ASSIGNMENT-4

Aman Dangi (21CS01027)
Manan Khanna (21CS01028)
Shubham Kaushik (21CS01030)
Soumyabrata Chaudhuri (21CS01032)

Relevant Knowledge

Cache Operation: It is based on the principle of locality of reference. There are two ways with which data or instruction is fetched from main memory and get stored in cache memory. These two ways are the following:

- **Temporal Locality –** Temporal locality means current data or instruction that is being fetched may be needed soon. So we should store that data or instruction in the cache memory so that we can avoid again searching in main memory for the same data. When CPU accesses the current main memory location for reading required data or instruction, it also gets stored in the cache memory which is based on the fact that same data or instruction may be needed in near future. This is known as temporal locality. If some data is referenced, then there is a high probability that it will be referenced again in the near future.
- **Spatial Locality –** Spatial locality means instruction or data near to the current memory location that is being fetched, may be needed soon in the near future. This is slightly different from the temporal locality. Here we are talking about nearly located memory locations while in temporal locality we were talking about the actual memory location that was being fetched.

Cache Performance: The performance of the cache is measured in terms of hit ratio. When CPU refers to memory and find the data or instruction within the Cache Memory, it is known as cache hit. If the desired data or instruction is not found in the cache memory and CPU refers to the main memory to find that data or instruction, it is known as a cache miss.

Multilevel caches are a common technique to reduce the memory access latency and improve the performance of computer systems. However, designing and optimizing multilevel caches involves trade-offs between the hit rate, the miss penalty, and the cost and complexity of the cache hierarchy.

The **hit rate** is the fraction of memory requests that are satisfied by the cache, without accessing the lower-level memory.

A **miss ratio** is the flip side of this where the cache misses are calculated and compared with the total number of content requests that were received.

The **miss penalty** is the additional time required to fetch the data from the lower-level memory when the cache misses.

The hit rate and the miss penalty are inversely related: a higher hit rate means a lower miss penalty, and vice versa. The overall cache performance depends on both the hit rate and the miss penalty, as well as the base access time of the cache.

$$\text{Hit} + \text{Miss} = \text{Total CPU Reference}$$

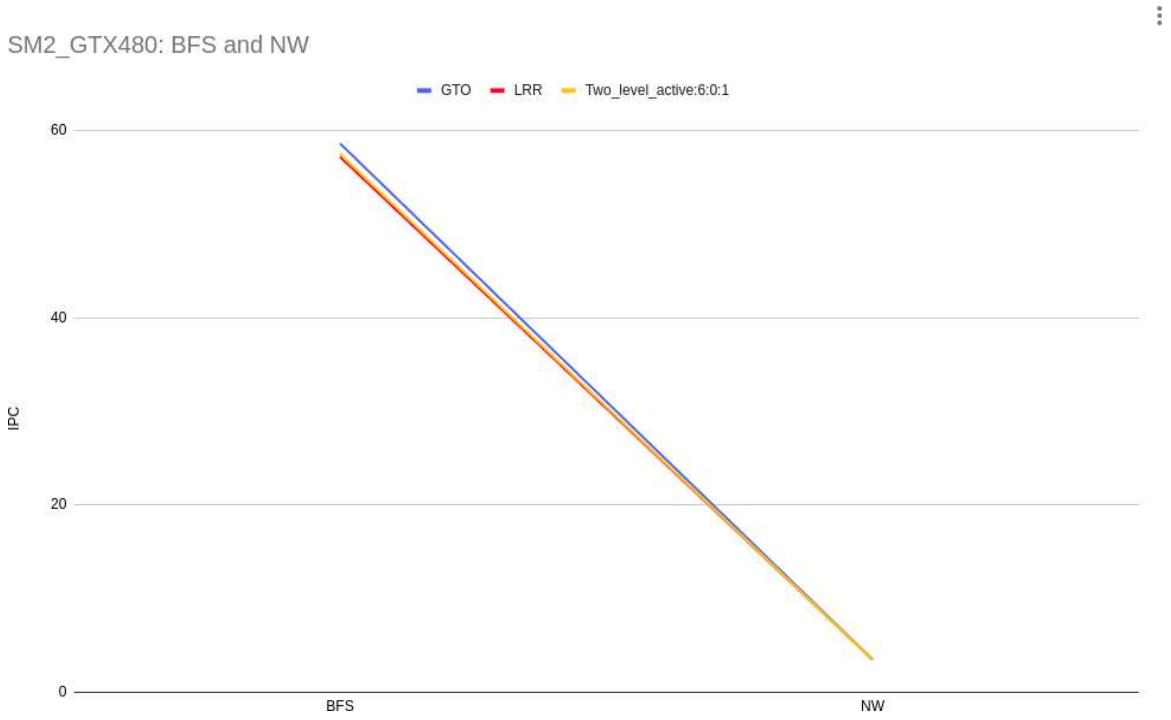
$$\text{Hit Ratio}(h) = \text{Hit} / (\text{Hit} + \text{Miss})$$

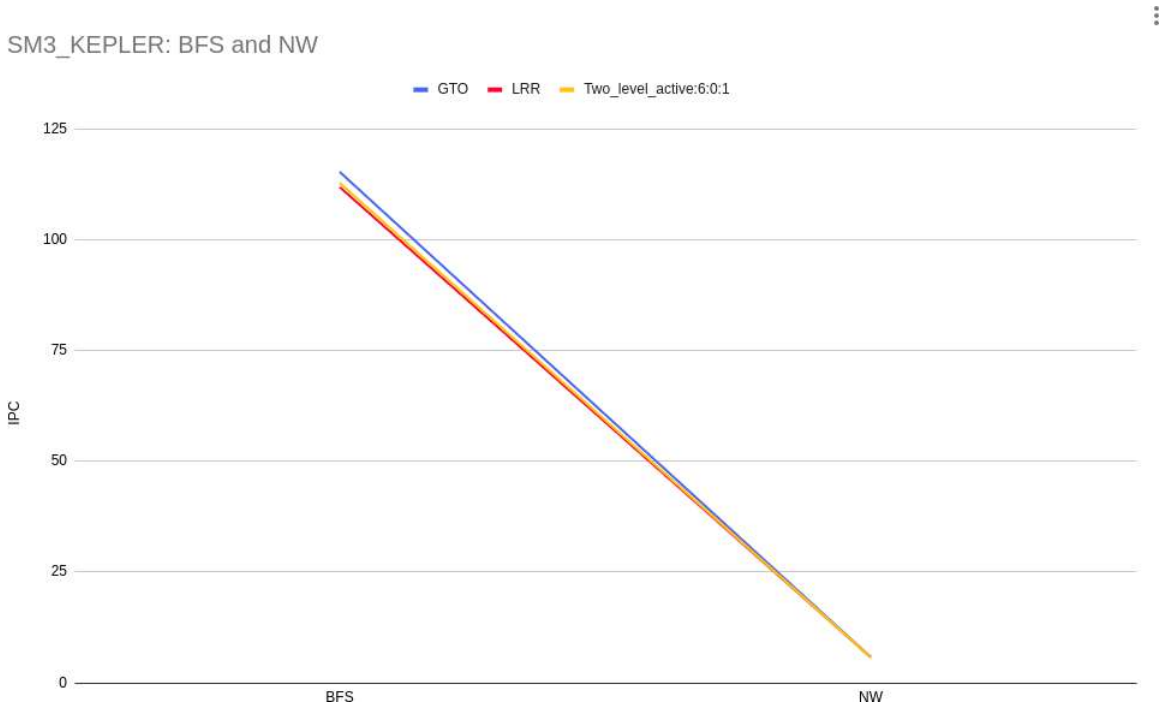
$$\text{Miss Ratio} = 1 - \text{Hit Ratio}(h)$$

