NC State University

Department of Electrical and Computer Engineering

ECE 786: Spring 2023

Final Project

Profile based Cache Bypass logic using GP-GPU Sim

by

RAGHUL SRINIVASAN

UnityID: rsriniv7

(200483357)

TASK 1: Cache Efficiency Analysis

Configuration Used -> SM2_GTX480

Assumptions ->

- In each of the kernel, if the % change in IPC between No cache-bypassing and cache-bypassing mode is
 - o Greater than 2%, then it is assumed as Cache Unfriendly Kernel
 - o Inbetween -2% and 2% then it is assumed as Cache Insenstive Kernel
 - o Less than -2%, then it is assumed as Cache friendlt Kernel
- If more than 50% of kernels in a benchmark is Cache Unfriendly, then the benchmark is considered as Cache Unfriendly Benchamark.
- Max. number of instructions executed to 100M by using below command in gpgpusim.config

-gpgpu_max_insn 100000000

bench mark name	kernel name	kernel_l aunch _uid	IPC with no cache bypassing	IPC with cache bypassing	percentage change of comparing the IPC with/without cache bypassing	kernel category	Benchmark Category	
	_Z6KerneIP4NodePiPbS2_S1_S2_i	1	217.5687	167.9066	-22.82593958	Cache Friendly (L1 cache bypassing will cause performance degradation)		
1	_Z6KerneIP4NodePiPbS2_S1_S2_i	2	206.0139	146.9099	-28.6893263	Cache Friendly (L1 cache bypassing will cause performance degradation)		
1	_Z6KerneIP4NodePiPbS2_S1_S2_i	3	165.9271	112.0179	-32.48968975	Cache Friendly (L1 cache bypassing will cause performance degradation)		
	_Z6KerneIP4NodePiPbS2_S1_S2_i	4	76.2236	61.3361	-19.53135249	Cache Friendly (L1 cache bypassing will cause performance degradation)	1	
BFS	_Z6KerneIP4NodePiPbS2_S1_S2_i	5	21.3021	36.1667	69.77997474	Cache Unfriendly (L1 cache bypassing can improve the IPC)	-	
	_Z6KerneIP4NodePiPbS2_S1_S2_i	6	22.5533	44.4395	97.042118	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z6KerneIP4NodePiPbS2_S1_S2_i	7	46.5675	86.5094	85.77205132	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
1	_Z6KerneIP4NodePiPbS2_S1_S2_i	8	354.4445	455.3303	28.46307391	Cache Unfriendly (L1 cache bypassing can improve the IPC)	1	
	_Z6KerneIP4NodePiPbS2_S1_S2_i	9	473.1056	486.792	2.892884802	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
LPS	_Z13GPU_laplace3diiiiPfS_	1	383.1095	408.8568	6.720611209	Cache Unfriendly (L1 cache bypassing can improve the IPC)	Cache Unfriendly	
NQU	_Z24solve_nqueen_cuda_kerneliiPj S_S_S_i	1	30.4185	30.7699	1.155218042	Cache Insensitive (IPCs have no significant difference when enabling /disabling the bypass logic)	Cache Insensitive	
	_Z17executeFirstLayerPfS_S_	1	345.3974	144.1486	-58.26586998	Cache Friendly (L1 cache bypassing will cause performance degradation)	Cache Friendly	
NN	_Z18executeSecondLayerPfS_S_	2	211.7879	97.8234	-53.81067568	Cache Friendly (L1 cache bypassing will cause performance degradation)		
] "" [_Z17executeThirdLayerPfS_S_	3	9.449	4.325	-54.22796063	Cache Friendly (L1 cache bypassing will cause performance degradation)		
	_Z18executeFourthLayerPfS_S_	4	6.8539	3.2347	-52.80497235	Cache Friendly (L1 cache bypassing will cause performance degradation)		
ВР	_Z22bpnn_layerforward_CUDAPfS_S S_ii	1	675.6067	671.3728	-0.626681174	Cache Insensitive (IPCs have no significant difference when cache bypassing)	Cache Insensitive	
HS	_Z14calculate_tempiPfS_S_iiiiiffffff	1	701.3718	707.6299	0.8922657	Cache Insensitive (IPCs have no significant difference when cache bypassing)	Cache Insensitive	
	_Z12lud_diagonalPfii	1	0.7026	0.7176	2.134927412	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13lud_perimeterPfii	2	9.2446	9.1103	-1.452739978	Cache Insensitive (IPCs have no significant difference when cache bypassing)		
	_Z12lud_internalPfii	3	501.2445	567.1572	13.14981012	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
]	_Z12lud_diagonalPfii	4	0.7558	0.7742	2.434506483	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	5	10.9464	11.8102	7.891178835	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	6	497.3745	574.7466	15.55610511	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
]	_Z12lud_diagonalPfii	7	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
] [_Z13lud_perimeterPfii	8	10.1697	10.9718	7.88715498	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
]	_Z12lud_internalPfii	9	473.0808	557.2787	17.79778423	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
] [_Z12lud_diagonalPfii	10	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
] [_Z13lud_perimeterPfii	11	9.3893	10.1287	7.874921453	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13lud_perimeterPfii	11	9.3893	10.1287	7.874921453	Cache Unfriendly (L1 cache bypassing can improve the IPC)		

