On the Average Case Complexity of Parallel Sublist Selection

Geppino Pucci
Wolf Zimmermann

TR-91-023
March, 1991

Abstract

The Sublist Selection Problem (SSP) is the following. Given an input list of nodes labeled True or False, extract the sublist of nodes labeled True. This paper analyzes the average case complexity of a parallel algorithm that solves SSP on the PRAM model of computation. The algorithm is based on the well-known recursive doubling technique. Doubly logarithmic upper and lower bounds are derived for the average number of iterations needed to produce the output list, under the assumption that all the nodes of the input list are marked False with probability $p$, independently of the other nodes. Finally, the exact number of iterations (up to lower order terms) is established in the case that the input list is drawn from the uniform distribution over all possible labelings.

Keywords: Analysis of Algorithms, Generating Function Technique, Parallel Algorithms, Pointer Jumping, PRAM Model.

---

1This work has been partly supported by the Ministero della Pubblica Istruzione of Italy and by the C.N.R. project “Sistemi Informatici e Calcolo Parallelo”

2On leave from the University of Pisa, Dipartimento di Informatica, Pisa, Italy.

3On leave from the University of Karlsruhe, Institut für Programmstrukturen und Datenorganisation, Karlsruhe, Germany.
1 Introduction

In sequential computation, the average case analysis of deterministic algorithms is a well established subject, its roots dating back to Knuth’s seminal work in *The Art of Computer Programming* [Knu73]. The average case complexity measure for an algorithm $A$ is based on the definition of a sample space $\Omega_A^n$, consisting of the inputs to $A$ of length $n$, and a probability measure over $\Omega_A^n$. The running time of the algorithm can be seen as a random variable $T_A(\omega)$ over $\Omega_A^n$ [VF90]. The average complexity of the algorithm is defined as $\bar{T}_A = \mathbb{E}[T_A]$.

In contrast with sequential computation, average case complexity of parallel deterministic algorithms is a largely unexplored field. In this paper we study the average case behavior of an algorithm designed for the PRAM (Parallel Random Access Machine) model of parallel computation [FW78]. A PRAM consists of a collection of sequential RAM machines and a global memory, with any processor able to access any memory cell in unit time. The algorithm that we study solves the Sublist Selection Problem (SSP). An SSP instance is a list $L$, whose nodes are labeled True or False. Nodes labeled True (resp., False) are called T-nodes (resp., F-nodes). The output is a list $T$ containing only the T-nodes in the same order in which they appear in $L$. Sublist selection is often found as a subroutine in many parallel algorithms for graph problems [GR88].

It is immediately seen that SSP can be efficiently solved in parallel by applying the recursive doubling paradigm, also known as pointer jumping [KR90]. Suppose that $L$ is stored in the PRAM shared memory as two vectors, $L[1, \ldots, n]$ and $S[1, \ldots, n]$. $L[i]$ contains the label of the node associated to location $i$, while $S[i]$ is a pointer to its successor (we assume that the last element of the list points to itself). The recursive doubling step is implemented by the following procedure:

```plaintext
procedure JUMP;
    begin
        forall $1 \leq i \leq n$ do in parallel
            if (not $L[S[i]]$) and ($S[i] \neq S[S[i]]$)
                then $S[i] := S[S[i]]$
            endif
        endfor
    end;
```

Using JUMP, we solve SSP as follows*:

```plaintext
procedure SSP;
    begin
        for $\lceil \log n \rceil + 1$ iterations do
            JUMP
        endfor
    end.
```

When the computation ends, each T-node points either to the next T-node, if any, or to the end of the list. A particular execution of the above algorithm is given in figure 1, where T-nodes are colored in black and F-nodes in white.

The pointer to the head of the output list can be obtained in two additional steps. In the first step, each node in the sublist notifies its successor. In the second step, the (only) node which

*Where not specified, the logarithms are taken to the base two
Figure 1: Pointer Jumping in Sublist Computation
was not notified provides the pointer to the head of the sublist. Note that the algorithm avoids concurrent access to the same memory cell by more than one processor. Therefore it runs correctly on the EREW (Exclusive-Read, Exclusive-Write) variant of the PRAM model.

