CMSC 341

Skip Lists

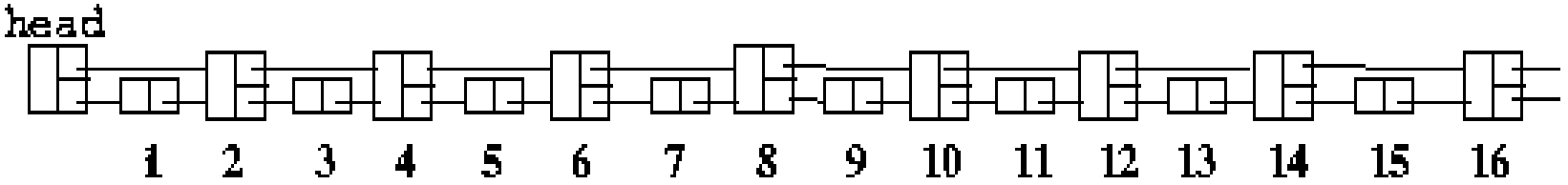
Looking Back at Sorted Lists

- **Sorted Linked List**
  What is the worst case performance of find( ), insert( )?

- **Sorted Array**
  What is the worst case performance of find( ), insert( )?

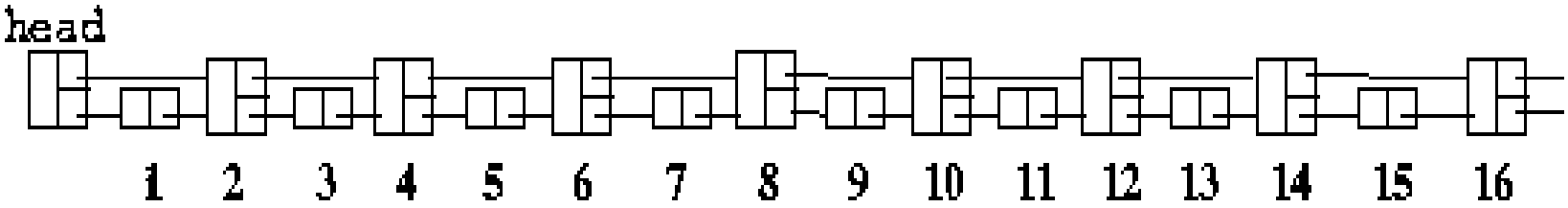
An Alternative Sorted Linked List

- What if you skip every other node?
  - Every other node has a pointer to the next and the one after that



- Find :
  - follow “skip” pointer until target < this.skip.element
- Resources
  - Additional storage
- Performance of find( )?

Skipping Every 2nd Node

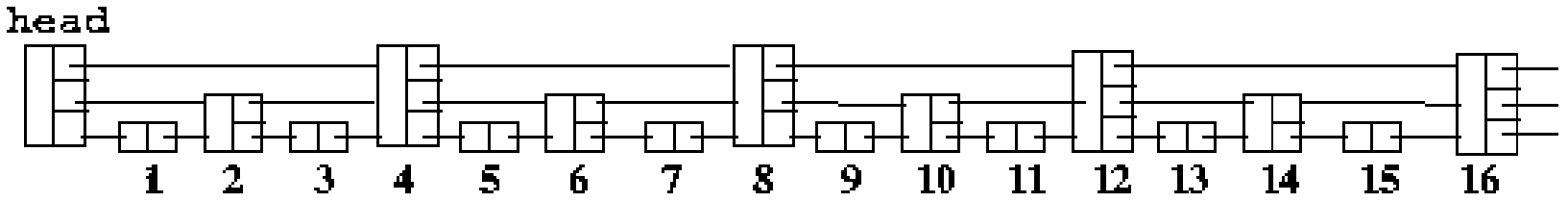


The value stored in each node is shown below the node and corresponds to the the position of the node in the list.

It's clear that find( ) does not need to examine every node. It can skip over every other node, then do a final examination at the end. The number of nodes examined is no more than $\lceil n/2 \rceil + 1$.

For example the nodes examined finding the value 15 would be
2, 4, 6, 8, 10, 12, 14, 16, 15 -- a total of $\lceil 16/2 \rceil + 1 = 9$.

