University of Pennsylvania
Department of Electrical and System Engineering
System-on-a-Chip Architecture

ESE5320, Fall 2022 — Midterm Solutions — Wednesday, October 5

- Exam ends at 11:45AM; begin as instructed (target 10:15AM)
  Do not open exam until instructed.
- Problems weighted as shown.
- Calculators allowed.
- Closed book = No text or notes allowed.
- Show work for partial credit consideration. All answers here.
- Unless otherwise noted, answers to two significant figures are sufficient.
- Sign Code of Academic Integrity statement (see last page for code).

I certify that I have complied with the University of Pennsylvania's Code of Academic Integrity in completing this exam.

Name: Solutions

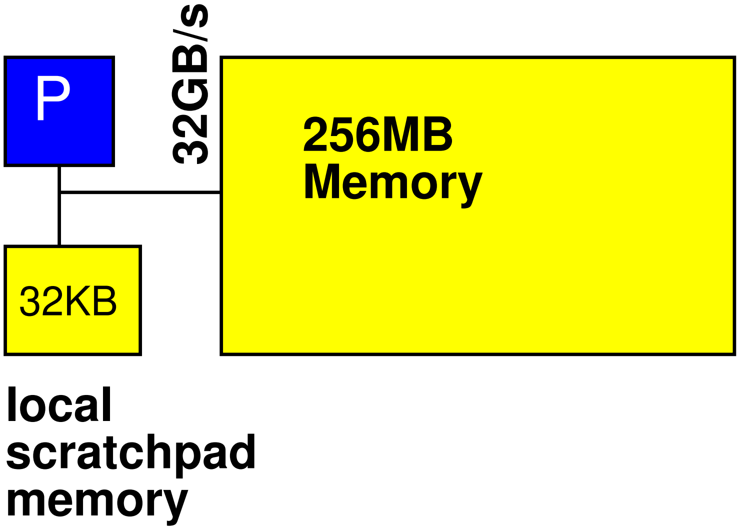
1	2a	2b	3	4	5	6	7	8	Total
10	5	5	10	10	20	10	10	20	100