$$\text{Miss Ratio} = \text{Miss} / (\text{Hit} + \text{Miss})$$

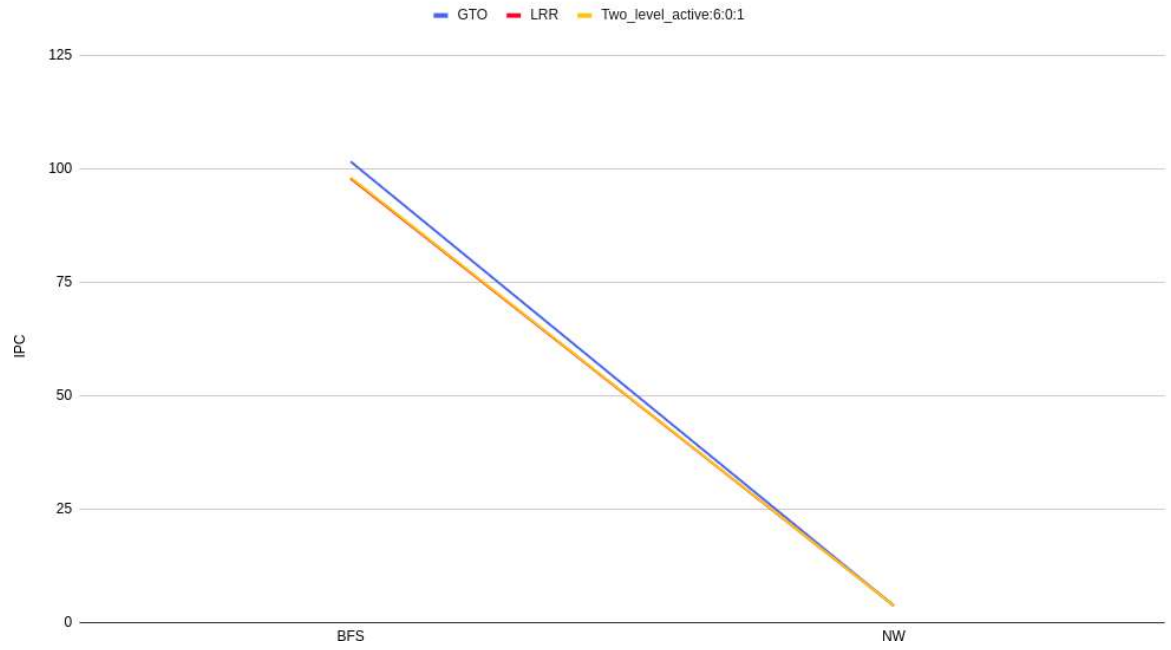
Question 1: Plot of IPC v/s Applications

BFS and NW are two different applications, and GTO/LRR/two_level_active:6:0:1 are three different warp schedulers