Skipping Every 2nd and 4th Node



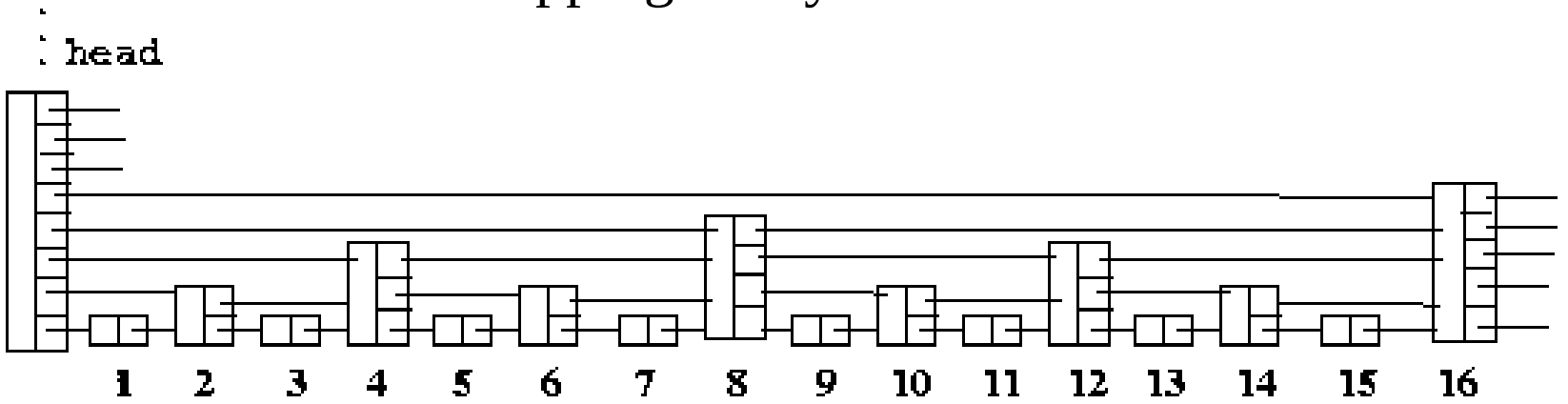
The find operation can now make bigger skips than the previous example. Every 4th node is skipped until the search is confined between two nodes of size 3. At this point as many as three nodes may need to be scanned. It's also possible that some nodes may be examined more than once. The number of nodes examined is no more than $\lceil n / 4 \rceil + 3$.

Again, look at the nodes examined when searching for 15.

New and Improved Alternative

- Add hierarchy of skip pointers
  - every 2^i-th node points 2^i nodes ahead
  - For example, every 2^{nd} node has a reference 2 nodes ahead; every 8^{th} node has a reference 8 nodes ahead

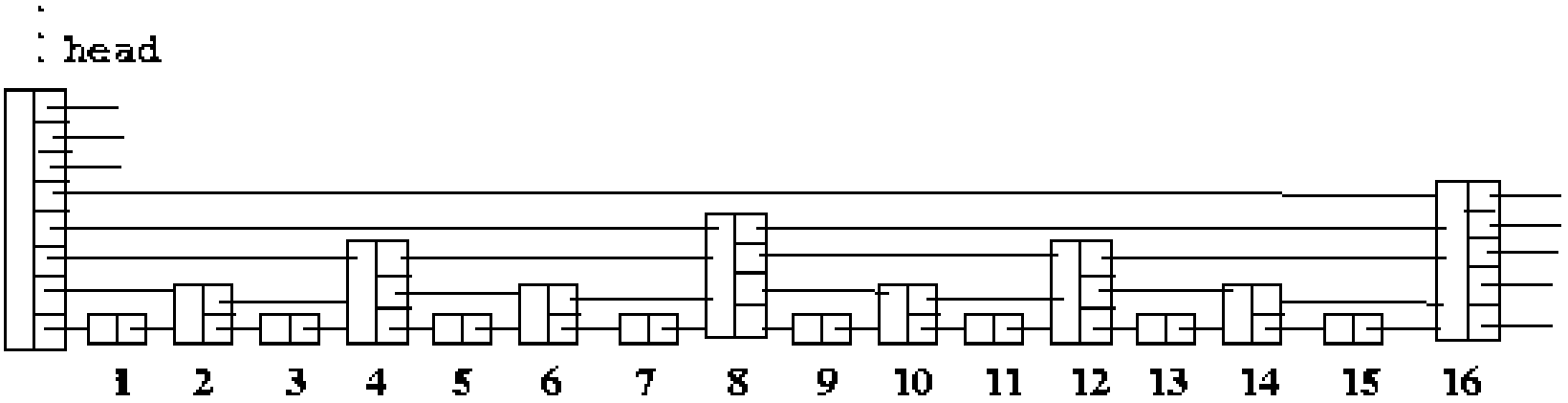
Skipping Every 2^i-th node



Suppose this list contained 32 nodes and we want to search for some value in it. Working down from the top, we first look at node 16 and have cut the search in half. When we look again one level down in either the right or left half, we have cut the search in half again. We continue in this manner until we find the node being sought (or not).

