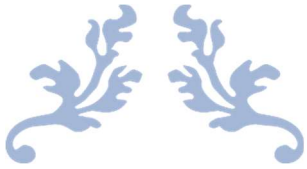


COA ASSIGNMENT 4

POWER PROFILING

Group 9:

Ponnanna A H

Priyam Saha

Arihant Garg

Shubh Modi

Tejas Singh Rajput

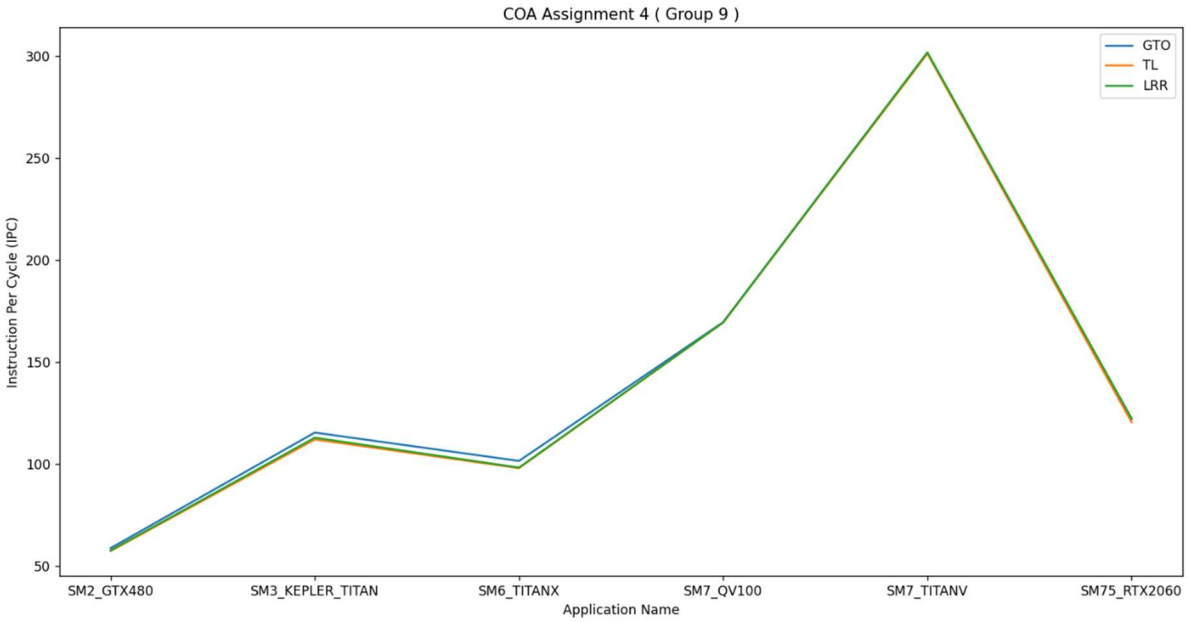
STEPS FOR THE ASSIGNMENT:

1. Go to https://github.com/yuhc/gpu-rodinia repository for the benchmark files.
2. We git cloned the benchmark files in a folder.
3. We downloaded the sample data files from the given link: https://www.dropbox.com/s/cc6cozpboht3mtu/rodinia-3.1-data.tar.gz
4. From the benchmark files we chose the BFS application.
5. Then we copied the sample data file for BFS into the data folder inside the gpgpu rodinia folder.
6. When copying the config files to the BFS code folder, open the gpgpusim.config file and enable the power profiling by setting -power_simulation_enabled attribute to 1.

1. **What is the runtime for each configuration?**
   Example: Plot showing the IPC on Y-axis and application name on X-axis, legend: different warp schedulers.

