CSE 431/531: Algorithm Analysis and Design (Fall 2021)

Greedy Algorithms

Lecturer: Shi Li

Department of Computer Science and Engineering
University at Buffalo
Def. In an optimization problem, our goal is to find a valid solution with the minimum cost (or maximum value).
Def. In an optimization problem, our goal of is to find a valid solution with the minimum cost (or maximum value).

Trivial Algorithm for an Optimization Problem
Enumerate all valid solutions, compare them and output the best one.
Def. In an optimization problem, our goal is to find a valid solution with the minimum cost (or maximum value).

**Trivial Algorithm for an Optimization Problem**
Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
Def. In an **optimization problem**, our goal is to find a valid solution with the minimum cost (or maximum value).

**Trivial Algorithm for an Optimization Problem**

Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in *exponential* time, as the number of potential solutions is often exponentially large.
Def. In an **optimization problem**, our goal of is to find a valid solution with the minimum cost (or maximum value).

**Trivial Algorithm for an Optimization Problem**
Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
- \( f(n) \) is a polynomial if \( f(n) = O(n^k) \) for some constant \( k > 0 \).
Def. In an optimization problem, our goal is to find a valid solution with the minimum cost (or maximum value).

Trivial Algorithm for an Optimization Problem

Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
- $f(n)$ is a polynomial if $f(n) = O(n^k)$ for some constant $k > 0$.
- Convention: polynomial time = efficient
Def. In an optimization problem, our goal is to find a valid solution with the minimum cost (or maximum value).

Trivial Algorithm for an Optimization Problem
Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
- \( f(n) \) is a polynomial if \( f(n) = O(n^k) \) for some constant \( k > 0 \).
- convention: polynomial time = efficient

Goals of algorithm design
Def. In an optimization problem, our goal of is to find a valid solution with the minimum cost (or maximum value).

Trivial Algorithm for an Optimization Problem
Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
- $f(n)$ is a polynomial if $f(n) = O(n^k)$ for some constant $k > 0$.
- convention: polynomial time = efficient

Goals of algorithm design
1. Design efficient algorithms to solve problems
Def. In an optimization problem, our goal of is to find a valid solution with the minimum cost (or maximum value).

**Trivial Algorithm for an Optimization Problem**
Enumerate all valid solutions, compare them and output the best one.

- However, trivial algorithm often runs in exponential time, as the number of potential solutions is often exponentially large.
- $f(n)$ is a polynomial if $f(n) = O(n^k)$ for some constant $k > 0$.
- convention: polynomial time = efficient

**Goals of algorithm design**
1. Design efficient algorithms to solve problems
2. Design more efficient algorithms to solve problems
Common Paradigms for Algorithm Design

- Greedy Algorithms
- Divide and Conquer
- Dynamic Programming
Common Paradigms for Algorithm Design

- Greedy Algorithms
- Divide and Conquer
- Dynamic Programming

- Greedy algorithms are often for optimization problems.
Common Paradigms for Algorithm Design

- Greedy Algorithms
- Divide and Conquer
- Dynamic Programming

Greedy algorithms are often for optimization problems. They often run in polynomial time due to their simplicity.
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an **irrevocable** decision using a “reasonable” strategy
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a "reasonable" strategy

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is "safe"
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an irreversible decision using a “reasonable” strategy

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a “reasonable” strategy

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” *(key)*
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem *(usually easy)*

**Def.** A strategy is safe: there is always an optimum solution that agrees with the decision made according to the strategy.
Outline

1 Toy Example: Box Packing
2 Interval Scheduling
3 Offline Caching
   • Heap: Concrete Data Structure for Priority Queue
4 Data Compression and Huffman Code
5 Summary
Box Packing

**Input:** \( n \) boxes of capacities \( c_1, c_2, \ldots, c_n \)
\( m \) items of sizes \( s_1, s_2, \ldots, s_m \)
Can put at most 1 item in a box
Item \( j \) can be put into box \( i \) if \( s_j \leq c_i \)

**Output:** A way to put as many items as possible in the boxes.
Box Packing

**Input:** $n$ boxes of capacities $c_1, c_2, \cdots, c_n$

$m$ items of sizes $s_1, s_2, \cdots, s_m$

Can put at most 1 item in a box

Item $j$ can be put into box $i$ if $s_j \leq c_i$

**Output:** A way to put as many items as possible in the boxes.

**Example:**

- Box capacities: 60, 40, 25, 15, 12
- Item sizes: 45, 42, 20, 19, 16
- Can put 3 items in boxes: 45 $\rightarrow$ 60, 20 $\rightarrow$ 40, 19 $\rightarrow$ 25
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an irreversible decision using a “reasonable” strategy
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an **irrevocable** decision using a “reasonable” strategy

Designing a Reasonable Strategy for Box Packing

- Q: Take box 1. Which item should we put in box 1?
**Greedy Algorithm**

- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a “reasonable” strategy

**Designing a Reasonable Strategy for Box Packing**

- Q: Take box 1. Which item should we put in box 1?
- A: The item of the largest size that can be put into the box.
Analysis of Greedy Algorithm

Prove that the reasonable strategy is “safe”

Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem

**Lemma** The strategy that put into box 1 the largest item it can hold is “safe”: There is an optimum solution in which box 1 contains the largest item it can hold.
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem

Lemma The strategy that put into box 1 the largest item it can hold is “safe”: There is an optimum solution in which box 1 contains the largest item it can hold.

- Intuition: putting the item gives us the easiest residual problem.
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem

Lemma The strategy that put into box 1 the largest item it can hold is “safe”: There is an optimum solution in which box 1 contains the largest item it can hold.

- Intuition: putting the item gives us the easiest residual problem.
- formal proof via exchanging argument:
Lemma There is an optimum solution in which box 1 contains the largest item it can hold.
**Lemma**  There is an optimum solution in which box 1 contains the largest item it can hold.

**Proof.**

- Let $j =$ largest item that box 1 can hold.
**Lemma** There is an optimum solution in which box 1 contains the largest item it can hold.

**Proof.**
- Let $j =$ largest item that box 1 can hold.
- Take any optimum solution $S$. If $j$ is put into Box 1 in $S$, done.
Lemma  There is an optimum solution in which box 1 contains the largest item it can hold.

Proof.

- Let $j =$ largest item that box 1 can hold.
- Take any optimum solution $S$. If $j$ is put into Box 1 in $S$, done.
- Otherwise, assume this is what happens in $S$:
**Lemma**  There is an optimum solution in which box 1 contains the largest item it can hold.

**Proof.**

- Let $j = \text{largest item that box 1 can hold}.$
- Take any optimum solution $S.$ If $j$ is put into Box 1 in $S,$ done.
- Otherwise, assume this is what happens in $S':$

  $S':$

  - box 1
  - item $j' \quad \text{item } j$

  $s_{j'} \leq s_j,$ and swapping gives another solution $S'$
**Lemma** There is an optimum solution in which box 1 contains the largest item it can hold.

**Proof.**
- Let $j = \text{largest item that box 1 can hold}$.
- Take any optimum solution $S$. If $j$ is put into Box 1 in $S$, done.
- Otherwise, assume this is what happens in $S$:

  $S'$:
  
  \[
  \begin{array}{cccc}
  \text{box 1} & \text{item } j' & \text{item } j & \ldots & \text{item} \\
  \text{item } j' & \text{item } j & \ldots & \text{item} \\
  \end{array}
  \]

  \[s_{j'} \leq s_j, \text{ and swapping gives another solution } S'\]
- $S'$ is also an optimum solution. In $S'$, $j$ is put into Box 1.
Notice that the exchanging operation is only for the sake of analysis; it is not a part of the algorithm.
Notice that the exchanging operation is only for the sake of analysis; it is not a part of the algorithm.

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem
Notice that the exchanging operation is only for the sake of analysis; it is not a part of the algorithm.

**Analysis of Greedy Algorithm**

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem

- Trivial: we decided to put Item $j$ into Box 1, and the remaining instance is obtained by removing Item $j$ and Box 1.
Generic Greedy Algorithm

1. **while** the instance is non-trivial **do**
2. make the choice using the greedy strategy
3. reduce the instance

Greedy Algorithm for Box Packing

1. \( T \leftarrow \{1, 2, 3, \ldots, m\} \)
2. **for** \( i \leftarrow 1 \) **to** \( n \) **do**
3. \( \text{if} \) some item in \( T \) can be put into box \( i \) **then**
4. \( j \leftarrow \text{the largest item in } T \text{ that can be put into box } i \)
5. \( \text{print(“put item } j \text{ in box } i”) \)
6. \( T \leftarrow T \setminus \{j\} \)
Generic Greedy Algorithm

1: while the instance is non-trivial do
2: make the choice using the greedy strategy
3: reduce the instance

Lemma  Generic algorithm is correct if and only if the greedy strategy is safe.
Generic Greedy Algorithm

1: while the instance is non-trivial do
2: make the choice using the greedy strategy
3: reduce the instance

Lemma  Generic algorithm is correct if and only if the greedy strategy is safe.

- Greedy strategy is safe: we will not miss the optimum solution
Generic Greedy Algorithm

1: while the instance is non-trivial do
2: make the choice using the greedy strategy
3: reduce the instance

Lemma  Generic algorithm is correct if and only if the greedy strategy is safe.

- Greedy strategy is safe: we will not miss the optimum solution
- Greedy strategy is not safe: we will miss the optimum solution for some instance, since the choices we made are irrevocable.
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an irrevocable decision using a “reasonable” strategy
Greedy Algorithm

- Build up the solutions in steps
- At each step, make an **irrevocable** decision using a “reasonable” strategy

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem
**Greedy Algorithm**
- Build up the solutions in steps
- At each step, make an **irrevocable** decision using a “reasonable” strategy

**Analysis of Greedy Algorithm**
- Prove that the reasonable strategy is “safe”
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem

**Def.** A strategy is “safe” if there is always an optimum solution that is “consistent” with the decision made according to the strategy.
Exchange argument: Proof of Safety of a Strategy

- let $S$ be an arbitrary optimum solution.
- if $S$ is consistent with the greedy choice, done.
- otherwise, show that it can be modified to another optimum solution $S'$ that is consistent with the choice.
Exchange argument: Proof of Safety of a Strategy

- let $S$ be an arbitrary optimum solution.
- if $S$ is consistent with the greedy choice, done.
- otherwise, show that it can be modified to another optimum solution $S'$ that is consistent with the choice.

