Adaptive DAG Pruning

Step 2: Two-Input DAG Regularization

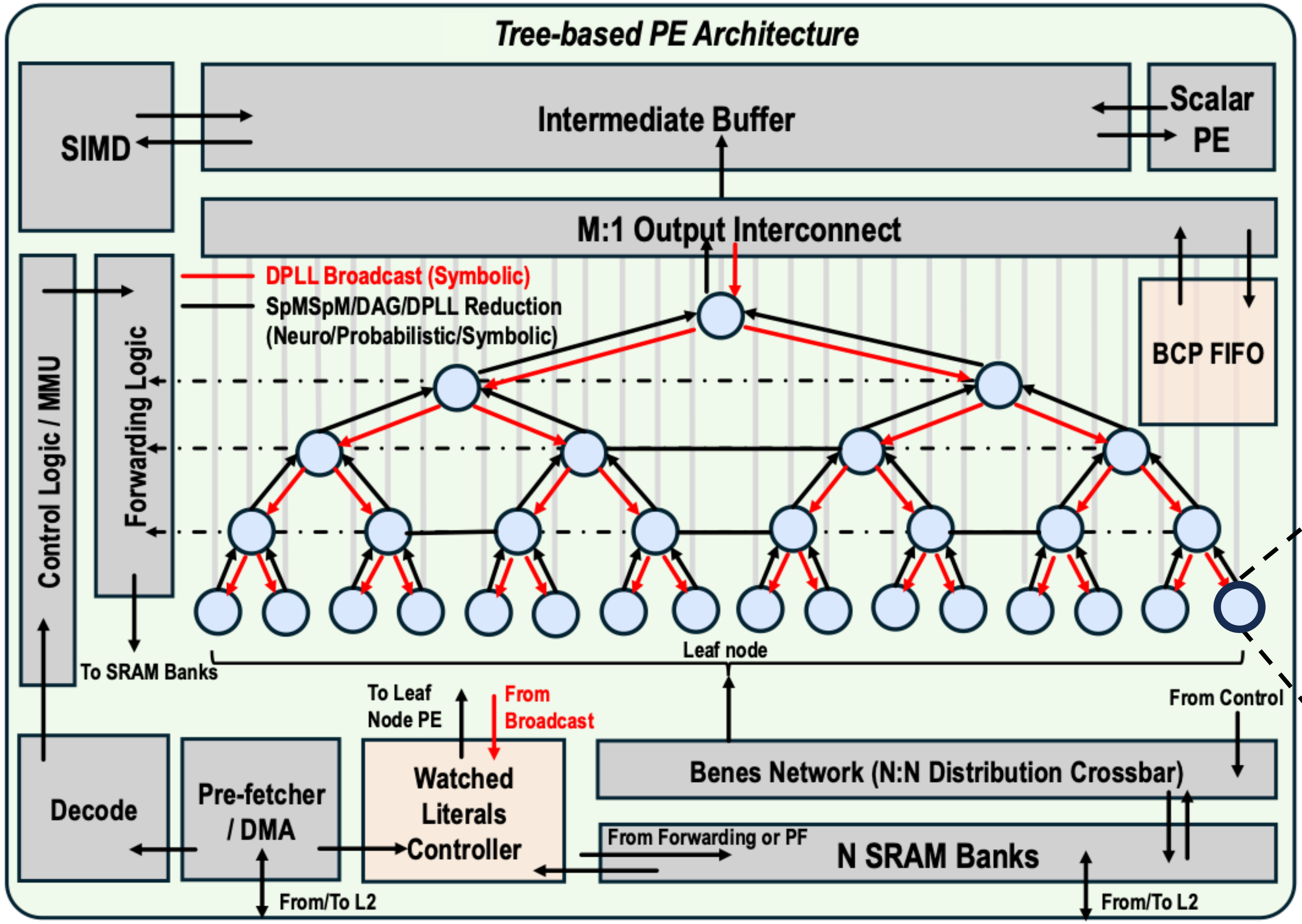
Step 3: DAG Dependency-aware scheduling

Instruction stream

Automated compilation process

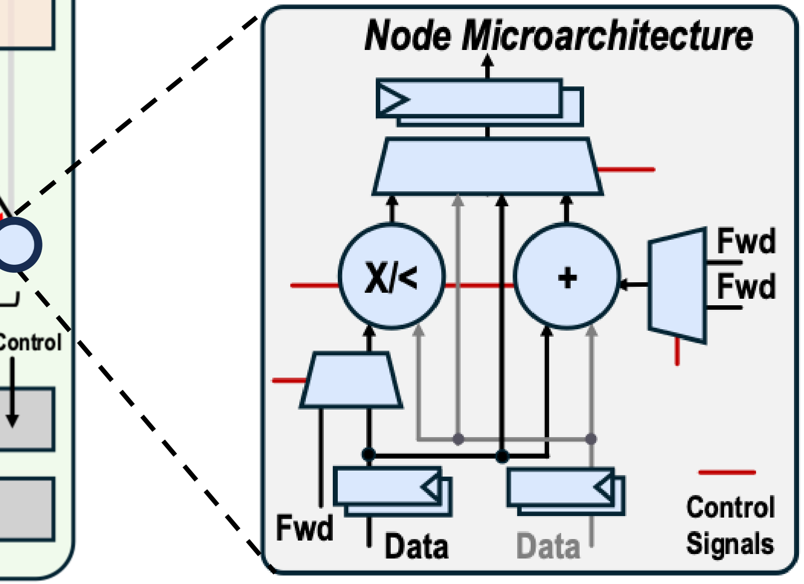
Compiler automates DAG pruning and regularization to generate hardware-friendly structures

Technique 2: Reconfigurable PE Architecture



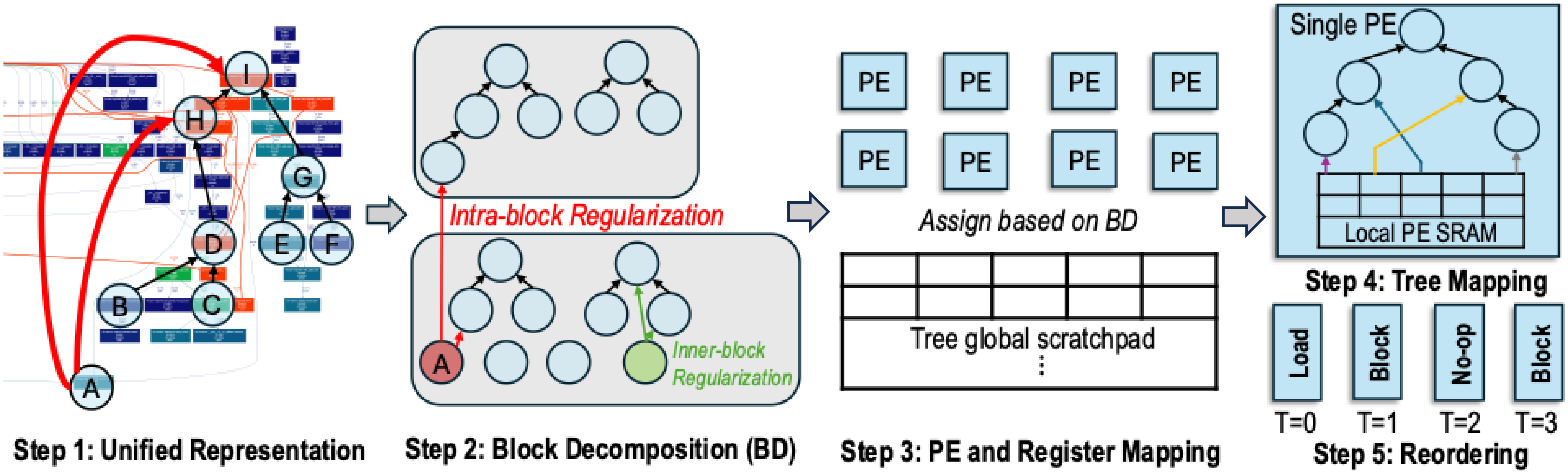
Three operational modes:

