



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SUBJECT: PARALLEL COMPUTER ARCHITECTURE AND PROGRAMMING (CSE-3252) (PCAP)

MID-SEMESTER EXAMINATION MARCH 2024

SCHEME OF EVALUATION

Time :2hrs

Max. Marks:30

Q. No	Questions	M	CLO	AHEP LO	Blooms Taxonomy level
1	The platform model defines a device as an array of compute units which can function independently and are further divided into _____. 1. Processing elements 2. kernels 3. memory units 4. cores	0.5	3	1,2	2
2	Creating buffer requires supplying the size of the buffer and a _____ in which the buffer will be allocated 1. Command queue 2. work item 3. context 4. kernel	0.5	3	1,2	2
3	Assume that a variable 'x' is declared inside a kernel function (OpenCL kernel) as int x. 'x' will be now allocated to _____ memory. 1. Private 2. local 3. global 4. constant memory	0.5	3	1,2	2
4	Give the work item's local id and global id which will be processing the shaded element of the 1D array in the following figure. Assume the work group size is 3.	0.5	3	1,2	3



	<table><tr><td>12</td><td>45</td><td>67</td><td>23</td><td>56</td><td>90</td><td>78</td><td>34</td><td>26</td></tr></table> <div><div>1. 2,5</div><div>2. 5, 2</div><div>3. 5,1</div><div>4. 5,0</div></div>	12	45	67	23	56	90	78	34	26				
12	45	67	23	56	90	78	34	26						
5	Identify the CORRECT statements with respect to the CUDA programming? i) A device function can be called from host function. ii) The second execution configuration parameter in CUDA specifies the dimensions of each block in number of threads. iii) All the threads in a block can access the block index using blockDim predefined variable in kernel. <div><div>1. i and ii are TRUE</div><div>2. i and iii are FALSE</div><div>3. ii and iii are TRUE</div><div>4. ii and iii are FALSE</div></div>	0.5	4	1,2	3									
6	The function cudaMemcpy() is used to copy data a) From one location of device memory to another location of device memory b) Between different GPUs in multi-GPU systems c) From one location of device memory to another location of host memory d) From one location of host memory to another location of device memory <div><div>1. a, c and d only</div><div>2. c and d only</div><div>3. b, c and d only</div><div>4. All of the above</div></div>	0.5	4	1,2	2									
7	For the given input array N = {10, 35, 52, 63, 7, 81, 90} and Mask array M={2, 3, 5}, the output array value P[6] calculated for the 1D convolution is _____ <div><div>1. 188</div><div>2. 7719</div><div>3. 432</div></div>	0.5	4	1,2	3									



	4. 707				
8	<p>In MPI, what is the purpose of MPI_Barrier function?</p> <ol style="list-style-type: none">1. To initialize MPI communication2. To synchronize processes within a communicator3. To broadcast data from one process to all others4. To reduce data across all processes	0.5	2	1	2
9	<p>The disadvantages in MPI_Bsend and MPI_Ssend are ____ and ____ respectively.</p> <ol style="list-style-type: none">1. extra copying and extra waiting2. extra waiting and extra copying3. extra scattering and extra gathering4. extra gathering and extra scattering	0.5	2	1	3
10	<p>In _____, all PEs receive the same instruction broadcast from the control unit but operate on different data sets from distinct data streams</p> <ol style="list-style-type: none">1. SISD2. SIMD3. MIMD4. MISD	0.5	1	1	2
11	<p>Write an opencl program that uses kernel function to put integer data into the 1d array of 128 locations. The program prompts the user to enter any number between 2 to 10(2 and 10 are included). If valid number is entered by the user, then only kernel is enqueued and produces the desired output. If a valid number is not entered, then program should prompt the user to enter valid number.</p> <p>Enter a number between 2 and 10. Number entered by user is 2 The following o/p is desired o/p</p>	4	3	1	4



