NC State University

Department of Electrical and Computer Engineering

ECE 463/563: Fall 2018 (Rotenberg)

Project #1: Cache Design, Memory Hierarchy Design

by

Emil Prisquilas Peter

NCSU Honor Pledge: "I have neither given nor received unauthorized aid on this test or assignment."

Student's electronic signature: __Emil Prisquilas Peter____________________________

(sign by typing your name)

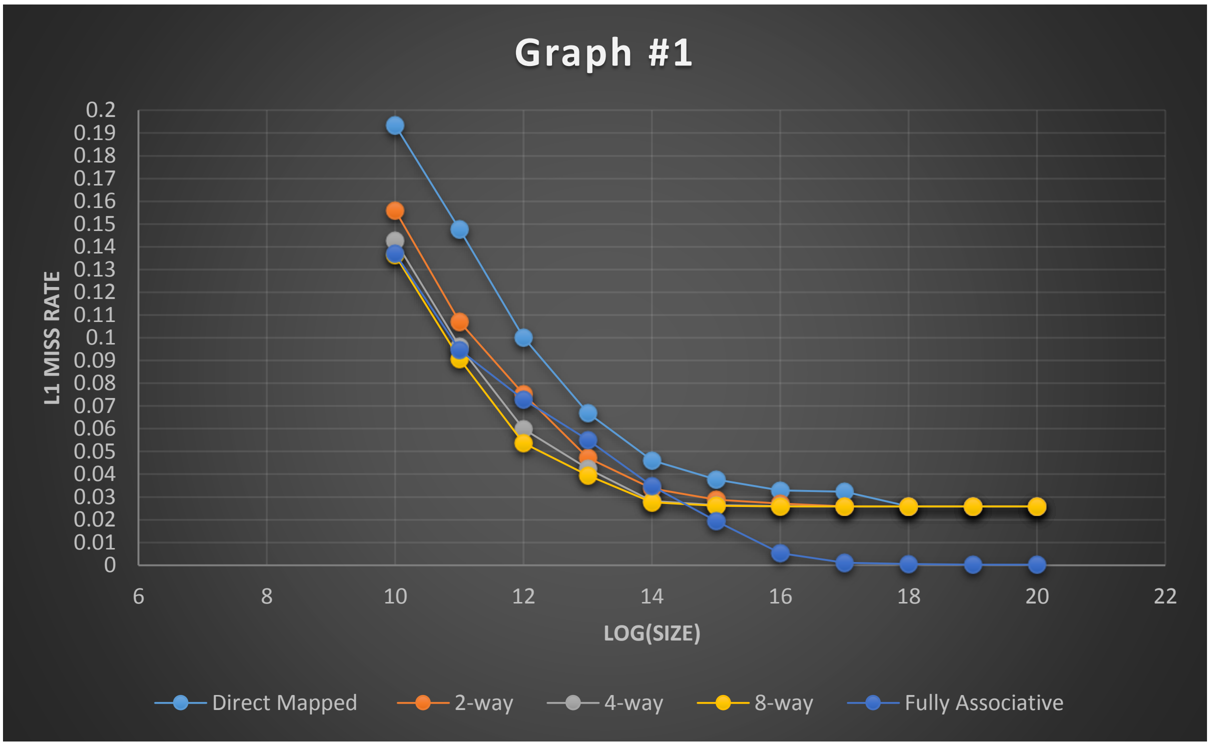
Course number: _____563______________

(463 or 563 ?)

Graph #1

Readings:

Size	Direct Mapped	2-way	4-way	8-way	Fully Associative	Log_2(Size)
1KB	0.1935	0.1560	0.1427	0.1363	0.137	10
2 KB	0.1477	0.1071	0.0962	0.0907	0.0947	11
4 KB	0.1002	0.0753	0.0599	0.0537	0.0729	12
8 KB	0.067	0.0473	0.0425	0.0395	0.0551	13
16KB	0.0461	0.0338	0.0283	0.0277	0.0349	14
32 KB	0.0377	0.0288	0.0264	0.0262	0.0194	15
64 KB	0.0329	0.0271	0.026	0.0259	0.0053	16
128 KB	0.0323	0.0259	0.0258	0.0258	0.0011	17
256 KB	0.0258	0.0258	0.0258	0.0258	0.0005	18
512 KB	0.0258	0.0258	0.0258	0.0258	0.0003	19
1 MB	0.0258	0.0258	0.0258	0.0258	0.0002	20



Discussions:

1. The miss rate decreases with increasing cache size for a given associativity. But after certain point there is hardly an decrease in the miss rate for the subsequent increase in cache size.
   The miss rate decrease with increase in cache size for a given cache size. But after certain point the is diminishing returns in terms of miss rate for any further increase in associativity.
2. The compulsory miss rate is approximately 0.025.
3. Conflict miss rates:
   a. For direct mapped cache: 0.195
   b. For 2-way set associative cache: 0.155
   c. For 4-way set associative cache: 0.145
   d. For 8-way set associative cache: 0.14

e. For fully associative cache: 0.14

Graph #2

We know,
$AAT=HT_{L1}+SRR*HT_{VC}+MR_{L1+VC}.Miss\_Penalty$
Blocksize=32
Miss_Penalty= 20.1
$HT_{VC}=0$
SRR=0
Therefore,
$AAT=HT_{L1}+MR_{L1+VC}.Miss\_Penalty$

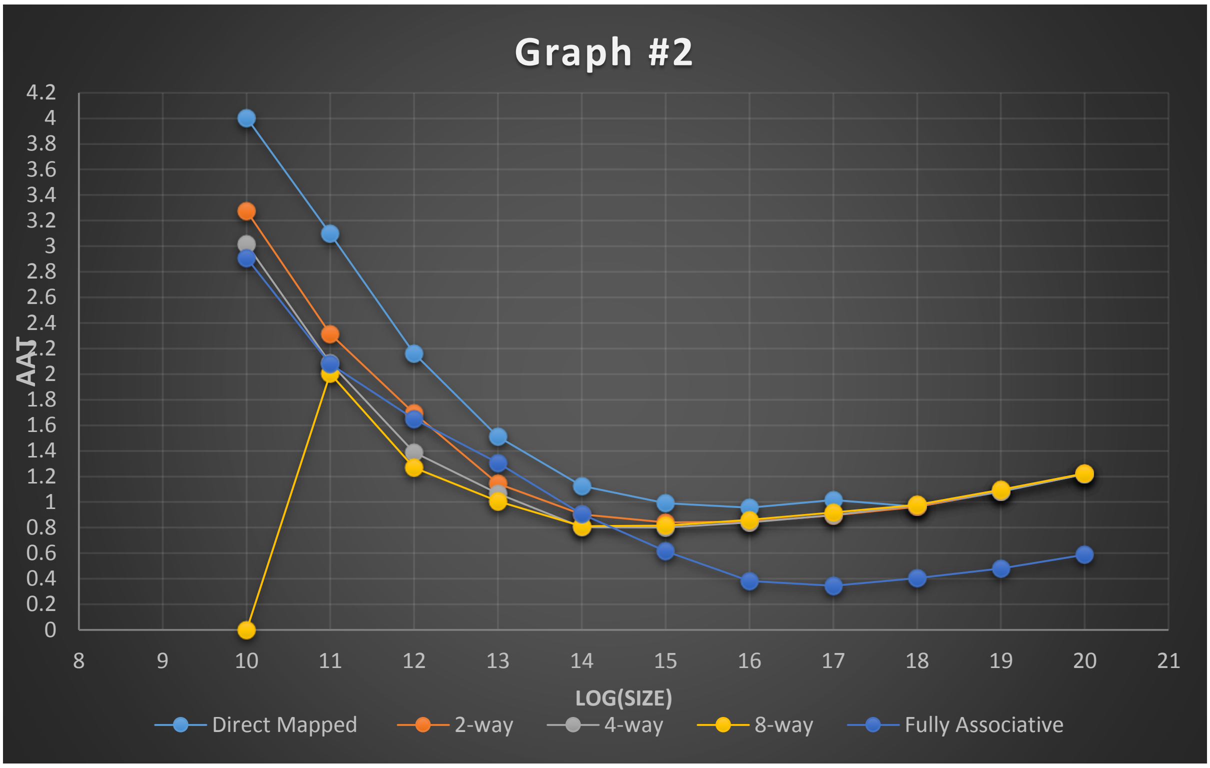
Readings:

Size	MR_{L1+VC} for direct Mapped (ns)	HT_{L1} for direct mapped (Access time(ns))	AAT (ns)	MR_{L1+VC} for 2-way (ns)	HT_{L1} for 2-way (Access time(ns))	AAT (ns)
1KB	0.1935	0.114797	4.004147	0.1560	0.140329	3.275929
2 KB	0.1477	0.12909	3.09786	0.1071	0.161691	2.314401
4 KB	0.1002	0.147005	2.161025	0.0753	0.181131	1.694661
8 KB	0.067	0.16383	1.51053	0.0473	0.194195	1.144925
16KB	0.0461	0.198417	1.125027	0.0338	0.223917	0.903297
32 KB	0.0377	0.233353	0.991123	0.0288	0.262446	0.841326
64 KB	0.0329	0.294627	0.955917	0.0271	0.300727	0.845437
128 KB	0.0323	0.3668	1.01603	0.0259	0.374603	0.895193
256 KB	0.0258	0.443812	0.962392	0.0258	0.445929	0.964509
512 KB	0.0258	0.563451	1.082031	0.0258	0.567744	1.086324
1 MB	0.0258	0.69938	1.21796	0.0258	0.706046	1.224626