- Logical mode
- Probabilistic mode
- SpMSpM mode



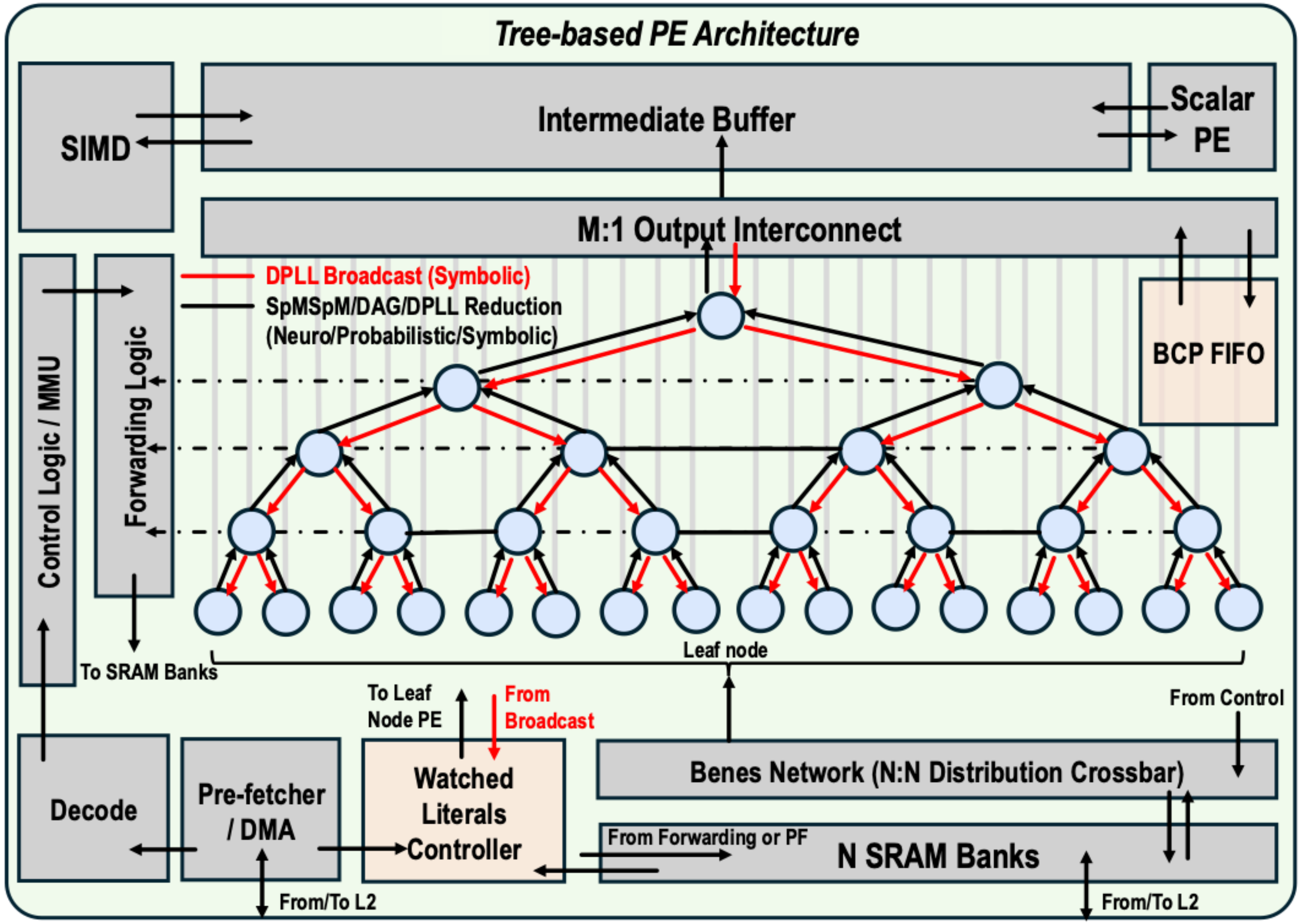
Reconfigurable PEs efficiently support heterogeneous reasoning operators at fine granularity

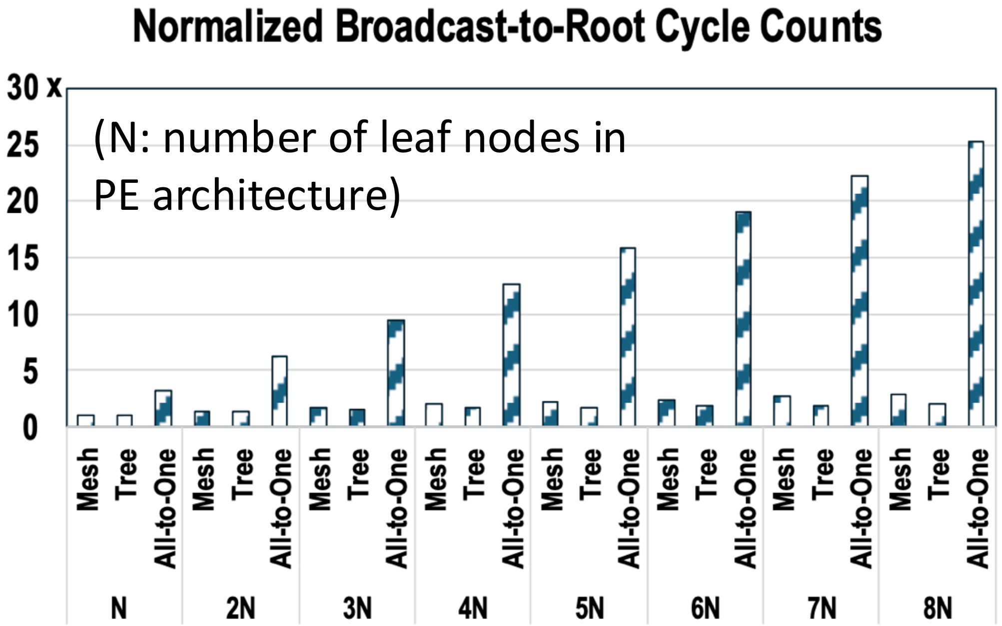
Technique 2: Compiler-Driven Hardware Mapping