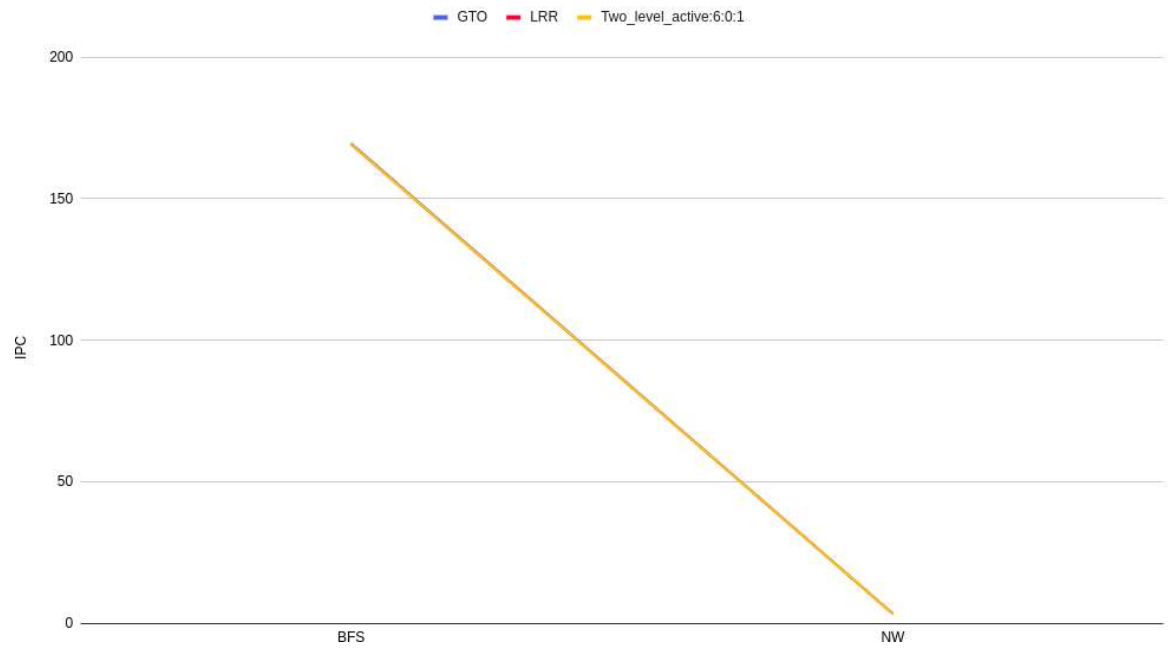


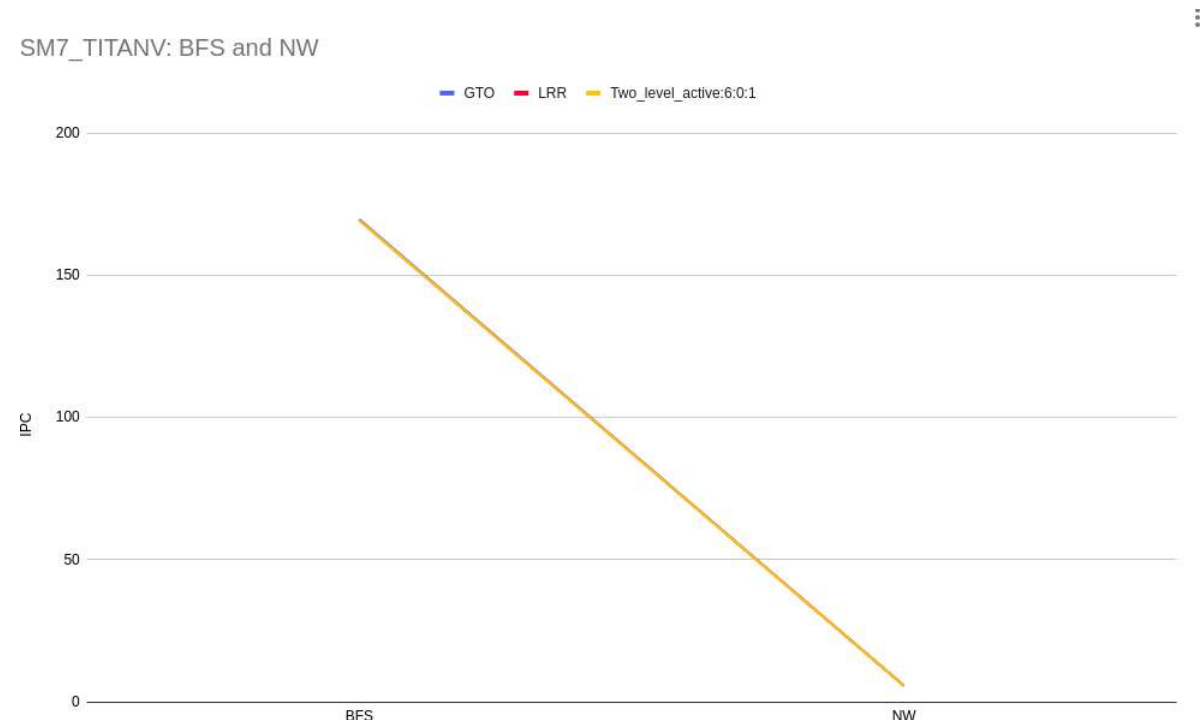


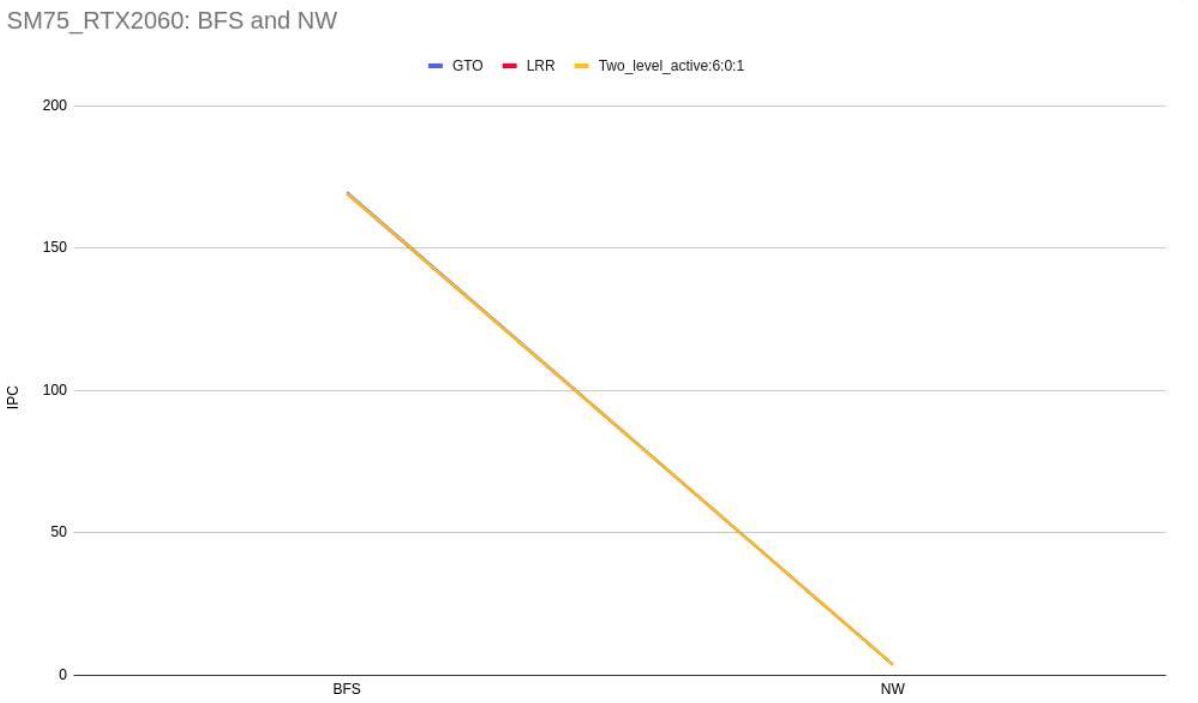
SM6_TITANX: BFS and NW



SM7_QV100: BFS and NW




SM7_TITANV: BFS and NW
GTO
LRR
Two_level_active:6:0:1
200
150
100
50
0
IPC
BFS
NW


SM75_RTX2060: BFS and NW
GTO
LRR
Two_level_active:6:0:1
200
150
100
50
0
IPC
BFS
NW

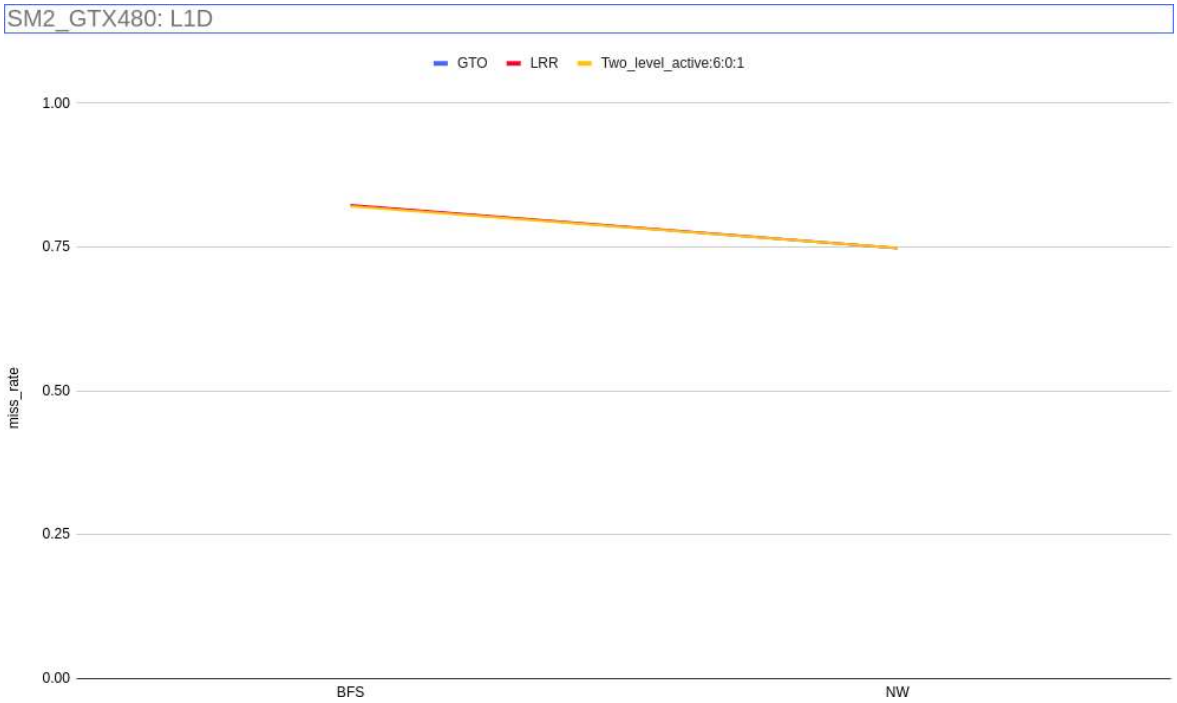
Excel Sheet showing IPC’s for different GPU graphics and different Warp schedulers

