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

Zishen Wan[1], Vijay Janapa Reddi[2], Tushar Krishna[1], Arijit Raychowdhury[1]
[1]Georgia Institute of Technology, Atlanta, GA [2]Harvard University, Cambridge, MA

DAC PhD Forum

Webpage

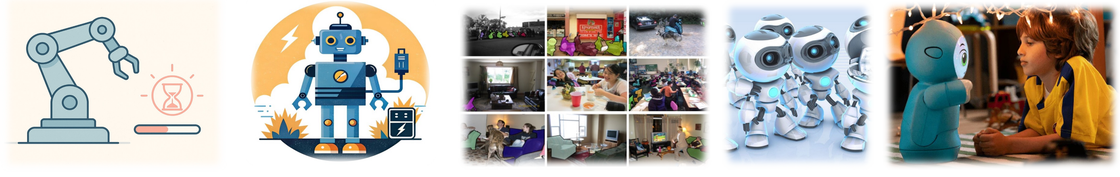
Introduction and Motivation

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

Manufacture | Smart home | Healthcare | Mixed-reality | Education

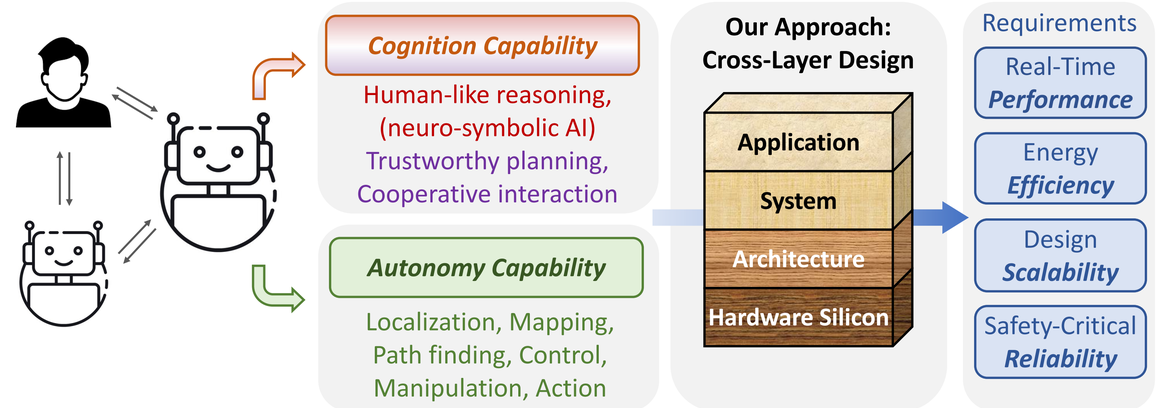
Challenges:

Real-time | Energy efficiency | Adaptability | Scalability | Safety



Research Overview

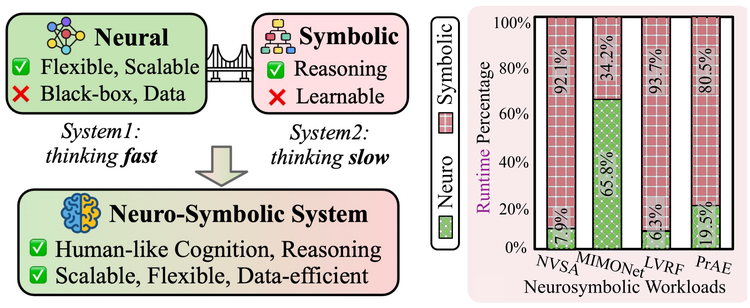
My research: **Tailored computing methodology** for **cross-layer co-optimization** of software, system, and hardware, enabling efficient, reliable, and adaptable architectures for **embodied intelligence**.



Software-System-Hardware Cross-Layer Design for Neuro-Symbolic (NeSy) Intelligence [ASPLOS26|HPCA25|DAC25|TCASAI24|DATE24|ISPASS24]

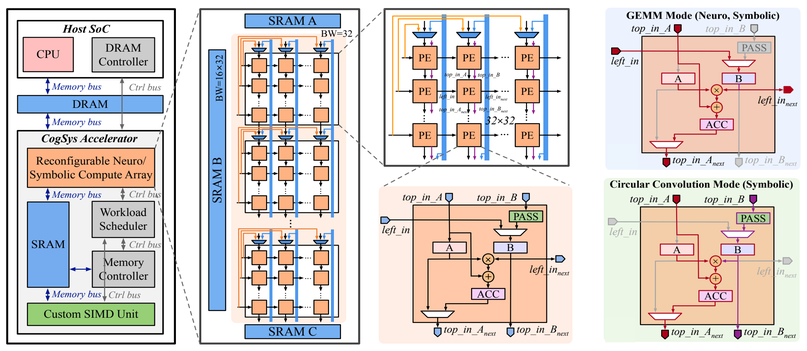
Problem: What is **system characteristics** of NeSy AI?
Insights:
- **Compositional system** bridges neural learning, symbolic reasoning, and probabilistic inference.
- **Compute**: heterogenous operational kernels.
- **System**: memory-bound, low ALU util, irregular access.