Compiler-driven DAG-to-hardware mapping in an automated heuristic process

Technique 2: Scalable Inter-node Topology



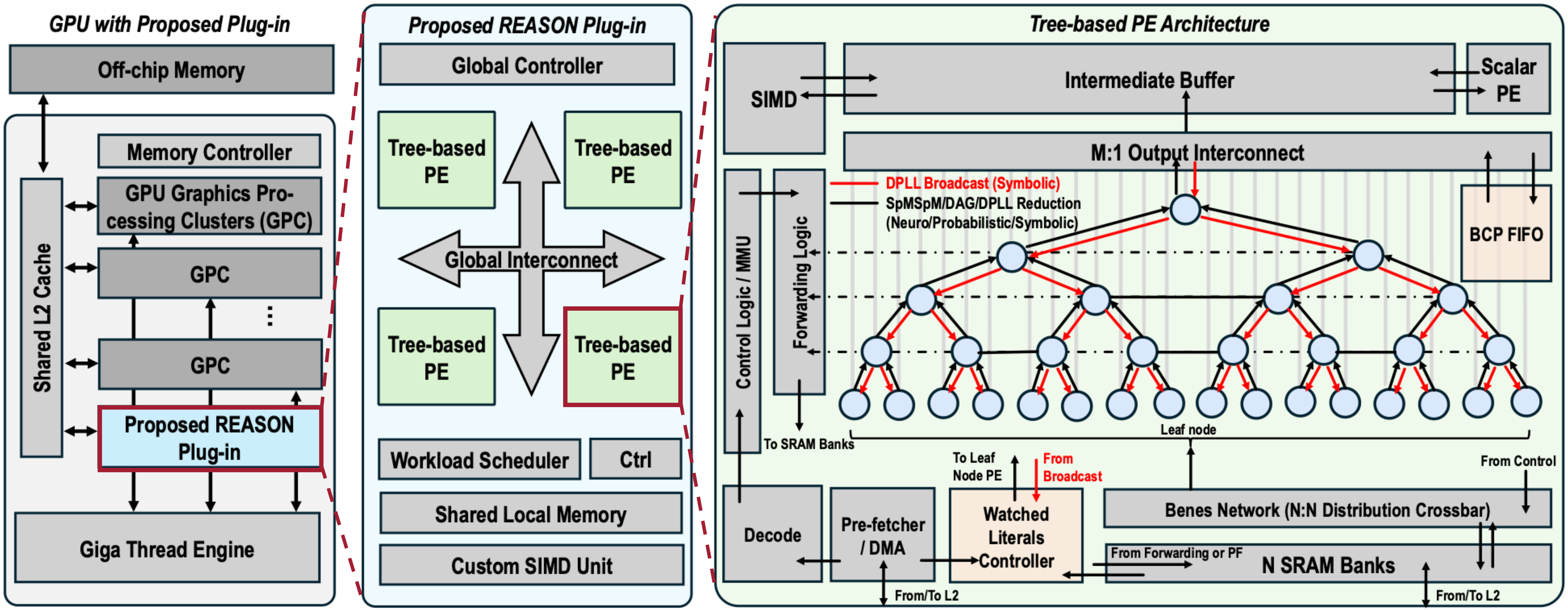


Root-to-leaf traversal latency

- Tree-based: O(log N)
- Mess-based: $O(N^{1/2})$
- Bus-based: O(N)

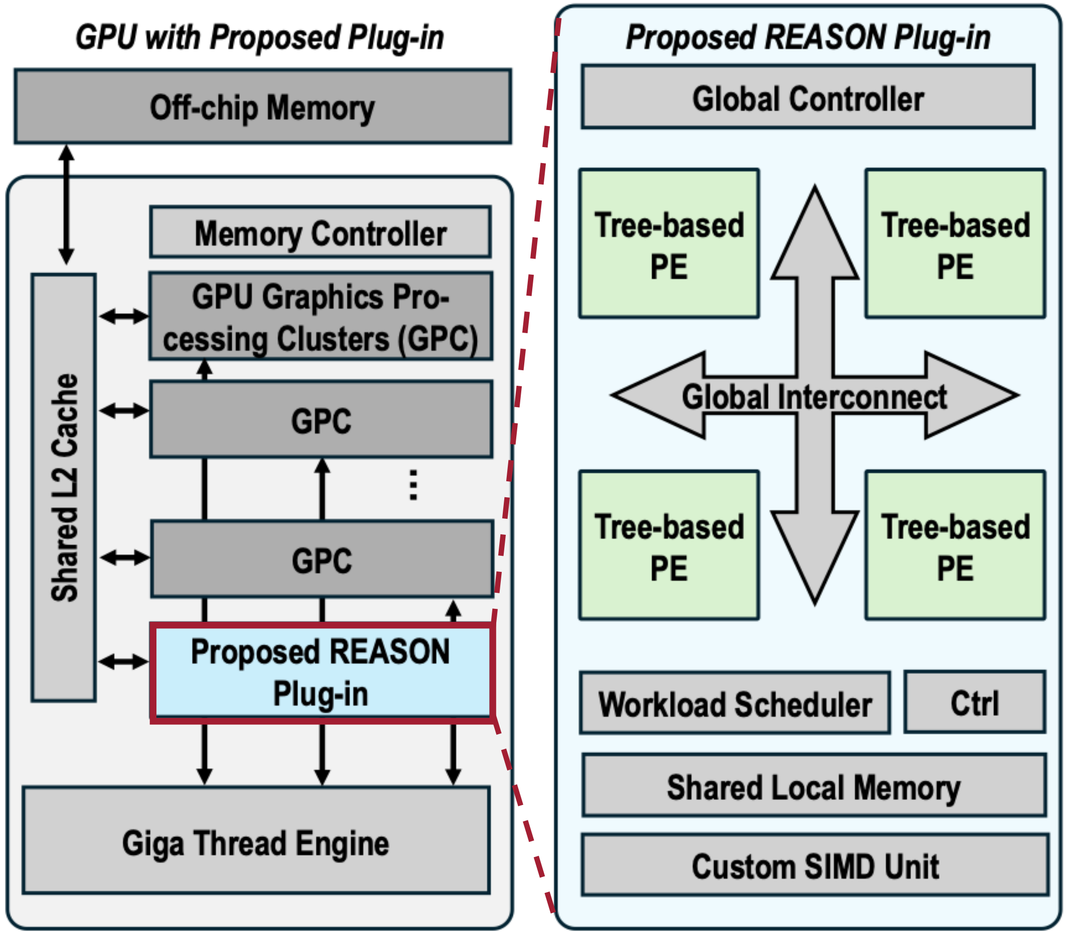
Scalable tree-based inter-node topology can support large-scale logic and probabilistic kernels

