

# GEORGIA INSTITUTE OF TECHNOLOGY

College of Computing

## CS6290/CS4290 — High-Performance Computer Architecture Spring 2003

---

CS6290/CS4290  
Project 3

Issued: April 4, 2003  
Due: April 18, 2003

---

**Purpose:** This project explores network performance in the context of an abstract parallel machine. It starts by introducing a simple performance model for parallel execution using a particular application as an example and then uses the model to estimate the effect of the network on performance. The goal is to develop intuition and a back-of-the-envelope strategy for evaluating the suitability of an architecture to an application.

**Reading:** H&P Chapter 8 for networks  
H&P Sections 6.1 and 6.2 for parallel applications

**Problems:**

1. Partitioning.
2. Communication Overhead.

## Introduction

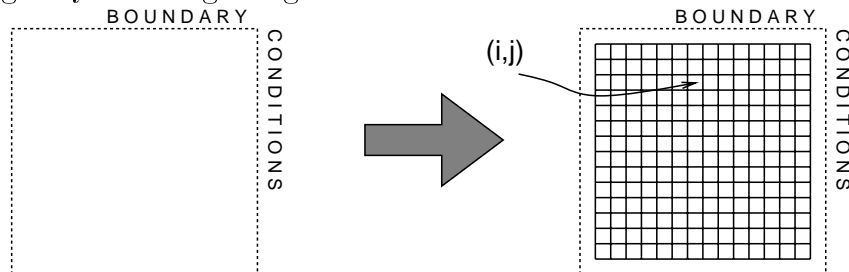
When developing multiprocessor applications, we attempt to exploit parallelism to achieve increased performance. With increased parallelism, however, comes increased interprocessor communication. Since real multiprocessors can provide only a finite amount of network bandwidth, this tension between parallelism and communication has significant ramifications which we need to keep in mind when designing and programming multiprocessors. In this exercise, you will examine several of these issues.

Multiprocessor programs also require the introduction of synchronization to coordinate multiple processes. While different languages offer varied forms of synchronization, in this homework you will explore producer-consumer style synchronization, implemented using `barrier()` constructs for shared-memory and via messages in message-passing.

The goal of these exercises is to familiarize you with the problems of partitioning parallel programs and data structures. In addition, you will develop a simple model of parallel program performance that gives insights into the demands of applications and the capabilities of machines.

## Example Application: Jacobi Relaxation

*Jacobi relaxation* is an iterative algorithm which, given a set of boundary conditions, finds (discretized) solutions to differential equations of the form  $\nabla^2 \mathcal{A} + \mathcal{B} = 0$ . As we've seen in lecture, we begin by choosing the grid which will form the basis of our discretization:



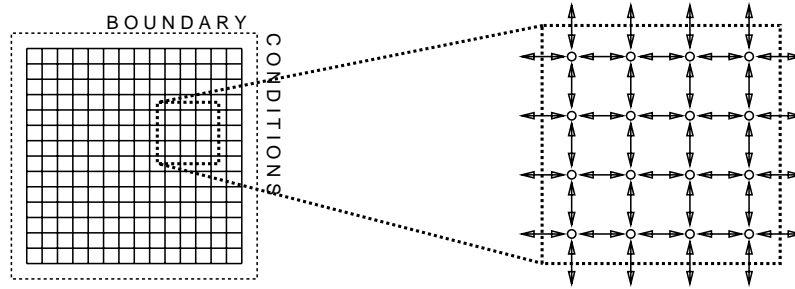
To find a solution on a grid, we repeatedly apply the following iterative step until we converge on a solution.

$$A_{i,j}^{k+1} = \frac{A_{i+1,j}^k + A_{i-1,j}^k + A_{i,j+1}^k + A_{i,j-1}^k}{4} + b_{i,j}$$

Jacobi differs from most other iterative relaxation algorithms in that the update of each point (at iteration step  $k + 1$ ) requires the *previous* values of the neighboring points (from iteration step  $k$ ).

