Distributed Storage Networks and Computer Forensics
8 Analysis of DHT

Christian Schindelhauer

University of Freiburg
Technical Faculty
Computer Networks and Telematics
Winter Semester 2011/12
Distributed Storage Networks
and Computer Forensics
Winter 2011/12

Distributed Hash-Table (DHT)

- Hash table
  - does not work efficiently for inserting and deleting
- Distributed Hash-Table
  - servers are „hashed“ to a position in an continuous set (e.g. line)
  - data is also „hashed“ to this set
- Mapping of data to servers
  - servers are given their own areas depending on the position of the direct neighbors
  - all data in this area is mapped to the corresponding server
- Literature

Pure (Poor) Hashing

DHT
Entering and Leaving a DHT

- **Distributed Hash Table**
  - devices are hashed to a position
  - blocks are hashed according to the ID

- **When a device is added**
  - only blocks from neighbors have to be moved

- **When a device is deleted**
  - blocks are moved only to the neighbors
Holes and Dense Areas
Size of Holes

› **Theorem**
  • If \( n \) elements are randomly inserted into an array \([0,1]\) then with constant probability there is a „hole“ of size \( \Omega(\log n/n) \), i.e. an interval without elements.

› **Proof**
  • Consider an interval of size \( \log n / (4n) \)
  • The chance not to hit such an interval is \( (1-\log n/(4n)) \)
  • The chance that \( n \) elements do not hit this interval is 
    
    \[
    \left(1 - \frac{\log n}{4n}\right)^n = \left(1 - \frac{\log n}{4n}\right)^{\frac{4n}{\log n} \cdot \frac{\log n}{4}} \geq \left(\frac{1}{4}\right)^{\frac{1}{4} \log n} = \frac{1}{\sqrt{n}}
    \]
  • The expected number of such intervals is more than 1.
  • Hence the probability for such an interval is at least constant.
Proof of Dense Areas

\[
\left(\frac{1}{4}\right) \frac{1}{4} \cdot \log n = 2 \left(\frac{1}{4} \cdot \log n\right) \left(\log \frac{1}{4}\right)
\]

\[
= 2 \left(-\frac{1}{3}\right) \cdot \log n
\]

\[
= \frac{2}{n^{\frac{1}{2}}}
\]

\[
\text{Expectation: } \frac{4n}{\log n} \cdot \frac{1}{\sqrt{n}} = \frac{4\sqrt{n}}{\log n}
\]
Dense Spots

- **Theorem**
  - If n elements are randomly inserted into an array [0,1] then with constant probability there is a dense interval of length 1/n with at least $\Omega\left(\log n/ (\log \log n)\right)$ elements.

- **Proof**
  - The probability to place exactly i elements in to such an interval is
  $$\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{n-i} \binom{n}{i}$$
  - for $i = c \log n / (\log \log n)$ this probability is at least $1/n^k$ for an appropriately chosen c and k<1
  - Then the expected number of intervals is at least 1
Proof of Dense Areas

\[ i = \frac{c \cdot \log n}{\log \log n} \]

Proof: \( n \) balls from \( n \) balls fall into an interval of size \( \frac{1}{n} \)

\[ \left( \frac{1}{n} \right)^i \left( 1 - \frac{1}{n} \right)^{n-i} \geq \frac{1}{n^\theta} \]

\( \theta \approx 0.63 \)
Proof of Dense Areas

\[ \frac{\Lambda}{q} \leq \left( 1 - \frac{\Lambda}{m} \right)^m \leq \frac{\Lambda}{e} \]

\[ \left( 1 - \frac{\Lambda}{m} \right)^n = \left( 1 - \frac{\Lambda}{n} \right) \]

\[ \geq \left( \frac{\Lambda}{q} \right)^{1 - \frac{1}{n}} \]

\[ \geq \frac{1}{q} \]
Proof of Dense Areas

\[
\binom{m}{i} = \frac{m!}{i!(m-i)!} = \frac{m \cdot (m-1) \cdot (m-2) \cdots (m-i+1)}{i!}
\]

\[
\frac{m}{n} \cdot \frac{m-1}{n} \cdot \frac{m-2}{n} \cdots \frac{m-i+1}{n} \geq \left(\frac{1}{4}\right)^i \left(1 - \frac{i}{n}\right)^{(i-1)}
\]

\[
\left(1 - \frac{i}{n}\right)^{\frac{m}{n}} \cdot \frac{m}{n} \cdot \frac{m-1}{n} \cdot \cdots \frac{m-i+1}{n} \geq \left(\frac{1}{4}\right)^i \left(1 - \frac{i}{n}\right)^{(i-1)} \geq \left(\frac{1}{4}\right)^i \left(\frac{1}{2}\right)^i = \left(\frac{1}{2}\right)^i
\]
Proof of Dense Areas

\[
\left(\frac{A}{2}\right)^{\frac{1}{T}} = 2^{\frac{A}{2}} \geq 2^{\frac{1}{\ln n}} \frac{\log^2 \ln n}{\log \ln n} \frac{1 + \ln n - \ln \log \ln n}{1 + \ln n} \\
\leq \frac{c \cdot \log n}{\log \log n} \left(1 + \ln c + \ln 2\right) \log \log n \\
= c \left(1 + \ln c + \ln 2\right) \log n
\]
Averaging Effect

