COA

LAB ASSIGNMENT

Group 5

Team members:

1. Meesala Aakanksha – 21CS01056

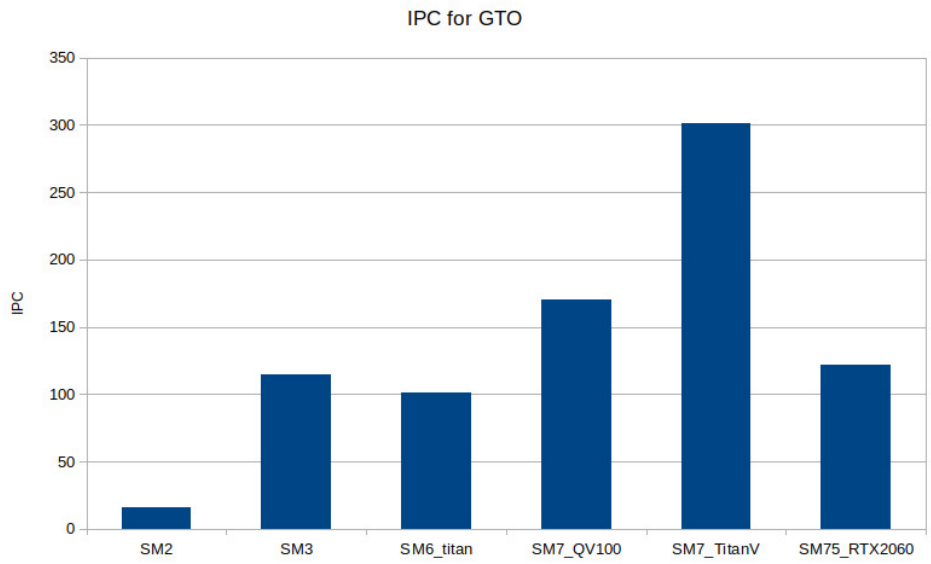
2. Prisha Srinidi – 21CS01057

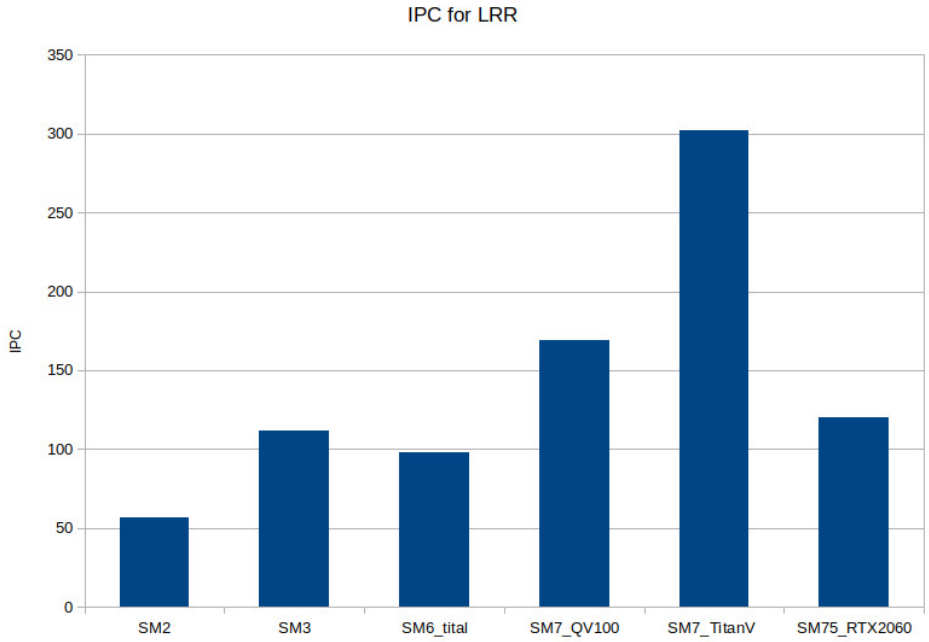
