From Cardinalities to Costs

Given: number of TIDs to dereference
Question: disk access costs?
Two step solution:

1. estimate number of pages to be accessed
2. estimate costs for accessing these pages

Parameters

Given a set of k TIDs after an index access:

How many pages do we have to access to dereference them?

Let R be the relation for which we have to retrieve the tuples. Then we use the following abbreviations

N	$\|R\|$	number of tuples in the relation R
m	$\|\|R\|\|$	number of pages on which tuples of R are stored
B	N/m	number of tuples per page
k		number of (distinct) TIDs for which tuples have to be retrieved

We assume that the tuples are uniformely distributed among the m pages. Then, each page stores $B = N/m$ tuples. B is called *blocking factor*.

Special Cases

Let us consider some border cases.
If $k > N - N/m$ or $m = 1$, then all pages are accessed.
If $k = 1$ then exactly one page is accessed.

General Case

The answer to the general question will be expressed in terms of

- *buckets* (pages in the above case) and
- *items* contained therein (tuples in the above case).

Later on, we will also use extents, cylinders, or tracks as buckets and tracks or sectors/blocks as items.

Different Settings

Outline:

1. random/direct access
   - 1.1 items uniformly distributed among the buckets
     - 1.1.1 request k distinct items
     - 1.1.2 request k non-distinct items
   - 1.2 non-uniform distribution of items among buckets
2. sequential access

Always: uniform access probability

Direct, Uniform, Distinct

Additional assumption:
The probability that we request a set with k items is

$$\frac{1}{\binom{N}{k}}$$

for all of the

$$\binom{N}{k}$$

possibilities to select a k-set.
[Every k-set is accessed with the same probability.]

Direct, Uniform, Distinct (2)

Theorem (Waters/Yao)

Consider m buckets with n items each. Then there is a total of $N = nm$ items. If we randomly select k distinct items from all items then the number of qualifying buckets is

$$\overline{\mathcal{Y}}_n^{N,m}(k) = m * \mathcal{Y}_n^N(k) \tag{17}$$

where $\mathcal{Y}_n^N(k)$ is the probability that a bucket contains at least one item.

Direct, Uniform, Distinct (3)

Theorem (Waters/Yao (cont.))

The probability is

$$\mathcal{Y}_n^N(k) = \begin{cases} [1-p] & k \leq N-n \\ 1 & k > N-n \end{cases}$$

where p is the probability that a bucket contains none of the k items. The following alternative expressions can be used to calculate p:

$$p = \frac{\binom{N-n}{k}}{\binom{N}{k}} \tag{18}$$

$$= \prod_{i=0}^{k-1} \frac{N-n-i}{N-i} \tag{19}$$

$$= \prod_{i=0}^{n-1} \frac{N-k-i}{N-i} \tag{20}$$

Direct, Uniform, Distinct (4)

Proof (1): The total number of possibilities to pick the k items from all N items is

$$\binom{N}{k}$$

The number of possibilities to pick k items from all items not contained in a fixed single bucket is

$$\binom{N-n}{k}$$

Hence, the probability p that a bucket does not qualify is

$$p = \binom{N-n}{k} / \binom{N}{k}$$

Direct, Uniform, Distinct (5)

Proof (2):

$$
\begin{aligned}
p &= \frac{\binom{N-n}{k}}{\binom{N}{k}} \\
&= \frac{(N-n)!\ \ k!(N-k)!}{k!((N-n)-k)!\ \ N!} \\
&= \prod_{i=0}^{k-1} \frac{N-n-i}{N-i}
\end{aligned}
$$

Direct, Uniform, Distinct (6)

Proof(3):

$$
\begin{aligned}
p &= \frac{\binom{N-n}{k}}{\binom{N}{k}} \\
&= \frac{(N-n)!\ \ k!(N-k)!}{k!((N-n)-k)!\ \ N!} \\
&= \frac{(N-n)!\ \ (N-k)!}{N!\ \ ((N-k)-n)!} \\
&= \prod_{i=0}^{n-1} \frac{N-k-i}{N-i}
\end{aligned}
$$

Direct, Uniform, Distinct (7)

Implementation remark:

> *The fraction $m = N/n$ may not be an integer.*
> *For these cases, it is advisable to have a Gamma-function based implementation of binomial coeffcients at hand*

Evaluation of Yao's formula is expensive. Approximations are more efficient to calculate.