Consider the following (very simplified) code to find paths in a gird with obstacles.

```
#define MAX_TARGETS 100
#define MAX_TIMESTEPS 100
#define WIDTH 1000
#define HEIGHT 1000
#include <stdint.h>
#include <stdlib.h>
#include <stdbool.h>

#define BLOCKED (1<<15)-1
#define USED (1<<14)
#define SOURCE 0
#define TARGET ((1<<14)|2)
#define NOPATH (MAX_TIMESTEPS+1)
#define FREE NOPATH

#define MASK16 ((1<<16)-1)

typedef struct pair_xy
{
  uint16_t y;
  uint16_t x;
} pair_xy;

uint16_t min(uint16_t a, uint16_t b); // assume single instruction
uint16_t max(uint16_t a, uint16_t b); // assume single instruction
void read_obstacles(uint16_t g[HEIGHT][WIDTH]);
            // marks obstacles BLOCKED in g
void read_sources_and_targets(uint16_t g[HEIGHT][WIDTH],
                              pair_xy source[MAX_TARGETS],
                              pair_xy target[MAX_TARGETS]);
   // loads targets into target, sources into source
   // marks target in g
void share_paths(pair_xy paths[MAX_TARGETS][MAX_TIMESTEPS]);
   // reports out results
void reset_tgrid(uint16_t tgrid[MAX_TIMESTEPS][HEIGHT][WIDTH],
                 pair_xy path[MAX_TARGETS][MAX_TIMESTEPS],
                 pair_xy target[MAX_TARGETS],pair_xy source[MAX_TARGETS],
                 int targ) {
  //   Need to know when path done to stop following
  bool pdone[MAX_TARGETS];
  for (int t=0;t<MAX_TARGETS;t++) { pdone[t]=false; } // loop F

  for (int step=0;t<MAX_TIMESTEPS;t++) { // loop G -- should be step++
    for (int y=0;y<HEIGHT;y++)  // loop H
      for (int x=0;x<WIDTH;x++) // loop I
        { tgrid[step][y][x]=NOPATH; }
    if (t==0) // should be (step==0)
      for (int s=0;s<MAX_TARGETS;s++) // loop J
        { tgrid[0][source[s].y][source[s].x]=0; }
    for (int t=0;t<targ;t++) // loop K
      if (!pdone[t]) {
          tgrid[step][path[t][step].y][path[t][step].x]=USED;
          if ((path[t][step].x==target[t].x)
              && (path[t][step].y==target[t].y)) // end of path
            { pdone[t]=true; }
      }
}
```

```
uint16_t new_cost(uint16_t grid[HEIGHT][WIDTH],
                     uint16_t tgrid[MAX_TIMESTEPS][HEIGHT][WIDTH],
                  uint16_t t, uint16_t y, uint16_t x){
  uint16_t below=max(grid[y-1][x],tgrid[t][y-1][x]);
  uint16_t above=max(grid[y+1][x],tgrid[t][y+1][x]);
  uint16_t left=max(grid[y][x-1],tgrid[t][y][x-1]);
  uint16_t right=max(grid[y][x+1],tgrid[t][y][x+1]);
  return(min(min(below,above),min(left,right))+1);
}

uint32_t predecessor(uint16_t tgrid[MAX_TIMESTEPS][HEIGHT][WIDTH],
                     uint16_t t, uint16_t y, uint16_t x) {
  if (tgrid[t-1][y-1][x]==t-1) { return((y-1)<<16 | x); }
  if (tgrid[t-1][y+1][x]==t-1) { return((y+1)<<16 | x); }
  if (tgrid[t-1][y][x-1]==t-1) { return(y<<16 | (x-1)); }
  if (tgrid[t-1][y][x+1]==t-1) { return(y<<16 | (x+1)); }
  abort(); // inconsistency: predecessor not find match
}

void find_paths () {
  uint16_t grid[HEIGHT][WIDTH];
  uint16_t tgrid[MAX_TIMESTEPS][HEIGHT][WIDTH];
  pair_xy target[MAX_TARGETS];
  pair_xy source[MAX_TARGETS];
  pair_xy path[MAX_TARGETS][MAX_TIMESTEPS];
  bool found;
  int found_time;

  read_obstacles(grid);
  read_sources_and_targets(grid,source,target);
  reset_tgrid(tgrid,path,target,source,0);

  for(int targ=0;targ<MAX_TARGETS;targ++) { // loop A
    found=false;
    for (int t=0;((!found) || (t<MAX_TIMESTEPS));t++) // loop B
      for (int y=0;y<HEIGHT;y++) // loop C
        for (int x=0;x<WIDTH;x++) { // loop D
          uint16_t cost=new_cost(grid,tgrid,t,y,x);
          if (grid[y][x]==FREE) { tgrid[t+1][y][x]=cost; }
          bool found_now=((target[targ].x==x)&&(target[targ].y==y)
                          &&(cost<NOPATH));
          found|=found_now;
          if (found_now) { found_time=t; }
        }
    path[targ][found_time].x=target[targ].x;
    path[targ][found_time].y=target[targ].y;
    for(int it=found_time;it>0;it--) { // loop E
      uint32_t pxy=predecessor(tgrid,it,path[targ][it].y,path[targ][it].x);
      path[targ][it-1].y=pxy>>16;
      path[targ][it-1].x=(pxy&MASK16);
    }
    reset_tgrid(tgrid,path,target,source,targ); // loop F-K inside
  }
  share_paths(path);
}
```

We start with a baseline, single processor system as shown.