3. Bharati Pradhan – 21CS01009

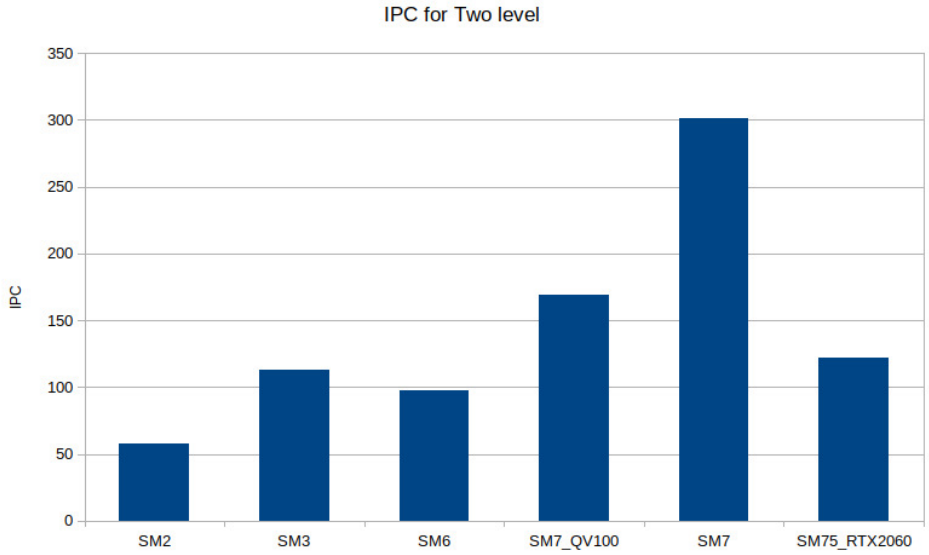
4. KSS Amrutha – 21CS02005

5.Meruva Yashna – 21CS01039

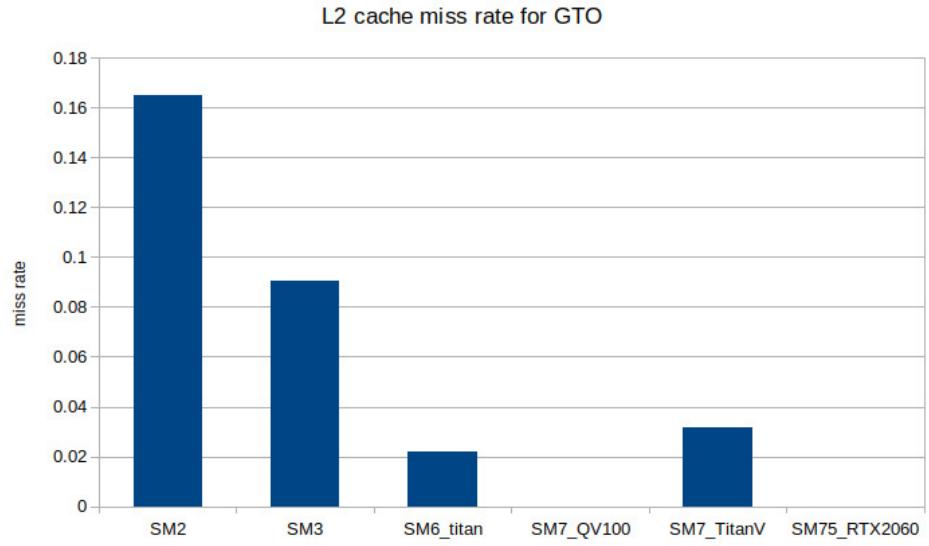
Q 1.)