The procedure is not a part of the algorithm.
Outline

1. Toy Example: Box Packing
2. Interval Scheduling
3. Offline Caching
   - Heap: Concrete Data Structure for Priority Queue
4. Data Compression and Huffman Code
5. Summary
Interval Scheduling

**Input:** \( n \) jobs, job \( i \) with start time \( s_i \) and finish time \( f_i \)

\( i \) and \( j \) are compatible if \( [s_i, f_i) \) and \( [s_j, f_j) \) are disjoint

**Output:** A maximum-size subset of mutually compatible jobs
Interval Scheduling

**Input:** $n$ jobs, job $i$ with start time $s_i$ and finish time $f_i$

$i$ and $j$ are compatible if $[s_i, f_i)$ and $[s_j, f_j)$ are disjoint

**Output:** A maximum-size subset of mutually compatible jobs
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
  - Schedule the job with the smallest size?
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
Greedy Algorithm for Interval Scheduling

Which of the following strategies are safe?

Schedule the job with the smallest size? No!
Greedy Algorithm for Interval Scheduling

Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs?
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!

![Diagram of intervals]

0 1 2 3 4 5 6 7 8 9
Greedy Algorithm for Interval Scheduling

• Which of the following strategies are safe?
• Schedule the job with the smallest size? No!
• Schedule the job conflicting with smallest number of other jobs? No!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
- Schedule the job with the earliest finish time?
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
- Schedule the job with the earliest finish time? Yes!
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
- Schedule the job with the earliest finish time? Yes!

![Schedule intervals diagram](image-url)
Greedy Algorithm for Interval Scheduling

- Which of the following strategies are safe?
- Schedule the job with the smallest size? No!
- Schedule the job conflicting with smallest number of other jobs? No!
- Schedule the job with the earliest finish time? Yes!
**Lemma**  It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

**Proof.**
**Greedy Algorithm for Interval Scheduling**

**Lemma** It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

**Proof.**
- Take an arbitrary optimum solution $S'$

$S$:  

```
[ ] [ ] [ ] [ ] [ ]
```
Lemma It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

Proof.
- Take an arbitrary optimum solution $S'$
- If it contains $j$, done
Greedy Algorithm for Interval Scheduling

**Lemma**  It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

**Proof.**
- Take an arbitrary optimum solution $S'$
- If it contains $j$, done
Lemma  It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

Proof.

- Take an arbitrary optimum solution $S'$
- If it contains $j$, done
- Otherwise, replace the first job in $S$ with $j$ to obtain another optimum schedule $S'$.
Greedy Algorithm for Interval Scheduling

**Lemma** It is safe to schedule the job \( j \) with the earliest finish time: There is an optimum solution where the job \( j \) with the earliest finish time is scheduled.

**Proof.**
- Take an arbitrary optimum solution \( S \)
- If it contains \( j \), done
- Otherwise, replace the first job in \( S \) with \( j \) to obtain another optimum schedule \( S' \).
Greedy Algorithm for Interval Scheduling

**Lemma** It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

- What is the remaining task after we decided to schedule $j$?
- Is it another instance of interval scheduling problem?
**Lemma** It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

- What is the remaining task after we decided to schedule $j$?
- Is it another instance of interval scheduling problem? Yes!
**Lemma** It is safe to schedule the job $j$ with the earliest finish time: There is an optimum solution where the job $j$ with the earliest finish time is scheduled.

- What is the remaining task after we decided to schedule $j$?
- Is it another instance of interval scheduling problem? Yes!
Greedy Algorithm for Interval Scheduling

**Lemma** It is safe to schedule the job \( j \) with the earliest finish time: There is an optimum solution where the job \( j \) with the earliest finish time is scheduled.

- What is the remaining task after we decided to schedule \( j \)?
- Is it another instance of interval scheduling problem? Yes!
Greedy Algorithm for Interval Scheduling

Schedule\((s, f, n)\)

1: \( A \leftarrow \{1, 2, \ldots, n\}, S \leftarrow \emptyset \)
2: while \( A \neq \emptyset \) do
3: \( j \leftarrow \arg\min_{j' \in A} f_{j'} \)
4: \( S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\} \)
5: return \( S \)
Greedy Algorithm for Interval Scheduling

Schedule\((s, f, n)\)

1: \(A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset\)
2: \textbf{while} \(A \neq \emptyset\) \textbf{do}
3: \(j \leftarrow \arg \min_{j' \in A} f_{j'}\)
4: \(S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\}\)
5: \textbf{return} \(S\)
Greedy Algorithm for Interval Scheduling

Schedule\((s, f, n)\)

1: \(A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset\)
2: \textbf{while} \(A \neq \emptyset\) \textbf{do}
3: \(j \leftarrow \arg \min_{j' \in A} f_{j'}\)
4: \(S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\}\)
5: \textbf{return} \(S\)
Greedy Algorithm for Interval Scheduling

\textbf{Schedule}(s, f, n)

1: \( A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset \)
2: \textbf{while} \( A \neq \emptyset \) \textbf{do}
3: \( j \leftarrow \arg \min_{j' \in A} f_{j'} \)
4: \( S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\} \)
5: \textbf{return} \( S \)
Greedy Algorithm for Interval Scheduling

Schedule($s, f, n$)

1: $A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset$
2: while $A \neq \emptyset$ do
3: $j \leftarrow \text{arg min}_{j' \in A} f_{j'}$
4: $S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\}$
5: return $S$
Greedy Algorithm for Interval Scheduling

\[ \text{Schedule}(s, f, n) \]

1. \( A \leftarrow \{1, 2, \ldots, n\}, S \leftarrow \emptyset \)
2. \textbf{while} \( A \neq \emptyset \) \textbf{do}
3. \( j \leftarrow \arg \min_{j' \in A} f_{j'} \)
4. \( S \leftarrow S \cup \{j\}; \ A \leftarrow \{j' \in A : s_{j'} \geq f_j\} \)
5. \textbf{return} \( S \)
Greedy Algorithm for Interval Scheduling

Schedule\((s, f, n)\)

1: \( A \leftarrow \{1, 2, \ldots, n\}, S \leftarrow \emptyset \)
2: \textbf{while} \( A \neq \emptyset \) \textbf{do}
3: \( j \leftarrow \arg\min_{j' \in A} f_{j'} \)
4: \( S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\} \)
5: \textbf{return} \( S \)

Running time of algorithm?
Greedy Algorithm for Interval Scheduling

Schedule($s, f, n$)

1: $A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset$
2: while $A \neq \emptyset$ do
3: \hspace{1em} $j \leftarrow \arg \min_{j' \in A} f_{j'}$
4: \hspace{1em} $S \leftarrow S \cup \{j\}; A \leftarrow \{j' \in A : s_{j'} \geq f_j\}$
5: return $S$

Running time of algorithm?

- Naive implementation: $O(n^2)$ time
Greedy Algorithm for Interval Scheduling

\textbf{Schedule}(s, f, n)

1: \( A \leftarrow \{1, 2, \cdots, n\}, S \leftarrow \emptyset \)
2: \textbf{while} \( A \neq \emptyset \) \textbf{do}
3: \( j \leftarrow \text{arg} \min_{j' \in A} f_{j'} \)
4: \( S \leftarrow S \cup \{j\}; \ A \leftarrow \{j' \in A : s_{j'} \geq f_j\} \)
5: \textbf{return} \( S \)

Running time of algorithm?

- Naive implementation: \( O(n^2) \) time
- Clever implementation: \( O(n \lg n) \) time
Clever Implementation of Greedy Algorithm

Schedule($s, f, n$)

1: sort jobs according to $f$ values
2: $t \leftarrow 0, S \leftarrow \emptyset$
3: for every $j \in [n]$ according to non-decreasing order of $f_j$ do
4: \hspace{1em} if $s_j \geq t$ then
5: \hspace{2em} $S \leftarrow S \cup \{j\}$
6: \hspace{1em} $t \leftarrow f_j$
7: return $S$
Clever Implementation of Greedy Algorithm

**Schedule**(*s*, *f*, *n*)

1. sort jobs according to *f* values
2. \( t \leftarrow 0, \; S \leftarrow \emptyset \)
3. for every \( j \in [n] \) according to non-decreasing order of \( f_j \) do
4. \hspace{1em} if \( s_j \geq t \) then
5. \hspace{2em} \( S \leftarrow S \cup \{j\} \)
6. \hspace{2em} \( t \leftarrow f_j \)
7. return \( S \)
Clever Implementation of Greedy Algorithm

**Schedule**(*s, f, n*)

1: sort jobs according to *f* values
2: \( t \leftarrow 0, S \leftarrow \emptyset \)
3: for every \( j \in [n] \) according to non-decreasing order of \( f_j \) do
   4: if \( s_j \geq t \) then
   5: \( S \leftarrow S \cup \{j\} \)
   6: \( t \leftarrow f_j \)
7: return \( S \)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
   4: \( \text{if } s_j \geq t \text{ then}\)
   5: \(S \leftarrow S \cup \{j\}\)
   6: \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4: \hspace{1em} if \(s_j \geq t\) then
5: \hspace{2em} \(S \leftarrow S \cup \{j\}\)
6: \hspace{2em} \(t \leftarrow f_j\)
7: return \(S\)
**Clever Implementation of Greedy Algorithm**

**Schedule**\((s, f, n)\)

1. sort jobs according to \(f\) values
2. \(t \leftarrow 0, S \leftarrow \emptyset\)
3. for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4. \hspace{1em} if \(s_j \geq t\) then
5. \hspace{2em} \(S \leftarrow S \cup \{j\}\)
6. \hspace{2em} \(t \leftarrow f_j\)
7. return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)
1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4: \(\text{if } s_j \geq t \text{ then}\)
5: \(S \leftarrow S \cup \{j\}\)
6: \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4: if \(s_j \geq t\) then
5: \(S \leftarrow S \cup \{j\}\)
6: \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4: \(\text{if } s_j \geq t \text{ then}\)
5: \(S \leftarrow S \cup \{j\}\)
6: \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1. sort jobs according to \(f\) values
2. \(t \leftarrow 0, S \leftarrow \emptyset\)
3. for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
   4. if \(s_j \geq t\) then
   5. \(S \leftarrow S \cup \{j\}\)
   6. \(t \leftarrow f_j\)
7. return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
   4: if \(s_j \geq t\) then
   5: \(S \leftarrow S \cup \{j\}\)
   6: \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

**Schedule**$\langle s, f, n \rangle$

1: sort jobs according to $f$ values
2: $t \leftarrow 0$, $S \leftarrow \emptyset$
3: **for** every $j \in [n]$ according to non-decreasing order of $f_j$ **do**
   4:     **if** $s_j \geq t$ **then**
   5:         $S \leftarrow S \cup \{j\}$
   6:         $t \leftarrow f_j$
7: **return** $S$
Clever Implementation of Greedy Algorithm