Technique 3: GPU Co-Processor Integration



GPU co-processor integration enables efficient and versatile LLM-symbolic processing

Technique 3: GPU Programming Model



Listing 1: C++ Programming Interface of REASON

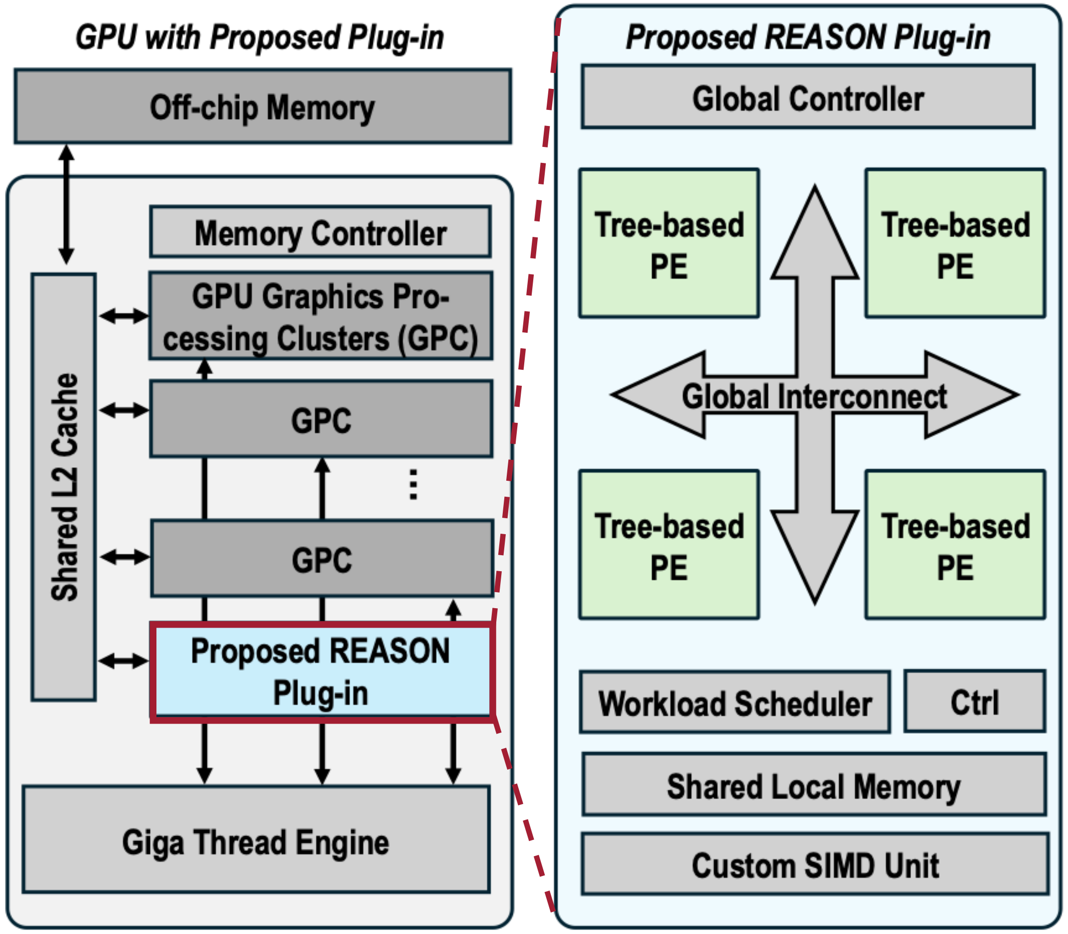
```
// Trigger symbolic execution for a single inference
void REASON_execute(
    int batch_id, // batch identifier
    int batch_size, // number of objects in the batch
    const void* neural_buffer, // neural results in
        shared memory
    const void* reasoning_mode, // mode selection
    void* symbolic_buffer // write-back symb. results
    );
```

```
// Query current REASON status for a given object
int REASON_check_status (
    int batch_id, // batch identifier
    bool blocking // wait till REASON is idle
    );
```

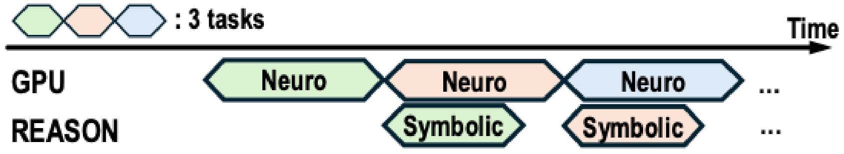
Coordination and synchronization between GPU SMs and REASON processor is handled through **shared-memory flag buffers and L2 cache**

Programming model enables flexibility and control for running diverse neuro-symbolic models

Technique 3: GPU-REASON Two-level Pipeline



GPU-REASON Pipeline:



Inter-REASON Pipeline (symbolic SAT example):

Module \ Cycle	T1-T4	T5	T6-T9	T10	T11	T15	T16	T17-T19	T22	T23	
Broadcast	Broadcast x1	x1 arrives		Broadcast x2	Broadcast x12	x2 arrives	x12 arrives	Broadcast x99	x99 arrives		Multiple parallelable CDCLs
Reduction			x2=1 propagate then x3=0	x3 arrives				Conflict propagate			
L2/DMA						DMA activated	DMA activated	DMA activated	DMA activated	Stop DMA	
PE Activity		Implication x2=1, x3=0					None		Conflicts		
BCP FIFO	[x12=0, x99=1]	[x12=0, x99=1]	[x12=0, x99=1]	[x12=0, x99=1]	[x99=1, x3=0]	[x99=1, x3=0]	[x3=0]	[x3=0]	[x3=0]	[NULL]	
Control	Decide x1=0				Push x3, Pop x12		Pop x99			FIFO Flush	
Watched Literals		No miss detected				No miss detected	Miss detected		conflicts!		

