Supplemental Material for
"Fine-Grained Memory Profiling of GPGPU Kernels"

Max von Buelow[1] , Stefan Guthe[1] and Dieter W. Fellner[1,2]

[1]Technical University of Darmstadt, Germany
[2]Fraunhofer IGD, Germany & Graz University of Technology, Institute of Computer Graphics and Knowledge Visualization, Austria

This supplemental material is organized as follows. In section A, we provide a complete comparison with actual cache hit rates on the set benchmark applications and section B describes the mistake of PPT-GPU-Mem in approximating the stack distance cache model (SDCM).

A. Cache Hit Rates

Table 1 shows bare cache hit rates estimated using our approach, PPT and the official NVIDIA profiler Nsight Compute on the Poly-Bench [GXS*12], Rodinia [CBM*09], Pannotia [CBRS13] and Tango [KKS*19] benchmark as well as our ray tracer.
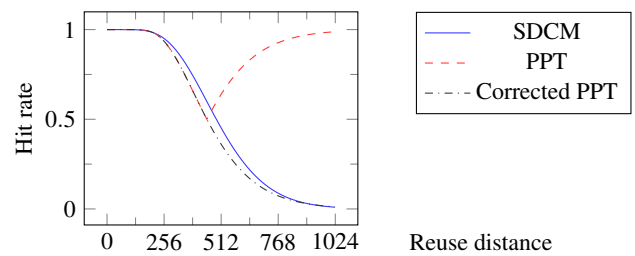
B. SDCM Approximation

In fig. 1, we visualize a plot of SDCM [AHH89] and compare it with the supposed approximation of PPT-GPU-Mem [ABE*21] on exemplary cache configurations. The original SDCM formulation is defined in eq. (1). Equation (2) is the approximation as defined and implemented in the work of ARAFA, BADAWY, ELWAZIR, et al. [ABE*21] using the *Q*-function. Note, that the function follows the opposite direction after reaching the discontinuity of the derivative. Equation (3) is a corrected version of eq. (2), where we removed the absolute value function.

$$p_{\text{sdcm}} = \sum_{a=0}^{A-1} \binom{D}{a} \left(\frac{A}{B}\right)^a \left(1 - \frac{A}{B}\right)^{D-a} \qquad (1)$$

$$p_{\text{ppt}} = 1 - Q\left(\frac{|A - 1 - D \cdot (A/B)|}{\sqrt{D \cdot (A/B) \cdot (1 - A/B)}}\right) \qquad (2)$$

$$p_{\text{pptcorr}} = 1 - Q\left(\frac{A - 1 - D \cdot (A/B)}{\sqrt{D \cdot (A/B) \cdot (1 - A/B)}}\right) \qquad (3)$$

Figure 1: *Plot of SDCM and its supposed approximations for an 8-way set associative cache with a capacity of 512 entries. "SDCM" denotes the direct implementation of the SDCM formula, "PPT" denotes the approximation used in the PPT-GPU-Mem implementation and "corrected PPT" is our correction to it.*

References

[ABE*21] ARAFA, YEHIA, BADAWY, ABDEL-HAMEED, ELWAZIR, AMMAR, et al. "Hybrid, scalable, trace-driven performance modeling of GPGPUs". *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '21. ACM, Nov. 2021. DOI: 10.1145/3458817.3476221 1.

[AHH89] AGARWAL, A., HENNESSY, J., and HOROWITZ, M. "An analytical cache model". *ACM Transactions on Computer Systems* 7.2 (May 1989), 184–215. DOI: 10.1145/63404.63407 1.

[CBM*09] CHE, SHUAI, BOYER, MICHAEL, MENG, JIAYUAN, et al. "Rodinia: A benchmark suite for heterogeneous computing". *2009 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, Oct. 2009, 44–54. DOI: 10.1109/iiswc.2009.5306797 1.

[CBRS13] CHE, SHUAI, BECKMANN, BRADFORD M., REINHARDT, STEVEN K., and SKADRON, KEVIN. "Pannotia: Understanding irregular GPGPU graph applications". *2013 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, Sept. 2013, 185–195. DOI: 10.1109/iiswc.2013.6704684 1.