Schedule$(s, f, n)$

1: sort jobs according to $f$ values
2: $t \leftarrow 0$, $S \leftarrow \emptyset$
3: for every $j \in [n]$ according to non-decreasing order of $f_j$ do
   4: if $s_j \geq t$ then
      5: $S \leftarrow S \cup \{j\}$
      6: $t \leftarrow f_j$
7: return $S$
Clever Implementation of Greedy Algorithm

Schedule($s, f, n$)

1: sort jobs according to $f$ values
2: $t \leftarrow 0$, $S \leftarrow \emptyset$
3: for every $j \in [n]$ according to non-decreasing order of $f_j$ do
4:     if $s_j \geq t$ then
5:         $S \leftarrow S \cup \{j\}$
6:         $t \leftarrow f_j$
7: return $S$
Clever Implementation of Greedy Algorithm

\[ \text{Schedule}(s, f, n) \]

1. sort jobs according to \( f \) values
2. \( t \leftarrow 0, S \leftarrow \emptyset \)
3. for every \( j \in [n] \) according to non-decreasing order of \( f_j \) do
4. \hspace{1em} if \( s_j \geq t \) then
5. \hspace{2em} \( S \leftarrow S \cup \{j\} \)
6. \hspace{2em} \( t \leftarrow f_j \)
7. return \( S \)
Clever Implementation of Greedy Algorithm

**Schedule**($s, f, n$)

1. sort jobs according to $f$ values
2. $t \leftarrow 0$, $S \leftarrow \emptyset$
3. for every $j \in [n]$ according to non-decreasing order of $f_j$ do
   4. if $s_j \geq t$ then
   5. $S \leftarrow S \cup \{j\}$
   6. $t \leftarrow f_j$
7. return $S$
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0\), \(S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4:    if \(s_j \geq t\) then
5:       \(S \leftarrow S \cup \{j\}\)
6:       \(t \leftarrow f_j\)
7: return \(S\)
Clever Implementation of Greedy Algorithm

Schedule\((s, f, n)\)

1: sort jobs according to \(f\) values
2: \(t \leftarrow 0, S \leftarrow \emptyset\)
3: for every \(j \in [n]\) according to non-decreasing order of \(f_j\) do
4: \(\text{if } s_j \geq t \text{ then}\)
5: \(S \leftarrow S \cup \{j\}\)
6: \(t \leftarrow f_j\)
7: return \(S\)
Outline

1. Toy Example: Box Packing
2. Interval Scheduling
3. Offline Caching
   - Heap: Concrete Data Structure for Priority Queue
4. Data Compression and Huffman Code
5. Summary
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests

Cache miss happens if requested page not in cache. We need to bring the page into cache, and evict some existing page if necessary.

Cache hit happens if requested page already in cache.

Goal: minimize the number of cache misses.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests

<table>
<thead>
<tr>
<th>page sequence</th>
<th>cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store \( k \) pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.

<table>
<thead>
<tr>
<th>Page Sequence</th>
<th>Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>5</td>
<td>✗</td>
</tr>
<tr>
<td>4</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store \( k \) pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need to bring the page into cache, and evict some existing page if necessary.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.

<table>
<thead>
<tr>
<th>Page Sequence</th>
<th>Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- Cache hit happens if requested page already in cache.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- Cache hit happens if requested page already in cache.
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- Cache hit happens if requested page already in cache.
## Offline Caching

- **Cache that can store** \( k \) **pages**
- **Sequence of page requests**
- **Cache miss happens if** requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- **Cache hit happens if** requested page already in cache.

<table>
<thead>
<tr>
<th>Page Sequence</th>
<th>Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✔</td>
</tr>
<tr>
<td>5</td>
<td>❌</td>
</tr>
<tr>
<td>4</td>
<td>❌</td>
</tr>
<tr>
<td>2</td>
<td>❌</td>
</tr>
<tr>
<td>5</td>
<td>❌</td>
</tr>
<tr>
<td>3</td>
<td>❌</td>
</tr>
<tr>
<td>2</td>
<td>✔</td>
</tr>
<tr>
<td>1</td>
<td>❌</td>
</tr>
</tbody>
</table>
Offline Caching

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- Cache hit happens if requested page already in cache.

<table>
<thead>
<tr>
<th>Page Sequence</th>
<th>Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x 1</td>
</tr>
<tr>
<td>5</td>
<td>x 5</td>
</tr>
<tr>
<td>4</td>
<td>x 4</td>
</tr>
<tr>
<td>2</td>
<td>x 4</td>
</tr>
<tr>
<td>5</td>
<td>x 4, 2, 5</td>
</tr>
<tr>
<td>3</td>
<td>x 4, 2, 3</td>
</tr>
<tr>
<td>2</td>
<td>✔ 4, 2, 3</td>
</tr>
<tr>
<td>1</td>
<td>x 1, 2, 3</td>
</tr>
</tbody>
</table>

misses = 7
**Offline Caching**

- Cache that can store $k$ pages
- Sequence of page requests
- Cache miss happens if requested page not in cache. We need bring the page into cache, and evict some existing page if necessary.
- Cache hit happens if requested page already in cache.
- Goal: minimize the number of cache misses.

<table>
<thead>
<tr>
<th>page sequence</th>
<th>cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗ 1</td>
</tr>
<tr>
<td>5</td>
<td>✗ 5</td>
</tr>
<tr>
<td>4</td>
<td>✗ 4</td>
</tr>
<tr>
<td>2</td>
<td>✗ 4</td>
</tr>
<tr>
<td>5</td>
<td>✗ 4</td>
</tr>
<tr>
<td>3</td>
<td>✗ 4</td>
</tr>
<tr>
<td>2</td>
<td>✔ 4</td>
</tr>
<tr>
<td>1</td>
<td>✗ 1</td>
</tr>
</tbody>
</table>

misses = 7
## A Better Solution for Example

<table>
<thead>
<tr>
<th>page sequence</th>
<th>cache</th>
<th>cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="#" alt="1" /></td>
<td><img src="#" alt="1" /></td>
</tr>
<tr>
<td>5</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
<tr>
<td>5</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="#" alt="✓" /></td>
<td><img src="#" alt="✓" /></td>
</tr>
<tr>
<td>1</td>
<td><img src="#" alt="x" /></td>
<td><img src="#" alt="x" /></td>
</tr>
</tbody>
</table>

- misses = 7
- misses = 6
Offline Caching Problem

**Input:**  
- $k$: the size of cache  
- $n$: number of pages  
- $\rho_1, \rho_2, \rho_3, \cdots, \rho_T \in [n]$: sequence of requests

**Output:**  
- $i_1, i_2, i_3, \cdots, i_T \in \{\text{hit, empty}\} \cup [n]$: indices of pages to evict ("hit" means evicting no page, "empty" means evicting empty page)

We use $[n]$ for \{1, 2, 3, \cdots , n\}.  

Offline Caching: we know the whole sequence ahead of time.  
Online Caching: we have to make decisions on the fly, before seeing future requests.

Q: Which one is more realistic?  
A: Online caching
Offline Caching Problem

**Input:**
- \( k \): the size of cache
- \( n \): number of pages

\( \rho_1, \rho_2, \rho_3, \ldots, \rho_T \in [n] \): sequence of requests

**Output:**
- \( i_1, i_2, i_3, \ldots, i_T \in \{ \text{hit}, \text{empty} \} \cup [n] \): indices of pages to evict ("hit" means evicting no page, "empty" means evicting empty page)

- Offline Caching: we know the whole sequence ahead of time.
- Online Caching: we have to make decisions on the fly, before seeing future requests.
**Offline Caching Problem**

**Input:**
- $k$: the size of cache
- $n$: number of pages

We use $[n]$ for $\{1, 2, 3, \cdots, n\}$.

- $\rho_1, \rho_2, \rho_3, \cdots, \rho_T \in [n]$: sequence of requests

**Output:**
- $i_1, i_2, i_3, \cdots, i_T \in \{\text{hit, empty}\} \cup [n]$:
  - indices of pages to evict ("hit" means evicting no page, "empty" means evicting empty page)

- Offline Caching: we know the whole sequence ahead of time.
- Online Caching: we have to make decisions on the fly, before seeing future requests.

**Q:** Which one is more realistic?
Offline Caching Problem

**Input:**
- $k$: the size of cache
- $n$: number of pages

**Output:**
- $\rho_1, \rho_2, \rho_3, \cdots, \rho_T \in [n]$: sequence of requests
- $i_1, i_2, i_3, \cdots, i_T \in \{\text{hit, empty}\} \cup [n]$: indices of pages to evict ("hit" means evicting no page, "empty" means evicting empty page)

- Offline Caching: we know the whole sequence ahead of time.
- Online Caching: we have to make decisions on the fly, before seeing future requests.

**Q:** Which one is more realistic?

**A:** Online caching
Offline Caching: we know the whole sequence ahead of time.

Online Caching: we have to make decisions on the fly, before seeing future requests.

Q: Which one is more realistic?

A: Online caching

Q: Why do we study the offline caching problem?
• Offline Caching: we know the whole sequence ahead of time.
• Online Caching: we have to make decisions on the fly, before seeing future requests.

Q: Which one is more realistic?

A: Online caching

Q: Why do we study the offline caching problem?

A: Use the offline solution as a benchmark to measure the “competitive ratio” of online algorithms
Offline Caching: Potential Greedy Algorithms

- FIFO (First-In-First-Out): always evict the first page in cache
Offline Caching: Potential Greedy Algorithms

- **FIFO (First-In-First-Out):** always evict the first page in cache
- **LRU (Least-Recently-Used):** Evict page whose most recent access was earliest

All the above algorithms are not optimum! Indeed all the algorithms are "online", i.e., the decisions can be made without knowing future requests. Online algorithms can not be optimum.
Offline Caching: Potential Greedy Algorithms

- FIFO (First-In-First-Out): always evict the first page in cache
- LRU (Least-Recently-Used): Evict page whose most recent access was earliest
- LFU (Least-Frequently-Used): Evict page that was least frequently requested
Offline Caching: Potential Greedy Algorithms

- FIFO (First-In-First-Out): always evict the first page in cache
- LRU (Least-Recently-Used): Evict page whose most recent access was earliest
- LFU (Least-Frequently-Used): Evict page that was least frequently requested

All the above algorithms are not optimum!