Size	MR_{L1+VC} for 4-way (ns)	HT_{L1} (Access time(ns))	AAT (ns)	MR_{L1+VC} for 8-way (ns)	HT_{L1} (Access time(ns))	AAT (ns)
1KB	0.1427	0.14682	3.01509	0.1363	-	-
2 KB	0.0962	0.154496	2.088116	0.0907	0.180686	2.003756
4 KB	0.0599	0.185685	1.389675	0.0537	0.189065	1.268435

8 KB	0.0425	0.211173	1.065423	0.0395	0.212911	1.006861
16KB	0.0283	0.233936	0.802766	0.0277	0.254354	0.811124
32 KB	0.0264	0.27125	0.80189	0.0262	0.288511	0.815131
64 KB	0.026	0.319481	0.842081	0.0259	0.341213	0.861803
128 KB	0.0258	0.38028	0.89886	0.0258	0.401236	0.919816
256 KB	0.0258	0.457685	0.976265	0.0258	0.458925	0.977505
512 KB	0.0258	0.564418	1.082998	0.0258	0.578177	1.096757
1 MB	0.0258	0.699607	1.218187	0.0258	0.705819	1.224399

Size	MR_{L1+VC} for fully Associative (ns)	HT_{L1} (Access time(ns))	AAT (ns)
1KB	0.137	0.155484	2.909184
2 KB	0.0947	0.176515	2.079985
4 KB	0.0729	0.182948	1.648238
8 KB	0.0551	0.198581	1.306091
16KB	0.0349	0.205608	0.907098
32 KB	0.0194	0.22474	0.61468
64 KB	0.0053	0.276281	0.382811
128 KB	0.0011	0.322486	0.344596
256 KB	0.0005	0.396009	0.406059
512 KB	0.0003	0.475728	0.481758
1 MB	0.0002	0.588474	0.588674



Discussions:
The fully associative cache configuration gives the best yield since the AAT is the least as seen from the graph.

Graph #3

$AAT=HT_{L1}+SRR*HT_{VC}+MR_{L1+VC}.(HT_{L2}+MR_{L2}*\ Miss\_Penalty)$
HT_{L2}= 0.578177 ns

SRR=0 & HT_{VC}=0

Therefore,

$AAT=HT_{L1}+MR_{L1+VC}.(HT_{L2}+MR_{L2}*\ Miss\_Penalty)$

Readings:

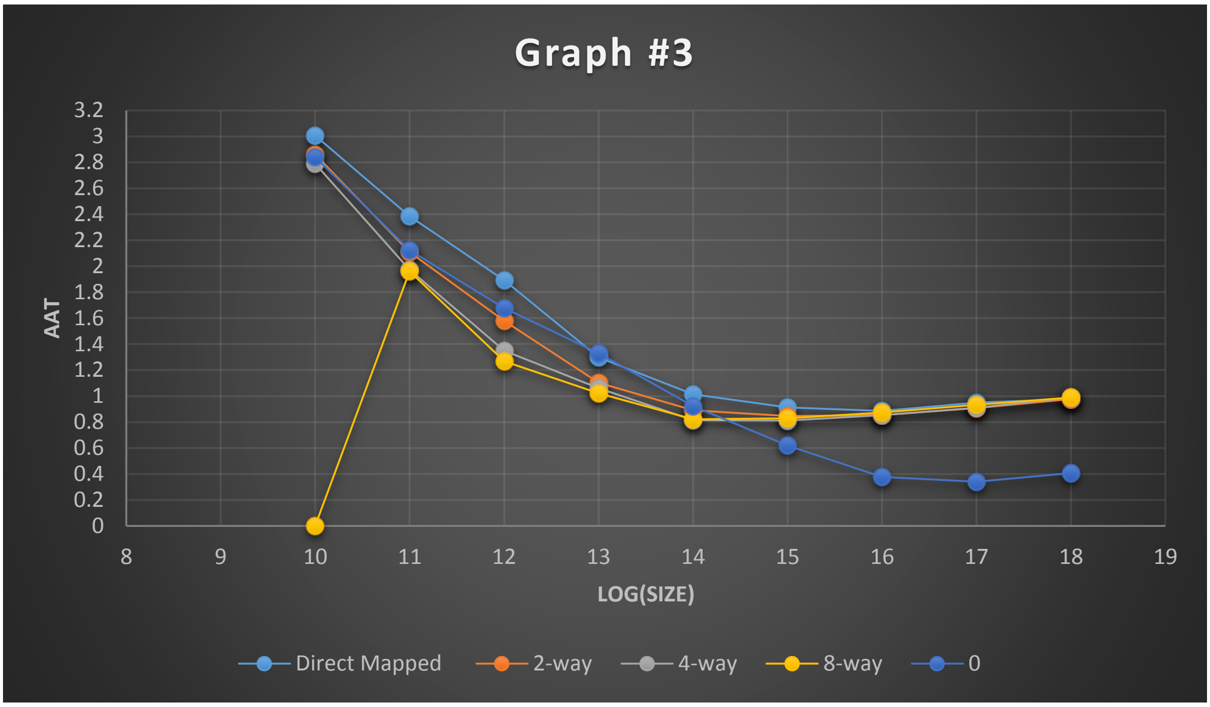
Size	MR_{L1} for Direct Mapped (ns)	MR_{L2} (ns)	HT_{L1} for direct mapped (Access time(ns))	AAT (ns)	MR_{L1} for 2-way (ns)	MR_{L2} (ns)	HT_{L1} (Access time(ns))	AAT (ns)
1KB	0.1935	0.7151	0.114797	3.0079	0.1560	0.8431	0.140329	2.860996612
2 KB	0.1477	0.7309	0.12909	2.3844	0.1071	0.8807	0.161691	2.110073157
4 KB	0.1002	0.8115	0.147005	1.89316	0.0753	0.8993	0.181131	1.579013528
8 KB	0.067	0.8131	0.16383	1.29756	0.0473	0.9324	0.194195	1.103593172
16KB	0.0461	0.8507	0.198417	1.01333	0.0338	0.9583	0.223917	0.891270183

