5 Parallel Algorithms

These notes are a distillation from a number of different parts of Mike Quinn’s book.

5.1 Basic Concepts

With ordinary computers and serial (=sequential) algorithms, we have one processor and one memory.

We count the number of operations executed by the processor.

We normally do the opcount and then pay lip service to issues like memory, etc.

With parallel computer algorithms, we have no choice but to make some assumptions up front, and these assumptions have a major affect on the kind of algorithms that we design, analyse, and use.

We clearly want algorithms to be scalable. This is a slightly vague term, as are many terms. Basically, we would like running time \( N \) with \( k \) processors to become time \( N/2 \) with \( 2k \) processors.

**Definition 5.1.** An algorithm is said to be scalable if the running time of the algorithm decreases roughly linearly with the number of processors.

In the real world, almost nothing is really scalable. Usually, good algorithms will scale to a certain point and then fail to scale beyond that point when overhead overtakes the part of the computation that is in fact getting useful work done.

And in the real world, the analysis of parallel algorithms can be difficult because we’d have to balance communication time with compute time and
with memory access time. And this is not easy.

5.1.1 Symmetric Multi Processors (SMPs)

(e.g., Cray vector machines, Sun Enterprise with caveats)

Many processors, large flat memory, all processors have genuine access to all memory locations with the same access time.

This is what we would like to have. But we can’t have it because it’s too expensive to interconnect processors with memory and to guarantee

Programs have access to all of the machine’s memory.

5.1.2 Parallel Computers with an Interconnect Structure

(e.g., Cray T3E, Thinking Machines CM-2 and CM-5, Meiko computing surface)

Many parallel compute nodes, each with its own memory.

The processors are connected to one another with a fixed interconnect.

E.g.:

T3E has 3-dimensional toroidal nearest neighbor connections.

CM-2 and CM-5 had a logarithmic network

With some machines (T3E), processors’ memories can be aggregated into a larger shared memory space.

With other machines (CM-2, CM-5), processors are really limited to using only their local memory, and communications from one processor to another must be explicit.
5.1.3 Cluster Computers with Non-Uniform Memory Access (NUMA)

(e.g., Sun Enterprise, SGI Origin, Pittsburgh Terascale)

Nodes of SMP computers connected to each other

Memory access on the node is faster than access off node through the network.

Programs see a shared memory space whose implementation may or may not act like a shared memory space, and programs definitely work better if computations are local.

5.1.4 Distributed Computers With Poor Interconnect

(e.g., Beowulf, Scalable Network of Workstations (SNOW))

Individual computers on a network.

Beowulf is a dedicated network.

SNOW is just “the usual” network linking ordinary workstations.

Very loose coupling of processors.

Computations must be primarily local.

Explicit message passing from processor to processor.
5.2 The Assumptions of Parallel Processing

Definition 5.2. A Parallel Random Access Machine, or PRAM, comprises

- a control unit
- an unbounded number of processors each with its own local memory
- a global memory
- an interconnect between the processors and the global memory

We assume that all processors execute the same operation with each time step, under the control of the control unit. Processors may be idle in a given time step rather than execute the operation. But the processors have access to the entire global memory.

The critical question with a PRAM is what happens when parallel processors choose to read or write the same memory locations in the same time step.

In decreasing order of restriction and increasing order of power, we have the following:

- **Exclusive Read Exclusive Write (EREW)**: Only one processor may read or write a given location in a given time step.

- **Concurrent Read Exclusive Write (CREW)**: Only one processor may write to a given location in a given time step, but any number of processors may read the same location in the same time step. This is the usual PRAM model.
• Concurrent Read Concurrent Write (CRCW):

  – **Common** write: So long as all processors are writing the same value to the same location, the write is permitted.

  – **Arbitrary** write: All processors can write to the same location. The processor that succeeds in writing the value (or the processor that happens to write last) is chosen arbitrarily.

  – **Priority** write: Only the processor with the highest priority (in some scheme of static priorities) gets to write.

• An EREW algorithm works on a CREW machine with no slowdown.

• An EREW algorithm works on a CRCW machine with no slowdown.

• An arbitrary algorithm works on a priority machine with no slowdown.

• A common algorithm works on an arbitrary machine with no slowdown.

**Lemma 5.3.** A $p$ processor EREW PRAM can sort a $p$-element array in global memory in $\Theta(\log p)$ time.

**Theorem 5.4.** An $p$-processor EREW machine can simulate a $p$ processor priority algorithm with no more than a $\Theta(\log p)$ slowdown.

**Proof.**

• The priority machine (PPRAM) uses processors $P_1,...,P_p$ and memory $M_1,...,M_m$.

• The EREW PRAM uses extra global locations $T_1,...,T_p$ and $S_1,...,S_p$ to simulate each step of the priority algorithm.
• When $P_i$ in PPRAM accesses location $M_j$, the EREW PRAM writes $(j, i)$ in extra location $T_i$.

• The EREW PRAM then sorts the extra location values in time $\log p$.