Indeed all the algorithms are “online”, i.e, the decisions can be made without knowing future requests. Online algorithms can not be optimum.
FIFO is not optimum

requests

FIFO

1
2
3
4
1
FIFO is not optimum

requests
1
2
3
4

FIFO

1

×
FIFO is not optimum

FIFO

requests

1
2
3
4

×
1

FIFO is not optimum
FIFO is not optimum
FIFO is not optimum

requests

FIFO

1
2
3
4
1
2
3
4
1
2
FIFO is not optimum

requests

1
2
3
4
1
2
3
4
1
2
3
FIFO

× 1
× 1 2
× 1 2 3
FIFO is not optimum

requests

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIFO

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIFO is not optimum

<table>
<thead>
<tr>
<th>requests</th>
<th>FIFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗ 1</td>
</tr>
<tr>
<td>2</td>
<td>✗ 1 2</td>
</tr>
<tr>
<td>3</td>
<td>✗ 1 2 3</td>
</tr>
<tr>
<td>4</td>
<td>✗ 4 2 3</td>
</tr>
</tbody>
</table>

1 FIFO requests are not optimum.
FIFO is not optimum

FIFO

requests

1
2
3
4

1

2

3

4

1
FIFO is not optimum
FIFO is not optimum

<table>
<thead>
<tr>
<th>requests</th>
<th>FIFO</th>
<th>misses = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>× 1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>× 1 2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>× 1 2 3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>× 4 2 3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>× 4 1 3</td>
<td></td>
</tr>
</tbody>
</table>
FIFO is not optimum

<table>
<thead>
<tr>
<th>requests</th>
<th>FIFO</th>
<th>Furthest-in-Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image" alt="FIFO request 1" /></td>
<td><img src="image" alt="Furthest-in-Future request 1" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image" alt="FIFO request 2" /></td>
<td><img src="image" alt="Furthest-in-Future request 2" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="image" alt="FIFO request 3" /></td>
<td><img src="image" alt="Furthest-in-Future request 3" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="image" alt="FIFO request 4" /></td>
<td><img src="image" alt="Furthest-in-Future request 4" /></td>
</tr>
</tbody>
</table>

misses = 5

misses = 4
**Furthest-in-Future (FF)**

- Algorithm: every time, evict the item that is not requested until furthest in the future, if we need to evict one.
- The algorithm is not an online algorithm, since the decision at a step depends on the request sequence in the future.
Furthest-in-Future (FF)

<table>
<thead>
<tr>
<th>requests</th>
<th>FIFO</th>
<th>Furthest-in-Future</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td>4</td>
</tr>
</tbody>
</table>

misses = 5

misses = 4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

×  ×  ×

1  1  1

5  5

4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3
Example

requests

1 5 4 2 1
5 3 2 4 5
3 1 5 4 2

1 1 1 2
5 5 5
4 4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

x  x  x  x

1  1  1  2

5  5  5

4  4
Example

requests

1 5 4 2 5 3 2 4 3 1 5 3

X X X X ✔

1 1 1 2 2

5 5 5 5

4 4 4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

\[\square  \ 1\  1\  1\  2\  2\]
\[\square  \ 5\  5\  5\  5\  5\]
\[\square  \ 4\  4\  4\]
Example

requests

\[
\begin{array}{cccccccc}
1 & 5 & 4 & 2 & 5 & 3 & 2 & 4 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
\checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\
\end{array}
\]

34/80
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

x  x  x  x  ✓  x

1  1  1  2  2  2

5  5  5  5  5  3

4  4  4  4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

X  X  X  X  ✓  X  ✓  ✓

[]  1  1  1  2  2  2  2

[]  []  5  5  5  5  3  3

[]  []  []  4  4  4  4  4
Example

requests

1 5 4 2 5 3 2 4 3 1 5 3

\[\times\ \times\ \times\ \times\ \checkmark\ \times\ \checkmark\ \checkmark\]
## Example

### Requests

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

- **X** **X** **X** **X** **✓** **X** **✓** **✓** **✓**
Example

requests

\[1\ 5\ 4\ 2\ 5\ 3\ 2\ 4\ 3\ 1\ 5\ 3\]

\[\times\ \times\ \times\ \times\ \checkmark\ \times\ \checkmark\ \checkmark\ \checkmark\]

\[
\square\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2
\]

\[
\square\ \square\ 5\ 5\ 5\ 5\ 3\ 3\ 3\ 3
\]

\[
\square\ \square\ \square\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4
\]
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

×  ×  ×  ×  √  ×  √  √  √  ×

1  1  1  2  2  2  2  2  2  1

5  5  5  5  3  3  3  3  3

4  4  4  4  4  4  4  4  4
Example

requests

1  5  4  2  5  3  2  4  3  1  5  3

✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓
Example

requests

\[
\begin{array}{cccccccccccccc}
1 & 5 & 4 & 2 & 5 & 3 & 2 & 4 & 3 & 1 & 5 & 3 \\
\times & \times & \times & \times & \checkmark & \times & \checkmark & \checkmark & \checkmark & \times & \times & \checkmark \\
\square & 1 & 1 & 5 & 2 & 2 & 2 & 2 & 2 & 1 & 5 & 5 \\
\square & \square & 5 & 5 & 5 & 5 & 3 & 3 & 3 & 3 & 3 & 3 \\
\square & \square & \square & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\
\end{array}
\]
Recall: Designing and Analyzing Greedy Algorithms

Greedy Algorithm
- Build up the solutions in steps
- At each step, make an irrevocable decision using a “reasonable” strategy

Analysis of Greedy Algorithm
- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)
Recall: Designing and Analyzing Greedy Algorithms

<table>
<thead>
<tr>
<th>Greedy Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Build up the solutions in steps</td>
</tr>
<tr>
<td>• At each step, make an <strong>irrevocable</strong> decision using a “reasonable” strategy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis of Greedy Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Prove that the reasonable strategy is “safe” (key)</td>
</tr>
<tr>
<td>• Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)</td>
</tr>
</tbody>
</table>
Offline Caching Problem

**Input:**
- \( k \): the size of cache
- \( n \): number of pages
- \( \rho_1, \rho_2, \rho_3, \ldots, \rho_T \in [n] \): sequence of requests

**Output:**
- \( i_1, i_2, i_3, \ldots, i_t \in \{ \text{hit}, \text{empty} \} \cup [n] \)
  - empty stands for an empty page
  - “hit” means evicting no pages
**Offline Caching Problem**

**Input:**
- $k$: the size of cache
- $n$: number of pages
- $\rho_1, \rho_2, \rho_3, \ldots, \rho_T \in [n]$: sequence of requests
- $p_1, p_2, \ldots, p_k \in \{\text{empty}\} \cup [n]$: initial set of pages in cache

**Output:**
- $i_1, i_2, i_3, \ldots, i_t \in \{\text{hit, empty}\} \cup [n]$
  - empty stands for an empty page
  - “hit” means evicting no pages
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is "safe" (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

**Lemma** Assume at time 1 a page fault happens and there are no empty pages in the cache. Let \( p^* \) be the page in cache that is not requested until furthest in the future. **It is safe to evict** \( p^* \) **at time 1.**
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

**Lemma** Assume at time 1 a page fault happens and there are no empty pages in the cache. Let \( p^* \) be the page in cache that is not requested until furthest in the future. There is an optimum solution in which \( p^* \) is evicted at time 1.
Proof.

1. \( S \): any optimum solution
2. \( p^* \): page in cache not requested until furthest in the future.
   - In the example, \( p^* = 3 \).
Proof.

1. \( S \): any optimum solution
2. \( p^* \): page in cache not requested until furthest in the future.
   - In the example, \( p^* = 3 \).
3. Assume \( S \) evicts some \( p' \neq p^* \) at time 1; otherwise done.
   - In the example, \( p' = 2 \).
Proof.

1. $S$: any optimum solution
2. $p^*$: page in cache not requested until furthest in the future.
   - In the example, $p^* = 3$.
3. Assume $S$ evicts some $p' \neq p^*$ at time 1; otherwise done.
   - In the example, $p' = 2$. 
Proof.

Create $S'$.

$S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.

After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S'$ contains $p^*(=3)$.

From now on, $S'$ will "copy" $S$. 

\[
S : \begin{array}{c|c}
1 & 1 \\
2 & 4 \\
3 & 3 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
4 & 2 & 3 \\
\end{array}
\]
**Proof.**

Create $S'$. $S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.
Proof.

4. Create $S'$.

$S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.
Proof.


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$. 
**Proof.**

1. Create $S'$. $S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.
2. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.
3. From now on, $S'$ will “copy” $S$.
Proof.


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S'$ contains $p^*(=3)$.

6. From now on, $S'$ will “copy” $S$. 
**Proof.**


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6. From now on, $S'$ will “copy” $S$. 
Proof.


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6. From now on, $S'$ will “copy” $S$. 
Proof.

1. Create $S'$. $S'$ evicts $p^* (=3)$ instead of $p' (=2)$ at time 1.

2. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p' (=2)$ and $S$ contains $p^* (=3)$.

3. From now on, $S'$ will "copy" $S$. 

Proof.

4 Create $S'$. $S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.

5 After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6 From now on, $S'$ will “copy” $S$. 
Proof.


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6. From now on, $S'$ will “copy” $S$. 
Proof.

4 Create $S'$. $S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.

5 After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6 From now on, $S'$ will “copy” $S$. 
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

$S$:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

$S'$:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Proof.


5. After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S$ contains $p^*(=3)$.

6. From now on, $S'$ will “copy” $S$. 
Proof.

4 Create $S'$. $S'$ evicts $p^*(=3)$ instead of $p'(=2)$ at time 1.

5 After time 1, cache status of $S$ and that of $S'$ differ by only 1 page. $S$ contains $p'(=2)$ and $S'$ contains $p^*(=3)$.

6 From now on, $S'$ will “copy” $S$. 

\[ \begin{array}{cccccc}
\hline
 & 4 & 5 & 4 & 6 & 2 & 3 \\
\hline
S: & \times & \times & \checkmark & \times \\
& 1 & 1 & 5 & 5 & 5 \\
& 2 & 4 & 4 & 4 & 4 \\
& 3 & 3 & 3 & 3 & 6 \\
\hline
S': & \times & \times & \checkmark & \times \\
& 1 & 1 & 5 & 5 & 5 \\
& 2 & 4 & 4 & 4 & 4 \\
& 3 & 2 & 2 & 2 & 6 \\
\hline
\end{array} \]
**Proof.**

If $S$ evicted page $p'$, $S'$ will evict the page $p^\ast$. Then, the cache status of $S$ and that of $S'$ will be the same. $S$ and $S'$ will be exactly the same from now on.
Proof.