	_Z12lud_internalPfii	12	462.4784	529.6388	14.52184578	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	13	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	14	8.6082	9.2874	7.890151251	Cache Unfriendly (L1 cache bypassing can improve the IPC)] [
	_Z12lud_internalPfii	15	378.4012	504.6895	33.37418063	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	16	0.7558	0.7742	2.434506483	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	17	7.8294	8.4467	7.884384499	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	18	357.2093	493.737	38.22064543	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	19	0.7558	0.7742	2.434506483	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	20	7.0473	7.604	7.899479233	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	21	338.0277	453.3258	34.10906858	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	22	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
LUD	_Z13Iud_perimeterPfii	23	6.264	6.7609	7.932630907	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
100	_Z12lud_internalPfii	24	324.1251	467.1097	44.11401647	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	25	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	26	5.4832	5.9163	7.898672308	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	27	290.9933	405.2074	39.2497353	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	28	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	29	4.7006	5.0733	7.92877505	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	30	246.8571	344.3503	39.49378	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	31	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	32	3.9172	4.2288	7.954661493	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	33	208.6225	253.2766	21.40425889	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	34	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	35	3.1348	3.3833	7.927140487	Cache Unfriendly (L1 cache bypassing can improve the IPC)	Cache Unfriendly	
	_Z12lud_internalPfii	36	142.2966	172.1319	20.96698024	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	37	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	38	2.3514	2.5387	7.965467381	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_internalPfii	39	111.9498	134.8471	20.45318527	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	40	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13Iud_perimeterPfii	41	1.5679	1.6926	7.953313349	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12Iud_internalPfii	42	39.4499	44.9208	13.86796925	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	43	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z13lud_perimeterPfii	44	0.8583	0.8467	-1.351508796	Cache Insensitive (IPCs have no significant difference when cache bypassing)		
	_Z12Iud_internalPfii	45	16.2623	16.6957	2.665059678	Cache Unfriendly (L1 cache bypassing can improve the IPC)		
	_Z12lud_diagonalPfii	46	0.7558	0.7741	2.42127547	Cache Unfriendly (L1 cache bypassing can improve the IPC)		

From the above table,

- BFS, LPS, LUD benchmarks -> Cache Unfriendly benchmarks
- NQU, BP, HS benchmarks -> Cache Insensitive benchmarks
- NN -> Cache Friendly Benchmark

TASK 2: Profile-based Cache Bypassing

Configuration Used -> SM7_QV100

Benchmarks Used -> BFS, LPS, LUD (Cache Unfriendly Benchmarks from Task 1)

Implimentation:

- The simulation runs in two modes depending on the existence of "ProfileDump.txt" file in the run directory.
 - Profiling Mode (when "ProfileDump.txt" is not present)
 - Profile Based Bypassed Mode (when "ProfileDump.txt" is present)
- The mode is maintained using an enum variable **profiling_mode**
- A hash table **cache_bypass_profiling_data** with three nested keys (SM, Kernel, Address) are created to store the reference counters for each of the memory access addresses without any duplicate entry.

Profiling Mode:

- The reference counters with appropriate keys are incremented for each memory access in ldst_unit::memory_cycle().
- At the end of simulation, all the keys are stroed in "ProfileDump.txt" which can be used as reference for subsequent simulation to run in Profile based Bypass mode.