The number of iterations required by the above algorithm is always fixed to be \([\log n] + 1\), regardless of the particular input instance. However, on many input lists, a much smaller number of iterations is needed to reach a fixed point, that is, an iteration where no doubling steps \((S'[i] := S[S[i]])\) are performed. Unfortunately, determining the occurrence of such a fixed point requires \(\Omega(\log n)\) time on an EREW PRAM but can be done in \(O(1)\) time by using a single cell in CRCW (Concurrent-Read, Concurrent-Write) mode [Kuc82]. Our average case analysis refers to the algorithm modified to run on an EREW PRAM augmented with a single CRCW-access cell:

```plaintext
procedure SSP;
    begin
        repeat
            JUMP
        until fixed point
    end.
```

In section 2 we derive upper and lower bounds for the average number of iterations, \(E[I]\), needed to reach a fixed point, under the assumption that the input list is drawn from a known distribution. More precisely, if any of the \(n\) nodes of the list has probability \(p\) to be labeled False, independently of the other nodes, we are able to show that \(E[I]\) is at most \([\log \log(n + 1) - \log \log \frac{1}{p}] + \frac{2-p}{p}\) and, for \(p \leq 1 - \frac{1}{n^\epsilon}\), for some positive \(\epsilon < 1/2\), \(E[I]\) is at least \([\log \log(n + 1) - \log \log \frac{1}{p}] - o(1)\). Note that the two bounds match, up to constant factors, for values of \(p\) as large as \(1 - \frac{1}{\log \log n}\). For the case \(p = \frac{1}{2}\), corresponding to a uniform distribution of the inputs, in section 3 we evaluate the exact value \(E[I]\) (up to lower order terms) by using a variation of the generating function method used in [Knu78]. Define \(\delta(n) = [\log \log n] - \log \log n\). For \(p = \frac{1}{2}\) we are able to show that

\[
E[I] = \begin{cases} 
\log \log n + 1 - e^{\frac{1}{2}} + O\left(\frac{\log^2 n}{n}\right) & \delta(n) = 0 \\
[\log \log n] + 1 + O\left(\frac{1}{\log(n)}\right) & \delta(n) > 0
\end{cases}
\]

where \(\gamma(n) = 2^{\delta(n)+1} - 1\).

2 Upper and lower bounds for the expected number of iterations

Let \(\mathcal{F}\) be the longest sublist of consecutive F-nodes in the input list \(\mathcal{L}\). A simple inductive argument shows that the number of iterations needed to reach the fixed point is \(I = [\log F] + 1\), where \(F = \text{length}(\mathcal{F})\). Note that, when the input set is regarded as a sample space, \(I\) and \(F\) are integer random variables. We assume that any list of length \(n\) with a total number of \(k\) F-nodes has probability \(p^k(1-p)^{n-k}\) to occur as an input, where \(p\) is a parameter of the analysis. Under this probability distribution, any node of the input list is an F-node with probability \(p\), independently of the labelings of the other nodes. We have:

**Theorem 1** \(E[I] \leq [\log \log(n + 1) - \log \log \frac{1}{p}] + \frac{2-p}{1-p}\).
Proof: We know that \( I = [\log F] + 1 \), hence
\[
E[I] = \sum_{k=1}^{[\log n]+1} \text{Prob}(I \geq k) = \sum_{k=1}^{[\log n]+1} \text{Prob}(F \geq 2^{k-1})
\] (1)

For a fixed \( s \), \( 1 \leq s \leq n \), let \( \{X_i^{(s)}\}, i = 1, \ldots, n - s + 1 \) be a collection of independent Bernoulli variables such that \( X_i^{(s)} = 1 \) if and only if the \( i \)th, \((i+1)\)th, \( \ldots \), \((i+s-1)\)th node of the list \( \mathcal{L} \) are F-nodes. Under the above probability distribution, \( \text{Prob}(X_i^{(s)} = 1) = p^s \) and, by the union bound,
\[
\text{Prob}(F \geq s) = \text{Prob}\left( \sum_{i=1}^{n-s+1} X_i^{(s)} \geq 1 \right) \leq \min\{1, (n-s+1)p^s\}
\]

Let \( k_0 = \max\{k| (n+1)p^{2k-1} \geq 1\} \). Clearly, \( k_0 = \lfloor \log \log(n+1) - \log \log \frac{1}{p} \rfloor + 1 \). We have:
\[
E[I] \leq k_0 + (n+1) \sum_{k > k_0} p^k \leq \lfloor \log \log(n+1) - \log \log \frac{1}{p} \rfloor + \frac{2 - p}{1 - p}
\] (2)

Note that for \( p \leq 1 - \frac{1}{\log \log n} \) relation (2) yields \( \text{E}[I] \in O(\log \log n) \), that is, on average, the number of iterations needed to reach a fixed point is exponentially smaller than in the worst case. However, for larger values of \( p \) \( (p \geq 1 - \frac{1}{\log \log n}) \), theorem 1 does not provide a significant bound, as (2) just gives \( \text{E}[I] \in O(\frac{1}{1-p}) \), with \( \frac{1}{1-p} \in \Omega(\log n) \).