If $S$ evicted the page $p'$, $S'$ will evict the page $p^*$. Then, the cache status of $S$ and that of $S'$ will be the same. $S$ and $S'$ will be exactly the same from now on.
Proof.

7. If $S$ evicted the page $p'$, $S'$ will evict the page $p^*$. Then, the cache status of $S$ and that of $S'$ will be the same. $S$ and $S'$ will be exactly the same from now on.

8. Assume $S$ did not evict $p'(=2)$ before we see $p'(=2)$. 

\[
\begin{array}{cccc}
4 & 5 & 4 & 2 \\
\times & \times & & \\
1 & 1 & 5 & 5 \\
2 & 4 & 4 & 4 \\
3 & 3 & 3 & 3 \\
\hline
S: & & & \checkmark \\
\hline
1 & 1 & 5 & 5 \\
2 & 4 & 4 & 4 \\
3 & 2 & 2 & 2 \\
\hline
S': & & & \checkmark \\
\end{array}
\]
Proof.

7 If $S$ evicted the page $p'$, $S'$ will evict the page $p^*$. Then, the cache status of $S$ and that of $S'$ will be the same. $S$ and $S'$ will be exactly the same from now on.

8 Assume $S$ did not evict $p'(=2)$ before we see $p'(=2)$.
Proof.

If $S$ evicts $p^*$ ($=3$) for $p^*'$ ($=2$), then $S$ won't be optimum. Assume otherwise. So far, $S'$ has 1 less page-miss than $S$ does. The status of $S'$ and that of $S$ only differ by 1 page.
Proof.

If $S$ evicts $p^* (=3)$ for $p' (=2)$, then $S$ won't be optimum. Assume otherwise.

So far, $S'$ has 1 less page-miss than $S$ does.

The status of $S'$ and that of $S$ only differ by 1 page.
Proof.

If $S$ evicts $p^*$ ($=3$) for $p^*$ ($=2$), then $S$ won't be optimum. Assume otherwise.

So far, $S'$ has 1 less page-miss than $S$ does.

The status of $S'$ and that of $S$ only differ by 1 page.

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$S'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Proof.
Proof.

If $S$ evicts $p^*(=3)$ for $p'(=2)$, then $S$ won’t be optimum. Assume otherwise.
**Proof.**

If $S$ evicts $p^*(=3)$ for $p'(=2)$, then $S$ won’t be optimum. Assume otherwise.
Proof.

If \( S \) evicts \( p^*(=3) \) for \( p'(=2) \), then \( S \) won’t be optimum. Assume otherwise.
Proof.

If $S$ evicts $p^*(=3)$ for $p'(=2)$, then $S$ won’t be optimum. Assume otherwise.
Proof.

9 If $S$ evicts $p^*(=3)$ for $p'(=2)$, then $S$ won’t be optimum. Assume otherwise.

10 So far, $S'$ has 1 less page-miss than $S$ does.
Proof.

9 If $S$ evicts $p^*(=3)$ for $p'(=2)$, then $S$ won’t be optimum. Assume otherwise.

10 So far, $S'$ has 1 less page-miss than $S$ does.

11 The status of $S'$ and that of $S$ only differ by 1 page.
Proof.

We can then guarantee that $S'$ make at most the same number of page-misses as $S$ does. Idea: if $S$ has a page-hit and $S'$ has a page-miss, we use the opportunity to make the status of $S'$ the same as that of $S'$. 

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>5</th>
<th>5</th>
<th>...</th>
<th>6</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>...</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>5</th>
<th>5</th>
<th>...</th>
<th>6</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S'$</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>...</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Proof.

We can then guarantee that $S'$ make at most the same number of page-misses as $S$ does.
Proof.

We can then guarantee that $S'$ make at most the same number of page-misses as $S$ does.

- Idea: if $S$ has a page-hit and $S'$ has a page-miss, we use the opportunity to make the status of $S'$ the same as that of $S$. 

\[ \begin{array}{cccccccc}
4 & 5 & 4 & \cdots & 2 & 3 \\
\text{x} & \text{x} & \checkmark & \text{x} & \\
1 & 1 & 5 & 5 & \cdots & 6 & 2 \\
2 & 4 & 4 & 4 & \cdots & 8 & 8 \\
3 & 3 & 3 & 3 & \cdots & 3 & 3 \\
\end{array} \]

\[ \begin{array}{cccccccc}
1 & 1 & 5 & 5 & \cdots & 6 & 6 \\
\text{x} & \text{x} & \checkmark & \checkmark & \\
1 & 1 & 5 & 5 & \cdots & 6 & 6 \\
2 & 4 & 4 & 4 & \cdots & 8 & 8 \\
3 & 2 & 2 & 2 & \cdots & 2 & 2 \\
\end{array} \]
Thus, we have shown how to create another solution $S'$ with the same number of page-misses as that of the optimum solution $S$. Thus, we proved

**Lemma** Assume at time 1 a page fault happens and there are no empty pages in the cache. Let $p^*$ be the page in cache that is not requested until furthest in the future. There is an optimum solution in which $p^*$ is evicted at time 1.
Thus, we have shown how to create another solution $S'$ with the same number of page-misses as that of the optimum solution $S$. Thus, we proved

**Lemma** Assume at time 1 a page fault happens and there are no empty pages in the cache. Let $p^*$ be the page in cache that is not requested until furthest in the future. It is safe to evict $p^*$ at time 1.
Thus, we have shown how to create another solution $S'$ with the same number of page-misses as that of the optimum solution $S$. Thus, we proved

**Lemma**  Assume at time 1 a page fault happens and there are no empty pages in the cache. Let $p^*$ be the page in cache that is not requested until furthest in the future. It is safe to evict $p^*$ at time 1.

**Theorem**  The furthest-in-future strategy is optimum.
1: \textbf{for } t \leftarrow 1 \text{ to } T \text{ do} \\
2: \quad \textbf{if } \rho_t \text{ is in cache } \textbf{then} \text{ do nothing} \\
3: \quad \textbf{else if } \text{ there is an empty page in cache } \textbf{then} \\
4: \quad \quad \text{evict the empty page and load } \rho_t \text{ in cache} \\
5: \quad \textbf{else} \\
6: \quad \quad p^* \leftarrow \text{page in cache that is not used furthest in the future} \\
7: \quad \quad \text{evict } p^* \text{ and load } \rho_t \text{ in cache}
Q: How can we make the algorithm as fast as possible?

A: The running time can be made to be $O(n + T \log k)$. For each page $p$, use a linked list (or an array with dynamic size) to store the time steps in which $p$ is requested. We can find the next time a page is requested easily. Use a priority queue data structure to hold all the pages in cache, so that we can easily find the page that is requested furthest in the future.
Q: How can we make the algorithm as fast as possible?

A:
- The running time can be made to be $O(n + T \log k)$. 

For each page $p$, use a linked list (or an array with dynamic size) to store the time steps in which $p$ is requested. We can find the next time a page is requested easily. Use a priority queue data structure to hold all the pages in cache, so that we can easily find the page that is requested furthest in the future.
Q: How can we make the algorithm as fast as possible?

A:
- The running time can be made to be $O(n + T \log k)$.
- For each page $p$, use a linked list (or an array with dynamic size) to store the time steps in which $p$ is requested.
Q: How can we make the algorithm as fast as possible?

A:

- The running time can be made to be $O(n + T \log k)$.
- For each page $p$, use a linked list (or an array with dynamic size) to store the time steps in which $p$ is requested.
- We can find the next time a page is requested easily.
Q: How can we make the algorithm as fast as possible?

A:

- The running time can be made to be $O(n + T \log k)$.
- For each page $p$, use a linked list (or an array with dynamic size) to store the time steps in which $p$ is requested.
- We can find the next time a page is requested easily.
- Use a priority queue data structure to hold all the pages in cache, so that we can easily find the page that is requested furthest in the future.
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

**P1:**

```
1 10
```

**P2:**

```
4 7
```

**P3:**

```
6 9 12
```

**P4:**

```
3 8
```

**P5:**

```
2 5 11
```

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

P1:  

| P1: | 1 | 10 |

P2:  

| P2: | 4 | 7 |

P3:  

| P3: | 6 | 9 | 12 |

P4:  

| P4: | 3 | 8 |

P5:  

| P5: | 2 | 5 | 11 |

priority queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

### Priority Queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P1: 

| 1 | 10 |

P2: 

| 4 | 7 |

P3: 

| 6 | 9 | 12 |

P4: 

| 3 | 8 |

P5: 

| 2 | 5 | 11 |
time | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12
---|---|---|---|---|---|---|---|---|---|---|---|---|---
pages | P1 | P5 | P4 | P2 | P5 | P3 | P2 | P4 | P3 | P1 | P5 | P3 |   

P1:  
- Pages: 1, 10

P2:  
- Pages: 4, 7

P3:  
- Pages: 6, 9, 12

P4:  
- Pages: 3, 8

P5:  
- Pages: 2, 5, 11

priority queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
</tbody>
</table>

46/80
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pages**

- **P1**: 1 10
- **P2**: 4 7
- **P3**: 6 9 12
- **P4**: 3 8
- **P5**: 2 5 11
time | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12
---|---|---|---|---|---|---|---|---|---|---|----|----|----
pages | P1 | P5 | P4 | P2 | P5 | P3 | P2 | P4 | P3 | P1 | P5 | P3

priority queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
</tbody>
</table>
The figure shows a priority queue with the following contents:

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
</tbody>
</table>

The priority queue is updated as follows:

<table>
<thead>
<tr>
<th>time</th>
<th>pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>P1</td>
</tr>
<tr>
<td>1</td>
<td>P5</td>
</tr>
<tr>
<td>2</td>
<td>P4</td>
</tr>
<tr>
<td>3</td>
<td>P2</td>
</tr>
<tr>
<td>4</td>
<td>P5</td>
</tr>
<tr>
<td>5</td>
<td>P3</td>
</tr>
<tr>
<td>6</td>
<td>P2</td>
</tr>
<tr>
<td>7</td>
<td>P4</td>
</tr>
<tr>
<td>8</td>
<td>P3</td>
</tr>
<tr>
<td>9</td>
<td>P1</td>
</tr>
<tr>
<td>10</td>
<td>P5</td>
</tr>
<tr>
<td>11</td>
<td>P3</td>
</tr>
<tr>
<td>12</td>
<td>P2</td>
</tr>
</tbody>
</table>

The values in the priority queue are:

- P1: 10
- P5: 5

The pages at times 1, 2, and 12 are marked with an x to indicate they are not the most recent page accesses.
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td></td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
</tr>
</tbody>
</table>

P1: 1 10  
P2: 4 7  
P3: 6 9 12  
P4: 3 8  
P5: 2 5 11  

Priority queue

```plaintext
<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
</tbody>
</table>
```
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

P1: 1 10
P2: 4 7
P3: 6 9 12
P4: 3 8
P5: 2 5 11

priority queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>10</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
<tr>
<td>time</td>
<td>P1</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>7</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

P1: \[1\ 10\]

P2: \[4\ 7\]

P3: \[6\ 9\ 12\]

P4: \[3\ 8\]

P5: \[2\ 5\ 11\]

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>7</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
</tbody>
</table>
The diagram shows a timeline with time intervals and corresponding pages for different processes. The priority queue is listed separately with pages and their associated priority values.