This is just like binary search in an array. Intuitively we can understand why the max number of nodes examined is O(lg N).

Some Serious Problems

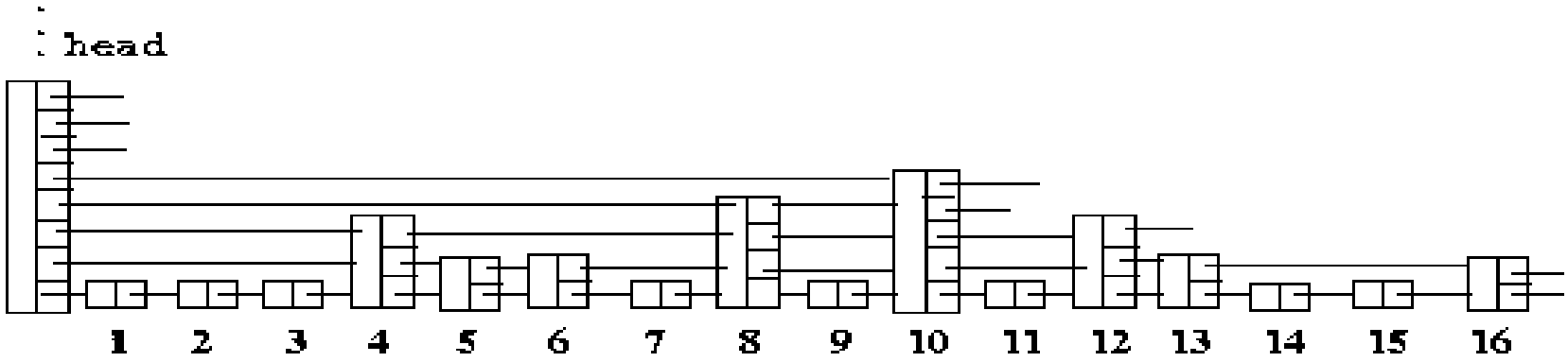


- This structure looks pretty good, but what happens when we insert or remove a value from the list? Reorganizing the the list is O(N).
- For example, suppose the first element of the list was removed. Since it's necessary to maintain the strict pattern of node sizes, it's easiest to move all the values toward the head and remove the end node. A similar situation occurs when a new node is added.

Skip Lists

- Concept:
  A skip list maintains the same **distribution** of nodes, but without the requirement for the rigid pattern of node sizes
  - 1/2 have 1 pointer
  - 1/4 have 2 pointers
  - 1/8 have 3 pointers
  - …
  - $1/2^i$ have i pointers
- It's no longer necessary to maintain the rigid pattern by moving values around for insert and remove. This gives us a ***high probability*** of still having O(lg N) performance. The probability that a skip list will behave badly is very small.

A Probabilistic Skip List



The number of forward reference pointers a node has is its “size”.

The distribution of node sizes is exactly the same as the previous figure, the nodes just occur in a different pattern.

Inserting a Node

- When inserting a new node, we choose the size of the node probabilistically.
- Every skip list has an associated (and fixed) probability, ***p***, that determines the distribution of node sizes. A fraction, ***p***, of the nodes that have at least ***r*** forward references also have ***r + 1*** forward references.

Skip List Insert

- To insert node:
  - Create new node with random size.
  - For each pointer, i , connect to next node with at least i pointers.

```
int generateNodeSize(double p, int maxSize)
{
   int size = 1;
   while (drand48() < p) size++;
   return (size >maxSize) ? maxSize : size;
}
```

An Aside on Node Distribution

- Given an infinitely long skip list with associated probability ***p***, it can be shown that **1 – p** nodes will have just one forward reference.
- This means that **p(1 – p)** nodes will have exactly two forward references and in general $\mathbf{p^k(1 - p)}$ nodes will have **k + 1** forward reference pointers.
- For example, with p = 0.5
  - 0.5 (1/2 of the nodes will have exactly one forward reference)
  - 0.5 (1 – 0.5) = 0.25 (1/4 of the nodes will have 2 references)
  - 0.5^2 (1 – 0.5) = 0.125 (1/8 of the nodes will have 3 references)
  - 0.5^3 (1 – 0.5) = 0.0625 (1/16 of the nodes will have 4 references)
- Work out the distribution for p = 0.25 (1/4) for yourself.

