CogSys: Efficient and Scalable Neuro-Symbolic Cognition System via Algorithm-Hardware Co-Design

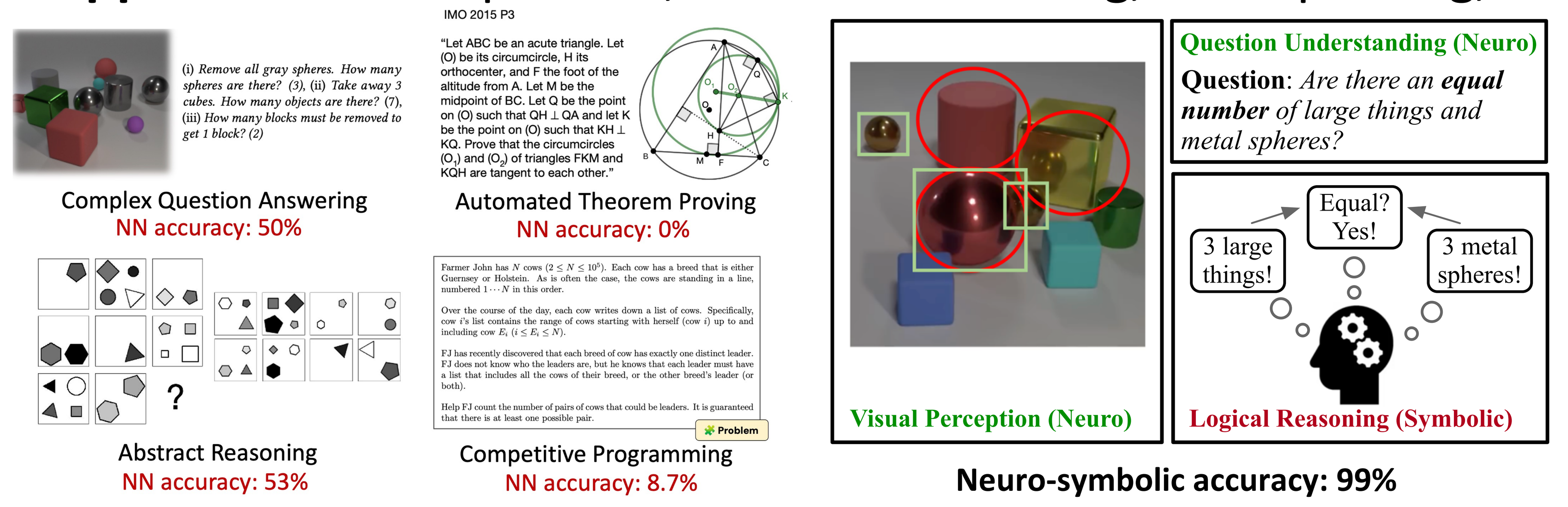
Zishen Wan[1*], Hanchen Yang[1*], Ritik Raj[1*], Che-Kai Liu[1], Anand Samajdar[2], Arijit Raychowdhury[1], Tushar Krishna[1]