Two-level execution pipeline maximizes concurrency across neuro and symbolic kernels

REASON Enables Efficient Neuro-Symbolic Cognition

Challenge 1:
Heterogeneous & diverse operators

Challenge 2:
Inefficient processing, Low hardware utilization

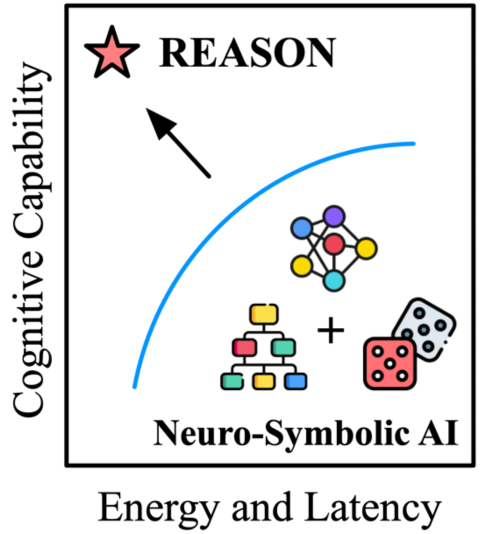
Challenge 3:
Complex control flow

Technique 1:
Unified intermediate representation of diverse symbolic kernels

Technique 2:
Flexible architecture & **compilation flow** to support diverse symbolic kernels

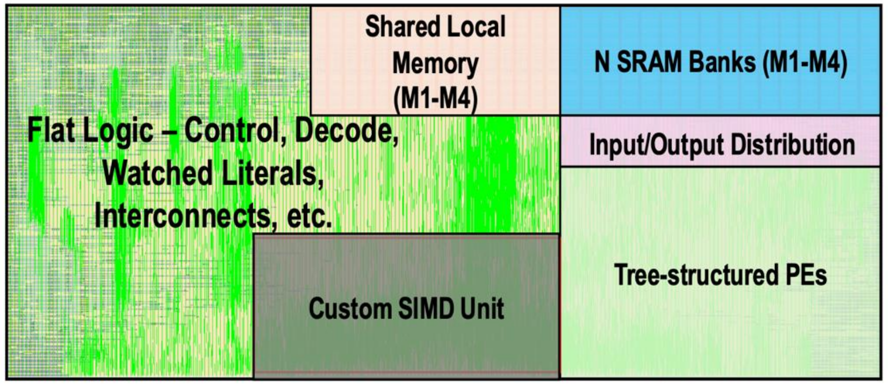
Technique 3:
Programming model with GPU **co-processor integration**

Goal:



Evaluation Setup

- **Task**: cognitive reasoning tasks
- **10 Datasets**: IMO, MiniF2F, TwinSafety, XSTest, CommonGen, News, AwA2, FOLIO, CoAuthor, ProofWriter
- **6 Models**: AlphaGeometry, R2- Guard, GeLaTo, Ctrl-G, NeuroPC, LINC
- **Hardware Baselines**: Orin NX, RTX GPU, CPU, ML accelerator (TPU, DPU)
- **REASON Implementation**: in Verilog, synthesize and PnR @ TSMC 28nm
- **Simulation Setup**: GPU co-processor integration modeled in Accel-Sim



Technology	28 nm
Core VDD	0.9 V
Power	2.12 W
SRAM	1.25 MB
# of PEs	12
# of Nodes	80
DRAM BW	104 GB/s
Area	6 mm^2

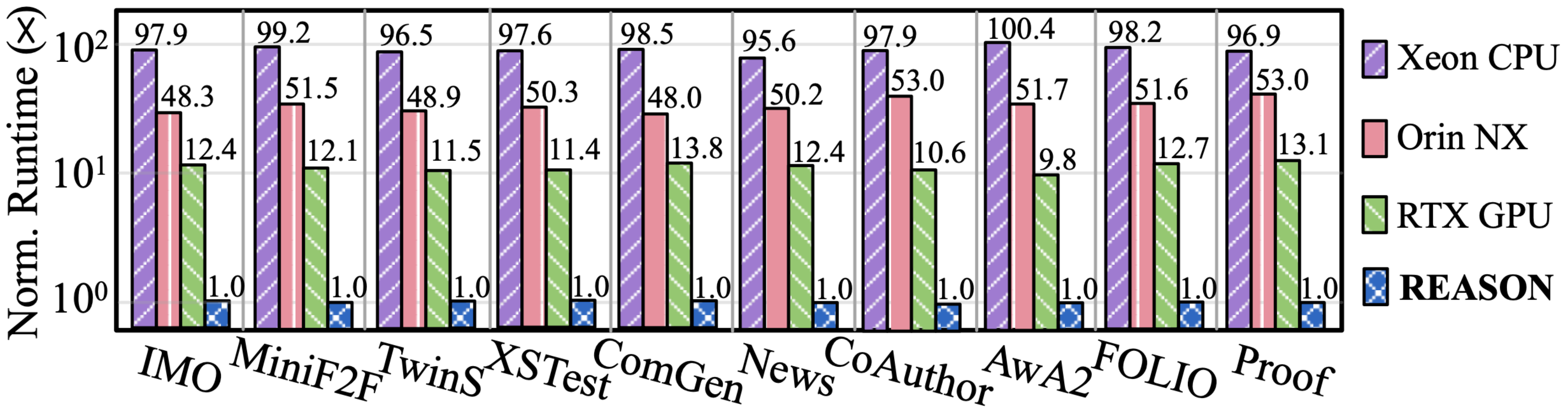
Configured for Orin NX architecture	
# SMs	8
Threads/warp	32
Shared memory	48 KB
L1 cache	128 KB
L2 cache	2 MB
LPDDR5 BW	104 GB/s (peak)

Evaluation – Memory Footprint Reduction

Workloads	Benchmarks	Metrics	Baseline Performance	After REASON Algo. Opt. Performance	After REASON Algo. Opt. Memory ↓
AlphaGeo	IMO	Accuracy (↑)	83%	83%	25%
	MiniF2F	Accuracy (↑)	81%	81%	21%
R^2-Guard	TwinSafety	AUPRC (↑)	0.758	0.752	37%
	XSTest	AUPRC (↑)	0.878	0.881	30%
GeLaTo	CommonGen	BLEU (↑)	30.3	30.2	41%
	News	BLEU (↑)	5.4	5.4	27%
Ctrl-G	CoAuthor	Success rate (↑)	87%	86%	29%
NeuroSP	AwA2	Accuracy	87%	87%	43%
LINC	FOLIO	Accuracy (↑)	92%	91%	38%
	ProofWriter	Accuracy (↑)	84%	84%	26%

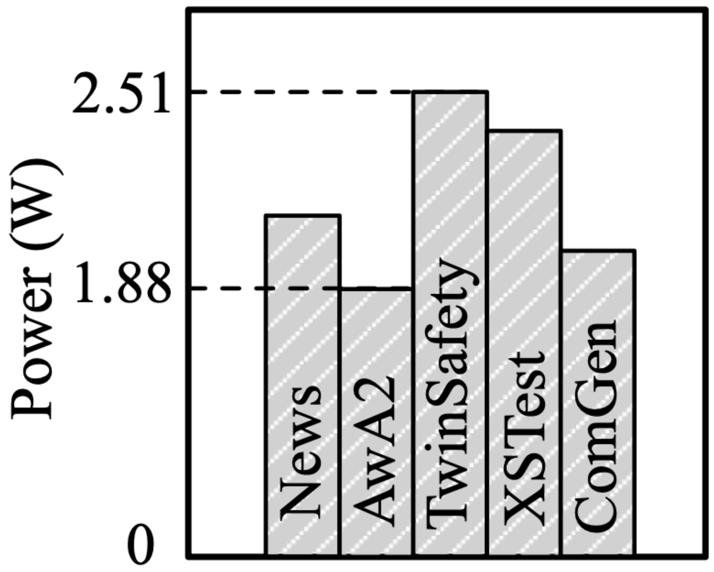
REASON enables 31.7% memory footprint savings on average across six workloads