• Constant time to find the highest priority processor for any particular location.

• Now processor $P_1$ reads $(i_1, j_1)$ from location $T_1$ and writes a 1 in location $S_{j_1}$.

• Processors $P_2$ through $P_p$ read locations $T_k$ and $T_{k-1}$.

• If $i_k \neq i_{k-1}$ then $P_i$ writes a 1 into $S_{j_k}$, else write a 0.

• Now the elements of $S$ with value 1 correspond to the highest priority processors accessing each memory location.

• For a write operation, the highest priority processor doing the write gets to write.

• For a read, the highest priority processor gets to read and then can communicate the value to the other processors in $\Theta(\log p)$ time.
5.3 Parallel Communications and Process Spawning

We first can observe that with any of the PRAM models, we can spawn $p$ processes on $p$ processors in $\lg p$ time using a binary tree. A control processor can spawn two processes, each of those can spawn two more, and so forth.

This works fine on PRAM machines, but not so well on machines with fixed communication structures.

**Definition 5.5.** The broadcast problem is the problem on a parallel computer of having one processor broadcast a value to all other processors.

**Definition 5.6.** The total exchange problem is the problem on a parallel computer of having every processor exchange a value with every other processor.

**Proposition 5.7.** In a CREW PRAM, broadcast takes constant time.

*Proof.* In a CREW PRAM, processor $P_0$ can broadcast to all other processors by writing in one time step to a specific memory location. In the next time step, all other processes can read from that location. \(\square\)

**Proposition 5.8.** In a CREW PRAM, total exchange takes time linear in the number of processors.

*Proof.* Similar to broadcast, all processors can write their information to their specific location in time step 1. Each processor now reads the $n - 1$ values in the next $n - 1$ time steps. \(\square\)

**Proposition 5.9.** In any architecture in which information is communicated one unit at a time, total exchange takes at least time linear in the number of processors.
Proof. The best possible situation is that of the CREW PRAM. Each processor must receive \( n - 1 \) units of information, and therefore regardless of how fast that information can be made available to be read by each processor, it cannot in fact be read by the processors in fewer than \( n - 1 \) time steps.

**Proposition 5.10.** In an architecture or with a problem in which information can be aggregated for multiple processors before being communicated, total exchange can be done at least as fast as time logarithmic in the number of processors.

Proof. Consider the following communication pattern. We have eight processors running across, exchanging information in time steps going down.

Processor \( i \) starts with an 8-bit string with only the \( i \)-th bit set, which we write in hex. The arcs going down represent addition (or XOR). After \( \lg n \)
steps, every processor has received information from every other processor.

The communication pattern described above is called a **butterfly** and is associated, among other things, with the Fast Fourier Transform (FFT).

Another situation in which this communication pattern could be useful would be a tree search to optimize an objective function.

However, in a two-dimensional mesh/grid connected machine, things are much different. Think about this.

**Definition 5.11.** The cost of a (PRAM) computation is the product of the number of processors and the time complexity of the computation.

For example, parallel merge (discussed below) takes time $\lg n$ on $n$ processors and thus has cost $n \lg n$.

**Definition 5.12.** A parallel computation (algorithm) is said to be cost optimal if the cost is in the same time complexity as the optimal sequential algorithm. of a (PRAM) computation is the product of the number of processors and the time complexity of the computation.

For example, parallel merge as discussed below is not cost optimal since the sequential time (1 processor times $n$ comparisons) is asymptotically smaller than the parallel $n \lg n$ cost.

As always with parallel algorithms, sometimes we might care (about cost-optimality), and sometimes we might not.
5.4 Complications and General Issues in Parallelism

- Processors are fast.

- Communications are either fast and expensive to build (the switch in an SMP) or inexpensive but slow (the network connection in a SNOW).

- Memory is slow and becoming slower relative to processors.

- We have had for 25 years the problem of getting “gigabits through punybaud”

- The problem in general is one of having local information on separate processors and needing global information distributed to all processors or else concentrated in one processor.

- If we are lucky, then there is a regular structure that will help us with the information flow (e.g., butterfly, parallel prefix).

- If we are not lucky, then we need to look at queueing theory or at parameter measurement (e.g., beowulf, SNOW)

- Either way, we must know what the rules are for communications among processors. Consider broadcast and total exchange as an example.

For example, on a 2-dimensional mesh, we can ask

- Can we write to all four links at once?
- Can we read from all four links at once?
- Can we read and write simultaneously on a link?
Or, for example, on a beowulf cluster or SNOW, we can ask

- What's the actual size of the memory block we can transfer?
- What does MPI (or PVM) do to reduce hot spots?
- What's the average transfer rate unloaded and loaded?

- Are we going in lockstep, or are we running separate processes at their own pace?

- What's the slowdown in going from one architecture to another?

E.g., how do we do a binary tree on a mesh?
5.5 A Systolic Sorting Algorithm and Machine

We shall loosely refer to an algorithm as **systolic** if it operates by pumping the data past processors in much the way that blood is pumped through the vascular system.