32 KB	0.0377	0.8681	0.233353	0.91315	0.0288	0.9823	0.262446	0.844902298
64 KB	0.0329	0.8648	0.294627	0.88553	0.0271	0.9926	0.300727	0.854384797
128 KB	0.0323	0.8692	0.3668	0.94978	0.0259	1	0.374603	0.907577784
256 KB	0.0258	1	0.443812	0.977308	0.0258	1	0.445929	0.976845967

Size	MR$_{L1}$ 4-way (ns)	MR$_{L2}$ (ns)	HT$_{L1}$ (Access time(ns))	AAT (ns)	MR$_{L1}$ 8-way (ns)	MR$_{L2}$ (ns)	HT$_{L1}$ (Access time(ns))	AAT (ns)
1KB	0.1427	0.8975	0.14682	2.790791	0.1363	0.9304	-	-
2 KB	0.0962	0.9166	0.154496	1.973655	0.0907	0.9526	0.180686	1.9611
4 KB	0.0599	0.9428	0.185685	1.349792	0.0537	0.975	0.189065	1.2672
8 KB	0.0425	0.9706	0.211173	1.060756	0.0395	0.9937	0.212911	1.0207
16KB	0.0283	0.9968	0.233936	0.814487	0.0277	0.9968	0.254354	0.8225
32 KB	0.0264	0.9955	0.27125	0.812138	0.0262	1	0.288511	0.8276
64 KB	0.026	1	0.319481	0.854514	0.0259	1	0.341213	0.8741
128 KB	0.0258	1	0.38028	0.911197	0.0258	1	0.401236	0.9321
256 KB	0.0258	1	0.457685	0.988602	0.0258	1	0.458925	0.9898

Size	MR$_{L1}$ Fully Associative (ns)	MR$_{L2}$ (ns)	HT$_{L1}$ (Access time(ns))	AAT (ns)
1KB	0.137	0.9516	0.155484	2.842078249
2 KB	0.0947	0.9993	0.176515	2.123942562
4 KB	0.0729	0.9956	0.182948	1.676681903
8 KB	0.0551	0.9964	0.198581	1.328471353

16KB	0.0349	0.9997	0.205608	0.923576977
32 KB	0.0194	0.9845	0.22474	0.617942634
64 KB	0.0053	0.9137	0.276281	0.376197538
128 KB	0.0011	0.7652	0.322486	0.339956395
256 KB	0.0005	1	0.396009	0.406298089



Discussions:

1. The 2-way set associative and direct mapped cache shows their AATs close to the best AAT in graph#2 after the addition of L2 cache.
2. The fully associative cache configuration yields the best AAT. Its AAT is 0.2 lower to the best AAT found in graph #2.
3. The optimal AAT configuration is 256KB fully associative.

