Embedded AI for Autonomous Vehicles:

Efficient Programmability of Cognitive Heterogeneous Systems (EPOCHS)

Pradip Bose, Sarita Adve, Vikram Adve, Sasa Misailovic, Luca Carloni, Ken Shepard, David Brooks, & Gu-Yeon Wei IBM, UIUC, Columbia, & Harvard

Architectures Thrust: Domain-Specific System on Chip (DSSoC)

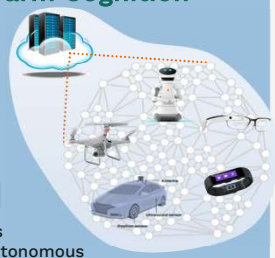
EPOCHS Project Statement

How to design and quickly implement an easily programmed domain-specific SoC that implements a smart, real-time cognitive decision engine within smart connected vehicles (e.g., cars or drones)

System & Application Context: Cloud-Backed Mobile Swarm Cognition

Swarm AI

- Mobile (swarm) computing *(cloud-backed)*
- Interaction over *ad hoc* wireless networks
- Resilient system reconfiguration
- Efficiency knobs within edge devices
  - Approximation, sampling, filtering
  - Machine learning acceleration
  - Dynamic voltage and frequency control



DoD and Commercial Vendor Impact Target

- ✓ Military combat vehicles, drone-swarms
- ✓ Smart connected cars: semi- or fully autonomous

Optimized Hardware for Embedded AI

Next generation cognitive IoT embedded systems

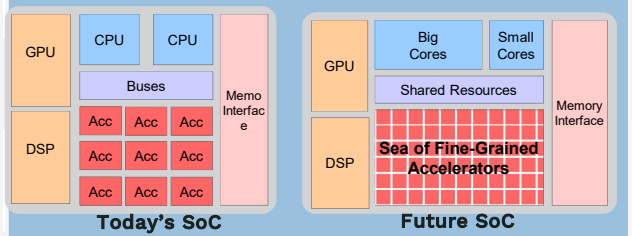
Today's cars are moving data centers with onboard sensors and computers that capture real-time information.

What is needed?

- On-device computational and inference capability
- Low Power
- Resilience to harsh environment
- Security against malicious attacks

Custom cognitive hardware with built-in resilience features

Edge Embedded Intelligent SoC Trends



Challenges: design integration complexity, validation, efficiency, reliability, programmability

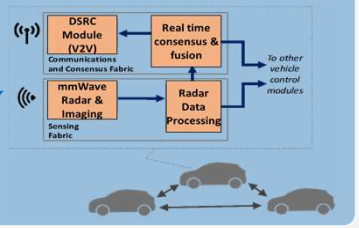
EPOCHS – IBM's DSSoC Project

EPOCHS: Efficient Programmability of Cognitive Heterogeneous Systems

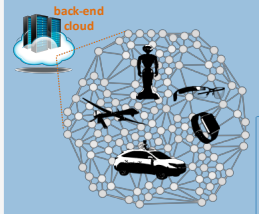
- Innovate novel methodologies for rapid development of multi-application systems through a single programmable, heterogeneous SoC
- Focus on autonomous vehicles with applications drawn from the domains of *Computer Vision and Software Defined Radio*

EPOCHS Reference Application (ERA)

- *Key driver of the cross-layer rapid development methodology*
- *Application for multi-vehicle (cooperative) sensor fusion*



EPOCHS: Key Innovations (Summary)



Backed by our PERFECT expertise in low power ML/DL accelerator design and tape-outs (16 nm TSMC & 14nm GF)