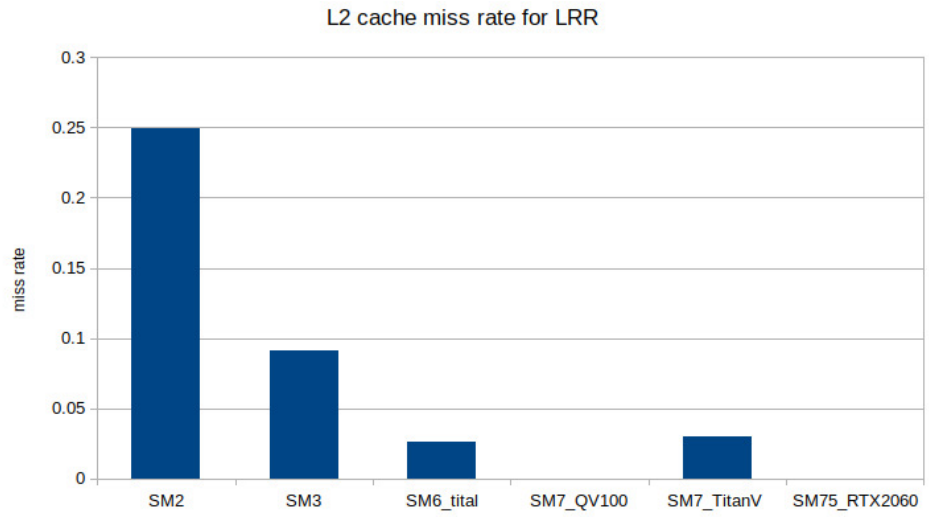


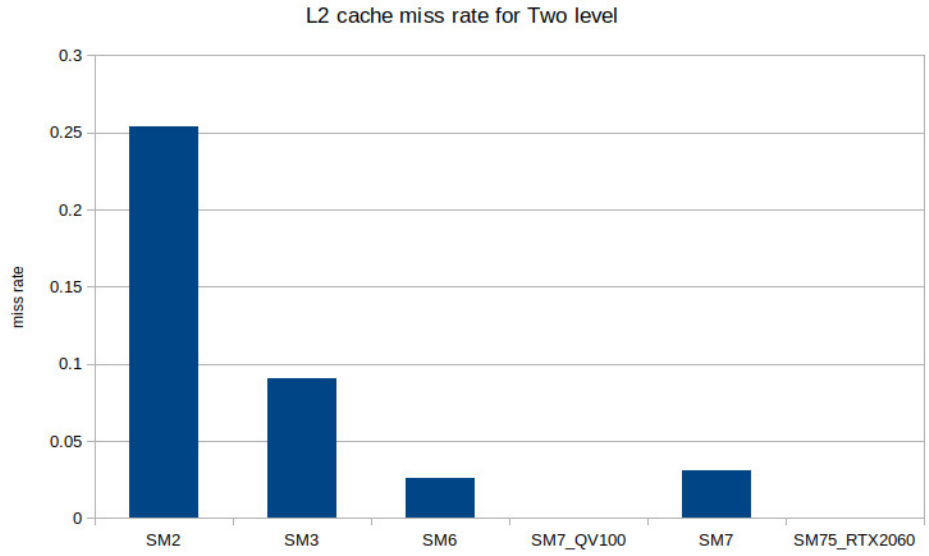

IPC for LRR
IPC
350
300
250
200
150
100
50
0
SM2
SM3
SM6_tital
SM7_QV100
SM7_TitanV
SM75_RTX2060