Profile Based Bypassed Mode:

- The reference counters with appropriate keys are updated with the datum present in "ProfileDump.txt" file.
- These counters are checked everytime, when accessing a memory. If the counter value is less than 3, bypassing logic enabled (Similar to PA3a) in ldst_unit::cycle())

Modifications made in the code: Shadder.h file

- Hash Table, Profile mode enum and DumpFileName are defined in shadder.h
- These files are defined using **extern** to avoid multiple declaratation when shader.h is called by may .cc files.

```
#define WRITE_PACKET_SIZE 8

#define WRITE_MASK_SIZE 8

#define WRITE_MASK_SIZE 8

extern std::map<unsigned, std::map<unsigned, unsigned>>> Cache_bypass_Profiling_data;
enum profiling_mode_enum

{

PROFILING = 0,

BYPASSING = 1

};

extern profiling_mode_enum profiling_mode;
extern std::string_DumpFileName;

class gpgpu_context;

enum_exec_unit_type_t

enum_exec_unit_type_t
```

Shadder.cc file

Hash Table, Profile mode enum and DumpFileName are initialized in shadder.cc

```
#define MIN(a, b) (((a) < (b)) ? (a) : (b))

std::map<unsigned, std::map<unsigned, unsigned>>> Cache_bypass_Profiling_data;

profiling_mode_enum profiling_mode;

std::string_DumpFileName = "ProfileDump.txt";

mem_fetch *shader_core_mem_fetch_allocator::alloc(
```

- In Idst_unit::memory_Cycle(), the hash table cache_bypass_profiling data is
 - o incremented if the simulator is in profiling mode
 - o read used to bypass L1D if the ref. counter is les than 3. (in bypassing mode)

• In **Idst_unit::cycle()**, if the simulator is in Bypasing mode, **bypassL1D** is set to true if the ref. counter in cache bypass profiling data is less than 3.

```
if (profiling_mode == profiling_mode_enum::BYPASSING)
{

if (Cache_bypass_Profiling_data[m_core->get_sid()][m_core->get_wernel()->get_uid()][mf->get_addr()] < 3)

bypassL1D = true;
}
</pre>
```

Gpgpusim entrypoint.cc file

- In **gpgpu_context::gpgpu_ptpx_sim_init_perf()**, enum variable profile_mode assigned depending on the presence of "ProfileDump.txt".
- If "ProfileDump.txt" is present, all the datum read, parsed and reference counters with keys are stored in **cache_bypass_profiling_data** table.

```
ifstream DumpFileHandler;
DumpFileHandler.open(DumpFileName);
if (!DumpFileHandler.is_open())
  profiling_mode = profiling_mode_enum::PROFILING;
cout << "***ProfileDump.txt not found. Entering PROFILING Mode***" << '\n';</pre>
 profiling_mode = profiling_mode_enum::BYPASSING;
 cout << "***ProfileDump.txt found. Entering BYPASSING Mode***" << '\n';</pre>
  string FileLine;
  int File_SM, File_Kernel, File_Address, File_Counter;
  while (getline(DumpFileHandler, FileLine))
    if (!FileLine.empty())
      string SubFileLine = FileLine.substr(0, 1);
if (SubFileLine == "1_")
        File_SM = stoi(FileLine.substr(5)); // 1_SM
      else if (SubFileLine == "2_")
        File Kernel = stoi(FileLine.substr(9)); // 2 KERNEL
      else if (SubFileLine == "3_")
        string SecondPartFileLine = FileLine.substr(FileLine.find(";") + 1);
        FileLine = FileLine.substr(0, FileLine.find(";"));
File_Address = stoi(FileLine.substr(10)); // 3_ADDRESS
        File_Counter = stoi(SecondPartFileLine.substr(10)); // 4_COUNTER
        Cache_bypass_Profiling_data[File_SM][File_Kernel][File_Address] = File_Counter;
return the_gpgpusim->g_the_gpu;
```

• In **gpgpu_context::print_simulation_time()**, if the sim is in profiling mode, the reference counters are dumped to "ProfileDump.txt" files in appropriate format.

Steps to be followed to run in no-bypass / profile based bypass mode:

- In the run directory, if the "profiledump.txt" file is not present, then the GPGPUsim will run in Profiling mode to get the reference counter values and dump it in "profiledump.txt" file, so that on subsequent runs, the GPGPUsim used this dump file to run in profile based bypassing mode.
- In the run directory, if the "profiledump.txt" file is present, then the GPGPUsim will run in Profile based bypassing mode by reading the reference counters from this file.
- One main advantage of this method is we don't have to recompile the GPGPUsim for switching between the two modes.