**Theorem 2** Let \( p \leq 1 - \frac{1}{n^\epsilon} \), for some positive \( \epsilon < 1/2 \). Then \( \text{E}[I] \geq [\log \log(n+1) - \log \log \frac{1}{p}] - o(1) \).

**Proof:** Using the same notation as in theorem 1 we have:
\[
\text{Prob}(F \geq s) = \text{Prob}\left( \sum_{i=1}^{n-s+1} X_i^{(s)} \geq 1 \right) = 1 - \text{Prob}(\{\forall i, 1 \leq i \leq n-s+1 : X_i^{(s)} = 0\})
\]

because the \( X_i^{(s)} \)'s are nonnegative variables. Note that
\[
\text{Prob}(\{\forall i, 1 \leq i \leq n-s+1 : X_i^{(s)} = 0\}) \leq \text{Prob}(\{\forall h, 0 \leq h \leq \lfloor \frac{n}{s} \rfloor - 1 : X_{h+s}^{(s)} = 0\})
\]
as the first event is strictly contained in the second. The variables \( X_{h+s}^{(s)} \) are mutually independent, as the associated node spans do not overlap. Therefore,
\[
\text{Prob}(F \geq s) \geq 1 - (1 - p^s)^{\frac{n}{s}-1}
\]

By the Markov inequality we have that, for any \( k > 0 \),
\[
\text{E}[I] \geq k \text{Prob}(I \geq k) = k \text{Prob}(F \geq 2^{k-1})
\]

Choose \( \bar{k} = [\log \log n - \log \log \frac{1}{p}] \). Then \( s = 2^{\bar{k}-1} \leq \frac{\log_1 n}{2} \) and \( p^s \geq \frac{1}{\sqrt{n}} \). As a consequence
\[
(1 - p^s)^{\frac{n}{s}-1} \leq \left(1 - \frac{1}{\sqrt{n}}\right)^{\left(\frac{\log_1 n}{2} - 1\right)} \leq \left(1 - \frac{1}{\sqrt{n}}\right)^{\sqrt{n}}\left(\frac{2\sqrt{n}}{\log_1 n} - 1\right) \leq \exp\left(-\frac{2\sqrt{n}}{\log_1 n} - 1\right)
\]

4
The thesis follows by noticing that, for \( p \leq 1 - \frac{1}{n^2}, \varepsilon < 1/2, \)

\[
\lim_{n \to \infty} \left( \log \log_{1/p} n \right) \exp \left( -\frac{2\sqrt{n}}{\log_{1/p} n} - \frac{1}{\sqrt{n}} \right) = 0
\]  

\[\blacksquare\]

3 Average number of iterations under the uniform distribution

In this section, we assume that the inputs are drawn from the uniform distribution, that is, any True/False labeling of the \( n \) nodes occurs with probability \( 2^{-n} \). Under this assumption we are able to evaluate the exact value of \( E[I] \) (up to lower order terms). The main result of this section is the following theorem.

**Theorem 3** Let \( n \) be the length of the input list, let \( \delta(n) = \lfloor \log \log n \rfloor - \log \log n \), and let \( \gamma(n) = 2^\delta(n) - 1 \). If the inputs are drawn according to the uniform distribution over all the labeled lists of \( n \) nodes then:

\[
E[I] = \begin{cases} 
\log \log n + 1 - e^{\frac{1}{2}} + O \left( \frac{\log^2 n}{n} \right) & \delta(n) = 0 \\
[\log \log n] + 1 + O \left( \frac{1}{n^{\frac{1}{4}(\delta(n))}} \right) & \delta(n) > 0 
\end{cases}
\]

In order to prove theorem 3, we need some notation and preliminary lemmas. Let \( p_{n,k} \) be the probability that a random sequence in \{True, False\}^n contains a subsequence of at least \( k \) consecutive False's; and let \( a_{n,k} \) be the number of such subsequences. Under the uniform distribution on the input space, we have \( p_{n,k} = \text{Prob}(F \geq k) \), hence \( E[I] = \sum_{k=1}^{\lfloor \log n \rfloor + 1} p_{n,2k-1} \). Moreover

\[
p_{n,k} = \frac{a_{n,k}}{2^n}
\]

We start with the following observation about the \( a_{n,k} \):

**Lemma 1** For any \( n \geq 0 \), \( a_{n,k} \) satisfies the following recurrence:

\[
a_{n,k} = \begin{cases} 
2^n & k = 0 \\
2a_{n-1,k} + 2^{n-1-k} - a_{n-1-k,k} & 1 \leq k < n \\
1 & k = n \\
0 & k > n 
\end{cases}
\]

**Proof:** Induction over \( n \).

\( n = 0 \): There is precisely one sequence of length 0, the empty sequence. Thus \( a_{0,0} = 1(= 2^0) \) and \( a_{0,k} = 0 \) if \( k > 0 \).