<p>The 1d array is 1 3 5 7 9 11 13.....251, 253, 255</p> <p>Enter a number between 2 and 10. Number entered by user is 3 The following o/p is desired o/p The 1d array is 1 4 7 10 13 16.....376, 379, 382</p> <pre>#include <stdio.h> #include <stdlib.h> // OpenCL includes #include <CL/cl.h> // #include <iostream> #include <string.h> #define MAX_SOURCE_SIZE 1000 // #include <conio.h> // OpenCL kernel to perform an element-wise // add of two arrays int main() { // This code executes on the OpenCL host // Host data char *source_str; size_t source_size; char* programSource; char str[1000]; FILE *fp; fp=fopen("E1.txt", "r"); if(!fp) { fprintf(stderr, "Failed to load kernel.\n"); getchar(); exit(1); } programSource = (char*)malloc(MAX_SOURCE_SIZE); source_size = fread(programSource, 1, MAX_SOURCE_SIZE, fp); fclose(fp); int *A = NULL; // Input array int n; int datasize=sizeof(int)*128; A=malloc(128*sizeof(int)); do { printf("enter number between 2 and 10\n"); scanf("%d",&n); }while(n<2 n>10); cl_int status; // STEP 1: Discover and initialize the platforms cl_uint numPlatforms = 0;</pre>				
--	--	--	--	--



	<pre>cl_platform_id *platforms = NULL; // Use clGetPlatformIDs() to retrieve the number of // platforms status = clGetPlatformIDs(0, NULL, &numPlatforms); printf("%d platform success %d ",status,numPlatforms); // Allocate enough space for each platform platforms = (cl_platform_id*)malloc(numPlatforms*sizeof(cl_platform_id)); // Fill in platforms with clGetPlatformIDs() status = clGetPlatformIDs(numPlatforms, platforms, NULL); char pform_vendor[40]; clGetPlatformInfo(platforms[0], CL_PLATFORM_NAME, sizeof(pform_vendor), &pform_vendor, NULL); printf(" the vendor %s",pform_vendor); // STEP 2: Discover and initialize the devices cl_uint numDevices = 0; cl_device_id *devices = NULL; // Use clGetDeviceIDs() to retrieve the number of // devices present status = clGetDeviceIDs(platforms[0], CL_DEVICE_TYPE_GPU, 0, NULL, &numDevices); // Allocate enough space for each device devices = (cl_device_id*)malloc(numDevices*sizeof(cl_device_id)); // Fill in devices with clGetDeviceIDs() status = clGetDeviceIDs(platforms[0], CL_DEVICE_TYPE_GPU, numDevices, devices, NULL); printf("%d Device success %d ",status,numDevices); char name_data[100]; int err = clGetDeviceInfo(devices[0], CL_DEVICE_NAME, sizeof(name_data), name_data, NULL); printf(" the device name %s",name_data); // STEP 3: Create a context cl_context context = NULL; // Create a context using clCreateContext() and // associate it with the devices context = clCreateContext(</pre>				
--	--	--	--	--	--



	<pre>NULL, numDevices, devices, NULL, NULL, &status); printf("%d context success %d ",status,numDevices); // STEP 4: Create a command queue l_command_queue cmdQueue; // Create a command queue using clCreateCommandQueue(), // and associate it with the device you want to execute // on cmdQueue = clCreateCommandQueue(context, devices[0], 0, &status); printf("%d CQ success %d ",status,numDevices); STEP 5: Create device buffers cl_mem bufferA; // Input array on the device cl_mem bufferB; // Output array on the device // Use clCreateBuffer() to create a buffer object (d_A) // that will contain the data from the host array A bufferA = clCreateBuffer(context, CL_MEM_WRITE_ONLY, datasize, NULL, &status); // Use clCreateBuffer() to create a buffer object (d_C) // with enough space to hold the output data bufferB = clCreateBuffer(context, CL_MEM_READ_ONLY, sizeof(cl_int), NULL, &status); / STEP 6: Write host data to device buffers // Use clEnqueueWriteBuffer() to write input array A to // the device buffer bufferA status = clEnqueueWriteBuffer(cmdQueue, bufferB, CL_FALSE, 0, sizeof(cl_int), &n, 0, NULL,</pre>				
--	---	--	--	--	--