Results:
- *First* automated NeSy AI **profiling tool**: program trace -> dataflow graph -> operator extraction



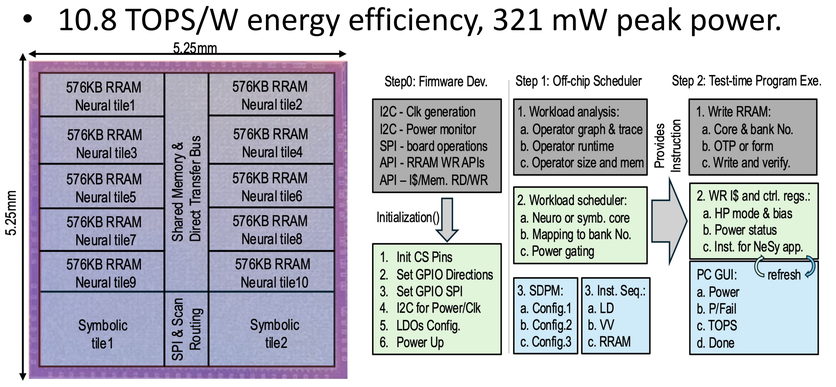
Problem: How to **optimize efficiency** of NeSy AI system?
Insights:
- **Processing element**: reconfigurable neuro/symbolic PE.
- **Architecture**: host + scalable neuro/symbolic PE array.
- **Dataflow**: bubble streaming dataflow.
- **FPGA prototype**: end-to-end automated design flow.

Results:
- *First* NeSy AI **architecture** and FPGA **prototype**.
- **75x** speedup over TPU; **4-96x** speedup over edge GPU.



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.

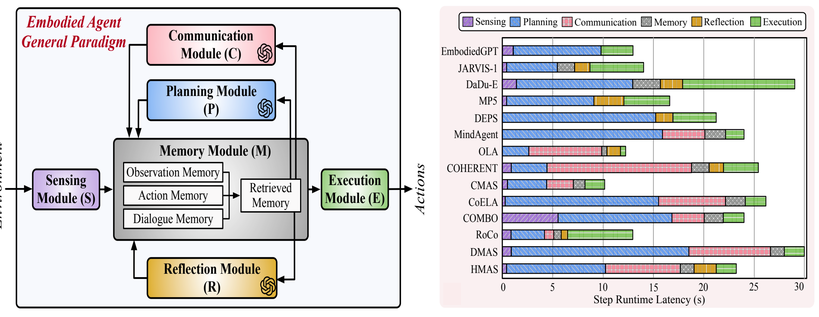
Results:
- *First* NeSy AI SoC **test chip**.
- 10.8 TOPS/W energy efficiency, 321 mW peak power.



Software-System-Hardware Cross-Layer Design for Cooperative Embodied Intelligence [ASPLOS25|ISPASS25|ICCAD24|CACM24|DAC23]

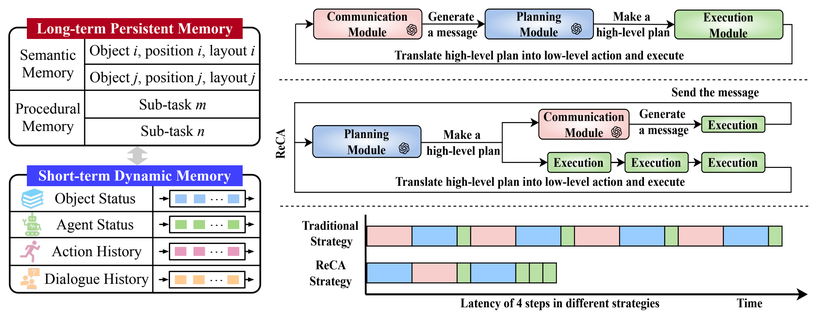
Problem: What is **sys. characteristics** of embodied agent?
Insights:
- **Compositional system** integrates perception, LLM-driven cognition, and physical actions for long-horizon tasks.
- **Source of inefficiency**: longed plan latency, redundant interaction, memory inconsistency, complex control.

Results:
- *First* **benchmark suite** for embodied AI system: 15 benchmarks, 4 paradigms, 4 key metrics.



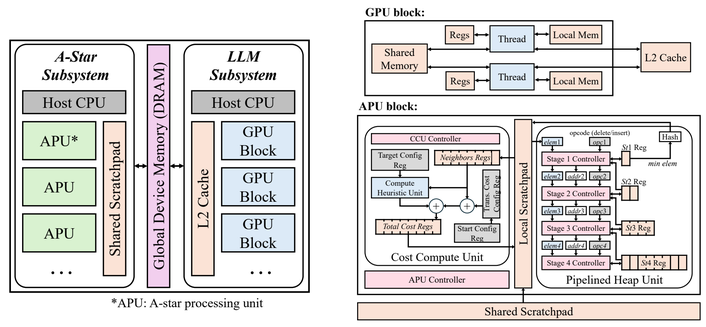
Problem: How to **optimize efficiency** of embodied system?
Insights:
- **Memory**: long-term persistent & short-term dynamic.
- **Scalability**: inter-cluster central & intra-cluster decentral.
- **Operation**: planning-guided multi-step execution.
- **System**: prioritizing system morphology brings 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.



Software-System-Hardware Cross-Layer Design for Physical Autonomy Intelligence [ASPLOS24|DATE23|TCAD23|MICRO22|ICCAD22|DATE22|DAC21]

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:
- *First* **benchmark suite** for robotics computing perf.

Problem: How to improve **energy efficiency** of auto machines?
Insights:
- **Algorithm**: robust low-voltage on/off-device learning.
- **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.

Problem: How to **deploy** physical autonomy **safely**?
Insights:
- **Safety characterization**: end-to-end fault analysis tool, autonomy kernels have inherent robustness variations.
- **Safety deployment**: vulnerability-adaptive protection, assign protection budget based on robustness level.

Results:
- *First* **fault analysis** framework for robotic systems.