<table>
<thead>
<tr>
<th>time</th>
<th>pages</th>
<th>priority queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>P1</td>
<td>pages</td>
</tr>
<tr>
<td>1</td>
<td>P5</td>
<td>P2</td>
</tr>
<tr>
<td>2</td>
<td>P4</td>
<td>P5</td>
</tr>
<tr>
<td>3</td>
<td>P2</td>
<td>P3</td>
</tr>
<tr>
<td>4</td>
<td>P4</td>
<td>P2</td>
</tr>
<tr>
<td>5</td>
<td>P5</td>
<td>P4</td>
</tr>
<tr>
<td>6</td>
<td>P3</td>
<td>P1</td>
</tr>
<tr>
<td>7</td>
<td>P5</td>
<td>P5</td>
</tr>
<tr>
<td>8</td>
<td>P1</td>
<td>P3</td>
</tr>
<tr>
<td>9</td>
<td>P5</td>
<td>P1</td>
</tr>
<tr>
<td>10</td>
<td>P5</td>
<td>P5</td>
</tr>
<tr>
<td>11</td>
<td>P3</td>
<td>P3</td>
</tr>
<tr>
<td>12</td>
<td>P3</td>
<td>P3</td>
</tr>
</tbody>
</table>

The page with the priority value 11 is highlighted, indicating it is the page with the highest priority value at that time.
A priority queue is used to manage the pages. The queue is organized by priority values, with higher values indicating higher priority. The pages and their corresponding priority values are as follows:

- **Pages:**
  - P1: 1 10
  - P2: 4 7
  - P3: 6 9 12
  - P4: 3 8
  - P5: 2 5 11

- **Priority Values:**
  - P2: 7
  - P5: 11
  - P4: 8
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

**P1:**
- 1
- 10

**P2:**
- 4
- 7

**P3:**
- 6
- 9
- 12

**P4:**
- 3
- 8

**P5:**
- 2
- 5
- 11

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>7</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>7</td>
</tr>
<tr>
<td>P3</td>
<td>9</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>9</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
</tbody>
</table>

**Process Pages**

- P1: 1 10
- P2: 4 7
- P3: 6 9 12
- P4: 3 8
- P5: 2 5 11

Green checks indicate correct page replacement, red checks indicate incorrect page replacement.
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td></td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
</tr>
</tbody>
</table>

- **P1:** 1 10
- **P2:** 4 7
- **P3:** 6 9 12
- **P4:** 3 8
- **P5:** 2 5 11

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>9</td>
</tr>
<tr>
<td>P4</td>
<td>8</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

P1: 1 10

P2: 4 7

P3: 6 9 12

P4: 3 8

P5: 2 5 11

priority queue

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>9</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>9</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>
The diagram illustrates a priority queue scenario with pages and times. The table on the right shows the priority queue with pages and their priority values:

<table>
<thead>
<tr>
<th>Pages</th>
<th>Priority Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>

The pages are tracked over time in the main table, and the priority queue is updated accordingly.
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P1:  

\[ 1 \] 10 

P2:  

\[ 4 \] 7 

P3:  

\[ 6 \] 9 12 

P4:  

\[ 3 \] 8 

P5:  

\[ 2 \] 5 11 

---

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>

---

**Table**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

Priority queue:

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>

- P1: 1 10
- P2: 4 7
- P3: 6 9 12
- P4: 3 8
- P5: 2 5 11
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

- **P1:** time 1, priority 10
- **P2:** time 4, priority 7
- **P3:** time 6, priority 9
- **P4:** time 3, priority 8
- **P5:** time 2, priority 5

**Priority Queue:**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>

**Pages Accessed**

- P1 at time 1 and 10
- P2 at time 4 and 7
- P3 at time 6, 9, and 12
- P4 at time 3 and 8
- P5 at time 2, 5, and 11

- Red X marks pages not accessed
- Green check marks pages accessed
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

**Process Pages**

- **P1**: 1 10
- **P2**: 4 7
- **P3**: 6 9 12
- **P4**: 3 8
- **P5**: 2 5 11
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>priority queue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pages</td>
</tr>
<tr>
<td>P1:</td>
<td>1</td>
</tr>
<tr>
<td>P2:</td>
<td>4</td>
</tr>
<tr>
<td>P3:</td>
<td>6</td>
</tr>
<tr>
<td>P4:</td>
<td>3</td>
</tr>
<tr>
<td>P5:</td>
<td>2</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
</tr>
<tr>
<td>pages</td>
<td>P1</td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5</td>
<td>(\infty)</td>
</tr>
<tr>
<td>P3</td>
<td>12</td>
</tr>
<tr>
<td>P4</td>
<td>(\infty)</td>
</tr>
</tbody>
</table>

**Pages**

| P1   | 1 10 |
| P2   | 4 7  |
| P3   | 6 9 12 |
| P4   | 3 8  |
| P5   | 2 5 11 |

**Priority Queue**

| P1   | 4 7 12 |
| P2   | 3 5 11 |
| P3   | 2 5 11 |
| P4   | 3 8  |
| P5   | 1 10 |

- **X** represents pages that need to be swapped.
- **✓** represents pages that are already in the correct order.
- **∞** represents pages with infinite priority.
<table>
<thead>
<tr>
<th>time</th>
<th>0</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages</td>
<td>P1</td>
<td>P5</td>
<td>P4</td>
<td>P2</td>
<td>P5</td>
<td>P3</td>
<td>P2</td>
<td>P4</td>
<td>P3</td>
<td>P1</td>
<td>P5</td>
<td>P3</td>
<td></td>
</tr>
</tbody>
</table>

**Priority Queue**

<table>
<thead>
<tr>
<th>pages</th>
<th>priority values</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5</td>
<td>∞</td>
</tr>
<tr>
<td>P3</td>
<td>∞</td>
</tr>
<tr>
<td>P4</td>
<td>∞</td>
</tr>
</tbody>
</table>
1: for every $p \leftarrow 1$ to $n$ do
2: \hspace{1em} $times[p] \leftarrow$ array of times in which $p$ is requested, in increasing order \hspace{1em} $\triangleright$ put $\infty$ at the end of array
3: \hspace{1em} $pointer[p] \leftarrow 1$
4: $Q \leftarrow$ empty priority queue
5: for every $t \leftarrow 1$ to $T$ do
6: \hspace{1em} $pointer[\rho_t] \leftarrow pointer[\rho_t] + 1$
7: \hspace{1em} $nexttime[\rho_t] \leftarrow times[\rho_t, pointer[\rho_t]]$
8: if $\rho_t \in Q$ then
9: \hspace{1em} $Q$.increase-key($\rho_t, nexttime[\rho_t]$), print “hit”, continue
10: if $Q.size() < k$ then
11: \hspace{1em} print “load $\rho_t$ to an empty page”
12: else
13: \hspace{1em} $p \leftarrow Q$.extract-max(), print “evict $p$ and load $\rho_t$”
14: \hspace{1em} $Q$.insert($\rho_t, nexttime[\rho_t]$) \hspace{1em} $\triangleright$ add $\rho_t$ to $Q$ with key value $nexttime[\rho_t]$
Let $V$ be a ground set of size $n$.

**Def.** A **priority queue** is an abstract data structure that maintains a set $U \subseteq V$ of elements, each with an associated key value, and supports the following operations:

- **insert**$(v, key\_value)$: insert an element $v \in V \setminus U$, with associated key value $key\_value$.
- **decrease\_key**$(v, new\_key\_value)$: decrease the key value of an element $v \in U$ to $new\_key\_value$
- **extract\_min**(): return and remove the element in $U$ with the smallest key value
- ...
Simple Implementations for Priority Queue

- $n = \text{size of ground set } V$

<table>
<thead>
<tr>
<th>data structures</th>
<th>insert</th>
<th>extract_min</th>
<th>decrease_key</th>
</tr>
</thead>
<tbody>
<tr>
<td>array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sorted array</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
$n = \text{size of ground set } V$

<table>
<thead>
<tr>
<th>data structures</th>
<th>insert</th>
<th>extract_min</th>
<th>decrease_key</th>
</tr>
</thead>
<tbody>
<tr>
<td>array</td>
<td>$O(1)$</td>
<td>$O(n)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>sorted array</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Simple Implementations for Priority Queue

- \( n = \text{size of ground set } V \)

<table>
<thead>
<tr>
<th>data structures</th>
<th>insert</th>
<th>extract_min</th>
<th>decrease_key</th>
</tr>
</thead>
<tbody>
<tr>
<td>array</td>
<td>( O(1) )</td>
<td>( O(n) )</td>
<td>( O(1) )</td>
</tr>
<tr>
<td>sorted array</td>
<td>( O(n) )</td>
<td>( O(1) )</td>
<td>( O(n) )</td>
</tr>
</tbody>
</table>
Simple Implementations for Priority Queue

- $n = \text{size of ground set } V$

<table>
<thead>
<tr>
<th>data structures</th>
<th>insert</th>
<th>extract_min</th>
<th>decrease_key</th>
</tr>
</thead>
<tbody>
<tr>
<td>array</td>
<td>$O(1)$</td>
<td>$O(n)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>sorted array</td>
<td>$O(n)$</td>
<td>$O(1)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>heap</td>
<td>$O(\log n)$</td>
<td>$O(\log n)$</td>
<td>$O(\log n)$</td>
</tr>
</tbody>
</table>
The elements in a heap is organized using a complete binary tree:

- Nodes are indexed as \( \{1, 2, 3, \cdots, s\} \)
- Parent of node \( i \): \( \lfloor i/2 \rfloor \)
- Left child of node \( i \): \( 2i \)
- Right child of node \( i \): \( 2i + 1 \)
A heap $H$ contains the following fields:

- $s$: size of $U$ (number of elements in the heap)
- $A[i], 1 \leq i \leq s$: the element at node $i$ of the tree
- $p[v], v \in U$: the index of node containing $v$
- $key[v], v \in U$: the key value of element $v$

$s = 5$

$A = (‘f’, ‘g’, ‘c’, ‘e’, ‘b’)$

$p[‘f’] = 1, p[‘g’] = 2, p[‘c’] = 3, p[‘e’] = 4, p[‘b’] = 5$
Heap

The following heap property is satisfied:

- for any two nodes $i, j$ such that $i$ is the parent of $j$, we have $key[A[i]] \leq key[A[j]]$.