	<pre>NULL); // STEP 7: Create and compile the program // Create a program using clCreateProgramWithSource() cl_program program = clCreateProgramWithSource(context, 1, (const char**)&programSource, (const size_t *)&source_size, &status); // Build (compile) the program for the devices with // clBuildProgram() status = clBuildProgram(program, numDevices, devices, NULL, NULL, NULL); printf("the build is %d\n",status); // STEP 8: Create the kernel cl_kernel kernel = NULL; // Use clCreateKernel() to create a kernel from the // vector addition function (named "vecadd") kernel = clCreateKernel(program, "E1", &status); // STEP 9: Set the kernel arguments // Associate the input and output buffers with the // kernel // using clSetKernelArg() status = clSetKernelArg(kernel, 0, sizeof(cl_mem), &bufferA); status = clSetKernelArg(kernel, 1, sizeof(cl_mem), &bufferB); // STEP 10: Configure the work-item structure // Define an index space (global work size) of work // items for // execution. A workgroup size (local work size) is not // required, // but can be used. size_t globalWorkSize[1]; // There are 'elements' work-items globalWorkSize[0] = 128; // STEP 11: Enqueue the kernel for execution // Execute the kernel by using</pre>				
--	---	--	--	--	--



	<pre>// clEnqueueNDRangeKernel(). // ' globalWorkSize' is the 1D dimension of the // work-items status = clEnqueueNDRangeKernel(cmdQueue, kernel, 1, NULL, globalWorkSize, NULL, 0, NULL, NULL); // STEP 12: Read the output buffer back to the host // Use clEnqueueReadBuffer() to read the OpenCL output // buffer (bufferC) // to the host output array (C) clEnqueueReadBuffer(cmdQueue, bufferA, CL_TRUE, 0, datasize, A, 0, NULL, NULL); // Verify the output for(int i=0;i<128;i++) printf("%d ",A[i]); // STEP 13: Release OpenCL resources // Free OpenCL resources clReleaseKernel(kernel); clReleaseProgram(program); clReleaseCommandQueue(cmdQueue); clReleaseMemObject(bufferA); clReleaseMemObject(bufferB); clReleaseContext(context); // Free host resources free(A); //free(C); free(platforms); free(devices); } __kernel void E1(__global int* A,__global int* n) { int idx=get_global_id(0);</pre>				
--	---	--	--	--	--



	<pre>if(idx==0) A[idx]=1; else A[idx]=idx*(n)+1; }</pre> <p>Scheme:Kernel 1M, steps 1to 4 1m, 5 to 8 1M , 9 to 12 1M</p>																													
2	<p>i)Assume that a grid has 128 blocks arranged in 2D and grid length (x direction) is 32. Threads in a block are arranged in 2D with block height(y direction) is 5. Each block contains 30 threads. Fill the following table with appropriate values. Show the calculations along with the required formulae for the last column. Note: Global and local thread indexing and block indexing shown in the table starts with 0. For blocks and threads(x,y,z) notation is used.</p> <table><tr><td>gridDim.x</td><td>gridDim.y</td><td>blockDim.x</td><td>blockDim.y</td><td>Global thread id of a thread (12,4) in block (2,3)</td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr></table> <p>ii) Write a CUDA kernel to multiply matrix A with size mxn and matrix B with size nXp and produce the resultant matrix C. In the host write the code snippet which launches the above kernel with a single block and required number of threads in each block. The thread should be created such that every column of the resultant matrix should be computed by one thread.</p> <table><tr><td>gridDim.x</td><td>gridDim.y</td><td>blockDim.x</td><td>blockDim.y</td><td>Global thread id of thread (3,4) block (2,3)</td></tr><tr><td>32</td><td>4</td><td>6</td><td>5</td><td>2967</td></tr></table> <p>Writing values in table – 0.5M</p>	gridDim.x	gridDim.y	blockDim.x	blockDim.y	Global thread id of a thread (12,4) in block (2,3)											gridDim.x	gridDim.y	blockDim.x	blockDim.y	Global thread id of thread (3,4) block (2,3)	32	4	6	5	2967	4	4	1,2	5
gridDim.x	gridDim.y	blockDim.x	blockDim.y	Global thread id of a thread (12,4) in block (2,3)																										
gridDim.x	gridDim.y	blockDim.x	blockDim.y	Global thread id of thread (3,4) block (2,3)																										
32	4	6	5	2967																										



