Full Proof

Here are proofs about space complexity to provide ϵN error guarantee with $1-\delta$ probability.

The target is to guarantee $\Pr[|ERR| > \epsilon N] \le \delta$

With Chernoff bound: $\Pr[|ERR| > t] \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$, for sub-Gaussian ERR with a variance of σ^2.

Disk-Bounded Chunk Sketches

The total variance is $\sigma^2 = \frac{N}{N_c}\sum_{h=1}^{h=H_c-1} 4^{h-1}\cdot 1 \le \frac{N}{3N_c}4^{H_c-1}$

To achieve $2\exp\left(-\frac{\epsilon^2 N^2}{2\sigma^2}\right) \le \delta$, there should be $-\frac{\epsilon^2 N^2}{2\sigma^2} \le \log\frac{\delta}{2}$. Thus $-\epsilon^2 N^2 \le 2\sigma^2 \log\frac{\delta}{2} \Rightarrow \epsilon^2 N^2 \ge 2\sigma^2 \log\frac{2}{\delta} \Rightarrow \sigma^2 \le \epsilon^2 N^2 / \log\frac{2}{\delta}/2$

When H_c is small enough, there will be $\frac{N}{3N_c}4^{H_c-1} \le \epsilon^2 N^2/\log\frac{2}{\delta}/2$ satisfying above. Thus $4^{H_c} \le \epsilon^2 6 N_c N \log\frac{\delta}{2}$, i.e., $2^{H_c} \le \sqrt{6N_c N}\epsilon\sqrt{\log\frac{\delta}{2}}$ can provide the required error guarantee.

Thus, the required chunk sketch size to guarantee $\Pr[|ERR| > \epsilon N] \le \delta$ is

$$\frac{N_c}{2^{H_c}} = \sqrt{6N_c N}\cdot\epsilon\sqrt{\log\frac{\delta}{2}} \tag{1}$$

The I/O cost, i.e., the total size of chunk sketches, is

$$\frac{N}{2^{H_c}} = \sqrt{\frac{N}{6N_c}}\cdot\frac{1}{\epsilon}\sqrt{\log\frac{2}{\delta}}$$

Disk-Bounded SSTable Sketches

When $\sqrt{T} < T_s < T$, the variance of error of SSTable sketches is at most $\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right)$ times of concatenated chunk sketches.

Now the target is to satisfy $\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \frac{N}{3C} 4^{H_c - 1} \leq \epsilon^2 N^2 / \log \frac{2}{\delta} / 2$

$$\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot 4^{H_c} \leq \epsilon^2 6 N_c N \log \frac{\delta}{2}$$

At level L, the size of top sketch is $\frac{N_c}{2^{H_c}} \cdot T_S^L$. Then the size of largest top sketch is $O\left(N^{\log_T T_s} / 2^{H_c}\right)$.

Let $T_s = T^{\frac{1}{2}(1+b)}, 0 < b < 1$. Then $T/T_S^2 = T^{-b}$, $\log_T T_s = \frac{1}{2}(1+b)$, $\frac{4^{H_c}}{(1 - T/T_S^2)T} = \frac{4^{H_c}}{(1 - T^{-b})T}$

$4^{H_c} \cdot 4^{H_c} = 2^{4H_c} = O\left(\epsilon^2 N_c N T (1 - T^{-b}) \log \frac{\delta}{2}\right)$

$2^{4H_c} = O\left(\epsilon^2 N_c N \log \frac{\delta}{2}\right)$

The space complexity is $O\left(\frac{N^{\log_T T_s}}{2^{H_c}}\right)$

$$= O\left((N_c T)^{-\frac{1}{4}} \cdot N^{\left(\frac{1}{4} + \frac{1}{2} b\right)} \cdot \left(\frac{(T^b - 1)}{T^b}\right)^{-\frac{1}{4}} \cdot \epsilon^{-\frac{1}{2}} \cdot \left(\log \frac{1}{\delta}\right)^{\frac{1}{4}} \right)$$

$$= O\left((N_c T)^{-\frac{1}{4}} \cdot N^{\left(\frac{1}{4} + \frac{1}{2} b\right)} \cdot \epsilon^{-\frac{1}{2}} \cdot \left(\log \frac{1}{\delta}\right)^{\frac{1}{4}} \right)$$

Memory-Constrained Chunk Sketches

To make the merge-and-compressed chunk sketches more accurate than streaming KLL, in other words, $\sum_{i=1}^{H_c-1} 2^{2(i-1)}\frac{n}{n_c} \leq \sum_{i=1}^{H_c-1} m_i\omega_i^2$.