[GXS*12] GRAUER-GRAY, SCOTT, XU, LIFAN, SEARLES, ROBERT, et al. "Auto-tuning a high-level language targeted to GPU codes". *2012 Innovative Parallel Computing (InPar)*. IEEE, May 2012, 1–10. DOI: 10.1109/inpar.2012.6339595 1.

[KKS*19] KARKI, AAJNA, KESHAVA, CHETHAN PALANGOTU, SHIVAKUMAR, SPOORTHI MYSORE, et al. "Detailed Characterization of Deep Neural Networks on GPUs and FPGAs". *Proceedings of the 12th Workshop on General Purpose Processing Using GPUs - GPGPU '19*. GPGPU '19. ACM Press, 2019. DOI: 10.1145/3300053.3319418 1.

Table 1: *Raw results from cache hit rate estimation of our model and PPT compared to measured values from NVIDIA Nsight Compute.*

Kernel	L1 our	L1 PPT	L1 meas.	L2 our	L2 PPT	L2 meas.	Kernel	L1 our	L1 PPT	L1 meas.	L2 our	L2 PPT	L2 meas.
cor1	50	50	50	75	50	50.25	sssp1	0	×	0	50	×	99.67
cor2	66.62	66.62	66.62	74.98	66.62	26.93	sssp2	0	×	0	50	×	51.46
cor3	77.56	76	77.54	83.95	79.56	68.04	sssp3	68.77	×	68.35	66.23	×	50.83
cor4	78.4	99.9	80.05	99.98	98.3	97.51	sssp4	0	×	0	50	×	2.93
cov1	50	50	50	75	50	50.25	DrtD	85.92	95.02	81.91	84.75	97.95	76.27
cov2	63.19	63.33	63.19	76.14	66.3	53.44	RrtD	74.02	92.32	71.12	74.48	96.84	61.22
cov3	78.45	99.91	79.82	99.98	98.32	97.75	IrtD	71.33	93.04	70.2	89.87	97.37	75.66
2mm1	92.74	94.13	93.53	99.87	99.42	99.97	MrtD	85.84	95.64	82.33	87.57	98.27	78.7
2mm2	92.76	94.13	93.54	99.87	99.42	99.85	SrtD	88.28	94.63	83.6	68.55	97.91	60.57
3mm1	93.29	96.31	93.63	99.74	99.58	99.79	DrtB	90.57	96.63	87.02	97.08	99.75	97.82
3mm2	93.28	96.31	93.64	99.74	99.58	99.65	RrtB	84.53	95.74	81.47	98.03	99.7	98.35
3mm3	93.29	96.31	93.62	99.74	99.58	99.66	IrtB	84.31	95.6	81.01	96.12	99.78	98.03
atax1	97.32	98.36	98.61	88.51	98.25	90.82	MrtB	89.52	96.74	85.29	97.65	99.77	97.1
atax2	90.99	93.35	94.35	94.06	92.86	55.48	SrtB	96.61	96.06	91.99	91.93	99.65	79.29
bicg1	55.27	55.37	55.38	75.06	55.54	50.1	DrtH	90.54	98.11	85.97	98.92	99.83	90.63
bicg2	88.32	89.14	86.48	65.01	89.18	76.42	RrtH	75.62	96.69	72.17	97.92	99.51	87.11
doitgen1	92.73	92.73	92.73	98.99	98.96	99	IrtH	70.38	97.26	68.8	98.96	99.76	92.08
doitgen2	0	0	0	50	0	58.2	MrtH	88.52	98.45	84.3	99.21	99.87	90.5
gemm	93.3	96.31	93.66	99.74	99.58	99.45	SrtH	91.14	97.9	86.06	92.24	99.1	74.83
gemver1	73.12	71.39	72.68	77.34	75.79	54.69	bptree1	70.37	68.03	72.47	81.92	81.7	66.74
gemver2	55.28	55.37	55.38	75.06	55.54	50.08	bptree2	66.63	59.37	69.05	81.74	78.64	66.75
gemver3	88.33	89.14	86.74	64.74	89.18	78.37	bfs1	0.04	0.25	0.04	49.98	0.37	18.98
gesummv	56.23	90.22	23.1	53.3	90.24	95.28	bfs2	0	0.03	0.01	49.99	0.03	20.02