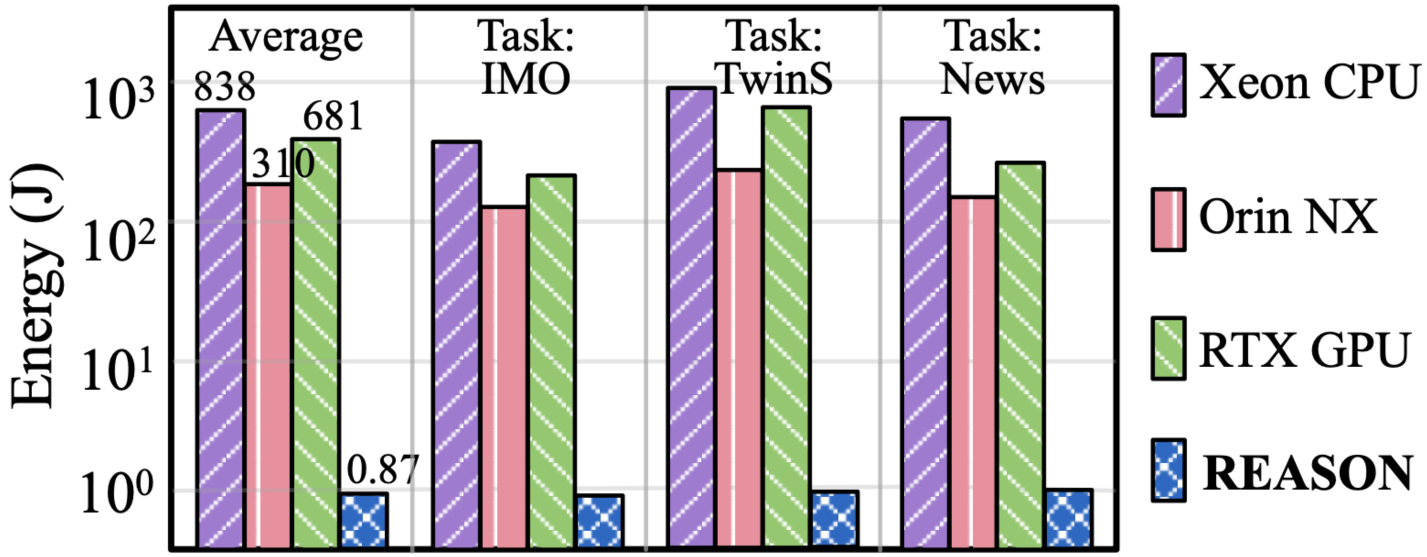
Evaluation – Runtime Performance Improvement



REASON achieves 50x speedup over Orin NX, 12x speedup over RTX GPU

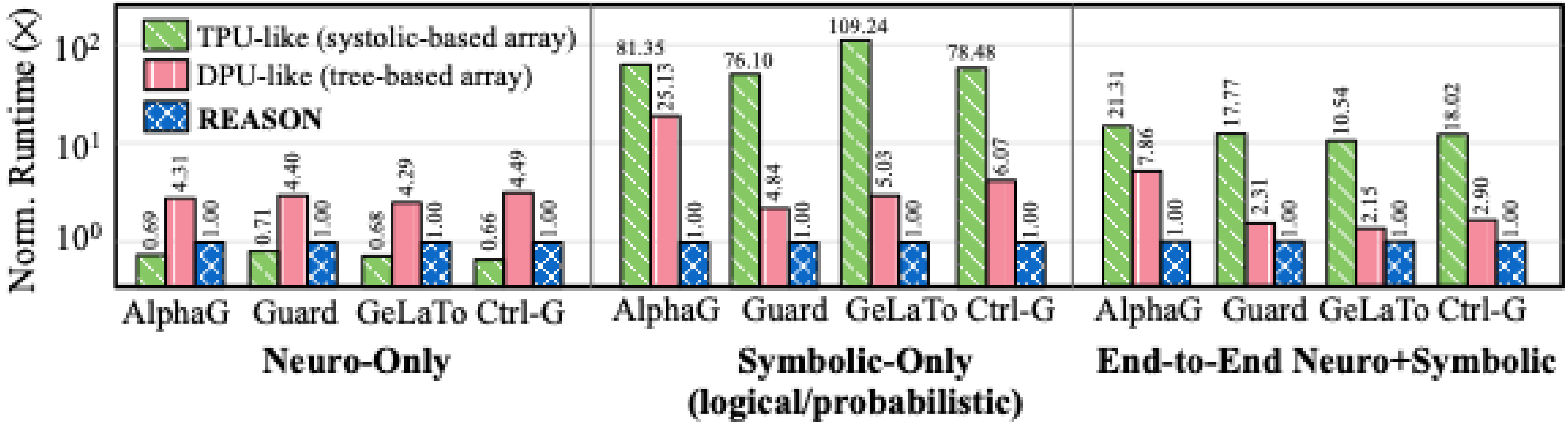
Evaluation – Energy Efficiency Improvement





REASON achieves two orders of magnitude higher energy efficiency compared to CPU/GPU

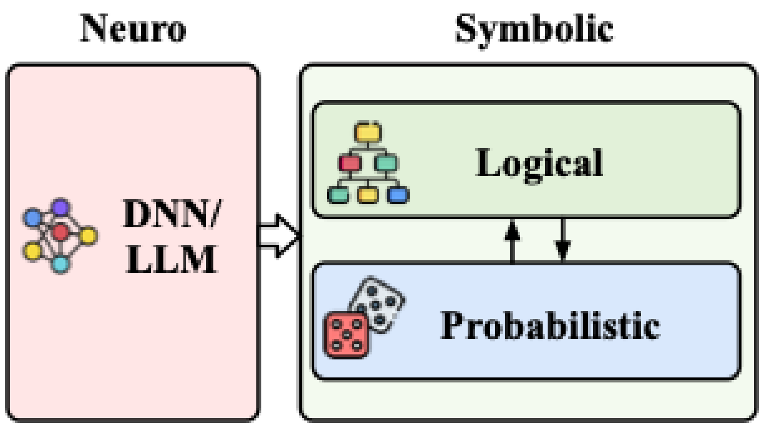
Evaluation – Compare with ML Accelerators