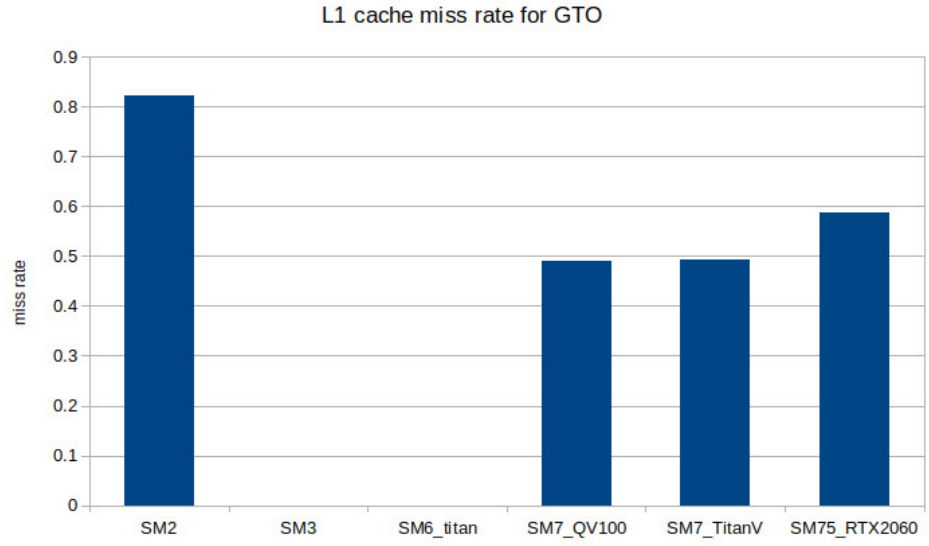

IPC for Two level
IPC
350
300
250
200
150
100
50
0
SM2
SM3
SM6
SM7_QV100
SM7
SM75_RTX2060

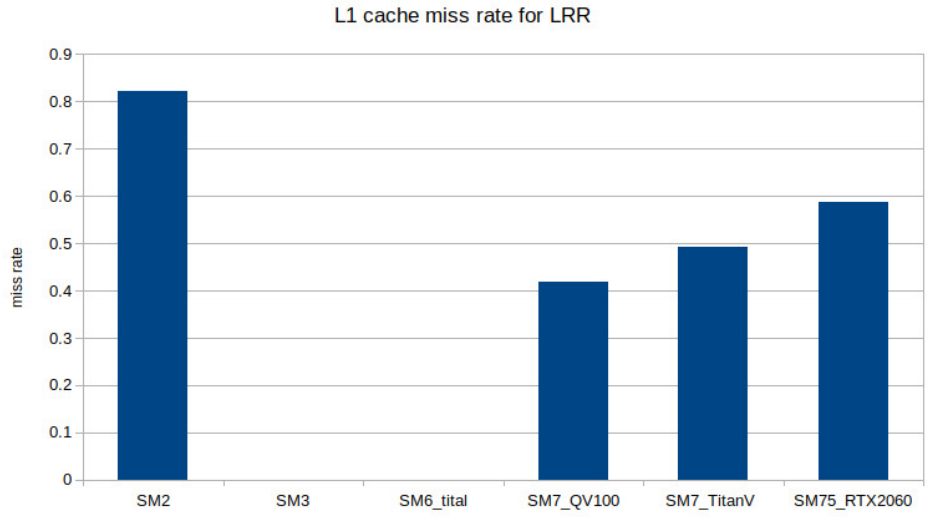
Q 2.)