We shall use this concept to construct a parallel “machine” with \( n \) processors that will sort \( n \) data elements in time \( 2n \).

The processors will be arranged in a one-dimensional linear array. Data will be pumped in at one end (the left end) of the array, and sorted lists will be pumped out the other end (the right end) of the array.

Each basic processor will consist of a **new** register for a data item input to the processor, a **current** register for the data item retained from the last time step, and a comparator. Processing will take place by having each processor receive an item from the processor to the left and place it in the new register, compare the new value with the current value, and pump the larger of the two to the processor to the right.

A precise implementation depends on the communication rules and operating procedures. (For example, can we receive data as new, compare it with current, and push the larger to the right in one time step, or will these separate operations take separate time steps?)

One such implementation is as follows, of four processors, to which we input the string 8, 5, 3, 7. We will label the new and current registers as \( N_i \) and \( C_i \), respectively. We will also subscript the otherwise identical dummy values so that they can be distinguished for this example.
<table>
<thead>
<tr>
<th>Tick</th>
<th>$P_0$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$N_1 = \infty_1$</td>
<td>$N_2 = \infty_3$</td>
<td>$N_3 = \infty_5$</td>
<td>$N_4 = \infty_7$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = \infty_2$</td>
<td>$C_2 = \infty_4$</td>
<td>$C_3 = \infty_6$</td>
<td>$C_4 = \infty_8$</td>
</tr>
<tr>
<td>1</td>
<td>$N_1 = 7$</td>
<td>$N_2 = \infty_2$</td>
<td>$N_3 = \infty_4$</td>
<td>$N_4 = \infty_6$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = \infty_1$</td>
<td>$C_2 = \infty_3$</td>
<td>$C_3 = \infty_5$</td>
<td>$C_4 = \infty_7$</td>
</tr>
<tr>
<td>2</td>
<td>$N_1 = 3$</td>
<td>$N_2 = \infty_1$</td>
<td>$N_3 = \infty_3$</td>
<td>$N_4 = \infty_5$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = 7$</td>
<td>$C_2 = \infty_2$</td>
<td>$C_3 = \infty_4$</td>
<td>$C_4 = \infty_6$</td>
</tr>
<tr>
<td>3</td>
<td>$N_1 = 5$</td>
<td>$N_2 = 7$</td>
<td>$N_3 = \infty_2$</td>
<td>$N_4 = \infty_4$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = 3$</td>
<td>$C_2 = \infty_1$</td>
<td>$C_3 = \infty_3$</td>
<td>$C_4 = \infty_5$</td>
</tr>
<tr>
<td>4</td>
<td>$N_1 = 8$</td>
<td>$N_2 = 5$</td>
<td>$N_3 = \infty_1$</td>
<td>$N_4 = \infty_3$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = 3$</td>
<td>$C_2 = 7$</td>
<td>$C_3 = \infty_2$</td>
<td>$C_4 = \infty_4$</td>
</tr>
<tr>
<td>5</td>
<td>$N_1 = \infty_1$</td>
<td>$N_2 = 8$</td>
<td>$N_3 = 7$</td>
<td>$N_4 = \infty_2$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = 3$</td>
<td>$C_2 = 5$</td>
<td>$C_3 = \infty_1$</td>
<td>$C_4 = \infty_3$</td>
</tr>
<tr>
<td>6</td>
<td>$N_1 = \infty_1$</td>
<td>$N_2 = 3$</td>
<td>$N_3 = 8$</td>
<td>$N_4 = \infty_1$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = \infty_1$</td>
<td>$C_2 = 5$</td>
<td>$C_3 = 7$</td>
<td>$C_4 = \infty_2$</td>
</tr>
<tr>
<td>7</td>
<td>$N_1 = \infty_1$</td>
<td>$N_2 = \infty_1$</td>
<td>$N_3 = 5$</td>
<td>$N_4 = 8$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = \infty_1$</td>
<td>$C_2 = 3$</td>
<td>$C_3 = 7$</td>
<td>$C_4 = \infty_1$</td>
</tr>
<tr>
<td>8</td>
<td>$N_1 = \infty_1$</td>
<td>$N_2 = \infty_1$</td>
<td>$N_3 = 3$</td>
<td>$N_4 = 7$</td>
</tr>
<tr>
<td></td>
<td>$C_1 = \infty_1$</td>
<td>$C_2 = 3$</td>
<td>$C_3 = 5$</td>
<td>$C_4 = 8$</td>
</tr>
</tbody>
</table>

At this point we start pumping the sorted array out the end of the array.

Some things to notice.

The steady state of “receive, compare, output larger” is easy to describe. The initialization conditions are slightly harder.

Since we have no guarantee that $\infty$ isn’t a legitimate value, it is necessary to pass along a bit with the data. The bit will indicate that the data is
in fact valid data, so when the 8 appears along with the validity bit, we will take the 8 seriously as genuine data.