We use a simple graphical representation to capture those features we are interested in and abstract away excess detail. In this representation, graph nodes represent fixed amounts of

computation; graph edges represent communication between computation nodes. In such a representation, a single iteration of Jacobi relaxation looks as follows:



In this graph, each node represents the computation required to compute a new value for a single grid point. To compute a new value for a grid point, Jacobi relaxation dictates that we average the previous values of each of the neighboring grid points – thus each node is connected to its four neighboring nodes with an edge that represents the communication of two grid values (one in each direction).

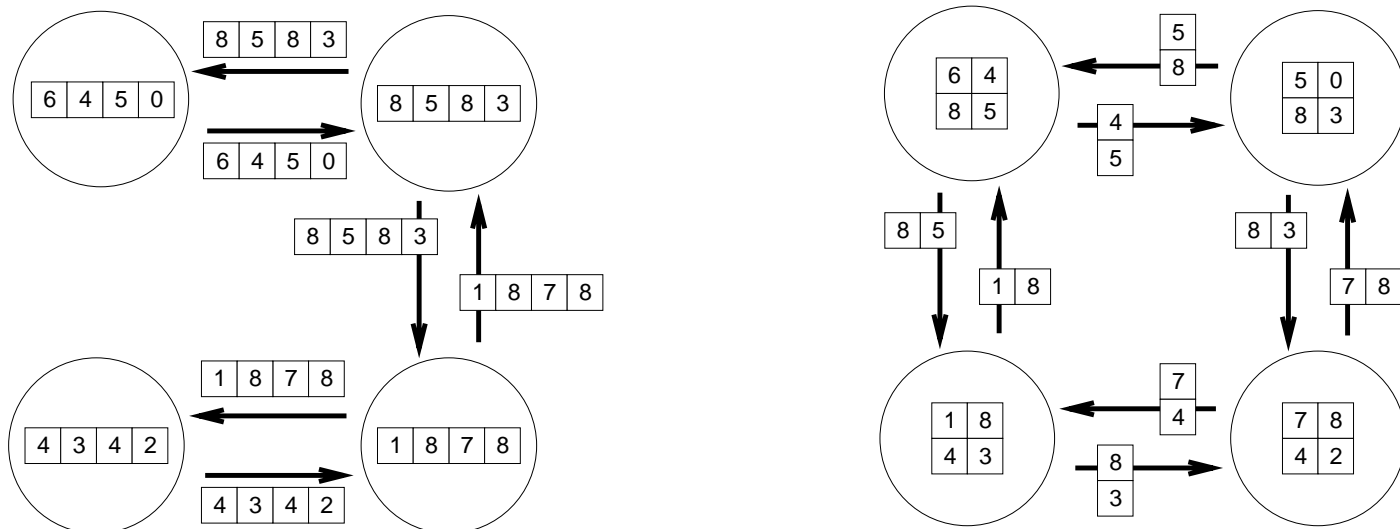
The following figure shows one step of a relaxation on a four by four grid. The  $A^k$  matrix contains the values of the grid at step  $k$ , and the  $A^{k+1}$  matrix has the values for the next step. The boundary conditions are shown on the edges of the  $A^k$  and  $A^{k+1}$  matrices. The  $B$  matrix, which is constant, contains the values that are added to the grid points on each step according to the equation above.