A heap. Numbers in the circles denote key values of elements.
$\text{insert}(v, key\_value)$
\text{insert}(v, \text{key\_value})
insert(v, key_value)
\text{insert}(v, key\_value)
$\text{insert}(v, key_value)$
**insert**($v, \text{key\_value}$)

1: $s \leftarrow s + 1$
2: $A[s] \leftarrow v$
3: $p[v] \leftarrow s$
4: $\text{key}[v] \leftarrow \text{key\_value}$
5: heapify\_up$(s)$

**heapify\_up**(i)

1: **while** $i > 1$ **do**
2: \hspace{1em} $j \leftarrow \lfloor i/2 \rfloor$
3: \hspace{1em} **if** $\text{key}[A[i]] < \text{key}[A[j]]$ **then**
4: \hspace{2em} swap $A[i]$ and $A[j]$
5: \hspace{2em} $p[A[i]] \leftarrow i, p[A[j]] \leftarrow j$
6: \hspace{1em} $i \leftarrow j$
7: \hspace{1em} **else** break
extract_min()
extract_min()
extract_min()
extract_min()
extract_min()
extract_min()
1: \textit{ret} \leftarrow A[1]
3: p[A[1]] \leftarrow 1
4: s \leftarrow s - 1
5: \textbf{if} \; s \geq 1 \; \textbf{then}
6: \; \textbf{heapify-down}(1)
7: \textbf{return} \; \textit{ret}

heapify-down(i)
1: \textbf{while} \; 2i \leq s \; \textbf{do}
2: \; \textbf{if} \; 2i = s \; \textbf{or}
\; \; \; key[A[2i]] \leq key[A[2i + 1]] \; \textbf{then}
3: \; \; j \leftarrow 2i
4: \; \textbf{else}
5: \; \; j \leftarrow 2i + 1
6: \; \; \textbf{if} \; key[A[j]] < key[A[i]] \; \textbf{then}
7: \; \; \; \textbf{swap} \; A[i] \; \textbf{and} \; A[j]
8: \; \; \; p[A[i]] \leftarrow i, \; p[A[j]] \leftarrow j
9: \; \; i \leftarrow j
10: \; \; \textbf{else} \; \textbf{break}

decrease_key(v, key_value)
1: key[v] \leftarrow key_value
2: \textbf{heapify-up}(p[v])
• Running time of heapify\_up and heapify\_down: $O(\lg n)$
- Running time of heapify\_up and heapify\_down: $O(\lg n)$
- Running time of insert, exact\_min and decrease\_key: $O(\lg n)$
- Running time of heapify\_up and heapify\_down: $O(\lg n)$
- Running time of insert, exact\_min and decrease\_key: $O(\lg n)$

<table>
<thead>
<tr>
<th>data structures</th>
<th>insert</th>
<th>extract_min</th>
<th>decrease_key</th>
</tr>
</thead>
<tbody>
<tr>
<td>array</td>
<td>$O(1)$</td>
<td>$O(n)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>sorted array</td>
<td>$O(n)$</td>
<td>$O(1)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>heap</td>
<td>$O(\lg n)$</td>
<td>$O(\lg n)$</td>
<td>$O(\lg n)$</td>
</tr>
</tbody>
</table>
Two Definitions Needed to Prove that the Procedures Maintain Heap Property

**Def.** We say that $H$ is almost a heap except that $key[A[i]]$ is too small if we can increase $key[A[i]]$ to make $H$ a heap.

**Def.** We say that $H$ is almost a heap except that $key[A[i]]$ is too big if we can decrease $key[A[i]]$ to make $H$ a heap.
Outline

1 Toy Example: Box Packing

2 Interval Scheduling

3 Offline Caching
   • Heap: Concrete Data Structure for Priority Queue

4 Data Compression and Huffman Code

5 Summary
Encoding Letters Using Bits

- 8 letters $a, b, c, d, e, f, g, h$ in a language
- need to encode a message using bits
- idea: use 3 bits per letter

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>000</td>
<td>001</td>
<td>010</td>
<td>011</td>
<td>100</td>
<td>101</td>
<td>110</td>
<td>111</td>
</tr>
</tbody>
</table>

$deacfg \rightarrow 011100000010101110$

Q: Can we have a better encoding scheme?

- Seems unlikely: must use 3 bits per letter

Q: What if some letters appear more frequently than the others?
Q: If some letters appear more frequently than the others, can we have a better encoding scheme?

A: Using variable-length encoding scheme might be more efficient.

Idea
- using fewer bits for letters that are more frequently used, and more bits for letters that are less frequently used.
Q: What is the issue with the following encoding scheme?

- $a$: 0
- $b$: 1
- $c$: 00

A: Can not guarantee a unique decoding. For example, 00 can be decoded to $aa$ or $c$.

Solution: Use prefix codes to guarantee a unique decoding.
Q: What is the issue with the following encoding scheme?

- $a$: 0
- $b$: 1
- $c$: 00

A: Can not guarantee a unique decoding. For example, 00 can be decoded to $aa$ or $c$. Use prefix codes to guarantee a unique decoding.
Q: What is the issue with the following encoding scheme?

- $a$: 0
- $b$: 1
- $c$: 00

A: Can not guarantee a unique decoding. For example, 00 can be decoded to $aa$ or $c$.

Solution
Use prefix codes to guarantee a unique decoding.
Def. A prefix code for a set $S$ of letters is a function $\gamma : S \rightarrow \{0, 1\}^*$ such that for two distinct $x, y \in S$, $\gamma(x)$ is not a prefix of $\gamma(y)$.
Def. A prefix code for a set \( S \) of letters is a function \( \gamma : S \to \{0, 1\}^* \) such that for two distinct \( x, y \in S \), \( \gamma(x) \) is not a prefix of \( \gamma(y) \).

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>( b )</td>
<td>( c )</td>
<td>( d )</td>
</tr>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>( e )</td>
<td>( f )</td>
<td>( g )</td>
<td>( h )</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

![Prefix Code Tree](image)

- \( \gamma(b) = 0 \)
- \( \gamma(c) = 1 \)
- \( \gamma(d) = 11 \)
- \( \gamma(e) = 01 \)
- \( \gamma(f) = 001 \)
- \( \gamma(g) = 0000 \)
- \( \gamma(h) = 0001 \)
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

Diagram: 

```
 b 0 1
  c 0 1
  a 0 1
 d 0 1
 h 0 1
 e
```

```
 f 0 1
 g 0 1
```

```
```

Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 00010011000000001011110100001001
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 0001/001100000001011110100001001
- c
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

0001/001/100000001011110100001001

c a
Prefix Codes Guarantee Unique Decoding

- **Reason**: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

0001/001/100/000001011110100001001

cad
Prefix Codes Guarantee Unique Decoding

Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

0001/001/100/0000/01011110100001001

cadb
Prefix Codes Guarantee Unique Decoding

Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

0001/001/100/0000/01/011110100001001

cadbh
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 0001/001/100/0000/01/01/1110100001001
- cadbh
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 0001/001/100/0000/01/01/11/10100001001
- cadbhhe
Prefix Codes Guarantee Unique Decoding

Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 0001/001/100/0000/01/01/11/1010/0001001
- cadbhhef
Prefix Codes Guarantee Unique Decoding

- Reason: there is only one way to cut the first code.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

- 0001/001/100/0000/01/01/11/1010/0001/001
- cadbhhefc
Prefix Codes Guarantee Unique Decoding

Reason: there is only one way to cut the first code.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0000</td>
<td>0001</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>11</td>
<td>1010</td>
<td>1011</td>
<td>01</td>
</tr>
</tbody>
</table>

0001/001/100/0000/01/01/11/1010/0001/001/
cadbhhefca
Properties of Encoding Tree

- Rooted binary tree
- Left edges labelled 0 and right edges labelled 1
- A leaf corresponds to a code for some letter
- If coding scheme is not wasteful: a non-leaf has exactly two children

Best Prefix Codes

- Input: frequencies of letters in a message
- Output: prefix coding scheme with the shortest encoding for the message
Properties of Encoding Tree

- Rooted binary tree

Best Prefix Codes

- Input: frequencies of letters in a message
- Output: prefix coding scheme with the shortest encoding for the message
Properties of Encoding Tree

- Rooted binary tree
- Left edges labelled 0 and right edges labelled 1

Best Prefix Codes

Input:
- frequencies of letters in a message

Output:
- prefix coding scheme with the shortest encoding for the message
Properties of Encoding Tree

- Rooted binary tree
- Left edges labelled 0 and right edges labelled 1
- A leaf corresponds to a code for some letter

Best Prefix Codes

Input:
- frequencies of letters in a message

Output:
- prefix coding scheme with the shortest encoding for the message
Properties of Encoding Tree

- Rooted binary tree
- Left edges labelled 0 and right edges labelled 1
- A leaf corresponds to a code for some letter
- If coding scheme is not wasteful: a non-leaf has exactly two children
Properties of Encoding Tree

- Rooted binary tree
- Left edges labelled 0 and right edges labelled 1
- A leaf corresponds to a code for some letter
- If coding scheme is not wasteful: a non-leaf has exactly two children

Best Prefix Codes

**Input:** frequencies of letters in a message

**Output:** prefix coding scheme with the shortest encoding for the message
example

<table>
<thead>
<tr>
<th>letters</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequencies</td>
<td>18</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

scheme 1 length: 2
scheme 2 length: 3
scheme 3 length: 2

total = 89

total = 87

total = 84

scheme 1

scheme 2

scheme 3
### Example

<table>
<thead>
<tr>
<th>letters</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequencies</td>
<td>18</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>scheme 1 length</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>scheme 2 length</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>scheme 3 length</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**Scheme 1**

```
(a) -> (b) -> (c)
```

**Scheme 2**

```
(a) -> (b) -> (c) -> (d) -> (e)
```

**Scheme 3**

```
(a) -> (d) -> (b) -> (c)
```
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)
Example Input: \((a: 18, \ b: 3, \ c: 4, \ d: 6, \ e: 10)\)

**Q:** What types of decisions should we make?
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)

**Q:** What types of decisions should we make?

- Can we directly give a code for some letter?
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)

Q: What types of decisions should we make?