	<p>BlockId = blockIdx.y * gridDim.x + blockIdx.x</p> <p>threadId = BlockId * blockDim.x * blockDim.y + threadIdx.y * blockDim.x + threadIdx.x 0.5M</p> <p>BlockId = 3 * 32 + 2 = 98 0.5M</p> <p>threadId = 98 * 6 * 5 + 4 * 6 + 3 = 2940 + 24 + 3 = 2967 0.5M</p> <p>multiplyKernel_b<<<1, p>>>(d_a, d_b, d_c, m,n); __global__ void multiplyKernel_colwise(int * a, int * b, int * c, int ha, int wa) { int cidB = threadIdx.x; int wb = blockDim.x; int sum, k; for(ridA = 0; ridA < ha; ridA++) 1M { sum = 0; for(k=0; k< wa; k++) { sum += (a[ridA * wa + k] * b[k * wb + cidB]); } c[ridA * wb + cidB] =sum; 1M } } Scheme: i) 2m ii) 2m</p>				
13	<p>1) Write a note on MPI_Scatter, MPI_Gather and MPI_Allgather collective communications along with syntax. Compare two MPI operations that are used for collective computations while one perform the operations across all the tasks in the group and place the result in one task and the other computes and puts the partial operation results on each processor. Scheme: 3 collectives with syntax: each 0.5M *3=1.5M Identifying Reduce and syntax=0.5M Identifying Scan and Syntax=0.5M Comparison of reduce and scan =0.5M Total=3M (images displaying the operation explanation can be considered along with syntax instead of writeup)</p> <p>Solution: MPI_Scatter (void *sendbuf, int sendcount, MPI_Datatype sendtype, void *recvbuf, int recvcount, MPI_Datatype recvttype, int root, MPI_Comm comm) • Sends individual messages from the root process to all other processes</p>	3	2	1	3



	<ul style="list-style-type: none"> • Inverse to MPI_Gather • sendbuf is ignored by all non-root processes <p>MPI_Gather (void *sendbuf, int sendcount, MPI_Datatype sendtype, void *recvbuf, int recvcnt, MPI_Datatype recvtype, int root, MPI_Comm comm)</p> <ul style="list-style-type: none"> • One process (root) collects data from all the other processes in the same communicator (i.e each process in comm (including root itself) sends its sendbuf to root.) • The root process receives the messages in recvbuf in rank order Must be called by all the processes with the same arguments <p>MPI_Allgather (void *sendbuf, int sendcnt, MPI_Datatype sendtype, void *recvbuf, int recvcnt, MPI_Datatype recvtype, MPI_Comm comm)</p> <ul style="list-style-type: none"> • All the processes collect data from all the other processes in the same communicator (i.e similar to MPI_Gather except now all processes receive the result.) • recvbuf is NOT ignored • Must be called by all the processes with the same arguments <p>MPI_Reduce(void *sendbuf, void *recvbuf, int count, MPI_Datatype datatype, MPI_Op op, int root, MPI_Comm comm)</p> <ul style="list-style-type: none"> • One process (root) collects data from all the other processes in the same communicator, and performs an operation on the data (i.e combines elements provided by input buffer of each process in the group using operation op.) • Returns combined value in the output buffer of process with rank root • MPI_Scan (void *sendbuf, void *recvbuf, int count, MPI_Datatype datatype, MPI_Op op, MPI_Comm comm) • A reduction means all processors get the same value while scan returns the partial operation results on each processor • For example: <ul style="list-style-type: none"> o if you had 10 processors and you were taking the sum of their rank, MPI_Reduce would give you the scalar 45 (0+1+2+3+4+5+6+7+8+9) on the root process, o while MPI_scan would give you the scalar of the reduction up to the rank of the processor on each processor. So processor 0 would get 0, processor 1 would get 1, processor 2 would get 3, and so on. Processor 9 would get 45 				
14	<p>2) Write an MPI program in MPI to get the following output.</p> <p>Input string read by root: MIT TOP HUT I am rank 0, my string is: TIM I am rank 1, my string is: POT POT POT I am rank 2, my string is: TUH TUH TUH TUH Output string displayed by root: TIM POT POT POT TUH TUH TUH TUH TUH TUH</p>	3	2	1	4