[1]Georgia Tech, Atlanta, GA, USA [2]IBM Research, Yorktown Heights, NY, USA

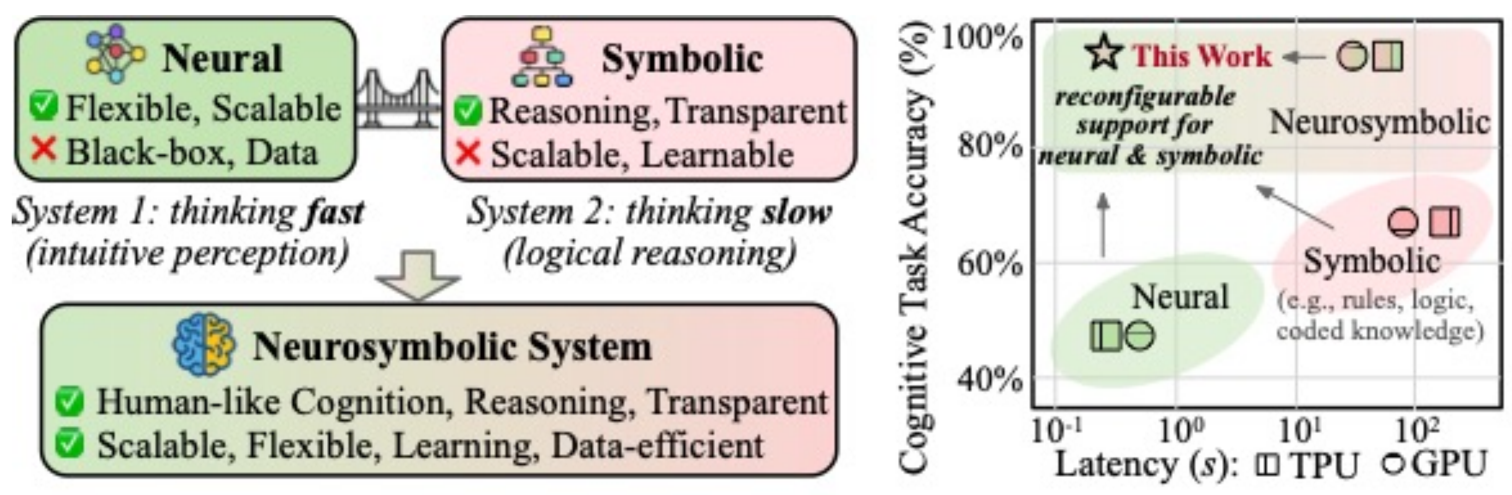
Motivation: Why Neuro-Symbolic AI?

❖ **Compositional system** to enhance cognitive capability

✓ **Applications**: complex QA, abstract reasoning, math proving, etc



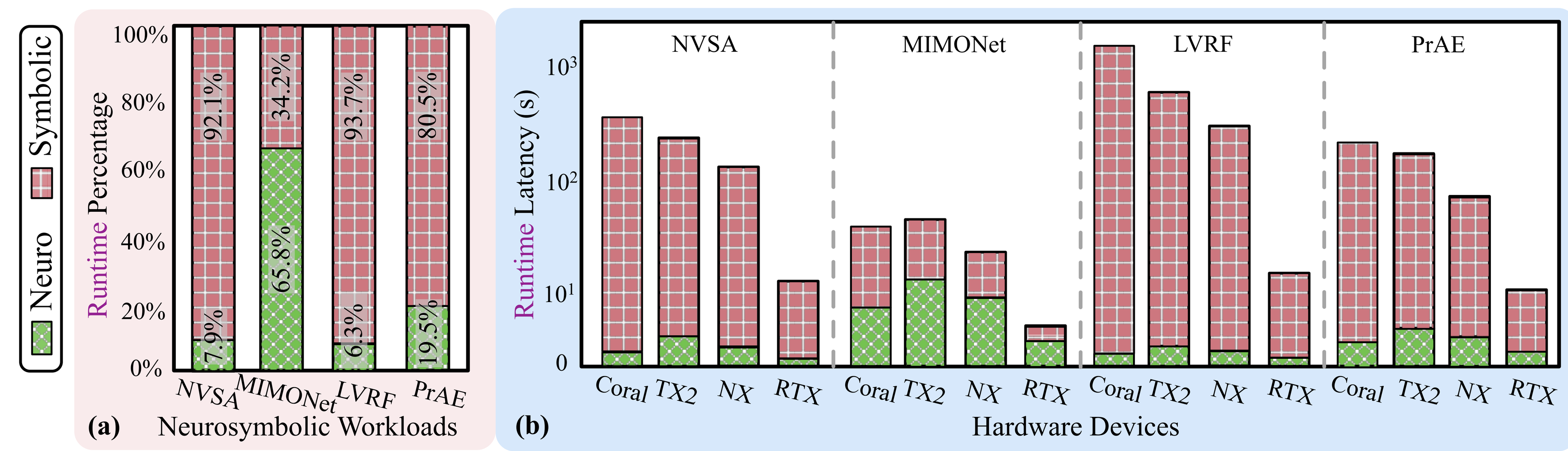
✓ **Neuro-Symbolic AI** bridges neural learning & symbolic reasoning



