Yhteentörmäysten generoiminen Rolling Hashille modulo 2^{64}.

Olkoon meillä merkkijono s, merkitään s_i sen i:nnen kirjaimen ASCII-arvoa. Merkkijonon s hajautusarvo H_s tekniikkaa "Rolling Hash" käyttäen, lasketaan kaavalla

$$H_s = \sum_{i=0}^{n_s-1} A^i s_i \bmod N,$$

missä A on sopiva ennalta valittu kokonaisluku ja n_s on merkkijonon s pituus. **Huomaathan, että potenssit menevät "eri suuntaan" tässä kuin miten yleensä rolling hashissa.** Tämän vaikutuksen voi kumota kääntämällä merkkijonon toisin päin.

Tässä keskitymme tapaukseen, jossa $N = 2^{64}$. Tällainen tilanne tulee vastaan muun muassa silloin, jos moduloa ei oteta ollenkaan laskiessa hajautusarvoa 64-bittiseen etumerkittömään kokonaislukuun, vaan välituloksen annetaan ylivuotaa.

Muistutuksena vielä mainittakoon, että $x \bmod N = y \bmod N$ täsmälleen siinä tapauksessa, että N jakaa luvun $x - y$. Notaation lyhentämiseksi merkitään $[x]$ luvun x jakojäännöstä luvulla $N = 2^{64}$. On tunnettua, että tällöin $[[x][y]] = [xy]$ ja $[[x] + [y]] = [x + y]$.

Oletetaan ensin, että A on parillinen, $A = 2B$. Jos merkkijonossa s on vähintään 64 kirjainta, saadaan

$$H_s = [\sum_{i=0}^{n_s-1} (2B)^i s_i] = \sum_{i=0}^{n_s-1} [2^i][B^i][s_i] = \sum_{i=0}^{63} [2^i][B^i][s_i],$$

sillä $[2^{64}] = 0$. Tästä nähdään, että kahdella vähintään 64-merkkisellä merkkijonolla, joilla on sama 64 merkkiä pitkä alkuosa, on sama hajautusarvo. Täten luvun A valitseminen parilliseksi mahdollistaan yhteentörmäysten erittäin helpon generoimisen.

Tarkastellaan sitten tapausta, jossa A on pariton. Tarkastellaan merkkijonoa s, jonka pituus n_s on $2048 = 2^{11} - 1$ merkkiä. Merkitään $cnt(i)$:llä numeron i bittiesityksessä olevien ykkösbittien määrää. Merkkijonon s i:nnes kirjain on 'a' mikäli $cnt(i)$ on pariton, muuten i:nnes kirjain on 'b'. Olkoon s' merkkijono, joka on saatu merkkijonosta s vaihtamalla kirjaimet 'a' ja 'b' keskenään. Nyt $H_s - H_{s'} = 0$, kuten näemme seuraavaksi.

Koska $H_s - H_{s'} = [\sum_{i=0}^{n_s-1} A^i(s_i - s'_i)]$, riittää osoittaa, että $\sum_{i=0}^{n_s-1} A^i(s_i - s'_i)$ on jaollinen luvulla 2^{64}. Koska kirjainten 'a' ja 'b' ASCII-arvot ovat peräkkäisiä kokonaislukuja, on tämä sama asia kuin $\sum_{i=0}^{n_s-1} A^i(-1)^{cnt(i)}$. Tämä johtuu siitä, että merkkijonossa s on kirjain 'a' tasan niissä indekseissa i, joille $cnt(i)$ on pariton, jolloin $s_i = s'_i - 1$. Toisaalta jos $cnt(i)$ on parillinen, on merkkijonossa s indeksissä i kirjain 'b' ja $s_i = s'_i + 1$.