<p>The root will distribute words in the string to the processes including itself and collects to print the final output. Each process is involved in reversing the word/partial string received and repeating the input word equal to (word length * rank) times. If its rank 0, at least 1 time the reversed string needs to be printed. Use collective communication to distribute the words from the input string and use point to point communication to send the repeated string to root process. Each process should print its computed string as well.</p> <p>Note: Take a string of equal word length to keep the problem simple.</p> <p>Scheme: Reversing and computing new length= 1 M Each process displaying its output =1M Final output and total correctness=1M</p> <pre>#include<string.h> #include <stdio.h> #define NPROCS 3 int main(int argc, char *argv[]) { int rank, numtasks; char str[50],rstr[10],temp[50],t[50]=" "; char t1[20]; int i,j,k,h,len,r; MPI_Init(&argc,&argv); MPI_Comm_rank(MPI_COMM_WORLD, &rank); MPI_Comm_size(MPI_COMM_WORLD, &numtasks); printf(" %d ",numtasks); if(rank==0) { strcpy(str, "MIT TOP HUT"); for(i=0;str[i]!='\0';i++) { if(str[i]==' ') { for(j=i;str[j]!='\0';j++) str[j]=str[j+1]; str[j]='\0'; } else continue; } Len =3; } MPI_Bcast(&len,1,MPI_INT,0,MPI_COMM_WORLD); MPI_Bcast(str,len,MPI_CHAR,0,MPI_COMM_WORLD); - rstr[len]='\0'; printf("im %d process and sting i got is %s\n",rank,rstr);</pre>				
--	--	--	--	--

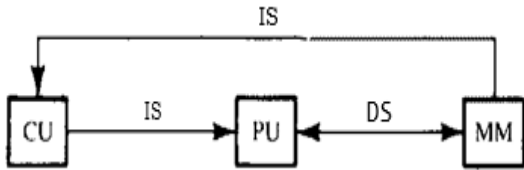


	<pre>for(i=0;i<=len/2;i++) { char temp=rstr[i]; rstr[i]=rstr[len-i-1]; rstr[len-i-1]=temp; } for(i=1;i<=rank+1;i++) { strcat(t,str); } printf("im %d process and string modified is got is %s\n",rank,t); int newlen; if (rank == 0) newlen=(rank+1)*len+1; else newlen=rank*len +1; if(rank!=0) { MPI_Send(&newlen,1,MPI_INT,0,rank,MPI_COMM_WORLD); MPI_Send(t,newlen,MPI_CHAR,0,rank+1,MPI_COMM_WORLD); } MPI_Barrier(MPI_COMM_WORLD); if(rank==0) { strcpy(temp,t); printf("%s",temp); for(i=1;i<=numtasks;i++) { MPI_Recv(&newlen,1,MPI_INT,i,i,MPI_COMM_WORLD,&st); MPI_Recv(t1,newlen,MPI_CHAR,i,i+1,MPI_COMM_WORLD,&st); strcat(temp,t1); for(j=0;j<20;j++) t1[j]='\0'; } printf("the new string is "); puts(temp); } MPI_Finalize(); return 0; }</pre> <p>Input string read by root: MIT TOP HUT I am rank 0, my string is: TIM I am rank 1, my string is: POT POT POT I am rank 2, my string is: TUH TUH TUH TUH TUH Output string displayed by root: TIM POT POT POT TUH TUH TUH TUH TUH TUH</p>				
15		3	3	1,2	4