System Characteristics and Challenges

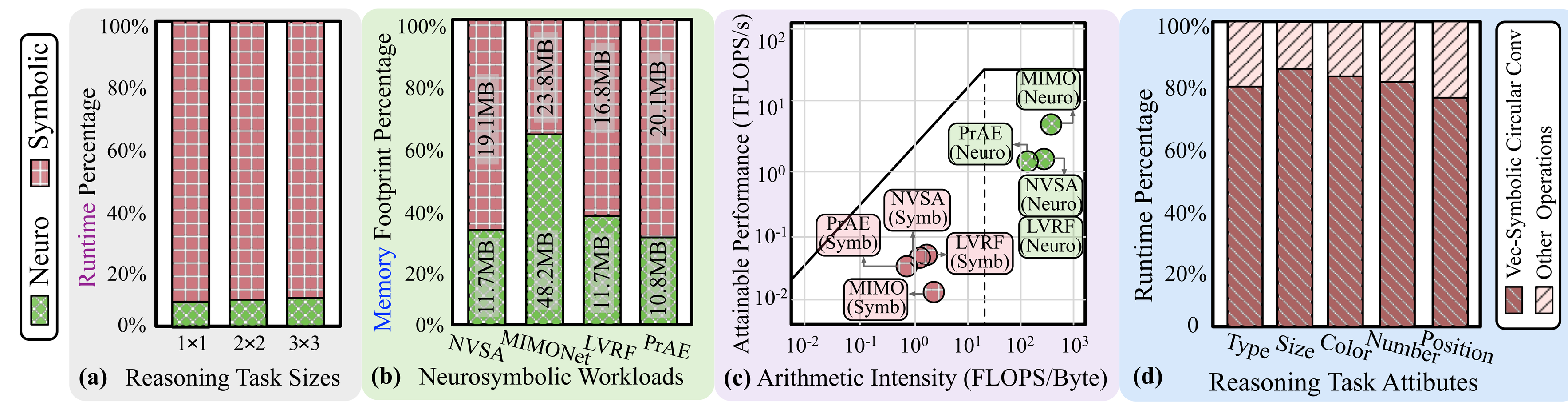
❖ **System Challenges**: Latency, memory, heterogeneity

✓ **Latency**: large end-to-end runtime; symbolic critical path bottleneck

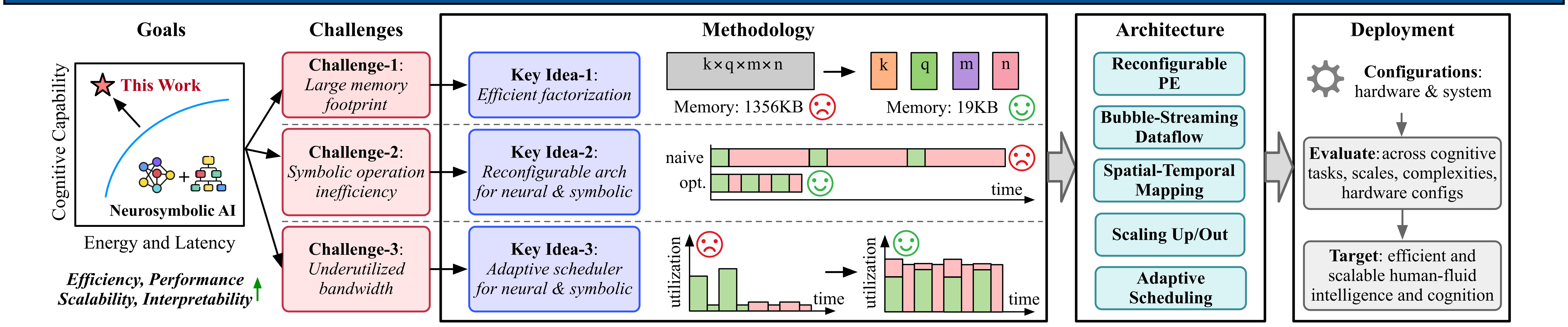


✓ **Memory:** symbolic memory-bound; low ALU util, low cache hit rate

✓ **Heterogeneity**: neuro (GEMM, conv), symbolic (vector, circular conv)



Methodology: Software-Hardware Co-design for Real-Time and Efficient Neuro-Symbolic AI