That requires $K\gamma^H \leq \frac{3}{2}\frac{(2\gamma)^{H_c}-2\gamma}{(2\gamma-1)(4^{H_c-1}-1)}n_c$

$$K\gamma^H \leq \Gamma\frac{\gamma^{H_c}}{2^{H_c}}n_c$$

For any top capacity K in the streaming KLL,

$$\gamma^H\left(\frac{2}{\gamma}\right)^{H_c} \leq \frac{\Gamma n_c}{K}$$

$$\left(\frac{2}{\gamma}\right)^{H_c} \leq \frac{\Gamma n_c}{\gamma^H K}$$

Note that $(AB)^C = (A^C)^{\log_A AB}$

$$\left(\frac{\mathbf{2}}{\boldsymbol{\gamma}}\right)^{\boldsymbol{H_c}} = (\mathbf{2}^{\boldsymbol{H_c}})^{\mathbf{log_2\,2/\boldsymbol{\gamma}}} \leq \frac{\Gamma n_c}{\gamma^H K}, \qquad \log_2 2/\gamma \in (1,2)$$

I/O cost is

$$\frac{n}{2^{H_c}} = O\left(\frac{n}{\left(\frac{\Gamma n_c}{\gamma^H K}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{n}{\left(\frac{\Gamma n_c}{\gamma^H K}\right)^{1/\log_2 \frac{2}{\gamma}}}\right) = O\left(\frac{N}{\left(N_c\left(\frac{1}{\gamma}\right)^{\log N}\right)^{1/\log \frac{2}{\gamma}}}\right)$$

$$= O\left(\frac{N}{\left(N_c (N)^{\log \frac{1}{\gamma}}\right)^{1/\log \frac{2}{\gamma}}}\right) = O\left(\frac{N^{1-\frac{\log \frac{1}{\gamma}}{\log \frac{2}{\gamma}}}}{N_c{}^{1/\log \frac{2}{\gamma}}}\right) = O\left(\frac{N^{1-\frac{\log \frac{1}{\gamma}}{1+\log \frac{1}{\gamma}}}}{N_c{}^{1/\log \frac{2}{\gamma}}}\right)$$

$$= O\left(\frac{N^{\frac{1}{1+\log \frac{1}{\gamma}}}}{N_c{}^{1/\left(1+\log \frac{1}{\gamma}\right)}}\right) = O\left(\left(\frac{N}{N_c}\right)^{1/\left(1+\log \frac{1}{\gamma}\right)}\right)$$

When $\gamma = 2/3$, that is $O\left(\left(\frac{N}{N_c}\right)^{0.631}\right)$

Memory-Constrained SSTable Sketches

The height of the top sketch in level L SSTable is $H_c + L \cdot \log \frac{T}{T_s}$

The target is to make the merge-and-compacted SSTable sketches more accurate than streaming KLL, invoke the lemmas bounding SSTable sketch error with Chunk sketch error and we have:

$$\left(1 + \frac{4^{H_c}}{T - \left(\frac{T}{T_s}\right)^2}\right) \cdot \sum_{i=1}^{H_c - 1} 2^{2(i-1)} \frac{N}{N_c} \leq \sum_{i=1}^{H_c - 1 + L \cdot \log \frac{T}{T_s}} m_i' \omega_i^2$$

Let $T_s = T^{\frac{1}{2}(1+b)}, 0 < b < 1$. Then $T/T_s^2 = T^{-b}$, $\log_T T_s = \frac{1}{2}(1+b)$

The target is to satisfy

$$\left(1+\frac{4^{H_c}}{T-\left(\frac{T}{T_s}\right)^2}\right)\cdot\frac{K\gamma^H}{2^{2*\frac{L}{2}(1+b)}}\le\Gamma\frac{\gamma^{H_c+\frac{L}{2}(1+b)}}{2^{H_c+\frac{L}{2}(1+b)}}n_c$$

$$\left(1+\frac{4^{H_c}}{T-\left(\frac{T}{T_s}\right)^2}\right)\cdot\frac{K\gamma^H}{2^{\frac{L}{2}(1+b)}}\le\Gamma\frac{\gamma^{H_c+\frac{L}{2}(1+b)}}{2^{H_c}}n_c$$

$$\left(1+\frac{4^{H_c}}{T-\left(\frac{T}{T_s}\right)^2}\right)\cdot K\gamma^H\le\Gamma\left(\frac{\gamma}{2}\right)^{H_c}(2\gamma)^{\frac{L}{2}(1+b)}n_c$$