- For simplicity throughout, we will treat non-memory indexing adds (subtracts count as adds), compares, min, max, abs, divides, multiplies, shifts, and logical operations (binary and bitwise) as the only compute operations. We'll assume the other operations take negligible time or can be run in parallel (ILP) with the adds, multiplies, and memory operations. (Some consequences: You may ignore loop and conditional overheads in processor runtime estimates; you may ignore computations in array indecies.)
- Baseline processor can execute one multiply, divide, compare, min, max, abs, or add per cycle and runs at 1 GHz.
- Data can be transfered from the 256 MB main memory at 32 GB/s when streamed in chunks of at least 256B. Assume `for` loops that only copy data can be auto converted into streaming operations.
- Non-streamed access to the main memory takes 20 cycles.
- Baseline processor has a local scratchpad memory that holds 32KB of data. Data can be streamed into the local scratchpad memory at 32 GB/s. Non-streamed accesses to the local scratchpad memory takes 1 cycle.
- By default, all arrays live in the main memory.
- Arrays `source`, `target`, and `pdone` live in local scratchpad memory.
- Assume scalar (non-array) variables can live in registers.
- Assume all additions are associative.

1. Simple, Single Processor Resource Bounds

   Give the single processor resource bound time for compute operations and memory access for each loop directly inside loop A and the total bound for loop A.

   `new_cost` is 8 large memory reads for 160; and 8 compute cycles. `predecessor` is 4 large memory reads for 80; and 16 compute cycles (counting t-1 only once, $y << 16$ only once).

loop	Compute	Memory
B	$100 \times 1000 \times 1000 \times (8+7)$ $= 1.5 \times 10^9$	$100 \times 1000 \times 1000 \times (160 + 2 \cdot 20 + 2 \cdot 1)$ $= 2.02 \times 10^{10}$
E	$100 \times (16+2)$ $=1{,}800$	$100 \times (80 + 4 \cdot 20)$ $=16{,}000$
F	0	100
G	0 $+100$ $+100 \times 100 \times 3$ $= 300100$	stream $\frac{100 \times 1000 \times 1000 \times 2}{32}$ $+100 \times (20+2)$ $+100 \times 100 \times (5 \cdot 20 + 1 \cdot 4)$ $\approx 7.3 \times 10^6$
A	$\approx 1.5 \times 10^9$	$\approx 2.0 \times 10^{10}$

2. Based on the simple, single processor mapping from Problem 1:

   (a) What loop is the bottleneck? Consider both compute and memory. (circle one)

   [✓ B]

   E

   F

   G

   (b) What is the Amdahl's Law speedup if you only accelerate the identified function? Consider both compute and memory.

   $\frac{2.17\times10^{10}}{7.294\times10^{6}} \approx 3000$

3. Parallelism in Loops

   (a) Classify the following loops as data parallel or not? (loop bodies could be executed concurrently)

   (b) Explain why or why not?

Loop	Data Parallel?	Why or why not?
A	N	depends on paths used by previous `targ`
B	N	`tgrid[t+1]` depends on `tgrid[t]`
C	Y or N	all grid spots are indepenent; associative reduce on found, which you can count as not data parallel now
D	Y or N	all grid spots are indepenent; associative reduce on found which you can count as not data parallel now
E	N	need path element from previous iteration to find current
F	Y	No dependencies
G	N	Need pdone calculated on each iteration
H	Y	No dependencies
I	Y	No dependenceies
K	Y	`targs` are independent of each other

4. What is the critical path for the body of loop A?

   Asking for critical path for this entire loop A was too much. Should have asked less.

(This page intentionally left mostly blank for answers.)