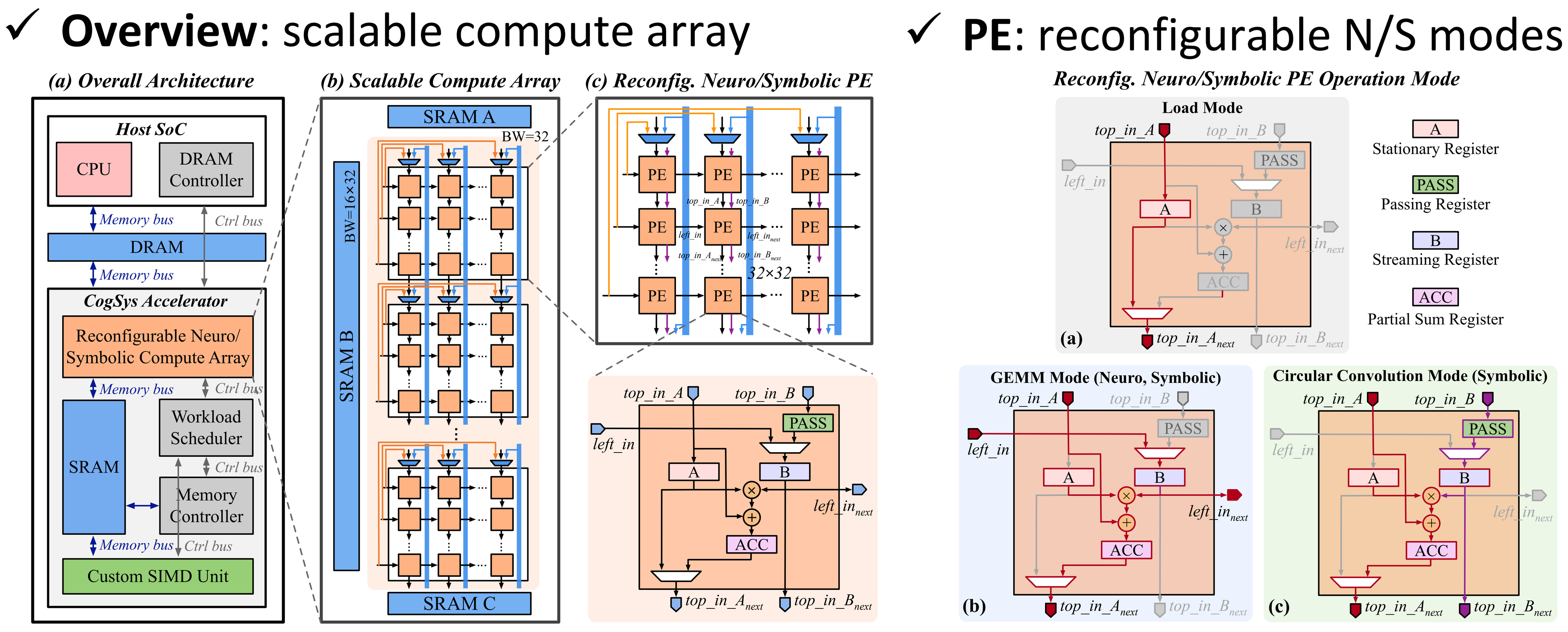


Cross-Layer Optimization: Algorithm, Hardware Architecture, System Scheduling

