Title - Application-aware Dynamic Thermal Management in 3D DRAM

Author - Shailja Pandey (2016CSZ8117)

Date/Time - July 23, Tuesday, 4 PM.

Abstract

3D DRAM is a promising technology to meet the bandwidth demands of data-hungry ML/AI applications and is being actively used in a wide range of hardware platforms such as GPUs, FPGAs, DNN accelerators, and general-purpose cores. Despite the potential of 3D DRAM to bridge the memory bandwidth gaps, the vertically stacked integration of DRAM dies brings a new set of system design challenges. One such challenge is the poor heat dissipation capability, causing excessive memory heating and eventually throttling the memory, slowing down the applications.

Heating in 3D DRAM, due to high power density and close proximity of memory cells in stacked architectures, impacts the data retention capabilities, increasing the refresh rates significantly. 3D stacking enables hundreds of banks/ranks and several independent channels per memory device such that each channel is connected to a separate memory controller; however, the off-chip memory bandwidth is often limited due to limited power budgets that might arise in practical scenarios. Due to this, it is often not possible to activate and run all memory ranks simultaneously, similar to dark silicon in multi-core systems. Systems running under a fixed power budget and a thermal constraint employ dynamic power budgeting (DPB) and dynamic thermal management (DTM) policies, resulting in performance and energy overheads due to thermal-induced CPU and memory throttling.

Existing solutions towards dynamic thermal management are application-agnostic and often lead to sub-optimal performances. Modern applications execute in distinct phases, exhibiting varying memory access behavior and showing different sensitivities to architectural parameters such as last level cache (LLC) size, prefetch settings, memory-level parallelism, and memory bandwidth. The indiscriminate approach of prior works for DTM in 3D DRAM may lead to avoidable performance degradation, as different application phases have varying contributions to memory temperature. To this end, this dissertation proposes application-aware DTM techniques targeted towards minimizing the associated performance penalties. We present novel micro-architecture-based approaches and system-level policies for managing heat in 3D DRAM and facilitate faster cooling, leveraging knowledge of the workload. We focus on deep neural networks (DNNs) in this thesis.

Deep neural network (DNN) implementations are typically characterized by huge data sets and concurrent computation, resulting in a demand for high memory bandwidth due to intensive data movement between compute units and off-chip memory. Performing DNN inference on a general-purpose processor is common, and multi-core CPUs are being used for DNN implementation in servers and commercial SoCs. 3D DRAM is becoming a key enabler for high memory-intensive applications such as DNNs running on general-purpose cores. This dissertation focuses on efficient

thermal management on a system consisting of general-purpose CPU cores and a thermally-constrained 3D DRAM.

Firstly, we present a memory temperature-aware prefetcher that jointly utilizes DNN application's memory behavior and 3D DRAM temperature to minimize DTM overheads. The customized prefetching mechanism dynamically computes the best prefetch settings for different DNN workload phases, utilizing the unused 3D DRAM bandwidth without adversely affecting the workload's thermal footprint.

Secondly, we present an efficient task mapping based DTM policy for multi-core processors to minimize memory throttling through proactive management of the potential thermal hotspots in 3D DRAM using application-aware optimizations in processing cores. An application-aware DVFS is used to reduce the thermal impact of the application’s phase when task mapping to an ideal core is not feasible.

Thirdly, we present a thermal and memory access pattern-aware data mapping based DTM policy to map a DNN layer’s data to 3D DRAM pseudo-channels. An intelligent mapping of the DNN application’s data across memory channels, pseudo-channels, and banks, leveraging the knowledge about each layer’s memory access pattern and date reuse, significantly reduces the effective memory latency and the thermal-induced application slowdown.

Finally, we present a reward-based and adjacency-aware dynamic power budgeting which periodically suggests the ideal set of channels that should be enabled under a given power budget and thermal constraint, performing a coordinated thermal and power management for 3D DRAM. The thermal characteristics of 3D DRAM make power budgeting a non-trivial task that requires run-time intervention to minimize the associated performance penalties.

To evaluate the proposed policies, an integrated performance-thermal simulation framework is used. Our results demonstrate significant DNN inference time and memory energy reductions over the state-of-the-art, with minimal run-time cost. This dissertation establishes the significance of leveraging application knowledge towards efficient thermal management in 3D DRAM for a class of well-structured workloads such as deep neural networks.

Teams Meeting Link

https://teams.microsoft.com/l/meetup-join/19%3ameeting_M2I2Mjc2MmYtN2IzZC00MjY5LWE2OTMtYjY2ZDNlY2ExMWJm%40thread.v2/0?context=%7b%22Tid%22%3a%22624d5c4b-45c5-4122-8cd0-44f0f84e945d%22%2c%22Oid%22%3a%2231dcd88c-777a-4c06-80be-1b1812075633%22%7d