Chapter 4

Distributed Sorting

“Indeed, I believe that virtually every important aspect of programming arises somewhere in the context of sorting [and searching]!”

– Donald E. Knuth, The Art of Computer Programming

In this chapter we study a classic problem in computer science—sorting—from a distributed computing perspective. In contrast to an orthodox single-processor sorting algorithm, no node has access to all data, instead the to-be-sorted values are distributed. Distributed sorting then boils down to:

**Definition 4.1 (Sorting).** We choose a graph with $n$ nodes $v_1, \ldots, v_n$. Initially each node stores a value. After applying a sorting algorithm, node $v_k$ stores the $k^{th}$ smallest value.

**Remarks:**
- What if we route all values to the same central node? This is not legal. With a star topology sorting finishes in $O(1)$ time!

**Definition 4.2 (Node Contention).** In each step of a synchronous algorithm, each node can only send and receive $O(1)$ messages containing $O(1)$ values, no matter how many neighbors the node has.

**Remarks:**
- Using Definition 4.2 sorting on a star graph takes linear time.

### 4.1 Array & Mesh

To get a better intuitive understanding of distributed sorting, we start with two simple topologies, the array and the mesh. Let us begin with the array:

#### Algorithm 16 Odd/Even Sort

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Given an array of $n$ nodes $(v_1, \ldots, v_n)$, each storing a value (not sorted).</td>
</tr>
<tr>
<td>2.</td>
<td>repeat</td>
</tr>
<tr>
<td>3.</td>
<td>Compare and exchange the values at nodes $i$ and $i+1$, $i$ odd</td>
</tr>
<tr>
<td>4.</td>
<td>Compare and exchange the values at nodes $i$ and $i+1$, $i$ even</td>
</tr>
<tr>
<td>5.</td>
<td>until done</td>
</tr>
</tbody>
</table>

**Remarks:**
- The compare and exchange primitive in Algorithm 16 is defined as follows: Let the value stored at node $i$ be $v_i$. After the compare and exchange node $i$ stores value $\min(v_i, v_{i+1})$ and node $i+1$ stores value $\max(v_i, v_{i+1})$.
- How fast is the algorithm, and how can we prove correctness/efficiency?
- The most interesting proof uses the so-called 0-1 Sorting Lemma. It allows us to restrict our attention to an input of 0’s and 1’s only, and works for any “oblivious comparison-exchange” algorithm. (Oblivious means: Whether you exchange two values must only depend on the relative order of the two values, and not on anything else.)

**Lemma 4.3 (0-1 Sorting Lemma).** If an oblivious comparison-exchange algorithm sorts all inputs of 0’s and 1’s, then it sorts arbitrary inputs.

**Proof.** We prove the opposite direction (does not sort arbitrary inputs $\Rightarrow$ does not sort 0’s and 1’s). Assume that there is an input $x = x_1, \ldots, x_n$ that is not sorted correctly. Then there is a smallest value $k$ such that the value at node $v_k$ after running the sorting algorithm is strictly larger than the $k^{th}$ smallest value $x(k)$. Define an input $x_j^* = 0 \leftrightarrow x_j \leq x(k)$, $x_j^* = 1$ else. Whenever the algorithm compares a pair of 1’s or 0’s, it is not important whether it exchanges the values or not, so we may simply assume that it does the same as on the input $x$. On the other hand, whenever the algorithm exchanges some values $x_i^* = 0$ and $x_j^* = 1$, this means that $x_i \leq x(k) < x_j$. Therefore, in this case the respective compare-exchange operation will do the same on both inputs. We conclude that the algorithm will order $x^*$ the same way as $x$, i.e., the output with only 0’s and 1’s will also not be correct.