Where	Detail	Cycles
B	B sequentialized, so multiply 100 by all things in B.	
	run all new_costs in C, D simultaneously; new_cost itself has all reads in parallel, all maxes in parallel then min+min+add one for 20+4;	100×24
	read grid[y][x] and tgrid[t+1][y][x] in parallel with new_cost grid, tgrid reads.	
	test grid[y][x] and tgrid write cost 20+1 (not strictly necessary to wait on tgrid write)	100×21
	found_now is 3 (reads with above, all comparisons in parallel, then two ands)	100×3
	found|=found_now is an associative reduce – can be done in $\log_2(1000 \times 1000)$ cycles	$\times 100 \times 20$
	Since we haven't done reduces, yet, can count this as sequential 10^6	or 100×10^6
after B	path[targ][found][time] and target[targ] 20+1 (run x and y in parallel)	21
E	read path[targ][it].x and y in parallel (technically can avoid reading) – 20	100×20
	read all predecessor tgrids in parallel for 20.	100×20
	predecessor: worst-case evaluate 4 ==t-1's then shift and add in parallel then or for return = 6 (could compute part of the return value calculations in parallel and reduce).	100×4
	path[targ][it-1] for x and y in parallel; each one compute and one write 21 (mostly do not need to wait for write to complete, so could omit)	100×21
F	clear pdone in parallel = 1	1
G	G must be sequentialized, so all multiplied by 100 below.	
	clear tgrid in H, I in parallel - 20	100×20
	read sources in parallel - 1	100×1
	set tgrid in parallel - 20 (only once for step==0)	20
K	K is parallel.	
	read pdone, target - 1	100×1
	!pdone (could count or not, we didn't explicitly say inversions should count)	0
	read path[t][step] x and y - 20	100×20
	write tgrid - 20 (but don't really need to wait for complete)	100×20
	already read path[t][step]	0
	compares in parallel - 1, and - 1 set pdone - 1	100×3
Total	(exploit reduce)	$\approx 2.2 \times 10^4$
	(not exploit reduce)	$\approx 10^8$

5. Revise the body of `loop B` to minimize the memory resource bound by exploiting the scratchpad memory and streaming memory operations.

   (a) Identify the array or arrays whose memory operations account for most of the time in the loop.

   tgrid, grid

   (b) How would you use the scratchpad memory to reduce the time required to access memory? (You don't need to give code, but you need to describe clearly how the code would change. You may show code if that is the most efficient way to communicate your changes.)

   Stream read one row (all x's for a y value) at a time in from **`grid`** and **`tgrid`**. Keep 3 rows of each in scratchpad (for -1, 0, and +1 y offset) for current t and one for next t (t+1, 0 offset). Stream read at top of loop C when y changes; stream write after completing each iteration of D. Inside the loop D, use local versions of tgrid, grid; adjust indexing and **`new_cost`** accordingly.

   (c) Account for total memory usage in the local scratchpad (use provided table)

Variable	Size (Bytes)
source	400
target	400
pdone	100
local_grid	$2 \times 3 \times 1000$=6000
local_tgrid_in	6000
local_tgrid_out	2000

   (d) Estimate the new memory resource bound for your optimized `loop B`.

   $100 \times 1000 \times 1000 \times (8 + 2 \cdot 1 + 2 \cdot 1)$ for local reads; plus $3 \times \frac{100 \times 1000 \times 1000 \times 2}{32}$ for streaming in tgrid and grid and out tgrid.

   $= 12 \times 10^8 + (3/16) \times 10^8 \approx 1.2 \times 10^9$

(This page intentionally left mostly blank for answers.)

6. Assume you have a vector processor that can provide 16 vector lanes for 16b (including uint16_t) operations. The vector processor can read or write 256b from its local scratchpad memory in one cycle using a vector read or vector write operation. Build on your memory optimizations in the previous question. If necessary, describe any additional memory optimizations you may do beyond the previous question for this vector case. Assuming perfect vectorization, what is the impact on the compute and memory resource bounds for `loop B`? (state new bounds; show work.)

   compute bound – vectorize all operations – $\frac{1.5\times10^9}{16} \approx 9.4 \times 10^7$

   memory bound – now perform vector reads to grid and tgrid and targ (streaming time remains unchanged) – $\frac{12\times10^8}{16} + \frac{3\times10^8}{16} \approx 9.4 \times 10^7$

7. Identify concurrency opportunities between loops.

   Which loops can run concurrently, as separate processes, to increase the **throughput** for loop A? If they cannot, explain what prevents concurrency. If they can, explain why and what conditions need to be met for the concurrency to work.