Configuration	Warp Scheduler	Instructions/ Second	Cycles/ Second	IPC
SM2_GTX480	gto	162771	2777	58.614
	lrr	174794	3058	57.159
	tl	193483	3366	57.481
SM3_KEPLER_TITAN	gto	209277	1815	115.304
	lrr	213637	1909	111.910
	tl	221322	1962	112.804
SM6_TITANX	gto	157763	1556	101.390
	lrr	161067	1646	97.853
	tl	158576	1616	98.128
SM7_QV100	gto	120171	709	169.493
	lrr	118322	699	169.273
	tl	114790	678	169.306
SM7_TITANV	gto	115220	382	301.623
	lrr	119239	396	301.186
	tl	117419	389	301.848
SM75_RTX2060	gto	199765	1639	121.882
	lrr	193483	1607	120.400
	tl	193483	1582	122.302

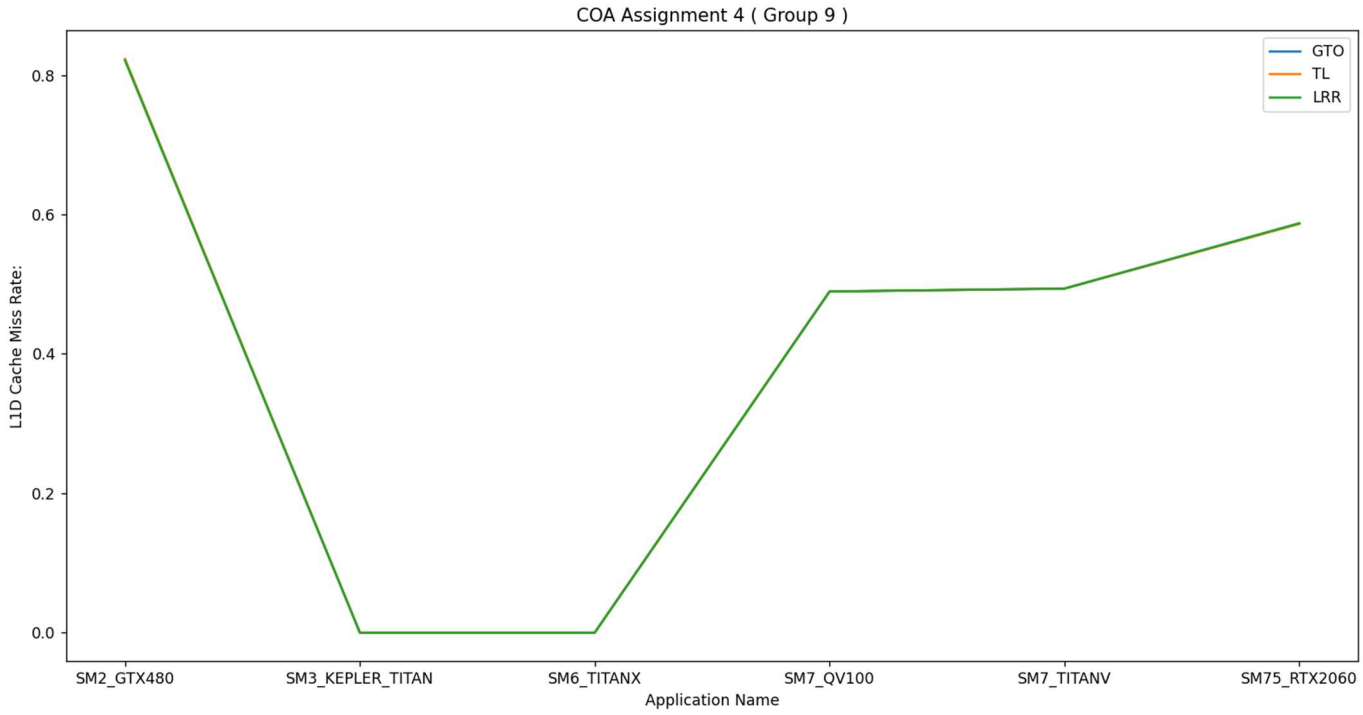
Graph for Instructions per cycle:



