SPARTA Toward A Stream-Based Processor And RunTime Architecture

University of Delaware | The University of Arizona | UCI University of California, Irvine

Our objectives

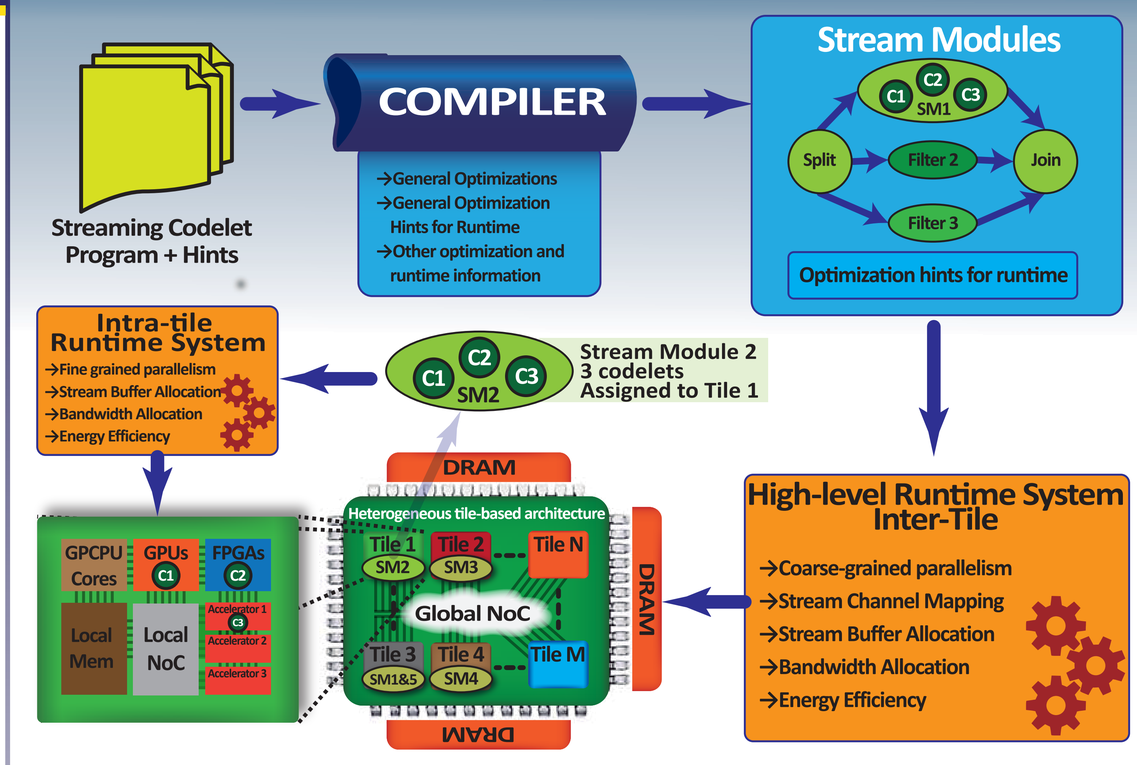
- Efficiently exploiting parallelism and delivering scalability in stream processing
- Address challenges in:
  - Programmability
  - Performance superiority
  - Energy efficiency
  - Overall scalability

Challenges and oportunities

- Exploit two levels of parallelism:
  - Coarse-grain parallelism
  - Fine-grain parallelism
- Heterogeneity of
  - Computation load. E.g. the MPEG decoding.
  - Computation type. E.g. vectorizable vs. scalar.
- Maximize locality and minimize data movement:
  - Continous streaming through the dataflow graph (FIFO)
  - Not streaming. Use shared memory hierarchy and exploit data locality

Application Domain of Interests

- Scientific Computations
  - Traditional computation intensive application
- Embedded Applications
  - Image/video processing
  - encoding/decoding
- Social-Media Processing
  - Graph Computing