	Concurrent?	How or Why not?
B + E	N	Must find path to target in B before can trace it back in E.
E + F	Y	pdone is not used in E, so fine to clear it.
F + B	Y	pdone is not used in F, so fine to clear it.

   Had actually meant to ask about E and G, which cannot run concurrently. Need to trace back path in E before can clear it in G.

   Hint: we're giving you that there is concurrency between F and B. Note that they both iterate over timesteps. Identify the constraints required for them to run concurrently.

   ...and had meant to ask about G+B. What matters here is that timesteps are cleared before they are used in B. So, as long as G is 3 rows (y values) ahead of B, it can still be clearing while B is running on the next target.

8. Map the `loop A` computation to a system composed of two simple processors (1 GHz as previously outlined), two fast processors (2 GHz, with everything running 2× as fast except data transfer from main memory), and four vector processors (Problem 6). Assume each processor has its own scratchpad and has a separate path to the large memory so they can all simultaneously stream at full rate.[1]

   (a) Describe how you would map the computation onto these heterogeneous computing resources.

   Divide the computation into 4 parts among rows (y in grid/tgrid); that will mean 250 rows per part. Use this to run B on the 4 vector processors, one for each of the 4 parts.

   Run E and F on a single, slow processor, synchronized to start after all 4 vector processors complete B. The time is small, so it could run anywhere. It is sequentialized with B, so doesn't really need to run on separate processor.

   Run G on remaining 3 processor. Since streaming time is the same, it won't make much difference running on the slow processor since streaming time dominates. G can mostly overlap with the next iteration B due to observation in 7 that it can run while B is running as long as it stays 3 rows ahead of B.

   (b) As necessary, describe how you would use the scratchpad memories as necessary beyond what you've already answered in Problems 5 and 6. [no further change is a possible answer here.]

   No change – each vector processor uses its own scratchpad, but same basic strategy from 6.

   (c) Estimate the performance your mapping achieves in cycles per `loop A` iteration.

What	Calculation	Cycles
B	$\frac{(9.4+9.4)\times 10^7}{4}$	4.7×10^7
E	1800+16000	17,800
F	100	100
G	$\frac{7.3\times 10^6}{3}\times\frac{3}{100}$	7.3×10^4
total		4.7×10^7

   Note: Sequentialling G (running on single processor and not overlapping with B) would bring total to roughly 5.4×10^7.

[1] Probably not realistic, but we'll use to simplify this problem.

(This page intentionally left mostly blank for answers.)

Code of Academic Integrity

Since the University is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the University community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.*

Academic Dishonesty Definitions

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

A. Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using a cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

B. Plagiarism Using the ideas, data, or language of another without specific or proper acknowledgment. Example: copying another person's paper, article, or computer work and submitting it for an assignment, cloning someone else's ideas without attribution, failing to use quotation marks where appropriate, etc.

C. Fabrication Submitting contrived or altered information in any academic exercise. Example: making up data for an experiment, fudging data, citing nonexistent articles, contriving sources, etc.

D. Multiple Submissions Multiple submissions: submitting, without prior permission, any work submitted to fulfill another academic requirement.

E. Misrepresentation of academic records Misrepresentation of academic records: misrepresenting or tampering with or attempting to tamper with any portion of a student's transcripts or academic record, either before or after coming to the University of Pennsylvania. Example: forging a change of grade slip, tampering with computer records, falsifying academic information on one's resume, etc.

F. Facilitating Academic Dishonesty Knowingly helping or attempting to help another violate any provision of the Code. Example: working together on a take-home exam, etc.

G. Unfair Advantage Attempting to gain unauthorized advantage over fellow students in an academic exercise. Example: gaining or providing unauthorized access to examination materials, obstructing or interfering with another student's efforts in an academic exercise, lying about a need for an extension for an exam or paper, continuing to write even when time is up during an exam, destroying or keeping library materials for one's own use., etc.

* If a student is unsure whether his action(s) constitute a violation of the Code of Academic Integrity, then it is that student's responsibility to consult with the instructor to clarify any ambiguities.