**Parallel Computing of Graph-based Functions in Re-RAM**

CMOS approaches its physical boundaries due to the continuous shrinking of feature size which leads to the search for promising emerging technologies beyond scaling limit. Resistive Random Access Memory (ReRAM) is a non-volatile memory technology and features low power consumption, has inherent computing capabilities with efficiency of logic synthesis.

Binary Decision Diagram (BDD) approach utilizes Multiply-Accumulate (MAC) operation instead of logic primitives. It directly maps the BDD nodes onto parallel MAC operations.

And-Inverter Graph (AIG) is an automated compiling procedure based on in-memory computer architecture. It translates arbitrary Boolean functions.

In Graph-based representations the edges represent wires between two-input AND gates that correspond to the nodes which can be complemented to represent inverters between the nodes.

ReRAM MAC Computation performs multiple MAC operations in parallel.

In Graph Based Computation MIGs need the smallest number of operations and devices and have become the state-of-the art graph structure for ReRAM-based synthesis.

In parallelism a computation is a wordline parallel if it uses one wordline and multiple bitlines, a bitline parallel if it uses one bitline and multiple wordlines and a mixed parallel if it uses multiple wordlines and multiple bitlines.

In BDD-based Parallel Computation every node must be realized as a 2x1 multiplexer.

In AIG-based Parallel Computation, all children of both nodes must be computed. There must not be any data dependencies between the nodes, and they must share a wordline operand, while host devices are placed in the same wordline, and the content must not be necessary for any other computations.

In M-AIG-based Parallel Computation each node represents an m-Input And Gate or a PI, each input edge can either be connected to the constant 1 or to a child node and if the input edge is connected to a node, then the edge may be complemented to indicate inversion.

Proposed approach significantly reduces the number of devices and operations needed. It reduces the number of operations by about 66% on average. BDD and AIG outperform in both area and operation. For small values of m, m-AIG has outperformed AIG.

**Power Aware Computing**

Processor designs are getting effected due to Power, energy, and temperature. Stagnation of CPU clock frequencies and reliance on parallelism increase energy efficiencies of future. Software design can bring huge improvements in addition to the hardware. Ability of measuring the power and energy consumption is mandatory step. PAPI library was used in this analysis that give generic and portable interface to hardware counters attached to CPU and other components. Experiments were executed on Intel Xeon Phi Knights Landing (KNL) architecture. Analysis was performed using Dense Linear Algebra (DLA) Kernels (BLAS kernels).

Kernels that can be found in high performance computing application were chosen to study and analyze the effect of application of power consumption and energy requirements. BLAS operations are categorized in three levels: level 1 addresses scalar and vector operations, level 2 addresses matrix-vector operations and level 3 addresses matrix-matrix operations. Level 1 and Level 2 belongs to Memory Bond Class, whereas Level 3 belongs to Compute Intensive Class.

Presented the study and to analyze the compute intensive routine dgemm and memory bound class dgemv.

PAPI library provides a coherent methodology to performance counter information varied for different hardware and software component. PAPI offers many components which enables to monitor power usage and consumption of energy through different interfaces.

Study of dgemm Kernel behavior shows that the FLAT mode does not use the MCDRAM as a cache, but as physical addressable memory space.

Study of dgemv Kernel behavior shows that the performance drops dramatically between the two storages and results are observed same as of Hybrid Mode except for DDR4. for which performance is 4 times slower as compared to MCDRAM.

This study shows that using high bandwidth MCDRAM on KNL is important for achieving high performances and minimal consumption of power and if the application is compute intensive than Hybrid mode is best.

**Temperature-Aware Computer Systems Opportunities and Challenges**

Power-aware design alone has failed to stem the tide of dealing with problems like heat density. Localized heating occurs much faster than chip-wide heating. Typical high-power applications still operate at 20 percent or more below the absolute worst case.

The need for Architecture-Level Thermal Management arises from the fact that architecture domain of the computing system is unique, and development of workload is required to control instruction level parallelism. Design manual provides hot spots and temperature gradients for computing system. In this scenario role of system-architecture and operating-system is important.

The need for architecture-level thermal modeling is there to avoid temporal and spatial nonuniformities in computing due to thermal effects.

The compact model of a parametric microarchitecture for a computing system must track temperatures at the granularity of individual microarchitectural units, must be parameterized in so that a new compact model for different microarchitectures, must be able to solve the RC circuit’s differential equations quickly and must be boundary and initial-condition independent.

For Temperature-tracking Dynamic Frequency Scaling, temperature versus average power density for gcc with a power averaging interval of 0.033 seconds is presented.

In Temperature-tracking Dynamic Frequency Scaling, the temperature dependence of carrier mobility in CMOS means that frequency is also linearly dependent on the operating temperature. When applications exceed the temperature specification, they can simply scale frequency down in response to the rising temperature.

Dynamic Voltage Scaling is a solution for thermal management. When changing the processor voltage, a processor must reduce frequency in conjunction with voltage, because circuits switch more slowly as the operating voltage approaches the threshold voltage.

From the results, Migrating Computation is the best DTM technique at 0.8 K/W because the floorplan used by itself is enough to reduce the operating temperature of the primary integer register file, MC can exploit ILP to hide the extra latency of the spare register file and complete elimination of activity in the primary register file allows it to cool quickly, minimizing the use of the slower secondary register file.