Calculating total area:

Without L2 cache-

Area= A_{L1}

= 0.766857492mm^2 {A_{L1} area= 0.766857492mm^2 $^+$ for 256KB fully-way associative}

With L2 cache

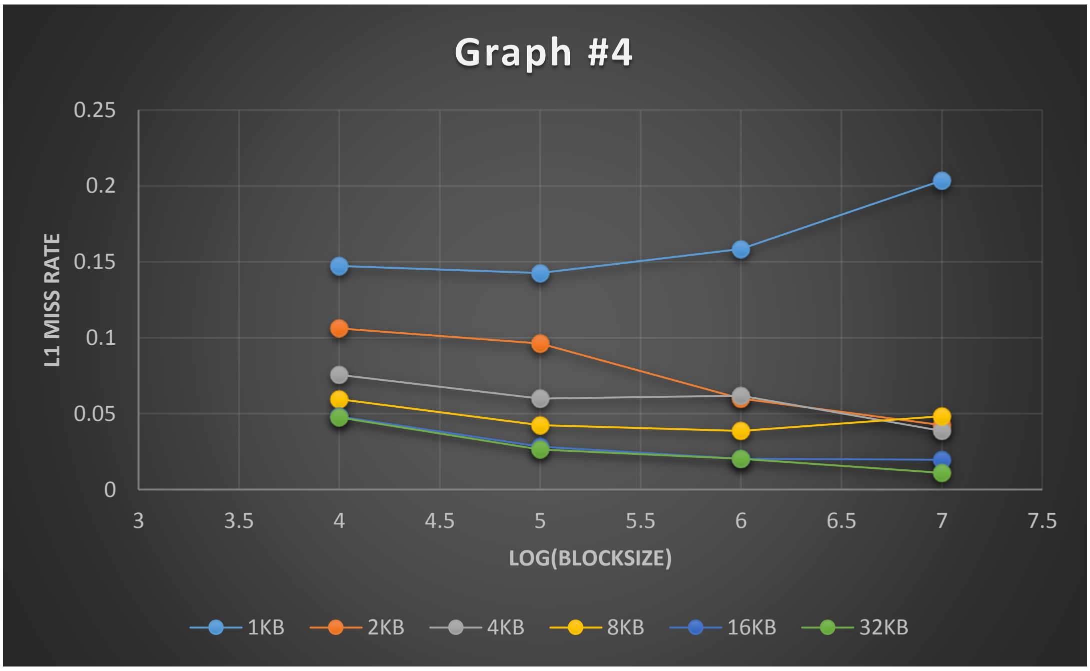
Area=A_{L1}+A_{L2}

= 0.766857492+2.640142073 {A_{L2} area=2.640142073mm^{2+} for 512KB 8-way associative}

=3.406999565mm^2

Graph #4

Size	1KB	2KB	4KB	8KB	16KB	32KB
MR_{L1} For blocksize 16 (ns)	0.1473	0.1062	0.0755	0.0595	0.0482	0.0475
MR_{L1} For blocksize 32 (ns)	0.1427	0.0962	0.0599	0.0425	0.0283	0.0264
MR_{L1} For blocksize 64(ns)	0.1584	0.0599	0.0619	0.0386	0.0204	0.0204
MR_{L1} For blocksize 128(ns)	0.2036	0.0425	0.0386	0.0483	0.0198	0.0111



Discussion:
From the graph it is evident that smaller caches prefer smaller block size whereas larger caches prefer larger block size. This occurs because for small caches for a larger block size, amount blocks in cache is eaten up by the larger blocksizes leading to cache pollution. Whereas for the same block size in a larger cache,the cache has sufficient amount of blocks to address the cache pollution issue.
Yes the balance shifts depending on the cache size. For smaller caches, cache pollution is more evident than the benefits of increased block size and vice-versa for larger caches.

Graph #5

$AAT=HT_{L1}+SRR*HT_{VC}+MR_{L1+VC}.(HT_{L2}+MR_{L2}* Miss\_Penalty)$
$HT_{L2}= 0.578177$