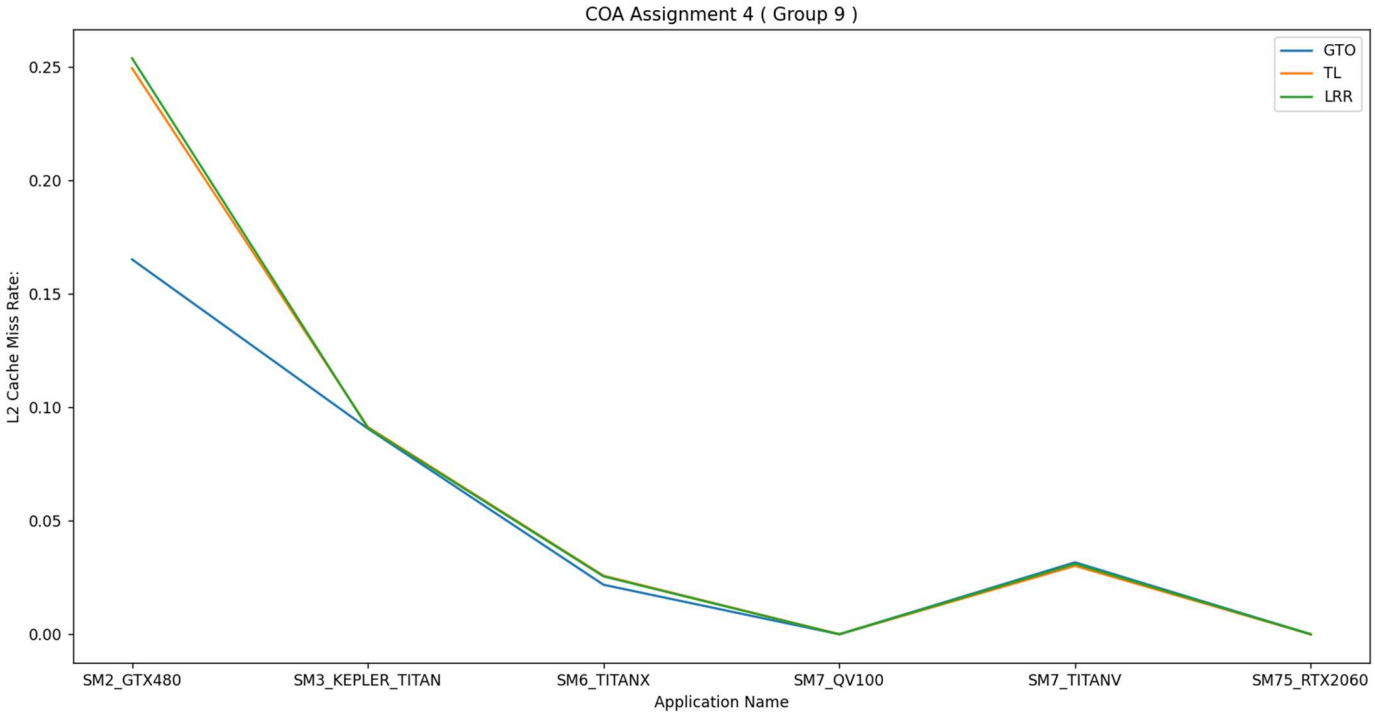
2. **Which warp scheduler has the best cache hit rate? Why?**
   Example: Plot showing the following on Y-axis and application name on X-axis,
   legend: different warp schedulers.
   (i) L1D miss rates
   (ii) L2 miss rates

Configuration	Warp Scheduler	L1D Miss Rate	L2 Miss Rate	L1D Hit Rate	L2 Hit Rate
SM2_GTX480	gto	0.8225	0.1651	0.1775	0.8349
	lrr	0.8230	0.2493	0.1770	0.7507
	tl	0.8209	0.2537	0.1791	0.7463
SM3_KEPLER_TITAN	gto	0	0.0906	1	0.9094
	lrr	0	0.0913	1	0.9087
	tl	0	0.0909	1	0.9091
SM6_TITANX	gto	0	0.0218	1	0.9782
	lrr	0	0.0258	1	0.9742
	tl	0	0.0255	1	0.9745
SM7_QV100	gto	0.4894	0	0.5106	1
	lrr	0.4895	0	0.5105	1
	tl	0.4895	0	0.5105	1
SM7_TITANV	gto	0.4936	0.0317	0.5064	0.9683
	lrr	0.4936	0.0301	0.5064	0.9699
	tl	0.4937	0.0310	0.5063	0.9690
SM75_RTX2060	gto	0.5869	0	0.4131	1
	lrr	0.5865	0	0.4135	1
	tl	0.5875	0	0.4125	1