Theorem

If \( \Theta(n \log n) \) elements are randomly inserted into an array \([0,1]\) then with high probability in every interval of length \(1/n\) there are \(\Theta(\log n)\) elements.
Excursion

- **Markov-Inequality**
  - For random variable $X > 0$ with $\mathbb{E}[X] > 0$:
    $$P[X \geq k \cdot \mathbb{E}[X]] \leq \frac{1}{k}$$

- **Chebyshev**
  - $P[|X - \mathbb{E}[X]| \geq k] \leq \frac{\mathbb{V}[X]}{k^2}$
    - for Variance $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

- **Stronger bound: Chernoff**
Theorem Chernoff Bound

- Let $x_1, \ldots, x_n$ independent Bernoulli experiments with
  - $P[x_i = 1] = p$
  - $P[x_i = 0] = 1-p$
- Let $S_n = \sum_{i=1}^{n} x_i$
- Then for all $c>0$
  $$P[S_n \geq (1 + c) \cdot E[S_n]] \leq e^{-\frac{1}{3} \min\{c, c^2\}pn}$$
- For $0 \leq c \leq 1$
  $$P[S_n \leq (1 - c) \cdot E[S_n]] \leq e^{-\frac{1}{2} c^2 pn}$$
Lemma

If \( m = k \cdot n \ln n \) Balls are randomly placed in \( n \) bins:

1. Then for all \( c > k \) the probability that more than \( c \ln n \) balls are in a bin is at most \( O(n^{-c'}) \) for a constant \( c' > 0 \).

2. Then for all \( c < k \) the probability that less than \( c \ln n \) balls are in a bin is at most \( O(n^{-c'}) \) for a constant \( c' > 0 \).

Proof:

Consider a bin and the Bernoulli experiment \( B(k \cdot n \ln n, 1/n) \) and expectation: \( \mu = m/n = k \ln n \)

1. Case: \( c > 2k \)
   \[
   P[X \geq c \ln n] = P[X \geq (1+(c/k-1))k \ln n] 
   \leq e^{-\frac{1}{3}(c/k-1)k \ln n} 
   \leq n^{-\frac{1}{3}(c-k)}
   \]

2. Case: \( k < c < 2k \)
   \[
   P[X \geq c \ln n] = P[X \geq (1+(c/k-1))k \ln n] 
   \leq e^{-\frac{1}{3}(c/k-1)^2k \ln n} 
   \leq n^{-\frac{1}{3}(c-k)^2}.
   \]

3. Case: \( c < k \)
   \[
   P[X \leq c \ln n] = P[X \leq (1-(1-c/k))k \ln n] 
   \leq e^{-\frac{1}{2}(1-c/k)^2k \ln n} 
   \leq n^{-\frac{1}{2}(k-c)^2/k}.
   \]
Concept of High Probability

Lemma
If A(i) holds with high probability, i.e. $1-n^{-c}$, then
(A(1) and A(2) and ... and A(n)) with high probability,
i.e. $1-n^{-(c-1)}$

Proof:
- For all i: $P[\neg A(i)] \leq n^{-c}$
- Hence: $P[\neg A(1) \lor \neg A(2) \lor ... \lor \neg A(n)] \leq n \cdot n^{-c}$
  $P[\neg (\neg A(1) \lor \neg A(2) \lor ... \lor \neg A(n))] \leq 1 - n \cdot n^{-c}$

DeMorgan:
$P[A(1) \land A(2) \land ... \land A(n)] \leq 1 - n \cdot n^{-c}$
Principle of Multiple Choice

- Before inserted check $c \log n$ positions
- For position $p(j)$ check the distance $a(j)$ between potential left and right neighbor
- Insert element at position $p(j)$ in the middle between left and right neighbor, where $a(j)$ was the maximum choice
- Lemma
  - After inserting $n$ elements with high probability only intervals of size $1/(2n)$, $1/n$ und $2/n$ occur.
Proof of Lemma

1. Part: With high probability there is no interval of size larger than $2/n$

follows from this Lemma

Lemma*

Let $c/n$ be the largest interval. After inserting $2n/c$ peers all intervals are smaller than $c/(2n)$ with high probability

From applying this lemma for $c=n/2, n/4, \ldots, 4$ the first lemma follows.
Proof

- 2nd part: No intervals smaller than $1/(2n)$ occur
  - The overall length of intervals of size $1/(2n)$ before inserting is at most $1/2$
  - Such an area is hit with probability at most $1/2$
  - The probability to hit this area more than $c \log n$ times is at least
    \[ 2^{-c \log n} = n^{-c} \]
  - Then for $c > 1$ such an interval will not further be divided with probability into an interval of size $1/(4m)$. 
Algorithms and Methods for Distributed Storage Networks
9 Analysis of DHT

Christian Schindelhauer

University of Freiburg
Technical Faculty
Computer Networks and Telematics
Winter Semester 2011/12