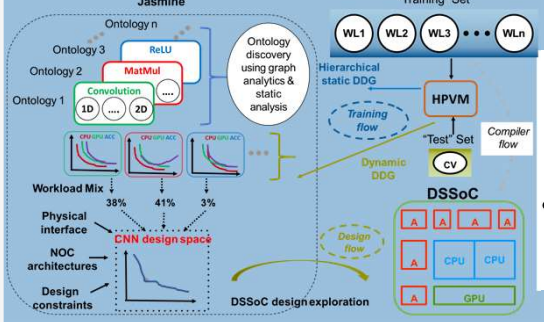
Flexibility
Programmability
Design Cost

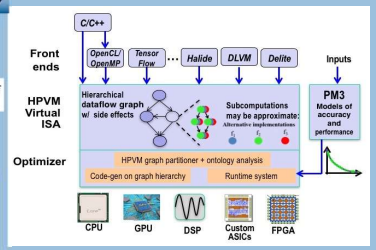
- EPOCHS Reference Application (ERA) – multi-domain application to drive the project + aid technology transfer
- Open-source tools + ecosystem for SoC arch. definition
- H/W assisted smart task scheduler, resource manager
- Retargetable open-source compiler (LLVM/HPVM)
- Accelerator IP identification and placement (Jasmine)
- GALS-driven NoC, swarm AI resource management
- Rapid system-level integration methodology (ESP)

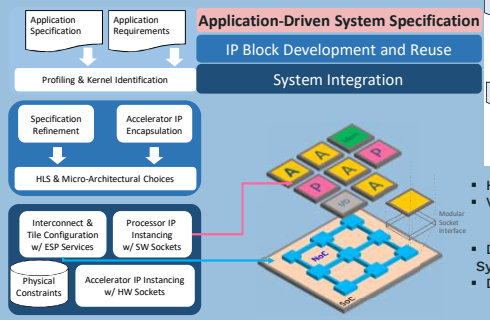
Technology Transition Impact Trajectory

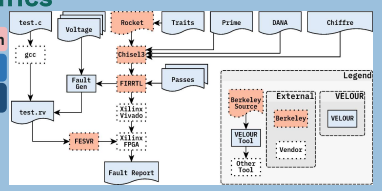
- **Commercialization Paths:**
  - IBM Watson IoT for Automotive:
    - https://www.ibm.com/internet-of-things/industries/iot-automotive/connected-cars
    - https://www.ibm.com/us-en/marketplace/watson-assistant-for-automotive
    - https://www.ibm.com/blogs/internet-of-things/iot-accessibleolli-drives-us-forward-at-ces/
  - Partner/collaborators like Honda, Samsung, BMW, Toyota, etc. leveraging also IBM's Research Frontier Institute (RFI)
    - http://www.research.ibm.com/frontiers/
- **DoD Impact Paths:**
  - Guidance from DoD military combat vehicle experts
  - Drone-swarm collaborative impact – Naval Research Lab (NRL)
  - Design methodology and IP contribution to IBM Albany Advanced Technology Center (ATC), with possible DoD support

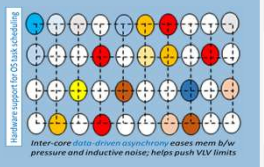
EPOCHS – IBM's DSSoC Project: Exemplary Innovation Specifics











- Systematic identification of accelerator components
- Rapid design space exploration
- Novel system-level integration
- Open-source compiler and programming model

- System-level design methodology promoting HW/SW codesign
- Flexible integration of heterogeneous components; rapid prototyping

- Heterogeneous cores organized in clusters of cores with local mem
- Very Low Voltage (VLV) operation for both logic and SRAM
  - Core-level voltage domains; circuit-level support for VLV SRAM[1]; cross-layer optimization[2]; aggressive under-volting with resilience guard
- Data-driven asynchrony between voltage domains – Globally Asynchronous Locally Synchronous (GALS)
- Decentralized resource self-management using swarm-AI learning
  - Distributed error recovery and reconfiguration

1. R. V. Joshi, M. M. Ziegler, et al., "A low voltage SRAM using resonant supply boosting," J. Solid-State Circuits 2017
2. E. Cheng et al., "CLEAR: Cross-layer exploration for architecting resilience: Combining hardware and software techniques to tolerate soft errors in processor cores," DAC 2016

PI: Pradip Bose (IBM)
Sarita Adve, Vikram Adve, Sasa Misailovic (UIUC)
Luca Carloni, Ken Shepard (Columbia)
David Brooks, Gu-Yeon Wei (Harvard)

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.
Distribution Statement A – Approved for Public Release, Distribution Unlimited

THE ELECTRONICS RESURGENCE INITIATIVE