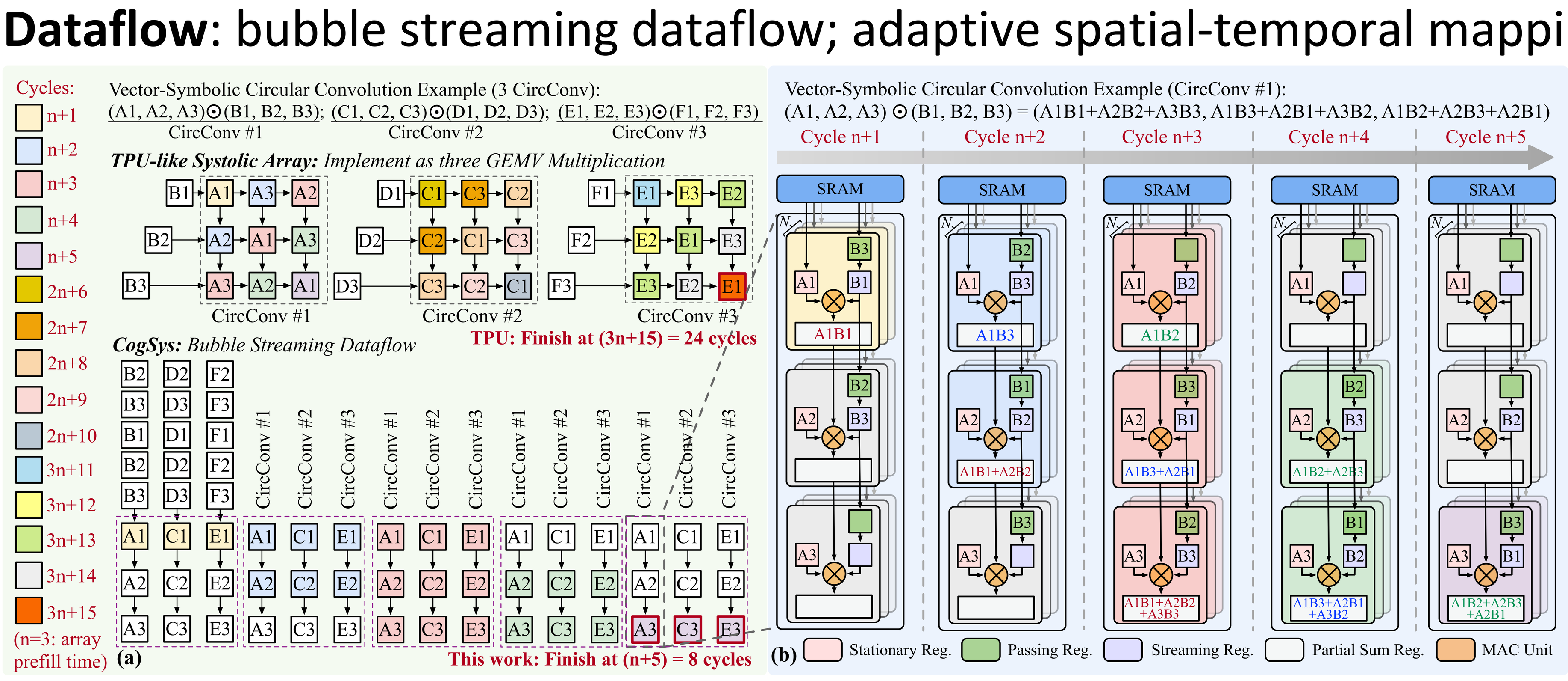
❖ **Hardware**: Reconfigurable neuro-symbolic architecture