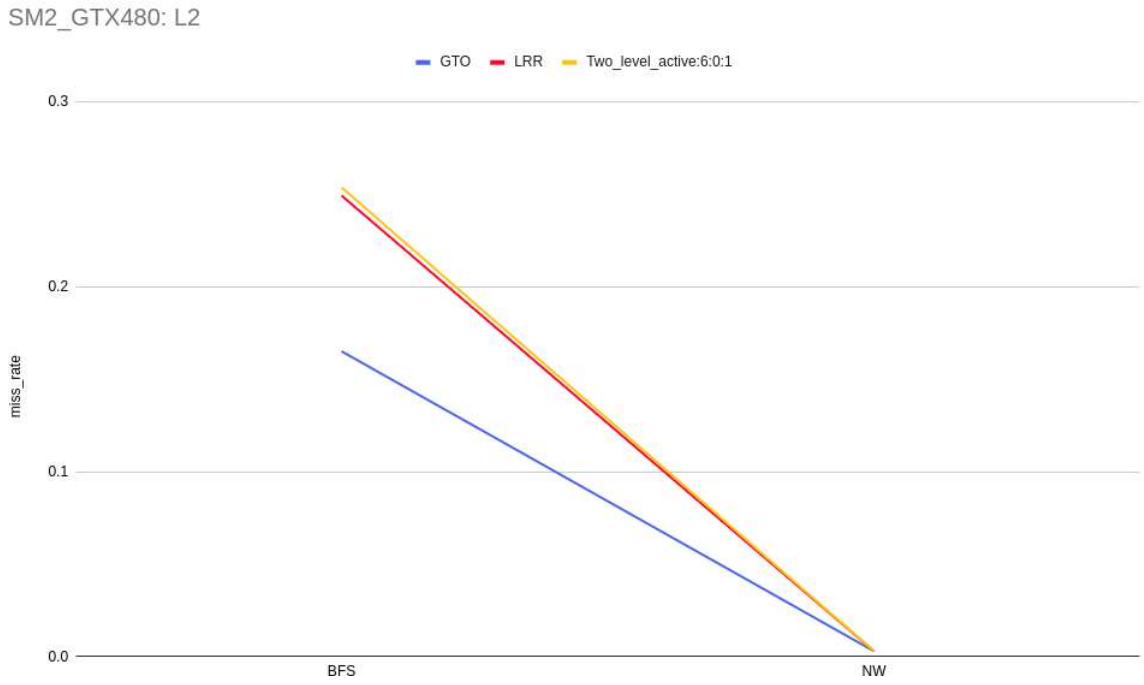
	GTO	LRR	Two_level_active:6:0:1	SM2
BFS	58.59	57.15	57.47	
NW	3.5	3.5	3.49	
	GTO	LRR	Two_level_active:6:0:1	SM3
BFS	115.3	111.88	112.81	
NW	5.77	5.77	5.75	
	GTO	LRR	Two_level_active:6:0:1	SM6
BFS	101.55	97.83	98.08	
NW	3.83	3.835	3.83	
	GTO	LRR	Two_level_active:6:0:1	SM7Q
BFS	169.55	169.21	169.24	
NW	3.66	3.65	3.65	
	GTO	LRR	Two_level_active:6:0:1	SM7T
BFS	169.63	169.27	169.27	
NW	5.74	5.75	5.73	
	GTO	LRR	Two_level_active:6:0:1	SM75
BFS	169.61	169.12	169.29	
NW	3.64	3.64	3.65	

Configuration runtime

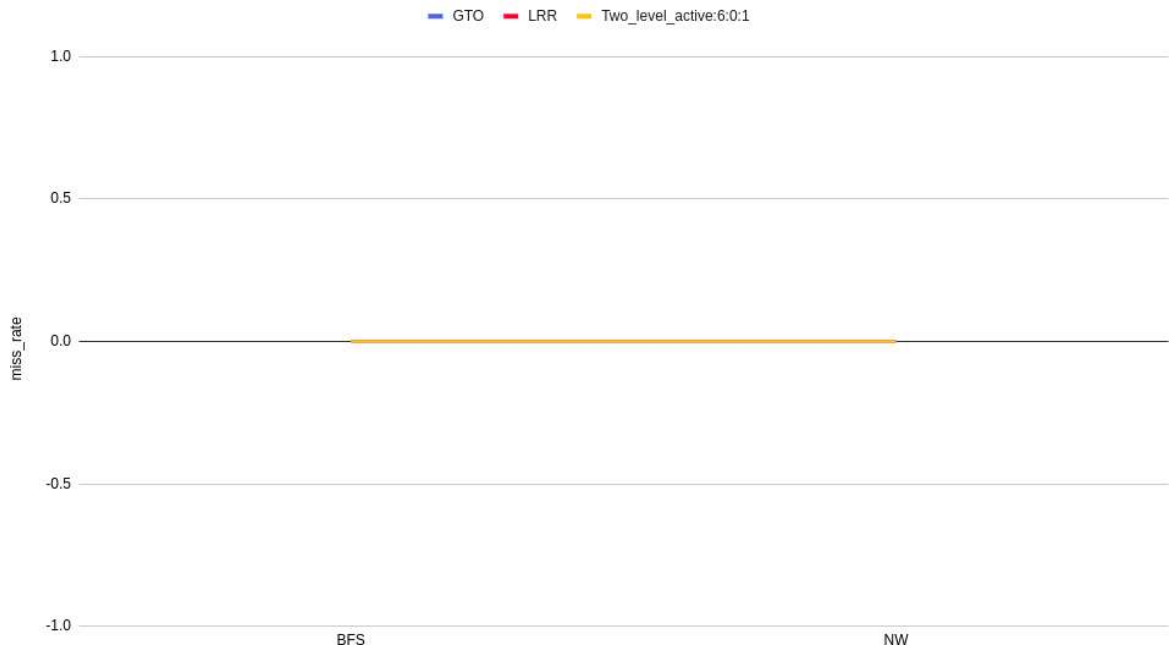
Configuration	Warp Scheduler	Runtime(sec)
SM2_GTX480	gto	169
	llr	135
	two_level	121
SM3_KEPLER_TITAN	gto	104
	llr	103
	two_level	105
SM6_TITANX	gto	212
	llr	202
	two_level	218
SM7_QV100	gto	188
	llr	180
	two_level	200
SM7_TITANV	gto	177
	llr	178
	two_level	190
SM75_RTX2060	gto	109
	llr	110
	two_level	119