\( n > 0 \): Consider a sequence \( S \triangleq (s_1, \ldots, s_{n-1}) \) of length \( n - 1 \). Suppose for now that \( k < n \).

**CASE 1:** The sequence \( S \) already contains \( k \) consecutive zeros. Then both sequences \((s_1, \ldots, s_{n-1}, 0)\) and \((s_1, \ldots, s_{n-1}, 1)\) contain at least \( k \) consecutive zeros. These sequences contribute the term \( 2a_{n-1,k} \) to the recurrence.

**CASE 2:** \( S \) has at most \( k - 1 \) consecutive zeros. In order to contribute to \( a_{n,k} \), \( S \) must have a suffix of \( k - 1 \) zeros. Moreover \( S \) must be extended as \((s_1, \ldots, s_{n-1}, 0)\). Note that it must be
\( s_{n-1} = \cdots = s_{n-k+1} = 0 \) and \( s_{n-k} = 1 \). Therefore the only elements which can be varied arbitrarily are \( s_1, \ldots, s_{n-k-1} \), subject to the constraint that the sequence \( s_1, \ldots, s_{n-k-1} \) contains at most \( k - 1 \) consecutive zeros. There are precisely \( 2^{n-k-1} - a_{n-k,k} \) such sequences.

The case \( n = k \) is immediate, as there is just one sequence of length \( n \) consisting only of zeros.

From (3), dividing by \( 2^n \), we obtain

**Corollary 1**

\[
p_{n,k} = \begin{cases} 
1 & k = 0 \\
\frac{p_{n-1,k} + 2^{-(k+1)}(1 - p_{n-1-k,k})}{2^{-n}} & 1 \leq k < n \\
2^{-n} & k = n \\
0 & k > n 
\end{cases}
\]

We consider now the probability generating function for \( p_{n,k} \):

**Lemma 2 (Probability Generating Function)** Define

\[
P_k(z) \triangleq \sum_{n=0}^{\infty} p_{n,k} z^n
\]  

(5)

Then the probability generating function \( P_k(z) \) satisfies

\[
P_k(z) = \frac{1}{1 - z} + \frac{(z/2)^k - 1}{Q_k(z)}
\]

(6)

where

\[
Q_k(z) = 1 - z + \left( \frac{z}{2} \right)^{k+1}
\]

(7)

**Proof:** Observe first, that by Corollary 1, it also holds

\[
P_k(z) = \sum_{n=k}^{\infty} p_{n,k} z^n \cdot
\]  

(8)

Multiplying both sides of the recurrence in Corollary 1 and summing up for \( n \geq k + 1 \) yields:

\[
\sum_{n\geq k+1} p_{n,k} z^n = \sum_{n \geq k+1} p_{n-1,k} z^n \frac{2^{-(k+1)} \sum_{n \geq k+1} z^n - 2^{-(k+1)} \sum_{n \geq k+1} p_{n-(k+1),k} z^n}{Q_k(z)}
\]

\[
\cong \{ \text{Note that } p_{k,k} = 2^{-k}, \text{ use (5) and (8)} \}
\]

\[
P_k(z) - \left( \frac{z}{2} \right)^{k+1} = z \left( P_k(z) + \left( \frac{z}{2} \right)^{k+1} \frac{1}{1 - z} - \left( \frac{z}{2} \right)^{k+1} \right) P_k(z)
\]

\[
P_k(z) = \frac{(z/2)^{k+1}}{Q_k(z)} + \frac{(z/2)^{k+1}}{(1 - z) Q_k(z)}
\]

\[
= \frac{1}{1 - z} + \frac{(z/2)^{k+1} - 1}{Q_k(z)}
\]

where the last equality uses the identity

\[
\frac{(z/2)^{k+1}}{(1 - z) Q_k(z)} = \frac{1}{1 - z} - \frac{1}{Q_k(z)}
\]
The next step is to find an asymptotic expansion for $p_{n,k}$ by an application of the Residue-Theorem. We use the following definitions, lemmas and theorems, which are well-known in the field of complex analysis (see e.g. [Evg66]):

**Definition 1 (Residue)** Let $f(z)$ be a complex-valued function with an isolated singularity at the point $z = a$ or regular\(^1\) in $z = a$. Then the residue of $f(z)$ in $z = a$ is defined as:

$$\text{Res}(f(z), z = a) \triangleq \frac{1}{2\pi i} \oint_{|z-a| = \varepsilon} f(z) \, dz$$

where $\varepsilon$ is sufficiently small (i.e. such that $f(z)$ is regular in then interior of the circle $|z - a| < \varepsilon$ with the possible exception of a itsself\(^2\).