- Can we directly give a code for some letter?
- Hard to design a strategy; residual problem is complicated.
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)

Q: What types of decisions should we make?

- Can we directly give a code for some letter?
- Hard to design a strategy; residual problem is complicated.
- Can we partition the letters into left and right sub-trees?
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)

Q: What types of decisions should we make?

- Can we directly give a code for some letter?
- Hard to design a strategy; residual problem is complicated.
- Can we partition the letters into left and right sub-trees?
- Not clear how to design the greedy algorithm
Example Input: \((a: 18, b: 3, c: 4, d: 6, e: 10)\)

**Q:** What types of decisions should we make?

- Can we directly give a code for some letter?
- Hard to design a strategy; residual problem is complicated.
- Can we partition the letters into left and right sub-trees?
- Not clear how to design the greedy algorithm

**A:** We can choose two letters and make them brothers in the tree.
Which Two Letters Can Be Safely Put Together As Brothers?

- Focus on the “structure” of the optimum encoding tree
Which Two Letters Can Be Safely Put Together As Brothers?

- Focus on the “structure” of the optimum encoding tree
- There are two deepest leaves that are brothers
Which Two Letters Can Be Safely Put Together As Brothers?

- Focus on the “structure” of the optimum encoding tree
- There are two deepest leaves that are brothers

best to put the two least frequent symbols here!
Which Two Letters Can Be Safely Put Together As Brothers?

- Focus on the “structure” of the optimum encoding tree
- There are two deepest leaves that are brothers

**Lemma** It is safe to make the two least frequent letters brothers.
**Lemma** There is an optimum encoding tree, where the two least frequent letters are brothers.
Lemma There is an optimum encoding tree, where the two least frequent letters are brothers.

- So we can irrevocably decide to make the two least frequent letters brothers.
**Lemma**  There is an optimum encoding tree, where the two least frequent letters are brothers.

- So we can irrevocably decide to make the two least frequent letters brothers.

**Q:** Is the residual problem another instance of the best prefix codes problem?
**Lemma** There is an optimum encoding tree, where the two least frequent letters are brothers.

- So we can irrevocably decide to make the two least frequent letters brothers.

**Q:** Is the residual problem another instance of the best prefix codes problem?

**A:** Yes, though it is not immediate to see why.
- \(f_x\): the frequency of the letter \(x\) in the support.
- \(x_1\) and \(x_2\): the two letters we decided to put together.
- \(d_x\): the depth of letter \(x\) in our output encoding tree.

\[
\begin{align*}
\sum_{x \in S} f_x d_x \\
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_{x_1} d_{x_1} + f_{x_2} d_{x_2} \\
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + (f_{x_1} + f_{x_2}) d_{x_1}
\end{align*}
\]
- $f_x$: the frequency of the letter $x$ in the support.
- $x_1$ and $x_2$: the two letters we decided to put together.
- $d_x$ the depth of letter $x$ in our output encoding tree.

\[
\sum_{x \in S} f_x d_x = \sum_{x \in S \setminus \{x_1, x_2\}} (f_x d_x + f_1 d_1 + f_2 d_2) = \sum_{x \in S \setminus \{x_1, x_2\}} (f_x d_x + (f_{x_1} + f_{x_2}) d_1)
\]
- $f_x$: the frequency of the letter $x$ in the support.
- $x_1$ and $x_2$: the two letters we decided to put together.
- $d_x$ the depth of letter $x$ in our output encoding tree.

\[
\sum_{x \in S} f_x d_x \\
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_{x_1} d_{x_1} + f_{x_2} d_{x_2} \\
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + (f_{x_1} + f_{x_2}) d_{x_1}
\]

Def: $f_{x'} = f_{x_1} + f_{x_2}$
- $f_x$: the frequency of the letter $x$ in the support.
- $x_1$ and $x_2$: the two letters we decided to put together.
- $d_x$ the depth of letter $x$ in our output encoding tree.

$$
\sum_{x \in S} f_x d_x = \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_1 d_1 + f_2 d_2
$$
$$
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + (f_1 + f_2) d_1
$$
$$
= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f' (d' + 1)
$$

Def: $f_{x'} = f_1 + f_2$
- $f_x$: the frequency of the letter $x$ in the support.
- $x_1$ and $x_2$: the two letters we decided to put together.
- $d_x$ the depth of letter $x$ in our output encoding tree.

**Def:** $f_{x'} = f_{x_1} + f_{x_2}$
• $f_x$: the frequency of the letter $x$ in the support.
• $x_1$ and $x_2$: the two letters we decided to put together.
• $d_x$ the depth of letter $x$ in our output encoding tree.

$$f_{x'} = f_{x_1} + f_{x_2}$$

$$\sum_{x \in S} f_x d_x$$

$$= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_{x_1} d_{x_1} + f_{x_2} d_{x_2}$$

$$= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + (f_{x_1} + f_{x_2}) d_{x_1}$$

$$= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_{x'} (d_{x'} + 1)$$

$$= \sum_{x \in S \setminus \{x_1, x_2\}} f_x d_x + f_{x'}$$

encoding tree for

$S \setminus \{x_1, x_2\} \cup \{x'\}$
In order to minimize
\[ \sum_{x \in S} f_x d_x, \]
we need to minimize
\[ \sum_{x \in S \setminus \{x_1, x_2\} \cup \{x'\}} f_x d_x, \]
subject to that \( d \) is the depth function for an encoding tree of \( S \setminus \{x_1, x_2\} \).

This is exactly the best prefix codes problem, with letters \( S \setminus \{x_1, x_2\} \cup \{x'\} \) and frequency vector \( f \).
Example

A 27  B 15  C 11  D 9  E 8  F 5
Example
Example
Example
Example
Example
Example

\[
\begin{align*}
A & : 00 \\
B & : 10 \\
C & : 010 \\
D & : 011 \\
E & : 110 \\
F & : 111
\end{align*}
\]
Def. The codes given the greedy algorithm is called the Huffman codes.
Def. The codes given the greedy algorithm is called the Huffman codes.

**Huffman**$(S, f)$

1: while $|S| > 1$ do
2: let $x_1, x_2$ be the two letters with the smallest $f$ values
3: introduce a new letter $x'$ and let $f_{x'} = f_{x_1} + f_{x_2}$
4: let $x_1$ and $x_2$ be the two children of $x'$
5: $S \leftarrow S \setminus \{x_1, x_2\} \cup \{x'\}$
6: return the tree constructed
Algorithm using Priority Queue

Huffman($S, f$)

1: $Q \leftarrow \text{build-priority-queue}(S)$
2: while $Q$.size $> 1$ do
3: $x_1 \leftarrow Q$.extract-min()
4: $x_2 \leftarrow Q$.extract-min()
5: introduce a new letter $x'$ and let $f_{x'} = f_{x_1} + f_{x_2}$
6: let $x_1$ and $x_2$ be the two children of $x'$
7: $Q$.insert($x'$, $f_{x'}$)
8: return the tree constructed
Outline

1. Toy Example: Box Packing
2. Interval Scheduling
3. Offline Caching
   - Heap: Concrete Data Structure for Priority Queue
4. Data Compression and Huffman Code
5. Summary
## Summary for Greedy Algorithms

### Greedy Algorithm
- Build up the solutions in steps
- At each step, make an **irrevocable** decision using a “reasonable” strategy
## Summary for Greedy Algorithms

**Greedy Algorithm**

- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a “reasonable” strategy

- Interval scheduling problem: schedule the job $j^*$ with the earliest deadline

---
Summary for Greedy Algorithms

**Greedy Algorithm**

- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a “reasonable” strategy

- Interval scheduling problem: schedule the job $j^*$ with the earliest deadline
- Offline Caching: evict the page that is used furthest in the future
Summary for Greedy Algorithms

**Greedy Algorithm**
- Build up the solutions in steps
- At each step, make an *irrevocable* decision using a “reasonable” strategy

- Interval scheduling problem: schedule the job $j^*$ with the earliest deadline
- Offline Caching: evict the page that is used furthest in the future
- Huffman codes: make the two least frequent letters brothers
Summary for Greedy Algorithms

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

Def. A strategy is “safe” if there is always an optimum solution that “agrees with” the decision made according to the strategy.
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S'$
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S$
- If $S$ agrees with the decision made according to the strategy, done
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S'$
- If $S$ agrees with the decision made according to the strategy, done
- So assume $S$ does not agree with decision
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S$
- If $S$ agrees with the decision made according to the strategy, done
- So assume $S$ does not agree with decision
- Change $S$ slightly to another optimum solution $S'$ that agrees with the decision
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S$
- If $S$ agrees with the decision made according to the strategy, done
- So assume $S$ does not agree with decision
- Change $S$ slightly to another optimum solution $S'$ that agrees with the decision
- Interval scheduling problem: exchange $j^*$ with the first job in an optimal solution
Take an arbitrary optimum solution $S$
If $S$ agrees with the decision made according to the strategy, done
So assume $S$ does not agree with decision
Change $S$ slightly to another optimum solution $S'$ that agrees with the decision
Interval scheduling problem: exchange $j^*$ with the first job in an optimal solution
Offline caching: a complicated “copying” algorithm
Proving a Strategy is Safe

- Take an arbitrary optimum solution $S$
- If $S$ agrees with the decision made according to the strategy, done
- So assume $S$ does not agree with decision
- Change $S$ slightly to another optimum solution $S'$ that agrees with the decision
  - Interval scheduling problem: exchange $j^*$ with the first job in an optimal solution
  - Offline caching: a complicated “copying” algorithm
  - Huffman codes: move the two least frequent letters to the deepest leaves.
### Analysis of Greedy Algorithm

- Prove that the reasonable strategy is "safe" (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)
### Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” *(key)*
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

- Interval scheduling problem: remove $j^*$ and the jobs it conflicts with
Summary for Greedy Algorithms

Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

- Interval scheduling problem: remove $j^*$ and the jobs it conflicts with
- Offline caching: trivial
Analysis of Greedy Algorithm

- Prove that the reasonable strategy is “safe” (key)
- Show that the remaining task after applying the strategy is to solve a (many) smaller instance(s) of the same problem (usually easy)

- Interval scheduling problem: remove $j^*$ and the jobs it conflicts with
- Offline caching: trivial
- Huffman codes: merge two letters into one