Question 2: Graphs showing miss rates of L1D and L2

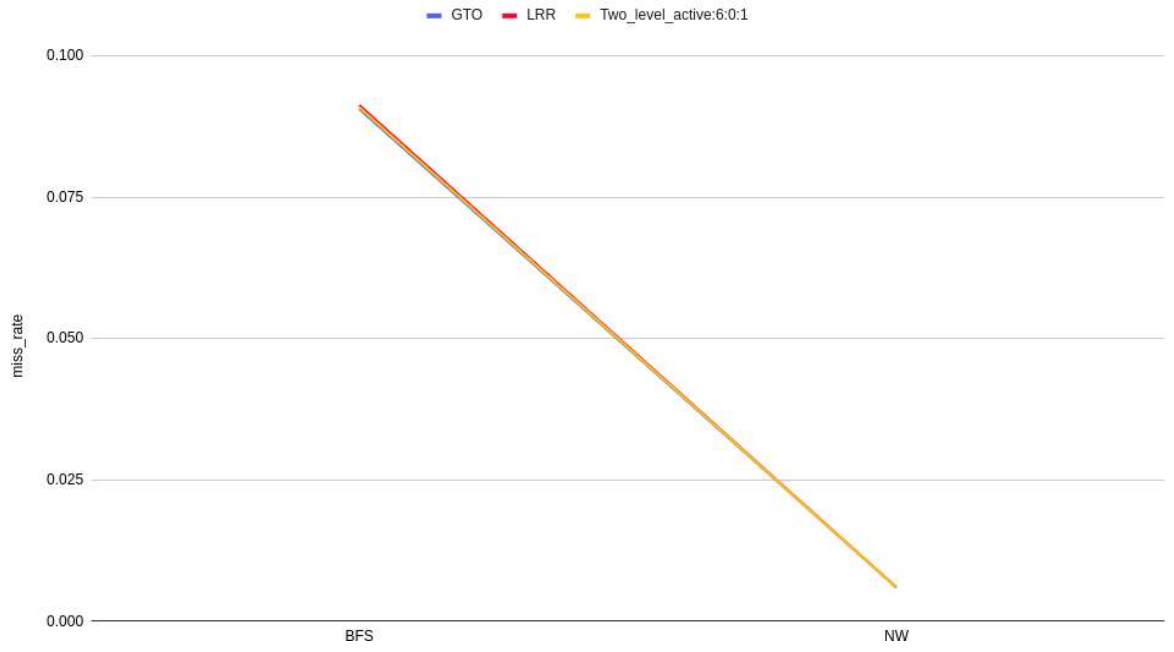




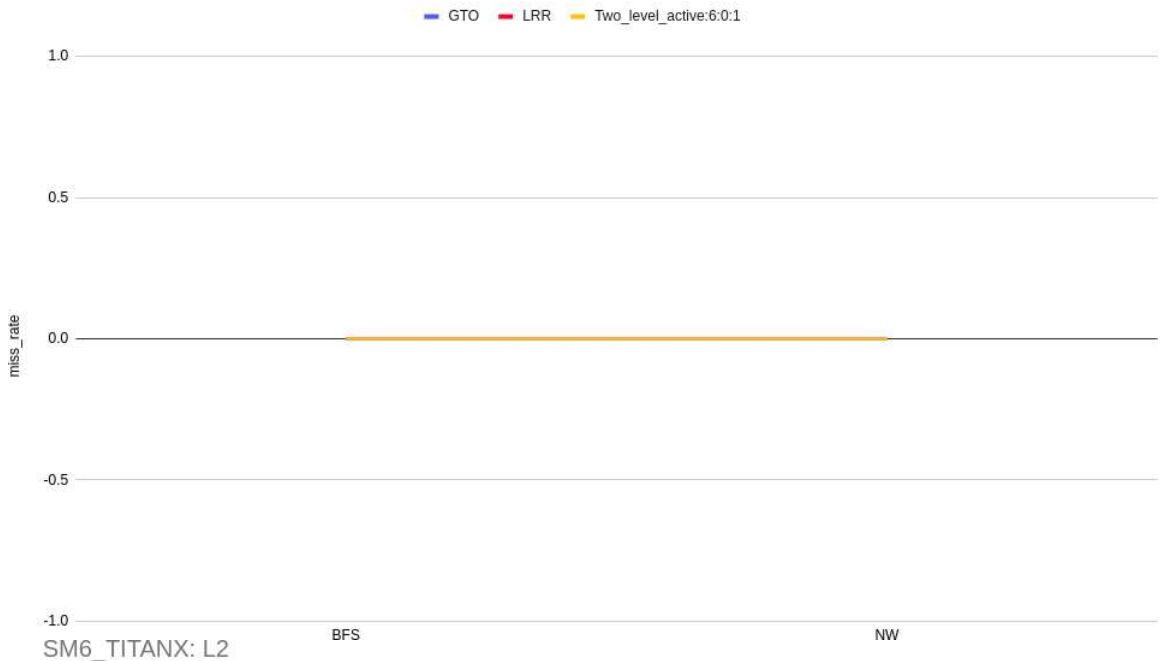
SM3_KEPLER_TITAN: L1D



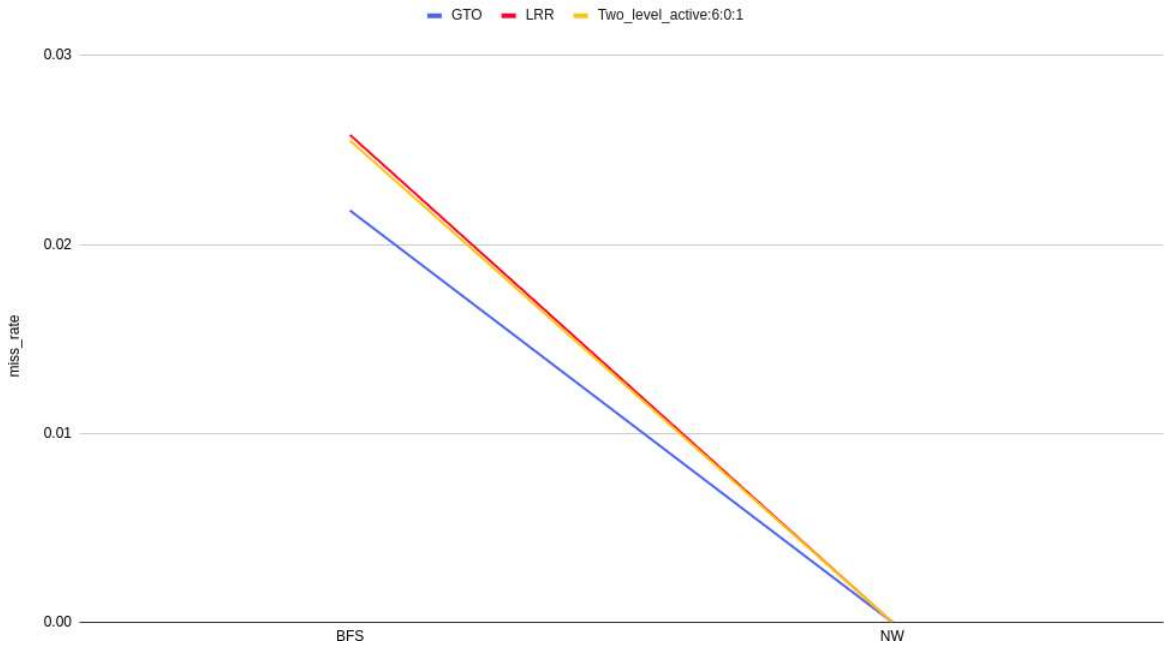