The next Lemma gives an important method to compute residues:

**Lemma 3 (Computation of Residues)** Let $F(z) = f(z)/g(z)$ where $f(z)$ and $g(z)$ are both regular in $z = a$. Furthermore let $z = a$ be a zero of order $n$ of $g(z)$. Then

$$\text{Res}(F(z), z = a) = \lim_{z \to a} \left( \frac{1}{(n-1)!} \frac{d^{n-1}}{dz^{n-1}} [z-a]^n F(z) \right)$$

**Proof:** See [Evg66], Theorem 4.2

Probably the most important theorem about integration over a contour in the complex plane is the following

**Theorem 4 (Residue Theorem)** Let $C$ be a contour and $f(z)$ a function regular in the interior of $C$ with the exception of points $a_1, \ldots, a_k$. Then

$$\frac{1}{2\pi i} \oint_C f(z) \, dz = \sum_{i=1}^{k} \text{Res}(f(z), z = a_i)$$

**Proof:** See [Evg66] Theorem 4.1

It is well-known that every analytic (regular in a disk around the origin) function $f(z)$ can be written as a power series. The residue theorem has an important corollary, which can be used to compute the coefficients of the taylor series.

**Corollary 2 (Cauchy's Integral Formula)** Let $f(z) = \sum_{n \geq 0} a_n \, z^n$ be an in the interior of the disk $D \triangleq \{ z : |z| \leq a \}$ analytic function. Then:

$$a_n = \frac{1}{2\pi i} \oint_{|z|=a} \frac{f(z)}{z^n} \, dz$$

\(^3\)\(^4\)

Finally, the last important theorem we use is:

\(^1\)A function $f(z)$ is regular at a point $z = a$ iff it is differentiable in $z = a$

\(^2\)The notation $\oint_{C} f(z) \, dz$ means the integration of $f(z)$ over the contour $C$
Theorem 5 (Rouché's Theorem) Let \( F(z) \) and \( f(z) \) be regular in the domain \( D \) and continuous on the boundary of \( D \). Suppose that on the boundary of \( D \) holds the inequality \(|f(z)| < |F(z)|\). Then the functions \( f(z) + F(z) \) and \( F(z) \) have the same number of zeros in \( D \). A zero with multiplicity \( k \) is counted \( k \) times.

Proof: See [Evg66], Theorem 6.2

Leaving this excursion to complex analysis, we will show how to use this theory for finding an asymptotic estimate of \( p_{n,k} \). The first important observation is:

**Lemma 4** For \( k > 1 \) the function \( Q_k(z) \) has exactly one zero point \( \gamma_k \) of order one inside \( A \triangleq \{ z : |z| < 3/2 \} \). Moreover, \( \gamma_k \) must be real.

**Proof:** Let be \( Q(z) \triangleq 1 - z \) and \( q_k(z) \triangleq \frac{(z/2)^{k+1}}{2^{k+1}} \). Clearly by (7), \( Q(z) + q_k(z) = Q_k(z) \). Obviously \( Q(z) \) has one zero of multiplicity one in \( A \). Furthermore \( |1 - z| \geq 1/2 \) for \( |z| = 3/2 \) and \( |q_k(z)| \leq 27/64 < 1/2 \) on the boundary of \( A \). Thus \( Q_k(z) = Q(z) + q_k(z) \) has by Rouché's Theorem precisely one zero of multiplicity one in \( A \). Since \( Q_k(z) \) is continuous in \( \mathcal{R} \), \( Q_k(1) > 0 \) and \( Q_k(\frac{3}{2}) < 0 \) this zero must be real.

The first step is to find an asymptotic expression for this zero. This can be done by the standard bootstrapping technique.