Finally, we note that with another bit for indicating breakpoint/initialization, we can follow the 8 with the beginning of another series of data items and keep all the processors busy all the time once we get going.
5.6 Parallel Summation

Clearly, we can apply the basic binary tree structure to anything like summation on a parallel machine. For example, we can add \( n \) items in \( \lg n \) time in parallel, in the obvious way.

\[
a + b + c + d + e + f + g + h
\]

And of course “add” can be extended to any associative operation.
5.7 Parallel Merge

**Theorem 5.13.** On a CREW PRAM, we can with \( n \) processors perform a merge of two sorted lists of \( n/2 \) elements in time \( \Theta(\lg n) \).

*Proof.* Consider two sorted lists, \( A \) and \( B \), each of length \( n/2 \). To each element \( a_i \) in list \( A \) we assign a processor \( P_{a_i} \). Using this processor, we can determine by binary search with \( \lceil \lg(n/2) \rceil \) comparisons that \( b_j \leq a_i \leq b_{j+1} \) for two elements \( b_j \) and \( b_{j+1} \) in list \( B \).

If we have indexed both lists from 1 through \( n/2 \), this means that \( a_i \) should be placed in location \( i + j \) in the merged list, because it should be preceded by the \( i - 1 \) elements \( a_1 \) through \( a_{i-1} \) and by the \( j \) elements \( b_1 \) through \( b_{j} \). (Note that duplicates don’t matter in this argument.)

Using the \( n/2 \) processors assigned to list \( A \) in this fashion, and using analogously the \( n/2 \) processors assigned to list \( B \), we can find in parallel the correct position in the sorted list of every element in each of lists \( A \) and \( B \) in \( \lceil \lg(n/2) \rceil \) steps.

All this involves concurrent reading, but no writing as yet. We now write, in parallel, each element into its proper location in the merged array. This can be done in parallel because each element goes into a different location.

(We have fudged ever so slightly on the question of duplicate values. To make this truly rigorous we must specify a consistent convention for breaking ties. Otherwise, in the case of \( a_i = a_{i+1} \) and \( b_j = b_{j+1} \), we would run the risk of matching \( i \) with \( j + 1 \) and \( i + 1 \) with \( j \), say, and writing the same value to location \( i + j + 1 \) but no value to \( i + j \).)

Note that we do \( n \lg n \) comparisons in the parallel algorithm instead of
the \( n \) of the serial algorithm. But we do them on \( n \) processors and hence \( n \) comparisons at a time, resulting in the \( \lg n \) execution time. So this is not cost optimal.
5.8 Parallel Prefix

The *prefix sums* of an array

\[ a_1, a_2, \ldots, a_n \]

are the sums

\[ a_1 \]
\[ a_1 \oplus a_2 \]
\[ a_1 \oplus a_2 \oplus a_3 \]
\[ \ldots \]
\[ a_1 \oplus a_2 \oplus \ldots \oplus a_n \]

where we can interpret \( \oplus \) to be any associative operation.

For example, we might have

<table>
<thead>
<tr>
<th>array</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>prefix sums</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>36</td>
</tr>
</tbody>
</table>
We can compute all the prefix sums in parallel in the following way.

<table>
<thead>
<tr>
<th>array</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_i \leftarrow a_{i-1} + a_i )</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>( a_i \leftarrow a_{i-2} + a_i )</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>18</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>( a_i \leftarrow a_{i-4} + a_i )</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>prefix sums</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>36</td>
</tr>
</tbody>
</table>

where in each of the intermediate steps we only do the addition if in fact the \( i - k \) subscript is positive.

**Theorem 5.14.** The prefix sums of an array of \( n \) elements can be computed on a CREW PRAM in \( \lfloor \log n \rfloor \) steps.

**Proof.** This should be obvious. \( \square \)
Now, why on earth would we want to do this? What does this rather odd looking prefix sum function do for us?

Look at in exactly the opposite way. Parallel prefix is a fast operation. Let’s look for ways in which it could be incorporated into some other algorithm.

If, for example, we are doing the “split” part of the quicksort, moving all the values less than the pivot to the first part of an array and all the values greater than the pivot to the last part of an array, then we can make use of parallel prefix in the following way.

Assume that the pivot element is 4, and that we have the array

\[
\begin{array}{cccccccc}
7 & 2 & 9 & 4 & 3 & 1 & 7 & 8 \\
\end{array}
\]

We assign a different processor to each element in the array, do all the comparisons against the pivot in parallel, and set a bit in a mask array for those elements less than or equal to the pivot.

\[
\begin{array}{cccccccc}
7 & 2 & 9 & 4 & 3 & 1 & 7 & 8 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
\end{array}
\]

Now do the parallel prefix.

\[
\begin{array}{cccccccc}
7 & 2 & 9 & 4 & 3 & 1 & 7 & 8 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 2 & 3 & 4 & 4 & 4 \\
\end{array}
\]

Now, if processor \( P_i \) reads locations \( i \) and \( i - 1 \) and finds different values in the prefix sums \( s_i \) and \( s_{i-1} \), it knows that it should write its value into location \( i \) in the “first part” array.
We create the “last part” of the array with an analogous operation, differing only in that we flip 0 and 1 in creating the mask array.