$$\begin{array}{c}
 \begin{array}{cccc}
 & 0 & 0 & 0 & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 6 & 4 & 5 & 0 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 8 & 5 & 8 & 3 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 1 & 8 & 7 & 8 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 4 & 3 & 4 & 2 \\ \hline \end{array} & 0 \\
 & 0 & 0 & 0 & 0
 \end{array}
 \end{array}
 \quad \longrightarrow \quad
 \begin{array}{c}
 \begin{array}{cccc}
 & 0 & 0 & 0 & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 8 & 4 & 4 & 4 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 4 & 4 & 4 & 4 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 8 & 4 & 4 & 4 \\ \hline \end{array} & 0 \\
 4 & \begin{array}{|c|c|c|c|} \hline 4 & 4 & 0 & 8 \\ \hline \end{array} & 0 \\
 & 0 & 0 & 0 & 0
 \end{array}
 \end{array}$$
  

$$\mathbf{B} = \begin{array}{|c|c|c|c|} \hline 4 & 0 & 1 & 2 \\ \hline 0 & -3 & -1 & 0 \\ \hline 2 & 0 & -3 & 1 \\ \hline 2 & 0 & -3 & 5 \\ \hline \end{array}$$

You might want to think about the values in the  $A^{k+2}$  matrix after the next relaxation step.

There are a number of ways to partition the data in this grid to a bunch of processors. Figure 1 shows two possible ways to partition the four by four grid to four processors. One partition allocates one row of the matrix to each processor, and the other allocates a two by two submesh of the matrix to each processor. For every step in the relaxation, *each processor will calculate the new values for its own matrix cells*. To accomplish this calculation, the processors need to get cell values from neighboring processors. The values

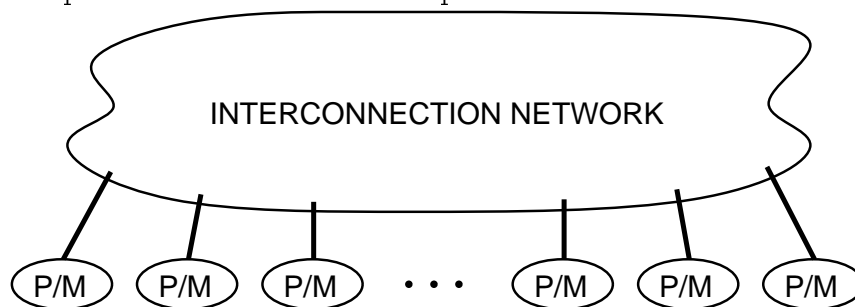


**Figure 1:** Partitioning by rows and by square tiles.

on the arrows between the nodes represent the data that must be communicated between processors.

### Machine Model

The machine model which we'll be using in this homework is relatively simple – a distributed-memory multiprocessor in which some number of processor/memory nodes communicate via an interconnection network. Each processor contains some amount of memory in which it stores the data partition for which it is responsible.



In this machine model, we assume the interconnection network provides uniform access – all interprocessor communication is equally expensive in terms of network resources consumed and communication latency!

Note that we don't make any assumptions about what programming model (e.g. shared memory, message passing, etc.) this machine provides to the end user (yet).

Given an application's graphical representation (such as that described above for Jacobi relaxation), we "program" a  $P$ -processor machine by deciding which processor should perform the computation represented by each of the nodes in the graph. Since we could

equivalently view this process as one of dividing the graph into (at most)  $P$  pieces, we usually refer to this as the *partitioning problem*.

**A:** (*Warmup – nothing to turn in*) For each of the two partitions in Figure 1 how should the  $B$  matrix be distributed to the processors? How should the boundary values be distributed?

**B:** (*Warmup – nothing to turn in*) Using total amount of communicated data as a metric, which of the two partitions in Figure 1 is better?

### Problem 1: Partitioning

The total running time of the program is the ideal metric of goodness – one partition of a program graph is better than another if it results in a shorter running time. But how do we determine running time for a particular partition running on a  $P$ -processor machine? If we assume that there is no overlap between computation and communication, we can estimate the running time as the sum of the computation time and the communication time.

$$T = (\text{time to compute}) + (\text{time to communicate})$$

We start by determining the following information from the program graph and partition:

$w_i$  – the total amount of computation for processor  $i$  (in abstract “computation units”)

$c_i$  – the total amount of communication invoked by processor  $i$  which cannot be resolved on that processor (in abstract “communication units”)

For simplicity, assume that we always partition things such that all processors get the same amount of work,  $w$ , and invoke the same amount of external communication  $c$ . ( $w = w_1 = \dots = w_i$  and  $c = c_1 = \dots = c_i$ )

Given  $w$  and  $c$ , we initially compute the running time  $T$  by summing the computation and communication times required by one processor. Since all the processors are running in parallel and doing the same amount of work, the running time of a single processor should be the same as the running time of the entire application.

