Architectural Support for Emerging Workloads

George Tzimpragos
Szilagyi Energy Breakthrough Fellowship
Single-Core Era

Enabled by:
✓ Moore’s Law
✓ Voltage Scaling

Constrained by:
× Power
× Complexity

Multi-Core Era

Enabled by:
✓ Moore’s Law
✓ Voltage Scaling
✓ SMP architecture

Constrained by:
× Power
× Parallel SW
× Scalability

Heterogeneous Systems Era

Enabled by:
✓ Data Parallelism
✓ Power efficient accelerators

Constrained by:
× Programming models
× Integration
× Comm. overhead
Overview of My Research

Productivity tools

Accelerator-rich systems

Emerging workloads and novel computing approaches

Stack of a Computing Problem

- **Problems**
  - Algorithms
  - Programming Languages
  - Runtime and Compilers
  - ISA
- **System Architecture**
  - Implementation
  - MicroArchitecture
  - Logic and Circuits
  - Transistors
  - Manufacturing

Heterogeneous Systems Era

Enabled by:
- ✓ Data Parallelism
- ✓ Power efficient accelerators

Temporarily Constrained by:
- ✗ Programming models
- ✗ Integration
- ✗ Comm. overhead

[Source: HHS Lee, Georgia Tech]
Productivity Tools

CHARM:
- Python-based DSL.
- Unified basis for the representation, execution, and optimization of closed-form high-level architecture models.

PyRTL: design HW pythonically

- Scripts
- Testbench
- HW design

CHARM: Python-based DSL.
Unified basis for the representation, execution, and optimization of closed-form high-level architecture models.

Accelerator-rich Systems

**Hardware challenges**
- programmability
- code portability
- scalability
- reliability
- OS support

**Software challenges**
- programmability
- code portability
- scalability
- reliability
- OS support

---

**Programmability vs Reusability**
- composable building blocks
- template (meta)programming

---

**App. Domain Expertise**
**Separation of Concerns**
**System & HW Expertise**

---

**Snapshots & Process Migration**

- Read and write “process” state
  - Identify possible switching states with minimum cost.
  - Use existing debug and fault-injection infrastructure to read and write circuit state.

---

A world of sensors and machine learning!

“The A11 Bionic neural engine is designed for specific machine learning algorithms and enables Face ID, Animoji and other features.”

Dynamic Vision Sensor

Time-of-Flight Sensing
**Novel Computing Approaches: “Race Logic”**

**Where does power go in CMOS?**

\[ P_{total} = P_{dynamic} + P_{static} \]
\[ P_{dynamic} = P_{switching} + P_{shortcircuit} \]
\[ P_{static} = (I_{sub} + I_{gate} + I_{junction} + I_{contention}) V_{dd} \]

**Ways to reduce power consumption**

- Voltage supply
- Load capacitance
- Switching activity
- Clock frequency

**Questions**

- What happens with the performance of “big data” applications?
- Can “Race Logic” lead to ultra-low-power solutions?

**Race Logic**

- What if “delay” could be used for computation?
- Can we encode data as “delay”?

(a) min  
\[ \min(x, y) \]

(b) max  
\[ \max(x, y) \]

(c) add-by-constant  
\[ \delta(x, k) \]

(d) inhibit: \[ inh(i,j) \]

\[ \delta(x, 1) \]
\[ inh(x, y) \]
\[ inh(y, x) \]
Boosted Race Trees for Low-Energy Classification

**User-defined Parameters**
- **Machine Learning:**
  - classification method
  - learning rate
  - # of estimators
  - # max depth
  - etc.
- **Hardware Specs:**
  - bits per input
  - memory cell bits
  - technology node
  - routing architecture
  - etc.

**ML Library**
- Abstract and reform
- Analytics & Preprocessing:
  - features importance
  - votes quantization
  - etc.

**Parser**
- Map to Hardware
- HW Template:
  - “glue” logic
  - memory gen.
  - etc.
- RTL Library:
  - energy/area estimation
  - Design Space Exploration

**Map to Hardware**

**High-level Arch. Modeling**

**Parser**

**ML Library**

**User-defined Parameters**

**Hardware Specs:**

**Boosted Race Trees**

**Map to Hardware**

**High-level Arch. Modeling**

**Parser**

**ML Library**

**User-defined Parameters**

**Hardware Specs:**

**Boosted Race Trees**

**Map to Hardware**

**High-level Arch. Modeling**

**Parser**

**ML Library**

**User-defined Parameters**

**Hardware Specs:**
Boosted Race Trees for Low-Energy Classification