**Theorem 4.4.** Algorithm 16 sorts correctly in $n$ steps.

**Proof.** Thanks to Lemma 4.3 we only need to consider an array with 0’s and 1’s. Let $j_k$ be the node with the rightmost (highest index) 1. If $j_k$ is odd (even) it will move in the first (second) step. In any case it will move right in every following step until it reaches the rightmost node $v_n$. Let $j_k$ be the node with the $k^{th}$ rightmost 1. We show by induction that $j_k$ is not “blocked” anymore (constantly moves until it reaches destination!) after step $k$. We have already anchored the induction at $k = 1$. Since $j_{k-1}$ moves after step $k-1$, $j_k$ gets a right 0-neighbor for each step after step $k$. (For matters of presentation we omitted a couple of simple details.)
Algorithm 17 Shearsort

1. We are given a mesh with $m$ rows and $m$ columns, $m$ even, $n = m^2$.
2. The sorting algorithm operates in phases, and uses the odd/even sort algorithm on rows or columns.
3. repeat
4. In the odd phases 1, 3, . . . we sort all the rows, in the even phases 2, 4, . . . we sort all the columns, such that:
5. Columns are sorted such that the small values move up.
6. Odd rows (1, 3, . . . , $m − 1$) are sorted such that small values move left.
7. Even rows (2, 4, . . . , $m$) are sorted such that small values move right.
8. until done

Remarks:
- Linear time is not very exciting, maybe we can do better by using a different topology? Let’s try a mesh (a.k.a. grid) topology first.

Theorem 4.5. Algorithm 17 sorts $n$ values in $\sqrt{\pi}n \log n + 1$ time in snake-like order.

Proof. Since the algorithm is oblivious, we can use Lemma 4.3. We show that after a row and a column phase, half of the previously unsorted rows will be sorted. More formally, let us call a row with only 0’s (or only 1’s) clean after a row and a column phase, half of the previously unsorted rows will be sorted. Since the algorithm is oblivious, we can use Lemma 4.3. We show that after a row and a column phase, half of the previously unsorted rows will be sorted.

Definition 4.6 (Sorting Networks). A comparator is a device with two inputs $x, y$ and two outputs $x', y'$ such that $x' = \min(x, y)$ and $y' = \max(x, y)$. We construct so-called comparison networks that consist of wires that connect comparators (the output port of a comparator is sent to an input port of another comparator). Some wires are not connected to comparator outputs, and some are not connected to comparator inputs. The first are called input wires of the comparison network, the second output wires. Given $n$ values on the input wires, a sorting network ensures that the values are sorted on the output wires. We will also use the term width to indicate the number of wires in the sorting network.

Remarks:
- The odd/even sorter explained in Algorithm 16 can be described as a sorting network.
- Often we will draw all the wires on $n$ horizontal lines (the “width” of the network). Comparators are then vertically connecting two of these lines.
- Note that a sorting network is an oblivious comparison-exchange network. Consequently we can apply Lemma 4.3 throughout this section. An example sorting network is depicted in Figure 4.1.

Definition 4.7 (Depth). The depth of an input wire is 0. The depth of a comparator is the maximum depth of its input wires plus one. The depth of an output wire of a comparator is the depth of the comparator. The depth of a comparison network is the maximum depth (of an output wire).

Definition 4.8 (Bitonic Sequence). A bitonic sequence is a sequence of numbers that first monotonically increases, and then monotonically decreases, or vice versa.
4.2. SORTING NETWORKS

Figure 4.1: A sorting network.

Remarks:
- $<1, 4, 6, 8, 3, 2>$ or $<5, 3, 2, 1, 4, 8>$ are bitonic sequences.
- $<9, 6, 2, 3, 5, 4>$ or $<7, 4, 2, 5, 9, 8>$ are not bitonic.
- Since we restrict ourselves to 0's and 1's (Lemma 4.3), bitonic sequences have the form $0^l1^r0^s$ or $1^r0^l1^s$ for $i, j, k \geq 0$.

Algorithm 18 Half Cleaner
1. A half cleaner is a comparison network of depth 1, where we compare wire $i$ with wire $i + n/2$ for $i = 1, \ldots, n/2$ (we assume $n$ to be even).

Lemma 4.9. Feeding a bitonic sequence into a half cleaner (Algorithm 18), the half cleaner cleans (makes all-0 or all-1) either the upper or the lower half of the $n$ wires. The other half is bitonic.