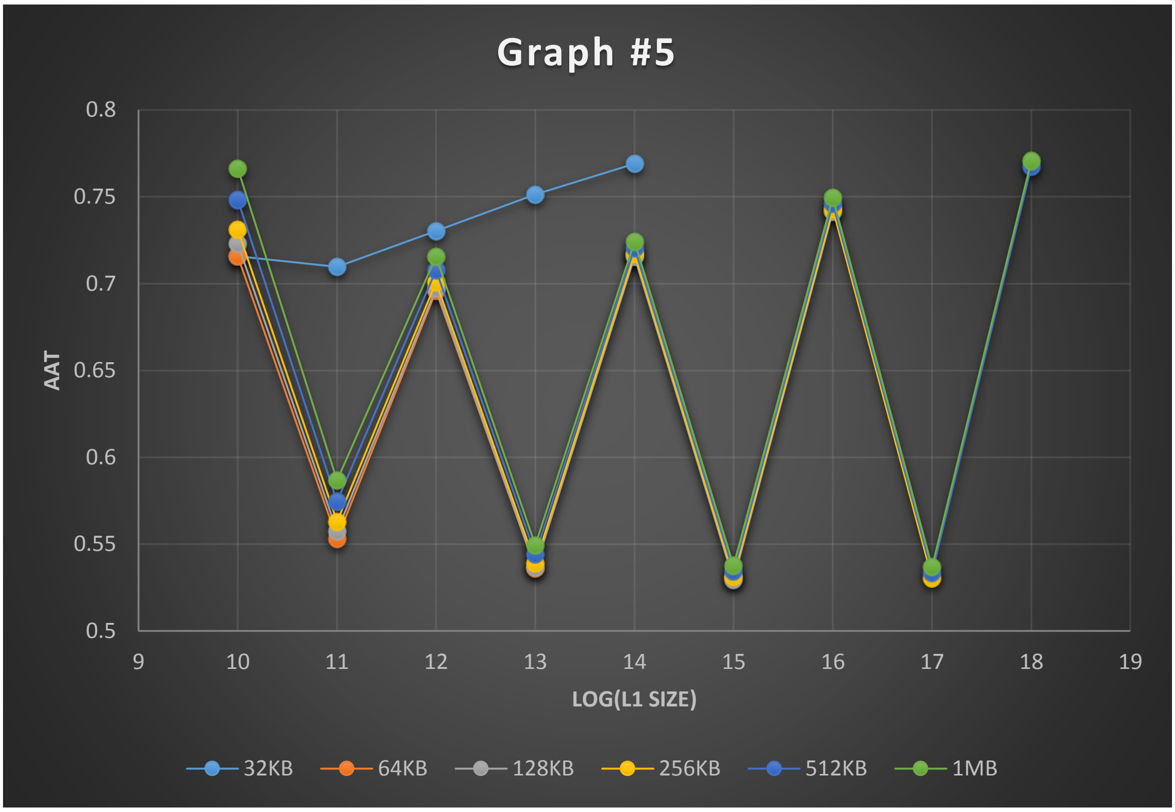
$AAT=HT_{L1}+MR_{L1+VC}.(HT_{L2}+MR_{L2}* Miss\_Penalty)$

Readings:

Size of L1 cache	HT$_{L1}$ for direct mapped (Access time(ns))	MR$_{L2}$ for 32KB (ns)	MR$_{L1}$ (ns)	AAT (ns)	Area	MR$_{L2}$ for 64KB (ns)	MR $_{L1}$ (ns)	AAT (ns)	Area
1KB	0.14682	0.1427	0.184	0.7157522	0.27285583	0.1427	0.1814	0.715815	0.511467091
2 KB	0.154496	0.0962	0.2728	0.70974	0.2608332994	0.0962	0.2691	0.553162	0.97597997
4 KB	0.185685	0.0599	0.4381	0.73043	0.279811188	0.0599	0.4326	0.695781	0.397958164
8 KB	0.211173	0.0425	0.6181	0.75144	0.310604791	0.0425	0.6105	0.536021	0.428751767
16KB	0.233936	0.0283	0.9266	0.76917	0.348112328	0.0283	0.9167	0.716788	0.466259304
32 KB	0.27125					0.0264	0.9841	0.531211	0.596965292
64 KB	0.319481								
128 KB	0.38028								
256 KB	0.457685								
HT$_{L2}$	0.288511					0.341213			

Size of L1 cache	MR_{L2} for 128KB (ns)	MR_{L1} (ns)	AAT (ns)	Area	MR_{L2} for 256KB (ns)	MR_{L1} (ns)	AAT (ns)	Area
1KB	0.1427	0.1809	0.722946	0.575048282	0.1427	0.1809	0.731179	1.308655484
2 KB	0.0962	0.2683	0.557389	0.578595693	0.0962	0.2683	0.562939	1.312202895
4 KB	0.0599	0.4309	0.697329	0.597573887	0.0599	0.4309	0.700785	1.331181089
8 KB	0.0425	0.608	0.536437	0.62836749	0.0425	0.608	0.538888	1.361974692
16KB	0.0283	0.9117	0.715642	0.665875027	0.0283	0.9117	0.717275	1.399482229
32 KB	0.0264	0.978	0.529559	0.796581015	0.0264	0.978	0.531082	1.530188217
64 KB	0.026	0.995	0.741592	0.862222704	0.026	0.995	0.743092	1.595829906
128 KB					0.0258	1	0.53042	1.960558502
256 KB								
HT_{L2}	0.401236				0.458925			