Compare this to the data movement part of the quicksort on a serial machine. There, we use $n$ comparisons to move $n$ elements. Here, we use $2 \lg n + 2$ for creating the two masks and doing the two parallel prefix operations.

So yes, prefix sums seem a little odd, but they can be useful!
5.9 Parallel List Ranking

Now consider the suffix sum problem in a linked list. Actually, we'll make
the problem slightly easier by making the values to be summed all 1 except
for the end of the list, to which we'll assign 0.

<table>
<thead>
<tr>
<th>Subscript</th>
<th>Value</th>
<th>Points to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

What we will do is iterate

\[
\text{suffix sum} += \text{value}[i] \\
\text{next}[i] = \text{next}[\text{next}[i]]
\]

\( \lg n \) times, provided that the \( \text{next}[\text{next}[i]] \) operation doesn't go beyond the end of the list. This gives us the following.
Iteration 0.

<table>
<thead>
<tr>
<th>Sub</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ptr</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Iteration 1.

<table>
<thead>
<tr>
<th>Sub</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ptr</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Iteration 2.

<table>
<thead>
<tr>
<th>Sub</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ptr</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Iteration 3.

<table>
<thead>
<tr>
<th>Sub</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ptr</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

When we’re done, all elements point to the end of the list, and the value held is the suffix sum. In this case, the suffix sum is the sequence position of that element measured from the end of the list.

Takes \( n \log n \) operations in \( \log n \) time.
5.10 Suffix Sums for Depth-First Search

Consider depth first search (this is also called preorder traversal) through a graph such as the one below.

```
  A
 /\  \
B   C
 /     \
D   E   F
   /\  /\  \
  G  H
```

The standard depth first search would be something like the following.

```
Depth-First Search(nodepointer)
{
    if(nodepointer != null)
    {
        nodecount += 1
        nodepointer.label = nodecount
        Depth-First Search(nodepointer.left)
        Depth-First Search(nodepointer.right)
    }
}
```

Let’s do this in parallel (CREW PRAM) by focusing not on the nodes but on the edges.

We have a tree with \( n \) nodes.

Then we have a tree with \( n - 1 \) edges.

We will use a separate processor for edge traversal going down from that for edge traversal going up.
Hence $2n - 2$ processors.

1. Every processor links its node to its successor node in the DFS traversal.

2. Processors for a downward edge write a 1 into a mask array. Processors for an upward edge write a 0 into a mask array.

3. Run suffix sum in parallel on the weights of step 2.

4. Use suffix sums in parallel to assign traversal values.

We use a data structure for each node as follows. Note among other things that this has constant space whether or not the tree structure is binary, ternary, or mixed.

\[
\begin{array}{cccccccc}
A & B & C & D & E & F & G & H \\
\text{parent} & \text{null} & A & A & B & B & C & E & E \\
\text{right-sibling} & \text{null} & C & \text{null} & E & \text{null} & \text{null} & H & \text{null} \\
\text{leftmost-child} & B & D & F & \text{null} & G & \text{null} & \text{null} & \text{null} \\
\end{array}
\]

**In greater detail:**

We the processors know whether we are matched with an upward edge or a downward edge for the following reason.

If the edge is $(i, j)$ linking nodes $i$ and $j$, and if the parent of node $i$ is $j$, then we are an upward edge.

Otherwise, we are a downward edge.

1. *Every processor links its node to its successor node in the DFS traversal.*

  向上端的边连接到父节点到右子节点的边，如果右子节点存在，否则连接到父节点到祖节点的边,
else to itself (at the end of the search). Downward edges link to the edge to the leftmost child node, if that child exists, else to the upward edge between the same nodes.

\[(A, B) \rightarrow (B, D) \rightarrow (D, B) \rightarrow (B, E) \rightarrow (E, G) \rightarrow (G, E) \rightarrow (E, H) \rightarrow (H, E) \rightarrow (E, B) \rightarrow (B, A) \rightarrow (A, C) \rightarrow (C, F) \rightarrow (F, C) \rightarrow (C, A)\]

2. **Processors for a downward edge write a 1 into a mask array. Processors for an upward edge write a 0 into a mask array.**

We know who we are, and write 0 or 1 accordingly.

\[(A, B), 1 \rightarrow (B, D), 1 \rightarrow (D, B), 0 \rightarrow (B, E), 1 \rightarrow (E, G), 1 \rightarrow (G, E), 0 \rightarrow (E, H), 1 \rightarrow (H, E), 0 \rightarrow (E, B), 0 \rightarrow (B, A), 0 \rightarrow (A, C), 1 \rightarrow (C, F), 1 \rightarrow (F, C), 0 \rightarrow (C, A), 0\]

3. **Run suffix sum in parallel on the weights of step 2.**

We saw how to do this in the previous section.

\[(A, B), 1, 7 \rightarrow (B, D), 1, 6 \rightarrow (D, B), 0, 5 \rightarrow (B, E), 1, 5 \rightarrow (E, G), 1, 4 \rightarrow (G, E), 0, 3 \rightarrow (E, H), 1, 3 \rightarrow (H, E), 0, 2 \rightarrow (E, B), 0, 2 \rightarrow (B, A), 0, 2 \rightarrow (A, C), 1, 2 \rightarrow (C, F), 1, 1 \rightarrow (F, C), 0, 0 \rightarrow (C, A), 0, 0\]
4. Use suffix sums in parallel to assign traversal values.