Recall that $\frac{4^{H_c}}{(1-T/T_s^2)T}=\frac{4^{H_c}}{(1-T^{-b})T}$

$$\frac{4^{H_c}}{(1-T^{-b})T}\cdot K\gamma^H\le\Gamma\left(\frac{\gamma}{2}\right)^{H_c}(2\gamma)^{\frac{L}{2}(1+b)}n_c$$

Again, the I/O cost is $O\left(N^{\log_T T_s}/2^{H_c}\right)$. However, L is determined by N and we need further analyze.

$n_c\cdot T^L=n;\ \ 2^L=(T*2/T)^L=(T^L)^{\log_T 2}$

$(2\gamma)^{\frac{L}{2}(1+b)}=\left((T^L)^{\log_T(2\gamma)}\right)^{\frac{1}{2}(1+b)}=\left(\frac{n}{n_c}\right)^{\frac{1}{2}(1+b)\log_T(2\gamma)}$

Now the target is to satisfy

$$\left(\frac{8}{\gamma}\right)^{H_c}\le\frac{\Gamma(1-T^{-b})T(2\gamma)^{\frac{L}{2}(1+b)}n_c}{K\gamma^H}$$

$$\left(\frac{8}{\gamma}\right)^{H_c}=(2^{H_c})^{\log_2 8/\gamma}\le\frac{(1-T^{-b})T\left(\frac{n}{n_c}\right)^{\frac{1}{2}(1+b)\log_T(2\gamma)}n_c}{K\gamma^H}$$

$$2^{H_c}\le\left(\frac{(1-T^{-b})Tn_c}{K\gamma^H}\right)^{\frac{1}{\log_2\frac{8}{\gamma}}}\left(\frac{n}{n_c}\right)^{\frac{1(1+b)}{2\log_2\frac{8}{\gamma}}\log_T(2\gamma)}$$

$$=\left(\frac{(1-T^{-b})Tn_c}{K\gamma^H}\right)^{1/\log_2\frac{8}{\gamma}}\left(\frac{n}{n_c}\right)^{\frac{(1+b)}{2}\cdot\frac{\log_T(2\gamma)}{3+\log_2\frac{1}{\gamma}}}$$

Now Hc is bounded and the I/O cost is

$$O\left(N^{\log_T T_s}/2^{H_c}\right) = O\left(\frac{N^{(1+b)/2}}{2^{H_c}}\right)$$

$$= O\left(\frac{N^{\frac{(1+b)}{2}}}{\left((1-T^{-b})Tn_c(N)^{\log\frac{1}{\gamma}}\right)^{\frac{1}{\log_2\frac{8}{\gamma}}}\left(\frac{n}{n_c}\right)^{\frac{1+b}{2\log T}\cdot\frac{\log 2\gamma}{\log\frac{8}{\gamma}}}}\right)$$

$$= O\left(\frac{N^{\frac{(1+b)}{2}-\frac{(1+b)}{2}\cdot\frac{1}{\log T}\cdot\frac{\log 2\gamma}{\log\frac{8}{\gamma}}-\log\frac{1}{\gamma}\cdot\frac{1}{\log\frac{8}{\gamma}}}}{(N_c)^{\frac{1}{\log\frac{8}{\gamma}}\left(1-\frac{1+b}{2\log T}\cdot\log 2\gamma\right)}}\right)$$

$$= O\left(\frac{N^{\frac{(1+b)}{2}-\frac{(1+b)}{2}\cdot\frac{1}{\log T}\cdot\frac{\log 2\gamma}{\log\frac{8}{\gamma}}-\log\frac{1}{\gamma}\cdot\frac{1}{\log\frac{8}{\gamma}}}}{(N_c)^{\frac{1}{\log\frac{8}{\gamma}}\left(1-\frac{1+b}{2\log T}\cdot\log 2\gamma\right)}}\right)$$

When $\gamma = \frac{2}{3}$, T=10, Ts=5, b=0.3979, it is $O\left(\frac{N^{0.5114}}{N_c^{0.32}}\right)$