**Lemma 5** Let \( \gamma_k \) be the zero of \( Q_k(z) \) in \( A \). Then, for \( k > 1 \), \( \gamma_k = 1 + \varepsilon_k \), where:

\[
\varepsilon_k = \frac{1}{2^{k+1}} \left( 1 + O(k2^{-k}) \right) \tag{10}
\]

**Proof:** Observe first that

\[
Q_k(1 + 1/k) = -1/k + \left( 1 + \frac{1}{k} \right)^{k+1} / 2^{k+1} < 0
\]

if \( k > 1 \). Now, using the "bootstrapping" technique, starting with \( \varepsilon_k = O(\frac{1}{k}) \) we get:

\[
0 = Q_k(1 + \varepsilon_k) = -\varepsilon_k + 2^{-k+1} \left( 1 + O(k^{-1}) \right)^{k+1}
\]

and therefore

\[
\ln \varepsilon_k = -(k + 1) \ln 2 + \ln \left( 1 + O(k^{-1}) \right)^{k+1} = -(k + 1) \ln 2 + O(1)
\]

where the last equality is obtained by the Taylor series of \( \ln(1 + z) \). Hence:

\[
\varepsilon_k = 2^{-(k+1)} \cdot O(1)
\]

Repeating the same process with this value of \( \varepsilon_k \) yields:

\[
\ln \varepsilon_k = -(k + 1) \ln 2 + (k + 1) \ln \left( 1 + 2^{-(k+1)} \cdot O(1) \right) = -(k + 1) \ln 2 + O(k2^{-k})
\]

where the last step again uses the Taylor series of \( \ln(1 + z) \). Exponentiating both sides yields the desired claim.
The following lemmas estimate $p_{n,k}$ as a function of $\epsilon_k$.

**Lemma 6** For $k > 1$ it holds: $p_{n,k} = 1 - \frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(1 - k\epsilon_k)} + O((3/2)^{-n})$

**Proof:** Consider the function $f(z) = \frac{1 - (z/2)^k}{z^{n+1}Q_k(z)}$. As before, let $A = \{z \in C \mid |z| < \frac{3}{2}\}$. By the residue theorem:

$$\frac{1}{2\pi i} \oint_{|z|=3/2} f(z)dz = \text{Res}(f(z), z = 0) + \text{Res}(f(z), z = 1 + \epsilon_k)$$

The residue of $f(z)$ at $z = 0$ is, by Cauchy’s integral formula, the coefficient of $z^n$ in the Taylor series of $\frac{1 - (z/2)^k}{Q_k(z)} = \frac{1}{z} - P_k(z)$. Hence $\text{Res}(f(z), z = 0) = 1 - p_{n,k}$. Moreover, $1 + \epsilon_k$ is a zero of multiplicity one for $Q_k(z)$. Therefore, by Lemma 3

$$\text{Res}(f(z), z = 1 + \epsilon_k) = \lim_{z \to 1 + \epsilon_k} (z - (1 + \epsilon_k)) \frac{1 - (z/2)^k}{z^{n+1}Q_k(z)}$$

The above limit can be solved by applying de l’Hôpital’s rule and using the following two facts:

$$0 = Q_k(1 + \epsilon_k) = -\epsilon_k + \left(\frac{1 + \epsilon_k}{2}\right)^{k+1} \implies \left(\frac{1 + \epsilon_k}{2}\right)^{k+1} = \frac{2\epsilon_k}{1 + \epsilon_k} \quad (11)$$

and

$$Q_k'(1 + \epsilon_k) = -1 + \frac{k + 1}{2} \left(\frac{1 + \epsilon_k}{2}\right)^{k} = -1 + (k + 1)\frac{\epsilon_k}{1 + \epsilon_k} \quad (12)$$

Hence,

$$\text{Res}(f(z), z = 1 + \epsilon_k) = \lim_{z \to 1 + \epsilon_k} (z - (1 + \epsilon_k)) \frac{1 - (z/2)^k}{z^{n+1}Q_k(z)} \quad \{\text{de l’Hôpital’s rule}\}$$

$$= \left[\lim_{z \to 1 + \epsilon_k} \frac{1 - (z/2)^k - k/2 (z/2)^{k+1} [z - (1 + \epsilon_k)]}{(n + 1) z^n Q_k(z) + z^{n+1} Q_k'(z)} \right]$$

$$= \frac{1 - [(1 + \epsilon_k)/2]^k}{(1 + \epsilon_k)^{n+1} Q_k'(1 + \epsilon_k)}$$

$$= \frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)} \quad \{\text{Facts (11) and (12)}\}$$