Direct, Uniform, Distinct (8)

Special cases:

If	then $\mathcal{Y}_m^N(k) =$
$n = 1$	k/N
$n = N$	1
$k = 0$	0
$k = 1$	B/N
$k = N$	1

Direct, Uniform, Distinct (9)

Let N items be distributed over N buckets such that every bucket contains exactly one item.
Further let us be interested in a subset of m buckets ($1 \leq m \leq N$).
If we pick k items then the number of buckets within the subset of size m that qualify is

$$m\mathcal{Y}_1^N(k) = m\frac{k}{N} \tag{21}$$

qualify.

Direct, Uniform, Distinct (10)

Proof:

$$
\begin{aligned}
\mathcal{Y}_1^N(k) &= (1 - \frac{\binom{N-1}{k}}{\binom{N}{k}}) \\
&= (1 - \frac{\frac{(N-1)!}{k!((N-1)-k)!}}{\frac{N!}{k!(N-k)!}}) \\
&= (1 - \frac{(N-1)!k!(N-k)!}{N!k!((N-1)-k)!}) \\
&= (1 - \frac{N-k}{N}) \\
&= (\frac{N}{N} - \frac{N-k}{N}) \\
&= \frac{N-N+k}{N} \\
&= \frac{k}{N}
\end{aligned}
$$

Direct, Uniform, Distinct (11)

Approximation of Yao's formula (1):

$$p \approx (1 - k/N)^n$$

[Waters]

Direct, Uniform, Distinct (12)

Approximation of Yao's formula (2):
$\overline{\mathcal{Y}}_n^{N,m}(k)$ can be approximated by:

$$\begin{aligned} m * [\quad & (1-(1-1/m)^k)+ \\ & (1/(m^2 b) * k(k-1)/2 * (1-1/m)^{k-1})+ \\ & (1.5/(m^3 b^4) * k(k-1)(2k-1)/6 * (1-1/m)^{k-1}) \quad] \end{aligned}$$

[Whang, Wiederhold, Sagalowicz]

Direct, Uniform, Distinct (13)

Approximation of Yao's formula (3):

$$\overline{\mathcal{Y}}_n^{N,m}(k) \approx \begin{cases} k & \text{if } k < \frac{m}{2} \\ \frac{k+m}{3} & \text{if } \frac{m}{2} \leq k < 2m \\ m & \text{if } 2m \leq k \end{cases}$$

[Bernstein, Goodman, Wong, Reeve, Rothnie]

Direct, Uniform, Distinct (14)

Upper and lower bounds for p:

$$p_{\text{lower}} = (1 - \frac{k}{N - \frac{n-1}{2}})^n$$

$$p_{\text{upper}} = ((1 - \frac{k}{N}) * (1 - \frac{k}{N - n + 1}))^{n/2}$$

for $n = N/m$.
Dihr and Saharia claim that the maximal difference resulting from the use of the lower and the upper bound to compute the number of page accesses is 0.224—far less than a single page access.

Direct, Uniform, Non-Distinct

Lemma

Let S be a set with $|S| = N$ elements. Then, the number of multisets with cardinality k containing only elements from S is

$$\binom{N+k-1}{k}$$

Proof: For a prove we just note that there is a bijection between the k-multisets and the k-subsets of a $N+k-1$-set.
We can go from a multiset to a set by f with

$$f(\{x_1 \leq \ldots \leq x_k\}) = \{x_1 + 0 < x_2 + 1 < \ldots < x_k + (k-1)\}$$

and from a set to a multiset via g with

$$g(\{x_1 < \ldots < x_k\}) = \{x_1 - 0 \leq x_2 - 1 \leq \ldots \leq x_k - (k-1)\}$$

Direct, Uniform, Non-Distinct (2)

Theorem (Cheung)

Consider m buckets with n items each. Then there is a total of $N = nm$ items. If we randomly select k not necessarily distinct items from all items, then the number of qualifying buckets is

$$\overline{Cheung}_n^{N,m}(k) = m * Cheung_n^N(k) \tag{22}$$

where

$$Cheung_n^N(k) = [1 - \tilde{p}] \tag{23}$$

Direct, Uniform, Non-Distinct (3)

Theorem (Cheung (cont.))

with the following equivalent expressions for $\tilde{p}$:

$$\tilde{p} = \frac{\binom{N-n+k-1}{k}}{\binom{N+k-1}{k}} \tag{24}$$

$$= \prod_{i=0}^{k-1} \frac{N-n+i}{N+i} \tag{25}$$

$$= \prod_{i=0}^{n-1} \frac{N-1-i}{N-1+k-i} \tag{26}$$

Direct, Uniform, Non-Distinct (4)

Proof(1):
Eq. 24 follows from the observation that the probability that some bucket does not contain any of the k possibly duplicate items is $\frac{\binom{N-n+k-1}{k}}{\binom{N+k-1}{k}}$.