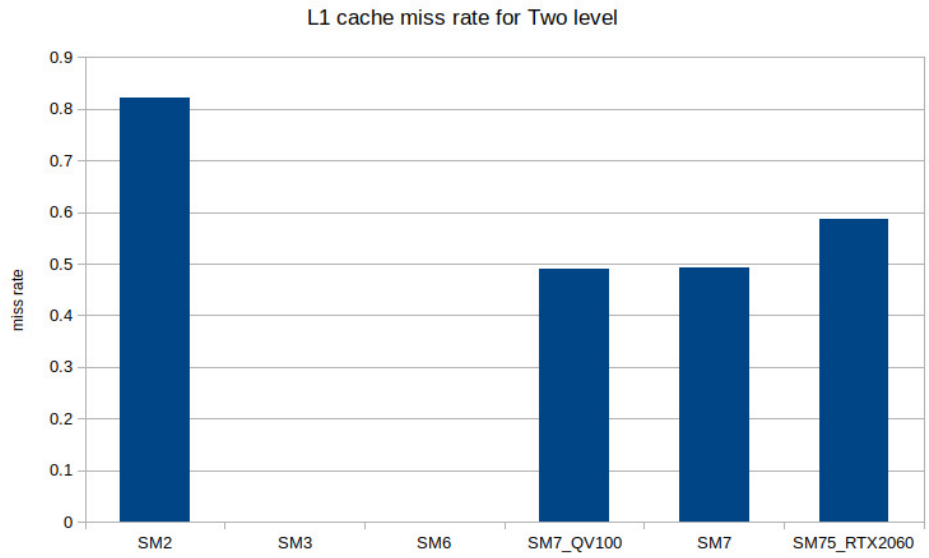

L2 cache miss rate for GTO
miss rate
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
SM2
SM3
SM6_titan
SM7_QV100
SM7_TitanV
SM75_RTX2060


L2 cache miss rate for LRR
miss rate
0.3
0.25
0.2
0.15
0.1
0.05
0
SM2
SM3
SM6_tital
SM7_QV100
SM7_TitanV
SM75_RTX2060


L2 cache miss rate for Two level
miss rate
0.3
0.25
0.2
0.15
0.1
0.05
0
SM2
SM3
SM6
SM7_QV100
SM7
SM75_RTX2060


L1 cache miss rate for GTO
miss rate
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SM2
SM3
SM6_titan
SM7_QV100
SM7_TitanV
SM75_RTX2060





Q 3.)

GTO WARP SCHEDULER	L2 HIT RATE	L1D HIT RATE
SM2_GTX480	0.52	0.67
SM3_KEPLER_TITAN	0.91	1
SM6_TITANX	0.98	1
SM7_QV100	1	0.51
SM7_TITANV	1	0.58
SM75_RTX2060	1	0.58

The utilization of General-Purpose Graphics Processing Units (GPGPUs) in computing systems has gained significant attention in recent years. One crucial parameter for GPGPU performance is cache hit rates, which greatly influence overall system efficiency.

The cache sizes within the GPU configuration file are delineated in the subsequent format. The <cache_type> parameter serves to specify the cache type that is being configured. Examples of cache types include dl1, dl2, il1, tex_cache, and const_cache.

The cache configuration parameters, denoted as <configuration_parameters>, encompass various aspects of the cache system. These parameters include <nsets>, <bsize>, <assoc>, <rep>, <wr>, <alloc>, <wr_alloc>, <set_index_fn>, <mshr>, <N>, <merge>, <mq>, and <fifo_entry>.

To modify the cache size, it is customary to adjust the parameters denoted as <nsets>, <bsize>, and <assoc>. The parameters serve to delineate the quantity of sets, the size of each block, and the level of associativity within the cache, correspondingly.

In the context of the overarching configuration file, as an illustrative instance, the topic of interest is the GPGPU cache, specifically the DL1 cache. The user has provided a set of data points, specifically N:32:128:4, L.

The variable N represents the number of sets, specifically 32 sets. The block size, denoted as 128, refers to the allocation of memory in units of 128 bytes. The topic under discussion is the concept of associativity, specifically in the context of a 4-way system.

In the present scenario, the cache is structured into 32 sets, with each set accommodating cache lines, or slots, equivalent to the designated associativity value of 4. This implies that within every given set, there exists a total of four cache lines that can be utilized for the purpose of storing data that has been retrieved from the main memory. Upon the loading of a memory block into the cache, it proceeds to occupy a cache line within the set that corresponds to it.

Enhanced associativity typically enhances cache hit rates by providing a larger number of slots to accommodate frequently accessed memory blocks. Nevertheless, it is important to note that a higher degree of associativity is accompanied by heightened intricacy and the possibility of experiencing performance drawbacks. A decrease in associativity has the potential to result in an increased occurrence of cache conflicts and a subsequent decrease in cache hit rates.

To modify the cache size, we endeavored to manipulate the parameters of the Number of Sets and Block Size. However, this adjustment resulted in the occurrence of an error. Therefore, the Associativity was modified in such a manner that the resultant value of the product of the Number of Sets, Block Size, and Associativity equates to 2^23 bytes, or 8 megabytes.

