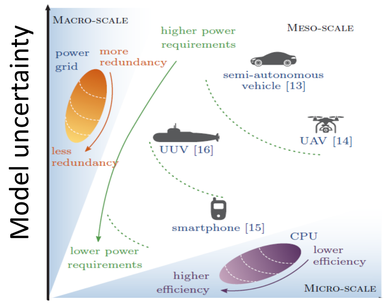
Modular Power Orchestration at the Meso-scale
Washington University in St. Louis

PI: Xuan Zhang; Co-PI: Christpoher Gill {xuan.zhang, cdgill} @wustl.edu
https://xzgroup.wustl.edu/; https://github.com/xz-group

Motivation for Meso-scale Orchestration

- The generation, storage, allocation, and distribution of power, as well as computational resources among the modules demands flexibility
- The semantics of precise resource management and the complexity of high-level tasks are mismatched
- High model uncertainty exists due to diverse sources of variability, especially at run-time

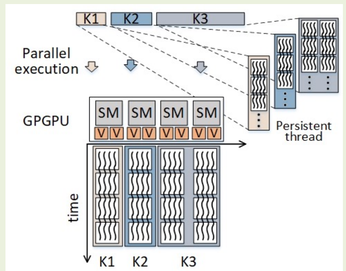


Semantic mismatch

Cross-layer Resource Management

Real-time GPU scheduling

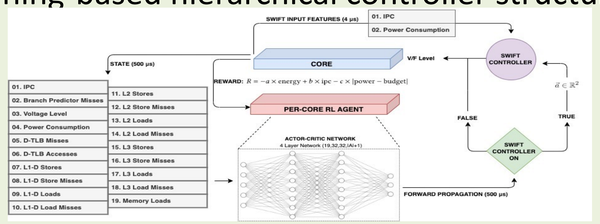
- System-level resource partitioning improves utilization
- OS-level real-time scheduling guarantees hard deadlines



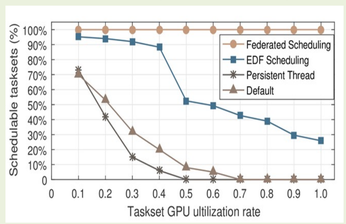
Resource partitioning

Learning-based energy management

- Fast microsecond timescale power management (DVFS)
- Learning-based hierarchical controller structure



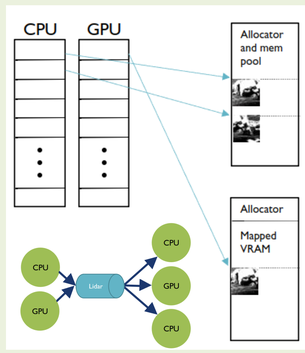
Reinforcement learning model as local controller



GPU schedulability

Heterogeneity-aware zero-copy memory management

- Minimizes memory bandwidth in pub/sub applications
- Data copied to different memory domains as-needed
- Prevents components from incurring memory overhead of hardware-accelerated computing
- Message queue to store passed messages with duplicate copies for relevant memory domains
- Custom allocator interface with new copy semantics



Research Approach

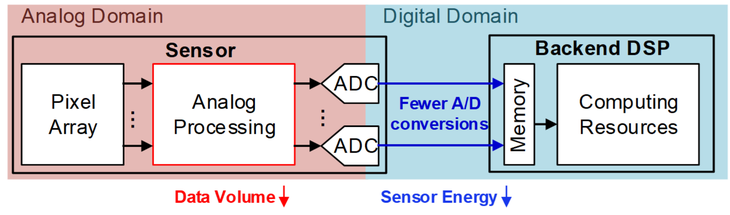
A bottom-up approach to developing resource (power/compute) resource management techniques for meso-scale systems across the system layers:

- T1: Efficient circuit-and-architecture-level power delivery
- T2: Architecture-and-operating-system-level real-time resource scheduling
- T3: Principled investigation and evaluation of mission-level CPS performance

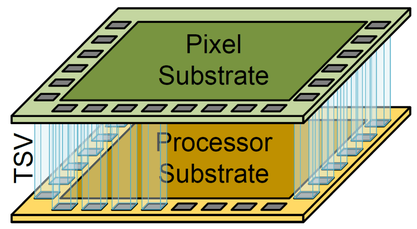
(T1 tasks have been presented in prior years; we focus on T2 and T3 tasks here)

Hardware Acceleration for Autonomous Systems

- Near-memory graph processing with communication-aware memory partition to reduce the latency and energy consumption
- In-sensor processing with analog-domain circuits to enable advanced machine perception tasks in resource- and energy-constrained edge devices
- Aggressive data compression can be realized by co-designing the in-sensor analog processing with the downstream vision algorithms.



In-sensor analog processing



RRAM-Based 3D-Stacked Image Sensor

Related Publications

- Boloor, A., et. al. (2020). "Attacking vision-based perception in end-to-end autonomous driving models". *J. of Sys. Arch.*
- Boloor, A., et. al. (2021) "Can Optical Trojans Assist Adversarial Perturbations?" ICCV.
- Yang, J., et. al. (2022). "Finding Physical Adversarial Examples for Autonomous Driving with Fast and Differentiable Image Compositing". *arXiv preprint*.
- Zou, A., et. al. (2022). "RTGPU: Real-Time GPU Scheduling of Hard Deadline Parallel Tasks with Fine-Grain Utilization". *arXiv preprint*.
- Zou, An, et al. (2022) "F-LEMMA: Fast learning-based energy management for multi-/many-core processors." TCAD.
- Zou, A., et. al. (2021). "System-Level Early-Stage Modeling and Evaluation of IVR-assisted Processor Power Delivery System". TACO.
- Ma, T., et. al. (2022). "HOGEye: Neural Approximation of HOG Feature Extraction in RRAM-Based 3D-Stacked Image Sensors". ISLPED, Best Paper Award
- Zhu, H., et. al. (2022). "PDNPulse: Sensing PCB Anomaly with the Intrinsic Power Delivery Network". arXiv preprint.
- Zhu, H., et. Al. (2022). "PowerScout: Security-Oriented Power Delivery Network Modeling for Side-Channel Vulnerability Analysis" IEEE TETC.

2022 NSF Cyber-Physical Systems Principal Investigators' Meeting
November 7-9, 2022

Award ID#: NSF CNS #1739643