Proof. Assume that the input is of the form $0^l1^r0^s$ for $i, j, k \geq 0$. If the midpoint falls into the 0’s, the input is already clean/bitonic and will stay so. If the midpoint falls into the 1’s the half cleaner acts as Shearsort with two adjacent rows, exactly as in the proof of Theorem 4.5. The case $1^r0^l1^s$ is symmetric.

Algorithm 19 Bitonic Sequence Sorter
1. A bitonic sequence sorter of width $n$ (n being a power of 2) consists of a half cleaner of width $n$, and then two bitonic sequence sorters of width $n/2$ each.
2. A bitonic sequence sorter of width 1 is empty.

Lemma 4.10. A bitonic sequence sorter (Algorithm 19) of width $n$ sorts bitonic sequences. It has depth $\log n$.

Algorithm 20 Merging Network
1. A merging network of width $n$ is a merger of width $n$ followed by two bitonic sequence sorters of width $n/2$. A merger is a depth-one network where we compare wire $i$ with wire $n - i + 1$, for $i = 1, \ldots, n/2$.

Remarks:
- Clearly we want to sort arbitrary and not only bitonic sequences! To do this we need one more concept, merging networks.

Algorithm 21 Batcher’s “Bitonic” Sorting Network
1. A batcher sorting network of width $n$ consists of two batcher sorting networks of width $n/2$ followed by a merging network of width $n$. (See Figure 4.2.)
2. A batcher sorting network of width 1 is empty.

Theorem 4.12. A sorting network (Algorithm 21) sorts an arbitrary sequence of $n$ values. It has depth $O(\log^2 n)$.

Proof. Correctness is immediate: at recursive stage $k$ ($k = 1, 2, 3, \ldots, \log n$) we merge $2^k$ sorted sequences into $2^{k-1}$ sorted sequences. The depth $d(n)$ of the sorting network of level $n$ is the depth of a sorting network of level $n/2$ plus the depth $m(n)$ of a merging network with width $n$. The depth of a sorter of level 1 is 0 since the sorter is empty. Since a merging network of width $n$ has the same depth as a bitonic sequence sorter of width $n$, we know by Lemma 4.10 that $m(n) = \log n$. This gives a recursive formula for $d(n)$ which solves to $d(n) = \frac{1}{2} \log^2 n + \frac{1}{2} \log n$.
Remarks:

- Simulating Batcher’s sorting network on an ordinary sequential computer takes time $O(n \log^3 n)$. As said, there are sequential sorting algorithms that sort in asymptotically optimal time $O(n \log n)$. So a natural question is whether there is a sorting network with depth $O(\log n)$. Such a network would have some remarkable advantages over sequential asymptotically optimal sorting algorithms such as heapsort. Apart from being highly parallel, it would be completely oblivious, and as such perfectly suited for a fast hardware solution. In 1983, Ajtai, Komlos, and Szemeredi presented a celebrated $O(\log n)$ depth sorting network. (Unlike Batcher’s sorting network the constant hidden in the big-$O$ of the “AKS” sorting network is too large to be practical, however.)

- It can be shown that Batcher’s sorting network and similarly others can be simulated by a Butterfly network and other hypercubic networks, see next chapter.

- What if a sorting network is asynchronous?!! Clearly, using a synchronizer we can still sort, but it is also possible to use it for something else. Check out the next section!

4.3 Counting Networks

In this section we address distributed counting, a distributed service which can for instance be used for load balancing.

Definition 4.13 (Distributed Counting). A distributed counter is a variable that is common to all processors in a system and that supports an atomic test-and-increment operation. The operation delivers the system’s counter value to the requesting processor and increments it.

Algorithm 22 Bitonic Counting Network.
1. Take Batcher’s bitonic sorting network of width $w$ and replace all the comparators with balancers.
2. When a node wants to count, it sends a message to an arbitrary input wire.
3. The message is then routed through the network, following the rules of the asynchronous balancers.
4. Each output wire is completed with a “mini-counter.”
5. The mini-counter of wire $k$ replies the value $k + i: w$ to the initiator of the $i^{th}$ message it receives.

Definition 4.14 (Balancer). A balancer is an asynchronous flip-flop which forwards messages that arrive on the left side to the wires on the right, the first to the upper, the second to the lower, the third to the upper, and so on.