SM3_KEPLER_TITAN: L2



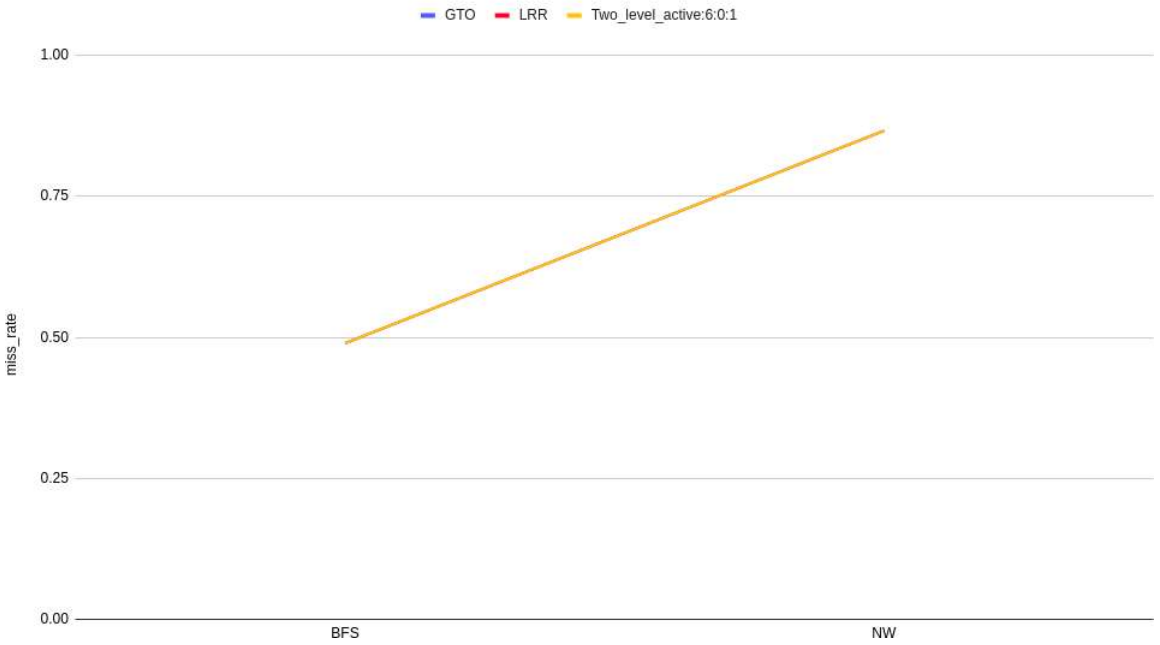
SM6_TITANX: L1D



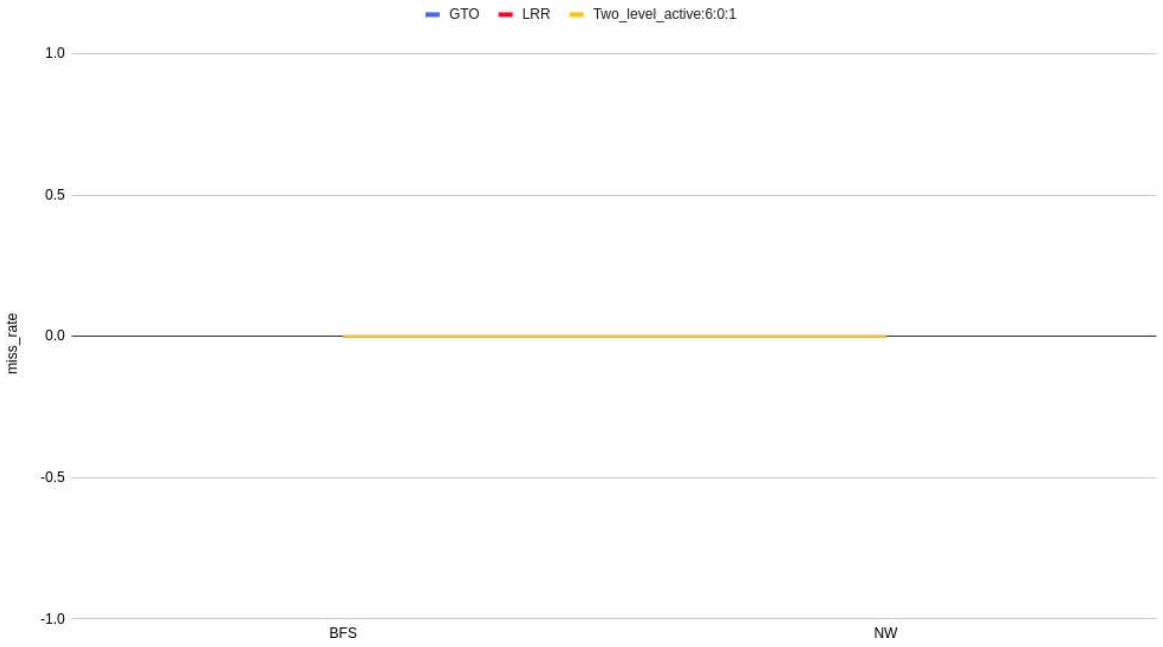
SM6_TITANX: L2



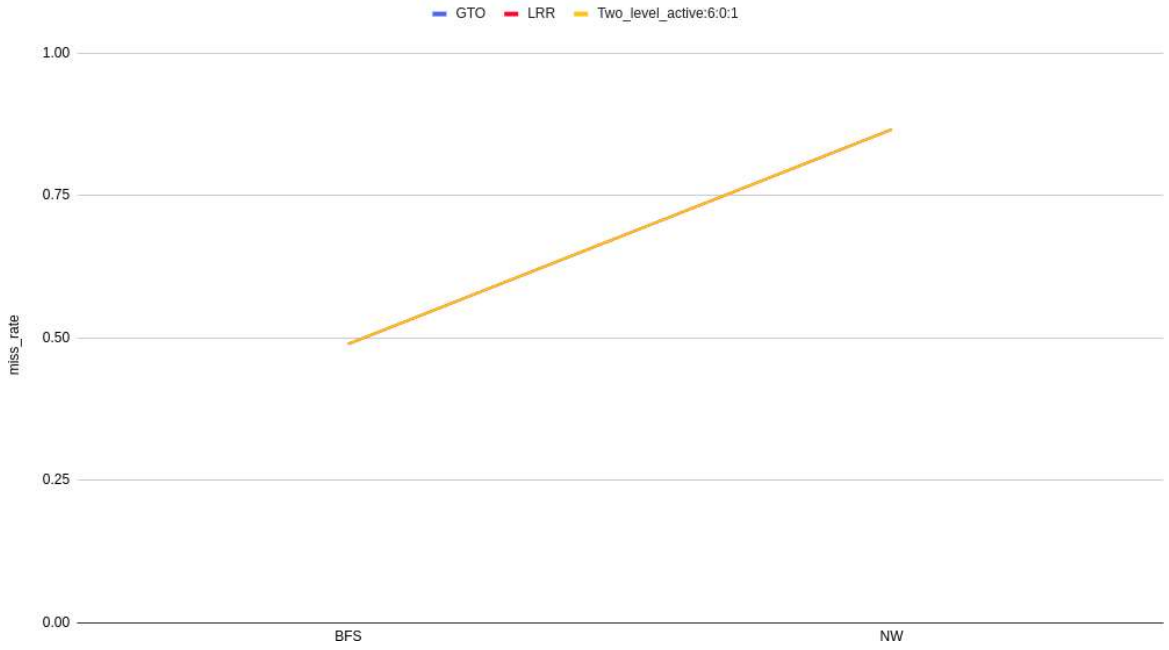
SM7_QV100: L1D