Note that for $|z| = 3/2$ and $k \geq 2$, we have $|Q_k(z)| \geq |1 - z| - (3/4)^{k+1} \geq \frac{16}{15}$. Therefore,

$$\left|1 - p_{n,k} + \frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)}\right| = \left|\frac{1}{2\pi i} \oint_{A} f(z)dz\right| \leq \frac{16}{\pi} \oint_{\frac{3}{2}} \left(\frac{3}{2}\right)^{-(n+1)} dz = 32(3/2)^{-n}$$

Hence

$$p_{n,k} = 1 - \frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(1 - k\epsilon_k)} + O((3/2)^{-n})$$

\(\blacksquare\)
From this result, we obtain immediately the following estimates for \( p_{n,k} \):

**Lemma 7** For \( 1 \leq k \leq \frac{\log n}{2} \), \( p_{n,k} = 1 - O(e^{-\sqrt{n}}) \). For \( k \geq 2 \log n \), \( p_{n,k} = O(n^{-1}) \)

**Proof:** For \( k = 1 \), \( p_{n,k} = 1 - 2^{-n} \). For \( 2 \leq k \leq \frac{\log n}{2} \), \( Q_k \left( 1 + \frac{1}{2^n} \right) > 0 \) and \( Q_k \left( 1 + \frac{1}{2^k} \right) < 0 \), therefore \( \frac{1}{4\sqrt{n}} \leq \frac{1}{2^n} < \epsilon_k < \frac{1}{2k} \). From these inequalities we have, for \( 2 \leq k \leq \frac{\log n}{2} \),

\[
\frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)} \leq \frac{1}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)} \leq \frac{2}{(1 + \epsilon_k)^{n+1}} \tag{13}
\]

Observe that:

\[-(n + 1) \ln(1 + \epsilon_k) \leq -n \ln \left( 1 + \frac{1}{2\sqrt{n}} \right) = -\sqrt{n} + O(1)\]

where the last step uses the Taylor series of \( \ln(1 + x) \). Exponentiating both sides of the inequality yields:

\[(1 + \epsilon_k)^{-(n+1)} = O(\exp(-\sqrt{n}))\]

Using this result with inequality (13) yields:

\[
\frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)} = O(e^{-\sqrt{n}})
\]

Therefore \( p_{n,k} = 1 - O(e^{-\sqrt{n}}) \).

Consider now \( k \geq 2 \log n \). From relation (10) we have \( \epsilon_k = O(n^{-2}) \). Hence

\[
1 - \frac{1 - \epsilon_k}{(1 + \epsilon_k)^{n+1}(k\epsilon_k - 1)} = 1 - (1 - \epsilon_k)(1 - n\epsilon_k + O(n^2\epsilon^2))(1 + k\epsilon_k + O(n^{-2})) = O(n^{-1})
\]

We are now ready to prove theorem 3.

**Proof of Theorem 3:** Define the sets:

\[
S_1 = \{ s \in \mathcal{N} \mid 0 \leq s \leq [\log \log n] - 1 \}
\]

\[
S_2 = \{ [\log \log n], [\log \log n] \}
\]

\[
S_3 = \{ s \in \mathcal{N} \mid [\log \log n] < s \leq [\log n] \}
\]

Note that \(|S_2| = 1\) for \( \delta(n) = 0 \). Clearly,

\[
\mathbb{E}[I] = \sum_{s \in S_1 \cup S_2 \cup S_3} p_{n,2^s}
\]

By lemma 7, \( p_{n,2^s} = 1 - O(e^{-\sqrt{n}}) \) for \( s \in S_1 \) and \( p_{n,2^s} = O(n^{-1}) \) for \( s \in S_3 \). Therefore,

\[
\mathbb{E}[I] = [\log \log n] + O\left( \frac{\log n \cdot n}{n} \right) + \sum_{s \in S_2} p_{n,2^s}
\]

10
Consider the case \( \delta(n) = 0 \), that is, \( n \) is an integral double power of 2. From (10) we have
\[
\epsilon_{\log n} = \frac{1}{2n} \left( 1 + O \left( \frac{\log^2 n}{n} \right) \right).
\]
Using this fact together with
\[
\frac{1 - \epsilon_k}{(1 + \epsilon_k)(1 - k\epsilon_k)} = 1 + O(k\epsilon_k)
\]
yields
\[
\frac{1 - \epsilon_{\log n}}{(1 + \epsilon_{\log n})(1 - k\epsilon_{\log n})} = 1 + O \left( \frac{\log n}{n} \right)
\]
The proof is now ready to be finished:
\[
(1 + \epsilon_{\log n})^{-n} = \begin{cases} \text{Using (10)} & \\
\exp(-1/2) \exp(-O(\log n/n)) & \\
\{\text{Taylor series of } \exp(x)\} & \\
\exp(-1/2) - O(\log n/n) & 
\end{cases}
\]
Hence
\[
\frac{1 - \epsilon_{\log n}}{(1 + \epsilon_{\log n})^{n+1} \epsilon_{\log n} \log n - 1} = e^{-1/2} - O \left( \frac{\log^2 n}{n} \right)
\]
Therefore \( p_{n, \log n} = 1 - e^{-1/2} + O \left( \frac{\log^2 n}{n} \right) \).