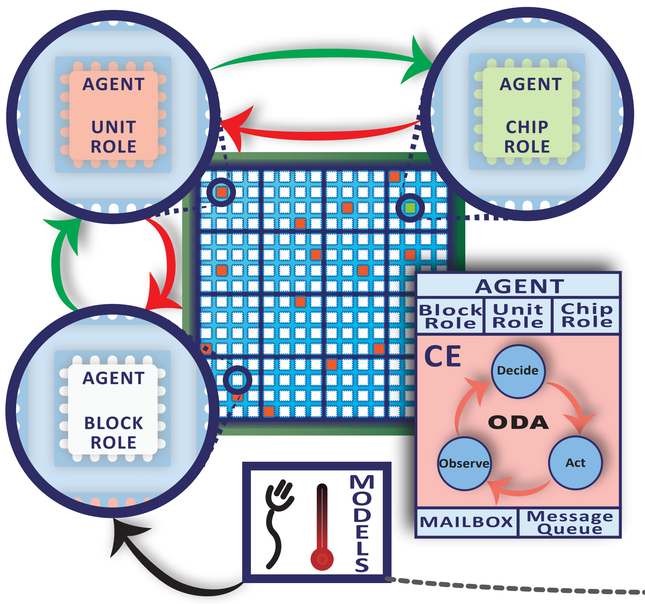


System Software Management Tasks

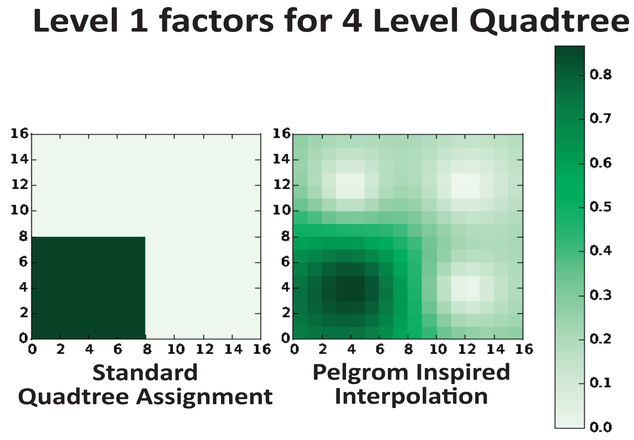
- Multigrain Parallelism
- Memory
- Power/Energy usage
- Network traffic
- Hierarchical NoC Bandwidth Usage
- Compiler-directed Asyncrony

Solution Methodology

- SPARTA Program Execution Model
- SPARTA Architecture Model
- SPARTA System Software Stack
- Co-design: Architecture and software



Level 1 factors for 4 Level Quadtree

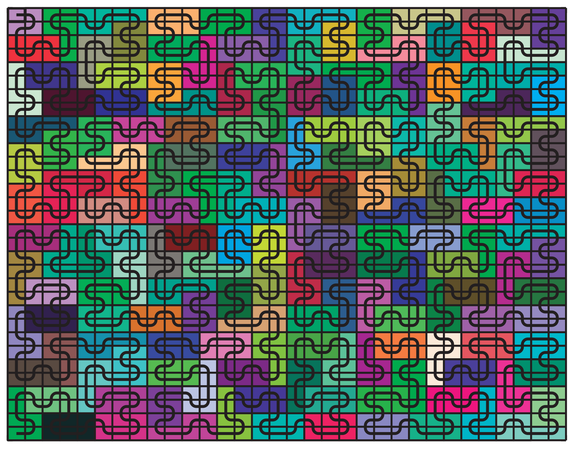


Power Model

$$P_{Leak}(x) = V_{DD} * \alpha * T^2 * e^{K_P}$$

$$K_P = \frac{\beta * V_{DD} - \lambda + x * Var_{Sys} * \lambda}{T} + \frac{(\lambda Var_{Rand})^2}{2 * T^2}$$

Tiling scheme using Hilbert Curves



Self-Aware Framework for Extremescale

- Fast parallel simulator of a simplified hierarchical organization of the SPARTA chip
- SAFE: Evaluate tread-off between performance, energy and power use and system goals
- Temperature and energy models based on instructions' energy and frequency of operation (Synthetic Loads)
- SAFE produce temperature maps, energy consumption and state changes logs

Variance Map

- Variance in the production process on NTV technology has huge effect in the frequency of operation
- Interpolate weight facturs using Pelgrom Model
- Once the variance is calculated off-line, it can be use as an input parameter for SAFE
- This model can be extended with hints or looser constraints to increase data locality