Seuraavaksi tärkeä havainto, joka perustellaan myöhemmin:

$$(1 - A^{2^0})(1 - A^{2^1})(1 - A^{2^2})...(1 - A^{2^{10}}) = \sum_{i=0}^{n_s-1} A^i(-1)^{cnt(i)}. \quad (1)$$

Koska A on pariton, on $1 - A$ jaollinen kahdella. Toisaalta, koska $1 - A^{2^i} = (1 - A^{2^{i-1}})(1 + A^{2^{i-1}})$, missä $1 + A^{2^{i-1}}$ on parillinen, nähdään induktiolla, että $1 - A^{2^i}$ on jaollinen luvulla 2 vähintään $i+1$ kertaa, eli 2^{i+1} jakaa luvun $1 - A^{2^i}$. Nyt siis $(1 - A^{2^0})(1 - A^{2^1})(1 - A^{2^2})...(1 - A^{2^{10}})$ on jaollinen luvulla 2 vähintään $1+2+3+...+11 = 66$ kertaa, joten $\sum_{i=0}^{n_s-1} A^i(s_i - s'_i)$ on jaollinen luvulla 2^{64}. Tämä tarkoittaa, että $H_s - H_{s'} = 0$, eli merkkijonot s ja s' saavat saman hajautusarvon.

Generoidut yhteentörmäykset riippuivat ainoastaan luvun A parillisuudesta, eikä ollenkaan sen tarkemmasta arvosta. Tämä tekee yhteentörmäysten generoinnista erityisen helppoa. Tarinan opetus: älä käytä Rolling Hash modulo kakkosen potenssi!

Kaavan (1) perustelu. Todistus perustuu induktioon. Väite:

$$(1 - A^{2^0})(1 - A^{2^1})...(1 - A^{2^j}) = \sum_{i=0}^{2^{j+1}-1} A^i(-1)^{cnt(i)}.$$

Kaava (1) saadaan tästä kun $j = 10$. Tilanne $j = 0$ on selvä. Oletetaan, että väite on selvää kun $j = 0..n$. Todistetaan, että väite pätee myös kun $j = n + 1$. Nyt

$$\begin{aligned}(1 - A^{2^0})...(1 - A^{2^n})(1 - A^{2^{n+1}}) &= (\sum_{i=0}^{2^{n+1}-1} A^i(-1)^{cnt(i)})(1 - A^{2^{n+1}}) \\ &= \sum_{i=0}^{2^{n+1}-1} A^i(-1)^{cnt(i)} - A^{2^{n+1}} \sum_{i=0}^{2^{n+1}-1} A^i(-1)^{cnt(i)} \\ &= \sum_{i=0}^{2^{n+1}-1} A^i(-1)^{cnt(i)} + \sum_{i=0}^{2^{n+1}-1} A^{2^{n+1}+i}(-1)^{cnt(i)+1}.\end{aligned}$$

Koska $i < 2^{n+1}$, on luvun $2^{n+1} + i$ binääriesityksessä täsmälleen yksi ykkösbitti enemmän, kuin luvun i binääriesityksessä. Täten $cnt(2^{n+1} + i) = cnt(i) + 1$. Kun i käy läpi luvut $0..2^{n+1} - 1$, käy $2^{n+1} + i$ läpi kaikki luvut $2^{n+1}..2^{n+1} + 2^{n+1} - 1 = 2^{n+2} - 1$. Täten

$$\sum_{i=0}^{2^{n+1}-1} A^{2^{n+1}+i}(-1)^{cnt(i)+1} = \sum_{i=2^{n+1}}^{2^{n+2}-1} A^i(-1)^{cnt(i)}$$

joten

$$\begin{aligned}(1 - A^{2^0})...(1 - A^{2^n})(1 - A^{2^{n+1}}) &= \sum_{i=0}^{2^{n+1}-1} A^i(-1)^{cnt(i)} + \sum_{i=2^{n+1}}^{2^{n+2}-1} A^i(-1)^{cnt(i)} \\ &= \sum_{i=0}^{2^{n+2}-1} A^i(-1)^{cnt(i)},\end{aligned}$$

mikä pitikin osoittaa.

Yleistys modulo p^n.

Samaa ideaa käyttäen voidaan myös generoida yhteentörmäyksiä modulo p^n, missä p on (pieni) alkuluku ja n on mielivaltainen luonnollinen luku. Tapaus, jossa A on jaollinen luvulla p menee täsmälleen samalla tavalla kuin aiemman tarkastelun tapaus, jossa A oli parillinen.