Size of L1 cache	MR_{L2} for 512KB (ns)	MR_{L1} (ns)	AAT (ns)	Area	MR_{L2} for 1MB (ns)	MR_{L1} (ns)	AAT (ns)	Area
1KB	0.1427	0.1809	0.748196	2.655257021	0.1427	0.1809	0.76641	4.889316353
2 KB	0.0962	0.2683	0.574411	2.658804432	0.0962	0.2683	0.58669	4.892863764
4 KB	0.0599	0.4309	0.707928	2.677782632	0.0599	0.4309	0.715574	4.911841964
8 KB	0.0425	0.608	0.543957	2.708483633	0.0425	0.608	0.549381	4.942542965
16KB	0.0283	0.9117	0.72065	2.746083766	0.0283	0.9117	0.724262	4.980143098
32 KB	0.0264	0.978	0.53423	2.819496348	0.0264	0.978	0.5376	5.05355568
64 KB	0.026	0.995	0.746193	2.942431443	0.026	0.995	0.749511	5.176490775
128 KB	0.0258	1	0.533497	3.307160039	0.0258	1	0.53679	5.541219371
256 KB	0.0258	1	0.767433	3.781436875	0.0258	1	0.770726	6.015496207
HT_{L2}	0.578177				0.705819			



Discussion:

1. 128KB L1 cache with 256 KB L2 cache yields the best AAT.
2. 32 KB L1 cache with 256 KB L2 cache gives the best yield with minimum area.

Graph #6

AAT=HT$_{L1}$+SRR*HT$_{VC}$+MR$_{L1+VC}$(HT$_{L2}$+MR$_{L2}$* Miss_Penalty)
HT$_{L2}$= 0.578177

AAT=HT$_{L1}$+MR$_{L1+VC}$(HT$_{L2}$+MR$_{L2}$* Miss_Penalty)

Readings:

For Direct mapped cache without victim cache:

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	AAT (ns)
1KB	0.14682	0.1935	0.1338	0.73324
2 KB	0.154496	0.1477	0.1752	0.725022
4 KB	0.185685	0.1002	0.2585	0.740499
8 KB	0.211173	0.067	0.3864	0.754399
16KB	0.233936	0.0461	0.5619	0.770328
32 KB	0.27125	0.0377	0.6882	0.805611

For Direct mapped cache with 2-entry victim cache:

HT_{VC} =0.131305 ns

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	SRR (ns)	AAT (ns)
1KB	0.14682	0.1608	0.161	0.1931	0.747407
2 KB	0.154496	0.1231	0.2104	0.1471	0.736409
4 KB	0.185685	0.0885	0.2913	0.0989	0.747047
8 KB	0.211173	0.0601	0.4295	0.0644	0.758976
16KB	0.233936	0.0416	0.6187	0.041	0.770846
32 KB	0.27125	0.0343	0.7455	0.0278	0.800574

For Direct mapped cache with 4-entry victim cache:

HT_{VC} =0.131305 ns

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	SRR (ns)	AAT (ns)
1KB	0.14682	0.1473	0.1758	0.1931	0.742932
2 KB	0.154496	0.1133	0.2285	0.1471	0.578344
4 KB	0.185685	0.0833	0.3096	0.0989	0.714278
8 KB	0.211173	0.0365	0.4561	0.0644	0.355528
16KB	0.233936	0.0401	0.6427	0.041	0.722774
32 KB	0.27125	0.0329	0.7789	0.0278	0.529955

For Direct mapped cache with 8-entry victim cache:

HT_{VC} =0.135715 ns

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	SRR (ns)	AAT (ns)
1KB	0.14682	0.1323	0.1957	0.1931	0.73858
2 KB	0.154496	0.1032	0.2509	0.1471	0.574975
4 KB	0.185685	0.0767	0.3362	0.0989	0.711963
8 KB	0.211173	0.0519	0.4967	0.0644	0.544317
16KB	0.233936	0.0378	0.6813	0.041	0.721604
32 KB	0.27125	0.032	0.8024	0.0278	0.530673

For Direct mapped cache with 16-entry victim cache:

HT_{VC} =0.142543 ns

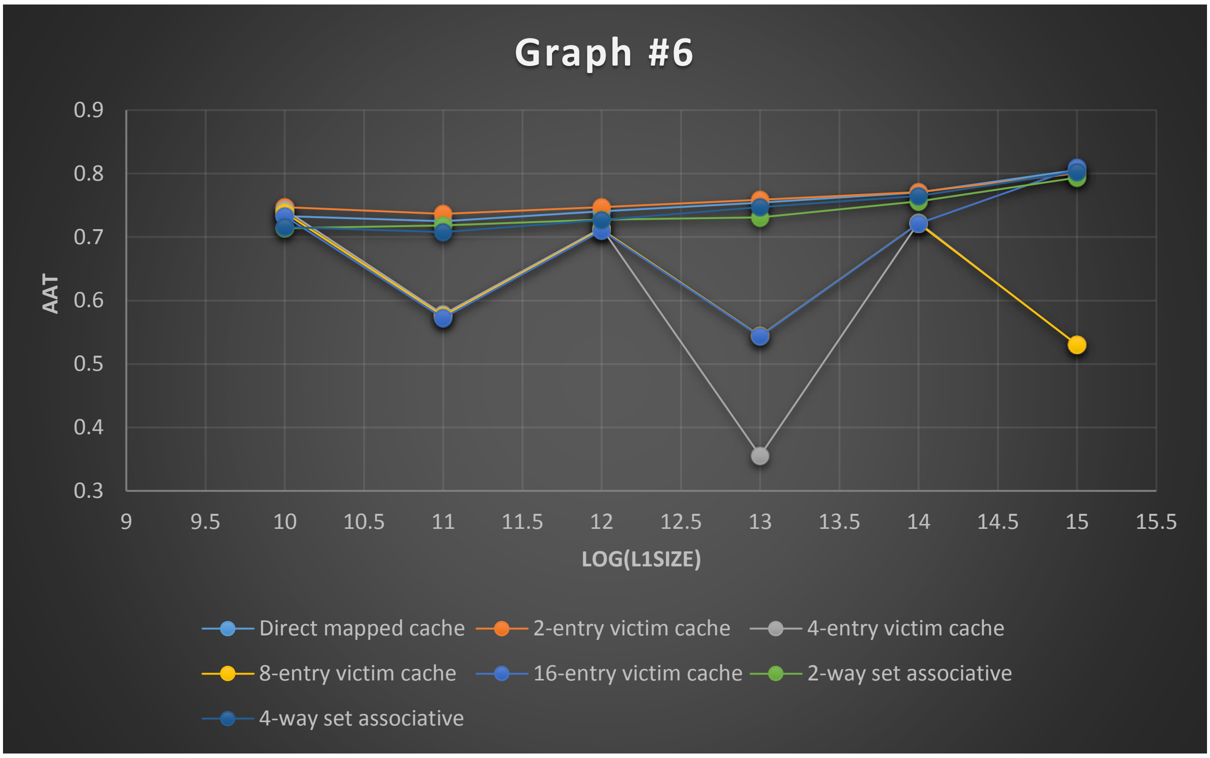
Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	SRR (ns)	AAT (ns)
1KB	0.14682	0.1125	0.2301	0.1931	0.733045
2 KB	0.154496	0.0903	0.2868	0.1471	0.57233
4 KB	0.185685	0.065	0.3977	0.0989	0.710367
8 KB	0.211173	0.0449	0.5753	0.0644	0.543703
16KB	0.233936	0.0357	0.7213	0.041	0.721294
32KB	0.27125	0.0306	0.8399	0.0278	0.80949

For a 2-way set associative L1 cache without Victim cache:

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	AAT (ns)
1KB	0.140329	0.156	0.1659	0.713754
2 KB	0.161691	0.1071	0.2416	0.71833
4 KB	0.181131	0.0753	0.3439	0.727327
8 KB	0.194195	0.0473	0.5477	0.731049
16KB	0.223917	0.0338	0..7668	0.756399
32 KB	0.262446	0.0288	0.9004	0.793496

For a 4-way set associative L1 cache without Victim cache:

Size	HT_{L1} for direct mapped (Access time(ns))	MR_{L1+VC} (ns)	MR_{L2} (ns)	AAT (ns)
1KB	0.14682	0.1427	0.1814	0.715815
2 KB	0.154496	0.0962	0.2691	0.707658
4 KB	0.185685	0.0599	0.4326	0.72697
8 KB	0.211173	0.0425	0.6105	0.747194
16KB	0.233936	0.0283	0.9167	0.765039
32 KB	0.27125	0.0264	0.9841	0.802461



Discussion:

1. Yes adding a victim cache to a direct-mapped cache make sits performance comparable to that of 2-way set associative. This is achieved for 32KB L1 caches and for 2-entry victim cache configuration.
2. 8KB L1 cache with 4-entry victim cache has the best yield.
3. 8KB L1 cache with 16-entry has smallest total area with yield within 5%of the best AAT.