Faculty EEMCS
Departments DIAM and ST

Examination IN4049TU

April 9, 2018

18:30—21:30 h

Question 1.

(a) State three reasons why simulations are to be preferred over doing experiments with real or prototype systems.

(b) What is the operational intensity of a program? Explain the basic roofline model for the attainable performance of a program.

(c) Explain the farmer/worker, the data parallel, and the task parallel models of parallel computation.

(d) Explain the pthreads, HPF, and MPI parallel programming models. How would you rate their levels of abstraction? Explain your answer.

Question 2.

Assume you have a system with four processors with four cores each.

(a) Give the definitions of the speedup and efficiency of a program.

(b) Assume that a sequential application A can be perfectly parallelized for a fraction of 80%. What is the speedup A achieves on the system mentioned above?

(c) Assume application A takes 200 seconds to run sequentially. How many processors are needed to have A run below 42 seconds?

(d) What is the efficiency in part (c)?

- See reverse side -

Question 3.

(a) What makes efficient parallelization of a Multi-Grid algorithm more complicated as compared to the Jacobi or Conjugate Gradient algorithm?

(b) Describe how parallel matrix-vector multiplication is related with graph partitioning. Give a small example (e.g., illustrate with a small example).

(c) Describe the important differences in performance characteristics between CPU and GPU. When should one use GPU for computing FFT?

(d) There are three levels in the Basic Linear Algebra Subroutines (BLAS) library: BLAS-1, BLAs-2 and BLAS-3. A function in BLAS-3 usually has a higher FLOPs than a function in BLAS-1 or BLAS-2, why? Please explain.

Question 4.

(a) Describe the "communication avoiding Jacobi" scheme for computing the matrix-vector products (for example, for a matrix corresponding to a regular 1-D grid). Explain how is communication overhead reduced with this scheme and what are the trade-offs?

(b) What are the challenges of implementating efficient SpMV on a GPU? Is there a sparse matrix storage scheme which is the best choice for SpMV on GPU? Explain your answer.

(c) Name two major categories of graph partitioning methods. Describe the main steps/phases of a multi-level graph partitioning algorithm.

(d) Consider the dense matrix-vector multiplication and dense matrix-matrix multiplication where the dimension of the vector is n and the dimension of the matrix is n by n. Assume we have P processors, describe a parallel algorithm for performing the matrix-vector multiplication and for the matrix-matrix multiplication respectively (please include the main steps and describe how you partition and distribute the matrix and the vector). What are the data locality ratio of each of these two parallel algorithms?

- End -