Thus,

$$T = s \cdot w + l \cdot c$$

where

$s$  is a measure of processor speed – a processor requires  $s$  time units to complete one unit of computation

$l$  is a measure of network latency – the network requires  $l$  time units to transport one unit of communication

Here is a table that summarizes the variables used in the above formulae:

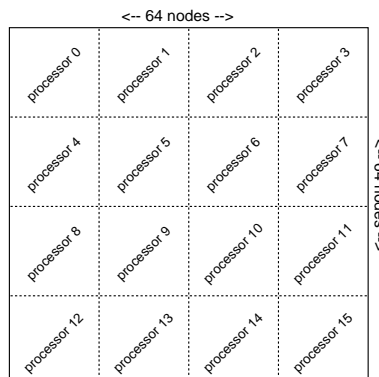
$T$	running time, the metric of a partition
$w$	amount of computation (work) for a processor
$s$	processor speed, in terms of time units to complete on unit of computation
$c$	amount of communication for a processor
$l$	network latency, in terms of time units to transport one unit of communication
$P$	number of processors

**Caveat:** The simple equation  $T = s \cdot w + l \cdot c$  models only processor and network *bandwidth*. In particular, if we use this equation to guide our partitioning decisions for Jacobi, we'll find that running a finer-grained partition on a larger number of processors is *always* preferable. Empirically, we know this isn't true – eventually the costs of communication will overwhelm the speedup so that the total running time is actually *longer* than it might be with a coarser partition. We will ignore the effects of overhead in this problem but come back to them in the next problem.

The following exercises consider the Jacobi algorithm using the model above. Unless stated otherwise, assume  $s = 10$ , and  $l = 1$ . Assume that the Jacobi problem grid is of size  $n \times n = 64 \times 64$ .

**A:** If you use  $P = 64$  processors and partition the Jacobi graph into 64 strips, each 64 nodes wide and one node high, what are the appropriate values of  $w$  and  $c$ , and  $T$ ? How much of a speedup is this over the sequential running time? Recall that a processor is responsible for updating the values in its partition.

**B:** Instead of partitioning the Jacobi graph into long, narrow strips, what if we partitioned it into square tiles? With 16 processors, each processor would get a 16 by 16 node tile, as shown below:



What are the appropriate values of  $w$  and  $c$ ? Using these values, what values do you get for  $T$ ? How much of a speedup is this over the sequential running time?

**C:** Assuming an  $n \times n$  Jacobi grid, derive an expression for the amount of communication for one partition, when the aspect ratio of each partition is given as  $a : b$ . The aspect ratio is specified as  $a : b$ , where  $a$  is the size of the  $x$  dimension of the partition, and  $b$  is the size of the  $y$  dimension of the partition. Furthermore, assume there are  $P$  processors and that each processor gets an equal amount of work.

**D:** Prove that the volume of communication per partition is minimized when  $a = b$ . Assume as before that there are  $P$  processors, and that each processor must get an equal amount of work.

**E: Optimal Rectangular Partitioning.** The classic Jacobi algorithm has a square “kernel”, but other computations on elements of an array may have arbitrary access patterns. Consider the following computation performed for each  $(i,j)$  on an  $n \times n$  grid.

$$A_{i,j}^{k+1} = \frac{A_{i+x,j}^k + A_{i-x,j}^k + A_{i,j+y}^k + A_{i,j-y}^k}{4}$$

Assume  $n \gg P$ ,  $n \gg x$  and  $n \gg y$ . Derive the aspect ratio that minimizes communication for the communication pattern inherent in the computation shown above.