Definition 4.15 (Step Property). A sequence $y_0, y_1, \ldots, y_{n-1}$ is said to have the step property, if $0 \leq y_i - y_j \leq 1$, for any $i < j$.

Remarks:

- A naive distributed counter stores the system’s counter value with a distinguished central node. When other nodes initiate the test-and-increment operation, they send a request message to the central node and in turn receive a reply message with the current counter value. However, with a large number of nodes operating on the distributed counter, the central processor will become a bottleneck. There will be a congestion of request messages at the central processor, in other words, the system will not scale.

- Is a scalable implementation (without any kind of bottleneck) of such a distributed counter possible, or is distributed counting a problem which is inherently centralized?!!

- Distributed counting could for instance be used to implement a load balancing infrastructure, i.e. by sending the job with counter value $i$ (modulo $n$) to server $i$ (out of $n$ possible servers).

4.3. COUNTING NETWORKS 38

Remarks:

- Is a scalable implementation (without any kind of bottleneck) of such a distributed counter possible, or is distributed counting a problem which is inherently centralized?!!

- Distributed counting could for instance be used to implement a load balancing infrastructure, i.e. by sending the job with counter value $i$ (modulo $n$) to server $i$ (out of $n$ possible servers).

Definition 4.14 (Balancer). A balancer is an asynchronous flip-flop which forwards messages that arrive on the left side to the wires on the right, the first to the upper, the second to the lower, the third to the upper, and so on.

Algorithm 22 Bitonic Counting Network.
1. Take Batcher’s bitonic sorting network of width $w$ and replace all the comparators with balancers.
2. When a node wants to count, it sends a message to an arbitrary input wire.
3. The message is then routed through the network, following the rules of the asynchronous balancers.
4. Each output wire is completed with a “mini-counter.”
5. The mini-counter of wire $k$ replies the value $k + i: w$ to the initiator of the $i^{th}$ message it receives.

Definition 4.15 (Step Property). A sequence $y_0, y_1, \ldots, y_{n-1}$ is said to have the step property, if $0 \leq y_i - y_j \leq 1$, for any $i < j$.

Remarks:

- If the output wires have the step property, then with $r$ requests, exactly the values $1, \ldots, r$ will be assigned by the mini-counters. All we need to show is that the counting network has the step property. For that we need some additional facts.

Facts 4.16. For a balancer, we denote the number of consumed messages on the $i^{th}$ input wire with $x_i$, $i = 0, 1$. Similarly, we denote the number of sent messages on the $i^{th}$ output wire with $y_i$, $i = 0, 1$. A balancer has these properties:

1. A balancer does not generate output-messages; that is, $x_0 + x_1 \geq y_0 + y_1$ in any state.
2. Every incoming message is eventually forwarded. In other words, if we are in a quiescent state (no message in transit), then $x_0 + x_1 = y_0 + y_1$. 

Remarks:
(3) The number of messages sent to the upper output wire is at most one higher than the number of messages sent to the lower output wire: in any state \( y_0 = \lfloor (y_0 + y_1)/2 \rfloor \) (thus \( y_1 = \lfloor (y_0 + y_1)/2 \rfloor \)).

**Facts 4.17.** If a sequence \( y_0, y_1, \ldots, y_{w-1} \) has the step property,

(1) then all its subsequences have the step property,

(2) then its even and odd subsequences satisfy

\[
\sum_{i=0}^{w/2-1} y_{2i} = \left\lfloor \frac{w-1}{2} \sum_{i=0}^{w-1} y_i \right\rfloor \quad \text{and} \quad \sum_{i=1}^{w/2} y_{2i} = \left\lceil \frac{w-1}{2} \sum_{i=0}^{w-1} y_i \right\rceil.
\]

**Facts 4.18.** If two sequences \( x_0, x_1, \ldots, x_{w-1} \) and \( y_0, y_1, \ldots, y_{w-1} \) have the step property,

(1) and \( \sum_{i=0}^{w-1} x_i = \sum_{i=0}^{w-1} y_i \), then \( x_i = y_i \) for \( i = 0, \ldots, w-1 \).

