MODELING AND ENERGY OPTIMIZATION OF MULTICORE PROCESSOR-CACHE SYSTEMS WITH EMERGING TECHNOLOGIES

By
Divya Praneetha Ravipati
(2017CSZ8061)

Venue: SIT 001 (Mathur Seminar Hall)

11AM, 19 July 2024

Technology advancements and shrinking transistor feature size have improved system performance and efficiency. However, with the transistor feature size shrinkage racing towards sub-nm regime, a multitude of challenges such as slowdown of Moore's law, halt of Dennard scaling, significant increase in leakage power consumption, and quantum effects have emerged, particularly at smaller geometries. To address these challenges, researchers and engineers are exploring new transistors and technologies. In this thesis, we explore two such emerging technologies, negative capacitance FinFET (NC-FinFET/ NCFET) and cryogenic computing, at system-level. Researchers have investigated the advantages of NCFET and cryogenic transistors at device-level and circuit-level. The research findings revealed that the unique characteristics of NCFET offer higher frequency of operation without the need to increase voltage along with reducing the leakage current. Similarly, the examination of cryogenic transistors at device- and circuit-levels has unveiled their potential to enhance system performance and mitigate leakage power concerns through operation at low temperatures. Both of these technologies address the challenge of rising leakage current in smaller geometries while simultaneously enhancing performance without the need for voltage increment. Recognizing the potential advantages of these two technologies, we make an attempt to bridge the research gap by exploring their system-level implications.

To comprehensively explore the trade-offs between performance and energy efficiency in CPU-memory systems at smaller geometries and with new technologies, it is essential to revise the processor and cache models used by instruction-level simulators. CACTI and McPAT are popular tools for system-level architectural studies to estimate power, performance and area of the system. However, the tools are primarily designed for CMOS technology using the data obtained from various projections. Furthermore, the models use various approximations at higher geometries which do not align with the current trends for the FinFET technology at lower geometries.

For the first time, we make an effort to revise the CACTI and McPAT models to suit the newer technologies, while respecting the overall modeling methodology of the tools for FinFET-/NCFET-based system estimates. Moreover, for the first time, we integrate the transistor characteristics from a 14 nanometer commercial FinFET technology within the tools. First, we revise CACTI models to support FinFET and NCFET technologies. We use the developed tool (**FN-CACTI**) to assess the energy efficiency of NCFET-based caches compared to FinFET-based caches. Additionally, we identify the optimal voltage to minimize cache energy consumption for FinFET- and NCFET-based caches at various access rates and cache sizes. Our investigations show that the optimal voltage for NCFET-based caches spans a range of voltages depending on the cache access rate, while the optimal voltage is within the lower range of applicable voltages for FinFET-based caches. As the next step towards revising the tools to estimate system-level delay, power and area estimates, we update the McPAT models. We first integrate FN-CACTI with our modeling tool, **FN-McPAT**, to derive estimates for FinFET- and NCFET-based memory structures. Then, we synthesize the BOOM CPU core with FinFET and NCFET technologies to revise the core component models in FN-McPAT. We use FN-CACTI and FN-McPAT to investigate the performance improvements and energy-efficiency of NCFET-based and FinFET-based systems. Our investigations provide novel insights into energy consumption of NCFET-based caches on systems running various workloads.

Motivated by our findings on the new trends in energy consumption behavior of NCFET-based caches, we present the first work towards optimizing energy in NCFET-based caches with minimal impact on performance. We leverage the unique characteristics offered by NCFETs and propose a dynamic voltage scaling policy, **CAPE**. Along with an approach that is suitable for both NCFET- and FinFET-based caches, we also introduce a novel metric to capture cache criticality. Our experimental evaluations indicate that the CAPE policy achieves 19.2% more last-level cache (LLC) energy savings compared to operating at the maximum available voltage.

Cryogenic circuits have applications in fields such as quantum computing, particle detectors, and magnetic resonance imaging. While cryogenic logic circuits are being addressed by the research community, there is limited work on designing with larger cryogenic systems at lower temperatures (10 K). Moreover, there is no tool to estimate delay, power and area of cryogenic systems at lower geometries. As a first step towards modeling cryogenic multi-core systems, we model DeepCryo-CACTI (**DC-CACTI**) for cryogenic caches due to their vital role in performance improvement and significant contribution to both area and power of the processor. Using DC-CACTI, we provide our insights on efficiency of cryogenic caches and propose new research directions.