Direct, Uniform, Non-Distinct (5)

Proof(2):
Eq. 25 follows from

$$
\begin{aligned}
\tilde{p} &= \frac{\binom{N-n+k-1}{k}}{\binom{N+k-1}{k}} \\
&= \frac{(N-n+k-1)!\;\; k!((N+k-1)-k)!}{k!((N-n+k-1)-k)!\;\; (N+k-1)!} \\
&= \frac{(N-n-1+k)!\;\; (N-1)!}{(N-n-1)!\;\; (N-1+k)!} \\
&= \prod_{i=0}^{k-1} \frac{N-n+i}{N+i}
\end{aligned}
$$

Direct, Uniform, Non-Distinct (6)

Proof(3):
Eq. 26 follows from

$$
\begin{aligned}
\tilde{p} &= \frac{\binom{N-n+k-1}{k}}{\binom{N+k-1}{k}} \\
&= \frac{(N-n+k-1)!\ \ k!((N+k-1)-k)!}{k!((N-n+k-1)-k)!\ \ (N+k-1)!} \\
&= \frac{(N+k-1-n)!\ \ (N-1)!}{(N+k-1)!\ \ (N-1-n)!} \\
&= \prod_{i=0}^{n-1} \frac{N-n+i}{N+k-n+i} \\
&= \prod_{i=0}^{n-1} \frac{N-1-i}{N-1+k-i}
\end{aligned}
$$

Direct, Uniform, Non-Distinct (7)

Approximation for $\tilde{p}$:

$$(1 - n/N)^k$$

[Cardenas]

Direct, Uniform, Non-Distinct (8)

Estimate for the number of distinct values in a bag:

Corollary

Let S be a k-multiset containing elements from an N-set T. Then the number of distinct items contained in S is

$$\mathcal{D}(N,k) = \frac{Nk}{N+k-1} \tag{27}$$

if the elements in T occur with the same probability in S.

Direct, Uniform, Non-Distinct (9)

Model switching:

$$\overline{\mathcal{Y}}_n^{N,m}(\mathit{Distinct}(N,k)) \approx \overline{\text{Cheung}}_n^{N,m}(k)$$

[for $n \geq 5$]

Direct, Non-Uniform, Distinct

So far:

1. every page contains the same number of records, and
2. every record is accessed with the same probability.

Now:

> *Model the distribution of items to buckets by m numbers n_i (for $1 \leq i \leq m$) if there are m buckets.*
> *Each n_i equals the number of records in some bucket i.*

Direct, Non-Uniform, Distinct (2)

The following theorem is a simple application of Yao's formula:

Theorem (Yao/Waters/Christodoulakis)

Assume a set of m buckets. Each bucket contains $n_j > 0$ items $(1 \leq j \leq m)$. The total number of items is $N = \sum_{j=1}^{m} n_j$. If we lookup k distinct items, then the probability that bucket j qualifies is

$$\mathcal{W}_{n_j}^{N}(k,j) = [1 - \frac{\binom{N-n_j}{k}}{\binom{N}{k}}] \quad (= \mathcal{Y}_{n_j}^{N}(k)) \tag{28}$$

and the expected number of qualifying buckets is

$$\overline{\mathcal{W}}_{n_j}^{N,m}(k) := \sum_{j=1}^{m} \mathcal{W}_{n_j}^{N}(k,j) \tag{29}$$

Direct, Non-Uniform, Distinct (3)

The product formulation in Eq. 20 of Theorem 2 results in a more efficient computation:

Corollary

If we lookup k distinct items, then the expected number of qualifying buckets is

$$\overline{W}_{n_j}^{N,m}(k) = \sum_{j=1}^{m}(1 - p_j) \tag{30}$$

with

$$p_j = \begin{cases} \prod_{i=0}^{n_j-1} \frac{N-k-i}{N-i} & k \leq n_j \\ 0 & N - n_j < k \leq N \end{cases} \tag{31}$$

Direct, Non-Uniform, Distinct (4)