If we are a downward edge, we assign value \( n + 1 - \text{suffix} \) to the destination node of the arc as the traversal value. Note that destination nodes of downward arcs are unique, so only one value per node gets assigned.

\[(A, B), 1, 7, B = 2 \rightarrow \quad (B, D), 1, 6, D = 3 \rightarrow \quad (D, B), 0, 5, * \rightarrow \]
\[(B, E), 1, 5, E = 4 \rightarrow \quad (E, G), 1, 4, G = 5 \rightarrow \]
\[(G, E), 0, 3, * \rightarrow \quad (E, H), 1, 3, H = 6 \rightarrow \]
\[(H, E), 0, 2, * \rightarrow \quad (E, B), 0, 2, * \rightarrow \]
\[(B, A), 0, 2, * \rightarrow \quad (A, C), 1, 2, C = 7 \rightarrow \]
\[(C, F), 1, 1, F = 8 \rightarrow \quad (F, C), 0, 0, * \rightarrow \]
\[(C, A), 0, 0, A = 1 \]
Depth First Search
/* Global variables */
n  /* number of nodes in the tree */
parent[1..n]  /* vertex number of parent node */
child[1..n]  /* vertex number of leftmost child node */
sibling[1..n]  /* vertex number of right sibling node */
successor[1..n]  /* edge number of successor edge */
position[1..n]  /* edge rank after suffix sum */
DFS_value[1..n]  /* DFS traversal number */
begin
  spawn one processor for each edge
  for all processors P(i,j)
  {
    /* phase 1: create linked list */
    if(parent[i] == j)  /* upward arc */
    {
      if(sibling[i] != null)
        successor[i,j] = (j,sibling[i])
      elseif(parent[i] != null)
        successor[i,j] = (j,parent[j])
      else
        successor[i,j] = (i,j)
        DFS[j] = 1  /* j is the root of the tree */
    }
    /* end if parent[i] == j */
    else /* parent[i] not j, downward arc */
    {
      if(child[j] != null)
        successor[i,j] = (j,child[j])
      else
        successor[i,j] = (j,1)
    } /* end if parent not j and downward arc */
  } /* end for all processors */
/* phase 2: weight the edges based on upward or downward */
if(parent[i] == j)
  position[(i,j)] = 0
else
position[(i, j)] = 1

/* phase 3: suffix sum */
for(k = 1; k < ceiling(log(2n-2)); k++)
{
    position[(i, j)] += position[successor[(i, j)]]
    successor[(i, j)] = successor[successor[(i, j)]]
} /* end for */

/* phase 4: assign preorder values */
if(i == parent[j])
    DFS_value[j] = n + 1 - position[(i, j)]
} /* end for all processors */
end
5.11 Brent’s Theorem

**Theorem 5.15.** Given a parallel algorithm $A$ that performs $m$ computation steps in time $t$, then $p$ processors can execute algorithm $A$ in time

$$t + \frac{m - t}{p}.$$

**Proof.** Let $s_i$ denote the number of computation steps performed in time step $i$, for $1 \leq i \leq t$. Then

$$\sum_{i=1}^{t} s_i = m.$$

Using $p$ processors we can simulate step $i$ in time $\lfloor s_i/p \rfloor$. So the entire algorithm can be simulated by $p$ processors in time

$$\sum_{i=1}^{t} \left\lfloor \frac{s_i}{p} \right\rfloor \leq \sum_{i=1}^{t} \frac{s_i + p - 1}{p}$$

$$= \sum_{i=1}^{t} \frac{p}{p} + \sum_{i=1}^{t} \frac{s_i}{p} - \sum_{i=1}^{t} \frac{1}{p}$$

$$= t + \frac{m - t}{p}.$$

$\Box$
5.12 Reducing Processor Count

We can apply Brent’s theorem as follows.

5.12.1 Parallel Summation

We can add $n$ items in time $\lg n$, but this requires $n$ processors doing $n \lg n$ additions, which is not cost optimal. Consider, however, that the $n$ processors are only needed in the first step. After that, we need fewer and fewer processors. So we balance the first part, needing lots of processors, with the latter part, needing several steps but fewer processors.