Determining the Size of the Header Node

- The size of the header node (the number of forward references it has) is the maximum size of any node in the skip list and is chosen when the empty skip list is constructed (i.e. it must be predetermined)

- Dr. Pugh has shown that the maximum size should be chosen as $\log_{1/p} N$. For p = ½, the maximum size for a skip list with 65,536 elements should be no smaller than $\log_2 65536 = 16$.

Performance Considerations

- The ***expected*** time to find an element (and therefore to insert or remove) is O(lg N).
- It is possible for the time to be substantially longer if the configuration of nodes is unfavorable for a particular operation.
- Since the node sizes are chosen randomly, it is possible to get a “bad” run of sizes.
- For example, it is possible that each node will be generated with the same size, producing the equivalent of an ordinary linked list.

Performance Considerations

- A “bad” run of sizes will be less important in a long skip list than in a short one.
- The probability of poor performance decreases rapidly as the number of nodes increases.

More performance

- The probability that an operation takes longer than expected is function of the associated probability ***p***. Dr. Pugh calculated that with p = 0.5 and 4096 elements, the probability that the actual time will exceed the expected time by more than a factor of 3 is less than one in 200 million.

- The relative time and space performance depends on p. Dr. Pugh suggests p = 0.25 for most cases. If the predictability of performance is important, then he suggests using p = 0.5 (the variability of the performance decreases with larger p).

- Interestingly, the average number of references per node is only 1.33 when p = 0.25 is used. A BST has 2 references per node, so a skip list is more space-efficient.

Skip List Implementation

```
public class
SkipList <Anytype extends Comparable<? super AnyType>>{
   private static class SkipListNode <AnyType>{
        void setDatum(AnyType datum){ }
        void setForward(int i, SkipListNode f){ }
        void setSize(int size){ }
        SkipListNode(){ }
        SkipListNode(AnyType datum, int size){ }
        SkipListNode(SkipListNode c){ }
        AnyType getDatum(){ }
        int getSize(){ }
        SkipListNode getForward(int level){ }

        private int m_size;
        private Vector <SkipListNode> m_forward;
        private Vector <AnyType> m_datum;

   }
```

Skip List Implementation (cont.)

```
    SkipList(){}
    SkipList(int max_node_size, double probab){}
    SkipList(SkipList<AnyType> ref) {}

    int getHighNodeSize(){}
    int getMaxNodeSize(){}
    double getProbability(){}
    void insert( AnyType item){}
    boolean find( AnyType item){}
    void remove( AnyType item){}

    private SkipListNode find(AnyType item, SkipListNode <AnyType>
    start){}
    private SkipListNode getHeader(){}
    private SkipListNode findInsertPoint( AnyType item, int nodesize){}
    private boolean insert( AnyType item, int nodesize){}

    private int m_high_node_size;
    private int m_max_node_size;
    private double m_prob;
    SkipListNode<AnyType> m_head;
}
```

find

```
boolean find(Comparable x)
{
   node = header node
   for(reference level of node from (nodesize-1) down to 0)
        while (the node referred to is less than x)
                node = node referred to
   if (node referred to has value x)
        return true
   else
        return false
}
```