✓ **Overview**: scalable compute array

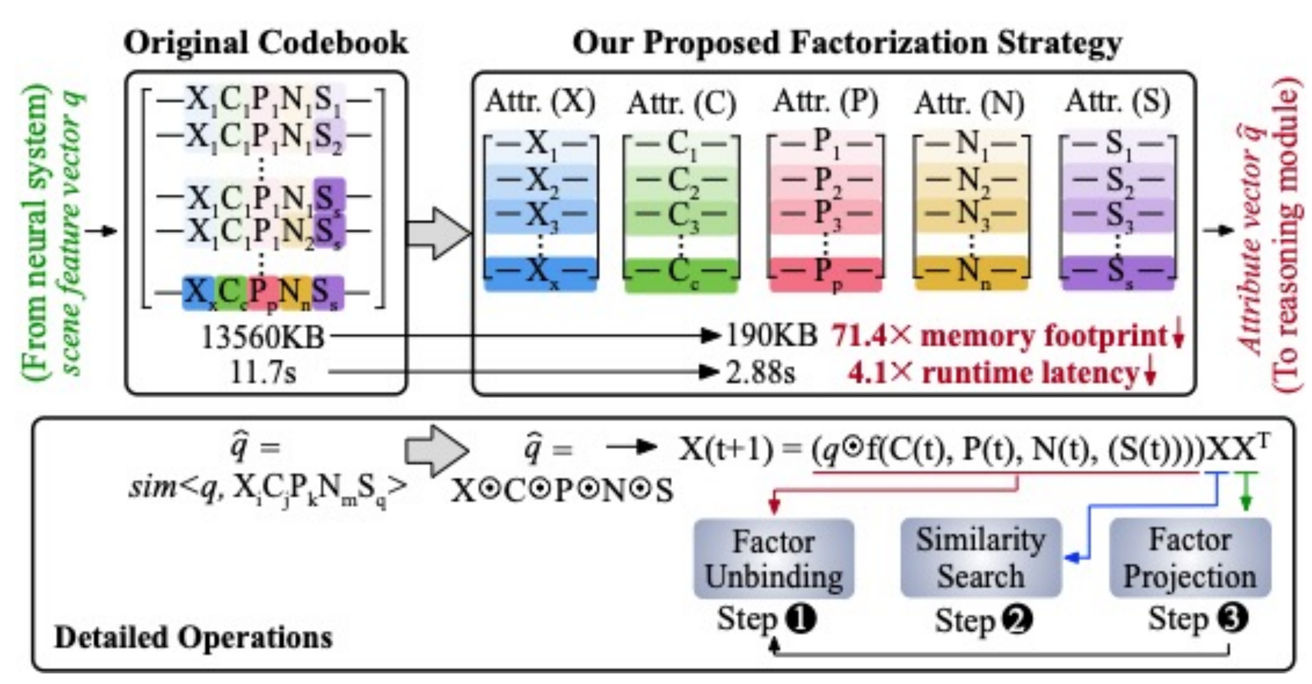
✓ **PE**: reconfigurable N/S modes



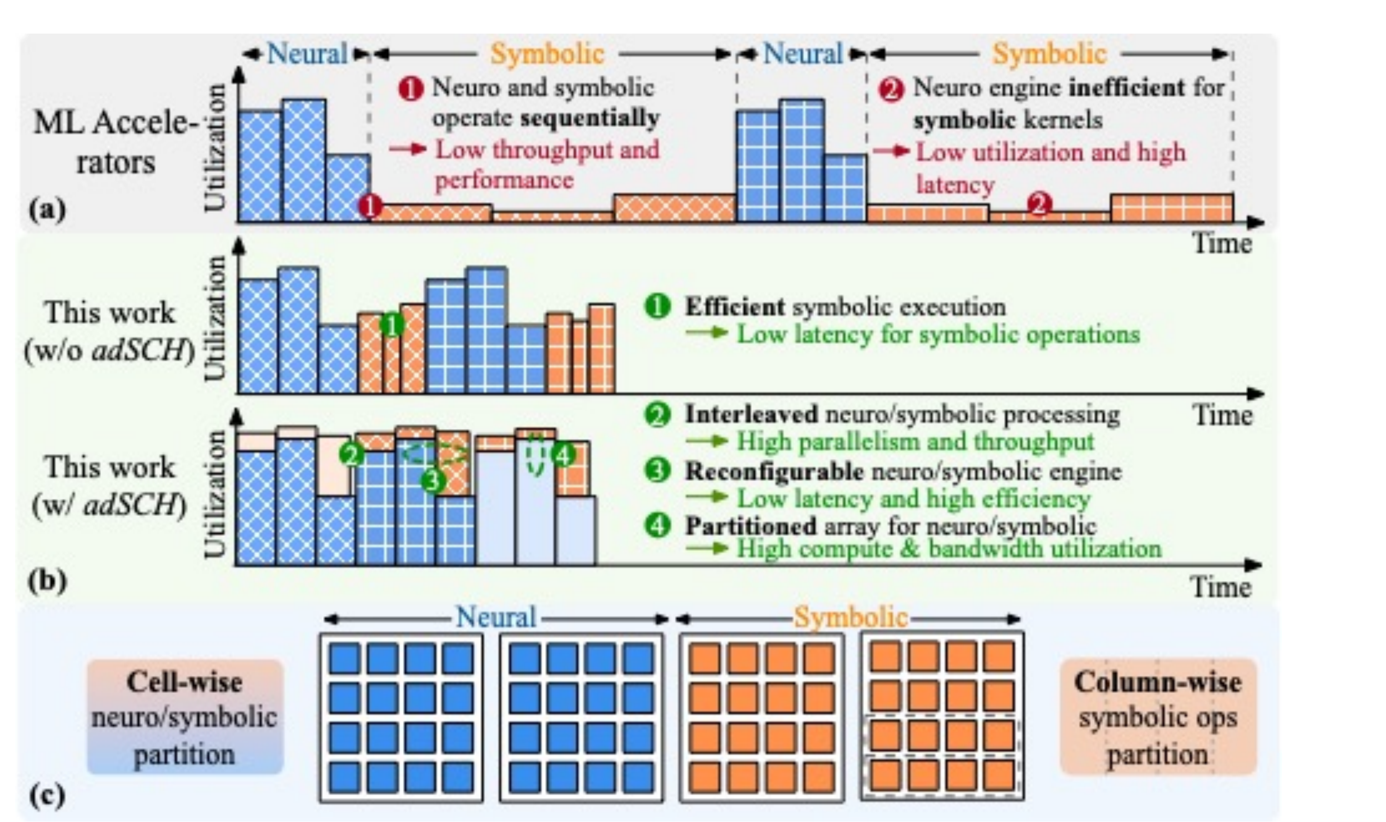
✓ **Dataflow**: bubble streaming dataflow; adaptive spatial-temporal mapping