Benchmark results – Without Bypassing vs Profile based bypassing

benchmark name	kernel name			IPC -> Profile based bypassin
	_Z6KerneIP4NodePiPbS2_S1_S2_i	1	139.74	104.468
	_Z6KerneIP4NodePiPbS2_S1_S2_i	2	124.5014	90.2066
	_Z6KerneIP4NodePiPbS2_S1_S2_i	3	115.862	77.9455
	_Z6KerneIP4NodePiPbS2_S1_S2_i	4	109.4502	72.6998
DEC.	_Z6KerneIP4NodePiPbS2_S1_S2_i	5	87.2392	66.722
BFS	_Z6KerneIP4NodePiPbS2_S1_S2_i	6	86.6631	66.8945
	_Z6KerneIP4NodePiPbS2_S1_S2_i	7	145.4857	111.03
	_Z6KerneIP4NodePiPbS2_S1_S2_i	8	229.1067	175.9334
	_Z6KerneIP4NodePiPbS2_S1_S2_i	9	161.7932	149.3736
	_Z6KerneIP4NodePiPbS2_S1_S2_i	10	193.9836	188.3223
LPS	_Z13GPU_laplace3diiiiPfS_	1	638.116	667.4357
	_Z12lud_diagonalPfii	1	0.7678	0.7782
	Z13lud_perimeterPfii	2	14.5785	15.359
	_Z12lud_internalPfii	3	712.5299	721.2603
	Z12lud_diagonalPfii	4	0.7678	0.7782
	_Z13lud_perimeterPfii	5	13.6126	14.3418
	_Z12lud_internalPfii	6	637.3949	642.5734
	_Z12lud_diagonalPfii	7	0.7678	0.7782
	_Z13lud_perimeterPfii	8	12.6459	13.3236
	_Z12lud_internalPfii	9	556.1233	557.6649
	_Z12lud_diagonalPfii	10	0.7678	0.7782
	_Z13lud_perimeterPfii	11	11.6784	12.3045
	_Z12lud_internalPfii	12	502.3962	506.7778
	_Z12lud_diagonalPfii	13	0.7678	0.7782
	_Z13lud_perimeterPfii	14	10.7269	11.3033
	_Z12lud_internalPfii	15	423.2689	429.6447
	_Z12lud_diagonalPfii	16	0.7678	0.7782
	_Z13lud_perimeterPfii	17	9.7561	10.2805
	_Z12lud_internalPfii	18	352.6066	357.2629
	_Z12lud_diagonalPfii	19	0.7678	0.7782
Ī	_Z13lud_perimeterPfii	20	8.7844	9.2568
Ī	_Z12lud_internalPfii	21	290.0794	293.3001
	_Z12lud_diagonalPfii	22	0.7678	0.7782
IIID	_Z13lud_perimeterPfii	23	7.8118	8.2322
LUD	_Z12lud_internalPfii	24	232.6988	235.2133

-I LU	IDPerimeten in		1.0220	U.LULL
LO	_Z12lud_internalPfii	24	232.6988	235.2133
	_Z12lud_diagonalPfii	25	0.7678	0.7782
	_Z13lud_perimeterPfii	26	6.8493	7.2186
	_Z12lud_internalPfii	27	178.1328	180.4473
	_Z12lud_diagonalPfii	28	0.7678	0.7782
	_Z13Iud_perimeterPfii	29	5.8735	6.1903
	_Z12lud_internalPfii	30	131.4149	132.7995
	_Z12lud_diagonalPfii	31	0.7678	0.7782
	_Z13Iud_perimeterPfii	32	4.8967	5.161
	_Z12lud_internalPfii	33	91.527	92.7536
	_Z12lud_diagonalPfii	34	0.7678	0.7782
	_Z13Iud_perimeterPfii	35	3.9191	4.1308
	_Z12lud_internalPfii	36	58.4963	59.3623
	_Z12lud_diagonalPfii	37	0.7678	0.7782
	_Z13Iud_perimeterPfii	38	2.9442	3.1034
	_Z12lud_internalPfii	39	33.1178	33.585
	_Z12lud_diagonalPfii	40	0.7678	0.7782
	_Z13Iud_perimeterPfii	41	1.966	2.0726
	_Z12lud_internalPfii	42	14.7418	14.943
		43	0.7678	0.7782
	Z13Iud_perimeterPfii	44	0.9846	1.0381
	_Z12lud_internalPfii	45	3.6923	3.7422
		46	0.7678	0.7782

From the above table,

- In LUD and LPS benchamrks, as expected IPCs increased when GPGPUsim is run in profile based bypass mode.
- But in case of BFS benchmark, there is decrease in IPC when rin in profile based bypass mode. This may be because there are some cache friendly kernels in BFS.