gs1	0	0	0	0	0	29.72	bfs3	56.37	58.26	57.52	67.35	63.42	76.87
gs2	1.34	1.35	1.35	0.17	1.51	58.11	bfs4	39.8	39.43	69.51	69.55	39.85	78.13
gs3	61.39	81.16	62.49	75.48	81.74	50.61	cfd	16.16	28.46	20	85.2	67.67	74.86
mvt1	97.34	98.38	98.61	88.4	98.25	89.94	dwt2d1	0	0	0	50	0	80.05
mvt2	91.05	93.25	94.35	94.05	92.81	55.27	dwt2d2	0.72	0.62	0.68	51.84	8.52	50.58
syr2k	86.47	98.43	97.62	94.63	99.87	94.86	dwt2d3	2.53	3.81	1.66	55.69	13.88	56.79
syrk	97.7	98.61	98.18	99.91	99.87	98.91	dwt2d4	6.77	12.63	11.75	59.98	25.14	63.05
adi1	50.18	95.02	36.67	73.32	95.45	97.19	heartwall	31.59	31.55	45.16	65.1	48.21	77.71
adi2	33.33	33.33	33.26	33.33	33.33	44.83	hotspot	1.7	0.92	1.25	77.06	50.23	76.19
adi3	41.89	90.52	48.38	65.01	92.49	94.43	hotspot3D	50.38	50.25	50.48	59.5	59.75	46.2
adi4	44.44	44.44	44.44	64.28	44.44	51.24	huff1	56.33	51.63	56.41	54.65	63.62	33.78
conv2D	75.15	74.86	75.1	64.95	81.78	71	huff2	56.32	51.63	56.41	54.65	63.62	33.76
conv3D	58.24	58.26	60.29	64.95	70.74	52.4	lud1	48.38	48.39	48.39	74.19	48.39	92.87
fdtd2D	46.91	46.47	45.53	67.97	49.22	35.97	lud2	39.24	39.24	39.24	79.07	58.14	69.75
jacobi1D	54.21	54.21	54.21	51.42	55.52	68.79	lud3	30.44	25	30.44	85.01	71.67	71.74
jacobi2D	58.2	57.79	58.21	63.06	68.8	65.94	nw1	13.39	13.39	41.07	41.44	13.39	55.04
bc1	0	0	0	50	0	99.23	nw2	13.39	13.22	21.62	41.89	13.66	65.44
bc2	0.51	0.49	0.52	50.25	0.52	99.36	pathfinder	18.99	22.09	23.96	55.74	30.9	18.66
bc3	69.76	10.37	59.38	83.14	80	34.63	sc	24.22	23.95	24	51.06	24.82	3.99
bc4	92.3	94.54	66.78	74.44	96.16	87.08	backprop1	59.2	61.99	62.72	77.89	64.14	59.86
bc5	92.62	95.77	92.92	54.31	96.01	56.68	backprop2	72.47	72.44	74.41	77.22	73.87	60.44
bc6	92.06	92.34	86.54	56.06	93.02	76.05	AlexNet1	75.02	91.11	76.9	99.84	99.93	99.7
bc7	88.58	88.44	88.38	61.29	91.74	51.59	AlexNet2	87.16	95.25	79.9	99.64	99.92	99.78
bc8	44.75	7.77	44.75	66.51	63.5	81.61	AlexNet3	90.71	95.08	82.45	99.54	99.92	98.54
color1	73.64	69.79	73.75	70.4	82.09	50.42	AlexNet4	94.1	97.32	86.04	99.18	99.91	99.68
color2	23.1	23.14	21.8	61.55	23.15	23.42	LSTM	1.41	3.56	1.72	51.03	3.56	16.69
mis1	58.64	57.88	59.31	68.49	73.25	48.13	ResNet1	29.83	97.48	51.79	99.93	99.78	99
mis2	42.64	42.9	44.5	66.56	60.22	59.65	ResNet2	93.91	93.96	93.38	94.44	93.96	94.39
pagerank1	0	0	0	50	0	99.98	ResNet3	93.91	93.96	71.13	94.44	93.96	96.13
pagerank2	85.7	83.54	50.67	54.2	86.88	88.59	ResNet4	90.28	93.75	89.06	84.98	93.75	84.97
pagerank3	33.33	33.33	33.32	66.66	33.33	67.22	ResNet5	93.31	93.66	94.38	79.61	93.64	83.23