Jos A ei ole jaollinen luvulla p, niin Fermat'n pienen lauseen mukaan $A^{p-1} = 1 \bmod p$. Jos asetamme $m = p - 1$, niin tämä tarkoittaa sitä, että $A^m = 1 + ap$, missä a on jokin kokonaisluku.

Olkoon $C = A^m = 1 + ap$. Nyt p^{i+1} jakaa luvun $C^{p^i} - 1$: tapaus $i = 0$ on triviaali, sillä $C - 1 = ap$. Toisaalta $C^{p^{i+1}} - 1 = (C^{p^i} - 1)(C^{p^{i+1}-p^i} + C^{p^{i+1}-2p^i} + ... + 1)$. Induktion nojalla tulon vasemmanpuoleinen alkio on jaollinen luvulla p^{i+1}, joten väitteen todistamiseksi riittää osoittaa, että oikeanpuoleinen alkio on jaollinen luvulla p. Koska $C = 1 \bmod p$, saadaan

$$C^{p^{i+1}-p^i} + C^{p^{i+1}-2p^i} + ... + 1 = 1^{p^{i+1}-p^i} + 1^{p^{i+1}-2p^i} + ... + 1 \bmod p.$$

Koska tässä summassa on tasan p termiä, on oikealla puolella oleva luku $0 \bmod p$, joten $C^{p^{i+1}-p^i} + C^{p^{i+1}-2p^i} + ... + 1 = 0 \bmod p$. Täten p^{i+1} jakaa luvun $C^{p^i} - 1$ kaikilla luonnollisilla luvuilla i.

Tarkastellaan seuraavaksi tuloa $(1 - C)(1 - C^{p^1})...(1 - C^{p^l})$. Samankaltaisesti, kuin miten kaava (1) perusteltiin, voidaan osoittaa, että

$$(1 - C)(1 - C^{p^1})...(1 - C^{p^l}) = \sum_i C^i (-1)^{cnt_p(i)},$$

missä $cnt_p(i)$ on luvun i esityksessä p-lukujärjestelmässä esiintyvien numeroiden summa ja i kulkee yli kaikkien sellaisten kokonaislukujen, joiden esitys p-lukujärjestelmässä on korkeintaan $l + 1$ merkkiä pitkä ja joiden esityksessä esiintyy ainoastaan numeroita 0 ja 1. Nyt $p^{(1+2+...+l+l+1)} = p^{(l+1)(l+2)/2}$ jakaa luvun $(1-C)^{p-1}(1-C^{p^1})^{p-1}...(1-C^{p^l})^{p-1}$, joten tarpeeksi isolla l tiedämme, että p^n jakaa luvun $\sum_{i=0}^{p^{l+1}-1} C^i(-1)^{cnt_p(i)}$.

Enää tarvitsee keksiä samanpituiset merkkijonot $s \neq s'$, joilla

$$\sum_{j=0}^{n_s-1} A^j(s_j - s'_j) = \sum_{i=0}^{p^{l+1}-1} C^i(-1)^{cnt_p(i)}.$$

Se onnistuu seuraavasti: merkkijonojen s ja s' pituus on mp^{l+1}. Jos j ei ole jaollinen luvulla m, niin asetetaan kumpaankin merkkijonoon indeksiin j kirjain 'a'. Toisaalta jos $j = mi$, asetetaan asetetaan indeksiin $j = mi$ toisessa merkkijonossa 'a' ja toisessa 'b', siten että $s_j - s'_j = s_{mi} - s'_{mi} = (-1)^{cnt_p(i)}$. Nyt

$$\begin{aligned}\sum_{j=0}^{mp^{l+1}-1} A^j(s_j - s'_j) &= \sum_{i=0}^{p^{l+1}-1} A^{mi}(s_{mi} - s'_{mi}) \\ &= \sum_{i=0}^{p^{l+1}-1} C^i(-1)^{cnt_p(i)}.\end{aligned}$$

Täten tarpeeksi isolla l saamme yhteentörmäyksen.

Viitteet

[1] http://codeforces.com/blog/entry/4898