The categorization of L1D cache hit rates can provide insights into the performance behavior of various GPU configurations:

1. Several configurations, such as SM2_GTX480 (gto, lrr, two_level), SM7_QV100 (gto, lrr, two_level), and SM7_TITANV (gto, lrr, two_level), demonstrate L1D cache hit rates that approach or surpass 0.8.

2. The SM75_RTX2060 (gto, lrr, two_level) configuration exhibits moderate L1D cache hit rates, which fall within the range of 0.5 to 0.6. Configurations exhibiting a diminished L1D cache hit rate, approximately 0.2, exemplify this category. Noteworthy instances include SM3_KEPLER_TITAN (gto, lrr, two_level) and SM6_TITANX (gto, lrr, two_level).

3. The phenomenon of low L2 cache hit rates is of particular interest, as certain configurations have been observed to achieve a complete absence of L2 cache misses. This observation presents a captivating aspect to the comprehension of L2 cache dynamics. The present study aims to categorize configurations by analyzing their L2 cache hit rates.

The configurations, namely SM2_GTX480, SM7_QV100, SM7_TITANV, and SM75_RTX2060, consistently exhibit a commendable L2 cache hit rate nearing unity, irrespective of the employed warp scheduler. The SM6_TITANX architecture exhibits a moderate L2 cache hit rate across all warp schedulers.

Q 4.)

GTO WARP SCHEDULER	Execution on units avg power	DRAM avg power	Register files avg power	Total avg power	% Execution	% DRAM	% Register files
SM2_GTX480	28.28	0.01	66	94.28	29.99	0.01	70.01
SM3_KEPLER_TITAN	72.65	0.09	6.13	78.79	92.22	0.09	7.78
SM6_TITANX	35.94	0.07	36.23	72.17	49.79	0.07	50.21
SM7_QV100	48.4	0.13	72.39	120.78	40.07	0.13	59.93
SM7_TITANV	111.39	0.78	20.78	132.19	84.27	0.78	15.73
SM75_RTX2060	40.51	0.05	20.9	61.41	65.97	0.05	34.03

The primary objective of this study is to investigate the potential advantages associated with achieving higher hit rates in the Level 1 data cache (L1D cache) while concurrently reducing power consumption. In a general context, increased L1D cache hit rates suggest that a larger proportion of memory accesses are effectively served from the cache, thereby diminishing the need to access the comparatively slower main memory. This phenomenon can lead to a reduction in power consumption related to memory operations, as cache access incurs a lower power cost compared to retrieving data from the primary memory.

Conversely, the implications of lower L1D cache hit rates and the possible resultant increase in power consumption should also be considered. When L1D cache hit rates decline, it indicates that a larger portion of memory accesses are resulting in cache misses, necessitating the retrieval of data from the main memory. Such cache misses can lead to extended memory access times and potentially heightened power consumption due to the increased data relocation involved.

The interplay between L1D cache hit rates and power consumption varies depending on the specific workload and the underlying architecture of the CPU or GPU. The diverse memory access patterns exhibited by different applications exert an influence on cache behavior and subsequently impact power consumption.

An important aspect to consider is the trade-off inherent in pursuing maximum cache hit rates. While doing so can lead to decreased power consumption during memory operations, it necessitates the use of larger and more energy-intensive caches. The net effect on power consumption depends on the balance between improvements in cache performance and the energy costs associated with maintaining larger caches.

However, it is crucial to acknowledge that the presence of correlation between cache hit rates and power consumption does not inherently indicate a causal relationship. While an observable correlation between these two variables may exist, it is essential to recognize that this correlation does not imply a direct cause-and-effect connection between them.

To precisely establish the correlation between L1D cache hit rates and power consumption for your specific applications, a comprehensive analysis utilizing empirical data is essential. If you could provide me with data regarding L1D cache hit rates and power consumption for various applications, I would be pleased to assist you in examining and elucidating the correlation between these variables.