Applying Brent’s theorem, if we use $\lceil \frac{n}{\lg n} \rceil$ processors, we can do parallel summation in time

$$\lfloor \lg n \rfloor + \frac{n - 1 - \lfloor \lg n \rfloor}{\lceil \frac{n}{\lg n} \rceil} = \Theta(\lg n + \lg n - \frac{\lg n}{n} - \frac{\lg^2 n}{n}) = \Theta(\lg n)$$

5.12.2 Parallel Prefix

Theorem 5.16. Using $\Theta(n/\lg n)$ processors, parallel prefix can be computed with optimal cost and with minimum execution time

$$\Theta(\frac{n}{p} + \lg p)$$

Proof. Parallel prefix as presented earlier takes $\lceil \log n \rceil$ iterations, and in each iteration $i$ we do $n - 2^i$ operations. So the total number of operations is

$$\sum_{i=0}^{\lceil \log n - 1 \rceil} (n - 2^i) = n \log n - (2^{\lceil \log n \rceil} - 1) = \Theta(n \log n).$$
This is not cost optimal, because a single processor can do all the prefix sums in \( n - 1 \) operations.

Let’s do the prefix sums of \( n \) things using \( p \) processors.

1. Break up the \( n \) items into \( p \) data blocks of less equal \( \lceil \frac{n}{p} \rceil \) things each.

2. Each processor (except the last) sequentially computes the overall sum of the values in its block of data. This is done in \( \lceil \frac{n}{p} \rceil - 1 \) time steps.

3. The \( p \) processors use the parallel algorithm to compute the prefix sums of the blocks in \( \lceil \log(p - 1) \rceil \) time steps.

4. The processors now sequentially update their local prefix sums with the global prefix sums of blocks, taking \( \lceil \frac{n}{p} \rceil \) time steps.

Total time is

\[
\left\lceil \frac{n}{p} \right\rceil - 1 + \lceil \log(p - 1) \rceil + \left\lceil \frac{n}{p} \right\rceil = \Theta \left( \frac{n}{p} + \log p \right).
\]

The total number of addition steps is \( \Theta(n + p \log p) \).

Now plug this in to Brent’s Theorem, which says we can get an execution time of

\[
\Theta \left( \frac{n}{p} + \log p + \frac{n + p \log p - \left( \frac{n}{p} + \log p \right)}{p} \right) = \Theta \left( \frac{n}{p} + \log p \right).
\]

Now, optimizing this isn’t entirely simple. The execution time function has a minimum at \( n = p \). The cost function has a min at \( \log p = -1 \), so that we would use fewer than one processor.
From looking at these two functions, it should be clear that the optimal value is going to be based on constraining each function with the other.

If we let \( p = n/g(n) \), then the time looks like

\[
g(n) + \log n - \log g(n)
\]

and is no larger than \( \log n \) provided \( g(n) \leq \log n \).

If we let \( p = n/g(n) \), then the cost looks like

\[
n + (n/g(n))(\log n - \log g(n))
\]

and is no larger than \( n \) provided \( g(n) \geq \log n \).

So we want \( p = \Theta\left(\frac{n}{\log n}\right) \) to minimize both time and cost together.
5.13 Lessons from Parallel Algorithms

On the one hand, PRAM models are unrealistic. Very often, PRAM algorithms have little real bearing on how we might choose to compute things on a parallel computer.

On the other hand, besides having four fingers and a thumb, sometimes we either can get some useful insight from a PRAM algorithm or else we could adapt it to be useful on a real machine.

One way in which we can apply parallel algorithms is to determine the cost in changing an unrealistic parallel algorithm into a more realistic parallel algorithm.
5.14 Embeddings and Dilations

A number of PRAM algorithms have already been seen to rely upon a binary tree structure. Computers, such as the Thinking Machines CM-2 and CM-5, in which the interconnect and communication structure were in fact that of a binary tree, permitted such "unrealistic" algorithms to be implemented on those "realistic" machines.

On other machines, such as the T3E, in which the interconnect supports a 2-dimensional mesh, it may be desirable to implement the communication structure of a binary tree.

**Definition 5.17.** An embedding of a graph $G = (V, E)$ into a graph $G' = (V', E')$ is a function $\phi$ from $V$ to $V'$.

**Definition 5.18.** Let $\phi$ be the function that embeds a graph $G = (V, E)$ into a graph $G' = (V', E')$. The dilation of the embedding is defined as follows:

$$dil(\phi) = \max\{\text{dist}(\phi(u), \phi(v)) \mid (u, v) \in E\}$$

where $\text{dist}(a, b)$ is the distance between vertices $a$ and $b$ in $G$.

Clearly, our goal is to find dilation-1 embeddings. These would permit an algorithm for communication structure $G$ to be implemented without slowdown on a machine with communication structure $G'$.

**Theorem 5.19.** There exists a dilation-1 embedding of a ring of $n$ nodes into a two-dimensional mesh of $n$ nodes if and only if $n$ is even.

**Proof.** Exercise. $\square$
Example

Theorem 5.20. A complete binary tree of height $> 4$ cannot be embedded into a 2-dimensional mesh without increasing the dilation beyond 1.