(2) and \( \sum_{i=0}^{w-1} x_i = \sum_{i=0}^{w-1} y_i + 1 \), then there exists a unique \( j \) (\( j = 0, 1, \ldots, w-1 \)) such that \( x_j = y_j + 1 \), and \( x_{i} = y_{i} \) for \( i = 0, \ldots, w-1, i \neq j \).

**Remarks:**

- An alternative representation of Batcher’s network has been introduced in [AH94]. It is isomorphic to Batcher’s network, and relies on a Merger network \( M[w] \) which is defined inductively. \( M[w] \) consists of two \( M[w/2] \) networks (an upper and a lower one) whose output is fed to \( w/2 \) balancers. The upper balancer merges the even subsequence \( x_0, x_2, \ldots, x_{w-2} \), while the lower balancer merges the odd subsequence \( x_1, x_3, \ldots, x_{w-1} \). Call the outputs of these two \( M[w/2] \)’s \( z \) and \( z’ \) respectively. The final stage of the network combines \( z \) and \( z’ \) by sending each pair of wires \( z_{2i} \) and \( z’_{2i} \) into a balancer whose outputs yield \( y_{2i} \) and \( y’_{2i} \).

- It is enough to prove that a merger network \( M[w] \) preserves the step property.

**Lemma 4.19.** Let \( M[w] \) be a merger network of width \( w \). In a quiescent state (no message in transit), if the inputs \( x_0, x_1, \ldots, x_{w-2} \) resp. \( x_{w/2}, x_{w/2+1}, \ldots, x_{w-1} \) have the step property, then the output \( y_0, y_1, \ldots, y_{w-1} \) has the step property.

**Proof.** By induction on the width.

For \( w = 2 \): \( M[2] \) is a balancer and a balancer’s output has the step property (Fact 4.16.3).

For \( w > 2 \): Let \( z \) resp. \( z’ \) be the output of the upper respectively lower \( M[w/2] \) subnetwork. Since \( x_0, x_1, \ldots, x_{w/2-1} \) and \( x_{w/2}, x_{w/2+1}, \ldots, x_{w-1} \) both have the step property by assumption, their even and odd subsequences also have the step property (Fact 4.17.1). By induction hypothesis, the output of both \( M[w/2] \) subnetworks have the step property. Let \( Z := \sum_{i=0}^{w/2-1} z_i \) and \( Z’ := \sum_{i=0}^{w/2-1} z’_i \). From Fact 4.17.2 we conclude that \( Z = \sum_{i=0}^{w/2-1} x_i + \frac{1}{2} \sum_{i=0}^{w/2-1} x_i \) and \( Z’ = \sum_{i=0}^{w/2-1} x_i + \frac{1}{2} \sum_{i=0}^{w/2-1} x_i \). Since \( |a| + |b| \) and \( |a| + |b| \) differ by at most 1 we know that \( Z \) and \( Z’ \) differ by at most 1.

**Counting Networks**

If \( Z = Z’ \), Fact 4.18.1 implies that \( z_i = z’_i \) for \( i = 0, \ldots, w/2 - 1 \). Therefore, the output of \( M[w] \) is \( y_0 = z_0 \) for \( i = 0, \ldots, w-1 \). Since \( z_0, \ldots, z_{w/2-1} \) has the step property, so does the output of \( M[w] \) and the lemma follows.

If \( Z \) and \( Z’ \) differ by 1, Fact 4.18.2 implies that there exists a \( j \) such that \( z_j = z_j’ + 1 \) for \( j = 0, \ldots, w/2 - 1 \). Let \( l := \min \{ j, z_j’ + 1 \} \). Then, the output \( y_l \) (with \( i < 2j \)) is \( l + 1 \). The output \( y_{l+1} \) (with \( i > 2j \) + 1) is \( l \). The output \( y_{2l+1} \) and \( y_{2l+1} \) are balanced by the final balancer resulting in \( y_{2l} = l + 1 \) and \( y_{2l+2} = l \). Therefore \( M[w] \) preserves the step property.