Pseudo-Hilbert Matrix Multiply

- Case of study. Dynamic tiling for matrix multiplication based on system state and core performance
- Accumulated lower level memory is higher than the higher level. It is necessary to improve prefetching of data into the memory
- Additionally, the high variance of the frequency when operating on NTV does not allow static optimizations of memory (since performance of the core is not guaranteed, but can be modeled)

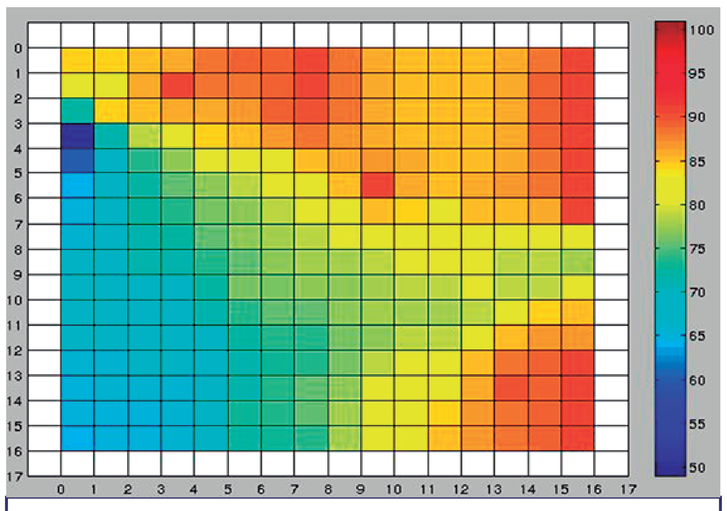
Results and publications

Preliminary Results

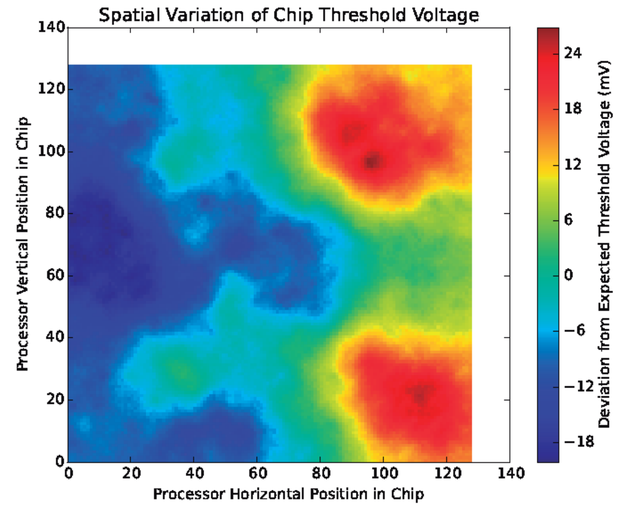
1. Accurate and precise way to model energy, heat, and power.
2. Design and simulate a self-aware behavior in the system software of the SPARTA chip.
3. Provide a path to allow a high-level, coarse-grain parallel control-flow driven programming model down to a fine-grain event-driven execution model
4. Design network-on-chip architectures using cognitive techniques with the main of scaling the overall architecture to a large number of cores and automating the resource optimization.

Some Publications

1. "A Holistic Dataflow-Inspired System Design," (DFM2014). Stéphane Zuckerman et al.
2. "Toward a Self-Aware Codelet Execution Model". (DFM2014). Stéphane Zuckerman et al.
3. PhD Thesis in preparation "Towards Co-Designed Energy Efficient Computing and Runtimes: Simulation Framework and Experiments" Kelly Livingston



Temperature map of the chip generated with SAFE. Each squeare represents a single tile. Red means 100 ºC and blue means 50 ºC. The syntetic load was designed with excessive memory operations to provide high energy consumption and show the behavior of the temperature model.



According to the variance in the production process modeled through the use of Pelgrom Model, it is possible to obtain a realistic Threshold Voltage variation of the chip. System hardware must have ability to efficiently provide introspect and tune operating points to reduce runtime overhead

This material is based upon work supported by National Science Fundation

CCF-1439097 CCF-1439142 CCF-1439165