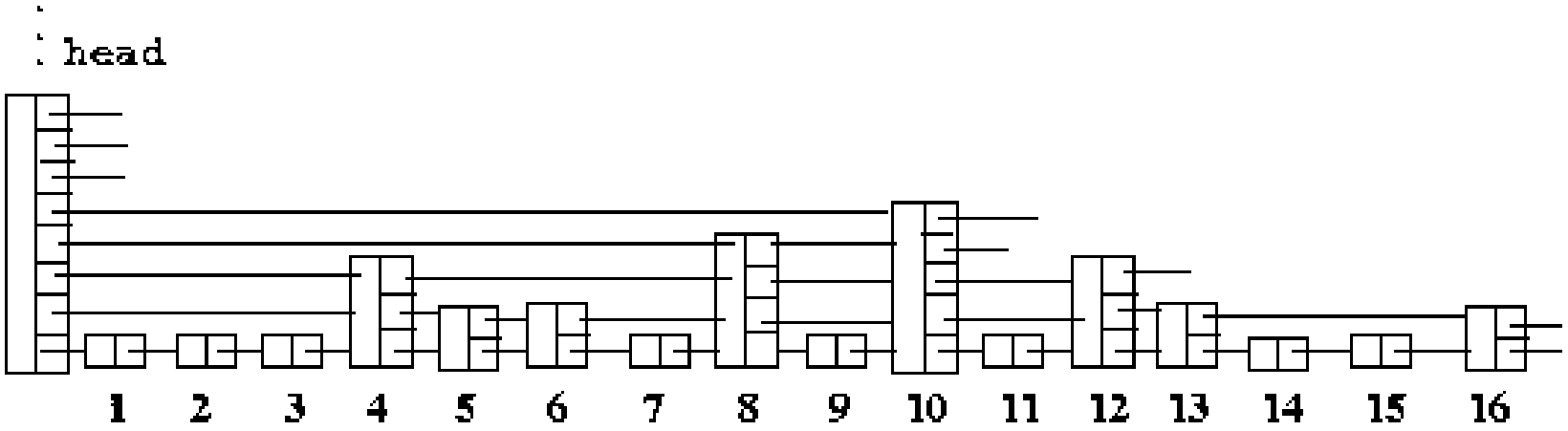
findInsertPoint

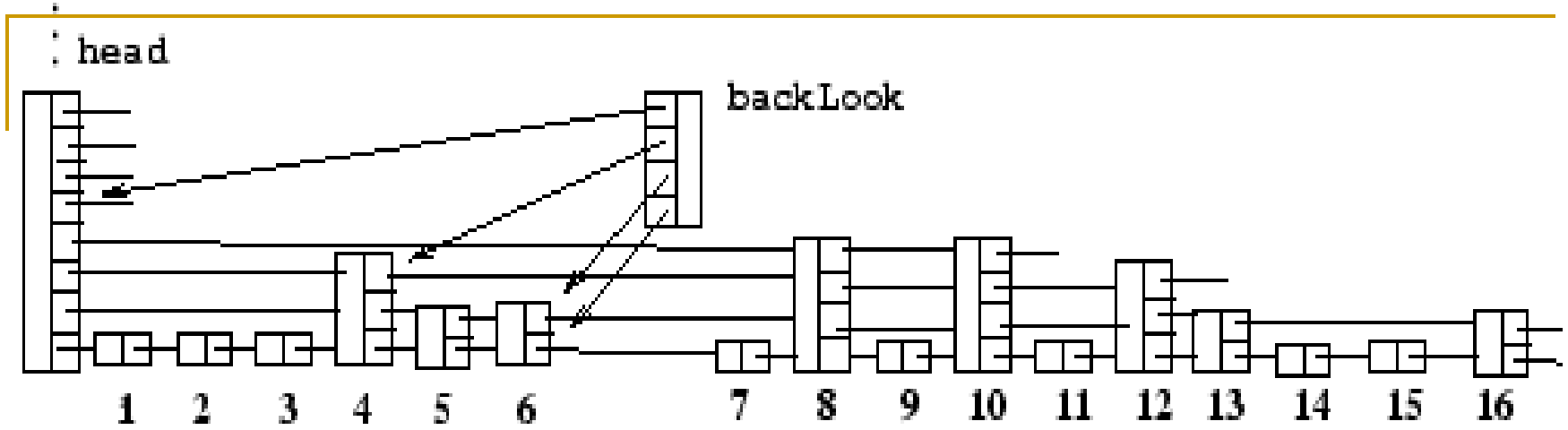
- Ordinary list insertion:

  Have handle (iterator) to node to insert in front of
- Skip list insertion:

  Need handle to all nodes that skip to node of given size at insertion point (all “see-able” nodes).
- Use `backLook` structure with a pointer for each level of node to be inserted

Insert 6.5





In the figure, the insertion point is between nodes 6 and 7. "Looking" back towards the header, the nodes you can "see" at the various levels are

level	node seen
0	6
1	6
2	4
3	header

We construct a "backLook" node that has its forward pointers set to the relevant "see-able" nodes. This is the type of node returned by the findInsertPoint method

insert Method

- Once we have the backLook node returned by findInsertPoint and have constructed the new node to be inserted, the insertion is easy.
- The public insert(AnyType x) decides on the new nodes size by random choice, then calls the overloaded private insert(AnyType x, int nodeSize) to do the work.
- Code in C is available in Dr. Anastasio's HTML version of these notes.