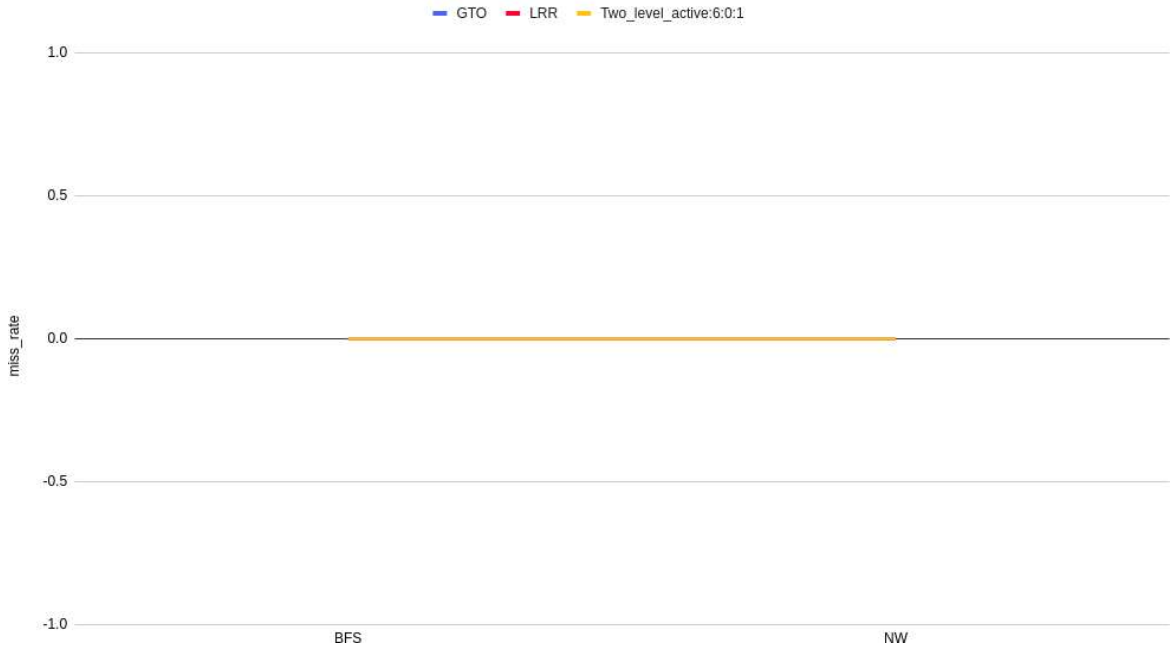
SM7_QV100: L2



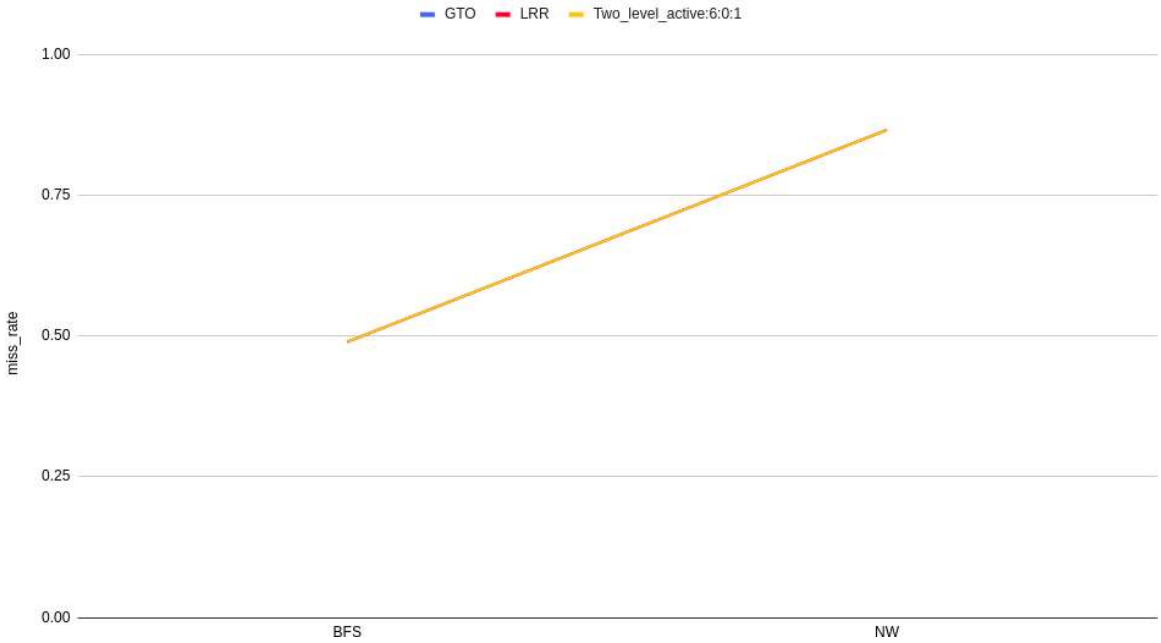
SM7_TITANV:L1D



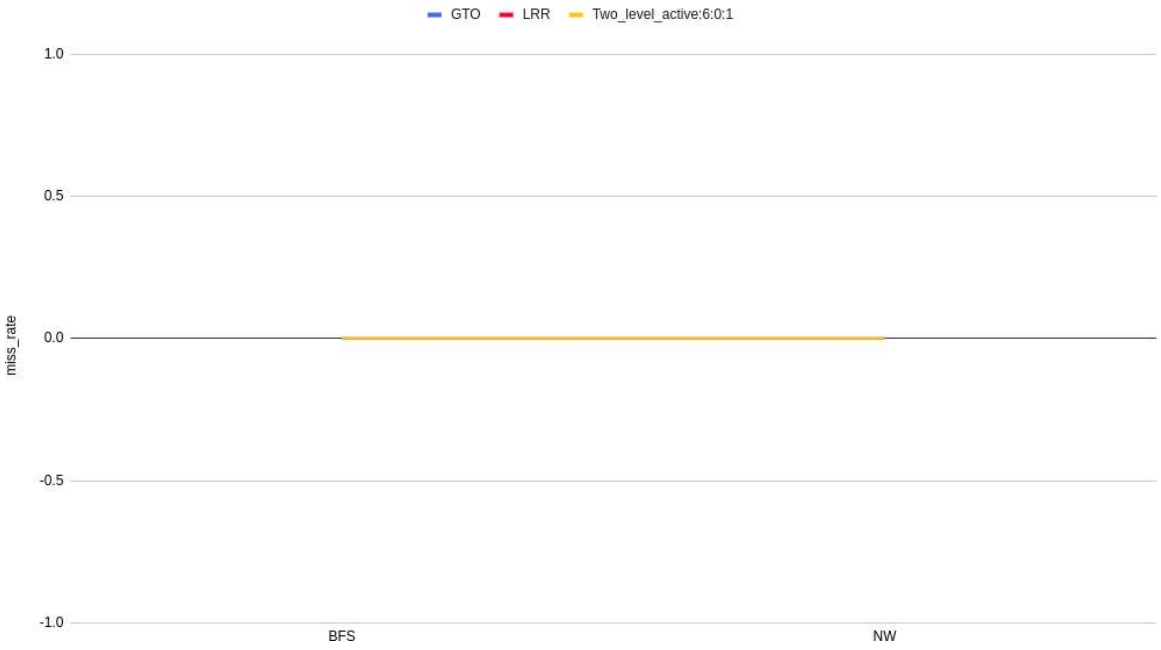
SM7_TITANV: L2



SM75_RTX2060: L1D



SM75_RTX2060: L2



L1D:

	GTO	LRR	Two_level_active:6:0:1	SM2
BFS	0.8225	0.823	0.8209	
NW	0.7482	0.7482	0.7484	
	GTO	LRR	Two_level_active:6:0:1	SM3
BFS	0	0	0	
NW	0	0	0	
	GTO	LRR	Two_level_active:6:0:1	SM6
BFS	0	0	0	
NW	0	0	0	
	GTO	LRR	Two_level_active:6:0:1	SM7Q
BFS	0.4894	0.4895	0.4895	
NW	0.8661	0.8661	0.8661	
	GTO	LRR	Two_level_active:6:0:1	SM7T
BFS	0.4894	0.4895	0.4895	
NW	0.8661	0.8661	0.8661	
	GTO	LRR	Two_level_active:6:0:1	SM75
BFS	0.4894	0.4895	0.4895	
NW	0.8661	0.8661	0.8661	

L2:

	GTO	LRR	Two_level_active:6:0:1	SM2
BFS	0.1651	0.2493	0.2537	
NW	0.003	0.003	0.0031	
	GTO	LRR	Two_level_active:6:0:1	SM3
BFS	0.0906	0.0913	0.0909	
NW	0.006	0.006	0.006	
	GTO	LRR	Two_level_active:6:0:1	SM6
BFS	0.0218	0.0258	0.0255	
NW	0	0	0	
	GTO	LRR	Two_level_active:6:0:1	SM7Q
BFS	0	0	0	
NW	0	0	0	
	GTO	LRR	Two_level_active:6:0:1	SM7T
BFS	0	0	0	
NW	0	0	0	
	GTO	LRR	Two_level_active:6:0:1	SM75
BFS	0	0	0	
NW	0	0	0	

Question 3

The categorization of L1D cache hit rates:

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

The SM75_RTX2060 (gto, llr, 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 particular category. Noteworthy instances include SM3_KEPLER_TITAN (gto, llr, two_level) and SM6_TITANX (gto, llr, two_level).

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.

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 cache sizes within the GPU configuration file are delineated in the subsequent format:

The utilization of General-Purpose Graphics Processing Units (GPGPUs) in computing systems has gained significant attention in recent years. One crucial

The <cache_type> parameter serves to specify the particular 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>.

In order to modify the cache size, it is customary to make adjustments to the parameters denoted as `<nsets>`, `<bsize>`, and `<assoc>`. The aforementioned 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
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.

In an attempt 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.

Configuration	Warp Scheduler	Runtime (sec)	L1D Miss Rate	L2 Miss Rate	L1D Hit Rate	L2 Hit Rate
SM2_GTX480	gto	86	0.3255	0.4795	0.6745	0.3257
SM3_KEPLER_TITAN	gto	65	0	0.0907	1	0
SM6_TITANX	gto	202	0	0.0217	1	0
SM7_QV100	gto	186	0.737	0	0.2628	0.7371
SM7_TITANV	gto	109	0.5869	0	0.4131	0.5869
SM75_RTX2060	gto	161	0.737	0.0013	0.2628	0.737

Question 4

Cache hit rates and power consumption can be correlated in certain ways, but it's important to note that this relationship is influenced by various factors and can vary based on the specific context and workload. Here's how cache hit rates might relate to power consumption:

Cache Hit Rates and Power Efficiency:

Higher cache hit rates generally lead to better cache utilization, which reduces the frequency of fetching data from higher-level memory (e.g., L2 or main memory). Fetching data from higher-levelmemory consumes more energy due to longer access latencies andhigher power requirements of those memory levels. Therefore, applications with higher cache hit rates are often more power- efficient because they minimize the need to access higher-level memory.

Cache Size and Power Consumption:

Increasing cache sizes, such as the L1D cache, can improve cachehit rates by providing more capacity to store frequently accessed data. However, larger caches also consume more power due to increased transistor count and higher access latencies. While largercaches can reduce overall memory access frequency, they might consume more power when compared to smaller caches.

Cache Access Patterns and Power Consumption:

The nature of cache access patterns, whether they exhibit good spatial and temporal locality, can influence both cache hit rates and power consumption. Applications with irregular memory access patterns or high cache contention might experience lower cache hit rates and potentially higher power consumption due to cache thrashing or contention-related delays.

Configuration	Warp Scheduler	Executi on Units Avg Power	DRAM Avg Power	Register Files Avg Power	Total Avg Power	% Execution	% DRAM	%Register Files
SM2_G TX480	gto	28.2801	0	66.0032	94.2833	29.994813 5	0	70.005186 5
SM3_K EPLER_ TITAN	gto	72.65407	0	6.13353	78.7876	92.215107 45	0	7.7848925 47
SM6_TI TANX	gto	35.9372	0	36.234	72.1712	49.794377 81	0	50.205622 19
SM7_Q V100	gto	48.3974	0	72.3856	120.783	40.069711 8	0	59.930288 2
SM7_TI TANV	gto	111.38912	0	20.79588	132.185	84.267594 66	0	15.732405 34
SM75_ RTX206 0	gto	40.51311	0	20.89769	61.4108	65.970659 88	0	34.029340 12

Application Characteristics and Power Consumption:

The workload characteristics of different applications play a crucial role in determining the relationship between cache hit rates and power consumption. Some applications might be inherently memory-bound and require frequent memory accesses, leading to potentially higher power consumption even with good cache hit rates.

1. Lower Power Consumption and Higher L1D Cache Hit Rates:
Higher L1D cache hit rates often mean that a greater percentage of memory accesses are being handled by the cache, obviating the need to visit slower main memory.
- Since accessing the cache uses less power than retrieving data from main memory, this could result in lower power consumption connected to memory.

2. Lower L1D Cache Hit Rates and Possibly Higher Power Consumption:
- Lower L1D cache hit rates imply that a greater percentage of memory requests are resulting in cache misses, necessitating the retrieval of data from main memory.
- Due to greater data flow, cache misses may result in slower memory access times and perhaps higher power consumption.

3. Workload and Architecture Impact:
- The correlation between L1D cache hit rates and power consumption can vary based on the workload and the architecture of the CPU or GPU.
- Different applications have varying memory access patterns, which can influence cache behavior and, consequently, power consumption.

4. Trade-offs:
- Optimizing for higher cache hit rates can lead to lower memory-related power consumption, but it might require larger and more power-hungry caches.
- The overall impact on power consumption depends on the balance between

cache performance improvements and the energy costs associated with maintaining larger caches.

5. Caveats:
- Correlation does not imply causation. While a correlation might exist, it doesn't necessarily mean that changes in cache hit rates directly cause changes in power consumption.

To determine the specific correlation between L1D cache hit rates and power consumption for your applications, you would need to perform a detailed analysis using the actual data you have. If you provide me with the L1D cache hit rates and power consumption data for different applications, I'd be happy to help you analyze and interpret the correlation between these factors.