If we compute the p_j after we have sorted the n_j in ascending order, we can use the fact that

$$p_{j+1} = p_j * \prod_{i=n_j}^{n_{j+1}-1} \frac{N-k-i}{N-i}.$$

Direct, Non-Uniform, Distinct (5)

Many buckets: statistics too big. Better: Histograms

Corollary

For $1 \leq i \leq L$ let there be l_i buckets containing n_i items. Then, the total number of buckets is $m = \sum_{i=1}^{L} l_i$ and the total number of items in all buckets is $N = \sum_{i=1}^{L} l_i n_i$. For k randomly selected items the number of qualifying buckets is

$$\overline{\mathcal{W}}_{n_j}^{N,m}(k) = \sum_{i=1}^{L} l_i \mathcal{Y}_{n_j}^{N}(k) \tag{32}$$

Direct, Non-Uniform, Distinct (6)

Distribution function. The probability that $x \leq n_j$ items in a bucket j qualify, can be calculated as follows:

- The number of possibilities to select x items in bucket n_j is

$$\binom{n_j}{x}$$

- The number of possibilites to draw the remaining $k - x$ items from the other buckets is

$$\binom{N - n_j}{k - x}$$

- The total number of possibilities to distributed k items over the buckets is

$$\binom{N}{k}$$

This shows the following:

Direct, Non-Uniform, Distinct (7)

Theorem

Assume a set of m buckets. Each bucket contains $n_j > 0$ items $(1 \leq j \leq m)$. The total number of items is $N = \sum_{j=1}^{m} n_j$. If we lookup k distinct items, then the probability that x items in bucket j qualify is

$$\mathcal{X}_{n_j}^{N}(k,x) = \frac{\binom{n_j}{x}\binom{N-n_j}{k-x}}{\binom{N}{k}} \tag{33}$$

Further, the expected number of qualifying items in bucket j is

$$\overline{\mathcal{X}}_{n_j}^{N,m}(k) = \sum_{x=0}^{\min(k,n_j)} x\mathcal{X}_{n_j}^{N}(k,x) \tag{34}$$

In standard statistics books the probability distribution $\mathcal{X}_{n_j}^{N}(k,x)$ is called *hypergeometric distribution*.

Direct, Non-Uniform, Distinct (8)

Let us consider the case where all n_j are equal to n. Then, we can calculate the average number of qualifying items in a bucket. With $y := \min(k, n)$ we have

$$\begin{aligned}
\overline{\mathcal{X}}_{n_j}^{N,m}(k) &= \sum_{x=0}^{\min(k,n)} x\mathcal{X}_n^N(k,x) \\
&= \sum_{x=1}^{\min(k,n)} x\mathcal{X}_n^N(k,x) \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} x \binom{n}{x} \binom{N-n}{k-x}
\end{aligned}$$

Direct, Non-Uniform, Distinct (9)

$$
\begin{aligned}
\overline{\mathcal{X}}_{n_j}^{N,m}(k) &= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} x \binom{n}{x} \binom{N-n}{k-x} \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} \binom{x}{1} \binom{n}{x} \binom{N-n}{k-x} \\
&= \frac{1}{\binom{N}{k}} \sum_{x=1}^{y} \binom{n}{1} \binom{n-1}{x-1} \binom{N-n}{k-x} \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \sum_{x=0}^{y-1} \binom{n-1}{0+x} \binom{N-n}{(k-1)-x} \\
&= \dots
\end{aligned}
$$

(cont.)

Direct, Non-Uniform, Distinct (10)

$$
\begin{aligned}
\overline{\mathcal{X}}_{n_j}^{N,m}(k) &= \dots \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \binom{n-1+N-n}{0+k-1} \\
&= \frac{\binom{n}{1}}{\binom{N}{k}} \binom{N-1}{k-1} \\
&= n\frac{k}{N} = \frac{k}{m}
\end{aligned}
$$

Direct, Non-Uniform, Distinct (11)

Let us consider the even more special case where every bucket contains a single item. That is, $N = m$ and $n_i = 1$. The probability that a bucket contains a qualifying item reduces to

$$\begin{aligned} \mathcal{X}_1^N(k,x) &= \frac{\binom{1}{x}\ \binom{N-1}{k-1}}{\binom{N}{k}} \\ &= \frac{\binom{N-1}{k-1}}{\binom{N}{k}} \\ &= \frac{k}{N}\ \ (= \frac{k}{m}) \end{aligned}$$

Since x can then only be zero or one, the average number of qualifying items a bucket contains is also $\frac{k}{N}$.