Let now \( \delta(n) > 0 \). Then \( S_2 = \{ s_1, s_2 \} \), with \( s_1 = \log \log n - \delta(n) \) and \( s_2 = s_1 + 1 \). Let \( k_i \triangleq 2^{s_i} \)
for \( i = 1, 2 \). It is \( 2^{2^{s_1}} = n^{2^{-\delta(n)}} \) and therefor by (10):
\[
\epsilon_{k_1} = \frac{1}{2 n^{2^{-\delta(n)}}} + O \left( \frac{\log n}{n^{2^{-\delta(n)}}} \right) = \frac{1}{2 n^{2^{-\delta(n)}}} + O(\log n/n)
\]
Hence:
\[
\frac{1 - \epsilon_{k_1}}{(1 + \epsilon_{k_1})(1 - k\epsilon_{k_1})} = 1 + O \left( \frac{\log n}{n^{2^{-\delta(n)}}} \right)
\]
\[
= \begin{cases} \{2^{-\delta(n)} \geq 1/2\} & \\
1 + O \left( \frac{\log n}{\sqrt{n}} \right) & 
\end{cases}
\]
These two facts yield:
\[
(1 + \epsilon_{k_1}) = \exp \left( -n \left[ \frac{1}{2} n^{2^{-\delta(n)}} + O \left( \frac{\log n}{n^{21-\delta(n)}} \right) \right] \right)
\]
\[
= \exp(-1/2 \ n^{1-2^{-\delta(n)}}) \exp(O(\log n \ n^{1-2^{1-\delta(n)}}))
\]
\[
= \begin{cases} \{2^{-\delta(n)} \geq 1/2\} & \\
O(\exp(-\sqrt{n})) & 
\end{cases}
\]
Hence: \( p_{n, 2^{s_1}} = 1 - O(e^{-\sqrt{n}}) \).
It follows from the definition of $s_2$ that $2^{s_2} = n^{2^{1-\varepsilon(n)}}$. Therefore $\varepsilon(n) \overset{\Delta}{=} 2^{1-\varepsilon(n)} > 1$. Hence:

$$
\epsilon_k = \frac{1}{2n\varepsilon(n)} + O\left(\frac{\log n}{n^{2\varepsilon(n)}}\right)
$$

Furthermore:

$$
\frac{1-\epsilon_k}{1-k_2\epsilon_k} = 1 + O(k_2\epsilon_k) = 1 + O\left(\frac{\log n}{n}\right)
$$

Together with

$$(1+\epsilon_k)^{-(n+1)} = 1 - O(n\epsilon_k) = 1 - O(n^{1-\varepsilon(n)})$$

we obtain $p_{n,2^{s_2}} = O(n^{-\gamma(n)})$, where $\gamma(n) \overset{\Delta}{=} 2^{1-\varepsilon(n)} - 1$. The thesis follows.

\[\square\]

4 Conclusions

We have analyzed the average case complexity of a parallel algorithm that solves the Sublist Selection problem on the PRAM model of computation. Let $p$ be the probability that a node is labeled False in the input list. The upper and lower bounds given in section 2 show that, for $p \leq 1 - \frac{1}{\log \log n}$, the average running time of the algorithm is $\Theta(\log \log n)$, which is exponentially smaller than the worst case running time. The average complexity becomes $\Theta(\log n)$ only for values of $p$ very close to 1, that is $p \geq 1 - \frac{1}{n^c}$, for some $c > 0$.

The generating function technique used in section 3 is a powerful tool, frequently employed for the average case analysis of sequential algorithms. Such a technique seems also to be a promising tool for analyzing parallel algorithms. Many parallel algorithms can be found in the literature which seem to exhibit a better behavior in the average than the worst case. Our future work will target the analysis of the average case of such algorithms, by making use of similar techniques to those employed in this paper.

Acknowledgements The authors are grateful to Peter Buergisser, Philippe Flajolet, Richard Karp and Thomas Lickteig for fruitful discussions.

References