Graph for L1D cache Miss Rate:



Graph for L2 cache Miss Rate:



3. **Categorize the applications w.r.t the L1D and L2 Cache hit rates. What changes do you observe w.r.t L1D and L2 cache hit rates when the L1D cache size is increased from 32KB to 8MB? (use warp scheduler GTO)**

Configuration	L1D		L2	
	32KB	8MB	32KB	8MB
SM2_GTX480	0.1775	0.6743	0.8349	0.5205
SM3_KEPLER_TITAN	1	1	0.9094	0.9094
SM6_TITANX	1	1	0.9782	0.9782
SM7_QV100	0.5106	0.5106	1	1
SM7_TITANV	0.5064	0.5064	0.9683	0.9683
SM75_RTX2060	0.4131	0.5768	1	1

In the provided configuration for the Fermi architecture, you have specified the cache settings for the Level 1 Data Cache (dl1). Let's break down the configuration parameters:

- `<nsets>:<bsize>:<assoc>`: These parameters define the basic properties of the cache.

-`<nsets>` is set to 32, which means there are 32 sets in the cache.

- `<bsize>` is set to 128, which implies that each cache block (line) is 128 bytes in size.

- `<assoc>` is set to 2048, indicating a high level of associativity. There are 2048 cache lines per set.

We observed that on increasing the L1D cache size from 32KB to 8MB, the hit rate increased due to the increased L1D cache size resulting in fewer misses .

4. **What percentage of power is consumed by Execution units, DRAM, Register Files in each application run? Do you notice any correlation between the L1D cache hit rates observed in Question 3 and the Power consumption between different applications?**

Configuration	Exec. Unit	DRAM	Reg. Files	Total Power	% EU	% DRAM	% RF
SM2_GTX480	28.28	0.01	66.01	94.28	29.99	0.01	70.01
SM3_KEPLER_TITAN	72.65	0.08	6.13	78.78	92.21	0.08	7.78
SM6_TITANX	35.93	0.07	36.23	72.17	49.79	0.07	50.20
SM7_QV100	48.39	0.13	72.38	120.78	40.06	0.13	59.93
SM7_TITANV	111.38	0.78	20.79	132.18	84.26	0.78	15.73
SM75_RTX2060	40.51	0.05	20.89	61.41	65.97	0.05	34.02

- Here, we'll analyze the changes that occur when the cache size is increased for the specified graphics processing unit (GPU) models from 32 kilobytes to 8 megabytes. This analysis's main goal is to investigate differences in the proportional representation of power consumption across various components, including execution units, DRAM, and register files.
- Execution Unit Power Consumption: In general, there is a positive relationship between execution unit power usage and cache size. The aforementioned finding is demonstrated in the context of specific GPU models, where a rise in cache capacity is accompanied by a rise in the fraction of power allocated to execution units.
- DRAM Power Consumption: - The administration of a larger cache needs increased power utilization, hence the power consumption of the DRAM shows a positive association with cache size. The aforementioned observation may be seen in certain graphics processing unit (GPU) models with an 8 MB cache by the perceptible increase in power consumption of dynamic random-access memory (DRAM) and the corresponding rise in the proportion of power allotted to DRAM.
- Power Consumption of Register Files: - Although it may not always follow a clear pattern, the power consumption of register files varies depending on the cache size.

- Aggregate Power: In certain cases, an increased cache size results in an increase in aggregate power consumption, which is mostly due to increased power consumption by execution units and DRAM modules.