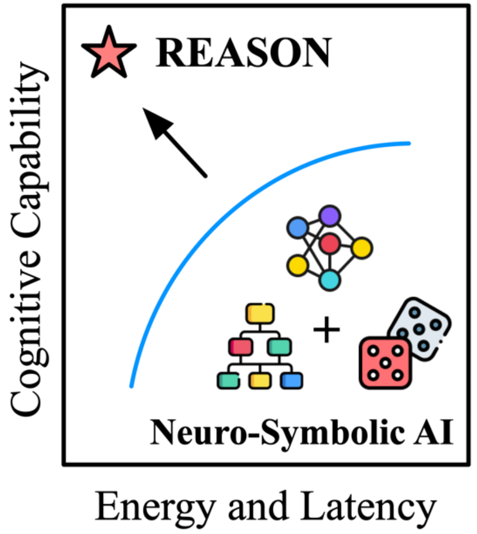


REASON achieves similar performance in neuro, while improved symbolic operation efficiency

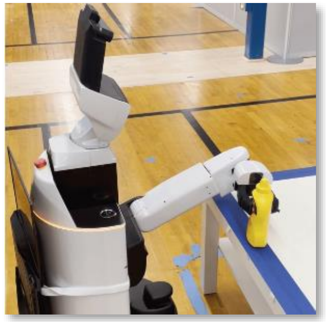
Summary: Key Insights from REASON

- **Neuro-symbolic AI** enables compositional reasoning beyond monolithic LLMs
- Core architectural insights:
  - Symbolic and probabilistic reasoning form a distinct workload class
  - Reasoning can be unified as dependency-driven DAG intermediate representation
  - Reconfigurable architecture enables low-latency reasoning across symbolic kernels
  - Compiler-driven workload-IR-hardware enables efficient mapping and scheduling
  - GPU co-processor integration enables efficient compositional systems

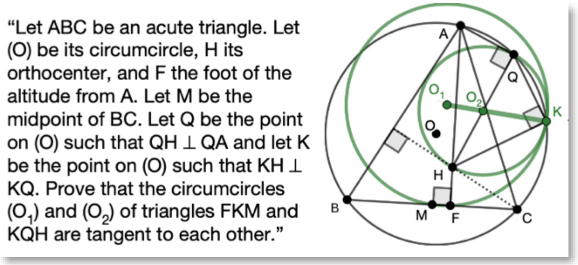




REASON: A Milestone in Building the Foundation of Compositional Intelligence



Embodied Planning



Scientific Reasoning

Trustworthy Trading

Verified Content Generation

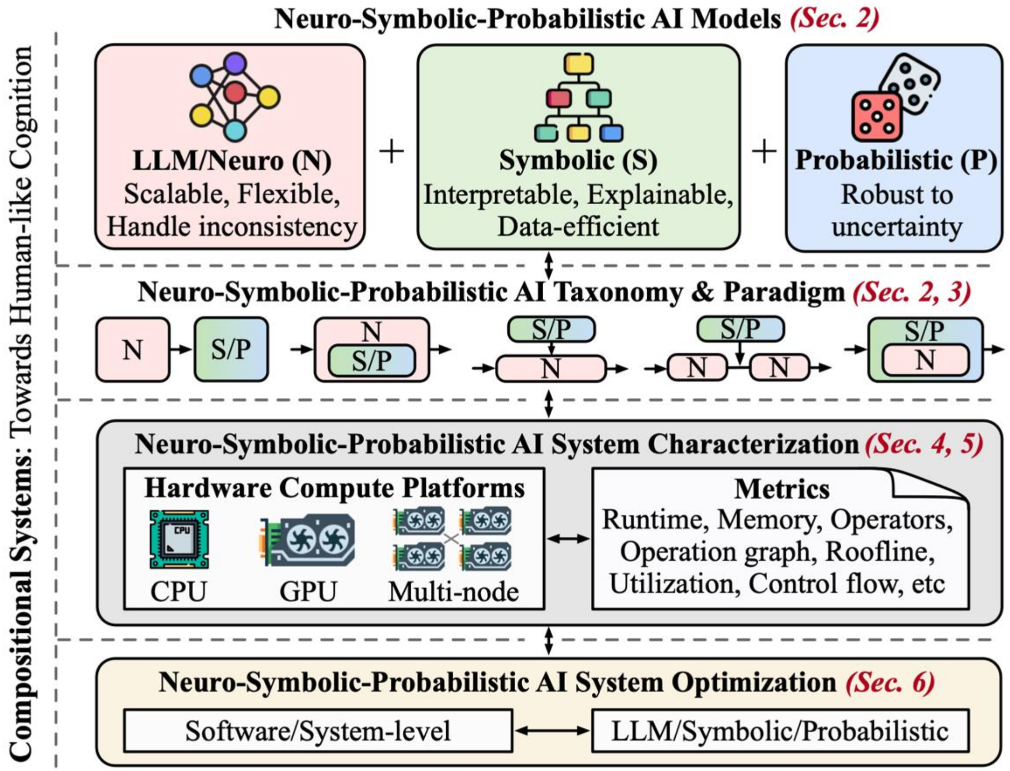
Compositional Reasoning Applications and Beyond

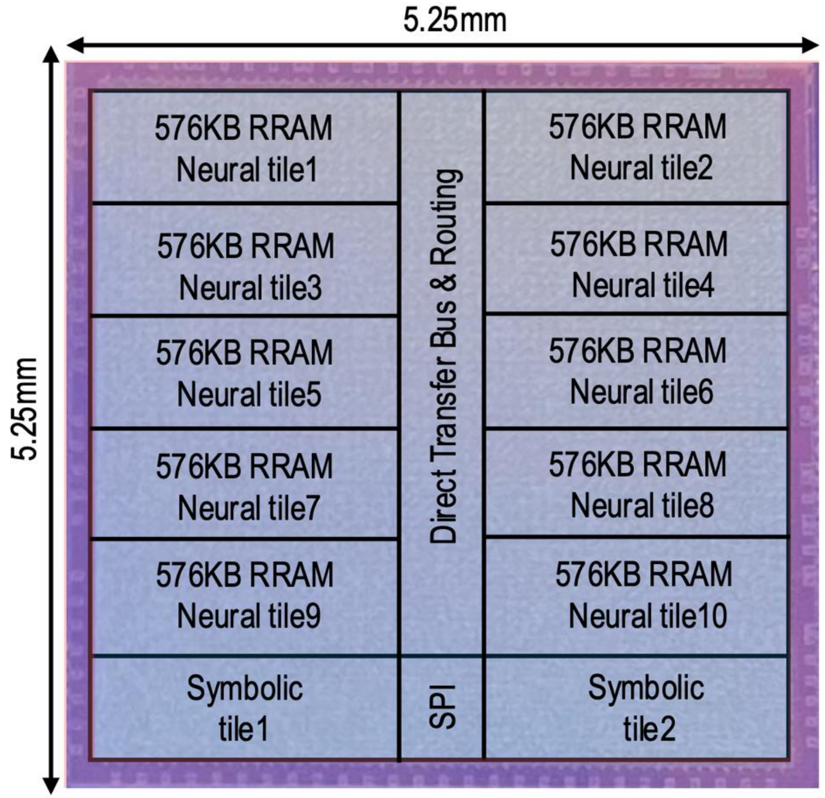
Real-time Cognition	Scalability	Efficient Reasoning