❖ **Algorithm**: Efficient symbolic factorization

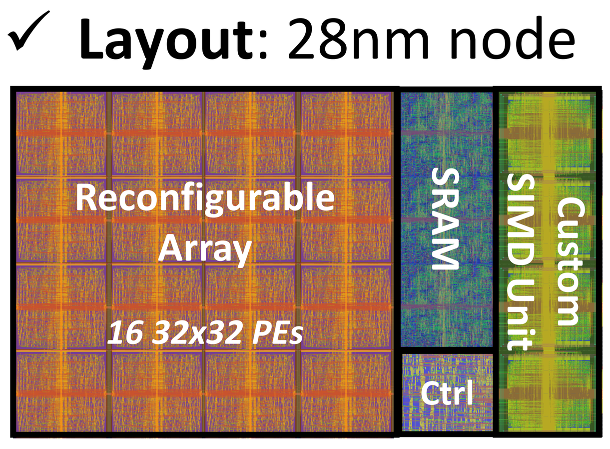


❖ **System**: Adaptive workload-aware scheduling

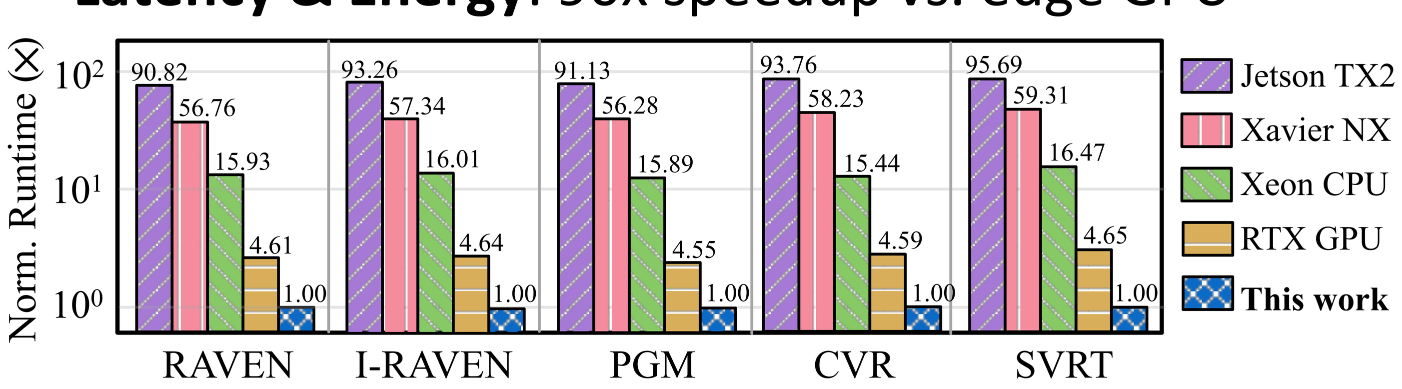


Evaluation Results

✓ **Layout**: 28nm node



✓ **Latency & Energy**: 90x speedup vs. edge GPU



✓ **Compared with ML accelerators**: 1.7-15.9x speedup