**F:** Does your answer to the previous question change if the some additional terms are included in the iteration step as shown below,

$$A_{i,j}^{k+1} = \frac{A_{i+x,j}^k + A_{i+x',j}^k + A_{i-x,j}^k + A_{i-x',j}^k + A_{i,j+y}^k + A_{i,j-y}^k}{6}$$

where  $x > x'$ ? Discuss briefly.

## Problem 2: Communication Overhead

In this problem, we'll look at the costs of communication and computation more closely and (somewhat) more realistically.

As discussed in class, the performance of an interconnect can be summarized in three parameters: the time-of-flight **latency** across the wire, the maximum **bandwidth** of the wire and the interface **overhead** (hardware and software) at the endpoints. Here are parameters that resemble some real networks:

Network	Wire Latency (S)	Bandwidth (bytes/S)	Overhead (S)
<b>T3E:</b>	$0.1 \cdot 10^{-6}$	$150 \cdot 10^6$	$0.5 \cdot 10^{-6}$
<b>Myrinet:</b>	$1 \cdot 10^{-6}$	$150 \cdot 10^6$	$2 \cdot 10^{-6}$
<b>ATM (uNet):</b>	$22 \cdot 10^{-6}$	$80 \cdot 10^6$	$5 \cdot 10^{-6}$
<b>ATM (TCP):</b>	$22 \cdot 10^{-6}$	$80 \cdot 10^6$	$50 \cdot 10^{-6}$

**A:** Endpoint overhead limits the effective bandwidth of an interconnect for “short” messages. How long (in bytes) must a message be to achieve half of the maximum bandwidth (the “3dB point”) in these networks? The overhead is for one endpoint (assume sending and receiving costs the same). Give two answers for each network: first, the bandwidth assuming two processors communicate completely synchronously (no messages may overlap) and, second, the bandwidth assuming messages may be overlapped/pipelined.

The previous problems used very simple model for compute time and communication time that ignored constant factors and overhead. With the numbers above, we can now nail down some constants, include overhead and get a much better estimate of performance:

- Use the Myrinet parameters above to answer the rest of the questions in this problem. Note that the Myrinet network is full-duplex: you can simultaneously send and receive data at  $150 \cdot 10^6$  bytes/S.
- Assume the processor is capable of 1 floating-point operation per cycle at a clock rate of 300MHz. Ignore all other instructions (i.e. assume they are covered by instruction-level parallelism). Jacobi requires 3 FADDs plus one FMUL/cell, so, for instance, a  $64 \times 64$  grid has a sequential execution time of

$$4 \cdot 64 \cdot 64 = 16384 \text{ cycles per iteration}$$

Aside: the machine parameters above approximately describe one of the CoC Intel clusters, “beetle” (somewhat old at this point).

**B:** Ignoring overhead/latency (i.e. considering only bandwidth), what is the execution time for one iteration of Jacobi on the  $64 \times 64$  array using 16 processors on square tiles?

**C:** Now, consider overhead and latency. Note that each processor will have to send four messages and receive four messages (paying overhead and bandwidth for all of them) but that latency can be overlapped somewhat. Since numeric answers alone are completely inscrutable, do the following:

- i. Draw a timeline for the execution of a loop iteration on one processor showing and labeling  $(x)$  compute time,  $(y)$  overhead for each message sent and received and  $(z)$  any latency you can't overlap.
- ii. What is the total execution time for one iteration of Jacobi on the  $64 \times 64$  array using 16 processors on square tiles?
- iii. What is the speedup over the sequential execution time?

**D:** In the spirit of the book's Figure 6.6 on p. 547, what is the "scaling of computation-to-communication" for Jacobi?