Proof. We count mesh points. In a 2-D mesh the number of nodes that are $k$ or fewer hops away from a given node is $2k^2 + 2k + 1$. The total number of nodes in a complete binary tree of height $k$ is $2^k - 1$. But $2^k - 1 > 2k^2 + 2k + 1$ for $k > 4$.  

We note that this is a proof that doesn’t tell us anything about the structure of embeddings or dilations.
Example

The following is a dilation-1 embedding of a binary tree of height 3 into a 2-D mesh.

The following is a dilation-1 embedding of a binary tree of height 4 into a 2-D mesh.
This is why the next one won’t work.

Count up the unused dots. We will need 32 dots for the next level of the binary tree, but there are only 30 dots unused. The structure is unimportant; no structure is going to provide enough dots to allow for the next tree level.
But now consider the following embedding using an H-tree.

**Theorem 5.21.** A complete binary tree of height $n$ has a dilation-$\lceil n/2 \rceil$ embedding into a 2-D mesh.

*Proof. *This is done recursively with an H-tree as above. 

### 5.15 Diameters and Bisections

**Definition 5.22.** The diameter of a network is the maximum distance between any two nodes.

**Definition 5.23.** The bisection width of a network is the minimum number of arcs that one must cut in order to separate the network into two halves ("half" to within one node).

In general we will want to compare and measure network communication capability based on four criteria.
1. diameter

2. bisection width

3. number of edges per node

4. the maximum edge length needed to realize the network

**Definition 5.24.** A hypercube is a network with $n$ nodes, numbered 0, 1, ..., $n-1$, in which nodes are connected if they differ by 1 in the Hamming distance between the binary representations of their node numbers.

**Definition 5.25.** A cube connected cycles network is a $k$-dimensional hypercube in which the nodes of the hypercube have been replaced by cycles of $k$ nodes with each node connecting along one edge of the hypercube and to two nodes of the $k$-cycle.
**Definition 5.26.** A shuffle exchange network consists of \( n = 2^k \) nodes with two kinds of connections. The exchange connections are bidirectional links between pairs of nodes that differ in the least significant bit. The shuffle connections are unidirectional links from node \( k \) and node \( 2k \pmod{n-1} \).

**Definition 5.27.** An omega network is a composition of \( k \) shuffle exchange networks.

<table>
<thead>
<tr>
<th>type</th>
<th>nodes</th>
<th>diameter</th>
<th>bisection width</th>
<th># edges ? constant?</th>
<th>constant edge length?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D mesh</td>
<td>( k )</td>
<td>( k - 1 )</td>
<td>1</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>2-D mesh</td>
<td>( k^2 )</td>
<td>( 2(k - 1) )</td>
<td>( k )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>3-D mesh</td>
<td>( k^3 )</td>
<td>( 3(k - 1) )</td>
<td>( k^2 )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>binary tree</td>
<td>( 2^k - 1 )</td>
<td>( 2(k - 1) )</td>
<td>1</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>butterfly</td>
<td>((k + 1)2^k)</td>
<td>( 2k )</td>
<td>( 2^k )</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>hypercube</td>
<td>( 2^k )</td>
<td>( k )</td>
<td>( 2^{k-1} )</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>cube conn cycles</td>
<td>( k2^k )</td>
<td>( 2k )</td>
<td>( 2^{k-1} )</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>shuffle exchange</td>
<td>( 2^k )</td>
<td>( 2k - 1 )</td>
<td>( \geq 2^{k-1}/k )</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

**Theorem 5.28.** A dilation-1 embedding of a complete binary tree of height \( n \) into a hypercube of dimension \( n + 1 \) does not exist for \( n > 1 \).

*Proof.* Note that a complete binary tree of height \( n \) has \( 2^{n+1} - 1 \) nodes, and an \( n + 1 \)-dimensional hypercube has \( 2^{n+1} \) nodes. So the optimal embedding, if it were to exist, would be of a height-\( n \) tree into an \( n + 1 \) dimensional cube. And the optimal embedding would have to use all the nodes of the hypercube except one.
Now, a tree of height $n$ for $n$ odd has more than half its nodes at an odd distance from the root. (height $n$, $2^{n+1} - 1$ nodes in all, and $2^n$ nodes that are the leaves at height exactly $n$). It is therefore not possible to embed the tree into the cube.

If, on the other hand, the height is even, then more than half the nodes are an even distance away from the root. Mutatis mutandis, the embedding isn’t possible.

We can’t do as well as we would like to. However, we can do almost as well.

**Theorem 5.29.** A dilation-2 embedding of a complete binary tree of height $n$ into a hypercube of dimension $n + 1$ exists for all $n > 1$.

**Theorem 5.30.** A dilation-1 embedding of a balanced binary tree of height $n$ into a hypercube of dimension $n + 2$ exists for all $n > 1$.

In a different direction, we cannot do as well as we would like for a mesh, but once again we can do the next best thing.

**Theorem 5.31.** Any 2-dimensional mesh with $n$ vertices can be embedded into a $\lceil \log n \rceil$-dimensional hypercube with dilation 2.