	<p>Explain in brief why the devices in a context (opencl context) is limited to a specific platform. Write a neat diagram of the abstract memory model defined by OpenCL. (2M+1m)</p> <p>Ans: i) Limiting the context to a given platform allows the programmer to provide context for multiple platforms and fully utilize s system comprising resources from a mixture of vendors.</p> <p>ii)In addition, the application developer can target his application to different devices from different vendors and thus benefit from larger customer base.</p> <p>Scheme: Diagram 1M ,explanation each point 1M,1*2= 2M</p> <p>Diagram illustrating the OpenCL abstract memory model:</p> <ul style="list-style-type: none">Global Memory: Shared across the entire device (CPU <-> GPU).Read only memory: Memory that is read-only for the kernel.Kernel-wide scope: The scope of the kernel, encompassing global and constant memory.Workgroup scope: The scope of a workgroup, encompassing local memory.Local memory: Memory shared among work-items within a workgroup.Work-item scope: The scope of a single work-item, encompassing private memory.Private memory: Memory private to a single work-item.				
16	<p>Compare and contrast temporal parallelism and spatial parallelism. (3M)</p> <p>Scheme: each [1.5M]</p> <p>Temporal parallelism:</p> <ul style="list-style-type: none">• Temporal parallelism or pipelining refers to the execution of a task as a 'cascade' of sub-tasks• There exists one functional unit to carry out each sub-task• All these successive units can work at the same time, in an overlapped fashion	3	1	1	3



	<ul style="list-style-type: none"> As data are processed by a given unit U_i, they are sent to the next unit U_{i+1} and the unit U_i restarts its processing on new data, analogously to the flow of work in a car production line. Each functional unit can be seen as a "specialized" processor in the sense that it always execute the same sub-task <p>spatial parallelism:</p> <ul style="list-style-type: none"> spatial parallelism refers to the simultaneous execution of tasks by several processing units At a given instant, these units can be executing the same task (or instruction) or different tasks. The former case is called SIMD (Single Instruction stream, Multiple Data stream), whereas the latter is called MIMD (Multiple Instruction stream, Multiple Data stream) 				
17	<p>. Discuss the following with neat diagrams: SISD computer organization, MISD computer organization.</p> <p style="text-align: center;">(3M)</p> <p>Scheme: Diagrams [0.5M+0.5M]+ Explanation [1M+1M]</p> <p>SISD:</p> <ul style="list-style-type: none"> This organization represents most serial computers available today Instructions are executed sequentially but may be overlapped in their execution stages (pipelining) An SISD computer may have more than one functional unit in it. All the functional units are under the supervision of one control unit Applications: Whatever we do with our personal computers today <div style="display: flex; align-items: center;">  <div style="margin-left: 20px;"> <p>CU: control unit PU: processing unit MM: memory module IS: instruction stream Ds: data stream</p> </div> </div> <p>SIMD:</p> <ul style="list-style-type: none"> In this organization, there are multiple processing elements supervised by the same control unit 	3	1	1	3



	<ul style="list-style-type: none"> • All PEs receive the same instruction broadcast from the control unit but operate on different data sets <p>from distinct data streams</p> <ul style="list-style-type: none"> • The SIMD model of parallel computing consists of two parts: <ul style="list-style-type: none"> ▪ A front-end computer of the usual von Neumann style ▪ And a processor array • Applications: <ul style="list-style-type: none"> ▪ Image processing ▪ Matrix manipulations ▪ Sorting 				
18	<p>Illustrate the compilation process of a CUDA C program with the help of a neat diagram</p> <ul style="list-style-type: none"> • Each CUDA source file can have a mixture of both host and device code. • By default, any traditional C program is a CUDA program that contains only host code. One can add device functions and data declarations into any C source file by marking them with special CUDA keywords. • The NVIDIA C Compiler (NVCC) separates the host code and the device code during compilation process. <ul style="list-style-type: none"> <input type="checkbox"/> The host code is compiled with the host's standard C compilers and runs as an ordinary CPU process. <input type="checkbox"/> The device code(kernels) is compiled by the NVCC and executed on a GPU device 	2	4	1,2	3