REASON System Achieves All The Three Requirements

Neuro-Symbolic-Probabilistic Form the Foundation of Compositional Reasoning

Our Vision: Full-Stack Design for Neuro-Symbolic AI





System-Level Optimization

“Compositional AI Beyond LLMs: System Implications of Neuro-Symbolic-Probabilistic Architectures”, in **ASPLOS 2026**

System-on-Chip (SoC) Tapeout

“A 40nm Programmable Heterogeneous SoC with 5.625MB/0.85MB RRAM/SRAM for Accelerating Neuro-Symbolic AI Models”, in **JSSC 2026**

Our Vision: Full-Stack Design for Neuro-Symbolic AI

Efficient Processing of Neuro-Symbolic AI: A Cross-Layer Co-Design Tutorial

Students: Zishen Wan[1], Che-Kai Liu[1], Hanchen Yang[1], Ritik Raj[1], Jiayi Qian[1]
Collaborators: Ananda Samajdar[2]
Principal Investigators: Arijit Raychowdhury[1], Tushar Krishna[1]
[1]Georgia Institute of Technology, [2]IBM Research

Recent News

- 01/21/2026 Our paper "A 40nm Programmable Heterogeneous SoC with RRAM/SRAM for Accelerating Neuro-Symbolic AI Models" accepted to **JSSC**.
- 11/07/2025 Our paper "REASON: Accelerating Probabilistic Logical Reasoning for Neuro-Symbolic Cognitive Intelligence" accepted to **HPCA 2026**.
- 06/25/2025 Our paper "Compositional AI Beyond LLMs: System Implications of Neuro-Symbolic-Probabilistic Architecture" accepted to **ASPLOS 2026**. [PDF]
- 04/15/2025 Our tutorial paper "Efficient Processing of Neuro-Symbolic AI: A Tutorial and Cross-Layer Co-Design Case Study" accepted to **NeuS 2025**. [PDF]
- 03/03/2025 Our paper "Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability" accepted to **ISPASS 2025**. [PDF]
- 02/15/2025 Our paper "NSFlow: An End-to-End FPGA Framework with Scalable Dataflow Architecture for Neuro-Symbolic AI" accepted to **DAC 2025**. [PDF]
- 01/29/2025 Our paper "ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents" accepted to **ASPLOS 2025**. [PDF]
- 11/02/2024 Our paper "CogSys: Efficient and Scalable Neurosymbolic Cognition System via Algorithm-Hardware Co-Design" accepted to **HPCA 2025**. [PDF]
- 08/04/2024 Our invited paper "Towards Efficient Neuro-Symbolic AI: From Workload Characterization to Hardware Architecture" accepted to **TCASAI**. [PDF]
- 07/01/2024 Our special session "Neuro-Symbolic Architecture Meets Large Language Models: A Memory-Centric Perspective" accepted to **ESWEEK 2024**. [PDF]
- 02/29/2024 Our paper "Towards Cognitive AI Systems: Workload and Characterization of Neuro-Symbolic AI" accepted to **ISPASS 2024**. [PDF]
- 11/07/2023 Our paper "H3DFACT: Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations" accepted to **DATE 2024**. [PDF]
- 04/10/2023 Our paper "Towards Cognitive AI Systems: A Survey and Prospective on Neuro-Symbolic AI" accepted to **SNAP workshop** at **MLSys 2023**. [PDF]

Recent Talks

- 10/20/2025 "Tailored Computing: Cross-Layer System, Architecture, and Silicon Co-Design for Physical Intelligence" at MICRO PhD Forum, Seoul, Korea.
- 09/05/2025 "System Implications and Opportunities for Compositional Neuro-Symbolic-Probabilistic AI" at Georgia Tech (Host: Dr. Alexey Tumanov), Atlanta, GA.
- 07/11/2025 "Demystifying Neuro-Symbolic AI for Software-Hardware Co-Design" at Purdue University (Host: Dr. Anand Raghunathan), West Lafayette, IN.
- 07/10/2025 "Demystifying Neuro-Symbolic AI for Software-Hardware Co-Design" at University of Notre Dame (Host: Dr. Ningyuan Cao), South Bend, IN.
- 06/24/2025 "Tailored Computing: Domain-Specific Hardware and Systems for Embodied Cognitive Intelligence" at DAC PhD Forum, San Francisco, CA. [Slide]
- 06/24/2025 "NSFlow: An End-to-End FPGA Framework with Scalable Dataflow Architecture for Neuro-Symbolic AI" at DAC, San Francisco, CA. [Slide]
- 06/21/2025 "Efficient and Safe Embodied Intelligence: From Benchmarking to Co-Design" at ISCA Arch4EAI Workshop, Tokyo, Japan. [Slide]
- 05/30/2025 "Efficient Processing of Neuro-Symbolic AI: A Tutorial and Co-Design Case Study" at NeuS, University of Pennsylvania, Philadelphia, PA.
- 05/11/2025 "Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability" at ISPASS, Ghent, Belgium. [Slide]
- 04/17/2025 "Demystifying Neuro-Symbolic AI for Software-Hardware Co-Design" at Google (Host: Dr. Suvinay Subramanian), Mountain View, CA. [Slide]
- 04/02/2025 "ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents" at ASPLOS, Rotterdam, the Netherlands. [Slide]
- 03/31/2025 "Demystifying Neuro-Symbolic AI for Software-Hardware Co-Design" at ASPLOS MLBench Workshop, Rotterdam, the Netherlands. [Slide]
- 03/22/2025 "Bridging Learning and Reasoning: A Cross-Layer Software-Architecture-FPGA-SoC Approach for Neuro-Symbolic AI" at DARPA JUMP2.0 CoCoSys Annual Review, Atlanta, GA.
- 03/13/2025 "Programmable Silicon Prototyping for Various Neuro-Symbolic Models" at CoCoSys Industry Meeting, Atlanta, GA.
- 03/04/2025 "CogSys: Efficient and Scalable Neurosymbolic Cognition System via Algorithm-Hardware Co-Design AI" at HPCA, Las Vegas, NV. [Slide]
- 03/01/2025 "Towards Cognitive AI Systems: Workload and Characterization of Neuro-Symbolic AI" at HPCA MLBench Workshop, Las Vegas, NV. [Slide]

Paper

Project Website

High Performance Computer Architecture

Georgia Tech

Semiconductor Research Corporation

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

REASON: Accelerating Probabilistic Logical Reasoning for Neuro-Symbolic Intelligence






Zishen Wan, Che-Kai Liu, Jiayi Qian, Hanchen Yang, Arijit Raychowdhury, Tushar Krishna

Georgia Institute of Technology, Atlanta, GA

Email: zishenwan@gatech.edu