A bitonic counting network is constructed to fulfill Lemma 4.19, i.e., the final output comes from a Merger whose upper and lower inputs are recursively merged. Therefore, the following theorem follows immediately.

**Theorem 4.20.** If a network is a counting network then it is also a sorting network, but not vice versa.

**Remarks:**

- Is every sorting network also a counting network? No. But surprisingly, the other direction is true!

**Theorem 4.21.** (Counting vs. Sorting) If a network is a counting network then it is also a sorting network, but not vice versa.

**Proof.** There are sorting networks that are not counting networks (e.g. odd/even sort, or insertion sort). For the other direction, let \( C \) be a counting network and \( I(C) \) be the isomorphic network, where every balancer is replaced by a comparator. Let \( I(C) \) have an arbitrary input of 0’s and 1’s; that is, some of the input wires have a 0, all others have a 1. There is a message at \( C \)”s input wire if and only if \( I(C) \)”s input wire is 0. Since \( C \) is a counting network, all messages are routed to the upper output wires. \( I(C) \) is isomorphic to \( C \), therefore a comparator in \( I(C) \) will receive a 0 on its upper (lower) wire if and only if the corresponding balancer receives a message on its upper (lower) wire. Using an inductive argument, the 0’s and 1’s will be routed through \( I(C) \) such that all 0’s exit the network on the upper wires whereas all 1’s exit the network on the lower wires. Applying Lemma 4.3 shows that \( I(C) \) is a sorting network.

**Definition 4.22.** A system is linearizable if the order of the values assigned reflects the real-time order in which they were requested. More formally, if there is a pair of operations \( a_1, a_2 \) where operation \( a_2 \) terminates before operation \( a_1 \) starts, and the logical order is ”\( a_2 \) before \( a_1 \)”, then a distributed system is not linearizable.

**Lemma 4.23.** (Linearizability). The bitonic counting network is not linearizable.
Proof. Consider the bitonic counting network with width 4 in Figure 4.3. Assume that two inc operations were initiated and the corresponding messages entered the network on wires 0 and 2 (both in light gray color). After having passed the second resp. the first balancer, these traversing messages “fall asleep”. In other words, both messages take unusually long time before they are received by the next balancer. Since we are in an asynchronous setting, this may be the case.

![Figure 4.3: Linearizability Counter Example.](image)

Remarks:
- Note that the example in Figure 4.3 behaves correctly in the quiescent state. Finally, exactly the values 0, 1, 2, 3 are allotted.
- It was shown that linearizability comes at a high price (the depth grows linearly with the width).

Chapter Notes

The technique used for the famous lower bound of comparison-based sequential sorting first appeared in [FJ59]. Comprehensive introductions to the vast field of sorting can certainly be found in [Knu73]. Knuth also presents the 0/1 principle in the context of sorting networks, supposedly as a special case of a theorem for decision trees of W. G. Bouricius, and includes a historic overview of sorting network research.

Using a rather complicated proof not based on the 0/1 principle, [Hab72] first presented and analyzed Odd/Even sort on arrays. Shearsort for grids first appeared in [SSS86] as a sorting algorithm both easy to implement and to prove correct. Later it was generalized to meshes with higher dimension in [SS89]. A bubble sort based algorithm is presented in [SS86]; it takes time $O(n \sqrt{\log n})$, but is fast in practice. Nevertheless, already [TK77] presented an asymptotically optimal algorithm for grid network which runs in $3n + O(n^{2/3} \log n)$ rounds for an $n \times n$ grid. A simpler algorithm was later found by [SS86] using $3n + O(n^{1/3})$ rounds.

Batcher presents his famous $O(\log^2 n)$ depth sorting network in [Bat68]. It took until [AKS83] to find a sorting network with asymptotically optimal depth $O(\log n)$. Unfortunately, the constants hidden in the big-O-notation render it rather impractical.

The notion of counting networks was introduced in [AHS91], and shortly afterward the notion of linearizability was studied by [HSW91]. Follow-up work in [AHS94] presents bitonic counting networks and studies contention in the counting network. An overview of research on counting networks can be found in [BH98].

Bibliography


