

# **Tailored Computing: Domain-Specific Systems and** Hardware for Embodied Autonomous Intelligence

**<u>Zishen Wan<sup>1</sup></u>**, Vijay Janapa Reddi<sup>2</sup>, Tushar Krishna<sup>1</sup>, Arijit Raychowdhury<sup>1</sup> <sup>1</sup>Georgia Institute of Technology, Atlanta, GA <sup>2</sup>Harvard University, Cambridge, MA



#### Introduction and Motivation

**Goals**: Develop **embodied systems** that can **perceive, reason, plan, and act** in the physical world, ensuring they are **efficient**, **intelligent**, **trustworthy**, **and robust**.

Healthcare

Manufacture





Smart home



Mixed-reality



Host SoC

Aemory bus

DRAM

CogSys Accelerato

Memory bus Ctrl

Memory bus Ctrl bu

Custom SIMD Unit

Workload

- † ‡ (

DRAM

Education

#### **Challenges**:



### **Research Overview**

<u>My research: Tailored computing methodology for cross-layer co-optimization of software, system,</u> and hardware, enabling efficient, reliable, and adaptable architectures for embodied intelligence.





### Software-System-Hardware Cross-Layer Design for Neuro-Symbolic (NeSy) Intelligence

**Problem**: What is **system characteristics** of NeSy AI? **Insights**:

- Compositional system bridges neural learning, symbolic reasoning, and probabilistic inference.
- **Compute:** heterogenous operational kernels.
- **System**: memory-bound, low ALU util, irregular access. **Results**:
- First automated NeSy AI profiling tool: program trace -> dataflow graph -> operator extraction



**Problem:** How to **optimize efficiency** of NeSy AI system? Insights:

- **Processing element**: reconfigurable neuro/symbolic PE.
- **Architecture**: host + scalable neuro/symbolic PE array.
- **Dataflow**: bubble streaming dataflow.
- **FPGA prototype**: end-to-end automated design flow. **Results**:
- *First* NeSy Al **architecture** and FPGA **prototype**.
- **75x** speedup over TPU; **4-96x** speedup over edge GPU.

top in B 🛡

	<u>ــــــــــــــــــــــــــــــــــــ</u>	5.25m	m
1	And a local of the local of the local	3555	
5.25mm	576KB RRAM Neural tile1		576KB RRAM Neural tile2
	576KB RRAM Neural tile3	ry & r Bus	576KB RRAM Neural tile4
	576KB RRAM Neural tile5	d Memo Transfe	576KB RRAM Neural tile6
	576KB RRAM Neural tile7	Share	576KB RRAM Neural tile8
	576KB RRAM Neural tile9		576KB RRAM Neural tile10
	Symbolic tile1	SPI & Scan Routing	Symbolic tile2

0: Firmware Dev.	Step 1: Off-cl	nip Scheduler	Ste	o 2: Test-time	Program Ex
Clk generation Power monitor board operations RRAM WR APIs	1. Workload a a. Operator b. Operator c. Operator	<ol> <li>Workload analysis:</li> <li>a. Operator graph &amp; trace</li> <li>b. Operator runtime</li> <li>c. Operator size and mem</li> </ol>		1. Write RRAM: a. Core & bank No. b. OTP or form c. Write and verify.	
zation()	<ul><li>2. Workload scheduler:</li><li>a. Neuro or symb. core</li><li>b. Mapping to bank No.</li><li>c. Power gating</li></ul>			2. WR I\$ and ctrl. regs. a. HP mode & bias b. Power status c. Inst. for NeSy app.	
et GPIO Directions et GPIO SPI C for Power/Clk DOs Config. ower Up	3. SDPM: a. Config.1 b. Config.2 c. Config.3	3. Inst. Seq.: a. LD b. VV c. RRAM	,	PC GUI: a. Power b. P/Fail c. TOPS d. Done	<sup>©</sup> refresh ≮

# [ASPLOS26 | HPCA25 | DAC25 | TCASAI24 | DATE24 | ISPASS24]

**Problem**: How to **deploy** and **program** NeSy hardware? **Insights**:

- **Chip tapeout**: programmable SoC @TSMC 40nm; integrated with RRAM/SRAM, NeSy tiles, and RISC-V cores.
- **Compiler**: programming support for various kernels.
- **Power management**: scheduler-informed power mgmt.

# **Results**:

- *First* NeSy Al SoC **test chip**.
- 10.8 TOPS/W energy efficiency, 321 mW peak power.

Software-System-Hardware Cros	s-Laver Desi	ign for Cool	perative Embodie	ed Intelligence

#### [ASPLOS25 | ISPASS25 | ICCAD24 | CACM24 | DAC23]

- **Problem**: What is **sys. characteristics** of embodied agent? Insights:
- **Compositional system** integrates perception, LLM-driven cognition, and physical actions for long-horizon tasks.
- **Source of inefficiency**: longed plan latency, redundant interaction, memory inconsistency, complex control. **Results**:
- *First* **benchmark suite** for embodied AI system: 15 benchmarks, 4 paradigms, 4 key metrics.



- **<u>Problem</u>**: How to **optimize efficiency** of embodied system? Insights:
- **Memory**: long-term persistent & short-term dynamic.
- **Scalability**: inter-cluster central & intra-cluster decentral.
- **Operation**: planning-guided multi-step execution.
- System: prioritizing system morphology brings adaptability. **Results**:
- *First* system-level embodied agent opt framework.
  - **3.4x** speedup over baseline agentic systems.



**Problem:** How to **deploy** embodied agent on suitable HW? **Insights**:

- **SoC**: heterogenous architecture with GPU for high-level planning and accelerator for low-level action.
- Interface: programming model for GPU-accelerator.
- Adaptability: design config via system requirement. **Results**:
- *First* embodied agent heterogenous **SoC prototype**.
- **10.3x** speedup over GPU-based agentic systems.



[ASPLOS24 | DATE23 | TCAD23 | MICRO22 | ICCAD22 | DATE22 | DAC21 ]



# Software-System-Hardware Cross-Layer Design for Physical Autonomy Intelligence

- **Problem**: How to **accelerate** low-level physical autonomy? Insights:
- Domain-specific architecture with system morphology. **Dataflow:** multi-level data reuse, time-multiplexing. **Memory optimization**: layout, sparsity, symmetry. **Spatial-aware computing** for environment dynamics. **Results**:
- **<u>Problem</u>**: How to improve **energy efficiency** of auto machines? Insights:
- **Algorithm:** robust low-voltage on/off-device learning.
- **<u>Problem</u>**: How to **deploy** physical autonomy **safely**? Insights:
- **Safety characterization**: end-to-end fault analysis tool, autonomy kernels have inherent robustness variations.

- *First* **benchmark suite** for robotics computing perf. U [5x5] X [5x3] Symmetry S matrix **W [3**×5] **V** [3x3] Coobservation U X w v Compute complexity:  $O(n^3) \rightarrow O(n)$ 4.1x reduct 720 kb
- **System**: collaborative spring-or-slack computing minimizes power across distributed resource-constrained nodes.
- Hardware: dynamic thermal-payload optimization. **Results**:
- *First* **perf-efficiency-robustness co-opt** framework.



- **Safety deployment**: vulnerability-adaptive protection, assign protection budget based on robustness level. **Results**:
- *First* **fault analysis** framework for robotic systems.

